SSP'2001

Sponsored by



DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

20011023 029

Proceedings of the

11TH IEEE SIGNAL PROCESSING WORKSHOP

on

STATISTICAL SIGNAL PROCESSING

6th – 8th August 2001, Orchid Country Club, Singapore

Sponsored by

The IEEE Signal Processing Society

With Support from



US AIRFORCE Research Laboratory, **USA**



US AIRFORCE Office for Scientific Research, **USA**



US AIRFORCE Asian Office of Aerospace R&D, USA



Defence Science & Technology Agency
SINGAPORE



Defence Science Organisation National Laboratories
SINGAPORE

REPORT DOCUMENTATION PAGE

Form Approved

OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
17-09-2001	Conference Proceedings	

4. TITLE AND SUBTITLE

11th IEEE Workshop on Statistical Processing (SSP2001)

5a. CONTRACT NUMBER

F6256201M9118

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Conference Committee

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

National University of Singapore
20 Science Park Road #02-34/37
Singapore 117674
Singapore

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AOARD
UNIT 45002
APO AP 96337-5002

10. SPONSOR/MONITOR'S ACRONYM(S)

AOARD

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

CSP-011024

12. DISTRIBUTUION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13 SUPPLEMENTARY NOTES

14. ABSTRACT

This is an interdisciplinary conference. Topics include:
□- Detection, Estimation, and Classification Theory, Time-Frequency signal analysis
□- Non-stationary signal processing, Multirate processing and wavelets
□- Signal and system modeling
□- Non-Gaussian and non-linear signal processing
□- Sensor array processing
□- Signal processing for communications
□- Space-time coding and modulation
□- Computer-based statistical methods for signal processing
□-Applications in areas including communications, GPS, radar, sonar, biomedicine, machine diagnostics, etc.

15. SUBJECT TERMS

Signal Processing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Tae-Woo Park, Ph.D.
U	U	U	UU	615	19b. TELEPHONE NUMBER *(Include area code)* +81-3-5410-4409

2001 IEEE Workshop on Statistical Signal Processing Proceedings

Copyright and Reprint Permission: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Operations Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. All rights reserved. Copyright ©2001 by the Institute of Electrical and Electronics Engineers, Inc.

IEEE Catalog Number: 01TH8563

ISBN: 0-7803-7011-2

Library of Congress: 01-087383

11th IEEE Workshop on Statistical Signal Processing

ORGANIZING COMMITTEE

General Co-Chairs
Tariq S Durrani
University of Strathclyde, UK

A. Rahim Leyman
Center for Wireless Communications
Singapore

Technical Chair
Abdelhak M Zoubir
Curtin University of Technology, Australia

Finance Chair
Saeid Sanei
Singapore Polytechnic Singapore

Local Arrangement Chair
Chong Meng Samson See
DSO National Laboratories, Singapore

Publications Chair
Ying-Chang Liang
Center for Wireless Communications
Singapore

Publicity Co-Chairs
Samir Attallah
Curtin University of Technology, Australia
Antonia Papandreou-Suppapola
Arizona State University, USA

North American Liaison
Brian Sadler
Army Research Lab., USA

South American Liaison
Daniel R. Fuhrmann
University of Washington, St.Louis, USA

Australian Liaison
Stuart Anderson
Defense Science & Technology Organization, Australia.

European Liaison
Thomas Kaiser
Fraunhofer Institute for Microelectronic Circuits and Systems, Germany

TECHNICAL COMMITTEE

S-I Amari
Riken, Japan
M G Amin
Villanova Univ., USA
J F Böhme
Ruhr Univ. Bochum, Germany
C-Y Chi
National Tsing Hua Univ., Taiwan
P-C Ching
Chinese Univ. of Hong Kong, China
C C Ko
National Univ. of Singapore,Singapore
A Cichocki
Riken, Japan
P Comon
University of Nice Sophia Antipolis, France
P M Djuric
SUNY, Stony Brook, USA
A B Gershman
McMaster Univ., Canada
G B Giannakis
University of Minnesota, USA
B Guoan
Nanyang Tech. Univ., Singapore
Y Hua
University of Melbourne, Australia
M Kaveh
University of Minnesota, USA
H Krim
North Carolina State Univ., USA
M A Lagunas
Poly. Univ. of Catalonia, Spain
H Messer-Yaron
Tel Aviv Univ., Israel
J M F Moura
Carnegie Mellon Univ., USA
A Swami
Army Research Lab., USA
M Viberg
Chalmers Univ. of Tech., Sweden
L B White
University. of Adelaide, Australia
G Vazquez
Poly. Univ. of Catalonia, Spain

Financial Support

The organizing committee of the 11th IEEE Workshop of Statistical Signal Processing (SSP2001) expresses their gratitude to the following for financial support

U.S. Air Force Research Laboratory (**AFRL**), Information Directorate, USA

U.S. Air Force Office of Scientific Research (**AFOSR**), USA

U.S. Air Force Research Laboratory, Asian Office of Aerospace R&D (**AOARD**), USA

Defence Science & Technology Agency (**DSTA**), Singapore

Defence Science Organisation National Laboratories (**DSO**), Singapore

Hewlett Packard Singapore (Pte) Ltd

Message from the Chairmen

Dear Delegates,

We take great pleasure in welcoming you to Singapore and to the 11th IEEE Signal Processing Society workshop on Statistical Signal Processing (SSP). This is a new experience for the IEEE Signal Processing Society as it represents the new series of workshops that will be organized biennially, taking over from the Statistical Signal and Array Processing(SSAPs); and the first occasion that the Society has held one of its international workshops in Singapore.

This Workshop series brings together leading authorities in the field and experts from keynote industries from all over the world; and offers an excellent vehicle for the exchange of latest research results, and the identification of emerging trends and future directions. To this end, spurred by the enviable successes of previous SSAPs, we are committed to providing the best possible platform based on an exciting technical and social program.

Our Call for Papers was met with overwhelming response. To maintain the collegial and convivial atmosphere of the Workshop, of necessity, the number of technical presentations had to be limited. Five plenary lectures will be presented at the Workshop, that in itself speaks volumes for the tremendous growth in statistical signal processing. In addition, we hope that the attendant all-poster sessions will create the vibrant exchanges much needed in a workshop setting.

It is our sincere wish that you make the most out of the opportunities that await you at SSP2001. We are certain that the gathering of among the world's best, in itself, serves as an impetus for you to further enrich each other's knowledge and experience.

It has been a fascinating and invaluable experience for us in organizing this Workshop. We would like to express our gratitude to our financial supporters whose contributions have helped immensely to enhance the quality of the Workshop.

Thank you for your tremendous support. We hope that you would make time to enjoy the food, fun and excitement of Singapore. We wish you a most intellectually stimulating and socially fulfilling Workshop.

Here's to a Workshop that spells success in every way and provides an avenue for even greater things to come.

Tariq S Durrani

A. Rahim Leyman

Message from the Technical Chairman

Dear Delegates

What was formerly known as the IEEE Workshop on Statistical Signal and Array Processing (SSAP) has taken on a new term. It is now called IEEE Workshop on Statistical Signal Processing.

The 11th IEEE Workshop on Statistical Signal Processing (SSP) is a forum for engineers, mathematicians and scientists to present and discuss issues ranging from theoretical and methodological developments to practical applications of statistical signal processing.

It thus offers an excellent opportunity for you to network and keep up with your counterparts from the world over. It therefore gives me great pleasure to welcome you to such a prestigious event.

This Workshop offers an outstanding technical program, covering diverse areas of statistical signal processing, stretching from wireless communications to biomedical signal processing. The technical programme comprises five plenary talks and 150 papers presented as posters in 19 technical sessions.

The response to the Call for Papers was overwhelming; we received over 200 submissions of the highest quality. All submissions underwent a thorough review by at least two members of the technical committee and/or other distinguished members of the statistical signal processing community. I wish to thank all reviewers for their tremendous effort and timely responses. To maintain a workshop atmosphere, as well as a technical program of the highest quality, we accommodated 64 percent of the regular submissions.

We have included 6 special sessions in the technical programme. These comprise regular as well as invited papers. I wish to thank the special sessions organizers for putting much effort in securing highly qualified speakers for their sessions.

A lot of effort has been put in by many quarters to ensure that this event runs smoothly. I would like to thank each and everyone for their tireless work. I would like to especially extend my gratitude to the plenary speakers for agreeing to participate and share their knowledge and expertise with the statistical signal processing community.

I am certain that you will find the technical program interesting and inspiring. Aside from serious work, I also hope that many of you will be able to take advantage of the many sightseeing and gastronomic opportunities Singapore has to offer.

A.M. Zoubir

Table of Contents

Plenary Session I
Aspects of Contemporary Statistical Methods 1
Peter Hall , Australian National University

Plenary Session II
A Geometric and Multiresolution Analysis Approach to Robust Detection 2
José M. F. Moura;, Carnegie Mellon University, USA

Plenary Section III
Entropy, Complexity and Chaos in Brain Rhythms 3
Nitish V. Thakor, The Johns Hopkins School of Medicine Baltimore, USA

Plenary Section IV
Statistical GNSS Carrier Phase Ambiguity Resolution: A Review 4
P.J.G. Teunissen, Delft University of Technology (The Netherlands)

Plenary Section V
On the Role of Linear Precoding in Signal Processing for Wireless 13
Georgios B. Giannakis, University of Minnesota, USA

Session MA1 : Resampling and Monte Carlo Methods for Statistical Signal Processing
Chair: Petar Djuric, State University of New York at Stony Brook(USA)
Time & Place: Monday, 10:30 – 12:30, Ruby Suite

Recursive Monte Carlo Algorithms for Parameter Estimation in General State Space Models 14
Christophe. Andrieu, University of Bristol (UK), and A. Doucet, The University of Melbourne(Australia)

Monte Carlo Smoothing with Application to Audio Signal Enhancement 18
William Fong and Simon Godsill, University of Cambridge(UK)

Parameter Estimation by a Markov Chain Monte Carlo Technique for the Candy Model 22
X. Descombes, INRIA(France), M.N.M. van Lieshout, R.Stoica, CWI(The Netherlands), and J.Zerubia, INRIA(France)

Resampling Based Techniques for Source Detection in Array Processing 26
Ramon F. Brcich and Abdelhak M. Zoubir, Curtin University of Technology(Australia)

The Sequential MCMC Filter: Formulation and Applications 30
Dominic S. Lee and Nicholas K. K. Chia, DSO National Laboratories(Singapore)

A Framework For Particle Filtering in Positioning, Navigation and Tracking Problems 34
F. Gustaffson, F.Gunnarsson, N. Bergman, U. Forssell, J. Jansson, P-J Nordlund and R. Karlsson, Linkoping University(Sweden)

Particle Filtering for Multiuser Detection in Fading CDMA Channels 38
E. Punskaya, University of Cambridge(UK), C. Andrieu, University of Bristol(UK), A. Doucet, The University of Melbourne(Australia), and W. J. Fitzgerald, University of Cambridge(UK)

The Rejection Gibbs Coupler: A Perfect Sampling Algorithm and its Application to Truncated Multivariate Gaussian Distributions 42
Yufei Huang, Tadesse Ghirmai and Petar M. Djuric, State University of New York at Stony Brook(USA)

Assessment of MCMC Convergence: A Time Series and Dynamical Systems Approach 46
Rodney C. Wolff, Queensland University of Technology(Australia), Darfiana Nur, Kerrie L. Mengersen, University of Newcastle(Australia)

Importance Sampling Analysis of Digital Phase Detectors with Carrier Phase Tracking 50
Francisco A. S. Silva and Jose M. N. Leitao, Instituto Superior Tecnico(Portugal)

Session MA2: Self-similarity, Long Range Dependence and Heavy Tails
Chair: Raghuveer Rao, Rochester Institute of Technology(USA)
Time & Place: Monday, 10:30 – 12:30, Jade 1 & 2

Long-Range Dependent Alpha-Stable Impulsive Noise in a Poisson Field of Interferers 54
Xueshi Yang and Athina P. Petropulu, Drexel University(USA)

Score Functions for Locally Suboptimum and Locally Suboptimum Rank Detection in Alpha-Stable Interference 58
Christopher L. Brown, Chalmers University of Technology(Sweden)

Kernel Approach to Discrete-Time Linear Scale-Invariant Systems 62
Seungsin Lee and Raghuveer Rao, Rochester Institute of Technology(USA)

Stochastic Discrete Scale Invariance and Lamperti Transformation 66
Pierre Borgnat, Patrick Flandrin, Ecole Normale Superieure de Lyon(France), and Pierre-Olivier Amblard, LIS-UMR CNRS(France)

The Viterbi Algorithm for Impulsive Noise with Unknown Parameters 70
Thomas Kaiser and Youssef Dhibi, Fraunhofer Institute for Microelectronic Circuits and Systems(Germany)

Self-Similar Traffic Sources: Modeling and Real-Time Resource Allocation 74
Krishnamurthy Nagarajan, Couth Infotech Pvt.Ltd.(India), and G. Tong Zhou, Georgia Institite of Technology(USA)

Super-Efficiency in Blind Signal Separation of Symmetric Heavy Tailed Sources 78
Yoav Shereshevski, Arie Yeredor and Hagit Messer, Tel Aviv University(Israel)

Kalman Filtering for Self-Similar Processes 82
Meltem Izzetoglu, Birsen Yazici, Banu Onaral, and Nihat Bilgutay, Drexel University(USA)

Bayesian Array Signal Processing in Additive Generalized Gaussian Noise 86
B. Kannan, Center for Wireless Communications(Singapore)

Nonlinear Image Filtering in a Mixture of Gaussian and Heavy-Tailed Noise 90
A. Ben Hamza and Hamid Krim, North Carolina State University(USA)

Session MA3: Multiaccess Communication Systems
Chair: Miguel Angel Lagunas Hernandez, Polytechnic University of Catalonia(Spain)
Time & Place: Monday, 10:30 – 12:30, Ruby Suite

A Mixed-Cost Blind Adaptive Receiver for DS-CDMA 94
Peerapol Yuvapoositanon, and Jonathon A. Chambers, University of Bath(UK).

A Reduced-Rank Decorrelating RAKE Receivers for CDMA Communications over Frequency Selective Channels 98
Ozgur Ozdemir and Murat Torlak, The University of Texas at Dallas(USA)

Multi-User Detection in Impulsive Noise 102
A. M. Zoubir and A. T. Lane-Glover, Curtin University of Technology(Australia)

Subspace Based Blind Adaptive Multiuser Detection for Multirate DS/CDMA Signals 106
L. Huang, F. C. Zheng and M. Faulkner, Victoria University of Technology(Australia)

A Decision Feedback CDMA Receiver with Partially Adaptive Interference Suppression 110
Gau-Joe Lin, Ta-Sung Lee, and Chan-Choo Tan, National Chiao Tung University(Taiwan)

Iterative Space-Time Soft Detection in Time-Varying Multiaccess Wireless Channels 114
Joaquín Míguez and Luis Castedo, Universidade da Coruna(Spain)

Blind Equalization using Cumulant Based MIMO Inverse Filter Criteria for Multiuser DS/CDMA Systems in Multipath 118
Chong-Yung Chi and Chii-Horng Chen, National Tsing Hua University(Taiwan)

Combined Downlink Beamforming and Channel Estimation for High Data Rates CDMA Systems 122
Sylvie Perreau, University of South Australia(Australia)

Session MP1: Radar Signal Processing
Chair: Stuart Anderson, DSTO(Australia)
Time & Place: Monday, 15:30 – 18:00, Jade 1 & 2

Feature Discovery and Sensor Discrimination in a Network of Distributed Radar Sensors for Target Tracking 126
S. Kadambe, HRL Laboratories(USA)

An Information Divergence Measure for ISAR Image Registration 130
A. Ben Hamza, Yun He, and Hamid Krim, North Carolina State University(USA)

An Advanced STAP Implementation for Surveillance Radar Systems 134
G. A. Fabrizio and M. D. Turley, Defence Science and Technology Organisation(Australia)

Monobit Receiver for Electronic Warfare 138
Jesús Grajal, Raúl Blázquez, Gustavo López, José M. Sanz and Mateo Burgos, Universidad Politecnica Madrid(Spain)

Neural Net Based Variable Structure Multiple Model Reducing Mode Set Jump Delay 142
Daebum Choi, Byungha Ahn, Kwangju Institite of Science and Technology(Korea) and Hanseok Ko, Korea University Anam-dong(Korea)

Session MP2: Signal Processing for Landmine Detection
Chair: Mats Viberg, Chalmers University of Technology(Sweden)
Time & Place: Monday, 15:30 – 18:00, Ruby Suite

Comparison of PCA and ICA based Clutter Reduction in GPR Systems for Anti-Personal Landmine Detection 146
Brian Karlsen, Jan Larsen, Helge B. D. Sørensen, Kaj B. Jakobsen, Technical University of Denmark(Denmark)

Elimination of Leakage and Ground-Bounce in Ground-Penetrating Radar 150
R. Abrahamsson, E. G. Larsson, J. Li, University of Florida, Gainesville(USA), J. Habersat, G. Maksymonko, US Army Night Vision and Electronic Sensor Directorate(USA), and M. Bradley, Planning Systems Inc(USA)

Towards Real-Time Detection of Landmines in FLIR Imagery 154
Mabo R. Ito, Sinh Duong, University of British Columbia(Canada), John E. McFee and Kevin L. Russell, Defense Research Establishment Suffield(Canada)

Signal Processing Techniques for Clutter Parameters Estimation and Clutter Removal in GPR data for Landmine Detection 158
L. van Kempen and H. Sahli, Vrije Universiteit Brussels(Belgium)

Model-Based Statistical Signal Processing using Electromagnetic Induction Data for Landmine Detection and Classification 162
Leslie Collins, Ping Gao and Stacy Tantum,Duke University(USA)

Polynomial Phase Signal Based Detection of Buried Landmines using Ground Penetrating Radar 166
Luke A. Cirillo, Christopher L. Brown and Abdelhak M. Zoubir, Curtin University of Technology(Australia)

Land Mine Detection in Rotationally Invariant Noise Fields 170
Magnus Lundberg and Lennart Svensson, Chalmers University of Technology(Sweden)

Session MP3: Signal Detection and Applications

Chair: Jose M. F. Moura, Carnegie Mellon University(USA)
Time & Place: Monday, 15:30 – 18:00, Ruby Suite

Unknown Signal Detection via Atomic Decomposition 174
Gustavo López-Risueño and Jesús Grajal, Universidad Politecnica de Madrid(Spain)

Multipath Detection of Stochastic Transient Processes 178
Francisco M. Garcia and Isabel M. G. Lourtie, IST(Portugal)

Detection of Independent Timing Jitter in Sinusoidal Measurements 182
Mark R. Morelande, Queensland University of Technology(Australia), and D. Robert Iskander, Griffith University(Australia)

Passive Signature Characterization and Classification by Means of Nonlinear Dynamics 186
Ron K. Lennartsson, Swedish Defence Research Agency(Sweden), James B. Kadtke and Aron Pentek, University of California La Jolla(USA)

Multichannel Detection and Spatial Signature Estimation with Uncalibrated Receivers 190
Amir Leshem, Metalink Broadband Access(Israel), and Alle-Jan van der Veen, Delft University of Technology(The Netherlands)

A New Wavelet-Based Tracking Algorithm for Rapidly Time-Varying Systems 194
Yuanjin Zheng, Institute of Microelectronics (Singapore) and Zhiping Lin, Nanyang Technological University(Singapore)

An Application of the Maximum Likelihood Principle to Semiblind Space-Time Linear Detection in Multiple-Access Wireless Communications. 198
Mónica F. Bugallo, Joaquín Míguez and Luis Castedo, Universidade da Coruna(Spain)

Recursive Bayesian Phase Estimation in Ranging And Mobile Communication 202
José M.N. Leitão, Institute Superior Tecnico(Portugal) and Fernando M.G. Sousa, Instituto Superior de Engenharia de Lisboa(Portugal)

A Study of a Time-Frequency Based Detectors for FSK Modulated Signals in a Flat Fading Channel 206
B. Barkat, Nanyang Technological University(Singapore), and S. Attallah, Center for Wireless Communications(Singapore)

MRC Receiver Performance with MQAM in Correlated Rician Fading Channels 210
Chunhua Yang, Guoan Bi, Nanyang Technological University(Singapore), and A.Rahim. Leyman, Center for Wireless Communications(Singapore)

Session MP4: Sensor Array Processing
Chair: Alex B. Gershman, McMaster University(Canada)
Time & Place: Monday, 15:30 – 18:00, Jade 1 & 2

Optimal Preprocessing for Source Localization by Fewer Receivers than Sensors 213
Joseph Tabrikian and Avi Faizakov, Ben-Gurion University of the Negev(Israel)

A Robust Technique for Array Interpolation Using Second-Order Cone Programming 217
Marius Pesavento, Ruhr University Bochum(Germany) Alex B. Gershman and Zhi-Quan Luo, McMaster University(Canada)

Performance Analysis of Mis-Modeled Estimation Procedures for a Distributed Source of Non-Constant Modulus 221
Jonathan Friedmann, Raviv Raich, Jason Goldberg and Hagit Messer, Tel-Aviv University(Israel)

A New Algorithm for Computing the Extreme Eigenvectors of a Complex Hermitian Matrix 225
Jonathan H. Manton, The University of Melbourne(Australia)

Locally Optimal Maximum-Likelihood Completion of a Partially Specified Toeplitz Covariance Matrix 229
Yuri I. Abramovich, Defence Science and Technology Organisation(Australia) and Nicholas K. Spencer, Cooperative Research Centre for Sensor Signal and Information Processing(Australia)

Session TA1: Biomedical Signal Processing

Chair: Nitish V. Thakor, The John Hopkins School of Medicine(USA)
Time & Place: Tuesday, 10:30 – 12:30, Jade 1 & 2

Biological Signal Filtering, Separation and Reconstruction

EEG Brain Map Reconstruction Using Blind Source Separation 233
S. Sanei, Singapore Polytechnic(Singapore), A. Rahim Leyman, Center for Wireless Communication(Singapore)

Extraction of Superimposed Evoked Potentials by Combination of Independent Component Analysis and Cumulant-Based Matched Filtering 237
A. Cichocki, R. R. Gharieb, RIKEN(Japan), and N. Mourad, South-Valley Faculty of Engineering(Egypt)

Estimating the Dynamics of Aberration Components in the Human Eye 241
D. R. Iskander, Griffith University(Australia), Mark R. Morelande, and Michael J. Collins, Queensland University of Technology(Australia)

A Time-Varying Model for DNA Sequencing Data 245
Nicholas M. Haan and Simon J. Godsill, University of Cambridge(UK)

Single Trial VEP Extraction Using Digital Filter 249
R. Palaniappan and P. Raveendran, University of Malaya(Malaysia)

Nonlinear Modeling in Biological Systems

Pathological Analysis of Myocardial Cell under Ventricular Tachycardia and Fibrillation Based on Symbolic Dynamics 253
Zhu Yi-sheng, Zhang Hongxuan, Shanghai Jiao Tong University(China) and Nitish V. Thakor, The John Hopkins School of Medicine(USA)

On the Application of Model Based Distance Metrics of Signals for Detection of Brain Injury 257
J. S. Paul, S. Tong, D. Sherman, A. Bezerianos, and N. V. Thakor, The John Hopkins School of Medicine(USA)

Entropy of Brain Rythms: Normal versus Injury EEG 261
N. V. Thakor, J. Paul, S. Tong, John Hopkins School of Medicine(USA), Y. Zhu, Shanghai Jian Tong University(China), and A. Bezerianos, University of Patras(Greece)

Session TA2: Blind System Identification

Chair: Karim Abed-Meraim, Telecom Paris(France)

Time & Place: Tuesday, 10:30 – 12:30, Ruby Suite

Blind Identification and Equalization of Minimum-Phase Channels 265
Senjian An, The University of Melbourne(Australia) and Yingbo Hua, University of California Riverside(USA)

On Blind Equalization of Rank Deficient Nonlinear Channels 269
Roberto Lopez-Valcarce, Universidad de Vigo(Spain) and Soura Dasgupta, University of Iowa(USA)

Multichannel Blind Deconvolution of Colored Signals via EigenValue Decomposition 273
Pando Georgiev and Andrzej Cichocki, RIKEN(Japan)

A Second Order Statistics Based Optimization Approach for Blind MIMO System Identification 277
Ivan Bradaric, Athina P. Petropulu, Drexel University(USA) and Konstantinos I. Diamantaras, Technological Education Institute(Greece)

Blind Channel Identification using Robust Subspace Estimation 281
S. Visuri, Helsinki University of Technology(Finland), H. Oja, University of Jyvaskyla(Finland), and V. Koivunen, Helsinki University of Technology(Finland)

Orthogonal Minimum Noise Subspace for Multiple Input Multiple Output System Identification 285
Anahid Safavi and Karim Abed-Meraim, Telecom Paris(France)

Recursive Semi Blind Equalizer For Time Varying Mimo Channels 289
Mihai Enescu, Marius Sirbu and Visa Koivunen, Helsinki University of Technology(Finland)

Blind Single-Input Multi-Output (SIMO) Channel Identification with Application to Time Delay Estimation 293
Chong-Yung Chi, Xianwen Chang and Chii-Horng Chen, National Tsing Hua University(Taiwan)

Multichannel System Identification Methods for Sensor Array Calibration in Uncertain Multipath Environment 297
Vijay Vardarajan and Jeffrey L. Krolik, Duke University(USA)

A Cooperative Maximum Likelihood MIMO Channel Estimator 301
Marc Chenu-Tournier, Thales-Coomunication(France) and Pascal Larzabal, LESiR/CNRS(France)

Session TA3: Non-Stationary Signal Analysis

Chair: Moeness G. Amin, Villanova University(USA)

Time & Place: Tuesday, 10:30 – 12:30, Ruby Suite

Nonstationary Signal Classification using Support Vector Machines 305
Arthur Gretton, Manuel Davy, University of Cambridge(UK), Arnaud Doucet, The University of Melbourne(Australia) and Peter J. W. Rayner University of Cambridge(UK)

Improved Auxiliary Particle Filtering: Applications to Time-Varying Spectral Analysis 309
Christophe Andrieu, University of Bristol(UK), Manuel Davy, University of Cambridge(UK) and Arnaud Doucet, The University of Melbourne(Australia)

Spatial and Time-Frequency Signature Estimation of Nonstationary Sources 313
Moeness G. Amin, Weifeng Mu and Yimin Zhang, Villanova University(USA)

Fast Computation of Discrete SLTF Transform 317
Osama A. Ahmed, KFUPM(Saudi Arabia)

Fractional-Fourier Domain Weighted Wigner Distribution 321
LJubiša Stankovic, University of Montenegro(Montenegro), Tatiana Alieva and Martin J. Bastiaans, Technische Universiteit Eindhoven (The Netherlands)

Wigner Distribution Reconstruction from Two Projections 325
Tatiana Alieva, Martin J. Bastiaans, Technische Universiteit Eindhoven (The Netherlands) and LJubiša Stankovic, University of Montenegro(Montenegro)

Prediction of Time Varying Composite Sources by Temporal Fuzzy Clustering 329
S. Policker and A. B. Geva, Ben-Gurion University of the Negev(Israel)

On the Use of a New Compact Support Kernel in Time Frequency Analysis 333
Adel Belouchrani, Ecole Nationale Polytechnique (Algeria), and Mohamed Cheriet, Ecole de Technologie Superieure(Canada)

A Weighted Decomposition of the Wigner Distribution 337
Junfeng Wang, Xiang Yan, Antonio H. Costa and Dayalan Kasilingam, University of Massachusetts Dartmouth(USA)

Session TA4: System Identification & Filter Design

Chair: Langford B. White, Adelaide University(Australia)

Time & Place: Tuesday, 10:30 – 12:30, Ruby Suite

Support Vector Regression for Black-Box System Identification 341
Arthur Gretton, University of Cambridge(UK), Arnaud Doucet, The University of Melbourne(Australia), Ralf Herbrich, Microsoft Research(UK), Peter J. W. Rayner, University of Cambridge(UK) and Bernhard Schölkopf, Microsoft Research(UK)

Time-Varying Quadratic Model Selection using Wavelet Packets 345
Matthew Green, Curtin University of Technology(Australia)

Internet Transport Layer System Identification 349
Langford B. White, Adelaide University(Australia)

Optimal Design of Variable Fractional-Delay Digital Filters 353
Tian-Bo Deng, Toho University(Japan)

Design of Digital Filters with Amplitude and Group Delay Specifications 357
Zhuquan Zang, Sven Nordholm, Curtin University of Technology(Australia), Sven Nordebo, Blekinge Institute of Technology(Sweden) and Antonio Cantoni, Curtin University of Technology(Australia)

Session TA5: Signal Processing Applications & Implementations

Chair: Alle-Jan van der Veen, Delft University of Technology(The Netherlands)

Time & Place: Tuesday, 10:30 – 12:30, Jade 1 & 2

Performance Analysis of Subspace Projection Techniques for Anti-Jamming GPS Using Spatio-Temporal Interference Signatures 361
Moeness G. Amin, Liang Zhao, Villanova University(USA), and Alan R. Lindsey, Air Force Research Labs(USA)

Gain Decomposition Methods for Radio Telescope Arrays 365
A. J. Boonstra, ASTRON(The Netherlands), and A. J. van der Veen, Delft University of Technology(The Netherlands)

Identification of Gear Mesh Signals by Kurtosis Maximisation and Its Application to CH46 Helicopter Gearbox Data 369
Wenyi Wang, Defence Science and Technology Organisation(Australia)

A Hyperbolic LMS Algorithm for CORDIC Based Realization 373
Mrityunjoy Chakraborty, Suraiya Pervin, and T. S. Lamba, Indian Institute of Technology(Kharagpur)

On External Cabliration of Analog-to-Digital Converters 377
Henrik Lundin, Mikael Skoglund and Peter Händel, Royal Institute of Technology (Sweden)

Session WA1: Multi-Carrier Communications Systems

Chair: Georgios B Giannakis, University of Minnesota (USA)
Time & Place: Wednesday, 10:30 – 12:30, Ruby Suite

Semi-Blind Channel Estimation for Block Precoded Space-Time OFDM Transmissions 381
Shengli Zhou, University of Minnesota(USA), Bertrand Muquet, Motorola Labs Paris(France) and Georgios B. Giannakis, University of Minnesota(USA)

Comparing DS-CDMA and Multicarrier CDMA with Imperfect Channel Estimation 385
Lucy L. Chong and Laurence B. Milstein, University of California San Diego(USA)

Asymptotic Performance Analysis for Redundant Block Precoded OFDM systems with MMSE Equalization 389
Merouane Debbah, Motorola Labs Paris(France), Walid Hachem, Service de radioelectricite(France), Philippe Loubaton, Universite de Marne la Vallee(France), and Marc de Courville, Motorola Labs Paris(France)

An Algorithm For Joint Symbol Timing And Channel Estimation For OFDM Systems 393
Erik G. Larsson, Jian Li, Guoqing Liu, University of Florida, Gainsville(USA) and Georgios B. Giannakis, University of Minnesota(USA)

A Distributed Algorithm for Dynamic Sub-Channel Assignment in a Multi-User OFDM Communication System 397
Alireza Seyedi and Gary J. Saulnier, Rensselaer Polytechnic Institute(USA)

A SOS Subspace Method for Blind Channel Identification and Equalization in Bandwidth Efficient OFDM Systems Based on Receive Antenna Diversity 401
Hassan Ali, Jonathan H. Manton, The University of Melbourne(Australia), and Yingbo Hua, University of California Riverside(USA)

A Channel Coded CP-OFDM Interpretation of TZ-OFDM Systems 405
Jonathan H. Manton, The University of Melbourne(Australia)

Session WA2: Parameter Estimation and Applications

Chair: Yuri I. Abramovich, DSTO(Australia)

Time & Place: Wednesday, 10:30 – 12:30, Ruby Suite

Frequency Estimation Utilizing the Hadamard Transform 409
Tomas Andersson, Mikael Skoglund and Peter Händel, Royal Institute of Technology(Sweden)

Best Quadratic Unbiased Estimator (BQUE) for Timing and Frequency Synchronization 413
Javier Villares and Gregori Vázquez, Polytechnic University of Catalonia(Spain)

An Efficient Hilbert Transform Interpolation Algorithm for Peak Position Estimation 417
Saman S. Abeysekera, Nanyang Technological University(Singapore)

Computationally Efficient Interative Refinement Techniques for Polynomial Phase Signals 421
Simon Sando, Dawei Huang and Tony Pettitt, Queensland University of Technology(Australia)

On Design of Correlation Based Frequency Estimators 425
Björn Völcker and Peter Händel, Royal Institute of Technology(Sweden)

Gaussian Particle Filtering 429
Jayesh H. Kotecha and Petar M. Djuric, State University of New York at Stony Brook(USA)

Bayesian Learning using Gaussian Process for Time Series Prediction 433
Sofiane Brahim-Belhouari and Jean-Marc Vesin, Swiss Federal Institute of Technology(Switzerland)

On the Estimation of Common Non-linearity Among Repeated Time Series 437
Adrian G. Barnett and Rodney C. Wolff, Queensland University of Technology(Australia)

Time Delay Estimation for Multipath CDMA-Systems based on a Fast Minimization Technique for Subspace Fitting 440
Patrik Bohlin, Anders Ranheim and Per Pelin, Chalmers University of Technology(Sweden)

Session WA3: Blind Source Separation

Chair: Hagit Messer-Yaron, Tel Aviv University(Israel)

Time & Place: Wednesday, 10:30 – 12:30, Jade 1 & 2

Blind Separation of Second-Order Non-stationary and Temporally Colored Sources 444
Seungjin Choi, POSTECH(Korea), Andrzej Cichocki, RIKEN(Japan) and Adel Belouchrani, Ecole Nationale Polytechnique(Algeria)

Blind Separation of Non-Stationary Sources Using Joint Block Diagonalization 448
Hicham Bousbia-Salah, Adel Belouchrani, Ecole Nationale Polytechnique(Algeria) and Karim Abed-Meraim, Telecom Paris(France)

Blind Source Separation of Audio Signals using Improved ICA Method 452
F. Sattar, M. Y. Siyal, L. C. Wee, and L. C. Yen, Nanyang Technological University(Singapore)

Weighted Closed-form Estimators for Blind Source Separation 456
Vicente Zarzoso, Frank Herrmann and Asoke K. Nandi, The University of Liverpool(UK)

Large Sample Performance Analysis of ACMA 460
Alle-Jan van der Veen, Delft University of Technology(The Netherlands)

Fast-Convergence Algorithm for ICA-Based Blind Source Separation Using Array Signal Processing 464
Hiroshi Saruwatari, Toshiya Kawamura and Kiyohiro Shikano, Nara Institute of Science and Technology(Japan)

Session WA4: Multidimensional Signal, Speech and Audio Processing
Chair: Andrzej Cichocki, RIKEN(Japan)
Time & Place: Wednesday, 10:30 – 12:30, Ruby Suite

Recognition of Facial Images using Support Vector Machines 468
K. I. Kim, J. Kim, Korea Advanced Institute of Science and Technology(Korea) and K. Jung, Sungkyunkwan University(Korea)

Higher Order Conditional Entropy-constrained Trellis-coded RVQ with Application to Pyramid Image Coding 472
Mohammad Asmat Ullah Khan, KFUPM(Saudi Arabia)

Texture Segmentation by Clustering the Phase of HOS Cepstra 476
S. Sanei, J. Li and S. H. Ong, Singapore Polytechnic(Singapore)

R-D Quantisation of Complex Coefficients in Zerotree Coding 480
T .H. Reeves and N.G. Kingsbury, University of Cambridge(UK)

Irregular Sampling Problems and Selective Reconstructions Associated with Motion Transformations 484
Jean-Pierre Leduc, University of Maryland(USA)

Nonlinear Perceptual Audio Filtering using Support Vector Machines 488
Simon I. Hill, Patrick J. Wolfe and Peter J. W. Rayner, University of Cambridge(UK)

A Probabilistic Framework for Subband Autoregressive Models Applied to Room Acoustics 492
James R. Hopgood and Peter J. W. Rayner University of Cambridge(UK)

Simple Alternatives to the Ephraim and Malah Suppression Rule for Speech Enhancement 496
Patrick J. Wolfe and Simon J. Godsill, University of Cambridge(UK)

A Spectral Distance Measure For Speech Detection In Noise and Speech Segmentation 500
K. Drouiche, Universite de Cergy-Pontoise (France), P. Gomez, A. Alvarez, R. Martinez, V. Rodellar, and V. Nieto, Universidad Politecnica de Madrid(Spain)

Session WP1: DOA Estimation and Source Localization
Chair: Johann F. Bohme, Ruhr University Bochum(Germany)
Time & Place: Wednesday, 15:30 – 18:00, Ruby Suite

A New Algorithm for Joint DOA and Multipath Delay Estimation: Separable Dimension Subspace Method 504
Jian Mao, Sigprowireless Inc.(Canada), Benoit Champagne, McGill University(Canada), Mairtin O'Droma, University of Limerick (Ireland) and Lijia Ge, Chongqin University(China)

High-Resolution Direction Finding using a Switched Parasitic Antenna 508
Thomas Svantesson, Chalmers University of Technology(Sweden) and Mattias Wennström, Uppsala University(Sweden)

Two-Step Music Algorithm for Improved Array Resolution 512
R. Chavanne, Office, National d'Etudes et de Recherches Aerospatiales(France), K. Abed Meraim,Telecom Paris(France) and D. Médynski, National d'Etudes et de Recherches Aerospatiales(France)

Optimization of Element Positions for Direction Finding with Sparse Arrays 516
Fredrik Athley, Chalmers University of Technology(Sweden)

High Resolution DF with a Single Channel Receiver 520
Chong Meng Samson See, DSO National Laboratories(Singapore)

Directions-of-Arrival Estimation of Cyclostationary Signals in Multipath Propagation Environment 524
Jingmin Xin, YRP Mobile Telecoms(Japan), and Akira Sano,Keio University(Japan)

Iterative Algorithm for the Estimation of Distributed Sources Localization Parameters 528
Antonio Pascual Iserte, Ana I. Pérez Neira and Miguel Ángel Lagunas Hernández, Polytechnic University of Catalonia(Spain)

Array Signal Processing for Recursive Tracking of Multiple Moving Sources based on LPA Beamforming 532
Vladimir Katkovnik and Yonghoon Kim, Kwangju Institute of Science and Technology(Korea)

Direction Finding in Partly Calibrated Arrays Composed of Nonidentical Subarrays: A Computationally Efficient Algorithm for the Rank Reduction (RARE) Estimator 536
Marius Pesavento, Ruhr University Bochum(Germany), Alex B. Gershman, Kon Max Wong, McMaster University(Canada) and Johann F. Böhme, Ruhr University Bochum(Germany)

Recursive EM and SAGE Algorithms 540
Pei Jung Chung and Johann F. Böhme, Ruhr University Bochum(Germany)

Session WP2: Channel Estimation & Equalization
Chair: Chong-Yung Chii,
Time & Place: Wednesday, 15:30 – 18:00, Jade 1 & 2

Convergence Performance of Subband Arrays for Spatio-Temporal Equalization 544
Yimin Zhang, Villanova University(USA), Kehu Yang, ATR Adaptive Communications Research Labs(Japan), and Moeness G. Amin, Villanova University(USA)

Simulation and Performance Bounds for Real-time Prediction of the Mobile Multipath Channel 548
Paul D. Teal and Rodney G. Vaughan, Industrial Research Limited(New Zealand)

Robustness of the Finite-length MMSE-DFE with respect to Channel and Second-order Statistics Estimation Errors 552
Athanasios P. Liavas,University of Ioannina(Greece)

Performance Evaluation of Blind Channel Estimation using a Frequency Domain Base-band Communication Model 556
Saman S. Abeysekera and Patrick K.S. Ong, Nanyang Technological University(Singapore)

Joint Channel Estimation and Decoding of Space-Time Trellis Codes 559
Jianqiu Zhang and Petar M. Djuric, State University of New York at Stony Brook(USA)

A Modified Constant Modulus Algorithm for Adaptive Channel Equalization for QAM Signals 563
Moeness G. Amin, Lin He, Villanova University(USA), Charles Reed Jr,, and Robert Malkemes, Sarnoff Corporation(USA)

A Performance Comparison of Fullband and Different Subband Adaptive Equalisers 567
Hafizal Mohamad, Stephan Weiss, University of Southampton(UK), Markus Rupp, Lucent Technologies(USA), and Lajos Hanzo, University of Southampton(UK)

Simulation of Wideband Mobile Radio Channels using Subsampled ARMA Models and Multistage Interpolation 571
Dieter Schafhuber, Gerald Matz and Franz Hlawatsch, Vienna University of Technology(Austria)

A Mixed MAP/MLSE Receiver for Convolutional Coded Signals Transmitted over a Fading Channel 575
Langford B. White, Adelaide University(Australia), and Robert J. Elliott, University of Alberta(Canada)

Session WP3: Multirate Signal Processing and Applications
Chair: Alan Lindsey, US Air Force Research Labs(USA)
Time & Place: Wednesday, 15:30 – 18:00, Ruby Suite

Orthogonal Extensions of AR Processes without Artifical Discontinuities for Size-limited Filter Banks 579
M. E. Dominguez Jiménez, Universidad Politecnica de Madrid(Spain), and N. Gonzalez Prelcic, Universidade de Vigo(Spain)

M-Band Perfect-Reconstruction Linear-Phase Filter Banks 583
X. M. Xie, S. C. Chan and T. I. Yuk, The University of Hong Kong(Hong Kong)

Improvement of Factorization for Two-Channel Perfect Reconstruction FIR Filter Banks 587
Shi Guangiming and Jiao Licheng, Xidian University(China)

Subband Adaptive Generalized Sidelobe Canceller for Broadband Beamforming 591
Wei Lui, Stephan Weiss and Lajos Hanzo, University of Southampton(UK)

An Efficient Design of Fractional-Delay Digital FIR Filters Using The Farrow Structure 595
Carson K. S. Pun, Y. C. Wu, S. C. Chan, and K. L. Ho, The University of Hong Kong(Hong Kong)

Efficient Design of a Class of Multiplier-Less Perfect Reconstruction Two-Channel Filter Banks and Wavelets with Prescribed Output Accuracy 599
Carson K. S. Pun, S. C. Chan, and K. L. Ho, The University of Hong Kong(Hong Kong)

Optimal Biorthogonal Filter Banks with Minimization of Quantization Noise Amplification 603
Arunkumar M, Sasken Communication Technologies Ltd(India) and Anamitran Makur, IIS(India)

Doubly Orthogonal Wavelet Packets for DS-CDMA Communication 607
Hongbing Zhang, H. Howard Fan, University of Cincinnati(USA) and Alan Lindsey, Air Force Research Lab(USA)

Author's Index 611

ABSTRACT

Aspects of Contemporary Statistical Methods

Peter Hall
Australian National University

Commenting on the development of statistics early in the 20th century, the UCLA historian Theodore Porter wrote that "the foundations of mathematical statistics were laid between 1890 and 1930", and argued that "the principal families of techniques for analyzing numerical data were established during the same period." There was of course a revolution in quantitative data analysis in the early part of last century, leading to the development of the subject we know today as Statistics. And at the time Porter wrote, 15 years ago, he would also have been correct in his second assertion. However, it would be difficult to justify the same remarks today. The speed and memory of computers have increased one thousand fold since 1986, and the second revolution in statistics, certainly motivated and perhaps driven by developments in computing, has begun to fundamentally change statistical methodology. It is a long way from running its course. Over the next few decades it will transform the subject into something that is quite different, in terms of its range and the emphases on types of problems that it treats, from that which we know today. If the development of statistics had taken place in the environment of contemporary advances in computing then the subject would most likely be less mathematical, and more of an experimental science, then it is today. The talk will discuss some of the changes, in areas of resampling and Monte Carlo methods, and outline new directions for at least the near future.

0-7803-7011-2/01/$10.00 ©2001 IEEE

ABSTRACT

A Geometric and Multiresolution Analysis Approach to Robust Detection

José M. F. Moura;
Carnegie Mellon University, USA

Detection algorithms whose design takes into account prior knowledge about the signals and the channel face a quandary: they provide marked improvement in performance when the field operating conditions match well this available knowledge; but they experience strong degradation when the actual conditions depart from the assumed ones. In other words, high resolution and robustness are commonly at odds. A third important variable affecting this tradeoff is the computational complexity of the solution. I will describe a geometric based approach to designing detectors that leads to a satisfying compromise: simple to implement detectors that are robust to mismatches and that exhibit good performance. The approach designs a representation subspace that is a good approximation (in the gap metric sense) to the signal set (à priori information), and uses multiresolution and wavelet analysis to design the representation subspace and implement the detector. I will illustrate the approach with multipath channels, and present detection results that illustrate the robustness of the geometric gap detector.

0-7803-7011-2/01/$10.00 ©2001 IEEE

ABSTRACT

Entropy, Complexity and Chaos in Brain Rhythms

Nitish V. Thakor
The Johns Hopkins School of Medicine Baltimore, USA

The classical approaches to analysis and interpretation of the brain rhythm, namely the EEG, are to employ non-parametric or parametric signal processing methods. These linear systems approaches to brain rhythm analysis have now given way to more advanced methodologies. These methods recognize that the brain rhythms are non-stationary and brain's responses to stimuli are non-linear. While spectral analysis has proved its value in sleep staging analysis, higher order spectral analysis has been useful in determining depth of anesthesia. Complexity analysis has been shown to discriminate neurological disorders such as schizophrenia. Chaotic dynamics have been observed in brain rhythms preceding or resulting from epileptic seizures. The concepts derived from information theory, including measures of entropy, have been useful in characterizing brain injury. Advanced signal processing has long been of interest in application areas such as diagnosis of brain disorders, epilepsy, sleep or anesthesia analysis, and more recently in brain-computer interfaces. An emerging application being developed by our group is monitoring brain's rhythm after neurological trauma or injury. Advanced quantitative analysis, based on the information and entropy analysis methods, has been used by our group to distinguish and characterize the injury response. This presentation will review the state of the art of brain rhythm analysis using the emerging signal processing methods and will especially help theoreticians targeting emergent, significant biomedical applications.

0-7803-7011-2/01/$10.00 ©2001 IEEE

STATISTICAL GNSS CARRIER PHASE AMBIGUITY RESOLUTION: A REVIEW

P.J.G. Teunissen

Department of Mathematical Geodesy and Positioning
Delft University of Technology
Thijsseweg 11
2629 JA Delft, The Netherlands
Fax: ++ 31 15 278 3711

Biography

Dr. Peter Teunissen is professor at the Delft University of Technology and Head of the Department of Mathematical Geodesy and Positioning. He is currently involved in the model development and signal processing of Global Navigation Satellite Systems such as GPS, Glonass and Galileo.

Abstract

Global Navigation Satellite System carrier phase ambiguity resolution is the key to high precision positioning and navigation. In this contribution a brief review is given of the probabilistic theory of integer carrier phase ambiguity estimation. Various ambiguity estimators are discussed. Among them are the estimators of integer rounding, integer bootstrapping, integer least-squares and the Bayesian solution. We also discuss the various relationships that exist between these estimators.

1. INTRODUCTION

Global Navigation Satellite System (GNSS) ambiguity resolution is the process of resolving the unknown cycle ambiguities of double difference (DD) carrier phase data as integers. The sole purpose of ambiguity resolution is to use the integer ambiguity constraints as a means of improving significantly on the precision of the remaining model parameters, such as baseline coordinates and/or atmospheric (troposphere, ionosphere) delays.

Ambiguity resolution applies to a great variety of current and future GNSS models. These models may differ greatly in complexity and diversity. They range from single-baseline models used for kinematic positioning to multi-baseline models used as a tool for studying geodynamic phenomena. The models may or may not have the relative receiver-satellite geometry included. They may also be discriminated as to whether the slave receiver(s) are stationary or in motion. When in motion, one solves for one or more trajectories, since with the receiver-satellite geometry included, one will have new coordinate unknowns for each epoch. One may also discriminate between the models as to whether or not the differential atmospheric delays (ionosphere and troposphere) are included as unknowns. In the case of sufficiently short baselines they are usually excluded.

Apart from the current Global Positioning System (GPS) models, carrier phase ambiguity resolution also applies to the future modernized GPS and the future European Galileo GNSS. An overview of GNSS models, together with their applications in surveying, navigation, geodesy and geophysics, can be found in textbooks such as [*Hofmann-Wellenhof et al.*, 1997], [*Leick*, 1995], [*Parkinson and Spilker*, 1996], [*Strang and Borre*, 1997] and [*Teunissen and Kleusberg*, 1998].

In this contribution we review the probabilistic theory for integer carrier phase ambiguity estimation. It is the key to high precision GNSS positioning and navigation. This contribution is organized as follows. In section 2 we introduce a general class of integer ambiguity estimators, determine their probability mass functions and show how their variability affect the uncertainty in the computed GNSS baselines. This theory is worked out in sections 3 and 4 for two of the most important integer ambiguity estimators. In section 3 we discuss the properties of integer bootstrapping and in section 4 those of integer least-squares. In the final section, section 5, we discuss the Bayesian solution to carrier phase ambiguity resolution. Although the Bayesian approach has not yet find a wide-spread use in any of the GNSS applications, the basic concepts involved are of interest in their own right. Where possible, the various ambiguity estimation principles are compared.

2. INTEGER AMBIGUITY RESOLUTION

2.1. The GNSS model

As our point of departure we will take the following system of linear(ized) observation equations

$$y = Aa + Bb + e \tag{1}$$

where y is the given GNSS data vector of order m, a and b are the unknown parameter vectors respectively of order n and p, and where e is the noise vector. In principle all the GNSS models can be cast in this frame of observation equations. The data vector y will usually consist of the 'observed minus computed' single- or dual- frequency double-difference (DD) phase and/or pseudorange (code) observations accumulated over all observation epochs. The entries of vector a are then the DD carrier phase ambiguities, expressed in units of cycles rather than range. They are known to be *integers*, $a \in Z^n$. The entries of the vector b will consist of the remaining unknown parameters, such as for instance baseline components (coordinates) and possibly atmospheric delay parameters (troposphere, ionosphere). They are known to be real- valued, $b \in R^p$.

The procedure which is usually followed for solving the GNSS model (1), can be divided into three steps. In the *first* step one simply disregards the integer constraints $a \in Z^n$ on the ambiguities

0-7803-7011-2/01/$10.00 ©2001 IEEE

and performs a standard least-squares adjustment. As a result one obtains the (real-valued) estimates of a and b, together with their variance-covariance (vc-) matrix

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}, \begin{bmatrix} Q_{\hat{a}} & Q_{\hat{a}\hat{b}} \\ Q_{\hat{b}\hat{a}} & Q_{\hat{b}} \end{bmatrix} \tag{2}$$

This solution is referred to as the 'float' solution. In the *second* step the 'float' ambiguity estimate $\hat{a}$ is used to compute the corresponding integer ambiguity estimate $\check{a}$. This implies that a mapping $S : R^n \mapsto Z^n$, from the n-dimensional space of reals to the n-dimensional space of integers, is introduced such that

$$\check{a} = S(\hat{a}) \tag{3}$$

Once the integer ambiguities are computed, they are used in the *third* step to finally correct the 'float' estimate of b. As a result one obtains the 'fixed' solution

$$\check{b} = \hat{b} - Q_{\hat{b}\hat{a}} Q_{\hat{a}}^{-1} (\hat{a} - \check{a}) \tag{4}$$

In the present review we will primarily focus our attention on the probabilistic properties of (3) and (4).

2.2. Admissible integer estimation

There are many ways of computing an integer ambiguity vector $\check{a}$ from its real-valued counterpart $\hat{a}$. To each such method belongs a mapping $S : R^n \mapsto Z^n$ from the n-dimensional space of real numbers to the n-dimensional space of integers. Due to the discrete nature of Z^n, the map S will not be one-to-one, but instead a many-to-one map. This implies that different real-valued ambiguity vectors will be mapped to the same integer vector. One can therefore assign a subset $S_z \subset R^n$ to each integer vector $z \in Z^n$:

$$S_z = \{x \in R^n \mid z = S(x)\}, \quad z \in Z^n \tag{5}$$

The subset S_z contains all real-valued ambiguity vectors that will be mapped by S to the same integer vector $z \in Z^n$. This subset is referred to as the *pull-in region* of z [*Jonkman*, 1998]. It is the region in which all ambiguity 'float' solutions are pulled to the same 'fixed' ambiguity vector z. Using the pull-in regions, one can give an explicit expression for the corresponding integer ambiguity estimator. It reads

$$\check{a} = \sum_{z \in Z^n} z s_z(\hat{a}) \tag{6}$$

with the indicator function

$$s_z(\hat{a}) = \begin{cases} 1 & \text{if } \hat{a} \in S_z \\ 0 & \text{otherwise} \end{cases}$$

Since the pull-in regions define the integer estimator completely, one can define classes of integer estimators by imposing various conditions on the pull-in regions. One such class is referred to as the class of admissible integer estimators. These integer estimators are defined as follows.

Definition 1
The integer estimator $\check{a} = \sum_{z \in Z^n} z s_z(\hat{a})$ is said to be *admissible* if

$$\begin{array}{ll} (i) & \bigcup_{z \in Z^n} S_z = R^n \\ (ii) & Int(S_{z_1}) \bigcap Int(S_{z_2}) = \emptyset, \ \forall z_1 \neq z_2 \in Z^n \\ (iii) & S_z = z + S_0, \ \forall z \in Z^n \end{array}$$

This definition is motivated as follows. Each one of the above three conditions describe a property of which it seems reasonable that it is possessed by an arbitrary integer ambiguity estimator. The first condition states that the pull-in regions should not leave any gaps and the second that they should not overlap. The absence of gaps is needed in order to be able to map any 'float' solution $\hat{a} \in R^n$ to Z^n, while the absence of overlaps is needed to guarantee that the 'float' solution is mapped to just one integer vector. Note that we allow the pull-in regions to have common boundaries. This is permitted if we assume to have zero probability that $\hat{a}$ lies on one of the boundaries. This will be the case when the probability density function (pdf) of $\hat{a}$ is continuous.

The third and last condition follows from the requirement that $S(x+z) = S(x) + z, \forall x \in R^n, z \in Z^n$. Also this condition is a reasonable one to ask for. It states that when the 'float' solution is perturbed by $z \in Z^n$, the corresponding integer solution is perturbed by the same amount. This property allows one to apply the *integer remove-restore* technique: $S(\hat{a} - z) + z = S(\hat{a})$. It therefore allows one to work with the fractional parts of the entries of $\hat{a}$, instead of with its complete entries.

With the division of R^n into mutually exclusive pull-in regions, we are now in the position to consider the distribution of $\check{a}$. This distribution is of the *discrete* type and it will be denoted as $P(\check{a} = z)$. It is a probability mass function, having zero masses at nongrid points and nonzero masses at some or all grid points. If we denote the *continuous* probability density function of $\hat{a}$ as $p_{\hat{a}}(x)$, the distribution of $\check{a}$ follows as

$$P(\check{a} = z) = \int_{S_z} p_{\hat{a}}(x) dx \ , \ z \in Z^n \tag{7}$$

This expression holds for any distribution the 'float' ambiguities $\hat{a}$ might have. In most GNSS applications however, one assumes the vector of observables y to be normally distributed. The estimator $\hat{a}$ is therefore normally distributed too, with mean $a \in Z^n$ and vc-matrix $Q_{\hat{a}}$. Its probability density function reads

$$p_{\hat{a}}(x) = \frac{1}{\sqrt{\det(Q_{\hat{a}})}(2\pi)^{\frac{1}{2}n}} \exp\{-\frac{1}{2} \parallel x - a \parallel^2_{Q_{\hat{a}}}\} \tag{8}$$

with the squared weighted norm $\parallel . \parallel^2_{Q_{\hat{a}}} = (.)^T Q_{\hat{a}}^{-1} (.)$. Note that $P(\check{a} = a)$ equals the probability of *correct* integer ambiguity estimation. It describes the expected success rate of GNSS ambiguity resolution.

2.3. The baseline solution

We are now in the position to determine the pdf of the 'fixed' baseline estimator (4). In order to determine this pdf, one needs to propagate the uncertainty of the 'float' solution, $\hat{a}$ and $\hat{b}$, as well as the uncertainty of the integer solution $\check{a}$ through (4). Should one neglect the random character of the integer solution and therefore consider the ambiguity vector $\check{a}$ as deterministic and equal to, say, z, then the pdf of $\check{b}$ would equal the conditional baseline distribution

$$p_{\hat{b}|\hat{a}}(x \mid z) = \frac{\exp\{-\frac{1}{2} \parallel x - b(z) \parallel^2_{Q_{\hat{b}|\hat{a}}}\}}{\sqrt{\det Q_{\hat{b}|\hat{a}}}(2\pi)^{\frac{1}{2}p}} \tag{9}$$

with conditional mean $b(z) = b - Q_{\hat{b}\hat{a}} Q_{\hat{a}}^{-1} (a - z)$, conditional variance matrix $Q_{\hat{b}|\hat{a}} = Q_{\hat{b}} - Q_{\hat{b}\hat{a}} Q_{\hat{a}}^{-1} Q_{\hat{a}\hat{b}}$ and $\parallel . \parallel^2_{Q_{\hat{b}|\hat{a}}} = (.)^T Q_{\hat{b}|\hat{a}}^{-1} (.)$. However, since $\check{a}$ is random and not deterministic, the conditional

baseline distribution will give a too optimistic description of the quality of the 'fixed' baseline. To get a correct description of the 'fixed' baseline's pdf, the integer ambiguity's pmf needs to be considered. As the following theorem shows this results in a baseline distribution, which generally will be multi-modal.

Theorem 1 (*Pdf of the 'fixed' baseline*)
Let the 'float' solution, $\hat{a}$ and $\hat{b}$, be normally distributed with mean $a \in Z^n$ and mean $b \in R^p$, and vc-matrix (2), let $\check{a}$ be an admissible integer estimator and let the 'fixed' baseline $\check{b}$ be given as in (4). The pdf of $\check{b}$ reads then

$$p_{\check{b}}(x) = \sum_{z \in Z^n} p_{\hat{b}|\hat{a}}(x \mid z) P(\check{a} = z) \tag{10}$$

Note that, although the model (1) is linear and the observables normally distributed, the distribution of the 'fixed' baseline is not normal, but multi-modal. As the theorem shows, the 'fixed' baseline distribution equals an infinite sum of weighted conditional baseline distributions. These conditional baseline distributions $p_{\hat{b}|\hat{a}}(x \mid z)$ are shifted versions of one another. The size and direction of the shift is governed by $Q_{\hat{b}\hat{a}} Q_{\hat{a}}^{-1} z$, $z \in Z^n$. Each of the conditional baseline distributions in the infinite sum is downweighted. These weights are given by the probability masses of the distribution of the integer bootstrapped ambiguity estimator $\check{a}$. This shows that the dependence of the 'fixed' baseline distribution on the choice of integer estimator is only felt through the weights $P(\check{a} = z)$.

2.4. On the quality of the 'fixed' baseline

In order to describe the quality of the 'fixed' baseline, one would like to know how close one can expect the baseline estimate $\check{b}$ to be to the unknown, but true baseline value b. As a measure of confidence, we take

$$P(\check{b} \in R) = \int_R p_{\check{b}}(x) dx \text{ with } R \subset R^p \tag{11}$$

But in order to evaluate this integral, we first need to make a choice about the shape and location of the subset R. Since it is common practice in GNSS positioning to use the vc-matrix of the conditional baseline estimator as a measure of precision for the 'fixed' baseline, the vc-matrix $Q_{\hat{b}|\hat{a}}$ will be used to define the shape of the confidence region. For its location, we choose the confidence region to be centered at b. After all, we would like to know by how much the baseline estimate $\check{b}$ can be expected to differ from the true, but unknown baseline value b. That is, one would like (11) to be a measure of the baseline's probability of concentration about b.

With these choices on shape and location, the region R takes the form

$$R = \{x \in R^p \mid (x-b)^T Q_{\hat{b}|\hat{a}}^{-1} (x-b) \leq \beta^2\} \tag{12}$$

The size of the region can be varied by varying β. The following theorem shows how the baseline's probability of concentration (11) can be evaluated as a weighted sum of probabilities of noncentral Chi-square distributions.

Theorem 2 (*The 'fixed' baseline's probability of concentration*)
Let $\check{b}$ be the 'fixed' baseline estimator, let R be defined as in (12), and let $\chi^2(p, \lambda_z)$ denote the noncentral Chi-square distribution with p degrees of freedom and noncentrality parameter λ_z. Then

$$P(\check{b} \in R) = \sum_{z \in Z^n} P(\chi^2(p, \lambda_z) \leq \beta^2) P(\check{a} = z) \tag{13}$$

with

$$\lambda_z = \| \nabla \check{b}_z \|^2_{Q_{\hat{b}|\hat{a}}} \text{ and } \nabla \check{b}_z = Q_{\hat{b}\hat{a}} Q_{\hat{a}}^{-1} (z - a)$$

This result shows that the probability of the 'fixed' baseline lying inside the ellipsoidal region R centered at b equals an infinite sum of probability products. If one considers the two probabilities of these products separately, two effects are observed. First the probabilistic effect of shifting the conditional baseline estimator away from b and secondly the probabilistic effect of the peakedness or nonpeakedness of the ambiguity pmf. The second effect is related to the expected performance of ambiguity resolution, while the first effect has to do with the sensitivity of the baseline for changes in the values of the integer ambiguities. This effect is measured by the noncentrality parameter λ_z. Since the tail of a noncentral Chi-square distribution becomes heavier when the noncentrality parameter increases, while the degrees of freedom remain fixed, $P(\chi^2(p, \lambda_z) \leq \beta^2)$ gets smaller when λ_z gets larger.

The two probabilities in the product reach their maximum values when $z = a$. The following corollary shows how these two maxima can be used to lower bound and to upper bound the probability $P(\check{b} \in R)$. Such bounds are of importance for practical purposes, since it is difficult in general to evaluate (13) exactly.

Corollary 1 (*Lower and upper bounds*)
Let $\check{b}$ be the 'fixed' baseline estimator and let R be defined as in (12). Then

$$P(\hat{b}_{|\hat{a}=a} \in R) P(\check{a} = a) \leq P(\check{b} \in R) \leq P(\hat{b}_{|\hat{a}=a} \in R) \tag{14}$$

with

$$P(\hat{b}_{|\hat{a}=a} \in R) = P(\chi^2(p, 0) \leq \beta^2)$$

Note that the two bounds relate the probability of the 'fixed' baseline estimator to that of the conditional estimator and the bootstrapped succes rate. The above bounds become tight when the ambiguity success rate approaches one. This shows, although the probability of the conditional estimator always overestimates the probability of the 'fixed' baseline estimator, that the two probabilities are close for large values of the success rate. This implies that in case of GNSS ambiguity resolution, one should first evaluate the success rate $P(\check{a} = a)$ and make sure that its value is close enough to one, before making any inferences on the basis of the distribution of the conditional baseline estimator. In other words, the (unimodal) distribution of the conditional estimator is a good approximation to the (multimodal) distribution of the bootstrapped baseline estimator, when the success rate is sufficiently close to one.

3. INTEGER BOOTSTRAPPING

3.1. The bootstrapped estimator

The distributional results presented so far hold for any admissible ambiguity estimator. The simplest way to obtain an integer vector from the real-valued 'float' solution is to round each of the entries

of $\hat{a}$ to its nearest integer. The corresponding integer estimator reads therefore

$$\breve{a}_R = ([\hat{a}_1], \ldots, [\hat{a}_n])^T \tag{15}$$

where '[.]' denotes rounding to the nearest integer. The pull-in region of this integer estimator equals the multivariate version of the unit-square.

Another relatively simple integer ambiguity estimator is the bootstrapped estimator. The bootstrapped estimator can be seen as a generalization of the previous estimator. It still makes use of integer rounding, but it also takes some of the correlation between the ambiguities into account. The bootstrapped estimator follows from a sequential conditional least- squares adjustment and it is computed as follows. If n ambiguities are available, one starts with the first ambiguity $\hat{a}_1$, and rounds its value to the nearest integer. Having obtained the integer value of this first ambiguity, the real-valued estimates of all remaining ambiguities are then corrected by virtue of their correlation with the first ambiguity. Then the second, but now corrected, real-valued ambiguity estimate is rounded to its nearest integer. Having obtained the integer value of the second ambiguity, the real-valued estimates of all remaining $n-2$ ambiguities are then again corrected, but now by virtue of their correlation with the second ambiguity. This process is continued until all ambiguities are considered. We thus have the following definition.

Definition 2 (*Integer bootstrapping*)
Let $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_n)^T \in R^n$ be the ambiguity 'float' solution and let $\breve{a}_B = (\breve{a}_{B,1}, \ldots, \breve{a}_{B,n})^T \in Z^n$ denote the corresponding integer bootstrapped solution. The entries of the bootstrapped ambiguity estimator are then defined as

$$\begin{aligned}
\breve{a}_{B,1} &= [\hat{a}_1] \\
\breve{a}_{B,2} &= [\hat{a}_{2|1}] = [\hat{a}_2 - \sigma_{21}\sigma_1^{-2}(\hat{a}_1 - \breve{a}_{B,1})] \\
&\vdots \\
\breve{a}_{B,n} &= [\hat{a}_{n|N}] = \\
&\quad [\hat{a}_n - \textstyle\sum_{j=1}^{n-1} \sigma_{n,j|J}\sigma_{j|J}^{-2}(\hat{a}_{j|J} - \breve{a}_{B,j})]
\end{aligned} \tag{16}$$

where $'[.]'$ denotes the operation of rounding to the nearest integer, and $\sigma_{i,j|J}$ denotes the covariance between $\hat{a}_i$ and $\hat{a}_{j|J}$, and $\sigma^2_{j|J}$ is the variance of $\hat{a}_{j|J}$. The shorthand notation $\hat{a}_{i|I}$ stands for the ith least-squares ambiguity obtained through a conditioning on the previous $I = \{1, \ldots, (i-1)\}$ sequentially rounded ambiguities.

Note that the bootstrapped estimator is not unique. Changing the order in which the ambiguities appear in vector $\hat{a}$ will already produce a different bootstrapped estimator. Although the principle of bootstrapping remains the same, every choice of ambiguity parametrization has its own bootstrapped estimator.

3.2. The bootstrapped pull-in regions

The pull-in regions for rounding are unit-cubes centred at integer grid points. For bootstrapping the shape of the pull-in regions will depend on the vc-matrix of the ambiguities. They will coincide with the unit- cubes only in case the vc-matrix is a diagonal matrix. Bootstrapping reduces namely to rounding in the absence of any correlation between the ambiguities. The following theorem gives a description of the bootstrapped pull-in regions in the general case.

Theorem 3 (*Bootstrapped pull-in regions*)
The pull-in regions of the bootstrapped ambiguity estimator $\breve{a}_B = (\breve{a}_{B,1}, \ldots, \breve{a}_{B,n})^T \in Z^n$ are given as

$$S_{B,z} = \{x \in R^n \,|\, |\, c_i^T L^{-1}(x-z) \,| \leq \frac{1}{2}\,,\ i = 1-n\} \tag{17}$$

$\forall z \in Z^n$ where L denotes the unique unit lower triangular matrix of the ambiguity vc-matrix' decomposition $Q_{\hat{a}} = LDL^T$ and c_i denotes the ith canonical unit vector having a 1 as its ith entry and zeros otherwise.

That the bootstrapped estimator is indeed admissible, can now be seen as follows. The first two conditions of Definition 1 are easily verified using the definition of the bootstrapped estimator. Since every real-valued vector $\hat{a}$ will be mapped by the bootstrapped estimator to an integer vector, the pull-in regions $S_{B,z}$ cover R^n without any gaps. There is also no overlap between the pull-in regions, since - apart from boundary ties - any real-valued vector $\hat{a}$ is mapped to not more than one integer vector. To verify the last condition of Definition 1, we make use of (17). From

$$\begin{aligned}
&S_{B,z} = \\
&\{x \in R^n \,|\, |\, c_i^T L^{-1}(x-z) \,| \leq \tfrac{1}{2}\,,\ i = 1,n\} = \\
&\{x \in R^n \,|\, |\, c_i^T L^{-1} y \,| \leq \tfrac{1}{2}\,,\ x = y+z\,,\ i = 1,n\} = \\
&S_{B,0} + z
\end{aligned}$$

it follows that all bootstrapped pull-in regions are translated copies of $S_{B,0}$. All pull-in regions have therefore the same shape and the same volume. Their volumes all equal 1. This can be shown by transforming $S_{B,0}$ to the unit cube centered at the origin. Consider the linear transformation $y = L^{-1}x$. Then

$$L^{-1}(S_{B,0}) = \{y \in R^n \,|\, |\, c_i^T y \,| \leq \frac{1}{2}\,,\ i = 1, \ldots, n\}$$

equals the unit cube centered at the origin. Since the determinant of the unit lower triangular matrix L^{-1} equals one and since the volume of the unit cube equals one, it follows that the volume of $S_{B,0}$ must equal one as well. To infer the shape of the bootstrapped pull-in region, we consider the two-dimensional case first. Let the lower triangular matrix L be given as

$$L = \begin{bmatrix} 1 & 0 \\ l & 1 \end{bmatrix}$$

Then

$$\begin{aligned}
S_{B,0} &= \{x \in R^2 \,|\, |\, c_i^T L^{-1} x \,| \leq \tfrac{1}{2}\,,\ i = 1,2\} \\
&= \{x \in R^2 \,|\, |x_1| \leq \tfrac{1}{2}, |x_2 - l x_1| \leq \tfrac{1}{2}\}
\end{aligned}$$

which shows that the two-dimensional pull-in region equals a parallellogram. Its region is bounded by the two vertical lines $x_1 = 1/2$ and $x_1 = -1/2$, and the two parallel slopes $x_2 = lx_1 + 1/2$ and $x_2 = lx_1 - 1/2$. The direction of the slope is governed by $l = \sigma_{21}\sigma_1^{-2}$. Hence, in the absence of correlation between the two ambiguities, the parallellogram reduces to the unit square. In higher dimensions the above construction of the pull-in region can be continued. In three dimensions for instance, the intersection of the pull-in region with the x_1x_2-plane remains a parallellogram, while along the third axis the pull-in region becomes bounded by two parallel planes.

3.3. The bootstrapped pmf

Since the integer bootstrapped estimator is defined as $\breve{a}_B = z \Longleftrightarrow \hat{a} \in S_{B,z}$, it follows that $P(\breve{a}_B = z) = P(\hat{a} \in S_{B,z})$. The pmf of $\breve{a}_B$ follows therefore as

$$P(\breve{a}_B = z) = \int_{S_{B,z}} p_{\hat{a}}(x)dx \, , \, z \in Z^n \tag{18}$$

Hence, the probability that $\breve{a}_B$ coincides with z is given by the integral of the pdf $p_{\hat{a}}(x)$ over the bootstrapped pull-in region $S_{B,z} \subset R^n$. The above expression holds for any distribution the 'float' ambiguities $\hat{a}$ might have. In most GNSS applications however, one usually assumes the vector of observables y to be normally distributed. For that case the following theorem gives an exact expression of the bootstrapped pmf.

Theorem 4 (*The integer bootstrapped pmf*)
Let $\hat{a}$ be distributed as $N(a, Q_{\hat{a}})$, $a \in Z^n$, and let $\breve{a}_B$ be the corresponding integer bootstrapped estimator. Then

$$\begin{aligned} P(\breve{a}_B = z) &= \prod_{i=1}^{n} [\Phi(\tfrac{1-2l_i^T(a-z)}{2\sigma_{\hat{a}_{i|I}}}) \\ &+ \Phi(\tfrac{1+2l_i^T(a-z)}{2\sigma_{\hat{a}_{i|I}}}) - 1] \, , \, z \in Z^n \end{aligned} \tag{19}$$

with

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}v^2\}dv$$

and with l_i the ith column vector of the unit lower triangular matrix L^{-T} and $\sigma^2_{\hat{a}_{i|I}}$ the variance of the ith least- squares ambiguity obtained through a conditioning on the previous $I = \{1, \ldots, (i-1)\}$ ambiguities.

The bootstrapped pmf equals a product of univariate pmf's and is therefore easy to compute. Note that the bootstrapped pmf is completely governed by the ambiguity vc-matrix $Q_{\hat{a}}$. The pmf follows once the triangular factor L and the diagonal matrix D of the decomposition $Q_{\hat{a}} = LDL^T$ are given. The above result also shows that the bootstrapped pmf is symmetric about the mean of $\hat{a}$. This implies that the bootstrapped estimator $\breve{a}_B$ is an unbiased estimator of $a \in Z^n$. Since the 'float' solutions, $\hat{a}$ and $\hat{b}$, are unbiased too, it follows from taking the expectation of (4) that the bootstrapped baseline is also unbiased.

For the purpose of predicting the success of ambiguity resolution, the probability of correct integer estimation is of particular interest. For the bootstrapped estimator this success rate is given in the following corollary.

Corollary 2 (*The bootstrapped success rate*)
The bootstrapped probability of correct integer estimation (the success rate) is given as

$$P(\breve{a}_B = a) = \prod_{i=1}^{n} [2\Phi(\frac{1}{2\sigma_{\hat{a}_{i|I}}}) - 1] \tag{20}$$

The method of integer bootstrapping is easy to implement and it does not need, as opposed to the method of integer least-squares (see next section), an integer search for computing the sought for integer solution. However, as it was mentioned earlier, the outcome of bootstrapping depends on the chosen ambiguity parametrization. Bootstrapping of DD ambiguities, for instance, will produce an integer solution which generally differs from the integer solution obtained from bootstrapping of reparametrized ambiguities. Since this dependency also holds true for the bootstrapped pmf, one still has some important degrees of freedom left for improving (20) or for sharpening the lower bound of (14).

In order to improve the bootstrapped success rate, one should work with decorrelated ambiguities instead of with the original ambiguities. The method of bootstrapping performs relatively poor, for instance, when applied to the DD ambiguities. This is due to the usually high correlation between the DD ambiguities. Bootstrapping should therefore only be used in combination with the decorrelating Z-transformation of the LAMBDA method [*Teunissen*, 1993, 1995]. This transformation decorrelates the ambiguities further than the best reordering would achieve and thereby reduces the values of the sequential conditional variances. By reducing the values of the sequential conditional variances, the bootstrapped success rate gets enlarged.

It may however happen that it is simply not possible to resolve the complete vector of ambiguities with sufficient probability. As an alternative of resolving the complete vector of ambiguities, one might then consider resolving only a subset of the ambiguities. The idea of partial ambiguity resolution is based on the fact that the success rate will generally increase when fewer integer constraints are imposed. However, in order to apply partial ambiguity resolution, one first will have to determine which subset of ambiguities to choose. It will be clear that this decision should be based on the precision of the 'float' ambiguities. The more precise the ambiguities, the larger the ambiguity success rate. It is at this point where the decorrelation step of the LAMBDA method and the bootstrapping principle can be applied. Once the transformed and decorrelated ambiguity vc-matrix is obtained, the construction of the subset proceeds in a sequential fashion. One first starts with the most precise ambiguity, say $\hat{z}_1$, and computes its success rate $P(\breve{a}_1 = z_1)$. If this success rate is large enough, one continues and determines the most precise pair of ambiguities, say $(\hat{z}_1, \hat{z}_2)$. If their success rate is still large enough, one continues again by trying to extend the set. This procedure continues until one reaches a point where the corresponding success rate becomes unacceptably small. When this point is reached, one can expect that the previously identified ambiguities can be resolved successfully.

Once the subset for partial ambiguity resolution has been identified, one still needs to determine what this will do to improve the baseline estimator. After all, being able to successfully resolve the ambiguities does not necessarily mean that the 'fixed' solution is significantly better than the 'float' solution. The theory presented in the previous sections provide the necessary tools for performing such an evaluation.

4. INTEGER LEAST-SQUARES

4.1. The ILS estimator

In this section we review some integer least-squares' theory for solving the GNSS model (1). When using the least-squares principle, the GNSS model can be solved by means of the minimization problem

$$\min_{a,b} \| y - Aa - Bb \|^2_{Q_y} \, , a \in Z^n \, , b \in R^p \tag{21}$$

with Q_y the vc-matrix of the GNSS observables. This type of least-squares problem was first introduced in [*Teunissen*, 1993] and has been coined with the term '*integer least-squares*'. It is a nonstandard least-squares problem due to the integer constraints $a \in Z^n$.

The solution of (21) is consistent with the three solution steps of section 1. This can be seen as follows. It follows from the orthogonal decomposition

$$\begin{aligned} \| y-Aa-Bb \|^2_{Q_y} &= \| \hat{e} \|^2_{Q_y} + \| \hat{a}-a \|^2_{Q_{\hat{a}}} \\ &+ \| \hat{b}(a)-b \|^2_{Q_{\hat{b}|\hat{a}}} \end{aligned} \quad (22)$$

with $\hat{e} = y - A\hat{a} - B\hat{b}$ and $\hat{b}(a) = \hat{b} - Q_{\hat{b}\hat{a}}Q_{\hat{a}}^{-1}(\hat{a}-a)$, that the sought for minimum is obtained when the second term on the right-hand side is minimized for $a \in Z^n$ and the last term is set to zero. The integer least-squares (ILS) estimator of the ambiguities is therefore defined as follows.

Definition 3 (*Integer least-squares*)
Let $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_n)^T \in R^n$ be the ambiguity 'float' solution and let $\breve{a}_{LS} \in Z^n$ denote the corresponding integer least-squares solution. Then

$$\breve{a}_{LS} = \arg\min_{z \in Z^n} \| \hat{a}-z \|^2_{Q_{\hat{a}}} \quad (23)$$

In contrast to integer rounding and integer bootstrapping, an integer search is needed to compute $\breve{a}_{LS}$. Although we will refrain from discussing the computational intricacies of ILS estimation, the conceptual steps of the computational procedure will be described briefly. The ILS procedure is mechanized in the GNSS LAMBDA (Least-squares AMBiguity Decorrelation Adjustment) method, which is currently one of the most applied methods for GNSS carrier phase ambiguity resolution. For more information on the LAMBDA method, we refer to e.g. [*Teunissen*, 1993], [*Teunissen*, 1995] and [*de Jonge and Tiberius*, 1996a] or to the textbooks [*Hofmann-Wellenhof*, 1997], [*Strang and Borre*, 1997], [*Teunissen and Kleusberg*, 1998]. Practical results obtained with it can be found, for example, in [*Boon and Ambrosius*, 1997], [*Boon et al.*, 1997], [*Cox and Brading*, 1999], [*de Jonge and Tiberius*, 1996b], [*de Jonge et al.*, 1996], [*Han*, 1995], [*Jonkman*, 1998], [*Peng et al.*, 1999], [*Tiberius and de Jonge*, 1995], [*Tiberius et al.*, 1997].

The main steps as implemented in the LAMBDA method are as follows. One starts by defining the ambiguity search space

$$\Omega_a = \{a \in Z^n \mid (\hat{a}-a)^T Q_{\hat{a}}^{-1}(\hat{a}-a) \leq \chi^2\} \quad (24)$$

with χ^2 a to be chosen positive constant. The boundary of this search space is ellipsoidal. It is centred at $\hat{a}$, its shape is governed by the vc-matrix $Q_{\hat{a}}$ and its size is determined by χ^2. In case of GNSS, the search space is usually extremely elongated, due to the high correlations between the ambiguities. Since this extreme elongation usually hinders the computational efficiency of the search, the search space is first transformed to a more spherical shape,

$$\Omega_z = \{z \in Z^n \mid (\hat{z}-z)^T Q_{\hat{z}}^{-1}(\hat{z}-z) \leq \chi^2\} \quad (25)$$

using the admissible ambiguity transformations $\hat{z} = Z^T\hat{a}$, $Q_{\hat{z}} = Z^T Q_{\hat{a}} Z$. Ambiguity transformations Z are said to be admissible when both Z and its inverse Z^{-1} have integer entries. Such matrices preserve the integer nature of the ambiguities. In order for the transformed search space to become more spherical, the volume-preserving Z-transformation is constructed as a transformation that decorrelates the ambiguities as much as possible. Using the triangular decomposition of $Q_{\hat{z}}$, the left-hand side of the quadratic inequality in (25) is then written as a sum-of-squares:

$$\sum_{i=1}^{n} \frac{(\hat{z}_{i|I} - z_i)^2}{\sigma^2_{i|I}} \leq \chi^2 \quad (26)$$

On the left-hand side one recognizes the conditional least-squares estimator $\hat{z}_{i|I}$, which follows when the conditioning takes place on the integers $z_1, z_2, \ldots, z_{i-1}$. Using the sum-of-squares structure, one can finally set up the n intervals which are used for the search. These sequential intervals are given as

$$\begin{aligned} (\hat{z}_1 - z_1)^2 &\leq \sigma^2_1 \chi^2 \\ (\hat{z}_{2|1} - z_2)^2 &\leq \sigma^2_{2|1}\left(\chi^2 - \tfrac{(\hat{z}_1 - z_1)^2}{\sigma^2_1}\right) \\ &\vdots \end{aligned} \quad (27)$$

In order for the search to be efficient, one not only would like the vc- matrix $Q_{\hat{z}}$ to be as close as possible to a diagonal matrix, but also that the search space does not contain too many integer grid points. This requires the choice of a small value for χ^2, but one that still guarantees that the search space contains at least one integer grid point. Since the bootstrapped estimator is so easy to compute and at the same time gives a good approximation to the ILS estimator (see section 4.4), the bootstrapped solution is an excellent candidate for setting the size of the ambiguity search space. Following the decorrelation step $\hat{z} = Z^T\hat{a}$, the LAMBDA-method therefore uses, as one of its options, the bootstrapped solution $\breve{z}_B$ for setting the size of the ambiguity search space as

$$\chi^2 = (\hat{z}-\breve{z}_B)^T Q_{\hat{z}}^{-1}(\hat{z}-\breve{z}_B) \quad (28)$$

In this way one can work with a very small search space and still guarantee that the sought for integer least-squares solution is contained in it.

4.2. The ILS pull-in region

The pull-in regions of integer rounding are unit cubes, while those of integer bootstrapping are multivariate versions of parallellograms. To determine the ILS pull-in regions we need to know the set of 'float' solutions $\hat{a} \in R^n$ that are mapped to the same integer vector $z \in Z^n$. This set is described by all $x \in R^n$ that satify $z = \arg\min_{u \in Z^n} \| x-u \|^2_{Q_{\hat{a}}}$. The ILS pull-in-region that belongs to the integer vector z follows therefore as

$$S_{LS,z} = \{x \in R^n \mid \| x-z \|^2_{Q_{\hat{a}}} \leq \| x-u \|^2_{Q_{\hat{a}}}, \forall u \in Z^n\} \quad (29)$$

It consists of all those points which are closer to z than to any other integer point in R^n. The metric used for measuring these distances is determined by the vc-matrix $Q_{\hat{a}}$. Based on (29), one can give a representation of the ILS pull-in regions that resembles the representation of the bootstrapped pull-in regions. This representation reads as follows.

Theorem 5 (*ILS pull-in regions*)
The pull-in regions of the ILS ambiguity estimator $\breve{a}_{LS} \in Z^n$ are given as

$$\begin{aligned} &S_{LS,z} = \\ &\cap_{c_i \in Z^n}\{x \in R^n \mid \mid c_i^T Q_{\hat{a}}^{-1}(x-z) \mid \leq \tfrac{1}{2} \| c_i \|^2_{Q_{\hat{a}}}\}, \\ &\forall z \in Z^n \end{aligned} \quad (30)$$

This shows that the ILS pull-in regions are constructed from intersecting half-spaces. One can also show that at most $2^n - 1$ pairs of such half spaces are needed for constructing the pull-in region.

The ILS pull-in regions are convex, symmetric sets of volume 1, which satisfy the conditions of Definition 1. The ILS estimator is therefore admissible. The ILS pull-in regions are hexagons in the two-dimensional case.

4.3. Maximizing the success-rate

Although various integer estimators exist which are admissible, some may be better than others. Having the problem of GNSS ambiguity resolution in mind, one is particularly interested in the estimator which maximizes the probability of correct integer estimation. This probability equals $P(\check{a} = a)$, but it will differ for different ambiguity estimators. The following theorem shows that the ILS estimator maximizes the probability of correct integer estimation.

Theorem 6 (*ILS is optimal*)
Let the pdf of the 'float' solution $\hat{a}$ be given as

$$p_a(x) = \sqrt{\det(Q_{\hat{a}}^{-1})}G(\| x-a \|^2_{Q_{\hat{a}}}) \tag{31}$$

where $G : R \mapsto [0,\infty)$ is decreasing and $Q_{\hat{a}}$ is positive-definite. Then

$$P(\check{a}_{LS} = a) \geq P(\check{a} = a) \tag{32}$$

for any admissible estimator $\check{a}$.

This theorem gives a probabilistic justification for using the ILS estimator. For GNSS ambiguity resolution it shows, that one is better off using the ILS estimator than any other admissible integer estimator. The family of distributions defined in (31), is known as the family of elliptically contoured distributions. Several important distributions belong to this family. The multivariate normal distribution can be shown to be a member of this family by choosing $G(x) = (2\pi)^{-\frac{n}{2}} \exp -\frac{1}{2}x, x \in R$. Another member is the multivariate t-distribution.

As a direct consequence of the above theorem we have the following corollary.

Corollary 3 (*The effect of the weight matrix*)
Let Σ be any positive-definite matrix of order n and define

$$\check{a}_\Sigma = \arg\min_{z \in Z^n} \| \hat{a} - z \|^2_\Sigma \tag{33}$$

Then $\check{a}_\Sigma$ is admissible and

$$P(\check{a}_{LS} = a) \geq P(\check{a}_\Sigma = a) \tag{34}$$

In order to prove the corollary, we only need to show that $\check{a}_\Sigma$ is admissible. Once this has been established, the stated result (34) follows from theorem 6. The admissibility can be shown as follows. The first two conditions of Definition 1 are satisfied, since the ILS-map produces - apart from boundary ties - a unique integer vector for any 'float' solution $\hat{a} \in R^n$. And since $\check{a}_\Sigma = \arg\min_{z \in Z^n} \| \hat{a} - u - z \|^2_\Sigma + u$ holds true for any integer $u \in Z^n$, also the integer remove-restore technique applies.

As the corollary shows, a proper choice of the data weight matrix is also of importance for ambiguity resolution. The choice of weights is optimal when the weight matrix equals the inverse of the ambiguity vc-matrix. A too optimistic precision description or a too pessimistic precision description, will both result in a less than optimal ambiguity success rate. In the case of GNSS, the observation equations (the functional model) are sufficiently known and well documented. However, the same can not yet be said of the vc-matrix of the GNSS data. In the many GNSS textbooks available, we will usually find only a few comments, if any, on this vc-matrix. Examples of studies that have been reported in the literature are: [*Euler and Goad*, 1991], [*Gerdan*, 1995], [*Gianniou*, 1996], and [*Jin and de Jong*, 1996], who studied the elevation dependence of the observation variances; [*Jonkman*, 1998] and [*Tiberius*, 1998], who considered time correlation and cross correlation; and [*Schaffrin and Bock*, 1988], [*Bock*, 1998] and [*Teunissen*, 1998a], who considered the inclusion of stochastic ionospheric constraints.

4.4. Bounding the ILS success-rate

A very useful application of theorem 6 is that it shows how one can *lower-bound* the ILS probability of correct integer estimation. This is particularly useful since the ILS success rate is usually difficult to compute. This is due to the rather complicated geometry of the ILS pull-in region. The bootstrapped success-rate is a good candidate for the ILS success-rates' lower-bound. The bootstrapped success-rate is easy to compute and it becomes a sharp lower-bound when applied to the decorrelated ambiguities $\hat{z} = Z^T \hat{a}$. In fact, at present, the bootstrapped success-rate is the sharpest available lower-bound of the ILS success-rate.

Apart from having a lower-bound, it is also useful to have an upper-bound available. For obtaining an upper-bound one can make use of the *geometric mean* of the ambiguity conditional variances. This geometric mean is referred to as the Ambiguity Dilution of Precision (ADOP) and it is given as

$$\text{ADOP} = \sqrt{\det Q_{\hat{a}}}^{\frac{1}{n}} \text{ (cycles)} \tag{35}$$

Note that this scalar measure of the ambiguity precision is invariant for the admissible volume preserving ambiguity transformations. With the ADOP one can obtain an upper-bound by making use of the fact that the probability content of the ILS pull-in region $S_{LS,a}$ would be maximal if its shape would coincide with that of the ambiguity search space, while its volume would still be constrained to 1. We have the following bounds for the ILS success-rate.

Theorem 7 (*Bounds on the ILS success-rate*)
The ILS success-rate $P(\check{a}_{LS} = a)$ is bounded from below and from above as

$$P(\check{z}_B = z) \leq P(\check{a}_{LS} = a) \leq P\left(\chi^2(n,0) \leq \frac{c_n}{\text{ADOP}^2}\right) \quad \text{with } c_n = (\tfrac{n}{2}\Gamma(\tfrac{n}{2}))^{2/n}/\pi \tag{36}$$

5. A BAYESIAN APPROACH

5.1. The Bayes estimate

The Bayesian approach to GNSS carrier phase ambiguity resolution starts from a set of assumptions which differs fundamentally from the one used in the previous sections, see e.g.[*Betti et al.*, 1993], [*Gundlich and Koch*, 2001]. In the Bayesian approach, not only the vector of observables, y, is assumed to be random, but the vector of unknown parameters, a and b, as well. Although the Bayesian approach has not yet find a wide-spread use in any of the GNSS applications, the basic concepts involved are of interest in their own right, also in their comparison with the nonBayesian theory of the previous sections.

Let us for the moment take the two type of parameter vectors a and b together in one vector $x = (a^T, b^T)^T$. If both y and x are random, we have according to Bayes' theorem

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} \tag{37}$$

Thus the posterior density $p(x \mid y)$ is proportional to the product of the likelihood function $p(y \mid x)$ and the prior density $p(x)$. Given the data vector y, that is, given the observations, the posterior density gives a complete description of the probabilistic properties of x. The idea of the Bayesian approach is therefore to use the posterior density for parameter estimation.

In the Bayesian approach to ambiguity resolution it is the so-called *Bayes estimate* which is used as the solution for the ambiguities and baseline. This estimate is defined as follows.

Definition 4 (*The Bayes estimate*)
The Bayes estimate $\hat{x}_{Bayes}$ of the random parameter vector x is defined as the conditional mean

$$\hat{x}_{Bayes} = \mathrm{E}\{x \mid y\} = \int x p(x \mid y) dx \tag{38}$$

This definition can be motivated as follows. In order to find a 'good' estimate $\hat{x}$ of the parameter vector x, we would like to determine a function of the data, say $\hat{x} = \hat{x}(y)$, which in a certain sense is close to x. Let $L(x,\hat{x}(y))$ be our measure of discrepancy, or our measure of loss, between x and $\hat{x}$. It then seems reasonable to take $\hat{x}$ as the solution which minimizes this discrepancy on the average. This amounts to solving the minimization problem

$$\min_{\hat{x}} \mathrm{E}\{L(x,\hat{x}(y)) \mid y\} = \min_{\hat{x}} \int L(x,\hat{x}) p(x \mid y) dx \tag{39}$$

This minimization problem is particularly easy to solve in case the loss function equals the quadratic form, $L(x,\hat{x}) = \| x - \hat{x} \|_Q^2$, with matrix Q being positive definite. From the decomposition

$$\begin{aligned} &\mathrm{E}\{L(x,\hat{x}(y)) \mid y\} = \\ &\int \| x - \hat{x} \|_Q^2 \, p(x \mid y) dx = \\ &\int \| x - \mathrm{E}\{x \mid y\} \|_Q^2 \, p(x \mid y) dx + \| \mathrm{E}\{x \mid y\} - \hat{x} \|_Q^2 \end{aligned}$$

it directly follows that the posterior expected loss is minimized when $\hat{x}$ is taken equal to the Bayes estimate. When the Bayes estimate is substituted into the loss function, the expected loss equals $\mathrm{E}\{L(x,\hat{x}_{Bayes})\} = \mathrm{trace}(Q_{x|y}Q^{-1})$.

5.2. The marginal posterior pdf's

In order to apply (38) to our ambiguity resolution problem, we first need an expression for the posterior density $p(x \mid y) = p(a,b \mid y)$. In the Bayesian approach to GNSS ambiguity resolution, a and b are assumed to be independent, with the following improper priors

$$\begin{cases} p(a) & \propto & \sum_{z \in Z^n} \delta(a - z) \ \text{(pulsetrain)} \\ p(b) & \propto & \textit{constant} \end{cases} \tag{40}$$

where δ denotes the Dirac function. From the orthogonal decomposition (22), the likelihood function can be seen to be proportional to $p(y \mid a,b) \propto \exp -\frac{1}{2}\{\| \hat{a} - a \|_{Q_{\hat{a}}}^2 + \| \hat{b}(a) - b \|_{Q_{\hat{b}|\hat{a}}}^2\}$. The posterior density follows therefore as

$$\begin{aligned} p(a,b \mid y) \propto &\exp -\tfrac{1}{2}\{\| \hat{a} - a \|_{Q_{\hat{a}}}^2 \\ &+ \| \hat{b}(a) - b \|_{Q_{\hat{b}|\hat{a}}}^2\} \sum_{z \in Z^n} \delta(a - z) \end{aligned} \tag{41}$$

The required marginal posterior densities, $p(a \mid y)$ and $p(b \mid y)$, follow from integrating the joint posterior density over the domains of respectively a and b. Note that in the present case, the domain of a is taken as R^n and not as Z^n. In the Bayesian approach, the discrete-like nature of a is thought to be captured by assuming the prior to be a pulsetrain. Once the integrations are carried out and the normalizing constants are restored, the marginals are obtained as follows.

Theorem 8 (*Marginal posterior pdf's*)
The posterior pdf's of the ambiguities and baseline are given as

$$\begin{cases} p(a \mid y) & = & w_a(\hat{a}) \sum_{z \in Z^n} \delta(a - z) \\ p(b \mid y) & = & \sum_{z \in Z^n} p_{b|a}(b \mid a = z, y) w_z(\hat{a}) \end{cases} \tag{42}$$

with the weight function

$$w_z(\hat{a}) = \frac{\exp -\frac{1}{2}\{\| \hat{a} - z \|_{Q_{\hat{a}}}^2\}}{\sum_{u \in Z^n} \exp -\frac{1}{2}\{\| \hat{a} - u \|_{Q_{\hat{a}}}^2\}}, \ z \in Z^n \tag{43}$$

and the conditional posterior

$$p_{b|a}(b \mid a,y) = \frac{1}{\sqrt{\det Q_{\hat{b}|\hat{a}}}(2\pi)^{\frac{1}{2}p}} \exp -\frac{1}{2} \| b - \hat{b}(a) \|_{Q_{\hat{b}|\hat{a}}}^2 \tag{44}$$

It is now interesting to observe how the above posterior marginal pdf for the baseline, $p(b \mid y)$, compares with the pdf of the 'fixed' baseline, $p_{\check{b}}(x)$, as given in (10). Both pdf's are very similar in structure. Both equal an infinite sum of weighted conditional baseline distributions. The two type of conditional baseline distributions, $p_{\hat{b}|\hat{a}}(x \mid z)$ and $p_{b|a}(b \mid z,y)$, have an identical shape but differ in their point of symmetry. The first is symmetric about $b(z) = b - Q_{\hat{b}\hat{a}}Q_{\hat{a}}^{-1}(a - z)$, while the second is symmetric about $\hat{b}(z) = \hat{b} - Q_{\hat{b}\hat{a}}Q_{\hat{a}}^{-1}$
$(\hat{a} - z)$. Also the weights share some resemblance. This can be seen if we consider the probability

$$P(\check{a} = a - z) = \int_{S_{a-z}} (\sqrt{\det Q_{\hat{a}}}(2\pi)^{\frac{1}{2}n})^{-1} \exp -\frac{1}{2} \| x - a \|_{Q_{\hat{a}}}^2 \, dx$$

This probability can be worked out to give

$$P(\check{a} = a - z) = \frac{\int_{S_0} \exp -\frac{1}{2} \| x - z \|_{Q_{\hat{a}}}^2 \, dx}{\sum_{u \in Z^n} \int_{S_0} \exp -\frac{1}{2} \| x - u \|_{Q_{\hat{a}}}^2 \, dx} \tag{45}$$

which shows the resemblance with (43).

5.3. The Bayes baseline

With the posterior baseline distribution available, one can now study the corresponding confidence regions as well as determine the Bayes estimate of the baseline,

$$\hat{b}_{Bayes} = \int b p(b \mid y) db$$

For a discussion on how to approximate the confidence regions of the posterior baseline, we refer to [*Gundlich and Koch*, 2001].

Using the results of theorem 8, the Bayes baseline follows as

$$\hat{b}_{Bayes} = \hat{b} - Q_{\hat{b}\hat{a}}Q_{\hat{a}}^{-1} \left(\hat{a} - \sum_{z \in Z^n} z w_z(\hat{a}) \right) \tag{46}$$

Again there is a striking resemblance with the results of section 2. From (4) and (2.2) it follows that the 'fixed' baseline can be written as

$$\check{b} = \hat{b} - Q_{\hat{b}\hat{a}} Q_{\hat{a}}^{-1} \left(\hat{a} - \sum_{z \in Z^n} z s_z(\hat{a}) \right) \qquad (47)$$

We thus see that the two solutions differ in the way the 'float' solution $\hat{a}$ is used to weigh all integer grid points $z \in Z^n$. In case of the Bayes baseline the smooth weights $w_z(\hat{a})$ are used, while in case of the 'fixed' baseline, the $0-1$ values of the indicator function $s_z(\hat{a})$ are used. Although both baseline solutions contain an infinite sum, the one of the 'fixed' baseline can be computed exactly, while the one of the Bayes baseline can only be approximated.

6. REFERENCES

Betti, B., M. Crespi, and F. Sanso (1993): A geometric illustration of ambiguity resolution in GPS theory and a Bayesian approach. *Manuscr Geod*, 18: 317-330.

Bock, Y. (1998): Medium distance GPS measurements. In: Teunissen, P.J.G., and A. Kleusberg (Eds), *GPS for Geodesy*, 2nd edition, Springer Verlag.

Boon, F., B. Ambrosius (1997): Results of real-time applications of the LAMBDA method in GPS based aircraft landings. *Proceedings KIS97*, pp. 339-345.

Boon, F., P.J. de Jonge, C.C.J.M. Tiberius (1997): Precise aircraft positioning by fast ambiguity resolution using improved troposphere modelling. *Proceedings ION GPS-97*, Vol. 2, pp. 1877-1884.

Cox, D.B. and J.D.W. Brading (1999): Integration of LAMBDA ambiguity resolution with Kalman filter for relative navigation of spacecraft. *Proc ION NTM 99*, pp. 739-745.

de Jonge P.J., C.C.J.M. Tiberius (1996a): The LAMBDA method for integer ambiguity estimation: implementation aspects. Publications of the Delft Computing Centre, *LGR-Series* No. 12.

de Jonge, P.J., C.C.J.M. Tiberius (1996b): Integer estimation with the LAMBDA method. *Proceedings IAG Symposium No. 115, 'GPS trends in terrestrial, airborne and spaceborne applications'*, G. Beutler et al. (eds), Springer Verlag, pp. 280-284.

de Jonge, P.J., C.C.J.M. Tiberius, P.J.G. Teunissen (1996): Computational aspects of the LAMBDA method for GPS ambiguity resolution. *Proceedings ION GPS-96*, pp. 935-944.

Euler, H.J., C. Goad (1991): On optimal filtering of GPS dual frequency observations without using orbit information. *Bull Geod*, 65: 130- 143.

Gerdan, G.P. (1995): A comparison of four methods of weighting double difference pseudo range measurements. *Trans Tasman Surv*, 1(1): 60- 66.

Gianniou, M. (1996): Genauigkeitssteigerung bei kurzzeit-statischen und kinematischen Satellitenmessungen bis hin zu Echtzeitanwendung. PhD- thesis, DGK, Reihe C, no. 458, Muenchen.

Gundlich, B., K-R. Koch (2001): Confidence regions for GPS baselines by Bayesian statistics. *Journal of Geodesy*, in print.

Han, S. (1995): Ambiguity resolution techniques using integer least- squares estimation for rapid static or kinematic positioning. Symposium *Satellite Navigation Technology: 1995 and beyond*, Brisbane, Australia, 10 p.

Hofmann-Wellenhof, B., H. Lichtenegger, J. Collins (1997): *Global Positioning System: Theory and Practice*. 4th edition. Springer Verlag.

Jin, X.X. and C.D. de Jong (1996): Relationship between satellite elevation and precision of GPS code measurements. *J Navig*, 49: 253-265.

Jonkman, N.F. (1998): Integer GPS ambiguity estimation without the receiver-satellite geometry. Publications of the Delft Geodetic Computing Centre, *LGR-Series*, No. 18.

Leick, A. (1995): *GPS Satellite Surveying*. 2nd edition, John Wiley, New York.

Parkinson, B., J.J. Spilker (eds) (1996): *GPS: Theory and Applications*, Vols 1 and 2, AIAA, Washington DC.

Peng, H.M., F.R. Chang, L.S. Wang (1999): Attitude determination using GPS carrier phase and compass data. *Proc ION NTM 99*, pp. 727-732.

Schaffrin, B., Y. Bock (1988): A unified scheme for processing GPS dual- band observations. *Bull Geod*, 62: 142-160.

Strang, G., K. Borre (1997): *Linear Algebra, Geodesy, and GPS*, Wellesley-Cambridge Press.

Teunissen, P.J.G., A. Kleusberg (eds) (1998): *GPS for Geodesy*, 2nd enlarged edition, Springer Verlag.

Teunissen, P.J.G. (1993): Least-squares estimation of the integer GPS ambiguities. Invited Lecture, Section IV Theory and Methodology, IAG General Meeting, Beijing, China, August 1993. Also in: *LGR Series*, No. 6, Delft Geodetic Computing Centre.

Teunissen, P.J.G. (1995): The least-squares ambiguity decorrelation adjustment: a method for fast GPS integer ambiguity estimation. *Journal of Geodesy*, 70: 65-82.

Teunissen, P.J.G. (1998a): The ionosphere-weighted GPS baseline precision in canonical form. *Journal of Geodesy*, 72: 107-117.

Teunissen, P.J.G. (1998b): Success probability of integer GPS ambiguity rounding and bootstrapping. *Journal of Geodesy*, 72: 606-612.

Teunissen, P.J.G. (1999a): The probability distribution of the GPS baseline for a class of integer ambiguity estimators. *Journal of Geodesy*, 73: 275-284.

Teunissen, P.J.G. (1999b): An optimality property of the integer least- squares estimator. *Journal of Geodesy*, 73: 587-593.

Teunissen, P.J.G. (2001): The probability distribution of the ambiguity bootstrapped baseline. *Journal of Geodesy*, in print.

Tiberius, C.C.J.M., P.J. de Jonge (1995): Fast positioning using the LAMBDA method. *Proceedings DSNS-95*, paper 30, 8p.

Tiberius, C.C.J.M., P.J.G. Teunissen, P.J. de Jonge (1997): Kinematic GPS: performance and quality control. *Proceedings KIS97*, pp. 289-299.

Tiberius, C.C.J.M. (1998): Recursive data processing for kinematic GPS surveying. Publ Geodesy, no. 45, Netherlands Geodetic Commision, Delft.

ABSTRACT

On the Role of Linear Precoding in Signal Processing for Wireless

Georgios B. Giannakis
University of Minnesota, USA

This talk introduces linear precoding (LP) as a useful signal processing tool for coping with frequency-selective propagation channels encountered with high-rate wireless block transmissions. The importance of LP will be presented for single- and multi-carrier (OFDM) systems, and its links with error-control coding will be delineated.

Its features will be described both for point-to-point and multiple access links with emphasis placed on the generalized multicarrier CDMA, and the novel ideas of block-spreading and chip-interleaving.

0-7803-7011-2/01/$10.00 ©2001 IEEE

RECURSIVE MONTE CARLO ALGORITHMS FOR PARAMETER ESTIMATION IN GENERAL STATE SPACE MODELS

Christophe Andrieu† - *A. Doucet* ‡

†Department of Mathematics, Statistics Group, University of Bristol
Bristol BS8 1TW, U.K.
‡Department of Electrical and Electronic Engineering, University of Melbourne
Parkville, Victoria 3052, Australia.
Email: c.andrieu@bris.ac.uk - doucet@ee.mu.oz.au

ABSTRACT

In this paper we present new algorithms that aim at estimating the "static" parameters of a latent variable process in an on-line manner. This new class of on-line algorithms is inspired by Monte Carlo Markov chain (MCMC) methods whose use has been mainly restricted to static problems, *i.e.* for which the set of observations is fixed. The main interest of this new class of algorithms is that it combines MCMC and particle filtering techniques, for which extensive know-how and literature are now available.

1 Introduction

1.1 Problem Statement

We consider here the problem of on-line estimation of the parameters for latent variable models, here modelled as Markov chains. More precisely we define two real vector-valued stochastic processes $\{x_t;\ t \in \mathbb{N}\}$ and $\{y_t;\ t \in \mathbb{N}^*\}$ where the process $x_t \in \mathbb{R}^{n_x}$ is usually called the *signal* process and the process $y_t \in \mathbb{R}^{n_y}$ is called the *observation* process. The *signal* process x_t is here assumed to be a Markov process with initial density $x_0 \sim \mu(x_0)$ and transition probability densities from state x_{t-1} to state x_t, $K_\theta(x_t|\, x_{t-1})$. The *observations* are conditional upon x_t independent and the conditional marginal density of y_t is $g(y_t|\, x_t, , \theta)$. The parameter $\theta \in \mathbb{R}^{n_\theta}$ is *unknown* and our aim is, given the observations $y_1, \ldots, y_t$, to *estimate sequentially in time the unknown parameter* θ. The problem is of interest in many applications of signal processing, and the importance of this class of problems has generated a vast literature. We review here some of the proposed solutions.

1.2 Brief Literature Review

Assume that an estimate of θ at time $t-1$, noted $\theta^{(t-1)}$, is available and that it is possible to compute exactly the optimal filtering density $p\left(x_t|\, y_{1:t}, \theta^{(t-1)}\right)$ and some of its associated statistics. Then one can use on-line Gradient/EM (Expectation-Maximization) algorithms to estimate θ [3], [5] so as to approach the maximum of $p(y_{1:t}|\,\theta)$ as $t \to +\infty$. In practice, at time t, a statistic $S\left(\theta^{(t-1)}, \theta\right)$ of the filtering density $p\left(x_t|\, y_{1:t}, \theta^{(t-1)}\right)$ is evaluated then the current value of the parameter $\theta^{(t)}$ is updated deterministically in order to maximize $S\left(\theta^{(t-1)}, \theta\right)$ with respect to θ, and so on. This has been proposed as a solution to our problem by many authors in both the statistical and signal processing literatures. Under some regularity assumptions, it can be shown that these algorithms are able to track the local maxima of the series of likelihoods $p(y_{1:t}|\,\theta), t = 1, \ldots$. Unfortunately, analytic expression of the quantity $S\left(\theta^{(t-1)}, \theta\right)$, which is an integral with respect to the filtering distribution $p\left(x_t|\, y_{1:t}, \theta^{(t-1)}\right)$ can only be obtained for restricted classes of processes. For many models that typically involve elements of non Gaussianity and nonlinearity in the dynamic one can in principle use any numerical technique to approximate it. The EKF (Extended Kalman Filter) is one of the earliest such approximation, valid for continuous state-spaces which are not "too" nonlinear. It should be added that in practice, even in favourable cases, these recursive parameter estimation methods are very sensitive to initialization and can easily get trapped in local maxima.

Another approach for parameter estimation consists of using a full Bayesian approach where θ is assumed to be random with a given prior density $p(\theta)$. Then it is possible to define the following filtering density $p(x_t, \theta|\, y_{1:t})$, on the "extended state" (x_t, θ). However this quantity can only rarely be evaluated analytically. The advent of powerful and cheap computers has permitted the development and the application of an efficient and versatile class of numerical methods that address the filtering problem, Sequential Monte Carlo aka Particle Filter [4]. Using such methods, one can compute the filter on the extended state (x_t, θ), that is $p(x_t, \theta|\, y_{1:t})$ and consequently perform inference on θ. Unfortunately, although attractive, this approach does not work in practice. Indeed the extended dynamic model is not ergodic and there is an accumulation of errors over time, whatever the particle filtering method used [1]. This is what motivates the following section, where we discuss a new class of algorithm to address static parameter estimation.

2 Recursive Monte Carlo Algorithms for Parameter Estimation

2.1 Principle

We first here recall the principle of MCMC methods and illustrate the Gibbs sampler on our problem, when the set of observations does not evolve with time. Then we show how it is possible to adapt this MCMC scheme for on-line estimation purposes. In our case MCMC methods will be designed to obtain samples $\left(x_{1:t}^{(i)}, \theta^{(i)}\right)$ $(i = 1, \ldots)$ from say the joint posterior distributions

0-7803-7011-2/01/$10.00 ©2001 IEEE

$p(x_{1:t}, \theta | y_{1:t})$. The samples can be used to evaluate integrals for example [6]. In many cases sampling from the joint distribution might be too complicated, whereas sampling from the conditional distributions $p(x_{1:t} | y_{1:t}, \theta)$ and $p(\theta | x_{1:t}, y_{1:t})$ might be routine. The Gibbs sampler, or data augmentation algorithm, exploits this fact and can be described as follows:

Data augmentation

- Initialization: $i = 0$ and $\theta^{(0)}$.
- Repeat iteration i
  - $x_{1:t}^{(i)} \sim p\left(x_{1:t} | y_{1:t}, \theta^{(i-1)}\right)$
  - $\theta^{(i)} \sim p\left(\theta | x_{1:t}^{(i)}, y_{1:t}\right)$
  - $i \leftarrow i + 1$

It can be shown that under mild conditions the homogeneous Markov chain described above is ergodic and produces - at least asymptotically - samples from the joint distribution $p(x_{1:t}, \theta | y_{1:t})$. More precisely, for a properly defined norm on distributions it can be shown that $\lim_{i \to +\infty} \left\| \mathbb{P}\left(x_{1:t}^{(i)}, \theta^{(i)}\right) - p(x_{1:t}, \theta | y_{1:t}) \right\| = 0$, or perharps more interestingly here $\lim_{i \to +\infty} \left\| \mathbb{P}\left(\theta^{(i)}\right) - p(\theta | y_{1:t}) \right\| = 0$. This algorithm is not designed to achieve our goals of on-line estimation. However it can be adapted to the on-line case in the simple following way:

Recursive Data Augmentation (RDA)

- Initialization: $t = 0$ and $\theta^{(0)}$.
- Iteration t
  - $x_{1:t}^{(t)} \sim p\left(x_{1:t} | y_{1:t}, \theta^{(t-1)}\right)$
  - $\theta^{(t)} \sim p\left(\theta | x_{1:t}^{(t)}, y_{1:t}\right)$
  - $t \leftarrow t + 1$

---

that is $i = t$ in this case. It should be stressed that, contrary to the Gibbs sampler presented above, the Markov chain here is non-homogeneous, due to the fact that the set of observations evolves with time. It can however be shown that at iteration t, the invariant distribution of the RDA is the joint posterior distribution $p(x_{1:t}, \theta | y_{1:t})$. One can therefore hope that as t becomes large, and if $p(x_{1:t+1}, \theta | y_{1:t+1})$ is not too different from $p(x_{1:t}, \theta | y_{1:t})$ (that is the evolution of this distribution is smooth enough in a certain sense) then $\mathbb{P}\left(x_{1:t}^{(t)}, \theta^{(t)}\right)$, the actual distribution of $\left(x_{1:t}^{(t)}, \theta^{(t)}\right)$ produced by the algorithm will track the series of distributions $p(x_{1:t}, \theta | y_{1:t})$. Similarly one can expect the chain $\theta^{(t)}$ to be asymptotically distributed according to $p(\theta | y_{1:t})$, *i.e.* $\lim_{t \to +\infty} \left\| \mathbb{P}\left(\theta^{(t)}\right) - p(\theta | y_{1:t}) \right\| = 0$. It is known that under regularity conditions the series of distributions $p(\theta | y_{1:t})$ converges in a certain sense to a mixture of delta functions located on the global maxima of $p(\theta | y_{1:t})$. Under consistency conditions it is known that these maxima correspond to the true values $\theta_\star$ of θ.

This algorithm shares many common features with the simulated annealing algorithm. Indeed both are non-homogeneous Markov chains that track series of distributions which concentrate themselves on a set of points. Motivated by this analogy, it should not be surprising that convergence of the chain towards the global maxima of the marginal distribution $p(\theta | y_{1:t})$ requires that this series of distributions does not concentrate itself too quickly on its set of asymptotic global maxima. It can be shown in many cases that the rate at which this concentration occurs (in terms of the rate at which the variance around global maxima goes to zero) is $1/t$, which is far more than what is required by the simulated annealing algorithm. This is why we modify our algorithm so as to change the rate mentioned above to $1/\log(t + t_0)$. More precisely we modify the two conditional distributions in order to define a new joint distribution and therefore define an alternative probabilistic model

$$\begin{aligned} \widetilde{p}(x_{1:t} | y_{1:t}, \theta) &= p(x_{1:t} | y_{1:t}, \theta) \\ \widetilde{p}(\theta | x_{1:t}, y_{1:t}) &\propto [q_t(\theta; x_{1:t}, y_{1:t})]^{\beta_t} p(\theta) \end{aligned}$$

where β_t is the inverse of the "temperature" and $q_t(\theta | x_{1:t}, y_{1:t})$ is recursively defined as follows,

$$\begin{aligned} l_1(\theta; x_1, y_1) &= p(x_1, y_1 | \theta) \\ l_i(\theta; x_{1:i}, y_{1:i}) &= [l_{i-1}(\theta | x_{1:i-1}, y_{1:i-1})]^{1-\gamma_i} \\ &\quad \times [p(x_i, y_i | x_{i-1}, \theta)]^{\gamma_i} \end{aligned}$$

for $i = 2, \ldots, t$ and a typically decreasing sequence of gains $\gamma_i \in (0, 1)$ (typically $\gamma_i = 1/i$). The idea behind the definition of $\widetilde{p}(\theta | x_{1:t}, y_{1:t})$ is the following. In view of the definition of a realistic sequential algorithm, it is clear that sampling from $p(x_{1:t} | y_{1:t}, \theta)$ will be practically impossible as $t \to +\infty$. Therefore a fixed number of hidden variables, say $x_{t-L:t}$, will effectively be sampled at each iteration. Consequently it is natural to give more weight to recent hidden variables and discard the information brought by old ones. This is what motivates the definition of $l_i(\theta; x_{1:i}, y_{1:i})$. Then intuitively the distribution proportional to $l_t(\theta; x_{1:t}, y_{1:t}) p(\theta)$ does not concentrate itself on its global maxima (contrary to $p(\theta | x_{1:t}, y_{1:t})$), and must be "annealed", therefore the power $\beta_t \to +\infty$.

Finally, we suggest the following improvement of the algorithm. As mentioned above, it will practically be impossible to sample a full trajectory from $p(x_{1:t} | y_{1:t}, \theta)$ as the computational complexity would increase over time. One can suggest to sample $x_{t-L:t}$ only, according to $p(x_{t-L:t} | y_{1:t}, \theta)$, the past values $x_{1:t-L-1}$ are not modified . The algorithm will proceed as follows.

(Approximate) RDA

- Initialization: $t = 0$ and $\theta^{(0)}$.
- Iteration t
  - $x_{t-L:t}^{(t)} \sim p\left(x_{t-L:t} | y_{1:t}, \theta^{(t-1)}\right)$
  - $\theta^{(t)} \sim \widetilde{p}\left(\theta | x_{1:t}^{(t)}, y_{1:t}\right)$
  - $t \leftarrow t + 1$

If $p(x_{1:t-L-1} | y_{1:t}, \theta) = p(x_{1:t-L-1} | y_{1:t-1}, \theta)$, as in the case of mixture distributions (see Section 3), then the algorithm is "exact". Otherwise one implicitely assumes that

$$p(x_{1:t-L-1} | y_{1:t}, \theta) \simeq p(x_{1:t-L-1} | y_{1:t-1}, \theta),$$

it is a valid assumption if the optimal filter has exponential forgetting properties.

2.2 Extensions and approximations

The algorithm that we have presented above assumes that we know how to sample exactly from the two conditional distributions. Although this assumption is reasonable for a large class of problems, as we shall see below, it does not encompass other complex processes of interest. This is why we point out here that modern simulation techniques can be used in order to apply the algorithm above to more complex scenarios.

Metropolis-Hastings (MH)/Reversible jump MCMC (RJMCMC): First it should be pointed out that perfect simulation from the conditional distributions might be replaced with standard MCMC techniques, that is one replaces sampling from the conditional distributions with sampling from a Markov transition kernel which admits π as invariant distribution. This includes as a very important special case RJMCMC, which allow for on-line model selection.

Analytical/Particle filtering approximation of $p(x_{t-L:t}|y_{1:t},\theta)$: It is important to notice that one of the distribution from which we want to generate samples is the distribution $p(x_{t-L:t}|y_{1:t},\theta)$. In order to make the algorithm practical it is first necessary to replace $p\left(x_{t-L:t}|y_{1:t},\theta^{(t-1)}\right)$ with $p\left(x_{t-L:t}|y_{1:t},\theta^{(1)},\ldots,\theta^{(t-1)}\right)$ (that is we do not restart the optimal filter from time $t=0$ but use the current approximation). This approach can be thought to be valid if $p(x_{t-L:t}|y_{1:t},\theta)$ evolves smoothly as a function of θ. When the filter does not possess nice analytical properties (Kalman filter, HMM), it is then possible to use Sequential Monte Carlo methods which have proved to be efficient and versatile numerical techniques.

2.3 Stochastic approximation, convergence issues and open problems

For the models for which it is practically possible to apply the RDA algorithm, the conditional distribution $\widetilde{p}(\theta|x_{1:t},y_{1:t})$ depends on some sufficient statistics $\Phi^{(t)}=\frac{1}{t}\sum_{i=1}^{t}\varphi(x_{i-1:i},y_i)$, and it is therefore possible to rewrite our algorithm in the following way:

(Approximate) RDA

- Initialization: $t=0$, $\widetilde{\Phi}^{(0)}=0$ and $\theta^{(0)}$.
- Iteration t
  - $x_{t-L:t}^{(t)}\sim p\left(x_{t-L:t}|y_{1:t},\theta^{(t-1)}\right)$.
  - $\widetilde{\Phi}^{(t)}=(1-\gamma_t)\,\widetilde{\Phi}^{(t-1)}+\gamma_t\varphi\left(x_{t-1:t}^{(t)},y_t\right)$.
  - $\theta^{(t)}\sim\widetilde{p}\left(\theta|\widetilde{\Phi}^{(t)}\right)$
  - $t\leftarrow t+1$

Clearly θ can be considered to be a dummy variable, and x_t is drawn from the distribution $\int p(x_t|y_{1:t},\theta)\,\widetilde{p}\left(\theta|\widetilde{\Phi}^{(t-1)}\right)d\theta$.

Then, the algorithm described above is clearly a stochastic approximation algorithm. It is interesting to notice that if we set $\beta_t=+\infty$, then $\widetilde{p}\left(\theta|\widetilde{\Phi}^{(t-1)}\right)$ is a mixture of delta function on its set of global maxima. This algorithm is therefore an on-line stochastic EM algorithm, and the RDA algorithm can then be thought of as a "noisy" version of this algorithm. This brings some light on the possible convergence properties of our algorithms. Stochastic approximation techniques can be used to prove the (local) convergence of this algorithm in the finite mixture case [2]. Although standard stochastic approximation for global optimization do not apply here, it can be thought that the RDA will provide consistent parameter estimation under some regularity conditions. However we would like to point out potential problems in a general setting. Several levels of approximation have been suggested in Section 2.2 in order to make the algorithms practical. Although we think that the analytical approximations proposed probably lead to valid algorithms, it seems to us that the use of particle filter techniques should lead to an accumulation of error, and in the most optimistic case lead to limited precision on the estimates of the true value of θ.

3 Applications

3.1 Mixture of normal distributions

We first start with a case which does not require any approximation. Proofs of convergence are given in [2]. Consider the special case of finite Gaussian mixture distributions as in [7]. In this scenario, we observe independent identically distributed data $y_1,\ldots,y_t,\ldots$

$$y_t|x_t\sim\mathcal{N}\left(\mu_{x_t},\sigma_{x_t}^2\right)$$

and $\Pr(x_t=j)=\pi_j$, independently of x_{t-1}. We want to estimate $\theta\triangleq\left\{\left(\mu_j,\sigma_j^2,\pi_j\right);j\in S\right\}$.

To complete the Bayesian model, we assume that the $\left(\mu_j,\sigma_j^2\right)$, $j\in S$, are distributed according to the conjugate priors

$$\mu_j|\sigma_j\sim\mathcal{N}\left(v_j,\sigma_j^2/\rho_j\right),\sigma_j^2\sim\mathcal{IG}\left(\frac{\rho_j}{2},\frac{\eta_j}{2}\right)$$

and $\Pi\triangleq(\pi_1,\ldots,\pi_s)$ is distributed according to a Dirichlet distribution

$$\Pi\sim\mathcal{D}_s(\alpha_1,\ldots,\alpha_s).$$

Note that the prior does not have any effect on the final results. It is just used to define a valid probability density for the parameter θ. In practice it proves to stabilize the algorithm (especially in the initial phase). The conditional distribution of the parameters given the missing data is given by

$$\begin{aligned}
\sigma_j^2 &\sim \mathcal{IG}\left(\frac{1}{2}\left(\beta_t n_{j,t}+\rho_j\right),\varsigma_{t,j}\right)\\
\text{with } \varsigma_{t,j} &= \frac{\rho_j v_j^2+\eta_j+\beta_t\Upsilon_{t,j}^2-\frac{\left(\rho_j v_j+\beta_t\Upsilon_{t,j}\right)^2}{\rho_j+\beta_t n_{t,j}}}{2}\\
\mu_j|\sigma_j^2 &\sim \mathcal{N}\left(\frac{\rho_j v_j+\beta_t\Upsilon_{t,j}}{\rho_j+\beta_t n_{t,j}},\frac{\sigma_j^2}{\left(\rho_j+\beta_t n_{t,j}\right)}\right)\\
\Pi &\sim \mathcal{D}_s\left(\beta_t n_{1,j}+\alpha_1,\ldots,\beta_t n_{s,j}+\alpha_s\right)
\end{aligned}$$

where

$$\begin{aligned}
n_{1,j} &= \delta_{x_1,j},\ n_{t,j}=(1-\gamma_t)\,n_{t-1,j}+\gamma_t\delta_{x_t,j},\\
\Upsilon_{1,j} &= \delta_{x_1,j}y_1,\ \Upsilon_{t,j}=(1-\gamma_t)\,\Upsilon_{t-1,j}+\gamma_t\delta_{x_t,j}y_t,\\
\Upsilon_{1,j}^2 &= \delta_{x_1,j}y_1^2,\ \Upsilon_{t,j}^2=(1-\gamma_t)\,\Upsilon_{t-1,j}^2+\gamma_t\delta_{x_t,j}y_t^2.
\end{aligned}$$

Sampling the missing data is routine, as it is a finite discrete distribution. For our simulation we generated 200000 samples from a mixture of three normals with means -2.0, 0.5 and 1.5, variances 3.0, 0.5 and 1.0 and proportions 0.35, 0.6 and 0.05. We applied our algorithm with $\beta_t\propto\alpha t+\beta$. The results are presented in Fig 1.

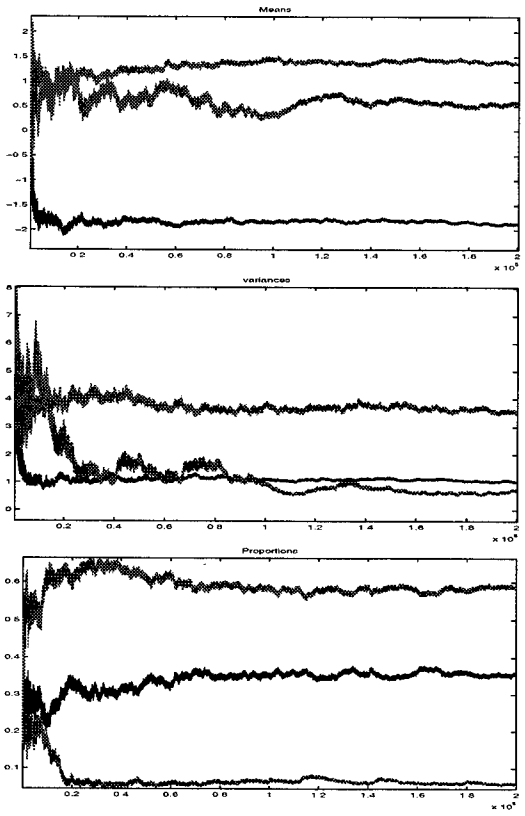


Figure 1: The first 200000 iterations of the algorithm, with from top to bottom, the means, the variances and the proportions

3.2 Noisy AR process

In this case the signal and observation processes are described by the two following equations

$$\begin{aligned} x_{t+1} &= a x_t + \sigma_v v_{t+1} \\ y_t &= x_t + \sigma_w w_t, \end{aligned}$$

that is for simplicity we restrict ourselves to a noisy AR(1) process. The unknown fixed parameters are $\theta = \left(a, \sigma_v^2, \sigma_w^2\right)$ the AR coefficient, the variance of the dynamic and observation noises respectively. We have used standard normal and inverse gamma distributions for the priors [1]. We sample from $p\left(x_{t-1:t} | \, y_{1:t}, \theta^{(1)}, \ldots, \theta^{(t-1)}\right)$ using the Kalman filter followed by a backward sampling step. Further comparison with other approximation techniques will be made in the future. We show our results in Fig. 2, which seem to lead to reasonable values of the parameters (the true values here were $a = .6$ and $\sigma_v = 0.5$). We however remind the reader of the potential validity problems of this approach pointed in Section 2.3.

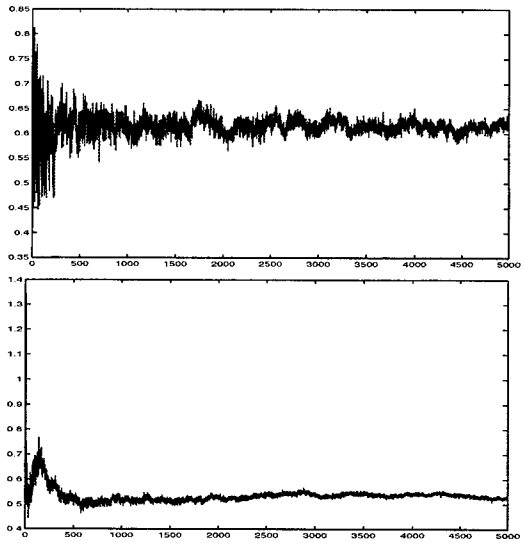


Figure 2: From top to bottom, the AR coefficient, the variance of the dynamic noise.

4 REFERENCES

[1] Andrieu, C., De Freitas, N., & Doucet, A. (1999) Sequential MCMC for Bayesian model selection. *in Proc. IEEE HOS'99, Israel.*

[2] Andrieu, C., Doucet, A. & Tadić, V. (2001) Recursive data augmentation for parameter estimation in finite mixture distributions, *forthcoming*.

[3] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* B 39, 1-38.

[4] Doucet A., de Freitas J.F.G. & Gordon N.J. (eds.) (2001) *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

[5] Quian, W. & Titterington, D.M. (1991). Estimation of parameters in hidden Markov models. *Phil. Trans. R. Soc. London* A 337, 407-28.

[6] Robert, C.P. & Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

[7] Titterington D. M., Smith A. F. M. and Makov U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*, London: Wiley.

MONTE CARLO SMOOTHING WITH APPLICATION TO AUDIO SIGNAL ENHANCEMENT

William Fong and Simon Godsill

Signal Processing Group, University of Cambridge
Cambridge, CB2 1PZ U.K.
E-mail: wnwf2@eng.cam.ac.uk, sjg@eng.cam.ac.uk

ABSTRACT

We describe methods for applying Monte Carlo filtering and smoothing for estimation of unobserved states in a non-linear state space model. By exploiting the statistical structure of the model, we develop a Rao-Blackwellised Particle Smoother. The suggested algorithm is tested with real speech and audio data and the results are shown and compared with those generated using the generic particle smoother and the extended Kalman filter. It is found that the suggested algorithm gives better results.

1. INTRODUCTION

Many problems in applied statistics, statistical signal processing and time series analysis can be stated in a state space form as follows,

$$\begin{array}{lcll} x_{t+1} & \sim & f(x_{t+1}|x_t) & \text{State evolution density} \\ y_{t+1} & \sim & g(y_{t+1}|x_{t+1}) & \text{Observation density} \end{array} \quad (1)$$

where $\{x_t\}$ are unobserved states of the system and $\{y_t\}$ are observations made over some time interval $t \in \{1, \ldots, T\}$. $f(.|.)$ and $g(.|.)$ are pre-specified state evolution and observation densities.

A primary concern in many state-space inference problems is sequential estimation of the filtering distribution $p(x_t|y_{1:t})$ and simulation of the entire smoothing distribution $p(x_{1:t}|y_{1:t})$, where $y_{1:t} \triangleq \{y_1, y_2, \ldots, y_t\}$ and $x_{1:t} \triangleq \{x_1, x_2, \ldots, x_t\}$. Updating of the filtering distribution can be achieved in principle using the standard filtering recursions

$$p(x_{t+1}|y_{1:t}) = \int p(x_t|y_{1:t}) f(x_{t+1}|x_t) dx_t$$

$$p(x_{t+1}|y_{1:t+1}) = \frac{g(y_{t+1}|x_{t+1}) p(x_{t+1}|y_{1:t})}{p(y_{t+1}|y_{1:t})}.$$

Similarly, smoothing can be performed recursively backwards in time using the smoothing formula

$$p(x_t|y_{1:T}) = \int p(x_{t+1}|y_{1:T}) \frac{p(x_t|y_{1:t}) f(x_{t+1}|x_t)}{p(x_{t+1}|y_{1:t})} dx_{t+1}.$$

In practice these filtering and smoothing computations can only be performed in closed form for linear Gaussian models using the Kalman filter / smoother and for finite state-space hidden Markov models.

For non-linear non-Gaussian models, there is no general analytic expression for the required density functions. The extended Kalman filter is a popular approach for non-linear models, which linearises the filtering distributions, so that the Kalman filter can be applied.

Another approximation strategy is that of sequential Monte Carlo methods, also known as Particle Filters [4, 3]. Within the particle filter framework, the filtering distribution is approximated with an empirical distribution formed from point masses, or particles,

$$p(x_t|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta(x_t - x_t^{(i)}), \quad \sum_{i=1}^{N} w_t^{(i)} = 1, \; w_t^{(i)} \geq 0$$

where $\delta(.)$ is the Dirac delta function and $w_t^{(i)}$ is a weight attached to particle $x_t^{(i)}$. In addition to the particle filter, particle smoothers, which are a simple and efficient method for generating realisations from the entire smoothing density $p(x_{1:T}|y_{1:T})$ using the particulate approximation has been developed [5].

2. AUDIO MODELS

Speech signals are inherently time-varying in nature, and any realistic representation should thus involve a model whose parameters evolve over time. One such model is the time-varying autoregression (TVAR)

$$u_t = \sum_{i=1}^{p} a_{t,i} u_{t-i} + e_t$$

Here $\{u_t\}$ is the audio signal process, $a_t = [a_{t,1}, \ldots, a_{t,p}]'$ is the pth order AR coefficient vector and e_t is a Gaussian excitation at time t having variance $\sigma_{e_t}^2$. A Gaussian random walk model is assumed for the log-variance $\phi_{e_t} = \log(\sigma_{e_t}^2)$,

$$f(\phi_{e_t}|\phi_{e_{t-1}}, \sigma_{\phi_e}^2) = \mathcal{N}(\mu_{\phi_t}, \sigma_{\phi_e}^2) \quad (2)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

where $\mu_{\phi_t} = \log(\alpha\sigma^2_{e_{t-1}})$ and α is a coefficient just less than 1.

For the time variation in a_t, we choose to work in the time-varying partial correlation (PARCOR) coefficient domain [5, 6]. Improved stability can be achieved by ensuring each reflection coefficient, ρ_t, is within the interval (-1,+1). The constrained PARCOR random walk model is

$$f(\rho_t|\rho_{t-1},\sigma_a^2) \propto \begin{cases} \mathcal{N}(\rho_{t-1},\sigma_a^2 I) & \text{if } \max\{|\rho_{t,i}|\} < 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\rho_t = [\rho_{t,1},\ldots,\rho_{t,p}]$. The TV-PARCOR is assumed because it provides a better physical representation of audio signals. This arises since the TV-PARCOR model can be regarded as a time-varying acoustical tube mechanism, which is a reasonable approximation for speech and many musical instruments.

The full specification of the state space model is then as follows. The state vector x_t is partitioned as $[z_t, \theta_t]$ with $z_t = u_{t-p+1:t}$ and θ_t being the signal state and the parameter state respectively. The signal is assumed to be submerged in white Gaussian noise (WGN) with known variance σ_v^2, *i.e.*

$$g(y_t|x_t) = \mathcal{N}(y_t; u_t, \sigma_v^2)$$

The parameter vector θ_t is further partitioned as $[a_t, \phi_{e_t}]$.

3. MONTE CARLO FILTERING AND SMOOTHING

Rearrange the filtering distribution $p(x_t|y_{1:t})$,

$$\begin{aligned} p(x_t|y_{1:t}) &\propto g(y_t|x_t)p(x_t|y_{1:t-1}) \\ &= \int g(y_t|x_t)f(x_t|x_{t-1})p(x_{0:t-1}|y_{1:t-1})dx_{0:t-1} \end{aligned} \tag{4}$$

Provided that a particle approximation to $p(x_{0:t-1}|y_{1:t-1})$ has already been generated,

$$p(x_{0:t-1}|y_{1:t-1}) \approx \sum_{i=1}^{N} \delta(x_{0:t-1} - x_{0:t-1}^{(i)})$$

Then, assuming we can evaluate $f(x_t|x_{t-1})$ and $g(y_t|x_t)$ pointwise, we generate, for each state trajectory $x_{0:t-1}^{(i)}$, a random sample from a proposal distribution, $q(x_t|x_{0:t-1}^{(i)}, y_{1:t})$. The filtering distribution (4) can then be approximated as

$$p(x_t|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta(x_t - x_t^{(i)})$$

with

$$w_t^{(i)} \propto \frac{g(y_t|x_t^{(i)})f(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t})}$$

In addition to the Monte Carlo filtering algorithm, Monte Carlo smoothing [5] methods have been developed to generate realisations from the joint smoothing density $p(x_{1:T}|y_{1:T})$. Sample realisations are obtained using the following factorisation

$$p(x_{1:T}|y_{1:T}) = p(x_T|y_{1:T}) \prod_{t=1}^{T-1} p(x_t|x_{t+1:T}, y_{1:T}) \tag{5}$$

where, given the particulate approximation to $p(x_t|y_{1:t})$ and using the Markovian assumptions of the model, we can write,

$$\begin{aligned} p(x_t|x_{t+1:T}, y_{1:T}) &\propto p(x_t|y_{1:t})f(x_{t+1}|x_t) \\ &\approx \sum_{i=1}^{N} w_{t|t+1}^{(i)} \delta(x_t - x_t^{(i)}) \end{aligned} \tag{6}$$

with the modified weights

$$w_{t|t+1}^{(i)} \propto w_t^{(i)} f(x_{t+1}|x_t^{(i)}) \tag{7}$$

This revised particle distribution can now be used to generate states successively in the reverse-time direction, conditioning upon future states.

4. RAO-BLACKWELLISED PARTICLE FILTERING AND SMOOTHING

One of the major drawbacks of any Monte Carlo filtering / smoothing strategy is that sampling in high-dimensional spaces can be inefficient. In some cases, however, the model has "tractable substructure" [2], which can be analytically marginalised out, conditional on other state variables. The advantage of this strategy is that it can drastically reduce the size of the space over which we need to sample, and hence the estimation variance [4].

Marginalising out some of the variables is an example of a standard statistical variance reduction strategy known as Rao-Blackwellisation, see [1] for a general discussion on the topic.

In this paper we focus on applying Rao-Blackwellisation to fixed-interval smoothing; that is, given $y_{1:T}$, we would like to simulate from the entire state density $p(x_{1:T}|y_{1:T})$. The reason for this is that a greater degree of smoothing can be important for the convincing reconstruction of audio signals.

4.1. Rao-Blackwellised Particle Filter

First we review the standard RB particle filter [2, 4]. Assume that the state vector $x_{1:t}$ can be partitioned as $[z_{1:t}, \theta_{1:t}]$ and $z_{1:t}$ can be marginalised out analytically. For instance, if conditional on $\theta_{1:t}$, $z_{1:t}$ reduces to a linear Gaussian state-space system, then all the integration can be performed analytically on-line using the Kalman filter and the prediction error decomposition.

Let us consider the marginal filtering distribution,

$$p(\theta_t|y_{1:t}) = \int p(z_t, \theta_t|y_{1:t})dz_t$$
$$\propto \int p(y_t|\theta_{1:t}, y_{1:t-1}) f(\theta_t|\theta_{t-1}) p(\theta_{0:t-1}|y_{1:t-1}) d\theta_{0:t-1}$$

Given the particle approximation to $p(\theta_{0:t-1}|y_{1:t-1})$, new particles $\theta_t^{(i)}$ are drawn from $f(\theta_t|\theta_{t-1}^{(i)})$, and $p(\theta_t|y_{1:t})$ is approximated by

$$p(\theta_t|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta(\theta_t - \theta_t^{(i)})$$

with

$$w_t^{(i)} \propto p(y_t|\theta_{1:t}^{(i)}, y_{1:t-1})$$

Under the assumption of a conditionally linear Gaussian structure, $p(y_t|\theta_{1:t}, y_{1:t-1})$ can be evaluated using the Kalman filter and the prediction error decomposition.

4.2. Rao-Blackwellised Particle Smoother

We modify the generic particle smoother [5] to incorporate Rao-Blackwellisation to form a RB Particle Smoother.

Given the particulate approximation for the parameter filtering distribution $p(\theta_t|y_{1:t})$, the marginal smoothing distribution $p(\theta_t|z_{t+1:T}, \theta_{t+1:T}, y_{1:T})$ is approximated by

$$p(\theta_t|z_{t+1:T}, \theta_{t+1:T}, y_{1:T}) \approx \sum_{i=1}^{N} w_{t|t+1}^{(i)} \delta(\theta_t - \theta_t^{(i)}) \quad (8)$$

with the modified weight $w_{t|t+1}^{(i)}$ being

$$w_{t|t+1}^{(i)} = w_t^{(i)} p(z_{t+1}, \theta_{t+1}|\theta_{1:t}^{(i)}, y_{1:t}) \quad (9)$$
$$= w_t^{(i)} f(\theta_{t+1}|\theta_t^{(i)}) p(z_{t+1}|\theta_{t+1}, \theta_{1:t}^{(i)}, y_{1:t})$$

Smoothed realisations for the parameters $\{\tilde{\theta}_t; t = 1, \ldots, T\}$ can be generated recursively backward in time using the equations above in a similar manner to the generic particle smoother [5]. We now prove the correctness of this Rao-Blackwellised approximation.

Proof: By partitioning the state vector as $x_{1:t} = [z_{1:T}, \theta_{1:T}]$, we factorise the smoothing density function $p(x_{1:T}|y_{1:T})$,

$$p(z_{1:T}, \theta_{1:T}|y_{1:T}) = p(z_T, \theta_T|y_{1:T}) \times$$
$$\times \prod_{t=1}^{T-1} p(z_t, \theta_t|z_{t+1:T}, \theta_{t+1:T}, y_{1:T}) \quad (10)$$

The conditional smoothing density $p(z_t, \theta_t|z_{t+1:T}, \theta_{t+1:T}, y_{1:T})$ can be further factorised,

$$p(z_t, \theta_t|z_{t+1:T}, \theta_{t+1:T}, y_{1:T}) = p(z_t|\theta_{t:T}, z_{t+1:T}, y_{1:T}) \times$$
$$\times p(\theta_t|z_{t+1:T}, \theta_{t+1:T}, y_{1:T}) \quad (11)$$

Using the particle approximation from the forward sweep of the RB particle filter

$$p(\theta_{1:t}|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta(\theta_{1:t} - \theta_{1:t}^{(i)})$$

and the Markovian assumptions of the model, we marginalise the current signal state z_t out from the joint density function (11) yielding the following approximation,

$$p(\theta_t|z_{t+1:T}, \theta_{t+1:T}, y_{1:T})$$
$$\propto p(z_{t+1}, \theta_{t+1}|\theta_t, y_{1:t}) p(\theta_t|y_{1:t})$$
$$\approx \int p(z_{t+1}, \theta_{t+1}|\theta_{1:t}, y_{1:t}) \sum_{i=1}^{N} w_t^{(i)} \delta(\theta_{1:t} - \theta_{1:t}^{(i)}) d\theta_{1:t-1}$$
$$= \sum_{i=1}^{N} w_{t|t+1}^{(i)} \delta(\theta_t - \theta_t^{(i)})$$

with the weight being $w_{t|t+1}^{(i)} = w_t^{(i)} p(z_{t+1}, \theta_{t+1}|\theta_{1:t}^{(i)}, y_{1:t})$, as required.

Given $\tilde{\theta}_{t:T}$ and $\tilde{z}_{t+1:T}$, $\tilde{\theta}_t$ is drawn using the approximate smoothing distribution (8) and the modified weight (9). Provided that $\tilde{\theta}_t = \theta_t^{(j)}$, the smoothed signal realisation $\tilde{z}_t$ is obtained by sampling from the conditional density function $p(z_t|\theta_t^{(j)}, \tilde{z}_{t+1:T}, \tilde{\theta}_{t+1:T}, y_{1:T})$, which is Gaussian.

Under the assumption of a conditionally Gaussian structure for the signal, the modified weight can be computed efficiently using the one-step ahead prediction equation from the Kalman filter.

By repeating the sampling process

$$\tilde{\theta}_t \sim \sum_{i=1}^{N} w_{t|t+1}^{(i)} \delta(\theta_t - \theta_t^{(i)})$$
$$\tilde{z}_t \sim p(z_t|\tilde{\theta}_t^{(j)}, \tilde{z}_{t+1:T}, \tilde{\theta}_{t+1:T}, y_{1:T})$$

recursively backward in time, approximate samples $[\tilde{z}_{1:T}, \tilde{\theta}_{1:T}]$ are drawn from $p(z_{1:T}, \theta_{1:T}|y_{1:T})$.

5. EXPERIMENTAL RESULTS

Extensive tests have been carried out to investigate the effectiveness of the suggested algorithms using a variety of audio datasets. It is found that our proposed Rao-Blackwellised (TV-PARCOR) particle smoother consistently outperforms the classical extended Kalman smoother and the generic particle smoother. Some representative examples of the tests conducted are described in detail.

In our simulations, the fixed hyperparameters used are $\sigma_a^2 = 10^{-4}$, $\alpha = 0.995$ and $\sigma_{\phi_c}^2 = 10^{-6}$. Finally, the TVAR model order is fixed to $p = 6$.

5.1. Test 1 — Rao-Blackwellised Particle Smoother

A section of speech data is used to compare the performance of the extended Kalman smoother, the generic particle smoother and the RB particle smoother.

For the generic particle smoother and the RB Particle Smoother, the TV-PARCOR random walk model is used. $N = 200$ particles are used and smoothing is applied to generate $M = 20$ realisations.The SNR in is 10.2 dB and the SNR out of different algorithms are summarised below:

Extended Kalman Smoother	12.1 dB
Generic Particle Smoother	10.8 dB
RB Particle Smoother	13.3 dB

As a result of this and other simulations on audio data, we conclude that the RB particle smoother outperforms the generic particle smoother and the Extended Kalman smoother. It confirms experimentally the theory that by marginalising out some of the state variables, the estimation performance will improve.

5.2. Test 2 — TV-PARCOR model

As the extended Kalman filter is a computationally cheap algorithm, the RB particle smoother has to show consistent improvement over the extended Kalman smoother.Therefore we re-run the test using two pieces of high quality music. Meanwhile, we include the RB (TVAR) particle smoother in the test in order to verify experimentally our suggestion that the TV-PARCOR model is a better physical representation of audio signals than the TVAR model.

The two pieces of music used is a section of violin playing (SNR in = 5.8 dB) and brass (SNR in = 10.4 dB). In each case, $N = 100$ particles are used and $M = 10$ smoothed trajectories are generated. The output SNR of different algorithms are summarised below:

	violin	**brass**
Extended Kalman Smoother	6.9 dB	6.3 dB
RB TVAR Smoother	10.0 dB	16.8 dB
RB TV-PARCOR Smoother	12.1 dB	17.0 dB

We conclude that the RB particle smoother outperforms the extended Kalman filter very dramatically in terms of SNR for these extracts, giving a significant noise reduction when the extended Kalman filter effectively fails. In addition, the TV-PARCOR model outperforms the standard TVAR model, with the amount of improvement depending on the type of input material.

5.3. Test 3 — Different input SNR

In our final test, we investigate the performance of the RB particle smoother algorithm at different input SNR levels and compare with those generated using the extended Kalman smoother. A section of piano music is used for this purpose.

The SNR out using the different algorithms given noisy signals at different input SNR levels are summarised below:

SNR in	**RB smoother**	**ex Kalman smoother**
0 dB	8.5 dB	4.2 dB
10 dB	13.9 dB	13.8 dB
20 dB	20.9 dB	20.9 dB

It is found that the Rao-blackwellised particle smoother performs significantly better than the extended Kalman filter at low SNR, while both algorithms perform equally well at high SNR.

6. CONCLUSIONS

In this article, we have applied sequential Monte Carlo smoothing methods to audio signal enhancement problems. A TV-PARCOR model was proposed for modelling the time variation of the AR coefficients, which has a good physical interpretation in terms of acoustical tube models. In cases where the models concerned have some "tractable substructure", Rao-Blackwellisation can be applied to reduce the estimation variance. A RB smoothing algorithm was developed for this purpose. Extensive tests have been carried out to investigate the effectiveness of the suggested algorithms applied to a variety of audio data. It is found that the RB (TV-PARCOR) particle smoother outperforms classical approaches such as the extended Kalman smoother and the generic particle smoother.

7. REFERENCES

[1] G. Casella and C. P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83:81–94, 1996.

[2] A. Doucet, N. de Freitas, K. Murphy and S. Russell. Rao-Blackwellised particle filtering for dynamic Bayessian networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, Stanford, California: Morgan Kaufmann, 2000.

[3] A. Doucet, N. de Freitas and N. J. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.

[4] A. Doucet, S. J. Godsill and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistic and Computing*, 10:197–208, 2000.

[5] A. Doucet, S. J. Godsill and M. West. Monte Carlo filtering and smoothing with application to time-varying spectral estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 701–704, 2000.

[6] B. Friedlander. Lattice filters for adaptive processesing. *Proc.IEEE*, 70(8):829–867, August 1982.

PARAMETER ESTIMATION BY A MARKOV CHAIN MONTE CARLO TECHNIQUE FOR THE CANDY MODEL

X. Descombes [1], *M.N.M. van Lieshout* [2],

R. Stoica [2], *J. Zerubia* [1]

[1] Ariana, CNRS/UNSA/INRIA joint research group
INRIA, PO Box 93,
06902 Sophia Antipolis Cedex,
France

[2] CWI, PO Box 94079,
1090 GB Amsterdam
The Netherlands

Note: This research was supported by NWO grant 'Inference for random sets' (613-03-045) and INRIA grant via ERCIM.

ABSTRACT

This paper presents a parameter estimation method for the Candy model based on Monte Carlo approximation of the likelihood function. In order to produce such an approximation a Metropolis-Hastings style algorithm [3] for simulating the Candy model [10, 11] is introduced.

1. SET-UP AND NOTATION

In the last decade in image processing, a few researchers moved away from pixel-based methods to more high-level image analysis based on point process models. In this spirit, Stoica, Descombes and Zerubia [11] introduced a marked point process model for line segments, dubbed *Candy*, as prior distribution for the image analysis problem of extracting linear networks such as roads or rivers from images obtained by aerial and high resolution satellite photography.

More formally, represent a line segment as a point in some compact subset $K \subset \mathbb{R}^2$ of strictly positive volume $0 < \nu(K) < \infty$ with an attached mark taking values in $[l_{\min}, l_{\max}] \times [0, \pi)$ for some $0 < l_{\min} < l_{\max} < \infty$. Each marked point (k, l, θ) can be interpreted as a line segment with midpoint k, length l, and orientation θ. When applying the model to road extraction, it is natural to include marks for characteristics such as width and color as well. A configuration of line segments is a finite set of marked points. The probabilistic model is defined by its density p with respect to a unit rate Poisson process on K with independently and uniformly distributed marks as follows. At $\mathbf{s} = \{s_1, \ldots, s_n\}$ with $s_i = (k_i, l_i, \theta_i) \in K \times [l_{\min}, l_{\max}] \times [0, \pi), i = 1, \ldots, n$,

$$p(\mathbf{s}) = \alpha \prod_{i=1}^{n} \exp\left[\frac{l_i - l_{\max}}{l_{\max}}\right] \times \gamma_f^{n_f(\mathbf{s})} \gamma_s^{n_s(\mathbf{s})} \gamma_d^{n_d(\mathbf{s})} \gamma_o^{n_o(\mathbf{s})} \gamma_r^{n_r(\mathbf{s})} \quad (1)$$

where $\gamma_f, \gamma_s, \gamma_d > 0$ and $\gamma_o, \gamma_r \in (0, 1)$, are the model parameters. Stoica et al. recommend $\gamma_f < \gamma_s < \gamma_d$, in order to favor configurations containing more connected segments than free ones. The sufficient statistics $n_f(\mathbf{s})$, $n_s(\mathbf{s})$, $n_d(\mathbf{s})$, $n_o(\mathbf{s})$, $n_r(\mathbf{s})$ respectively represent the number of 'free' segments, the number of segments with a single one of its endpoints near another segment endpoint, the number of segments with both extremities connected, the number of pairs of segments crossing at too sharp angles, and the number of pairs that are disoriented. Thus, there are penalties attached to each free and singly connected segment, as well as to each sharp crossing and to every disagreement in orientation. For more details on the model and its applications to network extraction see [11], and [9] where the authors prove existence and Ruelle stability of p and establish various Markov properties.

2. METROPOLIS–HASTINGS ALGORITHMS

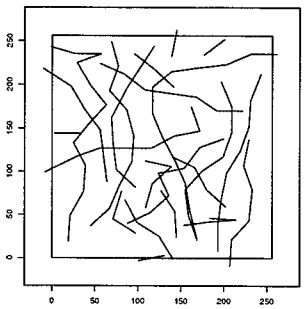
The Candy model (1) is too complicated to sample from directly. Hence, we apply Markov chain Monte Carlo techniques [6] to construct a Markov chain which has the Candy model π as its equilibrium distribution. Here we use the Metropolis–Hastings sampler, a flexible proposal-acceptance technique that is well adapted to point processes [3, 7]. In its generic form, the transition proposals are uniformly distributed births and deaths. The acceptance probabilities are based on the likelihood ratio of the new state compared to the old one. Due to the results in [2], the algorithm converges in total variation to π for π-almost all initial configurations provided. The theorem applies equally to any other pair of strictly positive birth and death kernels.

In order to improve mixing, we incorporate transitions that are tailor-made for the Candy model. Thus, we include a birth kernel that tends to add a segment in order to prolongate the current network. The idea is that when adding a segment, preference should be given to positions that 'fit' the current configuration. More specifically, a new seg-

0-7803-7011-2/01/$10.00 ©2001 IEEE

ment might be positioned in such a way that it is connected to an endpoint of a segment in the configuration, see [9]. For computational convenience, we only connect to segment endpoints that are sufficiently far from the boundary of K.

Another option is to include transition types other than births and deaths. For instance in [2] change transitions that do not alter the number of segments are described.There are many valid choices for the proposal kernel. For instance, we may shift a segment center a bit, modify the orientation and/or the length, or even discard a segment altogether and generate a new one randomly. For more details see [9].



Model parameters
$\gamma_f = 0.0002$
$\gamma_s = 0.05$
$\gamma_d = 12.2$
$\gamma_o = 0.08$
$\gamma_r = 0.08$

Sufficient statistics
$n_f = 4$
$n_s = 34$
$n_d = 63$
$n_o = 12$
$n_r = 9$

Fig. 1. Realization (top) of the Candy model given by the parameters in the middle table. The observed values of the sufficient statistics are listed below.

In Figure 1 we present a sample of the Candy model, its parameters and the observed values of the sufficient statistics. We carried out 2×10^7 iterations. The sufficient statistics were taken every 10^3 iterations. The point space is $K = [0, 256] \times [0, 256]$ while marks take values in $[30, 40] \times [0, \pi)$.The weights of the different transition kernels were fixed empirically. The Candy model is very complex, hence it is difficult to assess convergence. However, we may analyze the evolution of the cumulative means of the sufficient statistics during the simulation. These are plotted in Figure 2.

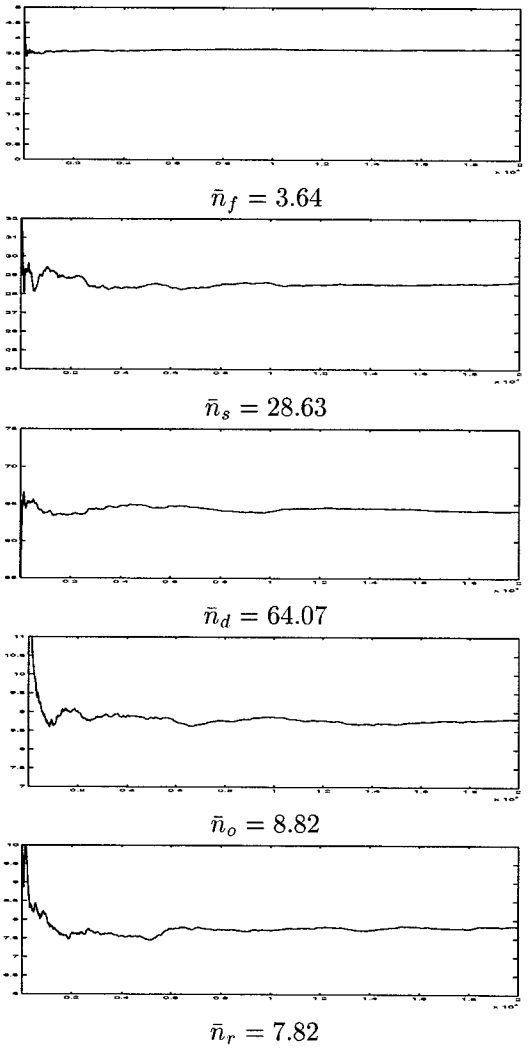


Fig. 2. Evolution of the empirical moments of the sufficient statistics during the simulation of the Candy model. The cumulative means $\bar{n}_f, \bar{n}_s, \bar{n}_d, \bar{n}_o, \bar{n}_r$ (from top to bottom) are plotted as a function of the number of iterations.

3. MAXIMUM LIKELIHOOD ESTIMATION

The Candy model (1) is a five–parameter exponential family

$$p_\omega(\mathbf{s}) = \alpha(\omega) \exp\left[t(\mathbf{s})^T \omega\right] h(\mathbf{s})$$

with canonical sufficient statistic

$$t(\mathbf{s}) = (n_f(\mathbf{s}), n_s(\mathbf{s}), n_d(\mathbf{s}), n_o(\mathbf{s}), n_r(\mathbf{s}))^T,$$

parameter vector

$$\omega = (\log\gamma_f, \log\gamma_s, \log\gamma_d, \log\gamma_o, \log\gamma_r)^T,$$

and $h(\mathrm{s}) = \prod_{i=1}^{n} \exp\left[\frac{l_i - l_{\max}}{l_{\max}}\right]$. Using the importance sampling ideas outlined in [4, 5] the ratio of normalizing constants can be expressed as

$$\alpha(\omega_0)/\alpha(\omega) = E_{\omega_0} \exp\left[t(S)^T(\omega - \omega_0)\right]$$

and the log likelihood ratio with respect to some reference value ω_0 can be written as

$$\begin{aligned} l(\omega) &= \log \frac{p_\omega(\mathrm{s})}{p_{\omega_0}(\mathrm{s})} \\ &= t(\mathrm{s})^T(\omega - \omega_0) - \log E_{\omega_0} \exp\left[t(S)^T(\omega - \omega_0)\right]. \end{aligned} \tag{2}$$

The score equations $\nabla l(\omega) = t(\mathrm{s}) - E_\omega T(S)$ and Fisher information matrix $I(\omega) = -\nabla^2 l(\omega) = \mathrm{Var}_\omega t(S)$ are easily derived, so that under the maximum likelihood estimator $\hat{\omega}$, the expected values of the sufficient statistics must be equal to the observed values. Now, since the covariance matrix of $t(S)$ is positive definite, (2) is concave in ω. Therefore, provided the score equations have a solution $\hat{\omega}$ in $\mathbb{R} \times \mathbb{R}^4_-$, a unique maximum likelihood estimator exists and equals $\hat{\omega}$. Otherwise, a maximum may be found on the boundary of the parameter space.

Numerically, the expectation in (2) can be approximated [4, 5] by its Monte Carlo counterpart

$$\frac{1}{n}\sum_{i=1}^{n} \exp\left[t(S_i)^T(\omega - \omega_0)\right]$$

based on a single sample $S_1, \ldots, S_n$ from p_{ω_0}.

Considering the true unknown MLE $\hat{\omega}$, due to [2, Theorem 7] the Monte Carlo maximum likelihood estimator is consistent and satisfies the central limit theorem :

$$\sqrt{(n)}\,(\hat{\omega}^n - \hat{\omega}) \to \mathcal{N}(0, I(\hat{\omega})^{-1}\Sigma I(\hat{\omega})^{-1})$$

where Σ is the asymptotic covariance matrix of the normalized Monte Carlo score $\sqrt{(n)}\nabla l_n(\hat{\omega})$ and $I(\hat{\omega})$ denotes the Fisher information matrix at the maximum likelihood estimator.

However, the method described above relies on a reference value ω_0 that is not too far from the maximum likelihood estimator. Here we used the iterative gradient method [1].

$$\begin{cases} l_n(\omega_k + \rho(\omega_k)\nabla l_n(\omega_k)) = \\ \quad = \max_{\rho\in\mathbb{R}} l_n(\omega_k + \rho\nabla l_n(\omega_k)) \\ \omega_{k+1} = \omega_k + \rho(\omega_k)\nabla l_n(\omega_k) \end{cases} \tag{3}$$

to find a reasonable value. Here $\rho(\omega_k)$ is the optimal step, which is computed using a one-dimensional minimization of the likelihood function.

We implemented the procedure for the data of Figure 1. Starting with sone arbitrary initial values (see Figure 3, first column) we ran (3) for 1000 steps to obtain the vector ω_0 listed in the second column of Figure 3. Based on a sample of size $n = 2\times10^7$ from p_{ω_0}, we calculated the Monte Carlo approximation $l_n(\omega)$, cross sections of which are shown in Figure 5. The maximum of $l_n(\omega)$ is located at $\hat{\omega}^n$ as listed in Figure 3 (third column).

In Figure 4 we show the asymptotic standard deviation of the true MLE, and the Monte Carlo Standard Error (MCSE) which approximates the difference between the unknown MLE and its Monte Carlo approximation. We notice that by increasing n, we can make the MCSE negligible.

Initial parameters	Iterative method	Monte Carlo MLE
$\omega_f^i = -9.5$	$\hat{\omega}_f^0 = -8.37$	$\hat{\omega}_f^n = -8.32$
$\omega_s^i = -4.0$	$\hat{\omega}_s^0 = -2.74$	$\hat{\omega}_s^n = -2.73$
$\omega_d^i = 1.5$	$\hat{\omega}_d^0 = 2.46$	$\hat{\omega}_d^n = 2.47$
$\omega_o^i = -3.5$	$\hat{\omega}_o^0 = -2.13$	$\hat{\omega}_o^n = -2.17$
$\omega_r^i = -3.5$	$\hat{\omega}_r^0 = -2.42$	$\hat{\omega}_r^n = -2.42$

Fig. 3. Estimation of the parameters for the data of Figure 1.

Asymptotic standard deviation of MLE	MCSE
0.51	0.002
0.23	0.003
0.17	0.001
0.30	0.002
0.31	0.005

Fig. 4. Estimation errors.

4. CONCLUSION AND FUTURE WORK

In practice, the main challenges in working with point processes are the following: to build appropriate moves, to find the optimal way of combining them into a simulation algorithm, and to carry out statistical inference. Here, we have built a Metropolis–Hastings sampler, that combines uniform birth and death proposals that guarantee the convergence of the Markov chain to the target equilibrium distribution (1) with transitions designed to exploit specific characteristics of the model, in our case connectivity properties.

The main application of the Candy model is that of thin network extraction. This was the topic of [10, 11], where results were obtained using fixed pareameters as well as approximations to the Metropolis–Hastings proposal kernels and acceptance probabilities. The results here, and in [9],

remove the need for approximate sampling, and may be a starting point for unsupervised network extraction.

5. REFERENCES

[1] X. Descombes, R. D. Morris, J. Zerubia and M. Berthod. Estimation of Markov random field prior parameters using Markov Chain Monte Carlo maximum likelihood. *IEEE Transactions on Image Processing* **8**, 954-963, 1999.

[2] C.J. Geyer. On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B* **56**, 261–274, 1994.

[3] C.J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics* **21**, 359–373, 1994.

[4] C.J. Geyer. Likelihood inference for spatial point processes. In O. Barndorff-Nielsen, W.S. Kendall, and M.N.M. van Lieshout, editors, *Stochastic geometry, likelihood, and computation*, CRC Press/Chapman and Hall, Boca Raton, 1999.

[5] C.J. Geyer and E.A. Thompson. Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B* **54**, 657–699, 1992.

[6] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, 1996.

[7] P.J. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711–732, 1995.

[8] M.N.M. van Lieshout. *Markov point processes and their applications*. Imperial College Press/World Scientific Publishing, London/Singapore, 2000.

[9] M.N.M. van Lieshout and R.S. Stoica. The Candy model revisited : Markov properties and inference. CWI Research Report, 2001.

[10] R. Stoica. *Processus ponctuels pour l'extraction des réseaux linéiques dans les images satellitaires et aériennes*. PhD Thesis (in French), University of Nice–Sophia Antipolis, 2001.

[11] R. Stoica, X. Descombes and J. Zerubia. A Gibbs point process for road extraction in remotely sensed images. Research Report 3923, INRIA Sophia Antipolis, 2000.

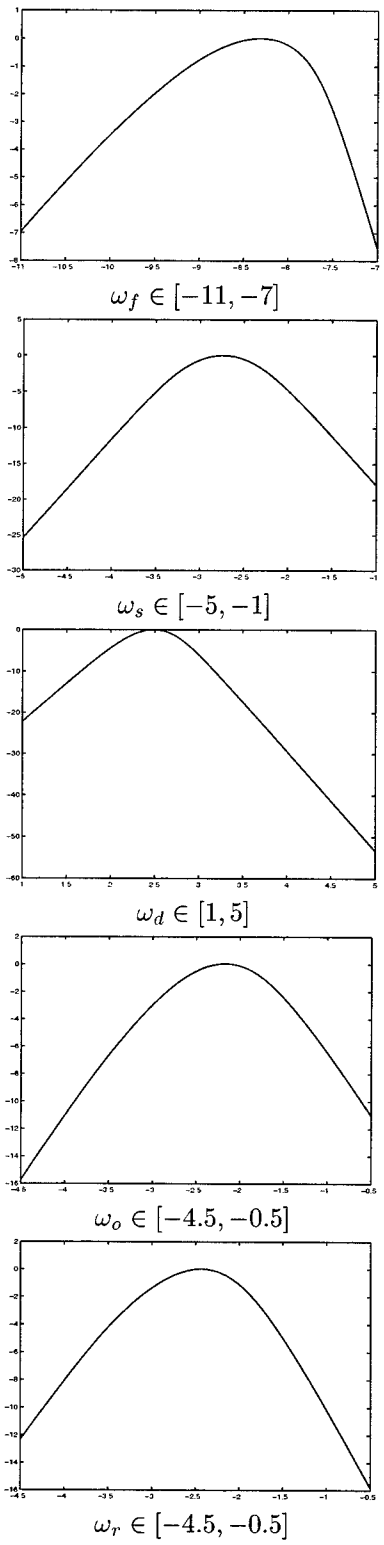


Fig. 5. Monte Carlo approximation of the log likelihood function. The X axis represents the variation of a single component. The Y axis represents the values of the Monte Carlo log likelihood with all other components of $\hat{\omega}^0$ fixed.

RESAMPLING BASED TECHNIQUES FOR SOURCE DETECTION IN ARRAY PROCESSING

Ramon F. Brcich and Abdelhak M. Zoubir

Australian Telecommunications Research Institute & School of Electrical and Computer Engineering
Curtin University of Technology, GPO Box U1987, Perth WA 6845, Australia.
Email : {r.brcich, zoubir}@ieee.org

ABSTRACT

The source detection problem in array processing can be considered a test for equality of eigenvalues. This approach is implemented through a multiple hypothesis procedure which compares all pairwise differences between eigenvalues. A resampling procedure is used to estimate the null distributions of the test statistics, an advantage for small sample sizes or non-Gaussian signals since traditional techniques such as the MDL assume Gaussianity. Simulations show the increased performance of the test compared to the MDL for small samples or non-Gaussian signals, with a noticeable improvement over the more accurate sphericity test.

Keywords : array source detection, resampling, bootstrap, multiple hypothesis tests, model selection

1. INTRODUCTION

Source detection is an important first step in array processing. Model order selection procedures based on information theoretic criteria such as Rissanen's minimum description length (MDL) [1] are well known. Alternatively, the problem can be cast as a hypothesis test for equality of the smallest sample eigenvalues, the sphericity test is such a procedure [2]. Both these methods are based on large samples and Gaussian signals, otherwise their behaviour may be unpredictable since the distribution of the sample eigenvalues can be sensitive to departures from Gaussianity [3].

Here a multiple hypothesis procedure is used to compare all pairwise differences of the eigenvalues to test for the number of sources. A similar approach has demonstrated significant potential for improved performance over the MDL [4]. Though conceptually similar to the sphericity test, we estimate the finite sample distributions of the test statistic by resampling, rather than using asymptotic approximations. In these cases we can then achieve improved performance.

Bias in the sample eigenvalues has a significant effect on this inference, notably when population eigenvalues are not well separated, as for small samples or low SNR. A bias correction based on the expectation of the sample eigenvalues is proposed. Performance is similar to resampling methods, but at a greatly reduced computational cost. With bias estimates incorporated into detection procedure it is possible to obtain an improvement in performance over the MDL and the sphericity test, itself a general improvement over the MDL.

This work was in part supported by the Australian Telecommunications Cooperative Research Centre (AT-CRC).

2. DATA MODEL

The setting for the source detection problem is as follows. n snapshots of i.i.d. zero mean complex data are received from a p element array,

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{v}(t), \qquad t = 1, \ldots, n$$

where $\boldsymbol{A}$ is the $p \times q$ array steering matrix, $\boldsymbol{s}_n$ is a q $(q < p)$ vector valued white source signal and $\boldsymbol{v}_n$ is noise with covariance $\sigma^2\boldsymbol{I}$. Assuming $\boldsymbol{s}(t)$ and $\boldsymbol{v}(t)$ are independent, the array covariance is

$$\boldsymbol{R} = \mathsf{E}\left[\boldsymbol{x}(t)\boldsymbol{x}^H(t)\right] = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H + \sigma^2\boldsymbol{I}$$

where $\boldsymbol{R}_s$ is the covariance of the sources. The eigenvalues of $\boldsymbol{R}$ are denoted

$$\lambda_1 > \cdots > \lambda_q > \lambda_{q+1} = \cdots = \lambda_p = \sigma^2, \tag{1}$$

so that the smallest $p - q$ population eigenvalues are equal. These will be referred to as multiple or noise eigenvalues, the others as distinct or source eigenvalues. The problem is then one of determining the multiplicity of the smallest sample eigenvalues,

$$l_1 > \cdots > l_p > 0,$$

of the sample array covariance,

$$\hat{\boldsymbol{R}} = \frac{1}{n-1}\sum_{t=1}^{n}\boldsymbol{x}(t)\boldsymbol{x}(t)^{\mathrm{H}}.$$

We now present the proposed multiple hypothesis procedure for source detection.

3. APPLICATION TO SOURCE DETECTION

From (1) it follows that we must test for equality of eigenvalues, by considering all possible pairwise comparisons between the eigenvalues we arrive at this set of hypotheses,

$$\begin{aligned} \mathsf{H}_{ij} &: \lambda_i = \lambda_j, \quad i = 1, \ldots, p-1, \quad j > i, \\ \mathsf{K}_{ij} &: \lambda_i \neq \lambda_j. \end{aligned}$$

A hypothesis test for equality of the smallest $p-k$ eigenvalues can be obtained by combining these pairwise comparisons to give the new hypotheses H_k, $k = 0, \ldots, p-2$,

$$\begin{aligned} \mathsf{H}_k &= \cap_{i,j}\mathsf{H}_{ij}, \qquad i = k+1, \ldots, p-1, \quad j > i, \\ \mathsf{K}_k &= \cup_{i,j}\mathsf{K}_{ij}. \end{aligned}$$

Acceptance of all pairwise comparisons in H_k implies the smallest $p-k$ eigenvalues are equal, or that there are k sources. Given p-values for the pairwise comparisons, a sequentially rejective Bonferonni (SRB) procedure [5] is employed to test the hypotheses H_k and estimate q in the following manner,

1. Set k=0.
2. Test H_k.
3. If H_k is accepted then set $\hat{q} = k$ and stop.
4. If H_k is rejected and $k < p-1$ then set $k \rightarrow k+1$ and return to step 2. Otherwise set $\hat{q} = p-1$ and stop.

The SRB procedure maintains the global level of significance, α, defined as the probability that at least one of the hypotheses is rejected, given all are true. In this test the global null is H_0, that all eigenvalues are equal.

P-values for H_{ij} are found using the bootstrap [6]. This resampling technique is used for several reasons, it avoids the need to know the distribution of the test statistic under the null and is valid for small samples and non-Gaussian data.

These advantages are quite important when working with eigenvalues since their distribution is too complex for general use [7], while asymptotic expansions [8] may not be valid for the small sample sizes considered. Asymptotic approximations developed for non-Gaussian cases require knowledge of the higher order moments of the data, which are difficult to estimate well for small sample sizes [3, 9].

The basic premise of the proposed test is that differences between noise eigenvalues are small. For finite samples the eigenvalues are biased, the amount of bias increasing as the separation of population eigenvalues decreases. Thus differences between multiple eigenvalues will be shifted away from zero. To correct for this shift we must estimate the bias of all eigenvalues. We now consider several methods for bias estimation.

4. BIAS ESTIMATION

4.1. Lawley's Expansion

Lawley developed an expression for the expectation of the distinct eigenvalues by considering the propagation of error from the sample covariance to the eigenvalues for Gaussian data [12]. From this the bias in l_i is estimated as

$$Bias_{\mathsf{LAW}}(l_i) = l_i \left(\frac{1}{n} \sum_{j=1,\, j\neq i}^{q} \frac{l_j}{l_i - l_j} + \frac{p-q}{n} \frac{\sigma^2}{l_i - \sigma^2} \right), \quad (2)$$

for $i = 1, \ldots, q$, where σ^2, the population multiple eigenvalue, is replaced with its maximum likelihood estimate under Gaussianity,

$$\hat{\sigma}^2 = \frac{1}{p-q} \sum_{j=q+1}^{p} l_j. \quad (3)$$

Bias in the distinct sample eigenvalues is of order $O(n^{-1})$. Though a similar expression for multiple eigenvalues does not exist, extensive simulations have shown that the bias is of order $O(n^{-1/2})$.

After applying this bias correction, the distinct eigenvalues have a bias of order $O(n^{-2})$, while $\hat{\sigma}^2$ is unbiased under Gaussianity. Note these corrections can only be applied if q is known and even then individual multiple eigenvalues cannot be corrected.

The estimate (2) is valid when the difference between successive distinct eigenvalues is large relative to the standard error, which is of order $O(n^{-1/2})$. If this condition is not fulfilled and $\lambda_i \approx \lambda_j$ for $i \neq j$, the variance of this estimator increases quickly. This follows intuitively by examining the denominator in the terms of the summation of (2). Similarly, if (2) was unknowingly applied to multiple eigenvalues the results will be unpredictable as the assumption of well separated population eigenvalues is invalid.

4.2. A Robust Bias Estimate

Based on Lawley's expansion we propose a bias estimate to overcome the aforementioned problems by taking a binomial expansion in the denominator of the summand of (2) and truncating to a finite number of terms. For simplicity, assume that all the population eigenvalues are distinct. Then the bias estimate for l_i becomes

$$Bias_{\mathsf{LBE}}(l_i) = \begin{cases} \frac{1}{n} \sum_{j=1,\, j\neq i}^{p} l_j \sum_{k=0}^{K} \left(\frac{l_j}{l_i}\right)^k, & l_j < l_i, \\ -\frac{1}{n} \sum_{j=1,\, j\neq i}^{p} l_i \sum_{k=0}^{K} \left(\frac{l_i}{l_j}\right)^k, & l_j > l_i, \end{cases}$$

for some suitable K. If required, the upper limit on the outer summation can be changed to q and the term corresponding to multiple eigenvalues from (2) included. Setting K to a moderate value will retain the bias correction properties while guarding against large increases in variance when the population eigenvalues are not well separated or multiple eigenvalues are present. A value of $K = 25$ was found to be acceptable. Hence $Bias_{\mathsf{LBE}}$ may be applied without any knowledge of the multiple eigenvalues by assuming q to be p and this bias estimate can be applied blindly to correct all eigenvalues irrespective of multiplicity issues.

An example with multivariate Gaussian data is shown in Figures 1 and 2 where the largest sample eigenvalue is considered and both corrections are applied. The data has a diagonal covariance matrix with population eigenvalues $(1.15, 1.1, 1.05, 1)'$. While the mean value of the corrected eigenvalues are very similar there is a notable decrease in the variance when using $Bias_{\mathsf{LBE}}$. For small sample sizes the decrease is significant as the separation of population eigenvalues is of the same order as the standard error.

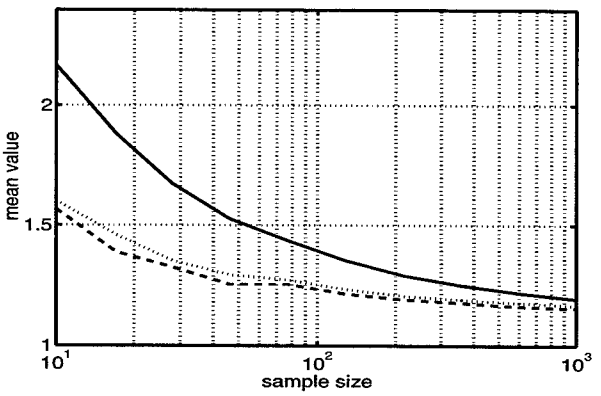


Figure 1: Mean of the largest sample eigenvalue with no bias estimation $(-)$, $Bias_{\mathsf{LAW}}$ $(--)$ and $Bias_{\mathsf{LBE}}$ $(\cdots)$, versus sample size for multivariate Gaussian data, $p = 4$, with diagonal covariance matrix and population eigenvalues $(1.15, 1.1, 1.05, 1)'$.

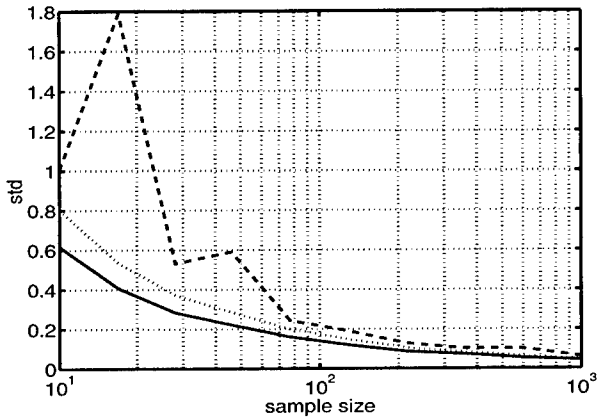


Figure 2: Standard Deviation of the largest sample eigenvalue with no bias estimation $(-)$, $Bias_{\text{LAW}}$ $(--)$ and $Bias_{\text{LBE}}$ $(\cdots)$, versus sample size for the same scenario as Figure 1.

4.3. Resampling Methods

Resampling methods are valid in small sample scenarios, for non-Gaussian data and may be applied blindly, that is, without having to know whether multiple eigenvalues are present. The two methods considered are the jackknife [6] and subsampling [10].

4.4. Jackknife Bias Estimation

In this jackknife procedure a jackknife data set is created by deleting a single sample from the original data set. Given a data set of size n there are n possible unique jackknife data sets. For each of these we recompute the statistic of interest yielding the estimates $l_i^*(b), b = 1, \ldots, n$. The jackknife estimate of bias in l_i is

$$Bias_{\text{JCK}}(l_i) = (n-1)\left(\frac{1}{n}\sum_{b=1}^{n} l_i^*(b) - l_i\right),$$

where l_i was estimated from the entire sample.

4.5. Subsampling Bias Estimation

Subsampling is a generalisation of the jackknife, where instead of methodically removing one sample at a time, d samples are removed to give a subsample of size $s = n - d$. The number of possible subsamples, $n!/s!(n-s)!$, may be very large, so a smaller number of B subsamples are chosen at random.

As the subsamples are smaller in size than the original data set, rescaling is required to correct the subsampling estimate. Assume that eigenvalue bias is proportional to $1/\tau_n$, τ_n being a function of n. Then the bias estimate from a subsample of size b is proportional to $1/\tau_b$, to apply this to the original statistic rescaling is required. While the function τ_n is problem dependent, it is usually of the form $n^\beta, \beta \in (0,1)$. The subsampling estimate of bias is

$$Bias_{\text{SUB}}(l_i) \approx \frac{\tau_r}{\tau_n}\left(\frac{1}{B}\sum_{b=1}^{B} l_i^*(b) - l_i\right),$$

where $r = sn/(n-s)$. If all possible subsamples are used then the approximation is replaced with equality. The reason τ_r is used instead of τ_b is because resampling is performed without replacement from a finite population, as opposed to with replacement from an infinite population, which is assumed with τ_b [11]. As an approximation $\beta \approx 1$ for distinct eigenvalues and $\beta \approx 0.5$ for multiple eigenvalues. Since the presence of either is unknown we must estimate β for each eigenvalue, however, a simple way to avoid this is to set $s = n/2$, so that $\tau_r/\tau_n = 1$, independent of β.

Note the jackknife and subsampling are valid in a wider variety of situations than other resampling techniques such as the bootstrap [6], with subsampling being the most widely applicable.

An example with multivariate Gaussian data is shown in Figures 3 and 4 for the largest sample eigenvalue where both $Bias_{\text{JCK}}$ and $Bias_{\text{SUB}}$ are applied. The data has diagonal covariance matrix with population eigenvalues $(4,3,2,1)'$. For $Bias_{\text{SUB}}$, $B = 100$ subsamples of size $s = n/2$ were chosen at random. Both methods behave very similarly and though $Bias_{\text{LBE}}$ is not shown here, it too produces very similar results in terms of the average bias estimate. Further experimentation has shown the subsampling estimator tends to have a slightly lower MSE.

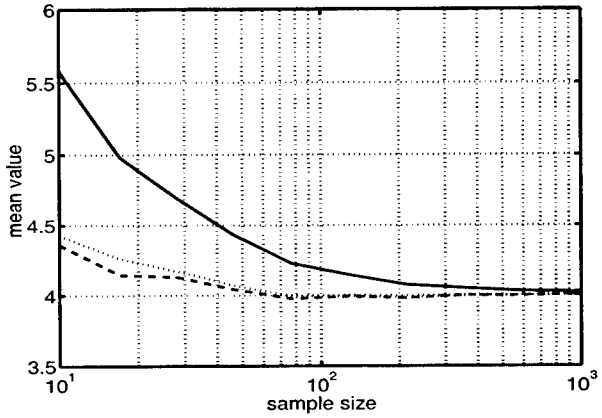


Figure 3: Mean of the largest sample eigenvalue with no bias estimation $(-)$, $Bias_{\text{JCK}}$ $(--)$ and $Bias_{\text{SUB}}$ $(\cdots)$, versus sample size for multivariate Gaussian data, $p = 4$, with diagonal covariance matrix and population eigenvalues $(4,3,2,1)'$.

5. SIMULATIONS

Figure 5 shows detection rates for a $p = 4$ element array with $q = 3$ sources at $10°$, $30°$ and $50°$ from broadside at SNR's of -2, 2 and 6 dB respectively. All signals are Gaussian and the global level for both the resampling and sphericity tests was set to $\alpha = 0.02$. Parameters for resampling bias correction were the same as those used in 4.3, for the bootstrap $B = 200$ resamples were taken. In this scenario there is an improvement in performance over both the MDL and sphericity test, most noticeable for small sample size. In Figure 6 we have a non-Gaussian scenario where the source is Laplacian and the SNR of the second source is varied for a sample size of $n = 50$. Again, there is an improvement over existing methods. Similar results are obtained for other non-Gaussian distributions, such as Gaussian mixtures.

Additional scenarios have shown that there is an improvement over the MDL in nearly all cases and a comparable or possibly superior performance to the sphericity test. This has to be weighed against the increase in computational complexity which increases

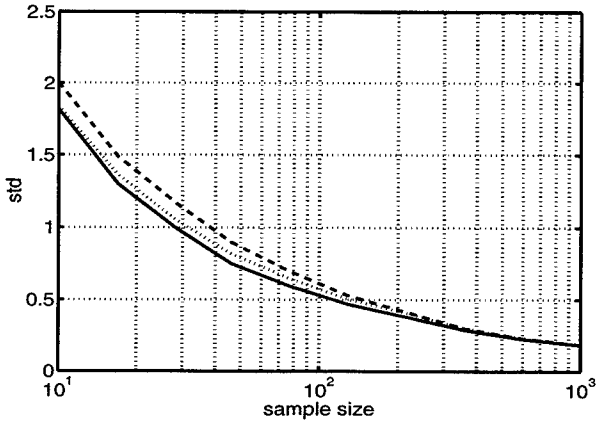


Figure 4: Standard Deviation of the largest sample eigenvalue with no bias estimation (−), $Bias_{\text{JCK}}$ (− −) and $Bias_{\text{SUB}}$ (· · ·), versus sample size for the same scenario as Figure 3.

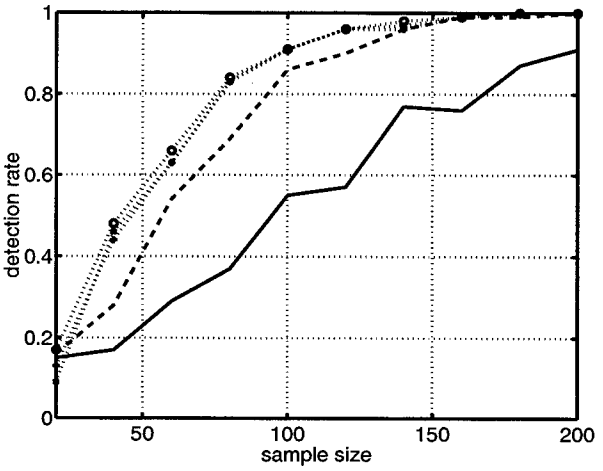


Figure 5: Detection rates versus sample size. MDL (−), sphericity test (− −), bootstrap methods using bias correction, $Bias_{\text{LBE}}$ (·⊙·), $Bias_{\text{JCK}}$ (·×·), $Bias_{\text{SUB}}$ (·+·).

in direct proportion to the number of times the data is resampled. Since $Bias_{\text{LBE}}$ does not use resampling it involves less computation than resampling methods. For the $B = 100$ resamples used here for bias estimation this represents a significant saving.

6. CONCLUSION

The source detection problem was approached as a multiple hypothesis test for equality of eigenvalues. It was shown that for the cases of interest, such as small sample size, bias in the sample eigenvalues is non-negligible and should be corrected for when carrying out the test. An improved bias estimate was proposed which overcomes the need to know the multiplicity of the eigenvalues, performing well in spite of their presence. It is less computationally intensive than resampling methods while achieving similar performance. Results show that the proposed procedure can yield improved performance compared to the MDL and sphericity test in the non-Gaussian and small sample cases.

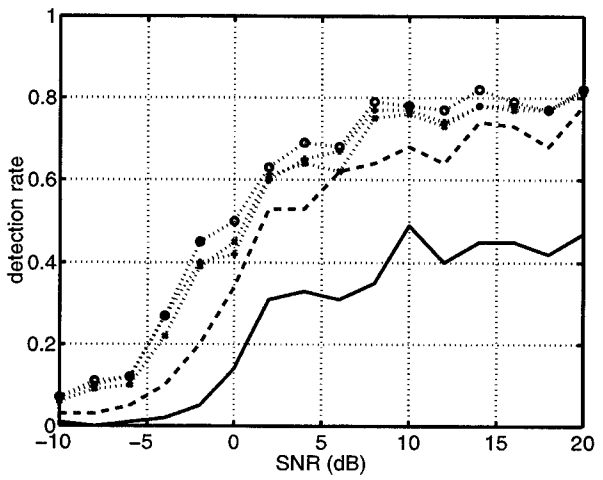


Figure 6: Detection rates versus SNR (dB) for Laplace sources and non-Gaussian noise. MDL (−), sphericity test (− −) and bootstrap methods using bias correction, $Bias_{\text{LBE}}$ (·⊙·), $Bias_{\text{JCK}}$ (·×·), $Bias_{\text{SUB}}$ (·+·).

7. REFERENCES

[1] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. ASSP.*, vol. ASSP-33, no. 2, pp. 387–92, April 1985.

[2] D. Williams and D. Johnson, "Using the sphericity test for source detection with narrow-band passive arrays," *IEEE Trans. ASSP.*, vol. 38, no. 11, pp. 2008–14, November 1990.

[3] C. Waternaux, "Asymptotic distribution of the sample roots for a nonnormal population," *Biometrika*, vol. 63, no. 3, pp. 639–45, 1976.

[4] R. Brcich, P. Pelin and A. Zoubir, "Detection of sources in array processing using the bootstrap," in *Proc. SSAP 2000*, Pocono Manor, PA, USA, August 2000, pp. 448–452.

[5] P. Westfall and S. Young, *Resampling-Based Multiple Testing*, John Wiley & Sons, 1993.

[6] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, 1993.

[7] A. James, "The distribution of the latent roots of the covariance matrix," *The Annals of Math. Stat.*, vol. 31, pp. 151–8, 1960.

[8] T. W. Anderson, "Asymptotic theory for principal component analysis," *Annals of Statistics*, vol. 34, pp. 122–48, 1963.

[9] Y. Fujikoshi, "Asymptotic expansions for the distributions of the sample roots under nonnormality," *Biometrika*, vol. 67, no. 1, pp. 45–51, 1980.

[10] J. Romano, D. Politis and M. Wolf, *Subsampling*, Springer-Verlag, 1999.

[11] D. Politis, "Computer-intensive methods in statistical analysis," *IEEE SP. Mag.*, pp. 39–55, January 1998.

[12] D. Lawley, "Tests of significance for the latent roots of covariance and correlation matrices," *Biometrika*, vol. 43, pp. 128–136, 1956.

THE SEQUENTIAL MCMC FILTER: FORMULATION AND APPLICATIONS

Dominic S. Lee and Nicholas K.K. Chia

DSO National Laboratories
20 Science Park Drive
Singapore 118230

ABSTRACT

We consider the general signal-processing problem of learning about certain attributes of interest from measurements. These attributes, which may be time-varying (dynamic) or time-invariant (static), can be anything that are relevant to the physical processes that produce the measurements. In statistical signal processing, imperfections or uncertainties in the physical processes are described using probability models, and the complete probabilistic solution to the problem is given by the distribution of the attributes conditioned on all available measurements (the posterior distribution).

We describe an algorithm for computing this solution, especially in situations with many measurements or low signal-to-noise ratios. The algorithm combines sequential importance sampling (SIS) and Markov chain Monte Carlo (MCMC) so as to achieve computational efficiency and stability. MCMC is performed sequentially for batches of measurements whose sizes are determined adaptively, hence the name *sequential MCMC filter*. For measurements within a batch, SIS is performed. Thus, bigger batch sizes mean that MCMC is performed less frequently. SIS is computationally efficient but with a finite Monte Carlo sample size, stability is not guaranteed indefinitely. MCMC is therefore needed from time to time to "refresh" the Monte Carlo sample, eliminating any errors that may have accumulated from the SIS steps. When MCMC is performed, it does not start from scratch but uses the most recent Monte Carlo sample from SIS to construct the proposal distribution. Adaptive batch sizing is based on a Kullback-Leibler distance that is easy to compute. By extending the algorithm to multiple models, the sequential MCMC filter can deal simultaneously with the dual pillars of statistical signal processing, namely detection (more generally, model selection) and parameter estimation.

We discuss general uses of the sequential MCMC filter, and demonstrate its use for simultaneous weak signal detection and parameter estimation in a real-data experiment.

1. INTRODUCTION

A goal of signal processing is to learn about certain attributes of interest from measurements. These attributes can be anything that is relevant to the physical processes that produce the measurements. For example, they may be signal attributes such as amplitude, frequency or phase; or noise attributes such as noise power; or image attributes such as intensities or edges; and so on. In statistical signal processing, imperfections or uncertainties in the physical processes are described using probability models. Whilst the complete description of probabilistic objects is provided by distributions, statistical signal processing solutions have predominantly been moment-based. Typically, these solutions make simplifying approximations, such as linearization and the use of convenient distributions, so as to be computable under various hardware constraints. Recently, interest in distribution-based statistical signal processing solutions has grown due to rapid advances in computer technology.

Together with the shift in focus from moments to distributions is a shift towards the Bayesian paradigm. This is natural because the Bayesian framework is the mathematically consistent and coherent framework for updating distributions. This theoretical impetus is steadily being reinforced by the increasing awareness of the tangible advantages of Bayesian techniques. For example, the ease with which prior information and domain knowledge can be incorporated; the finite-sample optimality properties of Bayesian solutions; the relevance of the solutions to the problems at hand (as opposed to "in the long run" or "on the average"); and the built-in Ockham effect (which penalizes model complexity and hence prevents model over-fitting) offered by Bayesian model selection. Furthermore, in certain applications, the Bayesian framework can unify signal-processing tasks that are conventionally regarded as separate. This unifying feature of the Bayesian framework may be referred to as *simultaneity*.

Bayesian ideas are not new but until about a decade ago, they remained largely of academic interest due to the difficulty of computing Bayesian solutions for real-world problems. Such problems often involve complex models that have been painstakingly derived by the domain experts. Today, the tables seem to have turned. In fact, the Bayesian approach is emerging as *the* one that can effectively handle complex models. Not only are these complex models no longer a hindrance, the Bayesian approach actually preserves them (no model simplification is made) and uses them to advantage (valuable domain knowledge). All this has come about because the availability of powerful computers has encouraged research into computer-intensive methods for Bayesian computations. These are essentially Monte Carlo or sample-based methods that represent a distribution of interest by a sufficiently large number of computer-generated sample points. Of these methods, the one with the greatest impact, making Bayesian computations with complex models possible, is Markov chain Monte Carlo (MCMC) (see [1] for a recent review). James Berger, a renowned and respected statistician, goes so far as to say [2], "The Bayesian 'machine,' together with MCMC, is arguably the most powerful mechanism ever created for processing data and knowledge." In signal processing, as in other application domains, we are learning how to utilize the full power of this "mechanism".

0-7803-7011-2/01/$10.00 ©2001 IEEE

2. THE SEQUENTIAL MCMC FILTER

We begin with data in the form of a sequence of measurements, $x_1, x_2, \ldots \in \Re^{d_x}$, where d_x is the measurement dimension. Our goal is to use the data to estimate, at each time-step, certain unknown attributes of interest, which are pertinent to the physical processes that produce the measurements. Some of these attributes may be time-invariant. We group these together and represent them by the *parameter* vector, $\psi \in \Re^{d_\psi}$, with dimension d_ψ. There may also be attributes that are time-varying. We represent these by a sequence of *state* vectors, $\theta_1, \theta_2, \ldots \in \Re^{d_\theta}$, where d_θ is the state dimension. We define the *cumulative state* at time-step k to be $\Theta_k = (\psi, \theta_1, \theta_2, \ldots, \theta_k)$, with $\Theta_0 = \psi$. By letting $X_k = (x_1, \ldots, x_k)$, our problem then is to estimate Θ_k using X_k, for $k = 1,2,\ldots$. For each time-step, the complete probabilistic solution is given by the joint distribution of Θ_k conditionally given X_k.

Henceforth, we assume that all distributions are continuous so that their associated densities exist. In the discrete case, mass functions should replace densities and summations should replace integrals wherever appropriate. For generic random vectors y and z, we use $p(y)$ to denote the density of y, and let $p(y|z)$ denote the conditional density of y given z. With this notation, the solutions that we seek can be written as $p(\Theta_1|X_1)$, $p(\Theta_2|X_2), \ldots$.

Formally, using Bayes' theorem, we have

$$\begin{aligned} p(\Theta_k|X_k) &\propto p(x_k|\Theta_k, X_{k-1}) p(\Theta_k|X_{k-1}) \\ &= p(x_k|\Theta_k, X_{k-1}) p(\theta_k|\Theta_{k-1}, X_{k-1}) p(\Theta_{k-1}|X_{k-1}), \end{aligned} \quad (1)$$

which indicates that the desired solution for the current time-step, k, can be obtained from the solution for the preceding time-step, $k - 1$, by incorporating new information brought in by the current measurement through the likelihood, $p(x_k|\Theta_k, X_{k-1})$. We refer to this sequential updating formula as the *general Bayesian filter* (GBF). An effective way to implement the GBF is to use adequately large Monte Carlo samples (random or weighted) to represent distributions. This approach is known as *Monte Carlo filtering* or *particle filtering* or *sequential Monte Carlo*. The ability of a sufficiently large sample to provide an arbitrarily close estimate of a distribution is established by the Glivenko-Cantelli Theorem, which states that the empirical distribution function converges almost surely and uniformly to the true underlying distribution function as the sample size increases (see, for example, [3]). Also, convergence of sample averages of integrable functions of the sample points to their respective expectations follows from the Laws of Large Numbers [3]. In many ways, a Monte Carlo sample representing $p(\Theta_k|X_k)$ makes it easier to conduct inference with $p(\Theta_k|X_k)$.

To illustrate the Monte Carlo filtering idea, suppose we have a weighted sample of size n, $\Theta_{k-1,1}, \ldots, \Theta_{k-1,n}$, with weights $\omega_{k-1,1}, \ldots, \omega_{k-1,n}$, representing $p(\Theta_{k-1}|X_{k-1})$. We denote such a weighted sample by $(\Theta_{k-1,1}, \omega_{k-1,1}), \ldots, (\Theta_{k-1,n}, \omega_{k-1,n})$. Then (1) suggests that one way to get a weighted sample, $(\Theta_{k,1}, \omega_{k,1}), \ldots, (\Theta_{k,n}, \omega_{k,n})$, that represents $p(\Theta_k|X_k)$ is to generate $\theta_{k,j}$ from $p(\theta_k|\Theta_{k-1,j}, X_{k-1})$, augment it to $\Theta_{k-1,j}$ to form $\Theta_{k,j}$, i.e.

$$\theta_{k,j} \sim p(\theta_k|\Theta_{k-1,j}, X_{k-1}), \ \Theta_{k,j} = (\Theta_{k-1,j}, \theta_{k,j}), \quad (2)$$

and then compute its updated weight by

$$\omega_{k,j} \propto p(x_k|\Theta_{k,j}, X_{k-1}) \omega_{k-1,j}. \quad (3)$$

We shall refer to this method of obtaining the desired Monte Carlo sample as *sequential importance sampling* (SIS) [4]. In practice, SIS is easy to use because $p(\theta_k|\Theta_{k-1}, X_{k-1})$ and $p(x_k|\Theta_k, X_{k-1})$ are usually readily available.

A problem that arises with SIS is that with a finite sample, the weights become increasingly skewed over time, adversely affecting the sample's ability to adequately represent the distribution. This phenomenon is known as *sample degeneration* and various schemes have been suggested to mitigate it. In [4], the authors propose a general framework that unifies many existing schemes. They also suggest a generic Monte Carlo filtering algorithm that first checks the skewness of the weights and then performs SIS if they are not too skewed, but otherwise performs SIS with resampling to counter sample degeneration. The skewness check is based on an "effective sample size", which is computed from the coefficient of variation of the weights [5]. With a finite Monte Carlo sample size, the generic algorithm (and hence all of its particular realizations as well) has been shown to delay degeneration but there is no guarantee that the problem is resolved entirely. We have not seen any demonstration of its stability for long measurement sequences or for low SNR, situations that are frequently encountered in signal processing.

Our sequential MCMC filter has a somewhat similar generic structure but the details differ. We perform SIS and check whether the resulting Monte Carlo sample provides an adequate representation of the distribution of interest. If it does, we proceed with SIS for the next time-step; otherwise, we perform MCMC with the desired distribution as target distribution and with a proposal distribution that is constructed from the Monte Carlo sample produced by SIS. Whilst the generic algorithm in [4] counters degeneration by resampling, we use MCMC to avoid the drawbacks of resampling such as increase in random variation in the resulting sample and decrease in diversity of the sample points. The need for a full MCMC from time to time has been alluded to in [6] – it "refreshes" the Monte Carlo sample and removes any approximation errors that may have accumulated from the SIS steps due to the finite sample size. Consequently, the sequential MCMC filter is guaranteed to be stable as long as there are enough resources to perform the MCMC properly. Unlike static MCMC schemes that perform MCMC from scratch at each and every time-step [7], we do not need MCMC at every time-step and we do not start MCMC from scratch but use the most recent Monte Carlo sample from SIS to construct the proposal distribution.

To summarize, our sequential MCMC filter has the following steps:

1. Start of time-step $k+1$: we have $(\Theta_{k,1},\omega_{k,1})$,.... $(\Theta_{k,n_k},\omega_{k,n_k})$ from $p(\Theta_k|X_k)$.
   Set $m=1$.
2. SIS: Obtain $(\Theta_{k+m,1},\omega_{k+m,1}),...,(\Theta_{k+m,n_k},\omega_{k+m,n_k})$ by
   $\theta_{k+m,j} \sim p(\theta_{k+m}|\Theta_{k+m-1,j},X_k)$,
   $\Theta_{k+m,j}=(\Theta_{k+m-1,j},\theta_{k+m,j})$,
   $\omega_{k+m,j} \propto p(x_{k+1},...,x_{k+m}|\Theta_{k+m,j},X_k)\omega_{k,j}$,
   for $j=1,...,n_k$.
3. If $(\Theta_{k+m,1},\omega_{k+m,1}),...,(\Theta_{k+m,n_k},\omega_{k+m,n_k})$ adequately represent $p(\Theta_{k+m}|X_{k+m})$, then
   Set $m=m+1$. Go to Step 2.
   Else
   (a) Construct $\hat{p}(\Theta_{k+m}|X_{k+m})$ using $(\Theta_{k+m,1},\omega_{k+m,1}),...,(\Theta_{k+m,n_k},\omega_{k+m,n_k})$.
   (b) MCMC: Obtain $(\Theta_{k+m,1},\omega_{k+m,1}),...,(\Theta_{k+m,n_{k+m}},\omega_{k+m,n_{k+m}})$ by MCMC with $p(\Theta_{k+m}|X_{k+m})$ as target density and $\hat{p}(\Theta_{k+m}|X_{k+m})$ as proposal density.
   (c) Set $k=k+m$. Go to Step 1.

Notice that in general the size of the Monte Carlo sample need not remain the same, hence the use of n_k and n_{k+m}. For checking whether $(\Theta_{k+m,1},\omega_{k+m,1}),...,(\Theta_{k+m,n_k},\omega_{k+m,n_k})$ adequately represent $p(\Theta_{k+m}|X_{k+m})$ in Step 3, we actually measure how well $p(\Theta_{k+m}|X_k)$ "predicts" $p(\Theta_{k+m}|X_{k+m})$ by computing the Kullback-Leibler distance between their two respective representative samples, $(\Theta_{k+m,1},\omega_{k,1})$, ..., $(\Theta_{k+m,n_k},\omega_{k,n_k})$ and $(\Theta_{k+m,1},\omega_{k+m,1}),...,(\Theta_{k+m,n_k},\omega_{k+m,n_k})$:

$$\kappa(\omega_{k+m},\omega_k)=\sum_{j=1}^{n_k}\omega_{k+m,j}(\log\omega_{k+m,j}-\log\omega_{k,j}). \quad (4)$$

We refer to the number m when MCMC is required as the *batch size*. It is determined adaptively by specifying a threshold for $\kappa(\omega_{k+m},\omega_k)$. Bigger batches mean that MCMC is performed less frequently. Thus, MCMC is performed sequentially for batches of measurements, hence the name sequential MCMC filter. Within a batch, SIS is performed.

Finally, any convenient MCMC procedure can be used in Step 3(b).

3. SIMULTANEOUS DETECTION AND ESTIMATION

We conducted an experiment in an acoustic anechoic chamber to record a weak ultrasonic acoustic chirp, and then processed it with our sequential MCMC filter. A linear chirp with chirp rate of 170.75 kHz/s was generated and transmitted through an electrostatic transducer. A microphone, placed some distance away (not exceeding 10 m) from the transmitter, received the acoustic chirp signal and recorded it at a sampling rate of 250 kHz. The noise in the recorded data comprised ambient noise and circuit noise. We suspected that the latter was dominant because acoustic noise in the anechoic chamber was very low. We analysed the measurement noise and found the Gaussian model to be adequate.

We knew that the received chirp was weak but it was not easy to estimate its SNR. Only after processing by the sequential MCMC filter did we realise that the SNR was about −14 dB. The raw recorded data required some pre-processing (scaling and bandpass filtering with pass band from 4 to 124 kHz) to remove certain hardware artefacts. After all these pre-processing, the only parameter of the real signal that is known exactly is the chirp rate.

To perform simultaneous detection and parameter estimation, we used the following models for the sequential MCMC filter:

$$H_1: x_t = w_t,\ w_t \sim N(0,\sigma_w^2). \quad (5)$$

$$H_2: x_t = \alpha\sin[2\pi(\beta t^2+\gamma t+\phi)]+w_t,\ w_t \sim N(0,\sigma_w^2). \quad (6)$$

Here, α is the amplitude, β is the chirp rate, and γ and ϕ are parameters that have the same dimension as frequency and phase shift respectively. The parameter vectors for the two models are

$$\psi^{(1)}=(\sigma_w^2), \quad (7)$$

$$\psi^{(2)}=(\alpha,\beta,\gamma,\phi,\sigma_w^2). \quad (8)$$

An alternative parameterization for model 2 is to use maximum frequency and minimum frequency, γ_{max} and γ_{min}, instead of β and γ, since it can be shown that

$$\gamma=\gamma_{min}, \quad (9)$$

$$\beta=\frac{\gamma_{max}-\gamma_{min}}{2T}, \quad (10)$$

where T, the duration of the data to be processed, is known. So we have

$$\psi^{(2)}=(\alpha,\gamma_{max},\gamma_{min},\phi,\sigma_w^2). \quad (11)$$

We started with equal model probabilities of 1/2 and with 1000 sample points, assigning 500 points to each model. Setting the maximum possible amplitude value to $\sqrt{2}$ V (corresponding to signal-to-noise ratio of 0 dB), the prior for amplitude, α, for H_2 was $U(0,\sqrt{2})$ (in V) to reflect the lack of prior information. Since the sampling rate was 250 kHz, the maximum instantaneous frequency permissible to avoid aliasing was 125 kHz. To reflect the lack of prior information, the prior distribution for the maximum frequency, γ_{max}, was chosen to be $U(0, 125)$ (in kHz). The prior for minimum frequency, γ_{min}, was modeled as $U(0,\gamma_{max})$, and the prior for ϕ was chosen to be $U(0, 2\pi)$. Lastly, the prior for noise power, σ_w^2, in both models was modeled using a lognormal distribution with a median of 1 and standard deviation of 0.2 to represent a noise power of around 1 W.

We processed the acoustic chirp data with the sequential MCMC filter and was able to detect the chirp after 1195 measurements. This is shown in Figure 1, which shows the posterior probability that a chirp is present reaching 1 at

measurement 1195. In comparison, we were not able to detect the chirp with the short-time Fourier transform (STFT) with the same 1195 measurements. However, techniques that are specially designed to detect linear chirps may fare better than the STFT. For example, the Radon ambiguity transform (RAT) [8] is able to detect the acoustic chirp, and provides an estimate of chirp rate only. In contrast, the sequential MCMC filter provides estimates of chirp rate (in terms of minimum frequency and maximum frequency), amplitude, initial phase and noise power. These are shown in Figure 2 as marginal medians and quartiles after processing 2048 measurements.

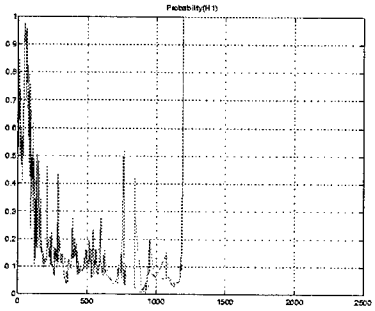


Figure 1. Probability of the chirp-plus-noise model in the acoustic chirp experiment.

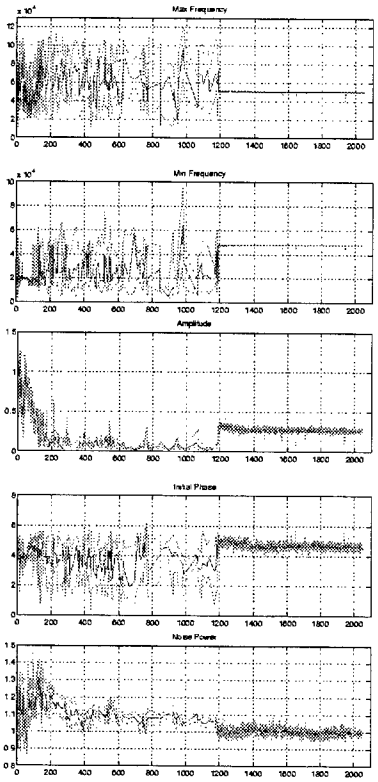


Figure 2. Marginal medians and quartiles of parameters in the chirp-plus-noise model from the sequential MCMC filter for 2048 measurements from the acoustic chirp experiment.

4. CONCLUSION

We have described an algorithm for Bayesian computations that combines SIS with MCMC. The algorithm performs MCMC sequentially on batches of measurements whose sizes are determined adaptively, hence the name sequential MCMC filter. Within a batch of measurement, SIS is used for computational efficiency. MCMC is needed from time to time to "refresh" the Monte Carlo sample and to remove any errors that may have accumulated from SIS due to the finite sample size. When MCMC is performed, it does not start from scratch but constructs its proposal distribution from the most recent Monte Carlo sample produced by SIS. Adaptive batch sizing is based on an easy-to-compute Kullback-Leibler distance. Bigger batches mean that MCMC is needed less often. For parameter-only problems that we have worked on, we have observed a trend of increasing batch size over time.

By incorporating multiple models, we have demonstrated the filter's ability to perform simultaneous model selection and parameter estimation. In the real-data experiment with the acoustic chirp, some degree of model mismatch is unavoidable, but the sequential MCMC filter performs reasonably well suggesting tolerance to model mismatch.

With today's computers, the sequential MCMC filter is computationally feasible for parameter-only problems. For problems with dynamic states, the growing dimension of the cumulative state is a severe obstacle to implementing the algorithm. We are exploring ways to overcome this, including a compression scheme that looks promising.

REFERENCES

[1] O. Cappe and C.P. Robert, "Markov chain Monte Carlo: 10 years and still running!" Journal of the American Statistical Association, vol. 95, no. 452, pp. 1282-1286, Dec. 2000.

[2] J.O. Berger, "Bayesian analysis: A look at today and thoughts of tomorrow," Journal of the American Statistical Association, vol. 95, no. 452, pp. 1269-1276, Dec. 2000.

[3] P. Billingsley, *Probability and Measure (second edition)*, John Wiley & Sons, 1986.

[4] J.S. Liu and R. Chen, "Sequential Monte Carlo methods for dynamic systems," Journal of the American Statistical Association, vol. 93, pp. 1032-1044, 1998.

[5] A. Kong, J.S. Liu and W.H. Wong, "Sequential imputations and Bayesian missing data problems," Journal of the American Statistical Association, vol. 89, pp. 278-288.

[6] C. Berzuini, N.G. Best, W.R. Gilks and C. Larizza, "Dynamic conditional independence models and Markov chain Monte Carlo methods," Journal of the American Statistical Association, vol. 92, pp. 1403-1412, 1997.

[7] B.P. Carlin, N.G. Polson and D.S. Stoffer, "A Monte Carlo approach to nonnormal and nonlinear state-space modeling," Journal of the American Statistical Association, vol. 87, pp. 493-500, 1992.

[8] M. Wang, A. Chan and C.K. Chui, "Linear frequency-modulated signal detection using Radon ambiguity transform," IEEE Transactions on Signal Processing, vol. 46, no. 3, pp. 571-586, Mar. 1998.

A FRAMEWORK FOR PARTICLE FILTERING IN POSITIONING, NAVIGATION AND TRACKING PROBLEMS

F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, P-J. Nordlund, R. Karlsson

Department of Electrical Engineering, Linköping University SE-581 83 Linköping, Sweden
Email: fredrik@isy.liu.se

ABSTRACT

A framework for positioning, navigation and tracking problems using particle filters (recursive Monte Carlo methods) is developed. Automotive and airborn applications, approached in this framework, have proven a numerical advantage over classical Kalman filter based algorithms. Here the use of non-linear measurement models and non-Gaussian measurement noise is the main explanation for the improvement in accuracy, and models for relevant sensors are surveyed.

1. INTRODUCTION

Recursive implementations of Monte Carlo based statistical signal processing [6] are known as *particle filters*, see [4, 3]. These may be a serious alternative for real-time applications classically approached by model-based Kalman filter techniques [9, 8]. The more non-linear model, or the more non-Gaussian noise, the more potential particle filters have, especially in applications where computational power is rather cheap and the sampling rate slow. The research has since the paper [7] steadily intensified.

The paper describes a general framework for a number of applications, where we have implemented the particle filter. The outline is as follows. We will start with a general framework of models covering all of our applications in Section 2, and in Section relevant sensors and their mesaurement models are surveyed. Section 4 describes how a number of applications we have studied fit into the framework, and the actual sensors we use. Conclusions, discussions and open questions of general interest are discussed in Section 5.

2. MODELS

Central for all navigation and tracking applications is the motion model to which various kind of model based filters can be applied. Models which are linear in the state dynamics and non-linear in the measurements and with additive noise are considered:

$$x_{t+1} = Ax_t + B_u u_t + B_w w_t, \tag{1a}$$
$$y_t = h(x_t) + e_t. \tag{1b}$$

The signals of primary interest in navigation and tracking applications are related to position, velocity and acceleration as summarized in Table 1. Depending on whether the signals are measureable or not, they may be components of either the state vector x_t or the input signal u_t.

Object	Position	Velocity	Acceleration
Own	$p^{(1)}$	$v^{(1)}$	$a^{(1)}$, $\delta a^{(1)}$ acc. bias
Other	$p^{(2)}$	$v^{(2)}$	-

Table 1. Interesting signals in navigation and tracking applications. The indexes (1) and (2) indicate signals related to one's own and another platform, respectively. All quantities can belong to either one, two or three-dimensional spaces, depending on the application.

Motion models (1a) are thoroughly discussed in literature, see *e.g.* [1, 8]. Note that the same kind of model can be used in all applications for both navigation and tracking. The main difference between the applications lies in the availability of measurements. Section 3 provides an extensive list of possible measurement equations (1b), that can be combined arbitrarily.

3. MEASUREMENT EQUATIONS

The main difference between the considered applications is the measurements available. Basically, the measurements are related to the positions of one's own platform $p^{(1)}$ and of the other object $p^{(2)}$. Therefore, the measurement equations can be categorized as depending on $p^{(1)}$ only, or depending on both $p^{(1)}$ and $p^{(2)}$:

$$y_t^{(1)} = h^{(1)}(p_t^{(1)}) + e_t^{(1)} \tag{2a}$$
$$y_t^{(2)} = h^{(2)}(p_t^{(1)}, p_t^{(2)}) + e_t^{(2)}, \tag{2b}$$

The current affiliations, of all but the first author, are, in order of appearance, Ericsson Radio, SaabTech Systems, NIRA Dynamics, Volvo CC, Saab Aircraft, Saab Bofors Dynamics, respectively.

0-7803-7011-2/01/$10.00 ©2001 IEEE

where the measurement noise contributions $e_t^{(1)}$ and $e_t^{(2)}$ are characterized by their distributions. If not explicitely mentioned, a Gaussian distribution is used.

In the studied applications, measurements from at least one of the categories above are available. It is important to note, that any combination of the sensors are possible. The presented applications are just a few examples.

3.0.1. *Measurements of Relative Distance*

As always, any position has to be related to a coordinate system and a reference position. Several types of sensors (*e.g.* GPS, RF) basically measure the distance relative to that reference point. One possibility is distance measurements of the own position relative to points of known positions p_i, $i = 1, \ldots, M$, which yields M measurement equations with

$$h_{a,i}^{(1)}(p_t^{(1)}) = \left| p_i - p_t^{(1)} \right|, \; i = 1, \ldots, M. \tag{2c}$$

This is also applicable when the position of another object is related to one's own position (*e.g.* radar, sonar, ultrasound):

$$h_b^{(2)}(p_t^{(1)}, p_t^{(2)}) = \left| p_t^{(2)} - p_t^{(1)} \right|. \tag{2d}$$

Some sensors do not measure the relative distance explicitly, but rather a quantity related to the same. One example is sensors that measure the received radio signal power transmitted from a known position p_i. This received power typically decays as $\sim K_1/r^\alpha$, $\alpha \in [2, 5]$, where K_1 and α are depending on the radio environment, antenna characteristics, terrain etc. In a logarithmic scale, the measurements are given by

$$h_{c,i}^{(1)}(p_t^{(1)}) = K - \alpha \log_{10} \left| p_i - p_t^{(1)} \right|, \; i = 1, \ldots, M. \tag{2e}$$

where $K = \log_{10} K_1$. Analogously, we can consider the situation when we focus on the power or intensity transmitted or reflected from an object and received at one's own position. The measurement is thus modeled by

$$h_d^{(2)}(p_t^{(1)}, p_t^{(2)}) = K - \alpha \log_{10} \left| p_t^{(1)} - p_t^{(2)} \right|. \tag{2f}$$

3.0.2. *Measurements of Relative Angle*

Similarly, the sensors can measure the relative angle between two positions (*e.g.* radar, IR, sonar, ultrasound). Given points of known positions p_i, $i = 1, \ldots, M$, the relative angle measurements can be described by

$$h_{e,i}^{(1)}(p_t^{(1)}) = \text{angle}\left\{ p_i, p_t^{(1)} \right\}, \; i = 1, \ldots, M. \tag{2g}$$

When relating the angle of an object to one's own position, we have

$$h_f^{(2)}(p_t^{(1)}, p_t^{(2)}) = \text{angle}\left\{ p_t^{(1)}, p_t^{(2)} \right\}. \tag{2h}$$

3.0.3. *Measurements of Relative Velocity*

Some sensors (*e.g.* radar) typically measure the Doppler shift of signal frequencies to estimate the magnitude of the relative velocity. This is essentially only applicable when relating the velocity of an object to one's own velocity. The measurements are categorized by

$$h_{g,i}^{(2)}(v_t^{(1)}, v_t^{(2)}) = \left| v_t^{(2)} - v_t^{(1)} \right|. \tag{2i}$$

3.0.4. *Map Related Measurements*

An aircraft can compute the ground altitude from radar measurements of height over ground and barometric measurements from which altitude is computed. The measured terrain height together with relative movement from the INS build up a height profile. Thus, $h_h(p^{(1)})$ denotes the height at point $p^{(1)}$ according to the Geographical Information System (GIS). Much effort has been spent on modeling the measurement error $e_t^{(1)}$ in a realistic way. It has turned out that a Gaussian mixture with two modes works well. One mode has zero mean, and the other a positive mean which corresponds to radar echos from the tree tops. The ground type in GIS can be used to switch the mean and variances in the Gaussian mixture. For instance, over sea there is only one mode with a small variance.

For map matching in the car positioning case, there is no real measurement. Instead, $h_j^{(1)}(p_t^{(1)})$ denotes the distance to the nearest road, and the measurement

$$y_t^{(1)} = h_j^{(1)}(p_t^{(1)}) + e_t^{(1)}$$

should therefore be equal to zero. A simple and relevant noise model is white and zero mean Gaussian noise.

4. APPLICATIONS

The problem areas are

- *Positioning,* where one's own position is to be estimated. This is a filtering problem rather than a static estimation problem, when an inertial navigation system is used to provide measurements of movement.
- *Navigation,* where besides the position also velocity, attitude and heading, acceleration and angular rates are included in the filtering problem.
- *Target tracking,* where another object's position is to be estimated based on measurements of relative range and angles to one's own position.

These problems are related in that they can be described by quite similar state space models. Traditional methods are based on linearized models and Gaussian noise approximations so that the Kalman filter can be applied. Research is

Application	State vector	Input	Measurement equations
Car positioning	$p_t^{(1)}$	$v_t^{(1)}$	Road map $h_j(p_t^{(1)})$, possibly GPS or base station distances $h_{a,i}^{(1)}(p_t^{(1)})$, base station powers $h_{c,i}^{(1)}(p_t^{(1)})$
Aircraft positioning	$p_t^{(1)}$	$a_t^{(1)}$	Altitude map $h_j(p_t^{(1)})$, GPS or other reference beacons $h_{a,i}^{(1)}(p_t^{(1)})$
Navigation in aircraft	$p^{(1)}, v^{(1)}, \delta a_t^{(1)}$	$a_t^{(1)}$	Altitude map $h_j(p_t^{(1)})$, GPS or other reference beacons $h_{a,i}^{(1)}(p_t^{(1)})$
Tracking	$p_t^{(2)}, v_t^{(2)}$		distance $h_b^{(2)}(p_t^{(1)}, p_t^{(2)})$, bearing $h_f^{(2)}(p_t^{(1)}, p_t^{(2)})$, doppler $h_g^{(2)}(p_t^{(1)}, p_t^{(2)})$, intensity $h_d^{(2)}(p_t^{(1)}, p_t^{(2)})$
Navigation and tracking in aircraft	$p_t^{(1)}, v_t^{(1)}, \delta a_t^{(1)}, p_t^{(2)}, v_t^{(2)}$	$a_t^{(1)}$	Altitude map $h_j(p_t^{(1)})$, GPS or other reference beacons $h_{a,i}^{(1)}(p_t^{(1)})$ distance $h_b^{(2)}(p_t^{(1)}, p_t^{(2)})$, bearing $h_f^{(2)}(p_t^{(1)}, p_t^{(2)})$, doppler $h_g^{(2)}(p_t^{(1)}, p_t^{(2)})$, intensity $h_d^{(2)}(p_t^{(1)}, p_t^{(2)})$
Navigation and tracking in cars	$p_t^{(1)}, v_t^{(1)}, \delta a_t^{(1)}, p_t^{(2)}, v_t^{(2)}$	$a_t^{(1)}$	Road map $h_j(p_t^{(1)})$, possibly GPS or base station distances $h_{a,i}^{(1)}(p_t^{(1)})$, base station powers $h_{c,i}^{(1)}(p_t^{(1)})$ distance $h_b^{(2)}(p_t^{(1)}, p_t^{(2)})$, bearing $h_f^{(2)}(p_t^{(1)}, p_t^{(2)})$, doppler $h_g^{(2)}(p_t^{(1)}, p_t^{(2)})$, intensity $h_d^{(2)}(p_t^{(1)}, p_t^{(2)})$

Table 2. List of considered applications with respective state vector (cf. Table 1), input signal and sensor information.

focused on how different state coordinates or multiple models can be used to limit the approximations. In contrast to this, the particle filter approximates the optimal solution numerically based on a physical model, rather than applying an optimal filter to an approximate model. The applications we have studied on real data are described below.

Car positioning by map matching. A digital road map is used to constrain the possible positions, where a dead-reckoning of wheel speeds is the main external input to the algorithm. By matching the driven path to a road map, a vague initial position (order of km's) can be improved to a meter accuracy. This principle can be used as a supplement to, or even replacement to, GPS (global positioning system).

Car positioning by Radio Frequency (RF) measurements. The digital road map above can be replaced by, or supplemented by, measurements from a terrestrial wireless communications system. For handover (to transfer a connection from one base station to another) operation, the mobile stations monitor the received signal powers from a multitude of base stations, and report regularly to the network. These measurements provide a power map which can be used in a similar manner as above. Mobile stations in a near future will moreover provide the possibility of monitoring the traveled distance of the radio signals from a number of base stations [5]. Such measurements can also be utilized in the same manner as with the power measurements.

Aircraft positioning by map matching or *terrain navigation.* A GIS contains, among other information, terrain elevation. The aircraft is equipped with sensors such that the terrain elevation can be measured. By map matching, the position can be deducted [2].

Integrated navigation. The aircraft's Inertial Navigation System (INS) uses dead-reckoning to compute navigation and flight data, *i.e.* position, velocity, attitude and heading. The INS is regarded as the main sensor for navitation and flight data due to being autonomous and having high reliability. However, small offsets cause drift and its output has to be stabilized. Here, terrain navigation is used today.

Target tracking. A classical problem in signal processing literature, where radar or IR measures relative angle, and for radar also relative range and range rate, to the object [1]. For the case of bearings only measuring IR sensor, either the state dynamics or measurement equation is very non-linear depending on the choice of state coordinates, so here the particle filter is particularly promising.

Combined navigation and tracking. Because the target tracking measurements are relative to one's own platform, positioning is an important sub-problem. Since the sensor introduces a cross-coupling between the problems, a unified treatment is tempting.

Car collision avoidance is very similar to the target tracking problem, here we are interested in predicting the own car's and other objects' future position. Based on the prediction, collision avoidance actions such as warning, braking and steering are undertaken when a collision is likely to happen. In order to have enough time to warn the driver the prediction horizon needs to be quite long. Therefore, utilizing knowledge about road geometry and infrastructure becomes important. One way to improve the prediction of possible maneouvres, is to use information in a digital map.

Thus, this is a specific project including all aspects of the problems listed above.

Typical state vectors, input signals and available (non-linear) sensor information are summarized in Table 2.

5. CONCLUSIONS AND DISCUSSION

We have given a general framework for positioning and navigation applications based on a flexible state space model and a particle filter. Five applications illustrate its use in practice. Evaluations in real-time, off-line on real data and in simulation environments show a clear improvement in performance compared to existing Kalman filter based solutions, where the new challenge is to find non-linear relations, state constraints and non-Gaussian sensor models that provide the most information about the position. Thus, *modeling* is the most essential step in this approach, compared to the various *implementations* of the Kalman filter found in this context (linearization issues, choice of state coordinates, filter banks, Gaussian sum filters, etc.).

General conclusions from the implementations are as follows: A choice of state coordinates making the state equation linear is beneficial for computation time and opens up the possibility for Rao-Blackwellization. This procedure enables a significant decrease in the particle state dimension. The evaluation of the likelihood one step ahead before resampling (APF[10], prior editing) is, together with adding extra state noise (jittering, roughening), crucial for avoiding divergence, and implies that the number of particles can be decreased further. Our implementations run in real-time (1–10 Hz), even in Matlab using several thousands of particles. Open questions for further research and development are listed below:

Divergence tests. It is essential to have a reliable way to detect divergence and to restart the filter (for the latter, see the transient below). For car positioning, the number of resamplings in the prior editing step turned out to be a very good indicator of divergence. Another idea, used in the terrain navigation implementation where the sampling rate is higher than necessary, is to split up the measurements to a filter bank, so that particle filter number i, $i = 1, 2, \ldots, n$ gets every n'th sample. The result of these n particle filters are approximately independent and voting can be used to restart each filter. This has turned out to be an efficient way to remove the outliers in data.

Transient improvement. The time it takes until the estimate accuracy comes down to the stationary value (the Cramer-Rao bound) depends on the number of particles. Given limited computational time, it may be advantageous to increase the number of particles N after a restart and discard samples in such a way that $N \cdot f_s$ is constant.

Since the particle filter has shown good improvement over linearization approaches, it is tempting to try even more accurate non-linear models. In particular, the flight dynamics of one's own vehicle is known and indeed used in model-based control, but is very rare in navigation applications. As a possible improvement, the particle filter may take full advantage of a more accurate model, where parts of the non-linear dynamics from driver/pilot inputs are incorporated.

Acknowledgment

The competence centre ISIS at Linköping University has brought all of the authors together and provides funding for Rickard and Per-Johan. We are very grateful to Christophe Andrieu and Arnoud Doucet for our fruitful discussions on the theoretical subjects. We want to acknowledge our gratitude to the master students Magnus Ahlström, Marcus Calais, who implemented the terrain navigation filter, and Peter Hall, who implemented the car positioning filter, and the supporting companies SAAB Bofors Dynamics and NIRA Dynamics, respectively.

6. REFERENCES

[1] Y. Bar-Shalom and X.R. Li. *Estimation and tracking: principles, techniques, and software.* Artech House, 1993.

[2] N. Bergman. *Recursive Bayesian Estimation: Navigation and Tracking Applications.* Dissertation nr. 579, Linköping University, Sweden, 1999.

[3] J.F.G. de Freitas, A. Doucet, and N.J. Gordon, editors. *Sequential Monte Carlo methods in practice.* Springer–Verlag, 2000.

[4] A. Doucet, S.J. Godsill, and C. Andrieu. On sequential simulation-based methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[5] C. Drane, M. Macnaughtan, and C. Scott. Positioning GSM telephones. *IEEE Communications Magazine*, 36(4), 1998.

[6] W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in practice.* Chapman & Hall, 1996.

[7] N.J. Gordon, D.J. Salmond, and A.F.M. Smith. A novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings on Radar and Signal Processing*, volume 140, pages 107–113, 1993.

[8] Fredrik Gustafsson. *Adaptive filtering and change detection.* John Wiley & Sons, Ltd, 2000.

[9] T. Kailath, A.H. Sayed, and B. Hassibi. *Linear estimation.* Information and System Sciences. Prentice-Hall, Upper Saddle Riber, New Jersey, 2000.

[10] M.K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, June 1999.

PARTICLE FILTERING FOR MULTIUSER DETECTION IN FADING CDMA CHANNELS

E. Punskaya.[1*] *C. Andrieu.*[2] *A. Doucet.*[3] *W.J. Fitzgerald* [1]

[1]Signal Processing Laboratory, Department of Engineering,
University of Cambridge, Trumpington Street, CB2 1PZ, Cambridge, UK.
[2]Department of Mathematics, Statistics Group,
University of Bristol, University Walk, Bristol, BS8 1TW, UK.
[3]Department of Electrical and Electronic Engineering,
The University of Melbourne, Victoria, 3010, Australia.

ABSTRACT

In this paper we address the problem of multiuser CDMA detection under fading conditions. The optimal detection problem can be reformulated as an optimal filtering problem for jump Markov linear systems; i.e. linear Gaussian state space models switching according to an unobserved finite state space Markov chain. Several approaches based on particle filtering techniques are reviewed to perform optimal filtering in this framework. A brief simulation study is carried out.

1. INTRODUCTION

Code division multiple access (CDMA) systems allow a significant increase of the capacity of cellular networks. It is likely they will be used not only in the 3G mobile but also in the following generations. This is why CDMA systems have recently been under intensive research. A significant thrust of this research has focused on the multiuser CDMA detection problem in multipath fading environments [2, 7, 11]. Multipath fading results in a significant increase of both the intersymbol interference (ISI) among the data symbols of the same user, and the multiple-access interference (MAI) among the data symbols of different users. These, added to a possibly non-Gaussian (impulsive) nature of the ambient noise in some physical channels such as urban and indoor radio channels, make the problem of symbol detection extremely difficult.

Under conditions of fading channels, the CDMA transmission model can be expressed in a state-space representation. Thus, in principle, general recursive expressions for the posterior distribution of the symbols may be derived, from which estimates of the symbols can be obtained. However, the problem has proved to be a difficult one. Indeed, the exact computation of these estimates involves a prohibitive computational cost exponential in the growing number of observations, and thus approximate methods must be employed.

In this paper, we concentrate on the problem of multiuser CDMA detection under conditions of Rayleigh flat (frequency-nonselective) fading channels. Our approach is based on particle filtering techniques, efficient simulation-based algorithms recently appeared in the literature (see [5] for a state-of-the-art in this field). The key idea of particle filters is to use an adaptive stochastic grid approximation of the conditional probability of the state vector with particles (values of the grid) evolving randomly in time according to a simulation-based rule. Depending on their ability to represent the different zones of interest of the state space which is dictated by the observation process and the dynamics of the underlying system, the particles can either give birth to offspring particles or die. The method uses several variance reduction techniques designed to make use of the structure of the model.

In [10], we applied particle filtering techniques to the problem of demodulation in fading channels. Here we develop a similar method in the more complex framework of CDMA systems. We also review and compare our approach with alternative deterministic and stochastic algorithms presented previously in the literature. Preliminary results indicate that the choice of the algorithm is very application dependent; a simple deterministic method can perform better than particle filtering techniques in some applications whereas it cannot even be realistically applied in other cases.

The remainder of the paper is organized as follows. The model specification and estimation objectives are stated in Section 2. It is shown that performing optimal estimation of symbols requires solving an optimal filtering problem. Section 3 introduces and reviews several deterministic and stochastic schemes to approximate the optimal filter. In Section 4 simulation results comparing various approaches are presented, and some conclusions are reached in Section 5.

2. PROBLEM STATEMENT AND ESTIMATION OBJECTIVES

Let us consider the downlink[1] of a synchronous CDMA system that is shared by L simultaneous users (see Fig. 1). Let us denote for any generic sequence κ_t, $\kappa_{i:j} \triangleq (\kappa_i, \kappa_{i+1}, \ldots, \kappa_j)^T$, and let $r_n^{(l)}$ be the nth information symbol from the lth user and $s_{\text{trans}}^{(l)}(\tau)$ be the corresponding equivalent lowpass signal waveform given by

$$s_{\text{trans}}^{(l)}(\tau) = \sqrt{E_l} s_n(r_n^{(l)}) u^{(l)}(\tau), \quad (n-1)T \leq \tau \leq nT,$$

where $s_n(.)$ performs the mapping from the digital sequence to waveforms and corresponds to the modulation technique employed,

*E. Punskaya is supported by the EPSRC grant, UK.

[1]Our method can equivalently be applied to the uplink multiuser problem with asynchronous CDMA.

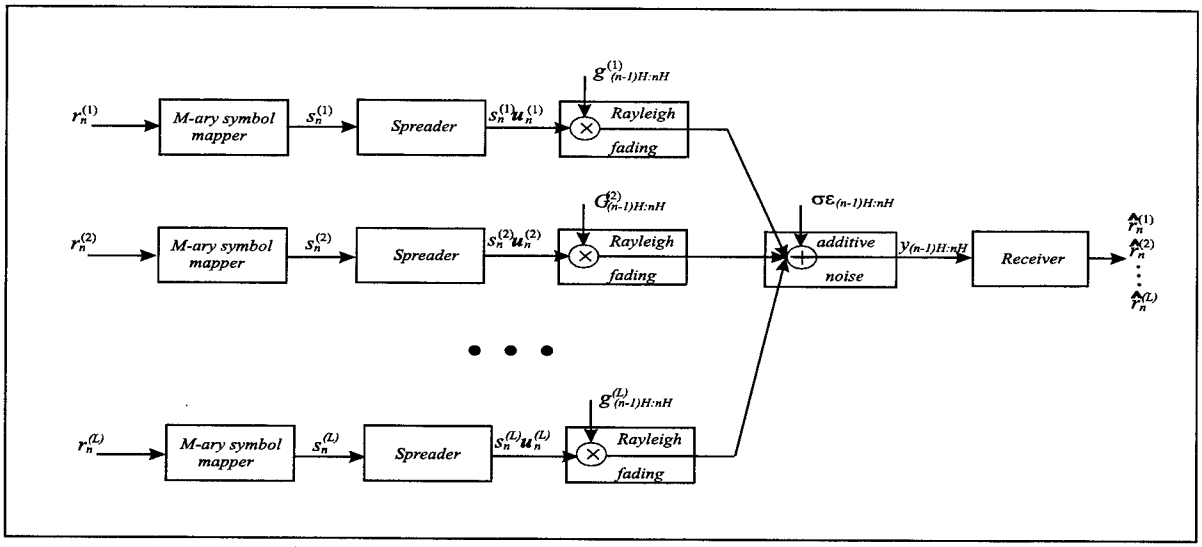


Fig. 1. Transmission of M-ary modulated signals in Rayleigh fading channels.

E_l is the signal energy per symbol (s_n is normalized to have unity power), and $u^{(l)}(\tau)$ is the signature waveform for the lth user, $u^{(l)}(\tau) = \sum_{h=1}^{H} a_h^{(l)} \eta(\tau - hT_c)$. Here, $a_{1:H}^{(l)}$ is a pseudo-noise (PN) code sequence consisting of H chips (with values $\{\pm 1\}$) per symbol, $\eta(\tau - hT_c)$ is a pulse of duration T_c, and T_c is the chip interval, $T_c = T/H$.

The waveform goes through a flat[2] Rayleigh fading channel and is corrupted by additive complex noise which is assumed to be Gaussian[3]. Thus, after matched filtering and sampling at the rate T_c^{-1}, the complex output of the channel at instant $t = (n-1)H + j_h$, $j_h = 1, \ldots, H$, corresponding to the transmission of the nth symbols can be expressed as

$$y|_{t=(n-1)H+j_h} = \mathbf{A}_{j_h} \mathbf{S}_n \left(r_n^{(l)}\right) \mathbf{g}_t + \sigma \varepsilon_t, \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}_c(0,1),$$

where $\mathbf{S}_n = diag(\left(\sqrt{E_1} s_n^{(1)}, \ldots, \sqrt{E_L} s_n^{(L)}\right)$, σ^2 being the noise variance, $\mathbf{A}_{j_h} = \left(a_{j_h}^{(1)}, \ldots, a_{j_h}^{(L)}\right)$, and $\mathbf{g}_t$ represents a multiplicative discrete time disturbance of the channels, $\mathbf{g}_t = \left(g_t^{(1)}, \ldots, g_t^{(L)}\right)^{\mathrm{T}}$, which is at instant t modelled as an ARMA(q,q) process (Butterworth filter of order q). The ARMA coefficients $\mathbf{a}$ (AR part) and $\mathbf{b}$ (MA part) are chosen so that the cut-off frequency of the filter matches the normalized channel Doppler frequency $f_d T_c$, which is known. Thus, the problem can be formulated in a linear Gaussian state space form (conditional upon the symbols), see [9] for details of representation.

The symbols $\mathbf{r}_n = \left(r_n^{(1)}, \ldots, r_n^{(L)}\right)^{\mathrm{T}}$, which are assumed i.i.d., and the channel characteristics $\mathbf{g}_t$ corresponding to the transmission of the nth symbol are unknown for $n > 0$. Our aim is to estimate $\mathbf{r}_n$ given the currently available data $\mathbf{y}_{1:n}$, $\mathbf{y}_n \triangleq y_{(n-1)H+1:nH}$. This can be done using the MAP (maximum *a posteriori*) criterion:

$$\hat{\mathbf{r}}_n = \arg\max_{\mathbf{r}_n} p(\mathbf{r}_n | \mathbf{y}_{1:n}).$$

[2]Frequency-selective channels can be considered in the same framework.

[3]The case of non-Gaussian noise can be easily treated using the techniques presented in [10].

However, this problem does not admit any analytical solution as computing $p(\mathbf{r}_n | \mathbf{y}_{1:n})$ involves a prohibitive computational cost exponential in the (growing) number of observations and, thus, approximate methods must be employed.

3. PARTICLE FILTERING

Given $\mathbf{y}_{1:n}$, all Bayesian inference on $\mathbf{r}_{1:n}$ relies on the posterior distribution $p(\mathbf{r}_{1:n} | \mathbf{y}_{1:n})$, which we propose to estimate using particle filtering techniques. The idea is to approximate $p(\mathbf{r}_{1:n} | \mathbf{y}_{1:n})$ by swarms of weighted points in the sample space $\left\{\mathbf{r}_{1:n}^{(i)}\right\}_{i=1}^{N}$, called particles. The particles evolve randomly in time in correlation with each other, and either give birth to offspring particles or die according to their ability to represent the different zones of interest of the state space.

A number of different algorithms of this type have been recently proposed in the literature (see [5] for the survey), some of them ([1, 3, 4, 6], for example) are specifically designed to make use of the structure of the model presented in Section 2. Here, we shall consider the essential features of these approaches, the details of the algorithms may be found in the appropriate references.

Sequential Importance Sampling and Resampling (SISR). The method is based on the following remark. Suppose that N particles $\left\{\mathbf{r}_{1:n}^{(i)}\right\}_{i=1}^{N}$ can be easily simulated according to an arbitrary convenient importance distribution $\pi(\mathbf{r}_{1:n} | \mathbf{y}_{1:n})$ (such that $p(\mathbf{r}_{1:n} | \mathbf{y}_{1:n}) > 0$ implies $\pi(\mathbf{r}_{1:n} | \mathbf{y}_{1:n}) > 0$). Then, using the importance sampling identity, an estimate of $p(\mathbf{r}_{1:n} | \mathbf{y}_{1:n})$ is given by the following point mass approximation

$$\hat{p}_N(\mathbf{r}_{1:n} | \mathbf{y}_{1:n}) = \sum_{i=1}^{N} \tilde{w}_{1:n}^{(i)} \delta_{(\mathbf{r}_{1:n}^{(i)})}(\mathbf{r}_{1:n}), \qquad (1)$$

where $\tilde{w}_{1:n}^{(i)}$ are the so-called *importance weights*

$$\tilde{w}_{1:n}^{(i)} = \frac{w_{1:n}^{(i)}}{\sum_{j=1}^{N} w_{1:n}^{(j)}}, w_{1:n}^{(i)} \propto \frac{p\left(\mathbf{r}_{1:n}^{(i)} \middle| \mathbf{y}_{1:n}\right)}{\pi\left(\mathbf{r}_{1:n}^{(i)} \middle| \mathbf{y}_{1:n}\right)}.$$

In order to propagate this estimate sequentially in time, $\pi(\mathbf{r}_{1:n}|\mathbf{y}_{1:n})$ has to admit $\pi(\mathbf{r}_{1:n-1}|\mathbf{y}_{1:n-1})$ as a marginal distribution. In addition, at each time step a selection step is included in the algorithm in order to discard particles with low normalized importance weights and multiply those with high ones. The choice of the importance distribution and a selection scheme is discussed in [4]; depending on those being employed, the computational complexity of the algorithm varies. As it is shown there, $\pi(\mathbf{r}_n|\mathbf{r}_{1:n-1},\mathbf{y}_{1:n}) = p(\mathbf{r}_n|\mathbf{r}_{1:n-1},\mathbf{y}_{1:n})$ is an importance distribution that minimizes the conditional variance of $w(\mathbf{r}_{1:n})$ and, therefore, is "*optimal*" in the framework considered (see [4] for details). However, for each particle it requires evaluation of the M^L H-step ahead Kalman filters for detection of the nth symbols since in this case

$$w_n \propto \sum_{m=1}^{M^L} p\left(\mathbf{y}_n|\mathbf{r}_{1:n-1}^{(i)}, \mathbf{r}_n = \boldsymbol{\rho}_m, \mathbf{y}_{1:n-1}\right),$$

where $\boldsymbol{\rho}_m$ corresponds to the mth ($m = 1, \ldots, M^L$) possible realization of $\mathbf{r}_n$ (see [9] for details). Thus, sampling from the optimal distribution is computationally extensive if M^L is large. In this case, the *prior* distribution can be used alternatively as the importance distribution, i.e. $\pi(\mathbf{r}_n|\mathbf{y}_{1:n},\mathbf{r}_{n-1}) = p(\mathbf{r}_n|\mathbf{r}_{n-1})$, so that, in total, for each particle at time t only one Kalman filter step is calculated. However, this method can be inefficient as it does not use the information carried by $\mathbf{y}_n$ to explore the state space. As far as the selection scheme is concerned, stratified sampling [8] employed in this paper can be implemented in $O(N)$ operations. The details of the algorithm are described in [3, 4, 9], a similar approach presented in [1].

Deterministic/Resample Low Weights approaches (RLW). An alternative approach to obtain the estimate of the posterior distribution $p(\mathbf{r}_{1:n}|\mathbf{y}_{1:n})$ is based on the following approximation:

$$\hat{p}_{N\times M^L}(\mathbf{r}_n|\mathbf{y}_{1:n}) = \sum_{i=1}^{N}\sum_{m=1}^{M^L} \tilde{w}_n^{(i,m)} \delta_{(\mathbf{r}_{1:n-1}^{(i)}),(\mathbf{r}_n=\boldsymbol{\rho}_m)}(\mathbf{r}_{1:n}),$$

where

$$w_n^{(i,m)} \propto p\left(\mathbf{y}_n|\mathbf{r}_{1:n-1}^{(i)}, \mathbf{r}_n = \boldsymbol{\rho}_m, \mathbf{y}_{1:n-1}\right). \quad (2)$$

Thus, we consider all possible "extensions" of the existing state sequences at each step n, and each particle has M^L offspring resulting in a set of $N \times M^L$ particles. These will each be assigned the weights, dependent on the weight of the parent at step $n-1$ and the likelihood term (2) that can be computed using the Kalman filter. In terms of calculations, this is equivalent to the use of the optimal distribution in SISR. However, when performing inference on the symbol $\mathbf{r}_n$, it is of course better to use $\hat{p}_{N\times M^L}(\mathbf{r}_n|\mathbf{y}_{1:n})$ than the standard SISR approximation; indeed one does not discard unnecessarily any information by selecting randomly one path out of the M^L available.

In order to avoid the exponentially increasing number of particles, a selection procedure has to be employed at each time step. The simplest way to perform such selection is just to choose the N most likely offspring and discard the others (as, for example, in [12]). A more complicated approach involves preserving the particles with high weights and resampling the ones with low weights, thus reducing their total number to N. In this particular case, a resampling scheme without replacement should be designed, i.e. each particle should appear at most once in the resulting set, as, indeed, there is no point in carrying along two particles evolving in exactly the same way. An algorithm of this type is presented in [6] but other selection schemes can be designed.

Whether we choose to preserve the most likely particles or employ the selection scheme proposed in [6], the computational load of the resulting algorithms at each time step t is that of $N \times M^L$ Kalman filters, and the selection step in both cases is implemented in $O(N \times M^L \log N \times M^L)$ operations. Of course, if M^L is large, which is the case in many applications (see Section 4, for example), both these methods are too computationally extensive to be used.

4. SIMULATION RESULTS

In order to demonstrate the bit-error-rate (BER) performance of our (SISR) algorithm it was, first, applied to the case of binary–phase-shift-keyed (BPSK) symbols transmitted over fast fading CDMA channels with $L = 3$, $H = 10$ and $f_d T_c = 0.05$. The results for different average signal to noise ratio (SNR) compared to those obtained in [2] are given in Fig. 2, where also the ideal channel state information (CSI) case is presented. As one can notice, even for just $N = 50$ particles, our algorithm outperforms substantially that of [2], especially when the signal-to-noise ratio (SNR) is large.

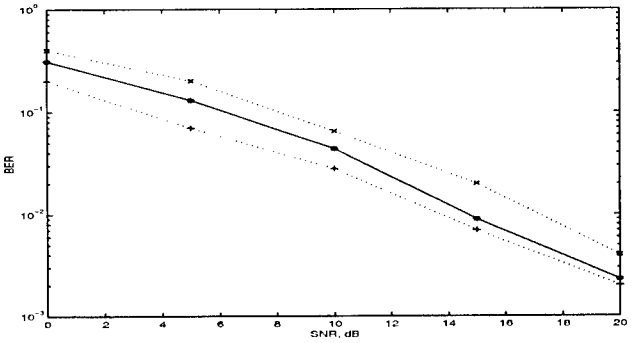


Fig. 2. Bit error rate (dotted line + (ideal CSI), solid line (SISR), dotted line x ([2])).

Then, computer simulations were carried out in order to compare the performance of the algorithms presented in Section 3. Some results for CDMA systems with the same parameters, SNR= 10 and $N = 50$ are presented in Table 1. They are interesting in the sense that, in this case, the deterministic approach preserving the N most likely particles (MLP) turned out to be the most efficient one! In order to achieve the same BER using a more complicated selection scheme (RLW) presented in [6], $N = 1000$ particles were required, and $N = 5000$ was needed with SISR. These results must indeed be interpreted cautiously. With other simulation parameters we found that the results between the different algorithms were much less pronounced. These issues need to be investigated further.

It should also be emphasized that if the number of users or processing gain in CDMA is large, a more complex modulation

	MLP	RLW	SISR
BER	2.51×10^{-2}	2.59×10^{-2}	2.70×10^{-2}

Table 1. Bit error rate for $N = 50$ and SNR=10 dB.

scheme is used and/or the additive noise is non-Gaussian (modelled as a mixture of Gaussians), both MLP and RLW are of no use due to their computational complexity, whereas our approach combined with Markov chain Monte Carlo (MCMC) methods ([4]) and employing the prior as an importance distribution leads to very good performance (see [9] for the details and further results).

5. DISCUSSION

In this paper, we consider the application of some particle filtering techniques to the problem of multiuser CDMA detection under fading conditions in the presence of possibly non-Gaussian additive noise. The results presented indicate quite small performance degradation compared to that of the receiver with ideal CSI. Moreover, additional simulations show that the algorithm exhibits good performance in the case of non-Gaussian additive noise, whereas other standard methods are not actually designed to treat this case (see [9]). Similar methods can also be applied to asynchronous CDMA systems and frequency-selective fading channels.

As was mentioned above, the problem addressed in this paper can be represented as a jump Markov linear system. We have reviewed several approaches to perform (approximate) optimal filtering in this framework. A simulation study has been carried out in order to compare these algorithms for the CDMA detection problem. Such a comparison has not been made before. In principle, all schemes are capable of providing optimal performance given a large number of particles. However, whenever it is applicable, we found out that a basic deterministic approach preserving the N most likely particles turned out to be the most efficient method! This deserves further study. This does not mean that particle filtering methods are of no use in communication systems. Indeed in most cases, the deterministic approach as well as the one proposed in [6] cannot be applied as they are too computationally extensive. In this case, particle filtering based on sampling importance resampling is relevant but requires the design of a "clever" importance distribution and/or the use of MCMC steps; see [4] for details.

6. ACKNOWLEDGMENT

The authors would like to acknowledge Simon Maskell for constructive and insightful comments which led us to a more critical evaluation of our methodology.

7. REFERENCES

[1] R. Chen, and J. Liu, "Mixture Kalman Filters," *J. Roy. Statist. Soc.* B, vol. 62, pp. 493-508, 2000.

[2] L.M. Davis and I.B. Collings, "Multi-user MAP decoding for flat-fading CDMA channels," in *Proc. Conf. DSPCS-99*, pp. 79-86, 1999.

[3] A. Doucet, S. Godsill and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197-208, 2000.

[4] A. Doucet, N.J. Gordon and V. Krishnamurthy, "Particle Filters for State Estimation of Jump Markov Linear Systems," *IEEE Trans. on Signal Processing*, vol. 49, no.3, pp. 613-624, 2001.

[5] A. Doucet, J.F.G. de Freitas and N.J. Gordon (eds.), *Sequential Monte Carlo Methods in Practice*, Springer-Verlag: New-York, 2001.

[6] P. Fearnhead, Sequential Monte Carlo methods in filter theory, PhD thesis, University of Oxford, 1998.

[7] W. Hou and B. Chen, "Adaptive detection in asynchronous code-division multiple access systems in multipath fading channels," *IEEE Trans. on Communications*, vol. 48, no. 5, pp. 863-873, 2000.

[8] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *J. Comp. Graph. Stat.*, vol. 5, no. 1, pp. 1-25, 1996.

[9] E. Punskaya, C. Andrieu, A. Doucet and W.J. Fitzgerald, "Multiuser CDMA detection under fading conditions using particle filtering," technical report, Cambridge University Engineering Department, CUED/F-INFENG/TR. 413, 2001.

[10] E. Punskaya, C. Andrieu, A. Doucet and W.J. Fitzgerald, "Particle Filtering for Demodulation in Fading Channels with Non-Gaussian Additive Noise," *IEEE Trans. on Communications*, vol. 49, no. 4, pp. 579-582, 2001.

[11] D. Raphaeli, "Suboptimal Maximum-Likelihood multiuser detection of synchronous CDMA on frequency-selective multipath channels," *IEEE Trans. on Communications*, vol. 48, no. 5, pp. 875-885, 2000.

[12] J.K. Tugnait and A.H. Haddad, "A detection-estimation scheme for state estimation in switching environment," *Automatica*, vol. 15, pp. 477-481, 1979.

THE REJECTION GIBBS COUPLER: A PERFECT SAMPLING ALGORITHM AND ITS APPLICATION TO TRUNCATED MULTIVARIATE GAUSSIAN DISTRIBUTIONS

Yufei Huang, Tadesse Ghirmai and Petar M. Djurić

Department of Electrical and Computer Engineering
State University of New York at Stony Brook
Stony Brook, NY 11794-2350
Tel: +1(631)632 8424 Fax: +1(631)632 8494
yfhuang, tadesse, djuric@ece.sunysb.edu

ABSTRACT

Recently, a new Markov chain based algorithm for drawing samples from a desired distribution has been proposed. This algorithm, also known as perfect sampling algorithm, can determine exactly when a Markov chain enters the equilibrium, and hence can output exact samples. In this paper, we introduce a perfect sampling algorithm called the rejection Gibbs coupler for perfect sampling from bounded multivariate distributions. We demonstrate an application of the rejection coupler for generation of samples from truncated multivariate Gaussian distributions.

1. INTRODUCTION

In the past decade, research in Markov chain Monte Carlo (MCMC) sampling has drawn much attention in the statistical and signal processing communities. In particular, the use of MCMC sampling has revived the interest in using the Bayesian methodology for solving various practical problems.

Diagnosis of convergence of Markov chains, however, remains a challenging problem. As a result, samples obtained by MCMC methods can only be considered approximately rather than exactly distributed according to a desired distribution. In 1996, Propp and Wilson [1] proposed a solution to the aforementioned problem of MCMC such that the convergence time of a Markov chain can be exactly determined. Thus the samples produced thereafter are exact samples from the desired distribution. This algorithm is named *coupling from the past* (CFTP). Since then, research on further development of CFTP algorithm has quickly picked up. The original CFTP was designed on discrete variable spaces. A successful extension of CFTP to continuous variable spaces was introduced by Murdoch and Green [2] where several algorithms such as the *multigamma coupler*, the *rejection coupler*, and the *Metropolis coupler* were proposed. In addition, the possibility of constructing a Gibbs-sampler-like perfect sampling algorithm was also demonstrated.

In [3], we have proposed a novel perfect sampling algorithms called the Gibbs coupler. The proposed algorithm on high dimensional binary spaces overcomes the obstacle of the original CFTP in that it can be efficiently implemented regardless of the existence of (anti-)monotonic Markov chains. Applications of the Gibbs coupler were shown for problems on variable selection [3] and multiuser detection of CDMA systems [4].

In this paper, we introduce a new version of the Gibbs coupler termed the rejection Gibbs coupler. The rejection Gibbs coupler combines the idea of the rejection coupler with the framework of the general Gibbs coupler and aims at sampling from bounded multivariate distributions. In this paper, first we outline the rejection Gibbs coupler, and then we discuss the partitioning technique which is important for practical implementation of the algorithm. Finally, we show how the Gibbs coupler can be applied to draw samples from truncated multivariate Gaussian distributions. Simulation results are also provided to show the performance of the rejection Gibbs coupler.

2. COUPLING FROM THE PAST

CFTP, similarly to the MCMC sampling methods, generates samples from a desired distribution by using Markov chains. However, CFTP constructs not a single but multiple Markov chains and utilizes the concept of *coupling*. In a coupling process, at any transition, the same update function and random number are assigned to all the Markov chains. In CFTP, the coupling process is implemented from the past to time 0. The CFTP algorithm can be described as an iterative scheme by the following pseudocode:

```
CFTP(T)
    t ← −T, B_t ← S
    while t < 0
        t ← t + 1
        B_t ← φ(B_{t−1}, U_t)
    if |B_t| = 1 then
        return(B_0)
    else
        CFTP(2T)
```

In the above pseudocode, $\mathcal{S}$ denotes a desired discrete state space with size $M = |\mathcal{S}|$, and $R^{(t)}$ and $\Phi(\cdot, R^{(t)})$ are the random seed and update functions, respectively. At the start of each iteration, CFTP initiates M Markov chains at every possible state of the state space $\mathcal{S}$ from some time $-T$ in the past, couples them together,

This work was supported by the National Science Foundation under Award CCR-9903120.

and runs towards time 0. Then at time 0, the coalescence of the chains is checked. It is noted that all the Markov chains should have the desired distribution as their stationary distribution. Now, if all the chains have coalesced to the same state at time 0, the coalesced state is then a perfect sample from the desired distribution. This is because if we started the algorithm from the infinite past but kept the existing random seeds of the transition from $-T$ to 0, the Markov chains would have coalesced into the same state at time 0. Apparently, since the chains would have been propagated from the infinite past, the coalesced state at time 0 is a steady state which follows the desired distribution exactly.

Notice that CFTP is proposed primarily for problems with finite discrete variable spaces. A direct extension of CFTP to continuous variable spaces is prohibitive since the size of a continuous variable space is infinite, and thus it will take CFTP infinite time to reach coalescence. To allow for perfect sampling from continuous variable spaces, special care must be taken to map infinite-size continuous variable spaces into finite discrete variable spaces [2]. In the next section, we introduce an algorithm that achieves perfect sampling from bounded multivariate distributions.

3. THE REJECTION GIBBS COUPLER

Suppose that we want to draw samples from an N dimensional multivariate distribution $p(\mathbf{x})$ defined on a variable space $\boldsymbol{\mathcal{S}}$. To apply the rejection Gibbs coupler to the problem, the full conditional distributions $p(x_i|\mathbf{x}_{-i})\ \forall i$ are required to be specified, where $\mathbf{x}_{-i}$ represents the vector of the $N-1$ variables in $\mathbf{x}$ except for the i-th variable. Moreover, we assume that an upper bound and a lower bound functions can be determined at every instant of time t such that

$$h_i^{(t)}(x_i) = \max_{\mathbf{x}_{-i} \in \boldsymbol{\mathcal{S}}_{-i}^t} g(x_i|\mathbf{x}_{-i}) \tag{1}$$

and

$$r_i^{(t)}(x_i) = \min_{\mathbf{x}_{-i} \in \boldsymbol{\mathcal{S}}_{-i}^t} g(x_i|\mathbf{x}_{-i}) \tag{2}$$

where $\boldsymbol{\mathcal{S}}_{-i}^{(t)} \subset \boldsymbol{\mathcal{S}}_{-i}$, and $g(x_i|\mathbf{x}_{-i})$ is a function proportional to $p(x_i|\mathbf{x}_{-i})$. Notice that detailed expression of $g(x_i|\mathbf{x}_{-i})$ can vary by including or removing terms with respect to $\mathbf{x}_{-i}$ (since these terms are considered as the proportional constant). A different expression of $g(x_i|\mathbf{x}_{-i})$ will eventually affect the complexity of the algorithm. Generally, there are two guiding principles for choosing the function $g(x_i|\mathbf{x}_{-i})$. First, $g(x_i|\mathbf{x}_{-i})$ should be in a form easy for the determination of $h_i(\cdot)$ and $r_i(\cdot)$. Second, the corresponding distribution $h_i(x)/\nu$ should be easy to sample from, where $\nu = \int h(x)dx$ is the normalizing constant. Now, once the bounded functions are determined, the algorithm of the rejection Gibbs coupler can be proceeded according to the outline displayed in Chart I.

In the algorithm, $\hat{h}_i^{(t)}$ and $\hat{r}_i^{(t)}$ are also determined according to (1) and (2). Notice that the general framework of the algorithm still follows that of CFTP. However the detailed coupling scheme is based on the rejection coupler. Typically, the ratio $\rho_i^{(t)} = r_i^{(t)}/h_i^{(t)}$ is a key factor in defining the speed of coalescence of the algorithm. This is because on average, the algorithm would generate $1/\rho_i^{(t)}$ samples at time t for the i-th component. Therefore, the larger the $\rho_i^{(t)}$, the less the number of samples the algorithm produces and hence the faster the coalescence.

Chart I.

Rejection Gibbs coupler(T):
- $t \leftarrow -T$, $\boldsymbol{\mathcal{S}}^{(t)} \leftarrow \boldsymbol{\mathcal{S}}$
- while $t < -T/2$
  - $t \leftarrow t+1$
  - for $i = 1, 2, \cdots, N$
    - determine $h_i^{(t)}(x_i)$ and $r_i^{(t)}(x_i)$ w.r.t. $\boldsymbol{\mathcal{S}}_{-i}^{(t)}$
    - $j \leftarrow 0$
    - repeat
      - $j \leftarrow j+1$
      - draw $U_{ij}^{(t)}$ from $U(0,1)$
      - draw X_{ij} from $h_i^{(t)}(\cdot)/\nu_i^{(t)}$
      - if $U_{ij}^{(t)} < r_i^{(t)}(X_{ij})/h_i^{(t)}(X_{ij})$ then
        - $J \leftarrow j$
        - exit repeat
    - $\mathcal{S}_i^{(t)} \leftarrow \{X_{i1}, X_{i2}, \cdots, X_{iJ}\}$
- while $t < 0$
  - $t \leftarrow t+1$
  - for $i = 1, 2, \cdots, N$
    - determine $\hat{h}_i^{(t)}(\cdot)$ and $\hat{r}_i^{(t)}(\cdot)$ w.r.t. $\boldsymbol{\mathcal{S}}_{-i}^{(t)}$
    - $j \leftarrow 0$
    - repeat
      - $j \leftarrow j+1$
      - for $X_j \in \mathcal{S}_i^{(t)}$
    - $k \leftarrow 0$
    - if $U_{ij}^{(t)} < \hat{h}_i^{(t)}(X_j)/h_i^{(t)}(X_j)$ then
      - $k \leftarrow k+1$
      - $\hat{X}_k = X_j$
      - if $U_{ij}^{(t)} < \hat{r}_i^{(t)}(X_j)/h_i^{(t)}(X_j)$ then
        - exit repeat
  - $K \leftarrow k$, $\mathcal{S}_i^{(t)} \leftarrow \{\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_K\}$
- if size of $\mathcal{S}_i^{(t)}$ is equal to 1, for $i = 1, 2, \cdots, N$ then
  - return($\boldsymbol{\mathcal{S}}^{(0)}$)
- else
  - Rejection Gibbs coupler($2T$)

4. THE PARTITIONING TECHNIQUE

In practice the ratio $\rho_i^{(t)}$ is often very small which, as a result, leads to vary slow convergence of the algorithm. To circumvent this difficulty, one can divide the variable space $\boldsymbol{\mathcal{S}}_{(-i)}^{(t)}$ into a collection of disjoint cells, or partitions [2], and specify the upper bound and lower bound functions for each partition. The partition is done in a way that the ratio for the l-th partition $\rho_{il}^{(t)}$ is large enough to guarantee a reasonable mean number of samples. For instance, one can impose M partitions on $\boldsymbol{\mathcal{S}}_{(-i)}^{(t)}$ and for each partition set $\rho_{il}^{(t)} = (\rho_i^{(t)})^{\frac{1}{M}}$. Then the average number of the produced samples at time t will be $M/(\rho_i^{(t)})^{\frac{1}{M}}$. As a simple illustration, if $\rho = 10^{-5}$ and $M = 5$, the value of $\rho_{il}^{(t)} = 0.1$, or, after partition, 50 samples are produced on average. If we compare this average with the

average of $1/\rho_i^{(t)} = 10^5$ samples before partitioning, the number of produced samples is reduced by 2000 times. Equivalently, we can say that the partitioned algorithm is 2000 times faster.

However, if M is determined at the first step of the algorithm or at $t = -T$, and then fixed afterwards, it is possible that at a certain time instant t, the mean number of samples $1/\rho_i^{(t)}$ is already less than M. Since the partitioning algorithm with fixed M would produce at least M samples, it introduces more samples at time t than the nonpartitioning algorithm. Consequently, in this case, the partitioning algorithm would be slower to coalesce. There are two remedies to this problem. With the first one, one can fix the detailed range for each of the M partitions at the first step of the algorithm, and the range of each partition will remain fixed later on in the algorithm. Then at any time t, the average number of samples produced for each partition would not change. With the second remedy, one can fix the value of $\rho_{il}^{(t)}$ for all t. In that case, from time to time the number of partitions M will change on different $\boldsymbol{\mathcal{S}}_{-i}^{(t)}$. Under this scheme, as long as $1/\rho_{il}^{(t)} > 1$, the proposed problematic scenario would no longer occur. This option leads to an adaptively partitioning algorithm.

5. PERFECT SAMPLING FROM TRUNCATED MULTIVARIATE GAUSSIAN DISTRIBUTIONS BY THE REJECTION GIBBS COUPLER

In this section, we demonstrate the use of the rejection Gibbs coupler for drawing perfect samples from truncated multivariate Gaussian distributions. First, let $\mathbf{x} = [x_1\ x_2 \cdots x_N]^\mathsf{T} \in \boldsymbol{\mathcal{S}}^N$ represent a vector of N random variables which is distributed according to the truncated multivariate Gaussian $TN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\mathcal{S}}^N)$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the $N \times 1$ mean vector and the $N \times N$ covariance matrix of the corresponding non-truncated Gaussian distribution, and $\boldsymbol{\mathcal{S}}^N = \cup_{i=1}^N [a_i, b_i]$. Next, we rearrange $\mathbf{x}$ by $\mathbf{x} = [x_i\ \mathbf{x}_{-i}^\mathsf{T}]^\mathsf{T}$, and partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly by

$$\boldsymbol{\mu} = [\mu_i, \boldsymbol{\mu}_{-i}]^\mathsf{T}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_i^2 & \boldsymbol{\Sigma}_{i1}^\mathsf{T} \\ \boldsymbol{\Sigma}_{i1} & \boldsymbol{\Sigma}_{-i} \end{bmatrix}.$$

Then, it can be shown that the full conditional distributions $p(x_i|\mathbf{x}_{-i})$ are also truncated Gaussians that can be expressed as

$$\begin{aligned} p(x_i|\mathbf{x}_{-i}) &= TN(\tilde{\mu}_i, \tilde{\sigma}_i^2, (a_i, b_i]) \\ &\propto \exp\{\frac{1}{2\tilde{\sigma}_i^2}(2\boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{x}_{-i} - \boldsymbol{\mu}_{-i})x_i + \mu_i x_i - x_i^2)\} \\ &= g(x_i|\mathbf{x}_{-i}) \end{aligned} \quad (3)$$

where $\tilde{\mu}_i = \mu_i + \boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{x}_{-i} - \boldsymbol{\mu}_{-i})$ and $\tilde{\sigma}_i^2 = \sigma_i^2 - \boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}\boldsymbol{\Sigma}_{i1}$. To apply the rejection Gibbs coupler, we need to determine the bounds on $g(x_i|\mathbf{x}_{-i})$ and specify the distribution that corresponds to the upper bound. First, define two $N \times 1$ vectors $\mathbf{x}^{(t)\max}$ and $\mathbf{x}^{(t)\min}$, whose j-th components are

$$x_j^{(t)\max} = \begin{cases} \arg\max_{x \in \mathcal{S}_j^{(t)}} x & \text{if } \beta_j \geq 0 \\ \arg\min_{x \in \mathcal{S}_j^{(t)}} x & \text{if } \beta_j < 0 \end{cases}$$

and

$$x_j^{(t)\min} = \begin{cases} \arg\min_{x \in \mathcal{S}_j^{(t)}} x & \text{if } \beta_j \geq 0 \\ \arg\max_{x \in \mathcal{S}_j^{(t)}} x & \text{if } \beta_j < 0 \end{cases}$$

where β_j represents the j-th element of $\boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}$, and $\mathcal{S}_j^{(t)}$ denotes the support of x_j at time t. Next, let $\mathbf{x}_{-i}^{(t)\max}$ and $\mathbf{x}_{-i}^{(t)\min}$ represent two $(N-1) \times 1$ vectors which consist of all except the i-th components in $\mathbf{x}^{(t)\max}$ and $\mathbf{x}^{(t)\min}$, respectively. Then the upper and lower bounds on $g(x_i|\mathbf{x}_i)$ are found to be

$$h_i^{(t)}(x_i) = \begin{cases} e^{\frac{1}{2\tilde{\sigma}_i^2}(2\boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{x}_{-i}^{(t)\max} - \boldsymbol{\mu}_{-i})x_i + \mu_i x_i - x_i^2)} & \text{if } x_i \geq 0 \\ e^{\frac{1}{2\tilde{\sigma}_i^2}(2\boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{x}_{-i}^{(t)\min} - \boldsymbol{\mu}_{-i})x_i + \mu_i x_i - x_i^2)} & \text{if } x_i < 0 \end{cases} \quad (4)$$

and

$$r_i^{(t)}(x_i) = \begin{cases} e^{\frac{1}{2\tilde{\sigma}_i^2}(2\boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{x}_{-i}^{(t)\min} - \boldsymbol{\mu}_{-i})x_i + \mu_i x_i - x_i^2)} & \text{if } x_i \geq 0 \\ e^{\frac{1}{2\tilde{\sigma}_i^2}(2\boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{x}_{-i}^{(t)\max} - \boldsymbol{\mu}_{-i})x_i + \mu_i x_i - x_i^2)} & \text{if } x_i < 0\,. \end{cases} \quad (5)$$

Furthermore, the distribution corresponding to the upper bound function can be shown as a mixture of two truncated Gaussian distributions and has the form

$$f_{h_i}^{(t)}(x_i) = w_{i1} TN(\tilde{\mu}_{i1}, \tilde{\sigma}_i^2, 0, b_i) + w_{i2} TN(\tilde{\mu}_{i2}, \tilde{\sigma}_i^2, a_i, 0) \quad (6)$$

where w_{i1} and w_{i2} are the weights assigned to the two mixands, $\tilde{\mu}_{i1} = \boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{x}_{-i}^{(t)\max} - \boldsymbol{\mu}_{-i}) + \mu_i$, and $\tilde{\mu}_{i2} = \boldsymbol{\Sigma}_{i1}^\mathsf{T}\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{x}_{-i}^{(t)\min} - \boldsymbol{\mu}_{-i}) + \mu_i$. The weights w_{i1} and w_{i2} are uniquely defined as

$$w_{i1} = c_1/(c_1 + c_2) \quad (7)$$

and

$$w_{i2} = c_2/(c_1 + c_2) \quad (8)$$

where

$$c_1 = \exp\{\tilde{\mu}_{i1}^2/(2\tilde{\sigma}_i^2)\}\,(Q((0 - \tilde{\mu}_{i1})/\tilde{\sigma}_i) - Q((b_i - \tilde{\mu}_{i1})/\tilde{\sigma}_i))$$

and

$$c_2 = \exp\{\tilde{\mu}_{i2}^2/(2\tilde{\sigma}_i^2)\}\,(Q((a_i - \tilde{\mu}_{i2})/\tilde{\sigma}_i) - Q((0 - \tilde{\mu}_{i2})/\tilde{\sigma}_i))\,.$$

Here, $Q(y)$ is the Q-function which represents the probability that a $\mathcal{N}(0,1)$ random variable exceeds y. Once we determine the weights of the two mixands, sampling from $f_h(x_i)$ is easy to accomplish and can proceed as follows:

draw u *from* $U(0,1)$;
if $u < w_1$
  draw x_i *from* $TN(\tilde{\mu}_{i1}, \tilde{\sigma}_i^2, 0, b_i)$;
otherwise
  draw x_i *from* $TN(\tilde{\mu}_{i2}, \tilde{\sigma}_i^2, a_i, 0)$;

Note that a sample $\tilde{x}$ from the univariate truncated Gaussian distribution $TN(\mu, \sigma^2, a, b)$ is obtained by the inverse transformation which computes

$$\tilde{x} = \sigma Q^{-1}(\tilde{u}(Q((a-\mu)/\sigma_i) - Q((b-\mu)/\sigma_i)) + Q((a-\mu)/\sigma_i)) + \mu$$

where $\tilde{u}$ is a sample from $U(0,1)$.

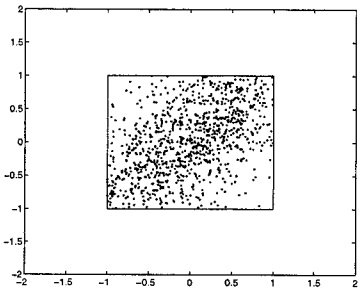


Figure 1: Scattergram of 1000 perfect samples from $TN(0, [1\ 0.8;\ 0.8\ 1], \cup_{i=1}^{2}(-1, 1])$ by the Gibbs coupler.

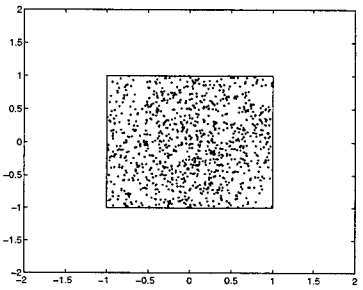


Figure 2: Scattergram of 1000 perfect samples from $TN(0, [1\ 0.2;\ 0.2\ 1], \cup_{i=1}^{2}(-1, 1])$ by the Gibbs coupler.

6. SIMULATION RESULTS

To demonstrate the performance of the proposed rejection Gibbs coupler, several experiments were performed. In the first experiment, 1000 perfect samples from the bivariate truncated Gaussian $TN(0, \Sigma, \cup_{i=1}^{2}(-1, 1])$ were collected, where $\Sigma = [1\ 0.8;\ 0.8\ 1]$. The scattergram of the samples is displayed in Figure 1.

In the second experiment, the covariance matrix was $\Sigma = [1\ 0.2;\ 0.2\ 1]$. Again, 1000 perfect samples were collected. The samples scattergram is shown in Figure 2.

In the third experiment, we examined the correlation between samples obtained through the rejection Gibbs coupler. By using the samples obtained in the first experiment, we calculated the estimate of the autocorrelation coefficients for first variable and the crosscorrelation coefficient between two variables. In addition, we also applied the Gibbs sampler [6] and generated samples from the bivariate truncated Gaussian with the same setting as that in the first experiment. Similarly we calculated the estimate of their autocorrelation and crosscorrelation coefficients. The results are demonstrated in Figure 3 and 4. The figures clearly show that the Gibbs sampler results in much larger correlations for adjacent samples than the Gibbs coupler. This indicates that any inferences carried out by the perfect samples generated through the rejection Gibbs coupler will have a smaller variance than that through the Gibbs sampler.

7. CONCLUSION

We proposed an algorithm called the rejection Gibbs coupler for perfect sampling from bounded multivariate distributions. As an application, we showed the implementation of the proposed algorithm on truncated multivariate Gaussian distributions. The advantage of the rejection Gibbs coupler over the Gibbs sampler is shown by the simulation results.

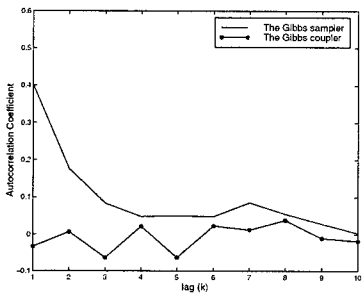


Figure 3: Plot of the sample autocorrelation coefficients of the first variable.

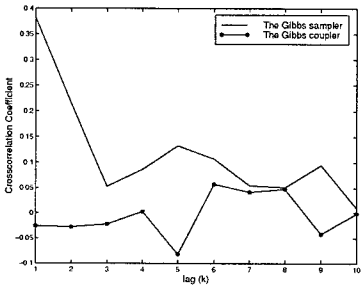


Figure 4: Plot of the sample crosscorrelation coefficients between two variable.

8. REFERENCES

[1] J. G. Propp and D. B. Wilson, "Exact sampling with coupled Markov chains and applications to statistical mechanics," *Random Structures Algorithms*, vol. 9, pp. 223–252, 1996.

[2] D. J. Murdoch and P. J. Green, "Exact sampling form a continuous state space," in *Scandinavian Journal of Statistics*, 1998, vol. 25, pp. 483–502.

[3] Y. Huang and P. M. Djurić, "The Gibbs coupler: A novel perfect sampling algorithm," submitted to *IEEE Transactions on Information Theory*.

[4] Y. Huang and P. M. Djurić, "Multiuser detection of synchronous Code-Division Multiple-Access signals by the Gibbs coupler," in the *IEEE* Proceedings of ICASSP, Salt Lake City, Utah, 2001.

[5] Y. Huang and P. M. Djurić, "Variable selection by perfect sampling," in the proceedings of the *IEEE - EURASIP Workshop on Nonlinear Signal and Image Processing*, Baltimore, Maryland, 2001.

[6] J. H. Kotecha and P. M. Djurić, "Gibbs sampling approach for generation of truncated multivariate Gaussian random variables," in the *IEEE* Proceedings of ICASSP, Phoenix, Arizona, 1999.

ASSESSMENT OF MCMC CONVERGENCE: A TIME SERIES AND DYNAMICAL SYSTEMS APPROACH

Rodney C. Wolff

School of Mathematical Sciences
Queensland University of Technology
GPO Box 2434
Brisbane QLD 4001, AUSTRALIA
email: r.wolff@qut.edu.au

Darfiana Nur, Kerrie L. Mengersen

Department of Mathematics
University of Newcastle
University Drive
Callaghan NSW 2308, AUSTRALIA

ABSTRACT

Important in the application of Markov chain Monte Carlo (MCMC) methods is the determination that a search run has converged. Given that such searches typically take place in high-dimensional spaces, there are many pitfalls and difficulties in making such assessments. In the present paper, we discuss the use of phase randomisation as tool in the MCMC context, provide some details of its distributional properties for time series which enable its use as a convergence diagnostic, and contrast its performance with a selection of other widely used diagnostics. Some brief comments on analytical results, obtained via Edgeworth expansion, are also made.

1. INTRODUCTION

MCMC methods support the application of Bayesian statistical methods through permitting complex distributions to be evaluated (specifically, by handling theoretically intractable integrals of high-dimensional probability density functions). Given the numerical and geometrical complexity of MCMC methods, assessment of convergence is a non-trivial task. Diagnostics for convergence are required in practical settings, and thus need to be accessible, accurate and fast.

In the theory of time series resampling, the method of phase randomisation has been used to generate so-called *surrogate* time series with the same first- and second-order properties as the original: see Theiler *et al.* (1992) and Timmer (1998), as well as Davison and Hinkley (1997) who use the term *phase scrambling*, and Braun and Kulperger (1997) who use the term *Fourier bootstrap*. In essence, one takes the discrete Fourier transform of a time series, replaces the phase with a new phase randomly chosen from the interval $(0, 2\pi)$, and back-transforms to obtain a new time series. Second-order properties are maintained by virtue of retaining the original amplitudes at their original locations in the original spectral estimate. If the original time series has an asymmetric marginal distribution then appropriate adjustments can be made in accordance with the so-called *rescaling methods* of Davison and Hinkley (1997), otherwise the standard algorithm suffices.

The algorithms as are follows. Denote the original series (of length n) as the array $x[t]$ with ranks r_t among the original unordered series.

Standard Algorithm

1. Compute the Discrete Fourier Transform $z[t] = DFT(x[t])$.
2. Randomise the phases; that is, randomly choose $\phi[t]$ from the uniform distribution of $(0, 2\pi)$, and put $z'[t] = z[t]\exp(i\phi[t])$.
3. Symmetrise the phases such that $Re(z''[t]) = Re(z'[t] + z'[n+1-t])/2$ and also $Im(z''[t]) = Im(z'[t] - z'[n+1-t])/2$.
4. Invert, putting $x'[t] = DFT^{-1}(z''[t])$.
5. The resulting series $x'[t]$ is the surrogate.

Rescaling Algorithm

1. Let $y_t = \Phi^{-1}\{r_t/(n+1)\}$, where Φ is the empirical distribution function of the original unordered series.
2. Apply the Standard Algorithm to $y_1, \ldots, y_n$, giving $Y_1^*, \ldots, Y_n^*$ (see above).
3. Set the surrogate series to be $X_t^* = x_{(r_t')}$, where r_t' is the rank of Y_t^* among $Y_1^*, \ldots, Y_n^*$.

0-7803-7011-2/01/$10.00 ©2001 IEEE

One can use surrogate time series to test a null hypothesis that the original series arises from a linear, stochastic, Gaussian stationary process. (Note that the rejection of this hypothesis covers a wide range of alternatives.) If a statistic from the original series is denoted as V_0 and the corresponding statistic from the j'th surrogate is denoted as V_j, with $E(V_j) = \mu_V$ and $\text{var}(V_J) = \sigma_V^2$, then one may use as the test statistic $|V_0 - \mu_V| / \sigma_V$, and calibrate against a Normal distribution, if appropriate. Timmer (1998) has illustrated this using the correlation dimension as the underlying statistic in the context of cyclostationary processes, demonstrating power to reject the null hypothesis in the presence of non-stationarity.

2. PHASE RANDOMISATION AND STATIONARITY

Second order properties, and some marginal shape properties, are known to be preserved under phase randomisation, the latter if the scaling method is used. We examine here the effect of phase randomisation on higher order moments and cumulants of a time series, in particular, to determine if conditions on linearity and stationarity are related to preservation of higher order properties under phase randomisation. In particular, for a time series $\{X_t\}$ with marginal mean μ, we treat higher central moments, of the form $E\{(X_t)\}^r$; higher order cumulants, of the form $E\left\{\prod_{j=1}^{r}(X_{t+k+j} - \mu)\right\}$; and higher order cross cumulants, of the lagged product form $E\{(X_t - \mu)^r (X_{t+k} - \mu)^r\}$. In each of these forms, $r = 3, 4, \ldots$, $k = 1, 2, \ldots$, and standard estimates were used in simulations.

Numerical experiments were based on some classical linear and non-linear time series models, including linear autoregression (AR), random walk (RW), bilinear stationary (BS), bilinear non-stationary (BN), GARCH stationary (GS), GARCH non-stationary (GN), threshold autoregression stationary (TS) and threshold autoregression non-stationary (TN). See Tong (1990) for a detailed discussion on the form and properties of these models.

Timeplots obtained from the numerical experiments showed broad agreement with the original data sets, and can be qualitatively compared as in the following table (using the rescaling method).

Model	AR	RW	BS	BN
Note	same	more symm.	same	same
Model	GS	GN	TS	TN
Note	larger vals.	same	same	same

When comparing the stationary with non-stationary models, it was sometimes possible to distinguish between them on the basis of higher order moments: the standard method produced zero values for odd moments; however, the rescaling method produced small values for the third moment for stationary series yet the same value as in the original series for non-stationary models. Thus, third order moments appear to have a reasonably good discriminatory ability for stationarity, and hence for convergence of MCMC procedures.

The behaviour of the higher order cumulants of the surrogates can be summarised according to the following table. The non-stationary models are shown in the last three rows of the table.

Model	Original	Standard	Rescaling
AR	small	odd near zero	all zero
BS	near zero	odd near zero	near zero
GS	near zero	near zero	near zero
TS	odd nr zero	odd near zero	small
RW	3rd, 4th small	odd smaller	small
BN	large	odd near zero	smaller
GN	large	smaller	smaller
TN	large	smaller	large

In addition, the distribution of higher order cumulants can be revealing in questions of stationarity, as the following table indicates. Modes refer to the number of modes of the empirical density function of the cumulants of surrogate series. (The standard method showed the same results as the rescaling method for all models.)

Model	Rescaling	Mode
AR	odd, even unimodal symm.	near zero
BS	unimodal	near zero
GS	unimodal, tails	near zero
TS	unimodal, tails	near zero
RW	multimodal	non-zero
BN	unimodal, tail	non-zero
GN	multimodal, tails	non-zero
TN	multimodal	non-zero

To summarise the results of these tables, we can comment as follows. In the case of the standard algorithm (i) higher order moments and cumulants are preserved for linear, Gaussian, stationary processes; (ii) higher moments are preserved for non-linear stationary processes, but not so for some higher order cumulants; (iii) second and cross-cumulants are not preserved for moderate and large lags if the process is linear non-stationary; and (iv) higher cumulants are not preserved for non-linear non-stationary processes. In the case of the rescaling algorithm (i) the method is inappropriate for linear, Gaussian, stationary processes as second

order cumulants are not preserved; (ii) higher order cumulants are not preserved for non-linear stationary processes; (iii) higher order cumulants and cross-cumulants are not preserved for linear non-stationary processes; (iv) higher cumulants are substantially different from the originals for non-linear non-stationary models; and (v) smoothing densities of higher order cumulants are multimodal, or at least unimodal with heavy tails, for non-stationary processes, while remaining unimodal for stationary processes.

It is on the above basis that convergence (i.e., stationarity) can be concluded from a run of an MCMC algorithm. Nur *et al.* (2001) give further details of the above methodology. In that paper, the methods were applied to some well-known data sets, and was found to reject convergence where some other less dynamically-driven methods concluded convergence of chains.

3. PHASE RANDOMISATION AS AN MCMC CONVERGENCE DIAGNOSTIC: AN EXAMPLE

There is a variety of tests for convergence of MCMC algorithms. Raftery and Lewis (1996) reduce the output of a chain to a two-state Markov chain and apply analytically explicit results to the modified output. Clearly, this is a form of discretisation and there is the possibility that important information about the original process may be lost. Heidelberger and Welch (1983) adopt a spectral analysis approach, as does Geweke (1992). These and other algorithms are available in the software package CODA (Best *et al.*, 1995).

We briefly describe an analysis of a widely-used 'benchmark' data set, and compare the relative performance of the existing methods with the present method.

The example concerns mortality rates in 12 hospitals performing cardiac surgery on babies: see Spiegelhalter *et al.* (1994). The authors proposed a random effects model for the number of deaths, r_j, in hospital j, with true unknown mortality probability p_j, as follows: $r_j \sim \text{Binomial}(p_j, n_j)$ $(j = 1, \ldots, 12)$, $\log p_j = b_j$, $b_j \sim N(\mu, \tau)$, $\tau = 1/\sigma^2$, $\mu \sim N\left(0, 10^{-6}\right)$, $\tau \sim \Gamma\left(10^{-3}, 10^{-3}\right)$. The analysis was restricted to a short run of 200 epochs. The timeplot of the MCMC run appeared to be similar to a bilinear stationary time series, based on the simulations we described in the previous section. Smoothing densities of the higher order cumulant estimates were plainly unimodal around zero, and standard quantile plots ascertained Normality of the surrogates' cumulants (supported strongly by the Shapiro-Wilks test). We can thus conclude that the MCMC algorithm has converged. This is supported by the *diag* assessment in BUGS, by Raftery and Lewis' test, and by Heidelberger and Welch's test. However, Geweke's test fails for this example because of the very short run, although it passes if a considerably longer run is used.

A detailed discussion of this analysis, along those of two other data sets, is given by D Nur, KL Mengersen and RC Wolff in an as yet unpublished manuscript. It indicates that phase randomisation performs at least as well as other existing methods in the assessment of MCMC convergence and, moreover, it is more informative about higher order statistical structures which in turn can classify stationarity and linearity. Their work also suggests that higher order cumulants from surrogate time series appear to be asymptotically Normally distributed, thus providing a route to robust formal testing of convergence (stationarity) hypotheses, and calibration thereof. There also appears to be evidence that the Metropolis-Hastings algorithm results in a Markov chain which is geometrically ergodic to the average when the target density is log-concave in the tails.

4. THEORETICAL ISSUES FOR PHASE RANDOMISATION

To give the above methodology a firm theoretical basis, it is required to prove that third (and higher order) cumulants of a stochastic process can be bootstrapped with accuracy $o\left(n^{-1/2}\right)$. Results of Götze and Hipp (1983) can be employed to verify this.

Let $\{\varepsilon_t\}$ be independent and identically distributed (iid) random variables. Generalising the Wold Decomposition Theorem for stationary processes, we write $X_t = \mu + \sum b_j \varepsilon_{t-j} + \sum\sum b_{kj} \varepsilon_{t-j}\varepsilon_{t-k} + \ldots$, and clearly X_t is non-linear if any of the higher order coefficients are non-zero.

We consider the formal Edgeworth expansion of order $s-2$ of the third cumulant of X_t, as follows. Define $Y_{jkt} = X_t X_{t-j} X_{t-k} - \sigma_{kj}$, where σ_{kj} is the theoretical third cumulant of X_t. Let Y_t denote the matrix form of Y_{jkt}. Götze and Hipp (1994) obtain valid formal Edgeworth expansions for sums of weakly dependent random vectors, with error of approximation $o\left(n^{-(s-2)/2}\right)$ if the moments of order $s+1$ are bounded, a conditional Cramer condition holds, and the random vectors can be approximated by other random vectors which satisy a strong mixing condition and a Markov-type condition. We extend their result, as follows.

Assume the following.

(A1) Let $\{\varepsilon_t\}$ be an iid sequence such that $E(\varepsilon_t) = 0$, $E\left(\varepsilon_t^2\right) = 1$, $E\left(\varepsilon_t^{3q(s+1)}\right) < \infty$, for some $s \geq 3$, $q \geq 1$.

(A2) For linear processes, $\sum_{r=m}^{\infty} |b_r| \leq c \exp(-\alpha m)$, $\alpha > 0$, for all m sufficiently large.

(A3) Let f denote a strongly contracting and continuous differentiable function, and let ε_t have density satisfying $E|f(\varepsilon_1, \ldots, \varepsilon_d)| < \infty$, f being positive and continuous.

(A4) $\Gamma = \lim_{n\to\infty} \left(n^{-1/2} \sum_{t=1}^{n} Y_t\right)$ exists and is positive definite. Denote the quantity under the limit as S_n.

Suppose that $|f(x)| \leq M(1 + |x|^{s_0})$ for every vector x. If the assumptions as set out in Götze and Hipp (1994) hold, then there exists $\delta > 0$ not depending on f and M, and, for any $k > 0$, the exists a constant $C = C(M) > 0$ not depending on f, such that

$$\left| Ef(S_n) - \int f d\psi_{n,s} \right| \leq Cw\left(f, n^{-k}\right) + o\left(n^{-(\delta-2)/2}\right),$$

where ψ is a functional of signed measures relating to the determinant of Γ, the term $o(.)$ depends on f through M only, and w is a supremum operator on a Lipschitz condition for f constraining y to be less than n^{-k} in norm.

Under conditions (A1) through (A4), the result holds for X_t, and the required Edgeworth expansion can be obtained.

In an as yet unpublished manuscript in preparation by D Nur, RC Wolff and KL Mengersen, the conditions for this theorem are being confirmed.

REFERENCES

Best, N, Cowles, MK and Vines, L (1995) CODA Manual v0.30. MRC Biostatistics Unit, Cambridge.

Braun, WJ and Kulperger, RJ (1997) Properties of a Fourier bootstrap method for time series. *Communications in Statistics – Theory and Methods* **26**, 1329-1336.

Davison, AC and Kinkley, DV (1997) *Bootstrap Methods and their Applications.* Cambridge University Press.

Geweke, J (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics* **4** (JM Bernardo, J Berger, AP Dawid and AFM Smith, eds). Oxford University Press, 169-193.

Götze, F and Hipp, C (1994) Asymptotic expansions for sums of weakly dependent random vectors. *Zeitschrift Wahrscheinlichkeitstheorie Verwandte Gebiete* **64**, 211-239.

Heidelberger, P and Welch, PD (1983) A spectral method for confidence interval generation and run-length control in simulation. *Communication ACM* **24**, 233-245.

Nur, D, Wolff, R C and Mengersen, KL (2001) Phase randomisation: numerical study of higher cumulants behaviour. *Computational Statistics and Data Analysis*, in press.

Raftery, A and Lewis, S (1992) Implementing MCMC. In *MCMC in Practice* (WR Gilks, ST Richardson and DJ Spiegelhalter, eds). Chapman and Hall: London, 115-130.

Spiegelhalter, DJ, Thomas, A, Best, NG and Gilks, WR (1994) BUGS: Bayesian Inference Using the Gibbs Sampler v40.4. Technical Report, MRC Biostatistics Unit, University of Cambridge.

Theiler, J, Eubank, S, Longtin, A, Galdrikian, B and Farmer, JD (1992) Testing for nonlinearity in time series: the method of surrogate data. *Physica D* **58**, 77-94.

Timmer, J (1998) Power of surrogate data testing with respect to non-stationarity. *Physical Review E* **58**, 5153-5156.

Tong, H (1990) *Non-linear Time Series: a Dynamical System Approach.* Oxford University Press.

IMPORTANCE SAMPLING ANALYSIS OF DIGITAL PHASE DETECTORS WITH CARRIER PHASE TRACKING

Francisco A. S. Silva *José M. N. Leitão*

Instituto de Telecomunicações - Instituto Superior Técnico 1049-001 Lisboa, PORTUGAL
{*sena, jleitao*}*@lx.it.pt*

ABSTRACT

In this work we introduce *importance sampling* techniques for the assessment of a class of open-loop digital phase modulation receivers with random carrier phase tracking in additive white Gaussian noise channels. We consider a symbol-by-symbol phase detector consisting of a bank of nonlinear stochastic filters tracking the random phase carrier and a decision algorithm driven by the filters' innovations. For the irreducible error floor assessment we use an importance sampling technique relying on large deviations principles that results in a multiple mode simulation density. The noisy operation of the receiver is addressed with an adaptive importance sampling technique. Simulations yield practically the same results obtained with conventional Monte Carlo with remarkable time gains.

1. INTRODUCTION

Symbol-by-symbol detection and random carrier phase tracking in additive white Gaussian noise (AWGN) channels is a particular scenario considered in reference [1]. Focusing on this particular problem, in this paper we are interested in the development of *importance sampling* (IS) techniques for fast simulation of the proposed receiver. Fast simulation, depends on the appropriate choice of a new simulation density which may be a difficult task particularly for highly nonlinear models that exhibit complicated error sets. This is the case analyzed in this paper. In [2] we presented IS results for the error floor operation of the receiver based on the error set knowledge and using *large deviations theory* (LDT) principles. To analyze the receiver behavior when observations are noisy, and adaptive importance sampling (AIS) technique must be used because the error set becomes unknown (see [3]).

The paper is structured as follows: Section 2 presents the communications model and some fundamental IS aspects. In Section 3 we derive the error set for density biasing using LDT, and present the main aspects of the AIS technique applied. Implementation aspects are also included in this section. In section 4 we show the results of our IS analysis.

This work was supported by Portuguese program Praxis XXI, under project 2/2.1/TIT/1583/95.

2. PROBLEM FORMULATION

2.1. Dynamics and observation model

Consider the discrete base-band received signal sampled N times per k^{th} symbol interval $[kT_s, (k+1)T_s]$ of duration T_s:

$$\mathbf{s}_n = \exp\left[j\left(\theta_n^{(s)} + \phi_n\right)\right] + \mathbf{v}_n, \qquad n = 1, \ldots, N$$

where $\theta_n^{(s)}$ is the digital phase sequence associated to one of M symbols, $\alpha_s \in \{\alpha_1, \cdots, \alpha_M\}$, ϕ_n is a discrete Brownian motion described by $\phi_n = \phi_{n-1} + \delta_n$, where δ_n is a zero mean white Gaussian sequence of variance σ_ϕ^2; $\mathbf{v}_n$ is a complex zero mean white Gaussian sequence of variance σ_v^2.

2.2. Receiver description

The receiver proposed in [1] consists of a bank of M 'matched' stochastic *nonlinear filters* (NLF) driven by the same input $\mathbf{s}_n$ and a decision algorithm driven by the filters innovations processes. The detector decides, at the end of the current symbol interval, according to a minimum Euclidean metric computed from those innovations. Parameters of the selected NLF are used as initial conditions to all NLFs for the next symbol interval (see [1] for details). This corresponds to a symbol aided decision criterion.

Matching to symbol α_s consists of eliminating the modulating sequence from the observation vector giving rise to observations denoted by $\mathbf{z}_n^{(s)}$. The NLF$^{(s)}$ propagates recursively probability densities of phase ϕ_n conditioned on the observations $\mathbf{z}_n^{(s)}$. Densities are represented, for this scalar phase process, as *Tikhonov* functions with mean $\hat{\phi}$ and concentration parameter γ. Propagation is accomplished in two steps, filtering (F) and prediction (P), implementing the following equations:

- Filtering

$$\hat{\phi}_n^F = \arctan \frac{z_{2,n}^{(s)} + \gamma_n^P \sigma_v^2 \sin \hat{\phi}_n^P}{z_{1,n}^{(s)} + \gamma_n^P \sigma_v^2 \cos \hat{\phi}_n^P} \quad (1)$$

$$\gamma_n^F = \left[\left\|\mathbf{z}_n^{(s)}\right\|^2 / \sigma_v^4 + \right.$$
$$\left. + \; 2\frac{\gamma_n^P}{\sigma_v^2}\left(z_{1,n}\cos\hat{\phi}_n^P + z_{2,n}\sin\hat{\phi}_n^P\right) + \left(\gamma_n^P\right)^2\right]^{1/2} \quad (2)$$

- Prediction

$$\hat{\phi}^P_{n+1} = \hat{\phi}^F_n \quad (3)$$

$$\gamma^P_{n+1} = \gamma^P\left(\gamma^F_n, \sigma^2_\phi\right) \quad (4)$$

For next symbol processing, all the M NLFs are initialized with parameters $\left(\hat{\phi}^P_{N+1}, \gamma^P_{N+1}\right)$ from the elected branch.

2.3. Modeling aspects and IS fundamentals

Fig. 1 is a schematic representation of the communication model considered in the previous subsections. In this figure $A_k = \alpha_s$ is

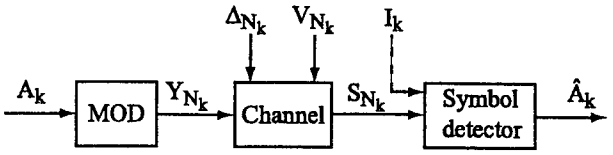


Fig. 1. Model for system simulation

the transmitted symbol with $\alpha_s \in \{\alpha_1, \cdots, \alpha_M\}$, and Y_{N_k} and S_{N_k} are the transmitted and received signal vectors respectively with N samples each during symbol interval $[kT_s, (k+1)T_s]$. V_{N_k} is the AWGN vector $V_{N_k} = [v_1, \ldots, v_N]_k$, and $\Delta_{N_k} = [\delta_1, \ldots, \delta_N]$ the phase increment vector. Since the symbol detector propagates density parameters from the $(k-1)^{th}$ symbol interval to the kth one, an observation record of size L

$$S^L = [S_k, S_{k-1}, \ldots, S_{k-L+1}], \quad L > 1 \quad (5)$$

$$S^1 = S_k, \quad L = 1 \quad (6)$$

influences estimate $\hat{A}_k$ through the initialization variable I_k. Accordingly, we define the associated transmitted signal and disturbance records Y^L and $U^L = [\Delta^L, V^L]$ respectively. In IS simulation, variable I_k that is normally hidden during *Monte Carlo* (MC) simulation will be generated to model the decision feedback mechanism.

The unbiased MC error probability estimator is

$$\hat{P}_e = \frac{1}{N_{MC}} \sum_{k=1}^{N_{MC}} g\left(A_k, \hat{A}_k\right)$$

where $g\left(A_k, \hat{A}_k\right)$ is one if $\hat{A}_k \neq A_k$ and zero otherwise, N_{MC} being the number of simulation runs. For *i.i.d.* errors $var\left\{\hat{P}_e\right\} = \sigma^2_{MC} = P_e(1-P_e)/N_{MC}$. IS simulation is intended to reduce high values of σ_{MC}/P_e associated with low P_e. For this, we must modify the simulation density $p([Y,U])$ obtaining $p^*([Y,U]^*)$ to generate the records $[Y,U]^*$ in order to obtain more frequent errors - the important and otherwise rare events. Although biased simulation densities lead naturally to biased error rate estimates, IS provides appropriate correction for each error event by means of the likelihood ratio $W([Y,U]^*_i) = p([Y,U]^*_i)/p^*([Y,U]^*_i)$. This yields the unbiased error rate IS estimator

$$\hat{P}^*_e = \frac{1}{N_{IS}} \sum_{i=1}^{N_{IS}} 1_E([Y,U]^*_i)\, W([Y,U]^*_i) \quad (7)$$

where $1_E(\cdot)$ is the indicator function for the error set E. When minimizing the IS estimator variance

$$\sigma^2_{IS} = \frac{1}{N_{IS}}\left[\int 1_E(y,u)W(y,u)p(y,u)dydu - P^2_e\right]$$

for a given N_{IS}, we act on $W(\cdot)$ through $p^*(\cdot)$.

Note that in (7) each simulation sample is in general a record with $U^L = [\Delta^L, V^L]$. This is not the case for $\hat{P}_e$ under conventional MC, and the difference between the approaches relies on the type of simulation - *stream simulation* for MC, and *error event simulation* for IS. Error event simulation introduced in [4] is considered as the method especially appropriate for IS in the presence of memory effects. It consists of generating independent realizations of U^L for a given information pattern A^L while testing $\hat{A}_k$ for error occurrence. We have modelled part of U^L through I making $U^1 = [\Delta^1, V^1, I]$. Bias done to the simulation density will be conditioned on each pattern A^L belonging to a finite denumerable set of configurations.

3. IMPORTANCE SAMPLING PROCEDURE

3.1. Error set derivation

Non-linear recursion in equations (1) to (4) along with the decision algorithm, preclude in general the error set analysis. Restricting our analysis to the space of the random phase increments denoted by $\mathcal{D}$, ($\sigma^2_\phi \neq 0, \sigma^2_v = 0$), we obtain a simpler model that is analytically tractable. For simplicity we derive here the error set for the binary case ($M = 2$). The filter equations in the error floor are:

$$\hat{\phi}^F_n = \arg\left(\mathbf{z}^{(s)}_n\right) = (\phi_n)_{2\pi} \qquad \gamma^F_n = \infty$$

$$\hat{\phi}^P_{n+1} = \hat{\phi}^F_n \qquad \gamma^P_{n+1} = \gamma^P\left(\sigma^2_\phi\right).$$

In the receiver, branch t decision metric conditioned on transmitted α_s (index t refers to *target* - the symbol α_t which is to be detected instead of α_s), becomes

$$\pi_{t|s} = \sum_{n=1}^{N} \left\| e^{j\left(\theta^{(s)}_n - \theta^{(t)}_n + \phi_n\right)} - e^{\left(-\sigma^2_\phi/2\right)} e^{j\hat{\phi}^{P(t)}_n} \right\|^2 \quad (8)$$

where $\hat{\phi}^{P(t)}_n = \theta^{(s)}_{n-1} - \theta^{(t)}_{n-1} + \phi_{n-1}$ and $\hat{\phi}^{P(t)}_1$ is the receiver initialization for all NLFs. In the error floor, $\hat{\phi}^{P(t)}_1$ is the sum of the random phase ϕ_0 (previous to ϕ_1) with an error term $I_{i|j} = \theta^{(j)}_N - \theta^{(i)}_N$ resulting from the $(k-1)^{th}$ decision feedback - erroneous detection of symbol α_i when the transmitted symbol was α_j.

We now define

$$E^D_{t|s,I_{i|j}} = \left\{\Delta_N \in R^N : \pi_{s|s} \geq \pi_{t|s}\right\}$$

as the error set in $\mathcal{D}$ conditioned on transmission of symbol α_s and initialization error $I_{i|j}$.

Equation $\pi_{s|s} = \pi_{t|s}$ defining the boundary $\partial E^D_{t|s,I_{i|j}}$ can not be solved in general. However we may identify $\partial\triangle$, as the infinite set of denumerable solutions satisfying

$$\cos\left(\delta_1 - I_{i|j}\right) = \cos\left(\epsilon^{t|s}_1 + \delta_1 - I_{i|j}\right) \quad (9)$$

$$\cos\delta_n = \cos\left(\epsilon^{t|s}_n + \delta_n\right), \quad n = 2, \ldots, N. \quad (10)$$

where

$$\epsilon_n^{t|s} = \theta_n^{(s)} - \theta_n^{(t)} - \left(\theta_{n-1}^{(s)} - \theta_{n-1}^{(t)}\right), n = 2, \ldots, N$$
$$\epsilon_1^{t|s} = \theta_1^{(s)} - \theta_1^{(t)}. \quad (11)$$

There is a finite collection of solutions $\partial\Delta_{2\pi} \subset \partial\Delta$ containing only 2^N elements obtained by the intersection $\partial\Delta_{2\pi} = \partial\Delta \cap \{\Delta_N \in [-\pi, \pi] \times \cdots \times [-\pi, \pi]\}$. Any single element in $\partial\Delta_{2\pi}$ allows the derivation of the remaining $2^N - 1$ elements and also of the symmetry center of the error set, which we designate by C_{Δ_N}. We are able to identify a point in $\partial\Delta_{2\pi}$ for each one of the 2^N quadrants Q_i wrt to C_{Δ_N}. In general C_{Δ_N} does not coincide with the origin of $\mathcal{D}$. As an example of such an error set, we show in Fig. 2 a diagram for $N = 2$ obtained by random generation of samples in $\mathbb{R}^2$. The solution set $\partial\Delta_{2\pi} = \{\Delta_{s_1}, \Delta_{s_2}, \Delta_{s_3}, \Delta_{s_4}\}$ is also represented. The error region presents a periodic structure generally non-connected and extending all over $\mathcal{D}$.

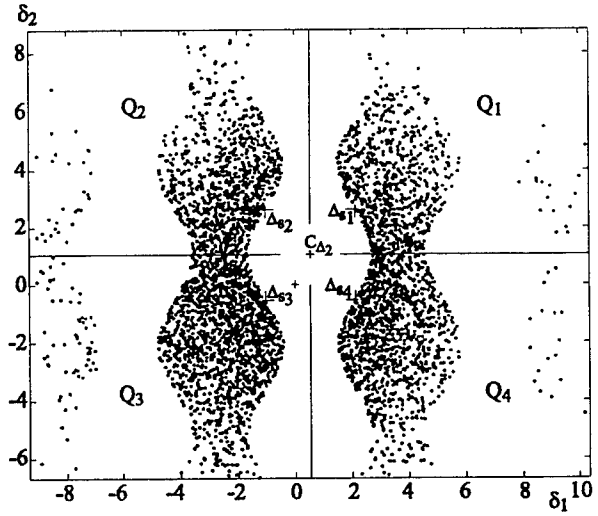


Fig. 2. Error set example for N=2. Gaussian sampled diagram

3.2. Large deviations and density biasing

Vector $\Delta_N \in \mathcal{D}$ is generated according to the simulation density

$$p(\Delta_N) = \left(\frac{1}{\sqrt{2\pi}\sigma_\phi}\right)^N \exp\left(-\sum_{n=1}^{N} \delta_n^2/(2\sigma_\phi^2)\right).$$

Identification of $E^D_{t|s,I_{i|j}} \subset \mathcal{D}$ shows that the *dominating point* (DP - see [5] - Definition 2), does not exist for the case under analysis. The DP would be used as a shift term for $p(\Delta_N)$ to obtain $p^*(\Delta_N)$ (a particular case of the exponential tilting). When there is no DP, Theorem 2 in [5] states the conditions for the use of a set of points $\{\nu_1, \ldots \nu_m\}$ for the mean translation of a finite number of terms that will constitute the biased density. This results in $p^*(\cdot)$ being a mixture of Gaussian terms for appropriate coverage of $E^D_{t|s,I_{i|j}}$. The set $\{\nu_1, \ldots \nu_m\}$ must contain at least all the *minimum rate points* (MRPs) of E (see also Def 2 in [5]) and other points, which, when used as bias vectors, will improve the coverage of E with the biased density $p^*(\cdot)$. Taking advantage of the symmetry shown by $E^D_{t|s,I_{i|j}}$ wrt C_{Δ_N}, we seek the quadrantwise minimization of the Euclidean norm of $\Delta_N \in \partial E$ in order to find all the MRPs wrt the sets $E_{Q_i} = E^D_{t|s,I_{i|j}} \cap Q_i$. Due to the big number of solutions, we selected for simulation biasing only the N_m solutions with the smaller Euclidean norms.

3.3. Density biasing using AIS

Optimization of IS density, consists now of biasing in the product space $\mathcal{D} \times \mathcal{V} \times \mathcal{I}$ since $(\sigma_v^2 \neq 0, \sigma_\phi^2 \neq 0)$. For the minimization of σ_{IS}^2 we use a stochastic search because we have no information about the error set $E^{DVI} \in \mathcal{D} \times \mathcal{V} \times \mathcal{I}$.

Considering the product space $\mathcal{D} \times \mathcal{V} \times \mathcal{I}$, with $\mathcal{V}$ being the $2N-$ dimensional noise sample space and $\mathcal{I}$ the one-dimensional space of the initialization phase error, we modelled the initialization error $\hat{\phi}_1^P$ as Gaussian its parameters being easily estimated in a short simulation preamble. The parameter γ_1^P was kept constant with its value in the error floor that is $\gamma_1^P = \gamma^P\left(\exp(-\sigma_\phi^2/2)\right)$.

We use a parametric AIS technique adapted from that proposed in [3]. We estimate the conditional mean $E\{(\Delta_N, V_N, I) \mid (\Delta_N, V_N, I) \in E^{DVI}\}$. Optimization must yield a multiple term solution that will constitute the modified density $p^*(\Delta_N, V_N, I)$. The proposed estimation cycle is increasingly repeated in our case, as σ_v^2 increases while σ_ϕ^2 is kept constant. The biased density is presented in the next subsection for $M > 2$. The major modifications we have done to the technique proposed in [3] consist in using as starting points for the search, the optimized biases in $\partial\Delta_\nu$ and no bias at all for V_N and I. This shortens the time required for starting the AIS algorithm. Quadrant separation in $\mathcal{D}$ is essential to keep the different bias terms separated during optimization.

3.4. Implementation aspects

In the error floor, $I_{i|j}$ depends on the result of estimate $\hat{A}_{k-1}$. However, the correct initialization ($I_{j|j} = 0$ in the error floor) happens naturally almost all the time for the modulation considered. We conducted tests with all the possible values for $I_{i|j}$, their a priori probability $P(I_{i|j})$ being estimated recursively, and the differences to use only $I_{j|j}$ were negligible. For the noisy channel, we modeled $I_{j|j}$ as Gaussian $(I \sim \mathcal{N}(0, s_I^2))$ as explained before with s_I^2 estimated in a short preamble due to its dependence from both σ_ϕ^2 and σ_v^2.

We considered until now binary signaling ($M = 2$). With an M-ary signaling scheme, the error set conditioned on α_s is the union

$$E_s = \bigcup_{\substack{i=1 \\ i \neq s}}^{M} E_{i|s}$$

which may render IS biasing suboptimal. To mitigate this, we introduce another level of multiple biasing in our IS simulation. Our biased density is then a Gaussian mixture of $(M-1) \cdot N_m$ terms for appropriate i target addressing and error set $E_{i|s}$ coverage respectively. The referred density becomes

$$p^*(\Delta_N, V_N, I|\alpha_s) = \sum_{m=1}^{N_m} \sum_{\substack{t=1 \\ t \neq s}}^{M} P(B(t,m)_{\alpha_s}) \times$$
$$\times p(\Delta_N, V_N, I|B(t,m)_{\alpha_s}) \quad (12)$$

where $p(\Delta_N, V_N, I|B(t,m)_{\alpha_s})$ is the $3N+1$ dimensional Gaussian term with mean $B(t,m)_{\alpha_s}$ - the bias vector in $\mathcal{D} \times \mathcal{V} \times \mathcal{I}$ if

α_s was transmitted. The $P\left(B(t,m)_{\alpha_s}\right)$ are the sampling probabilities for the bias terms. They were made inversely proportional to the Euclidean norms of the respective bias shift terms, but we do not know how much this option approaches optimality. Simulation density at the error floor, is a *mutatis mutandis* simplification of (12) since we are only generating $\Delta_N \in \mathcal{D}$.

4. RESULTS

We tested our IS methodologies in a practical example with 4-FSK modulation ($\Delta f = 2.6/(2\pi T_s)$ rads^{-1} between adjacent symbols). The number of samples per symbol was set to $N = 10$. Simulation gain, denoted by γ_s is defined by the ratio between the MC simulation time, T_{MC}, and the corresponding IS time, T_{IS}, ($\gamma_s = T_{MC}/T_{IS}$). Simulations were stopped when empirical precision reached a value lower than 10% (see for example [6]). Figure 3 represents simulation data corresponding to the error floors for values of σ_ϕ^2 ranging from 0.1 rad^2 $\left(1/\sigma_\phi^2 = 10\right)$ to 0.03 rad^2 $\left(1/\sigma_\phi^2 = 33.3\right)$. Density biasing was done in $\mathcal{D}$ according to LDT principles. The left vertical scale represents the estimate of symbol error probability, $\hat{P}_s$, whereas the right vertical scale represents simulation gain γ_s.

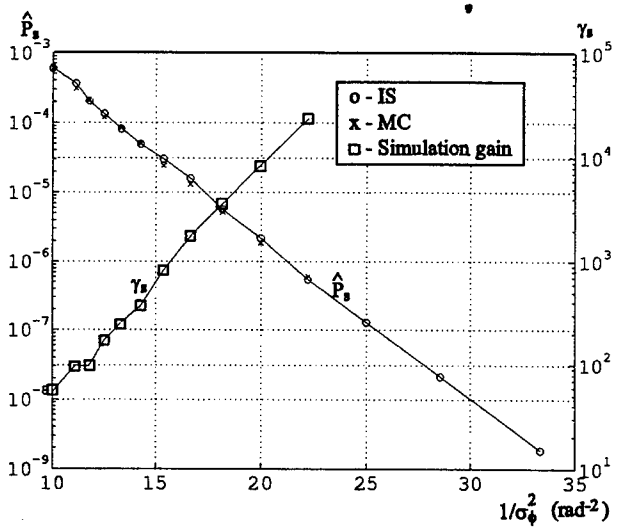


Fig. 3. Comparison of IS with MC in the error floor. $\hat{P}_s$ and γ_s versus σ_ϕ^{-2}

Notice the practical coincidence of $\hat{P}_s$ values of *IS* (mark o) with those of MC (mark $\times$) in the range $\left(1/\sigma_\phi^2 = 10\right)$ to $\left(1/\sigma_\phi^2 = 22.2\right)$. Simulation gains increase, in this range, from $\gamma_s = 55$ to $\gamma_s = 24000$. The value of T_{MC} for $\left(1/\sigma_\phi^2 = 22.2\right)$ is 13.7 hours using a PIII@450MHz computer.

The IS results presented in Figure 4 were obtained with AIS as they concern the receiver performance in a wide range of operating conditions (including the error floor). Also represented are the values of $\hat{P}_s$ obtained with MC for $\sigma_\phi^2 = 0.05$rad^2 and values of E_b/N_0 equal to 19, 22, 25 and 31 dB; the corresponding gains γ_s were 8.2, 54.3, 158.2 and 451 respectively; for $E_b/N_0 = 17$ dB, there is no practical simulation gain. Once again we stress the practical coincidence of the estimated values of $\hat{P}_s$ provided by both simulators. Points on the curve corresponding to $\sigma_\phi^2 = 0.045$ rad^2 took 10 minutes (with the above mentioned computer) in the range of [20, 50] dB, while the MC points in the same range would take an estimated 72.5 hours.

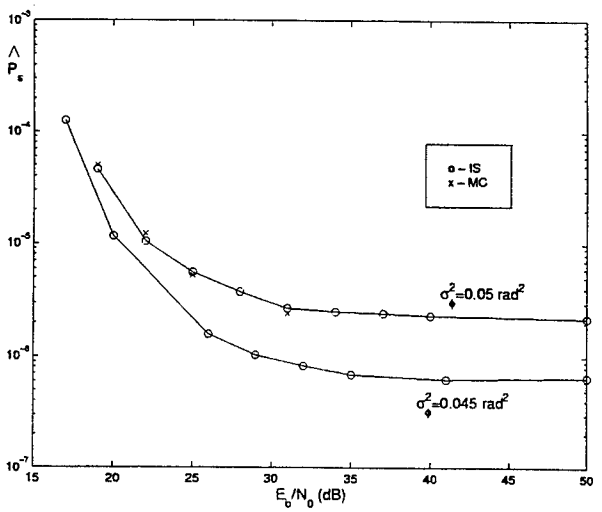


Fig. 4. $\hat{P}_s$ versus E_b/N_0 for two values of σ_ϕ^2

5. REFERENCES

[1] F. D. Nunes and J. M. N. Leitão, "A nonlinear filtering approach to estimation and detection in mobile communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 9, pp. 1649–1659, December 1998.

[2] F. Silva and J. Leitão, "Error floor assessment of digital phase detectors with carrier phase tracking. A large deviations approach," in *Proc. of the ISIT2001*, Washington, D.C., June 2001.

[3] J. S. Stadler and S. Roy, "Adaptive importance sampling," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 3, pp. 309–316, April 1993.

[4] D. Lu and K. Yao, "Improved Importance Sampling technique for efficient simulation of digital communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 1, pp. 67–75, January 1988.

[5] J. S. Sadowsky and J. A. Bucklew, "On Large Deviations Theory and asymptotically efficient Monte Carlo estimation," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 579–588, May 1990.

[6] Jyun-Cheng Chen, D. Lu, J. S. Sadowsky and K. Yao, "On Importance Sampling in digital communications - part I: Fundamentals," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 3, pp. 289–299, April 1993.

LONG-RANGE DEPENDENT ALPHA-STABLE IMPULSIVE NOISE IN A POISSON FIELD OF INTERFERERS

Xueshi Yang and Athina P. Petropulu

ECE Department, Drexel University
Philadelphia, PA 19104, USA
{yxs, athina}@ece.drexel.edu
Tel.215-895-2358 Fax.215-895-1695

ABSTRACT

In a wireless communication environment, interferers are often assumed to be Poisson distributed in space. While they are alive, they constantly emit pulses which interfere with signal reception at the observation point. Each transmitted pulse suffers a power-law attenuation with the distance to the receiver. It has been shown in the past that at baud rate sampling at the receiver, the interference is marginally α-stable. In this paper we consider the interference at different sampling points. We assume that the interferers have their own life of transmission sessions, the durations of which are heavy-tail distributed random variables. This assumption is consistent with the characteristics of multimedia traffic. We show that the resulting interference at different sampling points is a jointly α-stable process and exhibits long-range dependence in the generalized sense. Numerical simulations are consistent with our theoretical findings.

1. INTRODUCTION

In a wireless communications network, interference at the receiver is composed of interference due to other transmitting terminals (self-interference) and thermal noise.

Knowledge of the statistics of interference is important in signal detection. While the thermal noise term can be Gaussian distributed, the self-interference part significantly deviates from Gaussianity. The α-stable distribution [6] has been shown to be of particular interest for modeling self-interference.

According to a statistical-physical model, originating from the works of Furutsu and Ishida [2], and later advanced by Middleton [4] , Sousa [3], Nikias [6], [5], the contributions of a Poisson distributed field of interferers, subjected to power-law attenuation as they propagate to the receiver, add up to a marginally α-stable noise process.

In [3] [6] [5], the samples of the interference obtained at symbol rate, are assumed to be independent identically distributed. However, if one takes into account the characteristics of transmission periods of each interferer, the latter i.i.d. assumption does not hold. Once an interferer starts transmitting, it remains on for a certain period of time. This by itself introduces some correlation in the transmitted signal, which depends on the distribution of the session life.

In this paper we show that self-interference at any two sampling points are jointly α-stable distributed. By further assuming that the session duration of each interferer is heavy-tail distributed, we show that the self-interference sampling points is a long-range dependent process in the generalized sense. The assumption on heavy-tailed session durations (user holding times) is consistent with characteristics of high-speed wireline network measurements. Although no extensive studies have been conducted on the characteristics of wireless network traffic, assuming seamless connectivity between wireline and wireless networks would imply similar characteristics between wireless and wireline user holding times. The latter result suggests that the noise is strongly correlated and impulsive, which posts new challenges in signal detection at the receiver.

The paper is organized as follows. In next section, relevant mathematical background is provided. Following, in Section III, is the detail description of the noise model. The joint statistics are investigated in Section IV, followed, in Section V, by the proof of long-range dependence under the assumption that the session life of the interferers are heavy-tail distributed. Simulations and examples are given in Section VI.

2. MATHEMATICAL BACKGROUND

2.1. Multivariate α-Stable Distributions [8]

Vector $\mathbf{X} = (X_1, X_2, ..., X_d)$ is an α-stable random vector in $\mathbb{R}^d$ if and only if there exists a finite measure Γ on the unit sphere S_d of $\mathbb{R}^d$ and a vector μ in $\mathbb{R}^d$ such that $\Phi(\omega) = E\exp\{j\sum_{k=1}^{d}\omega_k X_k\}$, for $0 < \alpha \leq 2$ is given by

$$\Phi(\omega) = \exp\{-\int_{S_d}|(\omega,s)|^\alpha\left(1-j\text{sign}((\omega,s))\tan\frac{\pi\alpha}{2}\right)\cdot\Gamma(ds)+j(\omega,\mu)\Big\}, \quad (1)$$

if $\alpha \neq 1$; If $\alpha = 1$, $\tan\frac{\pi\alpha}{2}$ is replaced by $-\frac{2}{\pi}\ln|(\omega,s)|$. $\alpha \in (0,2]$ is the characteristic exponent.

Suppose $d = 1$, then S_1 consists of two points $\{-1\}$ and $\{1\}$, and the spectral measure Γ is concentrated on them. It becomes the univariate α-stable distribution. Denoted by $X \sim S_\alpha(\sigma,\beta,\mu)$, if $\alpha \neq 1$, its characteristic function is given by

$$\begin{aligned}\Phi(\omega) &= \exp\{-\|\omega|^\alpha[(\Gamma(\{1\})+\Gamma(\{-1\})) - j\text{sign}(\omega)\cdot(\Gamma(\{1\})-\Gamma(\{-1\}))\tan\frac{\pi\alpha}{2}] + j\mu\omega\Big\} \\ &\triangleq \exp\Big\{-\sigma^\alpha|\omega|^\alpha(1-j\beta\text{sign}(\omega)\tan\frac{\pi\alpha}{2}) + j\mu\omega\Big\}\end{aligned} \quad (2)$$

where σ is the scale parameter, β is the skewness parameter, and μ is the location parameter.

0-7803-7011-2/01/$10.00 ©2001 IEEE

2.2. Codifference

α-stable distributions are known for their lack of moments of order larger than α. In particular, for $\alpha < 2$, the second-order statistics do not exist. In such case, the role of covariance is played by the covariation or the codifference [8].

The codifference of two jointly $S\alpha S$, $0 < \alpha \leq 2$, random variables X_1 and X_2 equals:

$$R_{X_1,X_2} = \gamma_{X_1} + \gamma_{X_2} - \gamma_{X_1-X_2} \tag{3}$$

where γ_X is the scale parameter of the $S\alpha S$ variable X.

A quantity that is closely related to the codifference $R_{x(t+\tau),x(t)}$ is [8]:

$$\begin{aligned} I(\rho_1,\rho_2;\tau) = &- \ln E\{e^{\imath(\rho_1 x(t+\tau)+\rho_2 x(t))}\} \\ &+ \ln E\{e^{\imath\rho_1 x(t+\tau)}\} + \ln E\{e^{\imath\rho_2 x(t)}\}. \end{aligned} \tag{4}$$

The above quantity reduces to the codifference for the case of jointly $S\alpha S$ processes, i.e.

$$R_{x(t+\tau),x(t)} = -I(1,-1;\tau). \tag{5}$$

2.3. Long-Range Dependence

A second-order process $x(t)$ is called a (wide-sense) stationary process with long memory, or long-range dependence, if its autocorrelation function, $\rho(\tau)$, satisfies [9]:

$$\lim_{\tau\to\infty} \rho(\tau)/\tau^{\beta-1} = c \tag{6}$$

for some positive constant c and $\beta \in (0,1)$. From (6), it can be seen that a long-memory process is characterized by an autocorrelation that decays hyperbolically, as the lag τ increases. This is in contrast with the exponential decay corresponding to short memory processes, e.g. ARMA.

The following generalization of the concept of long memory process can be useful processes who lack autocorrelation. [10]

Definition 1 *Let $x(t)$ be a stationary process. We say that $x(t)$ is a long-memory process in a generalized sense, if $I(1,-1;\tau)$, as defined in (4), satisfies*

$$\lim_{\tau\to\infty} -I(1,-1;\tau)/\tau^{\beta-1} = c \tag{7}$$

where c is some real positive constant and $0 < \beta < 1$.

The notion of generalized long-memory has been used in [10] to study the dependence structure of the power-law shot noise. (See also [13])

3. THE INTERFERENCE MODEL

As in [3], we assume an infinite number of potential sources in the source domain. They are emitting pulses that may be seen at the receiver. Our basic unit of time is the symbol interval, or slot. In other words, we are considering a discrete time process. The fundamental assumptions of the interference model are the following.

1). At any given time slot, there are random number of emerging sources which begin to emit waveforms which interfere with the signal of interest. The number of emerging sources are a Poisson random variable. Moreover, their locations are also spatially Poisson distributed. In a two-dimensional space, the number of emerging sources in a region $\mathbb{R}$ of area A is Poisson distributed with density λA, i.e.

$$P[\text{Number of sources in } \mathbb{R} = k] = \frac{(\lambda A)^k}{k!} e^{-\lambda A}. \tag{8}$$

The parameter λ is not necessary constant. In a non-homogeneous case, a transformation can be performed to map the Poisson process from non-homogeneous to a homogeneous one, c.f.[3].

2). Once the interference source begin to transmit, it constantly emits waveforms for a random duration of times (*session life*). From the receiver's point of view, the resulted interference is a symmetrically distributed random variable.

3). The waveform propagation loss increases logarithmically with increasing distance between the source and the receiver. In terms of signal amplitude loss function, it can be written as

$$a(r) = \frac{1}{r^{\gamma/2}}, \tag{9}$$

where r is the distance and γ may vary from 1 to 6 in different environments.

4). The sources originated at different time slots are assumed to be independent of each other. The inception or termination of emission of a source will not affect any other sources.

The receiver using an omni-directional antenna is located in the center of the space (plane or volume) that we are interested. The received signal is given by

$$z(t) = s(t) + \sum_{i\in\text{active sources}} a(r_i)x_i(t), \tag{10}$$

where $s(t)$ is the signal of interest and the sum is the interference. We assume a standard correlation receiver, which correlate the received signal with a set of basis functions $\{\phi_k(t), k = 1,...,n\}$ and produce n-dimensional vectors $\mathbb{Z}, \mathbb{S}$ and $\mathbb{X}$, such that

$$\mathbb{Z}^l = \mathbb{S}^l + \sum_i \mathbb{X}_i^l \tag{11}$$

where the superscripts represent the l-th time slot, or symbol interval and the $\mathbb{X}_i$ represent the contribution from the i-th source.

Now consider the instantaneous interference at any symbol interval m. In order to calculate the instantaneous statistics, the number and locations of the active sources at any given symbol interval need to be specified.

Proposition 1 *Assuming the mean of the random session life of the interferers is finite, denoted by μ, and the density of the emerging sources in the space at every time slot is λ, asymptotically, the active number of sources in any time slot is Poisson distributed in the space with density $\lambda\mu$.* [1]

An immediate result follows the statement by use of the result in [3]

[1] Due to space limit, we omitted all the proofs in this paper.

Corollary 1 *If the received influence X_i for each interferer is spherically symmetric distributed*[2]*, the instantaneous interference is $S\alpha S$ distributed, with characteristic exponent $\alpha = 2/\gamma$.*

4. JOINT STATISTICS

The interference at different symbol time intervals constitutes a stochastic process. Its dependence depends on the transmitting characteristics of the co-channel users. It has already been shown that at any given symbol interval, the marginal distribution of the self-interference is alpha-stable distributed. It remains to be seen whether the interference at different time slots is jointly α-stable.

The interference at the m-th symbol interval is

$$\mathbb{Y}^m = \sum_i a(r_i)\mathbb{X}_i^m. \tag{12}$$

To simplify the presentation, we assume the vector $\mathbb{X}$ is one dimensional. To evaluate the joint statistics of $\mathbb{Y}$, we calculate the quantity

$$\mathrm{E}\left\{\exp[j\omega_1 Y^m + j\omega_2 Y^n]\right\}. \tag{13}$$

Specifically, we have the following result.

Proposition 2 *Let the mean of the random session life, L, be finite with complementary distribution*

$$\bar{F}_L(k) = P\left[L \geq k\right], \quad k = 1, 2, ... \tag{14}$$

Then the joint characteristic function of Y^m and Y^n is given by

$$\begin{aligned}\Phi_{m,n}(\omega_1,\omega_2) &= \exp\left\{-\sigma\left[H_1(\tau)|\omega_1|^\alpha\right.\right. \\ &+ \left.\left. H_2(\tau)|\omega_1+\omega_2|^\alpha + H_1(\tau)|\omega_2|^\alpha\right]\right\},\end{aligned} \tag{15}$$

where

$$\sigma = -\lambda\pi\int_0^\infty x^{-\alpha}d\Phi_0(x) \tag{16}$$

$$H_1(\tau) = \sum_{l=1}^{m-n}\bar{F}_L(l) \tag{17}$$

$$H_2(\tau) = \sum_{l=m-n+1}^{\infty}\bar{F}_L(l). \tag{18}$$

Here $\Phi_0(\cdot)$ is the characteristic function of X.

Remark 1 *By setting $\omega_2 = 0$, we obtain the first order characteristic function of the interference process, i.e.,*

$$\Phi(\omega_1) = e^{-\sigma\sum_{l=1}^{\infty}\bar{F}_L(l)|\omega_1|^\alpha}. \tag{19}$$

Recognizing that $\sum_{l=1}^{\infty}\bar{F}_L(l) = \mu$, which is the mean of the transmission life of the co-channel users, we get the same result as in the last section.

Remark 2 *As τ tends to infinity, $H_2(\tau)$ tends to zero, and $H_1(\tau)$ tends to the mean of transmission time, μ. The joint characteristic function may be simplified as*

$$\lim_{\tau\to\infty}\Phi_{m,n}(\omega_1,\omega_2) = e^{-\sigma\mu(|\omega_1|^\alpha+|\omega_2|^\alpha)}. \tag{20}$$

[2]A random vector $\mathbb{X}$ is spherically symmetric if its characteristic function depends only on the Euclidean norm of t, i.e. $\Phi_X(t) = \phi(||t||)$.

Eq.(20) implies that when the distance between two samples becomes asymptotically large, they are becoming independent α-stable random variables.

Corollary 2 *Indeed, the interference at two different time slots, which are separated by τ, are jointly α-stable distributed. They may be represented by different linear combinations of independent alpha-stable random variables.*

5. LONG-RANGE DEPENDENCE

Since the interference at any two symbol intervals are jointly $S\alpha S$, the conventional tools such as auto-correlation that measures the dependence structure of a stochastic process are not applicable. We adopt the codifference as previously defined in (4). The motivation behind it is that in many practical communication systems, the session life of the interferers are indeed heavy-tail distributed.

An example is given by the communication links in a spread spectrum packet radio networks. In a spread spectrum network, multiple access terminals use the same channel. The signals received at the receiver consists of superposition of the signals from all the users in the network. Assuming no multiuser detection and power-control, the interference from other users, or self-interference, falls into the scenario described in this paper. As more and more wireless users are equipped with data transmission enabled cellphone, the resource request holding time is found to exhibit more and more variation. In other words, the holding time are heavy-tail distributed. (cf. [7])

A simple but reasonable assumption on the distribution of the session life the interferers is that they are Zipf distributed. Zipf distribution is a discrete version of the more familiar Pareto distribution. A random variable X has a Zipf distribution [1] if

$$P\{X \geq k\} = [1 + (\frac{k-k_0}{\sigma})]^{-\alpha_L}, \quad k = k_0, k_0+1, k_0+2... \tag{21}$$

where k_0, the location parameter, is an integer, σ, the scale parameter, is positive and α, the tail index is positive. In this paper, to simplify presentation, we set $\sigma = k_0 = 1$, and $\alpha > 1$, which implies the mean of the session life is finite.

Proposition 3 *Asymptotically, the process formed by any components of the self-interference at different symbol intervals is a long-range dependent α-stable process, i.e.,*

$$\lim_{n\to\infty}\frac{-I(1,-1;n)}{n^{\alpha_L-1}} = c \tag{22}$$

where

$$\begin{aligned} n &: \text{ time distance between symbol intervals;} \\ \alpha_L &: \text{ tail index of the session life-time;} \\ c &: \text{ positive constant.}\end{aligned} \tag{23}$$

6. SIMULATIONS

In this section, we performed numerical simulations, of which settings are in accordance with afore-mentioned scenario. We assume standard correlator receivers for a communication link which is subjected to a Poisson field of interferers. The density of the interferers $\lambda = 20$, and and the session life is Zipf distributed with $\alpha = 1.2$. The random amplitude of the interference waveform are rectangular pulses with random amplitude of 1 or -1. Fig.1

shows the self-interference presented in the receiver. We apply the characteristic function based method [11] to estimate the tail index, and the result is 0.6269, which is very close to the theoretical value $2/\gamma$ ($\gamma = 3.$).

As shown in the last section, when the session life is heavy-tail distributed, the interference must exhibits long-range dependence. We calculated the codifference of the interference y_k and the result is shown in Fig.2 in a log-log scale. The linearity of the log-log plot confirms that the self-interference is long-range dependent in the generalized sense. The estimated slope is -0.1570, which is in good accordance to the theoretical value, $\alpha - 1 = -0.2$.

7. CONCLUSIONS

In this paper, we show that in a Poisson field of interferers, where the path loss is a power-law function, the interference at different time slots are jointly alpha-stable distributed. If we assume further that the transmission session life of the interferers are heavy-tail distributed, the resulted interference is long-range dependent in the generalized sense. Numerical simulations confirms our theoretical derivations.

8. REFERENCES

[1] B.C. Arnold, *Pareto Distributions*, International Co. Publishing House, Maryland, 1983.

[2] K. Furutsu and T. Ishida, "On the theory of amplitude distribution of impulsive random noise," *J. Applied Physics*, Vo.32, No.7, 1961.

[3] E.S. Sousa, "Performance of a spread spectrum packet radio network link in a Poisson field of interferers", *IEEE Trans. on Info. Theo.*, Vol. 38, No.6, Nov. 1992.

[4] D. Middleton, "Statistical-Physical Models of Electromagnetic Interference,", *IEEE Trans. Electromagnetic Compatibility*, Vol.EMC-19, No3, Aug. 1977.

[5] J. Ilow, D. Hatzinakos, and A.N. Venetsanopoulos, "Performance of FH SS radio networks with interference modeled as a mixture of Gaussian and Alpha-stable noise,", *IEEE Trans. on Comm.*, Vol.46, No.4, Apr. 1998.

[6] C. L. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications,* New York: Wiley, 1995.

[7] T. Kunz, T. Barry, X. Zhou *et al*, "WAP traffic: description and comparison to WWW traffic," *Proc. of 3rd ACM international Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Boston, USA, Aug. 2000.

[8] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random processes: Stochastic Models with Infinite Variance,* New York: Chapman and Hall, 1994.

[9] J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, New York, 1994.

[10] A.P. Petropulu, J-C. Pesquet, X. Yang, "Power-law shot noise and relationship to long-memory processes," *IEEE Trans. on Sig. Proc.*, Vol.48, No.7, July 2000.

[11] S. Kogon and D. Williams, "On the characterization of impulsive noise with α-stable distributions using Fourier techniques", *Proceedings of the 29th Asilomar Conference of Signals, Systems and Computing*, 1995.

[12] X. Yang, A.P. Petropulu, and J.-C. Pesquet, "Estimating long-range dependence in impulsive traffic flows,", *Proc. of ICASSP 01*, Salt Lake City, UT, 2001.

[13] X. Yang, A.P. Petropulu, "The extended alternating fractal renewal process for modeling traffic in high-speed communication networks", *IEEE Trans. on Sig. Proc.*, Vol.49, No.7, July, 2001.

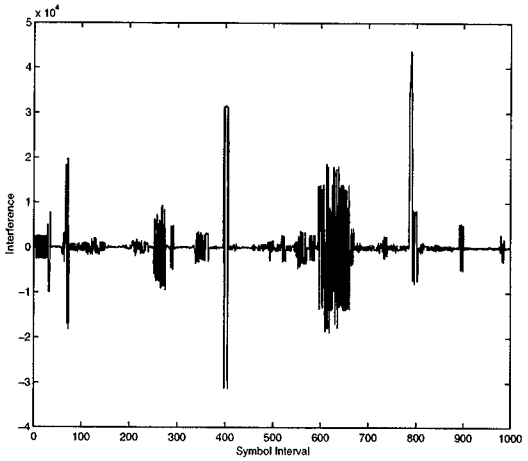


Figure 1: Self-interference presented in a communication link in a Poisson field of interferers.

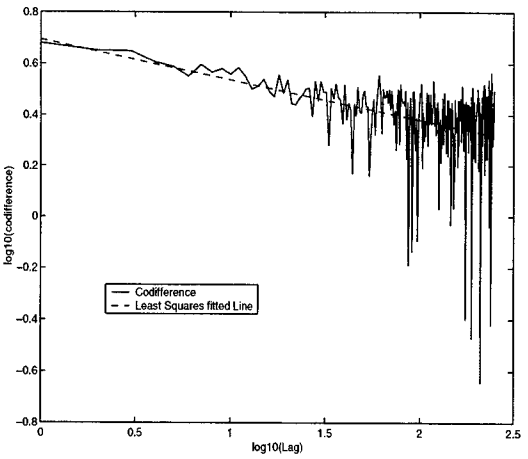


Figure 2: Log-log plot of generalized codifference of the Self-interference. The slope of the least squares fitted line is -0.1570.

SCORE FUNCTIONS FOR LOCALLY SUBOPTIMUM AND LOCALLY SUBOPTIMUM RANK DETECTION IN ALPHA-STABLE INTERFERENCE

Christopher L. Brown

Department of Signals and Systems
Chalmers University of Technology
SE-412 96 Göteborg, Sweden
chris.brown@ieee.org

ABSTRACT

Approximations to the locally optimum and locally optimum rank score functions for the detection of a known signal in additive symmetric α stable interference have recently been shown to introduce only slight performance loss. Here, the location of the apices of the nonlinearities is shown to follow an approximate linear relationship with the characteristic exponent, α, of the interference. The distribution of the corresponding test statistics is also found to be approximately Gaussian. These findings make implementation of the detectors more feasible and remove some expensive computational burden.

Keywords: locally optimum detection, nonparametric statistics, alpha-stable distribution.

1. INTRODUCTION

Detectors of a known signal in additive interference of unknown power have been formulated for a number of interference distributions – the most famous and widely used being the matched filter (MF) for Gaussian interference.

Sources of impulsive interference have presented some interesting problems for detection. Many conventional techniques perform poorly for heavy-tailed distributions. Additionally, the α-stable (αS) distribution, which can be used to describe some impulsive processes [1], has no general closed form expression for its pdf, thus making difficult the use of likelihood ratio procedures. For this reason, it is necessary to investigate alternative strategies for the detection of signals in αS interference.

Consider the model

$$\boldsymbol{X} = \theta\, \boldsymbol{s} + \boldsymbol{W} \tag{1}$$

where $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$ is the model for the real-valued observations, $\boldsymbol{s} = [s_1, \ldots, s_n]^T$ is the known, deterministic signal to be detected, $\boldsymbol{W} = [W_1, W_2 \ldots, W_n]^T$ is a stationary, iid, symmetric α-stable (SαS) interference process and θ is a non-negative, real, unknown parameter. To determine the presence of $\boldsymbol{s}$ in $\boldsymbol{X}$, the tested hypothesis is $\mathsf{H} : \theta = 0$ against $\mathsf{K} : \theta > 0$.

In the following sections, signal detection in SαS interference using correlation and rank-based detectors is discussed, including suggested approximations that will aid in the implementation of the said detectors. Following that, the distribution of the test statistics under consideration is discussed.

This work was conducted while the author was with the Australian Telecommunications Research Institute & School of Electrical and Computer Engineering, Curtin University of Technology, Australia

2. LOCALLY OPTIMUM DETECTORS

Locally optimum (LO) tests attain the best detection performance amongst the class of detectors of the same size for weak signal conditions, that is, they maximise the slope of the detector power function at $\theta = 0$. A LO detector for the detection of a signal in the model (1) uses the test statistic [2]

$$T_{\mathrm{LO}}(\boldsymbol{X}) = \sum_{i=1}^{n} s_i\; g_{\mathrm{LO}}(X_i) \tag{2}$$

where the nonlinear score function is

$$g_{\mathrm{LO}}(x) = -\frac{f'_W(x)}{f_W(x)}$$

and $f'_W(x)$ is the first derivative of $f_W(x)$, the pdf corresponding to the random process W.

Due to the absence of closed form expressions for $f_W(x)$ when W is αS distributed, $g_{\mathrm{LO}}(x)$ cannot be found exactly. In [1], g_{LO} was found numerically for some SαS cases and plotted for a number of values of α. These results are reproduced in Figure 1. Here, and throughout the rest of the paper, the scaling parameter of the SαS distribution is set to 1, $\gamma = c^\alpha = 1$. No generality is lost since if Y is an SαS random process with scaling parameter c, then $Z = Y/c$ is an SαS random process with scaling parameter of 1.

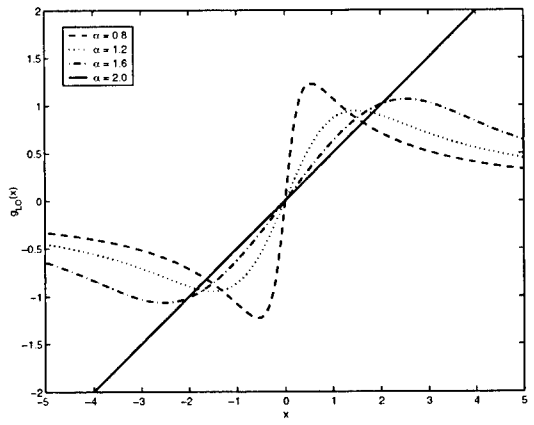


Fig. 1. Locally Optimum score functions for various SαS distributions.

While these numerical approximations can be made to be extremely accurate, their computational complexity requires that practical implementation uses coarser approximations. These approximations invalidate the "locally optimum" feature of the detector, and yield locally suboptimum (LSO) detectors.

Recently, the LSO-power nonlinearity was introduced [3, 4]

$$g_{\text{LSO-p}}(x) = \begin{cases} c\,x & ,|x| \le \lambda \\ \frac{\alpha+1}{x} & ,|x| > \lambda \end{cases}$$

where $c = \frac{\alpha+1}{\lambda^2}$ and $\lambda = \arg\max_x g_{\text{LO}}(x)$ is the location of the peak of g_{LO}. It was shown that the detector using this nonlinearity achieves near locally optimum detection. The nonlinearity decays at the same asymptotic rate as g_{LO} [1], as is shown in Figure 2. Other approximations to the computationally complex g_{LO} have been suggested in [5, 6, 7].

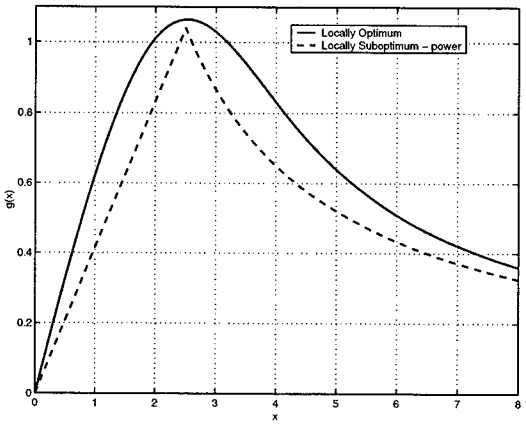


Fig. 2. Locally Optimum and Locally Suboptimum nonlinearities for $\alpha = 1.6$.

When the $g_{\text{LSO-p}}$ nonlinearity was introduced, the location of the apex, λ, was found by numerically locating the peak of the g_{LO} nonlinearity. No expressions exist for finding λ for a given α. However, when the location of the apex of $g_{\text{LO}}(x)$ is plotted against α, the relationship appears very linear. This is shown in Figure 3 along with the line of best fit. The equation of the line is

$$\lambda \approx 2.73\ \alpha - 1.75$$

Note that as $\alpha \to 2$, $g_{\text{LO}}(x)$ becomes a straight line and, therefore, the approximation breaks down.

With the aid of this relationship, it is now feasible to construct a LSO detector using the LSO-power nonlinearity without a heavy computational burden.

3. LOCALLY OPTIMUM RANK DETECTORS

It has long been accepted that by using a weak set of assumptions, rank-based tests can achieve robust performance while often only suffering slight losses in efficiency against parametric tests [8]. The locally optimum *rank* detectors (LOR) when W is symmetrically distributed uses the following test statistic [9]

$$T_{\text{LOR}}(\boldsymbol{X}) = \sum_{i=1}^{n} s_i\ \mathsf{sgn}(X_i)\ g_{\text{LOR}}(r_i) \qquad (3)$$

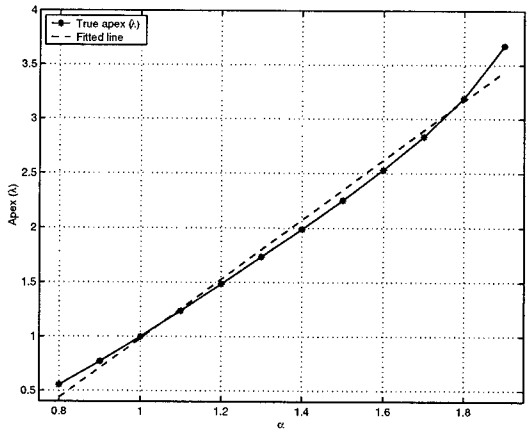


Fig. 3. Location of the apex of $g_{\text{LO}}(x)$ for varying α, as well as the fitted line of best fit.

where

$$g_{\text{LOR}}(i) = \mathsf{E}_{\mathsf{H}}\left[g_{\text{LO}}\left(|X|_{(i)}\right)\right] \quad ,$$

r_i is the rank of $|X_i|$ in the set $[\,|X_1|, |X_2|, \ldots, |X_n|\,]$ and $|X|_{(i)}$ is the set's ith smallest member. $\mathsf{E}_{\mathsf{H}}[\cdot]$ denotes the expectation operation under the hypothesis H. Asymptotically as $n \to \infty$, the LOR and LO detectors become equivalent.

The LOR nonlinearities, g_{LOR}, for a number of SαS distributions have been approximated numerically [3, 4] and are shown in Figure 4. In contrast to g_{LO} which has a slow rate of decay and infinite span, g_{LOR} need only be evaluated at n points. The results may be found *once* to a high degree of accuracy off-line and stored for on-line detection.

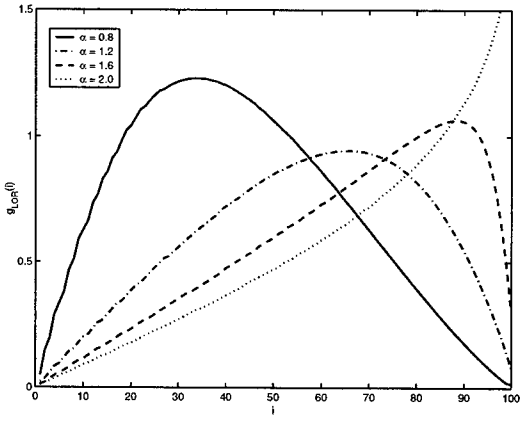


Fig. 4. Locally Optimum Rank score functions for various SαS distributions when $n = 100$.

If off-line estimation of g_{LOR} is not possible, then an appropriate locally suboptimal rank (LSOR) nonlinearity is a triangular score function [3] where, again, the location of the apex varies with α. These values, normalised by the sequence length n, are shown in Figure 5 with the line of best fit having the equation $0.5962\ \alpha - 0.0873$.

An interesting special case is when $\alpha = 1$, i.e. the Cauchy distribution. It has already been noted that an optimal Cauchy re-

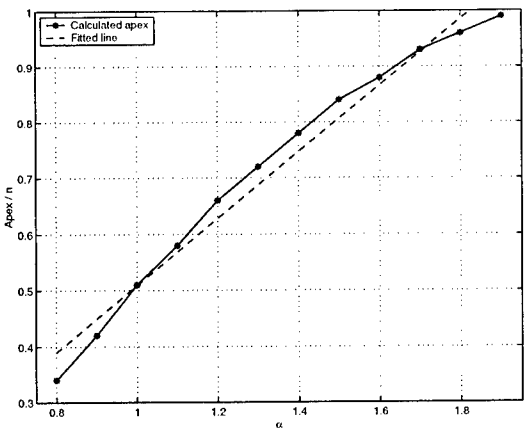


Fig. 5. Location of the apex, divided by the sequence length n, of $g_{\text{LOR}}(x)$ for varying α, as well as the line of best fit.

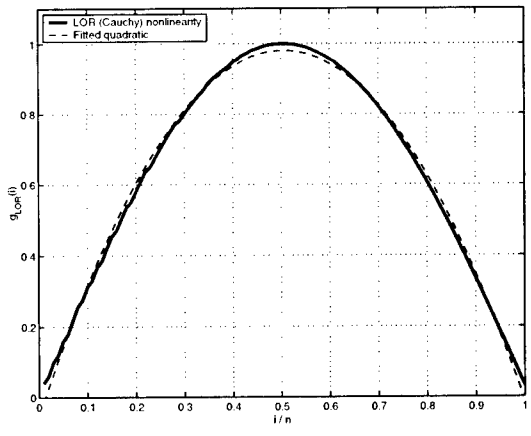


Fig. 6. The LOR score function and a quadratic of best fit, for the Cauchy distribution ($\alpha = 1$).

ceiver performs very well for a range of values of α, but particularly when it is close to 1 [7]. In Figure 6 it can be seen that g_{LOR} for $\alpha = 1$ is very well fitted by a quadratic score function

$$g(i) = -4.05\left(\frac{i}{n}\right)^2 + 4.09\left(\frac{i}{n}\right) - 5.29 \times 10^{-2} \quad .$$

It should be remembered that scaling a score function does not affect the corresponding detector's performance. Therefore, any centred quadratic function would suffice.

4. DISTRIBUTION OF TEST STATISTICS

Although the LO, LSO, LOR and LSOR detectors all use different score functions, the similarity in their structure as correlators means the distribution of their test statistics are very similar.

4.1. LO and LSO Detectors

Recall that the LO and LSO detector statistics have the form

$$T(\boldsymbol{X}) = \sum_{i=1}^{n} s_i \; g(X_i)$$

that is, the test statistic is the sum of independent random variables, assuming the X_i are iid and s_i is some bounded, known sequence. Under H, the summed variables have similar distributions, differing only in the non-constant scale, s_i. If $g(X)$ has finite variance, the Central Limit Theorem may be invoked, meaning $T(\boldsymbol{X})$ is asymptotically Gaussian.

Since only *symmetric* αS distributions are considered here then under H, $\mathsf{E}[g(X)] = 0$. Consequently,

$$\begin{aligned} \mathsf{E}[T(\boldsymbol{X})] &= 0 \\ \mathsf{var}[T(\boldsymbol{X})] &= \mathsf{var}[g(X)] \times \sum_{i=1}^{n} s_i^2 \quad . \end{aligned}$$

To determine if $g(X)$ has finite variance, an asymptotic expansion of the pdf of a standardised SαS random variable when $\alpha < 2$ and as $|x| \to \infty$, as given by [1] is used

$$\begin{aligned} f(x) &= \sum_{k=1}^{K} \frac{b_k}{|x|^{\alpha k+1}} + O\left(|x|^{-\alpha(K+1)-1}\right) \\ \text{where} \quad b_k &= -\frac{1}{\pi}\frac{(-1)^k}{k!}\Gamma(\alpha k+1)\sin\left(\frac{k\alpha\pi}{2}\right) \quad . \end{aligned}$$

This series may be approximated by its first term as this is the term with the slowest rate of decay for $|x| \to \infty$,

$$f(x) \approx \frac{b_1}{|x|^{\alpha+1}} \quad .$$

To determine if $\mathsf{var}[g_{\text{LO}}(X)] < \infty$, consider the integral

$$I = \int g_{\text{LO}}^2(x) f(x) dx$$

and its approximation using its highest order term as $x \to \infty$

$$\begin{aligned} I &\approx \int \frac{(\alpha+1)^2}{x^2}\frac{b_1}{x^{1+\alpha}}dx \qquad (4) \\ &= (\alpha+1)^2 \, b_1 \int x^{-3-\alpha} dx \\ &= -\frac{(\alpha+1)^2}{\alpha+2} \, b_1 \, x^{-2-\alpha} \quad . \end{aligned}$$

The highest order term of I does *not* diverge to $\pm\infty$ as $x \to \pm\infty$, and therefore, neither will I. Furthermore, both $g_{\text{LO}}(x)$ and $f(x)$ are bounded functions, therefore it can be concluded that evaluation of the integral, I, between $-\infty$ and $+\infty$, that is, the variance of $g_{\text{LO}}(X)$, is finite.

Any further terms taken from the asymptotic expansion of the pdf in (4) will have faster rates of decay, therefore it can be taken that the variance of $g_{\text{LO}}(X)$ and $g_{\text{LSO-P}}(X)$ are finite and the corresponding test statistics are asymptotically Gaussian (see Figure 7).

4.2. LOR and LSOR Detectors

Now consider the general form of the rank-based detector statistics considered here

$$T_{\text{R}}(\boldsymbol{X}) = \sum_{i=1}^{n} s_i \; \mathsf{sgn}(X_i) \; g_{\text{R}}(r_i)$$

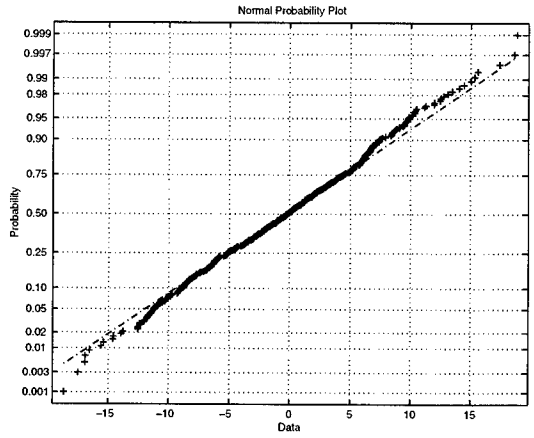


Fig. 7. Normal Probability Plot of 500 $T_{\text{LO}}(\boldsymbol{X})$ statistics calculated for $n = 100$ observations, $\alpha = 1.6$ and $s_i = 1$.

While $\mathsf{sgn}(X_i)$ is independent of $\mathsf{sgn}(X_j), i \neq j$, the same cannot be said of the ranks. If the possibility of ties is neglected, each r_i is an integer between 1 and n and each rank integer occurs only once, that is $r_i \neq r_j$ if $i \neq j$. Then, clearly, the ranking procedure introduces dependence between the terms. However, as $n \to \infty$, this dependence becomes negligible, and therefore *asymptotically* these terms are independent. Again, the Central Limit Theorem can be used to assert the asymptotic Gaussianity of $T_{\text{R}}(\boldsymbol{X})$. This is confirmed experimentally by the Normal Probability Plot in Figure 8.

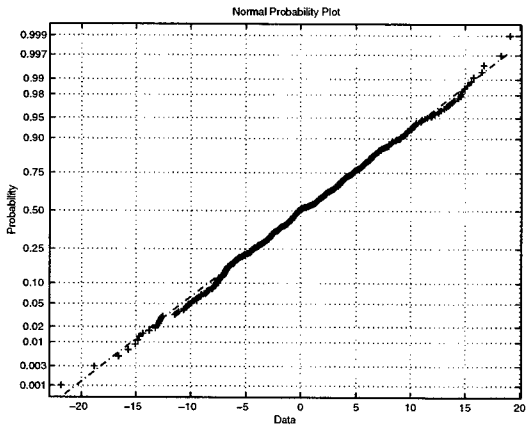


Fig. 8. Normal Probability Plot of 500 $T_{\text{LOR}}(\boldsymbol{X})$ statistics calculated for $n = 100$ observations, $\alpha = 1.6$ and $s_i = 1$.

If $|g_{\text{R}}(r_i)| < \infty$ for all $i = 1, \ldots, n$, then

$$\mathsf{E}\left[g_{\text{R}}(r)\right] = \frac{1}{n}\sum_{i=1}^{n} g_{\text{R}}(i) \quad \text{and}$$

$$\mathsf{E}\left[g_{\text{R}}^2(r)\right] = \frac{1}{n}\sum_{i=1}^{n} g_{\text{R}}^2(i)$$

will be finite for finite n.

The distribution of the test statistic $T_{\text{R}}(\boldsymbol{X})$ under H is independent of the distribution of X, provided it is symmetric. While its *exact* distribution may be calculated for any g_{R} and $\boldsymbol{s}$, in practice, this is tedious and a suitably accurate approximation can be made using the Gaussian distribution with

$$\begin{aligned} \mathsf{E}\left[T_{\text{R}}(\boldsymbol{X})\right] &= 0 \\ \mathsf{var}\left[T_{\text{R}}(\boldsymbol{X})\right] &= \frac{1}{n}\sum_{i=1}^{n} g_{\text{R}}^2(i)\sum_{j=1}^{n} s^2(j) \quad . \end{aligned}$$

Further discussion on the distribution of rank-based detection statistics can be found in [9].

5. CONCLUSIONS

Previous contributions have shown that approximations to LO and LOR detection can be achieved through other score functions that are more readily implementable. This concept has been extended here by finding an approximate linear relationship between the respective apices of g_{LO} and g_{LOR}, and α. Investigation of the distribution of the detection test statistics has also shown that, for finite sample size, they are approximately Gaussian. As a result of these findings, practical implementation of these detectors is easier.

6. REFERENCES

[1] C. L. Nikias and M. Shao. *Signal Processing with Alpha-Stable Distributions and Applications*. John Wiley & Sons, 1995.

[2] S. A. Kassam. *Signal detection in non-Gaussian noise*. Springer texts in electrical engineering. Springer-Verlag, New York, 1988.

[3] C. L. Brown and A. M. Zoubir. A nonparametric approach to signal detection in impulsive interference. *IEEE Transactions on Signal Processing*, 48(9):2665–2669, September 2000.

[4] C. L. Brown and A. M. Zoubir. Locally suboptimal and rank-based known signal detection in correlated alpha-stable interference. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'00*, volume I, pages 53–56, Istanbul, Turkey, June 2000.

[5] S. Ambike, J. Ilow, and D. Hatzinakos. Detection of binary transmission in a mixture of Gaussian noise and impulsive noise modeled as an alpha-stable process. *IEEE Signal Processing Letters*, 1(3):55–57, March 1994.

[6] E. E. Kuruoğlu, C. Molina, and W. J. Fitzgerald. Approximation of α-stable probability densities using finite Gaussian mixtures. In *Proceedings of EUSIPCO'98*, Rhodes, Greece, September 1998.

[7] G.A. Tsihrintzis and C.L. Nikias. Performance of optimum and suboptimum receivers in the presence of impulsive noise modeled as an alpha-stable process. *IEEE Transactions on Communications*, 43(2-4):904–914, March 1995.

[8] J. D. Gibson and J. L. Melsa. *Introduction to Nonparametric Detection with Applications*. Academic Press, 1975.

[9] T. P. Hettmansperger. *Statistical Inference Based on Ranks*. John Wiley, New York, 1984.

Kernel Approach to Discrete-Time Linear Scale-Invariant Systems

Seungsin Lee[†], Raghuveer Rao[‡]
[†]Center for Imaging Science
Rochester Institute of Technology
[‡]Department of Electrical Engineering
Rochester Institute of Technology
79 Lomb Memorial Drive
Rochester, NY 14623 USA

ABSTRACT

Zhao and Rao have proposed linear scale-invariant systems that operate with continuous dilation but in discrete-time. This was done through a discrete-time continuous-dilation operator which tacitly uses warping transforms such as bilinear transforms to implement conversion from discrete time frequency to continuous time frequency. This paper introduces a more general method based on kernels for effecting the dilation. It is shown that the warping function based scaling is a special case. The kernel approach results in an alternative formulation of discrete-time linear scale-invariant systems that possesses desirable properties not seen in the earlier formulation.

1. INTRODUCTION

The previous work of Zhao and Rao [10], [17]-[20] has shown that it is possible to formulate continuous dilation Linear Scale Invariant (DLSI) systems in discrete-time. The basis for their formulation is provided by a definition of scaling or dilation in discrete-time using warping and unwarping functions. Our subsequent work investigating self-similarity properties of signals generated by these systems with white noise inputs has produced results related to their suitability for synthesizing data with desired self-similarity parameters [4]. A motivation for studying self-similar signals has been provided by the seminal work of Leland *et al* [1] showing that Ethernet traffic is self-similar. Self-similarity has since been found in other types of network traffic including wireless networks [8], [11]. Self-similar traffic gives rise to buffering requirements that are different and usually higher from those predicted by Poisson assumptions [7]. Much of the theoretical foundation related to the characterization of statistical self-similarity was laid by Mandelbrot and Van Ness [6] in the context of describing fractional Brownian motion (fBm) and fractional noise. For simulating data such as, for example, network traffic we clearly require synthesis of *discrete-time* self-similar random processes. Several methods have been proposed for generating discrete-time self-similar signals [1], [2], [3], [8]. Our prior work has demonstrated that synthesis of self-similar signals using white noise inputs to our discrete-time LSI models produces data whose properties are consistent with that of network traffic.

The paper is organized as follows. Section 2 provides an overview of our earlier formulation of DLSI systems. The new kernel-based discrete-time continuous-dilation operator of the paper is introduced in Section 3. Section 4 describes the DLSI systems based on the new dilation operator. Concluding remarks are made in Section 5.

2. OVERVIEW OF ZHAO AND RAO'S DLSI SYSTEMS

2.1 Time-Scaling

The definition of self-similarity rests on the operation of time scaling or dilation. Whereas it is possible to dilate a continuous-time signal in a continuous fashion, the same cannot be done with discrete-time signals. To avoid this difficulty, Zhao and Rao [10], [17]-[20] define a scaling operator for discrete-time signals that can work with any real-valued scaling factor greater then zero based on a *warping transform* $f(\omega)$ which transforms a discrete-time frequency (ω) to continuous-time frequency (Ω). The inverse transform $f^{-1}(\,)$ defines the continuous-time frequency to discrete-time frequency or unwarping transform. One examples of the warping transform is bilinear transform (BLT)

$$\Omega = f(\omega) \equiv 2\tan(\omega/2). \tag{1}$$

Using the warping transform defined above and time-frequency scaling property of the continuous time Fourier transform, the scaling operator $S_a[\,]$ of discrete-time sequence $x(n)$ is defined by

$$y(n) = S_a[x(n)] = aG^{-1}\{X[\Lambda_a(\omega)]\} \tag{2}$$

where $y(n)$ is the output of the operator, G^{-1} is the discrete-time Fourier transform (DTFT), $\Lambda_a(\omega) = f^{-1}[af(\omega)]$. The scaling operator is shown in Figure 1.

For a stochastic input sequence, if the input $X(n)$ of the discrete-time scaling operator $S_a[\,]$ is a discrete-time wide-sense stationary random process with power spectral density $P_X(\omega)$, it was shown the output is also wide-sense stationary with power spectral density given by

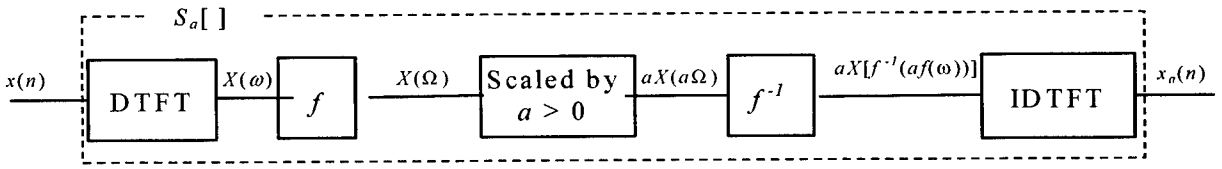


Figure 1. Block diagram of the discrete-time scaling function

$$P_Y(\omega) = \frac{a^2 P_X[\Lambda_a(\omega)]}{|\Lambda'_a(\omega)|} \quad (3)$$

where $\Lambda'_a(\omega)$ is the first derivative of $\Lambda_a(\omega)$ with respect to ω.

2.2 Discrete-Time Self-Similarity

Using the discrete-time continuous-dilation scaling operator $S_a[\ ]$ in (2), discrete-time stochastic self-similar signals can be defined as follows: a discrete-time random signal $X(n)$ is said to be self-similar with degree H in the wide-sense if it satisfies the following equations

$$E[S_a[X(n)]] = a^{-H} E[X(n)] \quad (4)$$

and

$$S_{a,a}[R_{XX}(n,n')] = a^{-2H} R_{XX}(n,n') \quad (5)$$

for any $a > 0$, where $R_{XX}(n, n')$ is the autocorrelation function of the sequence $X(n)$. For a discrete-time wide-sense stationary random process, the condition of self-similarity simply reduces to

$$\frac{P_X[\Lambda_a(\omega)]}{[\Lambda'_a(\omega)]} = a^{-2H-2} P_X(\omega) \quad (6)$$

where $P_X(\omega)$ is the power spectral density of the signal. Therefore, a stationary random process $X(n)$ whose power spectral density satisfies (6) is a self-similar signal in the statistical sense. Zhao and Rao suggested the next power spectrum for the density.

$$P_X(\omega) = \frac{|f(\omega)|^r}{|f'(\omega)|} \quad (7)$$

where $f'(\omega)$ is the first derivative of f with respect to ω.

From (6), (7) and $\Lambda_a(\omega) = f^{-1}[af(\omega)]$,

$$\frac{P_X[\Lambda_a(\omega)]}{[\Lambda'_a(\omega)]} = a^{r-1} P_X(\omega). \quad (8)$$

Thus, $X(n)$ is a self-similar random process with $H = -(r + 1)/2$.

If the power spectral density $Px(\omega)$ satisfies the *Paley-Wiener condition*, the density can be factorized as a product $L(\omega) L^*(\omega)$ and by passing white noise through a linear system with frequency response $L(\omega)$, the corresponding stochastic self-similar process can be generated.

The power spectral density for the BLT is

$$P_X(\omega) = \frac{|f(\omega)|^r}{|f'(\omega)|} = 2^r \left[\frac{1-\cos^2(\omega/2)}{\cos^2(\omega/2)}\right]^{r/2} \cos^2(\omega/2) \quad (9)$$

and this was known to satisfy the *Paley-Wiener condition.*

Let $z = e^{j\omega}$, then $P_X(\omega)$ transforms to

$$P_X(z) = L(z)L(z^{-1}) \quad (10)$$

where the causal part $L(z)$ is

$$L(z) = 2^{r/2-1}(1-z^{-1})^{r/2}(1+z^{-1})^{1-r/2} \quad (11)$$

Note that the spectrum is rational only for integer value of r.

The corresponding impulse response of is a causal filter whose coefficients are given by

$$l_1(n) = \begin{cases} 1 & n = 0 \\ (-1)^n (r/2) \sum_{k=0}^{n} \dfrac{(r/2-k+1)_{n-1}}{k!(n-k)!} & n > 0 \end{cases} \quad (12)$$

where $(\ )_n$ is the *Pochammer's symbol* defined as

$$(u)_0 \equiv 1 \quad (13)$$

and

$$(u)_v \equiv u(u+1)(u+2)\cdots(u+v-1) = \frac{\Gamma(u+v)}{\Gamma(u)} \quad (14)$$

The impulse response corresponding to $L_2(z)$ is a 2-tap filter with coefficients given by

$$l_2(0) = l_2(1) = 2^{r/2-1} \quad (15)$$

The overall impulse response $l(n)$ corresponding to the system transfer function given in (11) can be represented by two cascaded filters $l_1(n)$ and $l_2(n)$.

2.3 LSI System

A linear scale-invariant (LSI) system is a linear operator $L\{\}$ whose output is invariant to scale changes of the input signals, that is,

$$y(n) = L\{x(n)\} \Rightarrow S_a[y(n)] = L\{S_a[x(n)]\} \quad (16)$$

where $x(n)$ and $y(n)$ are the input and output sequence respectively.

A discrete-time causal LSI system for a given $x(n)$ can be defined similar to the continuous-time case [14]. Let $h(k)$ be any one-dimensional discrete-time sequence. The discrete-time causal LSI system is defined by the following relationship :

$$y(n) = \sum_{k=1}^{\infty} h(k) S_k[x(n)]/k \quad (17)$$

The output of the system is the sum of a series of dilation of the input sequence by k that are linearly weighted by $h(k)/k$.

If the input of the LSI system is a discrete-time stochastic self-similar signal with degree H, then the output is also a stochastic, self-similar signal with degree H [17]-[19]. In addition, if the input to a discrete-time LSI system is a discrete-time wide-sense stationary random process, the output of the system is non-stationary due to the fact that the system is time-varying. Using this property, a non-stationary self-similar random signal with parameter $H = -(r + 1) / 2$ can be generated by first generating a discrete-time self-similar random process with degree H by passing zero-mean white noise through a linear system with a frequency response given by (11), and then passing the signal thus obtained through a discrete-time LSI system. Note that the choice of the one dimensional function $h(k)$ in the discrete-time LSI system is arbitrary. This provides flexibility in signal construction. $h(k)$ can be chosen so that the output of the system has certain properties as desired.

3. KERNEL REPRESENTATION FOR DISCRETE-TIME CONTINUOUS-DILATION SCALING

Let $p(n,t)$ and $s(n,t)$ be linear operator kernels effecting transformation of signals between the discrete-time and continuous-time domains as

$$\begin{aligned} x(t) &= \sum_{n=-\infty}^{\infty} x(n)\, p(n,t) \\ x(n) &= \int_{-\infty}^{\infty} x(t)\, s(n,t)\, dt \end{aligned} \quad (18)$$

The warping defined in the previous section is a special case with

$$p(n,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp[j(\Omega t - f^{-1}(\Omega) n)] d\Omega \quad (19)$$

and

$$s(n,t)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\exp\left[j\left(\omega n-f(\omega)t\right)\right]d\omega \tag{20}$$

It follows that

$$\int_{-\infty}^{\infty}p(k,t)s(n,t)dt=\delta(n-k) \tag{21}$$

and

$$\sum_{n=-\infty}^{\infty}s(n,\tau)p(n,t)d\tau=\delta(t-\tau) \tag{22}$$

where δ stands for the discrete and continuous impulse functions depending on its context. The discrete-time scaling operation is now defined using the relations in Equation (18). Let $x_a(n)$ denote the time-scaling of a discrete-time signal $x(n)$ by a factor a. Let $x(t)$ be the continuous-time equivalent of $x(n)$ obtained through the transformation in Equation (18). We define

$$\begin{aligned}x_a(n)&=\int_{-\infty}^{\infty}x(t/a)s(n,t)dt\\&=\int_{-\infty}^{\infty}\left[\sum_{k=-\infty}^{\infty}x(k)p(k,t/a)\right]s(n,t)dt\\&=\sum_{k=-\infty}^{\infty}x(k)\int_{-\infty}^{\infty}p(k,t/a)s(n,t)dt\end{aligned} \tag{23}$$

Thus, with

$$g_a(k,n)=\int_{-\infty}^{\infty}p(k,t/a)s(n,t)dt \tag{24}$$

the scaling operator defined by the kernel is

$$x_a(n)=D_a\{x(n)\}=\sum_{k=-\infty}^{\infty}x(k)g_a(k,n) \tag{25}$$

When $a = 1$, the relationship becomes

$$x(n)=\sum_{k=-\infty}^{\infty}x(k)\int_{-\infty}^{\infty}p(k,t)s(n,t)dt=\sum_{k=-\infty}^{\infty}x(k)\delta(k-n) \tag{26}$$

One of the key properties of the scaling operator D_a is invertibility, that is D_a followed by $D_{1/a}$ yields the original discrete-time signal.

4. KERNEL BASED DLSI SYSTEMS

We now define scale-invariance for a discrete-time system as before (see Equation (16)) except that it must hold for the operator D_a. We will show that it is possible to effect DLSI systems simply by effecting discrete-time to continuous-time transformation using the $p(n,t)$ kernel, implementing a time-varying convolution corresponding to a continuous-time scale invariant system and then transforming the result to discrete-time using the $s(n,t)$ kernel.

Thus, given an input sequence $x(n)$ and a DLSI system characterized by a sequence $h(n)$, we first form

$$\begin{aligned}x(t)&=\sum_n x(n)p(n,t)\\h(t)&=\sum_n h(n)p(n,t)\end{aligned} \tag{27}$$

We then form

$$\begin{aligned}y(t)&=\int_\tau h(\tau)x(t/\tau)\frac{d\tau}{\tau}\\&=\int_\tau\sum_l h(l)p(l,\tau)\sum_m x(m)p(m,t/\tau)\frac{d\tau}{\tau}\end{aligned} \tag{28}$$

The output $y(n)$ of the DLSI system is obtained as

$$\begin{aligned}y(n)&=\int_\alpha y(\alpha)s(n,\alpha)d\alpha\\&=\int_\alpha\left[\int_\tau\sum_l h(l)p(l,\tau)\sum_m x(m)p(m,t/\tau)\frac{d\tau}{\tau}\right]s(n,\alpha)d\alpha\\&=\sum_l\sum_m h(l)x(m)\left[\int_\alpha\int_\tau p(l,\tau)p(m,t/\tau)s(n,\alpha)\frac{d\tau}{\tau}d\alpha\right]\\&=\sum_l\sum_m h(l)x(m)k_{l,m}(n)\end{aligned} \tag{29}$$

where

$$k_{l,m}(n)=\int_\alpha\int_\tau p(l,\tau)\,p(m,t/\tau)\,s(n,\alpha)\frac{d\tau}{\tau}d\alpha \tag{30}$$

Contrast this result with the expression in Equation (17). Unlike, the previous expression Equation (29) preserves symmetry between $x(n)$ and $h(n)$ much like the continuous-time LTI systems of Wornell.

There is another attractive property. Let

$$t_\alpha(n)=\int_t t^\alpha s(n,t)dt\,. \tag{31}$$

With $t_\alpha(n)$ as input to the DLSI system, we find the output is

$$y(n)=H(\alpha)t_\alpha(n) \tag{32}$$

where

$$H(\alpha)=\sum_l h(l)P_l(\alpha) \tag{33}$$

with

$$P_l(\alpha)=\int_\tau p(l,\tau)\tau^{-(\alpha+1)}d\tau \tag{34}$$

Thus $t_\alpha(n)$ is an eigenfunction of the DLSI system.

Suppose

$$X(\alpha)=\sum_l x(l)P_l(\alpha) \tag{35}$$

and

$$Y(\alpha)=\sum_l y(l)P_l(\alpha) \tag{36}$$

Then

$$Y(\alpha)=H(\alpha)X(\alpha) \tag{37}$$

We thus have the beginnings of a scale-domain transform operator similar to the Fourier transform for linear time-invariant systems.

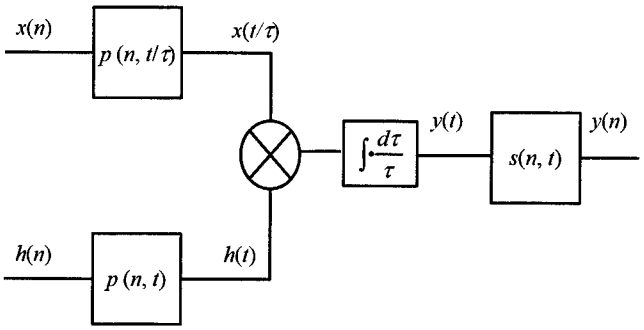


Figure 2. Block diagram of the kernel based the DLSI system

5. CONCLUSION

The discrete-time LSI systems proposed previously by Zhao and Rao provide a potential tool for the analysis and simulation of natural self-similar signals because of their scale invariant property (even though they are time-varying in general) in continuous scale. This earlier approach was based on a tacit warping in the frequency domain. The paper has provided an alternative approach based on linear kernels for transforming discrete-time signals to continuous time and vice versa. The resulting formulation of DLSI systems has several attractive properties. We have shown that such systems may be amenable to analysis using scale-domain transforms that are analogous to Fourier transforms. We believe the discrete-time LSI system formulation occupies a place in the study of scale-invariance and self-similarity that corresponds to the position of linear discrete-time time-invariant systems in the study of stationary random processes.

ACKNOWLEDGMENT

This work was supported in part by a grant from the Gleason Foundation and by a grant from NYSTAR.

REFERENCES

1. J. Feder. *Fractals*. Plenum Press, New York, 1988.
2. J. R. M. Hosking. Fractional differencing. *Biometrika*, 68(1):165-176, 1981.
3. M. S. Keshner. 1/f noise. *Proc. IEEE*, 70(3):212-218, Mar, 1982.
4. S. Lee, R.M. Rao and R. Narasimha. Characterization of self-similarity properties of discrete-time linear scale-invariant systems. *IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May 2001.
5. W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Trans. on Networking*, 2(1):1-15, Feb. 1994.
6. B. B. Madelbrot and J. W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4):422-437, Oct. 1968.
7. A. L. Neidhardt, F. Huebner and A. Erramilli. Shaping and policing of fractal traffic. *IEICE Transactions on Communications*, E81-B(4):858-869, May 1998.
8. V. Paxson. Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic, *Computer Communication Review*, 27(5):5-18, Oct. 1997.
9. A. Prasad, B. Stavrov and F. Schoute. Generation and testing of self-similar traffic in ATM networks. *1996 IEEE International Conference on Personal Wireless Communications, Proceedings and Exhibitions: Future Access*, 200-205, 1996.
10. R. M. Rao and W. Zhao. Image modeling with linear scale-invariant systems. *Proceedings of the SPIE-The International Society for Optical Engineering*, 3723:407-418, 1999.
11. W. Sheng, J. Rueda, W. Kinsner and D. C. Blight. Variance fractal dimension based wireless ATM LAN traffic estimation for network management. *Proceeding of the Applied Telecommunications Symposium*, 159-164, Apr. 1998.
12. M. S. Taqqu, V. Teverovsky and W. Willinger. Estimators for long-range dependence: an empirical study. *Fractals*, 3(4):785-798, 1995.
13. M. S. Taqqu, V. Teverovsky and W. Willinger. Is network traffic self-similar or multifractal?. *Fractals*, 5(1):63-73, 1997.
14. G. W. Wornell. *Signal Processing with Fractals*. Prentice-Hall, Upper Saddle River, New Jersey, 1996.
15. G. W. Wornell and A. V. Oppenheim. Estimation of fractal signals from noisy measurements using wavelets. *IEEE Trans. on Signal Processing, with Fractals*, 40:611-623, Mar. 1993.
16. G. W. Wornell and A. V. Oppenheim. Wavelet-based representations for a class of self-similar signals with application to fractal modulation. *IEEE Trans. on Information Theory*, 38(2):785-800, Mar. 1992.
17. W. Zhao. *Discrete-Time Continuous-Dilation Construction of Linear Scale-Invariant Systems and Multi-Dimensional Self-Similar Signals*. Ph.D. dissertation, Rochester Institute of Technology, Rochester, NY.
18. W. Zhao and R. M. Rao. Continuous-dilation discrete-time self-similar signals and linear scale-invariant systems. *Proceedings of the 1998 IEEE International Conference on Acoustics Speech and Signal Processing*, 3:1549-1552, 1998.
19. W. Zhao and R. M. Rao, Discrete-Time, Continuous-Dilation Construction of Self-Similar Signals and Linear Scale-Invariant Systems, *Tech. Report TR-Zharao-99-1*, Electrical Engineering Dept., Rochester Institute of Technology, NY.
20. W. Zhao and R. M. Rao. On modeling of self-similar random processes in discrete time. *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, 333-336, 1998

STOCHASTIC DISCRETE SCALE INVARIANCE AND LAMPERTI TRANSFORMATION

Pierre Borgnat, Patrick Flandrin

Laboratoire de Physique (UMR CNRS 5672)
École normale supérieure de Lyon
46, allée d'Italie 69364 Lyon Cedex 07, France

Pierre-Olivier Amblard

Laboratoire des Images et des Signaux
LIS-UMR CNRS 5083, ENSIEG-BP 46
38402 Saint Martin d'Hères Cedex, France

ABSTRACT

We define and study stochastic discrete scale invariance (DSI), a property which requires invariance by dilation for certain preferred scaling factors only. We prove that the Lamperti transformation, known to map self-similar processes to stationary processes, is an important tool to study these processes and gives a more general connection: in particular between DSI and cyclostationarity. Some general properties of DSI processes are given. Examples of random sequences with DSI are then constructed and illustrated. We address finally the problem of analysis of DSI processes, first using the inverse Lamperti transformation to analyse DSI processes by means of cyclostationary methods. Second we propose to re-write these tools directly in a Mellin formalism.

1. DISCRETE SCALE INVARIANCE

Scale invariance, also called self-similarity, is frequently called upon. Its central point is that the signal is scale invariant if it is equivalent to any of its rescaled versions, up to some amplitude renormalization [1]. More precisely, a function $X(t)$ is scale-invariant with exponent H, or H-ss, if for any $k \in \mathbb{R}$: $X(kt) = k^H X(t)$.

This definition is given here for a deterministic signal. The concept can be extended to stochastic signals when one thinks of the previous equality in a probabilistic way: the equality of the finite-dimensional probability distributions [1]. We will write $\stackrel{d}{=}$ this equality.

The strict notion of scale invariance, valid for all dilation factors above, is in some cases too rigid; the middle-third Cantor set is for example invariant only by dilations of a factor 3 (or a power of 3). Several weakened versions of self-similarity have been proposed to enlarge scale invariance's relevance and one is of special interest here: it is to require invariance by dilation for certain preferred scaling factors only, as it is the case for the Cantor set. This is known as *discrete scale invariance* (DSI), a concept which as been stressed upon by Sornette and Saleur [2, 3] as an efficient model in many situations (fracture, DLA, critical phenomena, earthquakes).

They studied DSI as a property of deterministic signals, and provided general arguments as why should DSI naturally occur: classical scenarii involve the existence of a characteristic scale, the apparition by instability of a preferred scale or more general arguments in non-unitary field theories [4]. They also found ways to estimate the preferred scaling ratio in this context, based on classical spectral analysis (Lomb periodogram).

As far as we know, this property has not been envisioned for stochastic processes, a framework which is often fruitful to dispose of when dealing with real measurements, as it allows to use statistical signal processing methods. The extension of DSI property to stochastic processes is straigthforward. We propose the following definition.

A process $\{X(t),\ t \in \mathbb{R}^+\}$ has discrete scale invariance with scaling exponent H and scale λ if

$$X(\lambda t) \stackrel{d}{=} \lambda^H X(t),\ t \in \mathbb{R}^+. \qquad (1)$$

We will refer to this property as (H, λ)-DSI. The equality here is the probabilistic equality. In the following only wide-sense property will be used (second-order statistical properties only).

2. LAMPERTI TRANSFORM : DSI AS AN IMAGE OF CYCLOSTATIONARITY

2.1. Lamperti transformation

A main issue is to find a way to study both theoretically and practically DSI processes. The answer is given by a transformation introduced by J. Lamperti in 1962 [5], which is an isometry between self-similar and stationary processes. It will be called the Lamperti transformation and is defined as follows.

For any process $\{Y(t),\ t \in \mathbb{R}\}$, its Lamperti transform $\{X(t),\ t \in \mathbb{R}^+\}$ and its inverse are given by

$$X(t) = (\mathcal{L}Y)(t) \ \hat{=}\ t^H Y(\ln t),\ t \in \mathbb{R}^+; \qquad (2)$$

$$Y(t) = \left(\mathcal{L}^{-1}X\right)(t) \ \hat{=}\ e^{-Ht} X(e^t),\ t \in \mathbb{R}. \qquad (3)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

The theorem in the paper of Lamperti is that a process $Y(t)$ is stationary if and only if its Lamperti transform $X = \mathcal{L}Y$ is H-ss. The central argument of the derivation is that the Lamperti transformation maps a time-shifted process to the dilated version of the Lamperti transform of the original process. Let $\left(\mathcal{D}_\lambda^H X\right)(t) \hat{=} \lambda^{-H} X(\lambda t)$ be the dilation operator and $\left(\mathcal{S}_\tau Y\right)(t) \hat{=} Y(t+\tau)$ the time-shift operator. The property is that

$$\left(\mathcal{L}^{-1}\, \mathcal{D}_\lambda^H\, \mathcal{L}\, Y\right)(t) \stackrel{d}{=} \left(\mathcal{S}_{\ln \lambda} Y\right)(t). \qquad (4)$$

Understanding this correspondence between time-shift and dilation operators, we can propose many variations around Lamperti's theorem, relaxing in some way the stationarity for Y and the self-similarity for X. We will only consider here the DSI property but some results about different classes of processes and their description are proposed in [6]. A useful property is that one can give the (potentially nonstationary) correlation function of the Lamperti transform X of a process Y:

$$\mathbb{E}\{X(t)X(s)\} \hat{=} R_X(t,s) \;=\; (st)^H R_Y(\ln t, \ln s). \qquad (5)$$

In the recent years some results have been obtained for H-ss processes with this transformation. Yazici and Kashyap proposed a general description of wide-sense self-similar processes and linear models for H-ss [7]. Burnecki *et al.* study α-stable and H-ss processes with this transform [8]. Nuzman and Poor give important results about the prediction, the whitening and the interpolation of H-ss processes, mainly applied to the fractional Brownian motion [9]. Finally Vidács and Virtamo [10] proposed a method of estimation of H for a fBm, based on the same idea. All these authors use the inverse Lamperti transformation (3) to map the question to a stationary problem and then use the known results for stationary issues in this context. Our objective is to show that nonstationary methods can be adapted in the same way, especially for DSI.

2.2. DSI and cyclostationarity

A process is called cyclostationary [11] or periodically-correlated [12, 13], if its correlation function is periodic in time. More precisely, if a period T is given, a process $\{Y(t),\ t \in \mathbb{R}\}$ is wide-sense cyclostationary if it satisfies for any times t, s

$$\begin{aligned} \mathbb{E}\, Y(t+T) &= \mathbb{E}\, Y(t), \\ \mathbb{E}\{Y(t+T)Y(s+T)\} &= \mathbb{E}\{Y(t)Y(s)\}, \end{aligned} \qquad (6)$$

The correlation function $R_Y(t, t+\tau)$ is then periodic in t of period T and one can decompose R_Y in a Fourier series

$$R_Y(t, t+\tau) \;=\; \sum_{n=-\infty}^{+\infty} C_n(\tau) e^{i2\pi n t/T}. \qquad (7)$$

Using the definitions of cyclostationarity and (H,λ)-DSI and the correspondance (4), we can state the following important result.

A process $\{Y(t),\ t \in \mathbb{R}\}$ is cyclostationary of period T if and only if its Lamperti transform of parameter H: $\{X(t) = t^H Y(\ln t),\ t \in \mathbb{R}^+\}$, is (H, e^T)-DSI.

This is one possible extension of Lamperti's theorem, one of importance in our study of DSI. A first consequence, using (5), is that the general form of covariance of (H,λ)-DSI processes is naturally expressed on a Mellin basis:

$$R_X(t,kt) \;=\; k^H t^{2H} \sum_{n=-\infty}^{+\infty} C_n(k) t^{i2\pi n/\ln\lambda}. \qquad (8)$$

Note that if the process X is real-valued, a necessary condition is imposed: $C_{-n}(k) = C_n^*(k)$. The Mellin function $t^{H+i2\pi n/\ln\lambda}$ in (8) is central in the study of DSI processes. This is not a surprise: Lamperti transformation maps the Fourier basis (invariant up to a phase by time-shift) to the Mellin basis (invariant up to a phase by dilation and having also the deterministic DSI property). We stress the fact the Mellin functions are a basis and that they have an associated transformation which can be numerically computed [14].

3. EXAMPLES OF PROCESSES AND SEQUENCES WITH DSI

Continuous-time systems with DSI property are easily constructed. Applying $\mathcal{L}$ to an ARMA(p,q) system, we obtain a generalization of the Euler-Cauchy (EC) system. It is a model for self-similar processes [7], driven by a multiplicative Gaussian noise $\eta(t)$, whose correlation is $\mathbb{E}\{\eta(t)\eta(s)\} = t\sigma^2\delta(t-s)$. The process $X(t)$ verifies

$$\sum_{n=0}^{p} b_n t^n \frac{\mathrm{d}^n}{\mathrm{d}t^n} X(t) = \sum_{m=0}^{q} a_m t^{m+H} \frac{\mathrm{d}^m}{\mathrm{d}t^m} \eta(t). \qquad (9)$$

In the same manner that a nonstationary ARMA model with periodic time-varying coefficients is cyclostationary [15], one obtains a DSI model when taking log-periodic time-varying coefficients a_m and b_n in the (EC) system. This will be not detailed further.

In order to obtain DSI processes in discrete time (random *sequences* with self-similarity and log-periodicity), a possibility is to consider a discrete-time system analog to (EC) (H-ss in a certain way), then introduce log-periodicity in the coefficients. We describe two approaches here.

A direct discretization in time of the (EC) system is given by the integration of its evolution between two instants. This was proposed in [16] for the first order. This nonstationary H-ss system is written as $X_k = a[k]X_{k-1} + e_k$, where $a[k] \simeq 1 - \alpha/k$ and e_k is a time-decorrelated

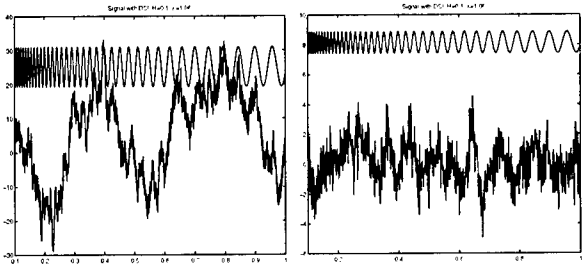


Fig. 1. Typical realizations of DSI random sequences. On the left the model is a (EC) system of order 2 discretized, in cascade with a log-periodic AR(1). On the right it is constructed on fractional difference (see text). The length is 5000 points, $H = 0.1$, $\lambda = 1.0$. The oscillations above the signals are indicative of the log-periodicity of the AR. $f_0 = 1/8$, $\rho = 0.6$, $\beta_r = 0.25$ and $\beta_f = 0$.

Gaussian noise with variance $\mathbb{E}\, e_k^2 \propto k^{2H-1}$, when k is large. The generalization to the discretization of (EC) of order n is straightforward. The result is of the form, for the large times k

$$(1-B)^n X_k + a_1 k^{-1}(1-B)^{n-1} X_{k-1} = k^{-2}\mathrm{AR}(n-1)X_{k-1} + e_k + \mathcal{O}(k^{-3}) \quad (10)$$

where B is the backward operator, and AR is an AR model.

Such a system with log-periodicity in the coefficient a_1 and in the AR, or in cascade with a log-periodic AR system (see for example the AR(1) proposed hereafter, equation 12), will present an approximate DSI property. The reader can see on the left of figure 1 a realization of such a process.

Another class of discrete-time self-similar systems is given by models constructed on the fractional difference operator. The usual method is to use its moving average representation written as a binomial expansion. We prefer to use the discretization proposed in [17], constructed with some generalization of the bilinear transformation in order to define a scaling operator for sequences. The fractional difference operator is then a filter $l_1[n]$ whose impulse reponse is

$$l_1[n] = \sum_{k=0}^{\infty} \frac{(-1)^k \Gamma(r+k)\Gamma(-r+n-k)}{\Gamma(k+1)\Gamma(n-k+1)\Gamma(r)\Gamma(-r)}. \quad (11)$$

This filter is in cascade with a nonstationary AR filter whose coefficients are log-perodic. For example we may limit ourselves to the first order (coefficient l_2), taking care that the filter is stable at each instant:

$$l_2 = \left(\rho + \beta_r \cos\frac{2\pi \ln t}{\ln \lambda}\right) e^{i2\pi f_0 (1+\beta_f \cos(2\pi \frac{\ln t}{\ln \lambda}))}. \quad (12)$$

We propose an example of such a signal fig. 1 on the right.

4. ANALYSIS BY DELAMPERTIZATION

In front of a general class of processes (or random sequences in the context of numerical processing) which are nonstationary, or of unknown structure, one has to find methods to analyse those. Given a sequence X_n suspected of DSI, the simplest way of analysis is to find the presumed cyclostationary process associated by applying $\mathcal{L}^{-1}$.

Generally speaking, classical stationary methods are useful to analyse self-similar process after "delampertization" of the signal. This was the essence of papers on H-ss processes cited before [7, 8, 9]. Nonstationary methods can then be used tu study classes of processes which have not proper self-similarity, but which have some kind of nonstationarity with regards to dilation - a nonstationarity in scale. DSI is then only a first interesting example of a precise kind of nonstationarity in scale.

Before using cyclostationary methods, a practical problem must be considered : how to compute in discrete time the inverse Lamperti transformation ? First, it needs a non linear sampling $t = q^n$ of the data (but such is not often the case with real signals), or an interpolation to find the data with this geometrical sampling, given a signal X_t with usual arithmetic sampling: the corresponding sequence Y_t is known for $t = \ln n$, with $n \in \mathbb{N}$ and we want it for $t = m$, $m \in \mathbb{Z}$. Figure 2 shows on the left the sequence Y constructed from the second process on figure 1.

A second difficulty is that H is a priori unknown. Using the transformation of parameter H seems tricky... In fact the tools used thereafter have not been found to be sensitive to this amplitude effect. The cyclostationary tools are found unaffected if one uses $H = 0.5$ to delampertize the process in place of the real H.

We tried the applicability of these ideas on synthetic sequences. As an example of a classical cyclostationary tool, we implemented the methods proposed in [18]. In a nutshell the algorithm to estimate a time-smoothed cyclic cross periodogram is as follows. First the signal is decomposed in N segments of length L in order to average on these parts. A filtered and decomposed version is computed, where h is a data tapering window:

$$\tilde{Y}_T(n,f) = \sum_{l=-N/2}^{N/2} h(l) Y(n-l) e^{-i2\pi f(n-l)T_e} \quad (13)$$

Then the spectral components $\tilde{Y}_T(n,\cdot)$ are correlated at frequencies $f - \nu_c/2$ and $f + \nu_c/2$ by a multiplier followed by a low-pass filter g:

$$S_Y^{\nu_c}(v,f) = \sum_n \tilde{Y}_T(n, f - \frac{\nu_c}{2}) \tilde{Y}_T^*(n, f + \frac{\nu_c}{2}) g(v-n).$$

This is an estimate of the spectral cross correlation. The usual spectrum is distributed on the main diagonal $\nu_c = 0$

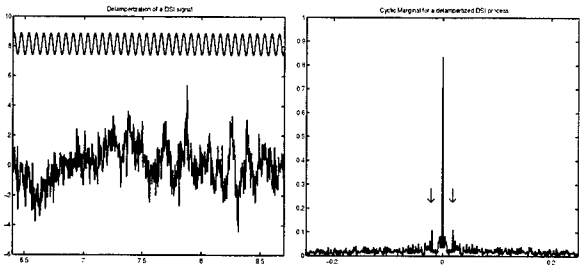


Fig. 2. On the left is shown the cyclostationary sequence after using $\mathcal{L}^{-1}$ on the signal plotted on the right of fig. 1. The marginal in cyclic frequency is represented on the right. The main peak on the center is the total energy of the signal. The two symmetric peaks (pointed on by arrows) are an indication of cyclostationarity and situated to frequencies $\pm 2\pi / \ln \lambda$.

and for cyclostationary sequences it presents non-zero distributions on $\nu_c = \pm 1/T$ (and eventually on higher harmonics). The marginal in cyclic frequency of this spectrum has then sharp peaks on $1/T$ where $\lambda = e^T$ for DSI and gives a reliable estimation of λ. See figure 2 the result of this procedure for the synthetic model described before.

5. TOWARD MELLIN-BASED TOOLS

Another way of thinking might be fecund to analyse DSI processes. We can formulate directly the methods in a Mellin formalism, with no geometrical resampling. That is to say that we oper a "lampertization" of the tools where the first way proposed to "delampertize" the signal studied.

By direct interpolation we have few details for the short times (in fact we can't reconstitute $m < 0$) and we ignore many details in the long times (taking one point among many). To obtain statistical relevance, one has to have a huge number of points in the original data to make some processing. The avantage, remarked in [8, 10], is that there are fewer points in X, then Y, after geometrical resampling and this keeps the computational cost low.

When one does not dispose of a large number of points, using a geometric sampling loose much information on the signal. As the Fourier transform of a process is related to the Mellin transform of the process transformed by $\mathcal{L}$, many methods for cyclostationary processes can be written with Mellin transformation and used on processes with DSI. For self-similar signals ($H = 0$), estimators constructed in this way were given in [19] and can be adapted to take into account an exponent H and DSI.

6. REFERENCES

[1] G. Samorodnitsky and M. Taqqu, *Stable Non-Gaussian Random Processes*, Chapman&Hall, 1994.

[2] D. Sornette, "Discrete scale invariance and complex dimensions," *Physics Reports*, vol. 297, pp. 239–270, 1998.

[3] H. Saleur and D. Sornette, "Complex exponents and log-periodic corrections in frustrated systems," *J. Phys. I France*, vol. 6, pp. 327–355, Mar. 1996.

[4] D. Sornette, "Discrete scale invariance," in *Scale Invariance and Beyond*, B. Dubrulle, F. Graner, and D. Sornette, Eds. 1997, pp. 235–247, Springer.

[5] J. Lamperti, "Semi-stable stochastic processes," *J. Time Series Anal.*, vol. 9, no. 2, pp. 62–78, 1962.

[6] P. Borgnat, P. Flandrin, and P.-O. Amblard, "Stochastic discrete scale invariance," subm. to *Signal Processing Lett.*, Apr. 2001.

[7] B. Yazici and R. L. Kashyap, "A class of second-order stationary self-similar processes for $1/f$ phenomena," *IEEE Trans. on Signal Proc.*, vol. 45, no. 2, pp. 396–410, 1997.

[8] K. Burnecki, M. Maejima, and A. Weron, "The Lamperti transformation for self-similar processes," *Yokohama Math. J.*, vol. 44, pp. 25–42, 1997.

[9] C. Nuzman and V. Poor, "Linear estimation of self-similar processes via Lamperti's transformation," *J. of Applied Probability*, vol. 37, no. 2, pp. 429–452, June 2000.

[10] A. Vidács and J. Virtamo, "ML estimation of the parameters of fBm traffic with geometrical sampling," in *IFIP TC6, Int. Conf. on Broadband communications '99*. Nov. 1999, Hong-Kong.

[11] W. Gardner and L. Franks, "Characterization of cyclostationary random signal processes," *IEEE Trans. on Info. Theory*, vol. IT-21, no. 1, pp. 4–14, Jan. 1975.

[12] E. Gladyshev, "Periodically and almost periodically correlated random processes with continuous time parameter," *Theory Prob. and Appl.*, vol. 8, pp. 173–177, 1963.

[13] H. Hurd, *An investigation of periodically correlated stochastic processes*, Ph.D. thesis, Duke Univ. dept. of Electrical Engineering, Nov. 1969.

[14] J. Bertrand, P. Bertrand, and J.P. Ovarlez, "The Mellin transform," in *The Transforms and Applications Handbook*, A.D. Poularikas, Ed. CRC Press, 1996.

[15] S. Lambert-Lacroix, "On periodic auro-regressive processes estimation," *IEEE Trans. on Signal Proc.*, vol. 48, no. 6, pp. 1800–1803, 2000.

[16] E. Noret and M. Guglielmi, "Modélisation et synthèse d'une classe de signaux auto-similaires et à mémoire longue," in *Proc. Conf. Delft (NL) : Fractals in Engineering*. 2000, pp. 301–315, INRIA.

[17] W. Zhao and R. Rao, "On modeling self-similar random processes in discrete-time," in *Proc. IEEE Time-Frequency and Time-Scale*, Oct. 1998, pp. 333–336.

[18] R. Roberts, W. Brown, and H. Loomis, "Computationally efficient algorithms for cyclic spectral analysis," *IEEE SP Magazine*, pp. 38–49, Apr. 1991.

[19] H. L. Gray and N. F. Zhang, "On a class of nonstationary processes," *Journal of Time Series Analysis*, vol. 9, no. 2, pp. 133–154, 1988.

THE VITERBI-ALGORITHM FOR IMPULSIVE NOISE WITH UNKNOWN PARAMETERS

Thomas Kaiser and Youssef Dhibi

Fraunhofer Institute for Microelectronic Circuits and Systems
Department of Wireless Chips & Systems
Finkenstrasse 61, 47057 Duisburg
phone: +49 203 3783 170, fax: +49 203 3783 299
email: thomas.kaiser@ims.fhg.de

ABSTRACT

In this paper we will propose a modification of the well-known VITERBI-Algorithm (VA) for communication channels distorted by *impulsive* instead of the often used GAUSSIAN noise. Here we assume that the parameters - e.g. the moments - of the noise are unknown. Instead of applying a recursive solution (see [2]) by repeated execution of the VA we will here directly embed the estimation of the unknown parameters into the structure of the VA itself. Such an approach is called *Per-Survivor Processing* (PSP) [8] which provides a general framework for the approximation of Maximum Likelihood Sequence Estimation (MLSE) whenever the presence of unknown quantities prevents the precise use of the classical VA. In addition, the classical VA will be modified so that it works optimally for some kind of impulsive noise. We will show by means of the modified VA, that the bit-error rate can be substantially decreased. In other words, only with minor technical modifications by minimizing an adequate nonlinear norm, the transmission becomes more reliable compared to the usual euclidian norm minimized by the *conventional* VA.

Keywords: Viterbi-Algorithm, Impulsive Noise, Per-Survivor Procesing, MLSE, Non-Euclidean Norms

1. INTRODUCTION

In the last decade, an increased interest in modeling of *impulsive* noise can be observed in the statistical signal processing community. The reasons are mainly twofold: 1. new insights in some special distributions, especially the *stable* distribution, are found, and, 2. the non-GAUSSIAN noise found in many applications, e.g. communication channels [6], [10], biology [4], sonar [1]. Here we will focus on the problem to suppress additive *impulsive* noise in data transmission, where the communication channel is assumed to be linear and preliminary time-invariant with impulse response $h(k)$.

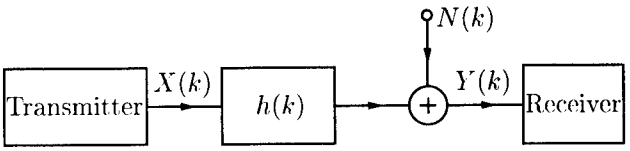


Figure 1: A simple discrete communication channel model

We have omitted sampling, coding and modulation for reasons of simplicity and deal only with sequences like $h(k)$, $k \in \mathbb{Z}$ instead of time-continuous signals like $h(t)$. The input sequence $X(k)$ (capital letters means random variables or random processes), which is assumed to be an i.i.d. random process, consists only of ± 1 for each k (a so-called BPSK-sequence). Before the data are sent, a short training sequence, being known to the receiver, is transmitted. Therefore, during the training period the receiver is able to estimate the impulse response. Note that in the PSP-approach the simultaneous estimation of the impulse response is also possible (see [8]). However, here we will focus on the simultaneous estimation only of the noise properties. Thus, in the following we always assume a known impulse response estimate $\hat{h}(k)$.

The most popular method for data reconstruction given $\hat{h}(k)$ is the well-known VITERBI-algorithm. If we assume a limited time interval $k = 0(1)K-1$ for transmission, exactly $M = 2^K$ different data sequences $x_m(k)$ (realizations of random process are written in small letters) could be sent. Given the received sequence $y(k)$, the VITERBI-algorithm reconstructs the transmitted data sequence $x_{m_0}(k)$, $m_0 \in (1, ..., M)$ by minimizing the euclidian norm

$$\sum_{k=0}^{K-1} \Big(y(k) - \hat{h}(k) * x_m(k) \Big)^2, \qquad (1)$$

regarding m in a recursive and therefore very efficient manner. We will denote this kind of VA as the *conventional* VITERBI-algorithm. The use of the euclidian norm can be mathematically substantiated by the *maximum-likelihood*-principle, if the channel noise is GAUSSIAN (see e.g. [7], pp. 249). However, despite of the intuitive nature of the euclidian norm, it is in no sense optimal in case of *impulsive* noise. Thus, deriving a modified VA also based on the maximum-likelihood-principle could be quite successful. Moreover, it is well-known from information theory, that the *channel capacity* $C|_{\text{GAUSS}}$ in case of GAUSSIAN noise is always lower than $C|_{\text{Non-GAUSS}}$ for non-GAUSSIAN noise

$$C|_{\text{GAUSS}} < C|_{\text{Non-GAUSS}}. \tag{2}$$

Since the channel capacity can be seen as an upper bound for the maximum data rate, this inequality means that on a non-GAUSSIAN channel a higher bit rate is principally possible. Of course, eq. (2) requires the same noise power in the GAUSSIAN as well as in the non-GAUSSIAN case. Because the noise power, which is simply a second order moment for zero-mean noise, does not necessarily exist for each random process – especially not for any kind of impulsive noise – we introduce here the *p-norm power* of $N(k)$ as

$$P_{N,p} = (\mathsf{E}\{|N(k)|^p\})^{\frac{2}{p}}, \tag{3}$$

where $\mathsf{E}\{...\}$ denotes expectation. Before we proceed with a modified VA, let us shortly present some popular models for impulsive noise.

2. MODELS FOR IMPULSIVE NOISE

Since the introduction of the so-called *stable* distribution by SHAO and NIKIAS [5], [9] in the signal processing community, this distribution has drawn a lot of attention. A stable *random variable* X is defined via the *characteristic function*

$$\phi_X(x) = \mathrm{e}^{\mathrm{j}\mu x - \gamma|x|^\alpha(1+\mathrm{j}\beta\,\omega(x,\alpha)\,\mathrm{sign}(x))} \tag{4}$$

where

$$\omega(x,\alpha) = \begin{cases} \tan\frac{\alpha\pi}{2} & \text{for } \alpha \neq 1 \\ \frac{2}{\pi}\log|x| & \text{for } \alpha = 1 \end{cases}$$

and

$$\mathrm{sign}(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x = 0 \\ -1 & \text{for } x < 0. \end{cases}$$

$\mu \in \mathbb{R}$ is called the *location parameter*, $\beta \in [-1,1]$ the *symmetry parameter*, $\alpha = (0,2]$ the *characteristic exponent* and $\gamma \in \mathbb{R}_+$ the scale parameter. The stable distribution enjoys many useful properties such as the *linear stability theorem* (see [5], p. 20, p. 24) and the *generalized central limit theorem* (see [5], p. 25). The only remarkable drawback of the stable distribution is that – even in the symmetric case with $\beta = 0$ – no closed form for the probability density function $p_X(x)$ (pdf) exists. Since it can be shown that a stable pdf has *algebraic* tails – in fact, the smaller the value of α, the thicker the tails – only the *fractional lower order moments*

$$\mathsf{E}\{|X|^p\} < \infty \quad \text{for} \quad p < \alpha,\ \alpha < 2 \tag{5}$$

exist (For $\alpha = 2$ any moment exist). This is the reason for introducing the p-norm power. Beside the stable distribution, the Generalized GAUSSIAN pdf ([3], p. 74), the Generalized CAUCHY pdf ([3], p. 78), the MIDDLETON's pdfs as well as the GAUSSian mixture pdf are popular models for impulsive noise. Other pdf's suitable to model impulsive noise, e.g. the LAPLACE- and *Student-t*-pdf, are special cases of the above density functions. In this paper we will concentrate on the CAUCHY-distribution; further investigations concerning the remaining distributions will be done in the near future. After the basics we can now proceed with the modification of the VITERBI-algorithm for additive CAUCHY-noise.

3. THE CAUCHY-VITERBI-ALGORITHM

A random variable X is CAUCHY-distributed if the pdf is given by

$$p_X(x) = \frac{\sigma}{\pi(\sigma^2 + x^2)},$$

which is a special case of the generalized CAUCHY-pdf. Note that the stable distribution is identical to the CAUCHY-distribution for $\gamma = \sigma$, $\alpha = 1$, $\beta = \mu = 0$ ([5], p.14). To emphasize this special case we will use γ instead of σ in the following. Observe that not only the variance but also the mean of a CAUCHY-random variable do not exist. To modify the VITERBI-algorithm consider the multivariate density function of $Y(k)$, $k = 0(1)K-1$

$$\begin{aligned} & f_{Y(0),..,Y(K-1)}(y(0),...,y(K-1)) \\ &= \prod_{k=0}^{K-1} f_{N(k)}(y(k) - \hat{h}(k) * x_m(k)) \\ &= \prod_{k=0}^{K-1} \frac{\gamma}{\pi\left(\gamma^2 + (y(k) - \hat{h}(k) * x_m(k))^2\right)}. \end{aligned} \tag{6}$$

Now, maximizing $f_{Y(0),..,Y(K-1)}(y(0),...,y(K-1))$ according to the *maximum-likelihood*-principle means to

minimize the following non-linear norm with respect to m

$$\min_{m} \sum_{k=0}^{K-1} \ln\left(\gamma + \frac{(y(k) - h(k) * x_m(k))^2}{\gamma}\right), \qquad (7)$$

which can be easily derived. Hence, the *optimal* receiver for additive CAUCHY-noise consists in a VA with a *non*-euclidian norm. We will denote this method as the CAUCHY-VITERBI-algorithm. Observe that this norm includes the parameter γ being not known a-priori. However, a data-aided estimation procedure of this unknown parameter based on tentative low-delay decisions at the VA-output can be applied. This approach is shown in Fig. 2. After the so-called low-

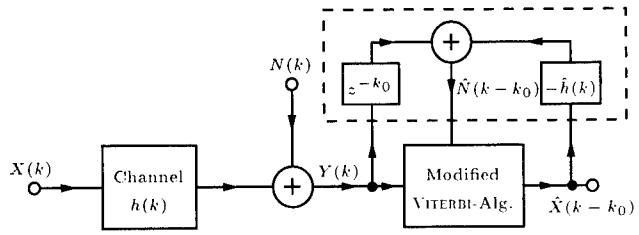


Figure 2: A data-aided estimation procedure based on tentative low-delay decisions at the VA-output

delay k_0, the VA has reconstructed the transmitted data $\hat{X}(k - k_0)$. These data will be filtered by the estimated impulse response $\hat{h}(k)$ and then substracted from the delayed received sequence $Y(k - k_0)$. Therefore, the additive noise $\hat{N}(k - k_0)$ can be reconstructed and the missing parameter γ can be estimated from the statistics of $\hat{N}(k - k_0)$. In fact, the following estimator for γ can be easily derived (along the lines in [5], p. 69)

$$\hat{\gamma}(K) = \prod_{k=0}^{K-1} |\hat{n}(k)|^{\frac{1}{K}} \qquad (8)$$

given a realization $\hat{n}(k)$, $k = 0(1)K - 1$ of the random process $\hat{N}(k - k_0)$. An alternative approach to the above blockwise data-aided estimation procedure consists in the application of the *per-survivor-principle* (PSP). This principle stems from the idea that data-aided estimation of unknown parameters can be embedded into the structure of the VA itself. This means that the estimation of γ is done in each trellis branch based on the above formula in a recursive manner

$$\hat{\gamma}(K) = (\hat{\gamma}(K-1))^{\frac{K-1}{K}} \; |\hat{n}(K)|^{\frac{1}{K}}. \qquad (9)$$

Alternatively, an exponential window can be used for recursive estimation of $\hat{\gamma}(K)$ to allow tracking of time-variant channels. With these formulas we are able to reconstruct the current noise value not in an external way as shown in Fig. 2, but directly in the VA. For a further introduction to PSP see [8]. At this point we are able to compare the conventional VA with the CAUCHY-VA as well as the PSP-CAUCHY-VA by numerical simulations.

4. NUMERICAL SIMULATION

In this simulation, we have carried out 500 MONTE-CARLO-run's with $K = 5$, $K = 10$, $K = 20$ transmitted BPSK-samples in each run. The impulse response has been assumed to be exactly known as $h(0) = 1/\sqrt{5}$, $h(1) = 2/\sqrt{5}$, $h(k) = 0 \forall k \neq [0, 1]$. The tentative delay has been chosen to $k_0 = K$. To measure some kind of signal-to-noise ratio, we define the *signal-to-noise p-norm ratio* as

$$\mathrm{SNR}_p = 10 \log_{10}\left(\frac{P_{X,p}}{P_{N,p}}\right).$$

Since p must be smaller than α (see eq. (5)) and α is equal to one for CAUCHY-noise, we have chosen p to $p = 0.9999$. Figure 3 (4,5) shows the results for $K = 5$ ($K = 10$, $K = 20$).

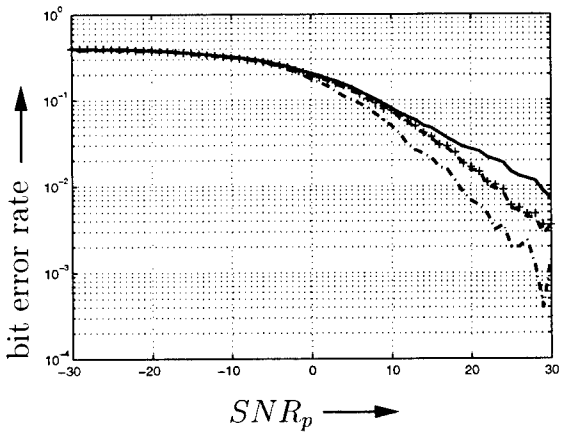


Figure 3: Bit error rate for $K = 5$ as a function of the signal-to-noise p-norm ratio for the conventional-VA (solid line), the CAUCHY-VA with true γ (dashed line), the CAUCHY-VA with estimated $\hat{\gamma}$ (plus-signs) and for the PSP-CAUCHY-VA with estimated $\hat{\gamma}$ (dashed-dotted line).

It can be seen that not only the CAUCHY-VA with true γ but also the CAUCHY-VA with estimated γ *clearly outperforms* the conventional VA. In particular, for an SNR_p around 10dB to 20dB the bit error rate is reduced more than 50 percent. Observe that the curve for true γ almost totally overlap the curve for estimated γ. We can conclude that the estimation error of γ, even if it is quite high, has not a large influence on the minimization procedure. In other words, the non-EUCLIDean

norm is very robust against parameter estimation errors. Observe also, that the PSP- CAUCHY-VA exhibits most often the lowest bit error rate.

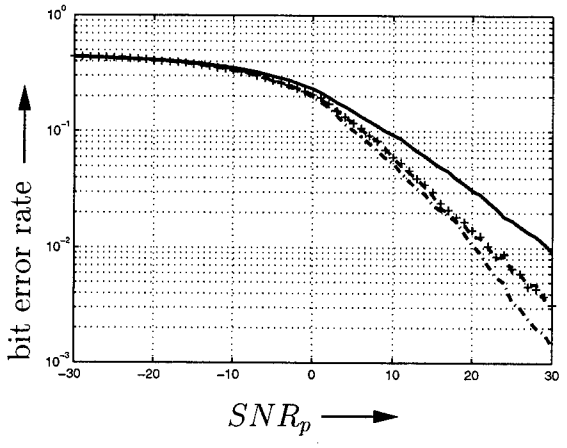


Figure 4: Bit error rate for $K = 10$.

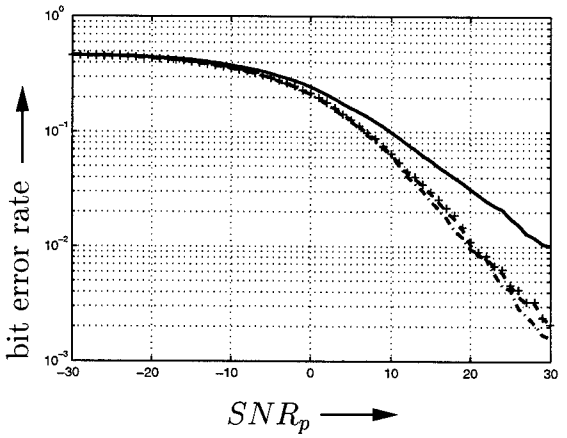


Figure 5: Bit error rate for $K = 20$.

Let us now consider the last three figures as a function of $K = 5, 10, 20$. It can be seen that for $K = 5$, so for very small data length, the PSP-approach has a much higher performance than the two other algorithms. If we now increase K, the bit error rate for the conventional VA and the PSP-VA remain almost unchanged, whereas the bit error rate of the CAUCHY-VA is decreasing. For $K = 20$ this bit error rate reaches the lower one of the PSP-approach. This means that by including the estimation of γ into the internal structure of the VA – so the PSP-approach – the estimation error does not play such a crucial role as for the blockwise approach (CAUCHY-VA). If K is increased, the estimation error of γ is reduced, so that the bit error rate of the CAUCHY-VA now starts to reach the one of the PSP-approach. Consequently, the PSP-approach seems to be very suitable for fastly time varying channels where only a small amount of data are available to estimate the noise statistics.

5. CONCLUSIONS

In this paper we have considered the estimation of only one parameter, namely γ to adapt the conventional VA to the noise statistics of the channel. To describe impulsive noise more precisely we will need at least two parameters – one is responsible for the height of the impulses, whereas the other for the probability of occurence of an impulse. Since future work will deal with the estimation of at least two parameters (e.g. two different noise variances in case of GAUSSian mixture processes), the PSP-approach is becoming more attractive even for a moderate data length. In general, all these results confirm the advantageous use of the channel noise statistics to enable a more reliable transmission.

REFERENCES

[1] M. Bouvet and S. C. Schwarz, *Comparison of Adaptive and Robust Receivers for Signal Detection in Ambient Underwater Noise*, IEEE Trans. on Acoustics, Speech and Signal Proc., Vol. 37, pp. 621-626, May 1989

[2] T. Kaiser, *How should the Viterbi-algorithm be modified for impulsive noise*, International Conference on Systemics, Cybernetics and Informatics, July 23-26, 2000, Orlando, Florida, USA

[3] S.A. Kassam, *Signal Detection in Non-Gaussian Noise*, Springer-Verlag, 1988

[4] X. Kong, T. Qiu, *Adaptive Estimation of Latency Change in Evoked Potentials by Direct Least Mean p-Norm Time Delay Estimation*, IEEE Transactions in Biomedical Engineering, Vol. 46, No. 8, August 1999

[5] C. L. Nikias and M. Shao, *Signal Processing with α-Stable Distributions and Applications*, John Wiley & Sons, Inc., 1995

[6] P. Mertz, *Model of Impulsive Noise for Data Transmission*, IRE Transactions on Communication Systems, Vol. 9, pp. 130-137, June 1961

[7] J. G. Proakis, *Digital Communications*, McGraw-Hill Series in Electrical and Computer Eng., 3rd ed., 1995

[8] R. Raheli, A. Polydoros and C-K. Tzou, *Per-Survivor Processing: A General Approach to MLSE in Uncertain Environments*, IEEE Trans. on Communications, Vol. 43, pp. 354-364, No. 2/3/4, Feb. 1995

[9] M. Shao and C. L. Nikias, *Signal processing with fractional lower order moments: stable processes and their applications*, Proceedings of the IEEE, vol. 81, No. 7, pp. 985-1010, July 1993

[10] B. W. Stuck and B. Kleiner, *A statistical analysis of telephone noise*, Bell Systems Technical Journal, Vol. 53, No. 7, pp. 1263-1320, 1974

SELF-SIMILAR TRAFFIC SOURCES: MODELING AND REAL-TIME RESOURCE ALLOCATION

Krishnamurthy Nagarajan[1] *G. Tong Zhou*[2]

[1]Couth Infotech Pvt. Ltd., 202 Vijayapuri, Secunderabad, AP 500017, INDIA. Email: krishna@couthit.com
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA.

ABSTRACT

Communication networks have to rely on efficient resource allocation schemes to share the network resources (bandwidth, buffer size, etc.) among users offering different types of traffic (e.g., voice, video and data). Existing schemes based on self-similar traffic models assume that the network traffic is Gaussian and exhibits long-term memory characteristics only. Certain classes of network traffic (e.g., MPEG video traces) are however, non-Gaussian and long-range dependent. In such cases, resource allocation based on simplified assumptions will be either excessive or fail to provide the specified guarantees on the quality of service (QoS). In an earlier work, we had presented an efficient resource allocation scheme for traffic sources having (i) Gaussian as well as non-Gaussian (log-normal) distributions and (ii) exhibiting short-term and/or long-term memory characteristics. In this paper, we assess the real-time performance of our as well as several existing schemes using a Texas Instruments TMS320C6701 DSP processor. The results show that (i) although our algorithm has a higher computational load, real-time implementation is still feasible, and (ii) the increased computational load is justified since the proposed algorithm is more reliable in providing QoS guarantees than existing simplified schemes.

1. INTRODUCTION

A stochastic process $\{S(n),\ n \geq 0\}$ is said to exhibit self-similar characteristics if it satisfies the scaling property $\{S(a \cdot n)\} \stackrel{d}{=} \{a^H\ S(n)\}$, with $a > 0, H > 0$; i.e., the statistical characteristics of $S(n)$ at time $a \cdot n$ is a scaled version of its characteristics at time n. Therefore, except for a scaling factor, $S(a \cdot n)$ and $S(n)$ are *similar*. Here, H is the Hurst parameter. The presence of self-similar characteristics in network traffic was first observed in aggregated Ethernet traffic measurements at Bellcore [7]. Since the original discovery, several independent measurements over different networks (including LAN, WAN, ATM and NSFNET) and traffic generated by commonly used applications such as TELNET, FTP, WWW browsers, and VBR video have been collected and analyzed. These studies clearly demonstrate the presence of self-similar characteristics in aggregated network traffic traces. Several different approaches have been proposed to deal with the self-similar nature of network traffic. The general consensus seems to be that self-similar characteristics must be taken into consideration if the traffic is serviced at medium to high operating load conditions. Load (ρ) is defined as the ratio between the mean traffic rate and the bandwidth at which it is serviced.

In this paper, we focus on the computational complexity of resource allocation schemes that are based on self-similar traffic models. A network link such as a router or switch is expected to support thousands of connections. Each connection carries user traffic requiring certain guarantees on the Quality of Service (QoS) such as loss, delay, delay jitter, etc. Based on the nature of the traffic and the specified QoS requirements, the network link has to decide whether to accept or reject the user connection. The mechanism that makes the accept/reject decision is called as Connection Admission Control (CAC). CAC relies on resource allocation algorithms to make its decision. It is important that the resource allocation schemes require minimal processing and memory overheads so that decisions can be made in real-time. Ideally, one would like to have schemes based on simple analytical expressions derived using parsimonious traffic models. However, to the best of our knowledge such schemes do not currently exist.

Recent advances in the semi-conductor industry have resulted in a tremendous increase in the computational capabilities of microprocessors. The access times as well as cost of memory devices have reduced dramatically. This trend is expected to continue for at least the next several years. Therefore, the focus of our work has been to develop real-time, numerically tractable CAC algorithms. These algorithms should exploit the available processing power to accurately allocate the resources to support the network traffic exhibiting complex behavior.

2. SELF-SIMILAR TRAFFIC MODELS

It is important that traffic models accurately capture both the marginal distribution as well as the autocovariance function exhibited by the network traffic. Although self-similar processes are inherently non-stationary, their increments can be stationary. Let $B(n)$ be a self-similar process with Gaussian marginal distribution and assume $H \in (0,1)$. If $B(n)$ has stationary increments, $x(n) = B(n+1) - B(n)$, then $B(n)$ is called fractional Brownian Motion (fBM), and the corresponding $x(n)$ is called fractional Gaussian Noise (fGN) [11]. The autocovariance function of a fGN process is given by:

$$c_{2x}(\tau) = \frac{E\{B(1)\}^2}{2}\left\{|\tau-1|^{2H} - 2|\tau|^{2H} + |\tau+1|^{2H}\right\}.$$

As $\tau \to \infty$, we have $c_{2x}(\tau) \sim K \cdot |\tau|^{2H-2}$, where K is a constant and $K \neq 0$. Traditional time-series analysis assume the "mixing" condition; i.e., $\sum_{|\tau|} |c_{2x}(\tau)| < \infty$, and hence $c_{2x}(\tau)$ decays at a rate faster than $1/|\tau|$. Note that for the fGN process, $c_{2x}(\tau)$ decays at a rate slower than $1/|\tau|$ implying that $\sum_{|\tau|} c_{2x}(\tau) = \infty$. Stationary processes whose autocovariance functions exhibit such behavior are called as long-memory or long-range dependent processes. For the rest of this paper, we will use the terms long-memory and self-similarity interchangeably.

FGN based traffic models are simple and amenable to mathematical analysis. However, the network traffic exhibit complex behavior requiring more sophisticated traf-

0-7803-7011-2/01/$10.00 ©2001 IEEE

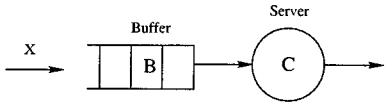


Figure 1. Single server queue.

fic models. A **fractional Autoregressive Integrated Moving Average** (fARIMA) process is an extension of the ARMA process and exhibits long-memory characteristics. A fARIMA process differs from a fGN process in the following manner. The cumulative sum of a fGN process $B(n) = \sum_{k=0}^{n} x(k)$ is *exactly* self-similar; the cumulative sum of a fARIMA process is *asymptotically* self-similar.

In the z-domain, the relation between the fARIMA(p,d,q) process $x(n)$ and the driving noise $w(n)$ is as follows

$$\mathbf{A}(z)X(z) = (1-z^{-1})^{-d}\,\mathbf{B}(z)W(z), \quad -0.5 < d < 0.5, \quad (1)$$

where $\mathbf{A(z)} = 1 + a(1)z^{-1} + \cdots + a(p)z^{-p}$, $\mathbf{B(z)} = 1 + b(1)z^{-1} + \cdots + b(q)z^{-q}$ and $w(n)$ is i.i.d. Gaussian. The presence of a fractional pole at $z = 1$ introduces long-memory. Note that the AR and MA parameters $\{a_i\}_{i=1}^{p}$ and $\{b_j\}_{j=1}^{q}$ give rise to short memory characteristics in the process. Hence, a fARIMA model offers flexibility in accurately capturing *both* the short- and long-memory characteristics exhibited by the network traffic. From the model parameters d ($d = H - 0.5$), $\mathbf{a} = [1,\ a(1), \cdots,\ a(p)]$ and $\mathbf{b} = [1,\ b(1), \cdots,\ b(q)]$, the theoretical $c_{2x}(\tau)$ can be calculated as follows:

$$c_{2x}(\tau) = \frac{(-1)^{\tau}\Gamma(1-2d)}{\Gamma(\tau-d+1)\Gamma(1-\tau-d)} \star \sum_{t} h(t)h(t+\tau), \quad (2)$$

where $\star$ denotes convolution, $H(z) = B(z)/A(z)$ and $\Gamma(k) = \int_0^{\infty} t^{k-1}\exp(t)dt$ is the familiar Gamma function.

Although multiplexed traffic tends to have a Gaussian distribution, individual traffic sources are seldom Gaussian. For example, MPEG traffic sources exhibit a heavier tail than Gaussian [10]. In [6], we proposed a **log-normal fARIMA** traffic model for MPEG traffic sources. Here, we model the log transformed data $y(n) = \ln x(n)$ as fARIMA. We then infer the autocovariance structure of the traffic source $x(n)$ through that of $y(n)$. The relationship between the autocovariance function $c_{2x}(\tau)$ of $x(n)$ and that of $y(n)$ is

$$c_{2x}(\tau) = \exp\{2\mu_y + \sigma_y^2\} \cdot (\exp\{c_{2y}(\tau)\} - 1)\,. \quad (3)$$

3. RESOURCE ALLOCATION FRAMEWORK

Our resource allocation framework provides statistical QoS guarantees based on the Effective Bandwidth Theory (EBT). EBT focuses on a network link such as a router or switch and gives a measure of bandwidth and buffer size required to achieve a trade off between different performance criteria such as the overflow probability, delay, jitter, etc.

Typically, the operation of a network link is modeled as a single server queue (Figure 1) and the allotted resources (bandwidth or capacity C and buffer size B) are expected to provide guarantees on the buffer overflow probability. Buffer overflow probability is amenable to mathematical analysis and is a commonly used performance index.

If an input traffic $x(n)$ is offered to a network link, then a burst of size k will cause an overflow if $\{x(1) + \cdots + x(k)\} > k \cdot C + B$. Denote the sample mean of $\{x(n)\}_{n=1}^{k}$ by $\bar{X}_k = \frac{1}{k}\{x(1) + \cdots + x(k)\}$. Then, if the user demands a loss probability no larger than ϵ, the capacity C and buffer size B should be such that

$$Pr\left[\bar{X}_k > C + \frac{B}{k}\right] \leq \epsilon, \quad \forall\ k. \quad (4)$$

Rigorous calculation of the network resources B and/or C based on (4) requires knowledge of the probability density function (PDF) of $\bar{X}_k$, the sample mean of a burst traffic of size k.

If $x(n)$ is a Gaussian i.i.d random process with mean μ and variance σ^2, then the required capacity or effective bandwidth C is given by [1]

$$C = \mu + \frac{\sigma^2}{2}\delta, \quad (5)$$

where $\delta = \left\lceil \frac{-\ln(\epsilon)}{B} \right\rceil$. Therefore, for a given loss probability ϵ and buffer size B, one can compute δ and then substitute into (5) to find the effective bandwidth C. Alternatively, if the network can only afford a service rate C, then δ can be obtained from (5) and the required buffer size $B = \left\lceil \frac{-\ln(\epsilon)}{\delta} \right\rceil$.

Reference [2] considered Gaussian sources with autocovariance function $c_{2X}(\tau)$ satisfying $\sum_{\tau} c_{2X}(\tau) < \infty$; e.g. ARMA processes, Markov (and its variants) processes, etc. Their *approximate* formula for effective bandwidth is

$$C = \mu + \frac{\delta}{2}\sum_{\tau} c_{2X}(\tau). \quad (6)$$

The presence of long-term memory in network traffic implies that $\sum_{\tau} c_{2x}(\tau)$ is unbounded. Therefore, (6) tends to be too conservative in allocating resources (B and C) when network traffic exhibits long-term memory. In [8], the Bellcore traffic traces were modeled as a fGN process. For a given buffer size B, the effective bandwidth is obtained as

$$C = \mu + \left(\frac{\sqrt{-2\ln\epsilon}}{(1-H)^{H-1}}\right)^{\frac{1}{H}} \cdot H \cdot c_{2x}(0)^{\frac{1}{2H}} \cdot B^{-\frac{1-H}{H}} \quad (7)$$

where $H = d + 0.5$. When $d = 0$ ($x(n) \equiv$ i.i.d) , it can be verified that (7) reduces to (5).

Several recent studies have confirmed that network traffic exhibits both short-term *and* long-term memory characteristics [3, 4]. Calculation of effective bandwidth based on a model that captures only the long-term memory property is inadequate when $x(n)$ exhibits both short-term and long-term memory characteristics. In [5], we presented a resource allocation scheme based on Gaussian fARIMA traffic model. For a specified buffer overflow probability ϵ, the optimal (B,C) pair is obtained by minimizing the cost function

$$J_{(B,C)}(k) \triangleq \frac{k}{2}\frac{(C-\mu+\frac{B}{k})^2}{\gamma_k} + \ln(\epsilon) - \ln(0.5) \geq 0, \quad \forall\, k. \quad (8)$$

Here, $\gamma_k = \sum_{|\tau|<k}(1 - \frac{|\tau|}{k})c_{2x}(\tau)$. Figure 2 shows a plot of $J_{(B,C)}(k)$ for different (B, C) pairs. It is seen that for each (B, C) pair, $J_{(B,C)}(k)$ has a unique minimum at $k = k_o$. If $J_{(B,C)}(k_o) \geq 0$, then (8) is satisfied $\forall\ k$. A (B, C) pair is optimum if $\min_k J_{(B,C)}(k) = J_{(B,C)}(k_o) = 0$.

For $x(n)$ non-Gaussian, the true PDF of the k-sample mean $\bar{X}_k$ does not usually have a closed form expression. In [6] we presented a resource allocation scheme based on log-normal fARIMA traffic model. We approximated $\bar{X}_k$ by a log-normal random variable $\tilde{X}_k$ for which $\ln \tilde{X}_k$ is Gaussian with mean $\tilde{\mu}_k$ and variance $\tilde{\sigma}_k^2$. Now going back to the resource allocation framework (4), we replace $\bar{X}_k$ by $\tilde{X}_k$ and write

$$Pr\left\{\ln \tilde{X}_k > \ln(C + B/k)\right\} \leq \epsilon, \quad \forall\ k.$$

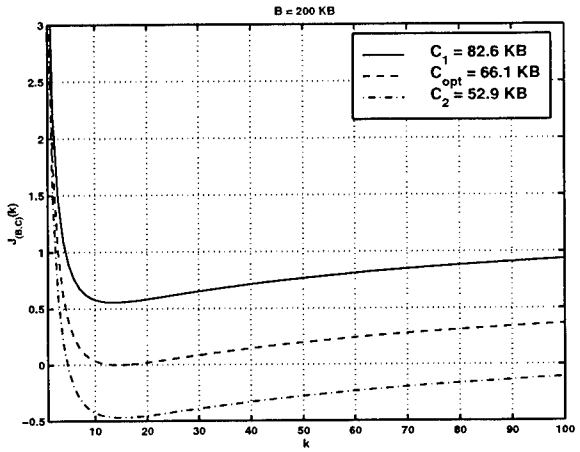


Figure 2. Plot of $J_{(B,C)}(k)$ for different (B, C) pairs.

Since $\ln \tilde{X}_k$ is a Gaussian random variable with mean $\tilde{\mu}_k$ and variance $\tilde{\sigma}_k^2$, we apply the Gaussian resource allocation framework to obtain the cost function

$$J_{(B,C)}(k) \triangleq \frac{(\ln(C + B/k) - \tilde{\mu}_k)^2}{2\tilde{\sigma}_k^2} + \ln(2\epsilon) \geq 0, \quad \forall\, k, \quad (9)$$

to compute the optimal (B, C) pair.

In the next section, we will analyze the computational complexity of our fARIMA based resource allocation schemes.

4. REAL-TIME IMPLEMENTATION

Resource allocation algorithms typically run as a software module in a network device such as a router or switch. In order to analyze the real-time performance of fARIMA model based resource allocation algorithms, we selected the Texas Instruments (TI) TMS320C6701 floating point DSP processor. The processor's dedicated multiplier-accumulator circuitry, multiple access and special memory addressing modes designed to speed up repetitive operations make it an ideal platform for CAC algorithm implementation. We operated the processor at a clock frequency of 133 MHz. The processor's on-chip program and data memory (64 KB each) were found to be sufficient for implementing our algorithm. The entire algorithm was implemented in the C-programming language.

4.1. Implementation Details

The algorithm first obtains the QoS requirement and the traffic model parameters from the user. Then, it calculates the autocovariance function $c_{2x}(\tau)$ from the model parameters. In our current implementation, we fixed the maximum lag value ($\tau_{\max}$) at 1024. The correlation values are stored in a double-precision floating-point array. An implication of using an upper limit on τ is that the algorithm can only compute those optimal (B, C) pairs for which the burst size $k_o = \arg\min_k J_{(B,C)}(k)$ is less than $\tau_{\max}$.

Based on the QoS parameter specified by the user and existing traffic conditions in that class of service, the network link allots a fixed buffer size B to the user traffic. Regulating the buffer size provides flexibility in handling delay requirements. Once B is fixed, the optimal bandwidth C_{opt} needed to support the connection is computed. Recall that when $C = C_{opt}$, we have $\min_k J_{(B,C)}(k) = 0$. From equations (8)-(9), we observe that $J_{(B,C)}(k)$ is monotonically increasing with C. Therefore if $C > C_{opt}$, we have that $\min_k J_{(B,C)}(k) > 0$ whereas if $C < C_{opt}$, we have that $\min_k J_{(B,C)}(k) < 0$. This prompts us to employ an iterative search algorithm to find C_{opt}. We first pick C_1 and C_2 such that $\min_k J_{(B,C_1)}(k) > 0$ and $\min_k J_{(B,C_2)}(k) < 0$. We can be sure that C_{opt} lies between C_1 and C_2, i.e., $C_1 > C_{opt} > C_2$, and we call C_1 and C_2 the bracket points. Next, we would like to narrow down this bracket. At the ith iteration, pick $C_1 > C_i > C_2$. If $\min_k J_{(B,C_i)}(k) > 0$, then we infer that $C_i > C_{opt} > C_2$ and we replace C_1 with C_i. On the other hand, if $\min_k J_{(B,C_i)}(k) < 0$, then we must have $C_1 > C_{opt} > C_i$ and we replace C_2 with C_i. By successively narrowing down the range, we bring the bracket points together and soon they converge to C_{opt}. Brent's method in particular provides superlinear convergence to the optimal solution [9].

In the process of searching for an optimal bandwidth, the algorithm needs to repeatedly search for the burst size k_o that results in the minimum cost function. For a given bandwidth C, the algorithm first obtains a three point bracket $(k_a,\ k_b,\ k_c)$ that captures the minimum [9, page 400]. Then, applying the "Golden Search" method, it identifies the burst size k_o within the interval $(k_a,\ k_c)$.

If the network link can afford to allocate C_{opt} to the user, it goes ahead and accepts the user connection. If it does not have the required bandwidth, the algorithm provides the user with an option to renegotiate its QoS requirement.

4.2. Performance Evaluation

We experimented with a MPEG video trace "Dino", obtained from [10]. "Dino" has a heavier tail than Gaussian. Its mean and standard deviation is equal to 20 KB and 8 KB respectively, long-memory parameter $d = 0.35$ and short-memory parameters $\hat{\mathbf{a}} = [1,\ -0.56]$ and $\hat{\mathbf{b}} = [1]$ respectively. Suppose this source is to be admitted to the network with a desired loss probability $\epsilon = 10^{-5}$. Based on this QoS parameter, the CAC algorithm allocates buffer size B and bandwidth C according to the assumed traffic model (c.f. Section 3).

If the traffic is assumed to be Gaussian and long-range dependent only, then the required bandwidth for a fixed buffer size can be calculated directly using (7). FGN model based resource allocation scheme has the lowest computational demand. If the traffic is assumed to be Gaussian having both short- and long-memory characteristics, then the optimal resources can be obtained by minimizing (8). If we model the traffic as a log-normal fARIMA process to accurately capture its marginal distribution, then the optimal resources can be obtained by minimizing (9).

Table 1 shows the CPU clock counts required for fARIMA model based resource allocation algorithms to converge. For B ranging from 50 KB to 400 KB, the algorithms converge within 100 milli-seconds. We observe that their rate of convergence depends on the following factors:

- Buffer size (B) - the execution time is directly proportional to the allotted buffer size. For large buffer sizes, the search space to compute the optimal C is higher resulting in an increase in execution time.
- Bracketing strategy - the execution time is depends on the number of times the algorithm evaluates different C's to obtain the bracket points.
- Strategy for obtaining C_{opt} within the bracket points.
- Strategy for obtaining the minimum of $J_{(B,C)}(k)$.

The execution time can be further reduced by (i) using a DSP processor with a higher MIPS (million instructions per second) rating, (ii) implementing key modules in assembly programming language, (iii) identifying modules that can be implemented in parallel and (iv) developing better bracketing and search strategies.

B (KB)	C_{opt} (KB)	Cycle Count	Execution Time (m-sec)
50	47.12	9,831,479	74
100	44.25	10,265,182	77
200	41.52	11,732,340	88
300	40.05	12,040,384	90
400	39.07	12,246,093	92

(a)

B (KB)	C_{opt} (KB)	Cycle Count	Execution Time(m-sec)
50	84.57	8,891,587	67
100	75.77	9,401,718	71
200	66.09	11,201,191	84
300	60.68	11,761,252	88
400	57.16	12,069,090	91

(b)

Table 1. Real-time implementation of fARIMA model based resource allocation schemes: (a) Gaussian (b) log-normal.

Is it important to have resource allocation schemes based on traffic models that accurately capture the marginal distribution and the autocovariance structure exhibited by the network traffic? Why can't we use fGN model based schemes for all traffic sources? Table 2 shows single server queue simulation results to illustrate the importance of having such schemes. Based on the log-normal fARIMA model parameters obtained from "Dino", we generated a synthetic traffic trace of length 10^7 samples to offer as input to the single server queue. Table 2(a) gives the buffer overflow probability when the network link allocates resources based on the fGN traffic model. Table 2(b) gives the overflow probability when the resources are allocated based on Gaussian fARIMA traffic model. From the results, we observe that resource allocation schemes based on inaccurate traffic models fail to maintain the specified guarantees on the overflow probability. When the resources are allotted based on log-normal fARIMA traffic model, which accurately captures marginal distribution as well as the autocovariance structure of the traffic, the allotted resources provide the specified guarantees; see results in Table 2(c).

5. CONCLUSIONS

FGN traffic model based resource allocation scheme has the lowest computational requirements for calculating the optimal (B, C) pair. However, it is not sufficient for accurately capturing the statistical characteristics of different traffic types. FARIMA based traffic models provide more flexibility in parsimoniously capturing the short- and long-memory characteristics of network traffic. Although, our fARIMA model based resource allocation schemes have a higher computational load, real-time implementation is still feasible. When implemented on a TI TMS320C67 DSP processor, the algorithms typically converge within 100 milli-seconds. Several optimization techniques have been identified which can further reduce the execution time significantly.

Acknowledgment: This work was supported in part by National Science Foundation grant MIP 9703312.

B (KB)	C_{opt} (KB)	$Pr\{Q(n) > B\}$
50	40.67	9.9×10^{-3}
100	38.27	9.7×10^{-3}
200	36.19	8.7×10^{-3}
300	35.10	8×10^{-3}
400	34.37	7.5×10^{-3}

(a)

B (KB)	C_{opt} (KB)	$Pr\{Q(n) > B\}$
50	47.12	2.5×10^{-3}
100	44.25	2.2×10^{-3}
200	41.52	1.8×10^{-3}
300	40.05	1.6×10^{-3}
400	39.07	1.4×10^{-3}

(b)

B (KB)	C_{opt} (KB)	$Pr\{Q(n) > B\}$
50	84.57	2.3×10^{-6}
100	75.77	2.5×10^{-6}
200	66.09	4×10^{-6}
300	60.68	5.7×10^{-6}
400	57.16	8.3×10^{-6}

(c)

Table 2. Queue simulations results ($\epsilon = 10^{-5}$) using a synthetic 'Dino' traffic trace of size 10^7 samples. For a fixed B, C_{opt} was computed based on (a) fGN, (b) Gaussian fARIMA and (c) log-normal fARIMA traffic models respectively. $Q(n)$ is the queue length process.

REFERENCES

[1] F. Kelly, "Notes on effective bandwidth," *Stochastic networks: theory and applications*, Oxford Univ. Press, 1996.

[2] C. Courcoubetis and R. Weber, "Effective bandwidths for stationary sources," *Prob. Eng. Inf. Sci.*, Vol. 9, pp 285-296, 1995.

[3] J. Beran, R. Sherman, M. Taqqu, and W. Willinger, "Long-range dependence in variable bit rate video traffic," *IEEE Trans. Comm.*, Vol. 43, pp 1566-1579, 1995.

[4] C. Huang, M. Devetsikiotis, I. Lambadaris, and A. Kaye, "Modeling and simulation of self-similar variable bit rate compressed video: A unified approach," *Proceedings ACM SIGCOMM*, 1995.

[5] K. Nagarajan and G. T. Zhou, "A new resource allocation scheme for Gaussian traffic sources," *Proc. ICASSP*, pp 2609-2612, Istanbul, Turkey, 2000.

[6] K. Nagarajan and G. T. Zhou, "A new resource allocation scheme for MPEG video sources," *Proc. 34th Asilomar Conference on Signals, Systems, and Computers*, pp 1245-1249, Pacific Grove, CA, October 2000.

[7] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic," *IEEE/ACM Trans. Net.*, pp 1-15, Vol. 2, No. 1, 1994.

[8] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE J. Selected Areas Comm.*, Vol. 13, No. 6, pp 953-962, 1995.

[9] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge Univ. Press, 1992.

[10] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems," *Tech. Rep. No. 101*, University of Wurzburg, 1995.

[11] G.Samarodnitsky and M.Taqqu. *Stable non-Gaussian Random Processes.* Chapman and Hall, 1994.

Super-efficiency in Blind Signal Separation of Symmetric Heavy-Tailed Sources

Yoav Shereshevski, Arie Yeredor, Hagit Messer
Dept. of Elect. Eng. - Systems, Tel-Aviv University

Abstract— **This paper addresses the Blind Source Separation (BSS) problem in the context of "heavy-tailed", or "impulsive" source signals, characterized by the nonexistence of finite second (or higher) order moments. We consider Pham's Quasi-Maximum Likelihood (QML) approach, a modification of the Maximum Likelihood (ML) approach, applied using some presumed distributions of the sources. We introduce a related family of suboptimal estimators, termed Restricted QML (RQML). A theoretical analysis of the asymptotic performance of RQML is presented. The analysis is used for showing that the variance of the optimal (non-RQML) estimator's error must decrease at a rate faster than $1/T$ (where T is the number of independent observations). This surprising property, sometimes called super-efficiency, has been observed before (in the BSS context) only for finite-support source distributions. Simulation results illustrate the good agreement with theory.**

I. Introduction

Heavy-tailed distributions assign relatively high probabilities to the occurrence of large deviations from the median, and are often used for modelling impulsive signals. A common characteristic property of many heavy-tailed distributions (such as the α-stable family), is the nonexistence of finite second (or higher) order moments, which will take an important roll in this work.

The simplest Blind Signal Separation (BSS) model assumes that N instantaneous linear mixtures of N independent stationary source signals are observed. They are formulated as $\mathbf{x}(t) = \mathbf{A}\cdot\mathbf{s}(t)$ where $\mathbf{x}(t)$ and $\mathbf{s}(t)$ are $N\times 1$ column vectors of the observed signals and source signals (respectively), and the square $N \times N$ "mixing matrix" $\mathbf{A}$ contains the mixing coefficients. The BSS problem consists of recovering the sources $\mathbf{s}(t)$ using only the observed data $\mathbf{x}(t)$, $t = 1, 2, \ldots T$ and the assumption of independence between the entries of $\mathbf{s}(t)$. It can be formulated as the computation of an $N \times N$ "separating matrix" $\mathbf{B}$ whose output $\mathbf{y}(t) = \mathbf{B}\cdot\mathbf{x}(t)$ is an estimate of $\mathbf{s}(t)$.

There are several well-known methods for BSS (see, e.g., [1] for a comprehensive survey), based on Information Theory, High-Order Statistics (HOS), etc. Some of these methods (e.g., using the whiteness constraint or HOS) assume that the sources have finite second (or higher) order moments. Other methods, such as Pham's Quasi Maximum Likelihood (QML) [2], avoid this assumption in the estimation algorithm; however it is still used for the error analysis.

Thus, when the source signals are "impulsive" with no finite second (or higher) order moments, the use and/or the error analysis of existing BSS algorithms become problematic in many respects. This paper is concerned with the asymptotic ($T \to \infty$) error analysis in such cases. To facilitate the analysis, the sources are assumed to be strictly white, i.e., each source signal is a sequence of independent, identically distributed (i.i.d.) random variables. Symmetric distributions are also assumed.

II. QML and RQML

A popular variant of the Maximum Likelihood (ML) estimator in "blind" contexts is the QML estimator [2], [3]. QML attempts to apply the ML approach using some given hypothetical model for the sources' probability distribution function (p.d.f.) as a substitute for the true unknown model. This leads to solving (with respect to the elements of $\mathbf{B}$) the following system of estimating equations, as outlined in [2]

$$\hat{E}\left[\Psi_i\left(y_i\right)y_j\right] = 0 \quad 1 \le i \ne j \le N \tag{1}$$

where:
$\hat{E}[\cdot]$ is the-time averaging operator, $\hat{E}[z] \triangleq \frac{1}{T}\sum_{t=1}^{T} z(t)$;
y_i is the i-th estimated source signal at the output of the separating matrix $\mathbf{B}$;
$\Psi_i(x)$ are some nonlinear "separation functions", chosen a priori, which would optimally be the (unknown) score functions of the i-th source.

The "small errors" analysis derived for this estimator in [2] is not valid for heavy-tailed sources which do not posses finite second moments. Note, for example, that even when $\mathbf{B}$ equals $\mathbf{A}^{-1}$ and we have $\mathbf{y} = \mathbf{s}$, the left-hand side of (1), $\hat{E}\left[\Psi_i\left(s_i\right)s_j\right]$, does not converge to 0 in L_2 sense (since $E\left[s_j^2\right]$ is infinite), which undermines the validity of a second-order-based analysis in this case. In order to mitigate this difficulty, we propose the following restricted estimator, which we term the "Restricted QML" (RQML):

$$\hat{E}[\Psi_i(y_i)y_j I(|y_j| < C)] = 0 \quad i \ne j \tag{2}$$

Where the Indicator Function $I(|x| \le C)$ equals 1 iff $|x| \le C$ (and equals 0 otherwise), and C is some arbitrary positive constant.

III. Small errors analysis of RQML

We now introduce a "small errors" analysis for the RQML estimate. Because of the permutation ambiguities, a "good" separation procedure needs not produce a

0-7803-7011-2/01/$10.00 ©2001 IEEE

$\mathbf{B}$ close to inverse of the true $\mathbf{A}$, but only that $\mathbf{B}\cdot\mathbf{A}$ be close to a permuted diagonal matrix. However, if we *permute and scale both* $\mathbf{B}$ and $\mathbf{A}$ *by the same convention*, we may expect that $\mathbf{B}\cdot\mathbf{A}=\mathbf{I}-\boldsymbol{\varepsilon}$, where $\mathbf{I}$ is the identity matrix, and $\boldsymbol{\varepsilon}$ is a "small" matrix. Now since $\mathbf{y}=\mathbf{B}\cdot\mathbf{A}\cdot\mathbf{s}$ we have $y_i=s_i-\sum_{j=1}^{N}\varepsilon_{ij}s_j$, where s_j denotes the j-th source and ε_{ij} denotes the general element of the error matrix $\boldsymbol{\varepsilon}$

To simplify the exposition, from now on we assume a two sensors - two sources model ($N=2$), using identical separation functions $\Psi_i=\Psi$. Generalization to more sensors and sources with individual Ψ_i-s is straightforward, but will not be pursued in here. Thus, $y_i \overset{N=2}{=} (1-\varepsilon_{ii})s_i-\varepsilon_{ij}s_j$. Under the "small errors" assumption, $\varepsilon_{ii}\ll 1$ (asymptotically), so that

$$y_i \approx s_i-\varepsilon_{ij}s_j \quad i,j=1,2;\ i\neq j. \tag{3}$$

Substituting (3) into (2) we have (for the $N=2$ case)

$$\hat{E}\left[\Psi(s_1-\varepsilon_{12}s_2)(s_2-\varepsilon_{21}s_1)I(|s_2-\varepsilon_{21}s_1|\leq C)\right]=0 \tag{4a}$$

$$\hat{E}\left[\Psi(s_2-\varepsilon_{21}s_1)(s_1-\varepsilon_{12}s_2)I(|s_1-\varepsilon_{12}s_2|\leq C)\right]=0 \tag{4b}$$

Eqns. (4a)&(4b) implicitly relate the off-diagonal elements of $\boldsymbol{\varepsilon}$ to $\mathbf{s}(1:T)$ (i.e., to $s_1(1),s_2(1),...,s_1(T),s_2(T)$). When ε_{ij} can be formulated as an explicit function of $\mathbf{s}(1:T)$, it is relatively straightforward to deduce the statistics of the errors from those of the sources. However, for arbitrary $\Psi(x)$, there is no explicit closed-form solution expressing ε_{ij} in terms of $\mathbf{s}(1:T)$. Consequently, the error analysis becomes more involved. We have to employ some general statistical assumptions on the errors ε_{ij} in order to find its more particular statistical properties. The general assumptions to be used are:

- A1: L_2 consistency: $\varepsilon_{ij}\to 0$ (in L_2 sense) as $T\to\infty$; and
- A2: Stronger convergence of higher powers of the error: i.e., ε_{ij}^n converges faster to zero as the order n increases; consequently, $\varepsilon_{ij}^n\ll\varepsilon_{ij}$ (asymptotically) for $n=2,3,\ldots$.

We restrict the discussion to separation functions $\Psi(x)$ which are differentiable, odd and bounded. In addition, we now further assume that $\Psi(x)$ satisfies the following condition:

- B: There exists some finite C_1 ($C_1\ll C$), such that $\Psi(x)$ (and its derivative $\Psi'(x)$) vanish outside the region $|x|<C_1$. This condition can be formulated as:

$$\Psi(x)=\Psi(x)\cdot I(|x|<C_1) \tag{5a}$$

$$\Psi'(x)=\Psi'(x)\cdot I(|x|<C_1), \tag{5b}$$

where $C1\ll C$.

Generally, this is a rather restrictive condition, used in here only to simplify the derivation. In [4] we show how this condition can be considerably relaxed. Note, however, that any differentiable estimating function $\Psi(x)$ can be tailored at the boundaries so as to smoothly roll-off to zero inside the region $|x|<C_1$. Since all useful $\Psi(x)$ usually decrease to zero at $\pm\infty$, this tailoring would not change $\Psi(x)$ by much if C_1 is large enough.

Substituting (5a) into (4a) we get

$$\hat{E}\left[\Psi(s_1-\varepsilon_{12}s_2)(s_2-\varepsilon_{21}s_1)I(|s_1-\varepsilon_{12}s_2|\leq C_1)\cdot \cdot I(|s_2-\varepsilon_{21}s_1|\leq C)\right]=0 \tag{6}$$

It is relatively straightforward to show, that under the "small errors" assumption and with $C_1\ll C$, the product $I(|s_1-\varepsilon_{12}s_2|\leq C_1)I(|s_2-\varepsilon_{21}s_1|\leq C)$ is nonzero only if $|s_2|\leq C$ (to some approximation). Thus, we may now use the Taylor expansion of $\Psi(s_1-\varepsilon_{12}s_2)$ about s_1. This is legitimate wherever the product of indicators is nonzero, since $\varepsilon_{12}s_2\ll s_1$ ($\varepsilon_{12}\to 0$) and $\varepsilon_{12}s_2$ is bounded (by $\varepsilon_{12}C$), and we have:

$$E\left[(\varepsilon_{12}s_2)^2\right]\leq E\left[(\varepsilon_{12}C)^2\right]=C^2E\left[\varepsilon_{12}^2\right]\overset{T\to\infty}{\longrightarrow}0. \tag{7}$$

So under assumption A1 we have $\varepsilon_{12}s_2\xrightarrow{L_2}0$, and it is sufficient to expand $\Psi(s_1-\varepsilon_{12}s_2)$ about s_1 up to first order, since higher orders of $\varepsilon_{12}s_2$ are negligible under assumption A2. It is important to observe, that although the QML estimator is equivalent to RQML when $C\to\infty$, we then have $E\left[(\varepsilon_{12}s_2)^2\right]\approx E\left[\varepsilon_{12}^2\right]\cdot E\left[s_2^2\right]\to\infty$ for any number of observations T. So, essentially, the reason we can use L_2 analysis of the Taylor expansion of $\Psi(s_1-\varepsilon_{12}s_2)$ about s_1 and neglect higher powers of $\varepsilon_{12}s_2$, is the presence of the restricting constant C used in RQML.

Expanding $\Psi(s_1-\varepsilon_{12}s_2)$ up to first order,

$$\Psi(s_1-\varepsilon_{12}s_2)\approx\Psi(s_1)-\Psi'(s_1)\varepsilon_{12}s_2, \tag{8}$$

and substituting into (4a) we get:

$$\hat{E}\left[\{\Psi(s_1)-\Psi'(s_1)\varepsilon_{12}s_2\}(s_2-\varepsilon_{21}s_1)\cdot \cdot I(|s_2-\varepsilon_{21}s_1|\leq C)\right]=0. \tag{9}$$

Using condition B we obtain

$$\hat{E}\left[\{\Psi(s_1)-\Psi'(s_1)\varepsilon_{12}s_2\}(s_2-\varepsilon_{21}s_1)I(|s_1|\leq C_1)\cdot \cdot I(|s_2-\varepsilon_{21}s_1|\leq C)\right]=0 \tag{10}$$

It is again relatively straightforward to show, under the assumptions of "small errors" and that $C\gg C_1$, that the following products of indicator functions are nearly equivalent:

$$I(|s_1|\leq C_1)I(|s_2-\varepsilon_{21}s_1|\leq C)\approx I(|s_1|\leq C_1)I(|s_2|\leq C)$$

So we may write

$$\hat{E}[\{\Psi(s_1)-\Psi'(s_1)\varepsilon_{12}s_2\}(s_2-\varepsilon_{21}s_1)\cdot \cdot I(|s_1|\leq C_1)I(|s_2|\leq C)]=0. \tag{11}$$

We use condition B again to obtain:

$$\hat{E}\left[\{\Psi(s_1)-\Psi'(s_1)\varepsilon_{12}s_2\}(s_2-\varepsilon_{21}s_1)I(|s_2|\leq C)\right]$$

$$= \underbrace{\hat{E}\left[\Psi(s_1)s_2 I(|s_2|\le C)\right]}_{A_1} - \varepsilon_{21}\underbrace{\hat{E}\left[\Psi(s_1)s_1 I(|s_2|\le C)\right]}_{A_2} - \varepsilon_{12}\underbrace{\hat{E}\left[\Psi'(s_1)s_2^2 I(|s_2|\le C)\right]}_{A_3} + \varepsilon_{12}\varepsilon_{21}\underbrace{\hat{E}\left[\Psi'(s_1)s_1 s_2 I(|s_2|\le C)\right]}_{A_4}$$

$$\triangleq A_1 - \varepsilon_{21}A_2 - \varepsilon_{12}A_3 + \varepsilon_{21}\varepsilon_{12}A_4 = 0 \qquad (12)$$

Note that A_1, A_2, A_3, A_4 are all time-averages of some bounded functions of random variables. As such, due to the i.i.d. assumption, they converge to their respective mean values, which are all finite. We may therefore neglect the term A_4 (which converges $E[A_4] = 0$ and is further multiplied by $\varepsilon_{12}\varepsilon_{21}$). A_2, A_3 converge to some nonzero value while A_1 converges to zero ($E[A_1] = E\left[\Psi(s_1)s_2 I(|s_2|\le C)\right] = E\left[\Psi(s_1)\right]\cdot E\left[s_2 I(|s_2|\le C)\right] = 0$ because s_1, s_2 are independent and have symmetric density and $\Psi(x)$ is odd). Finally, replacing A_2, A_3 with theirs means, (12) becomes:

$$A_1 - \varepsilon_{21}E[A_2] - \varepsilon_{12}E[A_3] = 0 \qquad (13)$$

where

$$\begin{aligned} E\left[A_2\right] &= E\left[\Psi(s_1)s_1 I(|s_2|\le C)\right] \\ &= E\left[\Psi(s_1)s_1\right]\cdot P(|s_2|\le C) \end{aligned}$$

and

$$\begin{aligned} E\left[A_3\right] &= E\left[\Psi'(s_1)s_2^2 I(|s_2|\le C)\right] \\ &= E\left[\Psi'(s_1)\right]\cdot E\left[s_2^2 I(|s_2|\le C)\right] \end{aligned}$$

since s_1, s_2 are independent. $P(|s_2|\le C)$ is the probability that $|s_2|\le C$. Note that $\lim_{C\to\infty} E\left[A_3\right] = \infty$ while $\lim_{C\to\infty} E\left[A_2\right] = E\left[\Psi(s_1)s_1\right] < \infty$. So increasing C would result in $E\left[A_3\right] \gg E\left[A_2\right]$ as $T\to\infty$ and we have $A_1 - \varepsilon_{12}E[A_3] \approx 0$ or equivalently

$$\varepsilon_{12} \approx \frac{A_1}{E[\Psi'(s_1)]\cdot E[s_2^2 I(|s_2|\le C)]}$$

substituting the definition of A_1 we have

$$\varepsilon_{12} \approx \frac{\hat{E}\left[\Psi(s_1)s_2 I(|s_2|\le C)\right]}{E[\Psi'(s_1)]E[s_2^2 I(|s_2|\le C)]} \qquad (14)$$

Note that when $T\to\infty$, A_1 is the sum of infinitely many i.i.d random variables (with finite variance). The central limit theorem guarantees that ε_{12} converges in distribution to a Normal random variable with mean :

$$E[\varepsilon_{12}] = \frac{E\left[A_1\right]}{E[\Psi'(s_1)]\cdot E[s_2^2 I(|s_2|\le C)]} = 0 \qquad (15)$$

(since $E\left[A_1\right] = 0$), and with variance :

$$E[\varepsilon_{12}^2] \approx \frac{E\left[A_1^2\right]}{E[\Psi'(s_1)]^2\cdot E[s_2^2 I(|s_2|\le C)]^2}. \qquad (16)$$

It can be further shown (using the i.i.d. assumption and the independence of the sources), that

$$E\left[A_1^2\right] = \frac{1}{T}E\left[\Psi(s_1)^2\right]\cdot E\left[s_2^2 I(|s_2|\le C)\right]. \qquad (17)$$

Substituting into (16), we obtain:

$$E[\varepsilon_{12}^2] \approx \frac{1}{T}\cdot\frac{E\left[\Psi(s_1)^2\right]}{E[\Psi'(s_1)]^2}\cdot\frac{1}{E[s_2^2 I(|s_2|\le C)]} \qquad (18)$$

The approximation $E\left[A_3\right] \gg E\left[A_2\right]$ was used under the assumption that C is "large" and $T\to\infty$. When this is not the case (C is not "large" enough), the error can be found by solving equation (13) together with (4b) derived in the same manner.

IV. Discussion

Obviously, the RQML estimate is a sub-optimal estimate, since the Indicator Functions discard occurrences of "outliers", which often reflect valuable information. Its only advantage lies with the fact, that in the RQML framework all moments exist, and therefore a "small errors" L_2 performance analysis is enabled. Moreover, the obtained analytic results carry implications on the attainable performance of better, non-RQML (e.g., QML) estimates. They allow quantification of the relative effect of the separating function $\Psi(x)$ on the RQML performance; this relative effect would be maintained as the constant C is increased, and it is therefore characteristic of QML as well. Furthermore, interesting implications on the attainable error convergence rate (with respect to the observation length T) can be deduced.

Specifically, the two following conclusions can be drawn from the analysis:

1. The asymptotic performance relates to the nonlinear separating function $\Psi(x)$ through the factor $\frac{E[\Psi(x)^2]}{E[\Psi'(x)]^2}$, where $E[\cdot]$ denotes the expectation with respect to the true distribution of the corresponding source (note that this factor is common to some other estimation problems using QML approach [3]). This relation provides some insight on how to choose the separation function $\Psi(x)$, or equivalently, predicts the performance degradation due to the use of suboptimal separation functions. When $\Psi(x)$ is the true score function, that is, $\Psi(x) = \frac{f'(x)}{f(x)}$ (where $f(x)$ is the true p.d.f. of the corresponding source), performance is optimized.
2. The variance of the optimal (non-RQML) estimator's error must converge to zero faster than the regular rate of $1/T$. This property, sometimes termed "super-efficiency" (e.g., [5]) has so far been observed (in the context of BSS) only for finite-support source distributions ([1]). It can be deduced here from (18) using the following argument: Assume that an optimal estimator exists, whose error-variance decreases as $1/T$, tending (asymptotically) to ρ/T where ρ is some constant. Now from (18), the RQML factor multiplying $1/T$ can be arbitrarily decreased below

any positive ρ by increasing C: for every "large enough" T_0, there exists some C_0 that would yield a smaller variance than ρ/T for all $T > T_0$. This contradict the optimality of the presumed optimal estimator. We therefore deduce, that the optimal estimator's error variance must decrease at a faster rate than $1/T$.

As mentioned earlier, as $C \to \infty$, RQML approaches QML. However, the effective value of C is closely related to the observation length T. With C fixed, the probability of occurrence of an outlier (strong enough to cause truncation) obviously increases monotonically with T. In other words, for any arbitrarily large C, the performance of RQML will only approach that of QML up to a certain value of T, and would then remain significantly worse as T increases.

In order to asymptotically improve RQML's performance, we can consider increasing C as the observation time T increases (hence denoting C as C_T). Note that if (18) still holds (with C substituted by C_T), then the RQML error variance would decrease faster than $1/T$. Thus, a key question here is, to what extent we can increase C_T with T, while maintaining the validity of (18).

The answer to this question naturally depends on the sources' distribution (more precisely, on the decreasing rate of their tails). For example, we prove elsewhere [4], that when the sources are Symmetric α-Stable ($S\alpha S$) signals with parameter α (which means that their tails decrease as $\frac{1}{|x|^{\alpha+1}}$ for "large" $|x|$), C_T can be increased as $C_T \sim T^{\frac{1}{\alpha}-\frac{\delta}{2}}$ (where δ is some arbitrarily small positive number). Consequently, the error variance will decrease at least as fast as $1/T^{\frac{2}{\alpha}-\delta}$.

The following figure demonstrates the agreement of some simulation results with our theoretical analysis. We used two symmetric α-stable sources, and applied the RQML approach using cauchy's distribution score function as the separating function $\Psi(x)$. The '+'-s indicate simulations results (as a function of T) in terms of the mean square error of ε_{12}. The solid line describes the expected performance (18). As expected, asymptotically the simulations results agree with our analytic results.

In addition, we present simulations results ('o') obtained by applying the QML algorithm (with the same $\Psi(x)$) to the same data. The dashed line indicates a slope corresponding to an error convergence rate of $1/T^{2/\alpha}$, which is seen to fit the QML simulations results, as expected from our discussion above. The vertical position of this line was determined manually, since we do not have a closed-form expression for the QML performance.

Note also, that with the chosen value of C, the RQML error follows the QML error for the smaller values of T, and then departs to values that are significantly worse than QML. By using an increased value for C, the departure point could be delayed (in T). Naturally, however, RQML (with C fixed) would always be asymptotically worse than QML, having a decrease rate of $1/T$ vs. $1/T^{2/\alpha}$.

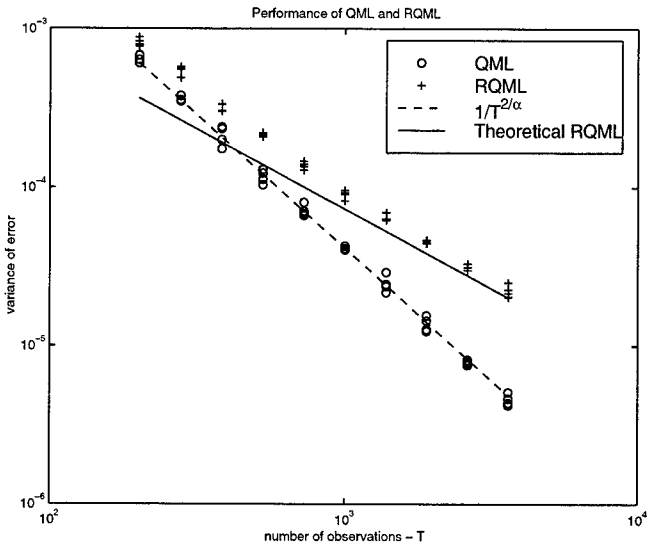


Fig. 1. Simulation results for two symmetric α - stable sources with parameters $\alpha = 1.2$ and $\sigma = 1.2$. $\Psi(x)$ is score function of a symmetric Cauchy ($\alpha = 1$) distribution with $\sigma = 1$. The solid line represents (18) (for $C = 120$), with which the RQML simulations results ('+') are seen to agree asymptotically. The QML results ('o') exhibit super-efficiency, with the predicted decrease rate of $1/T^{2/\alpha}$. Each point represents the average result of 500 independent experiments.

V. Conclusion

We addressed the performance analysis of BSS in the context of "heavy-tailed" signals. When these signals do not have finite second-order moments, standard analysis tools (formerly used, e.g., to analyze the QML estimate) are no longer useful.

To enable L_2 error analysis, we introduced the RQML sub-optimal estimate, which is parameterized by a limiting-constant C, such that when $C \to \infty$, RQML approaches QML. Using "small errors" analysis, we obtained expressions for the RQML performance.

Using these results, we concluded that the optimal estimator's performance must be super-efficient, in the sense that its mean squared error must decrease (asymptotically) at a rate faster that $1/T$. More specifically, we demonstrated that for symmetric α-stable sources, the decrease rate of QML is $1/T^{2/\alpha}$.

References

[1] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 9, pp. 2009–2025, oct 1998.

[2] D.-T. Pham and P. Garat, "Blind separation of mixture of sources through a quasi-maximum likelihood approach," *IEEE Transactions on signal processing*, vol. 45, pp. 1712–1725, July 1997.

[3] J. Friedmann, H. Messer, and J.-F. Cardoso, "Robust parameter estimation of a deterministic signal in stable noise," *IEEE Transactions on Signal Processing*, April 2000.

[4] Y. Shereshevski, "Blind signal separation of heavy-tailed sources (M.Sc. thesis)," 2001, Dept. of Elect. Eng. - Systems, Tel-Aviv University.

[5] S.-I. Amari, "Superefficiency in blind source separation," *IEEE Transactions on signal processing*, vol. 47, pp. 936–944, April 1999.

KALMAN FILTERING FOR SELF-SIMILAR PROCESSES

Meltem Izzetoğlu, Birsen Yazıcı, Banu Onaral, Nihat Bilgütay

Drexel University,
Electrical and Computer Eng. Dept.,
Philadelphia, PA, 19104

ABSTRACT

In our earlier work, we introduced a class of stochastic processes obeying a structure of the form, $E[X(t)X(t\lambda)] = R(\lambda), t, \lambda > 0$, and outlined a mathematical framework for the modeling and analysis for these processes. We referred to this class of processes as scale stationary processes. We demonstrated that scale stationarity framework leads to engineering oriented mathematical tools and concepts, such as autocorrelation and spectral density function and finite parameter ARMA models for modeling and analysis of statistically self-similar signals. In this work, we will introduce a state space representation for self-similar signals and systems based on scale stationary ARMA models. Such a representation provides a complete description of the inner and outer dynamics of a self-similar system or signal that can not be obtained from transfer function representation. We will introduce Kalman filtering techniques and Ricatti Equations for smoothing and prediction of self-similar processes.

1. INTRODUCTION

$1/f$ processes occur in a broad range of engineering and science applications including network traffic, noises in electronic devices, biomedical systems, burst error in communication channels to mention a few [1]-[9].

The major characteristics of these processes are their long term correlation structure, and their statistical self-similarity. These characterizations are apparent in the empirical $1/f^{\gamma}$ power spectrum. Typically, the parameter γ controls both the degree of long term correlations and the statistical self-similarity. Mathematical tools and concepts for such processes were first formulated and advocated in practice by Mandelbrot [2] within the context of "fractals". He proposed the well-known fractional Brownian motion (fBm) model to capture the long term correlation and statistical self-similarity of the $1/f$ processes. Given the elaborate fBm model, and the aura of "fractal science", a flurry of activity evolved around the modeling and analysis of $1/f$ processes in engineering literature [3]-[6]. However, these efforts never hold a strong ground in engineering applications, mainly due to the mathematical intractability of the fBm model, and the lack of foundational principles. In [1], Yazıcı et al. proposed a class of second order processes obeying a structure of the form $E[X(t)X(t\lambda)] = R(\lambda), t, \lambda > 0$ to model and analyze $1/f$ processes. These models, referred to as scale stationary, enjoy theoretical properties parallel to the ordinary wide sense stationary processes. Most importantly, their foundation is based on the extensions of the concept of stationarity on which powerful time series analysis tools are derived. Scale stationary processes come with the spectral analysis tools, and ARMA models just like the ordinary stationary processes. They are also directly linked to the linear scale invariant systems. Let us not forget to mention that, fBm model is simply a trended scale stationary model with stationary increments. It may be academically dissapointing! but true that the issue of "statistical self-similarity" can be managed to a large degree by the simple framework of "scale stationarity" .

In [1], authors introduced scale stationary ARMA models based on Euler-Cauchy system and showed that any scale stationary process can be captured by a finite parameter scale stationary autoregressive model. In this study, we extend the ARMA modeling to multiple input and multiple output (MIMO) systems and propose a state space representation for the self-similar processes. At first glance, both the state space representation and the Kalman filter may appear simply as time-varying models. However, with the proper definition of the derivative operation on the multiplicative group and the self-similarity, both the state space model and the Kalman filter are captured with constant matrix vector representation. This new definition of the derivative operation guides the implementation of the Kalman filter for self-similar processes, both in recursive update and the Ricatti equation, leading to superior performance than the ordinary time varying implementation.

The proposed state space representation and the Kalman filtering can be used in estimation, and prediction tasks involving $1/f$ phenomena. Applications include inverse filtering for communication channels and blurred images in which the blur or the channel is time varying and the underlying data and noise have $1/f$ characteristics. Another obvious application of the tool is in communication network traffic prediction which has potential implications in network management and quality service provisioning.

The organization of the paper is as follows: Section 2 covers the basic background on scale stationary processes and scale stationary ARMA modeling. Section 3 presents the state space representation and the derivative operator for functions defined on the multiplicative group. Section 4 introduces the Kalman filter for self-similar processes. Section 5 discusses the implementation of he Kalman filter and the Ricatti Equation and presents some simulation results. Section 6 discusses the applications of the proposed Kalman filter in various engineering problems. Finally, Section 7 concludes the discussion.

2. BACKGROUND ON SELF-SIMILAR PROCESSES

Before giving the derivation of our state space model and Kalman filtering, we like to summarize related background information on self-similar processes as introduced in detail in [1]. A linear system satisfying

$$S\{x(t\lambda)\} = \lambda^{-H} y(t\lambda) \tag{1}$$

is called a Linear Self-Similar (LSS) system with self-similarity parameter H. As it can be seen from this definition, analogous to LTI systems which are invariant to time shifts, LSS systems are invariant to scale changes within a constant parameter.

The output of the LSS system to any input is found by a scale convolution operation defined as:

0-7803-7011-2/01/$10.00 ©2001 IEEE

$$y(t) = \tilde{h}(t) * u(t) = t^H \int_0^\infty \tilde{h}(\frac{t}{\lambda})u(\lambda)dln\lambda, \qquad t0 \qquad (2)$$

where $t^H\tilde{h}(t)$ is the response of the system to the *unit driving force*, $\tilde{\delta}(t)$ [1] defined as: i) $\tilde{\delta}(t) = 0, t \neq 1 \quad t > 0$, ii) $\int_0^\infty \tilde{\delta}(t/\lambda)dln\lambda = 1, t > 0$, iii) $x(t) = \int_0^\infty x(\lambda)\tilde{\delta}(\frac{t}{\lambda})dln\lambda$.

A linear dynamical model for LSS systems is represented as time varying Euler-Cauchy type differential equations:

$$a_N t^N \frac{d^N}{dt_N}y(t) + \cdots + a_1 t\frac{d}{dt}y(t) + a_0 y(t) =$$
$$b_M t^{M+H}\frac{d^M}{dt_M}u(t) + \cdots + b_1 t^{1+H}\frac{d}{dt}u(t) + b_0 t^H u(t) \quad (3)$$

This type of system satisfies the self-similarity definition as in (1). The difference of the Euler-Cauchy system actually comes from the fact that the dynamics of the system is captured in scale derivatives defined on the multiplicative group [10] as:

$$t\frac{d}{dt}y(t) = \underset{\Delta \to 1}{lim} \frac{y(t\Delta) - y(t)}{ln\Delta} \qquad (4)$$

Since the model is invariant to scale changes, the memory of the system is stored in infinitesimal time scalings, similar to the Euler dynamical model for LTI systems where the memory of the system is stored in infinitesimal time shifts since these systems are time invariant.

In a probabilistic setting Euler-Cauchy system generates self-similar processes with $1/f$ spectrum [1]. Using the input-output relationship of the LSS system (2), the power spectrum of the Euler-Cauchy system in Fourier domain driven by a white noise having autocorrelation $R_u(t_1, t_2) = \sigma^2\tilde{\delta}(t_2/t_1)$ is shown to have power-law or a $1/f$ spectrum [1]. Therefore, any $1/f$ process can be approximated with a finite order Euler-Cauchy system which makes signal processing techniques for estimation and prediction of such processes possible. It is because of these facts that we use Euler-Cauchy systems in the derivation of the state space representation and Kalman filtering algorithm for LSS systems.

3. STATE SPACE REPRESENTATION OF SELF-SIMILAR PROCESSES

Beginning from the Euler-Cauchy system in (3), the general state space representation with states having different self-similarity parameters can be obtained as:

$$t\frac{d}{dt}\mathbf{x}(t) = t^{\mathbf{H}}(\mathbf{A} + \mathbf{H})t^{-\mathbf{H}}\mathbf{x}(t) + t^{\mathbf{H}}\mathbf{B}\mathbf{u}(t) \qquad (5)$$
$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \qquad (6)$$

where $\mathbf{x}(t) = [x_1(t)\ x_2(t)\ ...\ x_N(t)]^T$ ($[.]^T$ is the transpose operation) is the $Nx1$ state vector, $\mathbf{u}(t)$ is the $Rx1$ input vector, $\mathbf{y}(t)$ is the $Mx1$ output vector, $\mathbf{A}$ is a NxN matrix, $\mathbf{B}$ is a NxR matrix, $\mathbf{C}$ is a MxN matrix, $\mathbf{D}$ is a MxR matrix and $\mathbf{H}$ is a NxN diagonal matrix having values $H_1, H_2, ..., H_N$ in its diagonal entries.

In this representation, the self-similarity parameters, H_i for $i = 1, ..., N$ of the states can be equivalent, then the external system representation reduces to the Euler-Cauchy system in (3). However, for the states to have same self-similarities is not very realistic and it is a specific case of the general form. Therefore, we use the most general state space representation with the states having different H_is throughout the paper.

It can be argued that the state space representation for LSS systems can be expressed as first order time varying ordinary differential equations $\frac{d}{dt}\mathbf{x}(t) = \frac{t^{\mathbf{H}}(\mathbf{A}+\mathbf{H})t^{-\mathbf{H}}t}{t}\mathbf{x}(t) + \frac{t^{\mathbf{H}}\mathbf{B}}{t}\mathbf{u}(t)$ and time varying state space techniques can be used in their analysis. In this type of representation, the memory of the states are captured in infinitesimal time shifts as in the LTI systems. However, here for the LSS systems, expressing the inner dynamics of the whole system with first order self-similar Euler-Cauchy systems as states is more appropriate to the nature of the dynamics of the system. This is because of the fact that the states are also self-similar in nature therefore, their energy should be stored in infinitesimal time scalings as in (4).

In order to analyze the self-similar dynamics of the states more closely, let us consider the general kth state, $x_k(t)$ whose dynamical equation is:

$$t\frac{d}{dt}x_k(t) = (a_{k,k} + H_k)x_k(t) + t^{H_k}\sum_{\substack{l=1 \\ l\neq k}}^{N} a_{k,l}t^{-H_l}x_l(t) +$$
$$t^{H_k}\mathbf{B}\mathbf{u}(t) \qquad (7)$$

As can be seen from this equation, the dynamics of a state is affected by the state itself, other states and the input depending on the $\mathbf{A}$ and $\mathbf{B}$ matrices. The dependency on the state itself can be seen as an intrinsic self-similarity since the self-similarity parameter appears as a constant gain factor. If there is coupling to the other states, these states can be treated as inputs where the self-similarity of the state is provided with the fractional or self-similar leakage term t^{H_k}. Note here that the self-similarity of the coupled states $x_l(t)$ for $l = 1, ..., N$ and $l \neq k$ with parameters H_l does not have an effect on the self-similarity of the state $x_k(t)$ which guarantees a self-similar first order system for each state $x_k(t)$ with only one self-similarity parameter, H_k.

The solution of the states can be found using the state transition matrix $\Phi(t, \tau)$ as:

$$\mathbf{x}(t) = \mathbf{\Phi}(t, t_1)\mathbf{x}(t_1) + \int_{t_1}^{t} \mathbf{\Phi}(t, \tau)\tau^{\mathbf{H}}\mathbf{B}\mathbf{u}(\tau)dln\tau \qquad (8)$$

where $\mathbf{\Phi}(t, \tau)$ can be obtained using the fundamental matrix $\mathbf{\Phi}(t) = t^{\mathbf{H}}t^{\mathbf{A}}$ which is a solution of the homogeneous state equation in (5) as:

$$\mathbf{\Phi}(t, \tau) = \mathbf{\Phi}(t)\mathbf{\Phi}^{-1}(\tau) = t^{\mathbf{H}}(\frac{t}{\tau})^{\mathbf{A}}(\tau)^{-\mathbf{H}} \qquad (9)$$

Note here that the state transition matrix is also a solution of the homogeneous state equation and it satisfies the same properties as it LTI counterpart [12]. The unit driving force response is found as $\tilde{h}(t, \tau) = t^{\mathbf{H}}(\frac{t}{\tau})^{\mathbf{A}}\mathbf{B}$ using (8). Then the solution for the states and the outputs are:

$$\mathbf{x}(t) = t^{\mathbf{H}}t^{\mathbf{A}}\mathbf{x}(t_1) + t^{\mathbf{H}}\int_{t_1}^{t}(\frac{t}{\tau})^{\mathbf{A}}\mathbf{B}\mathbf{u}(\tau)dln\tau$$
$$\mathbf{y}(t) = \mathbf{C}t^{\mathbf{H}}t^{\mathbf{A}}\mathbf{x}(t_1) + \mathbf{C}t^{\mathbf{H}}\int_{t_1}^{t}(\frac{t}{\tau})^{\mathbf{A}}\mathbf{B}\mathbf{u}(\tau)dln\tau \quad (10)$$

Although each state $x_k(t)$ for $k = 1, ..., N$ is self-similar with self-similarity parameter H_k, depending on the matrix $\mathbf{C}$, the outputs $y_j(t)$ for $j = 1, ..., M$ can be expressed with either one self-similar state or a linear combination of self-similar states with different self-similarity parameters.

4. KALMAN FILTERING

In this section we will investigate the problem of estimating the state variables of a self-similar process by using noisy measurements of the linear combination of the states. Consider the LSS system in state space:

$$t\frac{d}{dt}\mathbf{x}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{w}(t) \qquad (11)$$
$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{v}(t) \qquad (12)$$

where $\mathbf{A}(t) = t^{\mathbf{H}}(\mathbf{A}+\mathbf{H})t^{-\mathbf{H}}$ and $\mathbf{B}(t) = t^{\mathbf{H}}\mathbf{B}$. Equation (11) is the system model where $w(t)$ is the system noise and equation (12) is the measurement model where $v(t)$ is the measurement noise. Both $w(t)$ and $v(t)$ are zero mean white Gaussian noise with covariances:

$$E\{\mathbf{w}(t)\mathbf{w}^T(\tau)\} = \mathbf{Q}(t)\tilde{\delta}(t/\tau)$$
$$E\{\mathbf{v}(t)\mathbf{v}^T(\tau)\} = \mathbf{R}(t)\tilde{\delta}(t/\tau) \quad (13)$$

They are also uncorrelated with each other and the states.

Assuming that $\mathbf{A}(t)$, $\mathbf{B}(t)$, $\mathbf{C}$ and $\mathbf{H}$ are completely known, the state estimate $\hat{\mathbf{x}}(t)$ is obtained by feeding a correction term back to the estimated system depending on the difference between the actual measurement and the estimated measurement as:

$$t\dot{\hat{\mathbf{x}}}(t) = t^{\mathbf{H}}(\mathbf{A}+\mathbf{H})t^{-\mathbf{H}}\hat{\mathbf{x}}(t) + \mathbf{K}(t)(\mathbf{y}(t) - \mathbf{C}\hat{\mathbf{x}}(t)) \quad (14)$$

where $\mathbf{K}(t)$ is the Kalman gain matrix that has to be estimated optimally.

Then the error states between the actual and estimated states $\tilde{\mathbf{x}}(t) = \hat{\mathbf{x}}(t) - \mathbf{x}(t)$ satisfy:

$$t\dot{\tilde{\mathbf{x}}}(t) = (\mathbf{A}(t) - \mathbf{K}(t)\mathbf{C})\tilde{\mathbf{x}}(t) + \mathbf{K}(t)\mathbf{v}(t) - \mathbf{B}(t)\mathbf{w}(t) \quad (15)$$

Using the solution of the error state in terms of its state transition matrix $\mathbf{\Phi}_{\tilde{\mathbf{x}}}(t,\tau)$, the covariance of the error state $\mathbf{P}(t)$ is:

$$\mathbf{P}(t) = E\{\tilde{\mathbf{x}}(t)\tilde{\mathbf{x}}^T(\tau)\} = \mathbf{\Phi}_{\tilde{\mathbf{x}}}(t,t_1)\mathbf{P}(t_1)\mathbf{\Phi}^T_{\tilde{\mathbf{x}}}(t,t_1) + \int_{t_1}^{t} \mathbf{\Phi}_{\tilde{\mathbf{x}}}(t,\tau)\mathbf{K}(\tau)\mathbf{R}(\tau)\mathbf{K}^T(\tau)\mathbf{\Phi}^T_{\tilde{\mathbf{x}}}(t,\tau)dln\tau + \int_{t_1}^{t} \mathbf{\Phi}_{\tilde{\mathbf{x}}}(t,\tau)\mathbf{B}(\tau)\mathbf{Q}(\tau)\mathbf{B}^T(\tau)\mathbf{\Phi}^T_{\tilde{\mathbf{x}}}(t,\tau)dln\tau \quad (16)$$

where $\mathbf{P}(t_1)$ is the initial error covariance matrix at initial time t_1.

For the estimated state to be optimal, the error should be minimized in time via $\mathbf{K}(t)$. In order to find the optimum Kalman filter gain, the cost function $J(t)$ related to the error state as:

$$J(t) = E\{\tilde{\mathbf{x}}^T(t)\tilde{\mathbf{x}}(t)\} = Trace\{\mathbf{P}(t)\} \quad (17)$$

should be minimized in the MMSE sense.

After some manipulations as explained in [11] the Kalman gain matrix that minimizes the cost function is found as:

$$\mathbf{K}(t) = \mathbf{P}\mathbf{C}^T\mathbf{R}^{-1}(t) \quad (18)$$

Then using this $\mathbf{K}(t)$ in (16), the change in the error covariance or the Riccati equation can be obtained as:

$$t\dot{\mathbf{P}}(t) = \mathbf{A}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A}^T(t) + \mathbf{B}(t)\mathbf{Q}(t)\mathbf{B}^T(t) - \mathbf{K}(t)\mathbf{R}(t)\mathbf{K}^T(t) \quad (19)$$

Here, the Kalman filtering algorithm has the same structure as its LTI counterpart. The major difference of our algorithm from the LTI case lies in the state update (14) and error covariance propagation equations (19). Here, the memory is captured in infinitesimal time scalings instead of time shiftings as opposed to the LTI case. Therefore, the self-similar nature of the state estimate and error covariance is satisfied.

5. IMPLEMENTATION AND SIMULATION RESULTS

We simulate a first order LSS system with parameters $H = -0.2$, $A = -0.1$, $B = 0.1$, $C = 1$, $Q = 1$ to test the performance of the proposed Kalman filter.

We generate the $1/f$ data, $x(t)$ via a covariance method that uses Karhunen-Loeve (KL) transform. The autocovariance of $x(t)$ for the first order system given above

$$C_{xx}(t_1,t_2) = \beta(t_1t_2)^{(A+H)}(max(t_1,t_2)^{(-2A)} - 1); \quad (20)$$

where $\beta = B^2Q/(-2A)$. Then using this covariance matrix and KL transform we generate the $1/f$ data, $x(t)$ for $1 < t < 20$. In the Kalman filtering algorithm, the estimated state in continuous time is approximated using the scale derivative definition (4) in geometric time intervals:

$$\hat{x}(\Delta t) = \hat{x}(t) + ln\Delta(A\hat{x}(t) + K(t)(y(t) - C\hat{x}(t))) \quad (21)$$

and the Riccati equation solution:

$$\mathbf{A}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A}^T(t) + \mathbf{B}(t)\mathbf{Q}(t)\mathbf{B}^T(t) - \mathbf{K}(t)\mathbf{R}(t)\mathbf{K}^T(t) = 0 \quad (22)$$

is obtained using the Schur algorithm as given in [13]. The continuous time approximation becomes more accurate when the scale step Δ is selected as close to 1 as possible. Here, in our application we select it as $\Delta = 1.01$.

We test the performance of the Kalman filter for two different SNRs of 20, and 10dBs using 100 Monte Carlo Runs (MCR). The SNR of the signal is calculated as:

$$SNR = var(x)/var(v) \quad (23)$$

A sample data $x(t)$ (solid line), $y(t)$ (dash-dot line) and the estimated data $\hat{x}(t)$ (dashed line) for $SNR = 20$ and 10dBs out of 100 MCR are given in Figure 1 and 2, respectively.

Then the estimation SNR' for each estimated signal $\hat{x}(t)$ is calculated as:

$$SNR' = var(x)/var(x - \hat{x}) \quad (24)$$

For the input $SNR = 20$ and 10dB, the range of estimation SNR's for 100 MCR in each case are found as 7.73dB$< SNR' <$ 8.15dB and 3.02dB$< SNR' <$ 3.69dB where the mean values of them are 7.92dB and 3.36db, respectively.

Let us mention that the proposed Kalman filter also suffers from the same problems as the usual Kalman filter, such as the building up of the "random walk" type error as the prediction time increases. This problem can be overcome with the usage of a backward smoother if the offline processing is possible.

6. APPLICATION AREAS

In this section, we will explain two potential application areas of the proposed Kalman filtering procedure 1) packet arrival estimation in self-similar network traffic and 2) time varying fading channel estimation during self-similar signal transmission in wireless communication applications.

Network traffic studies show that the aggregate of the packet arrival shows the same statistics of long range correlations which decays hyperbolically, the variance of the sample mean decays slowly and their power spectrum obey power law near the origin over different time scales. This observation is apparently valid for Ethernet traffic, ISDN packet networks, signaling (CCSN/SS7) networks for public telephone networks, [7, 8]. If x_1,x_2, x_3, ... denote the

number of arrivals in the first, second,... interval, the aggregate of these arrivals in consecutive, non-overlapping block of m intervals are calculated as follows: Let x_1^m denote the mean arrival rate of the first m intervals $(x_1 + x_2 + ... + x_m)/m$, x_k^m denote $(x_{m(k-1)+1} + ... + x_{km})/m$ and so on. Actually these aggregate processes give the whole arrival process in different time scales and since the aggregate arrival process shows the same long range statistics with slowly decaying variances and self similar characteristic, it is a self-similar process as opposed to the early assumptions of Poisson distribution. Therefore, analysis techniques for traffic density must consider this self-similar nature. Especially, since the buffering requirements for self-similar processes are larger than that are estimated with Poisson processes, the techniques should be selected carefully for the estimation of buffer size. Using the proposed Kalman filtering technique, the self-similar data traffic can be predicted recursively.

Another application area for the proposed Kalman filter is in communications. In present wireless communication applications such as radar, sonar, acoustics, etc., the transmission channel is usually modeled as a multipath fading channel having slowly time-varying characteristics. At the receiver end the transmitted signal through a multipath fading channel is further corrupted by noise. It is an important and a difficult task to deconvolve the original signal from this received data, especially when the transmitted signal and the corruption noise are nonstationary or $1/f$ type. To the best of our knowledge, there is only one work in the literature [9] that solves this problem optimally using a multiscale Wiener filter in wavelet domain. As an alternative, the proposed Kalman filter can be used for the estimation and prediction of the transmitted $1/f$ signal from the observation data in a recursive fashion where no extra steps of wavelet filtering is needed.

7. CONCLUSION

In this paper, we have developed continuous time state space representation and an optimal state estimation algorithm using Kalman filtering for self-similar processes. Beginning from the most general and mathematically tractable dynamical representation such as Euler-Cauchy type differential equation definition of $1/f$ processes, the dynamics of the states are represented with respect to the multiplicative group derivatives where the memory is captured in infinitesimal scalings of time.

Using this state space representation, we formulate the continuous time Kalman filter to estimate or predict the self-similar or $1/f$ data. Although the algorithm appears to be in the same form of LTI systems, the major difference is once again in the memory content or the dynamics of the estimated state and the error covariance matrix which is appropriate to capture the self-similar nature of the statistics.

This work can be extended to several further research areas. Here we assumed that the state space system parameters i.e. $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are available. However this may not be possible in some real time applications, such as in network traffic or in fading channels in communication networks. Therefore, a generalized Kalman filtering technique that estimates and updates the unknown system matrices can further be investigated. This framework can be extended and tested for 2D self-similar signals such as deblurring of textured images.

8. REFERENCES

[1] Birsen Yazici, R. Kashyap, "A Cl Processes for 1/f Phenomena," *IEEE Trans. on Signal Processing*, Vol.45, No. 2, 1997.

[2] B. Mandelbrot, "Some Noises with 1/f Spectrum: A Brass of Second-Order Stationary Self-Similar Processes for 1/f Phenomena," *IEEE Trans. on Signal Processing*, Vol.45, No. 2, 1997.

[3] M. S. Keshner, "1/f Noise," *Proc. IEEE*, Vol. 70, pp. 212-218, 1982.

[4] A. Van Der Ziel, "Unified Presentation of 1/f Noise in Electronic Devices: Fundamental 1/f Noise," *Proc. of IEEE*, Vol. 76, No. 3, 1988.

[5] G. W. Wornell, *Signal Processing with Fractals: A Wavelet-Based Approach*, Prentice Hall, 1996.

[6] M. M. Daniel, A. S. Willsky, " The Modeling and Estimation of Statistically Self-Similar Processes in a Multiresolution Framework," *IEEE Trans. on Information Theory*, Vol. 45, No. 3, pp. 955-970, 1999.

[7] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Trans. On Networking*, Vol. 2, No. 1, pp. 1-15, 1994.

[8] R. Rao, S. Lee, "Self-Similar Traffic Characterization Through Linear Scale Invariant System Models," *Proc. of IEEE International Conf. on Personal Wireless Communication*, pp. 138-142, 2000.

[9] B.S. Chen, Y.C. Chung, D.F. Huang, "Optimal Time-Frequency Deconvolution Filter Design for Nonstationary Signal Transmission Through a Fading Channel: AF Filter Bank Approach," *IEEE Trans. On Signal Processing*, Vol. 46, No. 12, pp. 3220-3234, 1998.

[10] G.S. Chirikjian, A.B. Kyatkin, *Engineering applications of noncommutative harmonic analysis*, CRC Press, 2001.

[11] A. Gelb, *Applied Optimal Estimation*, The M.I.T. Press, 1974.

[12] T. Kailath, *Linear Systems*, Prentice Hall, 1980.

[13] W. F. Arnold, A. J. Laub, "Generalized Eigenproblem Algorithms and Software for Algebraic Riccati Equations," *Proc. of IEEE*, Vol. 72, No. 12, pp. 1746-1754, 1984.

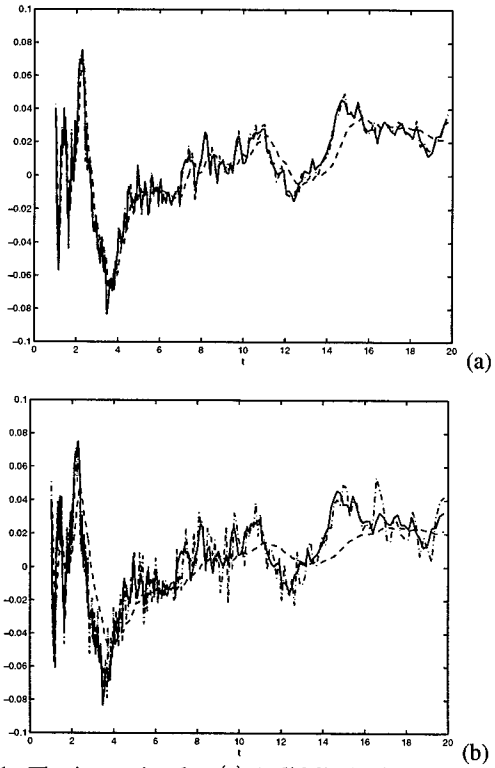


Figure 1: The input signal, $x(t)$ (solid line), the observed signal, $y(t)$ (dash-dot line) and the predicted signal, $\hat{x}(t)$ (dashed line) for a)input $SNR = 20$dB b) input $SNR = 10$dB.

BAYESIAN ARRAY SIGNAL PROCESSING IN ADDITIVE GENERALIZED GAUSSIAN NOISE

B. Kannan

Centre for Wireless Communications,
National University of Singapore,
20, Science Park Road,
Singapore 11764.
kannanb@cwc.nus.edu.sg

ABSTRACT

In this paper, we present a Bayesian approach for DOA and frequency estimation of narrow band signals in additive generalized Gaussian noise. Using Bayesian techniques, the posterior probability densities for DOA (Direction Of Arrival) and frequency parameters are derived from the signal and noise models. These posterior probabilities are then used in the Metropolis-Hastings (M-H) algorithm to derive the samples for the DOA and frequency parameters. The performances of our algorithms are studied by plotting the MSEs (Mean Square Errors) of the parameters for various SNRs. The MSEs of the parameters are compared with the CRLBs (Cramer Rao Lower Bound) for the generalized Gaussian models.

Keywords : Non-Gaussian signal processing, Sensor array processing, Bayesian estimation, M-H algorithm.

1. INTRODUCTION

Sensor array processing has found important applications in many areas such as radar, sonar, communications and seismic explorations. Determination of DOAs and the frequencies of the transmitted signals are two of the main problems in sensor array processing. Various methods have been proposed for estimating DOA and associated parameters for multiple plane-waves signals incident on array of sensors. These include subspace-based methods [5] and maximum likelihood techniques [4]. The signal processing literature has traditionally been dominated by Gaussian noise model assumptions. However, many classes of noise encountered in the real-world such as underwater acoustic noise, low frequency electro-magnetic disturbances [3] and atmospheric noise, exhibit outliers that will not fit into a Gaussian noise model.

In this paper, we will present a Bayesian approach to estimate the DOAs and frequencies of the signals in generalized Gaussian noise. The generalized Gaussian model has been applied successfully to a variety of physical phenomena. For instance, the density estimates of underwater acoustic returns of surface and bottom reverberation bear a strong resemblance to the members of the generalized Gaussian family with a wide range of values of the shape parameter (or decay rate parameter) corresponding to heavy as well as light-tailed distributions.

2. PROBLEM FORMULATION

In this section, we define the signal, noise models and array structure. We assume that M signal sources in the far field transmitting narrow band signals (with a centre frequency f_0) and the received data at a Uniform Linear Array (ULA) of sensors are corrupted by additive noise. The ULA has L ($> M$) sensors with an inter-distance d ($\leq \lambda/2$, where λ is the wavelength of the signal). In this paper, n denotes normalized time (with respect to the sampling interval T_s) and $\Re$ denotes the real part. We can represent the received data$\mathbf{x}(n)$ of M signals in terms of their steering matrix $\mathbf{A}(\boldsymbol{\theta})$, signal vector $\mathbf{s}(n)$ and noise vector $\mathbf{v}(n)$ as

$$\mathbf{x}(n) = \Re\{\mathbf{A}(\boldsymbol{\theta})\mathbf{s}(n)\} + \mathbf{v}(n) \quad \text{for } n = 1, 2, \ldots, N \tag{1}$$

where $\mathbf{x}(n)$ and $\mathbf{v}(n)$ are $L\times 1$ vectors denoting the received signals and the generalized Gaussian noise samples at time n respectively, and $\mathbf{A}(\boldsymbol{\theta})$ is a $L\times M$ matrix

$$\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1), \mathbf{a}(\theta_2), \ldots, \mathbf{a}(\theta_M)] \tag{2}$$

where $\mathbf{a}(\theta_i)$ is a $L\times 1$ steering vector of the ith signal and given by

$$\mathbf{a}(\theta_i) = [1, \exp(j\phi_i), \ldots, \exp(j(L-1)\phi_i)]^T \tag{3}$$

with

$$\phi_i = 2\pi f_0 d \sin(\theta_i)/c \quad \text{for } i = 1, 2, \ldots, M \tag{4}$$

In the above equation, c and θ_i are the speed of wave propagation in the medium and DOA of the ith signal respect ively. $\mathbf{s}(n)$ in (1) is a $M\times 1$ vector

$$\mathbf{s}(n) = [s_1(n), \ldots, s_M(n)]^{\mathrm{T}}; \tag{5}$$

Taking into account all the samples in $\mathbf{x}(n)$, $\mathbf{s}(n)$ and $\mathbf{v}(n)$ for $n = 1, \ldots, N$, we can modify (1) as

$$\mathbf{X} = \Re\{\mathbf{A}(\boldsymbol{\theta})\mathbf{S}\} + \mathbf{V} \tag{6}$$

where $\mathbf{X}$ and $\mathbf{V}$ are $L\times N$ matrices

$$\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(N)]; \tag{7}$$

$$\mathbf{V} = [\mathbf{v}(1), \ldots, \mathbf{v}(N)]; \tag{8}$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

and $\mathbf{S}$ is a $M \times N$ matrix

$$\mathbf{S} = [\mathbf{s}(1), \ldots, \mathbf{s}(N)]; \tag{9}$$

A generalized Gaussian pdf is given [2] by

$$p(v) = \frac{1}{2\sigma\Gamma(1+1/\alpha)B(\alpha)} \exp\left(\left[-\frac{\mid v-\mu \mid}{\sigma B(\alpha)}\right]^{\alpha}\right) \tag{10}$$

where $B(\alpha) = [\Gamma(1/\alpha)/\Gamma(3/\alpha)]^{1/2}$. In this model μ, σ ($>$ 0) and α (> 0) denote the mean, variance and decay rate of the density function, respectively. Smaller values of α correspond to heavier-tailed distributions, which, in turn, are indicative of impulsive noise environments. In this paper, we use only zero mean generalized Gaussian noise models. Assuming that the noise samples are statistically independent from one another both along the array sensors (spatially) and along time (temporally), a likelihood expression for the received data follows from (6) and (10)

$$p_{\mathrm{LGG}}(\mathbf{X} \mid \boldsymbol{\theta}, \mathbf{f}, \sigma, \alpha, M) = (2\sigma\Gamma(1+1/\alpha)B(\alpha))^{(-NL)} \times \prod_{l=1}^{L}\prod_{n=1}^{N} \exp\left(-\left[\frac{\mid \mathbf{X}(l,n) - \Re\{\mathbf{A}(l,:)\mathbf{S}(:,n)\} \mid}{\sigma B(\alpha)}\right]^{\alpha}\right) \tag{11}$$

3. DERIVATION OF BAYESIAN ESTIMATORS

3.1. Priors

Assigning various priors for signal and noise parameters, we use Bayesian principles to define a posterior density. Each DOA is assumed to be uniformly distributed between 0 and π. As we are dealing with narrow band signals with $\mathrm{f_{BW}}$ bandwidth, we assume that the frequency of each signal is also uniformly distributed in the interval [$\mathrm{f_0}$ - $\frac{\mathrm{f_{BW}}}{2}$, $\mathrm{f_0}$ + $\frac{\mathrm{f_{BW}}}{2}$]. Thus, the priors for the DOA vector $\boldsymbol{\theta}$ and frequency vector $\mathbf{f}$ are defined by

$$\begin{aligned} p(\boldsymbol{\theta} \mid M) &= \left(\frac{1}{\pi}\right)^{M} \\ p(\mathbf{f} \mid M) &= \left(\frac{1}{\mathrm{f_{BW}}}\right)^{M} \end{aligned} \tag{12}$$

A non-informative Jeffreys' prior $p(\sigma) = \frac{1}{\sigma}$ and a uniform prior $p(\alpha) = \frac{1}{2}$ are assigned for the parameters σ and α respectively.

3.2. Posterior Density Derivation

When the noise is model led by a generalized Gaussian pdf, a posterior density for the unknown parameters can be obtained from Bayes' theorem as

$$p_{\mathrm{GG}}(\boldsymbol{\theta}, \mathbf{f}, \sigma, \alpha \mid \mathbf{X}, M) \propto p_{\mathrm{LGG}}(\mathbf{X} \mid \boldsymbol{\theta}, \mathbf{f}, \sigma, \alpha, M) p(\boldsymbol{\theta} \mid M) \times p(\mathbf{f} \mid M) p(\sigma) p(\alpha) \tag{13}$$

Substituting (11), (12),$p(\sigma)$ and $p(\alpha)$ into (13), after some manipulation, we obtain

$$p_{\mathrm{GG}}(\boldsymbol{\theta}, \mathbf{f}, \sigma, \alpha \mid \mathbf{X}, M) \propto \sigma^{-(NL+1)} (2\Gamma(1+1/\alpha)B(\alpha))^{(-NL)} \times \exp\left(\sigma^{-\alpha} \sum_{l=1}^{L}\sum_{n=1}^{N} -\left[\frac{\mid \mathbf{X}(l,n) - \Re\{\mathbf{A}(l,:)\mathbf{S}(:,n)\} \mid}{B(\alpha)}\right]^{\alpha}\right) \times (\frac{1}{\pi})^{M} (\frac{1}{\mathrm{f_{BW}}})^{M} \frac{1}{2} \tag{14}$$

The noise parameter σ can be integrated out from (14) as

$$p_{\mathrm{GG}}(\boldsymbol{\theta}, \mathbf{f}, \alpha \mid \mathbf{X}, M) \propto \int_0^{\infty} p_{\mathrm{GG}}(\boldsymbol{\theta}, \mathbf{f}, \sigma, \alpha \mid \mathbf{X}, M)\, d\sigma \tag{15}$$

We can analytically perform this integration using the gamma integral. Thus, the marginalized posterior density is given by

$$p_{\mathrm{GG}}(\boldsymbol{\theta}, \mathbf{f}, \alpha \mid \mathbf{X}, M) \propto (2\Gamma(1+1/\alpha)B(\alpha))^{(-NL)} \Gamma(-NL/\alpha) \times \left(\sum_{l=1}^{L}\sum_{n=1}^{N} -\left[\frac{\mid \mathbf{X}(l,n) - \Re\{\mathbf{A}(l,:)\mathbf{S}(:,n)\} \mid}{B(\alpha)}\right]^{\alpha}\right)^{-NL/\alpha} \times (\frac{1}{\pi})^{M} (\frac{1}{\mathrm{f_{BW}}})^{M} \tag{16}$$

4. MCMC ALGORITHMS FOR PARAMETER ESTIMATION

We use an M-H algorithm [1] to estimate the parameters from the posterior density p_{GG}. We implement the algorithm in three M-H steps as shown below:

- Initialization: Assign initial values to the parameters: $\boldsymbol{\theta}^0$, $\mathbf{f}^0$, α^0.
- Iteration: *for* $i = 1$ to ite_{max}

  1) Update the frequency vector:

  Perform a M-H step with $p_{\mathrm{GG}}(\mathbf{f} \mid \boldsymbol{\theta}^{i-1}, \alpha^{i-1}, \mathbf{X}, M)$ as the invariant density. Sample $\mathbf{f}^n \sim \mathcal{N}(. \mid \mathbf{f}_k^{i-1}, \sigma_f^2)$ and accept with probability:

$$\beta_{\mathbf{f}} = \min\left\{1, \frac{p_{\mathrm{GG}}(\mathbf{f}^n \mid \boldsymbol{\theta}^{i-1}, \alpha^{i-1}, \mathbf{X}, M)}{p_{\mathrm{GG}}(\mathbf{f}^{i-1} \mid \boldsymbol{\theta}^{i-1}, \alpha^{i-1}, \mathbf{X}, M)}\right\} \tag{17}$$

  2) Update the DOA vector:

  Perform a M-H step with $p_{\mathrm{GG}}(\boldsymbol{\theta}^n \mid \mathbf{f}^i, \alpha^{i-1}, \mathbf{X}, M)$ as the invariant density. Sample $\boldsymbol{\theta}^n \sim \mathcal{N}(. \mid \boldsymbol{\theta}^{i-1}, \sigma_\theta^2)$ and accept with probability:

$$\beta_{\boldsymbol{\theta}} = \min\left\{1, \frac{p_{\mathrm{GG}}(\boldsymbol{\theta}^n \mid \mathbf{f}^i, \alpha^{i-1}, \mathbf{X}, M)}{p_{\mathrm{GG}}(\boldsymbol{\theta}^{i-1} \mid \mathbf{f}^i, \alpha^{i-1}, \mathbf{X}, M)}\right\} \tag{18}$$

  3) Update the noise parameter α:

  Perform M-H step with $p_{\mathrm{GG}}(\boldsymbol{\theta}^n \mid \mathbf{f}^i, \alpha^{i-1}, \mathbf{X}, M)$ as the invariant density. Sample $\alpha^n \sim \mathcal{R}(. \mid \alpha^{i-1}, \sigma_\alpha^2)$ and accept with probability:

$$\beta_{\alpha} = \min\left\{1, \frac{p_{\mathrm{GG}}(\alpha^n \mid \mathbf{f}^i, \boldsymbol{\theta}^i, \mathbf{X}, M)}{p_{\mathrm{GG}}(\alpha^{i-1} \mid \mathbf{f}^i, \boldsymbol{\theta}^i, \mathbf{X}, M)} \frac{\alpha^{i-1}}{\alpha^n}\right\} \tag{19}$$

- *end* iteration.

In the above M-H algorithm $\mathcal{N}(.)$ denotes the normal distribution. At the ith iteration, the pdf corresponding to the Rice distribution $\mathcal{R}(.)$ is defined by

$$p_{\mathcal{R}}(z) = \frac{z}{\sigma_\alpha^2} I_0\left(\frac{z\alpha^{i-1}}{\sigma_\alpha^2}\right) \exp\left(-\frac{z^2 + (\alpha^{i-1})^2}{2\sigma_\alpha^2}\right) \quad z > 0 \tag{20}$$

where the random variable $\mathbf{z}=\sqrt{\mathbf{a}^2 + \mathbf{b}^2}$. The random variable $\mathbf{a}$ is normal with mean α^{i-1} and σ_α^2 variance, and $\mathbf{b}$ is normal with zero mean and σ_α^2 variance. $I_0(.)$ in the above pdf is the modified Bessel function. When $\alpha^{i-1} = 0$, the above pdf is equivalent to a Rayleigh density.

Multivariate normal distributions are used as the proposal distributions for updating frequencies and DOAs. The mean values of these normal densities are the corresponding DOA and frequency estimates at the previous iteration. The invariant/target density at each M-H step can be formulated from the posterior density $p_{\mathbf{C}}$.

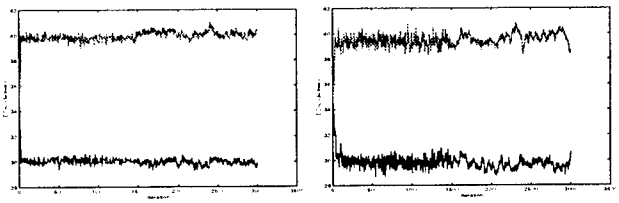
Similarly, we use normal distributions as the proposal distributions for DOAs and frequencies in the generalized Gaussian noise case. The noise parameter α is sampled from $\mathcal{R}(.)$. The M-H algorithm for the generalized Gaussian noise case is similar to the above algorithm with minor differences. In the above M-H algorithms, σ_f^2, σ_θ^2 and σ_α^2 are the variances of the corresponding proposal distributions. The constant ite_{max} denotes the maximum number of iterations. One must carefully choose these variances to obtain an algorithm with good mixing properties.

5. SIMULATIONS AND DISCUSSION

To test the performance of the Bayesian estimators in generalized Gaussian noise models, several experiments were designed. As the first sensor of the array was used as the reference element, the phases of the received data at the first sensor did not depend on the DOAs of the transmitted signals. Hence, the data at the first sensor was used to obtain initial estimates for the frequencies of the transmitted signals. This helped to ensure a fast convergence. In most of the experiments discussed below, the estimate of an unknown parameter was obtained by taking the mean of the corresponding samples after the M-H algorithm converged. Then, the MSE of an unknown parameter was computed by forming the mean of the MSEs from 50 Monte Carlo runs. In the following experiments, $M = 2$ (exponential signals), $d = 1$ m, $c = 300$ m/s $N = 64$ and the sampling frequency $f_s = 200$ Hz.

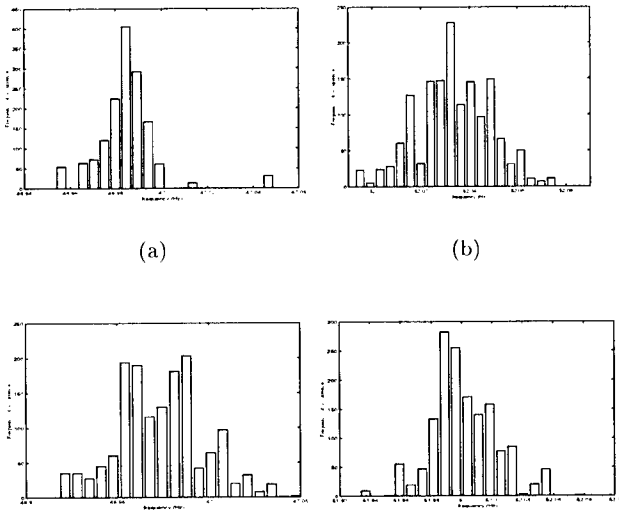
5.1. Performance and Convergence Properties of the Generalized Gaussian Estimator

In order to study the convergence properties of our algorithm, we used our algorithm to estimate the unknown parameters at smaller values of α corresponding to heavier-tailed distributions. In this experiment, $L = 8$ and the DOAs and the frequencies of the signals are [30° 40°] and [47 Hz 52 Hz] respectively. The number of iterations (ite_{max}) was 3,000. Fixing the value of σ at 0.5, the signal and noise parameters were evaluated at two different values of α: 0.8 and 1.3. Figure 1 shows the evolution of both DOAs with iteration number. We can see from these figures that the M-H sampler converged to the target DOA values, 30° and 40°, within 200 iterations. At the 1500th iteration we reduced the values of the variances of the proposal densities of DOAs. Consequently, the mixing properties of our sampler changed after the 1500th iteration (see figure 1). The histograms of the frequencies and α are given in figures 2 and 3 respectively. As expected, the samples are centred around the target frequencies, 47 Hz and 52 Hz, and the target α, 0.8 and 1.3. To study the performance of the Bayesian estimator in generalized Gaussian noise, the value of σ was varied from 0.1 to 1 while fixing the value of α at 1.5. For each σ, the number of iterations (ite_{max}) was 3,000. The MSEs of $\boldsymbol{\theta}$, $\mathbf{f}$ and α were computed from 50 Monte Carlo runs. The figures 4(a) and 4(b) show the MSEs and CRLBs of the DOAs and frequencies respectively. As expected, the MSE decreased with increasing SNR and approached the



(a) (b)

Figure 1: Evolution of DOAs with iteration number: (a) α = 0.8, (b) α = 1.3

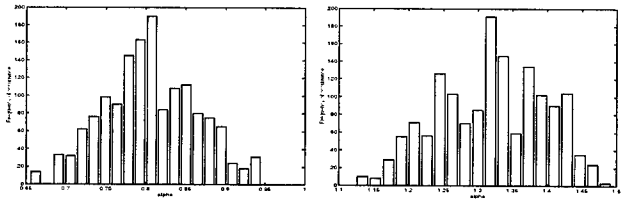


(a) (b)

(c) (d)

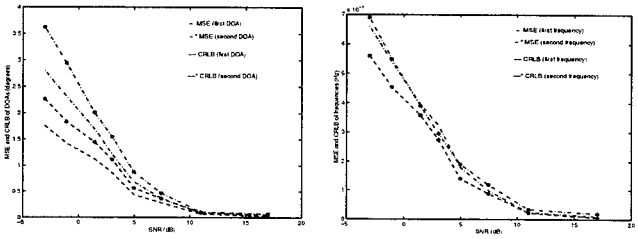
Figure 2: Histograms of Frequencies: (a) α = 0.8 (1st source), (b) α = 1.3 (2nd source), (c) α = 0.8 (1st source), (d) α = 1.3 (2nd source)

CRLB at high SNRs.



(a) (b)

Figure 3: Histograms of α: (a) $\alpha = 0.8$, (b) $\alpha = 1.3$



(a) (b)

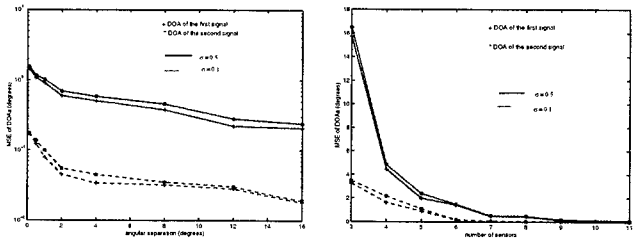
Figure 4: MSEs of the (a) DOAs, (b) frequencies in generalized Gaussian noise

5.2. Resolution Capabilities of the Estimators

In this section we analyzed the resolution capabilities of our generalized Gaussian estimator. We used two exponential signals whose frequencies were 50 Hz and assumed to be known. The data length (N), inter-sensor spacing (d), speed of the waveform (c) and sampling frequency (f_s) were as same as in the last experiment.

In the first experiment the DOA of the second signal was varied while the DOA of the first signal was kept constant at 10°. The number of sensors used in this experiment was 5. We did 50 Monte Carlo runs for each DOA of the second signal, running the M-H algorithm for 3,000 iterations. The MSEs of both DOAs were estimated using the generalized Gaussian estimators and the results are plotted against the angular separation in figure 5(a). As expected, the MSEs of both DOAs decreased with increasing angular separation. However, these figures show that for well-separated DOAs ($> 8°$), the resolution of the DOAs was only limited by the amount of noise.

In the second experiment we studied how the MSEs of the DOAs were affected by the array size (number of sensors). The DOAs of the signals were 10° and 15°. We did 50 Monte Carlo runs for each array size, running the M-H algorithm for 3,000 iterations. The MSEs of the DOAs were estimated using the generalized Gaussian estimators and the results are displayed in figure 5(b). As expected the MSEs of the DOAs decreased with increasing array size. It is also obvious from these figures that the resolution of the DOAs was affected more by the amount of noise than the array size (when $L > 7$).



(a) (b)

Figure 5: MSEs of the DOAs vs (a) angular separation, (b) array size

6. CONCLUSIONS

In this chapter we developed a Bayesian estimator for a generalized Gaussian noise model. We have shown that the MSEs of the estimates approached the CRLBs at high SNRs. Our simulation results demonstrated that our algorithm converged within 200 iterations.

Studying the resolution capabilities of our Bayesian estimator, we showed that the MSEs of the DOAs decreased with increasing angular separation and array size. However, our simulations showed that for well-separated DOAs ($> 8°$) or for a large array size ($L > 7$), the resolution of the DOAs was only limited by the amount of noise.

7. REFERENCES

[1] J. J. K. Ó Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian methods applied to signal processing.* Springer-Verlag, New York, Inc., 1996.

[2] M. K. Varanasi and B. Aazhang. Parametric generalised Gaussian density estimation. *J. Acoustic Soc. Am.*, 86(4):1404–1415, October 1989.

[3] E. J. Wegman, S. C. Schwartz, and J. B. Thomas. *Topics in non-Gaussian signal processing.* Springer-Verlag, 1989.

[4] K. M. Wong, J. P. Reilly, Q. Wu, and S. Qiao. Estimation of the direction of signals in unknown correlated noise, Part I: The MAP approach and its implementation. *IEEE Transactions on Signal Processing*, 40(8):2007–2017, August 1992.

[5] N. Yuen and B. Friedlander. Asymptotic Performance Analysis of ESPRIT, Higher Order ESPRIT and Virtual ESPRIT Algorithms. *IEEE Transactions on Signal Processing*, 44:2537–2550, October 1996.

NONLINEAR IMAGE FILTERING IN A MIXTURE OF GAUSSIAN AND HEAVY-TAILED NOISE

A. Ben Hamza *Hamid Krim*

Department of Electrical and Computer Engineering
North Carolina State University, Raleigh, NC 27695-7914, USA
E-mails: {abhamza, ahk}@eos.ncsu.edu

ABSTRACT

Inspired by robust estimation, nonlinear denoising methods combining the mean, the median, and the LogCauchy filters are proposed. Some statistical and asymptotic properties are studied, and comparisons with other nonlinear filtering schemes are performed. Experimental results showing a much improved performance of the proposed filters in the presence of Gaussian and heavy-tailed noise are analyzed and illustrated.

1. INTRODUCTION

A variety of models have been sources in modeling impulsive noise inluding the Laplacian model whose distribution has heavier tails than the Gaussian. Examples of impulsive noise include atmospheric noise, cellular communication, underwater acoustics, and moving traffic. Recently, it has been shown that α-stable ($0 < \alpha \leq 2$) distributions can approximate impulsive noise more accurately that other models [1]. The parameter α controls the degree of impulsiveness (heaviness of the tails), and the impulsiveness increases as α decreases. The Gaussian ($\alpha = 2$) and the Cauchy ($\alpha = 1$) distributions are the only *symmetric* α-stable distributions which have closed-form probability density functions. The two most important properties of α-stable distributions are the *stability property* and the *Generalized Central Limit Theorem* [1].

It is also known that in the presence of only Gaussian noise, the efficiency of a median filter leaves room for much improvement relative to that of a mean filter [2]. This led to a number of other proposed nonlinear schemes to attain a balance between the two. Among these proposed filters, figure Wilcoxon and Hodges-Lehmann filters [2].

Approaches to wavelet-based denoising have generally relied on the assumption on Gaussian noise, and are therefore sensitive to outliers, i.e., to noise distributions whose tails are heavier than the Gaussian distribution, such as Laplacian distribution. For independent ϵ-contaminated Gaussian distributions of the wavelet coefficients, Krim and Schick [4] derive a robust estimator of the wavelet coefficients based on minimax description length.

In the next section, we provide a brief review of Huber minimax approach, some basic sliding window filters and symmetric α-stable ($S\alpha S$) distributions. In Section 3, a nonlinear filtering structure called *Mean-Median* filter is introduced and its asymptotic analysis is performed. Section 4 is devoted to another class of nonlinear denoising techniques called *Mean-LogCauchy* filters. Finally, in Section 5, we provide experimental results to show a much improved performance of the proposed filters at removing noise from images corrupted by ϵ-contaminated Gaussian and heavy tailed noise, while preserving well image structures.

This work was supported by an AFOSR grant F49620-98-1-0190 and by ONR-MURI grant JHU-72798-S2 and by NCSU School of Engineering.

2. BACKGROUND

Consider the additive noise model

$$X_{\boldsymbol{i}} = S_{\boldsymbol{i}} + V_{\boldsymbol{i}}, \qquad \boldsymbol{i} \in \mathbb{Z}^m, \tag{1}$$

where $\{S_{\boldsymbol{i}}\}$ be a discrete m-dimensional deterministic sequence corrupted by the zero-mean noise sequence $\{V_{\boldsymbol{i}}\}$, and $\{X_{\boldsymbol{i}}\}$ is the observed sequence. The objective is to estimate the sequence $S_{\boldsymbol{i}}$ based on a filtering output $Y_{\boldsymbol{i}} = \mathcal{F}(X_{\boldsymbol{i}})$, where $\mathcal{F}$ is a filtering operator.

Here, we assume that the noise probability distribution is a scaled version of a *known* member of the family of ϵ-contaminated normal neighborhood proposed by Huber [3]

$$\mathcal{P}_\epsilon = \{(1-\epsilon)\Phi + \epsilon H : \quad H \in \mathcal{S}\},$$

where Φ is the standard normal distribution, $\mathcal{S}$ is the set of all probability distributions symmetric with respect to the origin (i.e. such that $H(-x) = 1 - H(x)$) and $\epsilon \in [0,1]$ is the known fraction of "contamination". The presence of outliers in a nominally normal sample can be modeled here by a distribution H with tails heavier than normal. Note that symmetry ensures the unbiasedness of the maximum likelihood estimator, making the expression for its asymptotic variance considerably simpler. Krim and Schick [4] proposed a robust wavelet thresholding technique based on the minimax description length (MMDL) principle, determining the least favorable distribution in $\mathcal{P}_\epsilon$ family as the member that maximizes the entropy. The MMDL approach results in a thresholding scheme that is resistant to heavy-tailed noise.

Let W be a sliding window of size $2N+1$. Define $W_{\boldsymbol{i}} = \{X_{\boldsymbol{i}+\boldsymbol{r}} : \boldsymbol{r} \in W\}$ to be the window centered at location $\boldsymbol{i}$. The output of the mean filter is given by

$$Y_{\boldsymbol{i}} = \overline{W_{\boldsymbol{i}}} = \operatorname*{argmin}_\theta \sum_{\boldsymbol{r}\in W} (X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2. \tag{2}$$

where $\overline{W_{\boldsymbol{i}}}$ is the sample mean of the window $W_{\boldsymbol{i}}$.
Denote by $[W_{\boldsymbol{i}}]_{(k)}$ the k-th order statistic of the samples in $W_{\boldsymbol{i}}$, that is $[W_{\boldsymbol{i}}]_{(1)} \leq [W_{\boldsymbol{i}}]_{(2)} \leq \cdots \leq [W_{\boldsymbol{i}}]_{(2N+1)}$.
The output of the standard median (SM) filter is given by

$$Y_{\boldsymbol{i}} = [W_{\boldsymbol{i}}]_{(N+1)} = \operatorname*{argmin}_\theta \sum_{\boldsymbol{r}\in W} |X_{\boldsymbol{i}+\boldsymbol{r}} - \theta|. \tag{3}$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

Such estimators are well founded and well known for a Gaussian and Laplacian distributions. Note that the mean and median filters are the maximum likelihood estimators of the location parameter for the Gaussian and Laplacian distributions, respectively.

The general class of α-stable distributions has also been shown to accurately model heavy-tailed noise [1]. A symmetric α-stable ($S\alpha S$) random variable is however only described by its characteristic function

$$\varphi(t) = \exp(j\theta t - \gamma|t|^\alpha),$$

where $j \in \mathbb{C}$ is the imaginary unit, $\theta \in \mathbb{R}$ is the location parameter (centrality), $\gamma \in \mathbb{R}$ is the dispersion of the distribution and $\alpha \in (0, 2]$ which controls the heaviness of the tails, is the characteristic exponent [1].

When $\alpha \in (0, 2)$, an $S\alpha S$ random variable has infinite variance, and the Cauchy ($\alpha = 1$) is the only distribution which has a closed-form for the probability density function. This is in fact useful when using the principle of maximum likelihood estimation.

The LogCauchy (LC_γ) filter [5] is the maximum log-likelihood estimator of the location parameter for a Cauchy density, and yields the following

$$Y_{\boldsymbol{i}} = \mathrm{LC}_\gamma(W_{\boldsymbol{i}}) = \arg\min_\theta \sum_{\boldsymbol{r}\in W} \log\left(\gamma^2 + (X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2\right), \quad (4)$$

where γ is the dispersion, and θ is the estimation parameter.

3. THE MEAN-MEDIAN FILTER

From Eqs. (2) and (3), it can easily be seen that the mean filter is optimal for Gaussian noise in the sense of mean square error while the standard median filter for Laplacian noise in the sense of mean absolute error. Assume that the noise probability distribution P is a scaled version of a member of $\mathcal{P}_\epsilon$, i.e. $P = (1-\epsilon)G + \epsilon L$, where G is Gaussian $\mathcal{N}(0, \sigma_G^2)$ with variance σ_G^2, and L is Laplacian (or double-exponential) $\mathcal{L}(0, \sigma_L^2)$ with variance σ_L^2 (clearly $L \in \mathcal{S}$). This assumption on the noise to be ϵ-contaminated Gaussian and Laplacian distributed is motivated by the fact that heavier tails than the Gaussian mixture are provided by the Laplace distribution, which is used as a contaminant of the Gaussian distribution. A convex combination of the mean and the median filters can be defined as follows.

Definition 1 *The output of the Mean-Median (*MEM*) filter is given by*

$$Y_{\boldsymbol{i}} = (1-\lambda)\overline{W_{\boldsymbol{i}}} + \lambda[W_{\boldsymbol{i}}]_{(N+1)}, \quad \lambda \in [0, 1].$$

As a suitable performance measure for a robust estimator, Huber suggests its asymptotic variance since the sample variance is strongly dependent on the tails of the distribution. Indeed, for any estimator whose value is always contained within the convex hull of the observations, the supremum of its actual variance is infinite. For this and other reasons, the performance of the mean-median filter is carried out using its asymptotic variance.

The asymptotic variance $V(T, F)$ of an estimator T at the distribution F is then given by [3]

$$V(T, F) = \int IF(x; T, F)^2 dF(x), \quad (5)$$

where $IF(x; T, F)$ is the influence function of T at F defined as

$$IF(x; T, F) = \lim_{t\to 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t},$$

at all points x where the limit exists, and Δ_x stands for delta distribution function, i.e. with unit mass at x. The influence function gives the effect of an infinitesimal perturbation to the data at the point x.

It can be shown that the influence function of the mean and the median filters are given by [3]

$$IF(x; \overline{W_{\boldsymbol{i}}}, F_\theta) = x - \theta,$$

and

$$IF(x; [W_{\boldsymbol{i}}]_{(N+1)}, F_\theta) = \frac{sign(x-\theta)}{2f(\theta)}.$$

Then it follows that the influence function of the MEM filter is given by

$$IF(x; \mathrm{MEM}, F_\theta) = (1-\lambda)(x-\theta) + \lambda\frac{sign(x-\theta)}{2f(\theta)}. \quad (6)$$

Using (5) and (6), the following result holds.

Proposition 1 *The asymptotic variance $V(\mathrm{MEM}, F_\theta)$ of the MEM filter at the distribution F*

$$V(\mathrm{MEM}, F_\theta) = (1-\lambda)^2\mu_2 + \frac{\lambda^2}{4f(\theta)^2} + \lambda(1-\lambda)\frac{\mu_1}{f(\theta)}, \quad (7)$$

where $\mu_k = E|X-\theta|^k, k = 1, 2.$

Remark: While the independence assumption of the filter input simplifies the tractability of the problem, it is not strictly valid.

Minimizing (7) over λ, we obtain the minimum attainable asymptotic variance, and the filter attaining that minimum asymptotic variance will then provide the best filtering performance.

Corollary 1 *The minimum value of $V(\mathrm{MEM}, F_\theta)$ is attained at $\lambda_{\min}$ given by*

$$\lambda_{\min} = \left(\mu_2 - \frac{\mu_1}{2f(\theta)}\right) \Big/ \left(\mu_2 + \frac{1}{4f(\theta)^2} - \frac{\mu_1}{f(\theta)}\right). \quad (8)$$

Example: If the input is i.i.d. $\mathcal{N}(\theta, \sigma^2)$, then using (8), we obtain $\lambda_{\min} \approx 2/(2+\pi)$.

4. MEAN-LOGCAUCHY FILTERS

The LogCauchy filter has been shown to outperform the standard median filter in removing highly α-stable noise [5], then the MEM filter can be improved replacing the median by the LogCauchy, and therefore a new class of nonlinear filters is derived.

Now we assume that the noise probability distribution P is a scaled version of a member of $\mathcal{P}_\epsilon$ such that $P = (1-\epsilon)G + \epsilon S$, where G is Gaussian $\mathcal{N}(0, \sigma_G^2)$ and S is $S\alpha S$ with location parameter θ and dispersion γ_S. The parameter α controls how impulsive the distribution is.

Suppose that G and S are the cumulative distribution functions of two independent random variables X_G and X_S respectively, then the characteristic function φ_ϵ of the random variable $(1-\epsilon)X_G + \epsilon X_S$ is given by

$$\varphi_\epsilon(t) = \exp\left(j\epsilon\theta t - (1-\epsilon)^2\frac{\sigma_G^2}{2}t^2 - \epsilon^\alpha\gamma_S|t|^\alpha\right), \ \epsilon \in [0, 1]$$

For $\alpha \in (1, 2]$, all $S\alpha S$ random variables have finite mean given by their location parameter θ. Moreover, it is shown in [6]

that an $S\alpha S$ distribution with zero mean can be approximated by a finite-Gaussian mixture. Assuming that S is zero mean $S\alpha S$ $(1 < \alpha \leq 2)$, then $P = (1-\epsilon)G + \epsilon S$ can be approximated by a finite-Gaussian mixture, and hence the noise model (1) becomes an ϵ-contaminated Gaussian mixture noise model.

For $\alpha \in (0, 1]$, all $S\alpha S$ random variables have a median and the only $S\alpha S$ distribution having closed-form probability density function is Cauchy distribution ($\alpha = 1$), thus the maximum log-likelihood principle can be applied to derive (4). A convex combination of the mean and the LogCauchy filters can then be defined as follows.

Definition 2 *The output of Mean-LogCauchy* (MLC_γ) *filter with parameter* γ *is given by*

$$Y_{\boldsymbol{i}} = \mathrm{MLC}_\gamma(W_{\boldsymbol{i}}) = (1-\lambda)\overline{W_{\boldsymbol{i}}} + \lambda\,\mathrm{LC}_\gamma(W_{\boldsymbol{i}}),\ \lambda \in [0,1], \quad (9)$$

where γ *is the dispersion of a Cauchy distribution.*

The output of the LogCauchy filter is defined as a solution of the following maximum log-likelihood estimation problem

$$\begin{aligned}\hat{\theta}_{\boldsymbol{i}} &= \operatorname*{argmax}_\theta \ell_\gamma(\theta; W_{\boldsymbol{i}}) \\ &= \operatorname*{argmax}_\theta \log \prod_{\boldsymbol{r}\in W} \frac{\gamma}{\pi}\left(\frac{1}{\gamma^2 + (X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2}\right), \quad (10)\end{aligned}$$

where $\ell_\gamma(\theta; W_{\boldsymbol{i}})$ is the log-likelihood function of a Cauchy distribution $\mathcal{C}(\gamma, \theta)$.

It is clear that for a given γ, solving (10) is equivalent to minimizing the function $\rho_\gamma(\theta; W_{\boldsymbol{i}})$ given by

$$\rho_\gamma(\theta; W_{\boldsymbol{i}}) = \prod_{\boldsymbol{r}\in W}\left(\gamma^2 + (X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2\right), \quad (11)$$

as well as to solving the problem (4) since the $\log(\cdot)$ function is strictly monotone. Thus the minimum of (4) is attained at the same place as that of $\rho_\gamma(\theta; W_{\boldsymbol{i}})$. This is very important because $\rho_\gamma(\theta; W_{\boldsymbol{i}})$ is a polynomial of degree $2(2N+1)$ in θ and its characteristics can then be obtained easily. It can be shown that $\rho_\gamma(\theta; W_{\boldsymbol{i}})$ is a convex function of θ if $\gamma \geq [W_{\boldsymbol{i}}]_{(2N+1)} - [W_{\boldsymbol{i}}]_{(1)}$, and therefore has a unique minimum $\theta_0 \in [[W_{\boldsymbol{i}}]_{(1)}, [W_{\boldsymbol{i}}]_{(2N+1)}]$. At $\gamma = 0$, the function $\rho_\gamma(\theta; W_{\boldsymbol{i}})$ has distinct minima at all the points $X_{\boldsymbol{i}+\boldsymbol{r}}$. If γ is increased, the number of minima decreases. After a certain limit of γ, there is only a unique minimum.

Proposition 2 *When* $\gamma \to \infty$, *the Mean-LogCauchy filter becomes the mean filter, i.e.*

$$\mathrm{MLC}_\gamma(W_{\boldsymbol{i}}) \to \overline{W_{\boldsymbol{i}}} \quad as \quad \gamma \to \infty.$$

Proof. Using basic properties of the argmin function, the output of the LogCauchy filter can be expressed as

$$\begin{aligned}\mathrm{LC}_\gamma(W_{\boldsymbol{i}}) &= \operatorname*{argmin}_\theta \sum_{\boldsymbol{r}\in W} \log\left(\gamma^2 + (X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2\right) \\ &= \operatorname*{argmin}_\theta \sum_{\boldsymbol{r}\in W} \gamma^2 \log\left(1 + \frac{(X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2}{\gamma^2}\right) \\ &= \operatorname*{argmin}_\theta \sum_{\boldsymbol{r}\in W} \log\left(1 + \frac{(X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2}{\gamma^2}\right)^{\gamma^2}\end{aligned}$$

Since

$$\lim_{\gamma\to\infty} \log\left(1 + \frac{(X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2}{\gamma^2}\right)^{\gamma^2} = \exp\left\{(X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2\right\},$$

and the exponentionl function $\exp\{\cdot\}$ is monotonically increasing, it follows that

$$\mathrm{LC}_\gamma(W_{\boldsymbol{i}}) \to \operatorname*{argmin}_\theta \sum_{\boldsymbol{r}\in W} (X_{\boldsymbol{i}+\boldsymbol{r}} - \theta)^2 \quad \text{as} \quad \gamma \to \infty.$$

This concludes the proof using (2) and (9). ■

Note that asymptotically, the tuning parameter γ transforms a nonlinear filter to a linear one.

5. EXPERIMENTAL RESULTS

This section presents simulation results where the proposed filters are applied to enhance images corrupted by mixed Gaussian and heavy tailed noise. The performance of a filter clearly depends on the filter type and its sliding window size, the properties of signals/images, and the characteristics of the noise. The choice of criteria by which to measure the performance of a filter presents certain difficulties. In particular, it is clear that a global performance measure such as the mean square error only gives a partial picture of reality: for instance, one filter may do vary well at the nominal model but badly at an outlier, while another do poorly at the nominal model but well at an outlier, and yet the two could have the same mean square value. Another important performance measure in the mean absolute error which is obviously tend to downplay the influence of large errors, compared to mean square error precisely in the presence of heavy-tailed noise.

Mean square error (MSE) between the filtered and the original image is evaluated to quantitatively compare the good performance of the proposed filters with other filtering techniques.

The scale-contaminated Gaussian and Laplace distributions are relatively light tailed. The $S\alpha S$ distributions are very heavy-tailed noise distributions. The Cauchy distribution is a member of this family ($\alpha = 1$), whose variance is infinite. To assess the performance of Mean-LogCauchy filters in mixed noise, the original image in Fig. 1(a) was contaminated by both Gaussian white noise ($\sigma^2 = 100$) and α-stable noise $S\alpha S(\alpha = 0.5)$. The ϵ-contaminated mixed noise corrupted image is shown in Fig. 1(b). The visual comparison with other techniques is shown in Fig. 1. The relaxed median filter [7] outperforms Wilcoxon and Hodges-Lehmann in suppressing highly α-stable noise, while the Mean-LogCauchy filter, with mixture parameter $\lambda = \pi/(2+\pi)$ and optimal tuning parameter $\gamma = 2.38$, achieves the best performance. In the simulation results of Fig.1, the contamination fraction ϵ is chosen to be equal to λ.

The high sensitivity of many specific filters to an accurate modeling of noise that is to be removed led us to investigate the proposed new techniques that include a number of filters whose optimality when given a specific noise distribution is attained by merely adjusting or optimizing the parameter λ. On the other hand, the filtering performance is also sensitive to the fraction of contamination ϵ. When $\epsilon = 0$ the mixed noise is purely Gaussian, and when $\epsilon = 1$ it is purely α-stable. Fig. 2 shows the influence of the parameter ϵ on the filtering performance.

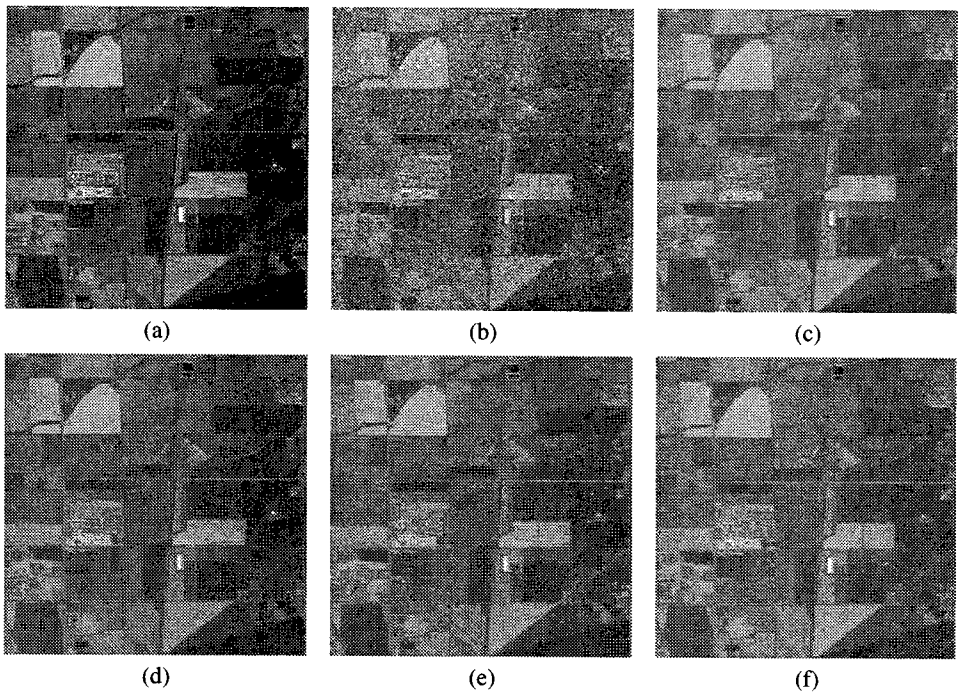


Fig. 1. Filtering results in the presence of ϵ-contaminated Gaussian and α-stable noise, and using a 3×3 square window: (a) Original image, (b) ϵ-mixed noisy image with $\mathcal{N}(0, 100)$ and $S\alpha S$, (c) Output of the MLC filter, $\lambda = 2/(2+\pi)$, (d) Output of the relaxed median filter, (e) Output of the Wilcoxon filter, and (f) Output of the Hodges-Lehmann filter.

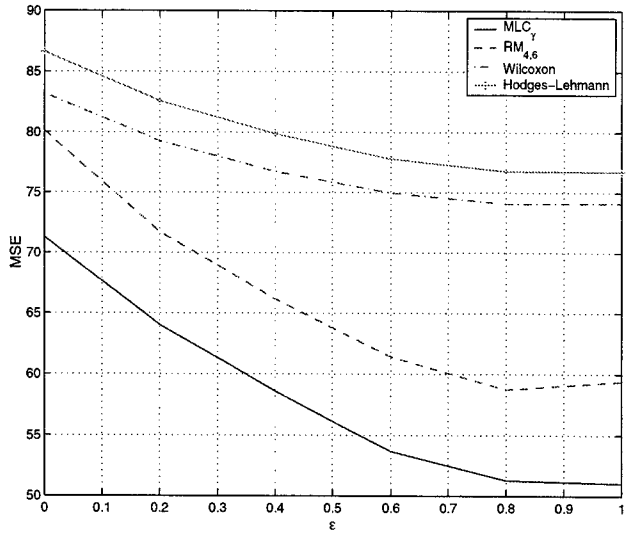


Fig. 2. Influence of the contamination fraction ϵ on filtering performance: MSE vs. ϵ.

6. REFERENCES

[1] M. Shao and C.L. Nikias, "Signal Processing with fractional lower order moments: Stable processes and their applications," *Proceedings of the IEEE*, vol. 81, no. 7, pp. 986-1010, July 1993.

[2] J. Astola and P. Kuosmanen, *Fundamentals of Nonlinear Digital Filtering*, CRC Press LLC, 1997.

[3] P. Huber, *Robust Statistics*, John Wiley, 1981.

[4] H. Krim and I.C. Schick, "Minimax description length for signal denoising and optimized representation," *IEEE Trans. Information Theory*, vol 45, no. 3, pp. 898-908, April, 1999.

[5] S. Ambike and D. Hatzinakos, "A new filter for highly impulsive α-stable noise," *Proc. 1995 Int. Workshop Nonlinear Signal Image Processing*, Greece, 1995.

[6] E.E. Kuruoglu, C. Molina, S.J. Gosdill and W.J. Fitzgerald, "A new analytic representation of the α-stable density function," *American Statistical Society Proceedings*, 1997.

[7] A. Ben Hamza, P.Luque, J.Martinez, and R. Roman "Removing noise and preserving details with relaxed median filters," *Journal of Mathematical Imaging and Vision*, vol. 11, no. 2, pp. 161-177, October 1999.

A MIXED-COST BLIND ADAPTIVE RECEIVER FOR DS-CDMA

Peerapol Yuvapoositanon and Jonathon A. Chambers

Signal and Image Processing Group,
Department of and Electronic and Electrical Engineering,
University of Bath,
Bath BA2 7AY, United Kingdom
Email: eeppay@bath.ac.uk

ABSTRACT

A new mixed-cost receiver for direct-sequence code-division multiple access (DS-CDMA) systems is proposed. An adaptive mixing function is introduced to combine the constrained minimum output energy (CMOE) and constant modulus (CM) criteria together. Simulations confirm the near-far resistance of the proposed receiver over a wide range of near-far situations.

1. INTRODUCTION

The constrained minimum output energy (CMOE) criterion [1, 2] is widely known as an effective interference cancellation scheme for code-division multiple access (CDMA) systems. This feature is emphasised when the channel exhibits a *near-far* environment: the situation when one or more interfering users have greater power than the desired user. The performance of the CMOE receiver degrades, however, in high signal to noise ratio (SNR) situations and by distortion of the received signals due to multipath fading channels [1]. In [2], the constraint proposed in [1] is replaced by a code constraint matrix to retain the output energy of the desired user at a particular path delay. Although this new scheme prevents the cancellation of the desired signal and sidesteps the use of an explicit constraint on the orthogonal vector [1], the performance of the CMOE receiver still degrades either in the case of low interference power or when the number of multipaths is extended [3].

The constant modulus algorithm (CMA) receiver performs better in (inverse) channel estimation and provides near-Weiner receiver performance [4]. However, since its cost surface is multimodal, the CM criterion possibly possesses some undesirable local minima which in some cases associate to the solutions of interfering users. Good initialisation for a CMA receiver can help evade these local minima and accelerates the convergence of the algorithm. In severe near-far environments, a pre-whitening process of the received signal is indispensable despite its excessive computational complexity [5].

This paper concerns the exploitation of the salient features of both criteria to produce a near-far resistant receiver which can be operated in multipath fading channels with a wide-range of near-far levels. The proposed algorithm jointly updates the receiver weight vector by adaptively minimising a mixed-cost function. The mixing parameter is also adapted according to the near-far level. Simulation results are provided to show the signal to interference plus noise ratio (SINR) performance of the proposed combining scheme compared to those of the existing algorithms. It is shown that the mixed-cost scheme is superior in terms of SINR levels over a wide-range of near-far levels in multipath fading channels.

2. SYSTEM MODEL

For the real system model, the baseband received signal for a K-user asynchronous CDMA channel is defined as

$$r(t) = \sum_{i=-\infty}^{\infty} \sum_{k=1}^{K} A_k b_k[i] c_k(t - iT - \tau_k) + v(t), \quad (1)$$

where A_k represents the received amplitude of the kth user. The data bits $b_k[i]$ are independent identically distributed (i.i.d.) and assumed to be drawn from the finite alphabet $\{-1, +1\}$. The symbol period is denoted by T. The spreading (or signature) waveform of the kth user $c_k(t)$ is L_c-dimensional and has unit energy property, i.e., $||c_k||^2 = 1$ and $\tau_k \in [0, T)$ are the relative offsets of the asynchronous signals at the receiver. The zero-mean additive white Gaussian channel noise $v(t)$ has constant power spectral density σ^2. If we incorporate the amplitude A_k and delay τ_k in the channel response $h_k(t)$, we can replace the spreading code sequence $c(t)$ with the discrete-time combined channel-code

This work was supported by the Mahanakorn University of Technology, Bangkok, Thailand.

response

$$g_k[l] = \sum_{i=0}^{L_c-1} c_k[i]h_k[l-i], \tag{2}$$

where $c_k[i]$ is the ith element of the code vector for the kth user $\mathbf{c}_k = (c_k[0], \ldots, c[L_c-1])^T$.

The continuous-time received signal $r(t)$ is sampled to form a length-L_f received signal vector at the nth observation, where L_f is the length of a receiver for the kth user with tap-weight vector $\mathbf{f}_k$,

$$\mathbf{r}[n] = (r[nN + L_f - 1], \cdots, r[nN])^T. \tag{3}$$

The received signal vector $r(t)$ in (1) can then be formulated in the matrix-vector form as

$$\mathbf{r}[n] = \sum_{k=1}^{K} \mathbf{r}_k[n] + \mathbf{v}[n] = \sum_{k=1}^{K} \mathbf{G}_k\mathbf{b}_k[n] + \mathbf{v}[n], \tag{4}$$

where $\mathbf{G}_k$ is the combined code-channel response matrix of the kth user and $\mathbf{b}_k[n] = (b_k[n+L_b-1], \cdots, b_k[n])^T$ with $L_b = \lceil \frac{L_f+L_h-1}{L_c} \rceil$ and $\mathbf{v}[n] = (v[nN]+L_f-1], \cdots, v[nN])^T$. Note that

$$\mathbf{G}_k = \mathbf{C}_k\mathbf{H}_k, \tag{5}$$

where the $L_f \times L_bL_h$ code matrix $\mathbf{C}_k$ and the $L_bL_h \times L_b$ channel matrix $\mathbf{H}_k$ are defined as

$$\mathbf{C}_k = \begin{pmatrix} c_k[L_c-1] & & \mathbf{0} \\ \vdots & \ddots & \\ c_k[0] & & c_k[L_c-1] \\ & \ddots & \vdots \\ \mathbf{0} & & c_k[0] \end{pmatrix}, \quad \mathbf{H}_k = \begin{pmatrix} \mathbf{h}_k & & \mathbf{0} \\ & \mathbf{h}_k & \\ \mathbf{0} & \ddots & \\ & & \mathbf{h}_k \end{pmatrix},$$

where the channel response vector for the kth user has length L_h, i.e., $\mathbf{h}_k = (h[L_h-1], \ldots, h[0])^T$. For brevity, we shall consider the first user as the desired user and drop the subscript k in all variables involving the first user.

3. MIXED-COST ALGORITHM

Consider a combined cost function

$$\begin{aligned} J(\mathbf{f},\lambda) &= \lambda\tilde{J}(\mathbf{f}) + (1-\lambda)\tilde{\tilde{J}}(\mathbf{f}), \\ \text{where} \quad \tilde{J}(\mathbf{f}) &= E\{\mathbf{f}^T\mathbf{R}\mathbf{f}\} \\ \tilde{\tilde{J}}(\mathbf{f}) &= E\{((\mathbf{f}^T\mathbf{r})^2-1)^2\} \end{aligned} \tag{6}$$

are the CMOE [2] and the CM [6] cost functions respectively and $\lambda \in [0,1]$ is the *mixing parameter* and $\mathbf{R} = E\{\mathbf{r}\mathbf{r}^T\}$. The CMOE criterion [2] is given by

$$\min_{\mathbf{f}} E\{\mathbf{f}^T\mathbf{R}\mathbf{f}\} \quad \text{subject to} \quad \mathbf{f}^T\mathbf{C} = \mathbf{1},$$

where $\mathbf{1} \triangleq (1, 0, \cdots, 0)^T$ since the first path is assumed to be the dominant path and the gradient of the CMOE cost is given by [2]

$$\frac{\partial\tilde{J}(\mathbf{f})}{\partial\mathbf{f}}\Big|_{\mathbf{f}=\mathbf{f}[n]} = z[n]\mathbf{\Pi}_{\mathbf{C}}^{\perp}\mathbf{r}[n]. \tag{7}$$

where $z[n] = \mathbf{f}^T[n]\mathbf{r}[n]$ is the output of the receiver and $\mathbf{\Pi}_{\mathbf{C}}^{\perp} = \mathbf{I} - \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T$ denotes the projection matrix onto the nullspace of $\mathbf{C}$. The CMA receivers, i.e., the locations in receiver parameter space of the local minima of $\tilde{\tilde{J}}(\mathbf{f})$, are found by means of the CMA algorithm [6] which searches adaptively for the zero of the gradient

$$\frac{\partial\tilde{\tilde{J}}(\mathbf{f})}{\partial\mathbf{f}}\Big|_{\mathbf{f}=\mathbf{f}[n]} = z[n](z^2[n]-1)\mathbf{r}[n]. \tag{8}$$

For the derivation of a CMA receiver, two important points need to be mentioned. First, it should be noted that the initialisation of a CMA receiver is crucial for the convergence to the solution of a desired user. In [5], a timing acquisition scheme of a desired user is proposed in order to be used as an initialisation of a CMA equaliser. Collectively, this acquisition-equalisation process is called the minimum-entropy CMA (ME-CMA) receiver [5]. When the received signals are not pre-whitened in the equalisation process, the ME-CMA receiver is essentially a conventional CMA receiver.

Second, the CMA receiver is shown to converge faster if a constraint is imposed on the received signals as shown in [7]. However, it can be shown that the CMA receiver still converges to a desired solution without the requirement of any constraint as long as an appropriate initialisation is used [5]. In the derivation of the mixed-cost receiver, therefore, we do not impose any constraint upon the constant modulus derivative.

3.1. Weight vector update equation

Following the derivation of the algorithms for both criteria, the update equation of the mixed cost CMOE-CMA receiver weight vector $\mathbf{f}[n]$ is given by

$$\begin{aligned} \mathbf{f}[n+1] &= \mathbf{f}[n] - \mu\frac{\partial J(\mathbf{f},\lambda)}{\partial\mathbf{f}}\Big|_{\mathbf{f}=\mathbf{f}[n]} \\ &= \mathbf{f}[n] - \mu\Big(\lambda z[n]\mathbf{\Pi}_{\mathbf{C}}^{\perp}\mathbf{r}[n] \\ &\qquad - \beta(1-\lambda)z[n](z^2[n]-1)\mathbf{r}[n]\Big), \end{aligned} \tag{9}$$

where μ is the stepsize of the mixed-cost algorithm. For best operation of this algorithm, it is necessary to weight the constant modulus derivative which in effect modifies the mixture in (6) and can be explained by the nonhomogeneity of the two costs. This is realised by introducing the scaling

factor β for the constant modulus derivative in (9). Note that for the case of $\lambda = 1$, equation (9) is the update equation of the CMOE receiver [2] and when $\lambda = 0$, the mixed-cost receiver is essentially the CMA receiver [6].

3.2. Update equations for the mixing parameter

The main objective of the derivation of the mixed-cost algorithm is to jointly exploit the features of the two criteria in various near-far environments. Therefore, the mixing parameter λ is replaced by the time-varying version $\lambda[n]$ in order to track the variation of the channel.

We adopt the multi-step method as described in [8] for the update of the mixing parameter $\lambda[n]$ in a similar manner as for adapting the gain in the adaptive gain algorithm. The adaptation of the mixing parameter $\lambda[n]$ is obtained by applying a second LMS-type algorithm to adaptively minimise $J(\mathbf{f}, \lambda)$ with respect to λ. The stochastic gradient update equation for $\lambda[n]$ is given by

$$\begin{aligned}\lambda[n+1] &= \Big[\lambda[n] - \alpha \frac{\partial J(\mathbf{f},\lambda)}{\partial \lambda}\Big|_{\lambda=\lambda[n]}\Big]_{\lambda_-}^{\lambda_+} \\ &= \Big[\lambda[n] - \alpha\Big(z[n]^2 - (z[n]^2-1)^2 \\ &\quad + \{2\lambda[n]z[n]\mathbf{r}^T[n] \\ &\quad + 4(1-\lambda[n])(z[n]^2-1)z[n]\mathbf{r}^T[n]\}\Psi[n]\Big)\Big]_{\lambda_-}^{\lambda_+}, \end{aligned} \tag{10}$$

where α is the adaptation rate and $[\cdot]_{\lambda_-}^{\lambda_+}$ denotes truncation to the limits of the range $[\lambda_-, \lambda_+]$ and $0 \le \lambda_- < \lambda_+$. $\Psi[n]$ represents the derivative $\partial \mathbf{f}[n]/\partial\lambda|_{\lambda=\lambda[n]}$. From (9), the update equation of $\Psi[n]$ is given by

$$\begin{aligned}\Psi[n+1] = &\Big[I - \mu(\lambda[n]\mathbf{\Pi}_{\mathbf{C}}^{\perp}\mathbf{r}[n]\mathbf{r}[n]^T \\ &- \beta(1-\lambda[n])(3z^2[n]-1)\mathbf{r}[n]\mathbf{r}[n]^T)\Big]\Psi[n] \\ &- \mu\Big(z[n]\mathbf{\Pi}_{\mathbf{C}}^{\perp}\mathbf{r}[n] - \beta(z^2[n]-1)z[n]\mathbf{r}[n]\Big). \end{aligned} \tag{11}$$

Equation (9) together with (10) and (11) constitute the new mixed-cost CMOE-CMA algorithm. The structure of the proposed receiver is shown in Fig. 1.

3.3. Computational complexity and convergence properties

The computational complexity of the algorithm is $L_f^2 + 16L_f + 12$ and $L_f^2 + 9L_f$ in terms of multiplications and additions. Since global convergence property of the CMOE has been given in [2] and local convergence of CMA has been shown in [4], with careful choice of μ and $\lambda[n]$, the combined algorithm should demonstrate at least similar convergence properties.

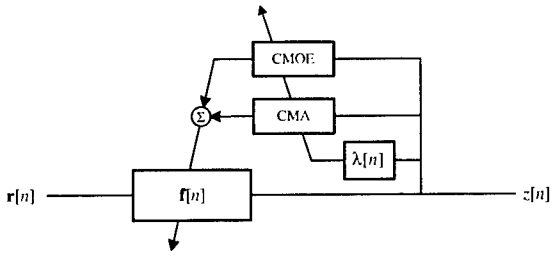
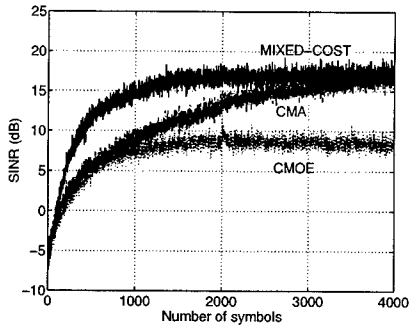


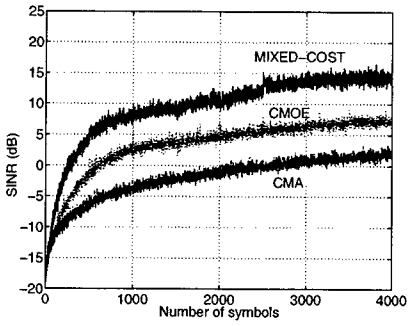
Fig. 1. The mixed-cost CMOE-CMA receiver.

4. SIMULATIONS

We considered a symbol-asynchronous system with processing gain $L_c = 31$ and number of users $K = 7$. User delays τ_k were uniformly distributed over $[0, 7T_c)$ and then kept fixed. The propagation channels are bandlimited with root-raised-cosine pulse shaping with excess bandwidth 0.2. The number of multipath rays was three, where the last two rays were uniformly distributed in delay over $[0, 7T_c)$. The channel length for all users was $10T_c$. We assumed without loss of generality that the first user is the user of interest with unity power. The timing of the first user was assumed to be known. The background noise was zero-mean AWGN with SNR=20 dB (referenced to the first user). Each receiver was length-$2L_c$. The initial value of $\lambda[n]$ was set to unity and λ_- and λ_+ were 0 and 1 respectively. The adaptation rate α was 5×10^{-4} and $\Psi[0]$ was $0.1[1, \cdots, 1]^T$. The performance measure was the averaged SINR in dB and all SINR plots were averaged over 100 Monte-Carlo runs. We compared the performances of the CMA receiver, the CMOE receiver [2] and the proposed mixed-cost CMOE-CMA receiver. We tested the performances of the three receivers in various settings of the near-far situations which can be quantified in terms of the near-far ratio (NFR) where NFR $= 10\log_{10}\frac{A_k^2}{A_1^2}, \forall k = \{2, \ldots, 7\}$. Fig. 2 (a) and (b) show the averaged SINR plots of the three receivers at NFR= 14 dB and 26 dB respectively. It is observed that the performance of the CMA receiver is degraded in high NFR cases because the attraction basin of the desired user is likely to reduce in dimension as the NFR increases. The CMOE receiver reveals the characteristic of near-far resistance as shown in Fig. 2 (b) but inferior to the CMA receiver in the low NFR cases as in Fig. 2 (a). In both cases, the mixed-cost receiver is superior to the two existing receivers. The steady-state averaged SINR plots at various NFRs are shown in Fig. 3. For the mixed-cost receiver, high SINR levels were maintained over a wide-range of near-far levels confirming its near-far resistance characteristic. Time evolution plots of $\lambda[n]$ for different NFRs are shown in Fig. 4. Notice that the relaxation rate is varied as a function of the NFR settings. For low NFRs, $\lambda[n]$ decays quickly to zero



(a)



(b)

Fig. 2. The comparison of SINR performances of three receivers at (a) NFR=14 dB and (b) NFR=26 dB.

whereas its magnitude is sustained at high levels for high NFRs.

5. CONCLUSION

We have presented a new mixed-cost receiver structure for DS-CDMA systems based on the CMOE and CM criteria. The multi-step method is exploited in the derivation of the adaptive mixing parameter algorithm. Simulations have confirmed that the averaged SINR performance of the proposed mixed-cost algorithm in various near-far situations is superior to the existing algorithms. On-going research is focused upon the evaluation of this method in the presence of time-varying interference.

6. REFERENCES

[1] M. L. Honig, U. Madhow, and S. Verdú, "Blind adaptive multiuser detection," *IEEE. Trans. Inform. Theory*, vol. 41, pp. 944–966, 1995.

[2] M. K. Tsatsanis, "Inverse filtering criteria for CDMA systems," *IEEE Trans. Signal Processing*, vol. 45, no. 1, pp. 102–112, 1997.

[3] P. Schniter and C. R. Johnson Jr., "On the robustness of blind linear receivers for short-code CDMA," in *Proc. IEEE 2nd SPAWC*, Annapolis, MD, pp. 13–16, 1999.

[4] H. H. Zeng, L. Tong, and C. R. Johnson Jr., "Relationships between the constant modulus and Wiener receivers," *IEEE Trans. Inform. Theory*, vol. 44, no. 4, pp. 1523–1538, 1998.

[5] P. Schniter and C. R. Johnson Jr., "Minimum-entropy blind acquisition/equalization for uplink DS-CDMA," in *Proc. Allerton Conf. on Commun., Contr., and Computing*, Monticello, IL, pp. 401–410, 1998.

[6] D. N. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. Commun.*, vol. 28, pp. 1867–1875, 1980.

[7] L. Li and H. Howard Fan, "Blind CDMA detection and equalization using linearly constraint," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Istanbul, Turkey, vol. 5, pp. 2905–2908, 2000.

[8] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations, Volume 22 of Applications of Mathematic*, Springer-Verlag, New York, 1990.

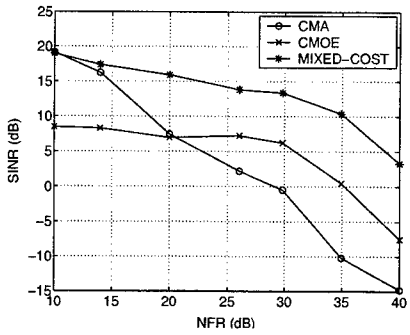


Fig. 3. The plots of SINR in dB as NFR varies from 10 dB to 40 dB.

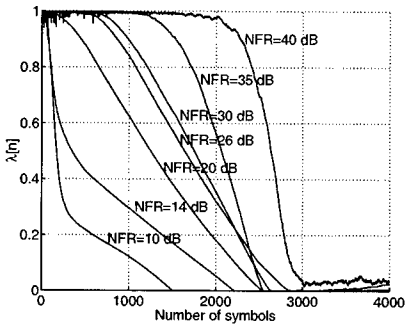


Fig. 4. The plots of $\lambda[n]$ as NFR varies from 10 dB to 40 dB.

A REDUCED RANK DECORRELATING RAKE RECEIVER FOR CDMA COMMUNICATIONS OVER FREQUENCY SELECTIVE CHANNELS

Ozgur Ozdemir and Murat Torlak

Wireless Information Systems Lab. (WISLAB)
The Department of Electrical Engineering
The University of Texas at Dallas, Richardson, TX 75083
ozdemir@utdallas.edu and torlak@utdallas.edu

ABSTRACT

CDMA systems need simultaneous multiple access interference suppression and adaptive interference suppression filters that may span three symbols. Thus, a large number of filter coefficients need to be estimated. By the use of reduced rank filtering, it is possible to lower the number of required filter coefficients with a small decrease in performance. In [1], a successful attempt for non-dispersive CDMA signals was made to develop a reduced rank algorithm based on the multistage Wiener (MSW) filter of Goldstein and Reed [2]. In this paper, motivated by MSW we propose a reduced rank decorrelating RAKE receiver for dispersive CDMA signals. The proposed receiver is blindly implemented in a lower dimensional space, relative to the full-rank receivers, without the aid of training sequences and the channel information. By exploiting the structure of the user signature waveform, the proposed receivers exhibit performance close to that of the reduced rank MMSE receiver implemented with the desired user's known channel information.

1. INTRODUCTION

In CDMA systems orthogonality between waveforms may not be protected because of the random delay of the users and the frequency-selective fading environment. It is well known that a simple matched filter suffers from increased cross correlations between user's signature waveforms in fading and results in a poor performance.

Multiuser detection provides an eminent solution to reinstate the CDMA systems' performance promised by the use of orthogonal waveforms. Most existing multiuser detectors such as the linear receiver require the knowledge of the users' (or at least the desired user's) signature waveforms. However, signature waveforms vary with the multipath channel characteristics. Although the signature waveform can be estimated periodically using training sequences, this may not be affordable in a fast changing wireless propagation environment.

Blind multiuser receivers based on subspace decomposition can mitigate possible multipath effects and channel dispersion [3]. However, these receivers require heavy computation due to the use of subspace decomposition. Motivated by the low complexity of RAKE receivers, Liu and Li [4] proposed a decorrelating RAKE receiver that provides a performance close to that of the MMSE receiver with the desired user's known channel information. Decorrelating RAKE and MMSE receivers is designed to suppress multiuser interference, therefore, they still require knowledge of the observed-data covariance matrix inverse to estimate a large number of filter coefficients.

In [1], a reduced rank minimum mean squared error (MMSE) receiver based on the multistage Wiener (MSW) filter is proposed for non-dispersive CDMA signals in order to lower the number of required filter coefficients especially for the cases where the processing gain is larger than the dimension of the signal subspace. Reduced rank receivers are concerned with reduction in dimensionality of the observed data. Thus, the purpose of a reduced rank receiver is to obtain near full-rank performance with a filter order smaller than the signal subspace. The important feature of the reduced rank MSW filter in [1] is that it does not rely on eigen decomposition and hence its low complexity. However, the algorithm is applicable when the signature waveform of the user of interest is available to the receiver or the training sequences are employed.

In this paper, we propose a reduced rank MSW decorrelating RAKE receiver which exploits the structure of the user's spreading waveform in multipath. The performance of the proposed RAKE receiver is compared by computer simulations to the eigen-based cross-spectral methods of rank reduction and the full rank decorrelating RAKE receiver. The simulation results demonstrate that the proposed method can outperform eigen-based methods without utilizing eigen decomposition and provide a performance similar to reduced-rank MSW receiver with the known signature waveform.

The rest of the paper is organized as follows. Section 2 gives a description of the signal model. Section 3 introduces the reduced rank decorrelating RAKE receiver. Section 4 discusses its implementation based on MSW. Simulation results and conclusions are given in Section 5 and 6, respectively.

2. DATA MODEL

We consider a P-user asynchronous CDMA (A-CDMA) baseband signal sampled at the chip rate:

$$x(l) = \sum_{i=1}^{P} \sum_{n=-\infty}^{\infty} s_i(n) g_i(l - nL_c) \qquad (1)$$

where $\{s_i(n)\}$ are the information symbols and the spreading waveform distorted by the time dispersive channel is

This work is supported in part by Texas Telecommunications Engineering Consortium (TxTEC).

given by

$$g_i(l) = \sum_{k=1}^{L} h_i(k)c_i(l-k+k_i), \quad l = 1, \cdots, 2L_c \qquad (2)$$

where k_i is the chip known delay index. The above signature waveform can be written in a vector form. We assume that $L < L_c$ so that a window spanning two symbols is enough to fully observe the signature waveform regardless of the delay index value.The vector $\mathbf{g}_i = \mathbf{C}_i\mathbf{h}_i$ is given by the channel vector $\mathbf{h}_i$ multiplied by $2L_c \times L$ Toeplitz filtering matrix constructed from the code c_i,

$$\mathbf{C}_i = \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} \\ c_i(1) & \ddots & \mathbf{0} \\ \vdots & \ddots & c_i(1) \\ c_i(L_c) & \ddots & \vdots \\ \mathbf{0} & \ddots & c_i(L_c) \\ \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}. \qquad (3)$$

It is seen that $\{\mathbf{g}_i\}$ are uniquely determined by the unknown channel vectors $\{\mathbf{h}_i\}$. Using MATLAB notation, define

$$\mathbf{g}_i(m) = \mathbf{g}_i((m-1)L_c + 1 : mL_c), \quad m = 1, 2. \qquad (4)$$

If we desire the first user's signal as a signal of interest, then, stacking data samples by a window that spans two symbols [3] provides

$$\mathbf{x}(n) = \mathbf{g}_1 s_1(n) + \mathbf{z}(n) \qquad (5)$$

where $\mathbf{z}(n)$ is given by

$$\begin{bmatrix} \mathbf{g}_1(2) & \mathbf{0} \\ \mathbf{0} & \mathbf{g}_1(1) \end{bmatrix} \begin{bmatrix} s_1(n-1) \\ s_1(n+1) \end{bmatrix} + \sum_{i=2}^{P} \mathbf{G}_i\mathbf{s}_i(n) + \mathbf{o}(n) \qquad (6)$$

and $\mathbf{G}_i$ and $\mathbf{s}_i(n)$ are defined by

$$\mathbf{G}_i = \begin{bmatrix} \mathbf{g}_i(2) & \mathbf{g}_i(1) & \mathbf{0} \\ \mathbf{0} & \mathbf{g}_i(2) & \mathbf{g}_i(1) \end{bmatrix}, \quad \mathbf{s}_i(n) = \begin{bmatrix} s_i(n-1) \\ s_i(n) \\ s_i(n+1) \end{bmatrix}. \qquad (7)$$

Note that $\mathbf{z}(n)$ contains the ISI components for the signal of interest, the MUI, and the background noise $\mathbf{o}(n)$.

3. REDUCED RANK DECORRELATING RAKE RECEIVER

The signature waveform of the desired user is a linear combination of the columns of the Toeplitz filtering matrix $\mathbf{C}_i$ that specify the constraints in the minimum output energy (MOE) receiver [4]. Omitting the desired user's subscript the problem is defined as

$$\mathbf{w}_l = \arg \min_{\mathbf{w}_l} \mathbf{w}_l^H \mathbf{R}_{\mathbf{xx}} \mathbf{w}_l \quad s.t. \quad \mathbf{C}^H \mathbf{w}_l = \mathbf{1}_l \qquad (8)$$

where $\mathbf{R}_{\mathbf{xx}} = \mathbf{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}$ is $2L_c \times 2L_c$ data covariance matrix, $\mathbf{w}_l$ is the weight vector corresponding to the lth arm of the RAKE receiver, and $(.)^H$ represents complex conjugate transpose. The closed form optimal solution to the problem can be obtained via Lagrange multipliers

$$\mathbf{w}_{l,opt} = \mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{C}(\mathbf{C}^H\mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{C})^{-1}\mathbf{1}_l. \qquad (9)$$

This solution is computationally complex in the sense that it requires $2L_c \times 2L_c$ matrix inversion. In order to simplify the computation of the weight vector given in (9), the projection matrix $\mathbf{P}_c = \mathbf{C}(\mathbf{C}^H\mathbf{C})^{-1}\mathbf{C}^H$ can be used to decompose this vector into two orthogonal components as follows:

$$\mathbf{w}_l = \mathbf{w}_{c,l} - \mathbf{M}^H\mathbf{w}_{a,l} \qquad (10)$$

where

$$\mathbf{w}_{c,l} = \mathbf{P}_c\mathbf{w}_{l,opt} = \mathbf{C}(\mathbf{C}^H\mathbf{C})^{-1}\mathbf{1}_l \qquad (11)$$

is not data dependent and $\mathbf{M}$ is a $(2L_c - L) \times 2L_c$ blocking matrix which is chosen to satisfy $\mathbf{MC} = \mathbf{0}$, therefore guarantees that second component is orthogonal to first component [5]. Note that $\mathbf{w}_{a,l}$ is the data dependent weight vector and the size of $\mathbf{w}_{a,l}$ dominates the complexity of the receiver. By projecting the blocked data $\mathbf{y}(n) = \mathbf{Mx}(n)$ onto a lower dimensional subspace, the number of filter taps needed to compute $\mathbf{w}_{a,l}$ can be reduced. As shown in Figure 1, let $\mathbf{Q}_l$ for each RAKE finger be a $D \times (2L_c - L)$ matrix where $D < (2L_c - L)$. In this case the rows of $\mathbf{Q}_l$ is a basis for the lower dimensional subspace. The new optimization problem involving the new reduced rank adaptive vector becomes

$$\mathbf{w}_{r,l} = \arg \min_{\mathbf{w}_{r,l}} E\{|d_l(n) - \mathbf{w}_{r,l}^H\mathbf{Q}_l\mathbf{Mx}(n)|^2\} \qquad (12)$$

where $d_l(n) = \mathbf{w}_{c,l}^H\mathbf{x}(n)$ and in turn $\mathbf{w}_{a,l}$ is computed by

$$\mathbf{w}_{a,l} = \mathbf{Q}_l^H\mathbf{w}_{r,l} \qquad (13)$$

the solution to the optimization problem in (12) is given by

$$\mathbf{w}_{r,l}^{opt} = \mathbf{R}_{\mathbf{Q}_l}^{-1}\mathbf{r}_{\mathbf{Q}_l} \qquad (14)$$

where

$$\mathbf{R}_{\mathbf{Q}_l} = \mathbf{E}\{\mathbf{Q}_l\mathbf{Mx}(n)\mathbf{x}^H(n)\mathbf{M}^H\mathbf{Q}_l^H\} \qquad (15)$$

$$= \mathbf{Q}_l\mathbf{MR}_{\mathbf{xx}}\mathbf{M}^H\mathbf{Q}_l^H \qquad (16)$$

and

$$\mathbf{r}_{\mathbf{Q}_l} = \mathbf{E}\{\mathbf{Q}_l\mathbf{Mx}(n)d_l^*(n)\} \qquad (17)$$

$$= \mathbf{Q}_l\mathbf{MR}_{\mathbf{xx}}\mathbf{w}_{c,l} \qquad (18)$$

This blind reception problem can be solved for each $\mathbf{Q}_l$ and $\mathbf{w}_{r,l}$ using the eigen-based reduced rank methods and the multistage decomposition [1]. The filter outputs $\mathbf{z}_l = \mathbf{w}_l\mathbf{x}$ can then be coherently combined to obtain the final signal estimate using an estimate of the principal vector of $\mathbf{R}_{\mathbf{zz}}$. The computational effort for coherent combining is on the order of $\mathcal{O}(L^2)$ which is insignificant effort due to $L \ll 2L_c$.

3.1. Eigen-Based Methods

Eigen-based methods have been extensively used in order to obtain a lower dimensional subspace for the received data [1, 3]. These methods are based on the eigen-decomposition of the covariance matrix

$$\mathbf{R}_{\mathbf{yy}} = \mathbf{E}\{\mathbf{y}(n)\mathbf{y}(n)^H\} = \mathbf{V}\Lambda\mathbf{V}^H \qquad (19)$$

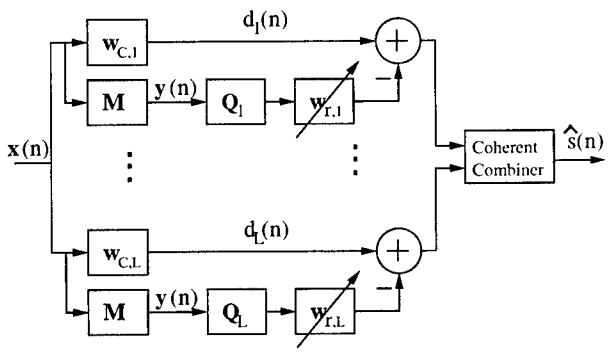


Figure 1: Reduced rank decorrelating RAKE receiver.

where $\mathbf{V}$ is the orthonormal matrix of eigenvectors of $\mathbf{R_{yy}}$ and Λ is the diagonal matrix of eigenvalues.

In the principle components (PC) method the rows of $\mathbf{Q}_l$ are determined by the $\mathbf{R_{yy}}$'s D principle eigenvectors which are associated with the D largest eigenvalues. The performance of this method degrades rapidly when D is smaller than the dimension of signal subspace in the blocked data $\mathbf{y}(n)$.

An alternative to PC is the Cross Spectral (CS) method. The CS method chooses a set of D eigenvectors from each branch to minimize the MSE on that branch. The full rank MMSE on the lth branch may be expressed in terms of eigen-decomposition of $\mathbf{R_{yy}}$ as follows:

$$\xi_l = \sigma_{d_l}^2 - \mathbf{r}_{\mathbf{y}d_l}^H \mathbf{R}_{\mathbf{yy}}^{-1} \mathbf{r}_{\mathbf{y}d_l} \quad (20)$$

$$= \sigma_{d_l}^2 - \mathbf{r}_{\mathbf{y}d_l}^H \mathbf{V} \Lambda^{-1} \mathbf{V}^H \mathbf{r}_{\mathbf{y}d_l} \quad (21)$$

$$= \sigma_{d_l}^2 - \sum_{i=1}^{2L_c-L} \frac{|\mathbf{r}_{\mathbf{y}d_l}^H \mathbf{v}_i|^2}{\lambda_i} \quad (22)$$

where $\sigma_{d_l}^2$ is the variance of $d_l(n)$, $\mathbf{r}_{\mathbf{y}d_l}$ is the cross correlation between $\mathbf{y}(n)$ and $d_l(n)$, $\mathbf{v}_i$ is the ith eigenvector of $\mathbf{R_{yy}}$, and λ_i is the associated eigenvalue. To minimize MSE, D eigenvectors with the largest values of $|\mathbf{r}_{\mathbf{y}d_l}^H \mathbf{v}_i|^2/\lambda_i$ are selected. This method can perform well even if D is smaller than the dimension of signal subspace.

The disadvantage of both PC and CS is that they include eigen-decomposition with a complexity of $\mathcal{O}(4L_c^3)$. In the next section, motivated by the MSW filter [1, 2], we present another reduced rank method which performs better than PC and CS and does not require matrix inversion or an eigen decomposition.

4. MULTISTAGE DECOMPOSITION OF DECORRELATING RAKE RECEIVER

Once we form $d_l(n) = \mathbf{w}_{c,l}^H \mathbf{x}(n)$ and the data is blocked by the blocking matrix $\mathbf{M}$, reduced rank MSW filtering can be applied to each arm to suppress the interference [1, 2]. A block diagram for a rank-3 MSW filter is shown in Figure 2.

The stages are associated with the sequence of nested filters $\mathbf{f}_{l,1}, \cdots, \mathbf{f}_{l,D}$ where $D \ll L_c$ is the order of the reduced rank filter. $\mathbf{B}_{l,m}$ denotes a blocking matrix such that:

$$\mathbf{B}_{l,m}^H \mathbf{f}_{l,m} = \mathbf{0} \quad (23)$$

Referring to Figure 2, $\{\mathbf{f}_{l,m}\}$ are determined by

$$\mathbf{f}_{l,m} = E\{\mathbf{y}_{l,m-1}(n) d_{l,m-1}^*(n)\}, \quad m = 1, \cdots, D \quad (24)$$

Here we assume that each blocking matrix $\mathbf{B}_{l,m}$ is $2L_c \times 2L_c$ so that each vector $\mathbf{f}_{l,m}$ is $2L_c \times 1$. And it will be convenient to normalize the filters $\mathbf{f}_{l,1}, \cdots, \mathbf{f}_{l,D}$ so that $\|\mathbf{f}_{l,m}\| = 1$.

The outputs of the filters $\mathbf{f}_{l,1}, \cdots, \mathbf{f}_{l,D}$ are linearly combined via the weights $w_l(m), \cdots, w_l(D)$ to obtain the filter output. This is accomplished stage by stage. Referring to Figure 2:

$$w_l(m) = \arg \min_{w_l(m)} E\{|\epsilon_{l,m-1}(n)|^2\}, \quad m = 1, \cdots, D \quad (25)$$

where

$$\epsilon_{l,m}(n) = d_{l,m}(n) - w_l(m+1)\epsilon_{l,m+1}(n). \quad (26)$$

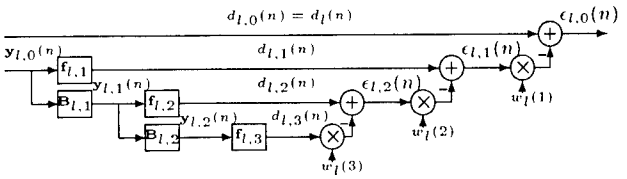


Figure 2: Multistage RAKE Receiver for lth arm.

The rank-D MSW filter is given by the following set of recursions.

Initialization:

$$d_{l,0}(n) = d_l(n), \quad \mathbf{y}_{l,0}(n) = \mathbf{M}\mathbf{x}(n) \quad (27)$$

For $m = 1, \cdots, D$ (Forward Recursion):

$$\mathbf{f}_{l,m} = \frac{E\{\mathbf{y}_{l,m-1}(n) d_{l,m-1}^*(n)\}}{\|E\{\mathbf{y}_{l,m-1}(n) d_{l,m-1}^*(n)\}\|} \quad (28)$$

$$d_{l,m}(n) = \mathbf{f}_{l,m}^H \mathbf{x}_{l,m-1}(n) \quad (29)$$

$$\mathbf{B}_{l,m} = \mathbf{I} - \mathbf{f}_{l,m}\mathbf{f}_{l,m}^H \quad (30)$$

$$\mathbf{y}_{l,m}(n) = \mathbf{B}_{l,m}^H \mathbf{y}_{l,m-1}(n) \quad (31)$$

Decrement $m = D, \cdots, 1$ (Backward Recursion):

$$w_l(m) = \frac{E\{d_{l,m-1}^*(n)\epsilon_{l,m}(n)\}}{E\{|\epsilon_{l,m}(n)|^2\}} \quad (32)$$

$$\epsilon_{l,m-1}(n) = d_{l,m-1}(n) - w_l(m)^* \epsilon_{l,m}(n) \quad (33)$$

where $\epsilon_{l,D}(n) = d_{l,D}(n)$. $\epsilon_{l,0}(n)$ is the final signal estimate of the lth arm and it is the input to the coherent combiner. The Coherent combiner uses the principal eigenvector of the estimated covariance matrix formed by the signal estimates of each arm. Each stage has a complexity of $\mathcal{O}(L_c^2)$ and multistage decomposition does not need the complete estimation of the covariance matrix.

5. SIMULATION RESULTS

Computer simulations have been conducted to examine the MSE performance and the convergence behavior of the proposed receiver. We consider a 10-user system with spreading gain of 16 and 400 symbols. Signals go through three-ray multipath fading with an SNR of 10 dB. The results in Figure 3 and Figure 4 are averaged over 100 simulations.

Figure 3 shows the output MSE as a function of rank for the proposed receiver, CS, and PC. Results show that PC degrades rapidly with the decreasing rank. On the other

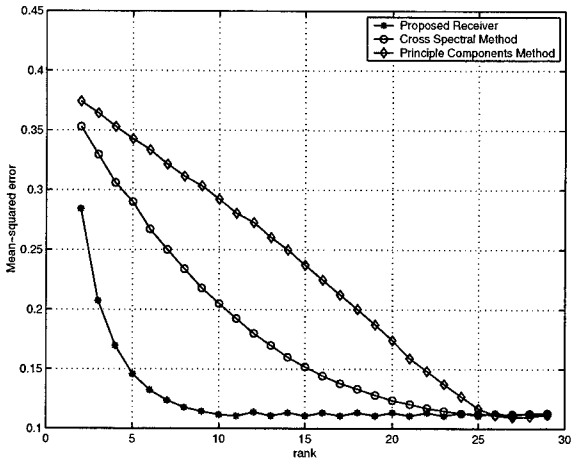


Figure 3: The comparison of proposed receiver with eigen-based methods.

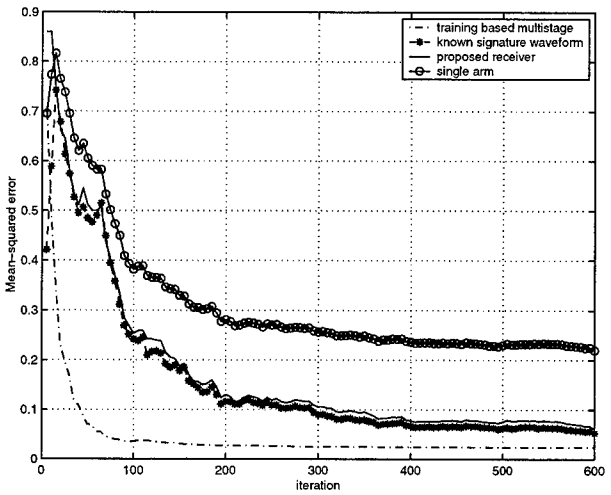


Figure 5: MSE vs. iteration for different receivers for $D = 8$.

hand for given subspace dimension D, CS performs much better than PC. We also observe that CS achieves near full-rank performance when $D = 22$. The proposed receiver outperforms both CS and PC. It achieves near full rank performance with less than half the number of weights used in the CS method.

Figure 4 plots the MSE for the following MSW-based receivers: training-based MMSE, MMSE with known signature waveform, the proposed decorrelating RAKE receiver, and the single arm receiver. The results show that our proposed method gives nearly identical MSE compared to the MMSE receiver with known signature waveform. The training-based MMSE performs best among all the receivers, whereas the single arm receiver [6] has the worst performance.

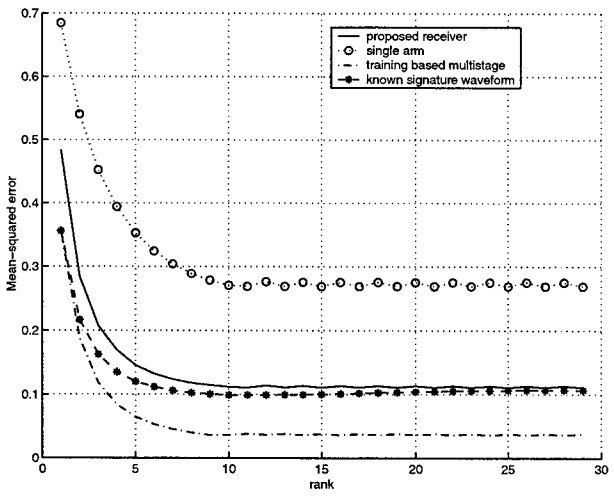


Figure 4: The comparison of different receivers with varying rank.

The convergence of the above methods is compared in Figure 5 for a similar scenario and $D = 8$. It is seen that the convergence rate of the proposed receiver is close to that of the MMSE receiver with known signature waveform. With large data size, the proposed method approaches the training-based MMSE receiver.

6. CONCLUSIONS

In this paper, we have proposed and demonstrated a reduced rank decorrelating RAKE receiver in the presence of frequency selective fading. The proposed RAKE receiver is based on the multistage Wiener (MSW) filter of Goldstein and Reed [2]. This receiver allows a large reduction in rank as well as in computational complexity relative to other reduced rank filters based on eigen decomposition methods. Finally, the proposed method offers a performance similar to Honig's reduced rank MMSE receiver with the known signature waveform [1].

7. REFERENCES

[1] M. Honig, "Adaptive reduced-rank residual correlation algorithms for DS-CDMA interference suppression", in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, Nov. 1998, pp. 1106–1110.

[2] J. Goldstein, I. Reed, and L. Scharf, "A multistage representation of the wiener filter based on orthogonal projections", *IEEE Tran. on Information Theory*, vol. 44, no. 7, pp. 2943–2959, Nov. 1998.

[3] M. Torlak and G. Xu, "Blind multi-user channel estimation in asynchronous CDMA systems", *IEEE Trans. on Signal Processing*, vol. 45, no. 1, pp. 137–147, Jan. 1997.

[4] H. Liu and K. Li, "A decorrelating RAKE receiver for CDMA communications over frequency-selective fading channels", *IEEE Tran. on Communications*, vol. 47, no. 7, pp. 1036–1045, July 1999.

[5] J. B. Schodorf and D.B. Williams, "Partially adaptive multiuser detection", in *Proc. IEEE Vehicular Tech. Conf.*, 1996, pp. 367–371.

[6] M. Tsatsanis, "Inverse filtering criteria for CDMA systems", *IEEE Tran. on Signal Processing*, vol. 45, pp. 102–112, Jan. 1997.

MULTIUSER DETECTION IN IMPULSIVE NOISE

A.M. Zoubir and A.T. Lane-Glover

Australian Telecommunications Research Institute
School of Electrical and Computer Engineering
Curtin University of Technology
GPO Box U 1987, Perth 6485, Western Australia.
e-mail: {zoubir, arran}@atri.curtin.edu.au

ABSTRACT

A new technique is proposed for robust multiuser detection in channels with impulsive noise. The method is a modification of traditional non-linear multiuser detection techniques, whereby the non-linearity is now positioned between the multiplier and the summation, within the correlator, instead of preceding it. This is then extended to provide a modification to the two-stage non-linear detector proposed in [1]. Simulation results show that the use of this technique increases multiple access interference (MAI) rejection. Near-far effects are also investigated.

1. INTRODUCTION

Conventional communication system models assume that the noise in the channel is Gaussian. However, in general, electromagnetic interference in channels displays impulsive behaviour, and is therefore non-Gaussian. Conventional detection methods are optimised for operation in Gaussian noise environments [2], and therefore, severe performance degradation occurs when the noise is non-Gaussian.

In Code-Division Multiple Access (CDMA) channels, this implies that fewer users can use the channel, for a given level of signal power. However, it has been shown [3] that, if properly treated, non-Gaussian noise can be beneficial to a system. It is necessary to design multiuser detection schemes that are robust to various levels of impulsive noise. In this paper, non-linear detection schemes are proposed in an attempt to improve the performance of multiuser detectors in impulsive noise channels.

The paper is structured as follows. In Section 2, the system model for DS-CDMA communications and the noise model for the impulsive noise channel are described. In Section 3, a modified multiuser detection model is proposed and simulation results are given. It is shown that such a scheme offers better MAI rejection than the conventional non-linear multiuser detection schemes and therefore offers better performance in a multiuser system. Based on this concept a modification to the two-stage nonlinear multiuser detector given in [1] is proposed in Section 4. It is shown that this model outperforms both the decorrelator and the two-stage non-linear detector. Finally, Section 5 contains some conclusions.

2. SYSTEM MODEL

Consider a K-user, baseband direct sequence - code division multiple access (DS-CDMA) system operating with a coherent binary phase shift keying (BPSK) modulation format, where signals are transmitted synchronously over the channel. The synchronous transmission model can be used without loss of generality. The continuous-time baseband signal received at the detector can be modelled as:

$$r(t) = \sum_{i=0}^{M-1} \sum_{k=1}^{K} A_k b_k(i) s_k(t - iT) + n(t) \qquad (1)$$

where M is the length of the data block in symbols, per user, in the observed data frame, K is the number of active users, T is the symbol period, $A_k, b_k \in \{-1, 1\}$ and $s_k(t)$ are the kth users received amplitude, symbol and normalised signature waveform, respectively, and $n(t)$ is the ambient channel noise, assumed to be identically and independently distributed. For the direct-sequence spread spectrum (DS-SS) multiple access format, the normalised signature waveforms take on the following form

$$s_k(t) = \sum_{j=1}^{N} c_k(j) P_{T_c}(t - (j-1)T_c) \qquad (2)$$

where N is the processing gain, $c_k(j) \in \{-1, 1\}$ are the signature bits for the kth user, and P_{T_c} is the normalised chip waveform with duration $T_c = \frac{T}{N}$. At the detector the received signal, $r(t)$, is filtered by a chip-matched filter and then sampled at the chip rate. The received signal, for a single symbol interval, is then given by

$$r_j = \sum_{k=1}^{K} A_k b_k s_{k,j} + n_j\,, \quad j = 1, \ldots, N\,, \qquad (3)$$

or in vector form,

$$\mathbf{r} = \mathbf{SAb} + \mathbf{n} \qquad (4)$$

where $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_K)$, with $\mathbf{s}_k = (s_{k,1}, \ldots, s_{k,N})'$, $\mathbf{A} = \mathrm{diag}(A_1, \ldots, A_K)$, $\mathbf{b} = (b_1, \ldots, b_K)'$, where $'$ denotes transpose and $\mathbf{n}$ is the noise. It is well known that there are two types of interference in the detection of a single users signal. The first type of interference is due to other users in the system. This is known as MAI, which is caused by the correlation between different users signature waveforms. The second type of interference is due to the noise, which has a non-Gaussian distribution.

To model impulsive noise phenomena, a noise model, that is a probability distribution function (PDF), with heavier tails, than the Gaussian model, is required. The Cauchy distribution is considered in this paper to model the non-Gaussian noise process. The PDF of the Cauchy distribution is

$$f_X(x) = \frac{1}{\pi} \frac{\gamma}{x^2 + \gamma^2} \qquad (5)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

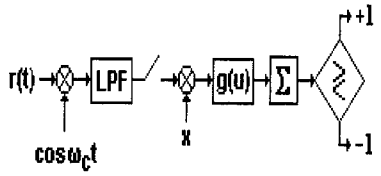


Fig. 1. Modified Non-Linear Multiuser Detection Model

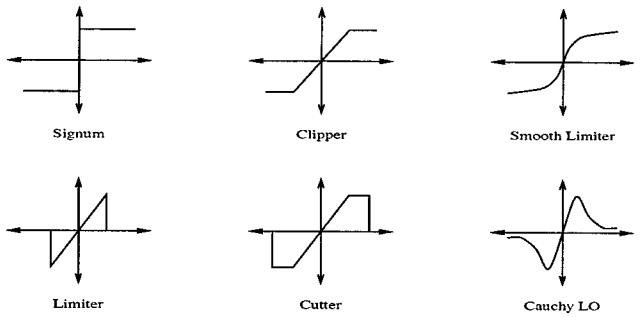


Fig. 2. Locally Optimal and Sub-Optimal Non-Linearities

where $\gamma > 0$ is the spreading factor, similar to variance. It should be noted that the Cauchy model has an infinite variance, but despite this, it is useful to model impulsive noise due to its heavy tail behavior compared to the traditional Gaussian model. The Cauchy model is a special case of the symmetric-alpha-stable ($s\alpha s$) parametric model. The signal to noise ratio (SNR) for an $s\alpha s$ process is defined as

$$SNR_{dB} = 20\log_{10}\left(\frac{A}{\gamma^{1/\alpha}\sqrt{2}}\right) \tag{6}$$

where $\alpha = 1$ for Cauchy noise and $\alpha = 2$ for Gaussian noise. When $\alpha = 2$ the SNR reduces to the conventional Gaussian SNR with $2\gamma = \sigma^2$.

3. MODIFIED NON-LINEAR MULTIUSER DETECTION

In this section, a modified non-linear detection model is proposed and analyzed. Firstly, the model is given and shown to be the solution to a modified least squares regression. Secondly, various non-linearities are introduced for testing of the system, and finally, the performance of the new detection technique is shown.

3.1. The Modified Detection Model and Least Squares

In an attempt to cause the conventional non-linear detectors to demonstrate greater levels of MAI rejection, and therefore better bit error rate (BER) performance in multiuser channels, we propose the modified non-linear multiuser detection model, shown in Figure 1. Note that, unlike in conventional non-linear detectors (see, for example, [3, 4, 5] and references therein), the non-linearity is placed between the multiplier and the summer within the conventional linear detection scheme, rather than before the multiplier. As a result, using the decorrelator [2], that is $\mathbf{x} = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'$ in Figure 1, the MAI can be removed before the non-linearity affects the data. The data is then passed through the equivalent of a non-linear summation. We will refer to the structure in Figure 1 as the modified non-linear detector.

For a non-linearity $g(\cdot)$, it can be shown, for a single bit period, that the output of the modified non-linear detection scheme, using the decorrelator, is the solution to

$$\sum_{j=1}^{N}\left(r_j - \sum_{l=1}^{K} s_{l,j}\, g^{-1}(\theta_l)\right)s_{k,j} = 0, \quad k = 1,\ldots,K \tag{7}$$

where $\theta_l = A_l b_l$, $l = 1,\ldots,K$, which is a modified least squares estimate of $\theta = \mathbf{Ab}$, that is

$$\hat{\theta} = \arg\min_{\theta} \|\mathbf{r} - \mathbf{S}\mathbf{g}^{-1}(\theta)\|^2 \tag{8}$$

The bit decision is made on the test statistic $T(\mathbf{r}) = g(\mathbf{xr})$.

3.2. The Non-Linearities

If the model of the noise in the channel is known to the receiver then, in conventional non-linear detection, a locally optimum non-linearity [4] can be used. However, this is not commonly the case and therefore sub-optimal non-linearities are required. Examples of these non-linearities can be seen in Figure 2. The parameters of the non-linearities were set for the highest SNR at the output of the non-linearity, using a first order approximation; the derivation is not included for brevity.

3.3. Simulations and Results

All simulations in this paper were carried out using random codes with a spreading gain of $N = 30$, in a Cauchy noise channel. Random codes were used to maximize the MAI so that the MAI rejection capabilities of the detectors could be tested. Figure 3 shows the BER against the number of users in the system, for $SNR \approx 5dB$ and $NFR = 0dB$. The first two curves are using the conventional non-linear detection scheme [4], while the next two curves are plotted using the modified detection scheme. It can be seen that the modified detection scheme has better MAI rejection ability since the rate of performance degradation, as the number of users increases, is less than that when using the conventional non-linear detection scheme, even when using a locally optimum non-linearity. In a realistic multiuser system, with many users, the modified detection scheme gives better BER performance.

Figure 4 shows the BER against the NFR for the two different non-linear detection schemes, for $K = 3$ users and $SNR \approx 5dB$. It can be seen that when optimizing in terms of the BER, neither scheme is near-far resistant, compare the clippers. However, there exists a trade-off in the modified detection scheme between BER performance and near-far resistance, as in the conventional scheme. This can be seen in the two plots of the modified smooth limiting non-linearity with different parameters.

The performance of the detectors in Gaussian noise, using the clipper non-linearity with comparative parameter settings, is shown in Figure 5. The simulation was conducted for $K = 3$ users. Using the conventional non-linear detection scheme performance degradation occurs in a Gaussian noise environment, while if the modified detection model is used, there is little or no performance loss.

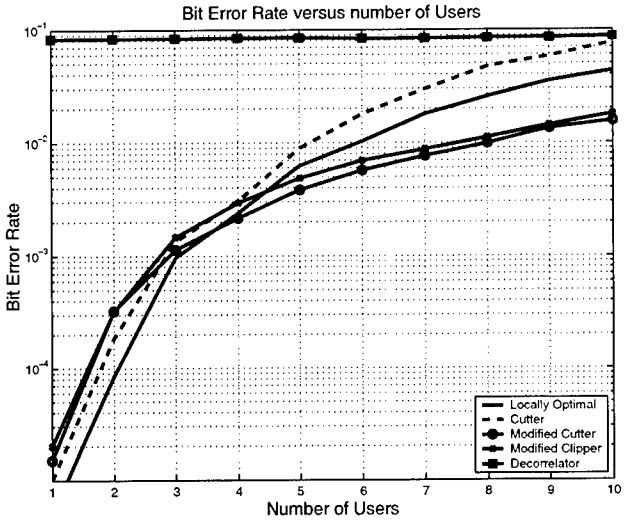


Fig. 3. Bit Error Rate versus the Number of Users

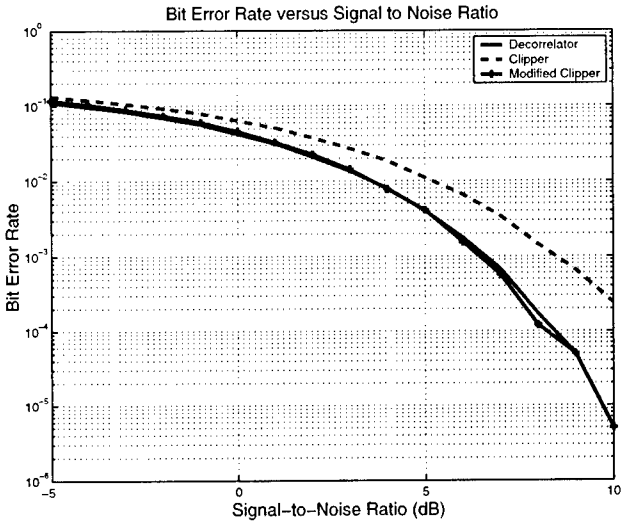


Fig. 5. Multiuser Detection in Gaussian Noise

4. MODIFIED TWO-STAGE NON-LINEAR MULTIUSER DETECTION

In this section we propose, and demonstrate the performance of, a modified version of the two-stage non-linear detector in [1]. Firstly, the modification, based on the modified non-linear multiuser detector proposed in Section 3, is shown and explained, and secondly, simulation results comparing its performance with the original two-stage non-linear detector and other detectors are presented.

4.1. Modified Two-Stage Non-Linear Detection Model

In [1] a two-stage non-linear detection scheme was proposed whereby the first stage consisted of a conventional non-linear multiuser detection scheme which estimated the MAI and removed it from the signal, and the second stage consisted of a single user conventional non-linear detector. A modification to this detector is proposed for two reasons; firstly, the problem with the near-far resistance in the modified non-linear multiuser detector can be reduced. Secondly, the modified non-linear detector has better BER performance in a high user system than the conventional non-linear scheme, and therefore the performance of the first stage of the two-stage non-linear detector can be improved using the modified non-linear scheme. It is also helpful since optimum performance in the modified scheme does not require a locally optimum non-linearity.

The modified two-stage non-linear multiuser detector can be seen in Figure 6. The first stage uses the modified detection model to estimate the MAI and then removes it from the received signal, leaving a single user detection problem, in impulsive noise. The conventional non-linear detector was used in the second stage since it performs better than the modified one in the single user case.

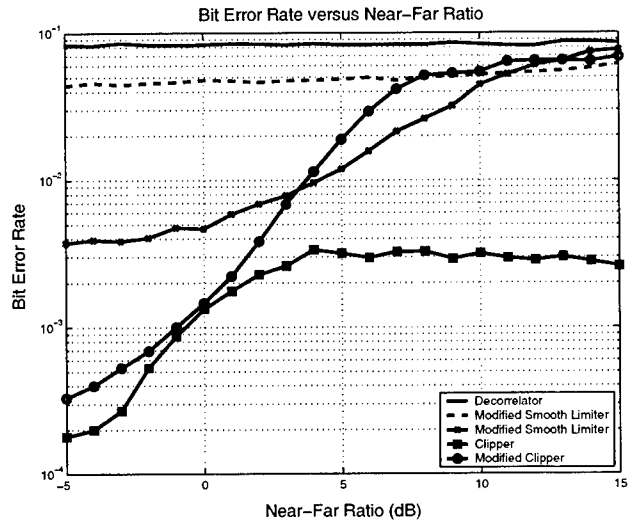


Fig. 4. Near-Far Characteristics

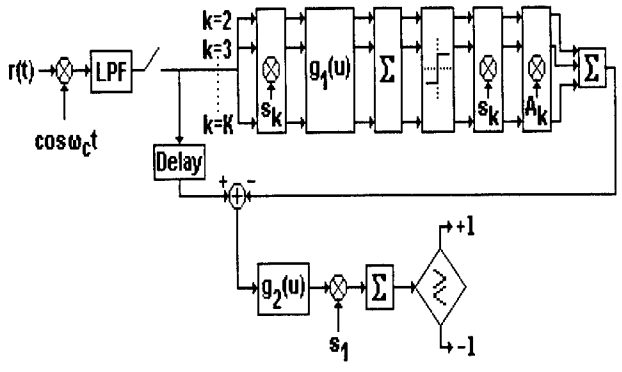


Fig. 6. Modified Two-Stage Non-Linear Detector

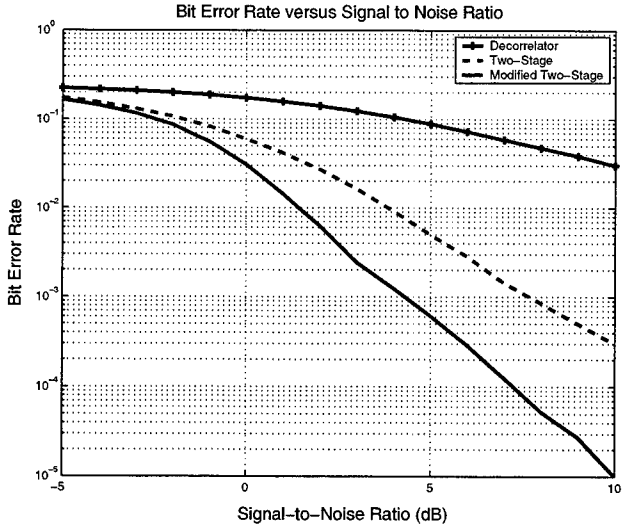


Fig. 7. Multiuser Detection in Cauchy Noise

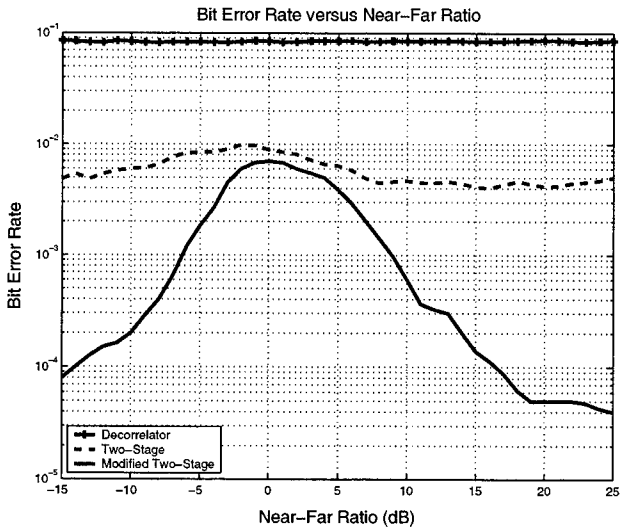


Fig. 8. Near-Far Performance Comparison

4.2. Simulations and Results

The choice of the non-linearities was as follows. The clipper was chosen for $g_1(u)$ due to its good MAI rejection capabilities and simplicity, and therefore is useful for detecting individual users signals in MAI. Since the second stage is essentially the detection of a single user in impulsive noise, a locally optimal non-linearity would give the optimum results. However, since it is unlikely that the exact noise model would be known at the receiver, a locally sub-optimum non-linearity is used. The clipper is also chosen for $g_2(u)$ since it performs very closely to the locally optimum case in a single user environment [6].

Figure 7 shows the results of simulations with $K = 6$ users and $NFR = 10dB$. The performance gain using the model of Figure 6 can be clearly seen. Figure 8 compares the near-far characteristics of the detectors, for $SNR \approx 5dB$ with $K = 6$ users. It can be seen that the near-far characteristic is similar to a decision-feedback detector. The modified two-stage non-linear detector outperforms the two-stage non-linear detector at all levels of the NFR, increasingly as the magnitude of the NFR increases.

5. CONCLUSIONS

The modified detection schemes investigated was shown to improve performance in terms of MAI rejection over the conventional non-linear detection schemes, in multiuser environments with impulsive noise. The scheme involves placing the non-linearity between the multiplier and the summation, rather than before the multiplier. This gain is achieved with little or no increase in computational cost and the requirement for locally optimum non-linearities is alleviated. The proposed model was also shown to improve performance in a Gaussian noise environment, over conventional non-linear detection techniques. Near-far characteristics were also improved by using the modified two-stage approach.

6. ACKNOWLEDGEMENTS

This work has been supported by the Australian Telecommunications Co-operative Research Centre under Program 3.5, "Multiuser Detection in W-CDMA"

7. REFERENCES

[1] J.F. Weng *et al*, *Two-Stage Nonlinear Detector in DS/SSMA Communications with Impulse Noise*, IEE Proceedings on Communications, Vol. 144, No. 6, December 1997, pp 387-394.

[2] S. Verdu, *Multiuser Detection*, Cambridge University Press, 1998.

[3] H.V. Poor and X. Wang, *Robust Multiuser Detection in Non-Gaussian Channels*, IEEE Transactions on Signal Processing, Vol. 47, No. 2, February 1999, pp 289-305.

[4] S.A. Kassam, *Signal Detection in Non-Gaussian Noise*, Springer-Verlag, 1988.

[5] C.L. Brown and A.M. Zoubir, *A Non-Parametric Approach to Signal Detection in Impulsive Interference*, IEEE Transactions on Signal Processing, Vol. 48, No. 9, September 2000, pp 2665-2669.

[6] A. Lane-Glover, *Multiuser Detection in Impulsive Noise*, Technical Report, Curtin University of Technology, 2001.

SUBSPACE-BASED BLIND ADAPTIVE MULTIUSER DETECTION FOR MULTIRATE DS/CDMA SIGNALS

L. Huang, F.-C. Zheng, M. Faulkner

School of Communications and Informatics
Victoria University of Technology
Melbourne, Vic 8001, Australia

ABSTRACT

The existing dual-rate blind linear detectors, which operate at either the low-rate (LR) or the high-rate (HR) mode, are not strictly blind at the HR mode and lack theoretical analysis. This paper proposes the subspace-based LR and HR blind linear detectors, i.e., blind decorrelating detectors (BDD) and blind MMSE detectors (BMMSED), for synchronous DS/CDMA systems. To detect an LR data bit at the HR mode, an effective weighting strategy is proposed. The theoretical analyses on the performances of the proposed detectors are carried out. It has been proved that the bit-error-rate of the LR-BDD is superior to that of the HR-BDD and the near-far resistance of the LR blind linear detectors outperforms that of its HR counterparts. The extension to asynchronous systems is also described. Simulation results show that the adaptive dual-rate BMMSED outperform the corresponding non-blind dual-rate decorrelators proposed in [2].

1. INTRODUCTION

The 3rd-generation wireless communications systems must be able to accommodate the heterogeneous traffic, such as voice, data and video, which have the different data rates and requirements of quality of service. This makes it imperative to develop multirate DS/CDMA receivers. As a result, several receivers originally proposed for single-rate DS/CDMA systems have been investigated for their use in multirate cases. Typical examples are the low-rate decorrelator (LRD) and the high-rate decorrelator (HRD) for dual-rate synchronous systems [1,2], which match to the bit interval of the low-rate (LR) users and the high-rate (HR) users, respectively. To overcome their requirements of the prior knowledge of the interfering users, the LR and HR blind linear MMSE detectors (LMMSED) were proposed for DS/CDMA systems in [3] and [4]. However, the performances of LR-LMMSED and HR-LMMSED are compared only by the numerical simulations. Also, to decode an LR data bit at the HR mode, the signal to interference-plus-noise ratio (SINR) of this LR user within each subinterval, which involves the knowledge of the noise level and the interfering users, is required to weight the partial result obtained in the corresponding subinterval. Thus, HR-LMMSED is not strictly blind.

This paper proposes the subspace-based LR and HR blind decorrelating detectors (BDD) and blind MMSE detectors (BMMSED), inspired by the single-rate counterpart [5]. A blind weighting scheme is proposed to detect the LR data bits at the HR mode. The theoretical analyses of the proposed blind detectors are carried out. The extension to asynchronous systems is also described.

2. SIGNAL MODEL

In this work, the baseband signal is assumed to be dual-rate binary DS/CDMA with variable spreading factor (VSF). However, all the results presented in this paper can be easily generalized to a general multirate system. We also assume that the system is synchronous in the formulation of the problem. The asynchronous case will be discussed in Subsection 3.4. We denote the processing gain of the LR users as N_0 and that of the HR users as N_1, where $N_0/N_1=M$ is an integer. The normalized signatures are denoted by $\underline{s}_{k,0}$ and $\underline{s}_{k,1}$ for the kth LR and HR user, respectively.

In a single LR bit interval, each HR user can be viewed as M virtual LR users. The kth LR user transmits bit $b_{k,0}$ with received amplitude $A_{k,0}$. The mth virtual user corresponding to the kth HR user transmits bit $b_{k,1}(m)$ with the received amplitude $A_{k,1}$ using the signature sequence $\underline{s}_{k,1}^{(m)}$, which is equal to $\underline{s}_{k,1}$ in the mth subinterval and otherwise is zero. The received signal in a single LR bit interval can be written as an N_0-vector

$$\begin{aligned}\underline{r} &= \sum_{k=1}^{K_0} A_{k,0} b_{k,0} \underline{s}_{k,0} + \sum_{k=1}^{K_1} \left\{ \sum_{m=0}^{M-1} A_{k,1} b_{k,1}(m) \underline{s}_{k,1}^{(m)} \right\} + \underline{n} \\ &= \mathbf{S}_0 \mathbf{A}_0 \underline{b}_0 + \mathbf{S}_1 \mathbf{A}_1 \underline{b}_1 + \underline{n} \\ &= \begin{bmatrix} \mathbf{S}_0 & \mathbf{S}_1 \end{bmatrix} \begin{bmatrix} \mathbf{A}_0 & 0 \\ 0 & \mathbf{A}_1 \end{bmatrix} \begin{bmatrix} \underline{b}_0 \\ \underline{b}_1 \end{bmatrix} + \underline{n} \\ &= \mathbf{S}\mathbf{A}\underline{b} + \underline{n} \end{aligned}, \quad (1)$$

where $\mathbf{S}_0$ consists of the signatures of K_0 LR users. $\mathbf{A}_0$=diag $\{A_{1,0},\cdots,A_{K_0,0}\}$ and $\underline{b}_0=[b_{1,0}\cdots b_{K_0,0}]^T$. $\mathbf{S}_1^{(m)}=[\underline{s}_{1,1}^{(m)}\cdots\underline{s}_{K_1,1}^{(m)}]$. $\mathbf{A}_1$= diag $\{\underbrace{\tilde{\mathbf{A}}_1,\cdots,\tilde{\mathbf{A}}_1}_{M}\}$ where $\tilde{\mathbf{A}}_1=\text{diag}\{A_{1,1},\cdots,A_{K_1,1}\}$, and $\underline{b}_1=[\underline{\tilde{b}}_{0,1}^T\cdots\underline{\tilde{b}}_{M-1,1}^T]^T$ where $\{\underline{\tilde{b}}_{m,1}^T\}_{m=0}^{M-1}$ is ordered as $\underline{\tilde{b}}_{m,1}^T=[b_{1,1}(m)\cdots b_{K_1,1}(m)]$. $\underline{n}$ is an additive white Gaussian noise (AWGN) vector with the covariance matrix $\sigma^2\mathbf{I}_{N_0}$. In addition, the received signal can also be modeled in each subinterval as an N_1-vector

$$\underline{r}^{(m)} = \mathbf{S}^{(m)}\tilde{\mathbf{A}}\underline{b}^{(m)} + \underline{n}^{(m)},\ 0 \le m \le M-1, \quad (2)$$

where $\mathbf{S}^{(m)}=[\tilde{\mathbf{S}}_0^{(m)}\ \tilde{\mathbf{S}}_1]$, and $\tilde{\mathbf{S}}_0^{(m)}$ is the portion of $\mathbf{S}_0$ within the mth subinterval, and $\tilde{\mathbf{S}}_1$ consists of the signatures of K_1 HR users. $\underline{b}^{(m)}=[\underline{b}_0^T\ \underline{\tilde{b}}_{m,1}^T]^T$, and $\tilde{\mathbf{A}}=\text{diag}\{\mathbf{A}_0,\tilde{\mathbf{A}}_1\}$. $\underline{n}^{(m)}$ is the corresponding noise vector.

3. DUAL-RATE BLIND LINEAR DETECTORS

This section will develop the LR-BDD, LR-BMMSED, HR-BDD, and HR-BMMSED.

3.1 The LR Blind Linear Detectors

In terms of model (1), in a single LR bit interval, the dual-rate system with K_0 LR users and K_1 HR users is equivalent to a single-rate system with $K_L=K_0+MK_1$ users. For convenience, we assume that the data bit, the signature and the received amplitude of the kth user are represented by b_k, $\underline{s}_k$ and A_k, respectively, whose physical meanings can be readily understood via (1).

Without loss of generality, we assume that $\text{rank}(\mathbf{S}) = K_L$. By performing an eigendecomposition, the covariance matrix of the received signal $\underline{r}$ can be represented by

$$\mathbf{C} = E\{\underline{r}\underline{r}^T\} = \mathbf{E}_s\mathbf{Y}_s\mathbf{E}_s^T + \mathbf{E}_n\mathbf{Y}_n\mathbf{E}_n^T, \quad (3)$$

where $\mathbf{E}_s$ is an orthonormal basis of the signal subspace and $\mathbf{E}_n$ is that of the noise subspace orthogonal to $\mathbf{E}_s$. $\mathbf{Y}_s$ contains the K_L largest eigenvalues of $\mathbf{C}$ and $\mathbf{Y}_n = \sigma^2\mathbf{I}_{N_0-K_L}$. Based on the subspace parameters, a linear detector for demodulating the kth user's data bit can be written as (by following the single-rate case in [5]):

$$\hat{b}_k = \text{sgn}(\underline{d}_k^T\underline{r}), \quad (4)$$

where

$$\underline{d}_k = \mathbf{D}\underline{s}_k \Big/ \underline{s}_k^T\mathbf{D}\underline{s}_k, \quad (5)$$

$$\mathbf{D} = \begin{cases} \mathbf{E}_s(\mathbf{Y}_s - \sigma^2\mathbf{I}_{K_L})^{-1}\mathbf{E}_s^T, & \text{LR-BDD} \\ \mathbf{E}_s\mathbf{Y}_s^{-1}\mathbf{E}_s^T, & \text{LR-BMMSED} \end{cases}. \quad (6)$$

Clearly, the two LR blind linear detectors become identical as $\sigma \to 0$. The scalar constant $1/\underline{s}_k^T\mathbf{D}\underline{s}_k$ has no effect on the signal detection and thus can be removed. Note that the implicit assumption that the exact signal covariance matrix and thus its eigencomponents are known is impractical. Generally, the subspace parameters must be estimated from the received signals using batch eigenvalue decomposition of sample covariance matrix, or batch singular value decomposition of sample matrix, or adaptive subspace tracking algorithms. For LR-BDD, the BER of the kth user can be expressed as (by following the single-rate case in [5])

$$P_k = Q\left(A_k \Big/ \sigma\sqrt{[\mathbf{R}^{-1}]_{k,k}}\right), \quad (7)$$

where $\mathbf{R}=\mathbf{S}^T\mathbf{S}$ and $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)dt$. The NFR $[\bar{\eta}_k]^L$ is thus

$$[\bar{\eta}_k]^L = 1/[\mathbf{R}^{-1}]_{k,k}. \quad (8)$$

LR-BMMSED has the same NFR as LR-BDD.

We also developed an adaptive algorithm for the estimation of LR-BMMSED. The signal subspace is first updated by the orthonormal PAST (OPAST) algorithm [6], and then LR-BMMSED can be estimated based on the updated signal subspace and some intermediate variables of the OPAST algorithm. Due to lack of space, the details of this adaptive algorithm are not given in this paper.

3.2 The HR Blind Linear Detectors

In terms of model (2), in the mth subinterval, each HR user k transmits one bit using the signature $\underline{s}_{k,1}$. Each LR user k transmits the mth segment $\underline{s}_{k,0}^{(m)}$ of the signature $\underline{s}_{k,0}$. Thus, it is equivalent to $K_H=K_0+K_1$ HR users simultaneously transmitting one data bit in each subinterval. For convenience, we enumerate all active users such that the kth LR user is numbered k while the kth HR user is numbered (K_0+k).

We assume that $\text{rank}(\tilde{\mathbf{S}}^{(m)}) = K_H$. The eigendecomposition is performed on the covariance matrix of the received signal $\underline{r}^{(m)}$ and the subspace parameters such as $\mathbf{E}_s^{(m)}$ and $\mathbf{Y}_s^{(m)}$ can be obtained. In the mth subinterval, a linear detector for detecting the kth HR user's data bit can be written as

$$\hat{b}_{k,1}(m) = \text{sgn}(\underline{d}_{k,1}^{(m)T}\underline{r}^{(m)}), \quad (9)$$

where

$$\underline{d}_{k,1}^{(m)} = \mathbf{D}^{(m)}\underline{s}_{k,1} \Big/ \underline{s}_{k,1}^T\mathbf{D}^{(m)}\underline{s}_{k,1}, \quad (10)$$

$$\mathbf{D}^{(m)} = \begin{cases} \mathbf{E}_s^{(m)}(\mathbf{Y}_s^{(m)} - \sigma^2\mathbf{I}_{K_H})^{-1}\mathbf{E}_s^{(m)T}, & \text{HR-BDD} \\ \mathbf{E}_s^{(m)}\mathbf{Y}_s^{(m)-1}\mathbf{E}_s^{(m)\text{T}}, & \text{HR-BMMSED} \end{cases}. \quad (11)$$

The scalar constant $1/\underline{s}_{k,1}^T\mathbf{D}^{(m)}\underline{s}_{k,1}$ can also be dropped. For HR-BDD, the BER of HR user k in the mth subinterval can be determined by

$$\hat{P}_{k,1}^{(m)} = Q\left(A_{k,1} \Big/ \sigma\sqrt{[\mathbf{R}^{(m)-1}]_{K_0+k,K_0+k}}\right), \quad (12)$$

where $\mathbf{R}^{(m)}=\mathbf{S}^{(m)T}\mathbf{S}^{(m)}$. The NFR $[\bar{\eta}_{k,1}^{(m)}]^H$ for the HR blind linear detectors is

$$[\bar{\eta}_{k,1}^{(m)}]^H = 1/[\mathbf{R}^{(m)-1}]_{K_0+k,K_0+k}. \quad (13)$$

Since the duration of an LR data bit spans M HR bit intervals, the following decision rule is used to estimate a data bit of LR user k:

$$\hat{b}_{k,0} = \text{sgn}\left[\sum_{m=0}^{M-1}\left(\underline{d}_{k,0}^{(m)T}\underline{r}^{(m)} \Big/ \underline{d}_{k,0}^{(m)T}\underline{d}_{k,0}^{(m)}\right)\right], \quad (14)$$

where

$$\underline{d}_{k,0}^{(m)} = \mathbf{D}^{(m)}\underline{s}_{k,0}^{(m)} \Big/ \underline{s}_{k,0}^{(m)T}\mathbf{D}^{(m)}\underline{s}_{k,0}^{(m)}. \quad (15)$$

Note that the employed weighting factors are the reciprocal of detector coefficients' energy (proportional to the output noise power), and need no knowledge of the noise level and the interfering users. This means that the contribution from a subinterval to the decision is inversely proportional to the output noise power in this subinterval. This strategy is reasonable since the output noise level is dominant over or comparable to multiple-access interference (MAI) after multiuser detection is applied, which is particularly true for HR-BDD where the MAI is

completely suppressed. For HR-BDD, since $\underline{d}_{k,0}^{(m)T}\underline{s}_{k,0}^{(m)}=1$, $\underline{d}_{k,0}^{(m)T}\underline{s}_{j,0}^{(m)}=0\ (j\neq k)$ and $\underline{d}_{k,0}^{(m)T}\underline{d}_{k,0}^{(m)}=[\mathbf{R}^{(m)-1}]_{k,k}$ [5], the BER of LR user k can be given by

$$\hat{P}_{k,0}=Q\left(\frac{A_{k,0}}{\sigma}\sqrt{\sum_{m=0}^{M-1}1/[\mathbf{R}^{(m)-1}]_{k,k}}\right). \tag{16}$$

The NFR $[\bar{\eta}_{k,0}]^H$ for the HR blind linear detectors is then

$$[\bar{\eta}_{k,0}]^H=\sum_{m=0}^{M-1}1/[\mathbf{R}^{(m)-1}]_{k,k}\,. \tag{17}$$

Note that the adaptive HR-BMMSED can also be derived in a similar manner to the adaptive LR-BMMSED.

3.3 A Comparison of LR and HR Blind Linear Detectors

It is easy to see that the use of the LR blind linear detectors incurs a detection delay for the HR users. In addition, the LR and HR blind linear detectors involve the computation of the subspace parameters. Also, the computational complexity of the former is much higher than that of the latter as the ratio M increases. We have the following proposition for the BER performance of LR-BDD and HR-BDD.

Proposition 1: For a general dual-rate synchronous system, if the exact subspace parameters are known, then

$$P_{b_{k,r}(m)}^{HR-BDD}\geq P_{b_{k,r}(m)}^{LR-BDD},\ r\in\{0,1\}, \tag{18}$$

where $P_{b_{k,r}(m)}^{HR-BDD}$ and $P_{b_{k,r}(m)}^{LR-BDD}$ are the BERs of HR-BDD and LR-BDD, respectively, for the LR or HR user's bits in any subinterval. Especially, both achieve the same BER for each LR user if the signatures for the LR users are the same in every subinterval, i.e., if the repetition code is employed.

Proof: Note that the proposed dual-rate BDD have the same BER expression as the corresponding dual-rate non-blind ones [1,2], which involve a special diagonal element of the inverse of signature correlation matrix $\mathbf{R}^{-1}$ or $\mathbf{R}^{(m)-1}$. They differ in that this element for the latter can be accurately calculated using the known signatures of all the users, while it can only be approximately estimated from the received signal for the former. Thus, if the exact subspace parameters are known, the former should have same BER as the latter. Since the similar proposition has been proven for the latter [1,2], inequality (18) must hold as well.

Based on the above proposition and the fact that Q function is monotonically descending, we can conclude that the NFR of the LR blind linear detectors is not inferior to that of its HR rivals. Furthermore, the above proposition and inference can be extended to the practical multirate scenario since a similar extension for dual-rate non-blind decorrelators has been proven in [1].

3.4 Asynchronous Case

Both LR and HR blind linear detectors can be applied to asynchronous systems. The same formulae in the synchronous case can be used for asynchronous systems with an increased dimension due to the fact that the number of virtual bits and their associated virtual signatures within the processing interval is larger than that in the synchronous case [3]. At the LR mode, if the desired user is an LR user, the desired LR user is viewed as the reference user, whose bit interval is taken as the processing interval. Otherwise an arbitrary LR user can be chosen as the reference user. For each user other than the reference one, there might be two virtual bits located at both ends of the processing interval, whose full-length signatures are partitioned into two virtual ones. Therefore, within the processing interval, the number of data bits (including the actual and virtual bits) is between K_0+MK_1 and $2K_0+(M+1)K_1-1$. To detect the data bit of the desired HR user, which is divided into two virtual ones, a similar strategy to (14) can be used to weight the partial estimates over two successive processing intervals before a final decision is made. It should be noted that at the LR mode, the number of data bits and their associated signatures are the same within the different processing intervals.

At the HR mode, if the desired user is an LR user, an arbitrary HR user can be chosen as the reference user, otherwise the desired HR user is viewed as the reference user. For each HR user other than the reference one, there might be two virtual bits located at both ends of the processing interval with full-length signatures being segmented into two virtual ones. For each LR user, there are either one or two virtual bits, whose associated virtual signatures are only a portion of the full-length signatures. Therefore, the number of data bits is between K_0+K_1 and $2K_0+2K_1-1$. A similar weighting strategy to (14) can be used to demodulate a data bit of the desired LR user. Note that at the HR mode, the number of the data bits and their associated signatures might change among the different processing intervals. Therefore, special attention should be paid to the implementation of the HR blind linear detectors for asynchronous systems.

Based on the results in the previous sections, it is obvious that the BER (or NFR) performances of the dual-rate blind linear detectors are related to the signature correlation matrix in the underlying processing interval. Considering that the signature correlation matrix depends on the relative delays of all the users and the choice of the reference user, it is impractical to make a theoretical comparison on the performances of the LR and HR blind linear detectors. However, by the numeric simulations in Section 4, it is shown that Proposition 1 may be invalid for a general dual-rate asynchronous system.

4. SIMULATIONS AND CONCLUSIONS

We assume that there are four active users in our simulations. The first two users are the LR users (K_0=2) and the others are the HR users (K_1=2). Other basic parameters are N_1=7, N_0=14 and M=2. Four Gold sequences with length 7 are intentionally chosen such that we have the worst cross-correlation of all signatures. The first two sequences are used by the LR users, whose signatures are generated via the repetition-code scheme proposed in [2]. The other sequences are assigned to the HR users. We assume that User 1 is the desired LR user and User 4 is the desired HR user. For Fig.s 1-2, the signal-to-noise ratio (SNR) of the desired user is fixed at 8dB and the SNRs of all the other

users change simultaneously. For asynchronous case, the delays for all the users are 0, 3, 1, and 3 chips, respectively.

We first investigate the performance of the proposed dual-rate blind linear detectors in both synchronous and asynchronous cases. The EVD is used to estimate the subspace parameters. As proven in Proposition 1, Fig. 1 shows that in synchronous case, the LR-BDD outperforms the HR-BDD for the HR users and they offer the same performance for the LR users. Note that in asynchronous case, the HR-BDD outperforms the LR-BDD for the HR users. This means that Proposition 1 is invalid for the underlying asynchronous case. In addition, it is obvious that the BMMSED are upper-bounded by the corresponding BDD. The performances of the adaptive dual-rate BMMSED in synchronous case are also evaluated. Fig. 2 indicates the BERs of adaptive dual-rate BMMSED after the adaptive algorithms have converged. As a comparison, the performances of the non-blind LRD and HRD are also plotted. It is shown that the BER performances of the adaptive dual-rate BMMSED are superior to that of the corresponding non-blind dual-rate decorrelators. Further work will focus on the multirate blind detectors for multipath fading channels.

5. REFERENCES

[1] J.X. Chen and U. Mitra, "Analysis of decorrelator-based receivers for multirate DS/CDMA communications," *IEEE Trans. Vehicular Technology*, vol. 48, pp. 1966-1983, 1999.

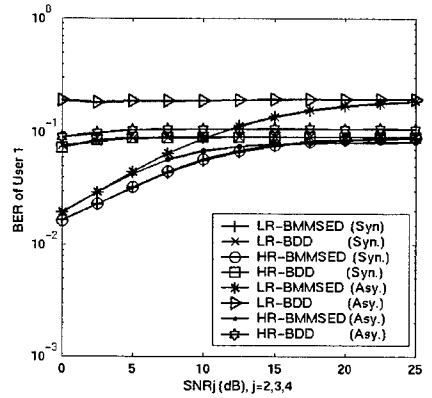
[2] M. Saquib, R. Yates, and N. Mandayam, "Decorrelating detectors for a dual rate synchronous DS/CDMA system," *Wireless Personal Communications*, vol. 9, pp. 197-216, 1998.

[3] H. Ge and J.W. Ma, "Multirate LMMSE detectors for asynchronous multi-rate CDMA systems," *Proc of ICC'98*, pp 714-18, vol.2, June 1998.

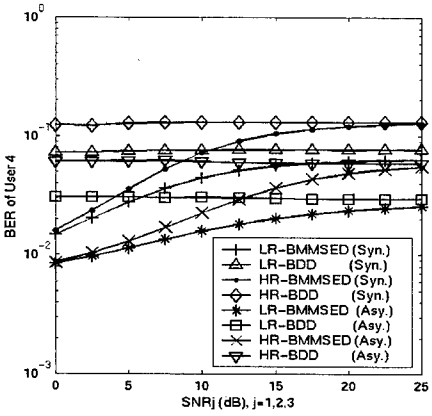
[4] H. Ge, "Multiuser detection for integrated multi-rate CDMA," *Proc. of the first int'l conf. on information, communications and signal processing*, pp 858-862, Singapore, Sep., 1997.

[5] X.D. Wang and H.V. Poor, "Blind multiuser detection: a subspace approach", *IEEE Trans. Information Theory*, vol. 44, pp. 677-690, 1998.

[6] K.A. Meraim, A. Chkeif, and Y. Hua, "Fast orthonormal PAST algorithm," *IEEE Signal Processing Letters*, vol. 7, pp. 60--62, 2000.

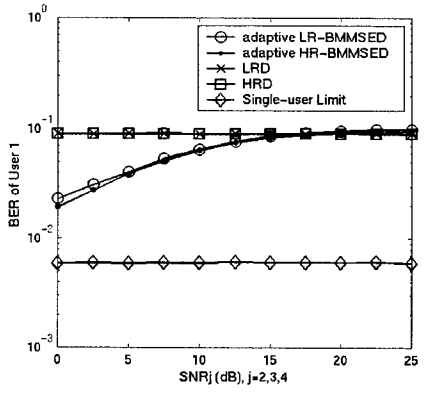


(a)

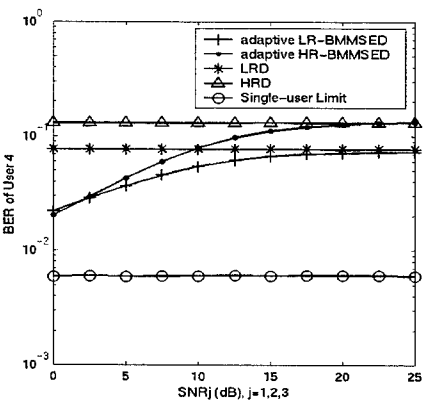


(b)

Figure 1. The BER curves of User 1 (a) amd User 4 (b) versus the SNRs of the other users



(a)



(b)

Figure 2. The BER curves of User 1 (a) and User 4 (b) versus the SNRs of the other users

A DECISION FEEDBACK CDMA RECEIVER WITH PARTIALLY ADAPTIVE INTERFERENCE SUPPRESSION

Gau-Joe Lin, Ta-Sung Lee and Chan-Choo Tan

Department of Communication Engineering and
Microelectronics and Information Systems Research Center
National Chiao Tung University
Hsinchu, Taiwan, R.O.C.
e-mail: lgj.cm88g@nctu.edu.tw

ABSTRACT

A novel CDMA receiver with adaptive multiple access interference (MAI) and narrowband interference (NBI) suppression is proposed for reverse link signal reception over multipath channels. The design of the receiver involves the following procedure. First, adaptive correlators are constructed based on the generalized sidelobe canceller (GSC) scheme to collect the multipath signals and suppress strong MAI and time-varying NBI. In particular, partial adaptivity is incorporated into the GSC for reduced complexity processing. A simple combiner with channel estimation then gives the symbol decisions. In order to enhance the performance of the adaptive correlators, a decision aided scheme is introduced which subtracts the reconstructed signal from the input data of the GSC, leading to improved output SINR and convergence. The proposed CDMA receiver is evaluated through simulations, and the results show that it can outperform the conventional MMSE receiver designed for time-invariant interference.

1. INTRODUCTION

A major limiting factor for a CDMA system is the multiple access interference (MAI). MAI causes the near-far problem and a reduction in system capacity. Adaptive multiuser detectors and interference cancellers have been suggested which provide immunity to the near-far effect [1]. On the other hand, time-varying narrowband interference (NBI) due to the presence of overlay systems can be treated as a group of equivalent MAI, and suppressed with a large adaptive degree of freedom [3]. With fully adaptive (FA) implementation, an adaptive CDMA receiver requires $(M-1)$-dim processing, where M is the processing gain. To reduce the complexity, partially adaptive (PA) implementation is suggested as an alternative in which the size of the adaptive weights is reduced by judiciously designed criteria [2]. The advantages of PA implementation include reduced complexity and faster convergence. However, performance of interference cancellation usually degrades as a result of a smaller processing dimension. A trade-off must thus be reached between complexity and performance.

In this paper, a novel CDMA receiver with enhanced joint MAI and NBI suppression is proposed for a reverse link pilot symbol-assisted system over multipath channels. The development of the new scheme involves the following procedure. First, a set of adaptive correlators are constructed based on the generalized sidelobe canceller (GSC) [4] scheme to collect the multipath signals, suppress strong MAI and time-varying NBI. In particular, partial adaptivity is incorporated into the GSC for reduced complexity MAI and NBI suppression. This is achieved by selecting a reduced dimension subspace of the column space of the blocking matrix of the GSC encompassing the MAI signatures obtained in a multi-user scenario and NBI signatures obtained in overlay applications. Next, a simple combiner with channel estimation gives the symbol decisions. In order to further enhance the convergence performance of the GSC adaptive correlators, a decision aided scheme is introduced in which the signal waveform is first estimated and then subtracted from the input data of the correlators. The proposed CDMA receiver is evaluated through simulations, and the results show that it can outperform the optimal MMSE receiver designed for time-invariant interference scenarios.

This work was sponsored jointly by the Ministry of Education and National Science Council, R.O.C, under the Contract 89-E-FA06-2-4

2. DATA MODEL

Suppose that there are K active users in a CDMA system. The kth user's contribution to the received signal can be written as

$$d_k(t) = \sum_i b_k(i) s_k(t - iT) \tag{1}$$

where $b_k(i)$ denotes the ith transmitted information bit, T is the bit duration, and $s_k(t)$ is the signature waveform given by

$$s_k(t) = \sum_{m=0}^{M-1} c_k[m] p(t - mT_c) \tag{2}$$

where $c_k[m]$ is the spreading sequence of the kth user, M is the spreading factor, $p(t)$ is the chip waveform, and T_c is the chip duration. The transmission channel is modeled as with L resolved Rayleigh fading paths. Putting the K user signals together, the received baseband data can be expressed in the following form:

$$x(t) = \sum_{k=1}^{K} \sum_{l=1}^{L} \alpha_{k,l} d_k(t - \tau_{k,l}) + i(t) + n(t) \tag{3}$$

where $\tau_{k,l}$ and $\alpha_{k,l}$ are the delay and complex gain, respectively, of the lth path of the kth user. The $i(t)$ is the NBI and $n(t)$ is the additive white noise with power σ_n^2. To fully exploit the temporal signature, $x(t)$ is chip matched filtered and sampled. Assuming

user 1 to be the desired user, the resulting chip-sampled data over the ith symbol can be put into the $(M+L-1)\times 1$ vector:

$$\begin{aligned}\mathbf{x}(i) &= [x(0), x(1), \cdots, x(M+L-2)]^T \\ &= \sum_{l=1}^{L} \alpha_{1,l}\mathbf{c}_{\tau_{1,l}} b_1(i) + \mathbf{i}(i) + \mathbf{n}(i) \end{aligned} \quad (4)$$

where $\mathbf{c}_{\tau_{1,l}}$ is the augmented signature vector associated with the lth path of user 1, $\mathbf{i}(i)$ is the interference vector, and $\mathbf{n}(i)$ is the noise vector. Depending on the delay $\tau_{1,l}$, $\mathbf{c}_{\tau_{1,l}}$ is given by one of the columns of the $(M+L-1)\times L$ matrix:

$$\mathbf{C} = [\mathbf{c}_{1,1}\mathbf{c}_{1,2}\cdots\mathbf{c}_{1,L}] \quad (5)$$

where $\mathbf{c}_{1,l}$ is the $(M+L-1)\times 1$ vector with $[c_k[0], c_k[1], \ldots, c_k[M]$ occupying the lth to $(l+M-1)$th entries. Note that $\mathbf{i}(i)$ includes ISI, NBI and MAI. For NBI, a more realistic assumption is to model it as a data-like signal, e.g., a BPSK signal with signaling rate much slower than the CDMA system chip rate [3]:

$$i(t) = A_I \sum_i b_I(i) p(t - iT_I) \quad (6)$$

where $b_I(i)$ denotes the ith NBI bit, T_I is the bit duration, and A_I is the complex gain. The bit rate $1/T_I$ is assumed Q times slower than the chip rate such that $T_I = QT_c$.

From (4), the effective "composite" signature vector of the kth user is given by:

$$\mathbf{h}_k = \sum_{l=1}^{L} \alpha_{k,l}\mathbf{c}_{\tau_{k,l}} \quad (7)$$

A receiver for user k is designed to identify and remove $\mathbf{h}_k$ to retrieve the data bits $b_k(i)$ from $\mathbf{i}(i)$ and $\mathbf{n}(i)$. In particular, a linear receiver combines $\mathbf{x}(i)$ using a weight vector $\mathbf{w}$ to obtain

$$b_k(i) = \mathbf{w}_k^H \mathbf{x}(i) \quad (8)$$

where H denotes the complex conjugate transpose.

3. PROPOSED RECEIVER

The proposed receiver is implemented with adaptivity and suitable for pilot symbol-assisted systems. Its overall schematic diagram is depicted in Figure 1. Without loss of generality, it is assumed that user 1 is the desired one and others are MAI.

3.1. GSC Realization of Adaptive Correlators

Conventionally, in order to restore the processing gain and retain the path diversity, $\mathbf{x}(i)$ is despread at each of the L fingers using a set of discrete-time correlators:

$$z_{1,l}(i) = \mathbf{w}_{1,l}^H \mathbf{h}_1 b_1(i) + \mathbf{w}_{1,l}^H \mathbf{i}(i) + \mathbf{w}_{1,l}^H \mathbf{n}(i) \quad (9)$$

for $l = 1, \ldots, L$, where $\mathbf{w}_{1,l}$ is the correlator weight vector at the lth finger. For an effective suppression of MAI, these weight vectors are determined in accordance with the linearly constrained minimum variance (LCMV) criterion. To avoid signal cancellation incurred with coherent multipaths, the LCMV correlators can be implemented in the form of GSC [4]. The concept of GSC is to decompose the weight vector $\mathbf{w}_{1,l}$ into two orthogonal components as $\mathbf{w}_{1,l} = \mathbf{c}_{1,l} - \mathbf{B}\mathbf{u}_{1,l}$. The matrix $\mathbf{B}$ is a is a "signal blocking" matrix which removes user 1's signal before filtering. Note that $\mathbf{B}$ must block signals from within the entire delay spread in order to avoid signal cancellation due to coherent multipaths. The goal is then to choose the adaptive weight vectors $\mathbf{u}_{1,l}$ to cancel the MAI and NBI. According to the GSC scheme, $\mathbf{u}_{1,l}$ is determined via the minimum mean square error (MMSE) criterion:

$$\begin{aligned}\min_{\mathbf{u}_{1,l}} \quad & E\{|\mathbf{c}_{1,l}^H \mathbf{x}(i) - \mathbf{u}_{1,l}^H \mathbf{B}^H \mathbf{x}(i)|^2\} \\ & \equiv \|\mathbf{R}_x^{1/2}\mathbf{B}\mathbf{u}_{1,l} - \mathbf{R}_x^{1/2}\mathbf{c}_{1,l}\|^2 \end{aligned} \quad (10)$$

where the data correlation matrix $\mathbf{R}_x = E\{\mathbf{x}(i)\mathbf{x}^H(i)\}$. Solving for $\mathbf{u}_{1,l}$ and putting $\mathbf{w}_{1,l}$'s in matrix form, we get

$$\begin{aligned}\mathbf{W} &= [\mathbf{w}_{1,1}, \mathbf{w}_{1,2}, \ldots, \mathbf{w}_{1,L}] \\ &= [\mathbf{I} - \mathbf{B}(\mathbf{B}^H\mathbf{R}_x\mathbf{B})^{-1}\mathbf{B}^H\mathbf{R}_x]\mathbf{C} \end{aligned} \quad (11)$$

The matrix $\mathbf{B}$ can be chosen to be a full rank $(M+L-1)\times(M-1)$ matrix whose columns are orthogonal to $\{\mathbf{c}_{1,1}, \ldots, \mathbf{c}_{1,L}\}$, i.e., $\mathbf{B}_l^H\mathbf{C} = \mathbf{O}$ and $\mathbf{B}^H\mathbf{B} = I$. With a large M, $\mathbf{u}_{1l}$ will have a large size too, leading to a high computational load and poor convergence in real-time implementation [4]. To alleviate this problem, the PA GSC is proposed which uses only a portion of the available degrees of freedom offered by the adaptive weights. Specifically, the PA techniques can be employed to reduce the size of $\mathbf{B}$.

3.2. Partially Adaptive Implementation

It is noteworthy from (10) that the optimal GSC weight vector $\mathbf{u}_{1,l}$ is the one lying in the subsapce $R\{\mathbf{R}_x^{1/2}\mathbf{B}\}$ that is closest to $\mathbf{R}_x^{1/2}\mathbf{c}_{1,l}$. In other words, $\mathbf{R}_x^{1/2}\mathbf{B}\mathbf{u}_{1,l}$ should be in the direction that exhibits maximum "correlation" with $\mathbf{R}_x^{1/2}\mathbf{c}_{1,l}$. It is therefore desired that the blocking matrix $\mathbf{B}$ be chosen such that the crosscorrelation $\rho = |\mathbf{c}_{1,l}^H\mathbf{R}_x\mathbf{B}\mathbf{u}_{1,l}|$ is large. Since $\mathbf{B}_l^H\mathbf{C} = \mathbf{O}$, the only way to maximize ρ is to retain as much MAI and NBI as possible. This in turn suggests that a suitable method for implementing the PA receiver is to find a reduced size $\mathbf{B}$ that can retain as much MAI and NBI as possible.

In a multi-user scenario, the MAI's effective signatures can be obtained by pilot symbol-assisted channel estimation and exploiting the corresponding spreading sequences. In particular, the effective signature vector of a user observed at the receiver is given by the convolution of the spreading codes with the FIR channel response. Accordingly, from (4), the estimated $(M+L-1)\times 1$ composite signature vector (CSV) of user i can be expressed as

$$\hat{\mathbf{h}}_i = \sum_{l=1}^{L} \hat{\alpha}_{i,l}\mathbf{c}_{\tau_{i,l}} \qquad i = 2, \ldots, K \quad (12)$$

with $i = 2, \ldots, K$, where $\hat{\alpha}_{i,l}$ is obtained by pilot symbol-assisted channel estimation at the lth finger. Given these CSV estimates, a reduced size blocking matrix $\mathbf{B}_{1m}$ can be constructed by projecting onto the column space of $\mathbf{B}$ the set of vectors $\{\hat{\mathbf{h}}_i\}$ readily obtained in a multi-user scenario. Therefore, the new blocking matrix of the PA GSC for MAI suppression can be obtained as

$$\hat{\mathbf{B}}_{1m} = \mathbf{B}\mathbf{B}^H\left[\hat{\mathbf{h}}_2 \cdots \hat{\mathbf{h}}_K\right] \quad (13)$$

On the other hand, the NBI signatures can be obtained in overlay applications. In particular, a narrowband linearly modulated signal can be treated as a group of related, virtual spread spectrum

signals with simple spreading codes [5]. Therefore, the NBI's effective signatures can be obtained by pre-estimating and exploiting these virtual spreading codes. Given these CSV estimates of NBI, a reduced size blocking matrix $\mathbf{B}_{1n}$ can be constructed by projecting onto the column space of $\mathbf{B}$ the set of virtual spreading codes, $\{\hat{\mathbf{g}}_i\}$, $i = 1, \ldots, \lfloor \frac{M+L-1}{Q} \rfloor + 1$. Therefore, the reduced size blocking matrix of the PA GSC for NBI suppression is given by

$$\hat{\mathbf{B}}_{1n} = \mathbf{B}\mathbf{B}^H \left[\hat{\mathbf{g}}_1 \cdots \hat{\mathbf{g}}_{\lfloor \frac{M+L-1}{Q} \rfloor + 1}\right] \quad (14)$$

For joint MAI and NBI supression, it is straightforward to construct the new blocking matrix $\hat{\mathbf{B}}_{1p} = \left[\hat{\mathbf{B}}_{1m}, \hat{\mathbf{B}}_{1n}\right]$. Note that $\mathbf{B}_{1p}$ can be regarded as the "smallest" blocking matrix with the number of columns equal to the number of "real" MAI and "virtual" MAI (due to NBI).

3.3. GSC Structure for Time-Varying NBI

Consider again the NBI in (6). It is noteworthy that if $G = T/T_I$ is an integer, then the NBI can be treated as G MAI's which can be suppressed with a time-invariant receiver. However, if the ratio G is not an integer, the detection rule would be time-varying. In general, it is plausible to assume that the ratio G is a rational number [3] such that the NBI sequence is a periodically time-varying one with a period equal to CT, where C is the smallest integer such that CT is an integer multiple of T_I. In other words, the NBI can be decomposed into C time-invariant parts, each representing G MAI's. This fact in turn implies that the data correlation matrix $\mathbf{R}_x$ is itself periodically time-varying with period C, thus resulting in the following periodically time-varying GSC processing:

$$\mathbf{w}_{1,1}(m) = [\mathbf{I} - \mathbf{B}(\mathbf{B}^H \mathbf{R}_x(m)\mathbf{B})^{-1}\mathbf{B}^H \mathbf{R}_x(m)]\mathbf{c}_{1,l} \quad (15)$$

for $m = 1, \ldots, C$, where $\mathbf{w}_{1,1}(m)$ is the weight vector of the mth part and $\mathbf{R}_x(m)$ is the correlation matrix constructed by collecting data samples over the $(qC + m)$th symbols, $q = 0, 1, 2, \ldots$. As a result, C different sets of correlator weights should be designed, with each set corresponding to a time-invariant component of the NBI. As shown in Figure 1, the input data is fed into the bank of C correlators, processed, and then combined back to a serial data stream corresponding to a finger.

3.4. RAKE Combining and Decision Aided Signal Reconstruction

With the time-varying PA correlator bank constructed, the next step is to perform a maximum ratio combining of the correlator outputs to collect the multipath energy. Since the MAI and NBI have been removed, channel estimation for the desired user can be done accurately, leading to improved performance as compared to the conventional RAKE receiver. However, the GSC is blind in nature and usually exhibits slow convergence due to the residual signal effect. To remedy this, a decision aided scheme is introduced in which the signal is estimated and then subtracted from the input data before the computation of GSC adaptive weights. First, assume that at the jth iteration, the mth part of the received data, $\mathbf{x}(m, i)$, is available and despread into:

$$\tilde{z}_{1,l}^{(j)}(m, i) = \mathbf{w}_{1,l}^{(j)}(m)^H \mathbf{x}(m, i) \quad (16)$$

where $\mathbf{w}_{1,l}^{(j)}(m)$ is estimated by (15), but using the "signal-subtracted" data $\mathbf{y}^{(j)}(m, i)$ as the input. With $\tilde{z}_{1,l}^{(j)}(m, i)$ available, we can obtain the channel estimate using a sequence of N_p pilot symbols:

$$\hat{\alpha}_{1,l}^{(j)}(m) = \frac{1}{N_p} \sum_{i=1}^{N_p} \tilde{z}_{1,l}^{(j)}(m, i) \quad (17)$$

With the channel estimate $\hat{\alpha}_{1,l}^{(j)}(m)$, the random phase of the lth finger output $\tilde{z}_{1,l}^{(j)}(m, i)$ is removed and coherent RAKE combining is achieved by

$$\hat{z}_1^{(j)}(m, i) = \sum_{l=1}^{L} \hat{\alpha}_{1,l}^{(j)*}(m) \tilde{z}_{1,l}^{(j)}(m, i) \quad (18)$$

which is then sent to the data decision device:

$$\hat{b}_1^{(j)}(m, i) = dec\{\hat{z}_1^{(j)}(m, i)\}$$
$$\hat{b}_1^{(j)}(k) = \text{P/S}[\hat{b}_1^{(j)}(m, i)] \quad k = (i-1)C + m \quad (19)$$

where P/S is parallel to serial transform. Second, signal reconstruction is done by exploiting the channel estimate $\hat{\alpha}_{1,l}^{(j)}(m)$, the desired user's signature $\mathbf{c}_{1,l}$ and data decisions $\hat{b}_1^{(j)}(i)$ as follows:

$$\hat{\mathbf{s}}_1^{(m,j)}(i) = \hat{b}_1^{(j)}(m, i) \sum_{l=1}^{L} \hat{\alpha}_{1,l}^{(j)}(m)\mathbf{c}_{1,l} \quad (20)$$

Finally, the reconstructed signal is subtracted from the data sent to the $j + 1$th iteration, which yields

$$\mathbf{y}^{(j+1)}(m, i) = \mathbf{x}(m, i) - \hat{\mathbf{s}}_1^{(j)}(m, i) \quad (21)$$

By using $\mathbf{y}^{(j+1)}(m, i)$ as the GSC input, the adverse slow convergence can be effectively improved. This above described procedure can be iterated several times to gain further improvement.

4. COMPUTER SIMULATIONS

As a performance index, we define the output SINR to be the ratio of the signal power to the interference-plus-noise power at the receiver output. The input SNR is defined to be $E\{|d_1(t)|^2\}/\sigma_n^2$ and the near-far-ratio (NFR) is the ratio of the MAI power to signal power before despreading. The path gains $\alpha_{k,j}$'s are assumed independent, identically distributed unit variance complex Gaussian random variables, the path delays $\tau_{k,j}$'s are assumed uniform over $[0, 3T_c]$, and the number of paths was $L = 4$ for all users. All simulations involved K equal power CDMA signals spread by the Gold code of length 31 and two equal power BPSK NBI's with $T_I = 12T_c$, $C = 6$ ($6 \times (31 + 4 - 1) = 17 \times 12$) and the ratio of the signal power to NBI power was -20 dB. The number of fingers was $L = 4$, and the input SNR was 0 dB. The PA dimension was chosen to be $P = (K - 1) + 4$, and each result was obtained by 100 independent trials. For comparison, we also included the results obtained with the MMSE [1], RAKE and proposed FA receivers. For all receivers, $N_p = 300$ pilot symbols were used for correlator weight vectors computation and channel estimation.

In Figure 2, with $K = 10$ and NFR = 10 dB, it is observed that the proposed PA receiver converges in three iterations, outperforms the MMSE receiver, and reaches the performance of the FA receiver. The system capacity is then evaluated in Figure 3

with NFR = 10 dB. Again, the proposed receiver outperforms the MMSE receiver and gives the performance of the FA receiver. Finally, the near-far resistance is evaluated in Figure 4 with $K = 10$. As observed, the proposed receiver achieves its excellent near-far resistance by successfully rejecting the MAI and NBI.

5. CONCLUSION

A decision aided receiver with partially adaptive interference suppression has been proposed. It is designed with the following procedure. First, a set of adaptive correlators implemented in the form of GSC is constructed to collect multipath signals, suppress strong MAI and time-varying NBI. In particular, partial adaptivity is incorporated into the GSC for reduced complexity which is achieved by selecting a reduced dimension subspace of the column space of the blocking matrix encompassing the MAI and NBI signatures. Next, a simple maximum ratio combiner gives the symbol decisions. In order to enhance the convergence performance of the adaptive correlators, a decision aided scheme is introduced which subtracts the reconstructed signal from the input of the GSC. The proposed CDMA receiver is evaluated through simulations, and the results show that it can outperform the conventional MMSE receiver designed for time-invariant interference.

6. REFERENCES

[1] G. Woodward and B. S. Vucetic, "Adaptive detection for DS-CDMA," *Proc. IEEE,* vol. 86, pp. 1413-1434, July 1998.

[2] E. G. Ström and S. L. Miller, "Properties of the single-bit single-user MMSE receiver for DS-CDMA systems," *IEEE Trans. Commun.,* vol. 47, pp. 416-425, March 1999.

[3] S. Buzzi, M. Lops and A. M. Tulino, "Time-varying narrow-band interference rejection in asynchronous multiuser DS/CDMA systems over frequency-selective fading channels," *IEEE Trans. Commun.,* vol. 47, pp. 1523-1536, October 1999.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, pp. 4-24, Apr. 1988.

[5] H. V. Poor, "Active interference supression in CDMA Overlay systems," *IEEE J. Select. Areas Commun.,* vol. 19, pp. 4-20, January 2001.

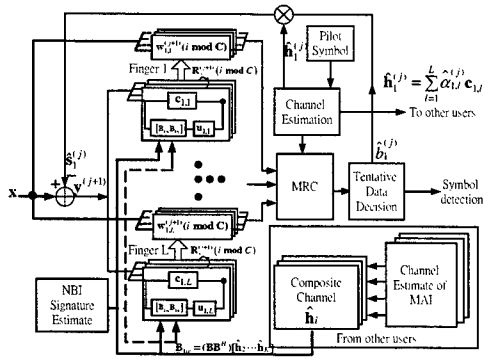


Figure 1: Structure of proposed CDMA receiver with partially adaptive interference suppression

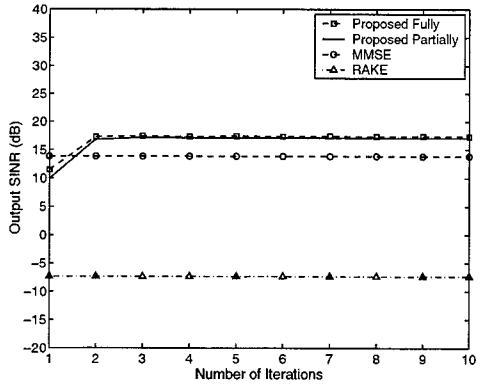


Figure 2: Output SINR versus iteration number with $K = 10$, NFR = 10 dB and SNR = 0 dB

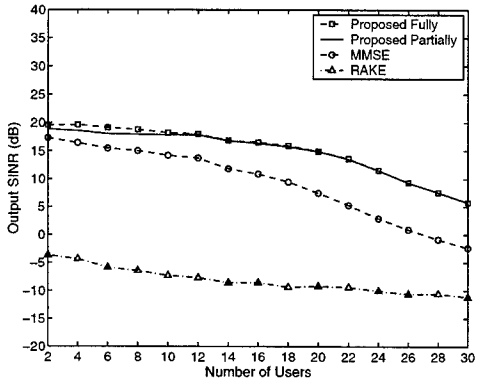


Figure 3: Output SINR versus user number with NFR = 10 dB and SNR= 0 dB

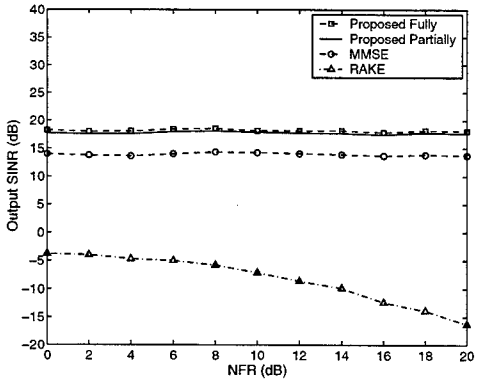


Figure 4: Output SINR versus NFR with $K = 10$ and SNR = 0 dB

ITERATIVE SPACE-TIME SOFT DETECTION IN TIME-VARYING MULTIACCESS WIRELESS CHANNELS

Joaquín Míguez, Luis Castedo

Departamento de Electrónica e Sistemas, Universidade da Coruña
Facultade de Informática, Campus de Elviña s/n, 15071 A Coruña (SPAIN).
Tel: +34 981167000 *Fax:* +34 981167160, *e-mail:* jmiguez@udc.es

ABSTRACT

This paper introduces an Iterative Space-time Soft Estimator (ISSE) that jointly performs linear channel estimation and soft data detection in time-varying Multiple-Input Multiple-Output (MIMO) channels, both according to the Minimum Mean Squared Error (MMSE) criterion. We introduce a scalar approximation of the channel autocovariance function that relies on the statistical homogeneity of the multipath scattering and allows to derive a globally-convergent burst-adaptive algorithm to estimate and track the channel statistics.

1. INTRODUCTION

It has been recently demostrated that deploying multiple transmitting and receiving antennae in a wireless link can lead to a significant capacity increase as long as multipath propagation is adequately exploited [1]. Further combination of vector coding techniques with signal processing methods at the receiver has led to the development of the so-called Space-Time Coding (STC) concept [2].

A relevant signal processing problem for the implementation of STC systems is the estimation of the Multiple Input Multiple Output (MIMO) channel. In a practical environment, the MIMO channel should be considered both as time-dispersive, leading to severe Inter-Symbol Interference (ISI), and time-varying. The conventional approach to address the channel variability is to use adaptive algorithms such as Least Mean Squares (LMS) and Recursive Least Squares (RLS) [3]. However, it has been shown [4, 5] that taking explicitly into account the time-varying nature of the channel leads to an improved performance. The most common approach is to use a block-adaptive procedure [4, 5] that consists of computing a set of snapshot estimates of the channel in different (e.g., equally spaced) observation windows using training data and, then, interpolating the channel coefficients between succesive snapshots.

In this paper, we propose a burst-iterative space-time scheme that alternates channel estimation and data detection. A linear channel estimator is derived, according to the Minimum Mean Squared Error (MMSE) criterion, that involves the second order statistics of the random channel process. Under a homogeneous scattering assumption, we show that the channel autocovariance can be characterized by a single scalar function, and we derive a globally-convergent burst-adaptive algorithm to estimate it. Although the latter result does not exactly hold for arbitrary scattering models, computer simulations show that there is only a slight performance loss, whereas a significant reduction in information requirements (i.e., knowledge of the full channel autocovariance matrices) is achieved. Since the linear channel estimator also depends on the transmitted symbols, we propose to iteratively alternate channel estimation and data detection until convergence and show that this approach achieves better performance than existing block-adaptive channel estimation methods.

In the next section, the system and signal model are introduced. The channel estimator is derived in section 3 and data detection is considered in section 4. Illustrative computer simulations are shown in section 5 and section 6 is devoted to the conclusions.

2. SYSTEM AND SIGNAL MODEL

Let us consider a wireless communication system with N antennae at the transmitter and L antennae at the receiver. The block diagram of such a system is depicted in figure 1. The information bits to be transmitted, $\{b(l)\}_{l=0,1,2...}$, are fed into a channel encoder and interleaver to yield a coded bit sequence. A Serial to Parallel (S/P) converter followed by a bank of N Waveform Encoders (WE) and transmitting antennae transforms this sequence into the identically-modulated information-bearing signals, $s_1(t), ..., s_N(t)$. Transmission is carried out in bursts of $NK\log_2(A)$ bits, i.e., K complex symbols per transmitting element, assuming that the WE's use a keying format with $\log_2(A)$ bits per symbol. Multipath propagation occurs betweeen each pair of transmitting and receiving antennae, resulting in a time-dispersive MIMO channel. At the receiver, a bank of L Matched Filters (MF) sampled at the symbol rate, $1/T$, are used to obtain $L \times 1$ vectors of observations, $\mathbf{x}(n) = [x_1(n), ..., x_L(n)]^T$, $n = 0, 1, ..., K-1$. These observations are sufficient statistics for an Iterative Space-time Soft Estimator (ISSE) device to obtain estimates, $\underline{y}(n) = [y_1(n), ..., y_N(n)]^T$, $n = 0, 1, ..., K-1$, of the complex transmitted symbols $\underline{s}(n) = [s_1(n), ..., s_N(n)]^T$, $n = 0, 1, ..., K-1$. A Parallel to Serial (P/S) converter produces the one-dimensional symbol sequence $\{y(k)\}_{k=0,1,2,...,NK-1}$ that is processed by a deinterleaver and channel decoder to obtain the final hard estimates of the information bits, $\{\hat{b}(l)\}_{l=0,1,...,NK\log_2(V)}$.

When a linear memoryless keying format for the WE's is employed, the following discrete-time signal model is obtained for

This work has been supported by FEDER funds (1FD97-0082) and Xunta de Galicia (PGIDT00PXI10504PR).

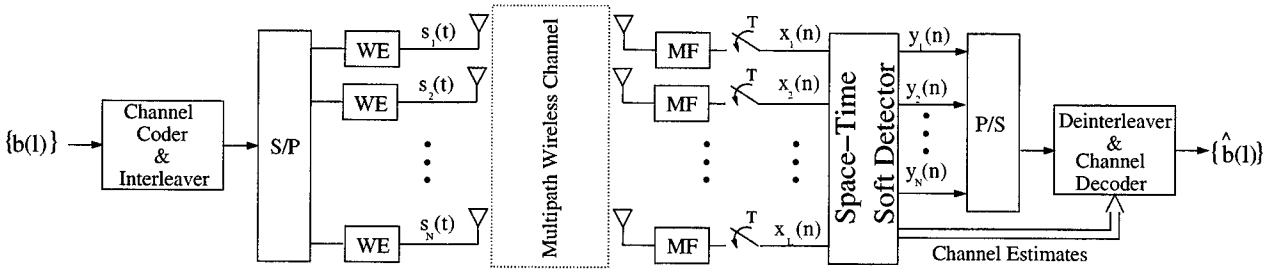


Fig. 1. Discrete-time model of a multiaccess wireless communication system.

the observations

$$\mathbf{x}(n) = \sum_{l=0}^{m-1} \underline{H}_l(n)\underline{s}(n-l) + \mathbf{g}(n) = \mathbf{H}(n)\mathbf{s}(n) + \mathbf{g}(n)$$

where $\underline{s}(n)$ is the n-th transmitted symbol vector, $\mathbf{s}(n) = [\underline{s}^T(n-m+1) \quad \cdots \quad \underline{s}^T(n)]^T$ is the n-th $Nm \times 1$ received symbol vector (note the existence of causal ISI), m is the maximum length of the time-varying discrete channel impulse response, $\mathbf{g}(n)$ is a $L \times 1$ vector of independent and identically distributed (i.i.d.) complex Gaussian components with zero-mean and variance σ_g^2, and

$$\mathbf{H}(n) = \left[\underline{H}_{m-1}(n) \quad \cdots \quad \underline{H}_0(n)\right]$$

is the n-th realization of the $L \times Nm$ random channel matrix. If we let vector

$$\mathbf{h}_{ij}(n) = \left[h_{ij}^0(n), h_{ij}^1(n), \ldots, h_{ij}^{m-1}(n)\right]^T$$

denote the (time-varying) channel impulse response between the j-th and the i-th transmitting and receiving elements, respectively, we can write down the $L \times N$ components of $\underline{H}_l(n)$, $l = 0, ..., m-1$ as

$$\underline{H}_l(n) = \begin{bmatrix} h_{11}^l(n) & h_{12}^l(n) & \ldots & h_{1N}^l(n) \\ h_{21}^l(n) & h_{22}^l(n) & \ldots & h_{2N}^l(n) \\ \vdots & \vdots & \ddots & \vdots \\ h_{L1}^l(n) & h_{L2}^l(n) & \ldots & h_{LN}^l(n) \end{bmatrix}.$$

According to the Gaussian Wide Sense Stationary Uncorrelated Scattering (GWSSUS) model, the channel coefficients in $\mathbf{H}(n)$ are assumed to be stationary independent Gaussian random processes with zero-mean, variance $\sigma_{h_{ij}^l}^2$ and autocovariance function $\phi_{ij}^l(k) = E\left[h_{ij}^l(n)h_{ij}^{l\,*}(n)\right]$, where superindex * denotes complex conjugation.

3. LINEAR MMSE CHANNEL ESTIMATION

The linear MMSE estimator for the channel coefficients at time n is obtained by processing a window of $M+1$ observation vectors (even M) as

$$\begin{aligned} \mathrm{MSE}(\mathbf{W}, i, n) &= E\left[\mathrm{Trace}\left[||\mathbf{X}(i)\mathbf{W} - \mathbf{H}(n)||^2\right]\right] \\ \hat{\mathbf{W}}(i,n) &= \arg \min_{\mathbf{W}} \{\mathrm{MSE}(\mathbf{W}, i, n)\} \quad (1) \\ \hat{\mathbf{H}}(i,n) &= \mathbf{X}(i)\hat{\mathbf{W}}(i,n), \quad (2) \end{aligned}$$

where $E[\cdot]$ denotes statistical expectation, $||\mathbf{M}||^2 = \mathbf{M}^H\mathbf{M}$ for an arbitrary matrix $\mathbf{M}$ and superdindex ^{H} meaning Hermitian transposition, $\mathbf{X}(i) = \left[\mathbf{x}(i-\frac{M}{2}) \quad \cdots \quad \mathbf{x}(i+\frac{M}{2})\right]$ is the $L \times (M+1)$ observation matrix used to estimate $\mathbf{H}(n)$, $\mathrm{MSE}(\cdot, i, n)$ is the associated mean squared-error cost funtion, $\hat{\mathbf{W}}(i,n)$ is the $(M+1) \times Nm$ Linear MMSE (LMMSE) filter and $\hat{\mathbf{H}}(i,n)$ is the MIMO channel estimate. Notice that there are several *optimum* filters for estimating $\mathbf{H}(n)$, depending on the observation window that is used (hence the two indices, i and n). Although for most practical channels, the best performance is achieved for $n = i$ we will consider the more general case of computing different channel estimates using the same window.

Problem (1) is purely quadratic and presents a closed-form solution

$$\begin{aligned} \hat{\mathbf{W}}(i,n) &= \left(E\left[\mathbf{X}^H(i)\mathbf{X}(i)\right]\right)^{-1} E\left[\mathbf{X}^H(i)\mathbf{H}(n)\right] \\ &= \left(\mathcal{R}(i) + L\sigma_g^2\mathbf{I}_{M+1}\right)^{-1}\mathcal{P}(i,n). \quad (3) \end{aligned}$$

In the above expression, $\mathcal{R}(i)$ is the $(M+1) \times (M+1)$ noiseless autocorrelation matrix, $\mathbf{I}_{M+1}$ is the $(M+1) \times (M+1)$ identity matrix and $\mathcal{P}(i,n)$ is the $(M+1) \times Nm$ cross covariance matrix. The element in the r-th row and c-th column of $\mathcal{R}(i)$ turns out to be

$$[\mathcal{R}(i)]_{r,c} = \mathbf{s}^H(i - \frac{M}{2} + r)\mathbf{R}_H(r-c)\mathbf{s}(i - \frac{M}{2} + c),$$

where $r, c = 0, ..., M$ and

$$\mathbf{R}_H(k) = E\left[\mathbf{H}^H(n)\mathbf{H}(n+k)\right]$$

is the MIMO channel autocovariance $Nm \times Nm$ matrix. Similarly, the r-th row vector of $\mathcal{P}(i,n)$ is

$$[\mathcal{P}(i,n)]_r = \mathbf{s}^H(i - \frac{M}{2} + r)\mathbf{R}(|n - i + \frac{M}{2} - r|), \quad r = 0, ..., M.$$

Since the channel coefficients are statistically independent, the autocovariance matrix can be further decomposed up to the diagonal form

$$\begin{aligned} \mathbf{R}_H(k) &= \sum_{i=1}^{L} \mathrm{diag}\left\{\phi_{i1}^{m-1}(k), ..., \phi_{iN}^{m-1}(k), ... \right. \\ &\quad \left. ..., \phi_{i1}^0(k), ..., \phi_{iN}^0(k)\right\}. \quad (4) \end{aligned}$$

Moreover, assuming a statistically homogeneous scattering we can state that

$$\rho(k) = \sum_{i=1}^{L} \phi_{ij}^l(k) \;\; \forall j, l, \quad (5)$$

and simplify (4) as

$$\mathbf{R}_H(k) = \rho(k)\mathbf{I}_{Nm}.$$

3.1. Block-Adaptive Estimation of the Channel Statistics

Under assumption (5), the time-varying channel estimation filter given by (3) is fully characterized by the single scalar function $\rho(k)$. We show, in this section, that $\rho(k)$ can be estimated by means of the globally-convergent burst-adaptive updating rule

$$\tilde{\rho}_i(k) = \frac{\sum_{n=0}^{K-k-1} \frac{\mathbf{x}^H(n)\mathbf{x}(n+k) - \delta(k)L\sigma_g^2}{\mathbf{s}^H(n)\mathbf{s}(n+k)} I(n)}{\sum_{n=0}^{K-1} I(n)} \quad (6)$$

$$\hat{\rho}_i(k) = (1-\mu)\hat{\rho}_{i-1}(k) + \mu\tilde{\rho}_i(k) \quad (7)$$

where $\mathbf{x}(0), ..., \mathbf{x}(K-1)$ is the i-th burst of observations, with associated data vectors $\mathbf{s}(0), ..., \mathbf{s}(K-1)$, $0 < \mu < 1$ is the step-size parameter, $\delta(\cdot)$ is Kronecker's delta function and $I(n) = 1 - \delta\left(\mathbf{s}^H(n)\mathbf{s}(n+k)\right)$ is an indicator function that avoids divisions by zero. The validity of algorithm (6)-(7) is granted by the statistical results stated below.

Hypotheses: Let $\mathbf{x}(n) = \mathbf{H}(n)\mathbf{s}(n) + \mathbf{g}(n)$ be a correlated stochastic process such that
(i) $\mathbf{H}(n)$ is a random $(L \times Nm)$-dimensional stationary process with autocovariance matrices

$$E\left[\mathbf{H}^H(n)\mathbf{H}(n+k)\right] = \rho(k)\mathbf{I}_{Nm}.$$

(ii) $\mathbf{s}(n)$ are vectors of deterministic symbols, that belong to a finite alphabet with constant-modulus and finite-valued elements,
(iii) For the desired delay k, $\mathbf{s}^H(n)\mathbf{s}(n+k) \neq 0, \forall n$.
(iv) $\mathbf{g}(n)$ is a temporally white Gaussian process with moments $E[\mathbf{g}(n)] = \mathbf{0}$ and $E\left[\mathbf{g}^H(n)\mathbf{g}(n+k)\right] = \delta(k)\sigma_g^2\mathbf{I}_L$.

Lemma 1 *Under assumptions (i)-(iv),*

$$\tilde{\rho}(k) = \frac{1}{K}\sum_{n=0}^{K-k-1} \frac{\mathbf{x}^H(n)\mathbf{x}(n+k) - \delta(k)L\sigma_g^2}{\mathbf{s}^H(n)\mathbf{s}(n+k)}$$

is an estimator of $\rho(k)$ with the following properties
(i) $\tilde{\rho}(k)$ is asymptotically unbiassed, i.e., $\lim_{K\to\infty} E[\tilde{\rho}(k)] = \rho(k)$ $\forall k$
(ii) $\tilde{\rho}(k \neq 0)$ is asymptotically consistent, i.e., $\lim_{K\to\infty} Var[\tilde{\rho}(k)] = 0$ $k \neq 0$, where $Var[\cdot]$ denotes variance, and
(iii) $\tilde{\rho}(0)$ is asymptotically consistent for high values of the SNR, i.e., $\lim_{K,\gamma\to\infty} Var[\tilde{\rho}(0)] = 0$ where γ is the signal-to-noise ratio in natural units.

Theorem 1 *Under assumptions (i)-(iv), the adaptive rule (7) yields an estimate, $\hat{\rho}_i(k)$, of the channel autocovariance function that verifies*
(i) $\lim_{i,K\to\infty} E[\hat{\rho}(k)] = \rho(k)$ $\forall k$,
(ii) $\lim_{i,K\to\infty} Var[\hat{\rho}(k)] = 0$, $k \neq 0$, and
(iii) $\lim_{i,K,\gamma\to\infty} Var[\hat{\rho}(0)] = 0$.

The proofs are necessarily skipped due to lack of space, but they can be checked in [6]. The above results state that the correlation features of the stationary channel process, which are required in order to build the matrix filter (3), can be adequately estimated from the available observations as long as the observation interval is large enough.

4. DATA DETECTION

The linear channel estimator (3) depends on the transmitted symbols and, therefore, data detection and channel estimation should be carried out jointly. In this section, we describe an ISSE scheme that alternates LMMSE channel estimation and Decision Feedback (DF) MMSE soft data detection. In order to describe the DFMMSE detector, it is convenient to define the following *stacked* model for the received signal,

$$\mathbf{x}_a(n) = \mathbf{H}_a(n)\mathbf{s}_a(n) + \mathbf{g}_a(n),$$

where a is a positive integer factor and

$$\mathbf{x}_a(n) = \left[\mathbf{x}^T(n) \quad \cdots \quad \mathbf{x}^T(n+a-1)\right]^T$$

$$\mathbf{H}_a = \begin{bmatrix} \underline{H}_{m-1}(n) & \mathbf{0} & \cdots & \mathbf{0} \\ \underline{H}_{m-2}(n) & \underline{H}_{m-1}(n) & \cdots & \mathbf{0} \\ \vdots & \underline{H}_{m-2}(n) & \ddots & \vdots \\ \underline{H}_0(n) & \vdots & \vdots & \underline{H}_{m-1}(n) \\ \mathbf{0} & \underline{H}_0(n) & \ddots & \underline{H}_{m-2}(n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \underline{H}_0(n) \end{bmatrix}^T$$

$$\mathbf{s}_a(n) = \left[\underline{s}^T(n-m+1) \quad \cdots \quad s^T(n+a-1)\right]^T$$

$$\mathbf{g}_a(n) = \left[\mathbf{g}^T(n) \quad \cdots \quad \mathbf{g}^T(n+a-1)\right]^T_{La\times 1}$$

are the $La \times 1$ stacked observation vector, $La \times N(m+a-1)$ channel matrix, $N(m+a-1) \times 1$ received symbol vector and $La \times 1$ AWGN vector, respectively.

Symbol estimates are computed as

$$\underline{y}(i,n) = \mathbf{F}^H(i,n)\mathbf{x}_a(i) + \mathbf{B}^H(i,n)\hat{\mathbf{s}}_B(i,n) \quad (8)$$

where $\mathbf{F}(i,n)$ and $\mathbf{B}(i,n)$ are *forward* and *backward* matrix filters with dimensions $La \times N$ and $(n+m-i-1) \times La$, respectively, and $\hat{\mathbf{s}}_B(i,n) = \left[\underline{\hat{s}}^T(i-m+1) \quad \cdots \quad \underline{\hat{s}}^T(n-1)\right]^T$ is a $N \times (n+m-i-1)$ vector containing past hard symbol estimates provided by a simple scalar threshold detector. Note that, in order to estimate $\underline{s}(n)$, index i must be in the range $n-a < i < n+m$. The forward and backward linear filters, $\mathbf{F}(i,n)$ and $\mathbf{B}(i,n)$, respectively, are selected according to the MMSE criterion as

$$\mathbf{F}(i,n), \mathbf{B}(i,n) = \arg \min_{\mathbf{F},\mathbf{B}} \left\{ E\left[||\underline{y}(i,n) - \underline{s}(n)||^2 \right] \right\}.$$

Assuming that the transmitted symbols are temporally uncorrelated and statistically independent of the AWGN, it is straightforward to derive closed-form solutions for $\mathbf{F}(i,n)$ and $\mathbf{B}(i,n)$ using the available channel estimates, $\hat{\mathbf{H}}(n)$ [6].

The proposed ISSE performs joint channel estimation and data detection by iterating equations (3), (2) and (8) until convergence for each data burst. The availability of training data to obtain initial channel estimates is assumed. As data bursts are processed, algorithm (6)-(7) is used to estimate and track the channel process statistics.

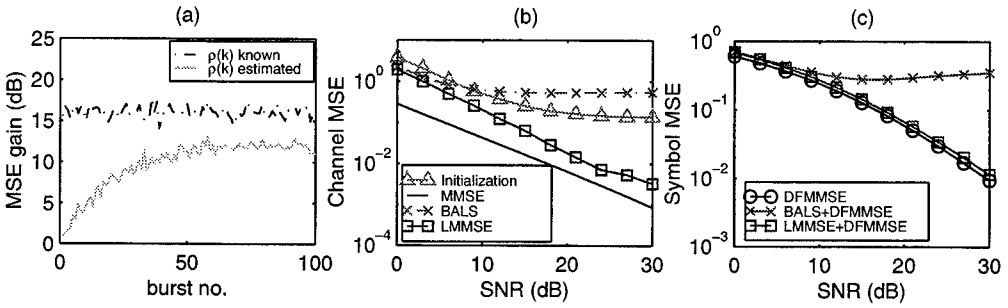


Fig. 2. **(a)** MSE improvement of LMMSE channel estimation with respect to BALS channel estimation as $\rho(k)$ is estimated. SNR=21 dB. **(b)** Channel estimation MSE for several values of the SNR, **(c)** Symbol estimation MSE for several values of the SNR. The forward and backward filters are $\mathbf{F}(n-2,n)$ and $\mathbf{B}(n-2,n)$, respectively, with $a=6$.

5. COMPUTER SIMULATIONS

Let us consider a system with $N = L = 3$ transmitting and receiving antennae, BPSK modulation and rectangular pulses. The burst length, per antenna, is $K = 800$ and the symbol period is $T = 4\ \mu s$. 16% equally-spaced pilot symbols are transmitted from every antenna. We assume a land mobile communication environment with the classical Rayleigh model of the power Doppler spectrum, which yields the autocovariance function, $\phi_{ij}^{l}(k) = \sigma_{h_{ij}^{l}}^{2}\, \mathcal{J}_0\left(2\pi f_D k T\right)$. Here, $\mathcal{J}_0(\cdot)$ is the zero-order Bessel function of the first kind, $f_D = \frac{v_m}{v_l} f_C$ is the maximum Doppler spread, $v_m = 120$ Km/h is the motion speed of the transmitter, v_l is the speed of light and $f_C = 2$ GHz is the carrier frequency. Using this model, we have simulated the type-B multipath channel defined by IMT-2000 for the vehicular environment [7] with 4 μs root mean-squared delay and a decreasing exponential delay power profile that yields a 16 μs maximum delay spread and significant ISI ($m = 4$). The homogeneous scattering assumption (5) does not hold in the described environment, but the simulation results show that the scalar autocovariance approximation, $\mathbf{R}_H(k) \approx \rho(k)\mathbf{I}_{Nm}$, still provides an adequate performance.

We compare the performance of the proposed ISSE with a combination of Block-Adaptive Least Squares (BALS) channel estimation [4, 5] and DFMMSE data detection. The performance limits, i.e., the MMSE for channel and symbol estimation, are also plotted as a reference. The $K = 800$ symbol frame is divided into 4 windows that only overlap in one symbol (hence, $M = 200$). Both LMMSE and BALS channel estimators are applied on these windows, so they have the same computational complexity. Indeed, with this setup, using BALS is the same as applying LMMSE with constant $\rho(k)$.

Figure 2(a) shows the Mean Squared Error (MSE) improvement of the proposed LMMSE channel estimator over the BALS estimator as $\rho(k)$ is adaptively estimated, starting with constant $\rho(k) = 1\ \forall k$. The MSE gain, in dB, of a sequence of channel estimates $\left\{\hat{\mathbf{H}}_0(n)\right\}_{n=0,\ldots,K-1}$ over another sequence $\left\{\hat{\mathbf{H}}_1(n)\right\}_{n=0,\ldots,K-1}$ is defined as $\varepsilon_{0,1} = \log_{10}\left(\frac{\varepsilon_1}{\varepsilon_0}\right)$ where $\varepsilon_i = \frac{1}{K}\sum_{n=0}^{K-1} \text{Trace}\left[||\hat{\mathbf{H}}_i - \mathbf{H}||^2\right]$ is the average channel MSE. It can be seen that the LMMSE estimates may attain up to a 15 dB improvement over the BALS estimates, when the Signal to Noise Ratio (SNR) value is 21 dB.

This improvement can also be observed in figure 2(b), which shows the channel MSE (after estimation of $\rho(k)$) for the iterative LMMSE and BALS methods, together with the MSE of the initial estimates (obtained via non iterative BALS using training data) and the correponding MMSE. The performance of the iterative BALS estimator degrades for higher SNR values due to severe error propagation.

Finally, figure 2(c) shows the symbol MSE attained by the ISSEs for several values of the SNR. It is observed that the proposed iterative method that combines LMMSE channel estimation and DFMMSE symbol detection practically attains the MMSE.

6. CONCLUSIONS

We have introduced an ISSE structure that performs joint LMMSE channel estimation and soft DFMMSE data detection in time-varying MIMO channels with ISI. A scalar approximation of the channel autocovariance function that relies on the statistical homogeneity of the multipath scattering is proposed, together with a globally-convergent burst-adaptive algorithm to estimate it.

7. REFERENCES

[1] G. J. Foschini, "Layered space-time arquitechture for wireless communications in a fading environment when using multi-element anntenas," *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 41–59, Autumn 1996.

[2] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communications: Performance analysis and code construction," *IEEE Trans. Information Theory*, vol. 44, pp. 744–765, March 1998.

[3] S. Haykin, *Adaptive Filter Theory, 3rd Edition*, Prentice Hall, Information and System Sciences Series, 1996.

[4] N. W. K. Lo, D. D. Falconer, and A. U. H. Sheikh, "Adaptive equalization and diversity combining for mobile radio using interpolated channel estimates," *IEEE Trans. Vehicular Technology*, vol. 40, no. 3, pp. 636–645, August 1991.

[5] H-N. Lee and G. J. Pottie, "Fast adaptive equalization/diversity combining for time-varying dispersive channels," *IEEE Trans. Commnunications*, vol. 46, no. 9, pp. 1146–1162, September 1998.

[6] J. Míguez and L. Castedo, "Space-time channel estimation and soft detection in time-varying multiaccess channels," *submitted to Signal Processing*, April 2001.

[7] V. K. Garg and J. E. Wilkes, *Principles & Applications of GSM*, Prentice Hall, Upper Saddle River, NJ 07458, 1999.

BLIND EQUALIZATION USING CUMULANT BASED MIMO INVERSE FILTER CRITERIA FOR MULTIUSER DS/CDMA SYSTEMS IN MULTIPATH

Chong-Yung Chi and Chii-Horng Chen

Department of Electrical Engineering &
Institute of Communications Engineering
National Tsing Hua University, Hsinchu, Taiwan, R.O.C.
Tel: 886-3-5731156, Fax: 886-3-5751787, E-mail: cychi@ee.nthu.edu.tw

ABSTRACT

Chi and Chen recently reported a blind equalization algorithm using cumulant based multi-input multi-output inverse filter criteria (MIMO-IFC) for mulituser DS/CDMA systems in multipath. Assuming that the user of interest is the weak user with signal power $\mathcal{E}_1$ and signal powers of all the interferers are identical, denoted $\mathcal{E}$, the performance of Chi and Chen's algorithm is superior to that of Tsatsanis and Xu's blind minimum variance (MV) equalizer for low near-far ratio (NFR) ($= \mathcal{E}/\mathcal{E}_1 \geq 1$). In this paper, two blind equalization algorithms, called Algorithms 2 and 3, also using cumulant based MIMO-IFC are proposed. The former (Algorithm 2) can improve the performance of the MV equalizer. The latter (Algorithm 3) based on the former performs as well as Chi and Chen's algorithm for low NFR and outperforms Chi and Chen's algorithm and the MV equalizer for high NFR. Some simulation results are presented to support the efficacy of the proposed algorithms.

1. INTRODUCTION

Blind equalization of a multi-input multi-output (MIMO) linear time-invariant (LTI) system, denoted $\mathbf{H}[n]$ ($P \times K$ matrix), is a problem of estimating the vector input $\mathbf{u}[n] = (u_1[n], u_2[n], ..., u_K[n])^T$ (K inputs) with only a set of non-Gaussian vector output measurements $\mathbf{x}[n] = (x_1[n], x_2[n], ..., x_P[n])^T$ (P outputs) as follows [1–4]

$$\mathbf{x}[n] = \sum_{k=-\infty}^{\infty} \mathbf{H}[k]\mathbf{u}[n-k] + \mathbf{w}[n] = \sum_{k=1}^{K} \mathbf{y}_k[n] + \mathbf{w}[n] \quad (1)$$

where $\mathbf{w}[n] = (w_1[n], w_2[n], ..., w_P[n])^T$ ($P \times 1$ vector) is additive noise and

$$\mathbf{y}_k[n] = \mathbf{h}_k[n] * u_k[n] = \sum_{l=-\infty}^{\infty} \mathbf{h}_k[n-l] \cdot u_k[l] \quad (2)$$

This work was supported by the National Science Council under Grant NSC 89-2219-E007-018.

(the contribution in $\mathbf{x}[n]$ from the input $u_k[n]$) in which $\mathbf{h}_k[n]$ is the kth column of $\mathbf{H}[n]$. Blind equalization of MIMO systems in multiuser detection of wireless communications includes suppression of multiple access interference (MAI) and removal of multipath effects that are crucial to the receiver design of multiuser communications systems.

Let $\mathbf{v}[n] = (v_1[n], v_2[n], ..., v_P[n])^T$ denote a linear FIR equalizer of length $L = L_2 - L_1 + 1$ for which $\mathbf{v}[n] \neq \mathbf{0}$ for $n = L_1, L_1+1, ..., L_2$. Then the output $e[n]$ of the FIR equalizer (inverse filter) $\mathbf{v}[n]$ can be expressed as

$$e[n] = \sum_{k=L_1}^{L_2} \mathbf{v}^T[k] \cdot \mathbf{x}[n-k] = \sum_{j=1}^{P} \boldsymbol{v}_j^T \boldsymbol{x}_j[n] = \boldsymbol{\nu}^T \tilde{\boldsymbol{x}}[n] \quad (3)$$

where $\boldsymbol{v}_j = (v_j[L_1], v_j[L_1+1], ..., v_j[L_2])^T$, $\boldsymbol{x}_j[n] = (x_j[n-L_1], x_j[n-L_1-1], ..., x_j[n-L_2])^T$, $\tilde{\boldsymbol{x}}[n] = (\boldsymbol{x}_1^T[n], \boldsymbol{x}_2^T[n], ..., \boldsymbol{x}_P^T[n])^T$ and

$$\boldsymbol{\nu} = (\boldsymbol{v}_1^T, \boldsymbol{v}_2^T, ..., \boldsymbol{v}_P^T)^T.$$

The design of the equalizer $\mathbf{v}[n]$ (or $\boldsymbol{\nu}$) such that $e[n] \rightarrow \alpha u_{j_0}[n-\tau]$ where $\alpha \neq 0$, τ is an unknown integer and $u_{j_0}[n]$ is the signal of interest (SOI), is a widely known signal processing problem in wireless communications.

2. REVIEW OF MIMO INVERSE FILTER CRITERIA (MIMO-IFC)

For ease of later use, let $\text{cum}\{y_1, y_2, ..., y_p\}$ denote the pth-order cumulant [5] of random variables y_1, y_2, ..., y_p,

$$\begin{aligned} \text{cum}\{y : p, ...\} &= \text{cum}\{y_1 = y, y_2 = y, ..., y_p = y, ...\} \\ C_{p,q}\{y\} &= \text{cum}\{y : p,\ y^* : q\} \end{aligned}$$

where y^* is complex conjugate of y. Assume that we are given a set of measurements $\mathbf{x}[n]$, $n = 0, 1, ..., N-1$, modeled by (1) with the following assumptions:

($\mathcal{A}1$) $u_i[n]$ is zero-mean, independent identically distributed (i.i.d.), non-Gaussian and statistically independent of $u_k[n]$ for all $k \neq i$, and $C_{p,q}\{u_i[n]\} \neq 0$, $i = 1, 2, ..., K$

for a chosen (p,q), where p and q are nonnegative integers and $p+q \geq 3$.

($\mathcal{A}2$) The MIMO system $\mathbf{H}[n]$ is exponentially stable.

($\mathcal{A}3$) The noise $\mathbf{w}[n]$ is zero-mean Gaussian (which can be spatially correlated and temporally colored) and statistically independent of $\mathbf{u}[n]$.

MIMO-IFC

Chi and Chen [1] proposed MIMO-IFC for the design of the equalizer $\mathbf{v}[n]$ by maximizing

$$J_{p,q}(\boldsymbol{\nu}) = \frac{|C_{p,q}\{e[n]\}|}{|C_{1,1}\{e[n]\}|^{(p+q)/2}} \tag{4}$$

where p and q are nonnegative integers and $p+q \geq 3$. The obtained optimum $e[n]$ turns out to be an estimate of $u_j[n]$, $j \in \{1,2,...,K\}$ except for an unknown scale factor and an unknown time delay. Chi and Chen's MIMO-IFC include Tugnait's MIMO-IFC [3] for $(p,q)=(2,1)$ and $(p,q)=(2,2)$ as special cases.

Efficient Algorithm for MIMO-IFC

Recently, Chi and Chen [2] proposed a fast gradient type iterative MIMO-IFC algorithm with convergence speed, computational load, and amount of multichannel intersymbol interference (MISI) similar to those of MIMO super exponential algorithm (MIMO-SEA) [6] as follows:

Algorithm 1. Given $\boldsymbol{\nu}^{(i-1)}$ and $e^{(i-1)}[n]$ obtained at the $(i-1)$th iteration, $\boldsymbol{\nu}^{(i)}$ at the ith iteration is obtained through the following two steps.

(S1) Obtain $\boldsymbol{\nu}^{(i)}$ by

$$\boldsymbol{\nu}^{(i)} = \frac{\widetilde{\boldsymbol{R}}^{-1} \cdot \widetilde{\boldsymbol{d}}^{(i-1)}}{\left\| \widetilde{\boldsymbol{R}}^{-1} \cdot \widetilde{\boldsymbol{d}}^{(i-1)} \right\|} \tag{5}$$

where $\widetilde{\boldsymbol{R}} = E[\widetilde{\boldsymbol{x}}^*[n]\widetilde{\boldsymbol{x}}^T[n]]$, $\|\mathbf{a}\|$ denotes the Euclidean norm of vector $\mathbf{a}$ and

$$\widetilde{\boldsymbol{d}}^{(i-1)} = \text{cum}\{e^{(i-1)}[n] : r, (e^{(i-1)}[n])^* : s-1, \widetilde{\boldsymbol{x}}^*[n]\} \tag{6}$$

where r and $s-1$ are nonnegative integers, $r+s=p+q$ as $\mathbf{x}[n]$ is real and $r=s=p=q$ as $\mathbf{x}[n]$ is complex.

(S2) If $J_{p,q}(\boldsymbol{\nu}^{(i)}) > J_{p,q}(\boldsymbol{\nu}^{(i-1)})$, go to the next iteration, otherwise update $\boldsymbol{\nu}^{(i)}$ through a gradient type optimization algorithm such that $J_{p,q}(\boldsymbol{\nu}^{(i)}) > J_{p,q}(\boldsymbol{\nu}^{(i-1)})$ and obtain the associated $e^{(i)}[n]$.

Channel Estimation and Signal Cancellation

With the obtained $e[n]$ (estimate of $u_j[n]$ up to a scale factor and a time delay where j is *unknown*) using Algorithm 1, $\mathbf{h}_j[k]$ can be estimated as [3]

$$\widehat{\mathbf{h}}_j[k] = \frac{E[\mathbf{x}[n]e^*[n-k]]}{E[|e[n]|^2]}. \tag{7}$$

Therefore, the contribution in $\mathbf{x}[n]$ due to $u_j[n]$ can be estimated as (see (2))

$$\widehat{\mathbf{y}}_j[n] = \widehat{\mathbf{h}}_j[n] * e[n]. \tag{8}$$

Removing $\widehat{\mathbf{y}}_j[n]$ from the data $\mathbf{x}[n]$ yields

$$\widetilde{\mathbf{x}}[n] = \mathbf{x}[n] - \widehat{\mathbf{y}}_j[n] = \mathbf{x}[n] - \widehat{\mathbf{h}}_j[n] * e[n] \tag{9}$$

that corresponds to the outputs of a $P \times (K-1)$ system driven by $(K-1)$ inputs $u_i[n]$, $i=1,...,j-1,j+1,...,K$.

3. MIMO CHANNEL MODELS FOR MULTIUSER DS/CDMA SYSTEMS IN MULTIPATH

Consider a K-user asynchronous DS/CDMA system. Assume that

$$\mathcal{R} = \{c_k[n],\ k=1,2,...,K,\ n=0,1,...,P-1\} \tag{10}$$

is the set of the K active users' signature sequences (binary sequences of $\{+1,-1\}$) with spreading factor equal to P ($\geq K$). Let $x[n]$ and $w[n]$ be discrete-time signals by sampling the received continuous time signal $x(t)$ and Gaussian noise $w(t)$ with sampling interval T_c (chip period), respectively. Two MIMO models are considered as follows.

MIMO Model I: Polyphase decomposition [1,4]

$$\mathbf{x}^{(1)}[n] = \sum_{k=-\infty}^{\infty} \mathbf{H}^{(1)}[k]\mathbf{u}[n-k] + \mathbf{w}^{(1)}[n] \tag{11}$$

where $\mathbf{x}^{(1)}[n] = (x[nP], x[nP+1], ..., x[nP+P-1])^T$, $\mathbf{w}^{(1)}[n] = (w[nP], w[nP+1], ..., w[nP+P-1])^T$ is a white Gaussian vector random process, $\mathbf{u}[n] = (u_1[n], u_2[n], ..., u_K[n])^T$ where $u_i[n]$ is the symbol sequence of user i, and $\mathbf{H}^{(1)}[n]$ is a $P \times K$ impulse response matrix with the ith column $\mathbf{h}_i^{(1)}[n]$ and the (i,k)th entry equal to

$$h_{i,k}^{(1)}[n] = h_k[nP+i-1] \tag{12}$$

in which $h_k[n]$ is the signature waveform of user k given by

$$h_k[n] = c_k[n] * g_k[n] = \sum_{l=0}^{P-1} c_k[l]g_k[n-l] \tag{13}$$

where $g_k[n]$ is an FIR multipath channel of order q_g for user k.

MIMO Model II:

Tsatsanis and Xu's minimum variance (MV) equalizer [4] estimates $u_i[n]$ by

$$\widehat{u}_{\text{MV},i}[n] = \mathbf{v}_{\text{MV},i}^H \overline{\boldsymbol{x}}[n] \tag{14}$$

where the superscript 'H' denotes complex conjugate transpose,

$$\overline{\boldsymbol{x}}[n] = (x[nP], x[nP+1], ..., x[nP+P+q_g-1])^T \tag{15}$$

$$\mathbf{v}_{\mathrm{MV},i} = \bar{\mathbf{R}}^{-1}\bar{\mathbf{C}}_i(\bar{\mathbf{C}}_i^H\bar{\mathbf{R}}^{-1}\bar{\mathbf{C}}_i)^{-1}\hat{\mathbf{g}}_i \quad (16)$$

in which $\bar{\mathbf{R}} = E[\overline{\mathbf{x}}[n]\overline{\mathbf{x}}^H[n]]$, $\bar{\mathbf{C}}_i$ is a $(P+q_g)\times(q_g+1)$ matrix constituted by $c_i[n]$ and $\hat{\mathbf{g}}_i$ is an estimate of $\mathbf{g}_i = (g_i[0], g_i[1], ..., g_i[q_g])^T$ obtained as the eigenvector of $\bar{\mathbf{C}}_i^H\bar{\mathbf{R}}^{-1}\bar{\mathbf{C}}_i$ associated with the smallest eigenvalue. Concatenating $\widehat{u}_{\mathrm{MV},i}[n]$, $i = 1, 2, ..., K$ yields

$$\begin{aligned}\mathbf{x}^{(2)}[n] &= (\widehat{u}_{\mathrm{MV},1}[n], \widehat{u}_{\mathrm{MV},2}[n], ..., \widehat{u}_{\mathrm{MV},K}[n])^T \\ &= \mathbf{H}^{(2)}[n] * \mathbf{u}[n] + \mathbf{w}^{(2)}[n] \quad (17)\end{aligned}$$

where $\mathbf{H}^{(2)}[n]$ and $\mathbf{w}^{(2)}[n]$ are $K\times K$ system and $K\times 1$ spatially correlated and temporally colored Gaussian noise, respectively. It can be shown that

$$\mathcal{E}_{j,j} = \sum_n |h_{j,j}^{(2)}[n]|^2 \gg \mathcal{E}_{i,j} = \sum_n |h_{i,j}^{(2)}[n]|^2, \ \forall\, i \neq j \quad (18)$$

where $h_{i,j}^{(2)}[n]$ is the (i,j)th entry of $\mathbf{H}^{(2)}[n]$. A worthy remark about the above two MIMO models is as follows:

(R1) Algorithm 1 can be employed to process either of $\mathbf{x}^{(1)}[n]$ and $\mathbf{x}^{(2)}[n]$ for obtaining one input estimate $\widehat{u}_j[n]$ where j is unknown. The identification of user number j associated with $\mathbf{x}^{(1)}[n]$ has been reported in [1], while that associated with $\mathbf{x}^{(2)}[n]$ is based on (18) as follows:

$$\widehat{j} = \arg\max_{1\le i\le K}\left\{\widehat{\mathcal{E}}_{i,j}\right\} \quad (19)$$

where $\widehat{\mathcal{E}}_{i,j}$ is obtained by (18) with $h_{i,j}^{(2)}[n]$ replaced by the ith entry $\widehat{h}_{i,j}^{(2)}[n]$ of the channel estimate $\widehat{\mathbf{h}}_j^{(2)}[n]$ (see (7)).

4. NEW ALGORITHMS FOR BLIND EQUALIZATION OF DS/CDMA SYSTEMS

Assuming that the SOI is $u_1[n]$, Chi and Chen's algorithm [1], a multistage successive cancellation algorithm, obtains $\widehat{u}_1[n]$ by processing $\mathbf{x}^{(1)}[n]$. This algorithm can be extended by processing either of $\mathbf{x}^{(1)}[n]$ given by (11) and $\mathbf{x}^{(2)}[n]$ given by (17) through the following three signal processing steps at each stage:

Algorithm 2. (with $l = 1$ or $l = 2$)

(V1) Process $\mathbf{x}^{(l)}[n]$ to obtain a local optimum $\boldsymbol{\nu}$ (and $\widehat{\mathbf{v}}[n]$) of $J_{p,q}(\boldsymbol{\nu})$ using Algorithm 1 and the associated $e[n]$ and $\widehat{\mathbf{h}}_j^{(l)}[n]$ (see (7)).

(V2) Update $\mathbf{x}^{(l)}[n]$ by $\mathbf{x}^{(l)}[n] - \widehat{\mathbf{h}}_j^{(l)}[n] * e[n]$ (see (9)) (i.e., signal cancellation).

(V3) Identify the user number j as presented in (R1). If $\widehat{j} = 1$, $\widehat{u}_1[n] = e[n]$ (i.e., $\widehat{u}_1[n]$ has been obtained).

Three worthy remarks with regard to Algorithm 2 are as follows.

(R2) The smaller the stage number $\Bbbk$ at which $\widehat{u}_1[n]$ is obtained, the better the performance of Algorithm 2 due to error propagation effects resulting from imperfect cancellation in (V2). However, the stage number $\Bbbk$ at which $\widehat{u}_1[n]$ is obtained is dependent upon the initial condition $\boldsymbol{\nu}^{(0)}$, which can be chosen as the least square solution of the decorrelating constraint as reported in [1] for $l = 1$ and as the one associated with

$$\mathbf{v}^{(0)}[n] = \mathbf{1}_{K,1}\cdot\delta[n-n_0], \ L_1 \le n_0 \le L_2 \quad (20)$$

for $l = 2$ where $\mathbf{1}_{K,1}$ is a $K\times 1$ unit column vector with the first entry equal to unity.

(R3) Algorithm 2 for $l = 1$ is exactly the same as Chi and Chen's algorithm. Algorithm 2 for $l = 2$ further processes the MV estimate $\widehat{u}_{\mathrm{MV},i}[n]$ (see (14) and (17)), and therefore, its performance is superior to that of the MV equalizer.

(R4) By our experience, the performance of Algorithm 2 is better for $l = 1$ than for $l = 2$ for low near-far ratio (NFR), whereas it is better for $l = 2$ than for $l = 1$ for high NFR.

Next, let us present a hybrid algorithm using Algorithm 2 with $l = 1$ and $l = 2$ based on (R4). The proposed algorithm obtains the SOI estimate $\widehat{u}_1[n]$ through the following two steps.

Algorithm 3.

(T1) Set $\Bbbk = 1$, perform all the steps of Algorithm 2 for $l = 1$ (identical with Chi and Chen's algorithm).

(T2) If $\widehat{j} \neq 1$ (i.e., $\widehat{u}_1[n]$ has not been obtained in (T1)), then perform all the steps of Algorithm 2 for $l = 2$ for all the ensuing stages $\Bbbk \ge 2$ until $\widehat{u}_1[n]$ is obtained.

A worthy remark regarding Algorithm 3 is as follows.

(R5) Assume that Algorithm 3 obtains $\widehat{u}_1[n]$ at the $\Bbbk_0$th stage. The obtained $\widehat{u}_1[n]$ for $\Bbbk_0 = 1$ (i.e., $\widehat{j} = 1$ in (T1)) (that usually happens for low NFR) is identical to that obtained by Chi and Chen's algorithm for $\Bbbk = 1$, while that for $\Bbbk_0 \ge 2$ (i.e., $\widehat{j} \neq 1$ in (T1)) (that usually happens for high NFR) is identical to the one obtained at the $(\Bbbk_0 - 1)$th stage of Algorithm 2 for $l = 2$. Therefore, Algorithm 3 performs better than Algorithm 2 regardless of l by (R4).

5. SIMULATION RESULTS

An asynchronous DS/CDMA channel for six users ($K = 6$) taken from [1] was considered. The users' spreading codes $c_j[n]$ were Gold codes of length $P = 31$. Input signals $u_i[n]$, $i = 1, 2, ..., K$ were assumed to be equally probable binary random sequences of $\{+1, -1\}$ whose amplitudes were adjusted such that $\mathcal{E}_i = E[\|\mathbf{h}_i^{(1)}[n] * u_i[n]\|^2] = \mathcal{E}$, $i = 2, 3, ..., 6$. The synthetic data $\mathbf{x}^{(1)}[n]$ for $N = 2500$, different

values of NFR ($= \mathcal{E}/\mathcal{E}_1 = 0, 10$ dB) and different values of SNR ($= \mathcal{E}_1/E[\|\mathbf{w}^{(1)}[n]\|^2] = 3, 5, 7, 9, 11, 13$ dB) were processed by both Algorithms 2 and 3 with $p = q = r = s = 2$, respectively, and with the length of the causal FIR inverse filter $\mathbf{v}[n]$ equal to three.

Figures 1(a) and 1(b) show the output signal-to-interference-plus-noise ratio (SINR) of user 1 (the weak user) for NFR = 0 dB and NFR = 10 dB, respectively, associated with the nonblind linear minimum mean square error (LMMSE) equalizer (which has maximum output SINR) (solid lines), Algorithm 3 ('$\triangle$'), Algorithm 2 for $l = 1$ ('$\bigcirc$') (identical with Chi and Chen's algorithm), Algorithm 2 for $l = 2$ ('X') and the MV equalizer ('$\square$'). All the results of Algorithm 3 were obtained at the stage $\mathbb{k}_0 = 1$ for NFR = 0 dB and $\mathbb{k}_0 = 2$ for NFR = 10 dB. One can observe, from these figures, that Algorithm 2 for $l = 2$ performs better than the MV equalizer (see (R3)), and that Algorithm 3 performs as well as Algorithm 2 with $l = 1$ (identical with Chi and Chen's algorithm) for NFR = 0 dB (low NFR) and Algorithm 2 with $l = 2$ for NFR = 10 dB (high NFR), respectively. These results are consistent with (R4) and (R5) and support that Algorithm 3 outperforms Chi and Chen's algorithm and the MV equalizer.

6. CONCLUSIONS

We have proposed two blind equalization algorithms, Algorithms 2 and 3, for multiuser asynchronous DS/CDMA systems in multipath. Algorithm 2, an extension of Chi and Chen's algorithm, can improve the performance of Tsatsanis and Xu's blind MV equalizer. Algorithm 3 based on Algorithm 2 outperforms Chi and Chen's algorithm and Tsatsanis and Xu's blind MV equalizer. Some simulation results were presented to support the efficacy of the proposed two algorithms.

7. REFERENCES

[1] C.-Y. Chi and C.-H. Chen, "Blind MAI and ISI suppression for DS/CDMA systems using HOS based inverse filter criteria," *IEEE Trans. Signal Processing* (in revision).

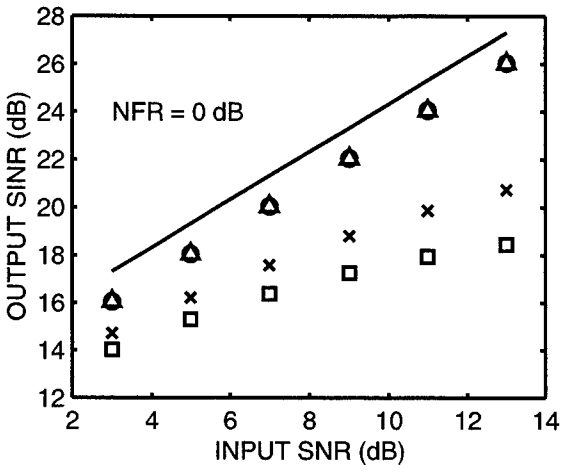
[2] C.-Y. Chi and C.-H. Chen, "Cumulant based inverse filter criteria for MIMO blind deconvolution: properties, algorithms, and application to DS/CDMA systems in multipath," to appear in *IEEE Trans. Signal Processing*, July 2001.

[3] J. K. Tugnait, "Identification and deconvolution of multichannel linear nonGaussian processes using higher-order statistics and inverse filter criteria," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 658-672, March 1997.

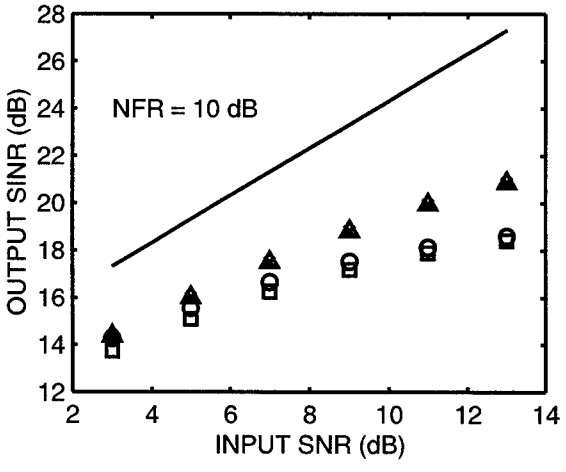
[4] M. K. Tsatsanis and Z. Xu. "Performance analysis of minimum variance CDMA receivers," *IEEE Trans. Signal Processing*, vol. 46, no. 11, pp. 3014-3022, Nov. 1998.

[5] C. L. Nikias and A. P. Petropulu, *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[6] K. L. Yeung and S. F. Yau, "A cumulant-based super-exponential algorithm for blind deconvolution of multi-input multi-output systems," *Signal Processing*, vol. 67, no. 2, pp. 141-162, 1998.



(a)



(b)

Fig. 1. Simulation results for a 6-user case with respective powers $\mathcal{E}_1$ and $\mathcal{E}_2 = \mathcal{E}_3 = \mathcal{E}_4 = \mathcal{E}_5 = \mathcal{E}_6 = \mathcal{E}$, including output SINR of user 1 (the weak user) associated with the LMMSE equalizer (solid line), Algorithm 3 ('$\triangle$'), Algorithm 2 for $l = 1$ ('$\bigcirc$') (i.e., Chi and Chen's algorithm), Algorithm 2 for $l = 2$ ('X') and the MV equalizer ('$\square$') for (a) NFR = 0 dB and (b) NFR = 10 dB, respectively.

COMBINED DOWNLINK BEAMFORMING AND CHANNEL ESTIMATION FOR HIGH DATA RATES CDMA SYSTEMS

Sylvie Perreau

Institute for Telecommunication Research
University of South Australia, MAwson Lakes SA 5095
email: Sylvie.Perreau@unisa.edu.au

1. INTRODUCTION

To obtain higher capacity, smart antenna techniques have been investigated for wireless communications . Most smart antenna techniques deal with signals from receive antenna arrays, but similar techniques can be used for transmitting signals using a transmit antenna array. In fact, the use of a transmit antenna array (TAA) at the base station for 3rd Generation (3G) CDMA systems has recently attracted a lot of attention [3], [6],[2]. Indeed, even if in theory the orthogonal property of Walsh codes allows mitigating multi user interference at the mobile station, the multipath propagation phenomenom introduces a non negligible interference level within the same cell, thus the advantage of downlink beamforming. In this paper, we focus our study on very high data rates 3G CDMA applications where spreading gains can be quite low (down to 4). In such a situation, the performance of the conventional Rake receiver is significantly degraded. This is because the Rake receiver is only effective in situations where multipath effects result in Inter Chip Interference (ICI) with small amount of Inter Symbol Interference(ISI). However, as the data rates increase (and the spreading gains decrease), the amount of ISI becomes less and less negligible which prevents the despreading operation part of the Rake receiver to be successful. Therefore, in such a situation, downlink beamforming can be very useful. Indeed, it can reduce the amount of multiuser interference at the receiver. In fact, the beamformer should transmit signals in the direction of multipaths less affected by multi user interference.

In order to construct downlink beampattern, one should utilize the knowledge of the downlink channel as well as the bearings of the downlink signal. However, in the frequency division duplexing (FDD) mode, since the downlink and uplink channels are different, the base station does not have access to the knowledge of the downlink channel. This requires downlink channel information to be fed back from the mobile station to the base station. Various issues associated with channel information feedback for downlink beamforming have been addressed in [6], where channel estimation was provided by the Rake receiver. For high data rates, we have seen previously that channel estimation cannot rely on the Rake receiver. Instead, it has to be provided by a scheme operating at the chip level. However, if the multiuser interference on the downlink is large, such a scheme may provide a poor channel estimate before the beamformer is correctly set up. In turn, one cannot expect the beamforming operation to be fully efficient if it relies on a poor channel estimate. This illustrates the fact that in this case, channel estimation and beamforming strongly depend on each other, mainly because the channel seen at the mobile station is in fact a combination of the propagation channel and the beamformer. Based on this simple observation, we propose in this paper a new method where channel estimation and beamforming operations are iterated several times until convergence to a fixed solution.

2. PROBLEM FORMULATION

In this paper, we consider the transmission of a CDMA signal through L transmit antennas. We consider here the baseband channel model (at the chip rate) where the channel memory M is greater than the processing gain (although the formulation still holds in the general case), as it is likely to occur for high data rate services. Let us define the notations used throughout this paper:
- $\mathbf{h}_q[l]$ is the $L \times 1$ channel vector at symbol time l for the qth path and is written as: $\mathbf{h}_q[l] = \beta_q[l] * \mathbf{a}(\theta_q[l])$ where $\mathbf{a}(\theta_q[l])$ is the $L \times 1$ known steering vector associated with the bearing $\theta_q[l]$ of the qth path and $\beta_q[l]$ is the attenuation factor for this path. Note that throughout this chapter, we will assume that the channel characteristics are time invariant. Therefore, the time index l will not be used hereafter. It is also important to note that with this model, $\beta_q[l]$ could very well be zero (no

0-7803-7011-2/01/$10.00 ©2001 IEEE

multipath component).
- $\mathbf{H}$ is the $L \times M$ channel matrix defined as:
$\mathbf{H} = [\mathbf{h}_1\ \mathbf{h}_2 \cdots \mathbf{h}_M]$
- $\mathbf{w}$ is the $L \times 1$ vector of coefficients of the beamformer.
-$x[m]$ is the sequence of chips after spreading.
- The general notation $\hat{v}^{(i)}$ represents the estimate of the quantity v after the ith iteration of the algorithm.
- Throughout this paper, the operator ^{H} denotes the Hermitian operator.

Using the above notations, the signal (at the chip rate) received at the desired mobile station is written as:

$$y[l] = \mathbf{w}^H \sum_{q=1}^{M} \mathbf{h}_q x[l-q] + n[l] \qquad (1)$$

where $n[l]$ is considered as a Gaussian noise of variance σ^2 comprising the thermal noise effects as well as the multiuser interference. It is worth pointing out that the channel viewed at the mobile station is in fact the combination of the beamforming coefficients and the propagation channel. Therefore, at the mobile station, the channel estimation will provide an estimate of $\mathbf{H}^H\mathbf{w} = \mathcal{H}(\mathbf{w})$.
The problem to be solved is the following:
assuming initial knowledge of the bearings, iteratively find the optimal set of coefficients $\mathbf{w}$ with the associated channel estimates such that the multi-user interference at the mobile is minimised.

3. THE ITERATIVE ALGORITHM

The proposed method involves the following steps:

1. Initialisation:
   - at the base station, $\mathbf{w}$ is initialised to $\mathbf{w}^{(0)}$ using the knowledge of the bearings.
   - at the mobile station, $\mathcal{H}(\mathbf{w})$ is initialised with a random value $\mathcal{H}(\mathbf{w})^{(0)}$.

2. at iteration i :
   - at the mobile station, estimate the downlink channel vector $\mathcal{H}(\mathbf{w}^{(i)})$. (see section 4)
   - feed $\mathcal{H}(\mathbf{w}^{(i)})$ back to the base station.
   - at the base station, compute $\hat{\mathbf{H}}^{(i+1)}$ using $\mathcal{H}(\mathbf{w}^{(i)})$ and $\mathbf{w}^{(i)}$
   - given $\hat{\mathbf{H}}^{(i+1)}$, compute at the base station the corresponding beamformer coefficients $\mathbf{w}^{(i+1)}$. (see section 5).

3. Termination
   When $\|\hat{\mathbf{H}}^{(i)} - \hat{\mathbf{H}}^{(i+1)}\| \leq \epsilon$, perform Maximum Aposteriori Probability (MAP) detection of non-coded bits.

4. BEAMFORMING

For the beamforming technique, we have selected the algorithm presented in [6] which maximises the Signal to Noise Ratio experienced at the mobile station. The beamformer coefficients are simply computed by:

$$\mathbf{w}^{(i)} = \arg\max_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{H}}^{(i)} \hat{\mathbf{H}}^{(i)^H} \mathbf{w} \qquad (2)$$

which means that $\mathbf{w}^{(i)}$ is the eigenvector associated to the largest eigenvalue of matrix $\mathbf{H}\mathbf{H}^H$.

5. CHANNEL ESTIMATION AND DATA DETECTION

We have selected a channel estimation scheme based on a Hidden Markov Model (HMM) and the Expectation Maximisation (EM) algorithm, as previously proposed for CDMA systems in [4]. This choice is motivated by the fact it is possible to include a measurement of the multi user interference as one of the parameters to be estimated (namely the variance of the noise $n[l]$ from equation 1. This is particularly useful in our case because it can be shown that this variance estimate drives the convergence of the EM algorithm in such a way that the only stable solution of the overall iterative scheme corresponds to a small value of the variance estimate. This is a key feature for this scheme since a small level of multi user interference means that the final beamformer has nulls in the directions where the interference is the highest. In this paper, we do not derive in full details the EM algorithm applied for channel estimation and data detection which can be found in [1]. We only report the main results which are useful for the understanding of the proposed iterative method.

Consider the vector $Y[l]$ containing N consecutive observations $Y[l] = [y[Nl]\ y[Nl-1] \cdots y[N(l-1)+1]]^T$. At chip time $Nl-j$, let us write the modulated chip $x[Nl-j]$ as a function of a spreading code chip and a non-coded symbol. Denoting by a and a' the integers such as $j = Na + a'$ with $a' \leq N$, we can easily check that

$$x[Nl-j] = s[l-a]c[a'] \qquad (3)$$

Therefore, if the Inter Chip Interference is of length M, the Inter Symbol Interference will be of length $m+1$, with m the integer such that: $M = Nm+m'$ with $m' < N$ Denoting by $\mathbf{s}[l]$ the vector $[\ s[l]\ s[l-1]\ \cdots s[l-m]\ ]$, one can rewrite equation 1 as

$$Y[l] = \mathcal{F}_c(\mathbf{s}[l])\mathcal{H}(\mathbf{w}) + N_k \qquad (4)$$

where the function $\mathcal{F}_c(\mathbf{s}[l])$, is only used for the purpose of showing that the observation vector $Y[l]$ depends on

the spreading code c and the non-coded sequence of transmitted symbols $[s[l]\ s[l-1]\ \cdots s[l-m]]$.

It is easy to show that the vector $\mathbf{s}[l]$ is a first order Markov process: it obeys the state equation:

$$\mathrm{s}[l+1] = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix} \mathrm{s}[l] + s[l+1] \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (5)$$

Therefore, equations 4 and 5 respectively correspond to the observation and state equations of a hidden Markov model. It is worth emphasising the fact that the hidden process (and therefore the detected sequence) is the sequence of bit before spreading. Consider a block of K vectors of observation data $\mathcal{Y} = (Y[1], Y[2], \cdots, Y[K])$. At iteration $i+1$ we process the whole block of data in order to produce the estimate of the combination of the beamformer and the propagation channel, namely $\mathcal{H}(\mathbf{w}^{(i)}) = \mathbf{H}^H \mathbf{w}^{(i)}$. Based on the HMM formulation, it can be shown on [1] that the solution provided by the block EM algorithm is given by:

$$\mathcal{H}(\mathbf{w}^{(i)})^{(i+1)} = \mathbf{R}^{-1}\mathbf{u} \quad (6)$$

where matrix $\mathbf{R}$ and vector $\mathbf{u}$ are expressed as

$$\mathbf{R} = \sum_{l=1}^{K} E\{\mathcal{F}_c(\mathrm{s}[l])^H \mathcal{F}_c(\mathrm{s}[l]) \,|\mathcal{Y}, \hat{\mathbf{H}}^{(i)}\} \quad (7)$$

$$\mathbf{u} = \sum_{l=1}^{K} E\{\mathcal{F}_c(\mathbf{s}[l])^H Y[l] \,|\mathcal{Y}, \hat{\mathbf{H}}^{(i)}\} \quad (8)$$

Note that the expectation $E\{\mathcal{F}_c(\mathbf{s}[l])|\mathcal{Y}, \hat{\mathbf{H}}^{(i)}\}$ is computed as:
$E\{\mathcal{F}_c(\mathrm{s}[l])|\mathcal{Y}, \hat{\mathbf{H}}^{(i)}\} = \sum_{j=1}^{2^{m+1}} Pr\{\mathrm{s}[l] = \xi_j|\mathcal{Y}, \hat{\mathbf{H}}^{(i)}\}\mathcal{F}_c(\xi_j)^H$
Here ξ_j is one among the 2^{m+1} possible realisations of the stochastic process $\mathbf{s}[l]$. The a-posteriori probability $Pr\{\mathrm{s}[l] = \xi_j|\mathcal{Y}, \hat{\mathbf{H}}^{(i)}\}$ is calculated by multiplying the so-called forward and backward variables of an HMM (see [1]). Note that in order to calculate these probabilities, we assume that the noise process $N[l]$ which corresponds to the multiuser interference and the thermal noise, is a white Gaussian process of unknown variance σ^2 which depends on the performance of the beamformer. The EM algorithm for estimating this parameter leads to:

$$\hat{\sigma}^{2(i+1)} = \frac{1}{K} \sum_{l=1}^{K} E\{\|Y[l] - \mathcal{F}_c(\mathrm{s})\hat{\mathcal{H}}(\mathbf{w}^{(i-1)})^{(i)}\|^2 |\mathcal{Y}, \hat{\mathbf{H}}^{(i)}\} \quad (9)$$

When the beamformer is not correctly set up, the variance estimate will be quite large. The a-posteriori probabilities will then be calculated with "flat" Gaussian functions and one can show that all $Pr\{\mathbf{s}[l] = \xi_j|\mathcal{Y}, \hat{\mathbf{H}}^{(i)}\}$ will have the same value. As a result, one can show that the channel estimates will converge to values close to zero. Therefore, the beamformer will not be able to use any particular information and will steer a beam in a random direction. In fact, as long as the variance estimate is large, the beamformer will go on steering beams in all possible directions until it finds one that corresponds to a lower multi user interference. In this case, the variance estimate will be small, resulting in an accurate channel estimate which will allow the beamformer to refine its configuration while keeping the same general steering. Indeed, the channel taps which correspond to the nulls will be estimated to zero at the mobile station, which means that the beamformer will further disregard these channel taps and will not transmit in their direction any longer. Therefore, the iterative algorithm will have converged.

6. SIMULATIONS

In this section, we highlight the potential benefits of the proposed method. The desired user is characterised by a spreading gain of 5 and the number of multipaths between the BS and the mobile station has been set to 9. We consider the transmission of the signal through an array of 10 antenna elements. In addition to the desired user, the system is supporting 10 additional users. Therefore, there will be a non-neglectible multiuser interference at the desired mobile station.

In figure 1, the upper graph shows the angular power spectra respectively for the interfering users (dashed curve) and the desired user (plain curve) as a function of their angle of arrival at the Base Station. Note that some multipath components for the desired user share common bearings with some interfering signals. Therefore, it is expected that the channel estimation operation at the first iteration will suffer from multiuser interference. This is confirmed by the graph which shows the beampattern of the beamformer after the first iteration of the algorithm. One can see that the beamformer is clearly steering a beam towards directions where the desired user' signal is weak compared to the interference level. During iterations 2, 3 and 4, one can see that the beamformer is steering beams towards directions where the multi user interference is quite high. It is interesting to observe that at iteration 2, the beamformer is steering a beam in a direction where both the desired user and interference signal are weak. Although the resulting estimated multi user interference variance is low (see on figure 2), the solution is not acceptable for the iterative algorithm: this is because at this stage, the beamformer will attempt to steer a beam in a direction where the SNR is maximised (see beampattern at iteration 3). However, the direction

where the SNR is maximised, is also a direction where the multi user interference is high. Therefore, this set up is unstable as at the next iteration, the interference variance will be high again, resulting in a poor channel estimate. The beamformer then keeps on scanning. Finally, at iteration 4, the beamformer finds a direction where the multi user interference is low and the SNR is maximised as well. After the beampattern is refined in this general direction at iteration 5, the algorithm reaches the convergence stage.
The graph on figure 2 plots the estimates of the fading coefficients and the interference variance, the latter playing a key role in the convergence of the iterative scheme.

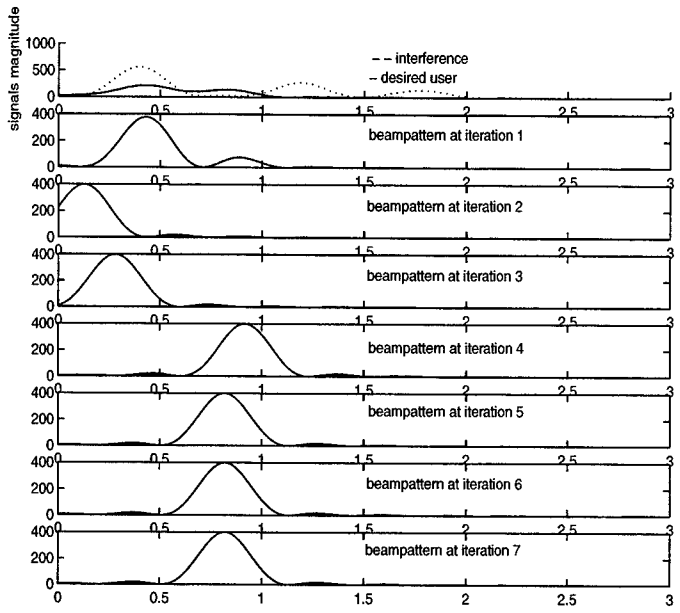


Figure 1: power spectrum of desired and interference signals and beampatterns during iterations

7. CONCLUSION

In this paper, we have presented a method for efficiently combining channel estimation and downlink beamforming for CDMA systems, in cases where the Rake receiver cannot be used for channel estimation purposes. This method relies on an iterative scheme which iterates between a channel estimation scheme which is only stable when the multi user interference is low and a beamforming operation which maximises the received Signal to Noise Ratio. Simulation results presented in this paper show that this iterative scheme seems to converge to solutions which maximise the Signal to Noise

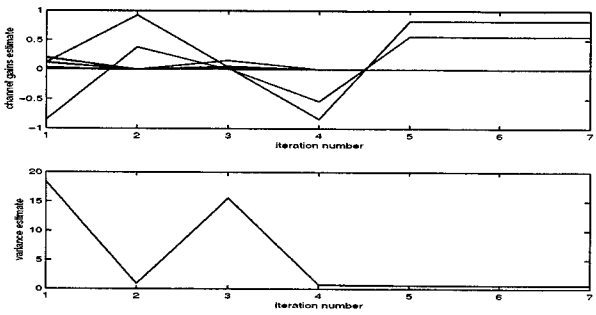


Figure 2: Fading coefficient and noise+interference variance estimates

plus Interference Ratio (SINR) which is an attractive feature since it is achieved without taking into account other user's statistics. Additional work is currently undertaken to analyse the convergence properties of this iterative scheme.

8. REFERENCES

[1] G.K. Kaleh and R. Vallet, "Joint parameter estimation and symbol detection for linear or nonlinear unknown dispersive channels," IEEE Trans. on Comms. January 1994.

[2] J.S. Thompson, J.E. Hudson, P.M. Grant, and B. Mulgrew, "CDMA downlink beamforming for frequency selective channels," Proceedings of PIMRC'99, Osaka, Japan, September 1999.

[3] P. Zetterberg and B. Ottersten, "The spectrum efficiency of a base station antenna array system for spatially selective transmission," in Proc. IEEE Vehicular Technology Conference, pp. 1517-1521, 1994.

[4] S. L. Perreau and L. B. White, "Channel estimation and symbol detection for multiuser CDMA receivers using HMMs", Proceedings of the 9th IEEE workshop SSAP, Portland, Oregon, USA, pp.272-275, September 1998,

[5] F. Swarts etal. CDMA techniques for third generation mobile systems, pp285-300 Kluwer academic publishers, 1999.

[6] Jinho Choi, Sylvie Perreau and Yong Lee Semiblind method for adaptive transmit antenna array for CDMA systems , proceedings of the IEEE VTC2000 fall conference, boston, sept. 2000.

FEATURE DISCOVERY AND SENSOR DISCRIMINATION IN A NETWORK OF DISTRIBUTED RADAR SENSORS FOR TARGET TRACKING[1]

S. Kadambe

HRL Laboratories, LLC, 3011 Malibu Canyon Road, Malibu CA 90265, USA
E-mail: skadambe@hrl.com

ABSTRACT

Spatially distributed network of radar sensors are being used for target tracking and for generating Single Integrated Aerial Picture (SIAP). In such a network generally each sensor sends whatever target track/association information it has to every other sensor. This has the disadvantage of requiring more communication bandwidth and processing power. One of the ways to reduce the communication bandwidth and the processing power is to discover features that would improve the target detection/track accuracy and activate those sensors that would provide the missing information and, form clusters of sensors that have consistent information. In this paper, we describe a minimax entropy based technique for feature discovery and within class entropy based technique for feature/sensor discrimination. After discovering the features, those sensors that can provide the discovered features are activated. The decision based on the sensor discrimination is used in cluster formation. The experimental details and simulation results that are provided here indicate that these metrics are efficient in discovering features and in discriminating sensors. The techniques described in this paper are dynamic in nature – as it acquires information it is making a decision on whether it is from a good sensor in terms of consistency. This has the advantage of discarding non-valid information dynamically and making progressive decision.

1. INTRODUCTION

Spatially distributed network of radar sensors are being used for target tracking and for generating Single Integrated Aerial Picture (SIAP). In such a network generally every sensor node has the same information. This is achieved by each sensor sending whatever target track/association information it has to every other sensor. This has the disadvantage of requiring more communication bandwidth and processing power. One of the ways to reduce the communication bandwidth and the processing power is to discover features that would improve the target detection/track accuracy and activate those sensors that would provide the missing information and, form clusters of sensors that have consistent information. The sensors that are part of a cluster will communicate using a high bandwidth by sending information that each sensor has to the other members of the cluster and the clusters themselves communicate with each other using low bandwidth communication network by transmitting only the fused information about a target track to other clusters. In this paper, we describe a minimax entropy based technique for feature discovery and within class entropy based technique for feature/sensor discrimination. After discovering the features, those sensors that can provide the discovered features are activated. The decision based on the sensor discrimination is used in cluster formation and mutual information metric is used in information fusion to improve the track accuracy. To the best knowledge of the author of this paper there is no study on sensor discrimination using within class entropy metric is reported even though, there is one study on using mutual information for selecting a subset of features from a bigger set that is described in [1]. The technique described in this paper uses within class entropy as a metric to discriminate good sensor vs. bad sensor. Unlike our technique, the technique in [1] is static in nature and cannot handle the case where the dimensionality of the feature set varies. In [2], the author shows that in general by fusing data from selective sensors the performance of a network of sensors can be improved. However, in this study, no specific novel metrics for the feature discovery and feature/sensor discrimination were developed unlike in this paper. In [3], techniques to represent Kalman filter state estimates in the form of information – Fisher and Shannon entropy are provided. In such a representation it is straightforward to separate out what is new information from what is either prior knowledge or common information. This separation procedure is used in decentralized data fusion algorithms that are described in [3]. However, to the best knowledge of this author no study has been reported on using minimax entropy principle for the feature discovery. In addition, the significance of this study is the possible application of feature discovery and sensor discrimination in the formation of a cluster of distributed sensors including radar sensors to improve the decision accuracy such as target tracking accuracy in the case of network of radars and reducing the communication bandwidth requirements. In the next section, proposed feature discovery and sensor discrimination techniques are described. The simulation description and experimental results are provided in section 3. Conclusions and future research directions are provided in section 4.

2. A BRIEF DESCRIPTION OF THE PROPOSED TECHNIQUES

2.1 Discovery of missing information:

In the case of (a) target detection, identification and tracking, (b) classification, (c) coalition formation, etc., applications, the missing information could correspond to feature discovery. This helps in only probing (awakening) the sensor node that can provide the missing information and thus save power and processing by not arbitrarily activating nodes. We apply the minimax entropy principle described in [4] for the feature discovery. The details of estimation of missing information in other words feature discovery using the minimax entropy principle are as follows.

[1] © 2001 HRL Laboratories, LC. All Rights Reserved

2.1.1 Minimax entropy principle:

Let N given values corresponds to n different information types. Let z_{ij} be the jth member of ith information type (where the information type is defined as a cluster of values that give similar information measures) so that

$$j = 1,2,\ldots, m_i; i = 1,2,\ldots, n; \sum_{i=1}^{n} m_i = N. \qquad \text{Eq. (1)}$$

Then the entropy for this type of classes of information is:

$$H = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{z_{ij}}{T} \ln \frac{z_{ij}}{T} \text{ where } T = \sum_{i=1}^{n}\sum_{j=1}^{m_i} z_{ij}. \qquad \text{Eq. (2)}$$

Let $T_i = \sum_{j=1}^{m_i} z_{ij}$.

Using this H can be written as:

$$H = \sum_{i=1}^{n} \frac{T_i}{T} H_i - \sum_{i=1}^{n} \frac{T_i}{T} \ln \frac{T_i}{T} = H_W + H_B \text{ where}$$

$$H_i = -\sum_{j=1}^{m_i} \frac{z_{ij}}{T} \ln \frac{z_{ij}}{T} \qquad \text{Eq. (3)}$$

the entropy of values that belong to information type i.

In the equation above, H_w & H_B are entropy of within classes (information types) and between classes, respectively. We would like types of information to be as distinguishable as possible and we would like the information within each type to be as homogenous as possible. The entropy is high if the values belonging to a type (class) represent similar information and is low if they represent dissimilar information. Therefore, we would like H_B to be as small as possible and H_w as large as possible. This is the principle of minimax entropy.

2.1.2 Application of minimax entropy principle for feature discovery:

Let z be the missing value (feature). Let T be the total of all known values such that the total of all values is $T + z$. Let T_I be the total of values that belong to information type to which z may belong. T_I+z then is the total of that particular type of information. This leads to:

$$H = -\sum{}' \frac{z_{ij}}{T+z} \ln \frac{z_{ij}}{T+z} - \frac{z}{T+z} \ln \frac{z}{T+z} \qquad \text{Eq. (4)}$$

$$H_B = -\sum{}'' \frac{T_i}{T+z} \ln \frac{T_i}{T+z} - \frac{T_1+z}{T+z} \ln \frac{T_1+z}{T+z}.$$

Here, Σ' denotes the summation over all values of i, j except that correspond to the missing information & Σ'' denotes the summation over all values of i except for the type to which the missing information belongs, respectively.

We can then estimate z by minimizing H_B/H_W or $H_B/(H-H_B)$ or H_B/H or by maximizing $(H-H_B)/H_B$ or H/H_B. The estimates of z provide the missing information values (features) and information (feature) type. From the above discussion we can see that we will be able to discover features as well as type of sensor from which these features can be obtained. This has the advantage of probing the appropriate sensor in a distributed network of sensors. The transfer of information and probing can be achieved in such a network by using network routing techniques. Before trying to use the newly acquired feature set, it is advisable to check the relevance of the feature set in terms of consistency/improving the accuracy to reduce the cost of processing. In a distributed network of sensors this has an added advantage of reducing the communication cost. We measure the relevance in other words discriminate feature set from a good sensor vs. bad sensor by using the within class entropy that is described below.

2.2 Measure of consistency:

We measure relevance by measuring consistency. For this we have developed a metric based on within class entropy that is described in this section. Let there are N events (values) that can be classified in to m classes and let an event x_{ij} be the jth member of ith class where i = 1,2,.....,m, j = 1,2,.....,n_i and $\sum_{i=1}^{m} n_i = N$. The entropy for this classification is:

$$H = \sum_{i=1}^{m}\sum_{j=1}^{n_i} p(i)\, p(x_{ij}) \log\left(\frac{1}{p(i)\, p(x_{ij})}\right)$$

$$= -\sum_{i=1}^{m}\sum_{j=1}^{n_i} p(i)\, p(x_{ij}) \log\left(p(i)\, p(x_{ij})\right)$$

$$= -\sum_{i=1}^{m} p(i) \sum_{j=1}^{n_i} p(x_{ij}) \log(p(x_{ij})) - \sum_{i=1}^{m} p(i) \log(p(i)) \sum_{j=1}^{n_i} p(x_{ij})$$

$$= \sum_{i=1}^{m} p(i) H_i - \sum_{i=1}^{m} p(i) \log(p(i))$$

since - $\sum_{j=1}^{n_i} p(x_{ij}) \log(p(x_{ij}))$ is the entropy of a class i

& $\sum_{j=1}^{n_i} p(x_{ij}) = 1$

$= H_W + H_b$ where H_W is called the entropy within classes

and H_b is called the entropy

between classes.

The entropy H_w is high if the values or events belonging to a class represent similar information and is low if they represent dissimilar information. This means H_w can be used as a measure to define consistency. That is, if two or more sensor measurements are similar then their H_w is greater than if they are dissimilar. Therefore, this measure can be used in sensor discrimination or selection. Note that even though the definitions of within class and between class entropy here are slightly different from section 2.1, they are similar in concept. Note also that the minimax entropy measure that uses both within and between class entropies was used earlier in the estimation of missing information; but here, within class entropy is defined as a consistency measure that can be used in sensor discrimination or selection. These two metrics have different physical interpretations and are used for different purposes.

3 SIMULATIONS

Above described feature discovery and sensor discrimination algorithm has been applied for the feature discovery, sensor discrimination and cluster formation in a network of radar sensors. Note that this simulation is simplified since the goal is to prove the concepts.

This network of sensors is used for tracking multiple targets. Each sensor node has a local and global Kalman filter based target trackers. These target trackers estimate the target states - position and velocity in Cartesian co-ordinate system. The local tracker uses the local radar sensor measurements to estimate the state estimates while the global tracker fuses target states obtained from other sensors if it is consistent and improves the accuracy of the target tracks.

For the purposes of testing the feature/sensor discrimination algorithm, a network of three radar sensors and a single moving target with constant velocity were considered. Two sensors were unbiased and thus considered as good and one was biased and thus considered as bad. The bias was introduced as the addition of a random number to the true position of a target. The bias was introduced this way because the biases in azimuth and range associated with a radar sensor translate into measured target position that is different from the true target position. In addition, currently in our simulations, we are assuming that the sensors are measuring the target's position in the Cartesian co-ordinate system instead of the polar co-ordinate system. The amount of bias was varied by multiplying the random number by a constant k i.e., measured position = (true position + k * randn) + measurement noise.

The measurements from a radar at each sensor node was used to estimate the target states using the local Kalman filter algorithm. The estimated target states at each sensor node were transmitted to other nodes. For this simulation, only estimated position was considered for simplicity.

We consider the estimated state vector as the feature set here. Since the goal of this simulation is proof of concept, the feature/sensor discrimination algorithm was implemented at sensor node 1 with the assumption it is a good sensor. Let the state estimate outputs of this node be A_g. Let the state estimate outputs of a second sensor correspond to B_g and a third sensor correspond to B_b.

For the computation of entropy probability values are needed as seen from the equation above. To obtain these values, ideally, one would need probability distribution functions (pdfs). However, in practice it is hard to obtain closed form pdfs. In the absence of knowledge of actual pdfs it is a general practice to estimate them by using histograms [5]. Researchers in signal and image processing use this technique most commonly [6]. Therefore, we use the histogram approach here. In order to obtain the histograms, initially, we need some data (features) to know how it is distributed. For this purpose, it was assumed that initially N state estimate vectors were accumulated at each sensor node and this accumulated vector was transmitted to other nodes. Note also that the accuracy of probability estimates using the histogram approach depends on the amount of accumulated (training) data. Also for non-stationary features, it depends on how often the histograms are updated. In practice, since the training data is limited we have set N to 10 in this simulation. To take care of the non-stationarity of the features, initially, we wait till N estimates are obtained at each node. From then on we update the histograms every time instant using the new state estimate and previous nine state estimates. At each time instant we discard the oldest feature (oldest state estimate).

To get the probability of occurrence of each feature vector, first the histogram was computed. For this, bin size N_{bin} of 5 was used. The center point of each bin was chosen based on the minimum and maximum feature values. In this simulation the bin centers were set as:

$$\text{min(feature values)} + \left(0 : N_{bin} - 1\right) * \frac{\text{max(feature values)} - \text{min(feature values)}}{N_{bin}}$$

Since the histogram provides the number of elements in a given bin, it is possible to compute the probabilities from the histogram. In particular it is computed as:

$$\frac{\text{\# elements in a particular bin}}{\text{total number of elements}}.$$

First, the minimax entropy principle was applied to find the missing information, the appropriate sensor was probed to obtain that information, then the consistency measure – within class entropy was applied to check whether the new information obtained from that particular sensor is consistent with the other sensors.

In the following two figures, within class entropy is plotted for feature discovered from two unbiased sensors and, one biased and one unbiased sensor. The measurement noise level was kept the same for all three sensors. However, k was set to 1.0 in Figure 1 and was set to 2 in Figure 2. The within class entropy was computed for different iterations using the definition provided in the previous section. The probability values needed in this computation were estimated using the histogram approach as described before. From these two figures, it can be seen that the within class entropy of two unbiased sensors is greater than

the within class entropy of one biased and one unbiased sensors. This indicates that the within class entropy can be used as a consistency measure to discriminate between sensors or to select sensors.

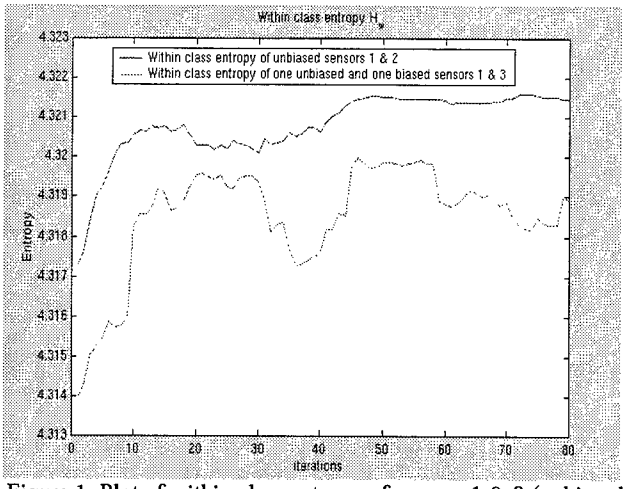


Figure 1: Plot of within class entropy of sensors 1 & 2 (unbiased sensors) and, 1 (unbiased) and 3 (biased). Bias constant k = 1

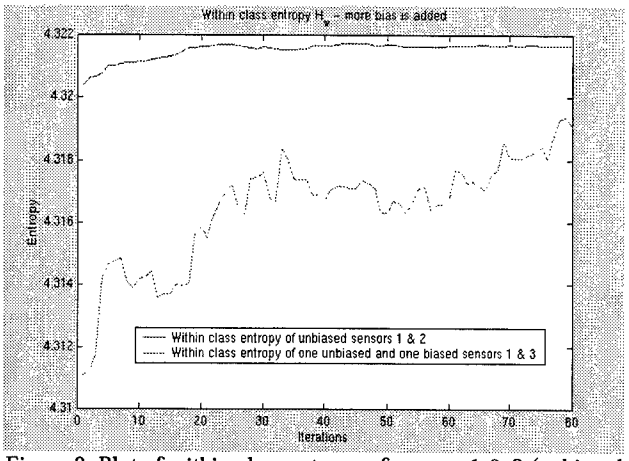


Figure 2: Plot of within class entropy of sensors 1 & 2 (unbiased sensors) and, 1 (unbiased) and 3 (biased). Bias constant k = 2

We then form a cluster of sensors that are consistent and apply the mutual information metric that we have developed in [7] to verify whether the mutual information increases. In [7], we have shown that by fusing information from sensors when the mutual information increases, the decision accuracy improves. We transmit the fused decision (which requires much lower bandwidth compared to the transmission of decision of each sensor to every other in the network) to other clusters of sensors and thus reduce the communication bandwidth requirement.

4 CONCLUSIONS

In this paper, we have described how minimax entropy principle can be used in feature (missing information) discovery. Further, in this paper a consistency measure is defined and it has been shown that this measure can be used in discriminating sensors. We also have shown how these two measures are used in the cluster formation and reduction of communication bandwidth requirement. The application of techniques in not restricted to network of radar sensors but can be applied to any other type of sensors. Hence, these techniques can be used in a distributed network of any type of sensors for hierarchical processing, for cluster formation and for data fusion. Future work warrant implementation of these techniques on distributed sensor hardware nodes with multiple sensors and test the capabilities of these algorithms in hierarchical processing and cluster formation in field.

5 References:

1. R. Battti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. On Neural Network*, vol. 5, no. 4, July 1994, pp. 537-550.
2. S. C. A. Thomopoulos, "Sensor selectivity and intelligent data fusion," *Proc. Of the IEEE MIT'94*, October 2-5, 1994, Las Vegas, NV, pp. 529-537.
3. J. Manyika and H. Durrant-Whyte, Data fusion and sensor management: An information theoretic approach, Prentice Hall, 1994.
4. J. N. Kapur, Measures of information and their applications, John Wiley, Eastern Limited, 1994.
5. G. A. Darbellay, I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4 , May 1999, pp. 1315-1321.
6. L. R. Rabiner and B-H. Juang, Findamentals of Speech Recognition, Prentice Hall, 1993, Chapter 6.
7. S. Kadmbe, "Information theoretic based sensor discrimination for information fusion and cluster formation in a network of distributed sensors," *to appear in Proc. Of FUSION '01*, August 8-10, 2001 Montreal, Canada.

AN INFORMATION DIVERGENCE MEASURE FOR ISAR IMAGE REGISTRATION

A. Ben Hamza, Yun He, Hamid Krim

Department of Electrical and Computer Engineering
North Carolina State University, Raleigh, NC 27695-7914, USA
E-mails: {abhamza, yhe2, ahk}@eos.ncsu.edu

ABSTRACT

Entropy-based divergence measures have shown promising results in many areas of engineering and image processing. In this paper, a generalized information-theoretic measure called *Jensen-Rényi* divergence is proposed. Some properties such as convexity and its upper bound are derived. Using the Jensen-Rényi divergence, we propose a new approach to the problem of ISAR (Inverse Synthetic Aperture Radar) image registration. The goal is to estimate the target motion during the imaging time. Our approach applies Jensen-Rényi divergence to measure the statistical dependence between consecutive ISAR image frames, which would be maximal if the images are geometrically aligned. Simulation results demonstrate a much improved performance of the proposed method in image registration.

1. INTRODUCTION

Image registration is an important problem in computer vision, remote sensing, data processing and medical image analysis. The objective of image registration is to find a spatial transformation such that a dissimilarity metric achieves its minimum between two or more images taken at different times, from different sensors, or from different viewpoints.

Inverse Synthetic Aperture Radar (ISAR) [1] is a microwave imaging system capable of producing high resolution imagery from data collected by a relatively small antenna. The ISAR imagery is induced by target motion, however, motion also blurs the resulting image. After conventional ISAR translational focusing process, image registration can be applied to estimate the target rotational motion parameter, then polar re-formating can be used to achieve a higher resolution image.

During the last three decades, a wide range of registration techniques have been developed for various applications. These techniques can be classified [2] into correlation methods, Fourier methods, landmark mapping, and elastic model-based matching.

In the work of Woods [3] and Viola [4], mutual information, a basic concept from information theory, is introduced as a measure for evaluating the similarity between images. When the two images are properly matched, corresponding areas overlap, and the resulting joint histogram contains high values for the pixel combinations of the corresponding regions. When the images are misregistered, non-corresponding areas also overlap and this will result in additional pixel combinations in the joint histogram. In case of misregistration, the joint histogram has less sharp peaks and is more dispersed than the correct alignment of the images. The registration criterion is then to find the transformation such that the mutual information of the corresponding pixel pair intensity values in the matching images is maximized. This approach is accepted by many [5] as one of the most accurate and robust registration measures.

In this paper, a novel generalized information theoretic measure, called *Jensen-Rényi* divergence and defined in terms of Rényi entropy [6] is introduced. Jensen-Rényi divergence is defined as the similarity measurement of any finite number of weighted probability distributions. Shannon mutual information is a special case of the Jensen-Rényi divergence. This generalization endows us the ability to control the measurement sensitivity of the joint histogram, which would end up with a better registration result.

In the section that follows, we give a brief statement of the problem. In section 3, we introduce the Jensen-Rényi divergence and its properties. Section 4 is devoted to the application of the Jensen-Rényi divergence in ISAR image registration. Finally, we provide some concluding remarks in the section 5.

2. PROBLEM STATEMENT

ISAR imagery registration can be applied to estimate the target motion during the imaging time. Let $\mathcal{T}_{(l,\theta,\gamma)}$ be a Euclidean transformation with translational parameter $l = (l_x, l_y)$, rotational parameter θ and scaling parameter γ. Given two ISAR images f and r, the objective of image registration is to determine the spatial transformation parameters $(l^*, \theta^*, \gamma^*)$ such that

$$(l^*, \theta^*, \gamma^*) = \arg\max_{(l,\theta,\gamma)} JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1(f, \mathcal{T}_{(l,\theta,\gamma)}r), \boldsymbol{p}_2(f, \mathcal{T}_{(l,\theta,\gamma)}r), \ldots, \boldsymbol{p}_n(f, \mathcal{T}_{(l,\theta,\gamma)}r)) \quad (1)$$

where $JR_\alpha^{\boldsymbol{\omega}}(\cdot)$ is the Jensen-Rényi divergence with order α and weight $\boldsymbol{\omega}$. Denote $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_k\}$ the sets of pixel intensity values of f and $\mathcal{T}_{(l,\theta,\gamma)}r$ respectively, then $\omega_i = P(X = x_i)$ and $\boldsymbol{p}_i(f, \mathcal{T}_{(l,\theta,\gamma)}r) = (p_{ij})_{1\le j\le k}$. $p_{ij} = P(Y = y_j | X = x_i)$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$ is the conditional probability of $Y = y_j$ given $X = x_i$ for the corresponding pixel pairs in f and $\mathcal{T}_{(l,\theta,\gamma)}r$. Here the Jensen-Rényi divergence acts as a similarity measure between images, which will be explained further in the next section.

ISAR imagery is induced by target motion, however, the target motion causes time-varying spectra of the received signals. Motion compensation has to be carried out to obtain a high resolution image. As the radar keeps tracking the target, the reflected signal is continuously recorded during the imaging time. By registration of a sequence of consecutive image frames, $\{f_i\}_{i=0}^N$, the target

This work was supported by an AFOSR grant F49620-98-1-0190 and by ONR-MURI grant JHU-72798-S2 and by NCSU School of Engineering.

motion during the imaging time can be estimated by interpolating $\{(l_i, \theta_i, \gamma_i)\}_{i=1}^{N}$. Then based on the trajectory of the target, translational motion compensation (TMC), and rotational motion compensation (RMC) [1] can be used to generate a clear image of the target.

3. THE JENSEN-RÉNYI DIVERGENCE

Let $k \in \mathbb{N}$ and $X = \{x_1, x_2, \ldots, x_k\}$ be a finite set with a probability distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$, i.e. $\sum_{j=1}^{k} p_j = 1$ and $p_j = P(X = x_j) \geq 0$, where $P(\cdot)$ denotes the probability.

Rényi entropy is a generalization of Shannon entropy, and is defined as [6]

$$R_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha} \log \sum_{j=1}^{k} p_j^\alpha, \quad \alpha > 0 \text{ and } \alpha \neq 1. \tag{2}$$

For $\alpha > 1$, the Rényi entropy is neither concave nor convex.

For $\alpha \in (0, 1)$, it is easy to see that Rényi entropy is concave, and tends to Shannon entropy $H(\boldsymbol{p})$ as $\alpha \to 1$. It can be easily verified that R_α is a non-increasing function of α, and hence

$$R_\alpha(\boldsymbol{p}) \geq H(\boldsymbol{p}), \quad \forall \alpha \in (0, 1). \tag{3}$$

Definition 1 *Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ be n probability distributions of X and $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_n)$ be a weight vector such that $\sum_{i=1}^{n} \omega_i = 1$ and $\omega_i \geq 0$. The Jensen-Rényi divergence is defined as*

$$JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n) = R_\alpha\left(\sum_{i=1}^{n} \omega_i \boldsymbol{p}_i\right) - \sum_{i=1}^{n} \omega_i R_\alpha(\boldsymbol{p}_i),$$

where $R_\alpha(\boldsymbol{p})$ is the Rényi entropy, $\alpha > 0$ and $\alpha \neq 1$.

Using the Jensen inequality, it is easy to check that the Jensen-Rényi divergence is nonnegative for $\alpha \in (0, 1)$. It is also symmetric and vanishes if and only if the probability distributions $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ are equal, for all $\alpha > 0$.

When $\alpha \to 1$, the Jensen-Rényi divergence is exactly the generalized Jensen-Shannon divergence [7].

Unlike other entropy-based divergence measures such as the well-known Kullback Leibler divergence, the Jensen-Rényi divergence has the advantage of being symmetric and generalizable to any finite number of probability distributions, with a possibility of assigning weights to these distributions.

In the sequel, we will restrict $\alpha \in (0, 1)$, unless specified otherwise, and will use a base 2 for the logarithm, i.e., the measurement unit is in *bits*.

The following result establishes the convexity of the Jensen-Rényi divergence of a set of probability distributions.

Proposition 1 *For $\alpha \in (0, 1)$, the Jensen-Rényi divergence $JR_\alpha^{\boldsymbol{\omega}}$ is a convex function of $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$.*

Proof. Recall that the mutual information between two finite sets $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_k\}$ is given by [8]

$$I(X;Y) = H(Y) - H(Y|X), \tag{4}$$

where $H(Y)$ is the Shannon entropy of Y and $H(Y|X)$ is the conditional Shannon entropy of Y, given X.

Instead of using Shannon entropy in (4), the mutual information can be generalized using Rényi entropy. Therefore, the α-mutual information can be defined as

$$I_\alpha(X;Y) = R_\alpha(Y) - R_\alpha(Y|X),$$

where R_α is the Rényi entropy of order $\alpha \in (0, 1)$.

Denote by $P(X = x_i) = \omega_i$, $P(Y = y_j | X = x_i) = p_{ij}$ and $P(Y = y_j) = q_j$, then it is easy to check that

$$R_\alpha(Y) - R_\alpha(Y|X) = JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n), \tag{5}$$

where $\boldsymbol{p}_i = (p_{ij})_{1 \leq j \leq k}$, for all $i = 1, \ldots, n$.

For fixed ω_i, the mutual information is a convex function of p_{ij} [8], then it can be verified that the α-mutual information is also a convex function of p_{ij}, leading to the Jensen-Rényi divergence a convex function of $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$. ■

Proposition 2 *The Jensen-Rényi divergence achieves its maximum value when $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ are degenerate distributions.*

Proof. The domain of $JR_\alpha^{\boldsymbol{\omega}}$ is a convex polytope in which the vertices are degenerate probability distributions. That is, the maximum value of the Jensen-Rényi divergence occurs at one of the degenerate distributions. ■

Since the Jensen-Rényi divergence is a convex function of $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$, then it achieves its maximum value when the Rényi entropy function of the $\boldsymbol{\omega}$-weighted average of degenerate probability distributions, achieves its maximum value too.

Next problem is to assign weights ω_i to the degenerate distributions $\Delta_1, \Delta_2, \ldots, \Delta_n$, that is to say, an assignment $\{\omega_i\} \longrightarrow \Delta_i = \{\delta_{ij}\}$ must be found, where $\{\delta_{ij}\}$ are probability mass functions, i.e. $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The following upper bound thus holds

$$JR_\alpha^{\boldsymbol{\omega}} \leq R_\alpha\left(\sum_{i=1}^{n} \omega_i \Delta_i\right). \tag{6}$$

Without loss of generality, consider the Jensen-Rényi divergence with equal weights $\omega_i = 1/n$ for all i, and denote it simply by JR_α. Using (6), the following holds

$$JR_\alpha \leq R_\alpha(\boldsymbol{a}) + \frac{\alpha}{\alpha - 1} \log(n), \tag{7}$$

where

$$\boldsymbol{a} = (a_1, a_2, \ldots, a_k) \quad \text{such that} \quad a_j = \sum_{i=1}^{n} \delta_{ij}. \tag{8}$$

Since $\Delta_1, \Delta_2, \ldots, \Delta_n$ are degenerate distributions, then we have $\sum_{j=1}^{k} a_j = n$. From (7), it is clear that JR_α achieves its maximum value when $R_\alpha(\boldsymbol{a})$ achieves its maximum too.

In order to maximize $R_\alpha(\boldsymbol{a})$, the concept of majorization will be used [9]. Let $(x_{[1]}, x_{[2]}, \ldots, x_{[k]})$ denote the non-increasing arrangement of the components of a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$.

Definition 2 *Let $\boldsymbol{a}$ and $\boldsymbol{b} \in \mathbb{N}^k$, $\boldsymbol{a}$ is said to be majorized by $\boldsymbol{b}$, written $\boldsymbol{a} \prec \boldsymbol{b}$, if*

$$\begin{cases} \sum_{j=1}^{k} a_{[j]} = \sum_{j=1}^{k} b_{[j]} \\ \sum_{j=1}^{\ell} a_{[j]} \leq \sum_{j=1}^{\ell} b_{[j]}, \quad \ell = 1, 2, \ldots, k-1. \end{cases}$$

Since R_α is Schur-concave function, then $R_\alpha(\boldsymbol{a}) \geq R_\alpha(\boldsymbol{b})$ whenever $\boldsymbol{a} \prec \boldsymbol{b}$.

The following result establishes the maximum value of the Jensen-Rényi divergence.

Proposition 3 *Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ be n probability distributions with*

$$\boldsymbol{p}_i = (p_{i1}, p_{i2}, \ldots, p_{ik}), \quad \sum_{j=1}^{k} p_{ij} = 1, \quad p_{ij} \geq 0.$$

If $n \equiv r \ (mod\ k)$, $0 \leq r < k$, then

$$JR_\alpha \leq \frac{1}{1-\alpha} \log \left(\frac{(k-r)q^\alpha + r(q+1)^\alpha}{(qk+r)^\alpha} \right), \qquad (9)$$

where $q = (n-r)/k$, and $\alpha \in (0,1)$.

Proof. It is clear that the vector

$$\boldsymbol{g} = (\overbrace{q+1, \ldots, q+1}^{r}, \overbrace{q, \ldots, q}^{k-r})$$

is majorized by the vector $\boldsymbol{a}$ defined in (8). Therefore, $R_\alpha(\boldsymbol{a}) \leq R_\alpha(\boldsymbol{g})$. This completes the proof using (7). ■

According to proposition 3, when $n \equiv 0 \ (\text{mod } k)$ the following inequality holds

$$JR_\alpha(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n) \leq \log(k).$$

4. ISAR IMAGE REGISTRATION

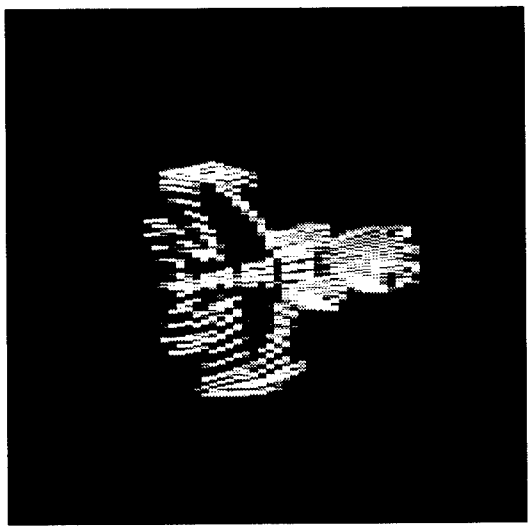


Fig. 1. ISAR image of moving target reconstructed by the Discrete Fourier Transform

To form a radar image, N bursts of received signals are sampled and organized burst by burst into a $M \times N$ two-dimensional array. This sample matrix is not uniformly spaced in the spatial frequency, instead, it is polar formatted data. The Discrete Fourier Transform processing of the polar formatted data would result in blurring at the edges of the target reflectivity image. Fig. 1 is a synthetic ISAR image of an aircraft MIG-25 [10]. The radar is assumed operating at $9GHz$ and transmits a stepped-frequency waveform. In each burst, 64 stepped frequency are used. The pulse repetition frequency is 15KHz. Basic motion compensation processing has been applied to the data. A total of 512 samples of the time history series are taken to reconstruct the image of this aircraft, which corresponds to $2.18s$ of integration time. As we can see, the resulting image is defocused due to the target rotation. In fact, the defocused image in Fig. 1 is formed by overlapping a series of MIG-25s at different viewing angles. By replacing the Fourier transform with the time varying spectral analysis techniques [11], we can take a sequence of snapshots of the target during the $2.18s$ of integration time. Fig. 2 shows the trajectory of the MIG-25, with 6 image frames taken at $t = 0.1280s$, $0.4693s$, $0.8107s$, $1.1520s$, $1.4933s$, $1.8347s$ respectively.

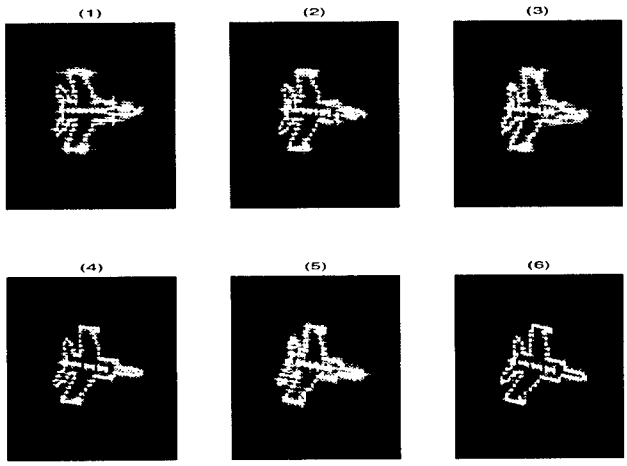


Fig. 2. Trajectory of a sequence of MIG-25 image frames

Then image registration can be applied to estimate the target motion in its trajectory. In this specific example, given a sequence of ISAR image frames $\{I_i\}_{i=0}^N$ which are observed in a time interval $[0,\ T]$, we search for the rotation angle $\{\theta_i\}_{i=1}^N$. Denote $r = I_{i-1}$ and $f = I_i$ for $i = 1, 2, \ldots N$, then by Equation (1), θ_i is given by

$$\theta_i^* = \arg\max_{\theta_i} JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1(f, \mathcal{T}_{\theta_i} r), \ldots, \boldsymbol{p}_n(f, \mathcal{T}_{\theta_i} r)).$$

Fig. 3 shows the rotation angles $\{\theta_i\}_{i=1}^N$ obtained by registering the 6 consecutive MIG-25 image frames. As we can see, α plays an important role in controlling the measurement sensitivity. When $\alpha < 1$, the peak of the JR-divergence vs θ is much shaper than the traditional Shannon entropy ($\alpha = 1$) based mutual information method. Obviously, a sharper peak would help to obtain a more accurate estimate of the true rotation angle.

By interpolating $\{\theta_i\}_{i=1}^N$, we obtain a trajectory of the MIG-25 rotational motion during the imaging time as shown in Fig. 3, then the polar re-formating [1] can be used to re-sample the received signal into rectangular format and generate a clear image of the MIG-25 based on all the received signals in the time interval $[0,\ 2.18s]$, as demonstrated in Fig. 4.

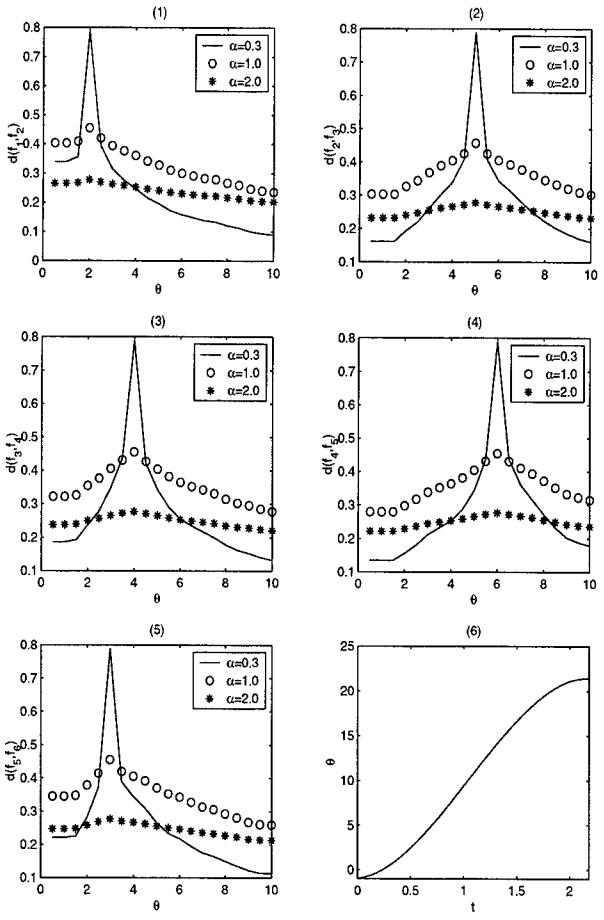


Fig. 3. Image registration of a MIG-25 Trajectory

5. CONCLUSIONS

A generalized information-theoretic divergence measure based on the Rényi entropy is proposed in this paper. We proved the convexity of this divergence measure and derived its maximum value. Using the Jensen-Rényi divergence, we propose a new approach to the problem of ISAR image registration. The ISAR imagery is induced by target rotation, which in turn causes time varying spectra of the reflected signals and blurs the target image. The goal of ISAR image registration is to estimate the target motion for further motion compensation processing. Our approach applies Jensen-Rényi divergence to measure the statistical dependence between consecutive ISAR image frames, which would be maximal if the images are geometrically aligned. Compared to the mutual information based registration techniques, the Jensen-Rényi divergence endows us the ability to control the measurement sensitivity of the joint histogram. This flexibility would result in a better registration accuracy. Maximization of the Jensen-Rényi divergence is a very general criterion, because no assumptions are made regarding the nature of this dependence and no limiting constraints are imposed on image contents. Simulation results demonstrate that our approach obtains an accurate estimation of target rotation automatically without any prior feature extraction.

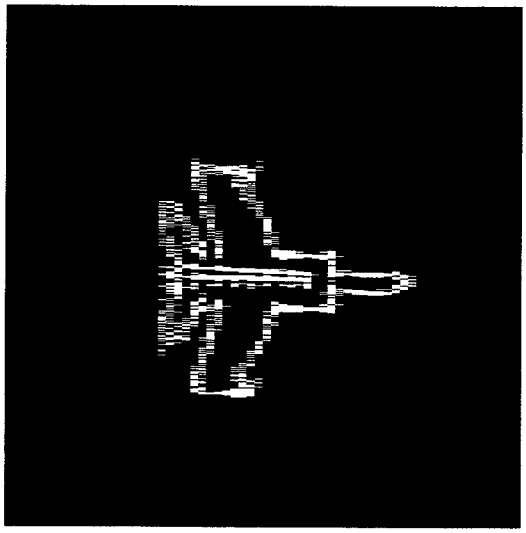


Fig. 4. Reconstructed MIG-25 by polar reformatting

6. REFERENCES

[1] D.R. Wehner, *High Resolution Radar, 2nd edition*,Artech House Inc., Norwood, MA 02062, USA, 1995.

[2] L. Brown, "A Survey of Image registration Techniques," *ACM Computing Surveys*, vol. 24, co. 4, pp. 325-376, 1992.

[3] R.P. Woods, J.C. Mazziotta, S.R. Cherry, "MRI-PET registration with automated algorithm," *J. Comput. Assist. Tomogr.*, vol. 17, no. 4, pp. 536-546, 1993.

[4] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 173-154, 1997.

[5] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. on Medical Imaging*, vol. 16, no. 2, pp. 187-198, 1997.

[6] A. Rényi, *On Measures of Entropy and Information*, Selected Papers of Alfréd Rényi, vol.2, pp. 525-580, 1961.

[7] J. Lin, "Divergence Measures Based on the Shannon Entropy," *IEEE Trans. Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[8] G. Gallager, *Information Theory and Reliable Communications*, John Willey Sons, 1968.

[9] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, 1979.

[10] V.C. Chen and S. Qian, "Joint Time Frequency Transform for Radar Range-Doppler Imaging," *IEEE Trans. Aerospace and Electronic Systems*, vol. 34, no. 2, pp. 486-499, 1998.

[11] Y. He, A. Ben Hamza, H. Krim, V.C. Chen, "An information theoretic measure for ISAR imagery focusing," *Proc. SPIE*, vol. 4116, 2000.

AN ADVANCED STAP IMPLEMENTATION FOR SURVEILLANCE RADAR SYSTEMS

G. A. Fabrizio and M. D. Turley

Radar Signal Processing Group, Surveillance Systems Division
Defence Science and Technology Organisation, Australia
e-mail: joe.fabrizio@dsto.defence.gov.au, mike.turley@dsto.defence.gov.au

ABSTRACT

Space-time adaptive processing (STAP) has emerged as a key technology for improving the performance of radar systems required to operate in the presence of severe and dynamic interference which generally includes clutter as well as jamming. While the theory of optimum STAP is well known, practical issues such as interference heterogeneity, finite sample support, mismatched signal models and computational load need to be overcome when it comes to implementing STAP in operational radar systems. This paper proposes an advanced STAP formulation which addresses important issues facing practical implementation and then tailors this general formulation for the case of interference rejection in over-the-horizon (OTH) radar to experimentally evaluate its target detection and localisation performance.

1. INTRODUCTION

The optimal STAP weight vector, which maximises the signal-to-interference plus noise ratio (SINR) at the radar output, is given in terms of the space-time interference covariance matrix and the desired signal response vector for each azimuth-range-Doppler bin processed [1,2]. In practice, the statistically expected interference covariance matrix is unknown and must be estimated from the received data, similarly the desired signal response vector is often estimated with a suitable model called a steering vector. The sample matrix inverse (SMI) technique directly substitutes the sample interference covariance matrix and the signal response vector model for their statistically expected or "true" forms in order to estimate the optimal STAP weight vector for the*adaptive* implementation.

The SMI technique is known to converge slowly towards the optimal solution when the "training data" used to form the sample covariance matrix contains the test cell(s) to be processed [3,4]. For this reason, the majority of STAP algorithms exclude the test cell(s), referred to as *primary data*, from the training set referred to as *secondary data*. Although this strategy avoids the potential for target self-nulling, degradations in output SINR may result if the second order statistics of the interference are heterogeneous over the primary and secondary data vectors, furthermore, the use of partitioned data sets inherits another latent but significant practical problem regarding unwanted detections or false alarms [5].

The presence of strong target signals or clutter discretes in the test cell but at an angle-Doppler different to the "look" angle-Doppler need to be suppressed from the output but may appear as or false alarms through the *sidelobes* of the adapted pattern because the secondary data does not contain information about them, and consequently, the resulting adaptive weights are unlikely to suppress such signals effectively. This problem has received insufficient attention in the literature, in fact, [6] states that the problem of target detection in a non-homogeneous test cell "has not been addressed" and proposes a method for addressing this "hitherto unsolved problem". The proposed method uses a non-statistical step which operates on the test cell only to cancel targets and clutter discretes in the sidelobes, followed by a statistical step employing secondary data to cancel residual interference in the test cell.

An alternative attack on this problem suggested by [7] involves the inclusion of a "piece" of the test cell in the training data set to "balance the risk of target self-nulling against the opportunity to capture the actual nonstationary interference present within the test cell". This intuitively appealing method proposes the use of a "deemphasis" factor ranging between 0 and 1 to scale the amount of the test cell included in the training data (0 = no test cell in training set, 1 = test cell completely in training set). A scheme for selecting the deemphasis factor was not proposed and the authors stated that "the best criterion for choosing this factor is an open problem and deserves further study".

This paper proposes a method for selecting the deemphasis factor for the test cell(s) in the primary data, moreover, this factor is dependent on the look angle-Doppler bin so that target self-nulling can effectively be avoided while *simultaneously* reducing false alarms and capturing the potentially different interference characteristics in the primary data. The proposed method is incorporated into an advanced STAP formulation which includes various features useful for practical implementation such as subspace models which take imperfect target signal coherence in angle and/or Doppler into account [8,9], localised processing for enhanced heterogeneous or nonstationary interference cancellation [10,11], rank reduction for increased convergence rate and computational speed [12] and alternative loading techniques to reduce degradations caused by finite sample support [13].

2. STAP ALGORITHM

The generalised sidelobe canceller (GSC) STAP implementation operates by forming a matched filter to a series of candidate signal models, denoted by steering vectors $\mathbf{s}(\psi)$ for different signal parameters $\psi = \psi_1, .., \psi_N$, to cover the search space quickly yet with some degree of reliability and subtracting an auxiliary adaptive filter $\mathbf{w}_a$ which has minimal impact on desired signals entering through the "main lobe" of $\mathbf{s}(\psi)$ but cancels as much interference as possible leaking through the "sidelobes" of $\mathbf{s}(\psi)$.

An adaptive GSC-STAP implementation which incorporates subspace signal models, localised processing, rank reduction, and steer dependent deemphasis factors is formulated in terms of the

The first author acknowleges A. Farina for useful dscussions on matched subspace detectors and B. White for replaying experimental data.

0-7803-7011-2/01/$10.00 ©2001 IEEE

minimiser $\mathbf{w}_a$ of the following criterion function;

$$f(\mathbf{w}) = \|\mathbf{D}\mathbf{X}(\mathbf{v}-\mathbf{w})\| + \lambda\|\mathbf{C}\mathbf{w}-\mathbf{f}\| + \kappa\|\mathbf{S}^{1/2}(\mathbf{v}-\mathbf{w})\| \quad (1)$$

where $\mathbf{v}$ is a (possibly tapered) steering vector with its dependence on ψ suppressed for notational convenience, $\|\cdot\|$ denotes Frobenius or Euclidean squared norm and the significance of other terms in (1) will be explained below. The *resultant* space-time adaptive weight vector $\mathbf{w}_r = \mathbf{v} - \mathbf{w}_a$ is then used to filter the primary space-time snapshot or test cell $\mathbf{x}$ to form a scalar GSC-STAP output $y = \mathbf{w}_r^H\mathbf{x}$. In general, the N-dimensional complex vector $\mathbf{x}$ is the sum of a signal component $\mathbf{s}$ with normalised space-time structure $\mathbf{s}^H\mathbf{s} = 1$, amplitude $\mu \geq 0$ and reference phase $e^{j\phi}$ and uncorrelated interference-plus-noise $\mathbf{n}$.

$$\mathbf{x} = \mu e^{j\phi}\mathbf{s} + \mathbf{n} \quad (2)$$

The interference-plus-noise in the test cell $\mathbf{n}$ is assumed to be zero-mean and multi-variate complex Gaussian distributed with statistically expected spatial covariance matrix $\mathbf{R}_n = E\{\mathbf{n}\mathbf{n}^H\}$. *Ideally*, the signal vector $\mathbf{s}$ can be represented by a space-time steering vector $\mathbf{s}(\psi_0)$ parameterised for example by $\psi_0 = [\theta_0, \omega_0]$ where θ_0 and ω_0 are the unknown signal angle-of-arrival and Doppler frequency respectively.

The first term (1) implements deemphasised localised processing by defining an $P \times N$ matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_P]^H$ composed of P test cell vectors $\mathbf{x}_p$ for $p = 1, .., P$ and a $P \times P$ diagonal matrix $\mathbf{D} = diag[\gamma_1 \cdots \gamma_P]$ containing the primary data deemphasis factors γ_p with their (ψ)-dependence momentarily suppressed. In essence, the role of the deemphasis factors is to "supervise" adaptive filter training by allowing $\gamma_p \rightarrow 1$ when the snapshot $\mathbf{x}_p$ is likely to contain interference only or interference plus a sidelobe signal (i.e. a *discrete interferer*) and $\gamma_p \rightarrow 0$ when $\mathbf{x}_p$ is likely to contain interference plus a target signal. A method for selecting the deemphasis factors $\gamma_p(\psi_n)$ for each system steer parameter ψ_n $(n = 1, .., N)$ is described later.

The second term in (1) concerns the application of linear constraints on the auxiliary adaptive weight vector $\mathbf{w}_a$ to prevent target cancellation when the criterion function $f(\mathbf{w})$ is minimised and to form optional "anticipatory" nulls or derivative constraints on the adapted pattern [14]. The $M < N$ linear constraints, specified by the $M \times N$ constraint matrix $\mathbf{C}$ and the M-dimensional constraint vector $\mathbf{f}$, can be enforced by letting $\lambda \rightarrow \infty$. To minimise target self-nulling caused by signal model mismatch (i.e. differences between the received signal $\mathbf{s}$ and the most closely matched steering vector $\mathbf{s}(\psi_n)$ where $n = 1, 2, ..N$), a low-rank linear subspace model can be adopted; $\mathbf{s}_\psi = \mathbf{H}(\psi)\mathbf{p}$ where the term $\mathbf{H}(\psi) \in C^{N\times M}$ is a pre-determined full rank mode matrix (e.g. a wave interference model $\mathbf{H}(\psi) = [\mathbf{s}(\psi + \Delta_1)\cdots\mathbf{s}(\psi + \Delta_M)]$ where the Δ_m for $m = 1, 2, .., M << N$ are positive or negative displacements closely clustered to the nominal system steer parameter ψ [8]) and $\mathbf{p} \in C^{M\times 1}$ is an unknown coordinate vector. In this case, setting $\mathbf{C}$ to the Hermitian of $\mathbf{H}(\psi)$, $\mathbf{f}$ to the zero vector and $\lambda \rightarrow \infty$ ensures that the auxiliary weight vector $\mathbf{w}_a$ estimated for sidelobe cancellation remains orthogonal to signals $\mathbf{s}_\psi$ spanned by the target subspace.

The final term in (1) represents the output power of the GSC (weighted by a factor κ) in a manner consistent with the *interference-plus-noise* sample covariance matrix $\mathbf{S}$ given by,

$$\mathbf{S} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{n}_k\mathbf{n}_k^H = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H \quad , \quad \mathbf{S}^{-1/2} = \mathbf{U}\mathbf{\Sigma}^{-1/2}\mathbf{U}^H \quad (3)$$

where $\mathbf{n}_k$ for $k = 1, .., K$ are judiciously chosen [15] *secondary* sample vectors assumed to contain interference-plus-noise only and $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ represents the eigen-decomposition of $\mathbf{S}$. This term stabilises the adaptive pattern so as to maintain effective interference cancellation when the number of primary data vectors in $\mathbf{X}$ that are likely to contain interference-plus-noise only is limited as well as to increase robustness against target self-nulling caused by deemphasis factor estimation errors.

When the target protection constraints are strictly enforced $(\lambda \rightarrow \infty)$ and no other constraints are applied $(\mathbf{f} = \mathbf{0})$, the solution for the resultant weight vector $\mathbf{w}_r$ is given by,

$$\mathbf{w}_r = \mathbf{Z}^{-1}\mathbf{C}^H[\mathbf{C}\mathbf{Z}^{-1}\mathbf{C}^H]^{-1}\mathbf{C}\mathbf{v} \quad (4)$$

where $\mathbf{Z} = \mathbf{X}^H\mathbf{D}^H\mathbf{D}\mathbf{X} + \kappa\mathbf{S}$ may be regarded as the deemphasised primary data covariance matrix "loaded" to a level κ by the secondary data covariance matrix $\mathbf{S}$. The resultant vector in (4) is calculated for each system steer parameter $\psi = \psi_1, .., \psi_N$.

3. DEEMPHASIS FACTOR

Using the theory of matched subspace detectors (MSDs) [16] and adaptive subspace detectors (ASDs) [17] the value of $\gamma(\psi) \in [0, 1]$ is adaptively generated based on the degree of confidence that a target with SNR greater than some prescribed value is present in the test cell. In the adaptive case, the unknown interference covariance matrix $\mathbf{R}_n$ is estimated by its sample covariance matrix $\mathbf{S}$ in (3) and it has been shown that when an unknown scaling σ exists between the interference-plus-noise in the primary data and that in the secondary, maximising the likelihood functions yields ASDs which are sample matrix versions of the corresponding MSDs [17].

As the interference may have heterogeneous statistical properties over the radar sampling grid, it is considered beneficial to use secondary sample vectors $\mathbf{n}_k$ located physically "close" to the test cell(s) on the basis that nearby data is statistically homogeneous or at least more nearly so. A rank-reduction linear transform $\mathbf{T} \in C^{N\times L}$ of full rank L may be incorporated for adaptive subspace detection to improve convergence rate (at the expense of reduced degrees of freedom) when the number of local secondary samples K is limited to improve the quality of interference homogeneity. Following this lead, the data $\mathbf{x}$ and signal model $\mathbf{H}(\psi)$ are transformed to a reduced-rank *quasi-whitened* space,

$$\mathbf{z} = \mathbf{\Phi}^{-1/2}\mathbf{T}^H\mathbf{x} \quad , \quad \mathbf{G}(\psi) = \mathbf{\Phi}^{-1/2}\mathbf{T}^H\mathbf{H}(\psi) \quad (5)$$

by defining the reduced-rank sample covariance matrix given by $\mathbf{\Phi} = \mathbf{T}^H\mathbf{S}\mathbf{T}$, its eigen-decomposition $\mathbf{\Phi} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H$ and the Hermitian square root $\mathbf{\Phi}^{1/2} = \mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{Q}^H$. It is then proposed to use the ratio of the energy in $\mathbf{z}$, per dimension, that lies in the resulting signal subspace (assumed to have full rank L) to the energy in $\mathbf{z}$, per dimension, that lies in the orthogonal subspace to indicate the "likelihood" of target signal presence within the test cell;

$$F(\psi) = \frac{\mathbf{z}^H\mathbf{P}_\mathbf{G}(\psi)\mathbf{z}/M}{\mathbf{z}^H\{\mathbf{I} - \mathbf{P}_\mathbf{G}(\psi)\}\mathbf{z}/(L-M)} \quad (6)$$

where $\mathbf{P}_\mathbf{G}(\psi) = \mathbf{G}(\psi)[\mathbf{G}^H(\psi)\mathbf{G}(\psi)]^{-1}\mathbf{G}(\psi)$ is the projector onto the transformed target signal subspace model. As the number of statistically homogeneous samples K increases $\mathbf{\Phi} \rightarrow \mathbf{T}^H\mathbf{R}_n\mathbf{T}$ the ratio in (6) tends to be F-distributed $F(\psi) \rightarrow F_{(\alpha,\beta)}[\nu]$ with $\alpha = 2M$ and $\beta = 2(L-M)$ degrees of freedom and non-centrality parameter $\nu = (\mu^2/\sigma^2)\tilde{\mathbf{s}}^H\tilde{\mathbf{R}}_n^{-1}\tilde{\mathbf{s}}$ where $\tilde{\mathbf{s}} = \mathbf{T}^H\mathbf{s}$ and

$\tilde{\mathbf{R}}_n = \mathbf{T}^H \mathbf{R}_n \mathbf{T}$ providing there is no signal mismatch, or stated mathematically $\mathbf{s} = \mathbf{H}(\psi)\mathbf{p}$. While this distribution is not strictly valid for the practical finite sample ASD, it tends to be approximately valid when sufficiently large and quasi-homogeneous training sets are used and strong enough target signals are considered. A soft decision approach which allows $\gamma(\psi)$ to vary continuously between 0 and 1 can be proposed as;

$$\gamma(\psi) = \int_{F(\psi)}^{\infty} p_F(f)\, df \quad , \quad F : F_{(\alpha,\beta)}[\nu] \tag{7}$$

where $p_F(f)$ is the probability density function of the appropriate non-central F-distribution for a design signal-to-noise ratio (SNR) ν. The selection of ν represents a tradeoff between higher sensitivity to weaker signals (lower ν) and enhanced localisation of stronger signals (higher ν). Although no optimality is claimed for this method, its performance will be experimentally demonstrated in the next section.

4. EXPERIMENTAL RESULTS

The data for this study were collected by 16 uniformly spaced narrowband receivers of the high frequency (3-30 MHz) Jindalee OTH radar, located near Alice Springs in central Australia. The linear aperture spanned by the 16 receivers is approximately 1.4 km with each receiver connected to a subarray composed of 28 dual-fan vertically polarised antenna elements, see [8] for further details regarding this facility. The data received in each subarray is range formed and Doppler processed using conventional (FFT-based) processing, in this section we are concerned with the adaptive filtering of space-time vectors constructed by stacking two array snapshots from consecutive range cells (the first being the test range cell) recorded at a particular Doppler bin in order to detect and azimuthally localise useful signals embedded in structured interference.

The data consists of a superposition of radio frequency interference (RFI) emitted by a source situated in the Darwin region (1250 km to the north of the receiver site and 22 degrees from the array boresight) at a carrier of 16.052 MHz and a relatively weaker coherent signal transmitted from a spatially separate source less than 2 minutes earlier at 16.106 MHz, both interference and signal were propagated by the ionosphere to the radar which was operated in *passive* mode (i.e. with transmitters switched off). The receive system processed a total of 42 ranges in each pulse repetition interval (PRI) with 128 PRI used for Doppler processing during a 4.2 second coherent processing interval (CPI). The coherent signal was localised at known ARD coordinates (namely, range cell 26, Doppler bin 33 and beam number 7) but could not be detected after conventional processing due to the presence of RFI which spread over the entire range-Doppler search space.

To perform STAP at test range cell 26, the secondary samples used to form the sample covariance matrix $\mathbf{S}$ were taken from all Doppler bins processed in range cells 24 and 28, which immediately neighbour a *guard cell* placed either side of the test cell in order to isolate the secondary data from range sidelobes of the coherent radar signal. Note that the number K of local training samples used to form $\mathbf{S}$ is $K = 2P = 8N$ where $P = 128$ Doppler bins and $N = 32$ (i.e 16 receivers and 2 stacked range cells). The target subspace model $\mathbf{H}(\psi)$ adopted was the wave interference model for 16 nominal steer directions corresponding to mutually orthogonal array steering vectors $\mathbf{s}(\psi_n)^H \mathbf{s}(\psi_{n'}) = \delta(n - n')$ and the target subspace dimension was chosen to be $M = 3$ with components $\Delta_1 = 0$ and $\Delta_2 = -\Delta_3$ equal to half the Rayleigh distance either side of ψ. No tapering of the matched filter, $\mathbf{v} = \mathbf{s}(\psi)$, was used to form the GSC-STAP output for $P = 128$ space-time vectors in the primary data matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_P]$ (one for each Doppler bin at the test range cell 26).

4.1. Traditional SMI filtering

The solid and dotted curves in Fig.1 relate to the left vertical axis and represent the Doppler spectra resulting in range cell 26 and beam number 7 when the standard SMI technique is used without the test cells ($\kappa = K$, $\mathbf{D} = \mathbf{0}$) and with the test cells ($\kappa = K$, $\mathbf{D} = \mathbf{I}$) included for training respectively (1). Evidently, the inclusion of the test cells (dotted line) is detrimental in the standard SMI approach as it causes target self-nulling of at Doppler bin 33. Note that a loss of 5-10 dB in target signal strength is apparent compared to the case where test cells are excluded (solid line).

The (+) and (∗) symbols in Fig.1 relate to the right vertical axis and represent the *normalised* beam spectra (i.e. noisefloor at 0 dB) resulting in range cell 26 and Doppler bin 33 when the test cells are excluded and included respectively. Although the maxima occurs in beam number 7 in both cases, the exclusion of the test cell (+) causes several false alarms (i.e. peaks in beamspace significantly above the noisefloor) due to the unregulated sidelobe response while the inclusion of the test cell (∗) results in a better sidelobe response but significantly degraded SINR (approximately 10 dB) at beam number 7 compared with the former due to target self-nulling. In a companion paper [18] that deals with spatial-only processing, it is shown that attempts to stabilise the sidelobes by diagonal loading [19] leads to intolerable degradations in interference cancellation performance and hence does not constitute a feasible solution for this problem.

4.2. Deemphasised SMI filtering

The solid line and (+) symbols in Fig.2 (same format as Fig.1) show the Doppler and normalised beam spectra corresponding to GSC-STAP filters adapted using the (ψ)-dependent deemphasis factors $\gamma(\psi)$ calculated according to (7) with a design SNR of ν=6 dB an $L = 16$ rank-reduction transform $\mathbf{T} = [\mathbf{I} \mid \mathbf{0}]^H$ which selects the test range cell (i.e. spatial-only processing) and a loading factor $\kappa = 1$. A constant design value of ν=6 dB was used for all azimuth-Doppler cells to avoid re-calculation of the cumulative non-central F-distribution.

The dotted curve and (∗) symbols in Fig.2 show the Doppler and normalised beam spectra resulting for a quiescent vector $\mathbf{v}(\psi) = \mathbf{t} \otimes \mathbf{s}(\psi)$ employing a cosine-pedestal taper $\mathbf{t}$ with *no* adaptive sidelobe cancellation. The adaptively deemphasised SMI output has a target SINR of almost 40 dB in beam number 7 which much larger than that of the standard SMI approaches considered in Fig.1, as well as having the added advantage of completely removing false alarms by effectively reducing all other peaks in beamspace to the noisefloor. A comparison between the adaptively deemphasised SMI output and the conventional one in Fig 3 demonstrates that adaptive sidelobe cancellation removes an extra 40 dB of the passively received interference that leaks through the sidelobes of the quiescent vector to detect the target which is otherwise not detected by $\mathbf{v}(\psi)$.

5. CONCLUSION AND FUTURE WORK

The effectiveness of the adaptively deemphasised GSC-STAP method proposed in this paper was demonstrated experimentally (with radar operated in passive mode) and shown to outperform traditional SMI methods both in terms of target output SINR and reduction of false alarms at moderately higher computational cost.

The derivation of the deemphasis factor is not claimed to be optimal and current research is being directed at the problem of formulating appropriate optimality criteria for this application.

6. REFERENCES

[1] Klemm, R: "Space-time adaptive processing - principles and applications" *IEE Publishers*, London, UK, 1998

[2] Ward, J: "Space-time adaptive processing for airborne radar" *Technical Report 1015*, Lincoln Laboratory, Massachusetts Institute of Technology, December 1994

[3] Reed, I S, Mallet and J D, Brennan, L E: "Rapid convergence rate in adaptive arrays", *IEEE Transactions on Aerospace and Electronic Systems*, Vol.10, No.6, November 1974, pp. 853-863

[4] Monzingo, R A and Miller, T W: "Introduction to adaptive arrays" *Wiley*, New York, 1980, 293-300.

[5] Pulsone, N B and Rader, M: "Adaptive beamformer orthogonal rejection test", *IEEE Transactions on Signal Processing*, Vol.49, No.3, March 2001, pp. 521-529

[6] Adve, R S, Hale, T B, and Wicks M C: "Practical joint domain localised adaptive processing in homogeneous and nonhomogeneous environments: Part1 and Part2", *IEE Proceedings - Radar, Sonar and Navigation*, Vol.147, No.2, April 2000, pp. 57-74

[7] Rabideau, D J and Steinhardt, A O: "Improved adaptive clutter cancellation through data-adaptive training" *IEEE Transactions on Aerospace and Electronic Systems*, Vol.35, No.3, July 1999, pp. 879-891

[8] Fabrizio, G A: "Space-time characterisation and adaptive processing of ionospherically-propagated HF signals" *Ph.D. dissertation*, Adelaide University, Australia, July 2000

[9] Ringelstein, J, Gershman, A B and Bohme, J F: "Sensor array processing for random inhomogeneous media" *Proceedings of the SPIE, Advanced Signal Processing: Algorithms, Architectures and Implementations*, IX 3807, Denver, Colorado, July 1999, pp. 267-276

[10] Fabrizio, G A, Abramovich, Y I, Gray, D A and Turley, M D: "Adaptive cancellation of nonstationary interference in HF antenna arrays", *IEE Proceedings- Radar, Sonar and Navigation*, Vol.145, No.1, February 1998, pp. 19-24

[11] Melvin, W L: "Space-time adaptive radar performance in heterogeneous clutter", *IEEE Transactions on Aerospace and Electronic Systems*, Vol.36, No.2, April 2000, pp. 621-633

[12] Guerci, J R, Goldstein, J S and Reed, I S: "Optimal and adaptive reduced-rank STAP", *IEEE Transactions on Aerospace and Electronic Systems*, Vol.36, No.2, April 2000, pp. 647-663

[13] Hughes, D T and McWhirter, J G: "Sidelobe control in adaptive beamforming using a penalty function", *International Symposium on Signal Processing and its Applications (ISSPA-96)*, Gold Coast, Australia, 25-30 August, 1996, pp. 200-203

[14] Griffiths, L J and Buckley, K M: "Quiescent pattern control in linearly constrained adaptive arrays", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol.35, No.7, July 1987, pp. 917-926

[15] Adve, R S, Hale, T B and Wicks, M C: "Transform domain localized processing using measured steering vectors and non-homogeneity detection", *Proceedings of the IEEE Radar Conference*, 1999, pp. 285-290

[16] Scharf, L L and Friedlander, B: "Matched subspace detectors", *IEEE Transactions on Signal Processing*, Vol.42, No.8, August 1994, pp. 2146-2157

[17] Kraut, S and Scharf, L L: "Adaptive subspace detectors", *IEEE Transactions on Signal Processing*, Vol.49, No.1, January 2001, pp. 1-16

[18] Fabrizio, G A and Turley, M D: "Adaptive spatial filtering with data-dependent training for improved radar signal detection and localisation in structured interference ", To appear in *Proceedings of the Defense Applications of Signal Processing (DASP) Workshop*, Adelaide, Australia 16-21 September, 2001,

[19] Carlson, B D: "Covariance matrix estimation errors and diagonal loading in adaptive arrays", *IEEE Transactions on Aerospace and Electronic Systems*, Vol.24, No.4, July 1988, pp. 397-401

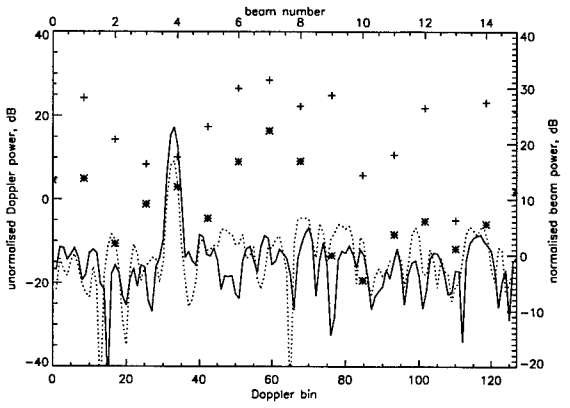


Fig. 1. Solid line and (+) symbol correspond to ($\kappa = K, \mathbf{D} = \mathbf{0}$), dotted line and (*) symbol correspond to ($\kappa = K, \mathbf{D} = \mathbf{I}$)

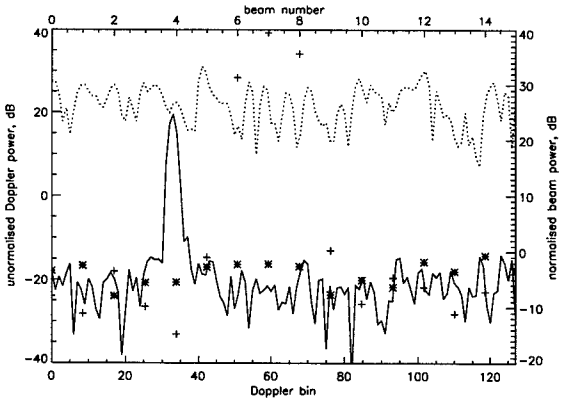


Fig. 2. Solid line and (+) symbol correspond to ($\kappa = 1, \mathbf{D}(\psi), \mathbf{T} = [\mathbf{I} \mid \mathbf{0}]^H, \nu = 6$ dB), dotted line and (*) symbol correspond to $\mathbf{v}(\psi) = \mathbf{t} \otimes \mathbf{s}(\psi)$

MONOBIT RECEIVER FOR ELECTRONIC WARFARE

Jesús Grajal, Raúl Blázquez, Gustavo López, José M. Sanz, Mateo Burgos

ETSIT, Universidad Politecnica Madrid
Ciudad Universitaria s/n, 28040 Madrid, Spain
email:jesus@gmr.ssr.upm.es

ABSTRACT

Two opposite requirements for digital broadband Electronic Warfare receivers are detection of simultaneous signals and real time operation. The monobit receiver represents an attempt to achieve both characteristics at the expense of low instantaneous dynamic range. This paper presents a detailed theoretical and experimental analysis and characterization of the performance of this promising receiver.

1. INTRODUCTION

The proliferation of electronic signals in modern combat environments requires the use of sophisticated Electronic Warfare (EW) receivers. Desirable characteristics of EW receivers include wide band frequency coverage, high sensitivity and dynamic range, high probability of intercept, simultaneous signal detection, frequency resolution, and full real-time operation. A classical receiver which accomplishes these requirements is a channelized receiver [1] which separates signals according to their frequencies.

Advancements in Analog-to-Digital Converters (ADC) technology and in the speed of digital processors have made possible to design relatively wide band digital channelized receivers. However, broadband digital channelized receivers, mainly based on Discrete Fourier Transform (DFT) related processing, are computation intensive and yet not suitable for real time applications in spite of the revolution of DSPs and FPGAs speed[1]. In an attempt to improve the real time operation, parallel processing can be considered. Another possibility is the reduction of the computational complexity of the signal processing algorithms by the simplification of the operations, e.g., avoiding complex multiplications in the calculations.

This is the philosophy in the monobit channelized receiver described in several US patents [2, 3] and papers [4]. As it was pointed out in [4], there are two possibilities in order to avoid multiplications in the calculation of the DFT: a single-bit digital representation of the input signal [2], which is equivalent to use a hard limiter, or a monobit representation of the kernel of the DFT [3]. Both schemes are possible, it is even possible to use both in the same processing algorithm.

The optimum scheme in terms of number of operations for the DFT is the Fast Fourier Transform (FFT). An FFT algorithm without multiplications is only possible with a monobit kernel.

This paper focuses on the theoretical and experimental evaluation of the performance, capabilities, and limitations of this receiver for the detection of multiple signals for radar applications.

[1]The aim of the proposed system is to cover hundreds of MHz with 100% real time.

2. DESCRIPTION OF THE SYSTEM

The system considered in this paper is depicted in figure 1. The radio-frequency front end is not included. The receiver uses a one bit ADC followed by a filter bank represented by a monobit DFT or FFT. Finally, a decision is made using the module of the different outputs of the DFT or FFT. In the following sections different important characteristics of the behaviour of this receiver will be explained.

This monobit receiver was implemented with a commercial module composed of an ADC SMT320 of 12 bits (we use only the sign bit) and a DSP from TI: TMS320C40. The experimental set-up is depicted in figure 2.

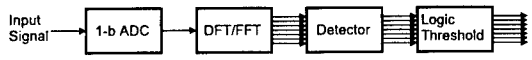


Fig. 1. Schematic diagram of the analysed monobit receiver.

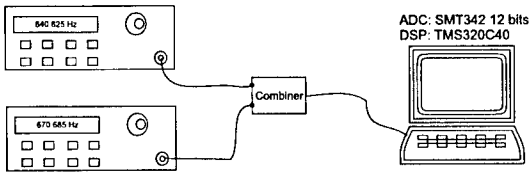


Fig. 2. Experimental set-up: Signal generators (noise and signals), combiner, ADC, and DSP.

3. MONOBIT DFT AND FFT

The concept of monobit DFT can lead to different implementations where the kernel function $e^{j\phi}$ is rounded to $1, -1, j$ or $-j$ through the function $G\left(e^{j\phi}\right)$. In this paper, the following function will be used [4]:

$$G\left(e^{j\phi}\right) = \begin{cases} 1 & \text{if } -\frac{\pi}{4} \leq \phi < \frac{\pi}{4} \\ j & \text{if } \frac{\pi}{4} \leq \phi < \frac{3\pi}{4} \\ -1 & \text{if } \frac{3\pi}{4} \leq \phi < \frac{5\pi}{4} \\ -j & \text{if } -\frac{3\pi}{4} \leq \phi < -\frac{\pi}{4} \end{cases} \quad (1)$$

Due to the use of equation (1) the property of the DFT that assured that if the input sequence was real-valued, the output se-

quence verified $X^*(k) = X(N-k)$ does not apply any more, as it is not always true that $G\left(e^{j\phi}\right) = G^*\left(e^{-j\phi}\right)$.

Three possibilities of implementing the monobit DFT have been considered, which lead to different filter banks:

- To replace directly the original kernel function with $G(\cdot)$. This implementation will be called monobit DFT. This is the slower implementation.
- To implement the DFT using the decimation in time algorithm, and replace the coefficients $e^{j\phi}$ by $G\left(e^{j\phi}\right)$. It will be called monobit FFT by decimation in time.
- To implement the DFT using the decimation in frequency algorithm, and replace the coefficients $e^{j\phi}$ by $G\left(e^{j\phi}\right)$. It will be called monobit FFT by decimation in frequency.

The use of the function $G(\cdot)$ modifies the coefficients of the filters in the three implementations, and so their frequency responses. However, the filters obtained from both monobit FFT implementations are nearly equal. As an example, the frequency response of the filter of channel 4 for a 32-point FFT and of its counterpart in a monobit FFT are compared in figure 3. A new sidelobe only 7.6 dB below the maximum in the monobit FFT appears. Windowing cannot improve this result.

It has been found that for the different implementations of the monobit DFT or FFT there are always two different kinds of channels. They are defined as:

- Type 1 channels: the kernel equals 1 or -1 for both the original and the monobit implementations. As the input signal is real valued, the output of this kind of channels is also real. The only channels of this kind are the channel 0 (DC component) and the channel $N/2$ (high frequency component). Quantization error due to the function $G(\cdot)$ is zero as the coefficients of the filters have not changed.
- Type 2 channels: the coefficients of the rest of the channels verify that half of them are real valued (1 or -1) and half of them are imaginary (j or $-j$). Channels $N/4$ and $3N/4$ verify also that their coefficients have not changed due to the use of $G(\cdot)$. The output of a type 2 channel is a complex number.

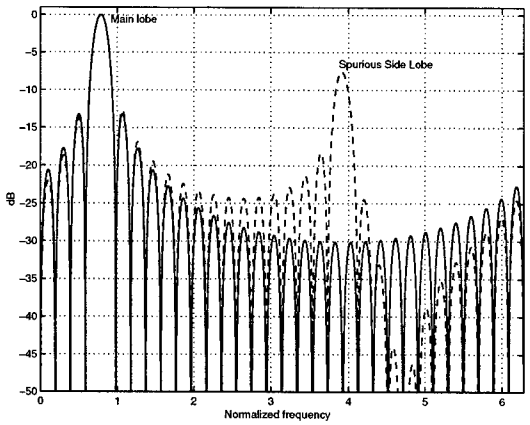


Fig. 3. Filter response for channel 4 from an original 32-point DFT and from a monobit 32-point FFT (- -).

4. FALSE ALARM PROBABILITY

The 1-bit ADC fixes the power at its output independently of the input power. Therefore, the false alarm probability is independent of the input power noise. As it was shown in the previous section, there are two different kind of channels in the receiver whose behaviour will be analysed separately.

In a type 1 channel, the filter coefficients are all equal to 1 or -1. Besides, the samples of the input signal are also 1 or -1. After multiplying them, a vector with k_+ elements equal to 1 and k_- elements equal to -1 is obtained. The sum of these elements renders $k = k_+ - k_-$, which is always an even number between $-N$ and N for a N-point DFT (FFT) with $N = 2^b$, b a natural number. Besides, $k_+ = \frac{N+k}{2}$, $k_- = \frac{N-k}{2}$.

After some reasoning based on combinatorial theory, we can calculate an analytical expression for the probability of getting an output k:

$$P(k) = \frac{\binom{N}{\frac{N-k}{2}}}{2^N} = \frac{\binom{N}{\frac{N+k}{2}}}{2^N} \tag{2}$$

After a linear detector at the output of each filter, we obtain an even value $k' = |k|$ with values $0 \leq k' \leq N$. The probability of getting a concrete k' is the sum of the probabilities of getting $k = k'$ and $k = -k'$ except for $k' = 0$. In short:

$$P(k') = \begin{cases} \binom{N}{\frac{N}{2}} \frac{1}{2^N} & \text{if } k' = 0 \\ \binom{N}{\frac{N-k'}{2}} \frac{2}{2^N} & \text{if } 0 < k' \leq N \end{cases} \tag{3}$$

If a threshold T_h is chosen at the output of the linear detector, the false alarm probability is obtained as:

$$P_{fa}(T_h) = \sum_{k' \geq T_h}^{N} P(k') \tag{4}$$

being P_{fa} a staircase function as it is clear in figure 4.

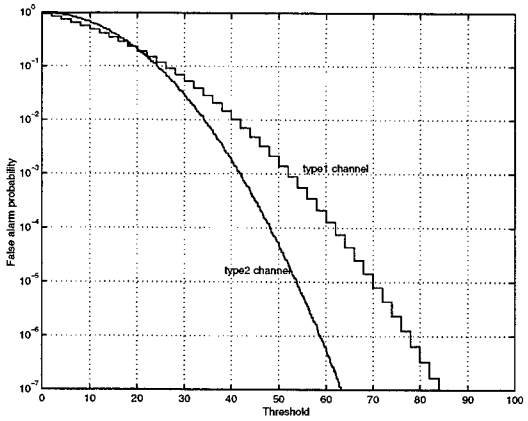


Fig. 4. False alarm probability for both type 1 and type2 channels and a 256-point DFT-FFT.

In a type 2 channel the output is a complex number. Its real and imaginary parts are both even numbers greater than $-N/2$

and lower than $N/2$. Following the same steps used for a type 1 channel, the probabilities of getting the real part of the complex number equal to k_r and its imaginary part equal to k_i are:

$$P_r(k_r) = \frac{\binom{\frac{N}{2}}{\frac{\frac{N}{2}-k_r}{2}}}{2^{\frac{N}{2}}} \qquad P_i(k_i) = \frac{\binom{\frac{N}{2}}{\frac{\frac{N}{2}-k_i}{2}}}{2^{\frac{N}{2}}} \tag{5}$$

Besides, if there is only white noise at the input of the system, k_i and k_r are independent and the probability of obtaining the complex number $k_r + j \cdot k_i$ is:

$$P(k_r + j \cdot k_i) = \binom{\frac{N}{2}}{\frac{\frac{N}{2}-k_r}{2}} \binom{\frac{N}{2}}{\frac{\frac{N}{2}-k_i}{2}} \frac{1}{2^N} \tag{6}$$

The output of these channels is converted to a real number using a linear detector. At its output, a threshold T_h is chosen, and the false alarm probability can be computed as:

$$P_{fa} = \sum_{|k_r+j\cdot k_i| \geq T_h} P(k_r + j \cdot k_i) \tag{7}$$

This mathematical expression can be applied to any channel of the three implementations as long as it is a type 2 channel.

As an example, in figure 4 the false alarm probability as a function of the threshold for a monobit 256-point DFT (FFT) is shown. This figure makes clear the following facts:

- Both functions are staircase functions for any of the three implementations.
- The number of different values of false alarm probability available is sensibly lower in a type 1 channel than in a type 2 channel.
- If a threshold $T_h > N$ for a type 1 channel or $T_h > N/\sqrt{2}$ for a type 2 channel is chosen, both the false alarm probability and detection probability are 0 because there are not any outcomes with that magnitude.

5. DETECTION PROBABILITY

A figure of merit of this system regarding detection capabilities can be its losses compared to a system without digitalization and with the original DFT for fixed detection and false alarm probabilities. Graphic 5 presents the average losses for centred sinusoidal signals, $P_d = 90\ \%$ and $P_{fa} = 10^{-3}$, and different lengths for the monobit DFT. Each point of the figure was calculated using Monte Carlo simulation with 5000 independent trials.

The detection capabilities can change depending on the implementation of the filter bank and the frequency of the sinusoidal signal to be detected.

All the channels of the monobit DFT have losses of nearly 1 dB with respect to the channels $i \cdot N/4$ $(i = 0, ..., 3)$ for sinusoidal signals centred in a channel. These filters are equal to the ones obtained with the original DFT. A monobit FFT have different losses depending on the selected filter. For example, maximum losses of 1.5 (2) dB appear for a 64 (256)-point monobit FFT.

Additional losses must be considered when the input signal is not at the centre frequency of the filter. The impact on the detection probability is not simply an increase in the signal power at the input of the system to compensate the attenuation of the filter. The reason is the non-linearity of the 1-bit ADC, which fixes the energy

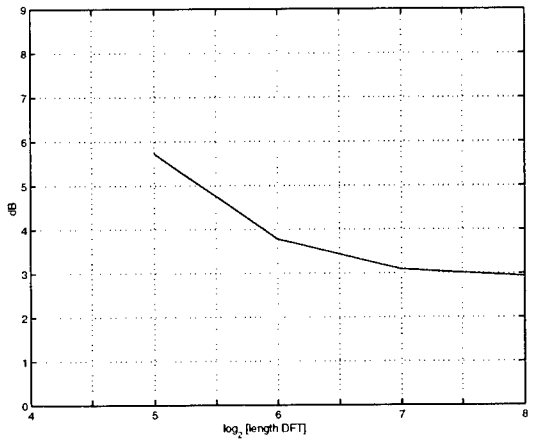


Fig. 5. Average losses for a monobit N-point DFT $P_{fa} = 10^{-3}$ and $P_d = 0.9$ for centred sinusoidal signals.

at its output independently of the input energy. As a consequence, a signal between to adjacent filters may not be detected for a fixed threshold because the energy is spread between these filters.

The simulated detection probabilities for sinusoidal signals with random frequencies within the bandwidth of each filter are shown in figures 6 and 7 for a $P_{fa} = 10^{-6}$ for $N = 64$ and $N = 128$, respectively. The dispersion of the traces is originated by the different amplitude response of the filters. The most important result of these figures is the impossibility to obtain simultaneously a mean detection probability per channel of 90 % for a fixed $P_{fa} = 10^{-6}$ by employing a monobit 64-point DFT-FFT. However, a monobit 128-point DFT-FFT can overcome this drawback.

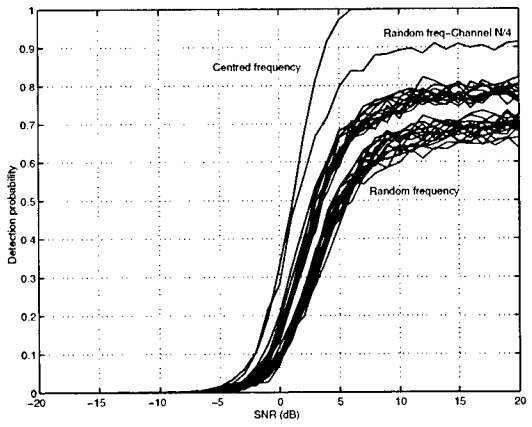


Fig. 6. Detection probability for sinusoidal signals with random frequencies in the band of each channel. Each trace represents a channel. 64-point *monobit* FFT and $P_{fa} = 10^{-6}$. These curves are compared to the average response of sinusoidal signals centred in different channels (*Centred frequency*).

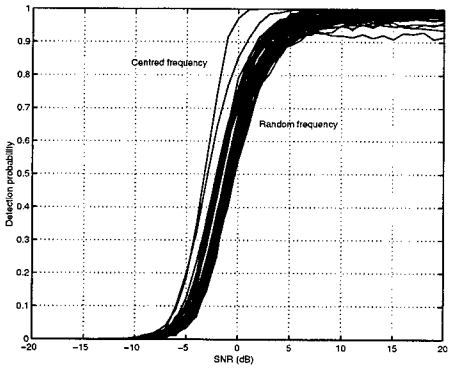


Fig. 7. Detection probability for sinusoidal signals with random frequencies in the band of each channel. Each trace represents a channel. 128-point *monobit* FFT and $P_{fa} = 10^{-6}$. These curves are compared to the average response of sinusoidal signals centred in different channels.

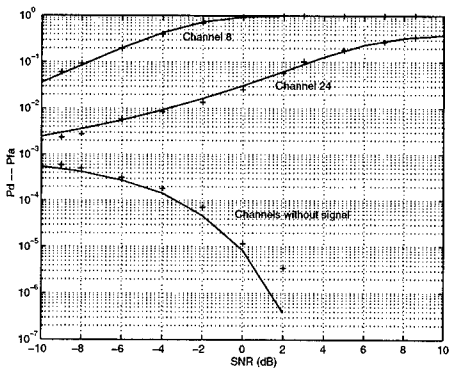


Fig. 8. Simulated and experimental results (+) for the detection probability with the input signal centred in channel 8. Signal in channel 24 is spurious. 64-point *monobit* FFT using decimation in frequency, and $P_{fa} = 10^{-3}$.

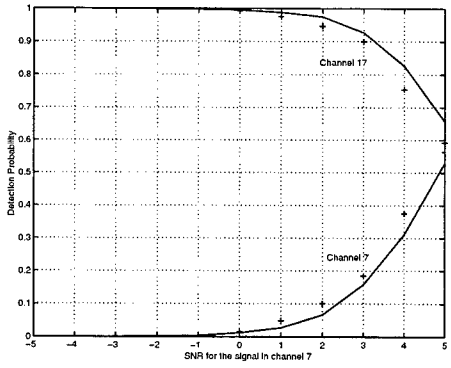


Fig. 9. Simulated and experimental results (+) for the instantaneous dynamic range. SNR for the signal in channel 17 is 5 dB and the power for the signal in channel 7 varies. 64-point *monobit* FFT using decimation in frequency. Threshold for $P_{fa} = 10^{-3}$.

6. DYNAMIC RANGE

One of the most important requirements for a channelized receiver is the detection of simultaneous signals. Due to the high non-linearity of the 1 bit ADC, there will be capture effect [5] and reduction of the instantaneous dynamic range (the ability to process concurrent signals of different amplitude) [6]. Therefore, a detailed study of the dynamic range with a single signal and with two simultaneous signals (with the same power or with different power) has been performed. The main conclusions can be summed up briefly.

Spurious generated by the non-linearity of the *ADC* can be detected with a significant detection probability, see figure 8, although it is possible to predict the channels where they appear once the original frequency is detected and to avoid them by blanking. Symbols (+) represents results obtained with the experimental set-up.

The probability of false alarm for the channels where neither the original frequency nor the spurious appear decreases with the power of the input signal due to the capture effect, figure 8.

For two signals with the same power, the detection probabilities for both signals decrease compared to the detection probability of one signal with the same power. This effect can be explained again by having in mind that power at the output of the ADC is constant and is spread among different channels.

The instantaneous dynamic range is less than 5 dB for a 64-point monobit FFT-DFT [2]. This result varies slightly with the position of the signals in the filter bank. Figure 9 shows a dynamic range of 3 dB for sinusoidal signals in channels 7 and 17. If the length of the FFT increases the instantaneous dynamic range also increases.

7. CONCLUSIONS

A monobit channelized receiver is analysed both theoretically and experimentally. Simplifications in the operations of the FFT allow to improve real time operation. However, one-bit digitalization and simplications in FFT result in losses and a rather low instantaneous dynamic range regarding the multibit approach.

[2] We have defined the instantaneous dynamic range when Pd=10 % for the weaker signal.

8. REFERENCES

[1] J. Tsui, *Microwave Receivers with Electronic Warfare Applications*. John Wiley & Sons, 1986.

[2] "US Patent us5793323,"

[3] "US Patent us5963164,"

[4] D.S. Pok, C.H. Chen, J.J. Schamus, C.T. Montgomery, J.B.Y. Tsui, "Chip Design for Monobit Receiver," *IEEE Transactions on Microwave Theory and Techniques*, vol. 45, pp. 2283–2295, Dec. 1997.

[5] S. Maas, *Nonlinear Microwave Circuits*. Artech House, 1988.

[6] J. Tsui, *Digital Techniques for Wideband Receivers*. Artech House, 1995.

Acknowledgements: This work was supported by CIDA (Centro de Investigación y Desarrollo de la Armada) and by Proyect TIC1999-1172-C02-01/02 of the National Board of Scientific and Technological Research (CICYT).

NEURAL NET BASED VARIABLE STRUCTURE MULTIPLE MODEL REDUCING MODE SET JUMP DELAY

*Daebum Choi** *Byungha Ahn** *Hanseok Ko***

*Dept. of Mechatronics, Kwangju Institute of Science and Technology
1 Oryong-dong, Puk-gu, Kwangju, 500-712, Korea
oner@moon.kjist.ac.kr
** Dept. of Electronics Engineering, Korea University
Anam-dong, Sungbuk-ku, Seoul,136-701 Korea

ABSTRACT

Variable structure multiple model (VSMM) is one of the most powerful algorithms for effectively tracking single maneuvering target. Although VSMM is developed specifically to improve the interactive multiple model (IMM) method focused to reducing computational cost and improving tracking performance, it presents an inherent limitation in the form of the presence of mode set jump delay (MJD). In this paper, MJD as an undesirable phenomenon in VSMM is described and analyzed. In order to eliminate the MJD, a neural network based VSMM that automatically selects the optimal mode set as achieved by supervised training is proposed. Through representative simulations we show the proposed algorithm outperforming over the conventional digraph switching VSMM in terms of tracking error.

1. INTRODUCTION

Target tracking is the process that estimates the state of moving object based on contaminated measurements. Kalman filter has been the most popular tool for tracking a moving object whose dynamics varies slowly. However, as the target changes their dynamics rapidly, the estimation error increases. To overcome the error during maneuvering, various research efforts in the past have merged essentially to the following four approaches: (1) single filter reactive adaptation, (2) variable dimension filtering, (3) cascaded filtering, and (4) multiple model (MM) filtering [1]. Among them, the MM method is known to be most promising [1]. MM method can be divided into two classes: (1) fixed structure multiple model (FSMM) and (2) VSMM. The most representative FSMM method is IMM, proposed in 1980's. In IMM, the mode jump is modeled as the Markov process and the input is statistically summarized or mixed from previous estimation in order to reduce the computational load exhibited by the GPB algorithm [2]. However, IMM has a drawback. As the number of models increases, the conflict among models causes increased estimation error. On the other hand, as the number of models is kept small, the tracking performance gets degraded since not all target movements including that of a maneuver can be adequately covered by a small mode set size [3]. X. Rong Li, et al. [3] proposed the VSMM algorithm to deal with this dilemma in a theoretical approach. However in VSMM, an important assumption is overlooked. VSMM presents a mode jump based on Markov process, which forms a moving pattern. A predetermined Markov transition matrix implies a pattern of mode jump. When a target maneuvers, maneuvering may lead the mode jump to one, which is unfamiliar to the predetermined one in Markov transition matrix. Due to the fixed jump pattern, wrong mode is selected, which in turn causes an estimation error. As search proceeds, the tracker eventually finds the correct mode but with additional scans. We call the additional scan time as 'mode jump delay (MJD)'.

In this paper, the MJD problem is first described and analyzed. We then propose an algorithm that reduces the effect of MJD in Section 3. Through representative simulations in Section 4, we show that the proposed algorithm outperforms that of either Digraph Switching VSMM (DSVSMM) or α–β filter.

2. VSMM AND ITS LIMIT

2.1 A Simple VSMM

For mode set matched filtering, discrete dynamic and measurement equations are given as

$$x_{k+1}(M_{k+1}(i)) = F_{k,k+1}x_k(M_k(j)) + v_k(M_{k+1}(i)) \quad (1)$$

$$z_k(M_k(j)) = H_k(M_k(j))x_k(M_k(j)) + w_k(M_k(j)) \quad (2)$$

where x_k is a state vector, z_k is a measurement, v_k is a process noise vector, w_k is a measurement noise vector, H_k is a measurement sensitive matrix, M_k is the *i-th* mode set in N mode sets at scan k, and $F_{k,k+1}$ is a state transition matrix from scan k to $k+1$.

An assumption in VSMM is that the mode set transition is modeled based on the Markov process, whose transition matrix T is given by:

$$T = \{t_{i,j}\}, for\ i,j = 1,...,N \quad (3)$$

$$t_{i,j} = P\{M_{k+1}(j) \mid M_k(i)\} \quad (4)$$

where the predetermined t_{ij} is the probability that mode set i transfers to mode set j after one scan.

Based on Markov process, the admissible mode set [3] is given by:

$$M_{k+1} = \{M \mid \exists x_{k+1}, P\{M_{k+1} \mid M_k, x_{k+1}\} > 0\} \quad (5)$$

where M is a mode set that is an element of total mode sets. Based on those two mode set jump assumptions, mode set matched estimation at scan $k+1$ is given as

$$\hat{x}_{k+1,k+1} = \sum_{j=1}^{N} P\{M_{k+1}(j)\}\hat{x}_{k+1,k+1}(M_{k+1}(j)) \quad (6)$$

where $\hat{x}$ is the overall state estimate, $P\{M\}$ is mode set probability, and $\hat{x}$ is the mode set matched estimate. In VSMM how to calculate $P\{M\}$ is the key.

2.2 MJD in VSMM

Before discussing MJD phenomenon, consider a problem in VSMM. In previous section, mode jump is based on Markov process. To design VSMM, we should define Markov transition matrix. However we don't know the general rule to define Markov transition matrix in VSMM. And other problem due to Markov process assumption is MJD.

Mode set sequence probability until scan $k+1$ in VSMM [3,4] is given by:

$$P\{M^{k+1} \mid Z^{k+1}\} = \frac{1}{c} P\{z_{k+1} \mid M^{k+1}, Z^k\} P\{M_{k+1} \mid M^k, Z^k\} P\{M^k \mid Z^k\} \quad (7)$$

where c is normalization constant, Z^k is the sequence of measurements and M^k is the sequence of the mode set until scan k.

In Equation (7), the first term, $P\{z_{k+1} \mid M^{k+1}, Z^k\}$ is the likelihood of mode set sequence M^{k+1} given z_{k+1}. That is, first term is the updated information form measurement. The second term, $P\{M_{k+1} \mid M^k, Z^k\}$ is the mode transition probability that is predetermined using Markov process assumption in VSMM. The meaning of second term is a priori that describes how a mode set jumps to another mode set in each scan. In practice, a target does not move in one predetermined mode jump pattern. As a result, Equation (7) leads to an erroneous mode set probability calculation due to the second term. The mode set for state estimation is selected among admissible mode sets that satisfy Equation (5). Wrong mode set probability calculation of each mode set causes incorrect selection of the admissible mode set. However as scans increase, the estimation mode set approaches the true mode set if the effect of the first term in Equation (7) becomes more dominant than that of the second term. This is the main essence of the MJD problem.

3. NEURAL NET BASED VSMM

3.1 One way of reducing MJD

To reduce the MJD effect, we develop a new mode set selection method instead of Equation (7). One simple way is that the tracker remembers possible mode set jump patterns.

In order to describe a jump pattern, three kinds of information are needed.

- Previous mode set information (PMI).
- Current measurement information (CMI).
- Current target's mode set for estimation.

PMI is the summary of the past information such as state estimation mode set at previous scan. CMI is the updated information from the current measurement such as measurement at current scan. CTI is the information of target at current scan, such as current mode set.

Before a measurement is obtained, PMI is available and after we have a measurement, CMI is obtained. The last term of the right side of Equation (7) implies PMI and the first term corresponds to CMI and left side of Equation (7) is similar to CTI. Therefore the mode selector should find CTI from PMI and CMI. The second term of left side of Equation (7) is not known. Assume a function f that maps {PMI, CMI} to {CTI} is known, and then we can determine the current CTI of a target using f. This mapping function is one solution that reduces MJD effect. However, since a target's movement depends on non-deterministic factors like that of pilot's will, situation around the target, or some commands to the target, f is not exactly deterministic.

3.2 Neural Net Based VSMM (NNVSMM)

How can the mapping function f be found? The function f is very complex and not expressible in a formula. In this case heuristic method is very useful. One of heuristic methods, neural network can be the suboptimal solution. Since backpropagation(BP) neural network can treat nonlinear mapping systems [5], we can establish the mode set selection logic using BP net. A simple mode set selector is shown in Figure 1.

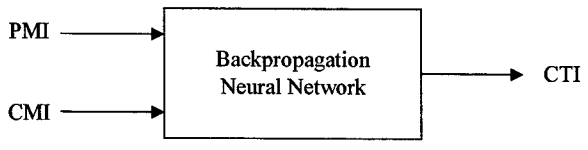


Figure 1 Mode set selection

Moreover, neural net is an associative memory [4], which can infer the most probable pattern for the untrained. In other words, for untrained pattern the neural net finds the most likely pattern in its memory. As a result, neural net mode set selection logic does not require more information for the unknown moving patterns and can treat the unknown by mixing trained patterns.

3.3 Implementation of a Simple NNVSMM

Let's consider 2-D single target tracking without clutter. A simple neural net mode set selector is shown in Figure 2. A mode is determined based on process noise variance, that is, mode parameter is process noise variance. As a result, PMI and CTI are process noise variance. Likelihood of a mode is used for CMI. In this case neural net compute measurement noise variance using previous measurement noise, and likelihood values.

For neural-net-training, independent multiple Kalman filters with different measurement noise variances are required. From a train scenario, we can obtain true position, state estimate and likelihood values of each filter, and the filter whose estimation error is the smallest. In this case training pair is constructed by

- Input

a. Measurement noise var. used in previous estimation

b. x-directional likelihood of the Kalman filter whose estimation error is smallest

c. y-directional likelihood of the Kalman filter whose estimation error is smallest

- Output

: Measurement noise var. for current estimation

This is one of the simplest implementations of neural net based mode selector. In this paper we propose NNVSMM, which exchange Markov process based mode set selection for neural net based.

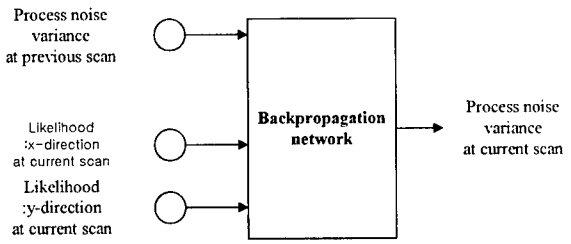


Figure 2 Realization of a simple neural net based mode selector

4. SIMULATION AND RESULT

In simulation the proposed algorithm, NNVSMM is compared to DSVSMM [3] and α-β filter [2]. Simulation parameters and environments are as follows.

- 2-D second order linear Kalman filter.
- Measurement noise covariance of each sensor - 12m for each coordinate.
- Varying process noise covariance determines the mode. Mode set is constructed based on maneuvering index.
- Range of process noise covariance: (1, 200) and spacing between neighbor modes: 5.
- 200 Monte-Carlo runs.
- Simulation Tool – Matlab 5.3.

Details on NNVSMM are given by:

- One hidden layer with 16 neurons for training scenario.
- For training data generation, 40 by 40 Kalman filters are running in parallel.
- Simulation scenario is given in Figure 3.

Test scenario is given in Figure 4.

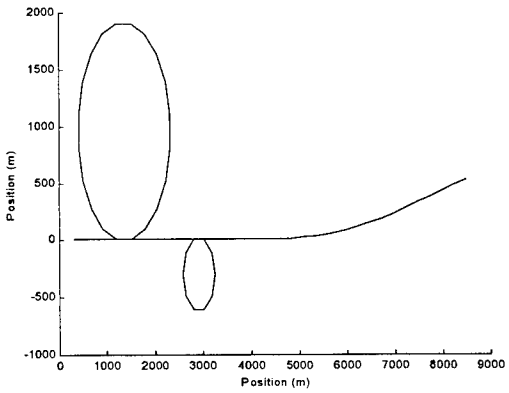


Figure 3 Training Scenario

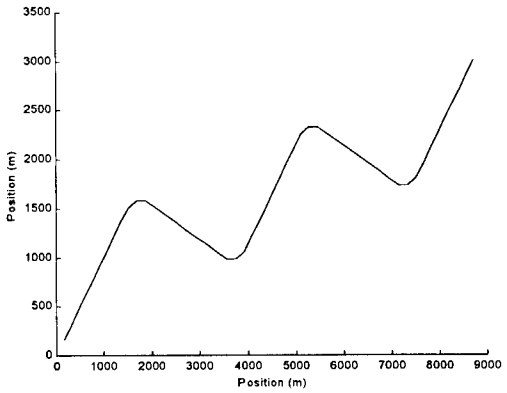


Figure 4 Test Scenario

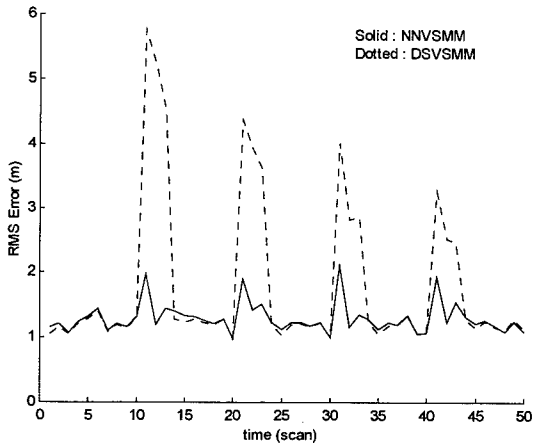


Figure 5 RMS errors of NNVSMM and DSVSMM

Scan Maneuvering periods	NN-VSMM	DS-VSMM	α-β filter
11-13	1.5791	4.6258	31.0747
21-23	1.6138	3.7418	31.1780
31-33	1.5083	2.9099	30.9601
41-43	1.5944	2.5968	31.1208

Table 1 Comparison of RMS errors in maneuvering periods

Figure 5 and Table 1 illustrate the result of simulation. In Figure 5 during maneuvering periods average error of NNVSMM is smaller than that of DSVSMM. Results indicate that the proposed NN-VSMM reduces the RMS error by decreasing the MJD phenomena.

Moreover the RMS error of NNVSMM maintains a level – around 1.6, but in DSVSMM RMS error varies from 2.6 to 4.6. As a result NNVSMM has a stable performance compare with DSVSMM.

Another strong point of NNVSMM is that it can track the movement that is not trained. In training scenario, three patterns are included: (1) Constant velocity, (2) Circular movement, and (3) Constant small acceleration. In test scenario, big acceleration changes happen. RMS error of DSVSMM during maneuvering is almost 2.6 to 4.6 times than during non-maneuvering but in NNVSMM only up to 1.6 times. This indicates that Markov process assumption in VSMM cannot cover all types of target motions and VSMM should be redesigned whenever the moving pattern is changed. However, NNVSMM can interpolate the unknown moving patterns using the training patterns. From this using neural network we can design VSMM without any information on mode transfer probability, which is not known in general. Therefore if there is a training set, which implies representative moving types, then we can design NNVSMM more easily than Markov based VSMM.

5. SUMMARY

In this paper, the MJD problem in VSMM is described. As a solution to the MJD problem, a new mode selection method based on neural net is developed and presented. Based on a neural net mode selector, NNVSMM that reduces MJD effect on VSMM is proposed. Through representative simulations, RMS error of NNVSMM is shown less than those of DSVSMM and α-β filter. For untrained moving patterns, NNVSMM is also shown achieving better performance over that of DSVSMM. Another promising feature about using the neural network based VSMM is that it can be designed without the prior information on the mode transition matrix derived from Markov process, which is not known in general.

6. REFERENCES

[1] S. Blackman and R. Popoli, *Design and Analysis of Mordern Tracking Systems*, Artech House, Boston, 1999..

[2] Bar-Shalom, Y. and Li, X. R., *Estimation and Tracking: Principles, Techniques, and Software*, Artech House, Boston, 1993.

[3] X. Rong Li and Bar-Shalom, Y., "Multiple model estimation with variable structure", *IEEE Transactions on Automatic Control*, 41(4): 478-493, 1996.

[4] Li, X. R.: "Hybrid Estimation Techniques." *In C. T. Leondes (Ed.), Control and Dynamic Systems: Advances in Theory and Applications*, Vol. 76. New York: Academic Press, 1996

[5] Philip D. Wasserman, *Neural Computing: Theory and practice*, Van Nostrand Reinhold, New York, 1989

COMPARISON OF PCA AND ICA BASED CLUTTER REDUCTION IN GPR SYSTEMS FOR ANTI- PERSONAL LANDMINE DETECTION

Brian Karlsen[1], *Jan Larsen*[2], *Helge B.D. Sørensen*[1] *and Kaj B. Jakobsen*[1]

[1]Ørsted•DTU, Technical University of Denmark
Ørsteds Plads, Building 348, DK-2800 Kongens Lyngby, Denmark
Web: http://www.oersted.dtu.dk, Email: brk,hbs,kbj@oersted.dtu.dk

[2]Informatics and Mathematical Modelling, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kongens Lyngby, Denmark
Web: http://eivind.imm.dtu.dk, Email: jl@imm.dtu.dk

ABSTRACT

This paper presents statistical signal processing approaches for clutter reduction in Stepped-Frequency Ground Penetrating Radar (SF-GPR) data. In particular, we suggest clutter/signal separation techniques based on principal and independent component analysis (PCA/ICA). The approaches are successfully evaluated and compared on real SF-GPR time-series. Field-test data are acquired using a monostatic S-band rectangular waveguide antenna.

1. INTRODUCTION

The development of techniques for automated detection of anti-personal landmines from sensor signal measurements is a significant problem. This paper focuses on improving signal-to-clutter ratio for detection systems based on ground penetrating radar (GPR) measurements. Clutter is characterized as signal components which are not directly correlated with primary scattering from mine objects. This comprises: measurement noise, disturbances from the antenna, inhomogeneities in the soil, scattering from rough surfaces, ground vegetation induced scattering, and to some extend multiple reflections. A number of recent clutter reduction approaches suggested in the literature cover: likelihood ratio testing [2], parametric system identification [3, 12, 15, 17], wavelet packet decomposition [4, 7], subspace techniques [8, 11, 18, 19], and simple mean scan subtraction [6].

We focus on unsupervised statistical based techniques for clutter reduction; in particular attenuation of surface disturbances. In Section 2 our previous suggested principal component analysis approach is revisited. Section 3 introduces a novel approach based on independent component analysis. Finally, Section 4 provides a comparative study on real GPR field test measurements.

JL is supported by the Danish Research Councils through the THOR Center for Neuroinformatics. BK acknowledges the Siemens Foundation for financial support. We thank Ole Nymann enthusiastic and steady support of our work in humanitarian mine detection. Furthermore, Staffan Abrahamson is acknowledged for collaboration on field data acquisition and Thomas Kolenda for valuable discussions on ICA.

2. PRINCIPAL COMPONENT ANALYSIS CLUTTER REDUCTION

Principal component techniques have previous been applied to GPR data analysis in [19] for detection of mines on preprocessed data using cross track-depth scans. In [18] clutter was reduced by reconstructing from the most significant eigenvectors, and [8] used a generalized singular values decomposition for separating noise and signal spaces. In [11] we took a different unsupervised approach where characteristics of the source signals (principal components) and associated eigenimages are used to determine the subspace for reconstruction.

Let $x_{ij}(t)$ denote the signal received at location x $= (i-1)$ cm, y $= (j-1)$ cm, where $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$. Traditional clutter reduction [6] consists in subtracting the mean scan across the xy-plane, $\bar{x}_{ij}(t) = x_{ij}(t) - (IJ)^{-1} \sum_{i,j} x_{ij}(t)$. This procedure removes the common signal across the xy-plane, which is mainly believed to originate from the very strong air-to-ground reflection. The approach taken here is inspired by explorative analysis of functional neuroimages and multimedia data [9, 13]. Define the $P \times N$ signal matrix: $\boldsymbol{X} = \{X_{p,t}\}$, $X_{p,t} = \bar{x}_{i,j}(t)$, where the pixel index $p = i + (j-1) \cdot I \in [1; P]$, $P = I \cdot J$. $t \in [1; N]$ is the time index with N being the total number of time samples. Column t of the matrix then represent the xy-plane scan image at time t reshaped into a vector, and the signal matrix represents the sequence of xy-plane images along the time or z-direction. Usually $P \gg N$ (in present experiments: $P = 51^2 = 2601$ and $N = 50$). Since the rank of $\boldsymbol{X}$ is at most N, the SVD reads

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \sum_{i=1}^{N} \boldsymbol{u}_i D_{i,i} \boldsymbol{v}_i^\top, \; X_{p,t} = \sum_{i=1}^{N} U_{p,i} D_{i,i} V_{t,i} \quad (1)$$

where the $P \times N$ matrix $\boldsymbol{U} = \{U_{p,i}\} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_N]$ and the $N \times N$ matrix $\boldsymbol{V} = \{V_{t,i}\} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N]$ represent the orthonormal basis vectors, i.e., eigenvectors of the symmetric matrices $\boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{X}^\top\boldsymbol{X}$, respectively. $\boldsymbol{D} = \{D_{i,i}\}$ is an $N \times N$ diagonal matrix of singular values ranked in decreasing order, as shown by $D_{i-1,i-1} \geq D_{i,i}$, $\forall i \in [2; N]$. The SVD identifies a set of uncorrelated time sequences, the principal components (PC's): $\boldsymbol{y}_i = D_{i,i}\boldsymbol{v}_i$, enumerated by the component index $i = 1, 2, \ldots, N$ and $\boldsymbol{y}_i = [y_i(1), \cdots, y_i(N)]^\top$. That is, we can

write the observed signal matrix (image sequence) as a weighted sum of fixed eigenvectors (eigenimages) $\boldsymbol{u}_i$ that often lend themselves to direct interpretation: some will contain mostly clutter, whereas others mainly mine reflections.

Consider the projection onto the subspace spanned by M selected PC's which mainly contain information about the mine object, i.e., $\boldsymbol{Y} = \widetilde{\boldsymbol{U}}^\top \boldsymbol{X}$, $\widetilde{\boldsymbol{U}} = [\boldsymbol{u}_{i_1}, \boldsymbol{u}_{i_2}, \cdots, \boldsymbol{u}_{i_M}]$, where $\boldsymbol{Y}$ is an $M \times N$ matrix. The selection can be done by inspecting the structure of the eigenimage or by the time course of $y_i(t)$. Ideally, if $y_i(t) = \delta(t - t_0)$ is a delta function, the structure of the eigenimage can be attributed to time t_0. The clutter is subsequently reduced by reconstructing $\boldsymbol{X}$ from the subspace, as given by $\widehat{\boldsymbol{X}} = \widetilde{\boldsymbol{U}}\boldsymbol{Y}$.

3. INDEPENDENT COMPONENT ANALYSIS CLUTTER REDUCTION

The spirit of the suggested method for independent component analysis (ICA) clutter reduction resembles that of the principal component based technique. The major difference is that the subspace formed by ICA is not orthogonal as in PCA. Moreover, the independent components (IC's), which are the counterparts of the PC's, are statistically independent. We thus expect the IC's to have a more distinct time localization.

Suppose that $\boldsymbol{X}$ first is projected to a subspace spanned by eigenvectors of non-zero eigenvalues, as we can not model from the null space [13]. Typically the dimension, d, of the signal subspace will be somewhat smaller than N. Let $\boldsymbol{U}$ be the $P \times d$ matrix of eigenvectors, and $\widetilde{\boldsymbol{X}} = \boldsymbol{U}^\top \boldsymbol{X}$ the projected signal matrix. The ICA problem is defined as: $\widetilde{\boldsymbol{X}} = \boldsymbol{A}\boldsymbol{S}$ where $\boldsymbol{A}$ is the $d \times M$, $M \leq d$, matrix of mixing coefficients and $\boldsymbol{S}$ is the $M \times N$ matrix of IC's – also referred to as source signals. That is, the original signal matrix is reconstructed as $\widehat{\boldsymbol{X}} = \boldsymbol{W}\boldsymbol{S} = \sum_{i=1}^{M} \boldsymbol{w}_i \boldsymbol{s}_i^\top$, where $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{A}$ is the matrix of eigenimages and $\boldsymbol{s}_i = [s_i(1), \cdots, s_i(N)]^\top$ is the i'th source signal. The literature provides a number of algorithms for estimating $\boldsymbol{A}$ and $\boldsymbol{S}$[1]. Basically they can be divided into two families in which the first deploy higher (or lower) order moments of non-Gaussian sources, whereas the other family uses the time correlation of the source signals. In the present case we expect that the sources are both non-Gaussian and colored. We deploy a member from each family: the widely used Bell-Sejnowski [1] algorithm using natural gradient learning, and the Molgedey-Schuster algorithm [9, 16]. They are both able to estimate $\boldsymbol{A}$ and $\boldsymbol{S}$ up to a scaling factors and permutations of the source signals.

4. EXPERIMENTS

A comparison of the PCA and ICA methods for clutter reduction in GPR signals were performed on field-test Stepped-Frequency GPR data. The field-test data are collected using a monostatic S-band waveguide antenna operating in the frequency range 2.65 – 3.95 GHz. The data were acquired using a HP8753C network analyzer. The bandwidth of the antenna determines the resolution which is approx. 11.5 cm. After antenna deembedding [11] the signals were down-mixed to the base band in order to remove the carrier [6]. The deployed sampling frequency is 5.12 GHz, which corresponds to a free-space sampling of 2.93 cm in the depth direction, which is below the resolution set by the antenna bandwidth.

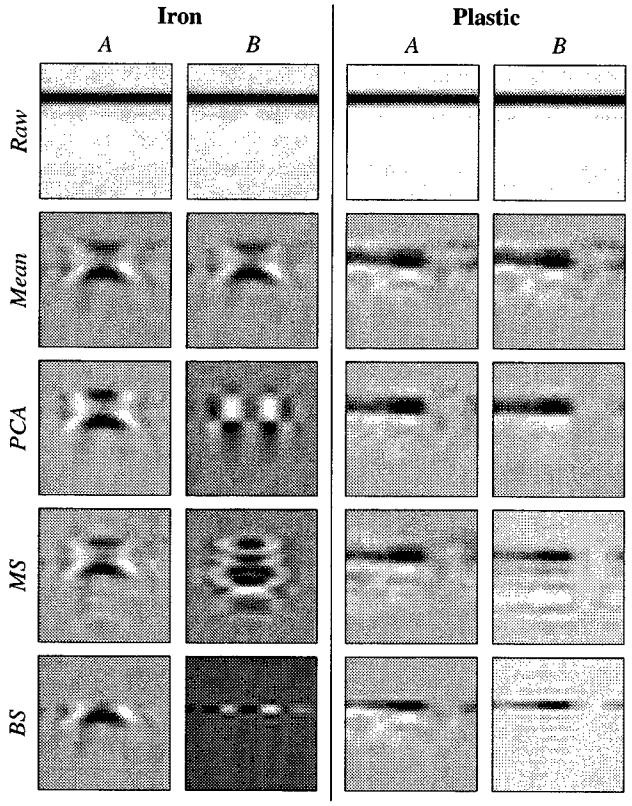


Fig. 1. Cross section (xt) images. The mine is located at the center in the x-direction and at $t = 16$ (2nd axis). The two left and right columns summarize results for iron and plastic mines, respectively. *A* columns correspond to reconstruction from components where only surface reflections are removed, and *B* to reconstruction from the strongest mine, see Figure 2. The rows are: *Raw* data, *Mean* subtraction method, *PCA*, Molgedey-Schuster ICA (*MS*), and Bell-Sejnowski ICA (*BS*). Raw data shows only air-to-ground reflection whereas *Mean* method helps somewhat in reducing the strong surface reflection. *PCA* seems to have a slight improvement over *Mean*, but *MS* does not provide much improvement, and further seems to enhance multiple reflections. *BS* on the other hand yields significant improvement, in particular when reconstructing from the strongest mine component only.

In a measurement area of 51 cm × 51 cm, M56 mine dummies[2] of iron and plastic (filled with bees wax) were buried in the center of the field in relatively dry sand 5 cm below the surface. The resulting signal matrices have $P = 51^2 = 2601$ and $N = 50$. The signal space dimension is $d = 22$ for the iron mine and 17 for the plastic mine. Using a smaller area resulted in signal matrices which have too low signal space dimension. When using the Bell-Sejnowski algorithm experiments show that appropriate learning rates are 10^{-4} and 10^{-3} for metal and plastic mine experiments, respectively. The lag value, τ, for the Molgedey-Schuster algo-

[1] For a resent review the reader is referred to [14].

[2] Dimensions are: diameter 5.4 cm, and height 4 cm.

rithm turned out to be quite sensitive, but $\tau = 1$ gave the best performance.

In Fig. 2 the eigenimages and associated PC's and IC's are depicted. ICA algorithms do not have any natural ordering. Since peak locations of the source signals determine the depth of scattering objects we choose to first rank according to peak locations occurring before the strong air-to-ground reflection at $t = 16$. Next, the components are ordered wrt. to variance contribution[3] in the reconstructed signal matrix [10], which for component i is $|\boldsymbol{w}_i|^2 \cdot \mathrm{Var}\{s_i(t)\}$.

The eigenimages of the iron mine experiments show nearly all very strong mine signatures, however, more clearly pronounced for the ICA algorithms. It should be noticed that the added contribution from more components can display surface like texture. For instance, the contributions from components 1 and 4 of PCA will add to a more blurred overall contribution. The source signals of PCA and Molgedey-Schuster do not possess good time localization[4], thus associated eigenimages cannot be attributed to a particular depth. This also makes the selection of components for reconstruction somewhat unclear. On the other hand, the Bell-Sejnowski algorithm produces very peaked source signals. E.g., component 5, which clearly peaks right after the surface reflection, also has a strong mine signature in its eigenimage. In addition, the width of the source peak is approximately 4 samples that corresponds to the resolution determined by the bandwidth of the antenna. Thus, source signals which have peak widths less than 4 samples do not make sense. The results for the plastic mine show that the mine signature is much less pronounced, i.e., signal-to-clutter ratio is low. Component 5 has a strong mine signature and is furthermore located at $t = 18$, which is at the mine location. Recall that the mine has an extension of approx. 5 cm which is half the resolution set by the antenna bandwidth. The reconstructed cross-section images are shown in Figure 1.

5. CONCLUSION

This paper provided a comparative study of PCA and ICA algorithms for clutter reduction. In particular the Bell-Sejnowski ICA showed significant improvement over PCA and Molgedey-Schuster ICA on real field GPR measurements. Future studies will focus on methods for automatic selection of subspace components and on convolutive ICA methods.

6. REFERENCES

[1] A. Bell & T.J. Sejnowski, "An Information-Maximation Approach to Blind Separation and Blind Deconvelution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[2] H. Brunzell: "Clutter Reduction and Object Detection in Surface Penetrating Radar," in *Proc. of IEE Radar'97*, issue 449, 1997, pp. 688–691.

[3] J.W. Brooks, L. van Kempen & H. Sahli: "Primary Study in Adaptive Clutter Reduction and Buried Minelike Target Enhancement from GPR Data," in *Proc. of SPIE, AeroSense 2000: Detect. and Rem. Techn. for Mines and Minelike Targets V*, vol. 4038, 2000, pp. 1183–1192.

[3] This measure is independent of the arbitrary scaling and permutation of the independent components.

[4] The Molgedey-Schuster algorithm most likely suffers from the fact that that the source signals are almost white.

[4] D. Carevic: "Clutter Reduction and Target Detection in Ground Penetrating Radar Data Using Wavelets," in *Proc. of SPIE Conference on Detect. and Rem. Techn. for Mines and Minelike Targets IV*, vol. 3710, 1999, pp. 973–97.

[5] P. Comon: "Independent Component Analysis: A New Concept," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[6] D.J. Daniels: *Surface Penetrating Radar*, IEE, 1996.

[7] H. Deng & H. Ling: "Clutter Reduction for Synthetic Aperture Radar Images Using Adaptive Wavelet Packet Transform," in *Proc. of IEEE Int. Antennas and Propagation Society Symposium*, vol. 3, 1999, pp. 1780–1783.

[8] A.H. Gynatilaka & B.A. Baertlein: "A subspace decomposition technique to improve GPR imaging of anti-personnel mines," in *Proc. of SPIE, AeroSense 2000: Detect. and Rem. Techn. for Mines and Minelike Targets V*, vol. 4038, 2000, pp. 1008–1018.

[9] L.K. Hansen, J. Larsen & T. Kolenda "On Independent Component Analysis for Multimedia Signals," in L. Guan, S.Y. Kung & J. Larsen (eds.) *Multimedia Image and Video Processing*, CRC Press, Ch. 7, pp. 175–199, 2000.

[10] L.K. Hansen, J. Larsen & T. Kolenda: "Blind Detection of Independent Dynamic Components," in *Proc. IEEE ICASSP'2001*, Salt Lake City, SAM-P8.10, vol. 5, 2001.

[11] B. Karlsen, J. Larsen, K.B. Jakobsen, H.B.D. Sørensen & S. Abrahamson: "Antenna Characteristics and Air-Ground Interface Deembedding Methods for Stepped-Frequency Ground Penetrating Radar Measurements," in *Proc. of SPIE, AeroSense 2000: Detect. and Rem. Techn. for Mines and Minelike Targets V*, vol. 4038, 2000, pp. 1420–1430.

[12] L. van Kempen, H. Sahli, E. Nyssen & J. Cornelis: "Signal Processing and Pattern Recognition Methods for Radar AP Mine Detection and Indentification," *Detection of Abandoned Land Mines*, no. 458, pp. 81–85, 1998.

[13] B. Lautrup, L.K. Hansen, I. Law, N. Mørch, C. Svarer & S.C. Strother: "Massive weight sharing: A Cure for Extremely Ill-posed Problems," in H.J. Herman *et al.*, (eds.) *Supercomputing in Brain Research: From Tomography to Neural Networks*, World Scientific Pub. Corp. 1995, pp. 137–148.

[14] T.W. Lee: *Independent Component Analysis: Theory and Applications* Kluwer Academic Publishers, ISBN: 0792382617, 1998.

[15] A. van der Merwe & I.J. Gupta:A Novel Signal Processing Technique for Clutter Reduction in GPR Measurements of Small, Shallow Land Mines," *IEEE Transactions on Geoscience and Remote Sensing* vol. 38, no. 6 , Nov. 2000, pp. 2627–2637.

[16] L. Molgedey & H. Schuster, "Separation of Independent Signals using Time-Delayed Correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.

[17] J.L. Salvati, C.C. Chen & J.T. Johnson: "Theoretical Study of a Surface Clutter Reduction Algorithm," in *Proc. of 1998 IEEE International Geoscience and Remote Sensing*, vol. 3, 1998, pp. 1460–1462.

[18] A.K. Shaw & V. Bhatnagar: "Automatic Target Recognition Using Eigen-Templates," in *Proc. of SPIE Conference on Algorithms for Synthetic Aperture Radar Imagery V*, vol. 3370, 1998, pp. 448-459.

[19] S.H. Yu & T.R. Witten: "Automatic Mine Detection based on Ground Penetratinig Radar," in *Proc. of SPIE Conference on Detect. and Rem. Techn. for Mines and Minelike Targets IV*, vol. 3710, 1999, pp. 961–972.

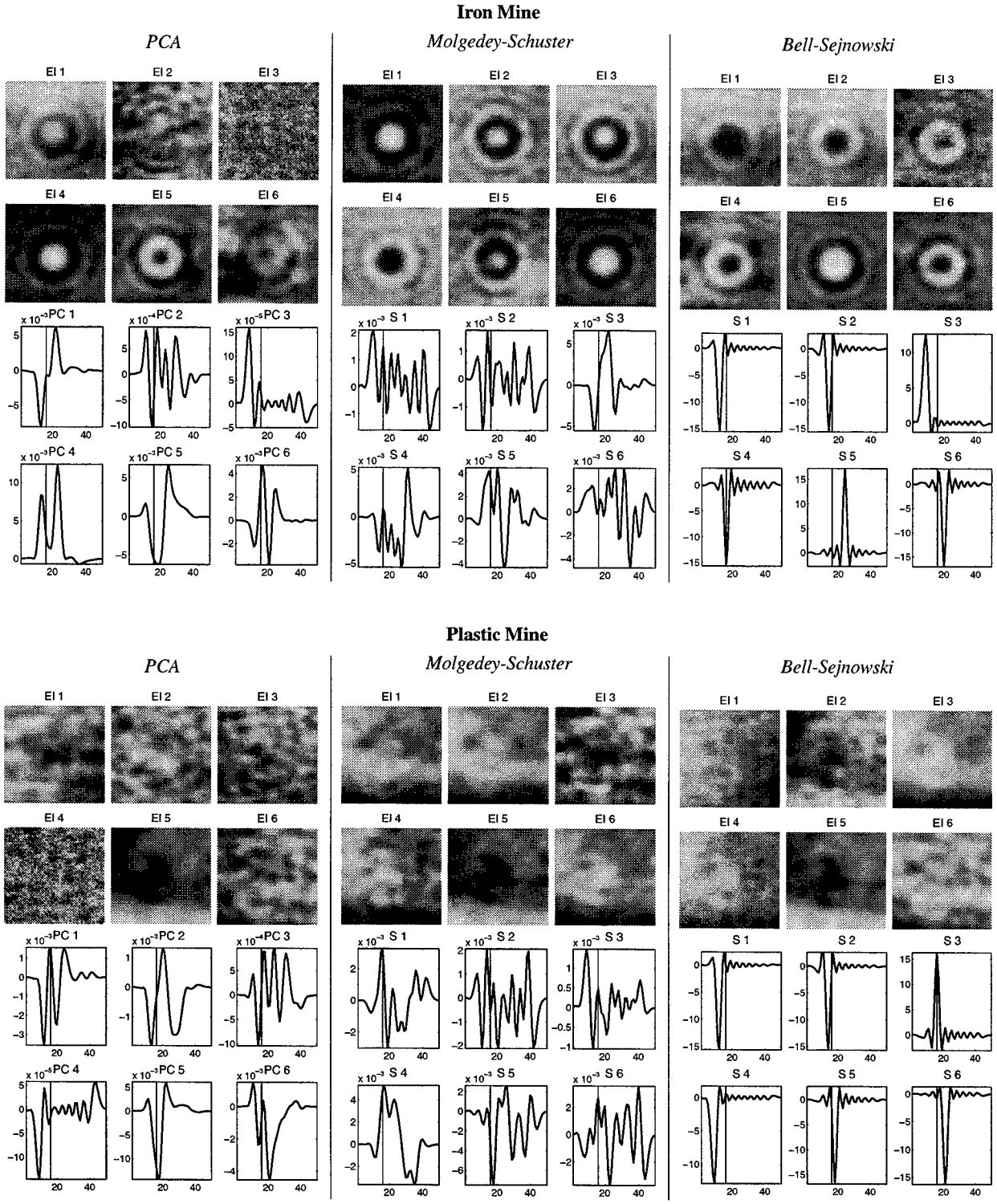


Fig. 2. Eigenimages and associated source signal, i.e., PC's or IC's. The vertical lines in the source signal pictures indicate the time corresponding to the position of the ground surface. Note that only the first 6 components are shown; the remaining source signals peak at later times and have smaller variance contributions.

ELIMINATION OF LEAKAGE AND GROUND-BOUNCE EFFECTS IN GROUND-PENETRATING RADAR DATA

R. Abrahamsson[a†], E. G. Larsson[a†], J. Li[a], J. Habersat[b], G. Maksymonko[b] and M. Bradley[c]

[a] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL.
[b] U.S. Army Night Vision and Electronic Sensor Directorate, Fort Belvoir, VA.
[c] Planning Systems Inc., Slidell, LA.

ABSTRACT

We address the problem of removing specular ground surface reflections and leakage/cross-talk from downward looking stepped frequency ground-penetrating radar (GPR) data. A new model for the ground-bounce and the leakage/cross-talk is introduced. An algorithm that jointly estimates these effects from collected data is presented. The algorithm has the sound foundation of a non-linear least squares (LS) fit to the presented model. The minimization is performed in a cyclic manner where one step is a linear LS minimization and the other step is a non-linear LS minimization where the optimum can efficiently be found using, e.g., the chirp-transform algorithm. The results after applying the algorithm to measured GPR data, collected at a U.S. army test range, are also shown.

1. INTRODUCTION

During the last decades, the enormous problems of buried landmines have raised the interest for subsurface exploration using ground-penetrating radar (GPR)[1]. Radar is an attractive type of sensor since it has the potential to detect anything with *electromagnetic contrast* (a term comprising permittivity and/or conductivity and/or permeability) to the surrounding medium as opposed to, e.g., metal detectors that can only detect objects with sufficiently high metal content. For an overview of some emerging technologies for mine detection, see, e.g., Chapter 2 of [2].

For a downward looking GPR system, which has the geometry that transfers the most radiated power to the subsurface, the collected data is severely contaminated by specular reflections from the ground surface. These reflections, usually referred to as the ground-bounce, can greatly surpass and hide the weak return of a shallowly buried plastic mine. The ground-bounce effect can be eliminated by positioning the antennas in direct contact with the soil so that no ground-bounce is allowed to form. For obvious reasons, this is not a feasible solution for the mine detection application of GPR. To be able to detect a possible mine in a reliable manner, we instead have to estimate and remove the ground-bounce from the data without impairing the return of the mine significantly.

† On leave from the Dept. of Systems and Control, Uppsala University, Uppsala, Sweden.

This work was partly supported by the U.S. Army under contract no. DAAB15-00-C-1024, the Swedish Foundation for Strategic Research (SSF) and SaabTech Electronics.

Please address all correspondance to Dr. Jian Li, Dept. of Electrical and Computer Engineering, University of Florida, P.O. Box 116130, Gainesville, FL 32611. Email: `li@dsp.ufl.edu`.

Another source of data degradation is leakage/crosstalk between the antennas and reflections from different parts of the platform itself; effects that in general cannot be eliminated completely by judicious system design. In contrast to the reflections from the ground surface, which may be subject to significant variations in radar cross section (RCS) and time of arrival, the leakage/cross-talk effects are mostly constant with antenna position. Nevertheless, these effects make the estimation of the ground-bounce difficult. It has been proposed that these leakage effects could be estimated by pointing the platform up in the sky and make a measurement without ground or subsurface reflections present. This measurement is then subtracted from the data prior to the ground bounce removal. However, since the antennas operate in their near-field range, the antenna gain patterns are affected by objects and material present within the antenna beams. For this reason the leakage/crosstalk and the platform reflections measured during the *sky-shot* can be quite different from those measured when the antennas are placed closely above the ground as in a normal operation mode.

In this paper we introduce a new model that takes both the leakage/cross-talk and the ground-bounce into account. Based on this model, we present a new algorithm named *DILBERT* (an acronym for *D*ecoupled *I*terative *L*eakage and ground-*B*ounce *E*stimation and *R*emoval *T*echnique) that jointly estimates the leakage and the ground-bounce.

2. DATA MODEL

For a stepped frequency radar system, a sequence of sinusoids with frequencies $\omega_k, k = 0, \ldots, K-1$, are transmitted at each antenna position $n = 0, \ldots, N-1$. For each ω_k and n the amplitude and phase of the return signal (as compared to the transmitted signal) are recorded as a complex value $x_{n,k}$. The propagation time, τ_n, associated with a specific scattering object then appears as the slope of the component of the recorded phase that is linear in ω_k . Hence, the range of the scattering object at a specific antenna position can be found using a Fourier transform w.r.t. ω_k. The scattering object will then show up in the range image as the Fourier transform of its complex RCS with respect to frequency, centered at the time $t = \tau_n$.

The model we propose for the collected data contaminated by leakage and ground-bounce can be written as

$$x_{n,k} = y_{n,k} + c_k + \alpha_n b_k e^{j\omega_k \tau_n}, \qquad (1)$$

where $y_{n,k}$ is the contribution from a possible mine, c_k comprises the unknown leakage/cross-talk and platform reflections that are

assumed to be constant with platform position, b_k is the frequency response of an unknown reference ground-bounce profile, α_n allows for variations in the complex valued ground surface RCS with platform position and τ_n accounts for a time shift of the reference ground-bounce profile and hence models surface height variations.

We note that the last term in (1) is ambiguous by a complex constant. However, since we are interested in the last two terms as a unit (for later subtraction from the data) rather than in the different parameters themselves, this is not necessarily a problem. The nuisance parameters $\{c_k, b_k, \alpha_n, \tau_n\}_{k=0,n=0}^{K-1,N-1}$ are only present to exploit the structure of the problem and even if a complex constant may move between the different parameters, the least squares fit that is presented in the following section will still be good. To guarantee that neither b_k nor α_n grows infinitely large with the other one vanishing and numerical problems as a result, one of these parameters needs to be fixed. In our case we choose $b_0 = 1$.

The model above is an extension of a frequency domain equivalent of the model proposed in [3] for an impulse based GPR system. The model in [3], however, did not take the leakage/cross-talk term into account and it also assumed that b_k was known.

3. ALGORITHM

With the assumption that the antenna main-lobe is narrow or that the mine response is significantly weaker than the ground surface reflection, a LS design criterion for a method based on our model is

$$\left\{\{\hat{c}_k\}_{k=0}^{K-1}, \left\{\hat{b}_k\right\}_{k=1}^{K-1}, \{\hat{\alpha}_n, \hat{\tau}_n\}_{n=0}^{N-1}\right\} = \\ = \arg\min_{\{c_k, b_k, \alpha_n, \tau_n\}} \sum_{k=0}^{K-1}\sum_{n=0}^{N-1} |x_{n,k} - c_k - \alpha_n b_k e^{j\omega_k \tau_n}|^2, \quad (2)$$

which is non-linear in the parameters to be estimated and cannot be minimized in closed form.

Our approach is to split (or decouple) the non-linear LS (NLS) criterion in (2) into two more tractable minimization problems by fixing two different subsets of the parameters and then use a cyclic algorithm that alternates between solving the two new optimization problems.

3.1. Minimization with respect to $\{\alpha_n, \tau_n\}$

By assuming that c_k and b_k are known and fixed, the model in (1) can be reduced to the frequency domain equivalent of that proposed in [3] by subtracting the leakage term from the data, $\tilde{x}_{n,k} = x_{n,k} - c_k$. In [4] the solution to the resulting NLS minimization problem

$$\{\hat{\alpha}_n, \hat{\tau}_n\}_{n=0}^{N-1} = \arg\min_{\{\alpha_n, \tau_n\}} \sum_{k=0}^{K-1} |\tilde{x}_{n,k} - \alpha_n b_k e^{j\omega_k \tau_n}|^2, \quad (3)$$

is shown to be

$$\{\hat{\tau}_n\}_{n=0}^{N-1} = \arg\max_{\tau_n} \left| \sum_{k=0}^{K-1} b_k^* \tilde{x}_{n,k} e^{-j\omega_k \tau_n} \right| \quad (4)$$

$$\{\hat{\alpha}_n\}_{n=0}^{N-1} = \frac{\sum_{k=0}^{K-1} b_k^* \tilde{x}_{n,k} e^{-j\omega_k \hat{\tau}_n}}{\sum_{k=0}^{K-1} |b_k|^2}, \quad (5)$$

where $(\cdot)^*$ denotes the complex conjugate. In the case of a vector or matrix operand, $(\cdot)^*$ will also denote the conjugate transpose.

Equations (4) and (5) can be interpreted as a correlation between $x_{n,k}$ and b_k in the frequency domain where $\hat{\tau}_n$ is the lag for the maximum of the correlation function and $\hat{\alpha}_n$ is the complex amplitude at that maximum. Hence, the parameter estimates can be found using Fourier transforms. For a coarse estimation of τ_n a zero-padded FFT can be used. For an estimate with higher resolution a chirp-transform algorithm may be used locally around the coarse estimate to find the maximum with higher precision or, since local convexity is likely, some fast search method can be applied.

3.2. Minimization with respect to $\{b_k, c_k\}$

By assuming that α_n and τ_n are known, the NLS criterion in (2) reduces to the linear LS criterion

$$\left\{\{\hat{c}_k\}_{k=0}^{K-1}, \{\hat{b}_k\}_{k=1}^{K-1}\right\} = \arg\min_{\{c_k, b_k\}} \|\mathbf{x} - \mathbf{A}\mathbf{b}\|^2, \quad (6)$$

where

$$\mathbf{x} = [\mathbf{x}_0^T \ \mathbf{x}_1^T \cdots \mathbf{x}_{N-1}^T]^T \in \mathbb{C}^{NK\times 1}$$

$$\mathbf{x}_n = [x_{n,0} - \alpha_n e^{j\omega_0\tau_n} \ \ x_{n,1} \ \cdots \ x_{n,K-1}]^T \in \mathbb{C}^{K\times 1}$$

$$\mathbf{b} = [c_0 \ \cdots \ c_{K-1} \ \ b_1 \ \cdots \ b_{K-1}]^T \in \mathbb{C}^{(2K-1)\times 1}$$

$$\mathbf{A} = [\mathbf{A}_1 \ \ \mathbf{A}_2] \in \mathbb{C}^{NK\times(2K-1)}$$

$$\mathbf{A}_1 = [\mathbf{I}_K \ \cdots \ \mathbf{I}_K]^T \in \mathbb{C}^{NK\times K}$$

$$\mathbf{A}_2 = [\alpha_0\mathbf{E}_0^T \ \cdots \ \alpha_{N-1}\mathbf{E}_{N-1}^T]^T \in \mathbb{C}^{NK\times(K-1)}$$

$$\mathbf{E}_n = \begin{bmatrix} 0 & \cdots & 0 \\ e^{j\omega_1\tau_n} & 0 & \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & e^{j\omega_{K-1}\tau_n} \end{bmatrix} \in \mathbb{C}^{K\times(K-1)},$$

where $\mathbf{I}_K$ is an identity matrix of size $K \times K$.

The solution to (6) can be found as (see, e.g, [5])

$$\hat{\mathbf{b}} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{x}. \quad (7)$$

The matrix $\mathbf{A}$ is built of diagonal blocks and zeros. This structure can be used to reduce the computational cost of finding the estimates. Writing (7) in terms of the blocks $\mathbf{A}_1$ and $\mathbf{A}_2$ gives us

$$\hat{\mathbf{b}} = \begin{bmatrix} \mathbf{A}_1^*\mathbf{A}_1 & \mathbf{A}_1^*\mathbf{A}_2 \\ \mathbf{A}_2^*\mathbf{A}_1 & \mathbf{A}_2^*\mathbf{A}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A}_1^* \\ \mathbf{A}_2^* \end{bmatrix} \mathbf{x} = \\ = \begin{bmatrix} N\mathbf{I}_K & \sum_{N=0}^{N-1} \alpha_n \mathbf{E}_n \\ \sum_{N=0}^{N-1} \alpha_n^* \mathbf{E}_n^* & \sum_{N=0}^{N-1} |\alpha_n|^2 \mathbf{I}_{K-1} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{N=0}^{N-1} \mathbf{x}_n \\ \sum_{N=0}^{N-1} \alpha_n^* \mathbf{E}_n^* \mathbf{x}_n \end{bmatrix}. \quad (8)$$

By applying the inversion on the partitioned matrix in (8) (see, e.g., [6] for the inversion of a partitioned matrix) we end up (after some straightforward algebra) with the estimates as

$$\hat{c}_0 = \frac{1}{N}\sum_{n=0}^{N-1} \left(x_{n,0} - \alpha_n e^{j\omega_0\tau_n}\right) \quad (9)$$

$$\{\hat{c}_k\}_{k=1}^{K-1} = \frac{\bar{x}_k \sum_{n=0}^{N-1} |\alpha_n|^2 - \breve{x}_k \sum_{n=0}^{N-1} \alpha_n e^{j\omega_k \tau_n}}{\sum_{n=0}^{N-1} |\alpha_n|^2 - \frac{1}{N} \left| \sum_{n=0}^{N-1} \alpha_n e^{j\omega_k \tau_n} \right|^2} \quad (10)$$

$$\{\hat{b}_k\}_{k=1}^{K-1} = \frac{\breve{x}_k N - \bar{x}_k \sum_{n=0}^{N-1} \alpha_n^* e^{-j\omega_k \tau_n}}{\sum_{n=0}^{N-1} |\alpha_n|^2 - \frac{1}{N} \left| \sum_{n=0}^{N-1} \alpha_n e^{j\omega_k \tau_n} \right|^2}, \quad (11)$$

where $\bar{x}_k$ is the mean of the data taken over platform positions,

$$\bar{x}_k = \frac{1}{N} \sum_{n=0}^{N-1} x_{n,k}, \quad (12)$$

and

$$\breve{x}_k = \frac{1}{N} \sum_{n=0}^{N-1} \alpha_n^* x_{n,k} e^{-j\omega_k \tau_n}. \quad (13)$$

3.3. Algorithm summary

The algorithm is initiated by assuming no leakage ($c_k = 0$) and the ground-bounce reference response as $b_k = x_{0,k}/x_{0,0}$. These values are used in (4) and (5) to obtain estimates of τ_n and α_n. These estimates are then plugged into (9), (10) and (11) to refine the estimates of c_k and b_k. The procedure is repeated until practical convergence is achieved (e.g., until the relative change in the cost function in (2) is less than some prespecified threshold, say 10^{-5}).

Since in each step we minimize (2) with respect to a subset of the parameters and (2) is bounded from below (the design criterion is positive), convergence is guaranteed. However, we cannot in general guarantee that the convergence is to the global minimum.

3.4. Reduced DILBERT

If we assume that no leakage/cross-talk is present ($c_k = 0$), (11) reduces to

$$\{\hat{b}_k\}_{k=1}^{K-1} = \frac{\breve{x}_n N}{\sum_{n=0}^{N-1} |\alpha_n|^2} \quad (14)$$

The estimates for τ_n and α_n are still found using (4) and (5) with the cosmetic modification that $\tilde{x}_{n,k} = x_{n,k}$. We will refer to this algorithm based on the "smaller" model as the *reduced* DILBERT since it does not account for the leakage term.

4. RESULTS

In this section we present images of data processed using the new algorithms as compared to images where no ground-bounce or leakage/cross-talk processing has been made and images where the along track mean has been removed (i.e., processing based on a constant ground-bounce model). We have used stepped frequency data with a frequency range from 0.5GHz to 4GHz. The data was provided by Planning Systems Incorporated and was recorded at a U.S. Army test range.

In the first example (Fig. 1) the measurements were made over a M19 plastic mine. To simulate larger ground surface variations, the antenna elevation was changed linearly as the antenna moved along the track. In the unprocessed image (Fig. 1(a)) we can see both the changing ground-bounce and the more stable leakage/cross-talk effects. In Fig. 1(b) the latter effects are satisfactorily removed by the mean subtraction while most of the ground bounce is left untouched. For the reduced DILBERT (Fig. 1(c)) more of the ground surface reflection has been removed. However, the algorithm is adversely affected by the presence of the leakage/cross-talk and cannot remove neither of the degrading effects satisfactorily. As expected, the DILBERT algorithm which takes both the leakage and the ground-bounce terms into account (Fig. 1(d)) performs much better than the reduced version.

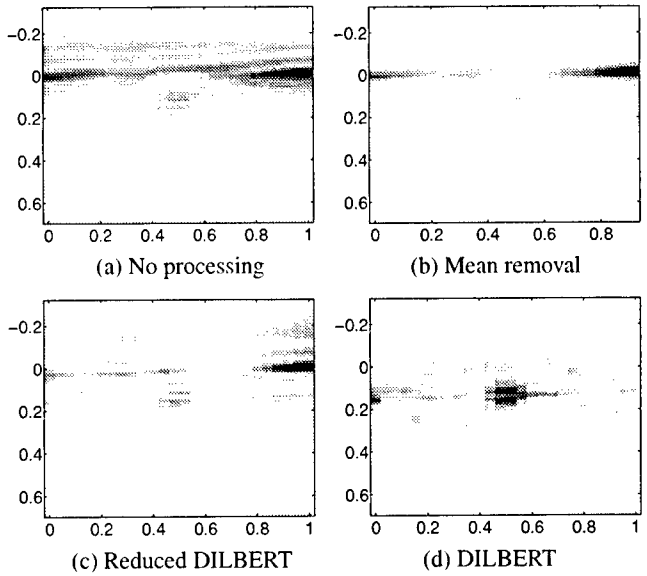


Fig. 1. Images of a M19 plastic mine buried at 5cm depth and along track position 0.5m.

In the next example (Fig. 2) the ground surface is more stationary with n. The roll-off observed at the along track edges is a result of the synthetic aperture radar (SAR) processing applied after the ground-bounce removal in the last two examples. Due to the more constant ground surface reflection the mean removal now works much better than in the first example, but still, large portions of the ground-bounce remains. Also the reduced version of our model is more accurate than in the previous example and consequently the reduced DILBERT algorithm performs acceptably. The full DILBERT removes even more of the unwanted effects. However, the extra degrees of freedom available also cause the top of the mine to be partly included in the estimate of the ground-bounce and is therefore reduced in intensity. The risk of including the mine in the ground-bounce estimate could be reduced by imposing a smoothness constraint for the time of arrival of the ground bounce. This can, e.g., be done by replacing the sub-algorithm in Section 3.1 by the procedure described in [7] where the discretized derivatives of τ_n are penalized. A more ad-hoc but computationally more appealing alternative would be to apply a window on (4) centered on $\hat{\tau}_{n-1}$ before finding the maximum to get $\hat{\tau}_n$.

In the last example (Fig. 3) we show results for data collected over a metal mine where the mine return is considerably stronger than for the plastic mines in the earlier examples. Again we see acceptable performance by both new algorithms. For DILBERT, Fig. 3(d), we even distinguish two dominant scatterers, which we

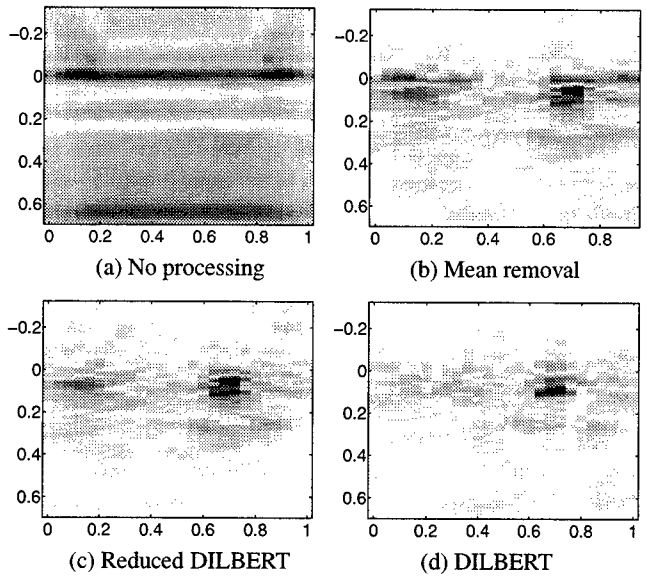


Fig. 2. Images of a VS16 plastic mine buried at 2.5cm depth and along track position 0.75m.

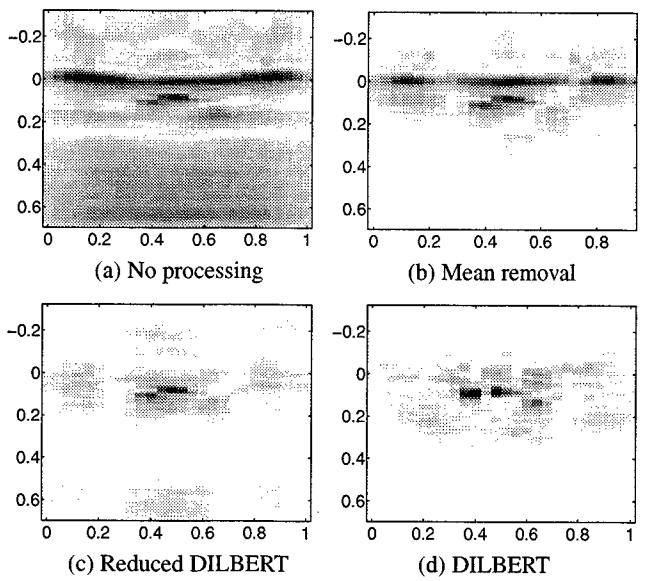


Fig. 3. Images of a M15 metal mine buried at 7.5cm depth and along track position 0.45m.

interpret as a part of the top structure of the mine.

The rate of convergence is very fast for the reduced DILBERT (typically less than 5 iterations, all of which are computationally very cheap). For the full version, the number of iterations needed is usually larger. This can be expected as there are many more parameters to estimate. Also, in the case of a ground surface that is nearly flat and with a surface RCS that is almost constant with antenna position, the problem of estimating both the ground-bounce and the leakage/cross-talk can become ill-conditioned or even ill-posed. However, during our data processing we have not encountered any numerical problems even with the datasets with the smoothest ground surfaces. Furthermore, even if the problem of estimating the parameters in (1) is ill-conditioned (or even singular, i.e., no *unique* solution exists), the result of applying the algorithm to remove the ground-bounce and leakage effects may still be satisfactory.

For a running version of the full DILBERT, the convergence speed should not be a problem. The algorithm can be allowed to converge by moving the platform slowly the first number of positions. Since the leakage and the ground-bounce profile are approximately constant, and hence previous values of $\hat{c}_k$ and $\hat{b}_k$ provide good initial estimates, a few iterations should be sufficient for convergence at the subsequent antenna positions.

5. CONCLUSIONS

We have introduced a new model for ground-bounce and leakage/cross-talk effects in stepped frequency ground-penetrating radar data. Based on the model, a novel least squares based cyclic algorithm (DILBERT) for removal of these effects has been presented together with a reduced version (reduced DILBERT) where the leakage/cross-talk is not taken into account. All the steps in the cyclic algorithms can be efficiently solved using FFTs and simple vector multiplications.

The results after the two algorithms have been applied to measured GPR data have been shown. In the case of a ground surface changing significantly in height, the full DILBERT outperforms the reduced version. When the ground surface is more constant, more of the ground bounce is eliminated also by the reduced DILBERT. The full DILBERT still removes more of the ground-bounce and leakage/cross-talk than the reduced version.

In our experience, despite the fact that it is not guaranteed that the global minimum is always achieved when solving (2), the algorithms provide good means for removing the ground-bounce and the leakage/cross-talk effects from ground-penetrating radar data.

6. REFERENCES

[1] D. J. Daniels, *Surface Penetrating Radar*, Institute of Electrical Engineers, 1996.

[2] H. Brunzell, "Detection of shallowly buried objects using impulse radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 2, pp. 875–886, 1999.

[3] H. Brunzell, *Signal Processing Techniques for Detection of Buried Landmines using Ground Penetrating Radar*, Ph.D. thesis, Chalmers University of Technology, Sweden, 1998.

[4] E. G. Larsson, R. Abrahamsson, J. Li, K. Gu, M. Bradley, J. Habersat, and G. Maksymonko, "Reducing the ground-bounce effects for mine detection with a ground-penetrating radar," in *Proceedings of the UXO/Countermine Forum*, New Orleans, LA, April 2001.

[5] P. Stoica and R. Moses, *Introduction to Spectral Analysis*, Prentice Hall, Upper Saddle River, NJ, 1997.

[6] R. A. Horn and C. A. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, England, 1985.

[7] E. G. Larsson, J. Li, M. Bradley, J. Habersat, and G. Maksymonko, "Removal of the ground-bounce effect in ground-penetrating radar data," in *Proceedings of SPIE*, Orlando, FL, 2001, vol. 4394.

TOWARDS REAL-TIME DETECTION OF LANDMINES IN FLIR IMAGERY

Mabo R. Ito, Sinh Duong
Department of Electrical Engineering, University of British Columbia
Vancouver, BC, Canada, V6T 1Z4

John E. McFee, Kevin L. Russell
Threat Detection Group, Defence Research Establishment Suffield
Box 4000, Medicine Hat, AB, Canada, T1A 8K6

ABSTRACT

A pipelined algorithm and parallel architecture is under development for real time detection of landmines. Our previous work has dealt with monochromatic images from airborne active infrared scanners and images from a low-altitude aircraft-mounted multi-spectral scanner. Because of the nature of the sensors and the aerial observation platform, the landmines were treated as small, sparse, discrete objects in a large clutter field. Our current work deals with passive infrared imagery obtained from cameras mounted on ground vehicle. In contrast to the previous work, although the targets are still relatively sparse, they are no longer small in the sense of occupying just a few pixels and the signal to noise ratio is considerably worse than in for the airborne active infrared and multi-spectral scanner problems. So significant changes to our detection algorithm are needed. The paper briefly describes the overall algorithm and the particular issues, such as irregular shapes, that need to be dealt with in FLIR imagery. Some early results are presented. In addition, changes in computer processing power and inter-processor communications has led to a rethink of the real-time hardware implementations of the system and these issues are discussed in the paper.

1. INTRODUCTION

For several years, the Canadian Defense Research Establishment in Suffield and the University of British Columbia have been jointly developing a pipe-lined algorithm and parallel architecture for real time detection of landmines in images. Minefields are a serious threat to land combat forces because they very effectively impede mobility and yield a high degree of vulnerability to stationary troops. Detecting minefields from distance, referred to as standoff or remote minefield detection, is a high priority among a number of nations including Canada. A typical minefield image contains a significant number of compact target objects that are sparsely distributed over a large area and may have particular spatial relationships to one another.

Since the research began on the minefield detection algorithm, it has become apparent that the algorithm might have applications other than the Airborne Active Infrared (AAIR) monochromatic imagery, which is obtained from a low-altitude aircraft-mounted multi-spectral scanner. For example, a high priority project within the Canadian Department of National Defense is the development of a vehicle mounted multiple mine detectors for use on roads. A passive infrared imager, which produces Forward Looking Infrared (FLIR) images, is one detector in this system. An automatic target recognition algorithm is being developed to assist the vehicle operator, and the Remote Minefield Detection Hierarchical algorithm, described in this paper, is a promising candidate.

This pipelined algorithm has been implemented on a network of transputers and tested using samples of AAIR imagery. Currently it has been adapted for use on FLIR images with success. Furthermore, the advances in computer processing power and inter-processor communications in recent years has led to a reconsideration of the system's hardware implementation.

This paper briefly presents the general system architecture of the Remote Minefield Detection Hierarchical. It then describes the significant changes to adapt this system to identify mine objects in FLIR imagery. And it finally outlines proposal real-time hardware architecture.

2. REMOTE MINEFIELD DETECTION HIERARCHICAL

2.1 Algorithm Structure: The general piped-lined algorithm for real-time detection of sparse small objects in images can be described using Figure 1. The major parts of the algorithm consists of the Low Level Target Cueing to reject non-suspect regions and thus drastically reduce the data rate, the Middle Level Target Shape Analysis to

0-7803-7011-2/01/$10.00 ©2001 IEEE

classify the suspect regions as target or non-target relying upon their morphological features, the High Level Target Spatial Analysis to extract the features of the spatial relationship between mine-like objects, and the Top Level Knowledge Integration to resolve whether or not the image contain mines using the spatial analysis results and external information resources.

Raw image data is acquired by a sensor and passed into the Image Correction (IC) stage. This IC module adjusts the raw image data to compensate for distortions, dropouts, overlapping swaths, misregistration, and other artifacts and imperfections due to the scanning process. After the IC stage, the data is passed to the Non-Suspect Region Rejection (NSRR) stage. The NSRR reduces the immense data flow down to a stream of small images (subimages) that are likely candidates to contain mines. To accomplish this task, the NSSR collects a block of scan lines of data image and then divides it into smaller non-overlapping square regions called subimages. Each subimage is further divided into smaller non-overlapping square regions, whose width is smaller than the dimensions of a small target but large enough to provide a stable average of the region. The contrast values of these small regions against the parent subimage are calculated and ranked. Only a few top values will be selected as suspect regions. Then the data of these suspect regions is transferred to the Local Region Thinning (LRT) module where redundant subimages are eliminated. The next stage, Local Region Segmentation (LRS), partitions the subimages into homogeneous regions. These regions are then passed to the Local Region Feature Extraction (LRFE) block. In this step, various morphological features such as pixel area, average pixel intensity and region compactness are measured for each candidate mine-like object. The next stage, Local Region Classification (LRC), categorizes the assembled feature vectors into different classes, to determine targets or non-targets with an estimated likelihood based on extracted features. The results of the classification are passed to the higher level, Target Spatial Analysis. Here, initially the relative positions of likely mines are analyzed and various spatial size and shape clusters are formed using a Clustering algorithm. Then, the clusters are delivered to Global Region Feature Extraction (GRFE) where both statistical measurements and pattern descriptors of each cluster, as well as spatial inter-relationship among clusters, are computed and extracted. Depending on the type of patterns encountered, either a statistical or a syntactic pattern classifier may be more appropriate. Scatterable minefields can be adequately characterized by a low-dimensional feature space and hence are sufficiently handled by statistical classification. While patterned minefields are more amenable to syntactic classification because of their recognizable arrangements. However, since the patterns are not known in advance, both classifiers are needed and operated in parallel. In the final level, Knowledge Integration, the expert system integrates the statistical and syntactical data output from GRFE with other sources from the user and the knowledge base to decide if the image likely contains a minefield.

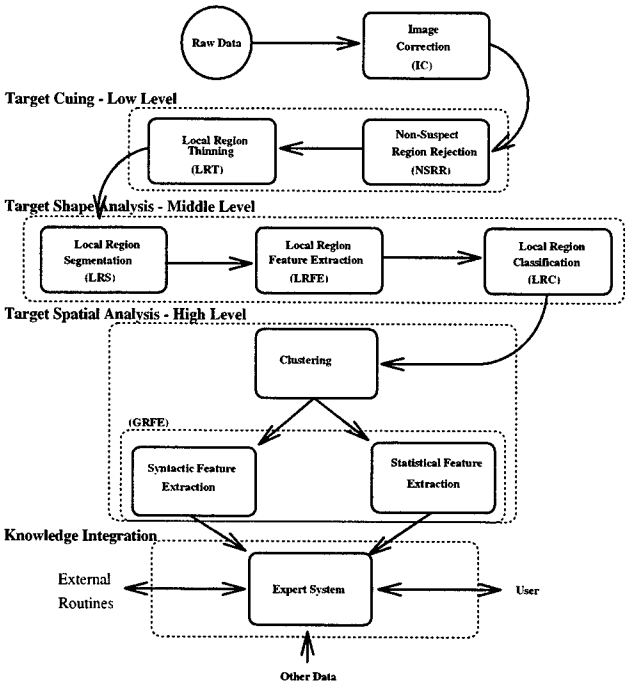


Figure 1: Remote Minefield Detection Hierarchical.

2.2 Algorithm Modification for FLIR Imagery: The system architecture described in Section 2.1 was initially designed for AAIR airborne imagery. Some modifications of algorithm and of hardware structure have been researched and developed to accommodate FLIR imagery.

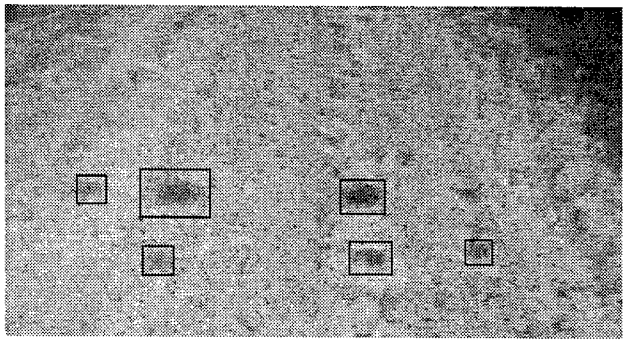


Figure 2: A FLIR image with visible mines.

FLIR images can record individual mines, but the whole minefield arrangement can not be seen. Therefore the Target Spatial Analysis component is not needed. However, the mine shapes are usually distorted because of

the aspect angle and the shallow depth of field of the camera. An additional software component may be required to restore the image perspective. However, to retain the hardware configuration and to achieve real-time speed, this problem might be corrected more efficiently by utilizing different thresholds and reference values for the mine and background feature vectors. Our study indicated that all of the 9 morphological feature components should be used in FLIR case.

Furthermore, because of the slow speed of the detection vehicle, which consequently produces a low input data rate, a simpler algorithm and hardware implementation for the Target Cueing level might be more appropriate. Also, the number of mines in FLIR image is small and the mine-like object sizes are rather large, implying that fewer computations are likely needed to identify suspect regions.

2.2.1 Non-Suspect Region Rejection (NSRR): A new averaging method was introduced to better distinguish the contrast between objects and background. This method also solved the problem that the upper region of a FLIR image is usually darker the lower part (Figure 2).

2.2.2 Local Region Thinning (LRT): Because of the high signal-to-noise ration in FLIR, more suspected objects were selected by NSRR. Thus the LRT was optimized to process the thinning function more quickly.

2.2.3 Local Region Segmentation (LRS): The targets in FLIR are quite large comparing to those in AAIR imagery. Thus shape analysis is the key to success. Currently, Region Split And Merge (RSAM) algorithm is used for segmentation due to it's speed and simpleness.

It is common that real FLIR images of mines have many tiny dots scattered around suspect objects, and these dots are usually much darker than the main targets. To save processing time for the segmenter, it is desired to clean up this clutter by using the mean of a scan window as the intensity threshold to filter out all objects that are darker.

Another enhancement is that LRS considers a target (or a background) as a homogenous region instead of a multi-region object. This helps to speed up the shape analysis and reduce the hardware requirements.

As shown in Figure 3, RSAM sometimes does not wholly separate the main target from its surrounding blots. The reason for this is RSAM was mainly designed for objects which have smaller object size/pixel size ratio, and objects should not connect to their adjacent pieces by narrow "bridges", which unfortunately is not the case in FLIR. To compensate this disadvantage, a Smooth function was utilized to blur out those small "bridges" as much as possible, and to produce a clean smooth surrounding edge. The degree of smoothing has to be tuned so that the segmented images do not lose their characteristic shapes and still satisfy the uniform surface requirement.



Figure 3: Scanned images of anti-personnel mines. Left: good image. Right: bad image with narrow bridges.

2.2.4 Local Region Feature Extraction (LRFE): Targets are identified by classification of patterns of morphological features extracted from the segmented regions. These features must be chosen according to the particular image type. For the minefield detection problem under discussion, it was decided to use well established morphological quantities to form the components of the feature vector of a region, since these have been extensively studied and are easily calculated.

Study of the FLIR feature vectors using On-line Pattern Analysis and Recognition System (OLPARS) software suggested the number of features to be used is 9. They are: the region area, intensity mean, intensity variance, maximum intensity, minimum intensity, 4-adjacency perimeter, 4-adjacency size, 4-adjacency compactness, and height/width ratio.

2.2.5 Local Region Classification (LRC): In this stage, a classifier will use a database of feature vectors of known mines to test against those of an unknown object. The classifier combines the feature values of an object into one or more values that will be compared with the database.

The Nearest Mean Classifier was used in the LRC. Using the pattern analysis tool OLPARS to analyze a number of feature vectors of known objects, it revealed that mine classes are clustered and well separated from background classes. Thus misclassification can ideally be minimized.

An important key for success is to design a good database that LRC will depend on to identify targets. A small database will leave many mime objects unrecognized, while an abundant database will cause the classifier to select many clutters or to be highly biased (over trained) on a certain set of test images.

3. HARDWARE IMPLEMENTATION

The current architecture, conceived in 1989, is based on a distributed network of transputer computing nodes,

connected to a workstation. The low and mid levels of the algorithm are implemented on the network, while the high and top levels are implemented on the Sparc workstation. . This system utilizes an array of vector processors i860s and transputers plus local memory (TRAMs), with varying topologies for each of the processing nodes. The network is highly scaleable, i.e., increasing the number of elements requires only minor changes of the programming code, and it utilizes a simple point-to-point serial communications protocol for interconnection.

Although the transputer was an attractive computing element, the technology is now over 15 years old. Despite of improvements have been made, in recent years the speed performance of an individual transputer has fallen badly behind other digital signal processor (DSP) boards, while the later have improved speed, ease of inter processor communication and price.

A study of an alternative hardware implementation of the Remote Minefield Detection system had concludes that the RMD system performs substantial image processing, but does not appear to fully utilize the enhanced abilities of a digital signal processor. Thus, it is not recommended that a digital signal processor such as the SHARC be utilized. Instead general-purpose processors such as the PowerPC, Pentium and Alpha should be considered. The generation of these processors should be selected such that those that have SIMD capabilities are utilized as appropriate. Besides being much faster at the image analysis than the SHARC, these general-purpose processors have the advantage of being much cheaper, as well as having much lower development costs. The SHARC is more optimized for execution of fast multiplies and accumulates, but this is not something that the RMD system does much of. NSRR and LRT are the two most computationally intensive functions that take a lot of hardware to run, while the remaining algorithms LRA, RSAM and LRFE scale linearly as the image size increases.

4. PRELIMINARY RESULTS

The algorithm has been developed up to the expert system level. The hardware configuration was implemented using a network of transputers. Non real-time studies using simulated, yet realistic, thermal infrared AAIR images and actual passive infrared FLIR imagery have demonstrated that the algorithm is successful in detecting individual targets.

In the recent tests, the algorithm up to and including the Local Region Classification was applied to a number of synthetic AAIR images, which consists of scattered minefields, patterned minefields and a combination of the two. The probability of detection of individual mines was estimated to be 90% and the probability of false alarm was 2%. For real FLIR images, the results are approximately 64.6% probability of detection and 25.6% probability of false alarm.

5. CONCLUSIONS

A pipelined algorithm and parallel architecture for real-time detection of mine objects in monochromatic AAIR imagery and passive infrared FLIR imagery has been described. The algorithm was implemented on a distributed computing system, and all the modules of the algorithm were tested on simulated and real passive infrared minefield images. Non real-time preliminary test results indicate the algorithm can reliably and consistently detect mines and minefields. Real-time operation should be achievable with a modestly sized, but special-purpose, parallel and pipelined computer system. Also modern computer architecture was recommended to upgrade the initially designed transputer hardware platform.

REFERENCES

[1] J. E. McFee, K. L. Russell, M. R. Ito, Detection of Surface-laid Minefields Using a Hierarchical Image Processing Algorithm, *Proceedings of SPIE Symposium, Volume 1567: Applications of Digital Image Processing XIV*, edited by A. G. Tescher, pages 42-52, 1991.

[2] J. E. McFee, K. L. Russell, Y. Das, R. C. Q. Vu and M. R. Ito, Analysis of Minefield Images Using a Transputer Network, *Proc. NATUG-6*, Vancouver, BC, pp. 99-114, May 1993.

[3] J. E. McFee, K. L. Russell, M. R. Ito, and R. C. Q. Vu, Cooperating Computer Architectures for Real-time Processing of Sparse Object Images, *in Proceedings of 3rd International Workshop on Parallel Image Analysis: Theory and Applications*, edited by A. Rosenfeld, pages 77-94, University of Maryland at College Park, MD, USA, 1994.

[4] S. Duong and M. R. Ito, Pipe-lined Algorithm and Parallel Architecture for Real-time Detection of Sparse Small Objects in Images, *SPIE Conference on Detection Technologies for Mines and Mine-like Targets*, Orlando, FL, USA, 1997.

[5] S. Duong and M. R. Ito, Extensions to Real-time Hierarchical Mine Detection Algorithm, Contract Report, Defense Research Establishment Suffield (Unclassified), 2001.

SIGNAL PROCESSING TECHNIQUES FOR CLUTTER PARAMETERS ESTIMATION AND CLUTTER REMOVAL IN GPR DATA FOR LANDMINE DETECTION

L. van Kempen, H. Sahli
Vrije Universiteit Brussel - Faculty of Applied Sciences
ETRO Dept. IRIS Research group
Pleinlaan 2, B-1050 Brussels - Belgium
E-mail:lmkempen,hsahli@etro.vub.ac.be

ABSTRACT

Ground Penetrating Radar (GPR) has become widely accepted as a major technique for subsurface investigations, mainly in civil engineering. Recently considerable efforts are put in the development of GPR systems for the detection of shallow buried landmines. However, GPR performs inadequally due to clutter, which dominates the data and obscures the mine information. The clutter varies with surface roughness and soil conditions and lead to uncertainty in the measurements. It is therefore necessary to overcome these surrounding effects when processing GPR data for detecting small, shallow buried objects.

In this paper we present improved signal processing techniques which can be used to reduce the clutter through data preprocessing. Several approaches are proposed for GPR clutter reduction techniques, most of them model the clutter statistically. The proposed clutter reduction technique models the clutter using parametric modeling. The clutter contained in the measurements is treated as an ARMA model.

The advantage of such approach lies therein that once the clutter is satisfactorily known, any target will show up as a small anomaly in against the known clutter background. This method suggests that the clutter shows a certain amount of correlation. Experimentally it is shown that the dominant interference in GPR data is correlated clutter, i.e., interference, which has a large correlation coefficient for lags greater than zero. However, the clutter environment cannot be considered completely stationary. An ideal filter would then be an adaptive filter, which estimates the slowly varying local clutter parameters, all the time ignoring the small parameter jumps caused by the buried targets to be detected. Kalman filtering is used for the estimation of the clutter parameters in the presence of random noise, detects jumps that occur at unknown points in time, and provides estimates of the new parameter values, without altering the target return.

1 INTRODUCTION.

The detection of minimum-metal anti-personnel land mines with GPR is encounters the problem of the extreme clutter environment within the first 5 cm of the soil surface. Almost anything under the surface of the ground presents a return signal, which may be confused with a valid (lethal) target. Since in humanitarian demining, it is mandatory that a lethal target be detected with nearly 100 per cent reliability in any soil type, the processing needs to 'clean up' the GPR data before any other tomographic or recognition algorithms can be applied.

This paper focuses on the pre-processing of GPR data, in order to reduce drastically the influence of the near-surface clutter. In order to estimate this clutter, a method of 'Clutter Parameter Estimation' is chosen. The advantage of such approach lies therein that once the clutter is satisfactorily known, any target will show up as a small anomaly in against the known clutter background.

The second section discusses the representation of the signal, what exactly is contained in the clutter, and the other signal parts that were removed out of the signal. Section three introduces an ARMA model for clutter estimation and consecutive reduction. Section four presents a method based on the Kalman filter to estimate the parameters of the clutter, and how to remove it. Section five shows the results of both methods. Finally the last section draws some conclusions.

2 SIGNAL REPRESENTATION.

The basic model for the GPR returns used in this work can be represented as:

$$\vec{E}_{rec'd}(k) = \vec{E}_{rad}(k) \circledast (h_c(k) + h_t(k)) + n \qquad (1)$$

This represents the relationship between the radiated electric field and the received one, where $h_c(n)$ and $h_t(n)$ are the impulse responses of the clutter and target, respectively, and n represents the measurement noise.

The removal of $\vec{E}_{rad}(k)$, the emitted signal, by deconvolution is the first step that has to be taken in the pre-processing of the data. This step can be performed before or after the clutter reduction is executed. The only difference is that the clutter to be removed will be in a different representation. Since our algorithms are based on a learning and estimating of the local clutter parameters the difference should be minimal. Experience showed however that the deconvolution process (extensively discussed in [1]) yields better results when performed after clutter reduction. This is why in this paper the clutter reduction will be performed directly on the raw data.

The clutter includes many components as there are the crosstalk from transmitter to receiver antenna, the initial ground reflection and the reflections resulting from non target scatterers within the soil. At this stage, a target signal can be either a mine or non-mine object of minelike size; it is the objective of the post-processing to make the decision between the two. The noise component specifically refers to the random measurement noise which

0-7803-7011-2/01/$10.00 ©2001 IEEE

adds to the composite signal, and will be considered to be dealt with during the deconvolution step.

3 ARMA MODEL FOR CLUTTER ESTIMATION

In this section the simple and straightforward implementation is discussed.

The process starts with a small amount of known clutter samples. The samples are then represented in a certain domain. Many choices of domains are available, in a system identification approach one of the most natural choices is the ARMA model. This model estimates the transfer function $H(z)$ which yields the signal to be represented, when excited with a unit input function. The system described by $H(z)$ may be written in transfer function format as

$$H(z) = \frac{B(z^{-1})}{A(z^{-1})} \tag{2}$$

where

$$A(z^{-1}) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_{n_a} z^{-n_a} \tag{3a}$$
$$B(z^{-1}) = b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{n_b} z^{-n_b} \tag{3b}$$

This is equivalent with a discrete Linear Time Invariant system described by the difference equation

$$y(n) + a_1 y(n-1) + a_2 y(n-2) + \ldots + a_{n_a} y(n-n_a) = b_0 u(n) + b_1 u(n-1) + \cdots + b_{n_b} u(n-n_b) \tag{4}$$

where $y(n)$ is the output sequence and $u(n)$ is the input sequence, and n_a,n_b are the *order* of the output and input processes, respectively. For causality, $n_b \leq n_a$.

The vector of coefficients $[a_i, b_i]$ is the ARMA representation of the original signal. The parameters of the target may be estimated only after the clutter/noise parameters are determined and removed.

The clutter model may be represented by

$$\hat{\theta}_c = [a_1, a_2, \ldots a_{n_a}, b_0, b_1, \cdots b_{n_b}]^T \tag{5}$$

and its block diagram is shown in here:

$$\delta(n) \rightarrow \boxed{\hat{H}_c(z) = \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})}} \rightarrow \hat{s}_c(n)$$

Representation of Clutter Parametric Model.

The input to the clutter block diagram is a δ-function under the hypothesis of an ideally deconvolved emitted signal.

When modeling real systems, it is usually necessary to include a noise model which will account for any random disturbances caused by the measurement equipment, etc. The inclusion of a (possibly time-varying) noise model is therefore indicated.

Such a noise process model can be described by

$$\hat{\theta}_\nu = [g_1, g_2, \ldots g_{n_g}, f_0, f_1, \cdots f_{n_f}]^T \tag{6}$$

The input to the noise model is a vector of independent, identically distributed (i.i.d.) samples with a Gaussian amplitude distribution of zero mean and constant spectral intensity. The input is applied to a filter which provides appropriate spectral and amplitude shaping to represent the measured noise from A-Scan to A-Scan.

Once the clutter estimate and noise processes are determined, the target may be easily extracted by simply subtracting the sum of those estimates from the measured signal.

4 KALMAN FILTER FOR PARAMETERS ESTIMATION.

In this approach the ARMA parameters of the clutter are estimated using a Kalman filter, where the parameters are considered as being constant with some fluctuations. In order to dynamically take into account the presence of a scatter from an object, abnormalities (parameter jumps) are detected, and the Kalman filter is restarted using the previously estimated parameters. The proposed approach is based on the ideas suggested in [2]

The Kalman filter is based on the following equations [4]:
the *system equation*

$$X_{k+1} = F_k X_k + B_k U_k + W_k \tag{7}$$

and the *observation equation*

$$Z_k = H_k X_k + V_k \tag{8}$$

where X is the state vector, F the state transition matrix, U the input, H the observation matrix, V some Gaussian noise process and Z the measured output.

In this model the ARMA representation is used so that:

$$H_k = [-Z_{k-1} \; -Z_{k-2} \; \cdots \; -Z_{k-n} \; U_{k-1} \; U_{k-2} \; \cdots \; U_{k-n}] \tag{9}$$

and

$$X_k = [a_1 \; a_2 \; \cdots \; a_n \; b_1 \; b_2 \; \cdots \; b_n] \tag{10}$$

In our application the input can be considered to be zero, so that the equations are simplified:

$$X_{k+1} = X_k + W_k \tag{11}$$

where W_k is the process noise.

The Kalman filter algorithm is given by the following equations:

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R)^{-1} \tag{12}$$
$$X_k = X_k^- + K_k (Z_k - H_k X_k^-) \tag{13}$$
$$P_k = (I - K_k H_k) P_k^- \tag{14}$$

Where P_k^- stands for $P_{k/k-1}$ and R represents the measurement noise covariance matrix.

This standard Kalman filter algorithm allows the estimation of the parameters in the presence of noise. In order to detect when the parameters deviate abruptly from their nominal value, due to

a scattering from a target, and readjust the filter gain so as to produce new estimates of the changed parameters, a hypothesis testing is applied to the normalized residual $(Z_k - H_k X_k^-)$

Let $Q_{zk} = E[Z_k Z_k^t]$ the covariance of Z_k. The normalised residual is given by:

$$Z_{nk} = \frac{Z_k - H_k * X_k^-}{\sqrt{Q_{zk}}} \tag{15}$$

with

$$Q_{zk} = H_k * P_k^- * H_k + R \tag{16}$$

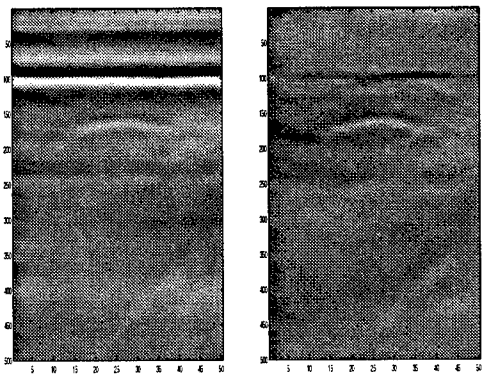
In this case Z_{nk} is a zero mean gaussian random variable. The comparison of Z_{nk} with a threshold suggests whether or not the signal parameters have effected a jump. On the basis of this the filter parameters are adjusted accordingly.

5 EXPERIMENTAL RESULTS AND DISCUSSION

The data used in this section were taken from two sources. The first was acquired at TUI [1], the second at RMA [2]. The data from the TUI is represented by a Bscan, acquired while scanning over a number of buried objects. The horizontal axis represents scanned distance with a scanning step of 2 cm, while the vertical axis shows measured time samples. The raw data used here are the data after subtraction of the measured cross talk. The data from RMA represent a Bscan over one buried object with a scanning step of 1 cm.

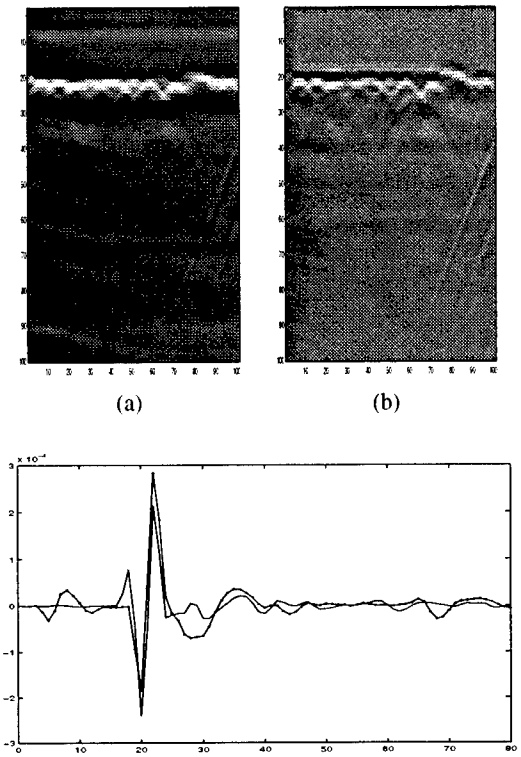
5.1 ARMA model for clutter estimation.

Here we show some results for the method described in 3.



(a) (b)

Figure 1: The results of the Dynamic method on the bistatic data. (a) Raw Data, (b) after processing



(a) (b)

(c)

Figure 2: The results of the first method on one of the channels of the array data. (a) Raw Data, (b) after processing, (c) representative Ascans from both (.-Raw and -Processed)

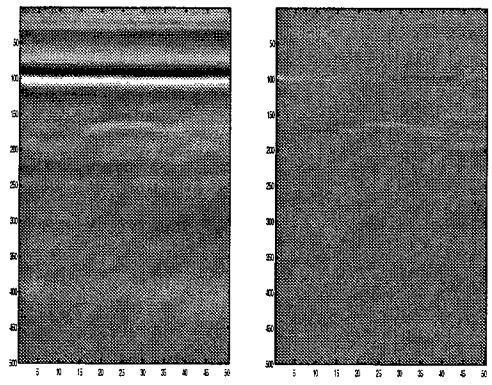
In the figures 1 and 2 the results for the dynamic method are shown for the bistatic data resp. the array data. As it can be seen, comparing the raw data bscans with the processed ones, the clutter both above and below the signals are reduced. The remaining signals are however somewhat distorted as can be seen in figure 2.c. Here the same representative Ascan is extracted from the raw and processed Bscans, and plotted on the same figure(The line with .- represents the raw data and the full line the processed data). It shows that the clutter, especially above the signal is reduced, but that the signal itself has undergone some distorting.

5.2 The Kalman Filter

Here we show some results for the method described in 4.

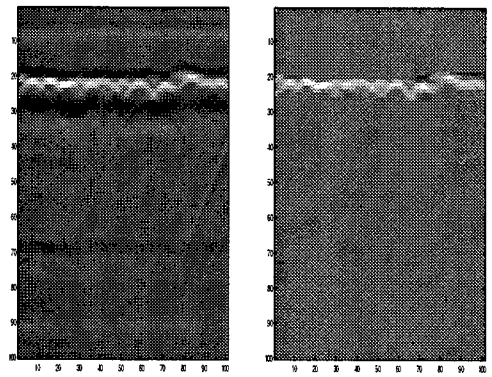
[1]The Technische Universität Ilmenau has a setup that simulates an array of 6 emitting and receiving antennae. The data is acquired in the frequency domain, between 1 and 6 GHz.

[2]The Royal Military Academy has an ultra wideband system with one emitting and one receiving antenna. For a more detailed description of the system, refer to [3]

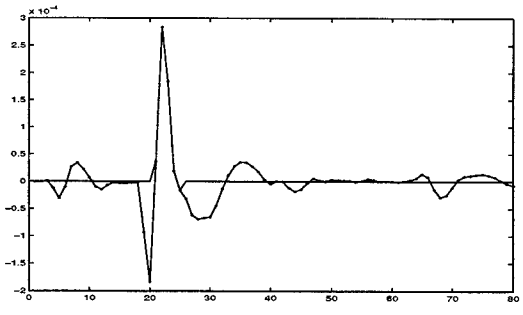


(a) (b)

Figure 3: The results of the Kalman method on the bistatic data. (a) Raw Data, (b) after processing



(a) (b)



(c)

Figure 4: The results of the Kalman method on one of the channels of the array data. (a) Raw Data, (b) after processing, (c) representative Ascans from both (.-Raw and -Processed)

In figures 3 and 4, The proposed Kalman Processing was performed on the representative Bscans. As explained in section 4, the proposed method is a 1 dimensional one. The clutter above and below the signals is almost completely nullified. In figure 3 some clutter remains above the signal, due to the fact that the raw data still contained the antenna crosstalk. Figure 4.c shows again the comparison between two representative Ascans (The line with .- represents the raw data and the full line the processed data). Here it is clear that the clutter above and below the signal is eliminated, while preserving the shape of the original signal quite accurately.

6 CONCLUSION

Two methods were theoretically introduced to reduce the 2D clutter in GPR Bscan images. They were applied to data originating from two different types of GPR. The Kalman method was found to give the better results, reducing most of the clutter to zero, while preserving the shape of the original signal.

ACKNOWLEDGEMENTS. A large part of the work presented in this work is performed in the framework of the DEMINE project. The authors wish to thank the people of the Technische Universität Ilmenau (TUI), one of the DEMINE partners, who provided part of the data used in this paper. The other data used, was acquired with a setup, placed in the Royal Military Academy (RMA) of Belgium. The authors wish to thank there especially Bart Scheers, for positive collaboration in the context of the HUDEM project.

7 REFERENCES

[1] L. van Kempen, H.Sahli and J. Brooks and J. Cornelis,"New Results on Clutter Reduction and Parameter Estimation for Landmine Detection using GPR", *GPR 2000, Eighth International Conference on Ground Penetrating Radar, Gold Coast, Australia*, May 23-26, 2000", pp. 872-879.

[2] F. Chowdhury, "Kalman Filter with Hypothesis Testing: A Tool for Estimating Uncertain Parameters", *Circuits Systems Signal Processing.*, 1996, Vol 15;3, pp. 291-311.

[3] B. Scheers and Y. Plasma and M. Piette and M. Acheroy and A. Vander Vorst, "Laboratory UWB GPR System For Landmine Detection", *GPR 2000, Eighth International Conference on Ground Penetrating Radar, Gold Coast, Australia*, May 23-26, 2000", pp. 747-757 .

[4] R. Brown and P. Hwang, "Introduction to Random Signals and Applied Kalman Filtering", *John Wiley & Sons, Inc. edts.*, 1996

MODEL-BASED STATISTICAL SIGNAL PROCESSING USING ELECTROMAGNETIC INDUCTION DATA FOR LANDMINE DETECTION AND CLASSIFICATION

Leslie Collins, Ping Gao, and Stacy Tantum

Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708-0291, USA

ABSTRACT

Traditionally, electromagnetic induction (EMI) sensors are operated in the time-domain and the response strength is related to the amount of metal present in the object. These sensors have been used almost exclusively for landmine detection. Unfortunately, there is often a significant amount of metallic clutter in the environment that also induces an EMI response. Consequently, EMI sensors employing detection algorithms based solely on metal content suffer from large false alarm rates. A second issue regarding processing of data collected on highly cluttered sites is that anomalies are often in close proximity, and the measured EMI signal consists of a weighted sum of responses from each anomaly. To mitigate the false alarm problem, statistical algorithms have been developed which exploit models of the underlying physics for mines with substantial metal content. In such models it is commonly assumed that the soil has a negligible effect on the sensor response, thus the object is modeled in "free space". To date, such advanced algorithms have not been applied specifically to the problem of detecting of low-metal mines in a cluttered environment. Addressing this problem requires considering the effects of soil on signatures, separating the multiple signatures constituting the measured EMI response as well as discriminating between landmine signatures and clutter signatures. In this paper, we consider statistically based approaches to the landmine detection and classification problem for frequency-domain EMI sensors. We also develop a preliminary statistical approach based on independent components analysis (ICA) for separating the signals of multiple objects that are within the field of view of the sensor and illustrate the performance of this approach on measured data.

1. INTRODUCTION

Land mines and unexploded ordnance (UXO) present a significant threat to individuals around the world. Currently deployed methods of clearing subsurface threat items are slow and less than 100% accurate. A variety of sensors for landmine and UXO detection have been proposed and utilized, each of which exploits a different fundamental phenomenology. The most commonly deployed sensor is an electromagnetic induction (EMI) sensor that operates by detecting the metal present in land mines. The high level of risk associated with the landmine detection problem requires 100% detection performance for any viable sensor for all possible targets. However, there are hundreds of varieties of land mines that vary in their construction from metal-cased varieties with a large mass of metal to plastic-cased varieties with very small amounts of metal. In addition, there is often a significant amount of metallic debris (clutter) present in the environment. Consequently, EMI sensors that utilize traditional detection algorithms based solely on the metal content operating at a high enough detection rate to satisfy performance requirements suffer from very high false alarm rates.

To address the false alarm issue, several groups have investigated target identification, or discrimination, using EMI sensors, while other groups have considered alternative sensor modalities and sensor fusion [1-8]. In [1,7], classification of metal targets using frequency-domain EMI sensors is considered, and a Bayesian approach is applied to address the inherent uncertainties concerning the target/sensor orientation. In [2], this same sensor is utilized to investigate the frequency-domain signatures of landmines. In the work presented here, a statistically-motivated approach is evaluated in a blind field trial, which is a better method of evaluating robustness than evaluating algorithms solely on synthesized, laboratory, or fully ground-truthed data.

Some statistical approaches to this problem may be ineffective since a statistical model is needed to describe the null hypothesis, and there is often insufficient data available to adequately develop an accurate statistical model [6]. Although the response to the ground can be characterized statistically, it is very likely that it is inappropriate to model discrete clutter in the same manner. Recent results from the signal processing community have indicated that an adaptive coherence detector [10] has optimality properties under a set of conditions that are applicable to the landmine detection problem considered here. Specifically, when detecting a signal of known form but unknown amplitude in zero-mean Gaussian noise with a known covariance structure but unknown gain, a correlation coefficient, or cosine statistic, provides a uniformly most powerful invariant test statistic [9,10]. Since the subspace algorithm is predicated on an assumption that the null hypothesis follows a zero mean Gaussian distribution, this particular detector is not directly applicable to the problem of landmine detection in the presence of anthropic clutter. However, the invariance class associated with this algorithm does address some of the issues associated with the detection problem and the sensor that we are considering.

In this paper, performance of an algorithm based on a subspace detector was evaluated on data collected in a blind field test. We have modified the original formulation of the detector to consider both a multi-component alternative hypothesis and a null hypothesis that does not follow a zero-mean multivariate normal model. We developed a set of features that could be used to robustly differentiate between landmines and discrete clutter and that could also be extracted from the EMI data in real time.

2. SENSOR AND DATA

The GEM-3 is a prototype wide-band frequency-domain EMI sensor manufactured by Geophex, Ltd. The GEM-3 uses a pair of concentric, circular coils to transmit a continuous, wideband, electromagnetic waveform [8]. The resulting field induces a secondary current in the earth as well as in any buried objects.

0-7803-7011-2/01/$10.00 ©2001 IEEE

The set of two transmitter coils has been designed so that they create a magnetic cavity at the center of the two coils. A third receiving coil is placed within the magnetic cavity so that it senses only the weak secondary field returned from the earth and buried objects. The frequency-domain in-phase and quadrature components are obtained from the received signal by convolving the received time-series with a sine time-series (for in-phase) and cosine time-series (for quadrature) at the frequency of interest.

When a mine is present, the response of the sensor consists of the sum of a response due to the mine and a response due to the background. To obtain the response due to the mine alone, it is necessary to determine the response of the sensor to the background alone. The GEM-3 sensor has some level of thermal drift in its background response [6], so the sensor noise cannot be treated with a simple statistical model (e.g. zero mean Gaussian). This drift must be tracked so that the background response can be removed from the measured signature.

Details of the data collection plan can be found in [11], however the most salient points are summarized here. A 50 meter by 20 meter plot of ground was selected for construction of the test grid, and a calibration area was created in an adjacent 5 meter by 25 meter plot. Initially, all indigenous clutter was removed from the site. Mine targets emplaced in the test grids were predominately "low metal" mines since these are the most challenging targets to detect using EMI sensors. Samples of the indigenous clutter were re-emplaced in the grids to provide discrete opportunities for false alarms. The ground truth associated with the calibration area is available; however, the ground truth associated with the main, or blind, test grid is sequestered. Algorithm developers provide the output of their algorithms for each grid square or "decision opportunity" to JUXOCO for scoring. The GEM-3 EMI sensor was programmed to measure responses at 20 frequencies spaced logarithmically between 270 and 23,790 Hz. Ten spatial positions in a '+' pattern were measured in each grid point, since spatial information has been shown to improved detection and discrimination performance [1]. Samples were taken every 2".

In order to track the background signature, background measurements were taken at one of four grid squares that had been set aside as known "blanks" by JUXOCO. The closest "blank" square was used as the background for each grid square while the measurements were taken in the lanes. A background measurement was taken before and after signature data was collected in each grid square. In order to compensate for sensor drift, a linear prediction algorithm was used to predict the background signature during each of the 10 measurements made in the grid square using the background data measured before and after data was collected in each square. These "corrected" data were the input to the algorithm described in the next section.

3. ALGORITHM DEVELOPMENT

Kraut, Sharf et al. have used the theory of generalized likelihood ratio tests to show that previously proposed matched subspace detectors can be employed in the case of unknown noise covariance and unknown scaling of a mean vector [9,10]. They have described what they term "adaptive" matched subspace detectors, where adaptive implies an estimation of the covariance structure using training data, and have investigated their optimality properties. The invariance to overall data scaling as well as the optimality when testing and training sets have a different scaling factor applied to the covariance matrix is one of the most appealing properties of this set of detectors for landmine detection problem. As the authors state, what these detectors sacrifice in high-SNR performance they gain in robustness to uncertain and *changeable* prior information regarding the signal and noise model and statistics.

The general detection problem addressed by Sharf and his colleagues is one in which a N-element signal, $\boldsymbol{s}$, is located in a signal subspace of dimension p. The signal is scaled by an unknown constant k and scaled noise, $g\mathbf{w}$ is added to the signal where, the scaling factor g is unknown. The measured signal, $\boldsymbol{r}$, is given by $\mathbf{r} = k\mathbf{s} + g\mathbf{w}$ which follows a complex normal (CN) distribution $\mathbf{r} \sim CN_N(k\mathbf{s}, g^2\Sigma)$. This formulation admits a solution ranging in complexity from a rank 1 matched filter (p=1) to rank p subspace detectors where $\mathbf{s} = \Psi\theta$, Ψ denotes the signal subspace and θ is the known or unknown parameter vector which locates the signal in the subspace. Under the signal hypothesis, H_1, $k \neq 0$ while under the null hypothesis, H_0, $k = 0$. Previous work has addressed this detection problem under various assumptions regarding the parameters Ψ, Σ, g^2, and θ.

For the mine detection problem, we know s, but do not know k, g, or Σ. It was shown in [9] that the optimal test under these uncertainties is given by a cosine-squared statistic

$$\cos^2 = \frac{(\mathbf{r}^T\hat{\Sigma}^{-1}\mathbf{s})^2}{(\mathbf{r}^T\hat{\Sigma}^{-1}\mathbf{r})(\mathbf{s}^T\hat{\Sigma}^{-1}\mathbf{s})}$$

and is termed a coherent CFAR adaptive subspace detector (ASD). In the formulation, Σ is estimated using maximum likelihood techniques and then used in the formulation of the detector. This approach is generally associated with a generalized likelihood ratio test, however it was shown in [9] that this detector is in fact optimal and uniformly most powerful.

To apply the subspace method described above we first consider only the data measured at the center point of each grid sqare, and if we assume the data follows the model

$$H_1 : r_i \sim N(k\mathbf{s}_i, g\Sigma)$$
$$H_0 : r_i \sim N(0, g\Sigma)$$

where the subscript denotes the ith target and k and g are arbitrary scaling factors on the mean and covariance matrix. For the mine detection problem considered here, the phenomenology inherent in the physical problem allows us to interpret the scalar k as uncertainty in object depth and the scalar g as uncertainty in the strength of the noise process resulting from the thermal drift. Under these assumptions, the detector for each spatial position is given by

$$\beta_i = \frac{(\mathbf{r}^T\hat{\Sigma}^{-1}\mathbf{s}_i)^2}{(\mathbf{r}^T\hat{\Sigma}^{-1}\mathbf{r})(\mathbf{s}_i^T\hat{\Sigma}^{-1}\mathbf{s}_i)}$$

This detector would be appropriate for finding a single mine type at the center position in a zero mean background. To extend the formulation to consider N possible mine types at the center

position in zero mean background (corresponding to prior knowledge regarding the form of θ) a bank of such detectors is used and the maximum output is selected:

$$\beta = \max_i \beta_i$$

This formulation also allows classification of the mine by type.

When clutter is encountered, the zero mean assumption on the measured signal under the null hypothesis is clearly invalid. There are two mechanisms by which to modify the detector described above to consider the case of discrete clutter. In the first, we considered an approach described in [10] wherein clutter is assumed to lie in a rank t "interference" subspace and is thus treated separately from the noise or background. Although this is probably an accurate model, in the application we are considering there was not enough data to accurately model the subspace. For example, for the M=20 clutter items contained in the data set, it is not usually possible to write one of the clutter signals as a linear combination of the other M-1 signals, i.e. the measured signals do not span the clutter subspace.

The second approach, which was adopted here, is to utilize the output of the bank of cosine-filters as a feature set and to develop statistics for that feature set under the two hypotheses. These statistics can then be used in the formulation of a likelihood ratio. Let $\boldsymbol{t}$ be the 11 x 1 data vector formed by concatenating the sorted β_i then the likelihood ratio for this data set is

$$\lambda(\mathbf{t}) = \frac{f(\mathbf{t}/H_1)}{f(\mathbf{t}/H_0)}$$

To complete the formulation of this detector, the pdfs of $\boldsymbol{t}$ under each hypothesis must be estimated. These estimates were obtained using the calibration data collected in conjunction with the JUXOCO experiment. A uniform distribution with independent components was used to model $f(t/H_1)$ and a Gaussian distribution (both the mean and the covariance structure were estimated) was used to model $f(t/H_0)$. A simple extension of this approach allows the incorporation of spatial data.

4. RESULTS

In Figure 1, ROC curves are presented for three algorithms. The performance of the baseline algorithm is shown with a solid line and represents the performance obtained when EMI detectors are operated in a traditional energy-detection mode. The energy of the signal measured at the center of each grid square was reported to JUXOCO for scoring. Clearly, energy, which is proportional to metal content and inversely proportional to distance between the object and the sensor, does not provide a good discriminator between landmines and clutter or ground at this site. The performance of the cosine detector operating on the signal measured at the center of each grid square is shown with the dashed line. This algorithm is the coherent CFAR ASD described in [9] modified to utilize the maximum statistic over each of the signals, $\mathbf{s_i}$. Also shown is the performance of the modified algorithm that utilizes the output of the coherent CFAR ASD for each potential mine signal as the input to the likelihood ratio. It is labeled with the notation 'clutter' to indicate that for this formulation the statistics of the discrete clutter objects are included in the formulation, which is not true for the standard coherent CFAR ASD. Including a clutter model improves the performance of this algorithm, and both approaches perform dramatically better than the baseline.

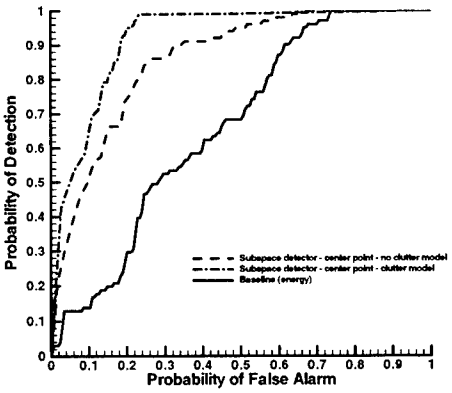


Figure 1. ROCs for the energy detector, CFAR ASD detector and the modified CFAR ASD detector on center point data.

In Figure 2, a similar set of curves is provided; however, in this figure the spatial information has been incorporated into the two CFAR ASD formulations. Again, including a clutter model improves the performance of this algorithm, and both approaches perform dramatically better than the baseline algorithm. In the lower false alarm rate region there is little difference between the performance of the two algorithms. This may be a result of the fact that spatial information for the low-metal mines falls off into the noise floor faster than for the larger metal mines. In the high Pd range, the modified ASD detector performs better than the standard ASD detector.

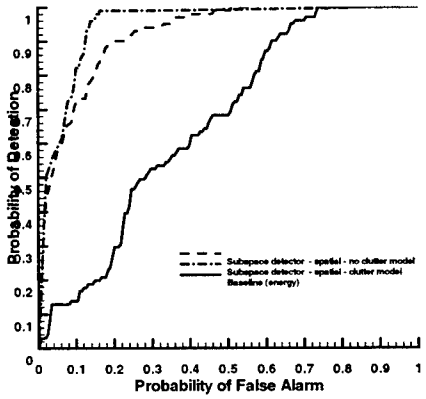


Figure 1. ROCs for the energy detector, CFAR ASD detector and the modified CFAR ASD detector on center point data.

As mentioned previously, it is possible to obtain classification information as the detector is implemented as a bank of processors, each tuned to a particular mine. We utilized the processor which has the maximum output to name the mine and this information was sent to JUXOCO for scoring. Using this approach, the mines were "named" correctly 65% of the time.

5. SIGNAL SEPARATION

One issue not considered in the above development is the case of overlapping signals. In actual field environments, objects are often in close proximity and thus the measured response will consist of a combination of the responses from the individual objects. For algorithms such as those described above, these signals must be separated prior to their application. Independent Components Analysis (ICA) has also been proposed as a viable solution for the problem of blind source separation [12-15], although it has not been explored in the subsurface sensing application. Therefore, an ICA algorithm has been implemented to test the feasibility of this approach for the problem of separating two landmine signatures from a set of measured responses.

We considered the signature of two ordnance items, where in-phase (solid black) and quadrature (dashed black) measured data as a function of frequency for the objects in isolation are shown in the top two panels of Figure 3. These signals were then "mixed", and the two of the resultant signals are plotted in the middle two panels of Figure 3. The mixing coefficients were selected so that the mixing is consistent with signatures that could be expected for EMI data measured at eleven different spatial locations. The bottom two panels show the two "independent components" extracted by a simple implementation of the ICA algorithm. As is typical with ICA algorithms, the signal with the highest energy is extracted first and with the highest fidelity. Clearly, this simple implementation, which has not been optimized for this problem, does an excellent job of extracting the signature of one of the objects and a reasonable job of extracting the second.

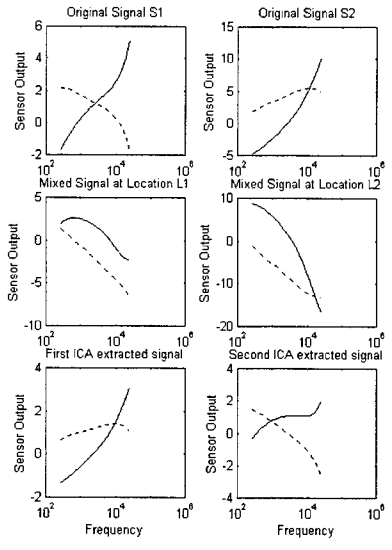


Figure 3. ICA Results. Top panel shows signals measured using the GEM3 from a Valmara landmine (left) and VS50 landmine (right). Middle panel shows two examples of the mixed signals that were supplied to the ICA algorithm. Bottom panel shows the signals as separated by the ICA algorithm.

6. ACKNOWLEDGEMENT

The authors would like to thank J. Cary, J. Moulton, L. Makowsky, D. Reidy, and D. Weaver for their help and support on the data collection and scoring effort. This work was supported by the Joint UXO Coordination Office and the Night Vision and Electronic Systems Directorate of the US Army.

7. REFERENCES

[1] P. Gao, L. Collins, P. Garber, N. Geng, and L. Carin, "Classification of Landmine-Like Metal Targets Using Wideband Electromagnetic Induction". IEEE Trans. Geosc. Remote Sens., Vol. 38, No. 3, May, 2000, 1352-1361.

[2] D.Keiswetter, I.J. Won, B.Barrow, and T.Bell, "Object identification using multifrequency EMI data," UXO Forum '99, Atlanta, GA, May 1999.

[3] L.S. Riggs, J.E. Mooney, and D.E. Lawrence, "Identification of metallic mine-like objects using low frequency magnetic fields," IEEE Trans. Geosc. Remote Sens., vol. 39, no. 1, pp. 56—66.

[4] G.D. Sower and S.P. Cave, "Detection and identification of mines from natural magnetic and electromagnetic resonances," Proceedings of SPIE, Orlando, FL, April 1995.

[5] A.H. Trang, P.V. Czipott, and D.A. Waldron, "Characterization of small metallic objects and non-metallic anit-personnel mines," Proceedings of SPIE, Orlando, FL, April 1997.

[6] L.M. Collins, P. Gao, S. Tantum, L. Makowsky, J. Moulton, D. Reidy, R.C, Weaverm=, "Improving Detection of Low-Metallic Content Landmines Using EMI Data" Proceedings of SPIE, Orlando, FL, April 2000.

[7] P. Gao and L. Collins, "A Comparison of Optimal and Sub-Optimal Processors for Classification of Buried Metal Objects," IEEE Signal Processing Letters, Volume 6, Issue 8, August, 1999, 216-218.

[8] I. J. Won, D. A. Keiswetter, and D. R. Hansen, "GEM-3: A Monostatic Broadband Electromagnetic Induction Sensor". Journal of Envir. Engin. Geophys., 2: 53-64, Aug. 1997.

[9] S. Kraut, L.L. Scharf, and T. McWhorter, "Adaptive Subspace Detectors", IEEE Trans. Sig. Proc., Vo. 49, January 2001, 1-16.

[10] L.L. Scharf and B. Friedlander, "Matched Subspace Detectors", IEEE Trans. Sig. Proc., Vol 42, August 1994, 2146-2157.

[11] "Hand Held Metallic Mine Detector Performance Baselining Collection Plan", JUXOCO, Ft. Belvoir, VA, December 1998.

[12] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," Neural Computation, vol 7, 1995.

[13] Hyvarinen, A., and Oja, E., "Independent Component Analysis: Algorithms and Applications," in press, Neural Networks.

[14] S.-I. Amari, A. Cichocki, and H.H. Yang. "A new learning algorithm for blind source separation", In Advances in Neural Information Processing Systems 8, pages 757-763. MIT Press, Cambridge, MA, 1996.

[15] 41. J.-F. Cardoso. "Infomax and maximum likelihood for source separation", IEEE Letters on Signal Processing, 4:112-114, 1997.

POLYNOMIAL PHASE SIGNAL BASED DETECTION OF BURIED LANDMINES USING GROUND PENETRATING RADAR

Luke A. Cirillo, Christopher L. Brown and Abdelhak M. Zoubir

Australian Telecommunications Research Institute &
School of Electrical and Computer Engineering
Curtin University of Technology
GPO Box U1987, Perth 6845, Australia
luke@atri.curtin.edu.au

ABSTRACT

Put simply, the global landmine problem is massive. Ground Penetrating Radar (GPR) is just one engineering solution currently being investigated. A polynomial amplitude - polynomial phase model is fitted to GPR returns. It is observed that the second order phase coefficient shows deviations from background-only levels when a buried target is present. A bootstrap-based detection scheme is proposed that tests for this change. The technique is applied to real GPR data, with encouraging results.

1. INTRODUCTION

Anti-personnel landmines have been used in war zones throughout the world, and have profound effects on civilian populations. Landmines have a very long life-span, rendering many post-war areas both useless and dangerous. These minefields can be found anywhere from agricultural fields, river banks, urban areas, transport routes and surrounding villages. The effect is a terrorised and demoralised local population.

In many post-war zones, landmines with little or no metal content have been found. They are often quite small; made using a plastic casing and very few, if any, metal parts. Consequently, conventional metal detectors are not effective countermeasures for these mines. Metal detectors also suffer from a high false alarm rate due to shrapnel and debris lodged below the surface.

Surface or ground penetrating radar (GPR) [1, 2] works by detecting discontinuities in the dielectric properties of the soil. The size and shape of targets made from materials such as plastic can potentially be determined using this technology. However, environmental conditions such as soil type and moisture content can heavily influence the performance of a GPR system. Therefore, signal processing techniques are needed in order to develop robust detection schemes which can compensate for changes in background conditions.

The complex nature of the physical scenario makes it very difficult to accurately model a GPR return. In this paper, investigations into a polynomial amplitude - polynomial phase model are made.

2. SIGNAL MODEL AND PRELIMINARY INVESTIGATIONS

It is difficult to define a complete, physically-motivated signal model of the GPR backscatter waveform. However, based on extensive data analysis, a polynomial amplitude - polynomial phase model has been proposed (following [3])

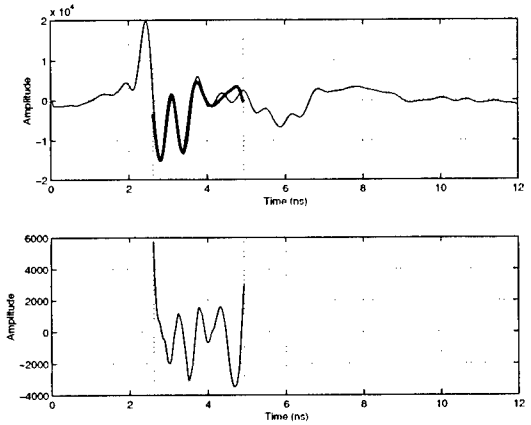


Fig. 1. Top: Fit of the polynomial amplitude - polynomial phase model (bold line) to data (fine line) from non-homogeneous soil. Bottom: Residual of the fit to data.

$$g_t = \left[\sum_{n=0}^{P_a} a_n t^n\right] \exp\left[j \sum_{m=0}^{P_b} b_m t^m\right] + z_t \qquad (1)$$

where z_t is assumed to be stationary interference. The amplitude and frequency modulation are described by polynomials of order P_a and P_b respectively.

To demonstrate the validity of this signal model, it has been applied to GPR returns from clay soil, containing a small plastic target, denoted ST-AP(1) – a surrogate for the M14 anti-personnel (AP) mine [4]. This target has a PVC casing and is filled with paraffin wax. A solid stainless steel cylinder 5 cm in diameter and length is also present, denoted by SS05x05. Phase and amplitude parameters are estimated using the DPT [5] and the method of least-squares respectively.

Results are shown for a single GPR return signal, in Figure 1. Here a window length of 100 samples has been chosen in a section of backscatter which contains a shallow buried AP mine. Phase and amplitude models of order 2 and 4 respectively have been used. The model is observed to closely approximate the GPR signal. As a result of extensive data analysis of returns from a variety

0-7803-7011-2/01/$10.00 ©2001 IEEE

of soil types, it was found that the polynomial phase may have order up to 3. It is proposed that a change in the model parameters will indicate the presence of a target.

The three phase coefficients from a second order phase model (i.e. chirp) are estimated for a full data set (B-scan) containing the two targets: ST-AP(1) and SS05x05. From the results in Figure 2, it is noted that, unlike b_0 and b_1, b_2 appears to show a deviation at *both* targets when compared to its underlying background-only value. This suggests that testing for a change in b_2 may be a target indicator. This was supported by results obtained from a variety of data sets.

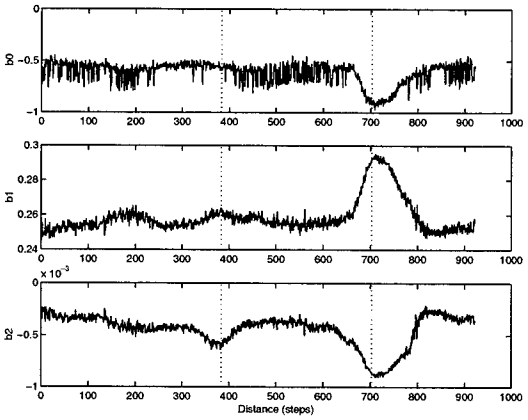


Fig. 2. Phase coefficients from an order 2 model. Vertical dotted lines indicate the approximate position of ST-AP(1) (left) and SS05x05 (right) targets.

3. TEST ON THE PARAMETERS

From a background region of 100 samples, estimates of b_2 are seen to be *approximately* Gaussian distributed. A normal probability plot ($Q-Q$ plot) of $\hat{b}_2$ is shown in Figure 3. It is suggested that the mean of this distribution changes in the presence of a target. Although, for the most part, the approximation appears to be valid, there does appear to be some deviation from Gaussianity in the tails.

For the data set shown in Figure 2, b_2 is seen to decrease in the presence of a target, however, in some cases, in particular when the soil type is loam, b_2 *increases* with the presence of a target. Therefore, two-sided hypotheses are considered

$$\begin{aligned} H_0 &: b_2 = b_{2,0} \\ H_A &: b_2 \neq b_{2,0} \end{aligned}$$

where $b_{2,0}$ denotes the value of b_2 under the null (no target present). The obvious test statistic is

$$T = \frac{\hat{b}_2 - b_{2,0}}{\hat{\sigma}_{\hat{b}_2}}$$

where $\hat{\sigma}_{\hat{b}_2}$ is an estimate of the standard deviation of $\hat{b}_2$. Since the exact distribution of $\hat{b}_2$ is unknown – it may deviate from Gaussianity in the tails and has unknown variance – it is proposed that the bootstrap be used to determine thresholds.

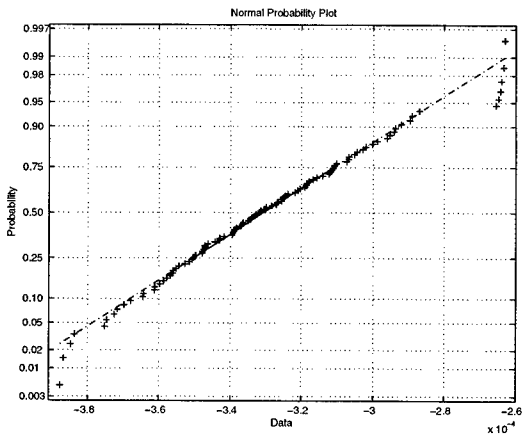


Fig. 3. Normal probability plot of third phase coefficient estimate, $\hat{b}_2$, from a background region.

4. BOOTSTRAP-BASED DETECTION

The bootstrap was introduced [6] as a tool for estimating the sample distribution of statistics when standard methods cannot be applied. Observations are randomly resampled and the statistics recomputed – mimicking the process of repeating the experiment.

When this is done a large number of times, the distribution of the re-computed values approximates the distribution of the statistic. Consequently, the test statistic can be compared to this bootstrap distribution, and a hypothesis test performed. More information on the use of the bootstrap for hypothesis testing can be found in [7].

The technique used here is described in Table 1. After fitting the polynomial amplitude - polynomial phase model to the data, the residuals are whitened using an AR model. A block bootstrap resampling technique is used due to remaining structure in the residuals – this ensured that the bootstrap signals were similar in form to the observed signals. See Figure 4 for examples of the generated bootstrap signals. The bound, $b_{2,0}$, is found from a region which is known to be target-free.

5. APPLICATION TO REAL DATA

The proposed detector has been applied to GPR data collected at the Defence Science and Technology Organisation, Salisbury, South Australia. An FR-127-MSCB Impulse GPR (ImGPR) system developed by the Commonwealth Scientific and Industrial Research Organisation (CSIRO, Australia) has been used for these measurements [4, 9]. The system collects 127 returns, or soundings, per second, each composed of 512 samples with 12 bit accuracy. The sounding range may vary from 4 ns to 32 ns. The GPR system uses bistatic bow-tie antennas which transmit wide-band, ultra-short duration pulses.

In this experiment, the antennas had a centre frequency of 1.4 GHz and 80% bandwidth. The GPR unit is suspended above the ground surface at a height of between 0.5 to 2 cm. Its motion is controlled by a stepper motor unit running along a track at a constant velocity, as shown in Figure 5. Since the motion of the GPR is controlled by a stepper motor, with constant speed, running on a straight track, these samples correspond to distances from the

1. If g_n for $n = 0, ..., N-1$ is a sampled GPR return signal, fit a polynomial amplitude - polynomial phase model to the data. From the model, an estimate of the second order phase parameter, $\hat{b}_2$, is made.
2. Form the residual signal $r_n = g_n - \hat{g}_n$, where $\hat{g}_n$ is the polynomial amplitude-polynomial phase model corresponding to estimated parameters.
3. Whiten r_n by removing an AR model of suitable to obtain the innovations z_n.
4. Re-sample from z_n N times using the Block Bootstrap [8] to obtain z_n^*.
5. Repeat step 4 B times to obtain $z_n^{*1}, ..., z_n^{*B}$.
6. Generate B bootstrap residual signals r_n^{*i}, for $i = 1, ..., B$ by filtering z_n^{*i} with the AR process obtained in 3.
7. Generate B bootstrap signals $g_n^{*i} = \hat{g}_n + r_n^{*i}$, for $i = 1, ..., B$.
8. Estimate the third phase coefficient from g_n^{*i} to obtain $\hat{b}_2^{*i}$ for $i = 1, ..., B$.
9. Calculate the bootstrap statistics $T^{*i} = \frac{\hat{b}_2^{*i} - \hat{b}_2}{\hat{\sigma}^*_{\hat{b}_2}}$ for $i = 1, ..., B$.
10. Compare the test statistic $T = \frac{\hat{b}_2 - b_{2,0}}{\hat{\sigma}_{\hat{b}_2}}$ to the empirical distribution of T^*. Reject H_0 if T is in the tail of the distribution, otherwise retain H_0.

Table 1. Bootstrap-based testing of b_2.

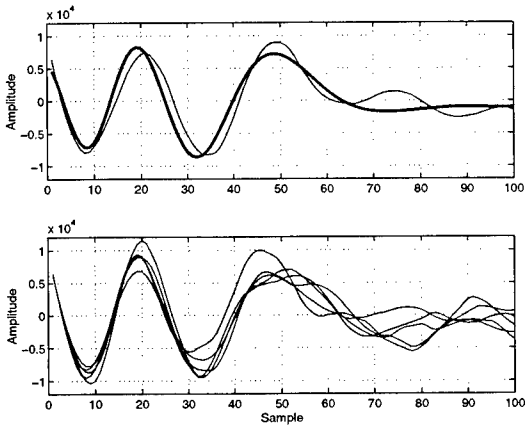


Fig. 4. Top: Fit of the polynomial amplitude - polynomial phase model (bold line) to data (fine line) in the region of interest. Bottom: Bootstrap signals generated using the model and blocked bootstrap resampling.

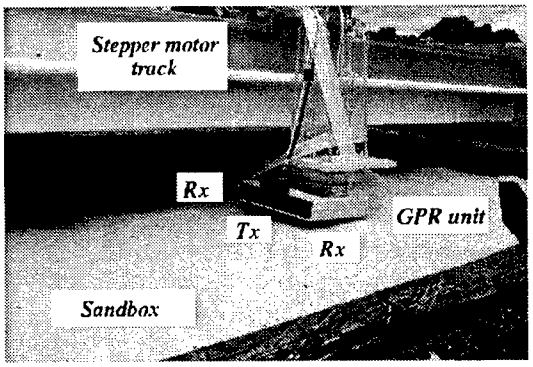


Fig. 5. The ImGPR unit running over a sandbox.

Target	Surrogate	Dimensions diam×hght	Orientation
ST-AP(1)	M14	52 × 42 mm	not critical
ST-AP(2)	PMN	118 × 50mm	sensitive
ST-AP(3)	PMN2	115 × 53mm	sensitive

Table 2. Minelike surrogate targets used.

starting point of the run.

Some of the targets used in the trial are listed in Table 2. The PMN and PMN-2 are AP mines with non-metallic casings. The M14 is an AP mine with almost no metal content and small size. As such it is a very difficult target for detection.

Shown in Figures 6,7 and 8 are sample results for three surrogate mines as well as various other "targets". The location of the first target in each scenario is around the 300'th trace, while the second target is around the 690'th trace. Overall, the results are very encouraging. It can be said that the targets have been correctly detected, while false alarms appear to be concentrated in areas near the targets – rather than *true* false alarms that occur far from the targets in background only areas. These near-target false alarms may be triggered by disturbances to the soil structure caused by the burying of the target.

All results included here were obtained from targets buried at approximately 5 cm below the surface. Results from shallow buried targets in the range 0.5 cm to 2 cm below the surface produce a more significant change in the test statistic, and the detection region is has greater spread. A comparative investigation of alternate models and detection schemes is continuing. Testing on different scenarios with different soil types is also ongoing.

6. CONCLUSIONS

From preliminary results it has been seen that when a polynomial amplitude - polynomial phase model is fitted to GPR returns, changes in the estimated second order phase parameter indicates the presence of a target. At present, the bootstrap is being utilised to estimate the distribution of the parameter. This has yielded a detector that has shown very encouraging results when run on real GPR data.

It should be stressed that the method presented in this paper is purely for detection. Classification is omitted. Classification

of detected targets would also be required for effective landmine clearance. Following a detection stage, techniques such as multiple test procedures [10] and time-frequency signatures [11] have been applied for this purpose.

7. ACKNOWLEDGEMENTS

The authors would like to thank Dr Ian Chant and Dr Canicious Abeynayake of the Defence Science and Technology Organisation, Salisbury, South Australia, for their assistance, particularly in the collection and analysis of data.

8. REFERENCES

[1] D. J. Daniels, *Surface-penetrating radar*, IEE, 1996.

[2] D. J. Daniels, D. J. Gunton, and H. F. Scott, "Introduction to subsurface radar," *Proceedings of IEE, F*, vol. 135, pp. 278–317, 1988.

[3] A. M. Zoubir, D. R. Iskander, I. Chant, and D. Carevic, "Detection of landmines using Ground-Penetrating Radar," in *Proc. SPIE: Detection and Remediation Technologies for Mines and Minelike Targets IV*, Orlando, USA, August 1999, vol. 3710, pp. 1301–1312.

[4] I. J. Chant and A. R. Rye, "Overview of current radar land mine detection research at the Defence Science and Technology Organisation, Salisbury, South Australia," in *Proceedings of IEE Conference on the Detection of Abandoned Land Mines*, Edinburgh, U.K., October 1996.

[5] S. Peleg and B. Friedlander, "The disctrete polynomial-phase transform," *IEEE Transactions on Signal Processing*, 1995.

[6] B. Efron, "Bootstrap methods: another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[7] A. M. Zoubir and B. Boashash, "The bootstrap and its application in signal processing," *IEEE Signal Processing Magazine*, vol. 15, no. 1, pp. 56–76, January 1998.

[8] D. N. Politis, "Computer-intensive methods in statistical analysis," *IEEE Signal Processing Magazine*, vol. 15, no. 1, pp. 39–55, 1998.

[9] A. M. Zoubir, B. Barkat, and C. L. Brown, "Comparison of signal processing techniques to the detection of landmines using Ground Penetrating Radar," Tech. Rep. 2/99-00, Curtin University of Technology/DSTO Project, Australia, 2000.

[10] H. Brunzell and A. M. Zoubir, "Multiple test procedures for radar-based detection of buried landmines," in *Proceedings of 33rd Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, USA, October 1999, vol. 2, pp. 831–834.

[11] A. M. Zoubir, F. Schulz, and C. L. Brown, "Ground penetrating radar target classification using time-frequency analysis," in *Proceedings of the Second Australian-American Joint Conference on the Technologies of Mine Countermeasures, MINWARA 2001*, Sydney, Australia, March 2001, The Mine Warfare Association.

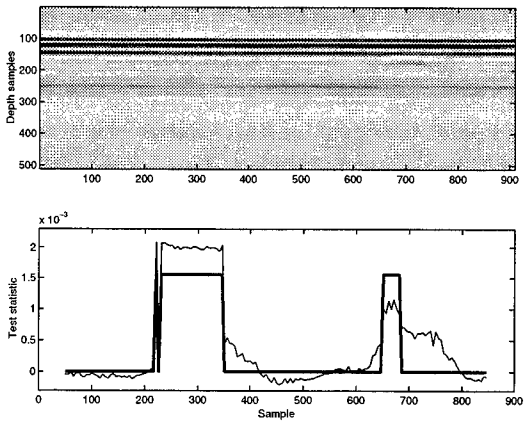


Fig. 6. Top: B-scan with ST-AP(1) and SS05x05 targets buried at 5 cm in clay. Bottom: Corresponding test statistics (thin line) and detection decision (thick line).

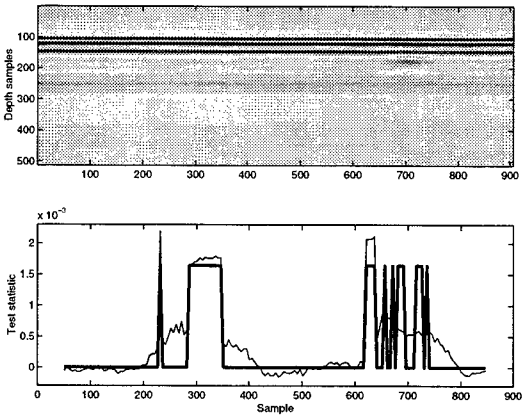


Fig. 7. Top: B-scan with ST-AP(2) target (parallel orientation) and an aluminium soft-drink can buried at 5 cm in clay. Bottom: Corresponding test statistics (thin line) and detection decision (thick line).

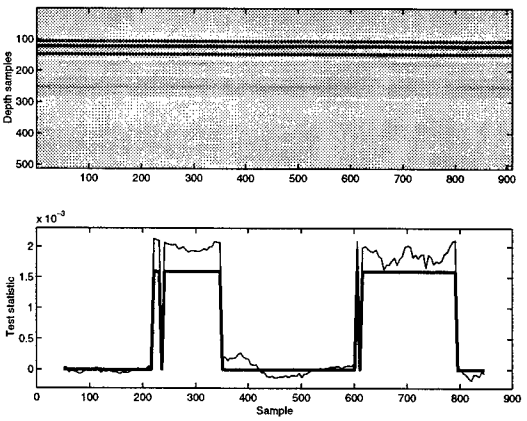


Fig. 8. Top: B-scan with ST-AP(3) target (perpendicular orientation) and shrapnel buried at 5 cm in clay. Bottom: Corresponding test statistics (thin line) and detection decision (thick line).

LAND MINE DETECTION IN ROTATIONALLY INVARIANT NOISE FIELDS

LENNART SVENSSON AND MAGNUS LUNDBERG

Department of Signals and Systems
Chalmers University of Technology
E-mail: {lennart.svensson,mlg}@s2.chalmers.se

ABSTRACT

This paper proposes a method to detect infrared land mine signatures embedded in rotationally invariant colored noise. A common problem in statistical image processing is high dimensionality. This causes a need for large sets of training data. To overcome this, an alternative formulation of the Generalized Likelihood Ratio Test (GLRT) is presented. This formulation makes it possible to utilize the circular-symmetry, rendering a substantial decrease in model dimensionality and consequently, in the amount of training data needed. Simulations indicate that a significant gain in performance can be achieved compared to both the non-parameterized detector and the matched filter.

1. INTRODUCTION

The presence of land mines is one of the worst environmental problems that faces humanity. Each year, 10000 people are killed and 30000 are injured in mine related accidents. Traditional techniques to detect and remove buried mines are both dangerous and time consuming, urging the need for more effective methods. One of the emerging techniques that has gained the most attention is infrared imaging [1]. Detecting buried mines using infrared imaging is possible since a buried object will interfere with the natural heat and mass transfer constantly taking place in the soil and at the surface. The result is a thermal signature at the surface that may be detected by an infrared imaging system.

The thermal signature will be embedded in noise caused by fluctuations in the soil structure and the surface. In order to design a detector it is necessary to model the characteristics of the noise. Such a model should be accurate enough to incorporate the vital features of the noise, while still simple enough to facilitate implementation. In practice, the distribution of the noise is not known, but may be estimated from off-line data. Assuming a Gaussian distribution, the issue reduces to finding the second order statistics.

The main problem when utilizing covariance-based methods in image processing is that the dimensionality is often very high. As a result the number of parameters that are to be estimated is considerable, requiring large sets of training data and memory expensive algorithms. An alternative approach is to use some basic features of the noise enabling parameterization of the covariance by means of a small set of parameters. Assuming spatial stationarity, one such approach is to model the colored noise as an autoregressive process [2]. This has previously been employed for land mine detection, see [5]. In addition to stationarity, it would also be desirable to exploit that the noise can be considered rotationally invariant for many relevant backgrounds such as gravel roads and sand. Further, for cases where there exists view angle dependence, the measured image can still have a rotationally invariant probability density function (pdf) if the camera capturing the surface has an unknown angle. Rotational invariance has been studied previously, mainly for the purpose of texture classification, see for instance [3]. To incorporate a parameterization of the covariance matrix that models the circular symmetry is a difficult task. To circumvent this problem, we present a reformulation of the detector where we train the linear detector directly.

2. BASIC ASSUMPTIONS

Let $s(x, y)$ denote an infrared image defined over an area of interest. The image is measured at the coordinates (x, y), where $x = -G, \ldots, G$, $y = -G, \ldots, G$, and the values are stacked into the column vector $\mathbf{s}$. Further, let the image, $s(x, y)$, be an outcome of the stochastic image, $\mathcal{S}(x, y)$, so that $\mathbf{s}$, is a realization of the stochastic vector $\mathbf{S}$.

The problem is to decide whether or not there is a buried land mine at the center of the image. Therefore, the detection problem, assuming additive noise, is to

* This work was supported by *the Swedish Research Council for Engineering Sciences* (TFR) and *the Swedish Defense Research Agency* (FOI).

0-7803-7011-2/01/$10.00 ©2001 IEEE

distinguish between the two hypotheses

$$\begin{aligned} \mathcal{H}_0: &\quad \mathbf{s} = \mathbf{n} \\ \mathcal{H}_1: &\quad \mathbf{s} = \mathbf{n} + \theta\,\mathbf{m}. \end{aligned} \tag{1}$$

Here, the noise vector, $\mathbf{n}$, is the sample vector of the background noise image, $n(x,y)$. As for the measured image, $n(x,y)$ is an outcome of the stochastic image $\mathcal{N}(x,y)$, while $\mathbf{n}$ is an outcome of the stochastic vector $\mathbf{N}$. The stochastic image, $\mathcal{N}(x,y)$, is assumed to have a rotationally invariant pdf, and $\mathbf{N}$ is assumed to be zero mean, Gaussian distributed with covariance matrix $\mathbf{R}$. The matrix $\mathbf{R}$ is regarded as unknown and deterministic. Moreover, the shape of the mine signature, $\mathbf{m}$, is the sample vector of the circularly symmetric image $m(x,y)$. The vector, $\mathbf{m}$, is assumed to be known, and of unit length ($\mathbf{m}^T\mathbf{m} = 1$). Finally, the magnitude, θ, is assumed to be a non-negative, but unknown, deterministic constant.

The Generalized Likelihood Ratio Test (GLRT) for the detection problem (1) is

$$\frac{f\left(\mathbf{s};\hat{\theta}^{ML},\mathcal{H}_1\right)}{f\left(\mathbf{s};\mathcal{H}_0\right)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma_1,$$

where $\hat{\theta}^{ML}$ is the Maximum Likelihood (ML) estimate of the parameter θ (assuming $\mathcal{H}_1$). For Gaussian noise, this test is equivalent to the test

$$\hat{\theta}^{ML} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{2}$$

for some constant γ. Unfortunately, the ML estimate can not be found since the covariance matrix of the noise is unknown. Therefore, we search for other estimates to be used in place of $\hat{\theta}^{ML}$. To our aid we have K noise samples (images) $\mathbf{n}_k$, $k = 1,2,\ldots,K$, that provide information about the noise distribution.

3. ESTIMATING THE AMPLITUDE, θ

We study three common estimators of θ (assuming $\mathcal{H}_1$) and propose a new estimator that utilizes the rotational invariance.

3.1. Known Covariance Matrix, R

This is the well known case when the noise is Gaussian with a known covariance matrix. The ML estimate

$$\hat{\theta}^{ML} = \frac{\mathbf{m}^T\mathbf{R}^{-1}\mathbf{s}}{\mathbf{m}^T\mathbf{R}^{-1}\mathbf{m}}, \tag{3}$$

can be easily derived from the definitions, see e.g. [4]. Although unable to be used in practice, it is included for comparison.

3.2. Unknown Covariance Matrix, R

Although the covariance matrix is unknown, the estimator structure in (3) can still be used if training data is used to estimate $\mathbf{R}$.

An unbiased, and commonly used, estimate of the covariance matrix, $\mathbf{R}$, is

$$\hat{\mathbf{R}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{n}_k\left(\mathbf{n}_k\right)^T.$$

Using this estimate instead of the true covariance matrix in (3), we obtain an alternative estimator of θ

$$\hat{\theta}^{\hat{R}} = \frac{\mathbf{m}^T\hat{\mathbf{R}}^{-1}\mathbf{s}}{\mathbf{m}\hat{\mathbf{R}}^{-1}\mathbf{m}}, \tag{4}$$

which is close to $\hat{\theta}^{ML}$ whenever $\hat{\mathbf{R}}$ is a good approximation of $\mathbf{R}$. The problem is that the number of parameters in $\mathbf{R}$ is very large (approximately $(2G+1)^4/2$ for an image of size $(2G+1)\times(2G+1)$). Therefore, to estimate it accurately a large number of training images is needed. It is also hard to utilize the circular symmetry in the model to reduce the number of parameters.

3.3. Matched Filter

The matched filter is the most frequently used estimator (and thus detector). It has the advantage of being both simple and relatively robust, though not optimal in general. Assuming $\mathcal{H}_1$, the matched filter estimate is (since $\mathbf{m}$ is of unit length)

$$\hat{\theta}^{MF} = \mathbf{m}^T\mathbf{s}. \tag{5}$$

This is the ML estimate if the noise is white, i.e. if $\mathbf{R} \propto \mathbf{I}$ ($\mathbf{I}$ represents the identity matrix), and we will argue in Section 3.4 that it is the best that can be done if we have an unknown covariance matrix and no training data.

3.4. Proposed Estimator

Again, we consider the case where the covariance matrix is unknown, and therefore the ML-estimate is non-trivial to calculate. As in Section 3.2, we use training data to estimate the unknown noise parameters, but here we search for a new formulation of the ML estimate, for which, in contrast to (4), the circular symmetry can be utilized to reduce the number of unknown parameters.

For notation, let Θ_n denote the amplitude of the stochastic noise in the mine direction ($\Theta_n = \mathbf{m}^T\mathbf{N}$), and let θ_n be an outcome of Θ_n, ($\theta_n = \mathbf{m}^T\mathbf{n}$). Also, let $\Pi^{\perp}$ be the projection matrix onto the orthogonal

complement of $\mathbf{m}$[1]. Then the measured image, $\mathbf{s}$, can be decomposed as

$$\mathbf{s} = \theta \cdot \mathbf{m} + \mathbf{n} = (\theta + \theta_n)\,\mathbf{m} + \Pi^{\perp}\mathbf{n}.$$

Therefore, as $\mathbf{m}^T\mathbf{s} = \theta + \theta_n$, a reasonable estimate of θ is

$$\hat{\theta} = \mathbf{m}^T\mathbf{s} - \arg\max_{\theta_n} f_{\Theta_n|\Pi^{\perp}\mathbf{N}}(\theta_n|\Pi^{\perp}\mathbf{n}). \qquad (6)$$

In order to prove that this is indeed the ML estimate, we state the following theorem:

Theorem 1 *Suppose we observe*

$$\mathbf{X} = \mathbf{M}_0(\theta + \mathbf{N}_0) + \mathbf{M}_1\mathbf{N}_1$$

where $\mathbf{M}_0$ *and* $\mathbf{M}_1$ *are matrices of size* $q \times p_0$ *and* $q \times p_1$ *respectively, such that* $\mathbf{M}_0^T\mathbf{M}_0 = \mathbf{I}$, $\mathbf{M}_1^T\mathbf{M}_0 = \mathbf{0}$ *and* $\mathbf{M}_1^T\mathbf{M}_1 = \mathbf{I}$. *Furthermore,* $\mathbf{N}_0$ *and* $\mathbf{N}_1$ *are stochastic vectors of size* $p_0 \times 1$ *and* $p_1 \times 1$ *respectively. Then the ML estimate of* θ *is*

$$\hat{\theta}^{ML} = \mathbf{M}_0^T\mathbf{x} - \arg\max_{\mathbf{n}_0} f_{\mathbf{N}_0|\mathbf{N}_1}(\mathbf{n}_0|\mathbf{M}_1^T\mathbf{x}),$$

where $\mathbf{x}$ *is the outcome of the stochastic vector* $\mathbf{X}$.

Proof:

$$\begin{aligned}
\hat{\theta}^{ML} &= \arg\max_{\theta} f_{\mathbf{X}}(\mathbf{x};\theta) \\
&= \arg\max_{\theta} f_{\mathbf{N}_0,\mathbf{N}_1}(\mathbf{M}_0^T\mathbf{x} - \theta, \mathbf{M}_1^T\mathbf{x}) \\
&= \arg\max_{\theta} f_{\mathbf{N}_0|\mathbf{N}_1}(\mathbf{M}_0^T\mathbf{x} - \theta|\mathbf{M}_1^T\mathbf{x}) \\
&= \mathbf{M}_0^T\mathbf{x} - \arg\max_{\mathbf{n}_0} f_{\mathbf{N}_0|\mathbf{N}_1}(\mathbf{n}_0|\mathbf{M}_1^T\mathbf{x})
\end{aligned}$$

■

Consequently, the ML estimate is obtained by correcting the matched filter estimate, $\mathbf{m}^T\mathbf{s}$, with an estimate of the noise contribution, θ_n, and (6) is indeed the ML estimate.

The problem then reduces to estimating θ_n given $\Pi^{\perp}\mathbf{n}$, i.e. $\arg\max_{\theta_n} f_{\Theta_n|\Pi^{\perp}\mathbf{N}}(\theta_n|\Pi^{\perp}\mathbf{n})$. Obviously, if we have no training data, i.e. no information about $\mathbf{R}$, the best estimate is zero (as Θ_n is zero mean). In other words, without training data we can do no better than the matched filter.

In order to derive a closed form expression for the ML estimate, we note that for a jointly Gaussian distribution we have

$$\arg\max_{\theta_n} f_{\Theta_n|\Pi^{\perp}\mathbf{N}}(\theta_n|\Pi^{\perp}\mathbf{n}) = \mathbf{a}^T\Pi^{\perp}\mathbf{n} = \mathbf{a}^T\Pi^{\perp}\mathbf{s}$$

for some vector $\mathbf{a}$. This in turn yields the estimate as

$$\hat{\theta}^{ML} = \left(\mathbf{m}^T - \mathbf{a}^T\Pi^{\perp}\right)\mathbf{s} \qquad (7)$$

where, from (3),

$$\mathbf{a} = \mathbf{m} - \frac{\mathbf{R}^{-1}\mathbf{m}}{\mathbf{m}^T\mathbf{R}^{-1}\mathbf{m}}.$$

The advantage of this formulation is that it enables us to invoke a parameterization that utilizes the circular symmetry to reduce the number of parameters.

According to the assumption that the pdf of $\mathcal{N}(x,y)$ is rotationally invariant, and that $m(x,y)$ is circular symmetric, the estimate

$$\hat{\theta}_n = \mathbf{a}^T\Pi^{\perp}\mathbf{n}$$

should be invariant to rotations of $n(x,y)$.

For this condition to be (at least almost) fulfilled, $\mathbf{a}$ should be the sample vector of some circular symmetric continuous image $a(x,y)$. If the images are sampled sufficiently often compared to their frequency content, the following Fourier-series like expansion is adequate

$$a(x,y) = \sum_{k=0}^{G} \alpha_k \cdot c_k(x,y)$$

where

$$c_k(x,y) = \begin{cases} \cos\left(\frac{2\pi k\sqrt{x^2+y^2}}{2N}\right) & , \sqrt{x^2+y^2} \le G \\ 0 & , \sqrt{x^2+y^2} > G. \end{cases}$$

Hence, we can find a linear parameterization using the unknown vector $\mathbf{d}$ as

$$\mathbf{a} = \mathbf{C}\mathbf{d},$$

where the k'th column of the matrix $\mathbf{C}$ is the sample vector of $c_{k-1}(x,y)$, and the k'th element of the vector $\mathbf{d}$ is α_{k-1}. Thus,

$$\hat{\theta}^{ML} = \left(\mathbf{m} - \mathbf{d}^T\mathbf{C}^T\Pi^{\perp}\right)\mathbf{s} \qquad (8)$$

and we have found a representation where the $G+1$ dimensional vector $\mathbf{d}$ is all that is to be estimated from training data. We have therefore reduced the number of parameters to $G+1$. To estimate $\mathbf{d}$, we utilize some measured noise samples $\mathbf{n}_k$, $k = 1, 2, \ldots, K$, and solve the least square problem

$$\hat{\mathbf{d}} = \arg\min_{\mathbf{d}} \sum_{k=1}^{K} \left(\mathbf{m}^T\mathbf{n}_k - \left(\Pi^{\perp}\mathbf{n}_k\right)^T\mathbf{C}\mathbf{d}\right)^2.$$

Using this estimate in (8), we obtain the following estimate for θ

$$\hat{\theta}^{\hat{d}} = \left(\mathbf{m} - \hat{\mathbf{d}}^T\mathbf{C}^T\Pi^{\perp}\right)\mathbf{s}. \qquad (9)$$

[1] $\Pi^{\perp} = \mathbf{I} - \mathbf{m}\mathbf{m}^T$

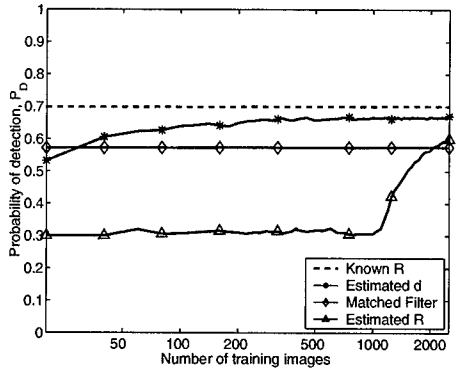


Figure 1: The performance of the different detectors for $P_{FA} = 0.3$, given the number of training data

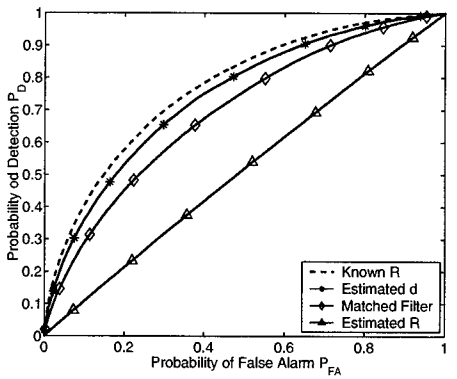


Figure 2: ROC for the four detectors using 240 training images

4. SIMULATION RESULTS

In order to evaluate the proposed approach, we use the detector as given by (2). In particular, we compare the detector performance when using the proposed estimator (8), here denoted *Estimated d*, to that of employing the three other estimates described in Section 3. The test employing the estimate of the full covariance matrix, $\mathbf{R}$, as given by (4) will be denoted *Estimated R*, while the one using the correlation estimate given by (5) will be denoted *Matched Filter*. To serve as a reference, although not possible to implement in reality, we also include the performance of the detector utilizing a known covariance matrix, here denoted *Known R*.

In the simulations we considered images of size 33×33. The background noise was created by passing white Gaussian noise through an FIR filter with low-pass characteristics. The rotational invariance was ensured by using an FIR filter with a circularly symmetric impulse response. Moreover, the mine signature had a radius of approximately 5 pixels. The signature is chosen as a smoothed version of the top view shape of a cylindrical shaped mine, see [5]. As both the target and the noise have low-pass characteristics, the detection problem is especially difficult.

The main contribution in this paper is a new formulation of the ML estimate, that greatly reduces the number of unknown parameters compared to when the full covariance matrix is estimated. Therefore, in Figure 1 we study how the probability of detection, P_D, depends on the number of training images, for a given probability of false alarm, $P_{FA} = 0.3$. Furthermore, to illustrate how the detectors perform for different probabilities of false alarm, we plot the *receiver operating characteristics* (ROC) using 240 training images in Figure 2. From the figures, it can be seen that the performance of the methods which estimate either d or R improves with the number of training images, whereas, of course, the performance of the *Known R* and the *Matched Filter* detectors do not depend on the amount of training data. More importantly, as expected, the proposed detector only needs a fraction of the number of training data as compared to the method that estimates the full covariance matrix. In particular, the proposed scheme only needs a few training images to outperform the detector employing the matched filter estimate. It is also interesting to note that for fewer than 1000 training images *Estimated R* does not seem to perform any better than a detector that totally disregards all measurements, i.e. $P_D = P_{FA}$.

5. CONCLUSIONS

In this paper we proposed a detector that takes advantage of the circular symmetry to drastically reduce the number of unknown parameters in the detector.

Simulations confirm that the need for training data is greatly reduced, and that the performance is improved for reasonable amounts of training data.

6. REFERENCES

[1] N. Del Grande. "Temperature Evaluated Mine Position Survey (TEMPS), Application of Dual-Band Infrared Methodology". In *Proc. IRIA IRIS*, Baltimore, MD, March 1990.

[2] A. Isaksson. "Frequency Domain Accuracy of Identified 2-D Causal Models". In *Proc. IEEE ICASSP 89*, pages 2294–2297, May 1989.

[3] R.L. Kashyap and A. Khotanzad. "A Model-Based Method for Rotation Invariant Texture Classification". *PAMI*, 8(4):472–481, July 1986.

[4] S.M. Kay. *"Fundamentals of Statistical Signal Processing: Estimation Theory"*. Prentice Hall International, Inc., 1993.

[5] M. Lundberg. "Infrared Land Mine Detection by Parametric Modeling". In *Proc. IEEE ICASSP 2001*, Salt Lake City, UT, May 2001.

UNKNOWN SIGNAL DETECTION VIA ATOMIC DECOMPOSITION

Gustavo López-Risueño, Jesús Grajal

Departamento de Señales, Sistemas y Radiocomunicaciones
ETSIT, Universidad Politécnica de Madrid
Ciudad Universitaria s/n, 28040 Madrid, Spain
email:risueno@gmr.ssr.upm.es

ABSTRACT

A detector using atomic decomposition with a chirplet dictionary is analyzed. It is derived from the generalized likelihood ratio test and has constant false alarm rate. The atomic decomposition is performed via a genetic algorithm and compared with previous approaches.

1. INTRODUCTION

The atomic decomposition [1], also known as matching pursuit [2] or adaptive Gabor representation [3], is an adaptive approximation technique providing a sparse, flexible, and physically meaningful representation of the signals. In spite of its suitability for the analysis of unknown signals [2], the detection of signals in noise via the atomic decomposition (AD) has not been stated yet. In this paper, we present an AD-based detector designed according to the generalized likelihood ratio test (GLRT), and analyze its performance in the Neyman-Pearson sense. We also prove its constant false alarm rate (CFAR) characteristic in zero-mean, complex, white, Gaussian noise (CWGN).

The AD sparseness is attained by a highly redundant dictionary of unit energy signals, called atoms. Let $D = \{h_{\underline{\gamma}}(n)\}$ be a dictionary of atoms, and $s(n)$ the signal under analysis, the AD is obtained as

$$\underline{\gamma}_p = \arg\max_{\underline{\gamma}} \left| \langle s_{p-1}(n), h_{\underline{\gamma}}(n) \rangle \right|^2 , \tag{1}$$

and

$$b_p \, e^{j\phi_p} = \left\langle s_{p-1}(n), h_{\underline{\gamma}_p}(n) \right\rangle . \tag{2}$$

$s_p(n)$ comes from

$$s_p(n) = s_{p-1}(n) - b_p \, e^{j\phi_p} \, h_{\underline{\gamma}_p}(n), \; p > 0 , \tag{3}$$

$$s_0(n) = s(n) . \tag{4}$$

The signal under analysis is approximated by

$$s(n) \approx \sum_p b_p \, e^{j\phi_p} \, h_{\underline{\gamma}_p}(n) \qquad p = 1, 2 \ldots . \tag{5}$$

As in [1, 4, 5], we employ a dictionary of chirplets, i.e. chirped Gabor functions, because linear frequency modulation is a very common feature of man-made signals and natural signals. In the following, it is assumed that the signal under analysis is a weighted sum of chirplets. The chirplet dictionary is parameterized through the 4-component vector:

$$\underline{\gamma} = [\alpha, \beta, T, f]^t , \tag{6}$$

and every chirplet is defined as[1]

$$h_{\underline{\gamma}}(n) = \left(\frac{\alpha}{\pi}\right)^{1/4} e^{-\frac{\alpha}{2}(n-T)^2} \cdot e^{j\left[2\pi f(n-T)+\pi\beta(n-T)^2\right]} . \tag{7}$$

Hence, every extracted atom is defined by means of the 6-component vector

$$\left[b_p, \underline{\gamma}_p^t, \phi_p\right]^t , \tag{8}$$

with b_p a positive real number, and b_p^2 the energy of the pth extracted atom.

2. AD USING A GENETIC ALGORITHM

The optimization procedure for (1) has to be carefully chosen because of the extremely complex structure of the objective function, with multiple local optima coming from the existence of noise and multi-component signals, and domain regions where it is nearly constant. Therefore, global search algorithms refined by descendent techniques are the most suitable strategies.

We use a genetic algorithm (GA) refined with a downhill simplex method [6]. The selected GA is the most popular described in the literature [7]. A detailed description of its parameter values is given in Table 1. The probability of crossover has been fixed to 1 and the population size to 200 in order to reduce the premature convergence to local optima [7]. The search range for the components of vector (6) is application dependent. The rest of GA parameters presents typical values [7].

Other authors propose different algorithms to optimize (1). In Table 2, we show the average time employed to find the first atom of an atomic decomposition using different approaches and their complexity. GAAD is our algorithm, and TFAD64 and TFAD512 are approaches by O'Neill and Flandrin using the ambiguity function with resolution 64 and 512 respectively [1][2]. GMP is a version of [5] with a subdictionary of more than 20000 signals. N is the number of signal samples ($N = 1024$). As can be noticed, the GAAD complexity is linear with regard to N. On the other hand,

This work was supported by the National Board of Scientific and Technology Research (CICYT) under project TIC-99-1172-C02-01/02, and by a FPU fellowship of the Ministry of Education.

[1] It is assumed that the sampling rate is one and the chirplets are time- and band-limited before sampling.

[2] Their MATLAB programs are available at *http://mdsp.bu.edu/jeffo*

GA parameter	Values
Population size	200
Number of generations	20
Number of enconding bits	16
Probability of crossover	1
Probability of mutation	0.03
Search range for α	$[10^{-6}, 10^{-1}]$
Search range for β	$[-0.1, 0.1]$
Search range for T	$[0, 1024]$
Search range for f	$[-0.5, 0.5]$

Table 1. Genetic algorithm parameters used in the simulations.

TFAD64 is the most efficient. However, it can suffer from worse detection performance than the GAAD, as will be shown. The MATLAB simulations were run on a Pentium II, 350 MHz, 256 MB RAM, with Windows 98.

Algorithm	Time in seconds	Complexity
GAAD	31	$O(N)$
TFAD512	40	$O(N\ \log_2(N))$
TFAD64	5	$O(N\ \log_2(N))$
GMP	150	$O(N^2\ \log_2(N))$

Table 2. Computation time and complexity of the first atomic extraction for several AD techniques.

3. DETECTION OF A CHIRPLET IN NOISE USING ATOMIC DECOMPOSITION

In this section, the detection of a chirplet in noise using the atomic decomposition is related to the GLRT. The detection is formulated in terms of a binary hypothesis test where the null and alternative hypothesis are:

$$H_0 \ : \ x(n) = r(n),\ n = 1, ...,\ N \tag{9}$$

$$H_1 \ : \ x(n) = b\, e^{j\phi}\, h_{\underline{\gamma}}(n) + r(n),\ n = 1, ...,\ N, \tag{10}$$

$r(n)$ is a CWGN with power σ^2. b is a positive real number, b^2 is the chirplet energy, and N is the number of samples. Using vector notation, the probability density function (pdf) under H_0 becomes

$$f_{H_0}(\underline{x}; \sigma) = \frac{1}{(\pi\, \sigma^2)^N}\ \exp - \frac{||\underline{x}||^2}{\sigma^2}\,, \tag{11}$$

and under H_1

$$f_{H_1}(\underline{x}; \sigma, b, \underline{\gamma}^t, \phi) = \frac{1}{(\pi\, \sigma^2)^N}\ \exp - \frac{\left|\left|\underline{x} - b\, e^{j\phi}\, \underline{h}_{\underline{\gamma}}\right|\right|^2}{\sigma^2}\,. \tag{12}$$

To evaluate the GLRT, the maximum likelihood estimate (MLE) under H_1 and H_0 must be calculated. Under H_1 the MLE is the first extracted atom of the atomic decomposition [1]. Namely, it leads to equations (1) and (2) for the parameters of the first extracted atom. Regarding σ^2, we can estimate it as

$$\widehat{\sigma}^2_{sn} = \frac{||\underline{x}||^2 - \widehat{b}^2}{N}\,. \tag{13}$$

In the following, we refer to the extracted atom parameters as $\widehat{b}$, $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{T}$, $\widehat{f}$, and $\widehat{\phi}$ to stress that they are estimates of the parameters of vector (8). For H_0, the MLE of σ^2, $\widehat{\sigma}^2_{on}$, becomes

$$\widehat{\sigma}^2_{on} = \frac{||\underline{x}||^2}{N}\,. \tag{14}$$

After some manipulations, the GLRT turns

$$L_{GLR}(\underline{x}) = \frac{\widehat{b}^2}{\widehat{\sigma}^2_{on}} \underset{H_0}{\overset{H_1}{\gtrless}} Th\,, \tag{15}$$

where Th is the threshold. Thus, for the one-chirplet-in-noise case the GLRT depends on the energy of the first extracted atom, $\widehat{b}^2$, and the noise power estimate $\widehat{\sigma}^2_{on}$.

3.1. Analytic model

The detector (15) allows an approximate analytic treatment if independence between the atom energy and the noise power estimates is assumed. Under H_0, $\widehat{\sigma}^2_{on}$ becomes a chi-square random variable of $2N$ degrees of freedom. Namely, its pdf is

$$f_{\widehat{\sigma}^2_{on}}(z) = \frac{N^N}{(N-1)!\ (\sigma^2)^N}\ z^{N-1} \exp(-\frac{N\, z}{\sigma^2}),\ z > 0\,. \tag{16}$$

On the other hand, it has been found through Monte Carlo analysis that $\widehat{b}^2$ has a lognormal distribution under H_0, i.e.

$$f_{\widehat{b}^2}(z; \mu_n, \sigma_n) = \frac{1}{\sqrt{2\pi\sigma_n^2 z}}\ \exp\left(-\frac{(\ln(z) - \mu_n)^2}{2\sigma_n^2}\right), z > 0\,, \tag{17}$$

This distribution gets a high significance level in the composite chi-square goodness-of-fit test. The lognormal model is valid for the GA of section 2 varying the population size, the number of generations, and the noise power. Parameters μ_n and σ_n are estimated using the maximum likelihood criterion. For the first extracted atom, their estimates follow

$$\widehat{\mu}_n = \ln(\sigma^2) + [1.9037 + 0.0104\ \ln(ngen) + 0.1050\ \ln(psize)]\,, \tag{18}$$

$$\widehat{\sigma}_n = -0.0239 \ln(ngen) - 0.0281 \ln(psize) + 0.4129\,. \tag{19}$$

$psize$ and $ngen$ are the GA population size and number of generations, respectively. As can be noticed, $\widehat{\sigma_n}$ does not depend on the noise power. Then, the probability of false alarm, Pfa, for a given threshold Th is expressed by

$$Pfa = \int\!\!\int_{\{y\ >Th\cdot z\}} f_{\widehat{b}^2}(y) \cdot f_{\widehat{\sigma}^2_{on}}(z)\, dydz\,, \tag{20}$$

Integral (20) can be calculated by numerical methods. It can be proved that it does not depend on the noise power, therefore, the GLRT has CFAR characteristic with respect to the noise power. Fig. 1 shows the Pfa curve versus the threshold Th obtained through simulations and using the analytic approximation (20). As can be appreciated, the agreement is very good. Mismatches are due to the fact that expressions (18) and (19) are obtained by fitting the lognormal model parameters within the range of the studied GA parameters, and due to the variance of the Pfa estimation by Monte Carlo analysis.

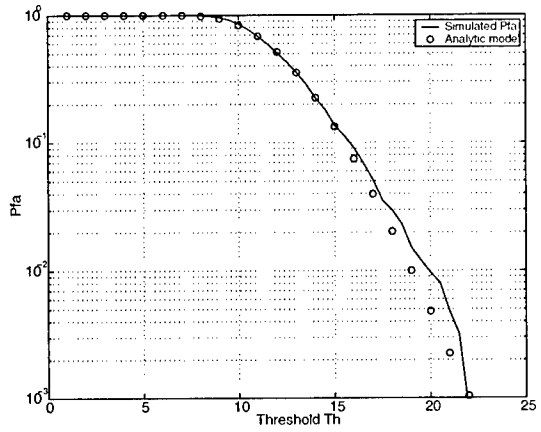


Fig. 1. Pfa simulated and calculated through the analytic approximation. GA population size of 200 and 20 generations. 1000 Monte Carlo trials.

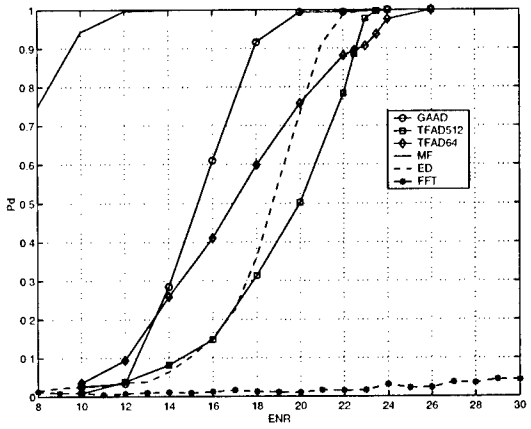


Fig. 2. Performance of the GLRT using GAAD, TFAD64, TFAD512. Also MF, ED and FFT detectors are depicted. The chirplet features are: $\alpha = 0.001$, $\beta = 0.003$, $T = 500$ and $f = 0.25$. $Pfa = 10^{-2}$ and 1000 trials.

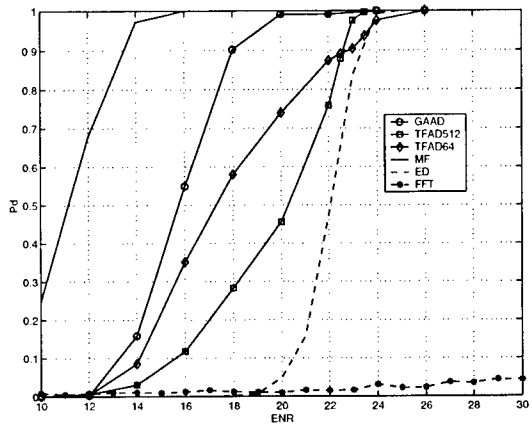


Fig. 3. Performance of the GLRT using GAAD, TFAD64, TFAD512. Also MF, ED and FFT detectors are depicted. The chirplet features are: $\alpha = 0.001$, $\beta = 0.003$, $T = 500$ and $f = 0.25$. $Pfa = 10^{-6}$ and 1000 trials.

It has been also checked that TFAD64 and TFAD512 follow the lognormal model in the only noise case. This fact allows to get the threshold required for low Pfa, as in the GAAD case.

In Figs. 2 and 3, probability of detection (Pd) curves of the GLRT using GAAD, TFAD64, TFAD512 are depicted for Pfa equal to 10^{-2} and 10^{-6} respectively, and a chirplet of features: $\alpha = 0.001$, $\beta = 0.003$, $T = 500$, and $f = 0.25$. 1000 trials have been employed. The greatest Pfa has been selected in order to check the agreement between our model and simulations. It has been seen that there is no difference in using the threshold from the model and from the simulations. The lowest Pfa is typical in radar applications, and the lognormal model is required to compute the threshold. The number of samples (N) is 1024. ENR means energy-to-noise power ratio, i.e. $ENR = 10\ \log_{10}\left(b^2/\sigma^2\right)$. It depends neither on the signal length, unlike the SNR in [1], nor on other chirplet features[3].

The performance of the energy detector (ED), a FFT of 1024 samples without windowing, and the matched filter (MF) are also depicted. The ED and the FFT represent classic techniques in signal detection. MF is the GLRT detector when the only unknown features of (10) are b and ϕ. In the ED and MF case it is assumed known noise power[4]. For $Pd = 90\%$, GAAD exhibits better performance than the TFADs. This indicates that the GA performs better in the search of the global optimum than the algorithm of TFADs. For the lowest Pfa, GAAD is approx. 4 dB worse than the MF and the ED is similar to the TFADs, although it does not provide chirplet feature estimation. Besides, TFAD512 is better than TFAD64, as expected due to its greater resolution. The extremely poor performance of the FFT is due to the chirplet chirp rate, i.e. the chirplet bandwidth is much greater than the FFT-based filter bandwidth.

Pd depends mainly on the chirplet α. If a longer chirplet is evaluated ($\alpha = 0.0001$, $\beta = 0$, $T = 400$, and $f = 0.25$) the sensitivity is degraded as Fig. 4 shows. In this case, the FFT is better than TFADs, although it does not give estimation of α. This is due to the fact that the chirplet is longer and does not have modulation. It behaves as a narrowband signal. It has been found that the GAAD performance strongly depends on the β search range (Table 1). Using an exploration range for β: $[-0.005, 0.005]$, the Pd becomes closer to TFADs behavior. GAAD with this modification is called GAAD2. For GAAD2, the lognormal model is still valid to describe the noise statistics. In Fig 3, an improvement of less than 1 dB would be obtained if GAAD2 were used instead of GAAD.

4. DETECTION OF MULTIPLE CHIRPLETS

Using the atomic decomposition, we propose a sequential detector consisting of a unitary decision test, i.e. the one-chirplet-case GLRT (15), for every extracted atom. When an extracted atom has an energy-to-noise power ratio estimate greater than the thresh-

[3] SNR in [1] is defined as $SNR = 10\ \log_{10}\frac{b^2}{N\sigma^2}$. It would be more suitable to use the chirplet mean power, i.e. $SNR = 10\ \log_{10}(\frac{b^2\sqrt{2\alpha}}{\sigma^2})$.

[4] The difference between known and unknown σ^2 for the GAAD and the TFADs has resulted in less than 0.3 dB for a fixed Pd and 1024 samples.

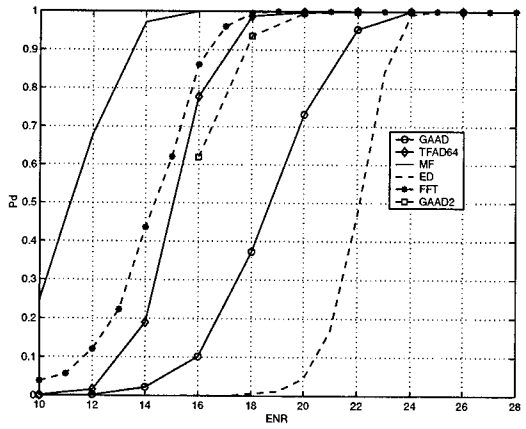


Fig. 4. Performance of the GLRT using GAAD, TFAD64, TFAD512. Also MF, ED and FFT detectors are depicted for a longer chirplet. GAAD2 means GAAD with a modified β range. The chirplet features are: $\alpha = 0.0001$, $\beta = 0.0$, $T = 400$ and $f = 0.25$. $Pfa = 10^{-6}$ and 1000 trials.

old, it is considered as one of the components forming the signal under analysis. Otherwise, it comes from noise. The threshold is obtained through the expressions of the one chirplet case. As an example, a 6-chirplet signal of 1024 samples is analyzed by GAAD. Chirplets 1 to 4 have $\alpha = 10^{-3}$, $\beta = 0$, $f = 0.0714$, and T equal to 150, 350, 600, and 800, respectively. Chirplets 5 and 6 share $T = 500$ and $f = 0.25$. In the case of chirplet 5, $\alpha = 10^{-3}$ and $\beta = 0.003$. For chirplet 6, $\alpha = 10^{-4}$ and $\beta = -0.0001$. All of them have $ENR = 18\ dB$ [5]. Fig. 5 shows the adaptive spectrogram (AS) [3] obtained by GAAD after detection. The chirplets are rightly recovered as Table 3 shows. The root mean square (RMSE) of the chirplet parameter estimation is shown in Table 4. $\widehat{\alpha}$ of the 6th chirplet is poorly estimated due to its low probability of detection at $ENR = 18\ dB$ [6].

5. CONCLUSIONS

The theoretical framework of a detector using AD, which is the GLRT in the one chirplet case, has been proposed, and its performance has been studied regarding classic techniques and the matched filter. Additionally, the advantages of the use of a GA for computing the AD have been pointed out with regard to other previous AD approaches in terms of complexity, efficiency and detection performance. For this algorithm, a useful statistical model of the pdf under the only-noise condition has been presented allowing an approximate analytic study of the detector. The detector dependency on the GA search range and signal characteristics has been shown through several examples.

6. REFERENCES

[1] J. O'Neill and P. Flandrin, "Chirp Hunting," in *IEEE International Symposium on Time-Frequency and Time-Scale Analysis*, 1998.

[2] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[3] S. Qian and D. Chen, *Joint Time-Frequency Analysis. Methods and Applications*. Prentice Hall, 1996.

[4] J. O'Neill and P. Flandrin, "Cramer-Rao Bounds for Atomic Decomposition," in *IEEE ICASSP*, 1999.

[5] A. Bultan, "A Four-Parameter Atomic Decomposition of Chirplets," *IEEE Trans. on SP*, vol. 47, no. 3, 1999.

[6] W.H. Press et al., *Numerical Recipies in Fortran*. Cambridge University Press, 1992.

[7] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, 1996.

[5] Equivalent to $SNR = -12\ dB$ according to [1], or $SNR = 4.5\ dB$ if signal mean power is used.

[6] Using GAAD, it would be expected $Pd \approx 40\%$ if chirplet 6 were alone, as Fig. 4 shows.

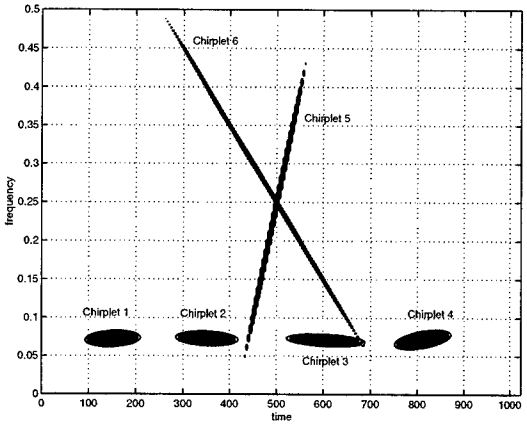


Fig. 5. AS of the multiple chirplet signal after GLRT detector using GAAD.

Chirplet	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{T}$	$\widehat{f}$
1	0.0010	0.0000	149.67	0.0714
2	0.0009	0.0000	348.74	0.0717
3	0.0009	0.0000	601.53	0.0714
4	0.0010	0.0000	800.38	0.0714
5	0.0014	0.0030	496.96	0.2498
6	0.0008	−0.0010	502.19	0.2476

Table 3. Chirplet parameters estimated by GAAD. 100 trials.

Chirplet	RMSE($\widehat{\alpha}$)	RMSE($\widehat{\beta}$)	RMSE($\widehat{T}$)	RMSE($\widehat{f}$)
1	0.0003	$4.85\ 10^{-5}$	4.83	0.0006
2	0.0003	$4.63\ 10^{-5}$	7.26	0.0011
3	0.0004	$4.63\ 10^{-5}$	5.41	0.0007
4	0.0003	$3.92\ 10^{-5}$	5.34	0.0007
5	0.0007	0.0001	3.28	0.01
6	0.0032	0.0001	29.51	0.0296

Table 4. Root mean square error of the chirplet parameters estimated by GAAD. 100 trials.

MULTIPATH DETECTION OF STOCHASTIC TRANSIENT PROCESSES

Francisco M. Garcia and Isabel M. G. Lourtie

ISR -Instituto de Sistemas e Robótica, IST - Instituto Superior Técnico
Torre Norte, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
E-mail : fmg@isr.ist.utl.pt

ABSTRACT

This paper reports on the detection of Gaussian stochastic transients in multipath environments described by random parameters. The solutions developed herein correspond to quadratic processors, with low computational cost and robust to changes in the statistical models of the channel. As a consequence, only a small amount of *a-priori* information is necessary to derive the parameters of the processor. A recursive form of the processor is also proposed, allowing for the recursive detection of the signal replicas arriving at the receiver.

1. INTRODUCTION

The classical solution for the detection of signals in a multipath environment is the generalized likelihood ratio test (GLRT), where the likelihood test is computed from estimates of the channel parameters in both hypotheses. In general, the estimation step in the GLRT is a heavy computational procedure. Furthermore, when the signals to detect are stochastic processes in low signal-to-noise ratio (SNR), which is the usual case in many passive detection applications, the variances of the estimates are large and the detector performance degrades. Recently, some authors have developed suboptimal processors to avoid the drawbacks of the GLRT. In [1], a suboptimal approximation for the detection of continuous-time, stationary processes in low SNR was proposed, assuming that the channel parameters are random variables. This processor was based on a Taylor series approximation of the likelihood ratio for the processor with known channel parameters. In [2], a geometric framework based on multiresolution techniques was proposed, where the set of all possible signals arriving at the receiver is approximated by a simpler linear subspace. The detectors proposed in [1] and [2] are both quadratic processors.

This paper extends the work of [1] to short-duration stochastic transients, which are nonstationary in nature. The following situation is considered: i) the processor is developed in discrete time; ii) the multipath channel is regarded as a "signal amplifier", instead of a nuisance; iii) the low SNR condition assumes that the signal eigenvalues are smaller than the noise ones. The solutions proposed have two major concerns. First, they must rely on a small *a-priori* amount of statistical information about the channel parameters. The basic idea is that this information should be easily inferred from local data (i.e., depth, salinity, temperature in an underwater media) and, when the local conditions change, the processor parameters must be recalculated fast. The simulation results show that the processors are robust to mismatches on the channel parameters, and thus only mild information about the range of the delay coefficients and approximate estimates of the attenuation coefficient means and variances are necessary. Second, the computational cost of the resulting processors must be low, allowing for real-time processing. Two possible processing structures are proposed: i) a quadratic form, which can be directly optimized in terms of a performance/computational complexity compromise using the methods proposed in [3]; ii) a recursive solution, where a sequence of tests are performed at the arrival of each signal replica. In most situations, this scheme reduces both the computational cost of the processor, and the mean time interval between the arrival of the first replica at the receiver and its detection time instant.

2. PROBLEM FORMULATION

The detection problem is formulated as a simple binary test. The channel is modeled such that the signal arriving at the receiver under hypothesis H_1 is a weighted sum of delayed replicas of the emitted signal. Thus, the observation process $r(t)$ is defined as

$$r(t) = \begin{cases} y(t) + n(t), & \text{under hypothesis } H_1 \\ n(t), & \text{under hypothesis } H_0, \end{cases} \tag{1}$$

where

$$y(t) = \alpha_1 s(t - T_0) + \sum_{k=2}^{N_q} \alpha_k s(t - T_0 - \tau_k) \tag{2}$$

$$\Leftrightarrow \qquad y(t + T_0) = \sum_{k=1}^{N_q} \alpha_k s(t - \tau_k), \qquad \text{with } \tau_1 = 0.$$

In (2), T_0, α_k and τ_k, $k = 1, \cdots, N_q$ denote, respectively, the time delay between the emission and the reception of the first replica of a signal $s(t)$, the attenuation coefficients (AC) and the delay coefficients (DC), where N_q represents the number of replicas arriving at the receiver. The emitted signal, $s(t)$, is a Gaussian, zero-mean transient with autocorrelation function $k_s(t_1, t_2)$. The noise, $n(t)$, is stationary, Gaussian distributed with zero-mean and with constant power spectrum up to a high frequency. The observation process is conveniently filtered and discretized at a sampling frequency T_s. For simplicity, we assume in the sequel that the sampled noise sequence is white, and that the DCs, τ_k, are integer multiples of T_s, i.e., $\tau_k = q_k T_s$, $k = 1, \cdots, N_q$. The optimal detector in this case corresponds to the likelihood test

$$\frac{E_{\boldsymbol{q},\boldsymbol{\alpha}}[p_{H_1}(\boldsymbol{r}|H_1, \boldsymbol{q}, \boldsymbol{\alpha})]}{E_{\boldsymbol{q},\boldsymbol{\alpha}}[p_{H_0}(\boldsymbol{r}|H_0, \boldsymbol{q}, \boldsymbol{\alpha})]} = E_{\boldsymbol{q},\boldsymbol{\alpha}}\left[\frac{p_{H_1}(\boldsymbol{r}|H_1, \boldsymbol{q}, \boldsymbol{\alpha})}{p_{H_0}(\boldsymbol{r}|H_0)}\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{3}$$

where $\boldsymbol{q} = \{q_1, \cdots, q_{N_q}\}$, $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_{N_q}\}$, $\boldsymbol{r} = [r(T_s + T_0) \cdots r(NT_s + T_0)]'$ corresponds to the observation vector over an interval of length N and $p_{H_i}(\boldsymbol{r}|H_i, \boldsymbol{q}, \boldsymbol{\alpha})$ represents the probability density function of $\boldsymbol{r}$ given $\boldsymbol{q}$ and $\boldsymbol{\alpha}$, under hypothesis H_i, $i = 0, 1$. It is assumed that the interval N is large enough to include all the replicas of an emitted signal arriving at the receiver.

3. LIKELIHOOD RATIO FOR KNOWN AC AND DC

Since $y(t)$ consists on a sum of zero-mean Gaussian-distributed signals, then the joint probability density function of $\boldsymbol{r}$ under both hypothesis is also Gaussian. Thus, the term inside brackets in (3) may be rewritten as

$$l(\boldsymbol{r}) = \frac{p_{H_1}(\boldsymbol{r}|H_1, \boldsymbol{q}, \boldsymbol{\alpha})}{p_{H_0}(\boldsymbol{r}|H_0)} = \sqrt{\frac{|\boldsymbol{C}_{H_0}|}{|\boldsymbol{C}_{H_1}|}} \exp\left(-\frac{1}{2}\boldsymbol{r}'[\boldsymbol{C}_{H_1}^{-1} - \boldsymbol{C}_{H_0}^{-1}]\boldsymbol{r}\right), \tag{4}$$

where $\boldsymbol{C}_{H_i}$ corresponds to the covariance matrix of dimension $(N \times N)$ of $\boldsymbol{r}$ under hypothesis H_i and $|\cdot|$ denotes the determinant. Let σ^2 be the variance of the discretized noise and $\boldsymbol{C}_y$ the covariance matrix of the discrete received signal, $\boldsymbol{y}$, under hypothesis H_1. Then, $\boldsymbol{C}_{H_0} = \sigma^2\boldsymbol{I}$ and $\boldsymbol{C}_{H_1} = \sigma^2\boldsymbol{I} + \boldsymbol{C}_y$ ($\boldsymbol{I}$ represents the identity matrix). Denote by $\boldsymbol{C}_s$ the $(N_1 \times N_1)$ covariance matrix of the discretized emitted transient signal $s(t)$, where N_1 represents the interval where most of the signal energy lies. Consider the decomposition $\boldsymbol{C}_s = \boldsymbol{V}_s\boldsymbol{D}\boldsymbol{V}_s'$, where $\boldsymbol{V}_s$ $(N_1 \times N_\lambda)$ and $\boldsymbol{D} = \text{diag}\{\lambda_1^s, \cdots, \lambda_{N_\lambda}\}$ are, respectively, the eigenvector and eigenvalue matrices of $\boldsymbol{C}_s$. Under these conditions, we have

$$\boldsymbol{C}_y = \boldsymbol{V}_y\boldsymbol{D}\boldsymbol{V}_y', \tag{5}$$

with

$$\boldsymbol{V}_y = \sum_{k=1}^{N_q} \alpha_k \boldsymbol{V}_s^{q_k} \tag{6}$$

and

$$\boldsymbol{V}_s^k = \begin{bmatrix} \boldsymbol{o}(k, N_\lambda) \\ \boldsymbol{V}_s \\ \boldsymbol{o}(N - N_1 - k, N_\lambda) \end{bmatrix}, \tag{7}$$

where $\boldsymbol{o}(n, m)$ stands for the $(n \times m)$ null matrix. It is easy to show [4] that the likelihood ratio (4) can be rewritten as

$$l(\boldsymbol{r}) = \exp\left[\frac{1}{2}\ln\left(|\boldsymbol{I} - \sigma^2\boldsymbol{V}_y'\boldsymbol{V}_y\boldsymbol{W}_y|\right) + \frac{1}{2}\boldsymbol{r}'\boldsymbol{V}_y\boldsymbol{W}_y\boldsymbol{V}_y'\boldsymbol{r}\right], \tag{8}$$

with

$$\boldsymbol{W}_y = \boldsymbol{D}_2(\boldsymbol{U}_1\boldsymbol{D}_2 + \boldsymbol{I})^{-1}/\sigma^2 \tag{9}$$

$$\boldsymbol{U}_1 = \sum_{k=1}^{N_q}\sum_{\substack{l=1 \\ l\neq k}}^{N_q} \alpha_k\alpha_l(\boldsymbol{V}_s^k)'\boldsymbol{V}_s^l \tag{10}$$

$$\boldsymbol{D}_2 = (c_1\boldsymbol{I} + \sigma^2\boldsymbol{D}^{-1})^{-1}, \tag{11}$$

where $c_1 = \sum_{k=1}^{N_q} \alpha_k^2$.

Remark that the optimal detector corresponds to taking the expected value in order to the ACs and DCs of (8). However, the resulting processor has no closed-form expression and requires a huge computational load. To overcome this inconvenient, we simplify (8) by taking the terms up to the first order of its Taylor series around a convenient working point. Under the conditions that the SNR is low (i.e., $\lambda_k < \sigma^2$, $\forall k = 1, \cdots, N_\lambda$) and the multipath channel amplifies the signal energy arriving at the receiver (i.e., $c_1 > 1$), then the elements of the diagonal of the matrix D_2 are small. Thus, the first-order approximation of the likelihood ratio around the point $D_2 = \boldsymbol{o}(N_\lambda, N_\lambda)$ is

$$l(\boldsymbol{r}) \simeq 1 - \frac{1}{2}tr\left\{\boldsymbol{V}_y\boldsymbol{D}_2\boldsymbol{V}_y'\right\} + \frac{1}{2\sigma^2}\boldsymbol{r}'\boldsymbol{V}_y\boldsymbol{D}_2\boldsymbol{V}_y'\boldsymbol{r} = \pi(\boldsymbol{r}). \tag{12}$$

When the DCs and the ACs are known, $\pi(\boldsymbol{r})$ still represents the optimum solution when there is no overlapping between the replicas arriving at the receiver, since $\boldsymbol{U}_1 = \boldsymbol{o}(N_\lambda, N_\lambda)$. This situation corresponds to the case where the duration of the emitted signal is small comparing to the channel delays. However, the examples presented in [4] show that, even when there is a large overlapping between replicas, the power of the signal arriving at the receiver increases and the performance of the processor does not degrade significantly.

In the next section the suboptimal processor is derived by taking the expected value in order to the ACs, $\boldsymbol{\alpha}$, and DCs, $\boldsymbol{q}$, of the approximated likelihood ratio $\pi(\boldsymbol{r})$ (12).

4. EXPECTED VALUE OF $\pi(\boldsymbol{r})$

In the sequel, the elements of $\{\alpha_1, \cdots, \alpha_{N_q}, q_1, \cdots, q_{N_q}\}$ are assumed to be mutually independent. By taking the expected value in order to the ACs, $\boldsymbol{\alpha}$, one gets

$$E_{\boldsymbol{\alpha}}[\pi(\boldsymbol{r})] = 1 + \frac{1}{2\sigma^2}\boldsymbol{r}'E_{\boldsymbol{\alpha}}[\boldsymbol{V}_y\boldsymbol{D}_2\boldsymbol{V}_y']\boldsymbol{r} - \frac{1}{2}\text{tr}\{E_{\boldsymbol{\alpha}}[\boldsymbol{V}_y\boldsymbol{D}_2\boldsymbol{V}_y']\}, \tag{13}$$

where $\text{tr}\{\boldsymbol{X}\}$ denotes the trace of $\boldsymbol{X}$. Since $\boldsymbol{D}_2$ depends on $\boldsymbol{\alpha}$ through c_1, the expected value in (13) should be evaluated using the joint probability density function of $\boldsymbol{\alpha}$. However, if we assume that $c_1 = c_1^c$ is approximately constant, it is only necessary to know the first and second order statistics of $\boldsymbol{\alpha}$, i.e.,

$$E_{\boldsymbol{\alpha}}[\boldsymbol{V}_y\boldsymbol{D}_2\boldsymbol{V}_y'] \simeq \sum_{k_1=1}^{N_q}\sum_{k_2=1}^{N_q} C_{\alpha_{k_1}\alpha_{k_2}}\boldsymbol{V}_s^{q_{k_1}}\boldsymbol{D}_2(\boldsymbol{V}_s^{q_{k_2}})', \tag{14}$$

where $C_{\alpha_{k_1}\alpha_{k_2}} = E[\alpha_{k_1}\alpha_{k_2}]$ represents the crosscorrelation between α_{k_1} and α_{k_2}. The constant c_1^c is for now left as a free parameter that will be chosen in order to maximize the performance of the final expression of the processor.

The final expression of the likelihood test is obtained taking the expected value of (13)with respect to the DCs, $\boldsymbol{q}$, i.e.,

$$\pi_p(\boldsymbol{r}) \underset{H_0}{\overset{H_1}{\gtrless}} \mu, \tag{15}$$

where the threshold μ includes the terms of (13) that do not depend on the observation process $\boldsymbol{r}$, and

$$\pi_p(\boldsymbol{r}) = \boldsymbol{r}'\left\{\sum_{k_1=1}^{N_q}\sum_{k_2=1}^{N_q} C_{\alpha_{k_1}\alpha_{k_2}}E_{q_{k_1},q_{k_2}}\left[\boldsymbol{V}_s^{q_{k_1}}\boldsymbol{D}_2(\boldsymbol{V}_s^{q_{k_2}})'\right]\right\}\boldsymbol{r}. \tag{16}$$

When $k = k_1 = k_2$ we have

$$\bar{\boldsymbol{C}}_k = E_{q_k}\left[\boldsymbol{V}_s^{q_k}\boldsymbol{D}_2(\boldsymbol{V}_s^{q_k})'\right] = \sum_{m=1}^{N_q} \boldsymbol{V}_s^m\boldsymbol{D}_2(\boldsymbol{V}_s^m)'P_k(m), \tag{17}$$

where $P_k(m)$ denotes the probability function of q_k, and for $k_1 \neq k_2$,

$$E_{q_{k_1},q_{k_2}}\left[\boldsymbol{V}_s^{q_{k_1}}\boldsymbol{D}_2(\boldsymbol{V}_s^{q_{k_2}})'\right] = \bar{\boldsymbol{V}}_{k_1}\boldsymbol{D}_2\bar{\boldsymbol{V}}'_{k_2}, \quad (18)$$

with

$$\bar{\boldsymbol{V}}_k = \sum_{m=1}^{N_q} \boldsymbol{V}_s^m P_k(m). \quad (19)$$

The probability $P_k(m)$ related to the DC q_k, corresponds to an uncertainty measure of the time interval between the arrival of the first and the k-th replica at the receiver. Thus, as noted before, $q_1 = 0$, and $P_1(m) = \delta_m$ (the kronecker delta). The quadratic form of the likelihood ratio is given by

$$\pi_p(\boldsymbol{r}) = \boldsymbol{r}'\left\{\sum_{k=1}^{N_q} C_{\alpha_k}\bar{\boldsymbol{C}}_k + \sum_{k_1=1}^{N_q}\sum_{\substack{k_2=1\\k_1\neq k_2}}^{N_q}\bar{\alpha}_{k_1}\bar{\alpha}_{k_2}\bar{\boldsymbol{V}}_{k_1}\boldsymbol{D}_2\bar{\boldsymbol{V}}'_{k_2}\right\}\boldsymbol{r}, \quad (20)$$

where $C_{\alpha_k} = C_{\alpha_k\alpha_k} = \bar{\alpha}_k^2 + \sigma_k^2$ ($\bar{\alpha}_k$ and σ_k^2 being, respectively, the mean and variance of α_k).

In (20), matrices $\bar{\boldsymbol{C}}_k$ and $\boldsymbol{D}_2$ depend on the constant c_1^c, that is tuned in order to maximize the detector performance, using the following procedure: i) for each possible value of c_1^c, let $\boldsymbol{M}$ be the matrix inside brackets in (20), and determine the covariance matrix $\boldsymbol{C}_{H_1}^*$ that corresponds to a hypothesis for which the processor (20) is optimal, i.e., $\boldsymbol{M} = \boldsymbol{C}_{H_0}^{-1} - (\boldsymbol{C}_{H_1}^*)^{-1}$ ($\boldsymbol{C}_{H_0} = \sigma^2\boldsymbol{I}$); ii) choose the value of c_1^c that maximizes either the Chernoff or the Bhattacharyya distances [5] between $\boldsymbol{C}_{H_1}^*$ and $\boldsymbol{C}_{H_0}$. This procedure corresponds to minimizing a bound for the probability of error between both hypotheses and is computationally efficient.

The processor presented in (20) is a quadratic form. Its computational cost is relatively low, comparing with the GLRT, because the matrix $\boldsymbol{M}$ can be computed off-line when the conditions of the multipath channel change. Although the processor depends on the probability functions of the delay coefficients which may not be easy to determine, the simulation studies show that the receiver performance is robust to mismatches on these probabilities. Thus, when the true $P_k(m)$ are unknown, the processor is designed assuming uniform probabilities and, in general, the resulting performance degradation is not important.

5. SEQUENTIAL DETECTION OF REPLICAS

In (20), the length N of the observation vector $\boldsymbol{r}$ must be large enough to include all replicas arriving at the receiver. Therefore, it is only possible to detect a signal when its last replica arrives. This section develops a sequential structure based on (20) that allows the detection of a signal without the need of waiting for all replicas. In this case, the detection is performed as soon as an arbitrary number of replicas possess a sufficient amount of energy to ensure with high probability that a signal was emitted by the source. Furthermore, in many situations, the sequential detection of replicas may reduce the computational cost of the processor.

In (20), matrices $\bar{\boldsymbol{C}}_k$ and $\bar{\boldsymbol{V}}_k$ have dimensions $(N \times N)$ and $(N \times N_\lambda)$. However, the dimension of the non-null terms of each of these matrices is significantly lower than N due either to the transient characteristic of the emitted signal and also because it is assumed that the probabilities $P_k(m)$ associated to each replica have compact support. Let $[\theta_k^i;\theta_k^f]$ be the support of $P_k(m)$ (with $\theta_1^i = 0$); then, the matrices $\bar{\boldsymbol{C}}_k$ and $\bar{\boldsymbol{V}}_k$ have, respectively, support $[\theta_k^i+1;\theta_k^f+N_1]\times[\theta_k^i+1;\theta_k^f+N_1]$ and $[\theta_k^i+1;\theta_k^f+N_1]\times[1;N_\lambda]$. Defining $N_{q_k} = \theta_k^f - \theta_k^i + N_1$, we denote by $\bar{\boldsymbol{C}}_k^r$ $(N_{q_k}\times N_{q_k})$ and $\bar{\boldsymbol{V}}_k^r$ $(N_{q_k}\times N_\lambda)$ the matrices that include only the elements belonging to the support of $\bar{\boldsymbol{C}}_k$ and $\bar{\boldsymbol{V}}_k$. If, at each time instant n, the vector of the last N_{q_k} elements of the observation process is represented by $\boldsymbol{r}_k(n) = [r(n-N_{q_k}+1)\cdots r(n)]'$, then expression (20) may be rewritten as

$$\pi_p(\boldsymbol{r}) = \sum_{k=1}^{N_q}(\bar{\alpha}_k^2+\sigma_k^2)l_k(\theta_k^f+N_1) + 2\sum_{k_1=2}^{N_q}\sum_{k_2=1}^{k_1-1}\bar{\alpha}_{k_1}\bar{\alpha}_{k_2}\boldsymbol{b}'_{k_1}(\theta_{k_1}^f+N_1)\boldsymbol{D}_2\boldsymbol{b}_{k_2}(\theta_{k_2}^f+N_1), \quad (21)$$

where

$$l_k(n) = \boldsymbol{r}'_k(n)\boldsymbol{Z}_k\boldsymbol{D}_k^z\boldsymbol{Z}'_k\boldsymbol{r}_k(n) \quad \text{and} \quad \boldsymbol{b}_k(n) = (\bar{\boldsymbol{V}}_k^r)'\boldsymbol{r}_k(n). \quad (22)$$

In (22), $\boldsymbol{Z}_k$ and $\boldsymbol{D}_k^z$ represent, respectively, the matrices of eigenvectors and eigenvalues of $\bar{\boldsymbol{C}}_k^r$ ($\bar{\boldsymbol{C}}_k^r = \boldsymbol{Z}_k\boldsymbol{D}_k^z\boldsymbol{Z}'_k$). Remark that, due to the fact that $q_1 = 0$ and $P_1(m) = \delta_m$, then $\boldsymbol{Z}_1 = \boldsymbol{V}_s$ and $\boldsymbol{D}_1^z = \boldsymbol{D}_2$. Under these conditions, the likelihood ratio may be rewritten in a recursive form:

$$\begin{aligned}\pi_h(n) &= \pi_{h-1}(n-\theta_h^f+\theta_{h-1}^f) + (\bar{\alpha}_h^2+\sigma_h^2)l_h(n)\\ &+ 2\bar{\alpha}_h\left[\sum_{k=1}^{h-1}\bar{\alpha}_k\boldsymbol{b}'_k(n-\theta_h^f+\theta_k^f)\right]\boldsymbol{D}_2\boldsymbol{b}_h(n), \quad (23)\end{aligned}$$

with $\pi_1(n) = (\bar{\alpha}_1^2+\sigma^2)l_1(n)$. For $h = N_q$, $\pi_{N_q}(n)$ corresponds to expression (20).

Comparing expressions (20) and (23) from a computational point of view, the solution that leads to a less expensive processor depends both on the signals to detect and on the multipath channel structure. In general, however, two situations may arise. First, when there is a large overlapping between replicas, using the recursive processor (23), it is necessary to perform N_q eigenvector decompositions of length N_{q_k}, while the non-recursive form (20) only needs one eigenvector decomposition of length N. Since, in this case, $N \ll \sum_k N_{q_k}$, we should expect that the non-recursive solution would be less expensive than the recursive one. However, this may not be true because, in most cases, the number of relevant eigenvalues is much more important in the non-recursive processor, thus increasing its computational cost. When the overlapping is smaller, then clearly the recursive solution becomes more attractive than the non-recursive one.

From the recursive processor (23), it is possible to derive a detection structure suited for sequential detection of replicas. For every $h = 1,\cdots,N_{q-1}$, two likelihood tests are performed. The first test compares $\pi_h(n)$ with a high-valued threshold, μ_h^{sup}; if $\pi_h(n) > \mu_h^{sup}$, then it is assumed that a signal arrived to the receiver with a low probability of false alarm, and there is no need to evaluate $\pi_{h+1}(n+\theta_{h+1}^f-\theta_h^f)$, that corresponds to the arrival of the next replica; if $\pi_h(n) < \mu_h^{sup}$, then another test is performed against a low threshold μ_h^{inf}; if $\pi_h(n) < \mu_h^{inf}$, we consider that no signal is present at the receiver (with low miss probability) and the procedure stops. Only in the case where $\mu_h^{inf} \leq \pi_h(n) \leq \mu_h^{sup}$, the processor waits for the next replica to make a decision. When $h = N_q$, only one final test is performed.

6. SIMULATION RESULTS

The received signal is a weighted sum of 10 delayed replicas of a chirplike stochastic transient with autocorrelation function shown in figure 1. The noise variance is $\sigma^2 = 5$. The mean value for the overlapping between consecutive replicas $\bar{\Delta}_q$, is 80% or 20%.

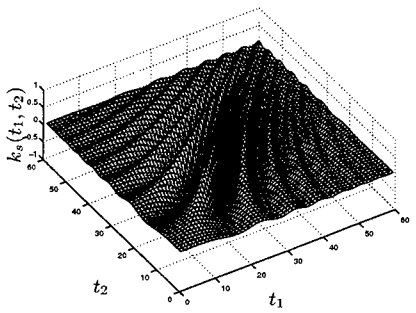


Fig. 1. Autocorrelation function of the emitted signal.

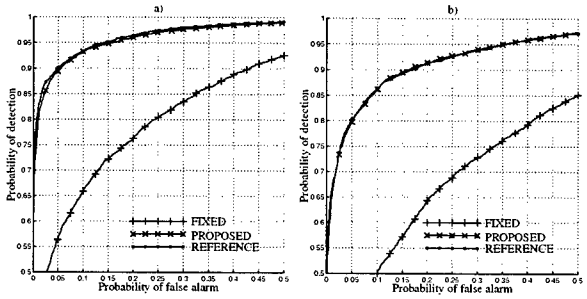


Fig. 2. ROCs: a) $\bar{\Delta}_q = 80\%$. b) $\bar{\Delta}_q = 20\%$.

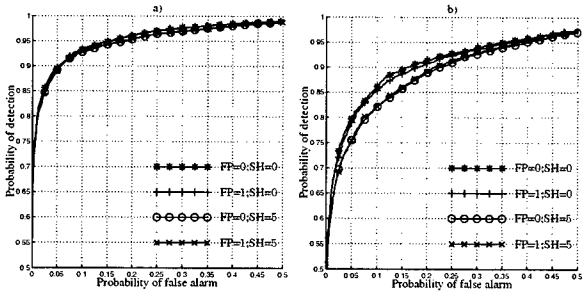


Fig. 3. Statistics mismatch: a) $\bar{\Delta}_q = 80\%$. b) $\bar{\Delta}_q = 20\%$.

The DCs are generated from a Gaussian probability function with 1/3rd of the length of the emitted signal, while the ACs have $\bar{\alpha}_k = 1$ and $\sigma_k = 0.2$, $\forall k = 1, \cdots, 10$. The receiver operating characteristics (ROC) are obtained by simulation with 5000 Monte-Carlo runs. In figures 2 a) and b), the performance of the proposed detector (PROPOSED) is compared with i) a reference one (REFERENCE), consisting on the best possible quadratic processor obtained from the covariance matrix estimated from the 5000 signals arriving at the receiver; and ii) a detector that assumes that the DCs and ACs take fixed values (FIXED), equal to the mean values of the real channel parameters. We conclude that, for both overlapping situations, the performance of the proposed solution is very close to the best quadratic processor and presents a huge gain comparing to the fixed parameters one. Figures 3 a) and b) show the performance degradation due to mismatch in the statistics of the DCs. For FP = 0 the DCs probability function is known, while for FP = 1 an uniform probability with the support of the true one is used. When SH = 5, a shift (25 %) of the support of the DCs probability function is considered in the temporal localization of the DCs. If SH = 0, no mismatch exists. For a large overlapping between replicas ($\bar{\Delta}_q = 80\%$), the proposed processor shows to be robust to statistics mismatch. However, for a smaller overlapping, figure 3 b) shows some sensitivity to shifts mismatch. In both cases, the assumption of an uniform probability function introduces only a small degradation on the processor.

Table 1. Sequential detection of replicas

Scenario	PFA	PD	CC
i)	0.1810	0.7102	18.16%
ii)	0.1064	0.8374	65.31%
iii)	0.1246	0.8140	41.59%

Regarding the sequential detection of replicas, and for mean overlapping between replicas of 50%, three scenarios are considered: i) small μ^{sup} and large μ^{inf}; ii) the opposite of i); and iii) an intermediate situation, between i) and ii). Table 1 shows the probabilities of detection (PD) and of false alarm (PFA), and the percentage of computational complexity (CC) needed, comparing with the processor that waits for the arrival of all the 10 replicas, for which PD = 0.8442 and PFA = 0.1. In situation i) the CC reduces drastically but the PD and the PFA also suffer an important degradation. In this case, detection is performed at the arrival of few replicas. In situation ii), although there is only a small loss of performance, the CC is still reduced to 65.31%.

7. CONCLUSION

The optimal processor for stochastic transient signal detection in a multipath environments is, in general, computationally untractable. This paper presents a computationally efficient suboptimal solution, where the multipath parameters are modelled as random variables. A structure for sequential detection of replicas is proposed, avoiding the need to wait for all replicas to make a decision. The proposed solution is robust to the multipath statistics mismatch, thus requiring only mild *a-priori* information about the channel.

8. REFERENCES

[1] I. M. G. Lourtie and G. C. Carter, "Signal detectors for random ocean media," *J. Acoust. Soc. Am.*, vol. 92, no. 3, pp. 1420–1427, September 1992.

[2] C. He and J. M. F. Moura, "Focused detection via multiresolution analysis," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 1094–1104, April 1998.

[3] F. M. Garcia and I. M. G. Lourtie, "Efficiency of real-time Gaussian transient detectors: Comparing the Karhunen-Loève and the wavelet decompositions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000.

[4] F. M. Garcia, *Detecção Passiva de Sinais Transientes*, Ph.D. thesis, Instituto Superior Técnico, September 2000.

[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition, Second Edition*, Academic Press, 1990.

DETECTION OF INDEPENDENT TIMING JITTER IN SINUSOIDAL MEASUREMENTS

Mark R. Morelande[1] *and D. Robert Iskander*[2]

[1] Centre for Eye Research, Queensland University of Technology
Victoria Park Road, Q 4059, Australia
[2] School of Engineering, Griffith University
PMB 50 Gold Coast Mail Centre, Q 9726, Australia

ABSTRACT

Two detectors of symmetrically distributed independent timing jitter in a data record composed of a complex harmonic in additive white Gaussian noise are proposed. The proposed detectors are computationally efficient and, although they are formulated using asymptotic results, they may be effectively used with small sample lengths under a wide range of conditions. The performances of the detectors are analysed using simulations and theoretical results.

1. INTRODUCTION

Nearly all sampling systems exhibit timing jitter wherein the spacing between sampling instants is not uniform but varies about the nominal sampling period in a random fashion. In most cases the deviations from the nominal sampling period are small enough that they may be ignored. However, there are cases in which the effects of jitter can become significant [1, 2]. The signal model considered in this paper is a complex harmonic in additive noise which may or may not have randomly spaced sampling instants. Specifically, we consider observations from

$$X_t = g_0 \exp[j\{\omega_0(t + U_t) + \psi\}] + W_t, \quad t \in \mathbb{Z}, \quad (1)$$

where $g_0 \in \mathbb{R}^+$, $-\pi \leq \omega_0 < \pi$, $\omega_0 \neq 0$, $-\pi \leq \psi < \pi$, $U_t, t \in \mathbb{Z}$ are zero-mean real-valued independent and identically distributed (iid) random variables with variance σ_U^2 and $\{W_t\}$ is a circular complex-valued white normal random process with variance σ_W^2, independent of $\{U_t\}$. It is assumed that $\{U_t\}$ is symmetrically distributed and has a characteristic function $\phi_U(s) = \mathsf{E} \exp(jsU_t)$ which converges for $s = k\omega_0$, $k = 1, \ldots, 4$. We aim to provide a solution to the following problem: Given observations $x_0, \ldots, x_{n-1}$ from (1) test the hypothesis $\mathsf{H} : \sigma_U = 0$ against the alternative $\mathsf{K} : \sigma_U > 0$. Under this formulation, a decision for the alternative indicates the presence of timing jitter. It is assumed that all signal and noise parameters and the values $u_0, \ldots, u_{n-1}$ of the timing offsets are unknown. A method of checking for the presence of timing jitter is desirable as a diagnostic tool for evaluating the performance of sampling systems. A jitter detector can also be used to select estimation procedures since the optimal estimation procedures will differ depending on whether or not jitter exists.

Previous work on detecting timing jitter has been performed by Sharfer and Messer [6, 7]. In that case the presence of jitter in the sampling of a band-limited zero-mean stationary random process with non-zero third-order cumulants was found to be characterised by non-nullity of the bispectrum in a certain region. This observation precipitated the formulation of an appropriate statistical test. The bispectrum jitter detector cannot be used for the problem considered here because the random process $\{X_t\}$ of (1) does not satisfy the required assumptions. No jitter detectors for the signal model considered here have been proposed in the literature. In this paper two test statistics based on estimators of $\omega_0^2\sigma_U^2$ are proposed. Starting from the same result, the test statistics are derived using different assumptions. In the first case the jitter is assumed to be normally distributed while in the second case a small jitter approximation is used. The resulting detectors are computationally efficient and maintain the nominal false alarm probability, although, for small sample lengths, mild conditions on the signal-to-noise ratio are required for the normal assumption detector. The effects of the assumptions made in the formulation of the two detectors are studied using theoretical and simulation results.

2. PROPOSED METHODS

In this section two detectors of independent timing jitter are proposed. The conditions under which the detectors are consistent are established and examined for various jitter distributions.

After estimating the frequency ω_0 and initial phase ψ as

$$\hat{\omega}_0 = \arg \max_{-\pi \leq \omega < \pi} |d_X(\omega)|, \quad (2)$$

$$\hat{\psi} = \angle d_X(\hat{\omega}_0), \quad (3)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

where $d_X(\omega)$ is the finite Fourier transform (FT) of the sequence $x_0, \ldots, x_{n-1}$,

$$d_X(\omega) = \sum_{t=0}^{n-1} x_t \exp(-j\omega t),$$

the observations are demodulated to form the sequence

$$y_t = x_t \exp\{-j(\hat{\omega}_0 t + \hat{\psi})\}, \quad t = 0, \ldots, n-1.$$

Although the estimators $\hat{\omega}_0$ and $\hat{\psi}$ are optimal only if no timing jitter exists, it has been shown in [3] that, even for large values of σ_U^2, these estimators have variances close to the Cramér-Rao bound for the case of normally distributed jitter. Under the given assumptions it is shown in [3] that $\hat{\omega}_0 = \omega_0 + O_m(n^{-3/2})$ and $\hat{\psi} = \psi + O_m(n^{-1/2})$ so that

$$Y_t = g_0 \exp(j\xi_t) + V_t + O_m(n^{-1/2}), \tag{4}$$

where $\xi_t = \omega_0 U_t$ and $V_t = W_t / \exp\{j(\omega_0 t + \psi)\}$. Since it is assumed that $\omega_0 \neq 0$, the variance σ_ξ^2 of ξ_t will be zero only if $\sigma_U^2 = 0$. Therefore, non-nullity of σ_ξ^2 can be used to test for the presence of jitter.

In the following, terms in (4) which disappear as the sample length $n \to \infty$ will be ignored. Such terms do not affect the asymptotic analysis. Let $r_{kY} = \mathsf{E}\,\mathrm{Re}(Y_t)^k$ and $i_{kY} = \mathsf{E}\,\mathrm{Im}(Y_t)^k$, $k = 1, 2, \ldots$. It is straightforward to show that

$$\sqrt{r_{2Y} - i_{2Y}} \Big/ r_Y = \sqrt{\phi_\xi(2)} \Big/ \phi_\xi(1), \tag{5}$$

where $\phi_\xi(s) = \mathsf{E}\,\exp(js\xi_t)$ is the characteristic function of $\{\xi_t\}$. Eq. (5) may be used as the basis of an asymptotically unbiased estimator of σ_ξ^2, provided that knowledge of the distribution of the jitter is available. Since the jitter distribution is assumed to be unknown, an estimator of σ_ξ^2 which is, in general, unbiased cannot be obtained from (5). However, since we are concerned only with jitter detection, an unbiased estimator of jitter variance is not a necessity. With this in mind, one approach is to assume a distribution for the jitter and derive an estimator of σ_ξ^2 on this basis. Another approach is to assume small amounts of jitter and replace the characteristic function $\phi_\xi(s)$ by its second-order approximation. A distribution-independent estimator can then be derived. Both of these approaches require that the performances of the resulting detectors are carefully studied for a range of jitter distributions. This analysis will be performed in Section 3.

2.1. Normal Assumption

We will proceed from (5) as if the jitter is normally distributed. The normal distribution is chosen for the simple derivation it affords. Importantly, this choice does not prohibit the use of the proposed detector for other jitter distributions, as will be demonstrated in Section 3. Substituting $\phi_\xi(s) = \exp(-s^2\sigma_\xi^2/2)$ into (5) gives

$$\exp(-\sigma_\xi^2/2) = \sqrt{r_{2Y} - i_{2Y}} \Big/ r_Y$$

Simple re-arrangements result in the following

$$\sigma_\xi^2 = 2\log(r_Y / \sqrt{r_{2Y} - i_{2Y}})$$

Since the moments of the real and imaginary parts of Y_t are unavailable, they are replaced by their sample estimators,

$$\hat{r}_{kY} = 1/n \sum_{t=0}^{n-1} \mathrm{Re}(Y_t)^k, \tag{6}$$

$$\hat{\imath}_{kY} = 1/n \sum_{t=0}^{n-1} \mathrm{Im}(Y_t)^k, \tag{7}$$

to form the estimator

$$\hat{\sigma}_\xi^2 = 2\log\left(\hat{r}_Y / \sqrt{\hat{r}_{2Y} - \hat{\imath}_{2Y}}\right)$$

Under H, $\sqrt{n}\hat{\sigma}_\xi^2 \overset{a}{\sim} \mathsf{N}(0, \sigma_W^4/g_0^4)$. Standardisation leads to the statistic $\hat{T} = \sqrt{n}\hat{r}_Y^2\hat{\sigma}_\xi^2/(2\hat{\imath}_{2Y})$ which is asymptotically standard normal under H.

2.2. Small Jitter Approximation

Under the small jitter approximation and assuming symmetrically distributed jitter, $\phi_\xi(s) = 1 - s^2\sigma_\xi^2/2$ is substituted into (5) giving

$$\begin{aligned} \sqrt{r_{2Y} - i_{2Y}} \Big/ r_Y &\approx \sqrt{1 - 2\sigma_\xi^2} \Big/ \left(1 - \sigma_\xi^2/2\right), \\ &\approx (1 - \sigma_\xi^2) \Big/ \left(1 - \sigma_\xi^2/2\right), \end{aligned} \tag{8}$$

where the second line is obtained by replacing the square root on the right hand side with its first-order Taylor series approximation. After replacing the moments with their sample estimators and performing some simple manipulations the following estimator of σ_ξ^2 is obtained:

$$\tilde{\sigma}_\xi^2 = \left(\hat{r}_Y - \sqrt{\hat{r}_{2Y} - \hat{\imath}_{2Y}}\right) \Big/ \left(\hat{r}_Y - \sqrt{\hat{r}_{2Y} - \hat{\imath}_{2Y}}/2\right).$$

Although $\tilde{\sigma}_\xi^2$ is, in general, a biased estimator of σ_ξ^2, it is asymptotically unbiased when $\sigma_U^2 = 0$, i.e. $\mathsf{E}\,\tilde{\sigma}_\xi^2 = 0$ under H. Using theorems given in [5] it can be shown that, under H, $\sqrt{n}\tilde{\sigma}_\xi^2 \overset{a}{\sim} \mathsf{N}(0, \sigma_W^4/g_0^4)$. It is not surprising to see that the asymptotic null distribution of $\tilde{\sigma}_\xi^2$ is the same as that of $\hat{\sigma}_\xi^2$. The statistic $\tilde{T} = \sqrt{n}\hat{r}_Y^2\tilde{\sigma}_\xi^2/(2\hat{\imath}_{2Y})$ is asymptotically standard normal under H.

Jitter detectors based on the statistics $\hat{T}$ and $\tilde{T}$ will be developed using asymptotic results. Since the test statistics are asymptotically standard normal under H, the null hypothesis is rejected, i.e. it is decided that jitter is present, if the test statistic exceeds $\Phi^{-1}(1-\alpha)$ where $\Phi(\cdot)$ is the standard normal distribution function and α is the prescribed false alarm probability. The asymptotic distributions of $\hat{T}$ and $\tilde{T}$ are derived in [4] under the alternative. It is shown that, for the mean values of the detectors to be real-valued it is required that $r_{2Y} - i_{2Y} > 0$ which corresponds to $\phi_\xi(2) > 0$. In practice, reliable use of the detectors requires that $\phi_\xi(2)$ is significantly non-zero so that the probability of $\hat{r}_{2Y} - \hat{i}_{2Y} < 0$ occurring for a particular realisation is small. This issue is considered in [4].

2.3. Consistency

The detectors will be consistent, i.e., the detection probability tends to one as $n \to \infty$ for fixed parameter values, if

$$\phi_\xi(1) > \sqrt{\phi_\xi(2)}, \quad \phi_\xi(2) > 0. \tag{9}$$

The inequality (9) holds for normally distributed jitter but not necessarily for uniformly distributed jitter. In particular, if $U_t \sim \mathsf{U}[-a, a]$, (9) becomes $\tan(|\omega_0|a) > |\omega_0|a$ with $|\omega_0|a \in 2\pi k + (0, \pi/2)$, $k = 0, 1, \ldots$. Of the values of a which satisfy this inequality for a given frequency ω_0, of significant interest are the subset $a < \pi/(2|\omega_0|)$. Other values of a which satisfy the required inequality will be too large to be of practical interest. In the most restrictive case, $|\omega_0| = \pi$ and the requirement for consistency becomes $a < 1/2$. This is not a strict condition since it allows values of a up to the point at which time-reversals are possible. Note that if $|\omega_0| < \pi$, it is possible to consistently detect uniformly distributed jitter even if time-reversals occur with non-zero probability.

Finally it is noted that, in addition to (9), the jitter detector derived under the small jitter approximation is consistent if, for $\phi_\xi(2) > 0$,

$$\phi_\xi(1) < 0 \bigcup \left\{ \phi_\xi(1) > 0 \bigcap 2\phi_\xi(1) < \sqrt{\phi_\xi(2)} \right\}.$$

Therefore the jitter detector based on the small jitter approximation is consistent under a wider range of conditions than the jitter detector based on the normal assumption.

3. PERFORMANCE ANALYSIS

This section contains a performance analysis and comparison of the proposed jitter detectors. We first verify the ability of the detectors to maintain the prescribed false alarm probability for finite data records. In the second part of this section the performances of the detectors are analysed for normal and uniform jitter distributions.

3.1. False Alarm Probability

The false alarm probabilities of the jitter detectors are estimated for various sample lengths and nominal false alarm probabilities of 1 % and 5 %. The frequency $\omega_0 = 1$, the initial phase $\psi = 1/2$ and the signal-to-noise ratio (SNR), defined as $\mathcal{S} = g_0^2/\sigma_W^2$, is set to 0 dB. For each scenario, 50 000 realisations of (1) are generated under the null hypothesis. The results, shown in Table 1, indicate that both detectors maintain the nominal false alarm probability in all cases. Although both detectors are excessively conservative for small sample lengths, it can be seen that the actual false alarm probabilities approach the nominal false alarm probabilities as the sample length increases.

Table 1: Estimated false alarm probabilities for jitter detectors based on the small jitter approximation (left) and the normal assumption (right). The SNR $\mathcal{S} = 0$ dB.

	α (%)			
$\log_2(n)$	1		5	
5	0	0.47	0.25	2.95
6	0.01	0.42	1.08	3.23
7	0.06	0.50	1.97	3.70
8	0.19	0.60	2.84	4.08
9	0.34	0.71	3.27	4.22
10	0.59	0.88	3.77	4.48
11	0.62	0.82	4.14	4.61

An issue of some importance is the effect of SNR on the false alarm probabilities. Simulation results given in [4] show that, for a given sample length, the actual false alarm probability of the detector based on the normal assumption increases as $\mathcal{S}$ decreases. In fact the false alarm probability exceeds the set level for SNRs below -3 dB although this exceedance becomes smaller as the sample length increases. The false alarm probability of the detector based on the small jitter approximation is less than the set level in all cases, but does exhibit a slight increase in false alarm probability as $\mathcal{S}$ increases.

3.2. Detection Probability

The power of the two detectors is now examined using theoretical and simulation results. In the first example the jitter is normally distributed. The frequency $\omega_0 = 3/4$, the initial phase $\psi = 1/2$ and the sample length $n = 512$. The variance σ_U^2 of the jitter is varied between 10^{-3} and 10^0 and the SNR $\mathcal{S}$ is varied between 0 dB and 10 dB. In all of these cases the probabilities of the test statistics being complex-valued are negligible. Simulation results, obtained using 1000 realisations of (1) for each scenario, are shown

in Figure 1 for $\alpha = 0.01$. The simulation results are accompanied by theoretical results derived in [4].

Table 2: Legend for Figures 1 and 2.

$\mathcal{S}$ (dB)	Empirical	Theoretical
0	*	—
5	o	- - -
10	△	···

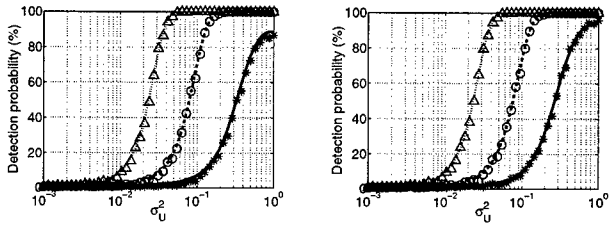


Figure 1: Detection probability (in %) of the jitter detectors based on a) the small jitter approximation and b) the normal assumption in the presence of normally distributed jitter. The false alarm probabilities is 1 %. The legend is given in Table 2.

It can be seen that the detection probabilities are high for small values of σ_U^2 when $\mathcal{S} = 5, 10$ dB. Notably, there is little difference between the performances of the detectors for these values of SNR. However, for $\mathcal{S} = 0$ dB it is clear that the detector based on the normal assumption outperforms the detector based on the small jitter approximation. Another results of interest is that, for $\mathcal{S} = 0$ dB, the detection probability does not tend to one as σ_U^2 increases. In fact, if σ_U^2 is increased beyond one, the detection probability actually decreases. This effect can be attributed to an increase in the variances of the test statistics accompanied by decreases in the mean. Additional simulations, not shown here, verify that for a given SNR this effect disappears as the sample length n increases.

The above experiments are repeated for uniformly distributed jitter with the results shown in Figure 2 for $\alpha = 0.01$. For a given jitter variance, it can be seen that the jitter detectors perform better for uniformly distributed jitter as compared to normally distributed jitter. This is particularly so for $\mathcal{S} = 0$ dB.

4. CONCLUSIONS

Two closely-related methods were proposed for detecting the presence of symmetrically distributed independent timing jitter in a complex harmonic. One detector was obtained

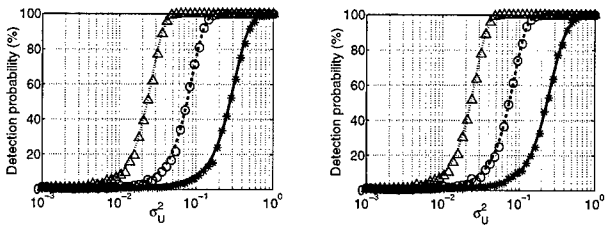


Figure 2: Detection probability (in %) of the jitter detectors based on a) the small jitter approximation and b) the normal assumption in the presence of uniformly distributed jitter. The false alarm probabilities is 1 %. The legend is given in Table 2.

by employing a small jitter approximation while the other was obtained through a normal assumption. The conditions required for consistency of the detectors were derived. Using these results it was shown that the detectors are consistent for the important special cases of normally distributed jitter and uniformly distributed jitter, although mild conditions apply for the case of uniformly distributed jitter. Simulation results showed that both detectors maintain the set level for an SNR of 0 dB, although they are excessively conservative for small sample lengths. The detectors also exhibit good performance under the alternative for relatively small sample lengths.

5. REFERENCES

[1] S.S. Awad. "The Effects of Accumulated Timing Jitter on Some Sine Wave Measurements". *IEEE Transaction on Instrumentation and Measurement*, 44(5):945–951, October 1995.

[2] A. Berkovitz and I. Rusnak. "FFT Processing of Randomly Sampled Harmonic Signals". *IEEE Transactions on Signal Processing*, 40(11):2816–2819, November 1992.

[3] M.R. Morelande. "Parameter Estimation of a Complex Harmonic Observed with Independent Timing Jitter". *IEEE Transactions on Instrumentation and Measurement.* (in review).

[4] M.R. Morelande and D.R. Iskander. "Formulation and Comparison of Two Detectors of Independent Timing Jitter in a Complex Harmonic". *IEEE Transactions on Instrumentation and Measurement.* (submitted).

[5] R.J. Serfling. *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York, 1980.

[6] I. Sharfer and H. Messer. "The Bispectrum of Sampled Data: Part 1- Detection of the Sampling Jitter". *IEEE Transactions on Signal Processing*, 41(1):296–312, January 1993.

[7] I. Sharfer and H. Messer. "The Bispectrum of Sampled Data: Part 2- Monte Carlo Simulations of Detection and Estimation of the Sampling Jitter". *IEEE Transactions on Signal Processing*, 42(10):2706–2714, October 1994.

PASSIVE SIGNATURE CHARACTERIZATION AND CLASSIFICATION BY MEANS OF NONLINEAR DYNAMICS

Ron K. Lennartsson[1], *James B. Kadtke*[2] *and Áron Péntek*[2]

[1]Swedish Defence Research Agency, SE 172 90 Stockholm, Sweden.
e-mail: ron@foi.se
[2]Marine Physical Laboratory, Scripps Institution of Oceanography
University of California, San Diego, La Jolla, CA 92093-0238, USA.
e-mail: {jkadtke, apentek}@ucsd.edu

ABSTRACT

We analyze sonar recordings of various boats as well as ambient sea noise using nonlinear dynamical signal models. Specifically, we discuss the estimation of the parameters of nonlinear delay differential equations from data. Using the model parameters as classification features we implement a three class Bayesian minimum-error-rate classifier and demonstrate almost perfect classification of the data set considered. This indicates that classifiers based on nonlinear dynamical models can be useful in sonar applications.

1. INTRODUCTION

An important task in underwater passive sonar signal processing is determination of target signatures based on the narrow-band signal content in the received signal. However identification of the harmonics and their interrelations for such sources is often difficult, particularly in shallow waters where multipath propagation is inevitable and the channel may be varying considerably with target distance. In this paper we present an alternative approach motivated by recent advances in nonlinear dynamical systems theory.

We present the theory and application of a recently developed algorithm for passive signature characterization and classification using nonlinear signal models [1]. This algorithm utilizes a robust method to estimate delay differential equation (DDE) models from data. The model parameters are estimated using generalized higher-order correlation functions, similar to the Yule-Walker equations in parametric signal processing. This method involves estimation of both higher-order statistical moments and dynamical moments. The dynamical moments are also of higher order, but involve the derivative of the signal [2].

Using this theory we outline the design for practical characterization and classification algorithms for applications in passive sonar signal processing. From the model coefficients, which reflect low dimensional dynamical information in a compact way, we define a feature space. These features can easily be used for classification purposes. The subsequent partitioning of the feature space can be done by employing any standard discrimination method such as Neyman-Person, Bayesian or neural networks. In this work we implement a Bayesian minimum-error-rate classifier [3].

Finally, we apply these ideas to the analysis of real-world passive sonar recordings from the Baltic Sea off the east coast of Sweden, in shallow water of an approximately constant depth of 40 meters. The data set consists of recordings of big and small boats as well as ambient sea noise.

2. ESTIMATION OF NONLINEAR DYNAMICAL SIGNAL MODELS

Here we present a brief description of our model estimation procedure using time-domain differential equation signal models. For a more detailed description the reader is referred to [1][2]. First, we hypothesize that that we observe a scalar data stream $x(t)$ generated by some measurement of some accessible observable of a physical process. We hypothesize that the process evolution itself can be approximated by a deterministic, relatively low-dimensional dynamics, but can include purely stochastic elements (i.e. noise) as well. We will also utilize up to D time-delayed copies of $x(t)$, written $x(t - d\tau)$ with $1 \leq d \leq D$. Hence our general model

JK AND AP WOULD LIKE TO ACKNOWLEDGE GENEROUS SUPPORT FOR THIS PROJECT THROUGH THE US OFFICE OF NAVAL RESEARCH, GRANTS NUMBER N00014-99-1-0072 AND N00014-97-1-0312

0-7803-7011-2/01/$10.00 ©2001 IEEE

form is

$$\dot{x}(t) = F[x(t), x(t-\tau), ..., x(t-D\tau)]. \quad (1)$$

The function F is often expanded in terms of some basis functions. Here we restrict our attention to a two-delay second order model

$$\dot{x} = a_1 x_{\tau_1} + a_2 x_{\tau_2} + a_3 x_{\tau_1} x_{\tau_2} \quad (2)$$

where the shorthand notations $\dot{x} \equiv \dot{x}(t)$ and $x_\tau \equiv x(t-\tau)$ have been introduced. The unknown model coefficients a_1, a_2, and a_3 are estimated for each observation window, and will comprise our feature space. Below we present an estimation method that is numerically robust and can explicitly preserve some of the nonlinear correlations possibly present in the original time series. Briefly, we multiply Eq. (2) by each basis term x_{τ_1}, x_{τ_2}, and $x_{\tau_1} x_{\tau_2}$, and average over an observation window of length T; the model coefficients are then computed by solving the following linear equation:

$$\mathbf{R} * \mathbf{A} = \mathbf{B} \quad (3)$$

where

$$\mathbf{R} = \begin{pmatrix} \langle x^2 \rangle & \langle x_{\tau_1} x_{\tau_2} \rangle & \langle x_{\tau_1}^2 x_{\tau_2} \rangle \\ \langle x_{\tau_1} x_{\tau_2} \rangle & \langle x^2 \rangle & \langle x_{\tau_1} x_{\tau_2}^2 \rangle \\ \langle x_{\tau_1}^2 x_{\tau_2} \rangle & \langle x_{\tau_1} x_{\tau_2}^2 \rangle & \langle x_{\tau_1}^2 x_{\tau_2}^2 \rangle \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} \langle \dot{x} x_{\tau_1} \rangle \\ \langle \dot{x} x_{\tau_2} \rangle \\ \langle \dot{x} x_{\tau_1} x_{\tau_2} \rangle \end{pmatrix}.$$

Where $\langle \star \rangle$ stands for the expectation value. Note that the correlation involving the signal derivative can be calculated from the derivative of the correlation function, i.e.

$$\langle \dot{x} x_{\tau_1} \rangle = \frac{d}{d\tau_1} \langle x x_{\tau_1} \rangle,$$

and (4)

$$\langle \dot{x} x_{\tau_1} x_{\tau_2} \rangle = \frac{d}{d\tau_1} \langle x x_{\tau_1} x_{\tau_2} \rangle + \frac{d}{d\tau_2} \langle x x_{\tau_1} x_{\tau_2} \rangle.$$

These formulas are valid in the long window limit for a bounded stationary signal $x(t)$. The main practical advantage of using Eq. (3) is that we can avoid computing the signal derivatives, which is the main difficulty for noisy signals. The expectation values on the left hand side of Eq. (3) can be expressed as standard higher-order data moment functions [4]. We also note that the dynamical moments involving $\dot{x}$ arise exactly because of the dynamical representation and express information not utilized in standard higher order methods.

3. MINIMUM-ERROR-RATE CLASSIFICATION

The formalism of the previous section allows us to estimate the parameters of a given dynamical data model using the standard and dynamical correlation functions. Our main aim here is to use these ideas to design detectors and classifiers using standard feature discrimination methods. First, we standardize the features a_1, a_2 and a_3 by requiring that all signal observation windows are normalized to zero mean and unit variance. Thus a particular observation window can be represented as a point in the three-dimensional feature space, and the set of all the observations form a distribution for a specific class in this space.

The subsequent partitioning of the feature space, i.e. classifier design, can be done by employing any standard discrimination method such as Neyman-Person, Bayesian or neural networks. We chose a Bayesian approach to build a minimum-error-rate classifier. Let $\mathbf{\Omega} = \{\omega_1, ..., \omega_s\}$ be our set of s data classes. For each class ω_n we make N_n observations, using a fixed window length. By estimating the model parameters a_1, a_2 and a_3 with the procedure described in the previous section, we obtain for each class a set of model coefficients $\left\{\mathbf{A}_i^{(\omega_n)}\right\}_{i=1}^{N_n}$ where the vector notation $\mathbf{A} = (a_1, a_2, a_3)$ has been introduced. There are many ways to represent a minimum-error-rate classifier, one way is in terms of a set of discriminant functions $g_i(\mathbf{A})$. The classifier is said to assign a feature vector $\mathbf{A}$ to class ω_i if

$$g_i(\mathbf{A}) > g_j(\mathbf{A}) \quad \text{for all} \quad j \neq i. \quad (5)$$

A Bayesian minimum-error-rate classifier can easily be represented in this way [3]. We can simply chose $g_i(\mathbf{A}) = P(\omega_i|\mathbf{A})$ so that the maximum discriminant function corresponds to the maximum a posteriori probability. This choice of discriminant function is in no way unique. We can replace every $g_i(\mathbf{A})$ with $f(g_i(\mathbf{A}))$, where f is a monotonically increasing function, without changing the classification ability. A particular useful form is

$$g_i(\mathbf{A}) = log\, p(\mathbf{A}|\omega_i) + log\, P(\omega_i). \quad (6)$$

Further, if we assume that the feature distributions are multivariate normal and that the a priori probabilities of all classes are equal Eq. (6) becomes

$$g_i(\mathbf{A}) = -\tfrac{1}{2}\left[\mathbf{A}^t \mathbf{\Sigma}_i^{-1} \mathbf{A} - 2\mathbf{A}^t \mathbf{\Sigma}_i^{-1} \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^t \mathbf{\Sigma}_i^{-1} \boldsymbol{\mu}_i\right] - \tfrac{1}{2} log\, |\mathbf{\Sigma}_i| \quad (7)$$

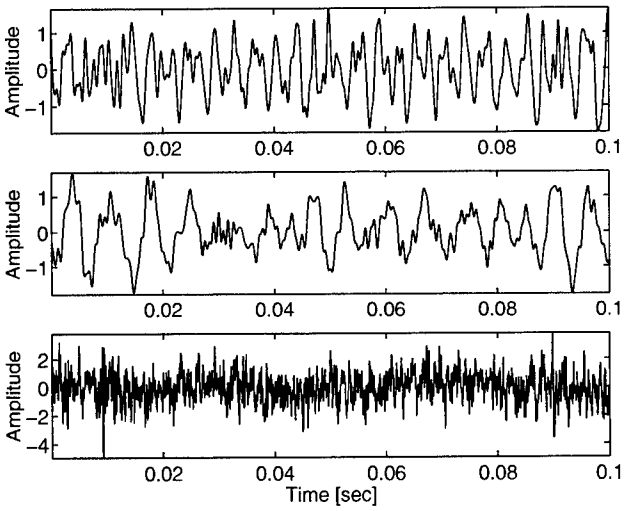


Figure 1: Typical time series from a small boat (top), a big boat (middle) and noise (bottom).

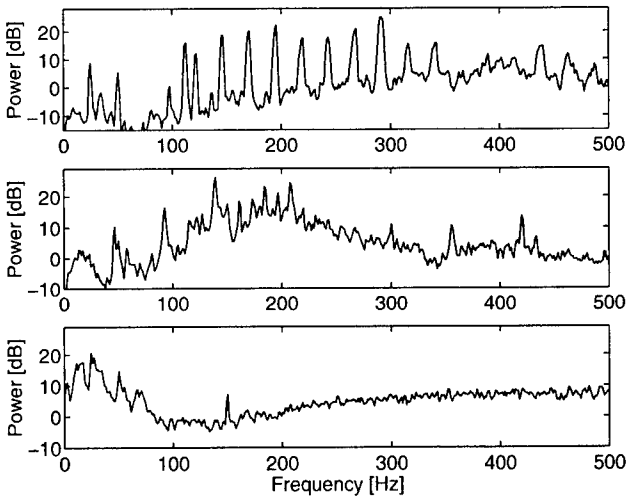


Figure 2: Typical power spectrum from a small boat (top), a big boat (middle) and noise (bottom).

where $\boldsymbol{\mu}_i$ is the mean and $\boldsymbol{\Sigma}_i$ is the covariance matrix of class i. In practical cases the means and the covariance matrices are unknown and have to be estimated from a training set, unless the analytic signal forms are known. While the distribution of the DDE model coefficients is typically not Gaussian, for low-SNR they are nearly so. We have found that alternative methods that do not rely on the normality assumption, e.g. logistic discrimination, do not show significant improvement.

4. DATA ANALYSIS

In this section we consider the analysis of a sonar data set from a sea trial, conducted by the Swedish Defence Research Agency. Our primary objective is to demonstrate robust classification using the dynamical classification method described in the previous sections.

4.1. The data set

Sonar recordings were performed in the Baltic Sea off the east coast of Sweden, in shallow waters of an approximately constant depth of 40 meters. The data set consists of hydrophone recordings from three small boats and one big boat passing over the hydrophone as well as ambient sea noise. All data were recorded with the sampling rate of 20 kHz. During the recordings the data were low-pass filtered 0 to 6 kHz, hence there is no problem with aliasing. In Fig. 1 typical time series from a small boat, big boat and noise are displayed. Corresponding power spectra are displayed in Fig. 2.

4.2. Classification of real-world sonar data

We now describe the classification of the real-world sonar data utilizing the two-delay second order DDE model given by Eq. (2). First we select one segment of length 40 seconds from each boat recording (the closest point of approach is included in the segment) and a noise segment of length 60 seconds. The data segments are then windowed into 1 second (20000 samples) observation windows, with a 0.5 second (10000 samples) window shift to provide independent samples. The three parameters a_1, a_2 and a_3 are estimated with equation Eq. (3). To solve Eq. (3) all the moments in the matrix equation have to be estimated, and we use an unbiased estimate defined as

$$\langle x^a(n)x^b(n-i)x^c(n-j)\rangle = \frac{1}{N-m}\sum_{n=m}^{N-1} x^a(n)x^b(n-i)x^c(n-j) \tag{8}$$

where m is equal to the largest of i and j; N is the window length and i and j are the discrete delays corresponding to τ_1 and τ_2 respectively; the powers a, b and c are set to 0, 1 or 2 corresponding to the moment that has to be calculate. The window length and the delays have to be tuned to the signal of interest. We use the window length $N = 20000$ samples and use a subset (12.5%) of the small boat recordings to select the two delays from the maximum significance of L, where L is given by

$$L = \sqrt{a_1^2(i,j) + a_2^2(i,j) + a_3^2(i,j)} \tag{9}$$

The significance of L is displayed in Fig. 3. From the figure we can identify a maxima at $i = 7$ and $j = 33$

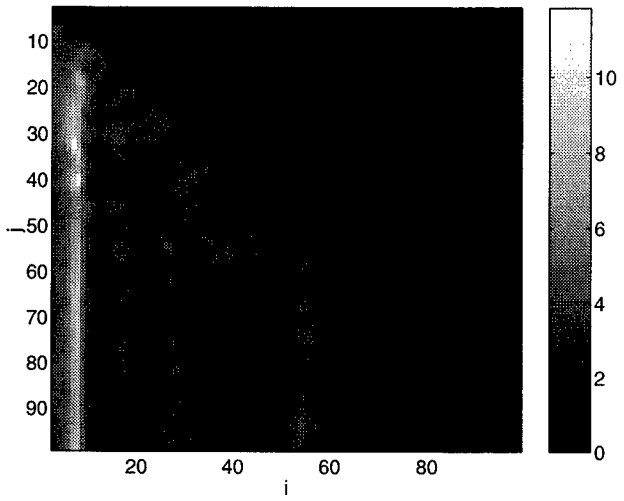


Figure 3: Significance of L estimated from a subclass of the small boats.

Output Decision⟹ True Classes⇓	Noise	Small boat	Big boat
Noise	100%	0%	0%
Small boat	0%	100%	0%
Big boat	0%	1%	99%

Table 1: The confusion matrix shows that the dynamical classifier provides the correct class decision in virtually all cases.

samples. The feature space spanned by the three model parameters is shown in Fig. 4. One can observe clear separation between the three data classes.

Next, we implement a three class minimum-error-rate classifier following the outline in section 3. The three classes are small boats, big boats and noise. A randomly selected subset (70%) from all classes were used as a training set (i.e. for estimating the mean and covariance matrices). The remaining 30% of the distributions were then used for testing of the classifier. The training and testing were repeated a hundred times to remove fluctuations in the classification performance. The results of the numerical analysis above is summarized in the confusion matrix of Table 1. This matrix consists of a table showing the true class of the input features, versus the output of the classification algorithm using the testing set of features. As can be seen this method provides almost perfect classification of the data set.

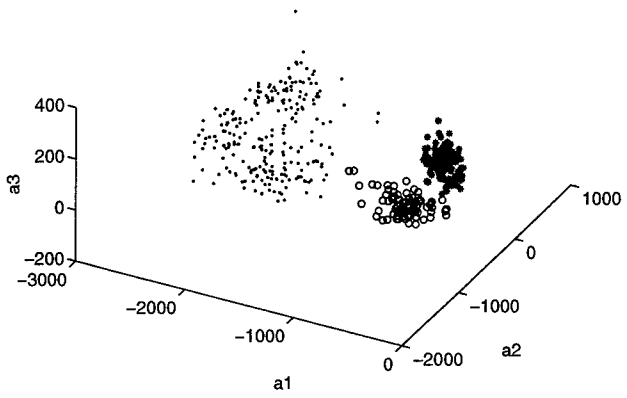


Figure 4: Parameter distributions for small boats (dots), big boats (circles) and noise (stars).

5. CONCLUSIONS

We have discussed a method for the estimation of DDE signal models motivated by the Yule-Walker equations, which provides computational speed, numerical stability and noise robustness. The model parameters can be used to represent a wide range of signals, further they can be used for detection and classification purposes. In this paper we presented a classification study of real-world sonar data, derived from a sea trial in the Baltic Sea. We implemented a three class minimum-error-rate classifier based on a two-delay second order DDE model, and showed almost perfect separation between the classes. This implies that classifiers based on nonlinear dynamical models can be useful in sonar applications.

REFERENCES

[1] Kadtke J., Pentek A., *Automated signal classification using dynamical signal models and generalized higher-order data correlations (U)*, USN Journal of Underwater Acoustics, in press (2000).

[2] Kadtke J., Kremliovsky M., *Estimating dynamical models using generalized moment functions*, Physics Letters A **260**, 203 (1999).

[3] Duda, R.O., Hart P.E.,*Pattern classification and scene analysis*, John Wiley & Sons, 1973

[4] Boashash, B., Edward J.P., Abdelhak, M.Z., eds., *Higher-order statistical signal processing*, Longman and Wiley Press, Melbourne, 1995.

MULTICHANNEL DETECTION AND SPATIAL SIGNATURE ESTIMATION WITH UNCALIBRATED RECEIVERS

Amir Leshem[1,2] *and Alle-Jan van der Veen*[1]

[1] Delft University of Technology, Dept. Electrical Engineering/DIMES, 2628 CD Delft, The Netherlands
[2] Metalink Broadband Access, Yakum Business Park 60972, Israel

Abstract A problem occuring in radio astronomy is the detection and cancellation of spatially correlated interfering signals entering via the sidelobes of the telescopes in an array. A complicating factor is that the noise powers can be different at each telescope. For the case that the sensors are uncalibrated, we formulate the detection problem as a test on the covariance structure, state the GLRT for this problem, and relate it to a simpler ad-hoc detector. We derive algorithms to estimate the noise powers and the subspace of interferer signature vectors. Once the subspace is estimated, the interference can be projected out. We compare this method to the conventional multichannel subspace detector and show its robustness to non-identical channels on data collected with the Westerbork radio telescope.

1. INTRODUCTION

In this paper we study the detection and suppression of spatially correlated signals impinging on an array of uncalibrated non-identical sensors, in the presence of spatially uncorrelated noise. The noise covariance matrix is diagonal but otherwise unknown.

The motivation for this study comes from an application in radio astronomy, where we wish to detect and suppress man-made interfering sources impinging on an array of telescopes. The output of the receiver after processing is essentially a sequence of short-term (~10 second) sample correlation matrices, composed of the contributions of astronomical sources in the pointing direction, the additive receiver noise, and the interference. The receiver noise is largely independent among the sensors, but the receiver gains are not identical, with differences of up to a few dB. Until now, calibration of this has been done separately and taken into account off-line. An interfering source is usually in the near field and received through the side-lobes of the parabolic dishes, hence the received signals are correlated but with arbitrary unknown gains. Our aim is to detect and cancel the interference online; this requires online calibration processing as well.

Two types of interference play a role: intermittent signals (e.g., TDMA signals as in the GSM system, certain radar signals) and continuously present signals (e.g., television signals, GPS). Our approach for intermittent signals is to detect their presence on-line on milli-second periods, and discard those periods which are deemed contaminated (temporal excision) [1]. For continuous interference, we also wish to estimate the signature (direction) vector, so that we can project out that dimension from the data. This is more ambitious, and also requires modifications to the way the astronomical data is processed after recording [2]. Note that the astronomical signals of interest are much weaker than the receiver noise and hence it is necessary to detect interference even if it is much below the noise power. The astronomical signals themselves are too weak to be detected at these short time scales.

When the interferers are weaker than the system noise and the receivers are non-identical, the change in eigenstructure of the sample covariance matrix is not detectable unless one of two steps is taken. The first is pre-calibration and whitening. The second which is easier to implement on-line is to use a different model where the noise covariance matrix is assumed diagonal but not necessarily equal to $\sigma^2\mathbf{I}$, and to detect deviation from this nominal model. This is the approach taken here. The Generalized Likelihood Ratio Test (GLRT) for this problem turns out to be the determinant of the sample correlation matrix, a fact which is not very well known in signal processing but has been used for a long time in certain other disciplines.

We demonstrate the results of the excision using the GLRT detector and compare it to a detector which assumes identical receivers. We also demonstrate the improvement in the estimate of the spatial signature as compared to the usual eigendecomposition technique.

2. PROBLEM FORMULATION

Assume that we have a set of q narrow-band Gaussian signals impinging on an array of p sensors. The received signal can be described in complex envelope form by

$$\mathbf{x}(k) = \sum_{i=1}^{q} \mathbf{a}_i s_i(k) + \mathbf{n}(k) = \mathbf{A}\mathbf{s}(k) + \mathbf{n}(k) \qquad (1)$$

where $\mathbf{x}(k) = [x_1(k), \cdots, x_p(k)]^T$ is a $p \times 1$ vector of received signals at sample times k, $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_q]$, where $\mathbf{a}_i$ is the array response vector for the i'th signal, $\mathbf{s}(k) = [s_1(k), \cdots, s_q(k)]^T$ is a $q \times 1$ vector of gaussian source signals at sample times k with covariance matrix $\mathbf{R}_s = E(\mathbf{s}\mathbf{s}^H)$, $\mathbf{n}(k)$ is the $p \times 1$ additive noise vector, which is assumed to have independent gaussian entries with unknown diagonal covariance matrix $\mathbf{R}_n = \mathrm{diag}\{\mathrm{v}_1, \ldots, \mathrm{v}_p\}$.

We would like to detect the presence of signals satisfying the above model, i.e., given data vectors $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ decide whether $q = 0$ or $q > 0$. Secondly, if $q > 0$, we would like to detect q and estimate the interfering subspace, i.e., span($\mathbf{A}$), so that we can project out this subspace from the data. We do not assume parametric knowledge of the array manifold (since the interferers enter in the side lobes) or a calibration of the noise power in each channel. Under these assumptions the only way to distinguish between signal and noise is to use the fact that the noise is spatially uncorrelated, hence has a diagonal covariance matrix.

The detection problem is thus given by a collection of hypotheses ($\mathcal{CN}(0, \mathbf{R})$ denotes the zero-mean complex normal distribution with covariance $\mathbf{R}$)

$$\begin{aligned} \mathcal{H}_q &: \mathbf{x}(k) \sim \mathcal{CN}(0, \mathbf{R}_q) \\ \mathcal{H}^I &: \mathbf{x}(k) \sim \mathcal{CN}(0, \mathbf{R}^I), \qquad q = 1, 2, \cdots \end{aligned} \qquad (2)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

where $\mathbf{R}_q$ is the covariance matrix of the model with q interferers,

$$\mathbf{R}_q = \mathbf{A}\mathbf{A}^{\mathrm{H}} + \mathbf{D}, \quad \text{where} \quad \mathbf{A}: p \times q, \quad \mathbf{D} \text{ diagonal}$$

and $\mathcal{H}'$ corresponds to a default hypothesis of an arbitrary (unstructured) positive definite matrix $\mathbf{R}'$. (Without loss of generality, we absorbed the interferer covariance matrix $\mathbf{R}_s$ in $\mathbf{A}$.)

As it turns out, this problem has been studied in the psychometrics, biometrics and statistics literature since the 1930s under the heading of *factor analysis* (but usually for real-valued matrices) [3, 4]. The problem has received much less attention in the signal processing literature. Related recent work includes e.g. direction estimation using two subarrays with mutually uncorrelated noise [5,6].

3. THE GLRT DETECTOR

In this section we give a short derivation of the GLRT for the detection problem $\mathcal{H}_q$ versus $\mathcal{H}'$. Note that both hypotheses are composite and we have to derive maximum likelihood estimates of the parameters for each of the hypotheses. Under $\mathcal{H}_q$, the likelihood function is given by

$$L(\mathbf{X}|\mathcal{H}_q) \equiv L(\mathbf{X}|\mathbf{R}_q) = \left(\frac{1}{|\mathbf{R}_q|} e^{-\mathrm{tr}(\mathbf{R}_q^{-1}\hat{\mathbf{R}})} \right)^N,$$

where $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(N)]$ and $\hat{\mathbf{R}} = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}(k)\mathbf{x}(k)^H$ is the sample covariance matrix, $|\cdot|$ denotes the determinant and $\mathrm{tr}(\cdot)$ the trace operator.

The ML estimate of $\mathbf{R}_q$ is found by maximizing $L(\mathbf{X}|\mathbf{R}_q)$ over the parameters of the model $\mathbf{R}_q = \mathbf{A}\mathbf{A}^{\mathrm{H}} + \mathbf{D}$, or equivalently the log-likelihood function

$$\mathcal{L}(\mathbf{X}|\mathbf{R}_q) = N\left(-\ln|\mathbf{R}_q| - \mathrm{tr}(\mathbf{R}_q^{-1}\hat{\mathbf{R}})\right).$$

Denote the estimate by $\hat{\mathbf{R}}_q = \hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathrm{H}} + \hat{\mathbf{D}}$. Under $\mathcal{H}'$ we obtain that the ML estimate of $\mathbf{R}'$ is given by $\hat{\mathbf{R}}$, the sample covariance matrix. The log-likelihood GLRT test statistic is thus given by

$$\ln \frac{L(\mathbf{X}|\mathcal{H}_q)}{L(\mathbf{X}|\mathcal{H}')} = -N\left(\mathrm{tr}(\hat{\mathbf{R}}_q^{-1}\hat{\mathbf{R}}) - \ln|\hat{\mathbf{R}}_q^{-1}\hat{\mathbf{R}}| - p\right).$$

A further result is that the ML estimate of $\hat{\mathbf{R}}_q$ is such that $\mathrm{tr}(\hat{\mathbf{R}}_q^{-1}\hat{\mathbf{R}}) = p$ so that we can base the test on

$$T_q(\mathbf{X}) := N \ln|\hat{\mathbf{R}}_q^{-1}\hat{\mathbf{R}}|. \tag{3}$$

If we generalize the results in [3,4] to complex data, we obtain the following.

Lemma 3.1 *If $\mathcal{H}_q$ is true and N is moderately large (say $N-q \geq 50$), then $2T_q(\mathbf{X})$ has approximately a χ^2_ν distribution with $\nu = (p-q)^2 - p$ degrees of freedom.*

In view of results of Box and Bartlett, a better fit is obtained by replacing N in (3) by [3]

$$N' = N - \frac{1}{6}(2p+5) - \frac{2}{3}q.$$

This provides a threshold for a test of $\mathcal{H}_q$ versus $\mathcal{H}'$ corresponding to a desired probability of false alarm P_{FA}. The test replaces the more familiar eigenvalue test on the rank of $\hat{\mathbf{R}}$ in the case of white noise, $\mathbf{D} = \sigma^2\mathbf{I}$. Note that before we can perform the test, we need to compute the ML estimates of $\mathbf{A}: p \times q$ and $\mathbf{D}$ (see section 5).

4. TEST FOR DIAGONALITY

Under $\mathcal{H}_0$ we can make the test more explicit. To estimate $\hat{\mathbf{R}}_0 = \hat{\mathbf{D}}$, we set the derivative of $\mathcal{L}$ with respect to the parameters of $\mathbf{D}$ to zero, which immediately gives $\hat{\mathbf{D}} = \mathrm{diag}(\hat{\mathbf{R}})$. Therefore the GLRT test statistic is given by

$$\frac{L(\mathbf{X}|\mathcal{H}_0)}{L(\mathbf{X}|\mathcal{H}_1)} = \frac{|\hat{\mathbf{R}}|^N}{\prod_{i=1}^{p} \hat{\mathbf{R}}_{ii}^N} = |\hat{\mathbf{C}}|^N, \tag{4}$$

where $\hat{\mathbf{C}}$ is the sample correlation matrix given by $\hat{\mathbf{C}} = \mathbf{W}\hat{\mathbf{R}}\mathbf{W}$ and $\mathbf{W} = \mathrm{diag}\{\hat{r}_{11}, \ldots, \hat{r}_{pp}\}^{-1/2}$. Note that $0 \leq |\hat{\mathbf{C}}| \leq 1$, where equality to 1 is obtained asymptotically for $N \to \infty$ if $q = 0$. Thus, for a certain threshold $\gamma = \gamma(N)$ between 0 and 1, the GLRT is

$$T_1 \equiv |\hat{\mathbf{C}}| \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \gamma \tag{5}$$

This result is identical to that in the real-valued case (see [4, p.137]). The expression is rather satisfactory since in the absence of sensor calibration data all the spatial information exists in the spatial correlation coefficients between the different sensors, and the GLRT suggests a proper way of combining these different correlations. It is also quite easy to implement and does not involve any eigenstructure computations. From lemma 3.1, under $\mathcal{H}_0$ we know that $-2N\ln|\hat{\mathbf{C}}|$ has asymptotically a chi-square distribution with $p^2 - p$ degrees of freedom. Again, a better asymptotic fit is obtained by replacing N by $N' = N - \frac{1}{6}(2p+11)$.

A related ad-hoc detector to which we can compare is based on the Frobenius-norm of the off-diagonal entries of $\hat{\mathbf{C}}$. Since the diagonal entries are equal to 1, it is equivalent to take the norm of $\hat{\mathbf{C}}$ itself, i.e.,

$$T_2 \equiv \|\hat{\mathbf{C}}\|_F \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma' \tag{6}$$

In fact, it is straightforward to prove that, for weak signals, the performance of this detector must be approximately equal to that of the GLRT. Indeed, for weak signals, the eigenvalues of $\hat{\mathbf{C}}$ are equal to $\lambda_i = 1 + \varepsilon_i$, for small ε_i. Note that $\mathrm{tr}(\hat{\mathbf{C}}) = p \Rightarrow \sum_i \lambda_i = p \Rightarrow \sum_i \varepsilon_i = 0$. We can write

$$\begin{aligned} T_1 &= \textstyle\prod_i \lambda_i = e^{\sum_i \ln\lambda_i} \\ \Rightarrow \quad \ln(T_1) &= \textstyle\sum_i \ln\lambda_i = \sum_i \varepsilon_i - \frac{1}{2}\varepsilon_i^2 + \mathcal{O}(\varepsilon_i^3) \\ &= \textstyle -\sum_i \frac{1}{2}\varepsilon_i^2 + \mathcal{O}(\varepsilon_i^3) \end{aligned}$$

whereas

$$\begin{aligned} T_2^2 &= \|\hat{\mathbf{C}}\|_F^2 = \textstyle\sum_i \lambda_i^2 = \sum_i 1 + 2\varepsilon_i + \varepsilon_i^2 \\ \Rightarrow \quad -\tfrac{1}{2}(T_2^2 - p) &= \textstyle -\sum_i \frac{1}{2}\varepsilon_i^2 \end{aligned}$$

Since a monotonic transformation of a test statistic does not change the outcome of the test if the threshold is modified accordingly,[1] the two detectors are equivalent up to third order. Computing the Frobenius-norm requires only $\mathcal{O}(p^2)$ operations, versus $\mathcal{O}(p^3)$ for the determinant test (implemented via a Cholesky factorization of $\hat{\mathbf{C}}$).

[1] Note that the decisions in (5) and (6) are opposite, hence the change of sign in the second transformation.

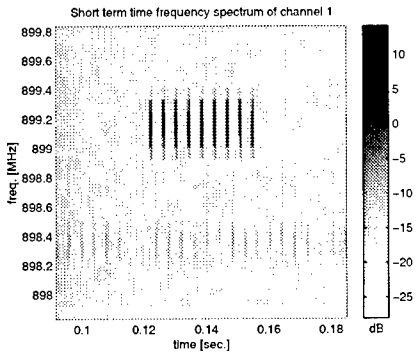


Figure 1. Time-frequency spectrum of channel 1, showing GSM interference

5. PARAMETER ESTIMATION

To enable the GLRT, we have to find ML estimates of the factors $\mathbf{A}: p \times q$ and $\mathbf{D}$, both dependent on the choice of q. The largest permissible value of q is that for which the number of degrees of freedom $\nu = (p-q)^2 - p \geq 0$, or $q \leq p - \sqrt{p}$. For larger q, there is no identifiability of $\mathbf{A}$ and $\mathbf{D}$: any sample covariance matrix $\hat{\mathbf{R}}$ can be fitted. Even for smaller q, $\mathbf{A}$ can be identified only up to a $q \times q$ unitary transformation at the right, i.e., we can identify span($\mathbf{A}$). This generalizes the white noise case (where span($\mathbf{A}$) would be given by the eigenvectors of $\hat{\mathbf{R}}$), and is sufficient for our application of interference cancellation.

For $q > 0$, there is no closed form solution to the estimation of the factors $\mathbf{A}$ and $\mathbf{D}$ in the ML estimation of $\hat{\mathbf{R}}_q = \hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathrm{H}} + \hat{\mathbf{D}}$. There are several approaches for this:

- Suppose that the optimal ML-estimate $\hat{\mathbf{D}}$ has been found. We can then whiten $\hat{\mathbf{R}}$ to $\tilde{\mathbf{R}} = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{R}}\hat{\mathbf{D}}^{-1/2}$, and similarly the model, giving $\tilde{\mathbf{R}}_q = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\mathrm{H}} + \mathbf{I}$. Note that $|\hat{\mathbf{R}}_q^{-1}\hat{\mathbf{R}}| = |\tilde{\mathbf{R}}_q^{-1}\tilde{\mathbf{R}}|$, which is the usual problem for white noise, solved via an eigenvalue decomposition of $\tilde{\mathbf{R}}$. This is equivalent to solving $\min \|\tilde{\mathbf{R}} - (\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\mathrm{H}} + \mathbf{I})\|_F^2$. Since $\hat{\mathbf{D}}$ is not known, this leads to an iteration where $\tilde{\mathbf{A}}$ is plugged back, $\hat{\mathbf{D}}$ is estimated, etc.

  A related technique is alternating least squares, where we alternatingly minimize $\|\hat{\mathbf{R}} - \mathbf{A}\mathbf{A}^{\mathrm{H}} + \mathbf{D}\|_F^2$ over $\mathbf{A}$ keeping $\mathbf{D}$ fixed, and over $\mathbf{D}$ keeping $\mathbf{A}$ fixed. (This is not equivalent to the determinant cost function unless a weighting by $\mathbf{D}^{-1/2}$ is used.) Both iterative techniques tend to be very slow.

- Gauss-Newton iterations on the original (determinant) cost function, or on the (weighted) least squares cost. This require an accurate starting point.

- Ad-hoc techniques for solving the least squares problem, possibly followed by a Gauss-Newton iteration. These techniques try to modify the diagonal of $\hat{\mathbf{R}}$ such that the modified matrix is low-rank q, hence can be factored as $\mathbf{A}\mathbf{A}^{\mathrm{H}}$. For this we can exploit the fact that submatrices away from the main diagonal with $q+1$ columns have rank q. See [7] for an example with $q = 1$.

More details on estimation algorithms will appear in an extended version of this paper.

6. APPLICATION TO RADIO ASTRONOMY

The main motivation for the detection and subspace estimation problem stems from applications to interference mitigation in radio

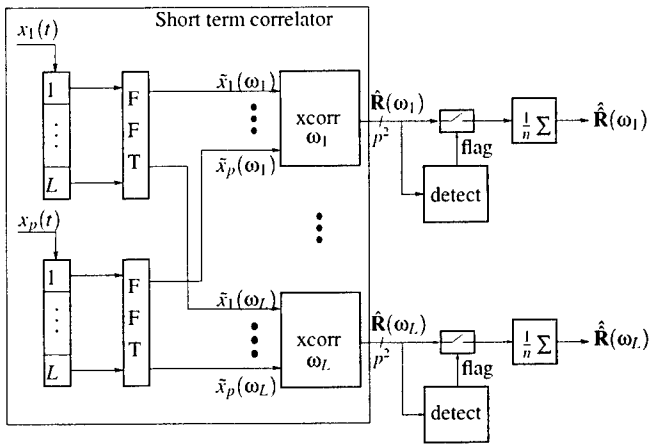


Figure 2. Computational structure of the blanking process

astronomy. We give two illustrations.

We first apply the detector for $\mathcal{H}_0$ to sample data collected with the Westerbork radio telescope. The data was recorded using the 8-channel NOEMI project data recorder [1]. We selected a bandwidth of 2 MHz, around 899 MHz, with a duration of 3 seconds. This band is contaminated with various GSM mobile telephony signals. Such signals are intermittent, occupying time slots of length 0.577 ms in frames of 4.6 ms. A segment of the data is shown in figure 1. The received data channels were split into subbands of 83 kHz by means of windowing and short-term FFTs, and subsequently correlated per frequency bin. Each covariance matrix is an average based on 21 samples and covers a period of 0.24 ms.

Our aim is to test for the presence of interference in each covariance matrix. Only if no interference is detected, the block is passed to a long-term correlator. Two detectors have been applied. The first is the detector of (4), and the other one is given by

$$T_3 \equiv \frac{|\hat{\mathbf{R}}|}{[\frac{1}{p}\mathrm{tr}(\hat{\mathbf{R}})]^p}. \tag{7}$$

This detector is a GLRT assuming identical channels (or $\mathbf{D} = \sigma^2\mathbf{I}$) [4].

Since $N = 21$ is small, we have not used the theoretical thresholds. Instead, we have excised the worst 10 percent of the data at each frequency channel and generated spectral estimates by further averaging the covariance matrices of the remaining 90 percent of the data. The processing structure is shown in figure 2.

Figure 3 shows the power spectrum of channel 1 and the cross-spectrum of channels 1 and 3, respectively, before and after blanking. Without excision, we can see that several interfering signals are present, most weak but one rather strong. We can clearly see that while both detectors excised properly the strong interference, the detector based on the $\mathbf{D} = \sigma^2\mathbf{I}$ assumption failed to excise the weak features of the interference.

In a second application, we wish to spatially filter out continuously present interference. The approach is to estimate span($\mathbf{A}$), and to apply a projector $\mathbf{P}_{\hat{\mathbf{A}}}^{\perp}$ onto the orthogonal complement of the span. Here, we describe only a limited-scope simulation on synthetic data, where we estimate a rank-1 subspace (*i*) using factor analysis, and for comparison (*ii*) using eigendecomposition assuming that $\mathbf{D} = \sigma^2\mathbf{I}$, or (*iii*) using eigendecomposition after whitening by $\mathbf{D}^{-1/2}$, assuming the true $\mathbf{D}$ is known from calibration. The algorithm used for factor analysis is a non-iterative ad hoc technique

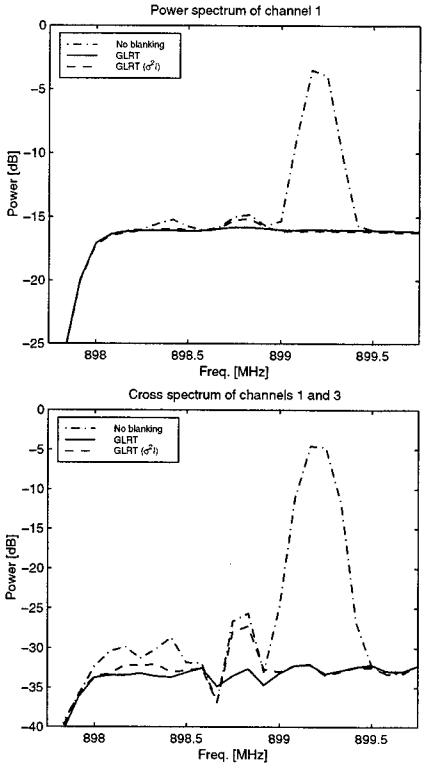


Figure 3. Power spectra and cross-spectra of channels 1 and 3, before and after interference excision

used to obtain a consistent initial estimate, followed by a Gauss-Newton optimization of the weighted least squares cost function (3 iterations). The weighting is by $\hat{\mathbf{D}}^{-1/2}$ as obtained from the ad hoc technique. We have generated covariance matrices based on the model (1) with $q = 1$, and show the residual interference power after projection, i.e., $||\mathbf{P}_{\hat{\mathbf{a}}}^{\perp}\mathbf{a}||$ as a function of number of samples N, mean noise power, and deviation in noise power. The noise powers are randomly generated at the beginning of the simulation, uniformly in an interval. Legends in the graphs indicate the nominal noise power and the maximal deviation. All simulations use $p = 8$ sensors and $q = 1$ interferer, and a nominal interference to noise ratio per channel of 0 dB.

The results are shown in figure 4. The first graph shows the residual interference power for varying maximal deviations, the second graph shows the residual for varying number of samples N, and a maximal deviation of 3 dB of the noise powers. The figures indicate that already for small deviations of the noise powers it is essential to take this into account. Furthermore, the estimates from the factor analysis are nearly as good as can be obtained via whitening with known noise powers.

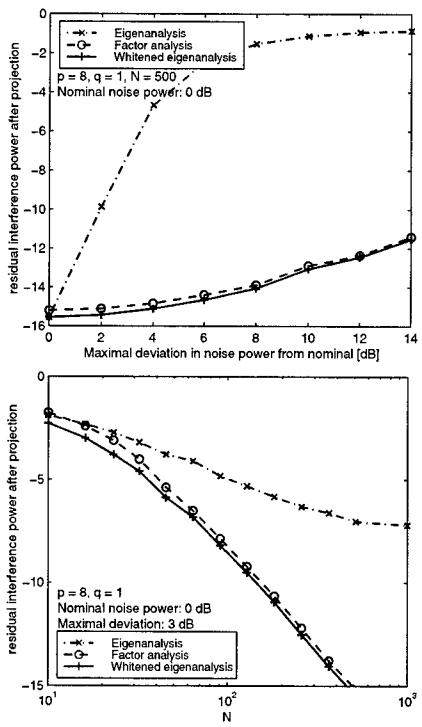


Figure 4. Residual interference power after projections

Acknowledgement

We are grateful to L.L. Scharf for pointing us to the topic of factor analysis and sharing his notes on this.

REFERENCES

[1] A. Leshem, A-J. van der Veen, and A-J. Boonstra, "Multichannel interference mitigation techniques in radio astronomy," *The Astrophysical Journal Supplements*, Nov. 2000.

[2] A. Leshem and A-J. van der Veen, "Radio-astronomical imaging in the presence of strong radio- interference," *IEEE Trans. Informat. Th.*, pp. 1730–1747, August 2000.

[3] D.N. Lawley and A.E. Maxwell, *Factor Analysis as a Statistical Method.* Butterworth and Co, 1963.

[4] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis.* Academic Press, 1979.

[5] Q. Wu and K.M. Wong, "UN-MUSIC and UN-CLE: An application of generalized correlation analysis to the estimation of the direction of arrival of signals in unknown correlated noise," *IEEE Trans. Signal processing*, vol. 42, pp. 2331–2343, Sept. 1994.

[6] P. Stoica and M. Cedervall, "Detection tests for array processing in unknown correlated noise fields," *IEEE Trans. Signal Processing*, vol. 45, pp. 2351–2362, Sept. 1997.

[7] A.J. Boonstra and A.J. van der Veen, "Gain decomposition methods for radio astronomy," in *Submitted to Proc. Workshop IEEE Stat. Signal Proc.*, (Singapore), Aug. 2001.

A New Wavelet-Based Tracking Algorithm for Rapidly Time-Varying Systems

Yuanjin Zheng[1] and Zhiping Lin[2]
[1]Institute of Microelectronics, 11 Science Park Road, Singapore Science Park II, SINGAPORE, 117685
[2] School of Electrical and Electronic Engineering, Nanyang Technological University, SINGAPORE, 639798
yuanjin@ime.org.sg[1], EZPLin@ntu.edu.sg[2]

Abstract— **In this paper, an RLS algorithm with selective modification of the system estimation covariance matrix is employed to track the rapidly changing components of system parameters. A new on-line wavelet detector is designed for accurately identifying the changing locations and the branches of changing parameters. Employing theses techniques, the tracking performance of the proposed algorithm to rapidly changing systems can be significantly improved.**

Keywords— **RLS algorithm, dyadic wavelet transform, recursive wavelet change detector, covariance matrix modification.**

I. Introduction

When a time-varying system is subject to rare but abrupt (jumping) changes, the estimated parameters by conventional adaptive algorithms cannot track the variations of true system parameters in the vicinity of these jumping locations, resulting in the so called 'lag' estimation. Three methods can be used to mitigate the effect of 'lag' estimation. One is to use variable forgetting factor RLS algorithms [1]. The second is to increase the system estimation covariance matrix at the jumping locations [2], [3]. The third includes various Bayesian Kalman filtering algorithms [4], [5]. In this paper, the second method will be adopted to track the abrupt changes of system parameters. One of the difficulty of this method is how to identify on-line the locations of the abrupt changes unknown to users. Some approaches have been developed towards this task [6]-[8]. However, the obvious trade-off between detection sensitivity and robustness exists in these methods. The design of a simple but efficient detection and modification algorithm need further investigation.

To identify the rapidly changing points effectively, a new on-line detection algorithm based on a multiscale product sequence in wavelet domain is proposed in this paper. The proposed wavelet detector can efficiently suppress background noise and enhance the abruptly changing components so that it is very robust to interferences and sensitive to jumping changes compared with the conventional detectors. A new algorithm for selectively modifying the elements of the covariance matrix is proposed.

II. Recursive Adaptive Algorithm for Rapidly Changing (Jumping) Systems

A. Tracking by changing point detection

A time-varying system commonly can be represented by a linear regression equation and the changes of system parameters can be modeled with a order one (first order) random walk model [2],

$$\theta_{t+1} = \theta_t + w_t, \quad (1)$$
$$y_t = \varphi_t^T \theta_t + e_t. \quad (2)$$

Where, θ_t is the true system parameter vector of size $N \times 1$, y_t is the scalar observation (output) signal, φ_t is the system (input) regressor vector of size $N \times 1$, w_t is the system noise vector of size $N \times 1$ and e_t is the measurement noise signal. When the variations of system parameters are slow enough, an RLS algorithm can be used to track the time-varying system [2],

$$\hat{\theta}_t = \hat{\theta}_{t-1} + G_t(y_t - \varphi_t^T \hat{\theta}_{t-1}), \quad (3)$$
$$G_t = P_t \varphi_t = \frac{P_{t-1}\varphi_t}{\lambda_t + \varphi_t^T P_{t-1}\varphi_t}, \quad (4)$$
$$P_t = \frac{1}{\lambda_t}(P_{t-1} - \frac{P_{t-1}\varphi_t\varphi_t^T P_{t-1}}{\lambda_t + \varphi_t^T P_{t-1}\varphi_t}), \quad (5)$$

where $\varepsilon_t = y_t - \varphi_t^T \hat{\theta}_{t-1}$ is the *a priori* prediction error, G_t is the filtering gain, P_t is the estimation covariance matrix, and λ_t is the forgetting factor.

When the RLS algorithm ((3)-(5)) is used for tracking a rapidly changing system, the estimation covariance matrix P_t or P_{t-1} can be increased at the locations of jumping points so that the filtering gain can be increased significantly to track the rapidly changing components [2]. When using this methods, the jumping points are needed to be known *a priori* and commonly this is unrealistic in practice. Therefore, an recursive parameter change detection algorithm is required to on-line identify the locations of jumping points. Some recursive change detection algorithms have been developed in [6]-[8]. An attractive method among them is the one used by Trigg and Leach (T & L) [6]. In this method, two filtering of the prediction error signal ε_t are used

$$\varepsilon_t^o = (1-\gamma)\varepsilon_{t-1}^o + \gamma\varepsilon_t, \quad (6)$$
$$\varepsilon_t^a = (1-\gamma)\varepsilon_{t-1}^a + \gamma|\varepsilon_t|, \quad (7)$$

where $|.|$ denotes absolute value and γ takes a small value of greater but very close to 0 (commonly $0.005 \leq \gamma \leq 0.05$). The T & L detection signal is defined as [6]

$$d_t = \frac{\varepsilon_t^o}{\varepsilon_t^a}. \quad (8)$$

According to the cental limiting theorem, d_t is asymptotically Gaussian distributed. It is shown in [6] and [7] that, for small γ, d_t is a zero mean signal with variance approximately as

$$\mathrm{Var}(d_t) = E(d_t^2) \approx \frac{\pi}{2}\frac{\gamma}{2-\gamma}. \quad (9)$$

The detection signal d_t will hence fluctuate around zero when no change has occurred. If the true parameters change, successive prediction errors are likely to have same sign and hence $|d_t|$ will increase. Assume a detection threshold is r (which can be evaluated using Chebyshev's inequality and (9), see[7]). When $|d_t| > r$ at time index t, it is considered a parameter change has happened, and P_t or P_{t-1} will increase a value by Δ_t at this time index [2].

The merit of this detector is that it is computationally simple, recursive, and the variance of detection signal is not relevant to the prediction error signal. However,there exists a trade-off between the false alarm probability and the detection probability of the T & L detector. If we want to decrease the false alarm probability, we must increase the detection threshold. This will

0-7803-7011-2/01/$10.00 ©2001 IEEE

increase the miss alarm probability and thus decease the detection probability. In the following subsection, we will develop a wavelet domain change detection algorithm which can achieve much higher detection probability.

B. Recursive DWT algorithm and multiscale product sequence

The fast algorithm of DWT has been implemented via discrete digital filters in [11], [12]. However, it is an *iterative* procedure which only can process a batch of signals. For detecting changing points on-line, a *recursive* DWT algorithm should be developed which can be summarized as the following theorem.

Theorem 1: A causal DWT coefficients $\bar{z}_t(j)$ of f_t at time t and scale j can be calculated as

$$\begin{aligned}\bar{z}_t(j) &= \bar{W}_{2^j} f_t \\ &= \sum_{k=0}^{2^j} h_k^e(j) f_{t-k}, \end{aligned} \tag{10}$$

where $h_k^e(j), k = 0, \cdots, 2^j$ is an equivalent causal filter for scale j. Further, $\bar{z}_t(j)$ can be recursively calculated as

$$\Omega^{(t)}(j) = \tilde{\Omega}^{(t-1)}(j) + f_t H^e(j), \tag{11}$$

$$\bar{z}_t(j) = \Omega_1^{(t)}(j). \tag{12}$$

Here $\Omega^{(t)}(j)$, $\tilde{\Omega}^{(t-1)}(j)$, and $H^e(j)$ are column vectors of size $2^j \times 1$ which are defined as:

$$\Omega^{(t)}(j) = [\bar{Z}_t^{(t)}(j), \cdots, \bar{Z}_{t+2^j-1}^{(t)}(j)]', \tag{13}$$

$$\tilde{\Omega}^{(t-1)}(j) = [\bar{Z}_{t+1}^{(t-1)}(j), \cdots, \bar{Z}_{t+2^j-1}^{(t-1)}(j), 0]', \tag{14}$$

$$H^e(j) = [h_0^e(j), \cdots, h_{2^j-1}^e(j)]', \tag{15}$$

and $\bar{Z}_k^{(t)}(j), 1 \le k \le t+2^j-2$ denote the causal DWT sequence of $Y_f^t = [f_1, \cdots, f_t]$ using filter $h_k^e(j)$.

The proof of the above theorem is omitted here to save space. See [14] for details. It shows that $\bar{z}_t(j)$ at time t can be recursively calculated on-line using (11) and (12) once a new data sample f_t arrives. In table 1, we have listed the filter coefficients of $h_k^e(j)$ for scale $j = 1$ to 4.

Multiscale product of the first K scale sequences in wavelet-domain at time index t is defined as

$$\xi_t^K = \prod_{j=1}^{K} \bar{z}_t(j). \tag{16}$$

The wavelet used for DWT in this paper is chosen as the first order derivative of a smooth function (a cubic spline function, see [11]). The multiscale product sequence ξ_t^K sharpens and enhances the modulus maxima which are dominated by the signal edges and at the same time suppresses the modulus maxima which are dominated by noises. It has been further shown that the probability density function (PDF) of a multiscale product sequence is heavy tailed compared with that of a Gaussian distributed one with the same variance [13]. Employing these characteristics, a DWT multiscale product sequence of an existing detection signal (for example, obtained from a T & L detector) can be used as a new detection signal. It will enhance the components representing possible abrupt changes in the original detection signal and thus a larger detection threshold can be used, which will lead to a smaller false alarm probability. At the same time, it will suppress the noise interference components in the original detection signal, which will decrease the miss alarm probability and thus increase the detection probability when using the same detection threshold as the one used by the original detection signal. Motivated by the above discussion, a new wavelet jump detector is proposed in the following subsection for on-line change detection.

C. Wavelet jump detector for on-line abrupt change detection

Denote $\tilde{z}_t(j)$ as the causal DWT of the T & L detection signal d_t (8) at time t and scale j. That is,

$$\tilde{z}_t(j) = \bar{W}_{2^j} d_t, \tag{17}$$

which can be recursively calculated as (11) and (12). The multiscale product signal of the first K scales thus can be calculated as

$$\tilde{\xi}_t^K = \prod_{j=1}^{K} \tilde{z}_t(j). \tag{18}$$

Define a new (multiscale product) detection signal ζ_t by filtering $\tilde{\xi}_t^K$ as follows

$$\zeta_t = (1-\eta)\zeta_{t-1} + \eta \tilde{\xi}_t^K, \tag{19}$$

where, η is an exponential smoothing factor which commonly takes a value in $0.05 \le \eta \le 0.13$. Although ξ_t^K is heavy-tailed non-Gaussian distributed, ζ_t obtained above is a Gaussian distributed signal according to the central limiting theorem. Now a new wavelet detector can be formed as

$$\tilde{d}_t = \sqrt{\frac{\mathrm{Var}(d_t)}{\mathrm{Var}(\zeta_t)}}\zeta_t, \tag{20}$$

Obviously, if d_t is a Gaussian distributed signal, $\tilde{d}_t$ is also a Gaussian distributed signal whose variance is the same as the one of d_t. However, if d_t has some local maxima (minima) which correspond to the abrupt changes of the original signal, these local maxima (minima) will be enlarged and sharpened in $\tilde{d}_t$. This characteristics undoubtedly can be employed to provide a more robust and accurate identification of the possible abrupt changes. That means if we choose the detection threshold $\tilde{r}$ of the wavelet detection signal $\tilde{d}_t$ equal to r of the T & L detection signal d_t, we can achieve much higher detection probability.

To get the new wavelet detector $\tilde{d}_t$, the variance of ζ_t ($\mathrm{Var}(\zeta_t)$) should be estimated in (20). One method is as follows. $\mathrm{Var}(\zeta_t)$ can be recursively estimated as

$$\hat{v}_t^2 = (1-\rho)\hat{v}_{t-1}^2 + \rho\zeta_t^2, \tag{21}$$

and the variance of d_t is asymptotically as (9), thus $\tilde{d}_t$ can be estimated as $\hat{d}_t$,

$$\hat{d}_t = \sqrt{\frac{\pi}{2}\frac{\gamma}{2-\gamma}}\frac{\zeta_t}{\hat{v}_t}, \tag{22}$$

where, ρ is another exponential smoothing factor which takes a value of greater but very close to 0 (commonly $0.01 \le \rho \le 0.03$). For estimating the variance of ζ_t, (22) is asymptotically efficient but very sensitive to the choice of ρ. Moreover, the variance estimation by (22) is heavily affected by the location density of jumping points. To overcome these problems, a better method is proposed below based on an empirical equation to estimate the variance of ζ_t.

Assuming the ratio of the variance of new wavelet multiscale product detection signal to the variance of the T & L detection signal (R-W-TL) as

$$\text{R-W-TL} = \frac{\mathrm{Var}(\zeta_t)}{\mathrm{Var}(d_t)} = \exp\left(\log\left(\frac{\eta}{2-\eta}\right) + \kappa\eta + \tau - \sum_{i=0}^{m}\nu_i\gamma^{m-i}\right), \tag{23}$$

and the wavelet detector (20) using empirical variance ratio estimation (23) can be represented as $\check{d}_t$,

$$\check{d}_t = \frac{\zeta_t}{\sqrt{\text{R-W-TL}}}, \tag{24}$$

where the values of κ, τ, and $\{\nu_i\}_{i=1,\cdots,m}$ can be estimated by applying least-squares method to experimental data through Monte Carlo simulations. A recommendation values for κ, τ are $\kappa = -5.1043$, $\tau = 1.0026$ and an order $m = 9$ polynomial coefficients are $\nu = \{1.272 \times 10^{11}, -5.9549 \times 10^{10}, 1.179 \times 10^{10}, -1.2877 \times 10^{9}, 8.4984 \times 10^{7}, -3.5005 \times 10^{6}, 9.0861 \times 10^{4}, -1.5622 \times 10^{3}, 2.5120 \times 10^{1}\}$. Extensive simulations have verified that the empirical equations (23) and (24) are effective and produce quite accurate results. See [14] for details.

D. Selectively tracking of rapidly changing systems using wavelet jump detectors

Normally, different branches of system parameters are not always subject to abrupt changes at the same time when a jump occurs in a time-varying system. When we modify the matrix P_{t-1} or P_t (in (4) or (5)) with Δ_t, it is common to select Δ_t as a diagonal matrix where each diagonal element reflects the change of the corresponding parameter branch. When one or several branches have changed rapidly at a specific time t, the corresponding elements in Δ_t should be increased while the remaining elements should keep unchanged [2]. This requires that the jump detector can not only identify the locations where the jumps have happened but also determine the branches producing these jumps. A priori prediction error signal is used to construct the jump detector [6], [7] which (named as *prediction detector*) only can determine where a jump happens for a time-varying system. To judge which branches this jump is produced by, a set of jump detectors can be constructed directly from the estimated filtering gains (named as *gain detectors*). Combining the *prediction detector* with *gain detectors*, a new *selective wavelet detector* is proposed in the following, which can determine not only the locations of jumping points but also the branches that have produced the jumps.

Assume a wavelet detector $\check{d}_t^{\varepsilon}$ (*prediction detector*) is obtained from the priori prediction error signal $\varepsilon_t = y_t - \varphi_t^T \hat{\theta}_{t-1}$ (3). Assume other N wavelet detectors $\check{d}_t^1, \cdots, \check{d}_t^N$ (*gain detectors*) are obtained from the estimated filtering gains $G_t(1), \cdots, G_t(N)$ (4) respectively. Without loss of generality, here we assume a system jumping change at a specific time is produced by an abrupt change of only one parameter branch (The case of several parameter branches changing at the same time is a simple extension). The proposed *selective wavelet detector* uses both the *prediction detector* and the *gain detectors* for parameter change detection. More explicitly, an abrupt change is considered to be detected at the ith ($i = 1, \cdots, N$) parameter branch at time t, if

$$|\check{d}_t^{\varepsilon}| > \tilde{r} \quad \text{and} \quad |\check{d}_t^{i}| > \tilde{r}, \tag{25}$$

where the detection threshold $\tilde{r}$ can be determined from (9).

In a summary, we list the complete RLS algorithm using estimation covariance matrix modification and selective wavelet detector (abbreviated as RLS-MSWD) at time t as follows:

- (a). **RLS algorithm**

Using (3)-(5) to calculate $\hat{\theta}_t$, ε_t, G_t and P_t;

- (b). **Selective wavelet detector for change detection**
  - (b1). From ε_t, calculating (6)-(8), (17) (implemented with (11) and (12)), (18), (23), and (24) to get the *predictive detector* $\check{d}_t^{\varepsilon}$,
  - (b2). For i=1:N {Using $G_t(i)$ instead of ε_t in (6) and (7), calculating equations as in (b1) to get the ith *gain detector* $\check{d}_t^i$} End
  - (b3). Using (25) to detect if a jumping change has happened. If yes, determine which parameter branch produces this change and got to (c) and set Δ_t, otherwise $t = t + 1$ and go to (a);
- (c). **Estimation covariance matrix modification**

Modify P_{t-1} or P_t in (4) or (5). $t = t + 1$ and go to (a).

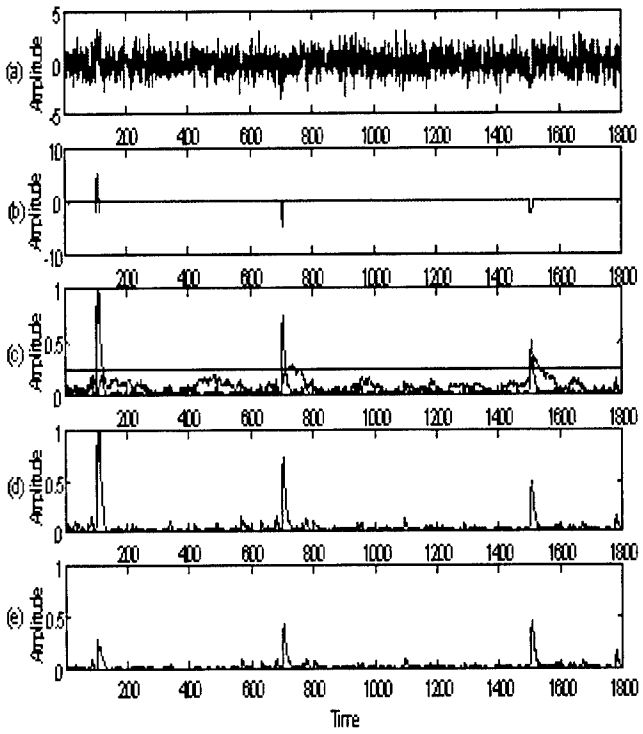


Fig. 1. Comparison of the wavelet detector with the T & L detector.

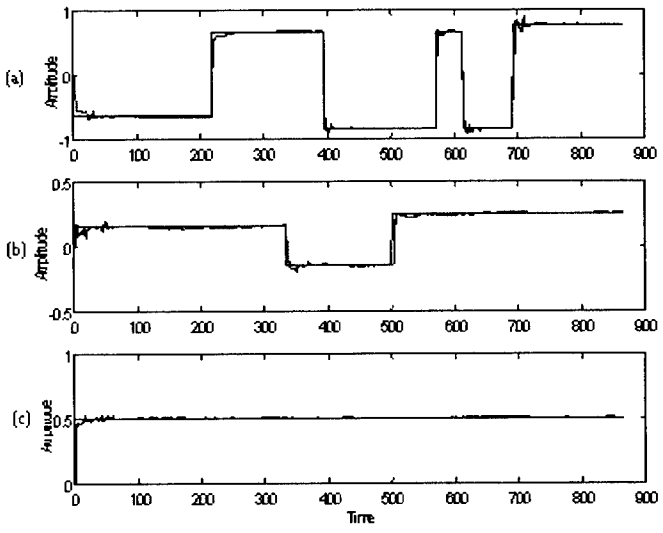


Fig. 2. An ARX(2,1) abruptly changing system identification by the proposed RLS-MSWD algorithm,(a)$b_t(1)$,(b)$b_t(2)$,(c)$c_t(1)$.

III. Simulation Results

In figure 1, a wavelet detector is compared with a T & L detector. (a) shows a stationary white Gaussian noisy signal

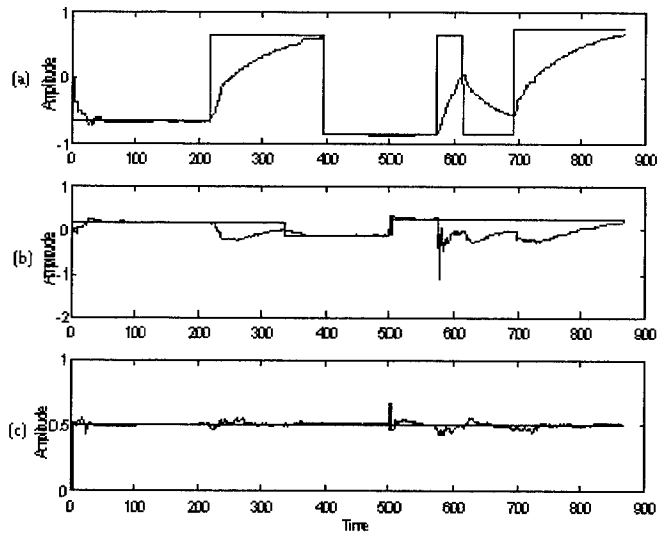


Fig. 3. An ARX(2,1) abruptly changing system identification by the RLS-MTLD algorithm,(a)$b_t(1)$,(b)$b_t(2)$,(c)$c_t(1)$.

which has three abrupt changes at the vicinity of time locations 100, 700, and 1500 respectively. The amplitudes and shapes of three abrupt changes are shown in (b). In (c), the solid line represents the T & L detection signal and the dotted line represents the wavelet detection signal (24) obtained using the theoretical R-W-TL (23). (d) shows the same trace as the one represented by the dotted line in (c), i.e., wavelet detection signal obtained using the theoretical R-W-TL (wavelet decomposition scale number $K = 3$). (e) shows the wavelet detection signal (22) obtained using the recursive variance estimation (21). Comparing the wavelet detection signal with the T & L detection signal in (c), the former can provide sharper and more accurate indication of the abrupt changing points and this is very important for detecting small amplitude or/and concentrated abrupt changes. Comparing (d) with (e), it can be seen that the detector in (e) is asymptotically consistent with the one in (d) when the recursive estimation of variance becomes more and more accurate.

An ARX(2,1) system

$$y_t = b_t(1)y_{t-1} + b_t(2)y_{t-2} + c_t(1)u_{t-1} + u_t,$$

is used for verifying the performance of the proposed abrupt change tracking algorithm. Here the system parameters $b_t(1)$ and $b_t(2)$ are both with abrupt changes and $c_t(1)$ is constant shown in figure 2. The identification results by the proposed RLS-MSWD are shown in figure 2, where $\gamma = 0.02$, $\eta = 0.10$, $K = 3$, and the empirical formulas (23) and (24) are used for producing the wavelet detectors.(Solid lines represent tracking results and dotted lines represent true values.) It can be seen that the estimation coincides with true parameter values very well. For comparison, identification results by the RLS algorithm using T & L detector (abbreviated as RLS-MTLD) are shown in figure 3. Since the T & L detector is not so sensitive to the abrupt changes as the selective wavelet detector, the identification results by the RLS-MTLD method can not track abrupt changes with small amplitude (see $b_t(2)$ between time index 330 and 500) and concentrated abrupt changes (see $b_t(1)$ between time index 570 and 620) in figure 3. Moreover, from figure 2 we can see that the proposed RLS-MSWD method can selectively track the abrupt changes of different parameter branches; while the estimation of different parameter branches by the RLS-MTLD method in figure 3 are disturbed and affected by each other.

IV. Conclusions

In this paper, the problem of tracking abruptly changing systems has been tackled. A new on-line wavelet detector has been proposed which is computationally simpler and can achieve much higher detection probability than commonly used abrupt detection methods. Selectively tracking the rapidly changing parameter branches via estimation covariance modification at the jumping points has been rigorously discussed.

References

[1] T. R. Fortescue, L. S. Kershenbaum, and B. E. Ydstie, "Implementation of self-tuning regulators with variable forgetting factor," *Automatica*, vol. 17, no. 6, pp. 831-835, 1981.

[2] L. Ljung and S. Gunnarsson, "Adaptation and tracking in system identification-a survey," *Automatica*, vol. 26, no. 1, pp. 7-21, 1990.

[3] M. J. Chen and J. P. Norton, "Estimation technique for tracking rapid parameter changes," *Int. J. of Control*, vol. 45, no. 4, pp. 1387-1398, 1987.

[4] P. Andersson, "Adaptive forgetting in recursive identification through multiple models," *Int. J. of Control*, vol. 42, no. 5, pp. 1175-1193, 1985.

[5] M. Niedzwiecki, "Identification of time-varying systems with abrupt parameter changes," *Automatica*, vol. 30, no.3, pp. 447-459, 1994.

[6] D. W. Trigg and A. G. Leach, "Exponential smoothing with an adaptive response rate," *Oper. Res. Quart.*, vol. 18, pp. 53-59, 1967.

[7] B. Carlsson, *Digital Differentiating Filters and Model Based Fault Detection*, Ph.D. Dissertations, Uppsala University, Sweden, 1989.

[8] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*, Englewood Cliffs, N.J.: Prentice Hall, 1993.

[9] J. Holst and N. K. Poulsen, "Self tuning control of plant with abrupt changes," *Proc. of IFAC 9th Triennial World Congress*, pp. 923-928, Budapest, Hungary, 1984.

[10] M. Niedzwiecki, "Identification of Nonstationary Stochastic systems using parallel estimation schemes," *IEEE Trans. on Automatic Control*, vol. 35, no. 3, pp. 329-334, 1990.

[11] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.14, no.7, 710-732, 1992.

[12] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. on Information Theory*, vol. 38, no. 2, 617-643, 1992.

[13] B. M. Sadler and A. Swami, "Analysis of multiscale products for step detection and estimation," *IEEE Trans. on Information Theory*, vol. 45, no. 3, pp. 1043-1051, 1999.

[14] Y. Zheng and Z. Lin, "Recursive adaptive algorithms for fast and rapidly time-varying systems: tracking performance improvement and best parameter selection," submitted to *IEEE Trans. on Circuit and System II* for publication.

Appendix

Tabel 1: Filter coefficients used for causal DWT ($1 \le j \le 4$)

$$
\begin{aligned}
h_k^e(1) &= \{-1.3333, 1.3333\};\\
h_k^e(2) &= \{-4.4643\times10^{-1}, -2.2321\times10^{-1}, 4.4643\times10^{-1},\\
&\quad 2.2321\times10^{-1}\};\\
h_k^e(3) &= \{-3.0941\times10^{-2}, -5.0278\times10^{-2}, -3.8676\times10^{-2},\\
&\quad -2.7073\times10^{-2}, 1.2136\times10^{-1}, 9.1019\times10^{-2},\\
&\quad 6.0680\times10^{-2}, 3.0340\times10^{-2}\};\\
h_k^e(4) &= \{-1.2136\times10^{-1}, -1.5170\times10^{-1}, -6.0680\times10^{-2},\\
&\quad 3.0340\times10^{-2}, -1.5470\times10^{-2}, -3.8676\times10^{-3},\\
&\quad 7.7351\times10^{-3}, 1.9338\times10^{-2}, 3.0941\times10^{-2},\\
&\quad 2.7073\times10^{-2}, 2.3205\times10^{-2}, 1.9338\times10^{-2},\\
&\quad 1.5470\times10^{-2}, 1.1603\times10^{-2}, 7.7351\times10^{-3},\\
&\quad 3.8676\times10^{-3}\}.
\end{aligned}
$$

AN APPLICATION OF THE MAXIMUM LIKELIHOOD PRINCIPLE TO SEMIBLIND SPACE-TIME LINEAR DETECTION IN MULTIPLE-ACCESS WIRELESS COMMUNICATIONS

Mónica F. Bugallo, Joaquín Míguez, Luis Castedo

Departamento de Electrónica e Sistemas, Universidade da Coruña
Facultade de Informática, Campus de Elviña s/n, 15071 A Coruña (SPAIN).
Tel: +34 981167000 *Fax:* +34 981167160, *e-mail:* monica@des.fi.udc.es

ABSTRACT

This paper introduces a novel semiblind approach to space-time linear detection in multiple-access systems. A new criterion for the selection of the linear receiver coefficients, based on the Maximum Likelihood (ML) principle, is derived and a practical implementation by means of a fast Expectation-Maximization (EM) algorithm is suggested. The semiblind criterion is obtained from a purely statistical point of view where the aim of training data is not to enhance performance but to eliminate misconvergence problems.

1. INTRODUCTION

It has been recently shown that deploying multiple transmitting and receiving antennae can substatially improve the capacity of multipath wireless channels if the rich time-scattering propagation is properly exploited. Space-Time Coding (STC) is a novel proposal that combines channel coding techniques suitable for multiple transmitting elements with signal processing algorithms that exploit the spatial and temporal diversity at the receiver [1, 2].

In the paper, we focus on the signal processing issue of soft detection as a prior step to channel decoding. We introduce a novel semiblind criterion based on the Maximum Likelihood (ML) principle that inherently exploits any existing spatio-temporal structure induced by the space-time encoder in order to linearly estimate the transmitted symbols. Note that estimating the stream of symbols transmitted from the j-th antenna involves removing the Inter-Symbol Interference (ISI), due to the channel time-scattering, and the Multiple Access Interference (MAI), due to the other symbol streams. The method is termed semiblind because not only exploits the *a priori* knowledge of the part of the transmitted symbols but also the information bearing symbols transmitted from a single antenna. A remarkable feature of the new criterion, when compared to other semiblind approaches which are basically *ad hoc* or heuristic methods [3], is that it is derived from a purely statistical point of view. A fast iterative algorithm, derived from the Expectation-Maximization (EM) [4] framework, is suggested as a means of practical implementation. In this algorithm, symbols *a priori* known are extremely useful to avoid the misconvergence problems so typical in ML methods.

The remaining of the paper is organized as follows. Section 2 describes the system model. The novel ML-based criterion is presented in section 3. The iterative EM algorithm is derived in section 4 and its performance is evaluated, through computer simulations, in section 5. Finally, section 6 is devoted to the conclusions.

This work has been supported by FEDER funds (1FD97-0082) and Xunta de Galicia (PGIDT00PXI10504PR).

2. SYSTEM AND SIGNAL MODEL

Figure 1 shows the basic building blocks of a wireless communication system with Space-Time (ST) coding capabilities [5]. The bit stream to be transmitted, $\{b(l)\}_{l=0,1,\cdots}$, is fed into a temporal coding stage followed by a Serial to Parallel (S/P) converter that creates some desired spatio-temporal structure. A bank of N identical Waveform Encoders (WE) and transmitting antennae yields the information bearing signals to be transmitted, $s_1(t), \cdots, s_N(t)$. Transmission is carried out in bursts of $NK \log_2 A$ bits, i.e., K complex symbols per antenna with $log_2 A$ bits per symbol. Multipath propagation between each transmitting and receiving antennae results in a Multiple Input Multiple Output (MIMO) channel with Inter-Symbol Interference (ISI). A bank of $L \geq N$ matched filters, sampled at the symbol rate, $\frac{1}{T}$, is employed at the receiver to obtain a set of sufficient statistics, $x_1(n), \cdots, x_L(n)$, $n = 0, \cdots, K-1$. An adequately chosen linear processor, consisting of a bank of linear Finite Impulse Response (FIR) filters, provides soft estimates, $y_1(n), \cdots, y_N(n)$, $n = 0, \cdots, K-1$, of the complex transmitted symbols, that we denote as $s_1(n), \cdots, s_N(n)$, $n = 0, \cdots, K-1$. A Parallel to Serial (P/S) converter and a channel decoder yield hard estimates of the transmitted information bits.

Assuming a linear memoryless modulation format is employed, we obtain a linear signal model for the discrete-time signals observed after the bank of symbol rate samplers during the n-th symbol period

$$\underline{x}(n) = \sum_{i=0}^{m-1} \underline{H}_i \underline{s}(n-i) + \underline{g}(n) = \mathbf{H}\mathbf{s}(n) + \underline{g}(n) \qquad (1)$$

where $\underline{x}(n) = [x_1(n), ..., x_L(n)]$ is the vector of observations obtained from the bank of receivers, $\underline{s}(n) = [s_1(n), ..., s_N(n)]$ is the n-th vector of transmitted symbols, $\mathbf{s}(n) = [\underline{s}^T(n-m+1) \ \cdots \ \underline{s}^T(n)]^T$ is a $Nm \times 1$ vector containing the data components received during the n-th symbol period due to the ISI, $\underline{g}(n)$ is a $L \times 1$ vector of Additive White Gaussian Noise (AWGN) components with zero-mean and covariance matrix $E[\underline{g}(n)\underline{g}^H(n)] = \sigma_g^2 \mathbf{I}_L$ (being $\mathbf{I}_L$ the $L \times L$ identity matrix), and $\mathbf{H} = [\underline{H}(m-1) \ \cdots \ \underline{H}(0)]$ is the $L \times Nm$ matrix that contains the discrete-time channel coefficients resulting from

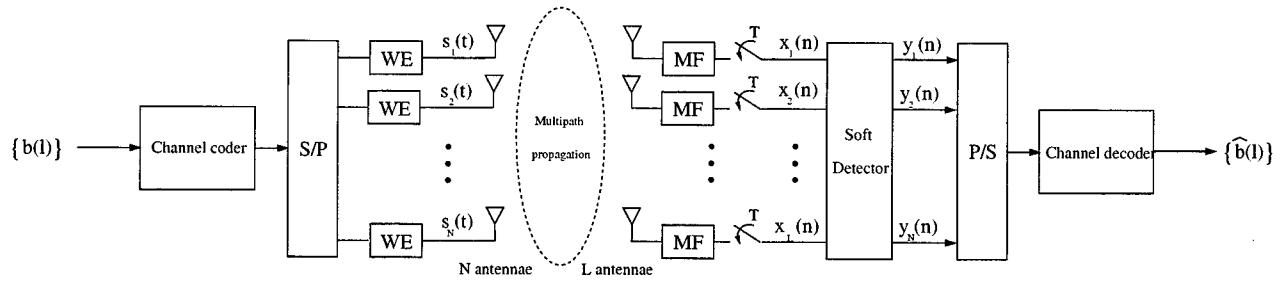


Fig. 1. Block diagram of the communication system.

symbol-rate sampling in the described structure. In more detail, let the $m \times 1$ vector $\mathbf{h}_{lp} = [h_{lp}(0), ..., h_{lp}(m-1)]^T$ represent the discrete-time channel impulse response between the p-th transmitting antenna and the l-th receiving antenna obtained after symbol rate sampling, where m is the maximum length of the impulse response and, as a consequence, the size of the ISI window. The $L \times N$ building submatrices in the MIMO channel, $\underline{H}(i)$ $i = 0, ..., m-1$, turn out to be

$$\underline{H}(i) = \begin{bmatrix} h_{11}(i) & h_{12}(i) & \dots & h_{1N}(i) \\ h_{21}(i) & h_{22}(i) & \dots & h_{2N}(i) \\ \vdots & \vdots & \ddots & \vdots \\ h_{L1}(i) & h_{L2}(i) & \dots & h_{LN}(i) \end{bmatrix}. \quad (2)$$

The symbols of interest to be estimated are those in $\underline{s}(n)$. In order to guarantee that the whole energy of this vector is processed, let us stack the observations from m consecutive symbol periods to obtain the extended signal model

$$\mathbf{x}(n) = \mathcal{H}\mathbf{s}_m(n) + \mathbf{g}(n) \quad (3)$$

where $\mathbf{x}(n) = [\underline{x}^T(n) \quad \cdots \quad \underline{x}^T(n+m-1)]^T$ is the $Lm \times 1$ observation vector, $\mathbf{s}_m(n) = [\underline{s}^T(n-m+1) \quad \cdots \quad \underline{s}^T(n+m-1)]^T$ is the vector of contributing symbols with dimensions $N(2m-1) \times 1$, $\mathbf{g}(n) = [\underline{g}^T(n) \quad \cdots \quad \underline{g}^T(n+m-1)]^T$ is an AWGN vector and the extended channel matrix has the block-diagonal form

$$\mathcal{H}^T = \begin{bmatrix} \underline{H}^T(m-1) & \mathbf{0} & \cdots & \mathbf{0} \\ \underline{H}^T(m-2) & \underline{H}^T(m-1) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \underline{H}^T(0) & \underline{H}^T(1) & \ddots & \underline{H}^T(m-1) \\ \vdots & \underline{H}^T(0) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \underline{H}^T(1) \\ \mathbf{0} & \mathbf{0} & \cdots & \underline{H}^T(0) \end{bmatrix}^T$$

and dimensions $Lm \times N(2m-1)$.

An $N \times 1$ vector of soft estimates, $\mathbf{y}(n) = [y_1(n), \cdots, y_N(n)]^T$, corresponding to the symbols in $\underline{s}(n)$, is obtained through linear processing as

$$\mathbf{y}(n) = \mathbf{W}^H\mathbf{x}(n) \quad (4)$$

where $\mathbf{W}$ is an $Lm \times N$ matrix filter and ^{H} denotes Hermitian transposition.

3. SELECTION OF THE RECEIVER COEFFICIENTS

The problem of selecting matrix $\mathbf{W}$ can be split into N simpler problems, i.e., $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_j, \cdots, \mathbf{w}_N]$, $j = 1, \cdots, N$ where $\mathbf{w}_j$ is the $Lm \times 1$ FIR filter that provides the estimate $y_j(n) = \mathbf{w}_j^H\mathbf{x}(n)$ corresponding to the symbol from the j-th transmitting antenna. In order to obtain this filter's coefficients, let us assume that $\mathbf{w}_{*,j}$ is the *optimum* value of the filter, meaning that it removes both the ISI and the MAI, leaving only a residual Gaussian interference. Hence, we can write

$$y_j(n) = \mathbf{w}_{*,j}^H\mathbf{x}(n) = s_j(n) + g_{f,j}(n) \quad (5)$$

where $s_j(n)$ is the desired symbol and $g_{f,j}(n)$ is a complex Gaussian random variable with zero mean and variance $\sigma_{f,j}^2 = \sigma_g^2\mathbf{w}_{*,j}^H\mathbf{w}_{*,j}$. For the sake of simplicity, the filtered noise variance, $\sigma_{f,j}^2$, will be considered a known constant in the subsequent derivations, but an easy-to-implement estimation algorithm will be proposed in the next section.

When a block of K observation vectors is available at the j-th receiver, and the symbols transmitted through the j-th antenna are i.i.d., it can be shown that the joint probability density function (p.d.f.) of the resulting soft estimates, $\mathbf{y}_j = [y_j(0), \dots, y_j(K-1)]^T$, is [6] (asuming white filtered noise)

$$f_{\mathbf{y}_j;\mathbf{w}_{*,j}} = \left(\frac{1}{\pi\sigma_{f,j}^2}\right)^K \prod_{n=0}^{K-1} E_s\left[e^{-\frac{|y_j(n)-s|^2}{\sigma_{f,j}^2}}\right] \quad (6)$$

where $E_s[\cdot]$ denotes the statistical expectation with respect to (w.r.t.) the desired symbol. Since the symbols belong to a finite alphabet of A elements, this expectation can be easily converted into an addition of A terms. Using (6), the ML estimate of $\mathbf{w}_{*,j}$ turns out to be

$$\hat{\mathbf{w}}_j = \arg \max_{\mathbf{w}_{*,j}} \left\{ \mathcal{L}(\mathbf{w}_{*,j}) = \sum_{n=0}^{K-1} \log E_s\left[e^{-\frac{|y_j(n)-s|^2}{\sigma_{f,j}^2}}\right] \right\} \quad (7)$$

where $\mathcal{L}(\mathbf{w}_{*,j})$ is the log-likelihood of $\mathbf{w}_j$ w.r.t. the observed soft-estimates $\mathbf{y}_j$. Since the linear filters, $\hat{\mathbf{w}}_j$ $j = 1, ..., N$, are chosen so as to ensure that the p.d.f. of $\mathbf{y}(n)$ is the desired one, the proposed criterion implicitly exploits any spatio-temporal structure created among the symbol substreams in order to remove the interferences.

Nevertheless, all the transmitting antennae use the same modulation format, and, as a consequence, solving the optimization problem (7) may lead to the *capture* of the j-th receiver by an interference, i.e., the estimation of a non desired

sequence, $s_{i\neq j}(n)$. This limitation can be easily circumvented in practice with the transmission of a short training sequence of $M < K$ symbols from each antenna. Conditioning the expectation in (7) to $\mathbf{s}_{j,t} = [s_j(0), \ldots, s_j(M-1)]^T$, we arrive at a *semiblind* receiver where the filter coefficients are computed as

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{*,j})|\mathbf{s}_{j,t} &= -\sum_{n=0}^{M-1} |y_j(n) - s_j(n)|^2 \\ &\quad + \sum_{n=M}^{K-1} \log E_s \left[e^{-\frac{|y_j(n)-s|^2}{\sigma^2_{f,j}}} \right] \\ \hat{\mathbf{w}}_j &= \arg \max_{\mathbf{w}_{*,j}} \{\mathcal{L}(\mathbf{w}_{*,j})|\mathbf{s}_t\}. \end{aligned} \quad (8)$$

The quadratic term in (8) reshapes the log-likelihood function by enhancing the local maximum corresponding to the desired j-th symbol stream and progressively removing (as M increases) the other non desired maxima.

It should be remarked that the semiblind criterion (8) is derived from a purely statistical point of view, whereas most semiblind criteria proposed so far are obtained in a rather heuristic manner by regularizing the Least Squares (LS) cost function for the training data using a different *blind* cost function [3].

4. ITERATIVE IMPLEMENTATION

Since it is not possible to find a closed form solution to problem (8), we resort to the EM algorithm [4] as a numerical optimization approach. Let the j-th sequence of soft estimates, $\{y_j(n)\}_{n=0,\ldots,K-1}$, be the *observed* or *incomplete* data and let the j-th stream of symbols, $\{s_j(n)\}_{n=0,\ldots,K-1}$, be the *hidden* data, according to the usual EM notation. Hence, the *complete* data are given by the sequence $\{y_j(n), s_j(n)\}_{n=0,\ldots,K-1}$ and taking similar steps as in the standard derivation of the EM algorithm [4] for the p.d.f. (6), we obtain the following iterative algorithm

$$\begin{aligned}\text{E step:} \quad & U(\mathbf{w}, \hat{\mathbf{w}}_j(i)) = -\sum_{n=0}^{M-1} |y_j(n) - s_j(n)|^2 \\ & - \sum_{n=M}^{K-1} E_{s_j(n)|y_j(n);\hat{\mathbf{w}}_j(i)} \left[|y_j(n) - s_j(n)|^2 \right] \\ \text{M step:} \quad & \hat{\mathbf{w}}_j(i+1) = \arg \max_{\mathbf{w}} \{U(\mathbf{w}, \hat{\mathbf{w}}_j(i))\} \end{aligned}$$

where $E_{s_j(n)|y_j(n);\hat{\mathbf{w}}_j(i)}[\cdot]$ denotes statistical expectation w.r.t. to symbol $s_j(n)$ conditioned upon the corresponding soft estimate, $y_j(n)$, and with parameter vector $\hat{\mathbf{w}}_j(i)$. Since function $U(\cdot,\cdot)$ is purely quadratic, it presents a single maximum that can be found analytically and it is possible to rewrite the above iteration as the single updating rule in eq. (9) shown at the top of next page. It can be proved by means of standard EM theory [4] that the sequence of filter updates obtained via (9) is non-decreasing in likelihood. Notice, also, that algorithm (9) reduces to the closed-form LS solution when $M = K$.

For the practical application of the iterative EM algorithm, the conditional expectation in (9) must be evaluated. This can be accomplished by means of the Bayes theorem as

$$E_{s_j(n)|y_j(n);\hat{\mathbf{w}}_j(i)} \left[s_j^*(n) \right] = \frac{E_{s_j(n)} \left[e^{-\frac{|y_j(n)-s_j(n)|^2}{\sigma^2_{j,f}}} s_j^*(n) \right]}{E_{s'} \left[e^{-\frac{|y_j(n)-s'|^2}{\sigma^2_{j,f}}} \right]}. \quad (10)$$

The expresion in the right-hand side of (10) depends on the filtered noise variance $\sigma^2_{j,f}$ which, in turn, is a function of the filter coefficients. A simple updating rule for this parameter is

$$\hat{\sigma}^2_{j,f}(i) = \sigma_g^2 \hat{\mathbf{w}}_j^H(i) \hat{\mathbf{w}}_j(i) \quad (11)$$

where the input AWGN variance, σ_g^2, (or, equivalently, the power spectral density of the channel noise) is assumed to be known *a priori*.

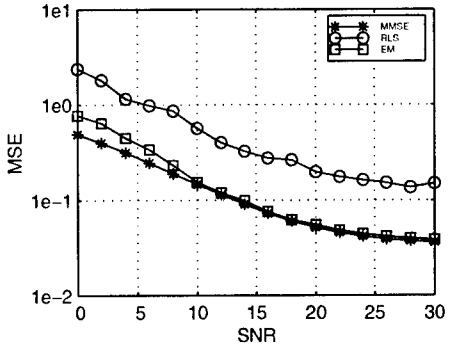


Fig. 2. MSE for several values of the SNR. Simulation parameters: $N = 3, L = 4, m = 2$.

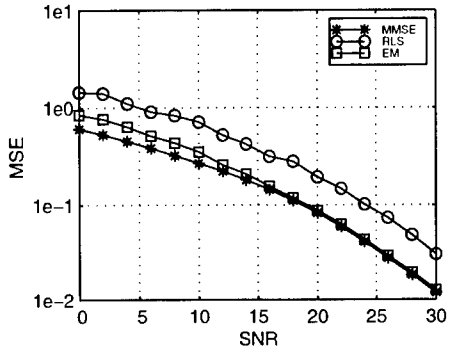


Fig. 3. MSE for several values of the SNR. Simulation parameters: $N = 6, L = 6, m = 1$.

5. COMPUTER SIMULATIONS

In this section, we present computer simulations to illustrate the performance of the proposed semiblind approach. As a figure of merit for soft-detection, we have chosen the Mean Squared Error (MSE) of the estimates, defined as

$$\text{MSE} = \frac{1}{N} \text{Trace} \left((\mathbf{y}(n) - \underline{s}(n))^H (\mathbf{y}(n) - \underline{s}(n)) \right) \quad (12)$$

$$\hat{\mathbf{w}}_j(i+1) = \left(\sum_{n=0}^{K-1} \mathbf{x}(n)\mathbf{x}^H(n) \right)^{-1} \left(\sum_{n=0}^{M-1} \mathbf{x}(n) s_j^*(n) + \sum_{n=M}^{K-1} E_{s_j(n)|y_j(n);\hat{\mathbf{w}}_j(i)} \left[s_j^*(n) \right] \mathbf{x}(n) \right). \quad (9)$$

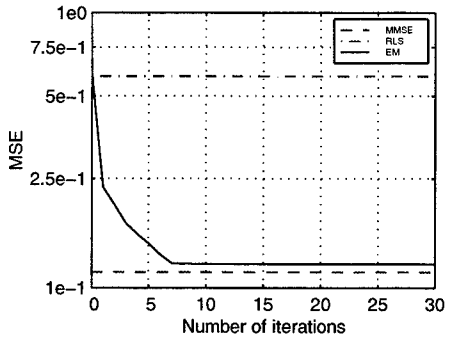


Fig. 4. Convergence rate of the EM algorithm. Simulation parameters: $N = 3$, $L = 4$, $m = 2$, SNR=12 dB.

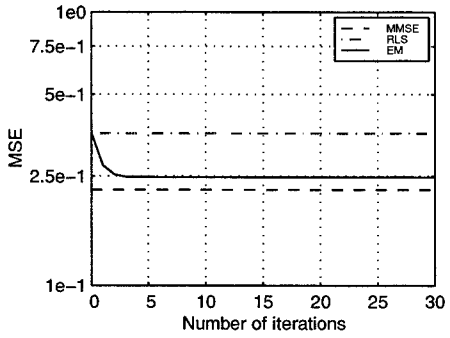


Fig. 5. Convergence rate of the EM algorithm. Simulation parameters: $N = 6$, $L = 6$, $m = 1$, SNR=12 dB.

where Trace$(\cdot)$ denotes the matrix trace operator. The MSE has been measured for different channels and different values of the average Signal to Noise Ratio (SNR) after sampling, given by $\text{SNR} = 10\log_{10} \frac{\sigma_s^2 \text{Trace}(\mathbf{H}\mathbf{H}^H)}{L\sigma_g^2}$, where $\sigma_s^2 = E_{s_j(n)}\left[|s_j(n)|^2\right]$ $\forall j, n$.

Figure 2 shows the MSE for several SNR values achieved in a system with $N = 3$ antennae, QPSK-modulated symbols, $L = 4$ receiving antennae and a maximum length of the discrete-time channel impulse response $m = 2$. The channel coefficients in matrix $\mathbf{H}$ are modelled as i.i.d. complex Gaussian random variables with zero-mean and standard deviation $\sigma_h = 0.5$. The results plotted in the figure have been obtained by averaging the performance over 20 independent realizations of the whole matrix $\mathbf{H}$. We have considered that transmission is carried out in bursts of $K = 100$ symbols per antenna, with training sequence length $M = 10$. It is apparent that the proposed semiblind approach performs close to the theoretical Minimum Mean Square Error (MMSE) limit. This is the performance limit that would be achieved by a linear MMSE detector constructed with perfect knowledge of the channel. We have also plotted the MSE achieved by a practical supervised MMSE soft detector implemented using the Recursive Least Squares (RLS) algorithm [7] that is run for the training sequences $\mathbf{s}_{j,t}$, $j = 1, \cdots, N$, in order to compute the filter coefficients. It can be seen that the performance of this practical receiver is considerably worse than the theoretical one because of the insufficient length of the training sequences.

These results are fully corroborated by an analogous simulation experiment carried out for a system with $N = 6$ transmitting antennae, $L = 6$ receiving antennae and channel length $m = 1$ (no ISI). The resulting curves are plotted in figure 3.

Finally, figures 4 and 5 illustrate the convergence rate of the EM algorithm for the two systems considered before. Very few iterations are enough to attain MSE convergence, which is an important advantage if real-time constraints have to be fulfilled.

6. CONCLUSIONS

We have presented a novel semiblind approach to space-time linear detection in wireless communication systems. A ML-based semiblind criterion is applied for the selection of the linear receiver coefficients, which are numerically computed by means of a fast iterative EM algorithm. Unlike other semiblind criteria, the proposed method is derived from a purely statistical point of view. Training data reveal themselves as extremely useful to avoid the typical misconvergence problems of ML methods.

7. REFERENCES

[1] G. J. Foschini, "Layered space-time arquitechture for wireless communications in a fading environment when using multi-element anntenas," *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 41–59, Autumn 1996.

[2] V. Tarokh and A.Naguib and N. Seshadri and A. R. Calderbank, "Space-time codes for high data rate wireless communications: Performance criteria in the presence of channel estimation errors, mobility and multiple paths," *IEEE Trans. Communications*, vol. 47, no. 2, pp. 199–207, February 1999.

[3] A. M. Kuzminskiy and D. Hatzinakos, "Semi-blind spatio-temporal processing with temporal scanning for short burst SDMA systems," *Signal Processing*, vol. 80, no. 10, pp. 2063–2073, October 2000.

[4] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 1997.

[5] A. R. Hammons, Jr., and H. El-Gamal, "On the theory of space-time codes for PSK modulation," *IEEE Trans. on Information Theory*, vol. 46, no. 2, pp. 524–542, March 2000.

[6] M. F. Bugallo, J. Míguez, and L. Castedo, "A maximum likelihood approach to blind multiuser interference cancellation," *IEEE Trans. Signal Processing*, vol. 49, no. 6, pp. 1228–1239, June 2001.

[7] S. Haykin, *Adaptive Filter Theory, 3rd Edition*, Prentice Hall, Information and System Sciences Series, 1996.

RECURSIVE BAYESIAN PHASE ESTIMATION IN RANGING AND MOBILE COMMUNICATION

José M. N. Leitão

Instituto de Telecomunicações
Instituto Superior Técnico
Av. Rovisco Pais
1049-001 Lisboa, Portugal
jleitao@red.lx.it.pt

Fernando M. G. Sousa

Instituto de Telecomunicações
Instituto Superior de Engenharia de Lisboa
R. Conselheiro Emídio Navarro, N. 1
1949-014 Lisboa, Portugal
fsousa@isel.pt

ABSTRACT

Mobile radio communication systems are generally designed without taking into account the relative emitter/receiver dynamics. In this paper we model this dynamics as a vector Markov process and formulate ranging and digital demodulation/detection as aspects of recursive absolute (not modulo 2π) phase estimation. Symbol-by-symbol detection and phase tracking within symbol interval are performed by a bank of 'matched' stochastic nonlinear estimators and a maximum a posteriori (MAP) decision algorithm. The approach applies to precision landing and communication with Low Earth Orbit (LEO) satellites or between rapid maneuvering platforms.

1. INTRODUCTION

Mobile radio communication systems are generally designed without taking into account the relative emitter/receiver dynamics. Doppler and Doppler rate estimates, necessary to cope with accelerative trajectories, are generally obtained with maximum likelihood (see [1] and references therein). In reference [2] we considered the problem of carrier tracking and symbol detection in Additive White Gaussian Noise (AWGN); phase dynamics is modelled as a vector linear Markov process, of which only the first component is observed. As in [3], symbol-by-symbol detection and phase tracking within symbol interval are performed by a bank of 'matched' stochastic nonlinear estimators and a maximum a posteriori (MAP) decision algorithm.

The results reported in [2] were limited to scalar phase dynamics, namely Brownian motion, and constrained to the interval $[-\pi\ \pi[$. In this paper we formulate ranging and digital demodulation/detection as aspects of recursive absolute (not modulo 2π) phase estimation. In doing this, we build on previous experience on this problem, see [4][5]. The proposed approach applies to precision landing and communication with Low Earth Orbit (LEO) satellites or between rapid maneuvering platforms, among others.

This work was partially supported by FCT, project 2/2.1/TIT/1583/95.

2. MODELLING ASSUMPTIONS

Consider mobile radio communications in AWGN channel and digital phase modulation. The received signal is $z(t) = \cos(\omega_0 t + \theta(t)) + v(t)$ (the carrier known amplitude is normalized to one), where ω_0 is the nominal carrier frequency (wavelength $\lambda = 2\pi c/\omega_0$), and $v(t)$ is white Gaussian noise with spectral density $N_0/2$. The phase process $\theta(t)$ is the sum of the digital information process $y(t)$, and the dynamics phase $x_1(t)$, which takes into account the Doppler phase shift due to relative emitter/receiver motion, and also oscillator phase drifts.

We are interested in applications where phase $x_1(t) = 2\pi R(t)/\lambda$ (proportional to range $R(t)$), varies significantly within digital symbol interval. We describe this process $x_1(t)$ as the first component of a vector Markov process $\mathbf{x}(t) \in \mathcal{R}^3$ modelled by

$$\dot{\mathbf{x}}(t) = \mathbf{A}_c\mathbf{x}(t) + \mathbf{B}_c u(t) \tag{1}$$

where $u(t)$ is white Gaussian driving noise with spectral density q_c. The components of vector $\mathbf{x}(t)$ are proportional to range, $x_1(t)$), velocity, $x_2(t)$, and acceleration, $x_3(t)$. Prior knowledge or side information about this dynamics is inserted in terms of matrices $\mathbf{A}_c$ and $\mathbf{B}_c$, noise variance q_c, and initial condition $p(\mathbf{x}(t_0))$.

The received signal $z(t)$ is down converted to baseband with reference to a local oscillator of nominal frequency ω_0. The sampled (normalized integration-and-dump) in phase and quadrature components form the observation vector

$$\mathbf{z}_n = [\cos(\theta_n)\ \sin(\theta_n)]^T + [v_{1,n}\ v_{2,n}]^T, \quad n = 1, 2, \cdots \tag{2}$$

where $v_{1,n}$ and $v_{2,n}$ are zero mean mutually independent white Gaussian sequences with variance $r = r_c/\Delta$ ($r_c = N_0$). The sampling interval Δ must be small enough to guarantee that both discrete and continuous models describe essentially the same process. For implementation and simulation purposes we adopt the discrete version of (1),

$$\mathbf{x}_{n+1} = \mathbf{A}\mathbf{x}_n + \mathbf{B}\mathbf{u}_n, \quad n = 1, 2, \cdots N \tag{3}$$

where $\mathbf{A} = \mathbf{I} + \mathbf{A}_c\Delta$, $\mathbf{B} = \mathbf{B}_c\Delta$, and u_n is a zero mean white Gaussian sequence with variance $q = q_c/\Delta$.

Signaling $y(t)$ will be presented in section 4.

3. RANGING

Optimal estimation of $\mathbf{x}_n$ involves the propagation of the probability density function $F_n = P(\mathbf{x}_n|\mathbf{Z}_n)$, *filtering density*, conditioned on the set of past and present observations $\mathbf{Z}_n = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$. This requires recursive application of Chapman-Kolmogorov equation and Bayes law

$$\text{Prediction:} \quad P_n = S_n * F_{n-1} \tag{4}$$

$$\text{Filtering:} \quad F_n = C_n H_n P_n \tag{5}$$

where $*$ denotes convolution, and C_n is a normalizing factor; the *convolution kernel* $S_n = P(\mathbf{x}_{n+1}|\mathbf{x}_n)$, which expresses the process dynamics (3), is Gaussian given by

$$S_n \propto \mathcal{N}\left(\mathbf{x}_{n+1} - \mathbf{A}\mathbf{x}_n, \mathbf{B}q\mathbf{B}^T\right) \tag{6}$$

where $\mathcal{N}(\mathbf{u}, \mathbf{V}) = \exp\left(-(1/2)\mathbf{u}^T\mathbf{V}^{-1}\mathbf{B}^T\right)$. The probability density function H_n (*observation factor*) is, according to model (2), given by

$$H_n \propto \exp\left(\lambda_n \cos(x_{1,n} - \eta_0^{H_n})\right) \tag{7}$$

with

$$\lambda_n = \frac{1}{r}\sqrt{z_{1,n}^2 + z_{2,n}^2}, \quad \eta_0^{H_n} = \arctan\frac{z_{2,n}}{z_{1,n}}. \tag{8}$$

To implement (4)(5) we need finite representations of the involved probability density functions. This concerns, in this problem, the periodic function H_n. As in [4], we represent H_n by

$$\tilde{H}_n \propto \sum_{i=-\infty}^{\infty} \mathcal{N}\left(x_{1,n} - \eta_i^{H_n}, \sigma^{H_n}\right), \quad \eta_i^{H_n} = \eta_0^{H_n} + 2\pi i \tag{9}$$

where σ^{H_n} is obtained according to a minimum Kullback distance criterion.

3.1. Tracking

Consider a *prediction density* $P_n \propto \mathcal{N}\left(\mathbf{x}_n - \eta^{P_n}, \mathbf{V}^{P_n}\right)$, and assume that only the mode of $\tilde{H}_n$ closest to η^{P_n} contributes significantly to the product (4). The *filtering density* F_n will be Gaussian, with mean η^{F_n} and covariance matrix $\mathbf{V}^{F_n}$. The optimal estimate is then given by $\hat{\mathbf{x}}_n = \eta^{F_n}$.

Fig. 1 shows the evolution of phase (range), phase rate and acceleration, generating phase as a double integrated Brownian motion with driving noise variance $q_c = 50\,\text{rad}\,\text{s}^{-6}$ (which is chosen to encompass the dynamics of a typical LEO satellite). Also shown are the estimates obtained by the filter in tracking conditions and with perfect matched parameters and observation noise variance $r = 0.3962\,\text{rad}^2$.

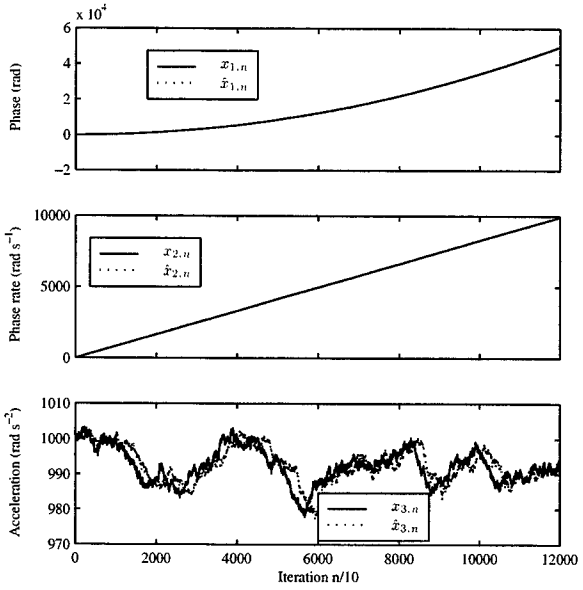


Fig. 1. Phase, phase rate and acceleration trajectories and their corresponding estimates ($r = 0.3962\,\text{rad}^2$ and $q_c = 50\,\text{rad}^2\,\text{s}^{-6}$).

3.2. Acquisition

In the preceding example the estimates were initialized at their nominal values. In general this is not exactly known and tracking has to be preceded by an acquisition period which, due to the multi-modal filtering density function induced by the sensor factor representation (9), corresponds to a phase ambiguity resolution. We apply to this problem the methodology developed in [4]. Fig. 2 illustrates this mechanism with scalar dynamics $\dot{x}(t) = a_c x(t) + u(t)$ (phase rate proportional to absolute phase). Starting with a multimodal density, the filter converges recursively to an essen-

tially uni-modal shape. The acquisition can be formalized by introducing an internal measure of dispersion; the first passage moment of this measure across a given threshold defines the acquisition time.

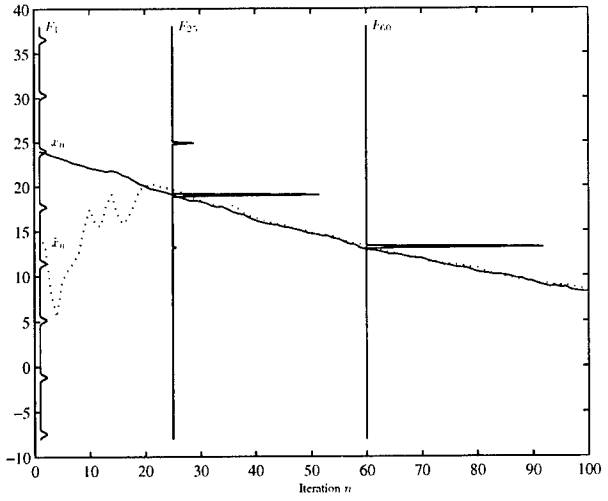


Fig. 2. Absolute phase acquisition ($x_{n+1} = ax_n + u_n$ with $q = 0.01$, $r = 0.5$, $a = 0.99$).

4. MOBILE COMMUNICATION

We now consider signaling $y(t)$ as the transient evolution of the first order linear dynamical system

$$\dot{y}(t) = -\beta y(t), \quad t \in [kT_s, (k+1)T_s[\tag{10}$$

randomly initialized at the beginning of each symbol interval, $[kT_s, (k+1)T_s[$, $k = 0, 1, \cdots$, according to the probability density function

$$P\left(y(kT_s)\right) = \sum_{j=1}^{M} \frac{1}{M} \delta\left(y(kT_s) - \frac{d}{2}\alpha^{(j)}\right), \quad j = 1, \cdots, M \tag{11}$$

where M is the number of distinct equiprobable symbols, $\alpha^{(j)} = 2j - M - 1$, and $a = \beta T_s$ and d are modulation parameters. Since the phase trajectories are not restricted to a 2π interval, we call this modulation scheme M-ary *absolute phase modulation* (M-APM). When $a \to 0$, two well-known digital schemes are produced: M-PSK, for $d = 2\pi/M$, and orthogonal M-FSK, for $ad = \pi$. Continuous phase modulation schemes can also be obtained by adjusting parameters a and d [3].

We adopt the discrete version of (10)

$$y_{n+1} = (1 - \beta\Delta)y_n, \quad n = 1, 2, \cdots N \tag{12}$$

where $N = T_s/\Delta$ is the number of samples per symbol interval.

The main task of the receiver is to acquire and track the dynamics process $\mathbf{x}_n$ (which provides a range solution). While tracking, it must decide, at the end of each symbol interval, which symbol was sent. We assume perfect symbol timing. Like in [2][3] the receiver is a parallel of M 'matched' nonlinear filters, each one preceded by a phase rotation block that eliminates the contribution of $y_n^{(j)}$ from the observation vector $\mathbf{z}_n$. The detector computes the weights associated with each filter block and decides according to a MAP criterion. The prediction density P_{kN} of the selected filter is used to set the initial condition of all filters to the next symbol interval. This corresponds to a symbol aided decision criterion.

4.1. LEO satellite example

Consider communication between an Earth station and a LEO satellite describing a circular orbit with an altitude of 780 Km [6]; the emitter and the receiver are both in the equatorial plane and the carrier nominal frequency $f_0 = 1.6$ GHz. Fig. 3 shows phase, phase rate and acceleration, during the entire visibility window (11.1 minutes) assuming a minimum elevation angle of 8.2°. Consider

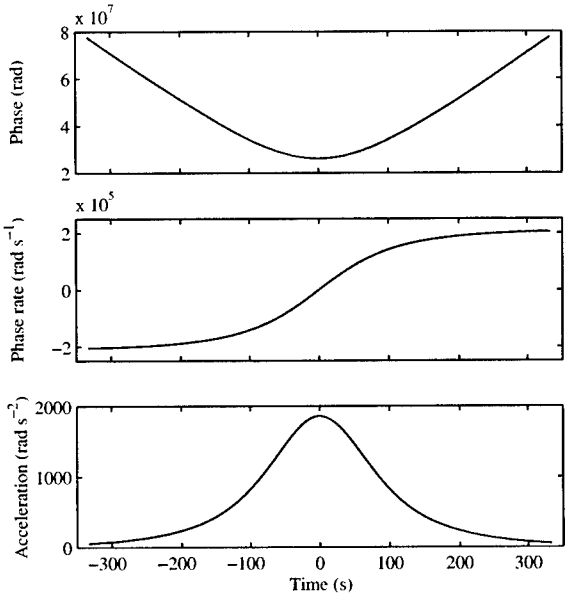


Fig. 3. Phase, phase rate and acceleration for a LEO satellite trajectory along the visibility window.

also quaternary phase modulation ($M = 4$) with signaling parameters $a = 1$, $d = 5.7$ [2], bit rate 2400 bit/s, ($T_s = 1/1200$ s), $N = 10$ samples per symbol ($\Delta =$

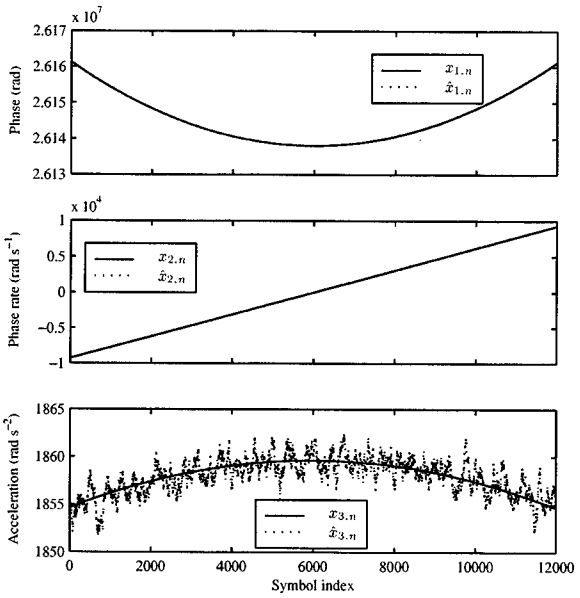


Fig. 4. Phase, phase rate and acceleration and their estimates for a LEO satellite trajectory ($E_b/N_0 = 8$ dB and $q_c = 50$ rad^2 s^{-6}).

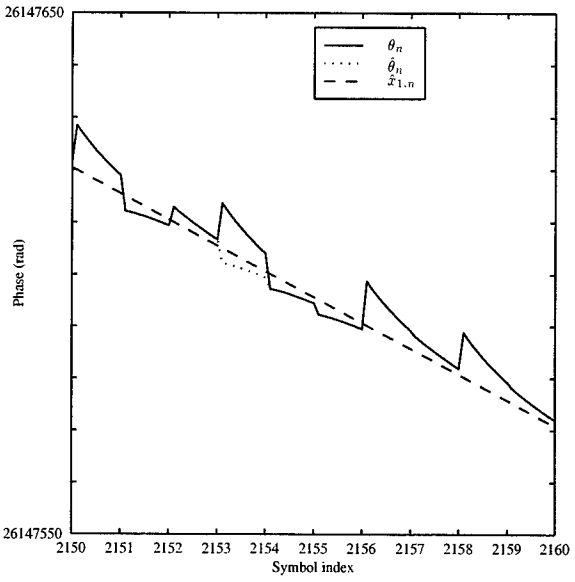


Fig. 5. Global phase process θ_n, its corresponding estimate $\hat{\theta}_n$ and dynamics phase $x_{1,n}$ ($E_b/N_0 = 8$ dB and $q_c = 50$ rad^2 s^{-6}). Notice that symbol 2153 is in error.

$T_s/N = 83$ μs), bit signal-to-noise ratio $E_b/N_0 = 8$ dB, and $r_c = T_s/(2E_b/N_0 \log_2 M)$.

Fig. 4 shows the satellite tracking ability, modelling phase as a double integrated Brownian motion with driving noise variance $q_c = 50$ rad s^{-6} as in Fig. 1. From 12000 transmitted symbols, corresponding to a time horizon of 10 seconds, only 9 symbols were detected in error. One of these situations can be seen in Fig. 5, where $\theta_n = x_{1,n} + y_n$. Notice the large phase variation along each symbol and the receiver recovering capacity after the false detection of symbol 2153.

5. CONCLUDING REMARKS

The proposed receiver is a parallel open-loop structure suited for DSP-based implementation. These allows to implement advanced algorithms required to optimally integrate all the available information. This was already the perspective of reference [7] in the beginning of the seventies, and the todays concept of *software radios*.

6. REFERENCES

[1] F. Giannetti, M. Luise, and R. Regiannini, "Simple carrier frequency rate-of-change estimators," *IEEE Transactions on Communications*, vol. 42, no. 9, September 1999.

[2] J. M. Leitão and F. D. Nunes, "A nonlinear filtering approach to carrier tracking and symbol detection in digital phase modulation," in *Proceedings of IEEE International Conference on Communications ICC'94*, New Orleans, 1994, pp. 386–390.

[3] F. D. Nunes and J. M. Leitão, "A nonlinear filtering approach to estimation and detection in mobile communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 9, December 1998.

[4] J. M. Leitão and J. M. Moura, "Acquisition in phase modulation: Application to ranging in radar/sonar systems," *IEEE Transactions on Aerosapace and Electronic Systems*, vol. 31, no. 2, 1995.

[5] J. M. Leitão and M. T. Figueiredo, "Absolute phase image reconstruction: A stochastic nonlinear filter approach," *IEEE Transactions on Image Processing*, vol. 7, June 1998.

[6] I. Ali, N. Al-Dhahir, and J. E. Hershey, "Doppler characterization for LEO satellites," *IEEE Transactions on Communications*, vol. 46, no. 3, March 1998.

[7] J. H. Painter and S. C. Gupta, "Recursive ideal observer detection on known m-ary signals in multiplicative and additive gaussian noise," *IEEE Transactions on Communications*, August 1973.

A STUDY OF TIME-FREQUENCY BASED DETECTORS FOR FSK MODULATED SIGNALS IN A FLAT FADING CHANNEL

B. Barkat

Nanyang Technological University
School of Electrical & Electronic Engineering
Block S2, Nanyang Avenue
Singapore 639798
Tel: +65-790-4386
E-mail: ebarkat@ntu.edu.sg

S. Attallah

National University of Singapore
Centre for Wireless Communications
20 Science Park Road
#02-34/37 TeleTech Park II
Singapore 117674
Tel: +65-870-9166

ABSTRACT

In this paper, we address the problem of channel fading in communication systems. In particular, we focus on the flat fading phenomenon. We study some time-frequency based techniques for the detection of frequency modulated signals subjected to flat fading channels. A comparison, based on bit error rate, of these techniques is also presented.

1. INTRODUCTION

In a wireless mobile communication system, a transmitted signal may experience *random* changes in its amplitude, phase and angle of arrival. These changes, referred to as fading, can be caused by multiple paths between the transmitter and receiver and/or by motion between the receiver and transmitter [1]. If the multiple paths are large in number and there is no line of sight signal component (no dominant component), the envelope of the received signal is statistically described by a Rayleigh probability density function [2].

Multipath fading results in two major degradation: frequency selective fading and frequency non-selective (or flat) fading [3]. Several techniques are available to combat fading [4].

In this paper, we focus on frequency shift keying (FSK) modulated signals transmitted through a channel subjected to flat fading. We review some time-frequency techniques used to retrieve such signals in a noisy environment. Also, we evaluate these techniques in terms of bit-error rate and compare them to standard methods used in telecommunications.

2. PRELIMINARIES

It can be shown that for a complex signal $z(t)$, transmitted over a flat fading channel, the received complex signal is given by [3]

$$y(t) = m(t) \cdot z(t) \tag{1}$$

where $m(t)$ is a complex Gaussian process. This channel has also been called a *multiplicative fading* channel. The above equation indicates that the original signal gets corrupted by a process $m(t)$ whose amplitude can be modeled by a Rayleigh density function while its phase is uniformly distributed.

The dramatic drop in power of the signal, due to fading, makes it very difficult to be detected. In some situations, such as FSK modulated signals, the signal information is contained in its instantaneous frequency (IF). Thus, by using an appropriate tool to estimate the IF of the modulated signal, we may be able to retrieve the signal without having to use expensive and very complex receivers. In this paper, we review some time-frequency techniques in order to detect the original transmitted signal.

The field of time-frequency signal analysis is one of the recent developments which provides suitable tools for analysing non-stationary signals, characterised by a time-varying spectral contents, occurring in many fields of engineering [5]. Time-frequency distributions (TFDs) are natural extensions of the Fourier transform. They map a one dimensional signal, function of time only, to a two dimensional quantity, function of time and frequency. One of the most popular TFD is the Wigner-

0-7803-7011-2/01/$10.00 ©2001 IEEE

Ville distribution (WVD) defined as [5]

$$W(t,f) = \int_{-\infty}^{+\infty} [z(t+\frac{\tau}{2}) \cdot z^*(t-\frac{\tau}{2})] e^{-j2\pi f\tau} d\tau \quad (2)$$

where $z(t)$ is the analytic form of the real-valued signal under investigation. The WVD first moment yields the IF of the analysed signal [6], defined as [7]

$$f_i(t) = \frac{1}{2\pi} \frac{d\phi(t)}{dt} \quad (3)$$

where $\phi(t)$ is the phase of the signal $z(t)$.

In practice, and considering orthogonal binary FSK (BFSK), the original real-valued signal is generated as

$$s(t) = \sqrt{\frac{2E_s}{T}} \cos(\omega_0 t + d(t)\ \Omega\ t)$$

where ω_0 is the carrier (angular) frequency, Ω a constant offset, $d(t) = 1$ or -1 (depending on whether the bit 1 or 0 has been transmitted), E_s is the symbol energy and T is the signaling period. At the receiver, we have

$$r(t) = \alpha \sqrt{\frac{2E_s}{T}} \cos(\omega_0 t + d(t)\ \Omega\ t + \phi) + n(t) \quad (4)$$

where α is the fading coefficient assumed to have a Rayleigh distribution, ϕ is a random phase uniformly distributed over $[0, 2\pi]$.

For a flat fading channel α and ϕ are assumed to be constant over one signaling period T and the additive noise $n(t)$ is assumed to be zero-mean white Gaussian with a variance equal to $\sigma_n^2 = N_0/2$. In this case, it can be shown that for an envelope or a square wave detector the bit error rate (BER) for a non-coherent detection is given by [8]

$$P_e = \frac{1}{2+\gamma} \quad (5)$$

and for a coherent detection it is [8]

$$P_e = \frac{1}{2}\left[1 - \sqrt{\frac{\gamma}{2+\gamma}}\right] \quad \text{with} \quad \gamma = \frac{E_s}{N_0} \cdot E[\alpha^2] \quad (6)$$

with $E[\cdot]$ being the expectation operator.

3. PROPOSED IF ESTIMATORS

The received signal $r(t)$, over one signaling period is assumed to be a constant amplitude sinusoid. The WVD, defined above, can be used to estimate the frequency of $r(t)$. This can be done by first, evaluating the time-frequency distribution of the received signal (or its analytic version) and, then, searching for the maximum of the distribution for every time instant. The WVD performance can be shown to degrade significantly at low signal-to-noise ratio (SNR).

In order to improve the statistical performance of the signal detection, one can use the B-distribution. The B-distribution is defined as[9]

$$W_B(t,f) = \int\int_{-\infty}^{+\infty} G(t',\tau)[z_r(t-t'+\frac{\tau}{2}) \cdot z_r^*(t-t'-\frac{\tau}{2})] \times e^{-j2\pi f\tau} dt' d\tau \quad (7)$$

where $G(t,\tau)$ is a function given by

$$G(t,\tau) = \left(\frac{|\tau|}{\cosh(t)}\right)^{\sigma}$$

and σ is a real parameter. We see that the B-distribution is similar to the WVD but instead of taking the FT of the product, we must first convolve (in the time variable t) the product $[z_r(t+\frac{\tau}{2}) \cdot z_r^*(t-\frac{\tau}{2})]$ with the function $G(t,\tau)$ and then take the FT of the result. In order to estimate the IF of the analysed signal, we can use the peak of the B-distribution to obtain it. As a quick qualitative comparison, consider the WVD and the B-distribution of a sinusoid in additive white Gaussian noise with 0 dB SNR. These two distributions are displayed in Figures 1 and 2 respectively. Observe the superiority of the B-distribution in suppressing the noise.

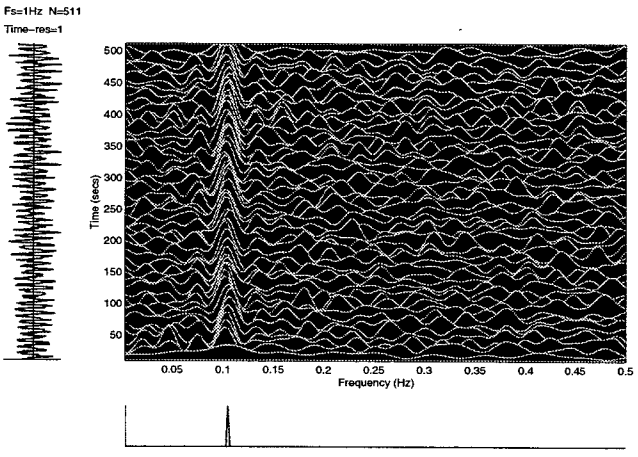


Figure 1: The WVD of a sinusoid in 0 dB noise.

In figures 3 and 4, we plot the IF estimates, of a sinusoid (normalised frequency=0.25) embedded in -4 dB noise, using the WVD and the B-distribution respectively. Once again, we can observe from these plots that the B-distribution gives a better result compared to the WVD.

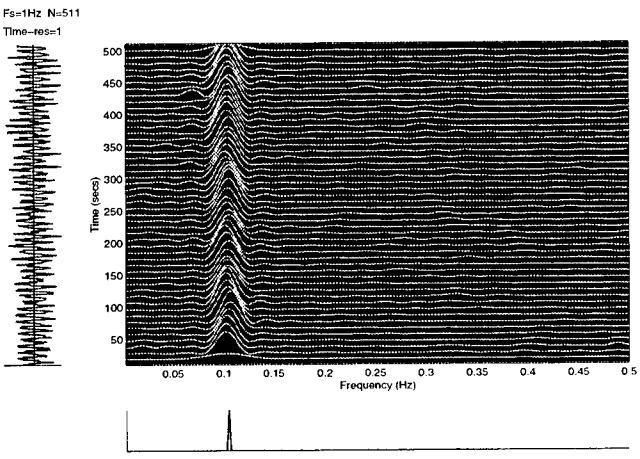


Figure 2: The B-distribution of a sinusoid in 0 dB noise.

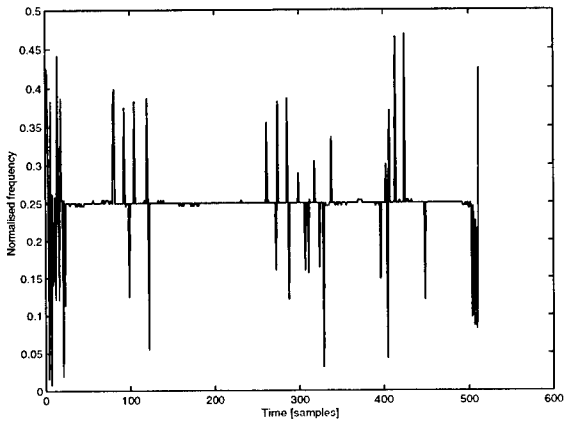


Figure 3: WVD based IF estimate of a sinusoid in -4 dB noise.

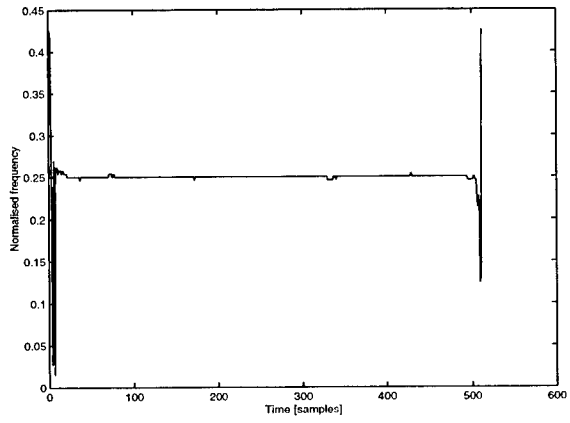


Figure 4: B-distribution based IF estimate of a sinusoid in -4 dB noise.

In what follows, we will use the above time-frequency distributions, namely, the WVD and the B-distribution in order to retrieve the IF of a signal transmitted over a flat fading channel. We will compare the performance of these distributions in terms of their respective bit error rates.

4. COMPARISONS

As stated above, in the comparisons, we limit our discussion to a flat fading channel only. We can have two cases: (i) the transmitted signals are totally unknown (exept that they are sinusoids) and (ii) the signals are known but we don't know which one was sent. For the first case, unknown frequencies of the transmit signals, we apply the time-frequency distributions directly on the received signal in order to decide which frequency is present. For the second case, the transmitted signals are known and we can incorporate this information in the time-frequency distribution in order to decide which frequency is present in the received signal.

Let us first consider the case of totally unknown transmitted signals. For that, we generate two orthogonal sinusoids $s_0(t)$ and $s_1(t)$. To account for the flat fading channel, we multiply each of these two signals by α (a value taken from a Rayleigh distribution such that $E(\alpha^2) = 1$). An initial random phase ϕ as well as some zero-mean Gaussian noise $n(t)$ are added to the signals, as suggested by Equation 4. The signal $r(t)$ (more precisely its analytic form) is then analysed using the time-frequency distributions and the peaks of these distributions will yield the corresponding frequency of the received signal. Based on this frequency, we decide which symbol ($s_0(t)$ or $s_1(t)$) was sent in that particular symbol interval. When the noise power increases, we tend to make more errors in our decision. Since we have a binary modulation, the number of errors divided by the total number of transmitted symbols constitutes the bit error rate. Figure 5 displays the BER versus the energy-to-noise ratio γ (in dB) for each time-frequency distribution. For comparison purposes, we have also analysed the transmitted signal using the periodogram. It is seen that the performance of the B-distribution is close to that of the periodogram (which is the optimal detector for a sinusoid). Note that since the transmitted signal is just a sinusoid, we expect a constant frequency over the whole symbol interval in the time-frequency plane (see for instance Figure 2). Thus, we average the time-frequency distribution (over time) and then search for the maximum to obtain an estimate of the transmitted frequency.

Now, we consider that we know the signals to be transmitted but we don't know which one has been transmitted at the particular time interval (symbol interval) of interest. In this case, we can incorporate this

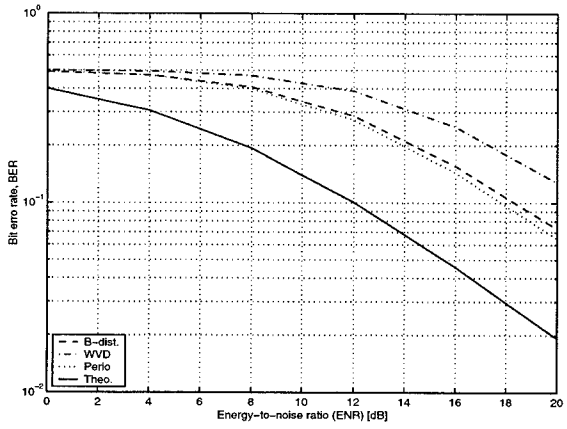


Figure 5: Performance of the various time-frequency distributions for a Rayleigh fading channel when there is no knowledge of the transmitted signals.

information in our time-frequency distributions in the detection process. An analytic version of the received signal $r(t)$ is multiplied by the analytic version of $s_0(t)$ and the analytic version of $s_1(t)$ respectively. Using the B-distribution or the WVD on the product, we can easily know the frequency present in $r(t)$. The BER of the time-frequency techniques, along with that of the periodogram, for this detection procedure are plotted in Figure 6.

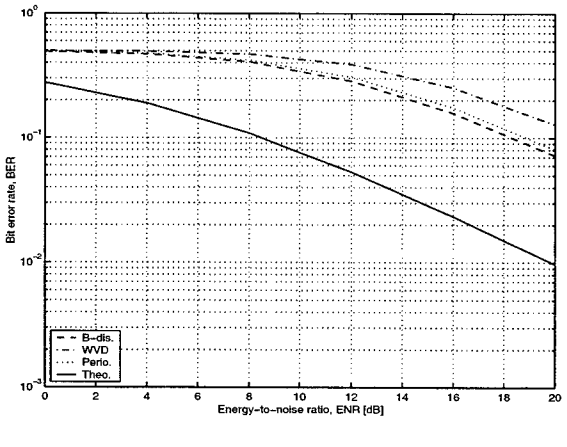


Figure 6: Performance of the various time-frequency distributions for a Rayleigh fading channel using the knowledge about the transmitted signals.

In our present situation, we see that the gain in using the knowledge of the transmitted signals in the detection does not improve significantly the performance of the detector.

We should note that, the time-frequency techniques are very robust and can be applied to other FM signals such as an M-ary FSK.

We also note that the proposed detector can be applied in global system for mobile communications (GSM) because this system uses Gaussian minimum shift keying (GMSK) modulation which can be non-coherently detected as simple FSK [2].

5. CONCLUSION

In this paper, we addressed the problem of retrieving FM signals transmitted over a flat or Rayleigh fading channel. In particular, we have applied some time-frequency tools to analyse the received signal. We have seen that these tools can be used whether we know the transmitted signals or not. A comparison, based on the bit-error rate, between these techniques has also been presented. The results show that the B-distribution gives a better detection performance compared to the WVD.

6. REFERENCES

[1] B. Sklar. Rayleigh Fading Channels in Mobile Digital Communication Systems Part I: Characterisation. *IEEE Communications Magazine*, pages 90–100, 1997.

[2] T.S. Rappaport. *Wireless Communications.* Prentice-Hall, Upper-Saddle River, NJ, USA, 1996.

[3] S. Stein. Fading Channel Issues in System Engineering. *IEEE Journal on Selected Areas in Communications*, SAC-5:68–89, Feb. 1987.

[4] B. Sklar. Rayleigh Fading Channels in Mobile Digital Communication Systems Part II: Mitigation. *IEEE Communications Magazine*, pages 102–109, 1997.

[5] L. Cohen. *Time-Frequency Analysis.* Prentice-Hall, 1995.

[6] F. Hlawatsch and G.F. Boudreaux-Bartels. Linear and quadratic time-frequency signal analysis. *IEEE Signal Processing Magazine*, 9 (2):21–67, 1992.

[7] J. Ville. Theorie et application de la notion de signal analytique. *Cables et Transmissions*, 2A(1):61–74, 1948.

[8] J.G. Proakis. *Digital Communications.* McGraw-Hill, third edition, 1995.

[9] B. Barkat and B. Boashash. A High-Resolution Quadratic Time-Frequency Distribution for Multicomponent Signals Analysis. *IEEE Trans. on Signal Processing*, 2001. (In print).

MRC RECEIVER PERFORMANCE WITH MQAM IN CORRELATED RICIAN FADING CHANNELS

Chunhua Yang, Guoan Bi

School of Electrical and Electronic Eng.,
Nanyang Technological University,
Singapore.

A. R. Leyman

Digital Comm. Strategic Research Group,
Center for Wireless Communication,
20, Science Park Road, Science Park II,
Singapore.

ABSTRACT

Due to difficulties in deriving the probability density function, performance of MRC diversity receiver in correlated Rician fading channels is rarely reported in the literature. This letter shows that the difficulty can be avoided by a linear transformation technique. General closed-form expressions of average symbol error rate for various modulation schemes can be easily derived. As an example, this letter derives the SER of MQAM over correlated Rician fading channels.

1. INTRODUCTION

Diversity is an effective technique to combat the detrimental effects of multipath fading. Previous work on performance analysis of diversity reception mainly focused on the case of independent fading with binary modulation schemes. In [1] the performance of an L-branch equal gain combiner on independent Rician fading channels was derived. In [3], the average bit error rate (BER) of a BPSK system with MRC on a general Rician fading channel was studied. Subsequently, the average BERs of M-ary modulated signals for non-diversity reception over Rician fading were presented in [2]. The exact expressions of SER for multilevel modulated signals with MRC over Rician fading channels are seldom reported in the literature possibly because the difficulties in deriving the probability density function (PDF).

In this letter, we show that the difficulty in deriving the PDF can be avoided by a linear transformation technique to obtain the required characteristic function (CF). The exact expressions of SER can be easily derived for the MRC diversity receiver with multilevel quadrature amplitude modulation (MQAM) in the correlated Rician fading channels. The method is simple and general enough to be used for any correlated signal model with arbitrary fading parameters.

2. CHARACTERISTIC FUNCTION

Consider an L-branch MRC over the correlated Rician fading channel and assume the received signals from the L-branch diversity system in complex Gaussian form to be $\mathbf{x}(t) = \mathbf{x_c}(t) + j\mathbf{x_s}(t)$, where the real part $\mathbf{x_c}(t) = [x_{c1}, \cdots, x_{cL}]$ and imaginary part $\mathbf{x_s}(t) = [x_{s1}, \cdots, x_{sL}]$ are Gaussian Random processes with $E[\mathbf{x_c}] = \mathbf{c} = [c_1, \cdots, c_L]$, $E[\mathbf{x_s}] = \mathbf{0}$, and covariance matrix $\mathbf{R}$. The ijth element of $\mathbf{R}$ is $\rho_{ij}\sigma_i\sigma_j$, where ρ_{ij} is the correlation coefficient, σ_i^2 is the variance of x_{ci} or x_{si}.

The resultant SNR at the output of L-branch MRC is

$$\gamma = \sum_{k=1}^{L} \frac{x_{ck}^2 + x_{sk}^2}{N_0}. \tag{1}$$

The characteristic function (CF) of γ was given in [5]:

$$\Phi(j\nu) = E[e^{j\nu\gamma}] = \frac{exp\left[j\frac{\nu}{N_0}\mathbf{c}^H\left(\mathbf{I} - j\frac{2\nu}{N_0}\mathbf{R}\right)^{-1}\mathbf{c}\right]}{det(\mathbf{I} - j\frac{2\nu}{N_0}\mathbf{R})} \tag{2}$$

which is difficult to be used in the performance analysis. Because the covariance matrix $\mathbf{R}$ is positive definite, it can be diagonalized with an orthonormal matrix $\mathbf{Q}$, defined as

$$\mathbf{QRQ}^T = \mathbf{Q}^T\mathbf{RQ} = \mathbf{\Lambda} \tag{3}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with elements $\lambda_k > 0$ ($k = 1, \cdots, L$), λ_k is the kth eigenvalue of the covariance matrix $\mathbf{R}$, and $\mathbf{Q}$ is the orthonormal matrix, composed of the eigenvectors for $\mathbf{R}$.

By using the orthonormal matrix $\mathbf{Q}$, the characteristic function of γ can be simplified as

$$\Phi(j\nu) =$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

$$exp\left[\frac{j\nu}{N_0}\mathbf{c}^H\mathbf{Q}^H\left(\mathbf{I}-\frac{2j\nu}{N_0}\mathbf{\Lambda}\right)^{-1}\mathbf{Qc}\right]$$
$$\prod_{k=1}^{L}\frac{1}{1-2j\nu\xi_k}$$
$$=exp\left[\frac{j\nu}{N_0}\eta^H\left(\mathbf{I}-\frac{2j\nu}{N_0}\mathbf{\Lambda}\right)^{-1}\eta\right]$$
$$\prod_{k=1}^{L}\frac{1}{1-2j\nu\xi_k}$$
$$=exp\left[\sum_{k=1}^{L}\frac{j\nu\mu_k}{1-2j\nu\xi_k}\right]\prod_{k=1}^{L}\frac{1}{1-2j\nu\xi_k} \quad (4)$$

where $\eta=\mathbf{Qc}$, $\xi_k=\frac{\lambda_k}{N_0}$, and $\mu_k=\frac{\eta_k^2}{N_0}$. It is seen that (4) is the CF of the Hermitian form for an independent complex Gaussian process $\mathbf{y}=\mathbf{Qx}$ with the mean η and covariance matrix $\mathbf{\Lambda}$. The CF of the output SNR of correlated branch signal $\mathbf{x}$ is obtained by deriving the CF of the output SNR of the new independent signal $\mathbf{y}$ obtained by multiplying $\mathbf{x}$ with $\mathbf{Q}$, where $\mathbf{Q}$ is considered to be the transformation matrix. The PDF of the resultant SNR γ can be easily obtained by taking inverse Fourier transform of (4)
The conditional SER for square QAM is given by [4]

$$P_s(\gamma)=2q\,erfc\left(\sqrt{p\gamma}\right)-q^2erfc^2\left(\sqrt{p\gamma}\right) \quad (5)$$

where $q=1-\frac{1}{\sqrt{M}}$, $p=1.5\,log_2\frac{M}{M-1}$.
By using the expressions [4]

$$erfc\left(\sqrt{b\gamma}\right)=\frac{2}{\pi}\int_0^{\frac{\pi}{2}}exp\left[-b\gamma csc^2\theta\right]d\theta$$
$$erf^2\left(\sqrt{b\gamma}\right)=\frac{4}{\pi}\int_0^{\frac{\pi}{4}}exp\left[-b\gamma csc^2\theta\right]d\theta, \quad (6)$$

for MQAM is derived by

$$P_{ave}=\int_0^{\infty}P_s(\gamma)p_\gamma(\gamma)d\gamma$$
$$=\int_0^{\infty}\left\{\frac{4q}{\pi}\int_0^{\frac{\pi}{2}}exp\left[-p\gamma csc^2\theta\right]d\theta\right.$$
$$\left.-\frac{4q^2}{\pi}\int_0^{\frac{\pi}{4}}exp\left[-p\gamma csc^2\theta\right]d\theta\right\}p_\gamma(\gamma)d\gamma. \quad (7)$$

Exchanging the order of integrals and using the definition of (2), (7) can be expressed as

$$P_{ave}=\frac{4q}{\pi}\int_0^{\frac{\pi}{2}}\Phi\left(-pcsc^2\theta\right)d\theta$$
$$-\frac{4q^2}{\pi}\int_0^{\frac{\pi}{4}}\Phi\left(-pcsc^2\theta\right)d\theta. \quad (8)$$

The SER in (8) is in a finite-range integral with an integrand, which is a product of polynomial and exponential functions. We assume that omni-directional antennas are arranged in one-dimension. The distance between the adjacent antennas is d and the spacial correlation follows an exponential function. The elements of the covariance matrix $\mathbf{R}$ for $\mathbf{x_c}$ or $\mathbf{x_s}$ are given by:

$$R_{ij}=\rho_{ij}\sigma_i\sigma_j=\sigma_i\sigma_j exp\left[-\frac{k}{2}(i-j)^2\left(\frac{d}{\lambda}\right)^2\right] \quad (9)$$

where λ is the carrier wavelength and $k=21.4$. Based on (8), numerical results, as shown in Figs.1 and 2, are calculated for different values of Rician factor K. Fig. 1 compares the SERs for $m=4,16$ and 64 with correlated or noncorrelated fading channels. Figure 2 shows that when $d/\lambda>0.4$, the SER becomes constant, which means that the distance between adjacent antennas can be as small as 0.4λ.

3. CONCLUSION

By transforming the correlated racian fading signals, we can easily obtain the characteristic function of the virtually independent racian fading signals. This letter illustrates the derivation of the SER with MQAM in correlated racian fading channels. The SERs with other multilevel modulation such as MPSK and MFSK can be derived in a similarly way.

4. REFERENCES

[1] Adnan A. Abu-Dayya, and Norman C. Beaulieu, "Microdiversity on Rician Fading Channels." *IEEE Trans. on Commun.* , Vol. 42, No. 6, pp. 2258-2267, Jun. 1994.

[2] Jonqyin Sun, and Irving S. Reed, "Performance of MDPSK, MPSK, and Noncoherent MFSK in Wireless Rician Fading Channels." *IEEE Trans. on Commun.* , Vol. 47, No. 6, pp. 813-816, Jun. 1999.

[3] Veeravalli V.V, and Mantravadi A., "Performance Analysis For Diversity Reception of Digital Signals over Correlated Fading Channels." in *VTC'99,* Vol. 3, pp. 1291-1295, May. 1999.

[4] A. Annamalai, and C. Tellambura,"Analysis of Maximal-Ratio and Equal-Gain Diversity Systems

for M-ary QAM on Generalized Fading Channels," *ICC'99,* Vol. 2, pp.848-852, June, 1999.

[5] M. Shwarts, W.R. Bennett, "Communication Systems and Techniques," New York: McGraw-Hill, 1969.

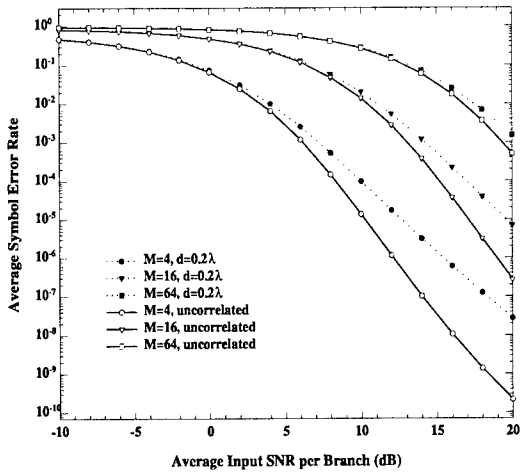


Figure 1: Average BER versus average input SNR for MQAM ($M = 4, 16, 64$) with MRC over the correlated Rician fading channels ($L = 3,\ d/\lambda = 0.2, \infty,\ K = 5$).

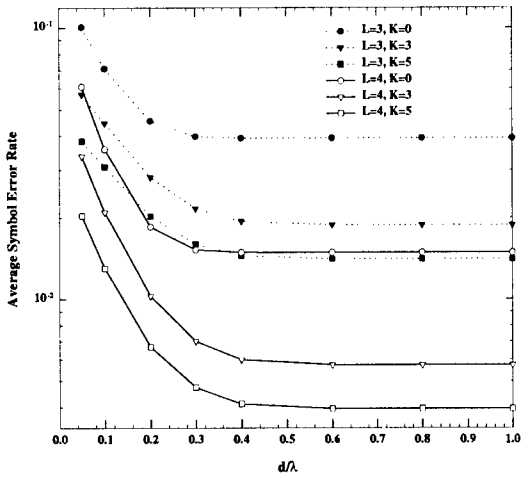


Figure 2: Average BER versus d/λ for MQAM receiver with MRC over the correlated Rician fading channels ($M = 16,\ L = 3, 4,\ K = 0, 3, 5$, and SNR/per channel =10dB)

OPTIMAL PREPROCESSING FOR SOURCE LOCALIZATION BY FEWER RECEIVERS THAN SENSORS

Joseph Tabrikian and Avi Faizakov

Dept. of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer Sheva 84105, ISRAEL
Email: joseph@ee.bgu.ac.il

ABSTRACT

In many source detection and localization applications, the number of receivers is smaller than the number of sensors, where the receivers are connected to the sensors via a preprocessing device. The preprocessing device can easily be implemented in hardware by linear transformation of the measured signals at the sensors. In this paper, the Cramér-Rao lower bound for this problem is developed and it is shown that by judicious choice of the preprocessing matrix, it is possible to reduce the bound on direction of arrival estimation errors. The results demonstrate the trade-off between azimuth and elevation estimation errors using a planar array divided to subarrays.

1. INTRODUCTION

Traditional array processing techniques assume that *simultaneous* measurements from *all* sensors are available. However, in many applications, such as radar and satellite, the number of sensors may be large, and using the same number of receivers as sensors, results in large number of receivers and A/D converters, which are expensive specially in wide-band applications with high sampling rate. Furthermore, source detection and localization algorithms which are applied to data received by large arrays is computationally expensive. Therefore, in practice, it is desired to reduce the number of receivers to be lower than the number of sensors. This solution enables to process lower amount of data without reducing the antenna aperture. This approach requires a transformation of the received signal at the array to data on which the source detection and localization algorithms can be applied. In [1], the Maximum-Likelihood estimator for source localization using fewer receivers than sensors has been presented. This approach assumes a linear, time-varying transformation unit as a preprocessing stage. That is

$$\mathbf{x}(t_l) = \mathbf{G}(t_l)\mathbf{y}(t_l) \ , \tag{1}$$

where $\mathbf{x}(t_l)$ is the received signal at the array at time t_l, $\mathbf{y}(t_l)$ is the input signal to the processor, and the matrix $\mathbf{G}(t_l)$ is the linear, time-varying transformation matrix.

In [2] this approach has been adopted with 2 receivers where the preprocessing unit contains two switches, i.e. the matrix $\mathbf{G}(t_l)$ contains zeros and ones. In [1] and [2], it is assumed that the transformation matrix is given.

Our goal in this paper is to obtain the optimal transformation matrix $\mathbf{G}(t_l)$ by means of the Cramér-Rao lower bound (CRLB). Two cases are examined. In the first, a two dimensional array with time-invariant transformation matrix is considered. A possible application for this problem is for phased-array radar systems which may be divided into several subarrays. A single receiver is assigned for the output of each subarray unit. In the second problem, we assume that multiple snapshots are available with time-varying transformation matrix, where the criterion for optimization is the CRLB on source direction estimation error variance.

2. PROBLEM FORMULATION

Consider a far-field source radiating a narrowband signal, received by a plane array of N sensors. The complex envelope of the vector of the received signal at the sensors is given by:

$$\mathbf{y}_l = \mathbf{a}(u,v)s_l + \mathbf{n}_l, \quad l = 1, \cdots, L, \tag{2}$$

where s_l is the complex envelope of the signal at the lth snapshot and $\mathbf{a}(u,v)$ is the array steering vector. The signal snapshots, $(s_1, \cdots, s_L)$, are assumed to be deterministic, unknown. The samples of the noise vector, $\{\mathbf{n}_l\}_{l=1}^{L}$ are assumed to be zero-mean, complex-Gaussian and i.i.d: $\mathbf{n}_l \sim N^c(\mathbf{0}, \sigma_n^2\mathbf{I})$, where the noise variance, σ_n^2 is known.

The source location parameters, u and v, are given by $u \triangleq sin\phi cos\theta$, and $v \triangleq cos\phi\cos\theta$, where ϕ and θ are the source azimuth and elevation, respectively. The elements of the steering vector are given by: $a_n(u,v) = e^{j\frac{2\pi}{\lambda}(d_{xn}u + d_{yn}v)}$ where the vector (d_{xn}, d_{yn}) denotes the nth sensor location.

Because of the limited number of receivers, the measurements $\mathbf{y}_l$ are linearly transformed to provide the input to the receivers according to (1). Now the data model is

$$\mathbf{x}_l = \mathbf{G}_l\mathbf{a}(u,v)s_l + \underbrace{\mathbf{G}_l\mathbf{n}_l}_{\mathbf{e}_l} . \tag{3}$$

The tranformed noise vector, $\mathbf{e}_l$, is now zero-mean, complex-Gaussian with covariance matrix,

$$\mathbf{R}_\mathbf{e} \triangleq \sigma_n^2\mathbf{G}_l\mathbf{G}_l^H . \tag{4}$$

In this model, it is assumed that the transformation matrix can be updated at each snapshot. A simple implementation of the preprocessing stage is by using subarrays, where the array is divided into groups of sensors which are linearly combined to provide the input of each receiver. In other words, the matrix $\mathbf{G}$ can be formed as a block-diagonal matrix whose blocks are row vectors which express the complex weights for the sensor of the subarray. Now, the number of parameters of the preprocessing stage to be determined is equal to the number of sensors at the array.

3. CRAMÉR-RAO LOWER BOUND

3.1. Case 1: Single Snapshot

In this section, we first develop the CRLB for a single snapshot case. To simplify the notation, the subscript $_l$ is dropped when considering a single snapshot case.

In the model of (3), the information on the unknown parameters is in the mean of the data, and therefore, the Fisher Information Matrix (FIM) for estimating the vector of unknown real parameters $\mathbf{\Psi}$ is given by [3]:

$$\mathbf{J}_\mathbf{\Psi} = 2Re\left\{\left(\frac{\partial(\mathbf{a}(u,v)s)}{\partial\mathbf{\Psi}}\right)^H \mathbf{G}^H\mathbf{R}_\mathbf{e}^{-1}\mathbf{G}\frac{\partial(\mathbf{G}\mathbf{a}(u,v)s)}{\partial\mathbf{\Psi}}\right\}, \tag{5}$$

and the CRLB on estimation errors of $\mathbf{\Psi}$ is given by $\mathbf{J}_\mathbf{\Psi}^{-1}$. The vector of unknown parameters, $\mathbf{\Psi}$ includes the source location parameters (u,v), in addition to the real and imaginary parts of the signal. The bound on the source azimuth and elevation, (ϕ,θ) can be obtained from $\mathbf{J}_\mathbf{\Psi}^{-1}$ by using the chain rule.

By using the expression for $\mathbf{R}_\mathbf{e}$ in (4), one obtains: $\mathbf{Q} \triangleq \mathbf{G}^H\mathbf{R}_\mathbf{e}^{-1}\mathbf{G} = \frac{1}{\sigma_n^2}\mathbf{G}^H\left(\mathbf{G}\mathbf{G}^H\right)^{-1}\mathbf{G}$. For subarrays preprocessing configuration, the matrix $\mathbf{G}$ can be written as:

$$\mathbf{G} = \begin{bmatrix} \mathbf{w}_1^H & \mathbf{0}^H & \cdots & \mathbf{0}^H \\ \mathbf{0}^H & \mathbf{w}_2^H & & \mathbf{0}^H \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^H & \cdots & \mathbf{0}^H & \mathbf{w}_{N_{SA}}^H \end{bmatrix}, \tag{6}$$

where $\mathbf{w}_n^H$ is a row vector which denotes the weight vector for the elements of the nth subarray. Thus, the matrix $\mathbf{Q}$ is a block-diagonal matrix whose mth block is $\frac{1}{\sigma_n^2}\frac{\mathbf{w}_m\mathbf{w}_m^H}{\mathbf{w}_m^H\mathbf{w}_m}$. Now derivation of the FIM from (5) is straight-forward.

3.2. Case 2: Multiple Snapshots

The measurements at different snapshots are independent, and therefore, it can easily be shown that the FIM in this case is given by:

$$\mathbf{J}_\mathbf{\Psi} = \sum_{l=1}^{L}\mathbf{J}_{\mathbf{\Psi}_l} \tag{7}$$

where $\mathbf{J}_{\Psi_l}$ is the FIM given the lth snapshot, and Ψ contain all the unknown parameters, that is (u, v) in addition to the real and imaginary parts of the unknown signal $(s_1, \cdots, s_L)$. Note that $\mathbf{J}_{\Psi_l}$ is a function of $\mathbf{G}_l$ only. However, the CRLB is a non-linear function of $\mathbf{G}_1, \cdots, \mathbf{G}_L$.

4. EXAMPLES

4.1. Example 1: Single snapshot

In order to demonstrate the trade-off between the azimuth and elevation estimation errors using a planar array, the following example is presented. Consider an array of two linear, horizontal subarrays (along x axis). The subarrays are set in parallel at different heights, and the array consists of 2x2=4 elements where each row/subarray consists of 2 elements. The horizontal and vertical distance between adjacent sensors is d. With no loss of generality we can impose the weight of a single element of each subarray to be one, and the matrix $\mathbf{G}$ becomes

$$\mathbf{G} = \begin{bmatrix} 1 & w_0 & 0 & 0 \\ 0 & 0 & 1 & w_1 \end{bmatrix}. \tag{8}$$

Now, the CRLB on the source location parameters (u, v) can be expressed as a function of (w_0, w_1). Assuming a single snapshot with known signal, it can be shown that the FIM for this problem is given by:

$$\mathbf{J}_{uv} = 2\left(\frac{2\pi d}{\lambda}\right)^2 \frac{|s|^2}{\sigma_n^2} \cdot \tag{9}$$

$$\begin{bmatrix} \frac{|w_0|^2}{1+|w_0|^2} + \frac{|w_1|^2}{1+|w_1|^2} & \frac{|w_1|^2 + Re\left(w_1 e^{-j\frac{2\pi d}{\lambda}u}\right)}{1+|w_1|^2} \\ \frac{|w_1|^2 + Re\left(w_1 e^{-j\frac{2\pi d}{\lambda}u}\right)}{1+|w_1|^2} & 1 + 2\frac{Re\left(w_1 e^{-j\frac{2\pi d}{\lambda}u}\right)}{1+|w_1|^2} \end{bmatrix},$$

where λ is the signal wavelength. For this example, the distance between adjacent sensors, d, was chosen to be half a wavelength. Fig. 1 presents the bound on both parameters u and v as a function of ξ, where ξ is related to w_1 via: $w_1 = e^{j\frac{2\pi d}{\lambda}\xi}$. This figure demonstrates the trade-off between the bounds on the two unknown parameters, u and v. Figs. 2 and 3 show the bounds on the two unknown parameters as a function of absolute value and phase of w_1 where w_0 was set to 1.

In cases where the number of sensors is larger than in the above example, or in multiple snapshots case, the number of unknown parameters (non-zero elements of the matrices $\mathbf{G}_l$) may be large and the multidimensional optimization problem is solved numerically.

4.2. Example 2: Multiple snapshots

Consider an array of two linear, horizontal subarrays (along x axis). The subarrays are set in parallel at different heights, i.e. the array consists of 2x4=8 elements where each row contains 4 elements. The distance between the sensors is half a wavelength in both axis. A source is located at azimuth $\phi = 30^o$ and elevation $\theta = 0^o$, that is $(u, v) = (\frac{1}{2}, \frac{\sqrt{3}}{2})$. The signal source is assumed to be unknown deterministic. Two snapshots with two different transformation matrices were assumed. The optimization criterion was to minimize the azimuth estimation error standard-deviation (STD) while ignoring the bound on source elevation error. The bound is minimized with respect to the unknown parameters of the matrices $\mathbf{G}_l$ using the Genetic Algorithm, and it is compared to the bound that is obtained using a transformation matrix for conventional beamforming in the direction of the source. The results show that by optimization of the preprocessing stage, the bound on azimuth error STD can be reduced from 1.82^o to 0.47^o. This improvement is achieved in the cost of greater bound on elevation error STD. Using this optimization procedure, one is able to control the accuracy in azimuth, elevation or any combination of them.

In the above examples, the bounds were calculated, assuming that the source direction is known. Although in some applications, such as radar[1], the source direction may be roughly known, this may not be the case in many others. For these problems the minimax criterion may be applied, that is,

[1]The illuminated targets are within the radar beam.

$$(\hat{\mathbf{G}}_1, \cdots, \hat{\mathbf{G}}_L) = \arg \min_{\mathbf{G}_1,\cdots,\mathbf{G}_L} \{\max_{u_o,v_o} CRLB(u_o, v_o)|_{\mathbf{G}_1,\cdots,\mathbf{G}_L}\} \quad (10)$$

where $CRLB(u_o, v_o)$ denotes the evaluated CRLB on source location parameter, (u, v), or any combination of them, where the source is located at (u_o, v_o).

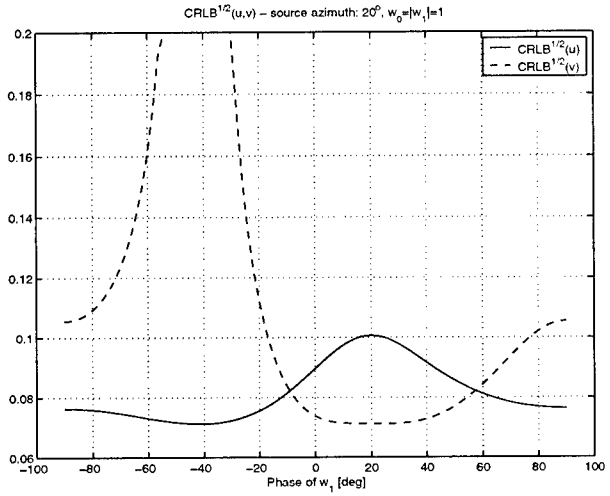


Figure 1: The CRLB on u and v as a function of ξ: $w_1 = e^{j\frac{2\pi d}{\lambda}\xi}$, $w_0 = 1$.

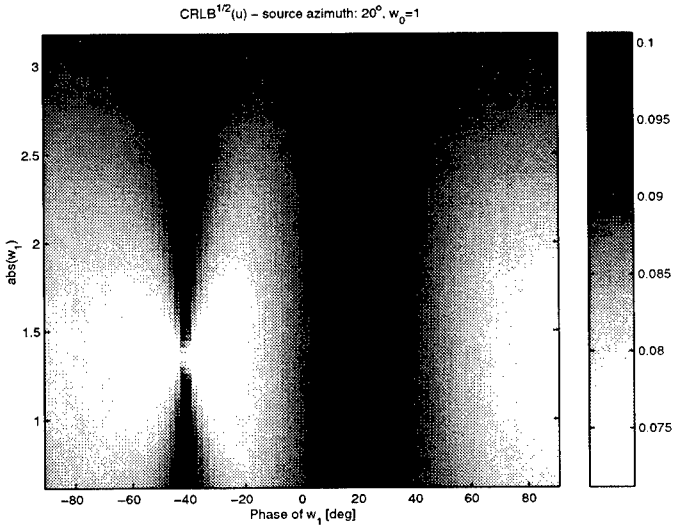


Figure 2: The CRLB on u as a function of $|w_1|$ and ξ: $w_1 = |w_1|e^{j\frac{2\pi d}{\lambda}\xi}$, $w_0 = 1$.

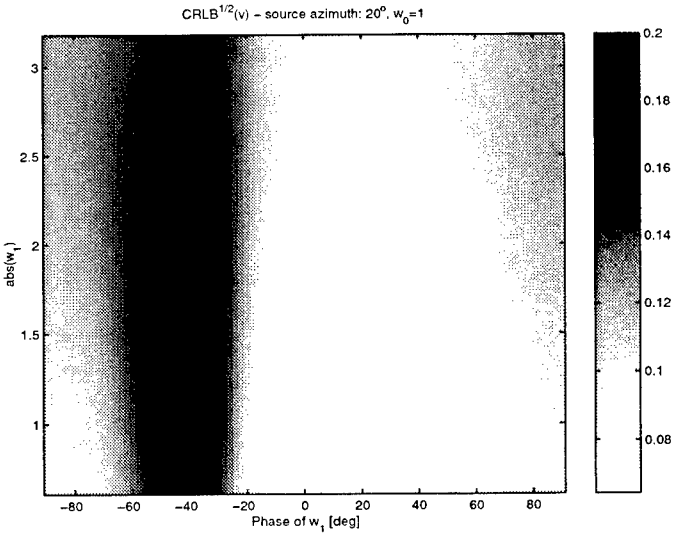


Figure 3: The CRLB on v as a function of $|w_1|$ and ξ: $w_1 = |w_1|e^{j\frac{2\pi d}{\lambda}\xi}$, $w_0 = 1$.

5. CONCLUSIONS

The problem of determination of the linear preprocessing matrix for source localization with fewer receivers than sensors is addressed. The CRLB for this problem is developed and used as the criterion for optimization. For a planar array divided into subarrays, the trade-off between azimuth and elevation error STD's is demonstrated. The problem of determination of linear, time-varying preprocessing stage is also investigated.

REFERENCES

[1] J. Sheinvald and M. Wax, "Direction finding with fewer receivers via time-varying preprocessing," *IEEE Trans. on SP*, vol. 47, 2-10, January 1999.

[2] E. Fishler and H. Messer, "Multiple source direction finding with an array of M sensors using two receivers," *Proc. of 10th IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, pp. 86-89, Pennsylvania, August 14-16, 2000.

[3] S. M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory*, Prentice Hall, 1993.

A ROBUST TECHNIQUE FOR ARRAY INTERPOLATION USING SECOND-ORDER CONE PROGRAMMING

Marius Pesavento[†] *Alex B. Gershman*[*] *Zhi-Quan Luo*[*]

[†]Signal Theory Group, Ruhr University, Bochum, D-44780, Germany
[*]Department of ECE, McMaster University, Hamilton, Ontario, L8S 4K1 Canada

ABSTRACT

We consider Friedlander's array interpolation technique, whose main shortcoming in multi-source scenarios is that it does not provide sufficient robustness against sources arriving outside specified interpolation sectors. In this paper, we propose a new interpolation approach which incorporates such robustness property. Our technique minimizes the interpolation error inside the sector of interest while setting multiple "stopband" constraints outside this sector to prevent performance degradation effects caused by out-of-sector sources. Based on such robust approach to array interpolation, we develop convex formulations of the interpolation matrix design problem using Second-Order Cone (SOC) programming.

1. INTRODUCTION

In array processing, specific array structures are frequently used to simplify implementations of subspace direction finding methods. For example, the Uniform Linear Array (ULA) structure is exploited to formulate computationally efficient search-free root-MUSIC and MODE algorithms [1], [2]. Uniform Circular Arrays (UCA's) and arrays with translational invariances also facilitate search-free formulations of subspace methods, such as conventional and multiple invariance ESPRIT [3], UCA root-MUSIC and UCA-ESPRIT [4], multiple invariance root-MUSIC [5], and RARE [6].

Unfortunately, all these methods are not straightforwardly applicable to arrays with an arbitrary geometry. In order to enable such application, Friedlander developed an elegant approach [7] based upon the idea of interpolating a virtual array of a required structure (e.g. ULA, UCA, etc.) using the original "non-structured" array. The interpolation is achieved by means of minimizing the error between the interpolated and desired array responses in some chosen interpolation sector. Although array interpolation approach has several attractive properties [8]-[9] and has been successfully applied to practical problems [10], an essential shortcoming of this method is that it does not provide sufficient robustness against sources which arrive outside specified interpolation sectors. In this paper, we propose a new interpolation approach which has such robustness property. Our technique minimizes the interpolation error inside the sector of interest under multiple "stopband" constraints outside this sector in order to prevent performance degradation effects caused by out-of-sector sources. Based on this robust array interpolation approach, we develop convex formulations of the interpolation matrix design problem using SOC programming.

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

2. THE CONVENTIONAL ARRAY INTERPOLATION APPROACH

The key idea of the conventional array interpolation approach is to interpolate virtual array observations inside a preliminary specified angular sector $\Theta = [\theta_{\min}, \theta_{\max}]$ using real array data. The $n \times n$ interpolation matrix $\boldsymbol{B}$ should obey

$$\boldsymbol{B}^H \boldsymbol{a}(\theta) \simeq \breve{\boldsymbol{a}}(\theta), \quad \theta \in \Theta$$

where $\boldsymbol{a}(\theta)$ and $\breve{\boldsymbol{a}}(\theta)$ are the $n \times 1$ steering vectors of the real and virtual arrays, respectively. Here, θ is the angle, n is the number of sensors, and $(\cdot)^H$ denotes the Hermitian transpose. Dividing the sector Θ into $M-1$ uniform subintervals of the width δ and defining the $n \times m$ matrices

$$\begin{aligned} \boldsymbol{C} &= [\boldsymbol{a}(\theta_{\min}), \boldsymbol{a}(\theta_{\min}+\delta), \ldots, \boldsymbol{a}(\theta_{\max}-\delta), \boldsymbol{a}(\theta_{\max})], \\ \breve{\boldsymbol{C}} &= [\breve{\boldsymbol{a}}(\theta_{\min}), \breve{\boldsymbol{a}}(\theta_{\min}+\delta), \ldots, \breve{\boldsymbol{a}}(\theta_{\max}-\delta), \breve{\boldsymbol{a}}(\theta_{\max})] \end{aligned}$$

the interpolation matrix $\boldsymbol{B}$ can be computed as a least squares solution to

$$\boldsymbol{B}^H \boldsymbol{C} = \breve{\boldsymbol{C}}$$

In its simplest form, this solution can be written as [7]

$$\boldsymbol{B} = (\boldsymbol{C}\boldsymbol{C}^H)^{-1}\boldsymbol{C}\breve{\boldsymbol{C}}^H \tag{1}$$

After noise prewhitening, the interpolated (virtual) array snapshots can be computed as

$$\boldsymbol{y}(i) = (\boldsymbol{B}^H\boldsymbol{B})^{-1/2}\boldsymbol{B}^H\boldsymbol{x}(i), \quad i = 1, \ldots, N \tag{2}$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

where $\boldsymbol{x}(i)$, $i = 1, \ldots, N$ are the real array observations, and N is the number of snapshots.

The application of any direction finding estimator to (2) and the generalization of this approach to the case of multiple angular interpolation sectors are straightforward [7].

The main shortcoming of this approach is that in multi-source scenarios, it does not provide sufficient robustness against sources impinging on the array outside Θ. The presence of such (sometimes quite powerful) sources may lead to a performance breakdown of the direction finding technique applied to interpolated array observations.

3. THE ROBUST ARRAY INTERPOLATION APPROACH

To incorporate robustness against out-of sector sources in the array interpolation approach, we reformulate the interpolation matrix design problem as the quadratic minimization problem with multiple inequality constraints

$$\min_{\boldsymbol{B}} \; ||\boldsymbol{B}^H \boldsymbol{C} - \check{\boldsymbol{C}}||$$

$$\text{subject to} \quad ||\boldsymbol{B}^H \boldsymbol{a}(\overline{\theta}_k)|| \leq \epsilon\,, \qquad (3)$$

$$\overline{\theta}_k \in \overline{\Theta}\,, \quad k = 1, 2, \ldots, K$$

where $\overline{\Theta}$ combines all directions lying outside the sector Θ, $\epsilon > 0$ is the parameter characterizing the "stopband" (out-of-sector) attenuation, and K is the number of inequality constraints.

Alternatively, another formulation can be used

$$\min_{\boldsymbol{B}} \max_{m} \; ||\boldsymbol{B}^H \boldsymbol{a}(\theta_m) - \check{\boldsymbol{a}}(\theta_m)||\,,$$

$$\theta_m \in \Theta\,, \quad m = 1, 2, \ldots M$$

$$\text{subject to} \quad ||\boldsymbol{B}^H \boldsymbol{a}(\overline{\theta}_k)|| \leq \epsilon\,, \qquad (4)$$

$$\overline{\theta}_k \in \overline{\Theta}\,, \quad k = 1, 2, \ldots, K$$

where the minimax criterion is employed.

4. CONVEX FORMULATIONS USING SECOND-ORDER CONE PROGRAMS

In this section, we present SOC formulations of the problems (3) and (4). Note that an efficient MATLAB toolbox is available that solves such problems in a computationally efficient way [11].

The general form of SOC program is given by

$$\max_{\boldsymbol{e}} \; \boldsymbol{d}^T \boldsymbol{e}$$

$$\text{subject to} \quad \boldsymbol{c}_q - \boldsymbol{F}_q^T \boldsymbol{e} \in \mathcal{SOC}^{C_q}\,, \quad q = 1, \ldots, p$$

Here, all vectors and matrices are real-valued, $\boldsymbol{e}$ is a vector containing the design variables, p is the number of SOC constraints and $C_q - 1$ is the dimension of the qth SOC defined as

$$\mathcal{SOC}^{C_q} = \left\{ (z_1, \bar{\boldsymbol{z}}) \in \mathbb{R} \times \mathbb{R}^{(C_q - 1)} |\; z_1 \geq \|\bar{\boldsymbol{z}}\| \right\}$$

where

$$\begin{aligned} \boldsymbol{z} &= \left[z_1, \bar{\boldsymbol{z}}^T \right]^T \\ &= \boldsymbol{c}_q - \boldsymbol{F}_q^T \boldsymbol{e}\,, \\ \bar{\boldsymbol{z}} &= \left[z_2, \ldots, z_{C_q} \right]^T \end{aligned}$$

Let us introduce the following notations

$$\begin{aligned} \boldsymbol{b} &= \text{vec}\left\{ \boldsymbol{B}^H \right\}\,, \\ \boldsymbol{c} &= \text{vec}\left\{ \boldsymbol{C} \right\}\,, \\ \check{\boldsymbol{c}} &= \text{vec}\left\{ \check{\boldsymbol{C}} \right\} \end{aligned}$$

Here, vec $\{\cdot\}$ denotes the vectorization operator, stacking the columns of a matrix on top of each other. The following property for arbitrary matrices $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ of conformable dimensions will be used frequently throughout the text

$$\text{vec}\left\{ \boldsymbol{X}\boldsymbol{Y}\boldsymbol{Z} \right\} = \left(\boldsymbol{Z}^T \otimes \boldsymbol{X} \right) \text{vec}\left\{ \boldsymbol{Y} \right\} \qquad (5)$$

where $\otimes$ denotes the Kronecker matrix product.

Note that

$$||\boldsymbol{X}|| = ||\text{vec}\{\boldsymbol{X}\}|| \qquad (6)$$

Making use of (5) we obtain that

$$\begin{aligned} \text{vec}\left\{ \boldsymbol{B}^H \boldsymbol{C} - \check{\boldsymbol{C}} \right\} &= \text{vec}\left\{ \boldsymbol{B}^H \boldsymbol{C} \right\} - \check{\boldsymbol{c}} \\ &= \text{vec}\left\{ \boldsymbol{I}\boldsymbol{B}^H \boldsymbol{C} \right\} - \check{\boldsymbol{c}} \\ &= \left(\boldsymbol{C}^T \otimes \boldsymbol{I} \right) \boldsymbol{b} - \check{\boldsymbol{c}} \qquad (7) \end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix. Similarly,

$$\begin{aligned} \boldsymbol{B}^H \boldsymbol{a}(\overline{\theta}_k) &= \text{vec}\left\{ \boldsymbol{I}\boldsymbol{B}^H \boldsymbol{a}(\overline{\theta}_k) \right\} \\ &= \left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I} \right) \boldsymbol{b} \qquad (8) \end{aligned}$$

4.1. Problem (3)

Using (6)-(8), we can reformulate (3) as

$$\min_{\boldsymbol{b}} \; || \left(\boldsymbol{C}^T \otimes \boldsymbol{I} \right) \boldsymbol{b} - \check{\boldsymbol{c}} \; ||$$

$$\text{subject to} \quad || \left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I} \right) \boldsymbol{b} \; || \leq \epsilon\,, \qquad (9)$$

$$\overline{\theta}_k \in \overline{\Theta}, \quad k = 1, 2, \ldots, K$$

Note that in (9), the vectors and matrices $\boldsymbol{b}$, $\boldsymbol{C}$, $\breve{\boldsymbol{c}}$ and $\boldsymbol{a}(\overline{\theta}_k)$ are complex valued. Apparently, the problem (9) can be replaced by the following equivalent optimization problem

$$\begin{aligned} &\max\ \tau \\ &\text{subject to}\ \ \| (\boldsymbol{C}^T \otimes \boldsymbol{I})\boldsymbol{b} - \breve{\boldsymbol{c}} \| \leq -\tau, \\ &\qquad\qquad \| (\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I})\boldsymbol{b} \| \leq \epsilon, \qquad (10) \\ &\overline{\theta}_k \in \overline{\Theta}, \quad k = 1, 2, \ldots, K \end{aligned}$$

The problem (10) can be reformulated in terms of real-valued variables. Let $(\cdot)_r$ and $(\cdot)_i$ hereafter denote the real and the imaginary parts of a matrix, respectively. Then (10) becomes

$$\begin{aligned} &\max\ \tau \\ &\text{subject to}\ \ \left\| \begin{bmatrix} ((\boldsymbol{C}^T \otimes \boldsymbol{I})\boldsymbol{b})_r \\ ((\boldsymbol{C}^T \otimes \boldsymbol{I})\boldsymbol{b})_i \end{bmatrix} - \begin{bmatrix} (\breve{\boldsymbol{c}})_r \\ (\breve{\boldsymbol{c}})_i \end{bmatrix} \right\| \leq -\tau, \\ &\qquad\qquad \left\| \begin{bmatrix} \left(\left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)\boldsymbol{b}\right)_r \\ \left(\left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)\boldsymbol{b}\right)_i \end{bmatrix} \right\| \leq \epsilon, \\ &\overline{\theta}_k \in \overline{\Theta}, \quad k = 1, 2, \ldots, K \end{aligned}$$

or, equivalently, maximize τ subject to

$$\begin{aligned} &\left\| \begin{bmatrix} (\boldsymbol{C}^T \otimes \boldsymbol{I})_r & -(\boldsymbol{C}^T \otimes \boldsymbol{I})_i \\ (\boldsymbol{C}^T \otimes \boldsymbol{I})_i & (\boldsymbol{C}^T \otimes \boldsymbol{I})_r \end{bmatrix} \begin{bmatrix} (\boldsymbol{b})_r \\ (\boldsymbol{b})_i \end{bmatrix} - \begin{bmatrix} (\breve{\boldsymbol{c}})_r \\ (\breve{\boldsymbol{c}})_i \end{bmatrix} \right\| \leq -\tau, \\ &\left\| \begin{bmatrix} \left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)_r & -\left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)_i \\ \left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)_i & \left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)_r \end{bmatrix} \begin{bmatrix} (\boldsymbol{b})_r \\ (\boldsymbol{b})_i \end{bmatrix} \right\| \leq \epsilon, \\ &\overline{\theta}_k \in \overline{\Theta}, \quad k = 1, 2, \ldots, K \end{aligned}$$

Defining the $(2n^2 + 1) \times 1$ vectors

$$\boldsymbol{d} = [1, 0, \ldots, 0]^T, \qquad (11)$$

$$\boldsymbol{e} = \left[\tau, (\boldsymbol{b}^T)_r, (\boldsymbol{b}^T)_i\right]^T \qquad (12)$$

and the matrices

$$\boldsymbol{M} = \begin{bmatrix} 1 & \boldsymbol{0}^T & \boldsymbol{0}^T \\ \boldsymbol{0} & \left(\boldsymbol{C}^T \otimes \boldsymbol{I}\right)_r & -\left(\boldsymbol{C}^T \otimes \boldsymbol{I}\right)_i \\ \boldsymbol{0} & \left(\boldsymbol{C}^T \otimes \boldsymbol{I}\right)_i & \left(\boldsymbol{C}^T \otimes \boldsymbol{I}\right)_r \end{bmatrix}, \qquad (13)$$

$$\boldsymbol{L}_k = \begin{bmatrix} 0 & \boldsymbol{0}^T & \boldsymbol{0}^T \\ \boldsymbol{0} & \left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)_r & -\left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)_i \\ \boldsymbol{0} & \left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)_i & \left(\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I}\right)_r \end{bmatrix} \qquad (14)$$

we can rewrite (3) as

$$\begin{aligned} &\max_{\boldsymbol{e}}\ \boldsymbol{d}^T\boldsymbol{e} \\ &\text{subject to}\ \ \begin{bmatrix} 0 \\ \breve{\boldsymbol{c}} \end{bmatrix} - \boldsymbol{M}\boldsymbol{e} \in \mathcal{SOC}^{2n^2+1}, \\ &\qquad\qquad \begin{bmatrix} \epsilon \\ \boldsymbol{0} \end{bmatrix} - \boldsymbol{L}_k\boldsymbol{e} \in \mathcal{SOC}^{2n+1}, \\ &\overline{\theta}_k \in \overline{\Theta}, \quad k = 1, 2, \ldots, K \end{aligned}$$

which is the SOC formulation of the optimization problem (3).

4.2. Problem (4)

Using (5), rewrite the optimization problem (4) as

$$\begin{aligned} &\max\ \tau \\ &\text{subject to}\ \ \| (\boldsymbol{a}^T(\theta_m) \otimes \boldsymbol{I})\boldsymbol{b} - \breve{\boldsymbol{a}}(\theta_m) \| \leq -\tau, \quad (15) \\ &\qquad\qquad \| (\boldsymbol{a}^T(\overline{\theta}_k) \otimes \boldsymbol{I})\boldsymbol{b} \| \leq \epsilon, \\ &\theta_m \in \Theta, \quad m = 1, 2, \ldots, M, \\ &\overline{\theta}_k \in \overline{\Theta}, \quad k = 1, 2, \ldots, K \end{aligned}$$

Defining the matrix

$$\boldsymbol{M}_m = \begin{bmatrix} 1 & \boldsymbol{0}^T & \boldsymbol{0}^T \\ \boldsymbol{0} & \left(\boldsymbol{a}^T(\theta_m) \otimes \boldsymbol{I}\right)_r & -\left(\boldsymbol{a}^T(\theta_m) \otimes \boldsymbol{I}\right)_i \\ \boldsymbol{0} & \left(\boldsymbol{a}^T(\theta_m) \otimes \boldsymbol{I}\right)_i & \left(\boldsymbol{a}^T(\theta_m) \otimes \boldsymbol{I}\right)_r \end{bmatrix}$$

and using (11), (12) and (14), we can reformulate (15) as the following SOC program

$$\begin{aligned} &\max_{\boldsymbol{e}}\ \boldsymbol{d}^T\boldsymbol{e} \\ &\text{subject to}\ \ \begin{bmatrix} 0 \\ \breve{\boldsymbol{a}}(\theta_m) \end{bmatrix} - \boldsymbol{M}_m\boldsymbol{e} \in \mathcal{SOC}^{2n+1}, \\ &\qquad\qquad \begin{bmatrix} \epsilon \\ \boldsymbol{0} \end{bmatrix} - \boldsymbol{L}_k\boldsymbol{e} \in \mathcal{SOC}^{2n+1}, \\ &\theta_m \in \Theta, \quad m = 1, 2, \ldots, M, \\ &\overline{\theta}_k \in \overline{\Theta}, \quad k = 1, 2, \ldots, K \end{aligned}$$

5. SIMULATIONS

In our simulations, the conventional and robust interpolated array approaches are compared in terms of DOA estimation Root-Mean-Square Errors (RMSE's) in the presence of an interfering out-of-sector source. We assume a linear array of ten sensors, $N = 100$, and two uncorrelated sources. One of the sources with the SNR $= 0$ dB is assumed to be the signal of interest whose DOA belongs to the interpolation sector $\Theta = [-15°, 15°]$ and the second one is the interfering (out-of-sector) source. In each simulation run, the DOA's of the signal and interfering sources are drawn uniformly from the intervals $[-15°, 15°]$ and $[-90°, -25°] \cup [25°, 90°]$, respectively. Furthermore, the sensor coordinates of the real array are drawn uniformly in each run from the interval $[0, 4.5\lambda]$, where λ is the wavelength and the coordinates of the leftmost and rightmost array sensors are fixed and

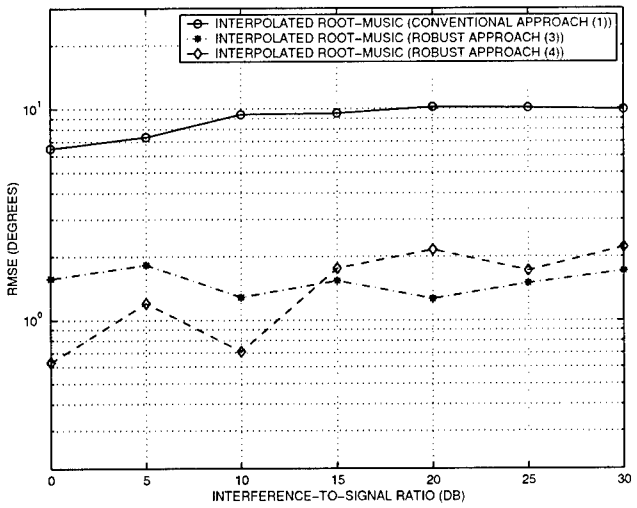


Fig. 1. DOA estimation RMSE's versus the ISR.

equal to 0 and 4.5λ, respectively. A virtual ULA with the half-wavelength spacing is interpolated using the conventional interpolation technique (1) and the robust approaches (3)-(4), respectively. In all these methods, the parameters $M = K = 12$ are chosen and a nonuniform grid is used for $\overline{\Theta}$ based on the output of the conventional beamformer.

The SeDuMi toolbox has been used to solve the corresponding SOC problems. Diagonal loading is used in the prewhitening step in order to guarantee stable inverse of the matrix $\boldsymbol{B}^H\boldsymbol{B}$ in (2). The interpolated root-MUSIC [7] is used to estimate the signal DOA. In total, 100 independent simulation runs are performed to estimate the RMSE's which are displayed in Fig. 1 versus the Interference-to-Signal Ratio (ISR). This figure validates essential performance improvements provided by the proposed robust approach.

6. CONCLUSIONS

A new robust approach to array interpolation has been proposed. Our technique minimizes the interpolation error inside the sectors of interest while setting multiple "stopband" constraints outside these sectors to prevent performance degradation effects caused by out-of-sector sources. Convex formulations of the interpolation matrix design problem have been proposed using second-order cone programming. Simulation results validate robustness of the proposed technique and demonstrate essential performance improvements of our approach relative to the conventional array interpolation method.

7. REFERENCES

[1] A.J. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Proc. ICASSP'83*, Boston, MA, pp. 336-339, May 1983.

[2] P. Stoica and K.C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1132-1143, July 1990.

[3] A.L. Swindlehurst, B. Ottersten, R. Roy, and T. Kailath, "Multiple invariance ESPRIT," *IEEE Trans. Signal Processing*, vol. 40, pp. 867-881, Apr. 1992.

[4] C.P. Mathews and M.D. Zoltowski, "Eigenstructure techniques for 2-D angle estimation with uniform circular arrays," *IEEE Trans. Signal Processing*, vol. 42, pp. 2395-2407, Sept. 1994.

[5] A.L. Swindlehurst, P. Stoica, and M. Jansson, "Application of MUSIC to arrays with multiple invariances," in *Proc. ICASSP'00*, Istanbul, Turkey, pp. 3057-3060, June 2000.

[6] M. Pesavento, A.B. Gershman, and K.M. Wong, "Direction of arrival estimation in partly calibrated time-varying sensor arrays," *Proc. ICASSP'01*, Salt Lake City, UT, May 2001.

[7] B. Friedlander, "The root-MUSIC algorithm for direction finding with interpolated arrays," *Signal Processing*, vol. 30, pp. 15-25, 1993.

[8] A.B. Gershman and J.F. Böhme, "A note on most favorable array geometries for DOA estimation and array interpolation," *IEEE Signal Processing Letters*, vol. 4, pp. 232-235, Aug. 1997.

[9] J. Eriksson and M. Viberg, "Data reduction in spatially colored noise using a virtual uniform linear array," in *Proc. ICASSP'00*, Istanbul, Turkey, pp. 3073-3076, June 2000.

[10] D.V. Sidorovitch and A.B. Gershman, "2-D wideband interpolated root-MUSIC applied to measured seismic data," *IEEE Trans. Signal Processing*, vol. 46, pp. 2263-2267, Aug. 1998.

[11] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones" *Optim. Meth. Software*, vol. 11-12, pp. 625-653, Aug. 1999, also see `http://www.unimaas.nl/~sturm/software/sedumi.html`.

PERFORMANCE ANALYSIS OF MIS-MODELED ESTIMATION PROCEDURES FOR A DISTRIBUTED SOURCE OF NON CONSTANT MODULUS

Jonathan Friedmann, Raviv Raich, Jason Goldberg and Hagit Messer

Department of Electrical Engineering–Systems, Tel Aviv University,
Tel Aviv 69978, Israel, Tel: +972 3 640 8275, Fax: + 972 3 640 7095,
EMAIL: jonaf@eng.tau.ac.il, raviv@ece.gatech.edu, {jason, messer}@eng.tau.ac.il.

ABSTRACT

The problem of estimating the bearing of a single, far-field, non constant modulus source, surrounded by local scatterers using passive sensor array measurements is addressed. An associated source bearing estimation problem is formulated, and the Cramér-Rao lower bound is evaluated. Estimation procedures that assume the source is of constant modulus and therefore suffer from mis-modeling errors are investigated. Specifically, the performance of the application of the maximum likelihood (ML) estimator designed for a constant modulus source performing on a non constant modulus source is evaluated, and the exact degradation in performance is quantified as a function of the source's empirical variance.

1. INTRODUCTION

Recently, bearing estimation for a so-called *distributed* or *scattered source* has begun to attract interest in the literature. A distributed source may arise due to the multipath scattering effects created by the presence of local scatterers about the emitter. The spatial extent of a distributed source is typically characterized by some type of parametric model. These models have formed the basis of a variety of recently reported bearing estimation techniques and performance studies, e.g., [1], [6].

One of the key assumptions found in much of the previously published work on distributed source bearing estimation is that the transmitted signal is deterministic, unknown and of *constant modulus* (CM). This assumption is particularly useful for the commonly cited complex normal (Rayleigh amplitude), temporally uncorrelated vector channel formed between the source and the receiving array. However, despite its convenience, the CM signal assumption may be inappropriate in many applications. The present work examines the performance of estimation procedures designed to perform under the CM assumption when such an assumption is untrue.

1.1. Problem Formulation

Assuming that the sampling interval is significantly greater than the channel coherence time, the received array measurement data for K snapshots may be described as a sequence of zero mean uncorrelated complex Gaussian random vectors with *time varying* covariance [2]:

$$\begin{aligned} &\mathbf{y}_t \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{y_t}) \\ &\mathbf{R}_{y_t} = P_t \mathbf{R}_b(\boldsymbol{\zeta}) + \sigma^2 \mathbf{I} \qquad t = 1, \ldots, K \end{aligned} \quad (1)$$

where $\mathbf{R}_b$ is the channel covariance matrix depending on the unknown spatial parameters, $\boldsymbol{\zeta}$ (typically including mean angle parameter, θ_0, corresponding to the source bearing). σ^2, is the noise variance, and $\mathbf{p} = [P_1, \ldots, P_K]^T$ is the vector of instantaneous source powers. If the source is assumed to be CM, then

$$\mathbf{y}_t \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_y) \qquad \mathbf{R}_y = P\mathbf{R}_b(\boldsymbol{\zeta}) + \sigma^2 \mathbf{I}. \quad (2)$$

The overall unknown parameter vector reduces to: $\boldsymbol{\psi} = \left[\boldsymbol{\zeta}^T, P, \sigma^2\right]^T$. Models of this latter type are very common in the literature, e.g., [1], [6].

In the sequel we study estimation procedures designed under the assumption that the source is CM when in fact the source is non constant modulus (NCM). Such model mis-match may arise due to modeling error, or alternatively, even if it is known that the source is NCM, it may be tempting to use a CM based algorithm such that the number of parameters to estimate is small.

The paper is organized as follows: Section 2 examines the limitations of the estimation problem for the general model of an NCM source (1). Then, Section 3 analyzes the performance of CM based algorithms performing on NCM sources. Lastly, Section 4 gives some simulation results and, Section 5 summarizes the paper.

0-7803-7011-2/01/$10.00 ©2001 IEEE

2. THE CRAMÉR RAO BOUND

2.1. General model

The Cramér Rao bound, $\mathbf{C}$, for the general model in (1) with parameter vector $\psi = \left[\zeta^T, \sigma^2, \mathbf{p}^T\right]^T$ may be written as:

$$\mathbf{C} = \mathbf{J}^{-1}; \qquad \mathbf{J} = \sum_{t=1}^{K} \mathbf{J}^{(t)};$$
$$[\mathbf{J}]_{ij}^{(t)} = \mathrm{Tr}\left(\mathbf{R}_{y_t}^{-1} \frac{\partial \mathbf{R}_{y_t}}{\partial \psi_i} \mathbf{R}_{y_t}^{-1} \frac{\partial \mathbf{R}_{y_t}}{\partial \psi_j}\right). \qquad (3)$$

where $\mathbf{J}$ is known as the Fisher Information Matrix (FIM), and $\mathrm{Tr}(\cdot)$ and $[\cdot]_{ij}$ denote the trace operator and the ijth element of a matrix, respectively. The FIM may be written in block form:

$$\begin{aligned} \mathbf{J} &= \begin{bmatrix} \mathbf{J}_{\eta\eta} & \mathbf{J}_{\eta\mathbf{p}} \\ \mathbf{J}_{\mathbf{p}\eta} & \mathbf{J}_{\mathbf{pp}} \end{bmatrix} \qquad \eta = [\zeta^T \sigma^2]^T \qquad (4) \\ \mathbf{J}_{\eta\mathbf{p}} &= [\mathbf{j}_{\eta P_1}, \cdots, \mathbf{j}_{\eta P_K}] \\ \mathbf{J}_{\mathbf{pp}} &= \mathrm{diag}\left[J_{P_1P_1}, \cdots, J_{P_KP_K}\right] \end{aligned}$$

where the notation $\mathrm{diag}[\cdot]$ denotes a diagonal matrix of specified diagonal elements. Note that $\mathbf{J}_{\mathbf{pp}}$ is a diagonal matrix since the array measurements are assumed to be statistically independent. This allows the CRB for the spatial parameters to be expressed as a matrix whose dimension does not increase with the number of measurements:

$$\mathbf{C} = \mathbf{J}^{-1} = \begin{bmatrix} \mathbf{C}_{\eta\eta} & \mathbf{C}_{\eta\mathbf{p}} \\ \mathbf{C}_{\mathbf{p}\eta} & \mathbf{C}_{\mathbf{pp}} \end{bmatrix}$$
$$\mathbf{C}_{\eta\eta} = \left(\mathbf{J}_{\eta\eta} - \sum_{t=1}^{K} \frac{1}{J_{P_tP_t}} \mathbf{j}_{\eta P_t} \mathbf{j}_{\eta P_t}^T\right)^{-1}. \qquad (5)$$

For the case of CM signals, the sum in (5) simplifies to: $\sum_{t=1}^{K} \frac{1}{J_{P_tP_t}} \mathbf{j}_{\eta P_t} \mathbf{j}_{\eta P_t}^T \rightarrow \frac{K}{J_{PP}} \mathbf{j}_{\eta P} \mathbf{j}_{\eta P}^T$.

It should be stressed that since the number of unknown parameters increases with the number of samples, the CRB is not necessarily attained by any estimator and may serve only as a lower bound (see e.g., [3]). In this sense, this problem is similar to the well know deterministic unknown point source bearing estimation problem [5].

2.2. Infinite SNR

Some simplification is possible even in a general model for infinite SNR, i.e., when there exists no additive noise. It is seen that:

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_{\zeta\zeta} & \mathbf{J}_{\zeta\mathbf{p}} \\ \mathbf{J}_{\mathbf{p}\zeta} & \mathbf{J}_{\mathbf{pp}} \end{bmatrix}; \qquad \mathbf{J}_{\zeta\zeta} = K\bar{\mathbf{J}}_{\zeta\zeta};$$
$$[\bar{\mathbf{J}}_{\zeta\zeta}]_{i,j} = \mathrm{Tr}\left(\mathbf{R}_b^{-1} \frac{\partial \mathbf{R}_b}{\partial \zeta_i} \mathbf{R}_b^{-1} \frac{\partial \mathbf{R}_b}{\partial \zeta_j}\right); \qquad (6)$$

$$\mathbf{j}_{\zeta\mathbf{p}} = [\mathbf{j}_{\zeta P_1}, \ldots, \mathbf{j}_{\zeta P_K}]^T; \qquad J_{P_tP_t} = \frac{M}{P_t^2};$$
$$\mathbf{j}_{\zeta, P_t} = \frac{1}{P_t}[\bar{j}_\zeta]_i = \mathrm{Tr}\left(\mathbf{R}_b^{-1} \frac{\partial \mathbf{R}_b}{\partial \zeta_i}\right). \qquad (7)$$

The CRB can now be written as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{\zeta\zeta} & \mathbf{C}_{\zeta\mathbf{p}} \\ \mathbf{C}_{\mathbf{p}\zeta} & \mathbf{C}_{\mathbf{pp}} \end{bmatrix}$$
$$\mathbf{C}_{\zeta\zeta} = \frac{1}{K}\left(\bar{\mathbf{J}}_{\zeta\zeta} - \frac{1}{M}\bar{j}_\zeta \bar{j}_\zeta^T\right)^{-1} \qquad (8)$$

where it is noted that the CRB for the spatial parameters in the noiseless case is independent of the sequence $\{P_t\}_{t=1}^K$ which would be expected intuitively. Furthermore, the CM bound at infinite SNR is of exactly the same form as (8). This property is intuitively expected since ζ is a vector of spatial parameters, and their estimation is not affected by measurement scaling parameters when there exists no additive noise. This means that for infinite SNR, the CRB does not depend on the sequence $\{P_t\}_{t=1}^K$.

3. CONSTANT MODULUS ALGORITHMS

Consider array measurement data arising from an NCM source as described by (1). We investigate the performance of algorithms designed to estimate the bearing of a CM source when applied to NCM data. Such a scenario may arise if, for example, the source is believed to be CM when in reality it is NCM. Alternatively, even if it is known that the source is NCM, it may be tempting to use a CM based algorithm since, under the CM model, the number of nuisance parameters does not increase with the number of measurements.

3.1. Consistency

Assume that the empirical first and second order moment of the instantaneous powers are finite:

$$A1) \qquad 0 < \overline{P} = \lim_{K\to\infty} \frac{1}{K} \sum_{t=1}^{K} P_t < \infty$$
$$A2) \qquad 0 < \overline{P^2} = \lim_{K\to\infty} \frac{1}{K} \sum_{t=1}^{K} P_t^2 < \infty.$$

Clearly this means that the empirical variance is also finite:

$$0 \le \overline{\Delta P^2} = \lim_{K\to\infty} \frac{1}{K} \sum_{t=1}^{K} \left(P_t - \overline{P}\right)^2 = \overline{P^2} - \overline{P}^2 < \infty.$$

Consider the class of covariance matching methods for bearing estimation. Examples include weighted least squares covariance matching and maximum likelihood (ML) estimation. Under mild conditions if the sample covariance

matrix $\widehat{\mathbf{R}}_y = \frac{1}{K}\sum_{t=1}^{K}\mathbf{y}_t\mathbf{y}_t^H$ is a consistent estimate of $\mathbf{R}_y = \overline{P}\mathbf{R}_b + \sigma^2\mathbf{I}$, then the resulting $\hat{\zeta}$ is a consistent estimate of the spatial parameters, ζ [4]. In order to show the consistency of the sample covariance matrix, it suffices to show that for all row and column indices i and j:

$$\lim_{K\to\infty} \mathrm{E}\left(\left|\frac{1}{K}\sum_{t=1}^{K}\mathbf{y}_t^i(\mathbf{y}_t^j)^* - [\mathbf{R}_y]_{ij}\right|^2\right) = 0 \qquad (9)$$

where $\mathbf{y}_t^i$ is the ith element of the vector $\mathbf{y}_t$ at an instant t and $(\cdot)^*$ denotes the complex conjugate.

Under assumptions A1 and A2, It is easily shown that:

$$\begin{aligned}
&\lim_{K\to\infty} \mathrm{E}\left(\left|\frac{1}{K}\sum_{t=1}^{K}\mathbf{y}_t^i(\mathbf{y}_t^j)^* - [\mathbf{R}_y]_{ij}\right|^2\right) = \\
&= \lim_{K\to\infty}\frac{1}{K^2}\sum_{t=1}^{K}[\mathbf{R}_{y_t}]_{ii}[\mathbf{R}_{y_t}]_{jj} \\
&= \lim_{K\to\infty}\frac{1}{K}[\mathbf{R}_y]_{ii}[\mathbf{R}_y]_{jj} + \frac{1}{K}\overline{\Delta P^2}[\mathbf{R}_b]_{ii}[\mathbf{R}_b]_{jj} \\
&= 0. \qquad (10)
\end{aligned}$$

Thus, $\widehat{\mathbf{R}}_y$ is a mean square error (MSE) consistent estimate of $\mathbf{R}_y$. Note that $\frac{1}{K}[\mathbf{R}_y]_{ii}[\mathbf{R}_y]_{jj}$ depends only on $\overline{P}$ and not on $\mathbf{p}$. The fact that the source is CM or NCM does not affect this term. The term $\overline{\Delta P^2}[\mathbf{R}_b]_{ii}[\mathbf{R}_b]_{jj}$ is proportional to the empirical variance of the source. It is equal to zero when the source is CM and increases as the source's empirical variance increases. Hence, as $\overline{\Delta P^2}$ increases, a larger observation time, K, is needed for the empirical correlation matrix to get "closer" to $\mathbf{R}_y$.

3.2. Small Error Performance Analysis

Attention is focused on the CM ML estimator due to its optimality under the CM model (2). This estimator assumes $P_t = P \quad \forall t$, i.e., it assumes the unknown parameters are: $\psi = [\zeta^T, \sigma^2, P]$. The ML estimate for ψ is then given as the solution of the following estimating equations:

$$-\mathrm{Tr}\left(\mathbf{R}_y^{-1}\frac{\partial\mathbf{R}_y}{\partial\psi_i}\right) + \mathrm{Tr}\left(\mathbf{R}_y^{-1}\frac{\partial\mathbf{R}_y}{\partial\psi_i}\mathbf{R}_y^{-1}\widehat{\mathbf{R}}_y\right) = 0 \qquad \forall i. \qquad (11)$$

The asymptotic behavior of the estimates can be determined by a first order expansion of the estimating equations (11) about the true parameter vector. Once again it is stressed that such an analysis yields the behavior of CM ML estimates when the data is NCM. The first order expansion is detailed in Appendix A in [2] and yields the following approximation for the estimation error:

$$\varepsilon = \psi - \hat{\psi} \approx \mathbf{Q}^{-1}\mathbf{v}$$

$$\begin{aligned}
\mathbf{Q}_{ij} &= \mathrm{Tr}\left(-\frac{\partial\mathbf{R}_y^{-1}}{\partial\psi_j}\frac{\partial\mathbf{R}_y}{\partial\psi_i}\right) = \mathrm{Tr}\left(\mathbf{R}_y^{-1}\frac{\partial\mathbf{R}_y}{\partial\psi_j}\mathbf{R}_y^{-1}\frac{\partial\mathbf{R}_y}{\partial\psi_i}\right) \\
\mathbf{v}_i &= \mathrm{Tr}\left(-\frac{\partial\mathbf{R}_y^{-1}}{\partial\psi_i}\left(\mathbf{R}_y - \widehat{\mathbf{R}}_y\right)\right) \\
&= \mathrm{Tr}\left(\mathbf{R}_y^{-1}\frac{\partial\mathbf{R}_y}{\partial\psi_i}\mathbf{R}_y^{-1}\left(\mathbf{R}_y - \widehat{\mathbf{R}}_y\right)\right). \qquad (12)
\end{aligned}$$

As expected, (12) shows the estimate is asymptotically unbiased. Also observe that matrix $\mathbf{Q}$ is simply the CM FIM. The asymptotic covariance of the estimates is given by (full evaluation in Appendix B in [2]):

$$\begin{aligned}
\widetilde{\mathbf{C}}_{\varepsilon\varepsilon} &= \mathrm{E}\left(\varepsilon\varepsilon^T\right) \approx \mathbf{Q}^{-1}\mathrm{E}\left(\mathbf{v}\mathbf{v}^T\right)\mathbf{Q}^{-1} \\
&= \frac{1}{K}\bar{\mathbf{C}} + \frac{1}{K}\overline{\Delta P^2}\bar{\mathbf{C}}\mathbf{K}\bar{\mathbf{C}}. \qquad (13) \\
\mathbf{K}_{ij} &= \mathrm{Tr}\left(\mathbf{R}_b\frac{\partial\mathbf{R}_y^{-1}}{\partial\psi_i}\mathbf{R}_b\frac{\partial\mathbf{R}_y^{-1}}{\partial\psi_j}\right) \qquad (14)
\end{aligned}$$

where $\bar{\mathbf{C}} = \mathbf{Q}^{-1}$ is the single snapshot CM CRB. The asymptotic performance consists of two terms. The first is the CM CRB. The second (which represents the degradation with respect to the CM CRB) is linear in the empirical variance of the instantaneous source powers $\{P_t\}_{t=1}^K$, $\overline{\Delta P^2}$. Note that the dependence of $\bar{\mathbf{C}}\mathbf{K}\bar{\mathbf{C}}$ on $\mathbf{p}$ manifests itself exclusively in terms of the mean power $\overline{P}$. Hence, for a given average source power, the performance of the ML estimate deteriorates linearly with the empirical variance of the instantaneous source power. In other words, performance deteriorates as P_t becomes "less constant".

3.2.1. Infinite SNR

For infinite SNR, it is seen that $\mathbf{K} = \frac{1}{\overline{P}^2}\mathbf{Q}$ such that the spatial parameters estimate's MSE can be expressed as:

$$\widetilde{\mathbf{C}}_{\varepsilon\varepsilon} \approx \frac{1}{K}\bar{\mathbf{C}}\left(1 + \frac{\overline{\Delta P^2}}{\overline{P}^2}\right). \qquad (15)$$

Thus the deterioration in performance with respect to the CRB is proportional to the empirical variance normalized by the empirical mean squared. Indeed, the empirical distribution of some sources may cause serious degradation in performance, especially those distributions with slowly decaying tails. Consider, for instance, a sequence $\{P_t\}_{t=1}^K$ which behaves as if it were a realization of independent identically distributed (IID) random variables governed by the Pareto distribution. For this distribution, when the variance is finite, $\frac{E(x^2)}{[E(x)]^2} = \frac{1}{a(a-2)}$ where $a > 2$. Clearly this term becomes arbitrarily large as a approaches two.

4. SIMULATIONS

To compare the derived analytical results to empirical ones we follow the uniform linear array model proposed in [1]. Simulations of 1000 Monte Carlo runs were carried each consisting of 100 snapshots. Figures 1 and 2 depict simulation results of the specific channel model given in [1] in two different scenarios. The first scenario assumes there exists no additive noise while the second does not make such an assumption. In both cases, half of the instantaneous source powers, $\{P_l\}_{l=1}^{K}$ are taken to be equal to five and half to one, i.e., $[P_1, \cdots, P_K]^T = [1, 5, \ldots, 1, 5]^T$. The spatial parameters, ζ, include the bearing which is set to $\theta_0 = \arcsin\left(\frac{1}{\pi}\right)$ (or by the alternative parameterization in [1], $\omega = 1$).

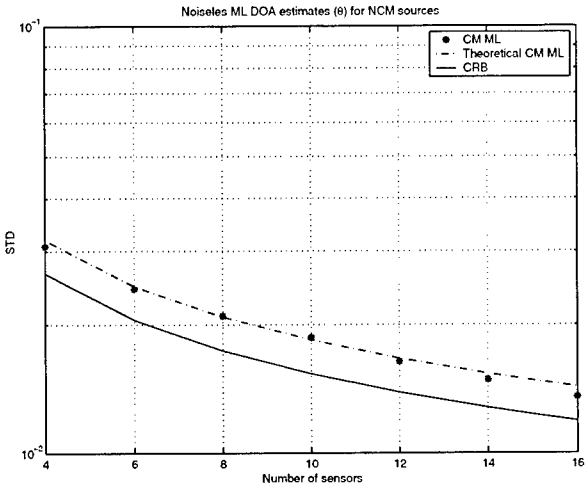


Figure 1: CM ML DOA estimates (θ) versus the number of sensors for NCM source with no additive noise $[P_1, \cdots, P_K]^T = [1, 5, \ldots, 1, 5]^T$, $\theta_0 = \arcsin\left(\frac{1}{\pi}\right)$.

Figure 1 depicts the performance of the CM ML for the bearing, θ_0, versus the number of sensors for the noiseless scenario. As a reference the NCM CRB is also shown. Figure 2 depicts the scenario that includes the additive noise. The figure shows the performance of the CM ML estimator for the bearing θ_0 versus the SNR. In these simulations, the number of sensors was taken to be 4. Once again, the CRB is shown for reference. For both figures, it is seen that theoretical results fit well the empirical results and that the degradation in performance compared to the CRB is relatively small.

5. SUMMARY

This paper examines the performance CM based bearing estimation algorithms for an NCM source. First, a general parametric description of an NCM source, is introduced and the CRB is evaluated. Then, the consistency of covariance matching methods which assume the source is of CM is proved. Focus is then set on CM based ML estimates and their performance is evaluated and shown to degrade linearily with the source's empirical variance Finally, theoretical results are shown to fit empirical results via simulations.

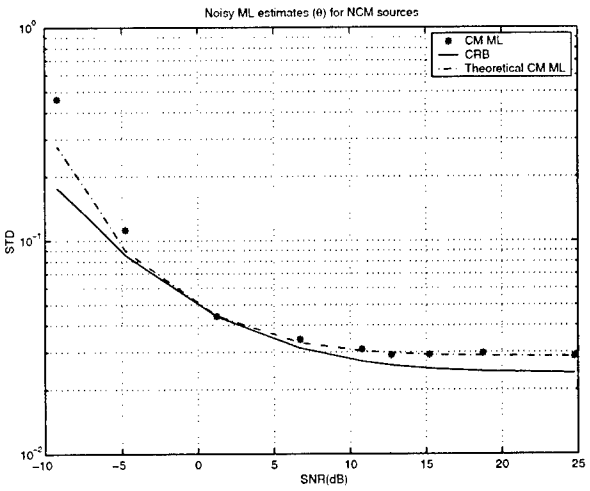


Figure 2: CM ML DOA estimates (θ) versus the SNR for NCM source, $\mathbf{p} = [1, 5, \ldots, 1, 5]$, $M = 4$, $\theta_0 = \arcsin\left(\frac{1}{\pi}\right)$.

6. REFERENCES

[1] O. Besson, F. Vincent, P. Stoica, and A.B. Gershman. Approximate maximum likelihood estimators for array processing in multiplicative noise environments. *IEEE Transactions on Signal Processing*, 48(9):2506–2518, September 2000.

[2] J. Friedmann, R. Raich, J. Goldberg, and H. Messer. Bearing estimation for a distributed source of non constant modulus. *submitted to IEEE Transactions on Signal Processing*, April 2001.

[3] J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, January 1948.

[4] A. Satora and P.M. Bentler. Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis*, 10:235–249, 1990.

[5] P. Stoica and A. Nehorai. Music, maximum likelihood, and Cramér rao bound. *IEEE Transactions on Signal Processing*, 37(5):720–741, May 1989.

[6] T. Trump and B. Ottersten. Estimation of nominal direction of arrival and angular spread using an array of sensors. *Signal Processing*, 50:57–69, April 1996.

A NEW ALGORITHM FOR COMPUTING THE EXTREME EIGENVECTORS OF A COMPLEX HERMITIAN MATRIX

Jonathan H. Manton

ARC Special Research Centre for Ultra-Broadband Information Networks
Department of Electrical and Electronic Engineering
The University of Melbourne, Parkville, Victoria 3010, Australia.
j.manton@ee.mu.oz.au

ABSTRACT

This paper presents a novel algorithm for computing the eigenvector associated with either the largest or the smallest eigenvalue of a complex Hermitian matrix. Necessary and sufficient conditions for convergence are proved, and simulations show the superior performance over traditional methods.

1. INTRODUCTION

Classical Direction of Arrival (DOA) and frequency estimation algorithms [10, 1] require the computation and subsequent tracking of the eigenvector associated with the smallest eigenvalue of a Hermitian matrix, henceforth referred to as a minimal eigenvector. Based on the novel optimisation algorithm developed in [6, 7] for optimising a cost function on the complex Grassmann manifold, this paper derives a new algorithm for computing a minimal eigenvector of a matrix. It is proved that the algorithm converges to a minimal eigenvector provided the initial vector is not orthogonal to the eigenspace spanned by the minimal eigenvectors.

The algorithm differs from traditional ones, such as the Power and inverse iteration methods [3], in two important ways. Firstly, whereas traditional methods can fail to converge in a reasonable number of iterations if two or more eigenvalues are closely spaced, the proposed algorithm continues to converge rapidly in such situations. The second difference is that, unlike traditional methods which converge to the eigenvector associated with the eigenvalue having the smallest or the largest *absolute* value, the proposed algorithm converges to the eigenvector associated with the eigenvalue having the smallest or the largest value. (Recall that a Hermitian matrix has real-valued eigenvalues.)

Notation: The superscripts T and H denote transpose and Hermitian transpose respectively. Throughout, the Frobenius norm $\|X\|^2 = \operatorname{tr}\{X^H X\}$ is used, where $\operatorname{tr}\{\cdot\}$ is the trace operator. The symbol I denotes the identity matrix whose size can be determined from its context.

2. COMPUTING AN EXTREME EIGENVECTOR

It is well known [3] that, for a symmetric matrix $A \in \mathbb{C}^{n\times n}$,

$$f(\boldsymbol{x}) = \frac{1}{2}\operatorname{tr}\left\{\boldsymbol{x}^H A \boldsymbol{x}\right\} \tag{1}$$

achieves its minimum, subject to $\boldsymbol{x}^H\boldsymbol{x} = 1$, when $\boldsymbol{x} \in \mathbb{C}^n$ corresponds to a minimal eigenvector, that is, an eigenvector associated with the smallest eigenvalue of A. This section specialises the steepest descent on the complex Grassmann manifold algorithm derived in [6, 7] to this particular cost function. (Note that by replacing A with $-A$, the same algorithm can be used to find an eigenvector associated with the largest eigenvalue of A.)

An attractive feature of this specialisation is that the optimal step size can be calculated at each iteration. Although steepest descent algorithms were derived in [2, 4] for computing a minimal eigenvector, none of the algorithms incorporated an optimal step size selection rule.

The key idea behind the algorithm is to rewrite the constrained optimisation problem as an unconstrained one on a complex Grassmann manifold. In general, the (n, p) complex Grassmann manifold is the collection of all p-dimensional subspaces of $\mathbb{C}^n$. Since $\boldsymbol{x}$ is a vector, the relevant manifold is the $(n, 1)$ complex Grassmann manifold, also known as complex projective space [5, 8]. It is standard to represent a point in complex projective space by a vector $\boldsymbol{x}$ where $\boldsymbol{x}$ is constrained to the unit ball, that is, $\boldsymbol{x}^H\boldsymbol{x} = 1$. Although both $\boldsymbol{x}$ and $-\boldsymbol{x}$ correspond to the same point on complex projective space, the cost function (1) is such that $f(\boldsymbol{x}) = f(-\boldsymbol{x})$. Thus, by treating $\boldsymbol{x}$ as a point in complex projective space, a minimal eigenvector of A can be found by minimising $f(\boldsymbol{x})$ on complex projective space.

As with all descent type algorithms, given a point $\boldsymbol{x}$, the aim is to compute a descent direction $\boldsymbol{z}$ and a step size γ such that $f(\boldsymbol{x} + \gamma\boldsymbol{z}) < f(\boldsymbol{x})$. Steepest descent algorithms in Euclidean space choose $\boldsymbol{z}$ to be the negative of the gradient of f. As shown in [7], this concept can be extended to the complex Grassmann manifold. Specifically, the steepest descent direction $\boldsymbol{z}$ of the cost function (1), when treated as a function on complex projective space, can be shown to be

$$\boldsymbol{z} = -(I - \boldsymbol{x}\boldsymbol{x}^H)A\boldsymbol{x} = -A\boldsymbol{x} + (\boldsymbol{x}^H A\boldsymbol{x})\boldsymbol{x} \tag{2}$$

provided $\boldsymbol{x}^H\boldsymbol{x} = 1$.

Having derived the steepest descent direction, all that remains is to determine the step size γ. It is expedient though to first state the whole algorithm and then explain how the formula for computing γ was derived.

Algorithm 1 (Minimal Eigenvector) *Let $A \in \mathbb{C}^{n\times n}$ be an arbitrary Hermitian matrix. The following algorithm converges to a minimal eigenvector of A with probability one (see Theorem 2 below).*

This work was supported by the Australian Research Council.

1. *Randomly choose an* $\boldsymbol{x} \in \mathbb{C}^n$ *with unit norm* ($\boldsymbol{x}^H\boldsymbol{x} = 1$).
2. *Compute the descent direction* $\boldsymbol{z} := \bar{\lambda}\boldsymbol{x} - A\boldsymbol{x}$ *where* $\bar{\lambda} := \boldsymbol{x}^H A\boldsymbol{x}$. *If* $\sqrt{\boldsymbol{z}^H\boldsymbol{z}}$ *is sufficiently small, then stop.*
3. *Compute* $\alpha := \boldsymbol{x}^H A^2\boldsymbol{x} - \bar{\lambda}^2$ *and* $\beta := \boldsymbol{x}^H A^3\boldsymbol{x} - 3\alpha\bar{\lambda} - \bar{\lambda}^3$. *Set* γ *to the positive root of* $\alpha^2\gamma^2 + \beta\gamma - \alpha = 0$.
4. *Set* $\boldsymbol{x} := \boldsymbol{x} + \gamma\boldsymbol{z}$. *Renormalise by setting* $\boldsymbol{x} := \frac{\boldsymbol{x}}{\sqrt{\boldsymbol{x}^H\boldsymbol{x}}}$. *Go to Step 2.*

Remark: When implementing Alg. 1, it is important to store α, β and $\bar{\lambda}$ as real-valued variables.

In order to derive the formula for γ in Step 3 of Alg. 1, it is necessary to obtain an expression for the decrease in cost caused by taking a step of size γ in direction $\boldsymbol{z}$. Since the step is performed in complex projective space, the point $\boldsymbol{x}$ goes to the point $\frac{\boldsymbol{x}+\gamma\boldsymbol{z}}{(\boldsymbol{x}+\gamma\boldsymbol{z})^H(\boldsymbol{x}+\gamma\boldsymbol{z})}$ rather than to the point $\boldsymbol{x} + \gamma\boldsymbol{z}$. Straightforward manipulation shows that the decrease in cost is given by

$$\begin{aligned} f(\boldsymbol{x}) - f\left(\frac{\boldsymbol{x}+\gamma\boldsymbol{z}}{\sqrt{(\boldsymbol{x}+\gamma\boldsymbol{z})^H(\boldsymbol{x}+\gamma\boldsymbol{z})}}\right) \\ = \frac{1}{2}\bar{\lambda} - \frac{1}{2}\left[(\boldsymbol{x}-\gamma\bar{A}\boldsymbol{x})^H(\boldsymbol{x}-\gamma\bar{A}\boldsymbol{x})\right]^{-1} \\ (\boldsymbol{x}-\gamma\bar{A}\boldsymbol{x})^H(\bar{A}+\bar{\lambda}I)(\boldsymbol{x}-\gamma\bar{A}\boldsymbol{x}) \\ = \frac{\gamma\left(\alpha - \frac{1}{2}\gamma\beta\right)}{1+\alpha\gamma^2} \end{aligned} \tag{3}$$

where

$$\bar{A} = A - \bar{\lambda}I, \qquad \alpha = \boldsymbol{x}^H\bar{A}^2\boldsymbol{x}, \qquad \beta = \boldsymbol{x}^H\bar{A}^3\boldsymbol{x}. \tag{4}$$

Note that α and β are real-valued since $\bar{A} = \bar{A}^H$. Also, since $\boldsymbol{x}^H\boldsymbol{x} = 1$, $\boldsymbol{x}^H\bar{A}\boldsymbol{x} = 0$.

Differentiating (3) with respect to γ and setting the result to zero shows that the greatest decrease in cost occurs when γ is the unique positive root of the quadratic equation given in Step 3 of Alg. 1.

Before proving global convergence, two properties of Alg. 1 are stated. Alg. 1 is invariant to shifts; replacing A with $A - \lambda I$ for any $\lambda \in \mathbb{R}$ has no effect. This supports the empirical evidence (see Section 3) that closely spaced eigenvalues, which are known to reduce severely the rate of convergence of Power methods [3], do not affect the performance of Alg. 1. Alg. 1 is also invariant to orthogonal changes of coordinates. That is, if Alg. 1 produces the sequence $\left\{\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \cdots\right\}$, then replacing A with QAQ^H and $\boldsymbol{x}^{(0)}$ with $Q\boldsymbol{x}^0$ will produce the sequence $\left\{Q\boldsymbol{x}^{(0)}, Q\boldsymbol{x}^{(1)}, \cdots\right\}$.

Theorem 2 (Convergence) *Let* $\boldsymbol{x}$ *be the initial vector chosen in Step 1 of Alg. 1. If* λ_1 *is the smallest eigenvalue of* A *and there exists an eigenvector* $\boldsymbol{v}_1$ *satisfying both* $A\boldsymbol{v}_1 = \lambda_1\boldsymbol{v}_1$ *and* $\boldsymbol{v}_1^H\boldsymbol{x} \neq 0$, *then Alg. 1 converges to an eigenvector* $\boldsymbol{v}$ *satisfying* $A\boldsymbol{v} = \lambda_1\boldsymbol{v}$.

PROOF. Referring to Alg. 1, since $\boldsymbol{z}^H\boldsymbol{z} = 0$ if and only if $\boldsymbol{x}$ is an eigenvector of A, it is clear that Alg. 1 converges to an eigenvector $\boldsymbol{v}$ of A. Let λ be the eigenvalue associated with $\boldsymbol{v}$. Assume to the contrary that $\lambda > \lambda_1$. Since $\boldsymbol{v}$ must then be orthogonal to $\boldsymbol{v}_1$, this implies $|\boldsymbol{v}_1^H\boldsymbol{x}| \to 0$. It will be shown below that one iteration of Alg. 1 increases $|\boldsymbol{v}_1^H\boldsymbol{x}|$ if the step size $\gamma > 0$ satisfies

$$\gamma\left[\alpha - (\bar{\lambda} - \lambda_1)^2\right] < 2(\bar{\lambda} - \lambda_1). \tag{5}$$

Since $\boldsymbol{x} \to \boldsymbol{v}$, it follows that $\bar{\lambda} = \boldsymbol{x}^H A\boldsymbol{x} \to \lambda$ and

$$\alpha = \boldsymbol{x}^H A^2\boldsymbol{x} - \bar{\lambda}^2 \to 0.$$

This means there will come a time when $\alpha - (\bar{\lambda} - \lambda_1)^2 < 0$, and hence (5) too, will hold for all subsequent iterations. This contradicts $|\boldsymbol{v}_1^H\boldsymbol{x}| \to 0$, proving that $\lambda = \lambda_1$.

To show (5) implies $|\boldsymbol{v}_1^H\boldsymbol{x}|$ will increase, note first that direct substitution proves that

$$\boldsymbol{v}_1^H\frac{\boldsymbol{x}+\gamma\boldsymbol{z}}{\sqrt{(\boldsymbol{x}+\gamma\boldsymbol{z})^H(\boldsymbol{x}+\gamma\boldsymbol{z})}} = \frac{1-\gamma(\lambda_1 - \bar{\lambda})}{\sqrt{1+\alpha\gamma^2}}. \tag{6}$$

Since $\alpha \geq 0$, it is readily verified that $\left[\frac{1-\gamma(\lambda_1-\bar{\lambda})}{\sqrt{1+\alpha\gamma^2}}\right]^2 > 1$ if and only if (5) holds. That is, (5) implies $|\boldsymbol{v}_1^H\boldsymbol{x}|$ will increase, unless $|\boldsymbol{v}_1^H\boldsymbol{x}| = 0$. However, the latter cannot occur because it is straightforward to show that $\lambda_1 \leq \bar{\lambda}$ always holds ($\bar{\lambda}$ is a weighted average of the eigenvalues of A), and so $1 - \gamma(\lambda_1 - \bar{\lambda})$ can never be zero. □

3. SIMULATIONS

This section studies the convergence rate of Alg. 1 and compares it with traditional methods for calculating extremal eigenvectors. It is demonstrated that the performance of Alg. 1 is relatively insensitive to the actual eigenvalue distribution.

The Inverse Iteration method [3] for finding an eigenvector of the matrix A associated with the eigenvalue having the smallest absolute value is to generate a sequence $\left\{\boldsymbol{x}^{(k)}\right\}$ of vectors according to the rule

$$\boldsymbol{x}^{(k+1)} = \frac{A^{-1}\boldsymbol{x}^{(k)}}{\|A^{-1}\boldsymbol{x}^{(k)}\|}. \tag{7}$$

Figures 1 to 4 compare the Inverse Iteration method (7) with the Steepest Descent method (Alg. 1). Figures 1 and 3 show that Alg. 1 outperforms (7) if the eigenvalues of A are closely spaced, while Figures 2 and 4 demonstrate that the converse holds too. This is now explained in more detail.

It is well-known that the convergence rate of the Power and Inverse Iteration methods [3] applied to the matrix A critically depends on the eigenvalue distribution of A. Indeed, replacing A with $A + \lambda I$ for some constant $\lambda \in \mathbb{R}$ (known as a shift in the literature) significantly alters the convergence rate of (7). In comparison, Section 2 shows that such shifts do not alter Alg. 1 at all. It is therefore expected that the Inverse Iteration method will exhibit convergence rates ranging from extremely poor to extremely good depending on the eigenvalue distribution of A whereas Alg. 1 is expected to achieve a steady rate of convergence over a wide range of eigenvalue distributions.

This hypothesis was tested by plotting the log of the error, defined as $\log\left(\left(\boldsymbol{x}^{(k)}\right)^H A\boldsymbol{x}^{(k)} - \lambda_{\min}\{A\}\right)$ where $\lambda_{\min}\{A\}$ is the smallest eigenvalue of A, against the iteration number k. (The fact that the resulting graphs in Figures 1 to 4 are essentially straight lines shows that both algorithms achieve a linear rate of convergence [9].) Figure 1 was generated by applying the algorithms to the matrix $A = \mathrm{diag}\{1, 1.01, 1.02, 1.03, 1.04\}$. (In all simulations, the initial starting vector was chosen to be $\boldsymbol{x}^{(0)} = [1\ 1\ 1\ 1\ 1]^T$.) Since the eigenvalues are closely spaced,

Alg. 1 significantly outperforms (7). Conversely, Figure 2 shows that (7) outperforms Alg. 1 when applied to the matrix

$$A = \text{diag}\{1, 2, 3, 4, 5\}.$$

Figures 3 and 4 suggest that this behaviour is typical. Figure 4 shows the performance of the two algorithms when applied to ten randomly generated 20-by-20 matrices with eigenvalues uniformly distributed between 0 and 1. The same ten matrices were then shifted so their eigenvalues lay between 10 and 11 (that is, each A was replaced with $A + 10I$), and the results plotted in Figure 3. Whereas the performance of Alg. 1 is unaltered, (7) performs badly in Figure 3 but exceptionally well in Figure 4.

The Steepest Ascent method, obtained by replacing A with $-A$ in Alg. 1, was compared with the Power method for converging to an eigenvector associated with the largest eigenvalue of A. The Power method updates $\boldsymbol{x}^{(k)}$ according to the rule (c.f., (7)) $\boldsymbol{x}^{(k+1)} = \frac{A\boldsymbol{x}^{(k)}}{\|A\boldsymbol{x}^{(k)}\|}$. Figures 5 and 6 were generated analogously to Figures 3 and 4. They demonstrate that Alg. 1 achieves a convergence rate which is much less sensitive to the location of the eigenvalues of A than the Power method does.

4. CONCLUSION

This paper applied the novel optimisation algorithm in [7] to the problem of finding an eigenvector associated with the smallest or the largest eigenvalue of a Hermitian matrix. The optimal step size was calculated and a global convergence proof was given. Simulations showed that, unlike classical algorithms for finding extremal eigenvectors, the convergence rate of the proposed method is relatively insensitive to the eigenvalue distribution.

5. REFERENCES

[1] P. Comon and G. H. Golub. Tracking a few extreme singular values and vectors in signal processing. *Proceedings of the IEEE*, 78(8):1327–1343, August 1990.

[2] S. C. Douglas, S. Amari, and S.-Y. Kung. On gradient adaptation with unit-norm constraints. *IEEE Transactions on Signal Processing*, 48(6):1843–1847, June 2000.

[3] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3rd edition, 1996.

[4] R. E. Mahony, U. Helmke, and J. B. Moore. Gradient algorithms for principal component analysis. *Journal of the Australian Mathematical Society, series B*, 37(4):430–450, 1996.

[5] J. H. Manton. An improved least squares channel identification algorithm. *IEEE Signal Processing Letters*, 2000. Submitted.

[6] J. H. Manton. Optimisation algorithms exploiting complex orthogonality constraints. *IEEE Transactions on Signal Processing*, 2000. Submitted.

[7] J. H. Manton. Steepest descent and Newton algorithms on the complex Grassmann manifold for orthogonally constrained optimisation problems. In *Sixth International Symposium on Signal Processing and Its Applications*, Kuala-Lumpur, Malaysia, August 2001. Submitted.

[8] J. H. Manton, Y. Hua, and X. Cao. Precoder assisted channel estimation in complex projective space. In *IEEE Signal Processing Advances in Wireless Communications*, Taiwan, March 2001. Accepted.

[9] E. Polak. *Optimization: Algorithms and Consistent Approximations*. Springer-Verlag, 1997.

[10] J.-F. Yang and M. Kaveh. Adaptive eigensubspace algorithms for direction or frequency estimation and tracking. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2):241–251, February 1988.

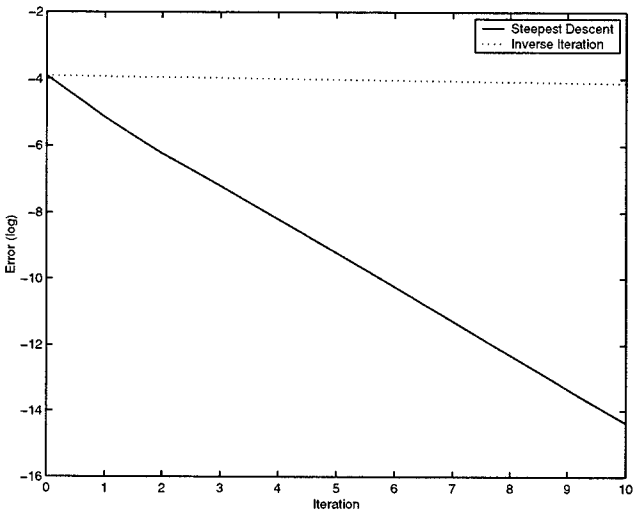


Figure 1: Graph comparing the convergence rates of the steepest descent and inverse iteration algorithms when applied to the matrix $A = \text{diag}\{1, 1.01, 1.02, 1.03, 1.04\}$.

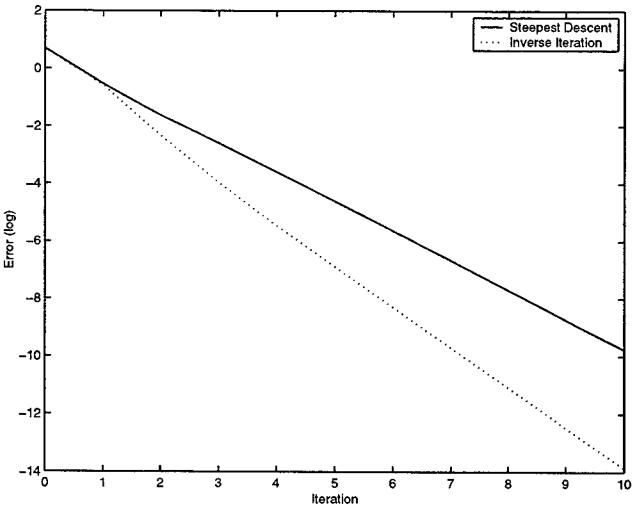


Figure 2: Graph comparing the convergence rates of the steepest descent and inverse iteration algorithms when applied to the matrix $A = \text{diag}\{1, 2, 3, 4, 5\}$.

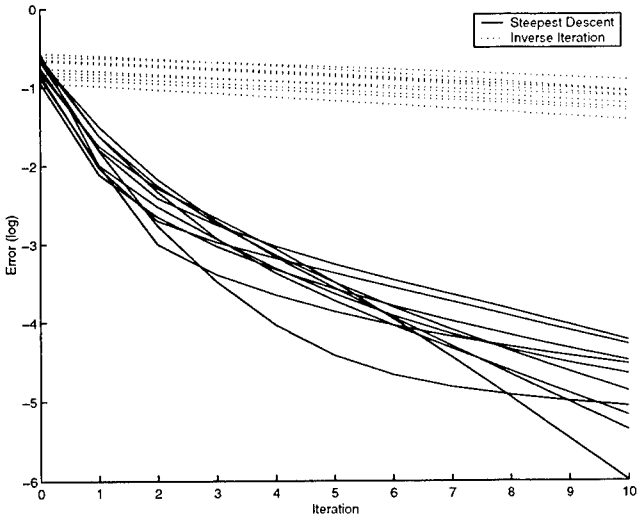


Figure 3: Graph comparing the convergence rates of the steepest descent and inverse iteration algorithms when applied to ten randomly generated 20-by-20 matrices with eigenvalues uniformly distributed between 10 and 11.

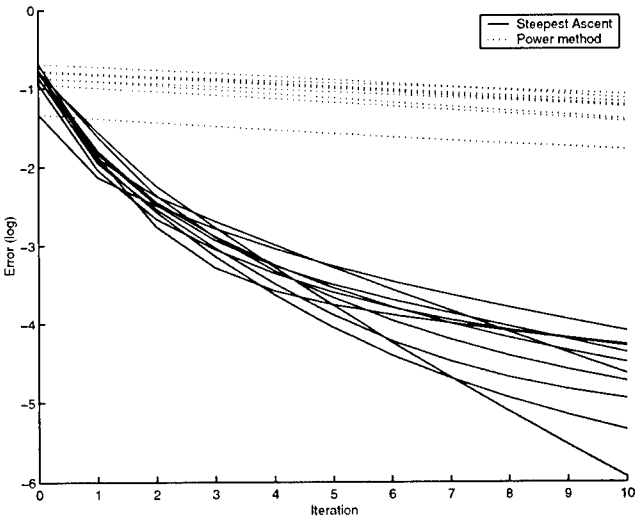


Figure 5: Graph comparing the convergence rates of the steepest ascent and Power method algorithms when applied to ten randomly generated 20-by-20 matrices with eigenvalues uniformly distributed between 10 and 11.

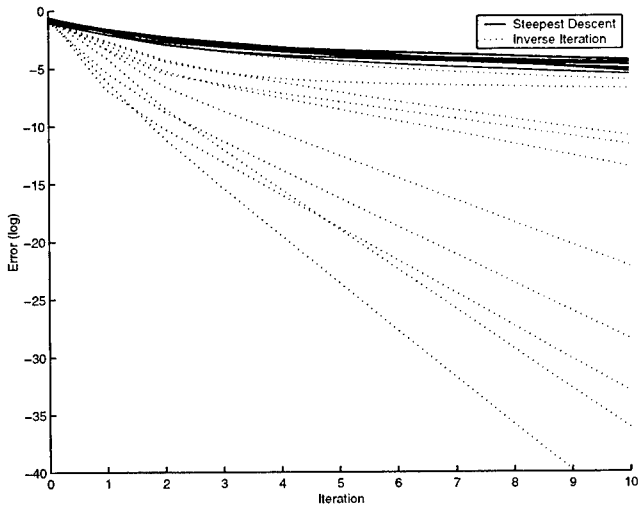


Figure 4: Graph comparing the convergence rates of the steepest descent and inverse iteration algorithms when applied to ten randomly generated 20-by-20 matrices with eigenvalues uniformly distributed between 0 and 1.

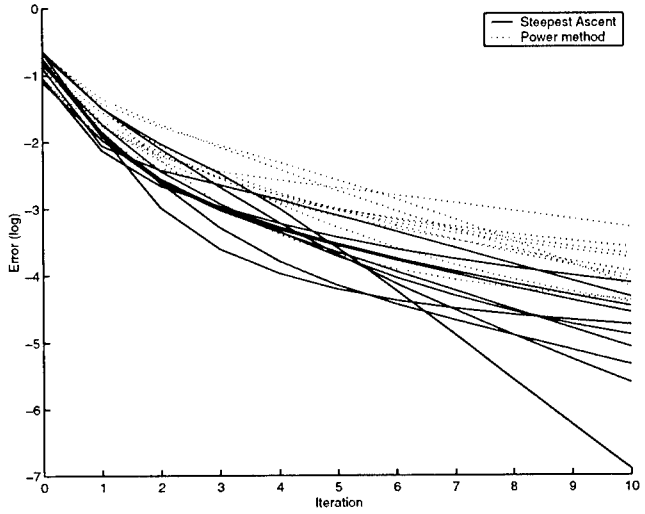


Figure 6: Graph comparing the convergence rates of the steepest ascent and Power method algorithms when applied to ten randomly generated 20-by-20 matrices with eigenvalues uniformly distributed between 0 and 1.

LOCALLY OPTIMAL MAXIMUM-LIKELIHOOD COMPLETION OF A PARTIALLY SPECIFIED TOEPLITZ COVARIANCE MATRIX

Yuri I. Abramovich[1,2] *and Nicholas K. Spencer*[2]

[1] Surveillance Systems Division, Defence Science and Technology Organisation (DSTO), P.O. Box 1500, Salisbury, South Australia, 5108, Australia

[2] Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP), SPRI Building, Technology Park Adelaide, Mawson Lakes, South Australia, 5095, Australia

yuri@cssip.edu.au nspencer@cssip.edu.au

ABSTRACT

The problem of maximum-likelihood (ML) completion of a partially specified Toeplitz covariance matrix is crucial in several applications, such as the detection and estimation of more independent Gaussian sources than sensors ($m > M$) in minimum-redundancy sparse linear antenna arrays. Given the sufficient statistic in the form of the M-variate direct data covariance matrix $\hat{R}$, we describe an algorithm that finds a positive-definite completed M_α-variate Toeplitz matrix ($M_\alpha \gg M$) with (locally) maximal likelihood ratio (LR). Simulations demonstrate a statistically high LR is achieved, compared with L_2 optimisation.

1. PROBLEM FORMULATION

Consider an M-element nonuniform linear array (NLA) with sensors located at positions $\boldsymbol{d}$, restricted to integer values measured in the inter-element spacing units of d, usually equal to half a wavelength:

$$\boldsymbol{d} = [d_1 \equiv 0,\, d_2,\, d_3,\, \ldots,\, d_M] \qquad (1)$$

We assume Gaussian processes are observed at the output of the sparse array as a combination of m uncorrelated plane waves with DOAs $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]^T$, powers $P = \mathrm{diag}\,[p_1, \ldots, p_m]$ and white noise of power p_0. Thus the M-variate vector of observed sensor outputs at time t (the "snapshot") is

$$\boldsymbol{y}(t) = B\,\boldsymbol{x}(t) + \boldsymbol{\eta}(t)\,, \qquad \text{for} \quad t = 1, \ldots, N \qquad (2)$$

where $\boldsymbol{y}(t) \in \mathcal{C}^{M\times 1}$, $\boldsymbol{x}(t) \in \mathcal{C}^{m\times 1}$ is the vector of Gaussian signal amplitudes with the property

$$\mathcal{E}\{\boldsymbol{x}(t_1)\,\boldsymbol{x}^H(t_2)\} = \begin{cases} P & \text{for} \quad t_1 = t_2 \\ 0 & \text{for} \quad t_1 \neq t_2\,, \end{cases} \qquad (3)$$

and $\boldsymbol{\eta}(t) \in \mathcal{C}^{M\times 1}$ is additive white Gaussian noise. The array-signal manifold matrix is $B = [\boldsymbol{b}(\theta_1), \ldots, \boldsymbol{b}(\theta_m)] \in \mathcal{C}^{M\times m}$, where each "steering vector" is

$$\boldsymbol{b}(\theta_j) = \left[1, \exp\left(i2\pi\frac{d_2}{\lambda}\sin\theta_j\right), \ldots, \exp\left(i2\pi\frac{d_M}{\lambda}\sin\theta_j\right)\right]^T \qquad (4)$$

and λ is the wavelength of incident radiation.

For a uniformly-spaced linear array (ULA), the array-signal manifold matrix $S = [\boldsymbol{s}(\theta_1), \ldots, \boldsymbol{s}(\theta_m)] \in \mathcal{C}^{M_\alpha\times m}$ is of Vandermonde structure, with

$$\boldsymbol{s}(\theta_j) = \left[1, \exp(i\omega_j), \ldots, \exp\left(i[M_\alpha - 1]\,\omega_j\right)\right]^T \qquad (5)$$

where the spatial frequency is $\boldsymbol{\omega} = 2\pi\frac{d}{\lambda}\sin\boldsymbol{\theta}$, and d is the inter-element unit spacing.

By definition, the M-element NLA is a subarray of the M_α-element ULA. Their relationship may be described by the $M \times M_\alpha$ binary selection (or incidence) matrix L, where L_{jk} is equal to unity in the jth row and d_k^{th} column, and zero otherwise. Thus the NLA manifold can be written $B = LS$, and the (virtual) ULA p.d. Toeplitz spatial covariance matrix is $T = SPS^H + p_0\, I_{M_\alpha}$ and is related to the p.d. Hermitian covariance matrix R of the actual NLA $R = BPB^H + p\, I_M$ by the crucial linear transformation $R = LTL^T$.

We restrict this study to the class of *identifiable* scenarios, *ie.* situations with a one-to-one correspondence between the set of signal parameters (m, $\boldsymbol{\theta}$, P, p_0) and the covariance matrix R. Consequently, there is a one-to-one correspondence between the M-variate (true) covariance matrix R and the M_α-variate covariance matrix T for any m of interest ($1 \leq m \leq M_\alpha - 1$). In fact, identifiability is not guaranteed for sparse arrays (identifiability issues are addressed in [1]).

Given N independent snapshots, the sufficient statistic for DOA estimation is the direct data covariance

0-7803-7011-2/01/$10.00 ©2001 IEEE

(DDC) matrix $R = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{y}(t)\,\boldsymbol{y}^H(t)$ that, simultaneously, is the ML estimate of the unstructured (*ie.* arbitrary p.d. Hermitian) covariance matrix.

Partially augmentable NLAs have one or more missing covariance lags [2], *eg.* the geometry

$$\boldsymbol{d}_5 = [0,\ 1,\ 4,\ 9,\ 11] \tag{6}$$

embodies all covariance lags except the single lag t_6. Therefore, if the sample covariance lags R_{jk} of R are used to construct an estimate of the augmented M_o-variate Toeplitz matrix, one or more diagonals of such an augmented matrix will be missing.

Traditionally, the straightforward direct augmentation approach (DAA) [3] is used to form the augmented Toeplitz matrix:

$$\hat{t}_{j-k=\kappa} = \frac{\sum_{j,k} R_{jk}\,\delta(\kappa,\, d_j - d_k)}{\sum_{j,k} \delta(\kappa,\, d_j - d_k)}, \quad j > k,\ \kappa \in S \tag{7}$$

where $\delta(a,b)$ is the generalised Kronecker delta function, and we have defined the complementary sets $S = \{\kappa : t_\kappa \text{ is specified}\}$ and $\overline{\mathcal{S}} = \{\kappa : t_\kappa \text{ is unspecified}\}$. For minimum redundancy arrays, such as $\boldsymbol{d}_5$, none of the elements in R are redundant (obviously except for R_{jj}), and in this case

$$\hat{t}_{j-k=\kappa} = \sum_{j,k} R_{jk}\,\delta(\kappa,\, d_j - d_k)\,. \tag{8}$$

Interestingly, (7) and especially (8) could be viewed as the optimum unconstrained solution that yields the minimum in the L_2 norm: $\|R - LTL^H\|_2$. This is important, since it is the L_2 norm (with an additional weighting matrix) that constitutes COMET [4].

In our approach, we are looking for the p.d. Toeplitz matrix estimate $\overline{T}$, and for its M-variate linear transformation $\overline{R} = L\overline{T}L^H$ that yields the (local) maximum to the (*eg.* sphericity test) likelihood ratio $\gamma(\overline{R})$:

$$\gamma(R) = \gamma_0^N(R), \qquad \gamma_0(R) = \frac{\det[\overline{R}^{-1}R]}{\left[\frac{1}{M}\,\mathrm{tr}\,[\overline{R}^{-1}R]\right]^M} \tag{9}$$

in the vicinity of $\overline{T}$, defined by (7),(8). Obviously, the problem of T estimation given the DDC matrix R, with some continuous (element-wise) transformation $R \to T$, $T \to R$, is important for many other applications (*eg.* see [5]).

2. PROBLEM SOLUTION

Observe that the LR (9) could be presented as

$$\gamma_0(R) = \frac{\prod_{k=1}^{M} \lambda_k^{-1}}{\left[\frac{1}{M}\sum_{k=1}^{M} \lambda_k^{-1}\right]^M} \tag{10}$$

where λ_k $(k = 1, \ldots, M)$ are eigenvalues of the matrix

$$G(R) = R^{-\frac{1}{2}}\,\overline{R}\,R^{-\frac{1}{2}}\,. \tag{11}$$

We present optimisation steps as a sequence of sufficiently small perturbations $T_0 \to T_1 \to \ldots \to \overline{T}$ where $T_{k+1} = T_k + \delta(T)$; specifically

$$T_{k+1}(\boldsymbol{z}) = T_k + \sum_k (\boldsymbol{z}_k E_+^k + \boldsymbol{z}_k^* E_-^k) \tag{12}$$

where

$$E_+ = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}, \quad E_- = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix} \tag{13}$$

or in terms of real variables

$$T_{k+1}(\boldsymbol{z}) = T_k + \sum_k [\mathcal{R}e\,(\boldsymbol{z}_k F_k^+) + i\,\mathcal{I}m\,(\boldsymbol{z}_k F_k^-)] \tag{14}$$

where

$$F_k^+ = E_+^k + E_-^k; \quad F_k^- = E_+^k - E_-^k\,. \tag{15}$$

It is well known that the set of Toeplitz Hermitian matrices is congruent to the set of real symmetric matrices via the unitary transformation:

$$Q = \mathcal{R}e\,(T) + J_p\,\mathcal{I}m\,(T) = HTH^H; \quad T = H^H Q H\,,$$

$$H = \frac{1}{2}[(I + J_p) + i(I - J_p)]; \quad H^H H = I \tag{16}$$

where

$$J_p = \begin{bmatrix} & & & 1 \\ & & \iddots & \\ & 1 & & \\ 1 & & & \end{bmatrix} \tag{17}$$

is the permutation matrix. Evidently

$$\det Q = \det(HTH^H) = \det T; \ \ \mathrm{eig}(Q) = \mathrm{eig}(T). \tag{18}$$

We may then transform to real-valued perturbations:

$$Q_{k+1} \equiv HT_{k+1}(\boldsymbol{z})H^H = HT_kH^H + \sum_{k=1}^{M_o} x_k' F_k^+ + x_k'' F_k^-,$$

$$Q_{k+1} =: Q_\kappa + \sum_{k=1}^{2M_o-1} x_k F_k \tag{19}$$

so for sufficiently small x_k, we can treat $\sum_{k=1}^{2M_o-1} x_k F_k = \delta Q$ as a perturbation of the matrix Q_k.

Supposing sufficiently small perturbations in T_k (and thus in Q_k), we have small perturbations in the iterate

R_k: $R_0 \to R_1 \to \ldots \to \overline{R}$ and therefore in the iterate $\hat{G}_k$: $\hat{G}_0 \to \hat{G}_1 \to \ldots \to \hat{\sigma}(\overline{R})$. Then we can write the first-order expansion of the eigenvalues of $\hat{G}_{k+1}$ as:

$$\lambda_\ell(k+1) = \lambda_\ell(k) + \delta_\ell(k+1) \tag{20}$$

with $|\delta_\ell(k+1)| < \lambda_\ell(k)$, so that the inverse $\lambda_\ell^{-1}(k+1)$ is also given by a first-order expansion:

$$\lambda_\ell^{-1}(k+1) = \lambda_\ell^{-1}(k)[1 - \delta_\ell(k+1)\lambda_\ell^{-1}(k)] \tag{21}$$

then the LR $\gamma(R_{k+1})$ could be also presented sufficiently accurately by its first-order expansion as:

$$\gamma_0(R_{k+1}) = \gamma_0(R_k) \frac{1 - \sum_{\ell=1}^{M} \delta_\ell(k+1)\,\lambda_\ell^{-1}(k)}{\left[1 - \frac{\sum_{\ell=1}^{M} \delta_\ell(k+1)\,\lambda_\ell^{-2}(k)}{\sum_{\ell=1}^{M} \lambda_\ell^{-1}(k)}\right]^M} \tag{22}$$

$$\approx \gamma_0(R_k) \exp\left[-\sum_{\ell=1}^{M} \delta_\ell(k+1)\left(\lambda_\ell^{-1}(k) - \lambda_\ell^{-2}(k)\frac{M}{\Lambda_k}\right)\right] \tag{23}$$

where $\Lambda_k = \sum_{\ell=1}^{M} \lambda_\ell^{-1}(k)$. Obviously local LR maximisation in its first-order approximation is equivalent to minimisation of the linear function

$$\sum_{\ell=1}^{M} \delta_\ell(k+1)\left[\lambda_\ell^{-1}(k) - \lambda_\ell^{-2}(k)\frac{M}{\sum_{\ell=1}^{M} \lambda_\ell^{-1}(k)}\right] \tag{24}$$

subject to some constraints that ensure the perturbed matrix Q_{k+1} (and thus T_{k+1}) are p.d., and that the perturbations (19) are sufficiently small for the validity of the first-order expansions. Given the perturbations of Q_{k+1} (19), the perturbations in $\hat{G}_{k+1}$ are

$$\hat{G}_{k+1} = \hat{G}_k + \sum_{\ell=1}^{2M_\alpha - 1} x_\ell \hat{R}^{-\frac{1}{2}} L H^H F_\ell H L^H \hat{R}^{-\frac{1}{2}} \tag{25}$$

and according to eigenvalue perturbation theory [6],

$$\lambda_j(\hat{G}_{k+1}) = \lambda_j(\hat{G}_k) + \sum_{\ell=1}^{2M-1} x_\ell \boldsymbol{g}_j^{(k)H} (\hat{R}^{-\frac{1}{2}} L H^H F_\ell H L^H \hat{R}^{-\frac{1}{2}}) \boldsymbol{g}_j^{(k)} \tag{26}$$

where $\boldsymbol{g}_j^{(k)}$ is the jth ($j = 1, \ldots, M$) eigenvector of the matrix $\hat{G}_k$. Similarly, the eigenvalues of the perturbed matrix Q_{k+1} (and T_{k+1} via (18)) can be written as:

$$\sigma_j(Q_{k+1}) = \sigma_j(Q_k) + \sum_{\ell=1}^{2M-1} x_\ell U_j^{(k)H} F_\ell U_j^{(k)}. \tag{27}$$

We now introduce the $M \times 2(M_\alpha - 1)$ matrix iterate $\mathcal{D}_k$

$$\mathcal{D}_k = \left[\boldsymbol{g}_j^{(k)H} \hat{R}^{-\frac{1}{2}} L H^H F_\ell H L^H \hat{R}^{-\frac{1}{2}} \boldsymbol{g}_j^{(k)}\right]_{j=1,\ldots,M}^{\ell=1,\ldots,2M_\alpha-1} \tag{28}$$

and the $M_\alpha \times 2(M_\alpha - 1)$ matrix iterate $\mathcal{P}_k$

$$\mathcal{P}_k = \left[U_j^{(k)H} F_\ell U_j^{(k)}\right]_{j=1,\ldots,M_\alpha}^{\ell=1,\ldots,2M_\alpha-1}. \tag{29}$$

According to (26), we may present the vector $\boldsymbol{\Delta}_{k+1}^T = [\delta_1(k+1), \ldots, \delta_M(k+1)]$ as $\boldsymbol{\Delta}_{k+1} = \mathcal{D}_k \boldsymbol{x}$ and perturbations to eigenvalues of the matrix Q_{k+1} as $\mathcal{P}_k \boldsymbol{x}$.

Finally, with all introduced notations, we can formulate the problem of the optimum perturbation as the linear programming problem:

$$\text{Find} \quad \min(L_k^T \mathcal{D}_k \boldsymbol{x}) \quad \text{subject to} \tag{30}$$

$$-\boldsymbol{\sigma}_k - \mathcal{P}_k \boldsymbol{x} < -\sigma_0 \mathbf{1} \tag{31}$$

$$-\varepsilon < \boldsymbol{x}_\ell < -\varepsilon, \quad \ell = 1, \ldots, 2M_\alpha - 1 \tag{32}$$

where

$$L_k = \left(\lambda_\ell^{-1}(k) - \lambda_\ell^{-2}(k)\frac{M}{\sum_{\ell=1}^{M} \lambda_\ell^{-1}(k)}\right) \tag{33}$$

for $\ell = 1, \ldots, M$; $\boldsymbol{\sigma}_k$ is the M_α-variate vector of eigenvalues of Q_k, and σ_0 is the minimum eigenvalue of the initial matrix T_0.

The M_α linear constraints (31) keep the optimised Toeplitz matrix T_{k+1} (and hence R_{k+1}) p.d., while the linear constraints (32) ensure the perturbations are sufficiently small. Of course, step size management is necessary to ensure that the solution obtained by both expansions (26) and (27) are valid, and that the LR is actually increased, whilst maintaining $T_{k+1} > 0$. If either condition is not satisfied, the "step size" ε is reduced, and the LP problem is solved again. If both conditions are met, we compute Q_{k+1} and use it for the next iteration T_k.

Finally, we need to specify the initialisation of our routine that produces an initial p.d. matrix T_0, given the DAA complete matrix $\hat{T}$. This step is explicitly described in [2], since we use maximum entropy (ME) completion as the starting point ($T_0 = T_{ME}$). In [2], we demonstrated that convex programming routines could be used to check the feasibility conditions (*ie.* that $\hat{T}$ could be p.d. completed), and to introduce L_2-minimal perturbations to the specified entries to actually achieve a p.d. completion. Next, the "missing" elements in T are optimised to achieve the ME condition. We emphasise that the missing lags are not presented in $R = LTL^H$, and thus they *do not affect* the optimised LR. Nevertheless, due to the p.d. condition, these lags need to be optimised in order to give more freedom to the specified-lag perturbations in pursuing the maximum LR. In our full detection-estimation algorithm (forthcoming paper), the solution $\hat{T}$ is the one that corresponds to the maximum number of independent

sources ($M_\alpha - 1$), thus yielding the maximum LR. Further equalisation of the last ($M_\alpha - \mu$) eigenvalues, also conducted with an iterated LP optimisation, generates "candidate models" T_μ with correspondingly degraded LR. Obviously, optimisation of the missing lags is crucial now for equalisation, and so the two-stage procedure is proposed. In the first stage, we optimise only the missing lags (that keep the LR unchanged), and only after convergence do we modify all entries for accurate equalisation, causing a degradation in LR.

3. SIMULATION RESULTS AND CONCLUSIONS

To demonstrate the efficiency of our method, we present the results of 1000 Monte-Carlo trials, conducted for the antenna array $\boldsymbol{d}_5$ and the identifiable six-source scenario $\boldsymbol{w}_0 = [-0.9, -0.68, -0.46, -0.24, -0.02, 0.20]$. Obviously, since the sixth lag is missing in T, no existing technique is directly applicable for detection and DOA estimation. Fig. 1 presents distributions of the LR calculated for:

(a) the true (exact) covariance matrix for $\boldsymbol{d}_5 + \boldsymbol{w}_0$,

(b) the ME completion $T_0 = T_{ME}$ [2],

(c) the locally optimal ML completion T_{ML}.

Note that T_{ME} experienced minimal perturbations to the specified R entries in order to achieve feasibility. Despite the perturbations in T_{ML} being minimal in the L_2 sense, the LR distribution is significantly worse than that for the optimised solution T. More importantly, comparison of the LR distributions for $\overline{T}$ and T demonstrate that in many cases the LR-optimised solution T_{ML} exhibits an LR greater than the true covariance matrix generated for the same sufficient statistics R! Similarly to the fully augmentable case [7], $LR(T_{ML})$ is clearly right-skewed compared with $LR(R)$. Though we still cannot prove that in every trial the global LR maximum has been achieved, this is a sufficient argument to treat further attempts to improve the LR (via local LR optimisation) as statistically unproductive in terms of detection-estimation efficiency. The introduced results also demonstrate, that even for sufficiently large sample volumes, the LR remains very sensitive and in many cases, optimisation according to some "related" criteria (such as minimum Euclidean norm in COMET [4]), could lead to results that are surprisingly far from the optimum.

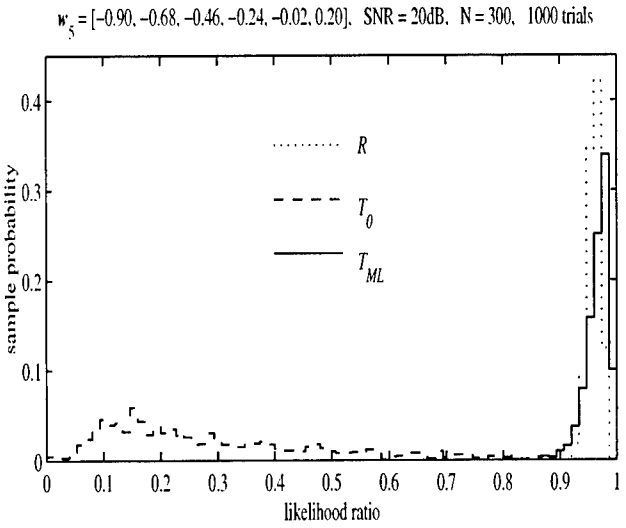


Figure 1: Comparison of likelihood ratio (LR) distributions.

REFERENCES

[1] Y.I. Abramovich, N.K. Spencer, and A.Y. Gorokhov, "Resolving manifold ambiguities in direction-of-arrival estimation for nonuniform linear antenna arrays," *IEEE Trans. Sig. Proc.*, vol. 47 (10), pp. 2629–2643, 1999.

[2] Y.I. Abramovich, N.K. Spencer, and A.Y. Gorokhov, "Positive-definite Toeplitz completion in DOA estimation for nonuniform linear antenna arrays — Part II: Partially augmentable arrays," *IEEE Trans. Sig. Proc.*, vol. 47 (6), pp. 1502–1521, 1999.

[3] S. Pillai, Y. Bar-Ness, and F. Haber, "A new approach to array geometry for improved spatial spectrum estimation," *Proc. IEEE*, vol. 73 (10), pp. 1522–1524, 1985.

[4] B. Ottersten, P. Stoica, and R. Roy, "Covariance matching estimation technique for array signal processing applications," *Digital Signal Processing*, vol. 8, pp. 185–210, 1999.

[5] A. Paulraj and T. Kailath, "Direction of arrival estimation by eigenstructure methods with imperfect spatial coherence of wave fronts," *J. Acoust. Soc. Am.*, vol. 83, no. 3, pp. 1034–1040, March 1988.

[6] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.

[7] Y.I. Abramovich, N.K. Spencer, and A.Y. Gorokhov, "Detection-estimation of more uncorrelated Gaussian sources than sensors in nonuniform linear antenna arrays — Part I: Fully augmentable arrays," *IEEE Trans. Sig. Proc.*, vol. 49 (5), pp. 959–971, 2001.

EEG BRAIN MAP RECONSTRUCTION USING BLIND SOURCE SEPARATION

S. Sanei[1], A. R. Leyman[2]

[1]School of Electrical and Electronic Engineering, Singapore Polytechnic
[2]Centre for wireless communication, NUS, Science Park II, Singapore

ABSTRACT

EEG-based brain maps are very useful in anatomical, functional and pathological diagnosis. These images are projection of energy of the signals in four different frequency bands. Joint Approximate Diagonalization of Eigenmatrices (JADE) is used as an effective tool in deconvolution of EEG signals prior to spectrum estimation. The algorithm also, restores the noise from the signal as a result of Higher Order Statistics (HOS) estimation. The spectrum is estimated using autoregressive (AR) modelling and pseudo-hot colours are used to represent brain activities. The results testify a great enhancement in diagnostic features in the reconstructed images. The overall system also enables real-time reconstruction of the images for patient monitoring purposes.

1. INTRODUCTION

EEGs project electrical activities of the brain [1][2][3]. The EEG is divided into 4 sub-bands; Delta activity is around 4 Hz or below. It tends to be the highest in amplitude. It is the dominant rhythm in infants and in stages 3 and 4 of sleep. It may occur focally with subcortical lesions and in general distribution with diffuse lesions, metabolic encephalopathy hydrocephalus or deep midline lesions. It is usually most prominent frontally in adults. Theta activity has a frequency of 4 to 8 Hz. It is abnormal in awaked adults but is perfectly normal in children up to 13 years and in sleep. It can be seen as a focal disturbance in focal subcortical lesions. Alpha waves are those between 8 and 14 Hertz. Alpha is usually best seen in the posterior regions of the head on each side, being higher in amplitude on the dominant side. It is brought out by closing the eyes and by relaxation, and abolished by eye opening or alerting by any mechanism. It is the major rhythm seen in normal relaxed adults. Beta activity is 'fast' activity. It has a frequency of 14 Hz and above (normally up to about 40 KHz). It is usually seen on both sides in symmetrical distribution and is most evident frontally. It may be absent or reduced in areas of cortical damage. It is generally regarded as a normal rhythm. It is the dominant rhythm in patients who are alert or anxious or who have their eyes open.

In this paper, we address the issue of finding an appropriate and accurate method to extract the spectrum of each actual EEG signal. This is done by blind deconvolution of the signals followed by AR-based spectrum estimation. Reconstruction of the brain map will then be more accurate and informative. Triangular cubic interpolation criterion is exploited in mapping the energy of each frequency band into the image of cross section of the brain. Four images represent the activity of the brain in above four frequency subbands. Each EEG signal is actually a combination of an unknown number of sources inside the brain plus various internal signals such as heart rate, cardiovascular, muscular, and external signals such as system noise. A suitable procedure for deconvolution and restoration of the signals prior to neuro-image reconstruction, is highly demanded.

2. BLIND SIGNAL SEPARATION

Blind signal separation is the process of extracting the unknown source signals from their combinations. The channel and permutation of the output signals are also unknown to us. The simple model is as follows:

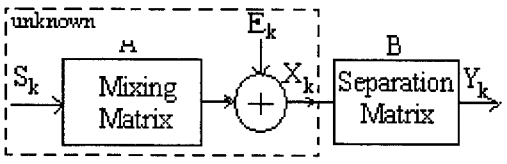


Figure 1. BSS block diagram

Noise normally appears as a disturbance to the signal. Because of the correlation between the two measurements, it is in principle possible to separate out the noise. Here, it is assumed that the EEG signals are linear, instantaneous combinations which are combined by using linear transformations. Under these assumptions, the problem can be rewritten in matrix formulation as

$$X_k = AS_k + E_k \qquad (1)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

and

$$Y_k = BX_k \tag{2}$$

where E_k is white Gaussian noise vector and A and B are unknown, constant matrices of sizes $m \times n$ and $n \times m$, respectively.

It is observed that producing output signals that are decorrelated is relatively easy whereas achievement of independent outputs require more work to be done. The mathematics of the separation task helps to explain this observation. We aim to produce independent outputs, i.e.,

$$p(y_i(k), y_j(k)) = p(y_i(k))\, p(y_j(k)) \tag{3}$$

for all pairs of outputs. A necessary but insufficient condition for having independent outputs is to have uncorrelated outputs, i.e., $E[y_i(k)y_j(K)] = E[y_i(k)]E[y_j(k)]$, for all pairs of outputs. A stronger condition will be $E[f(y_i(k))g(y_j(k))] = E[f(y_i(k))]E[g(y_j(k))]$, where $f(\cdot)$ and $g(\cdot)$ are nonlinear functions.

To form the decorrelated outputs $y(k) = Cx(k)$ where, C is a constant, linear transformation, let the covariance matrix of the measurements be $R = E[x(k)x(k)^T]$. Decorrelation of the output requires

$$E[y(k)y(k)^T] = E[Cx(k)x(k)^TC^T] = CRC^T = D \tag{4}$$

where D is a diagonal matrix. Let V be a matrix formed by assembling the eigenvectors of R into a matrix and W be a diagonal matrix whose main diagonal contains the eigenvalues of R.

Let $C = W^{1/2}V^T$. As the eigenvectors of R form orthogonal basis, it follows that:

$$CRC^T = W^{1/2}V^TVWV^TVW^{1/2} = I \tag{5}$$

However, the solution $C = UW^{1/2}V^T$, where U is an arbitrary unitary matrix, also decorrelates the outputs.

JADE algorithm was proposed by Cardoso [4]. This procedure uses matrices $Q_z(M)$ formed by the inner product of the fourth-order cumulant tensor of the outputs with an arbitrary matrix M, i.e.,

$$\{Q_z(M)\}_{ij} = \sum_{k=1}^{n}\sum_{l=1}^{n} Cum(z_i, z_j^*, z_k, z_i^*)m_{lk} \tag{6}$$

where the (l,k)th component of the matrix M is written as m_{lk} and $Z_k = CY_k$. The matrix $Q_z(M)$ has the important property that it is diagonalized by the correct rotation matrix U, i.e., $U^HQ_z(M)U = \Lambda_M$ where H denotes the complex (Hermitiam) transpose operator, and Λ_M is a diagonal matrix whose diagonal elements depend on the particular matrix M as well as Z_k. By using equation 6, For a set of different matrices M, a set of cumulant matrices $Q_z(M)$ can be calculated. The desired rotation matrix U then, jointly diagonalizes these matrices. The required procedure for estimation of $\hat{A}$ can be summarized into the following steps:

1. Form the sample covariance matrix

$$R_x = E\{X(t)X^T(t)\} \tag{7}$$

1. Compute a SVD of R_x

$$R_x = [U_s\ U_n]\begin{bmatrix} \lambda_1^2 & \cdot & \cdot & \cdot & 0 \\ 0 & \lambda_2^2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \lambda_m^2 \end{bmatrix}[U_s\ U_n]^T \tag{8}$$

Where $\lambda_i^2 = \sigma^2, i = N+1, \cdots m$ and σ^2 is the noise variance.

2. Estimate the number of the sources N by the number of singular values that do not equal to σ^2. The matrix U_s is composed of N singular vectors whose singular values do not equal to σ^2.
3. Whitened the data $x(t)$ by

$$y(t) = Wx(t) \tag{9}$$

where $W = D^{-1/2}U_s^T$ and the diagonal matrix is given as

$$D = diag\left(d_1^2, d_2^2, \cdots, d_N^2\right) \tag{10}$$

where $d_i^2 = \lambda_i^2 - \sigma^2 \quad i = 1,2,\cdots,m$.

4. Form the sample fourth-order cumulant matrix Q of the whitened data $y(t)$.
5. Compute the N most significant eigenpairs $\{\lambda_r, M_r : 1 \le r \le N\}$. This is done by first computing the eigen-decomposition of Q

$$Q = [U_s\ U_n] \begin{bmatrix} \lambda_1^2 & \cdot & \cdot & \cdot & 0 \\ 0 & \lambda_2^2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \lambda_m^2 \end{bmatrix} [U_s\ U_n]^T \quad (11)$$

Take each column of U_s and reshape the eigenvector into a $N \times N$ matrix M, Which is called an eigenmatrix.

7. Jointly diagonalize the set N_e defined as $N_e = \{\lambda_r, M_r : 1 \le r \le N\}$ by a unitary matrix V.
   The diagonalization is done by a modified Jacobi Rotation technique.
8. The estimated mixing matrix is then

$$\hat{A} = W^{\#} V \quad (12)$$

9. By inversing $\hat{A}$, the estimated source signal will be

$$\hat{Y} = \hat{A}^{-1} X \quad (13)$$

3. SPECTRUM ESTIMATION

The Yule-Walker autoregressive (AR) method [5] is a parametric method that estimates the autocorrelation function to solve for the AR model parameters. The method is superior to DFT since it avoids noise and blocking effects. The AR parameters, a_ks, are estimated by minimisation of the residual signal, $e(n)$.

$$e(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k) \quad (14)$$

where p is the prediction order whic can be obtained in order to achieve the minima for $e(n)$. Durbin algorithm [3] efficiently estimates a set of coefficients based on calculation of autocorrelation matrix and zeroing the error partial differentiation respect to a_ks. The minimum value for prediction order p can be identified using an iterative procedure. The procedure sets p for having an error average below a low threshold level [6].

4. IMPLEMENTATION AND RESULTS

AR is preferred to FFT due to suppression of noise and blocking effect especially in the theta and beta sub-bands where the activity is low when compared to the delta and alpha subbands. BSS is initially applied to decorrelate the EEG signals. Application of BSS in processing of EEGs greatly changes the results. The spectrum of each signal is estimated using AR method. The spectrums are divided into four subbands referred earlier. The energy of each subband is measured by using the following equation.

$$E_b = \sum_{band\ b} |X(m)|^2, b = Delta, Theta, Alpha, Beta \quad (15)$$

These energy values are allocated to the actual geometrical positions of the electrodes in the model image as the amplitudes of those points. Then the amplitudes are extrapolated to a surface and hot-colors are used to highlight the levels of activity in these surfaces. Figure 2 and 3 illustrate the effect of BSS followed by AR on these images.

The regions of activity are changed after application of BSS. This is more or less expected as each EEG can be assumed to be a combination of the electrode and its adjacent pin signals including noise. The BSS algorithm will serve to deconvolve the desired electrode signal from

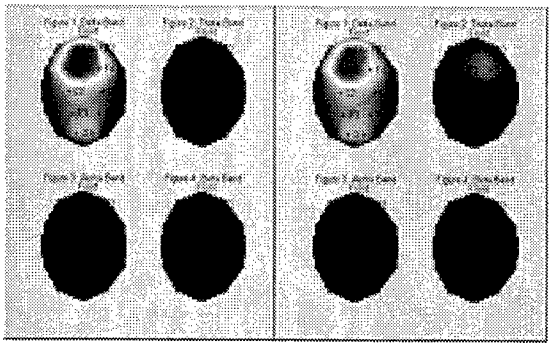


Figure 2: Reconstructed brain map of normal person using BSS (left) and without using BSS (right).

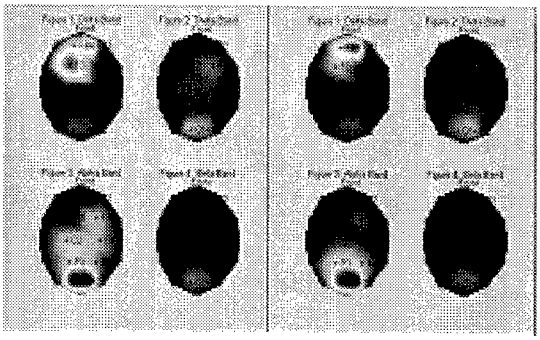


Figure 3: Reconstructed brain map of CJD patient using AR without BSS (left) with BSS (right).

the rest. In this case each signal carries the local information more accurately. This enhances visualisation of the activities too. A real time reconstruction of the maps will be more informative as it provides the functional information. Functional information introduces a remarkable time-frequency information about both rhythmic and arrhythmic brain abnormalities.

5. CONCLUSIONS

Application of BSS in reconstruction of brain maps from EEGs greatly enhances the neurodiagnostic information and localization of anatomical, functional and pathological abnormalities in the brain. JADE algorithm provides sufficient tools in estimation of number of sources and separation of EEGs. This obtains very accurate information whereas normal EEGs and brain maps usually fail in achieving that. Further work may be done to estimate the exact number of sources in the brain and adopt a more general BSS algorithm (handling non-linear and convolusive mixtures) to separate individual EEGs. The process can be extended to find the best location of the electrodes that inherently obtains the maximum decorrelation of the sources.

REFERNCES

1. Speekman E. J. and Elger C. E., "Introduction to the neurological basis of the EEG and DC Potentials", Electroencephalography, pages 4-9, April 1994
2. Thatcher RW, Hallett M, Zeffiro T, John ER, and Huerta M. *Functional Neuroimaging*. New York: Acedemic Press, 1994.
3. Access Health web page part Electroencephalogram (EEG) http://www.yourhealth.com/ahl/1537.html , 1996
4. Cardoso, J. Source separation using higher order moments. *Proc. Of International conference on Acoustics, Speech and Sig. Proc.*, 1989, pp. 2109-2112.
5. Ning T. and Bronzino D., "Autoregressive and bispectral analysis Techniques: EEG applications", IEEE Engineering in Medicine and Biology Magazine, pages 18-23, March 1990
6. Rabiner L. R. and Schafer R. W., Digital Processing of Speech Signals, prentice Hall, 1978

EXTRACTION OF SUPERIMPOSED EVOKED POTENTIALS BY COMBINATION OF INDEPENDENT COMPONENT ANALYSIS AND CUMULANT-BASED MATCHED FILTERING

A. Cichocki[1], *R. R. Gharieb*[1] *and N. Mourad*[2]

[1]Lab. for Advanced Brain Signal Processing
Brain Science Institute, Riken, Wako-Shi, Saitama, 351-0198, Japan
E-mail: rrgharieb@ieee.org, Fax: + 81-48-467-9694

[2] South-Valley Faculty of Engineering, Aswan, Egypt

ABSTRACT

A novel approach is proposed in order to efficiently separate mixed evoked potentials (EPs) presented simultaneously by different stimuli. This approach is developed as follows. We first apply a robust independent component analysis (ICA) approach to the observed sensor signals for the separation of the superimposed EP signals. Next, the desired EP components are estimated by matched-filtering the separated signals. Impulse response of such matched filter can be computed based on third-order cumulants of the filter input signal. Therefore, due to the tolerance of the third-order cumulants to both Gaussian and any symmetrical distributed non-Gaussian noise or interference, the filter impulse response will be matched with the desired signal alone. It is demonstrated by extensive computer simulations that applying the cumulant-based ICA and filtering improves dramatically the SNR of the final estimation of the EP components.

1. INTRODUCTION

The sensory brain evoked potentials (EPs) are electrical responses of the central nervous system to sensory stimuli applied in a controlled manner. The interest in these potentials arises from their utilization as clinical and research tools and for their contribution to the basic understanding of the functions of the brain [1]-[3], [5], [7]. Ensemble averaging and weighted ensemble averaging have been usually used to enhance the SNR [1]. Such techniques can be thought of as lowpass filtering of noise and a very large number of sweeps is required to obtain a suitable EP estimate. Wiener Filtering based techniques have also been extensively used for the enhancement and recovering of the EP [3], [7]. In one adaptive implementation of the Wiener filter, the noisy EPs are taken as the primary input while, the auxiliary reference input has been taken as constructed models of the EPs because the reference noise is in general not available. Various kinds of basis functions have been used to construct such models [4]. The performance of this approach is then dependent on how much the assumed model is close to the EP signal. In another approach, where multiple sweeps are available, the primary input is taken as the ensemble average while the reference input is taken as one sweep that is not included in the average, which keeps noise uncorrelation. Unfortunately, the Wiener filtering method deteriorates if both the signal and noise spectra are overlapped.

In the present work, separation of single trail EP components presented simultaneously by two or more different stimuli is considered. Because most blind signal separation techniques cannot handle additive noise, we propose to enhance the signal-to-noise ratio (SNR) of the independent components estimated by a robust (i.e., unbiased) ICA approach. This SNR enhancement is achieved by matched filtering the ICA output signals. Impulse response of such matched filter can be computed based on third-order cumulants of the filter input signal. Therefore, due to the tolerance of the third-order cumulants to both Gaussian and any symmetrical distributed non-Gaussian noise, the filter impulse response will be matched with the desired signal alone. In Section II we formulate the problem, give a brief review for the blind signal separation and the cumulant-based filtering approach. In Section III, the proposed approach is described. Section IV presents extensive simulation results and finally Section V gives the conclusions.

2. CONVENTIONAL METHODS

2.1. Signal Model and Problem Formulation

Multiple m observations of EP signals $x=[x_1\, x_2 \ldots x_m]^T$ can be modeled as a mixed of independent signals plus noise given by

$$x = As + v \tag{1}$$

where A is an $m\times n$ full-column rank matrix with $m>n$, $s=[s_1\, s_2 \ldots s_n]^T$ is an $n\times 1$ vector gathering the independent EP sources, and v is an $m\times 1$ vector for additive noise representing ongoing EEG of brain activity. We assume that the EP signals have no zero third-order correlations. The objective is to estimate the independent sources s_i given the observed noise mixed signals x, i.e., to find a separating matrix W so that

$$y = \hat{s} = Wx = WAs = PDs \tag{2}$$

where P is any permutation matrix and D is a diagonal scaling matrix.

It is often that blind signal separation methods assume noise free mixed observations, i.e., $v=0$ or negligible small. The challenging is then to achieve robust separation and to reduce noise either. In estimating the separating W, some separation

0-7803-7011-2/01/$10.00 ©2001 IEEE

methods are robust to noise, however, because the noise-free sources cannot directly be obtained based on W and the observed signals x.

In this paper our proposal is first to apply robust ICA and next to filter the independent components. In the last stage we can reconstruct clean or corrected sensor signals by using back projection $\hat{x}_i = W^{-1}\hat{S}_i$ (one by one).

2.2. Robust ICA in Gaussian noise

There are several efficient batch algorithms, which are theoretically insensitive to additive noise (if the number of available samples is sufficient large) including JADE (Joint Approximation Digitalization) [8], ROSBI (Robust Second Order Blind Identification) [9] and ERICA (Equivarent Robust ICA) [10], [11]. In this paper, we use family of ERICA algorithms based on third and/or fourth-order matrix cumulants. The ERICA algorithm developed by Cruces et al. [10], [11] estimates (unbiased in respect to Gaussian noise) the separating matrix $W(\ell+1)$ as follows

$$W(\ell+1)=W(\ell)+\eta(\ell)[I-C_{1,\beta}(y,y)Sig_y]W(\ell) \qquad (3)$$

where ℓ is the number of iteration or alternatively for prewhitened (sphere) data (using for example SVD or factor analysis) as

$$W(\ell+1)=W(\ell)+\eta(\ell)[Sig_y C_{\beta,1}(y,y) - C_{1,\beta}(y,y)Sig_y]W(\ell) \qquad (4)$$

where $c_\alpha(y_i)$ denotes the $\alpha-$order cumulant of the separated signals y_i, $C_{\alpha,\beta}(y,y)$ denotes a matrix cumulant whose (i,j)th element is given by the cross-cumulant function $c_{\alpha,\beta}(y_i,y_j)=Cum(\underbrace{y_i,...,y_i}_{\alpha},\underbrace{y_j,...,y_j}_{\beta})$ and Sig_y will be a short-hand notation to refer to the diagonal matrix containing the signs of the diagonal cumulants $Sig_y = diag(sign(diag(C_{1,\beta}(y,y))))$. For example, for the fourth-order matrix cumulants ($\beta=3$) $[Sig_y]_{ii}=[sign(Cum_4(y_i))]_{ii}$ (the actual kurtosis signs of the output signals) and $C_{1,3}(y,y)$ is the fourth-order cross-cumulant with elements $[C_{1,3}(y,y)]_{ij}=Cum(y_i,y_j,y_j,y_j)$; analogously is defined the cross-cumulant matrix $C_{3,1}(y,y)=(C_{1,3}(y,y))^T$.

Since the third-order cumulants are insensitive to additive Gaussian noise and any symmetrical distributed noise, it is recommended to use the third-order cumulants rather than the fourth-order cumulants.

2.3. Cumulant-Based Filtering

In [12], we have shown that it is possible based on third-order cumulants to design an FIR filter that is matched with the desired EP potential signal. Let the ith signal $x_i(n)$ be an input of such matched filter, the filter impulse response is taken as a replica of a selected one-dimensional third-order cumulant slice of $y_i(k)$. That is, $h_i(j)$ is given by [12]

$$h_i(j)=\begin{cases}\hat{c}_{y_i}(J-j), & j=0,1,....,J\\ \hat{c}_{y_i}(j-J), & j=J+1,J+2,...,2J\end{cases} \qquad (5)$$

where

$$\hat{c}_{y_i}(j)=\frac{1}{K}\sum_{k=0}^{K-1} y_i(k)y_i(k+j)y_i(k+\tau) \qquad (6)$$

where K is the number of available time samples and $\tau\geq 0$ is a positive time shift. The filter output, the enhanced version of the ith filter input signal $\hat{y}_i(k)$ is the convolution sum of the input and the impulse response given by

$$\hat{y}_i(n)=\gamma_i\sum_{j=0}^{2J} h_i(j)y_i(k-j) \qquad (7)$$

where γ_i is a constant chosen so as to provide unity Skewness gain for the filter.

3. THE PROPOSED APPROACH

The proposed approach is developed as follows. We first apply the robust ICA approach described in Subsection 2.2 to the observed sensor signals for the separation of the superimposed EP potentials. Unfortunately, although the robust ICA is capable of separating the independent components it is not capable of reducing additive noise. Therefore the idea is to enhance the SNR of the ICA outputs. This SNR enhancement is achieved by passing the ICA outputs through a bank of cumulant-based FIR filters of type described in Subsection 2.3. In this case the output $y_i(n)$ is taken as the input of the ith filter. The advantage of the filtering technique arises from the fact that the third-order cumulant of the evoked potential (modeled as a sum of damped sinusoidal signals) preserves the evoked potential structure in addition to its tolerance to additive Gaussian noise and other symmetrical distributed non Gaussian noise [12].

4. SIMULATION RESULTS

To examine the effectiveness of the proposed approach, extensive simulations have been carried out. Due to space limitations we present here only one illustrative example. In this example, two simulated models for the EP signals are given in Figure 1. Four observed evoked potentials are obtained using a mixing matrix of 4×2 random variables. The maximum of absolute value of each column is adjusted to unity. To generate additive colored Gaussian noise in order to simulate ongoing EEG, the spectrum of each sensor noise is generated using a moving average system of order 65. This moving average is computed as follows. First, a deterministic moving average is considered as the coefficients of a Hamming window finite impulse response (FIR) filter whose normalized frequency bandwidth is [0.05 0.1]. Next for each sensor we add to the deterministic impulse response additive Gaussian random variables so that the variance of the deterministic impulse response to this random variables is 10. The signal-noise-ration defined as the power of each observed signal to the additive noise is adjusted to 0.0 and −10.0 dBs. The independent component analysis procedure is first applied to the observed sensor (superimposed) signals. Next, the

cumulant-based FIR filter of order 32 is applied to the each independent component in order to obtain the desired decomposed EP signals.

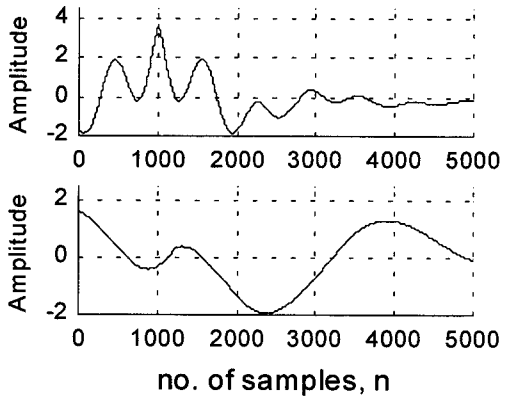


Figure 1. Two signal models for evoked potentials.

Results when SNR=0.0 dB have shown the ability of the proposed approach in separating the two EP signal models. The cumulant-based filtering approach has also reduced dramatically the additive noise, which confirms the beneficial of the filtering technique to improve the overall performance of the presented approach. Illustration Figures for the SNR=0.0 dB have been omitted for space limitation. Figure 2 shows the results when the SNR= –10.0 dB. Figure 2(a) shows the observed sensor signals. Figure 2(b) shows the independent components obtained using the ICA. Figure 2(c) shows the filtered version of the ICA outputs. It is apparent the filtered version of the independent components represent the desired EP signals. It can be mentioned that the final cumulant-based filter is necessary to reduce additive noise corrupting the desired EP signals.

Experimental Results- To examine the proposed approach for the real world data, simultaneous visual and auditory EPs from a male subject have been recorded using 64 electrodes EEG system. The visual stimulus was flash-like white circle in a black background that appears for 50 *msec* with rate 1/*sec*. The auditory stimulus was a click with width 50 *msec* modulated by a 1 KHz single tone with rate also 1/*sec*. The visual stimulus takes place first for 50 *msec* and then the auditory stimulus. The recorded data were filtered using 0.05-250 Hz bandpass filter and sampled using a sampling rate of 2000 Hz. In a single trial we recorded 2000 samples (1 *sec*) after the visual stimulus by electrodes 02, 01, 0Z, FC1, F2 and F1. The reference was the left mastoid. Figure 3(a) shows the raw data for 500 *msec* after the visual stimulus. Figure 3(b) shows the separated signals using the ICA. It is obvious that two evoked potentials appear. The P100 of the visual and the N100 of the auditory EPs are clearly observed. The time delay between both peaks is about 50 *msec*. After cumulant-based matched filtering SNR enhancement has been obtained as shown in Figure 3(c).

5. CONCLUSION

A novel approach has been described for the extraction of superimposed single trial evoked potentials components presented simultaneously in relative or short duration between them by different stimuli. In this approach, we first apply a robust independent component analysis (ICA) approach to the observed sensor signals for the separation of the superimposed EP components. Next, the desired EP components are estimated by FIR matched-filtering of noise corrupting the separated signals. The impulse response of this FIR filter is computed based on third-order cumulants of the filter input signal. Therefore, due to the tolerance of the third-order cumulants to any symmetrical distributed noise (white or colored), the proposed approach is capable of extracting EP source signals for very low SNR. Pre-simulation results have confirmed the efficiency of the presented approach.

6. REFERENCES

[1] C. E. Davila and M. S. Mobin, "Weighted averaging of evoked potentials," *IEEE Trans. Biomed. Eng.*, vol. 39, pp. 338-347, April 1992.

[2] P. A. Karjalainen, J. P. Kaipio, A. S. Koistinen and M. Vuhkonen, "Subspace regularization method for the single-trial estimation of evoked potentials," *IEEE Trans. Biomed. Eng.*, vol. 40, pp. 849-860, July 1999.

[3] P. G. Madhavan, "Minimal repletion evoked potentials by modified adaptive line enhancement," *IEEE Trans. Biomed. Eng.*, vol. 39, pp. 760-764, July 1992.

[4] P. Laguna *et al.*, "Adaptive filter for event-related bioelectrical signals using an impulse correlated reference input: comparison with signal averaging techniques," *IEEE Trans. Biomed. Eng.*, vol. 39, pp. 1032-1044, Oct. 1992.

[5] F.K. Lam, F.H. Y. Chan and P.W.F. Poon, "Visual evoked potential measurement by adaptive filtering," *Bio-Medical Materials and Engineering*, vol. 4, no. 6, pp. 409-417, 1994.

[6] T. Kobayashi and S. Kuriki, "Principle component elimination method for the improvement of S/N in evoked neuromagnetic field measurements," *IEEE Trans. Biomed. Eng.*, vol. 46, pp. 951-958, Aug. 1999.

[7] X. Yu, Z. He and Y. Zhang, "Time-varying adaptive filters for evoke potential estimation," IEEE Trans. Biomed. Eng., vol. 41, pp. 1062-1071, Nov. 1994.

[8] J-F. Cardoso and A. Souloumiac, "Blind beamforming for nonGaussian signals," *IEE Proceedings-F*, vol. 140, pp. 362-370, Dec. 1993.

[9] A. Belouchrani and A. Cichocki, "Robust whitening procedure in blind source separation context," *Electronics Letters*, vol. 36, pp. 2050-2053, 2000.

[10] S. Cruces, A. Cichocki and L. Castedo, "Blind source extraction in Gaussian noise," Proceeding of the Second International Workshop on ICA and BSS, ICA'2000, Helsinki, Finland, June 19-22, 2000, pp. 63-68.

[11] S. Cruces, L. Castedo and A. Cichocki, "Novel blind source separation using cumulant," ICASSP'2000 Istanbul, Turkey, June 2000, pp. 3152-3155.

[12] R. R. Gharieb and A. Cichocki," Noise reduction in brain evoked potentials based on third-order correlations," *IEEE Trans. Biomd. Eng*, pp. 501-512, May 2001.

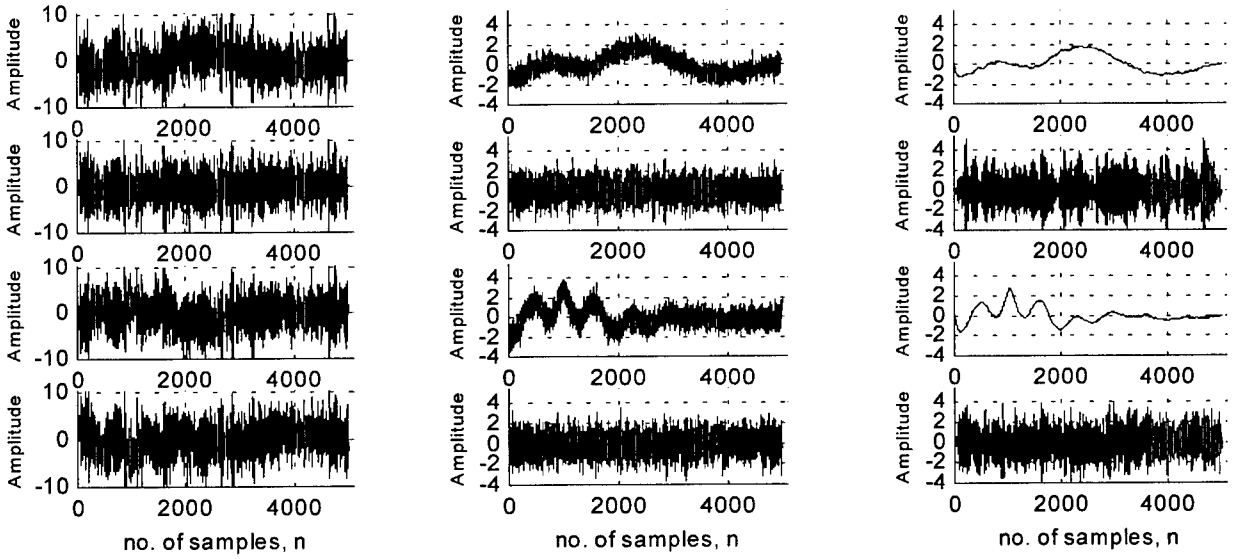


Figure 2. Results for the –10.0 dB SNR example: (a). the observed sensor signals: (b). independent (separated) signals; (c). the filtered version of the independent signals in (b).

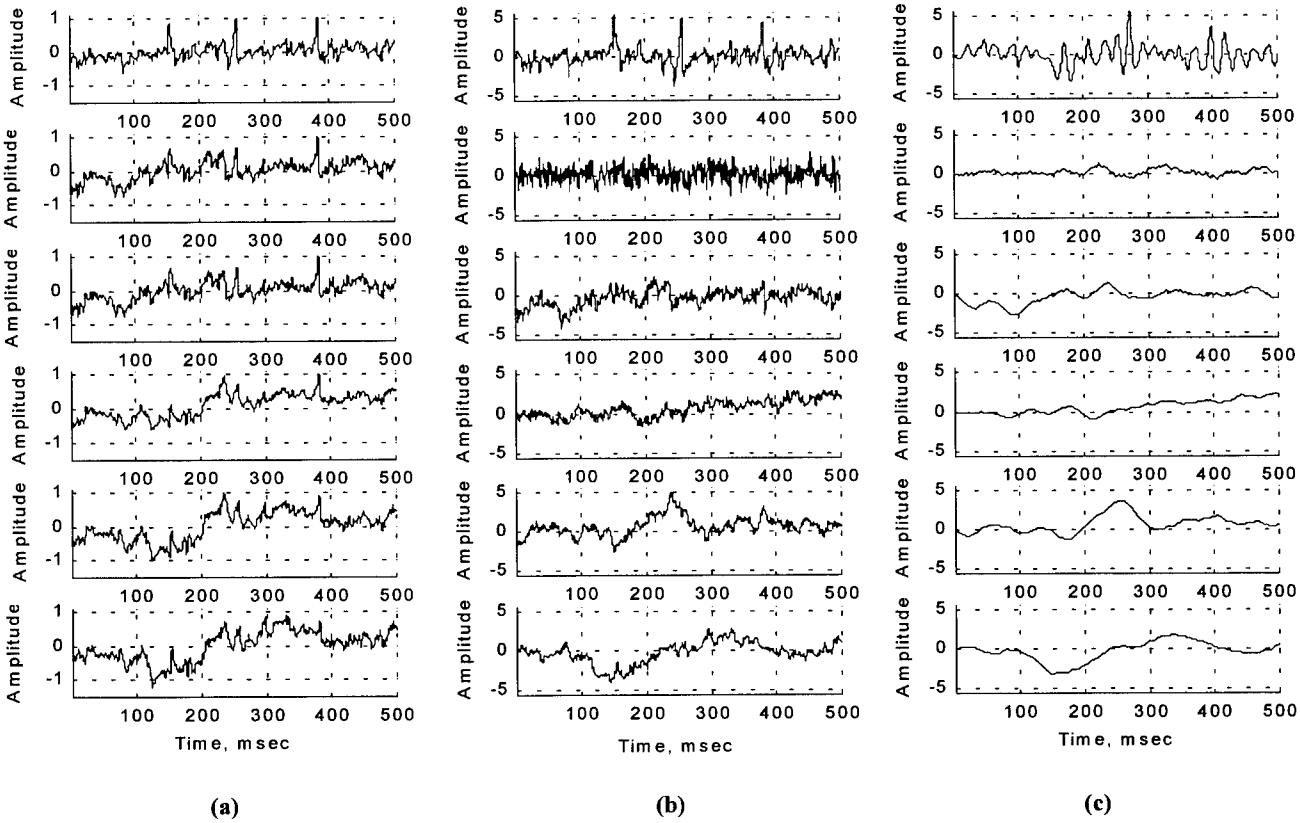


Figure 3. Results of the real world data: (a). the observed raw sensor signals: (b). independent (separated) signals; (c). the filtered version of the independent signals in (b).

ESTIMATING THE DYNAMICS OF ABERRATION COMPONENTS IN THE HUMAN EYE

D. Robert Iskander[1], *Mark R. Morelande*[2], *and Michael J. Collins*[2]

[1] School of Engineering, Griffith University,
PMB 50, Gold Coast, Q9726, Australia
[2] Centre for Eye Research, Queensland University of Technology
Victoria Park Road, Kelvin Grove, Q4059, Australia

ABSTRACT

To provide adequate information that would assist surgeons in performing advanced refractive corrections, it is essential to address the problem of microfluctuations in the eye's aberrations due to pulse and respiration. Although the effects of fluctuations in defocus are known and well described, very little is reported on modelling the fluctuations in other types of aberrations. We propose a methodology in which the dynamics of higher order aberration components are modelled by parametric AM-FM signals. Using our modelling approach, the effects of changes in these aberrations could be predicted and studied. In particular, we model the dynamics of components related to coma and spherical aberration. We provide a validation of the proposed modelling approach using aberration data from the eyes of six subjects.

1. INTRODUCTION

The interest in customised refractive surgery by optometrists and ophthalmologists as well as the patients is growing significantly. It is predicted that these advanced customised refractive surgeries will correct many aberrations of the eye, providing vision levels that currently cannot be achieved [1]. Such vision would be limited only by the resolution of the retinal photoreceptors and diffraction due to the pupil aperture [2]. However, these modern surgical procedures depend to a great extent on advances in eye measurement systems, such as wavefront sensors. One of the current problems in such systems is that the dynamics of the eye's aberrations, due to pulse and respiration [3, 4], are not taken into account.

The aberrations of the eye are usually described in terms of a scaled optical path difference between a ray passing through the optical system of the eye at a certain point in the pupil and the principal ray. This scaled optical path difference is referred to as the wavefront aberration or wavefront error.

The wavefront error of an eye can be measured with a Hartmann-Shack sensor [5]. It is an optical instrument equipped with a laser, an array of small lenses, and a CCD video camera. The light reflects from the retina, passes through the array of lenses, and forms a grid image that falls on the CCD. The displacements in the grid image, from the ideal square grid, are used to calculate transversal aberrations which are related to the wavefront error.

Wavefront error is often given a functional form in terms of basis functions. The discrete wavefront aberration data, denoted in polar coordinates as $W(r_d, \theta_d)$ can be modelled by a finite series of discrete basis functions

$$W(r_d, \theta_d; n) = \sum_{p=1}^{P} z_p(n) \Phi_p(r_d, \theta_d) + \varepsilon_d, \qquad (1)$$

where $z_p(n)$, $n = 0, \ldots, N-1$ are the time-varying aberration coefficients, $\Phi_p(r_d, \theta_d)$ is the p-th discrete basis function sampled from $\Phi_p(r, \theta)$ at discrete points $d = 1, \ldots, D$ and ε_d denotes the measurement noise. In some cases, such sampling may require further orthogonalisation using the Gram-Schmidt procedure. The most popular basis functions amongst vision researcher are the Zernike polynomials [6, 7], because each of the terms in the expansion can be related to a particular type of aberration. For example, the fourth term corresponds to defocus, the fifth and sixth to astigmatism, the seventh and eight to coma, and the 11th relates to spherical aberration. In most classifications, the first six Zernike terms are assigned as lower-order terms since they can be corrected with traditional spectacles.

In this work, we focus on dynamics of higher order aberrations, and in particular, of coma and spherical aberration. The values of the coefficients associated with each term vary in time [4], due primarily to changes in accommodation (focusing) which, in turn, are affected by pulse and respiration [3]. In order to predict and study the effects of dynamics of the eye's aberrations, it is desired to develop appropriate parametric models for the dynamics of each of the aberrations.

The paper is organised as follows. In the next section, we provide an overview of the protocol of aberration data acquisition. In Section 3, we describe the model for the

0-7803-7011-2/01/$10.00 ©2001 IEEE

components of the higher order aberrations. This is followed by the experimental results given in Section 4.

2. DATA ACQUISITION

A custom made Hartmann-Shack sensor was used for measuring the aberrations of the optical system of the eye. Six subjects were used in the study. For each subject, a series of grid images sampled at 10 Hz were taken within a period of 5 seconds. An example of typical grid image is shown in Figure 1. The sampling frequency was chosen well above the Nyquist rate for signals that exist in the cardiopulmonary system [3]. The limit of 50 images was due to the physical capacity of computer memory.

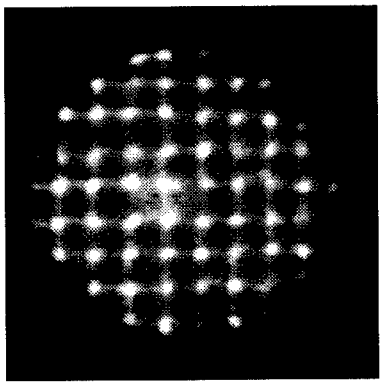


Fig. 1. Typical grid image of the Hartmann-Shack sensor.

The six considered subjects were aged between 20 and 30 years and had normal healthy eyes. They were asked to focus on the instrument's fixation target. All subjects were optically corrected for lower order aberrations with spectacle corrections. Ten series of 50 grid images where recorded for each subject. There was no blinking during the acquisition of each series. The study met the requirements of the university Human Research Ethics Committee.

For each grid image, a centroid detection algorithm was used to determine the transversal aberrations and wavefront slopes. Then, from this information, the wavefront error was derived. A series of the first 15 Zernike polynomials was then fitted to each wavefront error resulting in a 50 data point time-series for each type of aberrations. In the following, we focus on higher order terms since the first six Zernike terms correspond to aberrations that were optically corrected.

3. MODELLING APPROACH

Initially, we have performed a time, frequency, and time-frequency analysis of the data for each of the higher order aberrations. The time-frequency analysis has indicated that

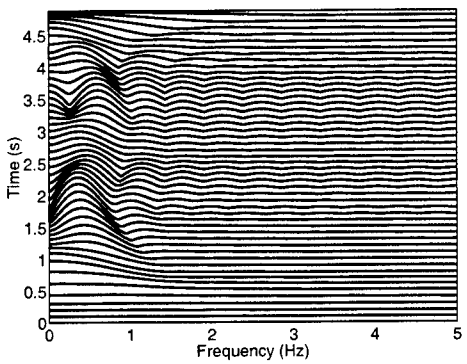


Fig. 2. The spectrogram of the dynamics measured for the eighth Zernike coefficient corresponding to horizontal coma.

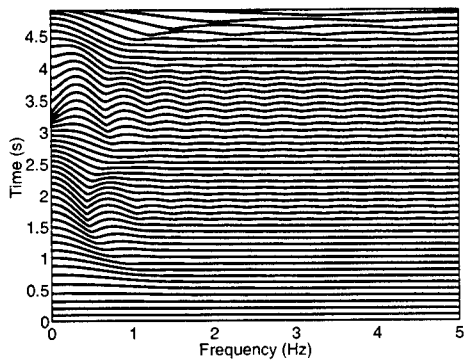


Fig. 3. The spectrogram of the dynamics measured for the eleventh Zernike coefficient corresponding to spherical aberration.

each of the Zernike coefficients can be modelled as

$$z_p(n) = \sum_{l=1}^{L} a_{p,l}(n), \qquad n = 0, \ldots, N-1,$$

where $a_{p,l}(n)$, $l = 1, \ldots, L$ denote aberration components that are well-separated in frequency. This allows us to band-pass filter each aberration to extract the components of interest corresponding to pulse and respiration.

We have also observed that the spectral characteristics of the aberration components $a_{p,l}(n)$ vary in time. This is demonstrated in Figures 2 and 3 which show the spectrogram of the horizontal coma and spherical aberration, respectively, for a subject with a significant amount of astigmatism. The non-stationary characteristics of the aberration components are clearly evident.

Motivated by these results, we propose to model each of the aberration components as an amplitude modulated-frequency modulated (AM-FM) signal. In particular, we

consider the following parametric model:

$$a_{p,l}(n) = \left(\sum_{m=0}^{M} g_m (n/N)^m\right) \cos\left(\sum_{q=0}^{Q} b_q n^q\right), \quad (2)$$

$n = 0, \ldots, N-1$. We form the analytic signal

$$c_{p,l}(n) = a_{p,l}(n) + j\mathcal{H}\{a_{p,l}(n)\}$$

which, under the assumption that the amplitude is low-pass, can be approximated by

$$c_{p,l}(n) \approx \left(\sum_{m=0}^{M} g_m (n/N)^m\right) \exp\left(j\sum_{q=0}^{Q} b_q n^q\right),$$

$n = 0, \ldots, N-1$. Estimation of the model parameters is performed in two steps. First, we perform order selection to estimate appropriate values of M and Q. We have experimented with two methods for model order selection. One procedure, described in [8] is based on a multiple hypothesis testing. The other procedure is based on the bootstrap [9]. As an example, we show the application of the parametric modelling procedure of [8] to the spherical aberration component shown in Figure 3. A significance level of 1 % was used in the order selection procedure which gave $M = 14$ and $Q = 4$. Using these orders, the fitted and the values estimated non-parametrically of the amplitude and phase are plotted against time in Figures 4 and 5.

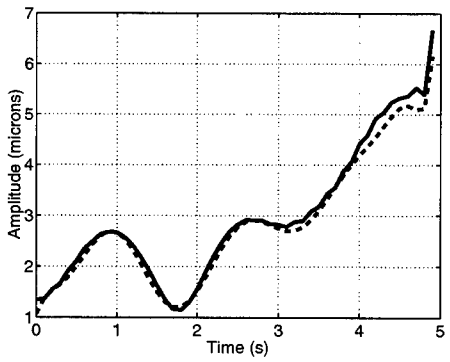


Fig. 4. Parametric modelling of the amplitude of the spherical aberration component. The non-parametric estimate (solid) and the fitted values (dashed) are shown.

Similar orders for the amplitude and phase were obtained with the bootstrap method. In this case, we find estimates $\hat{g}_0, \ldots, \hat{g}_M$, $\hat{b}_0, \ldots, \hat{b}_Q$ of the amplitude and phase parameters using linear regressions. It can be seen that the amplitude model provides a close fit to the actual amplitude while the phase model is reasonably close to the actual phase. Clearly, the AM-FM signal could be a valid model for the spherical aberration component. Validation of the proposed model is given in the next section.

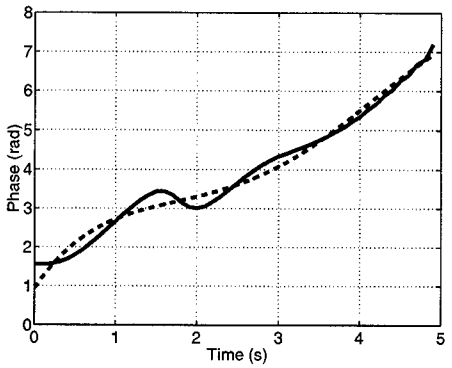


Fig. 5. Parametric modelling of the phase of the spherical aberration component. The non-parametric estimate (solid) and the fitted values (dashed) are shown.

4. EXPERIMENTAL RESULTS

As mentioned in Section 2, six subjects were used in the study. The best seven out of 10 data acquisitions have been selected for the analysis. In this way, we have ensured that the same size of grid was fitted to each grid image for calculation of the wavefront error. The variations in the grid size between the images were due to changes in the pupil size. Three higher order terms of the Zernike expansion were analysed. These were the horizontal coma, the vertical coma, and the spherical aberration. As noted before, the subjects were corrected for lower-order aberrations by wearing their usual spectacles.

For each of the terms, a low frequency single component was extracted by using a 1.5 Hz FIR low-pass filter of order 10. This was performed first manually on a number of sequences and then automated. All the data were detrended before filtering. Then, model order selection procedures were used to determine the parameter Q and M. It has been found that the procedure based on the bootstrap is more robust than the one based on multiple hypothesis testing. This is mainly due to the bootstrap ability to find a correct model for short data samples [9]. For each model a mean-square error (MSE) of the fit was calculated and then averaged across the seven records.

It has been observed that no single model can be fitted to the components of coma and the spherical aberration. We found that the model order of the phase ranges from $Q = 3$ to $Q = 5$. However, isolated cases in which both model selection procedures were giving higher orders for the phase were observed. A wider range of model order were recorded for the amplitudes (from $M = 8$ to $M = 20$). In Tables 1–3, we show the average standardised MSE for the amplitude and phase of the components associated with the coma the spherical aberration.

The results indicate that the dynamics of the considered

Table 1. Average values of the standardised MSE for the amplitude (top) and phase (bottom) of the components corresponding to vertical coma.

			M		
Q	8	11	14	17	20
3	0.0541	0.0249	0.0134	0.0072	0.0030
4	0.0504	0.0240	0.0131	0.0071	0.0030
5	0.0496	0.0253	0.0137	0.0073	0.0033
3	0.0047	0.0047	0.0047	0.0047	0.0047
4	0.0036	0.0036	0.0036	0.0036	0.0036
5	0.0028	0.0028	0.0028	0.0028	0.0028

Table 2. Average values of the standardised MSE for the amplitude (top) and phase (bottom) of the components corresponding to horizontal coma.

			M		
Q	8	11	14	17	20
3	0.0632	0.0300	0.0145	0.0060	0.0024
4	0.0528	0.0279	0.0135	0.0059	0.0024
5	0.0509	0.0253	0.0121	0.0060	0.0026
3	0.0037	0.0037	0.0037	0.0037	0.0037
4	0.0029	0.0029	0.0029	0.0029	0.0029
5	0.0024	0.0024	0.0024	0.0024	0.0024

Table 3. Average values of the standardised MSE for the amplitude (top) and phase (bottom) of the components corresponding to spherical aberration.

			M		
Q	8	11	14	17	20
3	0.0566	0.0271	0.0117	0.0052	0.0023
4	0.0512	0.0251	0.0116	0.0055	0.0023
5	0.0467	0.0233	0.0110	0.0056	0.0023
3	0.0046	0.0046	0.0046	0.0046	0.0046
4	0.0035	0.0035	0.0035	0.0035	0.0035
5	0.0028	0.0028	0.0028	0.0028	0.0028

aberration components can be well modelled by a polynomial phase signal of order less than $Q = 5$. Small variations in the average standardised MSE for the amplitude fit were observed. However, in the interest of parsimony, we would like to avoid choosing a model order that results in the number of parameters being a substantial fraction of the sample size. Therefore, we choose $M = 11$ which results in a fit whose average standardised MSE does not exceed 3%. It is possible that choosing a different set of basis functions would result in a lower order model.

5. CONCLUSIONS

We have addressed the problem of modelling the dynamic changes in the components of higher order aberrations in the human eye. We proposed a parametric AM-FM signal model for these dynamics. Although the data samples were short, we have shown that the proposed modelling approach is viable and could be used for prediction and study of higher order aberrations. Our methodology can be used in the design of protocols that deal with the measurement of aberrations in the human eye and take into account the dynamic characteristics of these aberrations.

6. REFERENCES

[1] C. Roberts, "Future challenges to aberration-free ablative procedures," *J. Refract. Surg.*, vol. 16, no. S, pp. 623–629, 2000.

[2] J. Liang, D. R. Williams, and D. T. Miller, "Supernormal vision and high-resolution retinal imaging through adaptive optics," *J. Opt. Soc. Am. A*, vol. 14, no. 11, pp. 2884–2892, 1997.

[3] M. J. Collins, B. Davis, and J. Wood, "Microfluctuations of steady-state accommodation and the cardiopulmonary system," *Vis. Res.*, vol. 35, pp. 2491–2502, 1995.

[4] H. Hofer, P. Artal, B. Singer, J. L. Aragon, and D. R. Williams, "Dynamics of the eye's aberration," *J. Opt. Soc. Am. A*, vol. 18, no. 3, pp. 497–506, 2001.

[5] J. Liang, B. Grimm, S. Goelz, and J. F. Bille, "Objective measurement of wave aberrations of the human eye with the use of a Hartmann-Shack wave-front sensor," *J. Opt. Soc. Am. A*, vol. 11, no. 7, pp. 1949–1957, 1994.

[6] L. N. Thibos, R. A. Applegate, J. T. Schwiegerling, R. Webb, and VSIA Standard Taskforce Members, "Standards for reporting the optical aberration of eye," www.osa.org/Homes/vision/resourses/intro.htm, 2000.

[7] D. R. Iskander, M. J. Collins, and B. Davis, "Optimal modeling of corneal surfaces with Zernike polynomials," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 1, pp. 87–95, 2000.

[8] M. R. Morelande, *Estimation, detection and model selection of random amplitude polynomial phase signals*, Ph.D. thesis, Curtin University of Technology, Perth, Australia, 2000.

[9] Zoubir, A. M. and Iskander, D. R.", Bootstrap modeling of a class of nonstationary signals, *IEEE Trans. Sig. Proc.*, vol. 48, no. 2, pp. 399–408, 2000.

A TIME-VARYING MODEL FOR DNA SEQUENCING DATA

Nicholas M. Haan and Simon J. Godsill

Signal Processing Group
Department of Engineering, University of Cambridge, U.K.
email: nmh28@cam.ac.uk

ABSTRACT

Methods for determining the letters of our genetic code, known as DNA sequencing, currently depend on clever use of electrophoresis to generate data sets indicative of the underlying sequence. Typically the subsequent off-line data processing is carried out using a combination of heuristic methods with little mathematical rigour. In this paper, we present a novel model which is able to accurately predict the effect of the many biological processes which are involved, and moreover, which is usable on-line. Off-line methods have been hampered by the need for processing in as little time as possible after the data is generated; performing the processing on-line has enabled a more advanced algorithm to be used with associated improved performance. The algorithm is framed within a Bayesian probabilistic framework, thereby allowing representation of the random nature of the generative process, and relies on new advances in the burgeoning field of Sequential Monte Carlo Methods to perform non-linear filtering and model selection operations.

1. INTRODUCTION

In the majority of living organisms, genetic information is encoded using a molecule known as Deoxyribonucleic acid (DNA) which may, for our purposes, be thought of as a sequence of chemical bases taken from a possible set of four: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). The determination of the genetic code, known as DNA sequencing, is important if genetic disease is to be properly understood.

In 1974, Sanger proposed a method for DNA sequencing which, with technical improvements, has since been almost universally accepted [10]. A simplified version of the Sanger sequencing process is now presented; for a more complete treatment see [4]. Initially, via a process of replication and truncation the DNA sequence of interest is used to form a large population of partial replicas. Each replica is identical to the sequence of interest over a range of bases, always commencing with the first base of the initial sequence, and terminating some random distance down the strand. That is, for the sequence ACGGG the population would contain a number of each of the following: A, AC, ACG, ACGG, and ACGGG. Each fragment is fluorescently labelled according to its terminating base. Subsequently, the entire population is aligned at the start of a large rectangular gel, and an electric field is applied. The fragments progress through the gel at rates approximately inversely proportional to their length, resulting in the various subpopulations arriving at the end of the gel in sequence order. A laser positioned near the end of the gel excites the fluorescent labels, allowing an emission detector to estimate the number of fragments terminated by a given base passing at each time instant.

After some preprocessing, four data sets are obtained (henceforth, *channels*), corresponding to the variation of fragment concentration with time for each of the four terminating bases. This collection of data is known as an electropherogram and is quite clearly indicative of the underlying base sequence; an example data set is shown in figure 2. The electropherogram is a mixture of peaks in four channels, with each base in the sequence associated with one major peak in the corresponding channel, and three secondary peaks in the remaining channels which result from leakage effects; the peaks corresponding to a particular base have common position and shape. A range of prior information, mainly detailing the effect of base sequence on the amplitudes and positions of the peaks, is available to constrain the problem; [11] provides a good review.

The current state-of-the-art from an off-line signal processing perspective, Phred, is described in [4], and uses a combination of heuristic, but effective, peak detection algorithms. In [5], an alternative block-based algorithm based on statistical modelling of the underlying process is proposed that outperforms Phred on some datasets, although at greater computational cost.

Here, we present a model similar to that of [5,6], which is capable of representing available prior information about the system. The major improvements of the model are that it allows removal of slowly varying background noise, and is also able to track nonstationarity in the various processes. In particular, variation in the spacing of peaks across the electropherogram is properly modelled, thereby reducing inferential ambiguity in more difficult data regions. The resulting algorithm can be run on-line and has immediate application to all data sets which comprise a series of peaks arriving sequentially in time (as encountered, for example, in some spectroscopy applications).

2. PROBLEM FORMULATION

As mentioned in the introduction, each fragment population generates peaks in all four channels as it passes the end of the gel. These are observed in a combination of slowly-varying and approximately i.i.d. white noise, suggesting the following general model for electropherogram data in the four channels at time $n \in \{1, \ldots, N\}$, $\mathbf{y}_n \triangleq$

0-7803-7011-2/01/$10.00 ©2001 IEEE

$\{y_{n,A}, y_{n,G}, y_{n,C}, y_{n,T}\}$:

$$\mathbf{y}_n = \mathbf{e}_n + \mathbf{t}_n + \sum_{i=1}^{k} a_i \boldsymbol{\omega}_i \phi_i(n)$$

where k denotes the total number of bases in the sequence, a_i is representative of the number of fragments in the population corresponding to the ith base in the sequence, $\mathbf{t}_n \triangleq \{t_{n,A}, \ldots, t_{n,T}\}$ denotes a background trend in the four channels, $\boldsymbol{\omega}_i \triangleq \{\omega_{i,A}, \omega_{i,G}, \omega_{i,C}, \omega_{i,T}\}$ is a vector defining the emission spectrum in the four channels, and $\phi_i(n)$ defines the peak shape. The uncorrelated noise in the system at time n, $\mathbf{e}_n \triangleq \{e_{n,A}, \ldots, e_{n,T}\}$, is assumed Normally distributed, $\mathbf{e}_n \sim \mathcal{N}(\mathbf{e}_n|\mathbf{0}, \sigma_\epsilon^2 \mathrm{I}_{4\times 4})$, where $\mathrm{I}_{4\times 4}$ denotes the identity matrix.

Here, the peaks are assumed truncated Gaussian in shape such that:

$$\phi_i(n) = (2\pi v_i)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2v_i}(n-p_i)^2\right\} \mathbb{I}(|n-p_i| \leq \epsilon)$$

where v_i denotes the *variance* of peak i, and p_i denotes its position. ϵ is defined a-priori to be sufficiently large that for the range of possible variances, the truncation effect is minimal.

The *base-state* of the system at base position i, $\mathbf{s}_i \triangleq \left\{s_{i,-2}, s_{i,-1}, s_{i,0};\ s_{i,j} \in \mathcal{B} \triangleq \{A,G,C,T\}\right\}$ is defined as a base triplet, e.g., $\{A,G,C\}$, with the last element corresponding to the ith base in the sequence, and the first two elements containing the previous two bases in the sequence (these are included to account for sequence dependent effects, see [8]). The state is assumed to evolve in a Markovian fashion:

$$s_{i,0} \sim p(s_{i,0}|s_{i-1,0}), \quad \mathbf{s}_{i,-2:-1} = \mathbf{s}_{i-1,0:-1}$$

where henceforth, the notation $\mathbf{r}_{i,j} \triangleq \{r_i, \ldots, r_j\}$ is used.

The peak position prior attempts to incorporate the idea that the mean peak spacing between bases $i-1$ and i, $\mu_{p,i-1}$, varies slowly across the electropherogram, while local peak jitter, represented by σ_p^2, can be substantially larger. Moreover, $\gamma_p(\mathbf{s}_i)$ is included to represent special sequence dependent effects.

$$\mu_{p,i-1} \sim \mathcal{N}(\mu_{p,i-2}, \sigma_{\mu_p}^2)$$
$$p_i \sim \mathcal{N}(p_{i-1} + \gamma_p(\mathbf{s}_i)\mu_{p,i-1}, \sigma_p^2)$$

The amplitude process similarly has a slowly varying mean, $\boldsymbol{\mu}_{a,i}$, with possibly substantial local jitter represented by σ_a^2:

$$\boldsymbol{\mu}_{a,i} \sim \prod_{j\in\mathcal{B}} \mathcal{N}(\mu_{a,i-1,j}, \sigma_{\mu_a}^2),$$
$$a_i \sim \mathcal{N}(\gamma_a(\mathbf{s}_i)\mu_{a,i,s_{i,0}}, \sigma_a^2)$$

where $\mu_{a,i,s_{i,0}}$ denotes the mean expected from a base of type $s_{i,0}$, with the effect of surrounding base sequence represented by $\gamma_a(\mathbf{s}_i)$.

The emission spectrum of a population with local sequence configuration $\mathbf{s}_i$ is assumed to have mean in channel $j \in \mathcal{B}$, $\mu_{\omega,j}(\mathbf{s}_i)$, and variance $\sigma_{\omega,j}^2(\mathbf{s}_i)$:

$$\boldsymbol{\omega}_i \sim \prod_{j\in\mathcal{B}} \mathcal{N}(\omega_{i,j}|\mu_{\omega,j}(\mathbf{s}_i), \sigma_{\omega,j}^2(\mathbf{s}_i))$$

The variances are assumed to evolve according to a slowly-varying random walk based on the Gamma distribution:

$$v_i \sim \mathcal{G}(v_i|\alpha_v(v_{i-1}), \beta_v(v_{i-1}))$$

where $\alpha_v(v_{i-1})$, $\beta_v(v_{i-1})$ are chosen such that the expectation of v_i is equal to v_{i-1}.

Here, the background trend is assumed to be independent between channels and locally linear as described in [7]:

$$\mathbf{t}_n \sim \mathcal{N}(2\mathbf{t}_{n-1} - \mathbf{t}_{n-2}, \sigma_t^2)$$

where σ_t^2 controls the smoothness of the trend.

2.1. State-Space Form

At each instant, the set of peaks affecting the data is limited as a result of the truncated peak shape. Defining $\mathcal{I}_n$ to denote the set of peak indices corresponding to bases affecting the data at time n, with the first future base not to affect the data also included, the observation equation becomes:

$$\mathbf{y}_n = \mathbf{e}_n + \mathbf{t}_n + \sum_{i\in\mathcal{I}_n} a_i \boldsymbol{\omega}_i \phi_i(n)$$

We define $\boldsymbol{\theta}_n \triangleq \{a_i, \boldsymbol{\omega}_i, p_i, v_i, \mathbf{s}_i\}$, $i \in \mathcal{I}_n$, to be the *system-state* at time n. The number of bases included in the system-state at time n, $k_n = \dim \mathcal{I}_n$, varies with time according to the number of peaks affecting the data. The resulting model is a Hidden Markov Model, with system-state evolution distribution $p\left(\boldsymbol{\theta}_n^{(i)}, k_n^{(i)} | \boldsymbol{\theta}_{n-1}^{(i)}, k_{n-1}^{(i)}\right)$. The translation of the priors of the previous section to time-indexed form is predominantly trivial. In brief, there are four possible birth / death scenarios at each time instant:

- The set of peaks affecting the data remains the same: $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1}$, $k_n = k_{n-1}$.
- A valid peak at time $n-1$ ceases to affect the data at time n, and there are no new peaks: $\boldsymbol{\theta}_{n,1:k_n} = \boldsymbol{\theta}_{n-1,2:k_{n-1}}$, $k_n = k_{n-1} - 1$.
- There is a new peak at time n, and no other peaks cease to affect the data: $\boldsymbol{\theta}_{n,1:k_n} = \{\boldsymbol{\theta}_{n-1,1:k_{n-1}}, \boldsymbol{\theta}_{n,k_n}\}$, $k_n = k_{n-1} + 1$. $\boldsymbol{\theta}_{n,k_n}$ denotes the parameters of the new base, as generated by the evolution equations of the previous section.
- There is a new peak at time n, and one peak ceases to affect the data, $\boldsymbol{\theta}_{n,1:k_n} = \{\boldsymbol{\theta}_{n-1,2:k_{n-1}}, \boldsymbol{\theta}_{n,k_n}\}$, $k_n = k_{n-1}$. $\boldsymbol{\theta}_{n,k_n}$ denotes the parameters of the new base.

2.2. Estimation Objectives

In a Bayesian framework the posterior distribution at time n, $p(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}|\mathbf{y}_{1:n})$, is used for inference, with the expected value of a function of interest $f(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n})$ under this posterior given by $\int\int f(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}) p(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}|\mathbf{y}_{1:n}) d\boldsymbol{\theta}_{1:n} d\mathbf{k}_{1:n}$. In most cases, including ours, the posterior is not amenable to closed form analysis owing to non-linearity in the parameters, and it is necessary to resort to numerical methods. Here, we develop a numerical algorithm to estimate the posterior distribution recursively in time for on-line estimation.

3. SEQUENTIAL SIMULATION

One means of performing the required integration is to represent the posterior at each time by a set of weighted particles [3, 9]:

$$\hat{p}\left(d\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n} | \mathbf{y}_{1:n}\right) = \sum_{i=1}^{P} \tilde{w}_n^{(i)} \delta_{\boldsymbol{\theta}_{1:n}^{(i)}, \mathbf{k}_{1:n}^{(i)}} \left(d\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}\right)$$

where P denotes the number of particles, $\tilde{w}_n^{(i)}$ denotes the normalised importance weight associated with the particle of value $\left\{\boldsymbol{\theta}_{1:n}^{(i)}, \mathbf{k}_{1:n}^{(i)}\right\}$, and $\delta_{\boldsymbol{\theta}_{1:n}^{(i)}, \mathbf{k}_{1:n}^{(i)}}(.)$ is the delta function. An algorithm for updating the particles as time progresses is [3, 9]:

Algorithm 1 - ***Sequential Monte Carlo Filter***

For $n = 2, \ldots, N$

For $i = 1, \ldots, P$:

- *Draw from the importance distribution*

$$\boldsymbol{\theta}_n^{(i)}, k_n^{(i)} \sim \pi\left(\boldsymbol{\theta}_n, k_n | \boldsymbol{\theta}_{n-1}^{(i)}, k_{n-1}^{(i)}, \mathbf{y}_{1:N}\right)$$

- *Evaluate the unnormalised importance weights:*

$$\overline{w}_n^{(i)} = \frac{p\left(\mathbf{y}_n | \boldsymbol{\theta}_n^{(i)}, k_n^{(i)}\right) p\left(\boldsymbol{\theta}_n^{(i)}, k_n^{(i)} | \boldsymbol{\theta}_{n-1}^{(i)}, k_{n-1}^{(i)}\right)}{\pi\left(\boldsymbol{\theta}_n^{(i)}, k_n^{(i)} | \boldsymbol{\theta}_{n-1}^{(i)}, k_{n-1}^{(i)}, \mathbf{y}_{1:N}\right)} \overline{w}_{n-1}^{(i)}$$

- *Normalise the importance weights:*

$$\tilde{w}_n^{(i)} = \left(\sum_{j=1}^{P} \overline{w}_n^{(j)}\right)^{-1} \overline{w}_n^{(i)}$$

- ***Optional:*** *Resample to obtain P samples approximately distributed according to $p\left(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n} | \mathbf{y}_{1:n}\right)$. Set the weights equal.*

- ***Optional:*** *Apply a Markovian transition kernel invariant to the posterior for each particle stream.*

End For

End For

In the above equations, neither $p\left(\boldsymbol{\theta}_n^{(i)}, k_n^{(i)} | \boldsymbol{\theta}_{n-1}^{(i)}, k_{n-1}^{(i)}\right)$ or $p\left(\mathbf{y}_n | \boldsymbol{\theta}_n^{(i)}, k_n^{(i)}\right)$ are directly available, but may be calculated by noting that the model is of linear Gaussian state-space form with respect to the background $\mathbf{t}_n$ and also the mean peak spacing and amplitude processes, and can therefore be marginalised using the Kalman filter (e.g. [1]). Note that the emission spectra can also be marginalised, although the computational burden introduced can be prohibitive.

3.1. Technical Details

As with the proposal density of MCMC methods, the selection of a suitable importance function is critically related to the performance of an SMC algorithm. Here, an importance distribution based on local linearisation of the model (similar to the Extended Kalman Filter) is used with the idea being to construct an importance distribution that approximates the true posterior; see [3] for similar. It is important to reinforce that this is not equivalent to making the assumption of linearity, as the importance distribution is merely used to generate proposals which are then weighted according to the true posterior.

The resampling step aims to multiply or discard particle trajectories according to how important they are to our approximation of the posterior distribution. When a resampling step is performed we use the standard residual method described in [9].

Our model is defined on a variable dimension space, with parameters fixed over moderate time intervals. Degeneracy of the standard SMC algorithm for such systems is commonly known. Here, since the interval of invariance is not large, MCMC transitions can be used to help replenish the particle set [2]. That is, a kernel invariant to the posterior distribution, $p\left(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n} | \mathbf{y}_{1:n}\right)$, is applied, the idea being that such a transition can only decrease the difference between the current approximate distribution and the invariant distribution. The kernel used consists of a fixed dimension part based on a modified Gibbs sampler and a variable dimension part based on a birth/death - split/merge Reversible Jump kernel; see [5] for more details. In order to reduce the computational load, transitions are applied on a subset of the total parameter space, corresponding to those peaks centered relatively near the current time.

4. RESULTS

In this paper, we restrict our discussion to two datasets, one demonstrating the algorithm's ability to discriminate between multiple interpretations when the data is ambiguous, and the other demonstrating its ability to track a slowly-varying baseline and changing mean peak spacing.

The middle plot of Figure 1 shows a typical example of baseline tracking, with the predicted baseline (obtained using the Kalman smoother) accurately tracking the true background. However, small discrepancies do occur on account of the inescapable ambiguity as to what is baseline, and what is signal; this is not a fault in the model itself. The bottom plot of the figure shows the true mean peak spacing process against the predicted. It can be seen that the algorithm is performing well.

In Figure 2, a dataset is shown where the standard deviation of the peak jitter process, σ_p, was set to one third of the mean, μ_p. The peaks overlap significantly, and the noise level is more than usually high. The data is shown superimposed on the prediction corresponding to two particle trajectories obtained from the algorithm. It is visually apparent that both particles provide a reasonable interpretation of the data. The priors on peak spacing and amplitude, however, skew the inference in favour of the more parsimonious 8 base model, which corresponds to the truth.

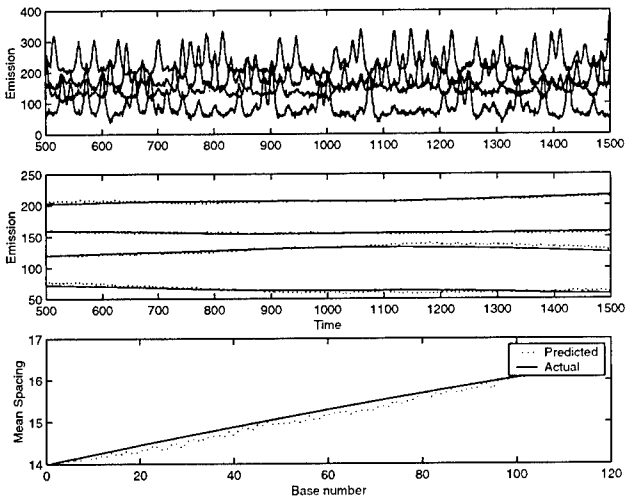


Figure 1: An example of base-line tracking. The top plot shows a the predicted signal (dotted) against the data (solid). The middle plot shows the true background (solid) against the predicted background (dotted). The bottom plot shows the true mean peak spacing process (solid) against the predicted process (dotted)

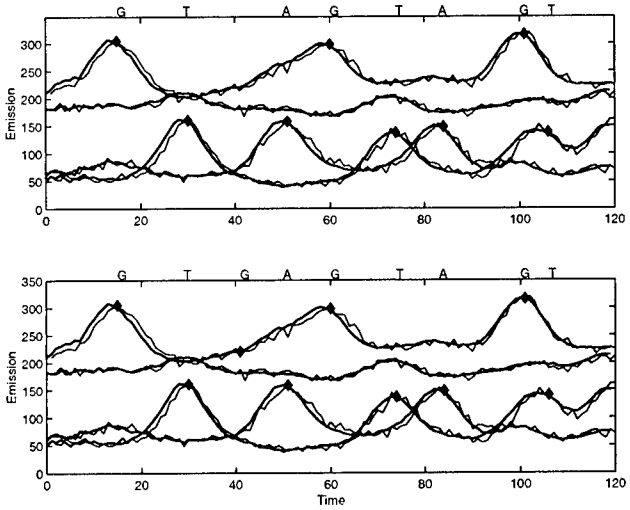


Figure 2: An example of base sequence ambiguity. The figures show the predicted data superimposed on the predictions of the SMC algorithm, with predicted sequence running along the top of each figure. The top figure corresponds to the correct interpretation.

The probabilities of the two models, as given by their frequency in the particle set, were .88 and .12 for the 8 and 9 base systems respectively; other model orders were not supported.

5. CONCLUSIONS

We have briefly introduced the DNA sequencing problem, and provided a meaningful statistical framework in which to represent available information. This framework was then translated into one suitable for sequential estimation of the posterior distribution of interest as it evolves in time. Results of the algorithm are promising, with tracking of the baseline and other nonstationarity leading to improved inference. In future, a more rigorous evaluation of the algorithm against Phred will be performed.

6. REFERENCES

[1] C. Andrieu and A. Doucet. Particle filtering for partially observed Gaussian state space models. Technical Report CUED/F-INFENG/TR393(2000), University of Cambridge, Dept. of Engineering, 2000.

[2] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.

[3] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[4] B. Ewing, L. Hillier, M. Wendl, and P. Green. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, 8:175–185, 1998.

[5] N. Haan and S. Godsill. Modelling electropherogram data for DNA sequencing using MCMC. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000. Paper no. 2573.

[6] N. Haan and S. Godsill. Sequential methods for DNA sequencing. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001. Paper no. 2227.

[7] G. Kitagawa and W. Gersch. *Smoothness Priors Analysis of Time Series*, volume 116 of *Lecture Notes in Statistics*. Springer-Verlag New York, 1996.

[8] R. Lipschutz, F. Taverner, K. Hennessy, G. Hartzell, and R. Davis. DNA sequence confidence estimation. *Genomics*, 19(417-424), 1994.

[9] J. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *J. Amer. Stat. Assoc.*, 93, 1998.

[10] F. Sanger, S. Nicklen, and A. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci*, 74:5463–5467, 1977.

[11] D. Thornley. *Analysis of Trace Data from Fluorescence Based Sanger Sequencing*. PhD thesis, University of London, Imperial College of Science, Technology, Medicine, Dept. of Computing, 1997.

SINGLE TRIAL VEP EXTRACTION USING DIGITAL FILTER

R.Palaniappan and P.Raveendran

Dept. of Electrical and Telecommunication
Faculty of Engineering
University of Malaya
50603 Kuala Lumpur Malaysia
Email: psar/ravee@fk.um.edu.my
Tel: (603)-79675253/5332

ABSTRACT

We describe a method to extract single trial Visual Evoked Potential (VEP) buried in ongoing Electroencephalogram (EEG) activity. The common method for separating VEP from EEG is to use signal averaging. But we use digital filters to extract VEP assuming that VEP spectra are in the gamma band. As an application, a Fuzzy ARTMAP (FA) neural network classifier with voting strategy is used with this extracted VEP to discriminate alcoholics from normal subjects. The VEP is extracted from subjects while seeing visuals of Snodgrass and Vanderwart picture set. The high FA classification of 96.5% shows the validity of the proposed method to successfully remove EEG contamination.

1. INTRODUCTION

VEP are signals generated in the brain in response to visual stimulus. Its analysis has become very useful for neuropsychological studies and clinical purposes. The VEP signal is embedded in the ongoing EEG with additive noise causing difficulty in detection and analysis of this signal. Furthermore, SNR of VEP to EEG is very low, approximately –5 dB [5], which complicates the situation further. The traditional technique of solving this problem is to use ensemble averaging [1]. However, this approach requires many trials and the averaged signal might tend to smooth out inter-trial information.

In addition, inter trial variation in latency and amplitude might serve to distort the VEP signal. In this paper, we propose a method to extract single trial VEP buried in the spontaneous EEG activity using digital filters and use it to discriminate alcoholics and control subjects.

2. VEP EXTRACTION FROM EEG

The extracted signal is first filtered to eliminate EEG signals since EEG signal spectra are in the range of 0 to 30 Hz. We assume that the spectra of the VEP signals lie in the gamma band centred at 40 Hz.

The z transform of the filter is

$$G(z) = (1 - z^{-1})^{2N} (1 + z^{-1})^{N}. \quad (1)$$

The integer value N can be increased to reduce the bandwidth of the filter. After some experimental simulation, we found that a value of 2 for N is sufficient for our purpose. This band-pass filter extracts spectra from 29 to 48 Hz (using 3 dB cutoff and rounded to nearest integer) with a sampling frequency of 128 Hz. The first half of (1) acts as high pass filter while the second half acts as low pass filter, which when combined gives a band-pass filter with a maximum gain at 39 Hz, which is close to the ideal gamma band centre of 40 Hz. This fact can be shown by replacing z with $e^{j2\pi fT}$ in (1) which gives us

$$G_N(f) = \left|2 \sin \pi fT\right|^{2N} (2 \cos \pi fT)^{N}. \quad (2)$$

As an example, consider a VEP segment buried in two EEG segments as shown in Figure 2. The signal is given by

$$x(n) = x_{VEP}(n) + x_{EEG1}(n) + x_{EEG2}(n), \quad (3)$$

where $x_{VEP}(n) = A_{VEP} \sin(2\pi n f_{VEP} / f_s)$,
$x_{EEG1}(n) = A_{EEG1} \sin(2\pi n f_{EEG1} / f_s)$ and
$x_{EEG2}(n) = A_{EEG2} \sin(2\pi n f_{EEG2} / f_s)$.

0-7803-7011-2/01/$10.00 ©2001 IEEE

We assume that f_{VEP}=40 Hz. We choose f_{EEG1}=15 Hz and f_{EEG2}=10 Hz, arbitrarily. SNR value of –5 dB corresponds to A_{EEG1}= A_{EEG2}=1.8 with A_{VEP}=1.0, approximately. Figure 1 shows the plot for $x(n)$ obtained by using these values for two seconds of data with a sampling frequency, f_s of 128 Hz. As shown in the figure, assume that the 3 signals exist at different points in time.

Figure 2 shows the data output from the filter with order N=2 for input $x(n)$ given by (1). For the filtered case, the SNRs of VEP/EEG1 and VEP/EEG2 are approximately 14 dB and 30 dB, respectively. This improved SNR values indicates the ability of the digital filter to remove EEG contamination successfully.

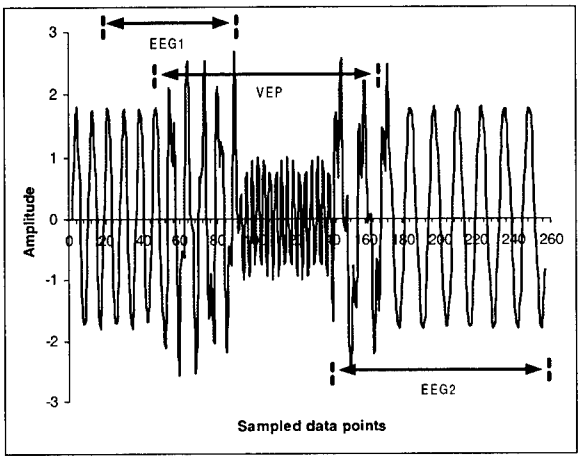


Figure 1. VEP segment buried in two EEG segments.

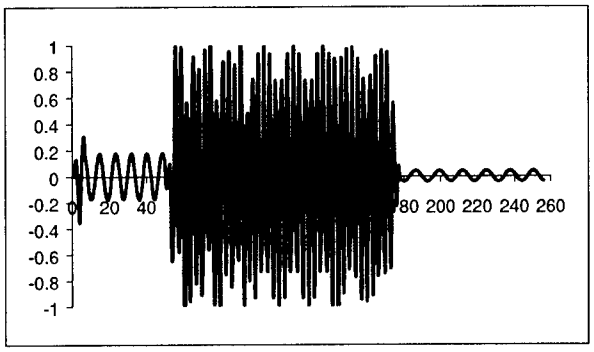


Figure 2. Filtered VEP.

The filter can be realised using only adder and delay circuits as shown in Figure 3.

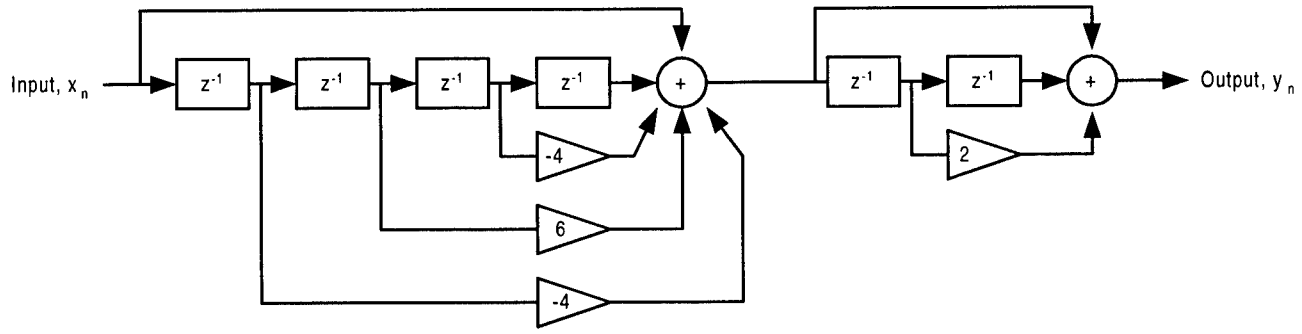


Figure 3. Filter with N=2.

3. EXPERIMENTAL METHOD

VEP signals are extracted from 20 (10 alcoholic and 10 normal) subjects with each completing 40 trials. Measurements are taken for one second from 64 electrodes placed on the subject's scalp, which are sampled at 256 Hz. The VEP signals are low pass filtered using

$$z(n) = y(n) + y(n-1), \tag{4}$$

where $y(n)$ is the output of filter discussed in Section 2 and $z(n)$ is the low pass filtered output. This will remove any frequency above 128 Hz. Next, the VEP signals are downsampled by half to obtain an equivalent sampling frequency of 128 Hz. This is since we are not interested in

frequencies higher than 64 Hz for evoked potential analysis. The electrode positions are located at standard sites (Standard Electrode Position Nomenclature, American Encephalographic Association). The electrode positions are as shown in Figure 4. The VEP data is extracted from subjects while being exposed to a single stimulus, which are pictures of objects chosen from the 1980 Snodgrass and Vanderwart picture set [6]. These pictures are common black and white line drawings like aeroplane, hand, banana, bicycle, ball, etc. executed according to a set of rules that provide consistency of pictorial representation.

The extracted signals are separated from EEG contamination by using the proposed digital filter. VEP signals with artefact contamination like eye blinks are removed in the preprocessing stage - VEP signals above 70μV denotes occurrence of eye blinks.

Periodogram (using Discrete Fourier Transform method) with Welch averaging [7] is used to obtain the power spectral density (PSD) of the extracted VEP. The Welch method is applied with 50% overlap.

The peak PSD from each channel is concanated into a single feature array to be used by a Fuzzy ARTMAP (FA) classifier to classify these VEP patterns as belonging to the alcoholic subjects class or normal subjects class. Fast learning method is employed to speed up training FA and voting strategy run with 10 simulations are used to improve FA classification [4]. FA vigilance parameter is varied from 0 to 0.9 in steps of 0.1.

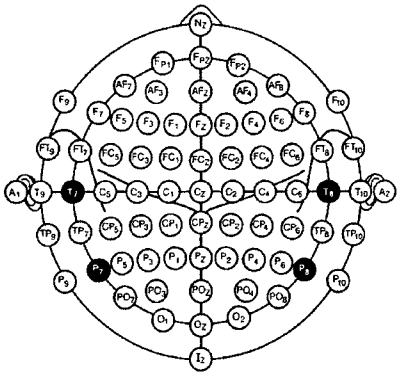


Figure 4. 64 channel electrode system.

Two experiments are simulated in the experimental study. First, the VEP signals are filtered and used in FA classification while the second classification experiment uses VEP data without filtering. This procedure is to show the advantage of using the filter to remove overlapping EEG from VEP.

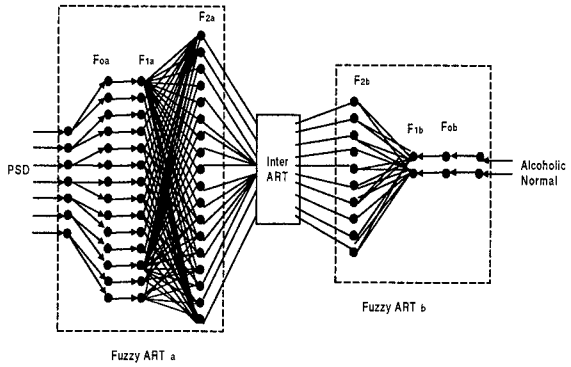


Figure 5. Fuzzy ARTMAP structure.

4. RESULTS

Table 1 shows the results of the experimental study. It can be seen that FA classification using the filtered data is higher than the case of without filtering. This is since VEP signals are contaminated with EEG and the filtering process successfully removes this contamination thereby allowing the visual stimulus to be represented in the VEP signal. In general, it can be seen that a better classification is obtained with a higher vigilance parameter with a maximum classification of 96.5% for vigilance parameter value of 0.9.

Table 1. FA classification results.

	VEP classification	
Vigilance parameter	**With filter**	**Without filtering**
0	87.25	73.25
0.1	87.75	72.25
0.2	88.25	73.00
0.3	87.00	67.75
0.4	86.75	71.75
0.5	86.50	77.00
0.6	88.00	77.25
0.7	88.50	79.00
0.8	90.25	79.50
0.9	96.50	76.25

5. CONCLUSION

This paper has proposed a method of detecting and extracting single trial VEP signals buried in EEG and noise using digital filters. FA classification using PSD of

the extracted VEP data obtained from subjects during the presentation of visuals from Snodgrass and Vanderwart picture set gives 96.5% accuracy in differentiating alcoholics from control subjects. The high classification shows that the proposed method is advantageous in single trial VEP detection and classification.

ACKNOWLEDGEMENTS

We thank Dr. Henri Begleiter at the Neurodynamics Laboratory at the State University of New York Health Center at Brooklyn, USA who generated the raw ERP data and Mr. Paul Conlon, of Sasco Hill Research, USA for making the data available to us.

REFERENCES

[1] J.I. Aunon, C.D. McGillem and D.G. Childers, "Signal Processing in Event Potential Research: Averaging and Modelling," CRC Crit. Rev. Bioeng., vol. 5, pp. 323-367, 1981.

[2] E. Basar., C.B. Eroglu, T. Demiralp and M. Schurman, "Time and Frequency Analysis of the Brain's Distributed Gamma-Band System", pp. 400-410, IEEE Eng. in Med. and Bio. Mag., July/Aug. 1995.

[3] J.P Burg, "A new analysis technique for time series data," in Modern Spectrum Analysis, New York, IEEE Press, 1978.

[4] G.A. Carpenter, S. Grossberg and J.H.Reynolds, "A Fuzzy ARTMAP Nonparametric Probability Estimator for Nonstationary Pattern Recognition Problems," IEEE Trans. on Neural Networks, vol. 6, no. 6, pp. 330-1336, 1995.

[5] H.O. Gulcur, M.Demirer and T. Demiralp, "An RBF Approach to Single Trial VEP Estimation," pp.54-56, IEEE EMBS Conference, 1997.

[6] J.G. Snodgrass and M. Vanderwart, "A Standardzed Set of 260 Pictures: Norms for Name Agreement, Image Agreement, Familiarity, and Visual Complexity," J. of Exp. Psychology: Human Learning and Memory, vol. 6, no.2, pp. 174-215, 1980.

[7] P.D. Welch, "The use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," IEEE Trans. Audio and Electroacoustics, vol. 15, pp.70-73, 1967.

PATHOLOGICAL ANALYSIS OF MYOCARDIAL CELL UNDER VENTRICULAR TACHYCARDIA AND FIBRILLATION BASED ON SYMBOLIC DYNAMICS

Zhu Yisheng* Zhang Hongxuan* Nitish V. Thakor**

* Department of Biomedical Engineering, College of Life Science and Biotechnology,
Shanghai Jiao Tong University, Shanghai, 200030, P. R. China

** Department of Biomedical Engineering, Johns Hopkins School of Medicine, Baltimore, MD 21205 USA

Abstract: Ventricular fibrillation (VF) is one of the most serious malignant arrhymias usually resulted from immediate degeneration of ventricular tachycardia (VT). In order to analyze the nonlinear dynamics of cardiac micro-mechanism under VT and VF rhythm, at the cellular level, myocardial cell action potentials (APs) are investigated under different rhythm, normal sinus rhythm, VT and VF. On the basis of nonlinear chaotic theory and symbolic dynamics, we forwarded some new definitions, complexity rate, etc, and obtained some useful properties for cellular electrophysiological analysis. The results of the experiments and computation show that the myocardial cellular signals under VT and VF rhythm are different kinds of chaotic signals that the cardiac chaos attractor under VF is higher than that under VT. The analytical complexity theory has a good promising in the clinical application.

I. INTRODUCTION

Study on the pathology and electrophysiology of the normal and abnormal heart rhythms is of great important clinical significance[1][2][3]. Currently, in this field the application of linear and nonlinear theories focus on the analysis of body surface ECGs and heart rate variability. They are all the studies conducted on the whole heart level. However, at the cellular level to study myocardial cell action potentials (APs) during different rhythm is also necessary, since Aps are the bases of the ECG. Through the analysis of Aps, we can more easily understand the electrophysiological mechanism of VT and VF. With the development of physiological experiment, it is realized that the ion channels of myocardial cell membrane possess nonlinearity. Aps reflect the nonlinear interaction of ion channels and contain all kinds of information of cell electrophysiology. Because of the exchange of material and energy with external environment through ion exchange and propagation of excitation between cell as well as the mechanism of the cell itself, various kinds of mechanical, electrical, thermal and chemical coupling exist among all the parts inside and outside the cells. Therefore, a myocardial cell can be treated as a nonlinear system, where chaos can occur under certain condition.

Quantitative chaos analysis has been a very effective method in the nonlinear biosignal processing [4][5][6][7]. But in recent study, some researchers have pointed out there appeared some application limitation in the traditional chaotic signal analysis method, such as Grassberger and Procaccia (GP) algorithm and Lyapunov exponent, because of the Takens' embedded theory which was constructed for the chaotic analysis of low-dimensional attractors[8][9][10]. Aiming at these limitations, this paper forwards some novel quantitative methods, such as complexity rate, complexity dispersity and complexity saturation, based on symbolic dynamics and nonlinear theory. To some extent, the extracted complexity information can reflect mechanisms of the body-fluid control and neurological adjustment. In our study, myocardial cell Aps were measured by floating microelectrode technique from isolated rabbit heart during ventricular tachycardia and fibrillation.

II. THEORY AND METHODOLOGY

2.1 L-Z complexity

From the viewpoint of dynamics, steady state and periodic motion is in order and not complexed, but the dynamical system becomes complex when it enters chaotic state[11]. For a system, it is important to characterize its complexity quantitatively[11][12][13][14][15]. It can be estimated from the measurable 1-dimensional signal reflecting the comprehensive interactions of components of the multi-dimensional system. From a given finite sequence $S=s_1s_2\cdots s_n$, Lempel and Ziv proposed one useful complexity measure $c(n)$ and offered the related mathematical definitions and deductions in detail[11]. $C(n)$ can characterize development of spatiotemporal patterns.

2.2 complexity rate and dispersity

According to our experiments and simulations, limited system information can be extracted from the coarse-grain symbol dynamic sequences and speed biosignal information cannot be obtained with only complexity measure computation. Clinic workers hope to get the accurate pathological reason in the abnormal cardiac signal analysis, for instance the body fluid and nerve control interdiction, apart from the abnormal cardiac signal features extraction. Complexity rate information extracted from the ECG data records can construct a correct and reasonable relationship between biosystem pathology and human cognition interface. On the basis of established complexity measure and complexity method of system features extraction [15][16][17], we put forward a new method for complexity study — the symbol dynamic system complexity rate and dispersity information. The inner reason of nonstationary biosystem dynamic change can be excavated with the help of this method[18].

Given a dynamic system time sequence $X=\{x_1, x_2, \cdots, x_i, \cdots\}$, there exists subsequence L_i,

$L_i=\{x_1, x_2, \cdots, x_i\}$, in which i=1, 2, $\cdots$, n;

Utilizing the L-Z complexity measure, corresponding complexity can be computed for each subsequence L_i; suppose L_i is correspond to complexity c_i.

2.2.1 Definition (Finite sequence complexity sequence)

Suppose sequence $X=\{x_1, x_2, \cdots, x_i, \cdots\}$, there exists subsequence L_i, $L_i=\{x_1, x_2, \cdots, x_i\}$, in which i=1, 2, $\cdots$, n; we difine $c_x=\{c_1, c_2, \cdots, c_n\}$ as the corresponding complexity measure sequence of the sequence X_n, in which c_i is the sequence complexity of the L_i, X_n is the finite time sequence of X.

2.2.2 Definition (Time sequence complexity rate)

Given a finite time sequence $X=\{x_1, x_2, \cdots, x_i, \cdots\}$, the corresponding finite complexity sequence is $c_x=\{c_1, c_2, \cdots, c_n\}$, we define complexity as follows:

$$cc(n)=\frac{c_{n_i}-c_{n_j}}{n_i-n_j} \quad (1)$$

in which n_i-n_j must at least larger than Takens' embedding

0-7803-7011-2/01/$10.00 ©2001 IEEE

dimension[8]. We can denote the complexity rate as: cc(n)=diff(n) ; intituled as time sequence instantaneous complexity. cc(n) reflects the speed of the complexity change of the definite time sequence.

According to this definition, the complexity rate of the whole time sequence X(n) can be calculated from slope rate of the sequence fitting polynomial:

$$cc[x(n)] = DIFF[x(n)] \tag{2}$$

2.2.3 Definition (Average complexity)

Given a limited dynamic time sequence $X=\{x_1, x_2, \cdots, x_n\}$, in which $n<\infty$;the corresponding complexity sequence is $c_x=\{c_1, c_2, \cdots, c_n\}$,

Then:

$$\overline{c_x} = \lim_{n\to\infty} \frac{1}{N} \sum_{i=1}^{n} c_i \tag{3}$$

2.2.4 Definition (Complexity dispersity)

$$FL = \frac{\|C_N - C_r\|}{\sigma_N} \tag{4}$$

in which, FL is the CD (complexity dispersity) of a given time sequence $\{x_n\}$; C_N is the complexity measure of the sequence; C_r is the complexity measure of surrogate data time sequence $\{x'_n\}$; σ_N is the mean square deviation of surrogate data time sequence $\{x'_n\}$ (Here we employ Gaussian surrogate method in our complexity dispersity computation[19][20][21]). The symbol $\|\bullet\|$ stands for Euclidean distance in the algorithm of complexity measure.

From the complexity dispersity definition, we can obtain the following. If the given time sequence $\{x_n\}$ or the main part of it is a stochastic process, the complexity measure of surrogate time sequence is proportional to that of the original time sequence, but the corresponding mean square deviation is bigger and that results in less complexity dispersity. On the other hand, if the given time sequence is deterministic chaotic signal, the corresponding complexity dispersity criterion is bigger[21].

2.3 complexity saturation

When a deterministic periodic dynamic system enters into chaos, then random, we observe that the complexity of periodic procedure is definite and not allied to sampling start point. When the dynamic system is a low-dimension chaotic system, the corresponding complexity is a definite value and there is saturation phenomenon during the complexity computation. When the dynamic system is in high-dimension chaos, the corresponding complexity increases with the length of the procedure and the complexity saturation phenomenon is not easy to observe in spit of the existence. When dynamic system diverges into random, there exist no saturation phenomenon and the complexity is proportional to the length of the sequence. The corresponding complexity rate of the high-dimension chaotic system is higher than that of low-dimension chaotic system. It can be concluded that complexity saturation is a significant parameter in the research of periodic, chaotic and random procedure.

III. EXPERIMENT RESULTS

1. Data acquisition

The experiments were accomplished on isolated heart experimental set-up in Johns Hopkins School of Medicine[22]. Large New Zealand rabbits (4-6 kg) were anesthetized and their hearts were rapidly exercised through median sternotomy and perfused in a Langendorff style preparation. During normal sinus rhythm, peak of retrograde perfusion pressure was maintained at 80 mm Hg. A pair of electrodes attached to the epicardium were used to pacing the heart or for inducing fibrillation by 60 Hz AC stimulation. Another separate pair, consisting of a large area patch electrode and an intraventricular catheter, was utilized for delivering defibrillation shocks. To record the Aps from single cell, a floating microelectrode technique was used[23]: the electrode was constructed of a thin, coiled silver wire and a standard capillary glass micropipette. By using a micromanipulator, as Fig 1, the microelectrode was lowered onto the heart surface until it achieved cell penetration due to natural gravitational force. The silver wire maintained the impalement despite heart motion and several minutes of continuous recordings were obtained in this manner. After filtered and amplified, the experimental data were collected on tape by a wide band FM recorder, and then digitized (200 samples per second) on a personal computer using a data acquisition system. The single cell Aps were obtained during the condition of NSR(normal sinus rhythm), VT(ventricular tachycardia) and VF(ventricular fibrillation). VT and VF were induced by 60 Hz AC electrical stimulation. During VF in vivo, the heart loses pumping function and the blood pressure in the coronary arteries is approximately zero. To stimulate the conditions in the isolated heart preparation, perfusion was stopped after the induction of VF. The action potential waveforms of the experiment were shown in the Fig 2.

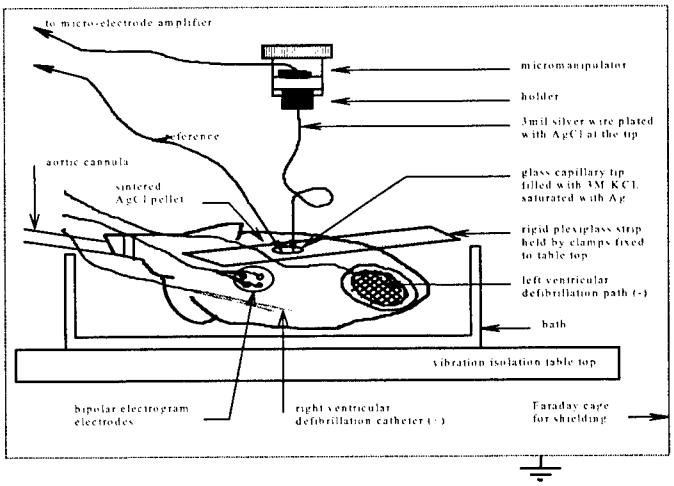


Fig 1. Schematic of the isolated animal heart experiment

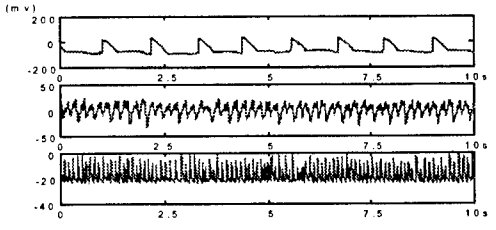


Fig 2. Time domain Waveform of myocardial cellular Aps under the rhythm of NSR, VT and VF

2. Complexity comparison of VT and VF

Figure 3 is the complexity comparison map of VT and VF, in which both have 15 groups record data. Every item data length of each record is 2000 and the signal-sampling rate is 200Hz. In the complexity comparison experiments, every group of Aps data under NSR, VT and VF rhythm are under similar condition to validate the complexity comparison. Figure 4. is the average complexity comparison of NSR, VT and VF.

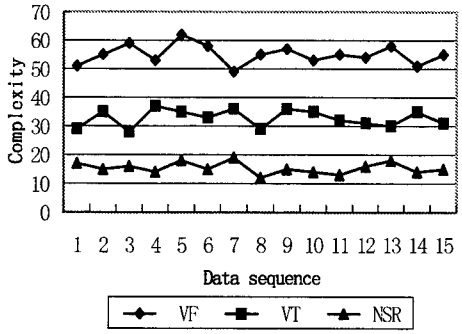


Fig 3. NSR, VT and VF complexity comparison

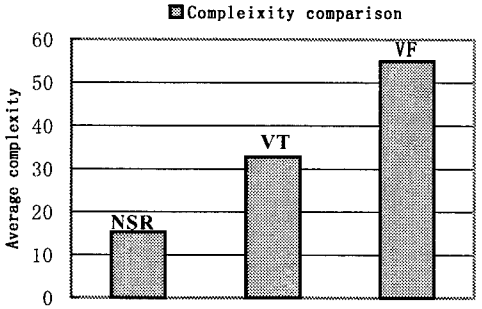


Fig 4. Average complexity comparison of AP

3. Complexity rate comparison of VT and VF

In our complexity computation and detection of VT/VF, we have utilized all experimental records (30 group data). Each individual data sample included one VF signal, one VT and one random. In our experiments and simulations, the random signal was a Gaussian distribution data generated by computer simulation. (See table 1 and table 2.)

Table 1. complexity information analysis(2000 sample data, two effective digit)

Investigation item		Maximum complexity	Average complexity	Complexity saturation
Group one	Random signal	98	53	No
	VF rhythm ECG	49	24	No
	VT rhythm ECG	14	7	Yes
Group two	Random signal	95	49	No
	VF rhythm ECG	45	22	No
	VT rhythm ECG	21	13	Yes
Group three	Random signal	97	45	No
	VF rhythm ECG	58	29	No
	VT rhythm ECG	19	11	Yes

Table 2. Complexity rate analysis of the three myocardial cell AP signals (2000 sample data, three effective digit)

Investigation item	Random signal	VF rhythm signal	VT rhythm signal
Group one	0.935	0.525	0.225
Group two	0.932	0.542	0.267
Group three	0.962	0.486	0.279
Average complexity rate	0.943	0.518	0.257

(We choose linear fit for the complexity computation.)

4. Complexity dispersity comparison of VT and VF

Table 3 Complexity dispersity analysis of myocardial cell AP signal under VT rhythm

Data	Original data complexity	Surrogate data complexity	Surrogate data mean square deviation	Complexity dispersity
Data 1(VT1)	49	87	3.56	10.67
Data 2(VT2)	55	89	3.49	9.74
Data 3(VT3)	47	78	3.45	8.99
Data 4(VT4)	59	84	3.55	9.86
Data 5(Gnoise1)	192	193	0.99	1.01
VT average	52.5	84.5	3.51	9.82

Table 4 Complexity dispersity analysis of myocardial cell AP signal under VF rhythm

data	Original data complexity	Surrogate data complexity	Surrogate data mean square deviation	Complexity dispersity
Data 1(VF1)	105	122	5.25	3.24
Data 2(VF2)	101	125	5.58	4.30
Data 3(VF3)	104	126	4.98	4.42
Data 4(VF4)	107	118	5.39	2.04
Data 5(Gnoise2)	194	195	0.99	1.01
VF average	104.3	122.8	5.30	3.50

Table 3 and Table 4 are complexity dispersity information of myocardial cell AP signals under VT and VF rhythm. The complexity and mean square deviation of VF signal in our experiments are higher than those of VT. On the standardized basis of stochastic analysis, the complexity dispersity constructed from chaotic dynamics and stochastic analysis exhibits better stability and practicability than GP algorithm. In the two tables, the complexity dispersity of Gauss stochastic process is about 1 and Gauss process cannot be a chaotic process but a random. Through electrophysiological experiments and data unification, we can confirm that the myocardial cell AP signals under VT and VF rhythm are chaotic. To our satisfaction, the complexity dispersity, to some extent, can be utilized quantitatively to identify low-dimension deterministic chaos from high-dimension deterministic chaos as well as cardiac chaotic qualitative analysis.

IV. DISCUSSION

From the experiments and computation above, we can see that a myocardial cell can be treated as a nonlinear chaotic system. With the alteration of conditions, the changes of nonlinear characteristics of the ion channels cause the dynamical behavior of the myocardial cell to change correspondingly, which is the basis of the change of dynamical behavior of the whole heart. This suggests that we should combine the studies at the cellular level with the ones on the whole heart level in order to acquire better understanding of the characteristics and mechanism of various rhythms. We can model the relationship between the activities of the epicardial

cell and the tissue or the whole heart for detecting different cardiac arrhythmia or developing new pacing mode for the patient with dysfunctions in conduction system. Moreover, the development of nonlinear dynamical information such as complexity rate and dispersity from 1-dimensional action potentials for studying the myocardial cell electrophysiology, which is of some instructiveness for bioelectrical modeling of the ion channels. These dynamical indicates can serve as clinical useful parameters for characterizing different cardiac rhythms. our study suggests that the ventricular tachycardia caused by some reasons might drive the myocardial cell into quasiperiodic motion from normal periodic motion and finally into chaos, however, chaotic rhythm can be terminated by some measures, such as drug administration and ICD (Implantable cardioverter defibrillator). We can further our investigation for the controllability of cardiac rhythms based on chaos and symbolic theory.

V. CONCLUSION

Based on symbolic dynamics and nonlinear theory, this paper forwards definitions of complexity sequence, sequence complexity rate, complexity dispersity and complexity saturation. In our experiments and computation of action potential signal, we observed credible and satisfactory results in analysis of myocardial cell AP signal under the rhythm of VT and VF. With complexity rate and dispersity, we can not only understand the cellular mechanism of the serious rhythm of VT and VF but also construct a basic link between physiology and detection parameters. Our analytical complexity information provides an effective and reliable method for analysis, detection of cardiac pathology.

ACKNOWLEDGEMENTS

The authors would like to thank the research group in John Hopkins Medical Institution for their kind help in the VT/VF research. This work was supported by National Natural Science Foundation of China (No. 69871019).

REFERENCE

[1] J. Kurths, A. Voss, P. Saparin, H.J. Kleiner and N. Wessel, "Quantitative analysis of heart rate variability," Chaos, vol.5, pp.88-94, 1995.

[2] Janice L. Jones and Oscar H. Tovar, "The mechanism of defibrillation and cardioversion," Proceeding of the IEEE, vol.84, pp.392-403, 1996.

[3] Jay A Warren, Robert D. Dreher, Ronald V. Jaworski, James J. Putzke and Renold J.Russie, "Implantable cardioverter defibrillation," Proceeding of the IEEE, vol.84, pp.468-479, 1996.

[4] A. Babloyantz and A. Destexhe, "Is the normal heart a periodic oscillator? " Biol. Cybern, vol. 58, pp.203-211, 1988.

[5] N.V. Thakor, "Chaos in the heart: Signal and Models," IEEE 2th International Biomedical Engineering Conference, Nov, 1998.

[6] Zhang Xusheng, "Nonlinear analysis of dynamical characteristics for ventricular fibrillation and study on its application," Ph.D dissertation of Shanghai Jiao Tong University, June, 1997. (in chinese)

[7] A. Bezerianos, T. Bountis, G, Papaioannou, and P. Polydoropoulos, "Nonlinear time series analysis of electrocardograms," Chaos, vol.5, pp.95-101, 1995.

[8] F. Takens, "Dynamical System and Turbulence (Lecture Notes in Mathematics), " Springer, Berlin, Heidelberg, New York, 1981

[9] James.Theiler, "Spurious dimension from correlation algorithm applied to limited time-series data," Phys.Rev.A, vol.34, pp.2427-2432, 1986.

[10]Rapp. PE and Vshore. TR, "Dynamics of brain electrical activity," Brain Topography, vol.2, pp.99-118, 1989.

[11]A. Lempel, J. Ziv, "On the complexity of finite sequences," IEEE trans on IT, vol.22, no.1, pp.75-81, 1976.

[12]P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," Physica D, vol.9, pp.189-208, 1983.

[13]P. Grassberger and I. Procaccia, "Characterization of strange attactors," Phys.Rev.Lett, vol.50, pp.346-349, 1983.

[14]B.-L.Hao, "Symbolic dynamics and characterization of complexity," Physica D, vol.51, pp.161-176, 1991.

[15]F. Kaspar, H. G.Schuster, "Easily calculable measure for the complexity of spatiotemporal patterns," Phys Rev.A, vol.36, pp.842-848, 1987.

[16]X.-S. Zhang, Y.-S. Zhu, and X.-J, Zhang, "New approach to studies on ECG dynamics: extraction and analyses of QRS complex irregularity time series," Medical & Biological Engineering & computing, vol.35, no.4, pp.467-474, 1997.

[17]Xu-Sheng Zhang, Yi-Sheng Zhu, and Nitish V. Thakor, "Detecting Ventricular Tachycardia and Fibrillation by Complexity Measure," IEEE Trans on Biomed. Eng, vol.46, pp.548-555, 1999.

[18]Zhang Hongxuan, Zhu Yisheng, and Wang Ziming, "Complexity Measure and Complexity Rate Information Based Detection of Ventricular Tachycardia and Fibrillation," Medical & Biological Engineering & computing, vol.38, no.5, pp.553-557, 2000.

[19]M.Plaus and D.Hoyer, "Detecting nonlinearity and phase synchronization with surrogate data, " IEEE EMBS Mag, vol.17, no.6, pp.40-45, 1998.

[20]James.Theiler, Stephen.Eubank, Andre Longtin, Bryan Galdrikian, and J. Doyne Farmer, "Testing for nonlinearity in time series: the method of surrogate data," Physica D, vol.58, pp.77-94, 1992.

[21]Zhang Hongxuan, Zhu Yisheng, Niu Jinhai and Tong Shanbao, "Complexity analysis of abnormal ECG rhythm: ventricular tachycardia and ventricular fibrillation," Acta Physica Sinica, vol.49, no.8, pp.1416-1422, 2000.

[22]A. Baykal, R. Ranjan, and N. V. Thakor, "Model based analysis of ECG during early stages of ventricular fibrillation," J. Electrocardiol., 27(suppl), pp.84-90, 1994.

[23]D. Hodgson, M. Fishler, R. Han, and N. V. Thakor, "Intracellular recordings during VF: Role of ion channels from experiments and computer models," Computers in Cardiology, pp.537-540, 1991.

On the Application of Model Based Distance Metrics of Signals for Detection of Brain Injury

J.S.Paul, S.Tong, D.Sherman, A.Bezerianos and N.V.Thakor
Department of Biomedical Engineering
Johns Hopkins school of Medicine
Baltimore MD 21205
Email: jpaul@bme.jhu.edu

ABSTRACT

In the basic and clinical research on brain's response to injury, electrical signals from the brain, namely EEG, is useful in providing an immediate signaling of the dysfunction. However, EEG signals have proven to be difficult to analyze and interpret due it its complex signal characteristic There is a critical need for developing robust and reliable measures that can be correlated with injury as well as survival. In this paper, we address a unique problem of characterizing quantitatively the electrical measures of brain injury for analysis of brain activity in animal and human subjects. The key objective is to model EEG spectra and its features so that signaling changes due to injury can be discovered. We do so with the method of autoregressive modeling and dominant frequency analysis. The trends in the electrical signaling following injury and following resuscitation are modeled using the cepstral distance derived from the AR model.

INTRODUCTION

About 70,000 persons per year are successfully resuscitated after cardiac arrest in both hospital and community settings in the United States. Around 60% of those persons subsequently die because of extensive brain injury and only 3 to 10% resume their former life-style [3]. The neurological recovery after successful resuscitation from cardiac arrest largely influences the morbidity and mortality of these patients [4]. Despite the magnitude of the problem, only clinical neurological assessment is used to monitor brain injury and no real time objective methods to detect and monitor brain injury exist at present time

EEG is a sensitive but nonspecific measure of brain function [15] and its use in cerebrovascular diseases is limited [5]. EEG has been used for prognostication in after resuscitation from cardiac arrest with some success]. In most of the applications, the EEG recording results in long traces with marked inter-observer variability [6]. QEEG has been used to reduce these difficulties. This technique has been confined to feature analysis, conventional power spectrum analysis, parametric description of EEG through linear auto regressive (AR) modeling, or frequency analysis based on clinically accepted δ, θ, α, and β waves [9]. Presently, the use of qEEG has very limited clinical utility, thus it is used mainly as an investigational tool. Power spectrum has been widely used to characterize EEG via the Fast Fourier Transform (FFT) and other power spectrum density estimation techniques. Linear AR modeling [7,8] has also been used and was able to reduce experimental data while preserving important features such as time-varying changes, dominant frequency components, as well as their amplitudes and powers.

We utilize AR modeling to investigate the transient properties of on going EEG, which can vital for the early detection of brain injuries. The brain's response to graded injury will be studied using quantitative characterizations of EEG signals based on distance measures, which are methods of differentiating spectra on the basis of a single continuous criterion. Our first goal is to show that the methods of distance measurement analysis determine significant variation in EEG, which reflects alteration in cerebral function during injury. The ability of the spectrum-based EEG distance measures Spectral Distance (SD) and Cepstral Distance (CD), are compared as they detect cerebral dysfunction after cardiac arrest. Our second goal is to determine if the distance measures are useful in providing prognosis of neurological recovery after global asphyxic injury.

METHODS

Q-EEG analysis – The specific goal of our preliminary work was to develop novel tools and technologies to analyze the brain's electrical signaling, as measured by Q-EEG. We utilize autoregressive (AR) modeling to investigate the transient properties of on going EEG signal characteristics. Such transient changes in EEG are vital for the early detection of brain injuries. In this model, the EEG is treated as a realization of the time-series $x[k]$ generated as follows:

0-7803-7011-2/01/$10.00 ©2001 IEEE

$$x[k]=\sum_{i=1}^{P} a_i x[k-i]+e[k]$$

where P is the order of the AR process and $e[k]$ is the unpredictable part of $x[k]$. The reason for choosing the AR models are that AR processes are capable of approximating the EEG spectrum. With the knowledge of AR model parameters which minimizes $E\{e^2[k]\}$, the EEG spectrum magnitude was estimated using

$$S_x(e^{j\Omega})=\frac{1}{\left|1-\sum_{n=1}^{P} a_n e^{-j\Omega n}\right|}$$

The mathematical scheme to build the AR model and to calculate the cepstral distance (CD) metric is illustrated in Fig. 1 [2,11]. Preliminary studies reported that the "distance" metric rigorously defines the spectral differences between the EEG during the control or preconditioning period and the subsequent injury and early recovery periods. The model order, and $w(k)$ is the error in prediction [12]

2. Spectrum generation, $P(z)$, from the autoregressive coefficients
   $P(z)=\sigma^2/\left|1+a(1)z^{-1}+a(2)z^{-2}+...+a(p)z^{-p}\right|$
   where $\sigma^2=E\{w(n)^2\}$ is the variance of the input noise. Peaks in spectrum show dominant frequencies.
3. The power in dominant peaks is given by the area under the peaks in the power spectrum [13] based on the residues
   $Power(\omega_{dominant})=$
   or
   $2\,\mathrm{Re}\{\text{Residue of } P(z)/z \text{ at } z=\exp(j\omega_{dominant})\}$

We examine the differences in the recovery pattern of all three dominant frequency bands by a parameter called Normalized Separation, NS, [1] defined as:

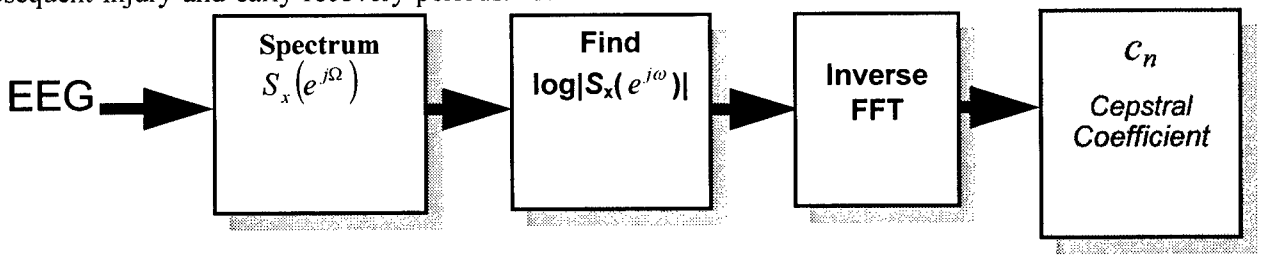


Fig. 1: Flowchart illustrating methodology for finding the cepstral coefficients of a time series. Low pass filtered EEG fast-Fourier transformed (FFT) (1). Afterwards the log of the FFT output is made and then the inverse FFT is taken. The resulting output are the cepstral coefficients. To find the cepstral distance or CD between two time series, the sum of magnitude squared differences between respective coefficients is made.

CD method has been shown to be more accurate in correlating the electrical signaling response to the Neuro Deficit Scoring (NDS). Using a stable EEG time series and the correct model order, dominant frequency analysis allows a spectral breakdown of the EEG from the data itself. Thus, it is unique to each patient. EEG may not cluster into traditional alpha, beta, delta, and theta bands, but this can be resolved since Dominant Frequency analysis allows for customization for the individual patient. An indication of balanced/proportional recovery in EEG Frequency Bands is obtained using the following steps:

1. Predict current sample by weighted past samples or
   $x(k)=w(k)+a_1x(k-1)+a_2x(k-2)+...+a_px(k-p)$
   where $x(k)$ is the data sequence, $a(i)$, $i = 1,...p$ are the autoregressive (AR) parameters, p is the

$$NS=(|P_{lf}-P_{mf}|+|P_{mf}-P_{hf}|+|P_{lf}-P_{hf}|)/(P_{lf}+P_{mf}+P_{hf})$$

Where P_{lf}, P_{mf}, and P_{hf} represent the power in the low, medium, and high dominant frequency components relative to their respective baseline values [1].

RESULTS

Graded Response to Injury was obtained using: the Cepstral distance measure. We analyzed EEG signal by constructing an AR signal model. As reviewed earlier, the AR model is appropriate for characterizing short EEG signal segments and yields its spectrum, which can then be used to identify its characteristic spectral peaks (Fig. 2 (a)). From the peaks of the AR spectra, dominant frequencies, where the power in EEG signal is concentrated, are identified.

This method was used extensively in ischemic brain injury studies to identify the duration and extent of electrical function change in response to different severity of insults, and to identify the various phases of recovery of electrical function [2, 10]. However, the measure is not specific, in that it is unable to predict the outcome beyond the first 20 minutes following the insult. Also, the measures do not show any indication of the short-lasting electro physiological changes such as bursts or seizure

For our experiment, we used the AR power spectrum to develop a new index of EEG recovery—the normalized separation (NS) (Fig. 3). This study showed that the HI injury causes a dispersion or redistribution of power in the dominant frequencies. NS monitors the rate of recovery for each band with respect to baseline. A high NS represents a disproportionate recovery of power, and vice versa. A high NS implies a poor recovery of the electrical function.

DISCUSSION

There are several limitations of our approach. A critical one is that a pre-injury baseline is needed to compare the distance measure against the post injury measure(s). The measure does not distinguish power in different frequency bands and different spectral energy evolutions in these bands. The measure is not specific, in that the relationship with the neurodeficit score is not high and that it is unable to predict the outcome beyond the first 20 minutes following the insult. Also, the measures do not show any indication of the short-lasting electrophysiological changes such as spindles, bursts or seizures. The metric tools can only partially characterize the static features of the post-injury EEG. We hypothesize that EEG, like many other biological phenomena, displays both 'static' and 'dynamic' features. The latter are formed as a pattern generated by numerous neuro-electrical events within the brain's complex structure. The static features represent the global characteristics of the system such as the parameters with an evolving trend suitable for monitoring long-term EEG especially during recovery from a trauma. The changes in the dynamical picture of various rhythmic and chaotic components of the EEG signal that make it possible to detect changes in the states of the underlying mechanisms [14].

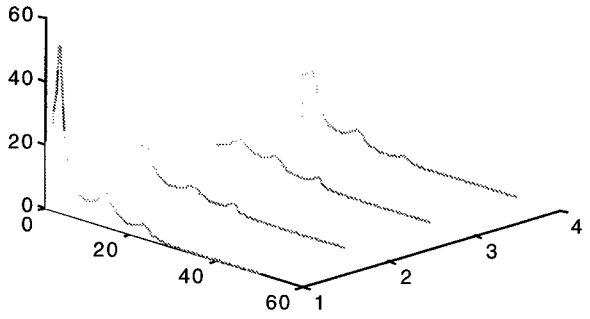


Fig. 2: (a) Autoregressive (AR) spectral plots taken during the course of HI injury and recovery in rat. (1) is the spectrum during baseline prior to the insult; (2) is the spectrum after 1.5 min of CA; (3) after 2 min of CA and (4) is after 15 min of recovery.

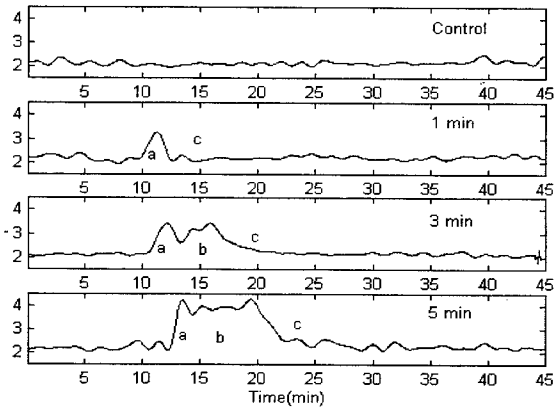


Fig. 2: (b) shows the CD measure plotted as a function of time. The CD measure demonstrably discriminates the injury duration and severity (Geocadin, et al. 2000).

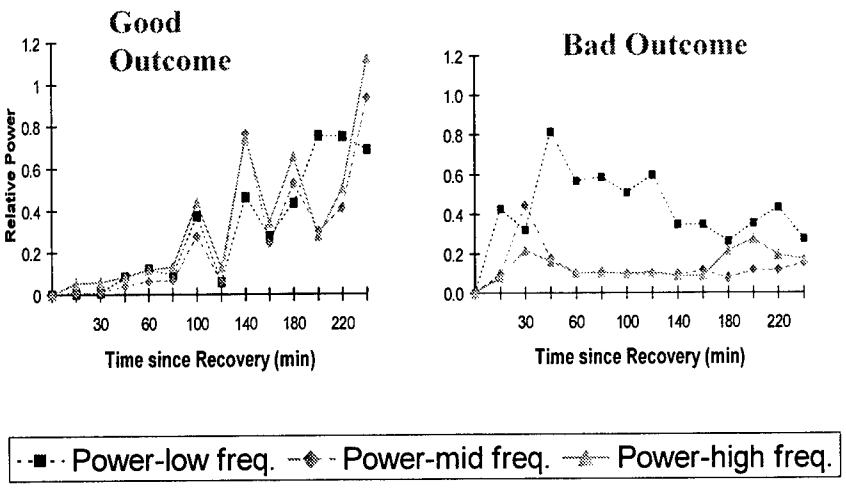


Fig. 3. Recovery of the relative power in the three dominant frequency bands for two animals. ' ': 1-5.5 Hz, '◇': 9-14 Hz, and '△': 18-21 Hz. Left: A uniform spectral recovery resulting in a low NS confirmed by a high NDS (good outcome). Right: Spectral recovery for an animal with a high NS indicating spectral dispersion or unequal recovery of different frequency bands. The Neuro Deficit Score (NDS) of this animal is low indicating a bad outcome.

References

[1] V.Goel, A.M. Brambrink, A.Baykal, R.C. Koehler, D.F. Hanley and N.V. Thakor, "Dominant Frequency Analysis of EEG Reveals Brain's Response During Injury and Recovery". *IEEE Trans. BME*, 45, 1996, 1083-1092

[2] R. G. Geocadin, R. Ghodadra, T. Kimura, H. Lei, D. L. Sherman, D. F. Hanley and N. V. Thakor, "A novel quantitative EEG injury measure of global cerebral ischemia", *Clinical Neurophysiology*, Volume 111, Issue 10, 1 October 2000, Pages 1779-1787.

[3] Krause, G.S., Kumar, K., White, B.C., Aust, S.D. & Wiegenstein, J.G. Ischemia, resuscitation, and reperfusion: mechanisms of tissue injury and prospects for protection. *Am Heart J.*, 1986; 111, 768-780.

[4] Earnest, M.P., Yarnell, P.R., Merrill, S.L. & Knapp, G.L. Long-term survival and neurologic status after resuscitation from out-of-hospital cardiac arrest. *Neurology.*, 1980; 30, 1298-1302.

[5] Nuwer, M. Assessment of digital EEG, quantitative EEG, and EEG brain mapping: report of the American Academy of Neurology and the American Clinical Neurophysiology Society. *Neurology.*, 1997; 49, 277-92.

[6] Williams, G.W., Luders, H.O., Brickner, A., Goormastic, M. & Klass, D.W. Inter-observer variability in EEG interpretation. *Neurology.*, 1985; 35, 1714-1719

[7] Madhvan, P.J., Stephens, B.E., Klinberg, D. & Morzorati, S. Analysis of rat EEG using auto regressive power spectra. *J. Neurosci. Method.*, 1991; 40, 91-100.

[8] Akay, M., Akay, Y.M. & Szeto, H.H. The effects of morphine on the relationship between fetal EEG, breathing and blood pressure signals using fast wavelet transform. *Biol Cybern.*, 1996; 74, 367-72.

[9] E. Niedermeyer and F.Lopes da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, Williams & Wilkins, 4th ed. Baltimore, MD 1998

[10] R.G. Geocadin, J. Muthuswamy, D. L. Sherman, N.V. Thakor, and D.F. Hanley Early Electrophysiological and Histological Changes after Global Cerebral Ischemia in Rats. *Movement Disorders*. 15, p. 14-21, 2000

[11] X. Kong, A. Brambrink, D. F. Hanley, and N. V. Thakor, "Quantification of injury-related EEG signal changes using distance and information measures," *IEEE Trans. Biomed. Eng.*, vol. 46, p. 899-901 1999.

[12] S. L. Marple, *Digital Spectral Analysis with Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[13] S. J. Johnson and N. Andersen, "On power estimation in maximum entropy spectral analysis," *Geophysics*, vol. 43, pp. 681-690, 1978.

[14] D. L. Sherman, A. M. Brambrink, R. N. Ichord, V. Dasika, R. C. Koehler, D. F. Hanley, R. C. Koehler, R. J. Traystman, and N. V. Thakor, "The EEG during Early Reovery From Hypoxic-Ischemic Injury In Immature Piglets: The Role of Bursting," . *Clin. Electroencephal.* Vol. 30, 175-183 1999.

[15] R. R. Clancy, "Differential Diagnosis and Contribution of EEG," presented at Report of the Workshop on Acute Perinatal Asphyxia in Term Infants, Rockville, MD, 1993.

Entropy of Brain Rhythms: Normal versus Injury EEG

N. V. Thakor[1], J. Paul[1], S. Tong[1,2], Y. Zhu[2], A. Bezerianos[1,3]

[1] Dept. of Biomedical Engineering, Johns Hopkins School of Medicine, Baltimore, MD21205, USA
[2] Dept. of Biomedical Engineering of Shanghai Jiaotong University, Shanghai, China
[3] Dept. of Biomedical Physics, School of Medicine , University of Patras, Patras, Greece

Abstract

In communication theory, information measures answer two fundamental questions, viz: the ultimate data compression (by entropy) and the ultimate transmission rate (by the channel capacity). In case of brain and the study of brain function analyzing EEG, the information measures help to show how entropy can be used to remove redundancy in EEG and consequently making it useful for monitoring of brain function in critical conditions and secondly on how information transmission measures describe normal e.g. sleep stages and divergence from normal e.g epilepsy or ischemic brain injury.

1. Introduction

Shannon's entropy [1] has been accepted as a method to characterize information content in a signal. Entropy is defined as a measure of uncertainty of information in a statistical description of a system [2]. In other words, the entropy is a measure of our ignorance about the system. Given a discrete random variable X with alphabet H=$\{x_i\}$ and probability function $p(x_i)=Pr(X=x_i)$, $x_i \in$ H the entropy is defined by

$$H = -\sum_{x_i} p(x_i) \ln p(x_i) \qquad (1)$$

The relative entropy or Kullback Leibler distance [3]between two probability functions $p(x)$ and $q(x)$ is defined as

$$D(p \parallel q) = \sum_{x_i} p(x_i) \ln \frac{p(x_i)}{q(x_i)}. \qquad (2)$$

Another measure on the correlationship between two systems is mutual information. Consider another discrete random variable Y with probability function $p(y_j)$. The joint probability between variables X and Y is $p(x_i,y_j)$. The mutual information I(X;Y) is the entropy transferred between the joint distribution and the product distribution $p(x_i)q(y_j)$, i.e.

$$I(X; Y) = \sum_{x_i} \sum_{y_j} p(x_i, y_j) \ln \frac{p(x_i, y_j)}{p(x_i) q(y_j)}. \qquad (3)$$

Recent years, a novel non logarithmic entropy (Tsallis entropy) was introduced as the generalization formalism of Shannon Entropy, which is parameterized and dependent on an entropic index r [4]

$$H_r = -(r-1)^{-1} \sum_{x_i} \left[p(x_i)^r - p(x_i) \right]. \qquad (4)$$

For $r \to 1$ Tsallis entropy coincides with Shannon entropy.

2. Application to Brain Rhythms

Entropy itself is a description of average uncertainty in the signal duration recorded. It is not useful for analyzing nonstationarity. To get a temporal evolution of entropy, an alternative time dependent entropy measure based on sliding temporal window technique is applied [5] [6]. Let $\{s(k): k = 1,\ldots,N\}$ denote the raw sampled signal. Now we define a sliding temporal window W determined by two parameters: the width $w \le N$, and the sliding step $\Delta \le w$. Then sliding windows are defined by: $W(n; w; \Delta) = \{s(i), i = 1 + n\Delta, \ldots, w + n\Delta\}$, $n = 0,1,2,\ldots,[N/\Delta] - w + 1$, where $[x]$ denotes the integer part of x. Within each window $W(n; w; \Delta)$, we introduce the set $\{I_i : i = 1,\ldots,L\}$ of disjoint amplitude intervals such that:

$$W = \bigcup I_i \qquad (5)$$

where L is the number of partitions of the amplitudes in window w. Then the entropy can be calculated by denoting $P^n(I_i)$ the probability that the signal $s(i) \in W(n; w; \Delta)$ belongs to the interval I_i. This

0-7803-7011-2/01/$10.00 ©2001 IEEE

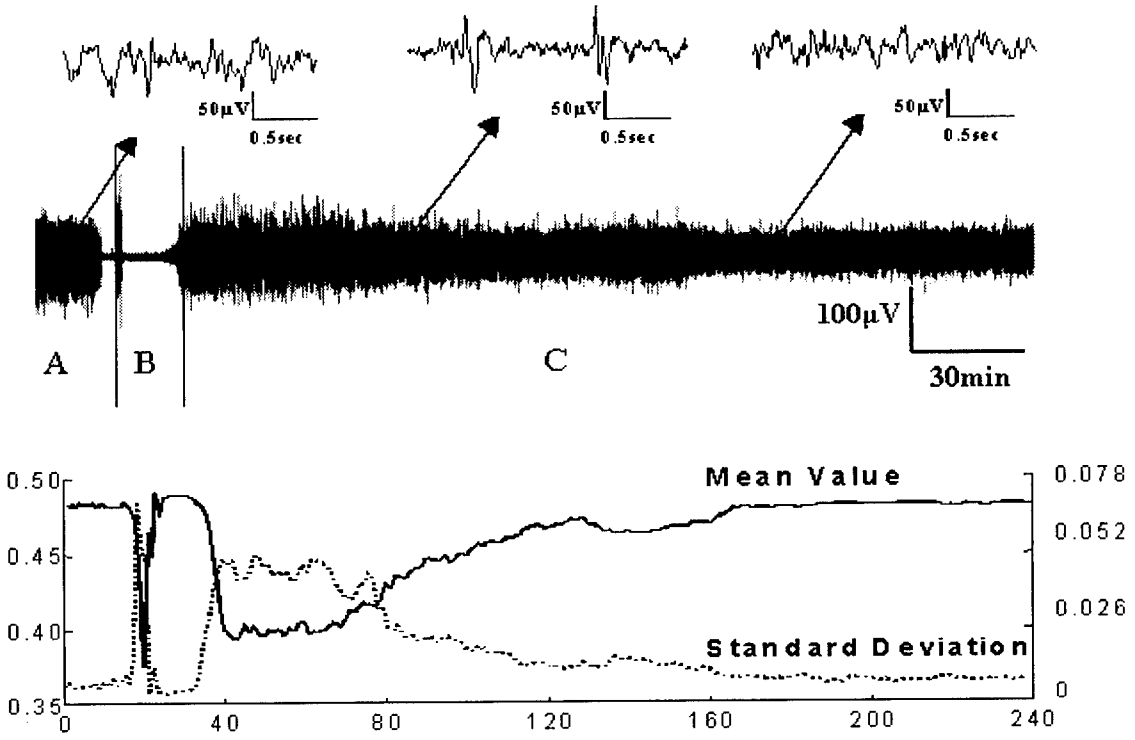


Fig.1. The mean value (MEAN) and standard deviation (SD) of 1min TE over time for an experiment with 5min of asphyxia (bottom panel) . The plots show the characteristic electrophysiological pattern of the EEG during the experiment (mid pannel). The MEAN reflects the changes in EEG spontaneous activity while the SD imprints the spike activity (top panel).

probability is the ratio between the number of *s(i)*-values of $W(n;w;\Delta)$ found within interval I_i and the total number of *s(i)*-values in $W(n;w;\Delta)$. By sliding the window *W*, we can explore the entropy evolution of the whole data $\{s(k): k=1,\ldots,N\}$ with Eqs. (1) and (4):

$$TDE\ (n) = -\sum_{i=1}^{L} p^n(I_i)\ln\ p^n(I_i) \qquad (6)$$

for Shannon, and

$$TDE_r(n) = -(q-1)^{-1}\sum_{i-1}^{L}\left[(p^n(I_i))^r - p^n(I_i)\right] \qquad (7)$$

for Tsallis entropy respectively.

Motivated by the belief that brain injury, such as caused by global ischemia from cardiac arrest, results in a reduction in the entropy of brain rhythm Bezerianos et al [7] calculated Shannon and Tsallis TDE in a group of animals recovered from brain asphyxia [8]. Their findings are in agreement with those of Martin et al [9] and even more they proved that they could be used for monitoring the recovery from brain asphyxia. The mean value (MEAN) and standard deviation (SD) of Tsallis TDE entropy was calculated every 1 min for a period of 4 hr (Fig 1). It seems that the MEAN and/or the SD can be used for quantitative assessment of brain recovery as in the past the cepstral distance has been used [10].

The problem of information flow in the study of brain dynamics is an essential one. Vastano and Swinney first applied the notion of mutual information (MI) to study the dynamics of spatiotemporal systems [11]. MI detects linear and nonlinear statistical dependencies between time series, whereas the more standard correlation function measures only their linear dependence. The MI between measurement x_i generated from system

X and measurement y_j generated from system Y is the amount of information that measurement x_i provides about y_j. Thus, MI is a measure of dynamical coupling or information transmission between X and Y, and when applied to EEG it may be postulated to be one measure of functional connectivity [12]. If one system is completely independent of another, then and only then MI between the time series generated from these dynamical systems is zero. The spatiotemporal relationships of multichannel EEG recordings to measure the information transition between various cortical areas have been studied in normals under differents physiological stages e.g awakening, sleep and simple arithmetic tasks [13] [14] in epilepsy [15] [14] [13] and mental diseases [16].

Xu et al [14] computed the MI between eight EEG channels and found that the differences between the waking state with open eyes and during sleep are very significant. However, the subjects with their eyes closed and light sleep display similar variations. In most cases showed large fluctuations that gradually decreased with time. In most cases a maximum peak in the MI appears in the time span 0-500ms [16] which can be considered as the time period after, the information generated in any position within the brain, has reached in every other place. The average MI $I_{X(t)Y(t+\tau)}$ (where τ is the time delay) between all electrodes over a time span of 500 ms were calculated to represent the information transmission across different cortical areas in normal subjects and Alzheimer disease patients. The average MI distribution is nearly symmetric, suggesting the presence of fast bidirectional transmission of information between brain areas. The MI in Alzheimer disease is lower than in normal controls suggesting the association of EEG abnormalities in Alzheimer disease patients with functional impairment of information transmission in long cortico-cortical connections.

Another EEG information study is using the information distance measure. In the research of cardiac arrest asphyxia, during the hypoxia and asphyxia phases, we are not only interested in the evolution of brain electric activity on each lead, but also in the relation between different sites. A time dependent relative entropy distance measure based on Kullback-Leibler entropy is a powerful solution to the problem. It is different from mutual information, which is based on conditional probability. Fig.2 shows entropy distance evolution of experimental EEG which including the (a) baseline, (b) hypoxia, (c) global asphyxia and early recovery, (d) later recovery. The information distances between each segment with the baseline EEG is clearly illustrated in the figure.

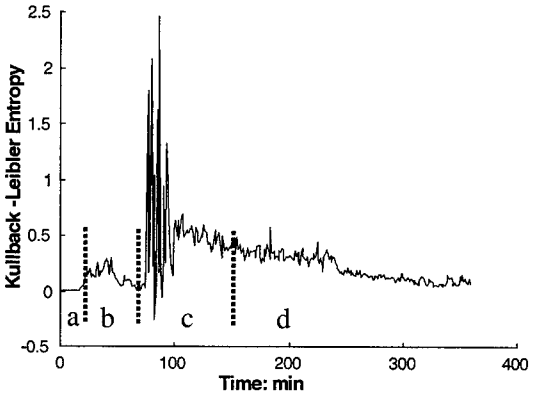


Fig. 2. Kullback-Leibler entroy of preconditioning rat. The baseline EEG was chosen to be the reference. There are evident increases during the preconditioning hypoxia and recovery phases. During the asphyxia, the ECG and artifacts dominate the EEG recordings. The fluctuations right after the asphyxia in the K-L entropy is due to the heart rate variation.

3. Discussion

We showed that the severity and the progression of cerebral ischemic injury can be evaluated by TDE of neuroelectrical activity in experimental brain animal models. We applied the same methodology in determining the outcome in human subjects after cardiac arrest and we hypothesized that TDE can be used in determining injury severity and outcome in human subjects. The use of the method for analysis of EEG in epileptic discharges is promising as also has been pointed out by others. In conclusion the Tsallis entropy is a non redundant information measure of brain dynamics and its application in different areas of interest is promising.

Acknowledgements: This work was supported by a grant NS24282 from teh NIH. The authors thank Drs D. Hanley and R. Geocadin for their collaborative efforts.

4. References

[1] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst Tech J*, vol. 27, pp. 623-656, 1948.

[3] S. Kullback, *Information Theory and Statistics*. New York: Dover Books, 1968.

[4] C. Tsallis, "Possible generalization of Boltmannn-Gibs statistics," *J Statistical Physics*, vol. 52, pp. 479-487, 1988.

[5] L. G. Gamero, A. Plastino, and M. E. Torres, "Wavelet analysis and nonlinear dynamics in a nonextensive setting," *Physica A*, vol. 246, pp. 487-509, 1997.

[6] M. Torres and L. Gamero, "Relative complexity changes in times series using information measures," *Physica A*, vol. 286, pp. 457-473, 2000.

[7] A. Bezerianos, S. Tong, A. Malhorta, and N. Thakor, "Information measures of Brain Dynamics," presented at Nonlinear Signal and Image Processing, Baltimore, USA, 2001.

[8] R. Geocadin, R. Ghodadra, K. Kimura, H. Lei, D. Sherman, D. Hanley, and N. Thakor, "A novel quantitative EEG injury measure of global cerebral ischemia," *Clinical Neurophysiology*, vol. 111, pp. 1779-1787, 2000.

[9] M. T. Martin, A. R. Plastino, and A. Plastino, "Tsallis-like information measures and the analysis of complex signals," *Physica A*, vol. 275, pp. 262-271, 2000.

[10] R. Geocadin, J. Muthuswamy, D. Sherman, N. Thakor, and D. Hanley, "Early electrophysiological and histologic changes after global cerebral ischemia in rats," *Movement Disorders*, vol. 15, pp. 14-21, 2000.

[11] J. A. Vastano and H. L. Swinney, "Information transport in spatiotemporal systems," *Physical Review Letters*, vol. 60, pp. 1773- 1776, 1988.

[12] O. Sporns, G. Tononi, and G. Edelman, "Connectivity and complexity: the relation between neuroanatomy and brain dynamics," *Neural Networks*, vol. 13, pp. 909-922, 2000.

[13] F. Chen, J. Xu, F. Gu, X. Yu, X. Meng, and Z. Qiu, "Dynamic process of information flow transmission complexity in human brains," *Biological Cybernetics*, vol. 83, pp. 355-366, 2000.

[14] J. Xu, Z. Liu, R. Liu, and Q. Yang, "Information transmission in human cerebral cortex," *Physica D*, vol. 106, pp. 363-374, 1997.

[15] N. K. Varma, R. Kushwaha, A. Beydoun, W. J. Williams, and I. Drury, "Mutual information analysis and detection of interictal morphological differences in interictal epileptiform discharges of patients with partial epilepsy," *Electroenceph Clin Neurophysiol*, vol. 103, pp. 426-433, 1997.

[16] J. Jeong, J. Gore, and B. Peterson, "Mutual Information analysis of the EEG in patients with Alzheimer' s disease,'*Clinical Neurophysiology*, vol. 112, pp. 827-835, 2001.

BLIND IDENTIFICATION AND EQUALIZATION OF MINIMUM-PHASE CHANNELS

Senjian An

Dept. of Electrical and Electronic Engineering
The University of Melbourne
Victoria 3010, Australia
senjian@ee.mu.oz.au

Yingbo Hua

Dept. of Electrical Engineering
University of California
Riverside CA 92521, USA
yhua@ee.ucr.edu

ABSTRACT

In this paper, a second order statistics based technique of blind identification and equalization is proposed for minimum phase channels driven by stochastically independent colored signals. Sufficient identifiability conditions are given. Unlike most existing blind identification methods, this method does not require the number of sensors to be greater than the number of source signals. Simulation result is given to demonstrate the performance of the proposed algorithm.

1. INTRODUCTION AND PROBLEM FORMULATION

Blind identification and equalization of FIR (finite impulse response) and MIMO (multi input and multi output) channels driven by colored signals are a fundamental problem in a wide range of applications such as speech enhancement, wireless communications and brain signal analysis. The existing works on FIR MIMO channels driven by colored signals include the subspace method [1], the matrix pencil method [2], the blind identification via decorrelating subchannels (BIDS) method [3,5] and blind identification via decorrelating the whole channel (BIDW) method [6]. These methods require the channel matrix to be irreducible and/or the output signal number is greater than the input signal number. In this paper, we will develop a blind method to identify square minimum-phase FIR channels.

Consider a FIR MIMO channel described by

$$\mathbf{y}(n) = \mathbf{H}(n) * \mathbf{x}(n) + \mathbf{w}(n) \doteq \sum_{l=0}^{q} \mathbf{H}(l)\mathbf{x}(n-l) + \mathbf{w}(n) \tag{1}$$

where $*$ denotes convolution, $\mathbf{x}(n), \mathbf{y}(n)$ are the sequences of the input and output vector of dimension m, $\mathbf{H}(n)$ is the sequence of the system's impulse response matrix of dimension $m \times m$, q is the length of the system's finite impulse response, and $\mathbf{w}(n)$ is the noise vector. An equivalent form of (1) is:

$$\mathbf{y}(n) = \mathbf{H}_z(z)\mathbf{x}(n) + \mathbf{w}(n) \tag{2}$$

where $\mathbf{H}_z(z) = \sum_{l=0}^{q} \mathbf{H}(l)z^{-l}$, which is the channel operator and also referred to as the channel matrix.

We assume that there are sufficient data so that the second-order statistics (SOS) of $\mathbf{y}(n)$ can be exploited. Then, we can write the autocorrelation function of $\mathbf{y}(n)$ as:

$$\mathbf{C}_{\mathbf{yy}}(\tau) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{y}(n)\mathbf{y}^T(n-\tau)$$

and the power spectral matrix of $\mathbf{y}(n)$ as:

$$\mathbf{S}_{yy}(z) = \sum_{\tau=-\infty}^{\infty} \mathbf{C}_{\mathbf{yy}}(\tau)z^{-\tau}$$

$$= \mathbf{H}_z(z)\mathbf{S}_{\mathbf{xx}}(z)\mathbf{H}_z^T(z^{-1}) + \mathbf{S}_{\mathbf{ww}}(z) \tag{3}$$

where all notations are defined in an obvious way. The above formulation assumes that the noise $\mathbf{w}(n)$ is uncorrelated with the desired signal $\mathbf{x}(n)$.

The aim here is to estimate the channel $\mathbf{H}_z(z)$ and/or recover the input signals using the power spectral matrix of $\mathbf{y}(n)$.

Our proposed method includes two main steps: 1) blind signal separation, which aims to find the separator $\mathbf{G}(z)$ such that $\mathbf{G}(z)\mathbf{H}(z) = diag$; 2) channel estimation and signal recovery, which aims to compute the channel from $\mathbf{G}(z)\mathbf{H}(z) = diag$ and recover the original input signals. The next two sections will deal with these two steps. In section 4, computer simulation results are shown to verify the effectiveness of the proposed algorithm.

2. BLIND SIGNAL SEPARATION

In this section, we apply the separation algorithm developed in [6] on the square channel case. In order to present the method clearly, we need introduce the concept *diversity* defined in [6]. Let $\mathbf{h}(z) = [h_1(z), h_2(z), \cdots, h_m(z)]^T$ be a polynomial vector with dimension m, and its greatest common divisor(GCD) be $c(z)$, then the mth diversity of $\mathbf{h}(z)$, denoted by $div_m(\mathbf{h}(z))$, is defined as

$$div_m(\mathbf{h}(z)) = deg(\mathbf{h}(z)) - deg(c(z)) \tag{4}$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

For $k = 2, 3, \cdots, m-1$, the kth diversity of $\mathbf{h}(z)$, denoted by $div_k(\mathbf{h}(z))$, is defined as the minimum of the kth diversity of any k dimensional subvectors of $\mathbf{h}(z)$, i.e.,

$$div_k(\mathbf{h}(z)) = \min \left\{ div_k \left([h_{i_1}(z), h_{i_2}(z), \cdots, h_{i_m}(z)]^T \right) : \right.$$
$$\left. 1 \le i_1 < i_2 < i_k \le m \right\} \quad (5)$$

Correspondingly, for the power spectral matrix, denoted by $\mathbf{S_{xx}}(z) = diag(s_1(z), s_2(z), \cdots, s_m(z))$, of m independent colored signals, the kth diversity of $\mathbf{S_{xx}}(z)$, denoted by $div_k(\mathbf{S_{xx}})$, is defined as half of the kth diversity of the polynomial vector $\mathbf{s}(z) = [s_1(z), s_2(z), \cdots, s_m(z)]^T$. $div_2(\mathbf{S_{xx}})$ is the diversity introduced in [3].

Now we present two technical Lemmas, which are crucial for the development of the separation algorithm:

Lemma 1. *Let $\mathbf{H}(z)$ be a $m \times m$ polynomial matrix with degree q, $\mathbf{S_{xx}}(z)$ be the diagonal input power spectra with $div_2(\mathbf{S_{xx}}(z)) > mq$ and $\mathbf{g}(z) = [g_1(z), g_2(z), \cdots, g_m(z)]^T$ be a polynomial vector with $deg(\mathbf{g}(z)) \le (m-1)q$. Then*

$$div_m[\mathbf{g}(z)^T \mathbf{S_{yy}}(z)] \le q \quad (6)$$

if and only if $\mathbf{g}^T(z)\mathbf{H}(z)$ has only one nonzero element.

Proof: Note that $\mathbf{S_{xx}}(z)$ is diagonal, the sufficiency is obvious. Now we show the necessity. Assume that $\mathbf{g}^T(z)\mathbf{H}(z)$ has L non-zero elements, we only need to show, if $L \ge 2$,

$$div_m(\mathbf{g}^T(z)\mathbf{S_{yy}}(z)) > q \quad (7)$$

Denote

$$\mathbf{g}^T(z)\mathbf{H}(z) = [c_1(z), c_2(z), \cdots, c_m(z)] \quad (8)$$

$$\mathbf{S_{xx}}(z) = diag\,(s_1(z), s_2(z), \cdots, s_m(z)) \quad (9)$$

$$\mathbf{d}(z) \doteq [c_1(z)s_1(z), c_2(z)s_2(z), \cdots, c_m(z)s_m(z)]^T \quad (10)$$

Then

$$\mathbf{g}^T(z)\mathbf{S_{yy}}(z) = \mathbf{d}(z)\mathbf{H}^T(z^{-1}) \quad (11)$$

Note that $s_i(z), 1 \le i \le m$ are double-side polynomials and $deg(c_i(z)) \le mq$, we have

$$div_m(\mathbf{d}(z)) \ge 2div_2(\mathbf{S_{xx}}(z)) - mq$$

Since any $L \times m$ submatrix of $\mathbf{H}^T(z^{-1})$ has at most Lq zeros (including the infinite), we have

$$div_m(\mathbf{g}(z)^T \mathbf{S_{yy}}(z)) \ge div_m(\mathbf{d}(z)) + q - Lq > q$$

The proof is completed.

This result is rather conservative. If any digonal function of $\mathbf{S_{xx}}(z)$ does not share common zeros with the other diagonals, which happens in most cases, the *diversity* condition can be much weaker.

Lemma 2. *Let $\mathbf{H}(z)$ be a $m \times m$ polynomial matrix with degree q, $\mathbf{g}(z) = [g_1(z), g_2(z), \cdots, g_m(z)]^T$ be a polynomial vector with $deg(\mathbf{g}(z)) \le (m-1)q$, $\mathbf{S_{xx}}(z)$ be the diagonal input power spectra with $div(\mathbf{S_{xx}}(z)) > (0.5m+1)q$ and any two diagonal elements of $\mathbf{S_{xx}}(z)$ do not share common zeros. Then*

$$div_m[\mathbf{g}(z)^T \mathbf{S_{yy}}(z)] \le q$$

if and only if $\mathbf{g}^T(z)\mathbf{H}(z)$ has only one nonzero element.

Proof: The proof is similar to that of Lemma 1. Assume that $\mathbf{g}^T(z)\mathbf{H}(z)$ has L non-zero elements, we only need to show, if $L \ge 2$,

$$div_m(\mathbf{g}^T(z)\mathbf{S_{yy}}(z)) > q$$

Since any two of $s_i(z), 1 \le i \le m$ do not share common zeros, we have

$$div_m(\mathbf{d}(z)) \ge 2div_2(\mathbf{S_{xx}}(z)) - \frac{mq}{L-1}$$

Note that $L \times m$ submatrix of $\mathbf{H}^T(z^{-1})$ has at most Lq zeros (including infinite), it follows

$$div_m(\mathbf{g}(z)^T \mathbf{S_{yy}}(z)) \ge div(\mathbf{d}(z)) + q - Lq > q$$

and then the proof is completed.

Remarks. 1). The result of Lemma 2 is still conservative. If $\mathbf{S_{xx}}(z)$ and $\mathbf{H}(z)$ are any fixed matrices, the identification condition can be further relaxed.

2). The channel degree may be unknown and it can be identified by minimizing q under the condition $\mathbf{g}(z)$ with degree $(m-1)q$ exists such that $div_m[\mathbf{g}(z)^T \mathbf{S_{yy}}(z)] \le q$.

It is known that there exists $\mathbf{G}(z)$ with degree $(m-1)q$ such that $\mathbf{G}(z)\mathbf{H}(z)$ is diagonal. If the input power spectra satisfy the conditions in Lemma 1 or Lemma 2, we can find the separator

$$\mathbf{G}(z) = [\mathbf{g}_1(z) \quad \mathbf{g}_2(z) \quad \cdots \quad \mathbf{g}_m(z)]^T$$

by searching for $\mathbf{g}_i(z)$ such that

$$div_m(\mathbf{g}_i^T(z)\mathbf{S_{yy}}(z) \le deg(\mathbf{H}(z)), i = 1, 2, \cdots, m \quad (12)$$

In [6], an efficient algorithm was proposed to find $\mathbf{g}_i(z)$ satisfying (12). Applying this algorithm, we can find the separator $\mathbf{G}(z)$ such that $\mathbf{G}(z)\mathbf{H}(z)$ is diagonal. In the following section, we will show that the channel can be computed directly from $\mathbf{G}(z)\mathbf{H}(z) = diag$.

3. CHANNEL ESTIMATION AND SIGNAL DECONVOLUTION

The following Lemma shows the channels can be identified up to scaling and permutation once its separator is obtained.

Lemma 3. Given a nonsingular polynomial matrix $\mathbf{G}(z)$, there exists a unique column-wise coprime polynomial matrix $\mathbf{H}(z)$ (up to column scaling) such that $\mathbf{G}(z)\mathbf{H}(z)$ is diagonal.

Proof: Suppose $\mathbf{G}(z)\mathbf{H}(z) = \Gamma(z)$ where $\mathbf{H}(z)$ is column-wise coprime and $\Gamma(z)$ is diagonal. Then

$$\mathbf{H}(z) = (\mathbf{G}(z))^{-1}\Gamma(z) = adj(\mathbf{G}(z))\frac{\Gamma(z)}{det(G(z))}$$

Denote $adj(\mathbf{G}(z)) = M(z)D(z)$ where $D(z)$ is a diagonal polynomial matrix and $M(z)$ is a column-wise coprime polynomial matrix. Then we have

$$\mathbf{H}(z) = M(z)\frac{\Gamma(z)D(z)}{det(\mathbf{G}(z))}$$

Note that $\mathbf{H}(z)$ is polynomial and $M(z)$ is column-wise coprime, $\frac{\Gamma(z)D(z)}{det(H(z))}$ must be a scalar diagonal matrix and the proof is completed.

Hence, if the channel is column-wise coprime, then from the separator $\mathbf{G}(z)$, we can compute the channel $\mathbf{H}(z)$. Now we present a time domain computation method.

Denote $l_g = (m-1)q$, $\mathbf{G}(z) = \sum_{k=0}^{l_g} \mathbf{G}_k z^{-k}$ and

$$\mathbf{W} = \begin{bmatrix} \mathbf{G}_0 & & \\ \mathbf{G}_1 & \ddots & \\ \vdots & \ddots & \mathbf{G}_0 \\ \mathbf{G}_{l_g} & \ddots & \mathbf{G}_1 \\ & \ddots & \vdots \\ & & \mathbf{G}_{l_g} \end{bmatrix} \in R^{m(q+1)\times m(q+l_g+1)}.$$

Let $\mathbf{W}_k (k = 1, 2, \cdots, m)$ equals $\mathbf{W}$ by deleting all its $((i-1)m+k)$th rows $(i = 1, 2, \cdots, l_g+q+1)$. Then from $\mathbf{G}(z)\mathbf{H}(z) = diag$, it follows

$$\mathbf{W}_k \mathbf{h}_k = 0$$

where $k = 1, 2, \cdots, m$ and $\mathbf{h}_k$ is the kth column of

$$\begin{bmatrix} \mathbf{H}_0 \\ \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_q \end{bmatrix}$$

If all the columns of $\mathbf{H}(z)$ have the same degree, then the solution of the above equation is unique (up to constant scaling) and we can get the channel parameters directly. If $\mathbf{H}(z)$ has different column degrees, some equations may have a solution space. Take any one solution and we can formulate a column polynomial $h(z)$, then excluding the common factor, the remainer is the corresponding column of $\mathbf{H}(z)$.

Once the channel $\mathbf{H}(z)$ is obtained, exclude the common row factors of $\mathbf{G}(z)$ and we get a row-wise coprime $\bar{\mathbf{G}}(z)$. Compute $\bar{\mathbf{G}}(z)\mathbf{H}(z) = diag(d_1(z), d_2(z), \cdots, d_m(z))$. Since $\bar{\mathbf{G}}(z)$ is row-wise coprime, any zeros of any $d_i(z)$ must be a zero of $\mathbf{H}(z)$. Hence all $d_i(z)$ are minimum phase if the channel is minimum phase and the source signals $\mathbf{x}_i(n), i = 1, 2, \cdots, m$ can be recovered uniquely (up to scaling and permutation) by the deconvolution of $\mathbf{u}_i(n) \doteq \mathbf{g}_i(z)\mathbf{y}(n) = d_i(z)\mathbf{x}_j(n)$ from $d_i(z)$. Another recovery method is to compute the inverse of the estimated channel $\hat{\mathbf{H}}(z)$ and recover the input signals by computing $\hat{\mathbf{x}}(n) = \hat{\mathbf{H}}(z)^{-1}\mathbf{y}(n)$ directly.

4. SIMULATIONS

In order to show how well this method performs, we consider a 3×3 FIR channel $\mathbf{H}(z)$ of degree 1 driven by three real speech signals (three sentences from the Linguistic Data Consortium): 1). "She had your dark suit in greasy wash water all year"; 2). "Don't ask me to carry on oily lag like that"; 3) "Draw every outer line first, then fill in the interior". Each sentence has about 46797 samples under sample rate 16000 Hz. In our simulation example, the first 30000 samples of them are applied. In order to guarantee the channel to be minimum-phase, the channel is selected in the following way: randomly select two nonsingular matrices H_0, D and let $\mathbf{H}(z) = \mathbf{H}_0\left(\mathbf{I} + \frac{\mathbf{D}}{2*\sigma_{max}(\mathbf{D})}z^{-1}\right)$, where $\sigma_{max}(\mathbf{D})$ denotes the maximal singular value of $\mathbf{D}$. The channel selected in our example is

$$\mathbf{H}(z) = \begin{bmatrix} 0.2241 & -1.5004 & 0.1029 \\ 0.3229 & 0.9436 & 0.4934 \\ -0.2951 & 1.8035 & 0.4264 \end{bmatrix} + \begin{bmatrix} -0.3148 & 0.0609 & 0.4503 \\ -0.0659 & 0.1353 & -0.2142 \\ 0.2936 & 0.0972 & -0.4530 \end{bmatrix} z^{-1}$$

Fig. 1-3 show the performance of the algorithm. The channel estimation relative error is 0.0591. The three original, mixed and recovered signals are shown in Figure 1, Figure 2 and Figure 3 respectively. The recovery is nearly perfect except for a slight noise on the first recovered signal.

5. CONCLUSION

In this paper, a second order statistics based blind system identification method for square minimum-phase channels has been developed. This method requires the channel to be column-wise coprime and the input power spectra are sufficiently diverse. Unlike most existing blind identification methods, this method does not require the output signal

number to be greater than the input signal number. Computer simulation has been shown to verify the effectiveness of the proposed algorithm. But the present algorithm is not very robust. The future work is to analyze the performance robustness and to improve the proposed algorithm to be more robust against noise and other possible uncertainties.

Acknowledgment

This work has been supported by the Australian Research Council.

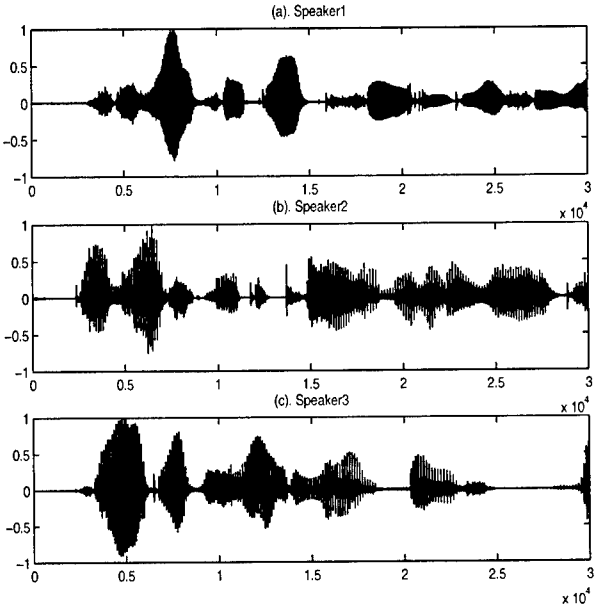


Fig. 1 The original speech signals

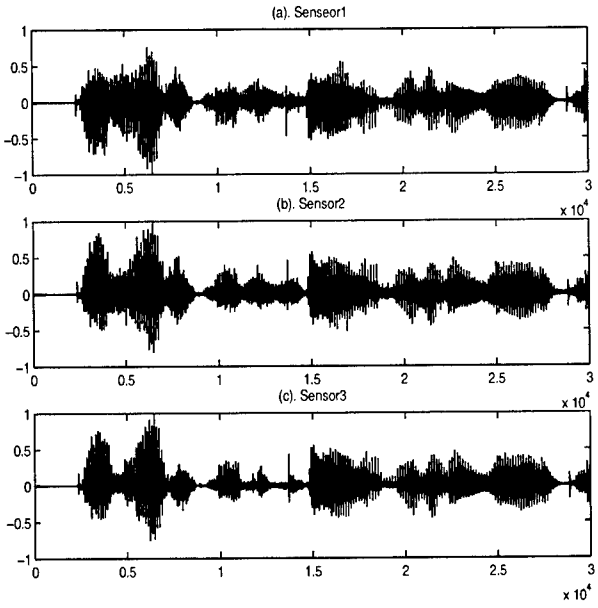


Fig. 2 The mixed speech signals

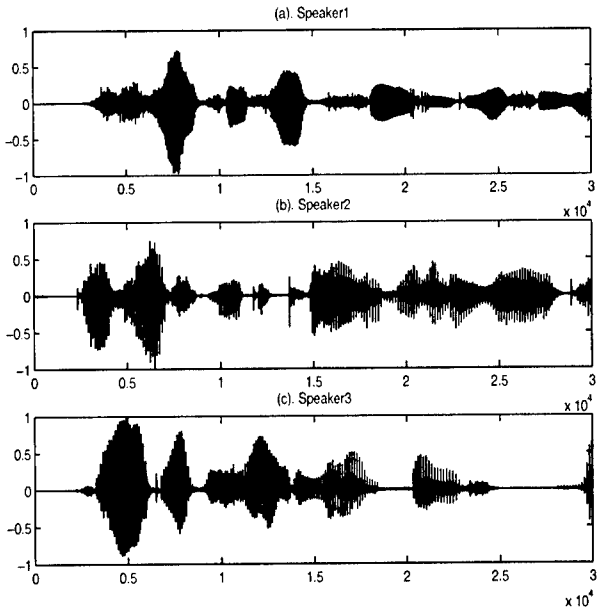


Fig. 3 The recovered speech signals

6. REFERENCES

[1] Gorokhov, A., and Loubaton, P., "Subspace-based techniques for blind separation of convolutive mixtures with temporally correlated sources", *IEEE Transactions on Circuits and Systems - I*, Vol. 44, No.9, pp. 813–820, Sept 1997.

[2] Ma, t., Ding, Z., and Yau, S. F., " A two-stage algorithm for MIMO blind deconvolution of colored input signals", *IEEE Transactions on Signal Processing*, Vol. 48, No.4, April 2000.

[3] Hua, Y., An, S. and Xiang, Y., "Blind identification and equalization of FIR MIMO channels driven by colored signals", *Submitted to IEEE Transactions on Signal Processing, Sept. 2000.*

[4] Hua, Y. and Tugnait, J., "Blind identifiability of FIR-MIMO Systems with colored input using second order statistics", *IEEE Signal Processing Letters* pp. 348-350, Vol. 7, No. 12, Dec. 2000.

[5] Hua, Y., An, S. and Xiang, Y., "Blind identification and equalization of FIR MIMO channels by BIDS", *Proc. of IEEE ICASSP'2001*, Salt Lake City, May 2001.

[6] An, S. and Hua, Y. "Blind signal separation and blind system identification of irreducible MIMO Channels", To appear in *Proceeding of Sixth International Symposium on Signal Processing and its Applications*, Kuala-Lumpur, Malaysia, Aug. 2001.

ON BLIND EQUALIZATION OF RANK DEFICIENT NONLINEAR CHANNELS

Roberto López-Valcarce

Dept. Tecnologías de las Comunicaciones
Universidad de Vigo, 36200 Vigo, Spain
valcarce@gts.tsc.uvigo.es

Soura Dasgupta

Dept. Electrical & Computer Engineering
University of Iowa, 52242 IA, USA
dasgupta@engineering.uiowa.edu

ABSTRACT

We consider the problem of blind equalization of nonlinear channels from the second-order statistics of the channel output. The channel model is linear in the parameters, with additive terms that are nonlinear functions of the transmitted symbols. All previous approaches assume that the corresponding channel matrix has full column rank, which ensures the existence of linear FIR zero forcing equalizers. We show that this assumption is not necessary, and that under certain circumstances linear FIR equalizers can be found despite the violation of this assumption. An important consequence of this fact is that equalization can be effected with a smaller level of diversity. In this paper necessary and sufficient conditions on the channel matrix are given. An algorithm for the computation of the equalizers is also given for those channels satisfying these conditions, assuming an i.i.d. symbol sequence and memory dominance of the linear part.

1. INTRODUCTION

Recently blind equalization of single-input multiple-output (SIMO) channels has received considerable attention, due to the fact that these channels can be perfectly equalized if the equalizer is long enough and the subchannels are coprime. This equalizer can be obtained from the second-order statistics (SOS) of the received signal [7].

With a few exceptions [1, 6, 9], almost all the available literature on blind equalization is devoted to the linear channel case. However, many real world communication systems, such as radio links with high power amplifiers, high-density magnetic and optical storage channels, etc., exhibit a considerable degree of nonlinearity. Thus it is of interest to consider blind equalization of *nonlinear channels*. Our 1-input, p-output channel model has the form

$$y(k) = \sum_{i=1}^{q} \sum_{j=0}^{l_i} h_{ij} s_i(k-j) + n(k), \qquad (1)$$

where $s_1(k) = a(k)$ is the scalar, stationary input, the terms $s_i(k) = f_i(a(k), a(k-1), \cdots)$ for $i = 2, \ldots, q$ are scalar nonlinear causal functions of $a(\cdot)$, h_{ij} are $p \times 1$ coefficient vectors, and $n(k)$, $y(k)$ are $p \times 1$ signal vectors representing an additive disturbance and the observed signal, respectively. $n(\cdot)$ and $a(\cdot)$ are assumed independent. This model accommodates polynomial approximations of nonlinear channels (Volterra models), but the 'basis functions' $s_i(\cdot)$ need not be monomials in principle.

Supported in part by NSF grants CCR-9973133 and ECS-9970105

We are interested in equalizer design for the class of channels (1) using only the SOS of $y(\cdot)$. As shown in [1], under certain conditions linear finite impulse response (FIR) filters can perfectly equalize nonlinear SIMO channels of the type (1). For those cases, [1] presented a blind, deterministic approach for equalizer design. However it has been shown in [2] that the conditions in [1] are in fact conservative. More general sufficient conditions on the channel and the input signal statistics for SOS-based blind equalizability were presented in [3].

Observe that (1) could be seen as a linear multiple-input multiple-output (MIMO) system if we regard the nonlinear terms $s_i(\cdot)$ as additional inputs. However, standard SOS-based equalization techniques for MIMO systems usually assume that the different inputs are uncorrelated (which is no longer true in our setting), and they only resolve the inputs to within a mixing matrix [7]. In addition, in our case only the term $s_1(\cdot)$ is of interest.

All previous approaches [1, 2, 3] assume that the so-called *channel matrix* constructed from the channel coefficients has full column rank. In that case linear FIR equalizers always exist. However, a consequence is that the number of subchannels required by these schemes must exceed the number of distinct kernels in (1). This level of diversity may at times be unacceptably high. In an earlier paper, [4], we had shown that in a linear multi-user multichannel setting, this full column rank condition can be relaxed, and a lower level of diversity can be tolerated. In particular suppose in (1) the $s_i(k)$ are independent users, and the goal is only to equalize $s_1(k)$. Then [4] gives a necessary and sufficient condition for equalization of that $s_1(k)$. Clearly this same condition will also ensure the existence of a linear FIR equalizer for the nonlinear setting of this paper. Should l_1 exceed all other l_i, *and the $s_i(k)$ are white and mutually uncorrelated* then [4] also provides an algorithm that permits the construction of the equalizer from the out put SOS alone. *However, in the nonlinear setting one cannot assume that the $s_i(k)$ are mutually uncorrelated even if $a(k)$ is white, as $s_i(k)$ are nonlinear functions of $a(k)$. Thus the algorithm of [4] cannot be applied to nonlinear channels.* The key contribution of this paper is to formulate an algorithm that provides the required equalizer from the output SOS, provided the equalizability condition of [4] are met, and even if the $s_i(k)$ are statistically dependent. This algorithm assumes that the memory of the nonlinear part is strictly less than that of the linear part. Simulation results are given as evidence of the feasibility of this procedure.

In our notation, $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^\#$ denote transpose, conjugate transpose and pseudoinverse respectively; I_n, J_n denote respectively the $n \times n$ identity matrix and the shift matrix with ones in the first subdiagonal and zeros elsewhere, and e_n denotes the n-th

0-7803-7011-2/01/$10.00 ©2001 IEEE

unit vector.

2. CONDITIONS FOR THE EXISTENCE OF LINEAR FIR ZF EQUALIZERS

By stacking m consecutive samples of $y(\cdot)$ into

$$Y(k)^T = [\ y(k)^T \quad y(k-1)^T \quad \cdots \quad y(k-m+1)^T\ ].$$

one gets

$$Y(k) = \mathcal{H}S(k) + N(k). \tag{2}$$

with $N(k)^T = [\ n(k)^T \quad n(k-1)^T \quad \cdots \quad n(k-m+1)^T\ ]$, $S(k)^T = [\ S_1^T(k) \quad S_2^T(k)\ ]$ the noise and signal vectors,

$$S_1^T(k) = [\ a(k) \quad \cdots \quad a(k-l_1-m+1)\ ]. \tag{3}$$

$$S_2^T(k) = [\ s_2(k) \quad \cdots \quad s_2(k-l_2-m+1) \quad | \quad \cdots \quad | \quad s_q(k) \quad \cdots \quad s_q(k-l_q-m+1)\ ]. \tag{4}$$

and the channel matrix $\mathcal{H} = [\ \mathcal{H}_1 \quad \mathcal{H}_2 \quad \cdots \quad \mathcal{H}_q\ ]$, with every $\mathcal{H}_i$ block Toeplitz:

$$\mathcal{H}_i = \begin{bmatrix} h_{i0} & \cdots & h_{il_i} & & \\ & \ddots & & \ddots & \\ & & h_{i0} & \cdots & h_{il_i} \end{bmatrix} \qquad pm \times (m+l_i).$$

For convenience, let $d_1 = m + l_1$, which is the size of $S_1(k)$, the linear part of the regressor; and $d_2 = l_2 + \cdots + l_q + (q-1)m$, which is the size of $S_2(k)$ (thus $S(k)$ is $(d_1+d_2)\times 1$).

Observe that if the channel matrix $\mathcal{H}$ has full column rank, then its pseudoinverse $\mathcal{H}^{\#}$ satisfies $\mathcal{H}^{\#}\mathcal{H} = I_{d_1+d_2}$. In that case, in the noiseless case ($N(k) = 0$) one obtains from (2) $\mathcal{H}^{\#}Y(k) = S(k)$. Thus the first d_1 rows of $\mathcal{H}^{\#}$ provide zeroforcing (ZF) equalizers with associated delays 0 through $d_1 - 1$. However, this also shows the existence of vectors (the last d_2 rows of $\mathcal{H}^{\#}$) that recover all the nonlinear terms $s_i(k)$ and their delays, which is clearly not necessary since these terms are of no interest to the receiver. This leads us to ask for necessary and sufficient conditions on $\mathcal{H}$ for the ZF equalizers to exist. First, let us introduce the following partition of the channel matrix:

$$\mathcal{H} = [\ \mathcal{H}_1 \quad \mathcal{H}_{nl}\ ] \quad \text{with} \quad \mathcal{H}_{nl} = [\ \mathcal{H}_2 \quad \cdots \quad \mathcal{H}_q\ ]. \tag{5}$$

That is, $\mathcal{H}_{nl}$ comprises the 'nonlinear part' of the channel matrix. Recall that $\mathcal{H}_1$ and $\mathcal{H}_{nl}$ have sizes $pm \times d_1$ and $pm \times d_2$ respectively. We shall make the following assumption:

A1: $\mathcal{H}_1$ has full column rank, and with $r_1 = \operatorname{rank}(\mathcal{H}_1)$, $r_2 = \operatorname{rank}(\mathcal{H}_{nl})$, $\mathcal{H}$ satisfies $\operatorname{rank}(\mathcal{H}) = r_1 + r_2 \le pm$.

Observe that if $\mathcal{H}$ has full column rank, then Assumption **A1** is satisfied but not conversely. The significance of this condition is reflected in the following result from [4]:

Theorem 1 *There exists a $pm \times d_1$ matrix $\mathcal{G}$ such that*

$$\mathcal{G}^H\mathcal{H} = [\ I_{d_1} \quad 0_{d_1\times d_2}\ ] \tag{6}$$

if and only if Assumption **A1** *holds.*

The columns of $\mathcal{G}$ constitute the desired ZF equalizers. The geometrical interpretation of Theorem 1 is as follows.

Lemma 1 *The condition* $\operatorname{rank}([\ \mathcal{H}_1 \quad \mathcal{H}_{nl}\ ]) = \operatorname{rank}(\mathcal{H}_1) + \operatorname{rank}(\mathcal{H}_{nl})$ *is equivalent to* $\operatorname{range}(\mathcal{H}_1) \bigcap \operatorname{range}(\mathcal{H}_{nl}) = \{0\}$, *with* $\operatorname{range}(A)$ *the subspace spanned by the columns of* A.

Thus linear FIR ZF equalizers exist iff $\mathcal{H}_1$ has full column rank and no nonzero vector lies in the range space of both $\mathcal{H}_1$ and $\mathcal{H}_{nl}$.

3. SOS-BASED EQUALIZER DESIGN

We turn our attention now to the problem of extracting the equalizers from the SOS of the received signal, assuming that $\mathcal{H}$ satisfies the relaxed rank condition **A1**. From (2), the covariance of the received vector $Y(\cdot)$ is given by

$$C_y(l) = \operatorname{cov}[Y(k), Y(k-l)] = \mathcal{H}C_s(l)\mathcal{H}^H + C_n(l). \tag{7}$$

with $C_s(l) = \operatorname{cov}[S(k), S(k-l)]$, $C_n(l) = \operatorname{cov}[N(k), N(k-l)]$ the signal and noise covariance matrices. In addition to **A1**, we adopt the following standard assumptions:

A2: $n(\cdot)$ is zero-mean, white, with covariance $\sigma_n^2 I_p$.

A3: The covariance matrix $C_s(0)$ is positive definite.

Observe that [4] assumes that $C_s(0)$ is diagonal. This assumption is not needed here. Under **A1** and **A2**, σ_n^2 can be estimated as the smallest eigenvalue of $C_y(0)$. Thus the effect of the noise can be removed from $C_y(l)$; henceforth we shall assume that $C_y(l) = \mathcal{H}C_s(l)\mathcal{H}^H$. **A3** is a 'persistent excitation' condition on $a(\cdot)$, which allows us to write

$$C_s(0) = QQ^H \quad \text{with } Q \text{ invertible.} \tag{8}$$

Now let Q be a square root of $C_s(0)$ as in (8), and define the *normalized* channel and source covariance matrices respectively as

$$H = \mathcal{H}Q. \qquad \bar{C}_s(l) = Q^{-1}C_s(l)Q^{-H}. \tag{9}$$

Using (9), the matrices $C_y(l)$ become

$$C_y(l) = H\bar{C}_s(l)H^H. \qquad \text{with } \bar{C}_s(0) = I_{d_1+d_2}. \tag{10}$$

The following result relates the ZF equalizers to the normalized channel matrix H.

Lemma 2 *Under* **A1-A3**, *let the square root of $C_s(0)$, Q, be block lower triangular:*

$$Q = \begin{bmatrix} Q_{11} & 0 \\ Q_{21} & Q_{22} \end{bmatrix}. \qquad \textit{with } Q_{ij} \textit{ of size } d_i \times d_j. \tag{11}$$

Then the matrix $\mathcal{G}$ satisfying (6) (ZF equalizers) is given by

$$\mathcal{G}^H = Q_{11}[\ I_{d_1} \quad 0_{d_1\times d_2}\ ]H^{\#}. \tag{12}$$

Thus if $H = U_1\Sigma V^H$ is an SVD of H, with U_1: $pm \times (r_1+r_2)$, Σ: $(r_1+r_2)\times(r_1+r_2)$, V: $(r_1+r_2)\times(d_1+d_2)$, and partitioning V as

$$V = [\ V_1 \quad V_2\ ]. \qquad V_i \text{ of size } (r_1+r_2)\times d_i, \tag{13}$$

then the equalizers are given by

$$\mathcal{G}^H = Q_{11}V_1^H\Sigma^{-1}U_1^H. \tag{14}$$

Observe that Q_{11} is known to us from the source statistics, and that Σ, U_1 can be obtained from an SVD of $C_y(0)$ since

$$C_y(0) = \mathcal{H}C_s(0)\mathcal{H}^H = HH^H = U_1\Sigma^2U_1^H. \tag{15}$$

Therefore if V_1 could be somehow estimated, the ZF equalizers could be computed. Note that $VV^H = V_1V_1^H + V_2V_2^H = I_{r_1+r_2}$. An additional property is shown by the next result.

Lemma 3 *Under* **A1-A3**, *let Q be block lower triangular as in (11). Let $H = \mathcal{H}Q$ have a singular value decomposition $H = U_1 \Sigma V$, and partition V as in (13). Then*

$$V_1^H V_1 = I_{d_1}, \qquad V_1^H V_2 = 0_{d_1 \times d_2}. \tag{16}$$

This property is obvious in the full column rank case (for which V is square), but it is somewhat surprising that it still holds even under the relaxed rank condition **A1**. Now consider the matrix

$$R(1) = \Sigma^{-1} U_1^H C_y(1) U_1 \Sigma^{-1}. \tag{17}$$

which satisfies $R(1) = V \bar{C}_s(1) V^H$. From (16), this gives

$$R(1)V_1 = V\bar{C}_s(1)\begin{bmatrix} I_{d_1} \\ 0 \end{bmatrix}, \quad R^H(1)V_1 = V\bar{C}_s^H(1)\begin{bmatrix} I_{d_1} \\ 0 \end{bmatrix}. \tag{18}$$

These relations will allow us to estimate V_1 under the following additional assumptions:

A4: The symbol sequence $a(\cdot)$ is a zero-mean i.i.d. process with $\mathrm{cov}[a(k), a(k)] = \sigma_a^2$.

A5: $S_2(k)$ satisfies $S_2(k) = f(a(k), a(k-1), \ldots, a(k-d_1+2))$, with $f(\cdot, \cdot, \ldots, \cdot)$ a memoryless mapping.

Basically **A5** amounts to saying that the memory of the nonlinear part of the channel is strictly shorter than that of the linear part. One has the following result:

Lemma 4 *Under* **A1-A5**, *a lower block triangular square root Q as in (11) exists such that $Q_{11} = \sigma_a^2 I_{d_1}$ and*

$$\bar{C}_s(1) = \begin{bmatrix} J_{d_1} & 0 \\ 0 & C \end{bmatrix} \text{ for some } d_2 \times d_2 \; C, \tag{19}$$

$$\bar{C}_s(d_1 - 1) = e_{d_1} e_1^H. \tag{20}$$

Substituting (19) in (18) one obtains the Jordan chains

$$R(1)V_1 = V_1 J_{d_1}, \qquad R^H(1)V_1 = V_1 J_{d_1}^H. \tag{21}$$

which show how V_1 can be estimated once its first or last column is available. Partition $V_1 = [\, v_1 \;\; v_2 \;\; \cdots \;\; v_{d_1} \,]$ columnwise, and consider the matrix

$$R(d_1 - 1) = \Sigma^{-1} U_1^H C_y(d_1 - 1) U_1 \Sigma^{-1}, \tag{22}$$

which satisfies $R(d_1 - 1) = V \bar{C}_s(d_1 - 1) V^H$. Using (20),

$$R(d_1 - 1) = V e_{d_1} e_1^H V^H = v_{d_1} v_1^H. \tag{23}$$

Thus $R(d_1 - 1)$ is a rank one matrix and its only nonzero singular value equals 1. The vectors v_1, $v_{d_1 - 1}$ can be obtained up to a constant of the form $e^{j\theta}$ from an SVD of $R(d_1 - 1)$, or alternatively they can be estimated as

$$\hat{v}_{d_1} = \frac{R(d_1 - 1) e_{i_{\max}}}{\| R(d_1 - 1) e_{i_{\max}} \|}, \quad \hat{v}_1^H = \frac{e_{j_{\max}}^H R(d_1 - 1)}{\| e_{j_{\max}}^H R(d_1 - 1) \|}, \tag{24}$$

where

$$i_{\max} = \arg\max \{ \| R(d_1 - 1) e_i \|, \; 1 \le i \le d_1 + d_2 \},$$
$$j_{\max} = \arg\max \{ \| e_j^H R(d_1 - 1) \|, \; 1 \le j \le d_1 + d_2 \}.$$

In this way these estimates $\hat{v}_{d_1}$, $\hat{v}_1$ are related to the true quantities v_{d_1}, v_1 by some complex constants with unit modulus. From these, the remaining columns of V_1 can be estimated via either of the Jordan chains (21), thus obtaining an estimate $\hat{V}_1$ satisfying $\hat{V}_1 = e^{j\theta} V_1$ for some real θ. Therefore the matrix $\hat{\mathcal{G}}_{\mathrm{ZF}} = \sigma_a U_1 \Sigma^{-1} \hat{V}_1$ satisfies $\hat{\mathcal{G}}_{\mathrm{ZF}}^H \mathcal{H} = e^{j\theta} I_{d_1}$, providing equalization up to an unknown phase rotation. This is acceptable since the need for a phase reference can be sidestepped by differentially encoding the data. Finally, it is possible to obtain the Minimum Mean-Squared Error (MMSE) equalizers in the spirit of [5]:

Lemma 5 *Under Assumptions* **A1-A3**, *the MMSE equalizers $\mathcal{G}_{\mathrm{MMSE}}$ minimizing* $\mathrm{trace}\, E[|\mathcal{G}^H Y(k) - S_1(k)|^2]$ *are related to the ZF equalizers by*

$$\mathcal{G}_{\mathrm{MMSE}} = [I - \sigma_n^2 C_y^{-1}(0)] \mathcal{G}_{\mathrm{ZF}} \tag{25}$$

where now $C_y(0) = \mathcal{H} C_s(0) \mathcal{H}^H + \sigma_n^2 I_{pm}$ represents the unde-noised *channel output covariance matrix.*

The resulting algorithm is summarized next.

Blind equalization algorithm

1. Compute estimates $\hat{C}_y(0)$, $\hat{C}_y(1)$, $\hat{C}_y(d_1 - 1)$.
2. Estimate $\hat{\sigma}_n^2$ as the smallest eigenvalue of $\hat{C}_y(0)$ and substract the noise effect from $\hat{C}_y(\cdot)$.
3. Perform an SVD of $\hat{C}_y(0)$ as in (15) to obtain U_1, Σ.
4. Compute $R(1)$, $R(d_1 - 1)$ as in (17), (22) respectively.
5. Form the estimates $\hat{v}_{d_1}$, $\hat{v}_1$ via (24).
6. For $i = 2, 3, \ldots, d_1$, let $\hat{v}_i = R(1)\hat{v}_{i-1}$. Alternatively, for $j = d_1, d_1 - 1, \ldots, 2$, let $\hat{v}_{j-1} = R^H(1)\hat{v}_j$.
7. ZF equalizers: $\hat{\mathcal{G}}_{\mathrm{ZF}} = \sigma_a U_1 \Sigma^{-1} [\, \hat{v}_1 \;\; \cdots \;\; \hat{v}_{d_1} \,]$.
8. Compute the MMSE equalizers via (25).

4. SIMULATION RESULTS

We present now a numerical example of the results obtained by the algorithm. For illustration purposes, the phase ambiguity inherent to the method was removed before computing the error rates. Averages were computed based on 100 independent runs.

The channel we consider is real with $q = 3$, $l_1 = 4$, $l_2 = l_3 = 1$ and i.i.d. symbols taking the values ± 1 with equal probabilities. The number of subchannels is $p = 4$; the coefficients are given in table 1. The nonlinear terms are $s_2(k) = a(k)a(k-1)$, $s_3(k) = a(k)a(k-2)$. The resulting linear-to-nonlinear distortion ratio for this channel is 8 dB. The equalizer length that we consider is $m = 6$. The corresponding channel matrix $\mathcal{H}$ (of size 24×24) is not full column rank but it satisfies the relaxed rank condition **A1**:

$$23 = \mathrm{rank}(\mathcal{H}) = \mathrm{rank}(\mathcal{H}_1) + \mathrm{rank}([\, \mathcal{H}_2 \;\; \mathcal{H}_3 \,]) = 10 + 13.$$

Note that this channel satisfies **A5**, and that σ_n^2 can still be estimated as the smallest eigenvalue of $C_y(0)$ even though the channel matrix $\mathcal{H}$ is square, since $\mathcal{H}$ is rank deficient. Figure 1(a) shows the symbol error rate (SER) vs. SNR using $K = 2000$ samples for covariance estimation, while figure 1(b) shows the variation of the SER with K for a fixed value of SNR = 24 dB, for the equalization delays 0, 3, 8 and 9. In this case the equalizer with maximal delay ($d = 9$) provides the poorest performance of all. The best results are obtained with the equalizer of delay $d = 3$.

channel	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{20}	h_{21}	h_{30}	h_{31}
1	1.0	0.5	0.4	0.2	-0.2	0.2	0.5	0.1	-0.1
2	0.1	0.6	1.0	-0.4	0.2	0.1	0.25	0.2	-0.2
3	-0.2	0.6	0.6	0.1	-0.3	0.2	0.5	0.2	-0.2
4	0.3	1.0	0.7	-0.5	0.2	0.1	0.25	0.1	-0.1

Table 1: Coefficients of the Volterra channel used in the simulations

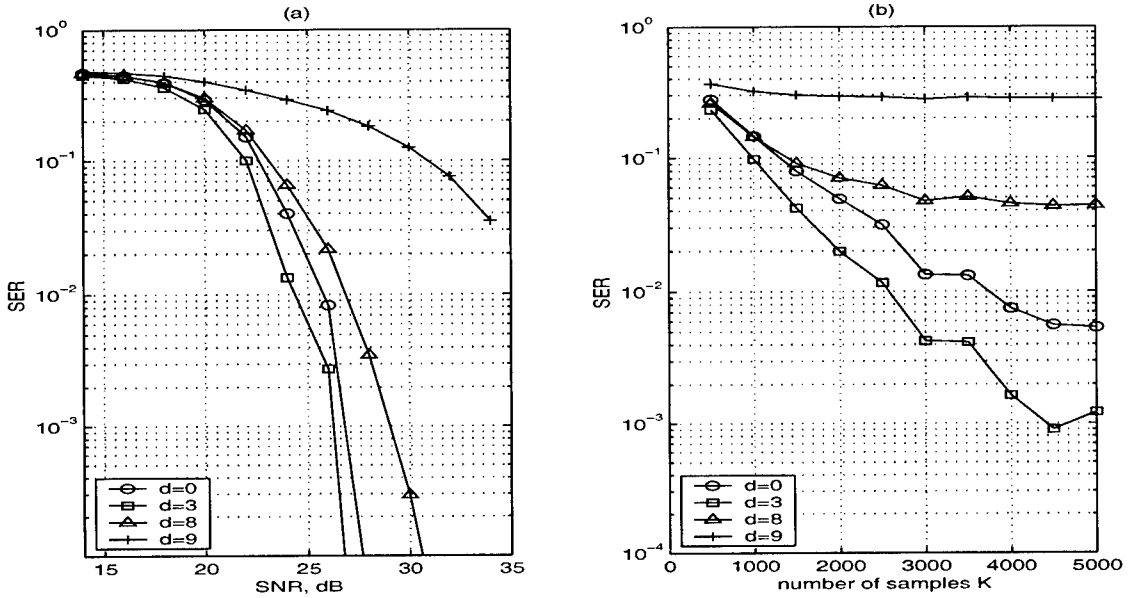


Figure 1: MMSE equalizer performance, $m = 6$. (a) SER vs. SNR, $K = 2000$ symbols. (b) SER vs. sample size K, SNR = 24 dB.

5. CONCLUSIONS

In contrast with the linear channel case, for equalizability of nonlinear channels with linear FIR filters it is not necessary that the channel matrix have full column rank. We have given necessary and sufficient conditions on the channel matrix for this property to hold. If in addition the input symbol sequence is i.i.d. and the memory of the nonlinear part of the channel is strictly shorter than that of the linear part, a blind algorithm based on the second-order statistics of the channel output provides the equalizers.

6. REFERENCES

[1] G. B. Giannakis and E. Serpedin, "Linear multichannel blind equalizers of nonlinear FIR Volterra channels", *IEEE Trans. Signal Processing*, vol. 45 no. 1, pp. 67-81, Jan. 1997.

[2] R. López-Valcarce and S. Dasgupta, "Blind identifiability/equalizability of single input multiple output nonlinear channels from second order statistics", *Proc. 2000 IEEE ICASSP*, Istanbul, Turkey.

[3] R. López-Valcarce and S. Dasgupta, "The role of second-order statistics in blind equalization of nonlinear systems", *Proc. 2000 IEEE SSAP Workshop*, pp. 211-215, Pocono Manor, PA.

[4] Z. Ding, S. Dasgupta and R. López-Valcarce, "Interference cancellation and blind equalization for linear multi-user systems", *Proc. 2000 IEEE ICASSP*, Istanbul, Turkey.

[5] C. Papadias and D. T. M. Slock, "Fractionally spaced equalization of linear polyphase channels and related blind techniques based on multichannel linear prediction", *IEEE Trans. Signal Processing*, vol 47 no. 3, pp. 641-654, March 1999.

[6] G. M. Raz and B. D. Van Veen, "Blind equalization and identification of nonlinear and IIR systems–A Least Squares approach", *IEEE Trans. Signal Processing*, vol. 48 no. 1, pp. 192-200, January 2000.

[7] L. Tong and S. Perreau, "Multichannel blind identification: from subspace to maximum likelihood methods", *Proc. IEEE*, vol. 86 no. 10 , pp. 1951-1968, Oct. 1998.

[8] L. Tong, G. Xu and T. Kailath, "Blind identification and equalization based on second-order statistics: a time-domain approach", *IEEE Trans. Information Theory*, vol. 40, no. 2, pp. 340-350, March 1994.

[9] M. Tsatsanis and H. Cirpan, "Blind identification of nonlinear channels excited by discrete alphabet inputs", *Proc. 1996 IEEE SSAP Workshop*, vol. 1, pp. 176-179, Corfu, Greece.

MULTICHANNEL BLIND DECONVOLUTION OF COLORED SIGNALS VIA EIGENVALUE DECOMPOSITION

Pando Georgiev[†] *and Andrzej Cichocki*[‡]

Brain Science Institute, RIKEN, Wako-shi, Saitama 351-01, Japan

[†]-On leave from the So£a University "St. Kl. Ohridski", Bulgaria

E-mail: georgiev@bsp.brain.riken.go.jp.

[‡]-On leave from the Warsaw University of Technology, Poland

E-mail: cia@bsp.brain.riken.go.jp.

ABSTRACT

We prove that a MIMO (multiple input multiple output) blind deconvolution problem for n colored uncorrelated signals can be converted to n SIMO (single input multiple output) problems, using eigenvalue decomposition of a special covariance matrix, depending on L-dimensional parameter $\mathbf{b}$, if appropriate covariance matrices have sets of eigenvalues with empty pairwise intersection. We present a suf£cient condition for this conversion and discuss how to £nd such parameters. We prove that the parameters $\mathbf{b}$ for which this is possible, form an open subset of $\mathbb{R}^L$, whose complement has a Lebesgue measure zero.

1. INTRODUCTION

The problems of independent component analysis (ICA), blind source separation (BSS) and multichannel blind deconvolution (MBD) of source signals have received wide attention in various £elds such as biomedical signal analysis and processing (EEG, MEG, ECG), geophysical data processing, data mining, speech and image recognition and enhancement and wireless communications. In such applications a number of observations are available, of signals or data that are £ltering superposition of separate signals from different independent sources, and it is desired to process the observations so that the outputs correspond to the separate primary source signals.

Acoustic applications include the signals from several microphones in a sound £eld that is produced by several speakers (the so-called cocktail-party problem) and the signals from several acoustic transducers in an underwater sound £eld from the engine noises of several ships (sonar problem). Radio and wireless communication examples include the observations corresponding to outputs of array antenna elements in response to several transmitters, and the observations may also include the effects of the mutual couplings of the elements. Other radio communication examples arise in the use of polarization multiplexing in microwave links; the maintenance of the orthogonality of the polarization cannot be perfect and there is interference between the separate transmissions. Radar examples include a superposition of signals from different target modulating mechanisms as observed by multiple receivers, such as elements sensitive to different polarizations.

To £nd the original sound source that was recorded with microphones in a conference room, we must cancel out, or deconvolve, the room impulse response from the original sound source. Since we have no prior knowledge of what this room impulse response is, we call this process the multichannel blind deconvolution or cocktail party problem.

Most of the existing algorithms for MBD assume that source signals are white and usually the additive noise is assumed negligible small (see for instance [4], [8], [9]).

The main objective of this paper is to present a procedure for conversion of a MIMO deconvolution problem to several SIMO deconvolution problems in presence of additive white noise.

We note that another idea for converting a MIMO problem into SIMO problems is contained in [3]. We refer to [7] and references therein for solving SIMO problems.

Here we develop an idea in [5], where a method for deconvolution of colored signals is presented, using matrix pencils.

2. PROBLEM FORMULATION

Consider a convoluted mixture $\mathbf{x}(k) = (x_1(k), ..., x_m(k))$ of uncorrelated colored source signals $s_i(k), i = 1, ..., n$ with $m > n$:

$$\mathbf{x}(k) = \sum_{p=0}^{M} \mathbf{H}(p)\mathbf{s}(k-p) + \mathbf{n}(k) \tag{1}$$

or

$$\mathbf{x}(z) = \mathbf{H}(z)\mathbf{s}(z) + \mathbf{n}(z) \tag{2}$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

where $\mathbf{H}(z) = \sum_{p=1}^{M} \mathbf{H}(p) z^{-p}$. We assume that the order M of the £lters is known or can be estimated.

The problem is to recover the original signals up to arbitrary scaling, permutation and delays.

We introduce the matrices $\mathbf{H}_{ij} \in \mathbb{R}^{(N+1)+(M+N+1)}$ by

$$\mathbf{H}_{ij} = \begin{bmatrix} h_{ij}(0)h_{ij}(1)... & h_{ij}(M) & 0... & 0 \\ 0 \quad h_{ij}(0)... & ... & h_{ij}(M)... & 0 \\ ... \quad . & . & . & . \\ ... \quad . & . & . & . \\ ... \quad . & . & . & . \\ 0 \quad . & 0 & h_{ij}(0)... & h_{ij}(M) \end{bmatrix},$$

where h_{ij} is the (i,j)-th element of the matrix $\mathbf{H}(p)$, and the matrix $\tilde{\mathbf{H}} \in \mathbb{R}^{m(N+1)\times n(M+N+1)}$ by

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12}... & \mathbf{H}_{1n} \\ \mathbf{H}_{21} & \mathbf{H}_{22}... & \mathbf{H}_{2n} \\ ... & . & . \\ ... & . & . \\ ... & . & . \\ \mathbf{H}_{m1} & \mathbf{H}_{m2}... & \mathbf{H}_{mn} \end{bmatrix}.$$

Denote:

$$\begin{aligned} \mathbf{s}_i(k) &= (s_i(k), s_i(k-1), ..., s_i(k-M-N))^T, \\ \tilde{\mathbf{s}}(k) &= (\mathbf{s}_1(k)^T, ..., \mathbf{s}_n(k)^T)^T, \\ \mathbf{x}_i(k) &= (x_i(k), x_i(k-1), ..., x_i(k-M))^T, \\ \tilde{\mathbf{x}}(k) &= (\mathbf{x}_1(k)^T, ..., \mathbf{x}_n(k)^T)^T. \\ \mathbf{n}_i(k) &= (n_i(k), n_i(k-1), ..., n_i(k-M))^T, \\ \tilde{\mathbf{n}}(k) &= (\mathbf{n}_1(k)^T, ..., \mathbf{n}_n(k)^T)^T. \end{aligned}$$

Then the convolution problem can be written as

$$\tilde{\mathbf{x}}(k) = \tilde{\mathbf{H}}\tilde{\mathbf{s}}(k) + \tilde{\mathbf{n}}(k). \tag{3}$$

Under the following assumptions the matrix $\tilde{\mathbf{H}}$ has full column rank (see for instance [2], [8] for a proof):

(H_1) $\mathbf{H}(z)$ is irreducible (i.e. rank$\mathbf{H}(z) = n, \forall z \neq 0, z = +\infty$);

(H_2) $\mathbf{H}(z)$ is column reduced (i.e. its highest column-degree coef£cient matrix has full rank).

The above assumptions are natural and used in many papers: see, for example, [4], [5], [8].

3. ROBUST ORTHOGONALIZATION

In this section we assume that $\tilde{\mathbf{H}}$ is nonsingular square matrix. This is true, if we have freedom to choose N and m such that $m(N+1) = n(M+N+1)$ and $(H_1), (H_2)$ to be satis£ed.

We use a preprocessing procedure, which is not sensitive to the white noise $\mathbf{n}(k)$ and which allows us to de£ne a new orthogonal mixing matrix for the preprocesed data $\tilde{\mathbf{x}}$. The idea is to use time-delayed correlation matrices that are not sensitive to additive white noise and construct a positive de£nite matrix from their linear combination (for suf£ciently large number of samples), a problem solved in [1] for instanteneous mixures by a £nite-step global convergence algorithm [10].

Let us de£ne a a time-delayed correlation matrix of the vector $\tilde{\mathbf{x}}(k)$ by

$$\mathbf{R}_{\tilde{\mathbf{x}}}(p) = E\{\tilde{\mathbf{x}}(k)\tilde{\mathbf{x}}^T(k-p)\} \tag{4}$$

and a symmetric matrix $\tilde{\mathbf{R}}_{\tilde{\mathbf{x}}}(p)$ by

$$\tilde{\mathbf{R}}_{\tilde{\mathbf{x}}}(p) = \frac{1}{2}\left\{\mathbf{R}_{\tilde{\mathbf{x}}}(p) + \mathbf{R}_{\tilde{\mathbf{x}}}^T(p)\right\}. \tag{5}$$

Similarly we de£ne analogous matrices $\mathbf{R}_{\tilde{\mathbf{s}}}(p)$ and $\tilde{\mathbf{R}}_{\tilde{\mathbf{s}}}(p)$ for the source signals $\tilde{\mathbf{s}}(k)$.

The time-delayed correlation matrices of the observation vector $\mathbf{x}(k)$ for any $p \neq 0$ satisfy (due to the assumption of white noise) $\tilde{\mathbf{R}}_{\tilde{\mathbf{x}}}(p) = \tilde{\mathbf{H}}\tilde{\mathbf{R}}_{\tilde{\mathbf{s}}}(p)\tilde{\mathbf{H}}^T$.

The robust orthogonalization algorithm can be summarized as follows.

Algorithm Outline: Robust Orthogonalization

1. Find (by the method described in [1]), i.e. choose or estimate a set of parameters $\{\alpha_i\}_{i=1}^{K}$ such that the matrix $\mathbf{C}_{\tilde{\mathbf{x}}}(\boldsymbol{\alpha}) = \sum_{i=1}^{K} \alpha_i \tilde{\mathbf{R}}_{\tilde{\mathbf{x}}}(p_i)$ is positive de£nite.
2. Perform an eigenvalue-decomposition (EVD) of $\mathbf{C}_{\tilde{\mathbf{x}}}(\boldsymbol{\alpha})$, $\mathbf{C}_{\tilde{\mathbf{x}}}(\boldsymbol{\alpha}) = \mathbf{U}_{\tilde{\mathbf{x}}}\boldsymbol{\Lambda}_{\tilde{\mathbf{x}}}\mathbf{U}_{\tilde{\mathbf{x}}}^T$, where the entries of diagonal matrix $\boldsymbol{\Lambda}_{\tilde{\mathbf{x}}}$ are the positive eigenvalues of $\mathbf{C}_{\tilde{\mathbf{x}}}(\boldsymbol{\alpha})$ and compute the preprocessing matrix $\mathbf{Q} = \boldsymbol{\Lambda}_{\tilde{\mathbf{x}}}^{-\frac{1}{2}}\mathbf{U}_{\tilde{\mathbf{x}}}^T$.
3. Compute the preprocessed data $\mathbf{z}(k) = \mathbf{Q}\tilde{\mathbf{x}}(k) = \mathbf{Q}\tilde{\mathbf{H}}\mathbf{s}(k)$.

Remark 1 By de£ning a new mixing matrix as $\mathbf{A} = \mathbf{Q}\tilde{\mathbf{H}}\mathbf{D}^{\frac{1}{2}}$, where $\mathbf{D} = \sum_{i=1}^{K} \alpha_i \tilde{\mathbf{R}}_{\tilde{\mathbf{s}}}(p_i)$ is a block diagonal positive de£nite matrix, it is easy to show that $\mathbf{C}_{\mathbf{z}}(\boldsymbol{\alpha}) = \mathbf{A}\mathbf{A}^T = \mathbf{I}_r$ ($(r \times r)$ unit matrix, $r = m(N+1)$), so $\mathbf{A}$ is orthogonal. This orthogonality condition is necessary for performing conversion to SIMO deconvolution problems using symmetric EVD. It should be noted that in contrast to the standard prewhithening procedure, for our robust orthogonalization generally $E\{\mathbf{z}\mathbf{z}^T\} \neq \mathbf{I}_r$. Also, we have $\mathbf{z}(k) = \mathbf{A}\bar{\mathbf{s}}(k) + \bar{\mathbf{n}}(k)$, where $\bar{\mathbf{s}}(k) = \mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{s}}(k)$, $\bar{\mathbf{n}}(k) = \mathbf{Q}\tilde{\mathbf{n}}(k)$, so $\bar{\mathbf{s}}(k)$ are £ltered (distorted) versions of the source signals $\mathbf{s}(k)$.

Remark 2 It is easy to see that the function φ_K which assighs to every $\boldsymbol{\alpha} \in \mathbb{R}^K$ the minimal eigenvalue of the matrix $\mathbf{C}_{\tilde{\mathbf{x}}}(\boldsymbol{\alpha})$, is concave, so point 1 in the above algorithm can be realized by any algorithm which searches for a maximum (which is global) of φ_K. The robust orthogonalization is possible, if the maximum value of φ_K is positive (which is not known a priori).

4. EXTRACTION OF FILTERED (DISTORTED) VERSIONS OF THE INPUT SIGNALS BY A SYMMETRIC EIGENVALUE PROBLEM

In this section we assume that robust orthogonalization is possible to be performed, so our model is $\mathbf{z}(k) = \mathbf{A}\mathbf{s}(k) + \overline{\mathbf{n}}(k)$ ($\overline{\mathbf{n}}(k) = \mathbf{Q}\tilde{\mathbf{n}}(k)$).

De£ne a covariance matrix of sensor signals by

$$\mathbf{R}_{\mathbf{z}}(p) = E\{\mathbf{z}(k)\mathbf{z}(k-p)^T\}$$

and similarly, a covariance matrix of source signals by

$$\mathbf{R}_{\mathbf{s}}(p) = E\{\mathbf{s}(k)\mathbf{s}(k-p)^T\}.$$

We recall that the source signals are *uncorrelated*, if $\mathbf{R}_s(p)$ are diagonal matrices for every p. If the source signals are statistically independent, then this condition is satis£ed, but the converse assertion is not always true. We say that the sources are *colored* if for some $p_0 \geq 1$ the matrix $\mathbf{R}_{\mathbf{s}}(p_0)$ is nonzero (diagonal) matrix.

For a vector $\mathbf{b} \in \mathbb{R}^L$ de£ne

$$\mathbf{Z}(\mathbf{b}) = \sum_{p=1}^{L} b_p \tilde{\mathbf{R}}_{\mathbf{z}}(p), \quad \mathbf{S}(\mathbf{b}) = \sum_{p=1}^{L} b_p \tilde{\mathbf{R}}_{\overline{\mathbf{s}}}(p). \qquad (6)$$

Then

$$\mathbf{Z}(\mathbf{b}) = \mathbf{A}\mathbf{S}(\mathbf{b})\mathbf{A}^T, \qquad (7)$$

and

$$\mathbf{Z}(\mathbf{b}) = \mathbf{A}diag\{\mathbf{S}_1(\mathbf{b}), ..., \mathbf{S}_n(\mathbf{b})\}\mathbf{A}^T, \qquad (8)$$

where $\mathbf{S}_i(\mathbf{b}) = \frac{1}{2}\sum_{p=1}^{L} b_p(E\{\overline{\mathbf{s}}_i(k)\overline{\mathbf{s}}_i(k-p)^T\} + E\{\overline{\mathbf{s}}_i(k-p)\overline{\mathbf{s}}_i(k)^T\}) \in \mathbb{R}^{M+N+1+M+N+1}$ is full matrix. Note that the matrix $\mathbf{S}_i(\mathbf{b})$ is symmetric and each diagonal of it has equal elements.

Let $\mathbf{V}(\mathbf{b})$ be a set of $n(M+N+1)$ orthonormal eigenvectors of the matrix $\mathbf{Z}(\mathbf{b})$. Denote by $\mathbf{L}_i(\mathbf{b})$ the set of all eigenvalues of $\mathbf{S}_i(\mathbf{b})$ and by $\mathbf{v}_{i,j}(\mathbf{b}), j = 1, ..., M+N+1$ these eigenvectors in $\mathbf{V}(\mathbf{b})$, which correspond to the eigenvalues from the set $\mathbf{L}_i(\mathbf{b})$.

We introduce the following condition:

$$\mathbf{L}_i(\mathbf{b}) \cap \mathbf{L}_j(\mathbf{b}) = \emptyset \quad \forall i \neq j. \quad (\mathbf{DEV}(\mathbf{b}))$$

Theorem 1 *Assume that condition* $(\mathbf{DEV}(\mathbf{b}))$ *is satis£ed for some* $\mathbf{b} \in \mathbb{R}^L$ *and the noise* $\mathbf{n}$ *is white. Then for any* $i = 1, ..., n$, *every signal*

$$y_{i,j}(k) = \mathbf{v}_{i,j}(\mathbf{b})^T\mathbf{z}(k), j = 1, ..., M+N+1$$

is a sum of £ltered (distorted) versions of the i-th signal s_i *plus noise* $n_{i,j}(k) = \mathbf{v}_{i,j}(\mathbf{b})^T\overline{\mathbf{n}}(k)$.

Proof. Let $\mathbf{v}_{i,j}(\mathbf{b})^T\mathbf{A} = (\mathbf{u}_{i,j,1}^T, ..., \mathbf{u}_{i,j,n}^T)$, where $\mathbf{u}_{i,j,l} \in \mathbb{R}^{M+N+1}$. We have

$$\mathbf{Z}(\mathbf{b})\mathbf{v}_{i,j}(\mathbf{b}) = \lambda\mathbf{v}_{i,j}(\mathbf{b}),$$

for some $\lambda \in \mathbf{L}_i(\mathbf{b})$. Hence, by (7), $\mathbf{S}(\mathbf{b})\mathbf{A}^T\mathbf{v}_{i,j}(\mathbf{b}) = \lambda\mathbf{A}^T\mathbf{v}_{i,j}(\mathbf{b})$, therefore, by (8), $\mathbf{S}_l(\mathbf{b})\mathbf{u}_{i,j,l} = \lambda\mathbf{u}_{i,j,l}$. By condition $(\mathbf{DEV}(\mathbf{b}))$ we obtain $\mathbf{u}_{i,j,l} = 0$ for $l \neq i$, therefore, for every $j = 1, ..., M+N+1$,

$$y_{i,j}(k) = \mathbf{v}_{i,j}(\mathbf{b})^T\mathbf{z}(k) = \mathbf{u}_{i,j,i}^T\overline{\mathbf{s}}_i(k) + n_{i,j}(k).$$

Since the components of the vector $\overline{\mathbf{s}}_i(k)$ are distorted versions of the original signal $s_i(k)$ (see Remark 1), their linear combinations are again distorted versions of the original signal $s_i(k)$, so the theorem is proved. ■

We introduce the following conditions for sources:

$$\forall i, j \neq i \quad \exists p_{i,j} \geq 1 :$$

$$E\{s_i(k)s_i(k-p_{i,j}) \neq E\{s_j(k)s_j(k-p_{i,j}) \quad (\mathbf{DAF})$$

i.e. the sources have different autocorrelation functions.

The following theorem is an extension (with more complicated proof) of that one contained in [6], which considers instantaneous mixtures.

Theorem 2 *Assume that condition* $(\mathbf{DAF})$ *is satis£ed. Then there exists* L *such that the condition* $(\mathbf{DEV}(\mathbf{b}))$ *is satis£ed for any* $\mathbf{b}$ *from an open subset* $B \subset \mathbb{R}^L$, *whose complement has a Lebesgue measure zero.*

Remark 3 The correlation matrices $E\{\mathbf{z}(k)\mathbf{z}(k-p)\}$ and consequently $\mathbf{Z}(\mathbf{b})$ are unbiased by the additive noise $\mathbf{n}(k)$ under condition that it is white (i.i.d.) and independent from the source signals.

5. EXTRACTION OF FILTERED (DISTORTED) VERSIONS OF THE INPUT SIGNALS BY A GENERALIZED EIGENVALUE PROBLEM

In this section we shall consider the case when the robust orthogonalization is not possible, so either the matrix $\tilde{\mathbf{H}}$ is nonsquare, or the functions φ_K (see Remark 2) has nonpositive maximum value.

Let $\mathbf{V}(\mathbf{b}, \mathbf{c})$ be a set of maximum number of unit linearly independent generalized eigenvectors of the matrix pencil $(\mathbf{Z}(\mathbf{b}), \mathbf{Z}(\mathbf{c}))$. Denote by $\mathbf{L}_i(\mathbf{b}, \mathbf{c})$ the set (possibly empty) of all generalized eigenvalues of the matrix pencil $(\mathbf{S}_i(\mathbf{b}), \mathbf{S}_i(\mathbf{c}))$ and by $\mathbf{v}_{i,j}(\mathbf{b}, \mathbf{c}), j = 1, ..., m_i$ the set of these eigenvectors in $\mathbf{V}(\mathbf{b}, \mathbf{c})$, which correspond to the eigenvalues from $\mathbf{L}_i(\mathbf{b}, \mathbf{c})$.

Theorem 3 *Assume that the condition*

$$\mathbf{L}_i(\mathbf{b}, \mathbf{c}) \cap \mathbf{L}_j(\mathbf{b}, \mathbf{c}) = \emptyset \quad \forall i \neq j \qquad (\mathbf{DEV}(\mathbf{b}, \mathbf{c}))$$

is satis£ed for some vectors $\mathbf{b} \in \mathbb{R}^L, \mathbf{c} \in \mathbb{R}^L$. *Then for any* $i = 1, ..., n$, *every signal*

$$y_{i,j}(k) = \mathbf{v}_{i,j}(\mathbf{b}, \mathbf{c})^T \mathbf{z}(k), j = 1, ..., m_i$$

is a sum of £ltered (distorted) versions of the i-th signal s_i *plus noise* $n_{i,j}(k) = \mathbf{v}_{i,j}(\mathbf{b}, \mathbf{c})^T \bar{\mathbf{n}}(k)$.

Proof. Let $\mathbf{v}_{i,j}(\mathbf{b}, \mathbf{c})^T \tilde{\mathbf{H}} = (\mathbf{u}_{i,j,1}^T, ..., \mathbf{u}_{i,j,n}^T)$, where $\mathbf{u}_{i,j,l} \in \mathbb{R}^{M+N+1}$. We have

$$\mathbf{Z}(\mathbf{b})\mathbf{v}_{i,j}(\mathbf{b}, \mathbf{c}) = \lambda \mathbf{Z}(\mathbf{c})\mathbf{v}_{i,j}(\mathbf{b}, \mathbf{c}),$$

for some $\lambda \in \mathbf{L}_i(\mathbf{b}, \mathbf{c})$. Hence

$$\tilde{\mathbf{H}}(\mathbf{S}(\mathbf{b}) - \lambda \mathbf{S}(\mathbf{c}))\tilde{\mathbf{H}}^T \mathbf{v}_{i,j}(\mathbf{b}, \mathbf{c}) = 0.$$

Since $\tilde{\mathbf{H}}$ has full column rank,

$$\mathbf{S}_l(\mathbf{b})\mathbf{u}_{i,j,l} = \lambda \mathbf{S}_l(\mathbf{c})\mathbf{u}_{i,j,l} \ \forall l = 1, ..., n.$$

By condition $(\mathbf{DEV}(\mathbf{b}))$ we obtain $\mathbf{u}_{i,j,l} = 0$ for $l \neq i$, therefore, for every $j = 1, ..., m_i$,

$$y_{i,j}(k) = \mathbf{v}_{i,j}(\mathbf{b}, \mathbf{c})^T \mathbf{z}(k) = \mathbf{u}_{i,j,i}^T \bar{\mathbf{s}}_i(k) + n_{i,j}(k),$$

and the conclusion follows as in the proof of Theorem 1. ■

Theorem 4 *Assume that condition* $(\mathbf{DAF})$ *is satis£ed. Then there exists L such that the condition* $(\mathbf{DEV}(\mathbf{b}, \mathbf{c}))$ *is satis£ed for any* $(\mathbf{b}, \mathbf{c})$ *from an open subset* $B \subset \mathbb{R}^{2L}$, *whose complement has a Lebesgue measure zero.*

Remark 4 One situation when we can check whether the condition $(\mathbf{DEV}(\mathbf{b}, \mathbf{c}))$ is satis£ed, is when the corresponding generalized eigenvalues of the matrix pencil $(\mathbf{Z}(\mathbf{b}), \mathbf{Z}(\mathbf{c}))$ are distinct. This case is considered in [5] for a matrix pencil (for single delays, so in our presentation in (6) only one coef£cient b_p is nonzero). Our condition $(\mathbf{DEV}(\mathbf{b}, \mathbf{c}))$ includes, in particular, this case. When a robust orthogonalization is possible, the check of condition $(\mathbf{DEV}(\mathbf{b}))$ is straitghforward, and due to Theorem 2 (assuming that the condition $(\mathbf{DAF})$ is satis£ed), we can choose randomly vector $\mathbf{b}$ until this condition is satis£ed.

6. CONCLUSIONS

We have proved that a MIMO blind deconvolution problem can be converted to multiple SIMO blind deconvolution problems using either symmetric EVD (after robust ortogonalization when it is possible), or generalized eigenvalue problem for a matrix pencil. If we have freedom to choose large number of observations $m = n(M + 1)$, under some conditions, this conversion is possible by a $m \times m$ matrix, where n is the number of the observed signals and M is the numbers of the delays. In both cases our method is robust to additive white noise.

7. REFERENCES

[1] A. Belouchrani and A. Cichocki, "Robust whitening procedure in blind separation context". *Electronics Letters*, Vol. 36, No. 24, pp. 2050-2051, 2000.

[2] R. Bitmead, S. Kung, B.D.O. Anderson and T. Kailath, "Greatest common division via generalized Sylvester and Bezout matrices", *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 1043-1047, 1978.

[3] Y. Hua and K. Abed-Meriam "Blind identi£cation of colored signals distorted by FIR channels", *Proc. ICASSP* , pp. 3124-3127, 2000.

[4] A. Mansur, C. Jutten and P. Loubaton, " Adaptive subspace algorithm for blind separation of independent sources in convolutive mizture", *IEEE Transactions on signal processing*, Vol. 48, No.2, pp. 583-586, 2000.

[5] C.T. Ma, Z. Ding and S.F. Yau, "A Two Stage Algorithm for MIMO Deconvolution of Nonstationary Colored Signals", *IEEE Trans. Signal Processing*, Vol. 48, No.4, pp. 1187-1192, 2000.

[6] P. Gr. Georgiev and A. Cichocki, "Blind Source Separation via Symmetric Eigenvalue Decomposition", ISSPA 2001.

[7] A. Gorokhov, "Robust blind second-order deconvolution", *IEEE Signal Proc. Letters*, Vo. 6, No. 1, pp. 13-16, 1999.

[8] D. Slock, "Blind fractionally-spaced equalization, perfect reconstruction £lter bankc and multichannel linear prediction", in *Proc. ICASSP*, Vol. 4, pp. 585-588, 1994.

[9] J.K. Tugnait and B. Huang, "On a whitening approach to partial channel estimation and blind equalization of FIR/IIR multiple-input multiple-output channels," *IEEE Trans. Signal Processing*, vol. SP-48, pp. 832-845, 2000.

[10] L. Tong, Y. Inouye and R. Liu, " A £nite-step global algorithm for the parameter estimation of multichannel MA processes". *IEEE Trans. Signal Processing*, Vol. 40, No. 10, pp. 2547-2558, 1992.

A SECOND-ORDER STATISTICS-BASED OPTIMIZATION APPROACH FOR BLIND MIMO SYSTEM IDENTIFICATION

*Ivan Bradaric, Athina P. Petropulu and Konstantinos I. Diamantaras**

Dept. of Electrical and Computer Engineering,
Drexel University, Philadelphia, PA 19104
ivan@cbis.ece.drexel.edu athina@artemis.ece.drexel.edu

*Department of Informatics, Technological Education Institute
Sindos, GR-54101, Greece
kdiamant@it.teithe.gr

ABSTRACT

We consider the problem of identifying a Multiple-Input Multiple-Output (MIMO) finite impulse response system excited by colored inputs with known statistics. Among other applications this problem appears in the context of CDMA communications systems with spatial and temporal diversity. We propose a novel approach that optimizes a criterion involving spectra and cross-spectra of the system output. Simulation results indicate that the proposed scheme works well, even for large order systems, and is robust to noise and channel length mismatch.

1. INTRODUCTION

The blind identification of a $m \times n$ Multiple-Input Multiple-Output (MIMO) system is of great importance in many applications, such as communications, biomedical engineering, seismology, etc.. The goal of blind system identification is to identify an unknown system $\mathbf{H}(z)$, driven by n unobservable inputs, based on the m system outputs ($n \leq m$), and subsequently use the system estimate to recover the input signals (sources).

In this paper we deal with the case of $m \times n$ MIMO system with colored inputs. Many of the existing methods address the problem using higher-order statistics [8], [10], [12], [14], [4], [2]. There are, however, a few methods that under certain conditions, address the problem using second-order statistics only [6], [11], [16], [7], [3]. Antenna-array CDMA system with spatial and temporal diversity can be formulated as MIMO system, where the system describes multipath and the input statistics depend on the user codes, which are known. Several algorithms have been proposed that take advantage of that knowledge [13], [15], [9].

Most of these methods are based on the time-domain analysis and depend on channel length information. In [3], [5] a method was proposed that uses frequency domain second-order correlations to recover the system frequency response within a frequency dependent phase ambiguity diagonal matrix. The advantage of a frequency domain approach for channel estimation is low sensitivity to channel length mismatch. In this paper, like in [5] we employ spectrum and cross-spectrum operations, but rather than using singular value decomposition, we optimize a criterion involving the aforementioned quantities. Our experiments indicated that the proposed approach results in much better channel estimates in terms of overall normalized mean-square error (ONMSE), while it does not yield phase ambiguity.

This work was supported by NSF under grant MIP-9553227.

2. PROBLEM FORMULATION

Let us consider an $m \times n$ FIR MIMO system with colored inputs. Let $\mathbf{e}(k) = [e_1(k) \cdots e_n(k)]^T$ be a vector of n statistically independent zero mean stationary sources, $\mathbf{h}(l)$ the impulse response matrix with elements $\{h_{ij}(l)\}$, and $\mathbf{x}(k) = [x_1(k) \cdots x_m(k)]^T$ the vector of observations. Then, the MIMO system output equals:

$$\mathbf{x}(k) = \sum_{l=0}^{L-1} \mathbf{h}(l)\mathbf{e}(k-l) \tag{1}$$

where L is the length of the longest $h_{ij}(k)$, and

$$e_i(k) = \sum_{l=0}^{L_c-1} c_i(l)s_i(k-l) \tag{2}$$

where $s_i(k)$ is a white signal with unit power, and $c_i(k)$, $k = 0, .., L_c-1$ is the corresponding color. L_c represents the maximum length in case the colors have different lengths.

For the quantities shown in the above two equations we will make the following assumptions.

- **(A1)** The inputs $\{s_j(k)\}$ are unknown, wide-sense stationary or cyclostationary, temporally white, and pairwise uncorrelated. For simplicity we will assume that they have equal variances.
- **(A2)** The input colors are known and pairwise non-identical.
- **(A3)** The mixing channels $h_{ij}(k)$ are in general complex.
- **(A4)** Let $\mathbf{H}(\omega)$ be a $m \times n$ matrix whose ij-th element is the N-point DFT of the unknown filter $h_{ij}(l)$, $l = 0, ..., L-1$ evaluated at frequency $\omega = \frac{2\pi}{N}k$, where k takes values in $[0, ..., N-1]$. We will assume that $\mathbf{H}(\omega)$ is full column rank for all ω's.

0-7803-7011-2/01/$10.00 ©2001 IEEE

The ultimate goal of blind system estimation/source separation is to estimate the channel matrix and use the estimate to subsequently recover the input sources.

By taking the length-N DFT ($N > L$) of Eq.(1), we obtain its frequency domain representation:

$$\mathbf{x}(\omega) \approx \mathbf{H}(\omega)\mathbf{e}(\omega) \tag{3}$$

where $\mathbf{e}(\omega)$ is the N-point DFT of the corresponding segment of $\mathbf{e}(n)$. Here ω denotes denotes discrete frequency of the form $\omega = \frac{2\pi}{N}k,\ k = 0, ..., N-1$.

The approximate equality above would be replaced with equality if the sequence $e(k)$ is periodic with period N.

The covariance matrix of the complex stochastic DFT process $\mathbf{x}(\omega)$ equals:

$$\begin{aligned} \mathbf{R}_x(\omega_1, \omega_2) &= E\{\mathbf{x}(\omega_1)\mathbf{x}(\omega_2)^H\} &(4)\\ &= \mathbf{H}(\omega_1)\mathbf{R}_e(\omega_1, \omega_2)\mathbf{H}(\omega_2)^H &(5)\end{aligned}$$

where the superscript ^{H} denotes Hermitian transpose, and $\mathbf{R}_e(\omega_1, \omega_2)$ is the covariance of $\mathbf{e}(\omega)$.

Since the inputs are assumed independent, $\mathbf{R}_e(\omega_1, \omega_2)$ is diagonal matrix, complex in general except for $\omega_1 = \omega_2$ when it is real. Since the input colors are assumed known, matrix $\mathbf{R}_e(\omega_1, \omega_2)$ can be predetermined.

Proposition 1: Under the assumptions (A1)-(A4), the channel matrix $\mathbf{H}(\omega)$ can be reconstructed up to a complex diagonal matrix based on $\mathbf{R}_x(\omega, \omega)$ and $\mathbf{R}_x(\omega, \omega + \alpha)$, $\alpha \neq 0$.

The proof of this proposition can be found in [1]. It is important to note that the residual ambiguity matrix is diagonal, meaning that the sources are decoupled, and that it doesn't depend on frequency.

In the next section we propose an iterative algorithm for blind identification of the channel matrix $\mathbf{H}(\omega)$ that is based on the estimates of $\mathbf{R}_x(\omega, \omega)$ and $\mathbf{R}_x(\omega, \omega + \alpha)$.

3. PROPOSED ALGORITHM

Our goal is to determine the channel matrix $\mathbf{H}(\omega)$ by using the knowledge of $\mathbf{R}_e(\omega, \omega)$ and $\mathbf{R}_e(\omega, \omega + \alpha)$ and estimates of $\mathbf{R}_x(\omega, \omega)$ and $\mathbf{R}_x(\omega, \omega + \alpha)$ that can be obtained based on the system outputs. Let us consider the time domain representation of the channel matrix in the following form:

$$\mathbf{H}(z^{-1}) = \mathbf{h}(0) + \mathbf{h}(1)z^{-1} + ... + \mathbf{h}(L-1)z^{-(L-1)} \tag{6}$$

where $\mathbf{h}(m) = \{h_{ij}(m)\}$. Although the channel length L appears in (6), as it will demonstrated in the simulations part, overestimation of L is not very critical.

We propose an iterative method for obtaining $\mathbf{H}(\omega)$ that is based on minimizing the following quantity:

$$\Gamma(l) \triangleq \sum_{k=0}^{N-1} \| \mathbf{D}_1(k) \|_F^2 + \sum_{k=0}^{N-1} \| \mathbf{D}_2(k;l) \|_F^2 \tag{7}$$

where $\mathbf{D}_1(k)$, $\mathbf{D}_2(k;l)$ are samples of $\mathbf{D}_1(\omega)$, $\mathbf{D}_2(\omega)$ obtained at $\omega = \frac{2\pi}{N}k$, $k \in [0, N-1]$, with

$$\begin{aligned} \mathbf{D}_1(k) &\triangleq \hat{\mathbf{R}}_x(k,k) - \mathbf{H}(k)\mathbf{R}_e(k,k)\mathbf{H}(k)^H &(8)\\ &= \hat{\mathbf{R}}_x(k,k) - \sum_{m=0}^{L-1}\sum_{n=0}^{L-1} \mathbf{h}(m)\mathbf{R}_e(k,k)\mathbf{h}(n)^H e^{[-j(m-n)k]} &(9)\end{aligned}$$

$$\begin{aligned} \mathbf{D}_2(k;l) &\triangleq \hat{\mathbf{R}}_x(k,k+l) - \mathbf{H}(k)\mathbf{R}_e(k,k+l)\mathbf{H}(k+l)^H &(10)\\ &= \hat{\mathbf{R}}_x(k,k+l) - \sum_{m=0}^{L-1}\sum_{n=0}^{L-1} \mathbf{h}(m)\mathbf{R}_e(k,k+l)\mathbf{h}(n)^H e^{[-j(m-n)k]+jnl} &(11)\end{aligned}$$

where $||.||_F$ denotes the Frobenius norm and l is an integer in $[0, \cdots, N-1]$ defined as $\alpha = \frac{2\pi}{N}l$.

Let us denote with $\mathbf{D}_{1R}(k)$ and $\mathbf{D}_{1I}(k)$ the real and imaginary parts of $\mathbf{D}_1(k)$, respectively, and similarly with $\mathbf{D}_{2R}(k;l)$ and $\mathbf{D}_{2I}(k;l)$, the real and imaginary parts of $\mathbf{D}_2(k;l)$, respectively. Then we can write:

$$\begin{aligned} \Gamma(l) &= \sum_{k=0}^{N-1} Tr(\mathbf{D}_1(k)\mathbf{D}_1^H(k)) + Tr(\mathbf{D}_2(k;l)\mathbf{D}_2^H(k;l))\\ &= \sum_{k=0}^{N-1} Tr(\mathbf{D}_{1R}(k)\mathbf{D}_{1R}^T(k) + \mathbf{D}_{1I}(k)\mathbf{D}_{1I}^T(k))\\ &+ Tr(\mathbf{D}_{2R}(k;l)\mathbf{D}_{2R}^T(k;l) + \mathbf{D}_{2I}(k;l)\mathbf{D}_{2I}^T(k;l)) &(12)\end{aligned}$$

The derivative of $\Gamma(l)$ with respect to $\mathbf{h}(i)$ can be computed in closed form [1], and can be used in any gradient based algorithm for minimizing Eq.(7). In our experiments the steepest-descent method was used, i.e.:

$$\tilde{\mathbf{h}}(i)^{k+1} = \tilde{\mathbf{h}}(i)^k - \mu_k \frac{\partial \Gamma(l)}{\partial \mathbf{h}(i)^k} \tag{13}$$

where $\tilde{\mathbf{h}}(i)^k$ denotes the updated estimate of $\mathbf{h}(i)$ at k-th iteration and μ_k is the step size.

Notice that not all the frequencies in Eq.(7) are required for successful channel reconstruction. The reconstruction can be based on as few as $2L$ DFT samples for complex channels. Using fewer frequencies reduces complexity (the complexity of the algorithm is proportional to the number of discrete frequencies used).

4. SIMULATION RESULTS

One of the very important applications that can be studied from the point of view of MIMO system estimation with colored inputs and known colors is an antenna-array CDMA system. If the system outputs are taken to be the received signals sampled at the chip-rate, then the system inputs are oversampled versions of the modulated information-bearing signals, each one colored by the corresponding user spreading code. The system response represents multipath between each input and output pair. In particular, for an n-user CDMA system the i-th receiver baseband signal can be described by the following equation:

$$x_i(t) = \sum_{j=1}^{n}\sum_{l=-\infty}^{\infty} g_{ij}(t - lT_s)s_j(l) \tag{14}$$

where j is the user index, $s_j(l)$ is the transmitted symbol sequence, and T_s is the symbol duration. For user j each symbol is multiplied by the pre-assigned spreading code sequence $\{c_j(0), \cdots, c_j(L_c - 1)\}$ at L_c times the symbol frequency T_s. The signature $g_{ij}(t)$ couples the j-th user with the i-th receiver. It incorporates the known sequence $c_j(k)$ and the unknown channel

$h_{ij}(t)$ which represents the multipath fading environment between the j-th user and the i-th receiver and can be described as:

$$g_{ij}(t) = \sum_{m=1}^{L_c} h_{ij}(t - mT_c)c_j(m) \tag{15}$$

The i-th receiver baseband signal $x_i(t)$ is sampled at the chip rate $1/T_c$ to obtain the following discrete time system:

$$x_i(k) = \sum_{j=1}^{n} \sum_{l=-\infty}^{\infty} g_{ij}(k - lL_c)s_j(l) \tag{16}$$

$$g_{ij}(k) = \sum_{m=0}^{L_c-1} h_{ij}(k - m)c_j(m) \tag{17}$$

It can easily be shown that the last expression can be rewritten as:

$$x_i(k) = \sum_{j=1}^{n} \sum_{m=k-L}^{k} h_{ij}(k - m)e_j(m) \tag{18}$$

where the process $e_j(k)$ can be viewed as the convolution between $c_j(k)$ and an oversampled by L_c version of $s_j(k)$.

In this section we will apply the proposed algorithm on the antenna array CDMA system and analyze its performances regarding both the system identification and channel equalization.

4.1. System Reconstruction

We considered a 5-user 5-antenna CDMA system with 4-level QAM inputs. The channels were generated according to the complex Gaussian distribution with zero means and unit variances, and normalized with respect to the zero-delay component. The number of multipaths was selected to be $L = 5$ (5 chip intervals long). The spreading codes were taken to be random sequences of length $L_c = 16$. The number of samples used was $M = 4096$ (256 symbols), the DFT size was $N = 128$ and the signal to noise ratio was selected to be SNR=$10dB$. The estimates $\hat{\mathbf{R}}_x(k,k)$ and $\hat{\mathbf{R}}_x(k,k+l)$ were obtained by segmenting the received data into $\frac{M}{N}$ segments, computing the DFT of each segment and averaging over all segments as in Eq.(4). The number of frequencies used for the optimization was $F = 32$. The frequency spacing used was $l = 8$. The selection of the frequency spacing plays an important role. As it has already been discussed, $\mathbf{R}_e(k,k)$ and $\mathbf{R}_e(k,k+l)$ are assumed to be known for all frequencies $k = 0, \cdots, N-1$. However, since the number of DFT points, N, is in general larger than the length of the colors L_c, the resolution doesn't allow us to use any frequency spacing l with the same accuracy. By selecting $l = \frac{N}{L_c}$ this problem is successfully resolved.

As the measure of the performance, the normalized mean-square error ($NMSE$) was used. The overall ($ONMSE$) was obtained by averaging over all cross-channels:

$$ONMSE = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} NMSE_{ij}}{mn} \tag{19}$$

The simulations were repeated for the 20 randomly selected 5×5 channel realizations, for various data lengths and $SNRs$. For each channel set $ONMSE$ was computed based on 50 independent input realizations. The $ONMSEs$ corresponding to the 20 different channels are shown in Fig. 1.

Figure 2 shows the performance of the algorithm for 3 different data lengths and various $SNRs$. Results are based on the 50 Monte Carlo runs and averaged over 10 different 5×5 channel realizations.

In order to show the robustness of the proposed algorithm on channel order mismatch we computed the $ONMSE$ for the same example (5×5 system with $4-QAM$ inputs, $SNR = 10dB$, $M = 2048$ and 20 different channel sets with $L = 5$) for different amount of length mismatch. The assumed channel lengths, L_a, were 5, 6 and 7. The results based on the 50 Monte Carlo runs are shown in Fig. 3.

4.2. Equalization

Based on the obtained system estimate a zero-forcing block linear equalizer was used to recover the inputs. Let us denote with $\mathbf{s}$ the combined data symbol vector:

$$\mathbf{s} = [\mathbf{s}_1^T, \mathbf{s}_2^T, \cdots \mathbf{s}_n^T] \tag{20}$$

with

$$\mathbf{s}_i = [s_i^{(1)}, s_i^{(2)}, \cdots, s_i^{(P)}]^T \tag{21}$$

where P is the number of symbols per user.

Let $\mathbf{x}_i$, $i = 1, \cdots, m$ be the data vector at the ith receiver. It is easy to show that the following expression holds:

$$\mathbf{x}_i = \mathbf{A}^{(i)}\mathbf{s} + \mathbf{n}_i, \quad i = 1, \cdots, m \tag{22}$$

where $\mathbf{n}_i$ is the noisy vector at the ith receiver and $\mathbf{A}^{(i)} = \{A_{kj}^{(i)}\}$ is the $(PL_c + L - 1) \times (nP)$ matrix defined as:

$$A^{(i)}_{L_c(q_1-1)+q_2,q_1+P(q_3-1)} = \begin{cases} g_{iq_3}(q_2), & \begin{matrix} q_1 = 1, \cdots, P \\ q_2 = 1, \cdots, L_c + L - 1 \\ q_3 = 1, ..., n \end{matrix} \\ 0, & elsewhere \end{cases} \tag{23}$$

where $g_{ij}(k)$ was defined in Eq.(17).

The zero-forcing block linear equalizer can now be implemented as (assuming $\mathbf{R}_{\mathbf{n}_i} = E\{\mathbf{n}_i\mathbf{n}_i^H\} = \sigma_n\mathbf{I}$):

$$\hat{\mathbf{s}}_{ZF} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{A}^{(i)H}\mathbf{A}^{(i)})^{-1}\mathbf{A}^{(i)H}\mathbf{x}_i \tag{24}$$

For a 5×5 system, typical signal at the output of the equalizer for $4-QAM$ inputs, $SNR = 10dB$, $L = 5$ and data length $M = 4096$ is shown in Fig. 4. The bit-error rate (BER) of the recovered signal is shown in Fig. 5. The results were obtained based on 50 Monte Carlo runs for two different data lengths used for system identification. Solid lines correspond to the case with no mismatch ($L_a = L = 5$), while dashed lines represent the case when $L_a = 7$.

5. REFERENCES

[1] I. Bradaric, A.P. Petropulu and K.I. Diamantaras, "Blind MIMO FIR Channel Identification Based on Second-Order Statistics," *IEEE Trans. on Signal Processing*, submitted in 2001.

[2] B. Chen and A.P. Petropulu, "Frequency Domain Blind MIMO System Identification Based on Second and Higher-Order Statistics," *IEEE Trans. on Signal Processing*, to appear in August 2001.

[3] K.I. Diamantaras, A.P. Petropulu and B. Chen, "Blind Two-Input-Two-Output FIR Channel Identification Based on Frequency Domain Second-Order Statistics," *IEEE Trans. on Signal Processing*, February 2000.

[4] P. Comon, "Contrasts for Multichannel blind deconvolution,"*IEEE Signal Processing Letters*, vol. 3, pp. 209-211, July 1996.

[5] K. Diamantaras and A.P. Petropulu, "Blind Equalization of Multiuser CDMA Channels: A Frequency-Domain Approach," *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP'99*, Phoenix, Arizona, USA, March 1999.

[6] A. Gorokhov. P. Loubaton, and E. Moulines, "Second Order Blind Equalization in Multiple Input Multiple Output FIR Systems: A Weighted Least Squares Approach", in *Proc. ICASSP-96*, pp. 2415-2418, 1996.

[7] Y. Hua and J.K Tugnait, "Blind Identifiability of FIR-MIMO Systems with Colored Input Using Second Order Statistics,"*IEEE Signal Processing Letters*, vol. 7(12), pp. 348-350, Dec. 2000.

[8] Y. Inouye and K. Hirano, "Cumulant-Based Blind Identification of Linear Multi-Input-Multi-Output Systems Driven by Colored Inputs", *IEEE Trans. on Signal Processing*, vol. 45 (6), pp. 1543–1552, June 1997.

[9] H. Liu and M. Zoltowski, "Blind equalization in antenna array CDMA systems," *IEEE Trans. on Signal Processing*, vol. 45(1), pp. 161-172, Jan. 1997.

[10] E. Moreau and J.-C. Pesquet, "Generalized contrasts for multichannel blind deconvolution of linear systems," *IEEE Signal Processing Letters*, vol. 4, pp. 182-183, June 1997.

[11] L. Parra and C. Spence, "Convolutive Blind Separation of Non-Stationary Sources," *IEEE Tr. Speech and Audio Processing*, Vol. 8, No. 3, pp.320-327, May 2000.

[12] S. Shamsunder and G.B. Giannakis, "Multichannel blind signal separation and reconstruction," *IEEE Trans. on Speech & Audio Processing*, vol. 5(6), pp. 515-527, November 1997.

[13] M. Torlak and G. Xu, "Blind Multiuser Channel Estimation in Asynchronous CDMA Systems", *IEEE Tr. Signal Processing*, pp. 137–147, vol. 45, no. 1, Jan. 1997.

[14] J.K. Tugnait, "Identification and Deconvolution of Multichannel Linear Non-Gaussian Processes Using Higher Order Statistics," *IEEE Trans. on Signal Processing*, vol. 45(3), pp. 658-672, March 1997. *IEEE Trans. on Signal Processing*, vol. 43(7), pp. 1602-1612, July 1995.

[15] X. Wang and H.V. Poor, "Blind Equalization and Multiuser Detection in Dispersive CDMA Channels", *IEEE Trans. on Communications*, vol. 46(1), pp. 91-103, Jan. 1998.

[16] J. Xavier, V. Barroso, and J.M.F. Moura, "Closed Form Blind Identification of MIMO Channels", in *Proc. ICASSP-98*, vol. 6, pp. 3165-3168, Seattle WA, 1998.

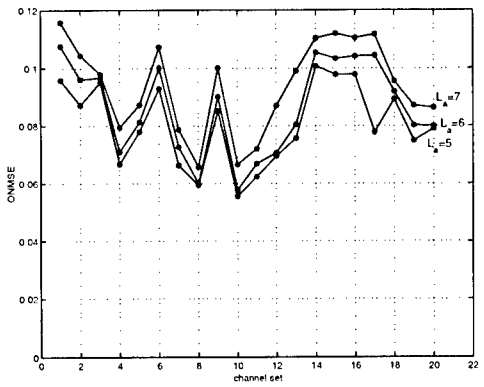


Figure 3: $ONMSE$ for different channel order mismatches ($L_o = 5$ is the true length)

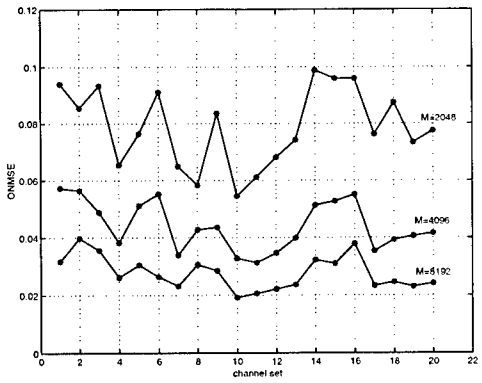


Figure 1: $ONMSE$ for different data lengths and $SNR = 10dB$

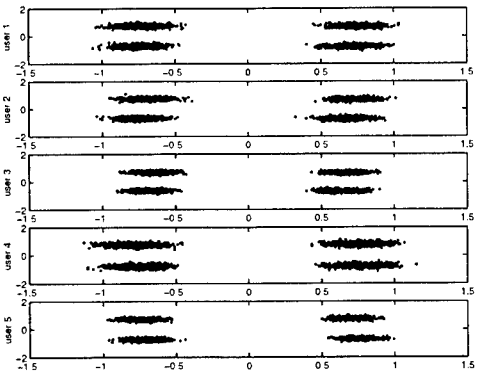


Figure 4: The output of the equalizer for $4 - QAM$ input signals, $SNR = 10dB$ and data length $M = 4096$

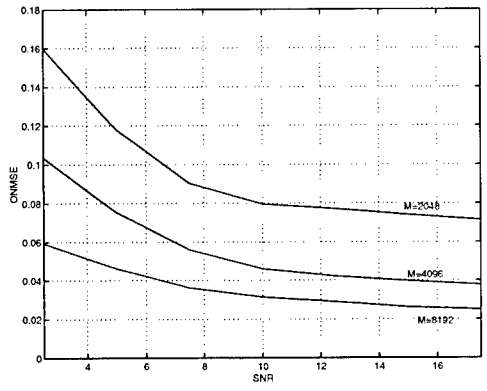


Figure 2: $ONMSE$ for different $SNRs$ and data lengths

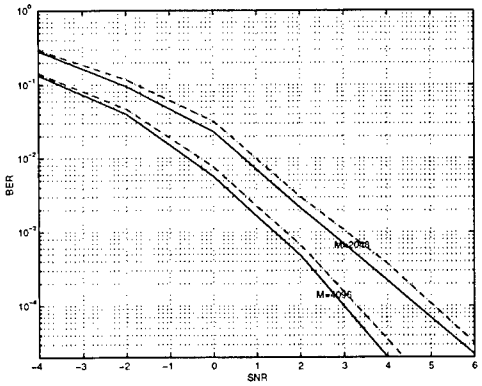


Figure 5: BER corresponding to a 4-level QAM inputs

BLIND CHANNEL IDENTIFICATION USING ROBUST SUBSPACE ESTIMATION

S. Visuri[1] *H. Oja*[2] *V. Koivunen*[1]

[1]Signal Processing Laboratory
Helsinki Univ. of Technology
P.O. Box 3000, FIN-02015 HUT
Finland

[2]Dept. of Statistics
University of Jyväskylä
P.O. Box 35, FIN-40351 Jyväskylä
Finland

ABSTRACT

The paper introduces a robust approach to subspace based blind channel identification. The technique is based on estimating the noise subspace from the sample sign covariance matrix. The theoretical motivation for the technique is shown under the white Gaussian noise assumption. A simulation study is performed to demonstrate the robust performance of the algorithm both in Gaussian and non-Gaussian noise. The results indicate that when the noise is Gaussian, the proposed method has similar good performance as the standard subspace method. When the noise is heavy-tailed, the proposed method outperforms the conventional subspace technique.

1. INTRODUCTION

Blind channel identification allows for improving spectral efficiency. It may be achieved using only second order statistics by employing Single-Input Multi-Output (SIMO) model resulting from fractional sampling or the use of an antenna array [1]. A subspace method performing the identification of the channel from the eigenvalue decomposition of the covariance matrix was proposed in [2].

Typically noise in the received signal are assumed to be spatially and temporally white Gaussian noise and the eigenvalues and corresponding noise subspace eigenvectors, needed in channel identification, are computed from the sample covariance matrix. Sample covariance matrix is known to perform poorly in the face of heavy-tailed noise. This is of concern in wireless communication applications, in particular in urban and indoor radio channels, where the ambient noise has been shown to be decidedly non-Gaussian [3]. Consequently, the estimated eigenvectors and eigenvalues may significantly deviate from the true ones.

In this paper, we propose a robust subspace identification method that performs almost optimally in Gaussian noise and highly reliably in non-Gaussian heavy-tailed noise. The sample covariance matrix used in [2] is replaced by a sample sign covariance matrix, which uses a multivariate generalization of the univariate sign function. Theoretical motivation of the method is shown under the white Gaussian noise assumption. The simulation results demonstrate that the performance is almost equal to that of the original method in Gaussian noise, and it remains highly reliable even in heavy-tailed noise such as Cauchy. The performance of the original method deteriorates significantly and it may completely fail in such noise conditions. The additional robustness of the proposed method is achieved without any significant increase in computational complexity.

Financial support for this work was provided by the Academy of Finland

The paper is organized as follows. The sample sign covariance matrix to be employed in blind identification is defined in section 2. Then the signal model used in SIMO model is given in section 3. Section 4 briefly describes the original blind subspace identification method by Moulines et al. [2]. The method is based on noise subspace eigenvectors. In section 5 we show how these eigenvectors may be estimated in a robust manner, hence yielding robust estimates of the channel coefficients. In section 6, simulation examples where the received signal is contaminated by Gaussian and heavy-tailed noise are presented. Finally, section 7 concludes the paper.

2. SIGN COVARIANCE MATRIX

We begin by defining the sample Sign Covariance Matrix (SCM) used in this article. For a M-variate complex vector $\mathbf{x}$, the *spatial sign function* is defined as

$$\mathbf{S}(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{||\mathbf{x}||}, & \mathbf{x} \neq \mathbf{0} \\ \mathbf{0}, & \mathbf{x} = \mathbf{0}, \end{cases}$$

where $||\mathbf{x}|| = (\mathbf{x}^H\mathbf{x})^{1/2}$. For a M-variate complex data set, $\mathbf{x}_1, \ldots, \mathbf{x}_K$, the sample SCM is

$$S_1 = \frac{1}{K}\sum_{i=1}^{K} \mathbf{S}(\mathbf{x}_i)\mathbf{S}^H(\mathbf{x}_i).$$

The (theoretical) SCM for the distribution F is defined by

$$\Sigma_1 = E_F\{\mathbf{S}(\mathbf{x})\mathbf{S}^H(\mathbf{x})\},$$

where $\mathbf{x}$ is distributed according to F. Various properties of the SCM have been discussed in [4, 5].

3. SIGNAL MODEL

Let s_n be a symbol emitted at the time nT, where T is the symbol duration. We assume the standard SIMO baseband signal model [6], in which the received signal $\mathbf{x}_n$ having P components arranged as a column vector is of the form

$$\mathbf{x}_n = \sum_{k=0}^{L} \mathbf{h}_k s_{n-k} + \mathbf{v}_n. \quad (1)$$

Here $\{\mathbf{h}_k\}$ is the channel impulse response sequence, L is the channel order and $\mathbf{v}_n$ is the noise. This SIMO model result either by sampling the received signal from P sensors by a symbol rate or by oversampling the signal received by a single sensor by a factor $\Delta = T/P$ [1].

0-7803-7011-2/01/$10.00 ©2001 IEEE

By stacking $N+1$ observations of (1) into an $NP \times 1$ vector $\mathbf{X}_n = [\mathbf{x}_n^T, \mathbf{x}_{n-1}^T, \ldots, \mathbf{x}_{n-N}^T]^T$ we may write

$$\mathbf{X}_n = \mathcal{H}_N \mathbf{S}_n + \mathbf{V}_n. \tag{2}$$

Here, $\mathbf{S}_n = [s_n, s_{n-1}, \ldots, s_{n-N-L}]^T$, $\mathbf{V}_n = [\mathbf{v}_n^T, \mathbf{v}_{n-1}^T, \ldots, \mathbf{v}_{n-N}^T]^T$ and $\mathcal{H}_N$ is the $(N+1)P \times (L+N+1)$ channel convolution matrix given by

$$\mathcal{H}_N = \begin{bmatrix} \mathbf{h}_0 & \mathbf{h}_1 & \cdots & \mathbf{h}_L & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{h}_0 & \mathbf{h}_1 & \cdots & \mathbf{h}_L & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{h}_0 & \mathbf{h}_1 & \cdots & \mathbf{h}_L \end{bmatrix}$$

4. CHANNEL IDENTIFICATION

We now review the basic steps of a subspace-based identification method first introduced in [2]. Assume that $\mathbf{x}_n$ given in (2) is a wide-sense stationary process and the signal $\mathbf{S}_n$ and noise $\mathbf{V}_n$ are mutually independent. The covariance matrix of $\mathbf{X}_n$ is

$$E\{\mathbf{X}_n \mathbf{X}_n^H\} = \Sigma_0 = \mathcal{H}_N \Sigma_s \mathcal{H}_N^H + \Sigma_v,$$

where $\Sigma_s = E\{\mathbf{S}_n \mathbf{S}_n^H\}$ is the signal covariance matrix and $\Sigma_v = E\{\mathbf{V}_n \mathbf{V}_n^H\}$ is the noise covariance matrix. We assume that the standard assumptions hold (see [2] for detail) for ensuring the channel identifiability. These assumptions require that the subchannels resulting from forming the SIMO model do not share common zeros. The transmitted symbols are assumed to be i.i.d. between the successive time instants and the noise covariance matrix is assumed to be $\Sigma_v = \sigma^2 I$, where σ^2 is the noise power. The maximum channel order L is assumed to be known as well.

The covariance matrix Σ_0 of the received signal can be represented in terms of its eigenvector decomposition. Based on the pattern of the eigenvalues one can perform the decomposition to signal and noise subspaces. The signal subspace spanned by eigenvectors corresponding to the $L+N+1$ largest eigenvalues spans the same space as columns of the channel matrix $\mathcal{H}_N$. The remaining $r = (P-1)N + P - L - 1$ eigenvectors span the noise subspace. The corresponding eigenvalues are all equal to the noise variance σ^2. Denote the noise subspace eigenvectors by $\mathbf{g}_i$, $i = 1, \ldots, r$. It is a standard result that

$$\mathcal{H}_N^H \mathbf{g}_i = \mathbf{0}, \; i = 1, \ldots, r.$$

This orthogonality of the signal and noise subspaces allows for identification of the channel coefficient vector

$$\mathbf{h} = [\mathbf{h}_0^T, \mathbf{h}_1^T, \ldots, \mathbf{h}_L^T]^T.$$

To illustrate how the identification is done, partition the noise subspace eigenvectors as

$$\mathbf{g}_i = [\mathbf{g}_0^{(i)^T}, \mathbf{g}_1^{(i)^T}, \ldots, \mathbf{g}_N^{(i)^T}]^T, \tag{3}$$

where $\mathbf{g}_k^{(i)}, k = 0, 1, \ldots, N$ are of size $P \times 1$. Define

$$\mathcal{G}_i = \begin{bmatrix} \mathbf{g}_0^{(i)} & \mathbf{g}_1^{(i)} & \cdots & \mathbf{g}_N^{(i)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{g}_0^{(i)} & \mathbf{g}_1^{(i)} & \cdots & \mathbf{g}_N^{(i)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{g}_0^{(i)} & \mathbf{g}_1^{(i)} & \cdots & \mathbf{g}_N^{(i)} \end{bmatrix} \tag{4}$$

It can be shown [2] that

$$\mathbf{g}_i^H \mathcal{H}_N \mathcal{H}_N^H \mathbf{g}_i = \mathbf{h}^H \mathcal{G}_i \mathcal{G}_i^H \mathbf{h}, \; i = 1, \ldots, r.$$

Therefore

$$\mathbf{h}^H \left(\sum_{i=1}^{r} \mathcal{G}_i \mathcal{G}_i^H \right) \mathbf{h} = 0.$$

In [2] it is also shown that the dimension of the null space of the matrix

$$\mathcal{C} = \sum_{i=1}^{r} \mathcal{G}_i \mathcal{G}_i^H$$

is one. This implies that the channel impulse response may be determined from the eigenvector of $\mathcal{C}$ corresponding to the eigenvalue that is equal to zero, and the solution is unique up to a multiplicative constant. Signal subspace eigenvectors may be employed as well [2].

5. ROBUST SUBSPACE ESTIMATION

In practice the noise subspace eigenvectors have to be estimated from the available measurements $\mathbf{X}_1, \ldots, \mathbf{X}_K$. The estimation is conventionally done by using the eigenvectors of the sample covariance matrix

$$S_0 = \frac{1}{K} \sum_{i=1}^{K} \mathbf{X}_i \mathbf{X}_i^H.$$

Let $\hat{\mathbf{g}}_{i,0}, i = 1, \cdots, r$ be eigenvectors of S_0 corresponding to the r smallest eigenvalues. An estimate of the channel vector may then be chosen to be the eigenvector corresponding to smallest eigenvalue of

$$\hat{\mathcal{C}} = \sum_{i=1}^{r} \hat{\mathcal{G}}_i \hat{\mathcal{G}}_i^H, \tag{5}$$

where $\hat{\mathcal{G}}_i$ are defined from equations (3)-(4) with $\hat{\mathbf{g}}_{i,0}$ used in place of $\mathbf{g}_i$.

Let $\hat{\mathbf{g}}_{i,1}, i = 1, \ldots, r$ be the eigenvectors corresponding to r smallest eigenvectors of the sample SCM. We now prove, assuming Gaussian noise, that these eigenvectors are convergent estimates of the noise subspace basis vectors. Therefore they may be used in any subspace based identification method.

Theorem 1 *Assume $\mathbf{X}_1, \ldots, \mathbf{X}_K$ distributed as given in (2) and assume that the SIMO identifiability conditions hold. Assume further that the multivariate noise in (2) is complex circular Gaussian distributed and denote the SCM of $\mathbf{X}_i$s by Σ_1. Let S_1 be the sample SCM of the data. Set $\hat{\mathbf{g}}_{i,1}, i = 1, \ldots, r$ to be the eigenvectors of S_1 corresponding to r smallest eigenvalues. Then:*

(i) The r smallest eigenvalues of $\Sigma_1 = E\{S_1\}$ are equal and the corresponding eigenvectors are orthogonal to the columns of the matrix $\mathcal{H}_N$.

(ii) As $K \to \infty$,

$$S_1 \overset{w.p.1}{\to} \Sigma_1.$$

(iii) As $K \to \infty$,

$$\mathcal{H}_N^H \hat{\mathbf{g}}_{i,1} \overset{w.p.1}{\to} \mathbf{0}, \; i = 1, \ldots, r.$$

Proof. Result (i) follows from Theorem 2 in [4]. By using the i.i.d. assumption of the symbol sequence the result (ii) follows

from Theorem 1.8.E in [7]. Result (iii) now follows from Theorem 3 in [4].

Note that the part (i) of the above theorem proves that the channel coefficient vector may be identified from the theoretical SCM of $\mathbf{X}$ in (2). Part (ii) then gives the convergence of the sample SCM to the theoretical SCM. Finally, part (iii) states the convergence of the noise subspace eigenvectors. The efficiency and robust performance of the sample SCM based subspace estimation technique, also in non-Gaussian noise, is shown using simulations in the following section.

6. SIMULATION RESULTS

In this section, we present simulation results illustrating the robustness of the channel identification using the noise subspace estimate obtained from the sample SCM. Moreover, we compare the performance to that of the identification method where the noise subspace estimate is obtained from the sample covariance matrix. The channel is estimated using the noise subspace method described earlier. In order to study robustness, ε-contamination and complex isotropic symmetric α-stable ($S\alpha S$) noise models are considered.

The characteristic function of a complex isotropic $S\alpha S$ distribution is

$$\rho(\omega) = \exp(-\gamma|\omega|^{\alpha}).$$

The smaller the characteristic exponent $\alpha \in [0, 2]$, the heavier the tails of the density (the case $\alpha = 2$ corresponds to Gaussian distribution). The positive valued scalar γ is the dispersion of the distribution. The dispersion plays a role analogous to that of the variance for second order processes [8].

In the ϵ-contaminated noise model, the noise is given by

$$v = (1 - b)v_1 + bv_2$$

where $b \sim Bin(1, \epsilon)$, $\mathbf{v}_1 \sim \mathcal{N}_C(\mathbf{0}, 1)$, $\mathbf{v}_2 \sim \mathcal{N}_C(\mathbf{0}, 1000)$.

As in Moulines et al. [2], the emitted signal is a random 4-QAM signal and the symbols are independent between the successive time instants. The noise is independent of the signals and i.i.d. between the samples. The number of virtual channels is $P = 4$; the width of the temporal window is $N = 10$; the degree of the ISI is $L = 4$. The channel coefficients are given by [2]

$$\begin{aligned}
\mathbf{h}_0^T &= [(-0.049 + 0.359j), (0.443 - 0.0364j), \\
&\quad (-0.221 - 0.322j), (0.417 + 0.030j)] \\
\mathbf{h}_1^T &= [(0.482 - 0.569j), (1), (-0.199 + 0.918j), (1)] \\
\mathbf{h}_2^T &= [(-0.556 + 0.587j), (0.921 - 0.194j), (1), \\
&\quad (0.873 + 0.145j)] \\
\mathbf{h}_3^T &= [(1), (0.189 - 0.208j), (-0.284 - 0.524j), \\
&\quad (0.285 + 0.309j)] \\
\mathbf{h}_4^T &= [(-0.171 + 0.061j), (-0.087 - 0.054j), \\
&\quad (0.136 - 0.19j), (-0.049 + 0.161j)]
\end{aligned}$$

The number of independent Monte-Carlo runs used in the simulations is 100. Since the correct channel vector can be estimated only up to an arbitrary multiplicative constant, the performance criterion used in our simulations is the canonical angle between the estimated channel vector $\hat{\mathbf{h}}$ and the correct channel vector $\mathbf{h}$.

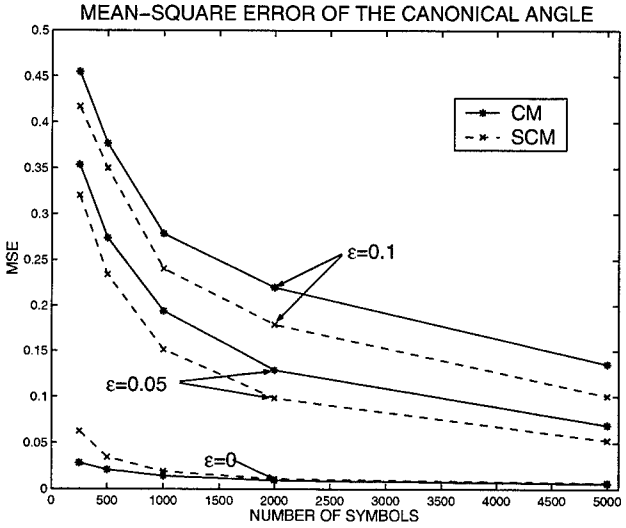


Figure 1: MSE of the canonical angle (in radians) in ϵ-contaminated noise. Solid lines: noise subspace estimated from the sample covariance matrix. Dashed lines: noise subspace estimated from the sample SCM.

The canonical angle is defined as

$$\sphericalangle(\mathbf{h}, \hat{\mathbf{h}}) = \text{acos}\left(\text{abs}\left[\frac{\hat{\mathbf{h}}^H \mathbf{h}}{||\hat{\mathbf{h}}||||\mathbf{h}||}\right]\right)$$

where $||\cdot||$ is the Euclidean vector norm. Note that $0 \leq \sphericalangle(\mathbf{h}, \hat{\mathbf{h}}) \leq \pi/2$. Moreover, $\sphericalangle(\mathbf{h}, \hat{\mathbf{h}}) = 0$ if and only if $\hat{\mathbf{h}} = c\mathbf{h}$, where c is a scalar constant.

In our first simulation we compare the behavior of the two algorithms in ϵ-contaminated noise. The output SNR (as defined in [9]) between the signal part and the nominal noise part v_1 is 20 dB. Figure 1 shows mean squared error of the canonical angle for cases $\epsilon = 0$,$\epsilon = 0.05$ and $\epsilon = 0.1$ and number of symbols $N_d = 250, 500, 1000, 2000, 5000$. The MSE is calculated by

$$\text{MSE} = \frac{1}{N_m}\sum_{i=1}^{N_m} \sphericalangle(\mathbf{h}, \hat{\mathbf{h}}_i)^2,$$

where N_m is the number of Monte-Carlo realizations and $\hat{\mathbf{h}}_i$ is the estimate from ith realization. In the Gaussian case, the behavior of the two methods is in practice equal for $N_d \geq 2000$. For small sample sizes the method based on sample covariance has smaller MSE. As expected, when $\epsilon > 0$, the method based on the sample SCM has better performance than the method based on the sample covariance matrix.

Figure 2 shows the simulation results for α-stable noise. The values used for the characteristic exponent are $\alpha = 2$, $\alpha = 1.5$ and $\alpha = 1$. The value for the dispersion is $\gamma = 1$ (in the Gaussian case the output SNR is 20 dB). Similarly to the previous simulation the number of symbols used in the estimation task is $N_d = 250, 500, 1000, 2000, 5000$. Naturally the simulation results for Gaussian noise are the same as in the previous simulation. When the noise is more heavy tailed than Gaussian noise, the method based on the SCM clearly outperforms the method based

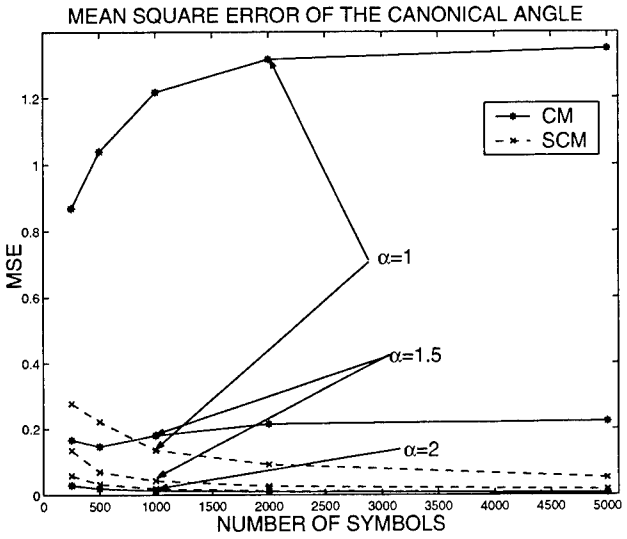


Figure 2: MSE of the canonical angle (in radians) in α-stable noise. Solid lines: noise subspace estimated from the sample covariance matrix. Dashed lines: noise subspace estimated from the sample SCM.

on the covariance matrix. Note that when $\alpha < 2$ the probability of having extremely deviating noise samples in data grows as a function of the number of samples N_d. Therefore also the MSE of the method employing the sample covariance matrix grows as a function of N_d.

7. CONCLUSION

In the paper we show how blind channel identification may be done in a robust manner by using the sample SCM. The simulation results imply that the proposed method performs reliably also in heavy-tailed noise, whereas the method based on the sample covariance matrix is sensitive to the deviations from Gaussian noise. The calculation of the sample SCM is straightforward and therefore the methods based on the SCM have approximately the same computational complexity as the methods based on the sample covariance matrix.

8. REFERENCES

[1] L. Tong, G. Xu, and T. Kailath, "Blind identification based on second-order statistics: A time domain approach," *IEEE Transactions on Information Theory*, vol. 40, no. 2, pp. 340–349, 1994.

[2] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Transactions on Signal Processing*, vol. 43, no. 2, pp. 516–525, 1995.

[3] D. Middleton, "Man-made noise in urban environments and transportation systems: Models and measurements.," *IEEE Transactions on Communications*, vol. 21, pp. 1232–1241, 1973.

[4] S. Visuri, H. Oja, and V. Koivunen, "Subspace-based direction of arrival estimation using nonparametric statistics," *IEEE Transactions on Signal Processing*, to appear, 2001.

[5] S. Visuri, *Array and Multichannel Signal Processing Using Nonparametric Statistics*. D.Sc. thesis, Helsinki University of Technology, 2001. www.hut.fi/Yksikot/Kirjasto/Diss/2001/isbn951225364X/.

[6] P. Loubaton, E. Moulines, and P. Regalia, "Subspace method for blind identification and deconvolution," in *Signal Processing Advances in Wireless & Mobile Communications* (G. Giannakis, Y. Hua, P. Stoica, and L. Tong, eds.), vol. 1: Trends in Channel Estimation and Equalization, pp. 63–112, Prentice Hall, 1998.

[7] R. Serfling, *Approximation Theorems of Mathematical Statistics*. New York: Wiley, 1980.

[8] M. Shao and C. L. Nikias, *Signal Processing with Alpha-Stable Distributions and Applications*. New-York: Wiley, 1995.

[9] L. Tong, G. Xu, and T. Kailath, "Fast blind equalization via antenna arrays," in *Proc. of Int'l Conf. on Acoust. Speech and Signal Proc.*, vol. 4, pp. 272–275, 1993.

ORTHOGONAL MINIMUM NOISE SUBSPACE FOR MULTIPLE-INPUT MULTIPLE-OUTPUT SYSTEM IDENTIFICATION

Anahid Safavi, Karim Abed-Meraim
Telecom Paris. 46 rue Barrault, 75634 Paris Cedex 13, France
email: safavi,abed@tsi.enst.fr

ABSTRACT

This contribution deals with a particular family of blind system identification techniques, referred to as Minimum Noise Subspace (MNS) method. MNS method is a computationally fast version of Subspace method. Here, we develop an orthogonal version of MNS method. Orthogonal Minimum Subspace (OMNS) method is more efficient in computation than a standard subspace method, and is more robust to channel noise than MNS.

I. INTRODUCTION

Recently, a new subspace method called Minimum Noise Subspace (MNS) has been proposed for Multiple-Input Multiple-Output (MIMO) system identification [1], [2]. This method computes the noise subspace via a set of noise vectors which are computed in parallel from a set of combinations of system outputs that form a basis of the rational noise subspace.

In this contribution, an orthogonal version of MNS called Orthogonal Minimum Noise Subspace (OMNS) is proposed. Here, the noise subspace is formed through computation of noise vectors that correspond to an orthogonal set of noise polynomial vectors (orthogonal basis of the rational noise subspace). The OMNS is more robust to channel noise than MNS method.

This paper is organized as follows: System model and general assumptions are introduced in section II. Section III describes the general subspace method. In section IV we derive the basic ideas of both MNS and OMNS method applying rational subspace formalism. OMNS algorithm is described in section V. In section VI, both methods are compared in terms of computation complexity and estimation accuracy. In section VII, the computer simulations are presented.

II. SYSTEM MODEL

Let $\mathbf{y}(n)$ be a q-variate discrete time stationary time series given by:

$$\mathbf{y}(n) = \sum_{k=0}^{M} \mathbf{H}(k)\mathbf{s}(n-k) + \mathbf{w}(n) \triangleq [\mathbf{H}(z)]\mathbf{s}(n) + \mathbf{w}(n) \quad (1)$$

where

$$\mathbf{H}(z) = \sum_{k=0}^{M} \mathbf{H}(k)z^{-k} \triangleq \begin{bmatrix} h_{1,1}(z) & \cdots & h_{1,p}(z) \\ \vdots & \ddots & \vdots \\ h_{q,1}(z) & \cdots & h_{q,p}(z) \end{bmatrix}.$$

$\mathbf{H}(z)$ is an unknown causal FIR $q \times p$ transfer function with $q > p$. $\mathbf{s}(n) = [s_1(n), ..., s_p(n)]^T$ is a p-dimensional unknown process and $\mathbf{w}(n)$ is an additive q-dimensional white noise, i.e. $E[\mathbf{w}(n)\mathbf{w}^*(n)] = \sigma^2\mathbf{I}_q$. Otherwise, (1) describes a p-input and q-output system.

In the communication context, the input sequence $\mathbf{s}(n)$ denotes the transmitted symbols, and the unknown FIR transfer function $\mathbf{H}(z)$ models the propagation channel between sources and sensors.

We study here the estimation of $\mathbf{H}(z)$ from the observation $\mathbf{y}(n)$ under the following assumptions:

$$rank(\mathbf{H}(z)) = p \quad \text{for each } z \quad (2)$$

$$\mathbf{H}(M) \quad \text{is full column rank} \quad (3)$$

In fact, (3) can be relaxed by simply assuming $\mathbf{H}(z)$ to be column-reduced [5].

III. SUBSPACE METHOD

Here, we present a brief review of original subspace method. Let $\mathbf{y}_i(n)$ be a vector of N successive samples from the i-th output of the system. According to (1), it can be written as:

$$\begin{aligned} \mathbf{y}_i(n) &= [y_i(n), ..., y_i(n-N+1)]^T \\ &= \mathcal{T}_N(\mathbf{H}_{i,:})\bar{\mathbf{s}}(n) + \mathbf{w}_i(n) \end{aligned} \quad (4)$$

$\bar{\mathbf{s}}(n)$ denotes the vector of input samples, i.e. $\bar{\mathbf{s}}(n) = [\mathbf{s}_1^T(n), ..., \mathbf{s}_p^T(n)]^T$ where $\mathbf{s}_j(n) = [s_j(n), ..., s_j(n-N-M+1)]^T$ for $1 \le j \le p$ and $\mathbf{w}_i(n) = [w_i(n), ..., w_i(n-N+1)]^T$. $\mathcal{T}_N(\mathbf{H}_{i,:})$ is the $N \times p(N+M)$ block Sylvester matrix given by:

$$\mathcal{T}_N(\mathbf{H}_{i,:}) = [\mathcal{T}_N(h_{i,1}), ..., \mathcal{T}_N(h_{i,p})]$$

where, $\mathcal{T}_N(h_{i,j})$ denotes the $N \times (N+M)$ Sylvester matrix associated to $h_{i,j}$ [3]. Considering all of the outputs of the system and putting them in to a vector called $\bar{\mathbf{y}}(n)$, we obtain:

$$\begin{aligned} \bar{\mathbf{y}}(n) &= [\mathbf{y}_1^T(n), ..., \mathbf{y}_q^T(n)]^T \\ &= \mathcal{T}_N(\mathbf{H})\bar{\mathbf{s}}(n) + \bar{\mathbf{w}}(n) \end{aligned} \quad (5)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

with

$$\begin{aligned}\bar{\mathbf{w}}(n) &= [\mathbf{w}_1^T(n),...,\mathbf{w}_q^T(n)]^T \\ \mathcal{T}_N(\mathbf{H}) &= [\mathcal{T}_N^T(\mathbf{H}_{1,:}),...,\mathcal{T}_N^T(\mathbf{H}_{q,:})]^T \end{aligned} \quad (6)$$

$\mathcal{T}_N(\mathbf{H})$ is $qN \times p(N+M)$ generalized Sylvester matrix of order N associated to $\mathbf{H}(z)$. Let $\overline{\mathbf{R}}_N$ be the $qN \times qN$ covariance matrix of $\bar{\mathbf{y}}(n)$:

$$\begin{aligned}\overline{\mathbf{R}}_N &\triangleq E[\bar{\mathbf{y}}(n)\bar{\mathbf{y}}^*(n)] \\ &= \mathcal{T}_N(\mathbf{H})\bar{\mathbf{S}}\mathcal{T}_N^*(\mathbf{H}) + \sigma^2\mathbf{I}_{qN} \end{aligned} \quad (7)$$

where $\bar{\mathbf{S}} \triangleq E[\bar{\mathbf{s}}(n)\bar{\mathbf{s}}^*(n)] > 0$ under assumptions (2) and (3) for $N > pM$, $\mathcal{T}_N(\mathbf{H})$ has full column rank $p(N+M)$. Therefore, $\overline{\mathbf{R}}_N$ can be written as follows:

$$\overline{\mathbf{R}}_N = \mathbf{U}_s\mathbf{\Lambda}_\mathbf{s}\mathbf{U}_\mathbf{s}^* + \sigma^2\mathbf{U}_\mathbf{n}\mathbf{U}_\mathbf{n}^* \quad (8)$$

where $\mathbf{U}_s = [\mathbf{u}_1,...,\mathbf{u}_{p(N+M)}]$ denotes the signal eigenvectors and $\mathbf{U}_n = [\mathbf{u}_{p(N+M)+1},...,\mathbf{u}_{qN}]$ denotes the noise eigenvectors and $\mathbf{\Lambda}_s = diag(\lambda_1,...,\lambda_{p(N+M)})$ with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{p(N+M)} > \sigma^2$ are the signal eigenvalues. It is shown that, range($\mathbf{U}_s$) = range($\mathcal{T}_N(\mathbf{H})$) and range($\mathbf{U}_n$) = range($\mathcal{T}_N(\mathbf{H})$)$^\perp$, i.e. the orthogonal complement subspace to the range space of $\mathcal{T}_N(\mathbf{H})$. Using orthogonality relation between noise and signal subspace leads to:

$$\mathbf{U}_n^*\mathcal{T}_N(\mathbf{H}) = 0 \quad (9)$$

In the original version of subspace method [3], [4], the computation of the hole $qN - p(N+M)$ noise vectors is required to estimate the channel parameters by minimizing $\|\mathbf{U}_n^*\mathcal{T}_N(\mathbf{H})\|^2$ under a suitable constraint.

IV. RATIONAL SUBSPACE AND POLYNOMIAL BASES

The Subspace identification method can be recast in a more general framework by resorting to the concept of rational subspaces. As we shall see below, one can express the signal and noise subspaces in the field of rational functions to get more insights into the subspace method.

A $qN \times 1$ vector $\mathbf{g} = [\mathbf{g}^T(0),\ldots,\mathbf{g}^T(N-1)]^T$ (where each vector $\mathbf{g}(k)$ is $q \times 1$) belongs to the noise subspace of $\overline{\mathbf{R}}_N$ if and only if $\mathbf{g}^*\mathcal{T}_N(\mathbf{H}) = 0$. The orthogonality condition is conveniently rewritten as:

$$\mathbf{g}^*\mathcal{T}_N(\mathbf{H}) = 0 \Longleftrightarrow \mathbf{g}^*(z)\mathbf{H}(z) = 0 \quad \text{for each } z$$

$$\mathbf{g}(z) = \sum_{k=0}^{N}\mathbf{g}(k)z^{-k} \quad \text{and} \quad \mathbf{g}^*(z) = \sum_{k=0}^{N}\mathbf{g}^*(k)z^{-k}$$

We denote by $\mathcal{C}^q(z)$ the set of all q-dimensional rational functions or, in other words, the q-dimensional vector space on the field of all scalar rational functions. Such a vector subspace is referred to as a *rational space*. Let $\mathcal{S}$ be the p-dimensional rational subspace of $\mathcal{C}^q(z)$ spanned by the column vectors of $\mathbf{H}(z)$ ($\mathcal{S}$ = range($\mathbf{H}(z)$). Let $\mathcal{B} \subset \mathcal{C}^q(z)$ denote the orthogonal complement of $\mathcal{S}$ (i.e., the subspace of all q-dimensional rational transfer functions $\mathbf{g}(z)$ satisfying $\mathbf{g}(z)\mathbf{f}^*(z) = 0$ for each $\mathbf{f}(z) \in \mathcal{S}$). It then follows that $\mathcal{B}$ has dimension $q-p$.

The subspace method can now be seen as a method of finding $\mathbf{H}(z)$ such that $\mathbf{H}(z) \perp \mathcal{B}$. However, $\mathcal{B}$ can be uniquely spanned by a basis of $q-p$, q-dimensional polynomial vectors. Therefore, to identify $\mathbf{H}(z)$, it suffices to find a *polynomial basis* $\mathbf{V}(z) = [\mathbf{v}_1(z),...,\mathbf{v}_{q-p}(z)]$ of $\mathcal{B}$ and to express the orthogonality between $\mathbf{v}_i$ and $\mathbf{H}(z)$ for $i = 1,\ldots,q-p$, i.e.

$$\mathbf{v}_i^*(z)\mathbf{H}(z) = 0.$$

V. ORTHOGONAL MINIMUM NOISE SUBSPACE METHOD

In [1], [2], an estimation method called MNS[1] has been introduced to compute the polynomial basis of $\mathcal{B}$. Each polynomial noise vector is obtained from the least eigenvector of a covariance matrix computed from (properly chosen) $(p+1)$-dimensional sub-system outputs.

In this contribution, an alternative method to compute the noise polynomial basis $\mathbf{V}(z) = [\mathbf{v}_1(z),...,\mathbf{v}_{q-p}(z)]$ is proposed. The noise vectors are computed (i) recursively (contrary to the MNS vector in [2] that can be computed in a parallel scheme), (ii) using all system outputs (in [2] each vector was computed using only $p+1$ system outputs), and (iii) in such a way to form an orthogonal basis of $\mathcal{B}$ (this is not the case in [2]), i.e.

$$\mathbf{v}_i^*(z)\mathbf{v}_j(z) = 0 \qquad \text{for} \quad i \neq j \quad (10)$$

At the i-th step, we compute a q-dimensional polynomial noise vector $\mathbf{v}_i(z)$ orthogonal to $\mathbf{H}(z)$ and to the previously computed q-dimensional polynomial noise vectors. Each noise vector is obtained by computing the least eigenvector of a $qN_i \times qN_i$, $(i = 1,...,q-p)$ matrix which is a function of the channel outputs and the previously computed polynomial noise vectors. N_i is chosen in order to obtain a tall block matrix at each step.

More precisely, we have the following algorithm.

1. Initialization:

- Choose N_1 a window length such that $qN_1 > p(M+N_1)$ and estimate the covariance matrix $\overline{\mathbf{R}}_{N_1}$ from the observations.
- Compute $\mathbf{v}_1$ as the least eigenvector of $\overline{\mathbf{R}}_{N_1}$, the latter satisfies:

$$\mathbf{v}_1^*\mathcal{T}_{N_1}(\mathbf{H}) = 0 \Longleftrightarrow \mathbf{v}_1^*(z)\mathbf{H}(z) = 0 \quad (11)$$

[1] It is minimum in the sense that $q-p$ is the minimum number of noise vectors needed to uniquely estimate $\mathbf{H}(z)$.

$\mathbf{v}_1(z) = \sum_{k=0}^{N_1-1} \mathbf{v}_1(k)z^{-k}$ with $\mathbf{v}_1 = [\mathbf{v}_1^T(0), ..., \mathbf{v}_1^T(N_1-1)]^T$.

2. for $i = 2, \ldots, q-p$:

• Choose N_i a window length such that:

$$qN_i > p(M+N_i) + \sum_{j=1}^{i-1}(N_j - 1) \tag{12}$$

and then compute the matrix:

$$\mathbf{M}_i = \overline{\mathbf{R}}_{N_i} + \sum_{j=1}^{i-1} \mathcal{T}_{N_i}(\mathbf{v}_j)\mathcal{T}_{N_i}^*(\mathbf{v}_j) \tag{13}$$

• Compute $\mathbf{v}_i$ as the least eigenvector of $\mathbf{M}_i$. The latter satisfies:

$$\begin{aligned} \mathbf{v}_i^* \mathcal{T}_{N_i}(\mathbf{H}) &= 0 \\ \mathbf{v}_i^* \mathcal{T}_{N_i}(\mathbf{v}_j) &= 0 \quad \text{for} \quad j = 1, \ldots, i-1 \end{aligned} \tag{14}$$

or equivalently :

$$\begin{aligned} \mathbf{v}_i^*(z)\mathbf{H}(z) &= 0 \\ \mathbf{v}_i^*(z)\mathbf{v}_j(z) &= 0 \quad \text{for} \quad j = 1, \ldots, i-1 \end{aligned} \tag{15}$$

$\mathbf{v}_i(z) = \sum_{k=0}^{N_i-1} \mathbf{v}_i(k)z^{-k}$ and $\mathbf{v}_i = [\mathbf{v}_i^T(0), ..., \mathbf{v}_i^T(N_i-1)]^T$.

3. Once the $(q-p)$ noise vectors are computed, estimate the channel matrix $\mathbf{H}(z)$ (up to a constant nonsingular $p \times p$ matrix) as:

$$\begin{aligned} \widehat{\mathbf{H}}(z) &= \text{argmin}_{\mathbf{H}(z)} \sum_{i=1}^{q-p} \|\mathbf{v}_i{}^* \mathcal{T}_{N_i}(\mathbf{H})\|^2 \\ &= \text{argmin}_{\mathbf{H}(z)} \|\widetilde{\mathbf{V}}^* \mathcal{T}_{N_{q-p}}(\mathbf{H})\|^2 \end{aligned} \tag{16}$$

where $\widetilde{\mathbf{V}}_i = [\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_{q-p}]$ with $\widetilde{\mathbf{v}}_i = [\mathbf{v}_i^T, \mathbf{0}_{1,N_{q-p}-N_i}]^T$ ($\mathbf{0}_{i,j}$ being the $i \times j$ all-zero matrix). The minimization in (16) is done under a suitable constraint as shown in [7].

VI. DISCUSSION

Computational complexity:

The computational cost of the MNS method is $\mathcal{O}((q-p)(p+1)^2(N)^2)$ flops comparing to $\mathcal{O}(\sum_{i=1}^{q-p}(N_i q)^2)$ flops for OMNS method when it is $\mathcal{O}((qN)^3)$ for the subspace method[2]. Therefore, MNS method has the least computational complexity. The above computation does not take into account the parallel structure of MNS method, which is an additional advantage of this method. However, the OMNS method remains less complex than subspace method in term of computational complexity. It is shown latter that for a large number of sensors and for small M, the value of N_i for OMNS is small and it remains constant for several iterations, which results a computational cost comparable or sometimes much less than that of the MNS method. Table I provides some examples for the values of the window lengths used in MNS and OMNS in function of the the system parametrs q, p and M.

[2]In this computational costs we didn't include the cost of covariance matrix estimations, i.e. the estimation of $E[\mathbf{y}(n+k)\mathbf{y}^*(n)] \quad k = 0, \ldots, M$ which is same for all considered methods. Also, it assumes that the algorithm implementations are optimized in the sense that they take advantage of the underlying Toeplitz structures to reduce computational complexity.

Performance:

As mentioned before, noise polynomial vectors in MNS method are obtained using only $p+1$ outputs for each of them [2], while in OMNS method each noise vector is computed from all the q system outputs. This leads to an improved (a more robust) channel estimation especially when the number of system outputs is much larger than the number of system inputs. Futhermore, the orthogonality of noise subspace might improve the quality of the parameter estimation. This has been demonstrated for othe subspace based applications, e.g. source localization [8], but performance analysis needs to be performed to asses this point in the context of MIMO system identification.

VII. SIMULATION RESULTS

In this section, the performance of the MNS method is compared with that of the OMNS method via simulation results. We consider $p = 2$ inputs where each input sequence is an i.i.d., zero-mean, unit-variance QAM4 process. Both MNS and OMNS methods estimate the polynomial matrix $\mathbf{H}(z)$ up to a $p \times p$ constant matrix $\mathbf{Q}$. The output observation noise is a sequence of i.i.d., zero-mean, gaussian variables and the number of samples is held constant ($T = 500$). For each experiment $N_r = 100$ independent Monte-Carlo runs are performed. The performance is measured by the mean-square-error (MSE) defined by:

$$MSE = [\sum_{r=1}^{N_r} \|\widehat{\mathbf{H}}_r\mathbf{Q}_r - \mathbf{H}\|^2/N_r]^{\frac{1}{2}}$$

Where $\widehat{\mathbf{H}}_r$ is an estimate of $\mathbf{H} \triangleq [\mathbf{H}(0)...\mathbf{H}(M)]^T$ at the r-th run, and $\mathbf{Q}_r$ is chosen so that $\|\widehat{\mathbf{H}}_r\mathbf{Q}_r - \mathbf{H}\|$ is minimum (This is to get rid of the constant matrix indeterminency). The channel transfer function associated with the first output corresponds to the same impulse reponse:

$$\begin{aligned} &h_{i,1}(k) = h_{i,2}(k) = ... \\ &= \lambda_0 g(kT_s) + \lambda_1 g(kT_s - \tau_1) + ... + \lambda_L g(kT_s - \tau_L) \end{aligned}$$

L denotes the number of paths and $g(t)$ is generated from the raised cosine spectrum pulse with the roll-off factor

	Channel parameters	N_{OMNS}	N_{MNS}
Fig 1	10×2 and $M = 2$	$(1, 1, 1, 1, 1, 2, 2.4)$	5
Fig 2	6×2 and $M = 2$	$(2, 2, 4, 10)$	5
Fig 3	4×2 and $M = 2$	$(3, 7)$	5
Fig 4	4×2 and $M = 5$	$(6, 16)$	11

TABLE I

CHANNEL PARAMETERS AND LENGTH OF THE PROCESSING WINDOW USED IN EXPERIMENTS.

equal to $1/2$. Then it is delayed and sampled at the rate of 270 kb/sec (T_s = 3.7 μs). The resulted channel impulse response is windowed such that the polynomial degree is M. τ_i denotes the delay and λ_i the attenuation. Attenuation is considered equal to -5 dB for all of the paths and the delay τ_i is a multiple of path number. i.e. $\tau_i = i \times 3.2\mu$sec (for $i = 1, \ldots, L$).

The other channel transfer functions are generated by assuming a plane propagation model of each path with corresponding electric angles uniformely distributeed in $[0, \pi/2]$, (i.e. $h_{i,l}(z) = \sum_{t=0}^{L} \lambda_L g(kT_s - \tau_t) e^{jl\theta_t} z^{-t}$ with $\theta_{tl} \in [0, \pi/2]$).

Figures (1) to (4) show the comparative performances of MNS and OMNS for different choices of channel parametrs q, p and M. In the figures the MSE of channel parameter estimates are plotted against the SNR, defined as the inverse noise power. As expected, the performance gain of OMNS is more significient when $q - p$ is large. For $q - p$ small and large channel degree (Fig. 4) the performance of OMNS is slightly deteriored in comparision with that of MNS. This is possibly due (but need to be certified by a theoritical study) to the large window sizes that are needed to compute the OMNS basis leading to a large number of parameters (here the noise vector coefficients) to be estimated.

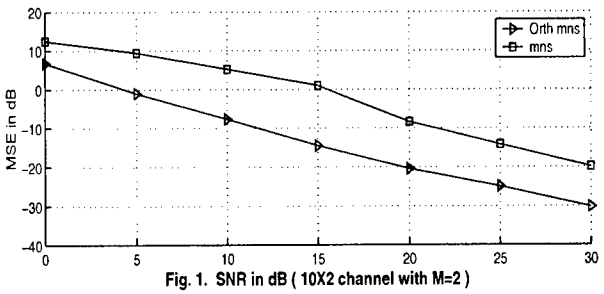


Fig. 1. SNR in dB (10X2 channel with M=2)

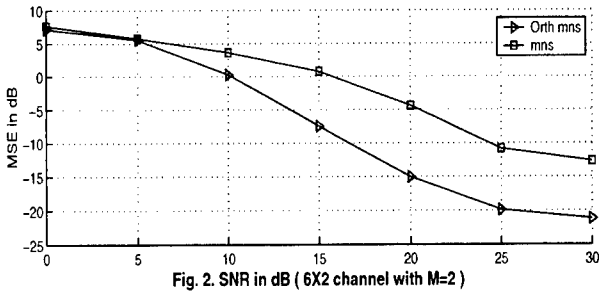


Fig. 2. SNR in dB (6X2 channel with M=2)

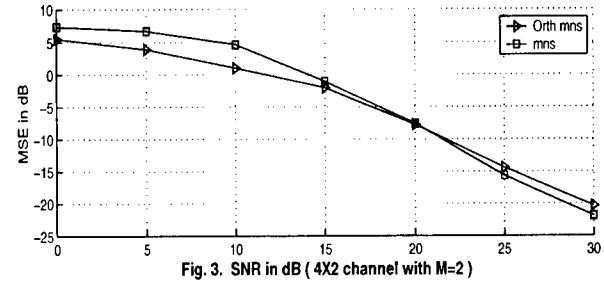


Fig. 3. SNR in dB (4X2 channel with M=2)

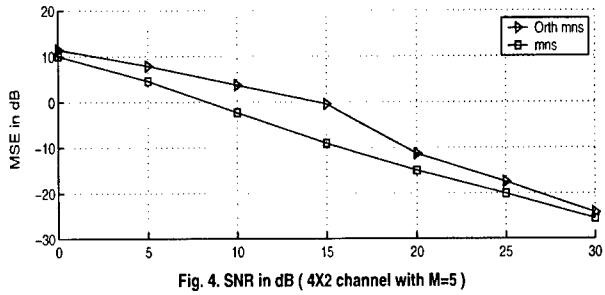


Fig. 4. SNR in dB (4X2 channel with M=5)

REFERENCES

[1] Y. Hua, K. Abed-Meraim and M. Wax "Blind system identification using minimum noise subspace" *IEEE Trans. on Signal processing*, vol. 45, pp. 770-773. no. 3, March 1997.

[2] K. Abed-Meraim and Y. Hua "Blind identification of Multi-Input Multiple-Output system using Minimum Noise Subspace" *IEEE Trans. on Signal processing*, vol. 45, pp. 254-258. no. 1, January 1997.

[3] E. Moulines, P. Duhamel, J. Cardoso and S. Mayrargue "Subspace methods for the blind identification of the multichannel FIR filters" *IEEE Trans. on Signal processing*, vol. 43, pp. 516-525. February 1995.

[4] K. Abed-Meraim, Ph. Loubaton and E. Moulines "A Subspace algorithm for certain blind identification problems" *IEEE Trans. on Information Theory*, vol. 43, pp. 499-511. March 1997.

[5] T. Kailath, *Linear Systems* . Englewood Cliffs, NJ:Prentice-Hall, 1980.

[6] G. Forney "Minimal Bases of rational vector spaces, with applications to multivariable linear systems" *SIAM J. Contr.*, vol. 13, pp. 493-520. Mar. 1996.

[7] A. Gorokhov and P. Loubaton, "Subspace based techniques for second order blind sepqrqtion of convolutive mixtures with temporally corrolated sources," *IEEE Trans. Circuits Syst.*, vol. 44, pp. 813-820. Sept. 1997.

[8] S. Marcos and M. Benidir, "Sorce bearing estimation and sensor positioning with the propagator method." *in Proc. SPIE (San Diego, USA)*, vol. 1348, pp. 312-323. Sept. 1990.

RECURSIVE SEMI-BLIND EQUALIZER FOR TIME-VARYING MIMO CHANNELS

Mihai Enescu, Marius Sirbu* and Visa Koivunen*

Signal Processing Laboratory
Helsinki University of Technology
P.O. Box 3000, FIN-02015 HUT, Finland

ABSTRACT

This paper addresses the problem of semi-blind Multi-Input Multi-Output (MIMO) equalization of time-varying channels by employing recursive filtering methods. Based on a realistic channel model described in COST 207 project, we derive a state-space model that characterizes the behavior of the channel in time. The channel estimation and tracking are performed using a Kalman filter method, and a decision feedback equalizer derived using MMSE criterion is used to perform the equalization.

1. INTRODUCTION

MIMO channels with Intersymbol Interference (ISI) and Inter-User Interference (IUI) arise in many applications including wireless communications. In addition, the time-varying nature of the wireless channels makes the equalization even more difficult to achieve. Deriving a model that describes the system's time evolution can be a very difficult task taking into account that the time-varying parameters are not directly observed and the model has to be realistic.

In this paper we derive a semi-blind MIMO algorithm capable of identifying, tracking and equalizing a Time-Varying Channel (TVC). Semi-blind algorithms need some training data for channel acquisition and then they run blindly. The advantages of the estimation and tracking stages can be summarized as follows: the estimator is akin to the usual Kalman filter, it is thus an exact solution to the estimation problem. Combining this structure with a multichannel Decision Feedback Equalizer (DFE) we get a true real-time algorithm in the sense that it is recursive in time and the storage space needed to evaluate the estimates remains constant, as time progresses and the amount of received data increases. This means that it is also feasible to cope with a large number of parameters. Another method combining DFE and Kalman was proposed in [2]. In our paper the measurement and process noise variances used in Kalman filter are estimated using the received data [1] and a new MIMO structure for DFE is derived based on MMSE criterion. Simulations are carried out using realistic channels.

This paper is organized as follows. The system model is presented first. Then a description of the proposed algorithm is given. In section 4, simulation results of equalization for realistic MIMO channels are presented.

*This work was supported by Nokia and Academy of Finland.

2. SYSTEM MODEL

Let consider a MIMO system with m source signals and n sensors at the receiver. The received observations from sensor j (with j=1,...,n) at time t are given by:

$$y_j(t) = \sum_{i=1}^{m} \sum_{l=0}^{L_{ij}-1} h_{ij}(l) x_i(t-l) + v_j(t) \qquad (1)$$

where $x_i(t-l)$ is the symbol drawn from a constellation $\mathcal{X}$ of the i-th user at time $t-l$, $h_{ij}(l)$ is the impulse response of the TVC, $y_j(t)$ is the received signal, and $v_j(t)$ is the additive Gaussian noise with variance σ_v^2. Setting $L = \max L_{ij}$, the channel length, we obtain the following vector form:

$$\mathbf{y}(t) = \sum_{l=0}^{L} \mathcal{H}_l(t)\mathbf{x}(t-l) + \mathbf{v}(t) \qquad (2)$$

where $\mathbf{y}$ is a column vector of n received signals, $\mathbf{x}$ is a column vector of m transmitted signals, $\mathcal{H}_l(t)$ is a $n \times m$ matrix containing the channel taps and $\mathbf{v}$ is an additive noise vector. For simplicity, a 2×2 MIMO model is presented in Figure 1.

Let assume that the j-th received signal is a superposition of N_p paths. The resulting channel impulse response can then be described using the Gaussian distributed *Wide-Sense Stationary with Uncorrelated Scattering* (WSSUS) model [6]:

0-7803-7011-2/01/$10.00 ©2001 IEEE

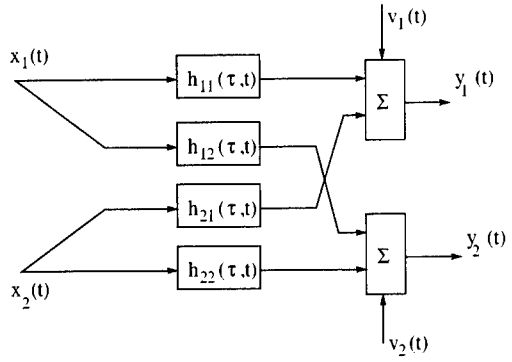


Figure 1: The (2,2) MIMO system model

$$h_{ij}(t,\tau) = \frac{1}{\sqrt{N_p}} \sum_{p=1}^{N_p} e^{j(2\pi f_{d,p} t + \theta_p)} h_{RF}(\tau - \tau_p) \quad (3)$$

where $f_{d,p}$ is the Doppler spread, θ_p is the angular spread, τ_p is the delay spread of the path p and $h_{RF}(t)$ is the impulse response of the receive filter.

Four propagation environments are widely used: Typical Urban (TU), Bad Urban (BU), Hilly Terrain (HT) and Rural Area (RA), each of them having specific parameter values. This model is suitable for many channels of practical interest in mobile wireless communications, which is our concern in this paper.

Considering that the m-dimensional transmitted sequence is a white sequence drawn from a PSK constellation and if the n-dimensional received signal is sampled at symbol rate we get the following discrete-time model:

$$\mathbf{y}(k) = X(k)\mathbf{h}(k) + \mathbf{v}(k) \quad (4)$$

where X is a $n \times nmL$ data matrix defined as:

$$\begin{aligned} X(k) &= [x_1(k)I_m \dots x_n(k)I_m \dots \\ &\quad x_1(k-L+1)I_m \dots x_n(k-L+1)I_m] \end{aligned} \quad (5)$$

I_m is an $m \times m$ identity matrix and

$$\begin{aligned} \mathbf{h}(k) &= \left[h_{11}^0(k) \dots h_{1n}^0(k) \dots h_{m1}^0(k) \dots h_{mn}^0(k) \dots \right. \\ &\quad \left. h_{11}^{L-1}(k) \dots h_{1n}^{L-1}(k) \dots h_{m1}^{L-1}(k) \dots h_{mn}^{L-1}(k)\right]^T \end{aligned} \quad (6)$$

is a vector of length nmL containing channel coefficients.

3. RECURSIVE ALGORITHM

The proposed algorithm consists of two stages: first we deal with channel aquisition using training data and in the second stage we perform channel tracking and equalization. This type of structure allows for real-time implementation.

3.1. Channel aquisition

In this section we are interested in estimating the channel coefficients $h_{ij}(l,k)$ using limited training data. Our algorithm is based on the well known Kalman filter [5]. In matrix notation we have the following state space equations:

$$\begin{aligned} \mathbf{y}(k) &= X(k)\mathbf{h}(k) + \mathbf{v}(k) & (7) \\ \mathbf{h}(k) &= A\mathbf{h}(k-1) + \mathbf{w}(k) & (8) \end{aligned}$$

where $X(k)$ contains transmitted symbols, $\mathbf{h}(k)$ are the channel taps at time instant k and A is the state transition matrix, in our case an identity matrix. Noises $\mathbf{v}$ and $\mathbf{w}$ are mutually uncorrelated, white noise sequences with covariance matrices R and Q. These covariance matrices may be estimated prior to perfoming the equalization based on the whiteness property of the innovation sequence in optimum Kalman filtering [1].

During the training period the transmitted symbols are known to the receiver. Let us denote them by $X_{training}$ according to (5). The Kalman filter equations can be summarized as follows:

$$\begin{aligned} \hat{\mathbf{h}}(k|k-1) &= A\hat{\mathbf{h}}(k-1|k-1) & (9) \\ P(k|k-1) &= AP(k-1|k-1)A^T + Q \\ K(k) &= \frac{P(k|k-1)X_{training}}{X_{training}{}^T P(k|k-1)X_{training} + R} \\ \hat{\mathbf{h}}(k|k) &= \hat{\mathbf{h}}(k|k-1) + K(k)(\mathbf{y}(k) - \\ &\quad X_{training}{}^T \hat{\mathbf{h}}(k|k-1)) \\ P(k|k) &= P(k|k-1) - K(k)X_{training}{}^T P(k-1|k-1) \end{aligned}$$

Thus, the filtered estimates of channel taps at time instant k are given by $\hat{\mathbf{h}}(k|k)$.

3.2. Equalization

In the previous section, we described how to estimate the channel taps. The remaining task is to perform equalization in order to get estimates of the desired symbols. A MIMO DFE based on MMSE criterion is derived. Let us start by defining the channel convolution matrices $\hat{H}_{ij}$ of dimension $N_{ch} \times N_f$, where $N_{ch} = L + N_f - 1$ and N_f is the feedforward filter

length.

$$\hat{H}_{ij}(k) = \begin{pmatrix} \hat{h}_{ij(0;k)} & \cdots & 0 \\ \hat{h}_{ij(1;k)} & \ddots & \vdots \\ \hat{h}_{ij(2;k)} & & 0 \\ \vdots & & \hat{h}_{ij(0;k)} \\ \hat{h}_{ij(L_h-1;k)} & \ddots & \hat{h}_{ij(1;k)} \\ 0 & & \hat{h}_{ij(2;k)} \\ \vdots & & \vdots \\ 0 & & \hat{h}_{ij(L_h-1;k)} \end{pmatrix} \quad (10)$$

where $\hat{h}_{ij}$ are obtained from the Kalman filter part.

Applying the feedforward filter to the past N_f received observations and the feedback filter to the past N_d estimated symbols for each output we get the soft estimate:

$$\hat{z}_i(k) = \sum_{q=1}^{N_f} \mathbf{f}_{iq} y_i(k-q) - \sum_{q=1}^{N_d} \mathbf{d}_{iq} \hat{x}_i(k-q) \quad (11)$$

Equalization is achieved via feedforward $\mathbf{f}_i = (f_{i1}, \ldots, f_{iN_f})^T$, and feedback $\mathbf{d}_i = (d_{i1}, \ldots, d_{iN_d})^T$ filters. These filters are obtained by minimizing the MSE cost function with respect to $\mathbf{f_i}$ and $\mathbf{d_i}$:

$$\mathcal{J}_i = E\{(x_i(k-\delta) - \hat{z}_i(k))^2\} \quad (12)$$

where δ is the equalization delay.

For illustration purposes, let us consider the simplest case of a 2×2 MIMO system. The soft estimates at receivers 1 and 2 are:

$$\begin{aligned} \hat{z}_1 &= \mathbf{x}_1^T H_{11}\mathbf{f}_1 + \mathbf{x}_2^T H_{21}\mathbf{f}_1 + \mathbf{v}_1^T\mathbf{f}_1 - \hat{\mathbf{x}}_1^T\mathbf{d}_1 \quad (13) \\ \hat{z}_2 &= \mathbf{x}_1^T H_{12}\mathbf{f}_2 + \mathbf{x}_2^T H_{22}\mathbf{f}_2 + \mathbf{v}_2^T\mathbf{f}_2 - \hat{\mathbf{x}}_2^T\mathbf{d}_2 \end{aligned}$$

where $\mathbf{x}_i = [x_i(k), \ldots, x_i(k-N_{ch}-1)]^T$ is a vector of transmitted symbols from user i, $i = \{1,2\}$, $\hat{\mathbf{x}}_j = [\hat{x}_j(k-1), \ldots, \hat{x}_j(k-N_d)]^T$ is a vector of estimated symbols from receiver j, $j = \{1,2\}$, $\mathbf{v}_j = [v_j(k-1), \ldots, v_j(k-N_f)]^T$ is the noise vector at the receiver j. The past decisions of the equalizer are assumed to be correct.

The gradient of $\mathcal{J}_1$ with respect to $\mathbf{d}_1$ is:

$$\begin{aligned} \nabla_{\mathbf{d}_1}\mathcal{J}_1 = 2E\{&\hat{\mathbf{x}}_1 x_1(k-\delta) - \hat{\mathbf{x}}_1\mathbf{x}_1^T H_{11}\mathbf{f}_1 - \quad (14) \\ &\hat{\mathbf{x}}_1\mathbf{x}_2^T H_{21}\mathbf{f}_1 - \hat{\mathbf{x}}_1\hat{\mathbf{v}}_1^T\mathbf{f}_1 + \hat{\mathbf{x}}_1\hat{\mathbf{x}}_1^T\mathbf{d}_1\} \end{aligned}$$

Assuming that the input sequences are uncorrelated with each other and with the noise, the above expression simplifies to:

$$\nabla_{\mathbf{d}_1}\mathcal{J}_1 = -2\sigma^2 M H_{11}\mathbf{f}_1 + 2\sigma^2 I\mathbf{d}_1 \quad (15)$$

where M is an $N_d \times N_{ch}$ matrix having the structure $M = (0_{Nd\times\delta}\ I_{N_d\times N_d}\ 0_{N_d\times N_{ch}-N_d-\delta})$, I is an identity matrix, $E\{\hat{\mathbf{x}}_1\mathbf{x}_1^T\} = \sigma^2 M$ and σ^2 is the variance of the input signal. The MMSE feedback filter can be written as:

$$\mathbf{d}_1 = M\hat{H}_{11}\mathbf{f}_1 \quad (16)$$

In a similar way we find $\mathbf{d}_2 = M\hat{H}_{22}\mathbf{f}_2$. Note that during the theoretical derivation, the notation H_{ij} was used. However, in the receiver we do not have knowledge of the real channel, thus an estimate $\hat{H}_{ij}$ is used instead.

Similarly, for the feedforward parameters we have :

$$\begin{aligned} \nabla_{\mathbf{f}_1}\mathcal{J}_1 = 2\,(&\sigma^2 H_{11}^T H_{11}\mathbf{f}_1 - \sigma^2 H_{11}^T M^T\mathbf{d}_1 + \quad (17) \\ &\sigma^2 H_{21}^T H_{21}\mathbf{f}_1 + \sigma_v^2\mathbf{f}_1 - \sigma^2 H_{11}^T e_\delta) \end{aligned}$$

where $e_\delta = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is the standard basis vector, with one at the position δ, $0 \le \delta \le N_f$. Substituting $\mathbf{d}_1 = M\hat{H}_{11}\mathbf{f}_1$ and denoting with $P_{DFE} = (I - M^T M)$, the MMSE feedforward filter is given by:

$$\mathbf{f}_1 = (\hat{H}_{11}^T P_{DFE}\hat{H}_{11} + \hat{H}_{21}^T\hat{H}_{21} + \lambda I)^{-1}\hat{H}_{11}^T e_\delta \quad (18)$$

For the second receiver we have: $\mathbf{f}_2 = (\hat{H}_{22}^T P_{DFE}\hat{H}_{22} + \hat{H}_{12}^T\hat{H}_{12} + \lambda I)^{-1}\hat{H}_{22}^T e_\delta$. A comprehensive derivation for a $m \times n$ MIMO case is given in [1].

Finally, the symbol estimate at time k is obtained by:

$$\hat{x}_i(k) = \arg\min_{\alpha\in\mathcal{X}} |\alpha - \hat{z}_i(k)| \quad (19)$$

where $\mathcal{X}$ is a finite alphabet.

3.3. Practical implementation of the algorithm

The algorithm operates in two modes:

Training mode. In the training mode only the Kalman filter is running.

Step 1. Obtain the observations $\mathbf{y}(k)$ and generate the local training data sequence $X(k)_{training}$, $k = 0, \ldots, N_{train} - 1$, where N_{train} is the length of the training sequence.

Step 2. Estimate the channel coefficients $\hat{\mathbf{h}}(k)$ and noise statistics [1] by running the Kalman algorithm described by the set of Equations (9).

Blind mode. In the blind mode both DFE and Kalman algorithms are running in an alternating manner. We assume that the $\hat{\mathbf{h}}(k)$ has been estimated during the training period and we use estimated symbols instead, $X_{training} = \hat{X}$.

Step 1. Run DFE algorithm and estimate transmitted sequence $\hat{X}(k)$.

Step 2. Having $\hat{X}(k)$ run Kalman algorithm and obtain $\hat{\mathbf{h}}(k)$. The channel estimate $\hat{\mathbf{h}}(k)$ is used at next step $k+1$ by the DFE.

4. SIMULATIONS

In the simulations we use a linearized GMSK signal [4]. The pulse shape of this modulation is used as the receive filter impulse response. A training sequence of 50 symbols is used for the algorithm initialization. After the training stage, the algorithm keeps the track without additional training data. For each simulation we consider 100 Monte Carlo realizations.

The downlink connection in a cellular communication system with the carrier frequency of 900 MHz is considered. The simulations are done for 'Hilly Terrain' propagation environment with the receiver speed of 100 km/h (HT 100). The corresponding maximum Doppler shift is 83.3 Hz. The estimated channel magnitudes for HT100 direct channels are presented in Figures 2 and Figure 3, respectively. Note that Kalman algorithm needs only few observations in order to find the true channel coefficients.

The symbol error rate (SER) for each user at different SNR is shown in Figure 4. We note that the quality of reception is different for the two users. This is due to the fact that we have different IUI powers for the two users.

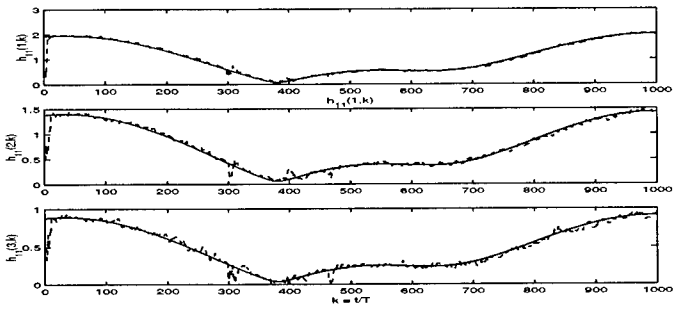


Figure 2: The estimated channel magnitude (dashed line) and the true channel magnitude (solid line) for the main path h_{11}, HT at 100 km/h, SNR=15dB.

5. CONCLUSIONS

A real-time MIMO TVC equalization technique is derived. The channel tracking is performed using a Kalman filter and the transmitted symbol estimation is done by a novel MIMO DFE structure. The channel is a fast TVC with the coherence time equal to the symbol period. The channel model fits very well to the wireless communication problem, when the signal arrives

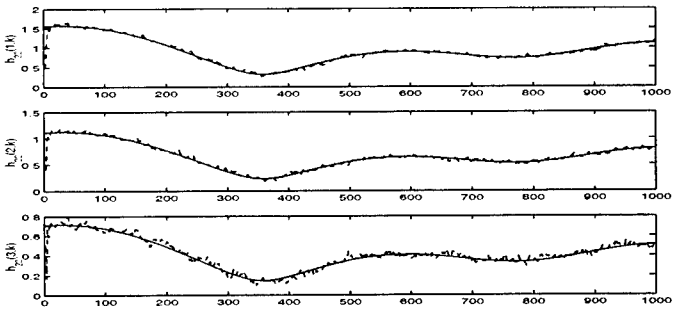


Figure 3: The estimated channel phase (dashed line) and the true channel phase (solid line) for the main path h_{22}, HT at 100 km/h, SNR=15dB.

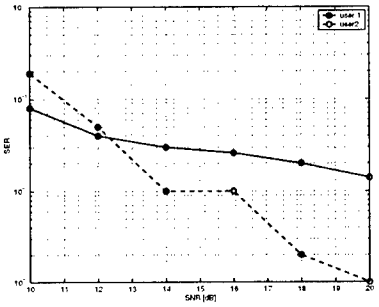


Figure 4: SER vs. SNR

to the receiver from different paths with different delay spread, angular spread and Doppler spread. The simulation results show that the algorithm achieves a good performance even in very demanding channel conditions. The main improvement is that it needs only a single training period at the beginning of the transmission, after that it runs blindly.

REFERENCES

[1] M. Enescu, M. Sirbu, V. Koivunen, "Adaptive Equalization of time-varying MIMO channels", submitted to Signal Processing.

[2] Komninakis, C.; Fragouli, C.; Sayed, A.H.; Wesel, R.D., "Adaptive multi-input multi-output fading channel equalization using Kalman estimation", IEEE International Conference on Communications, vol. 3 , 2000, 1655 -1659

[3] J. Salz,"Optimum mean-square DFE", Bell system Technical Journal, Vol.52, No.8, October 1973.

[4] Z. Ding, G. Li,"Single-Channel Blind Equalization for GSM Cellular Systems", IEEE J. Select. Areas. Commun., Vol. 16, 1493 - 1505, Oct. 1998.

[5] S. Haykin, "Adaptive Filter Theory", 3rd edition, Prentice-Hall, 1996.

[6] P. Hoeher, "A Statistical Discrete-Time Model for the WSSUS Multipath Channel", IEEE Trans. on Vehicular Technology, vol. 41, no. 4, 1992, 461-468.

BLIND SINGLE-INPUT MULTI-OUTPUT (SIMO) CHANNEL IDENTIFICATION WITH APPLICATION TO TIME DELAY ESTIMATION

Chong-Yung Chi, Xianwen Chang and Chii-Horng Chen

Department of Electrical Engineering &
Institute of Communications Engineering
National Tsing Hua University, Hsinchu, Taiwan, R.O.C.
Tel: +886-3-5731156, Fax: +886-3-5751787, E-mail: cychi@ee.nthu.edu.tw

ABSTRACT

In this paper, with a given set of non-Gaussian measurements, a cumulant based single-input multi-output (SIMO) blind channel estimation (BCE) algorithm is proposed that uses multi-input multi-output (MIMO) inverse filter criteria (blind deconvolution criteria using higher-order cumulants) proposed by Tugnait, and Chi and Chen. Then a time delay estimation (TDE) algorithm is proposed that estimates $P-1$ time delays from the phase information of the estimated single-input P-output ($P \geq 2$) system obtained by the proposed SIMO BCE algorithm. Some simulation results are presented to support the efficacy of the proposed SIMO BCE and TDE algorithms.

1. INTRODUCTION

Blind channel estimation (BCE) for single-input multi-output (SIMO) systems is a problem of estimating a $P\times 1$ linear time-invariant (LTI) system, denoted $\mathbf{h}[n] = (h_1[n], h_2[n], ..., h_P[n])^T$, with only a set of non-Gaussian vector output measurements $\mathbf{x}[n] = (x_1[n], x_2[n], ..., x_P[n])^T$ as follows

$$\mathbf{x}[n] = \sum_{k=-\infty}^{\infty} \mathbf{h}[k]u[n-k] + \mathbf{w}[n] \qquad (1)$$

where $u[n]$ is the non-Gaussian driving input signal and $\mathbf{w}[n] = (w_1[n], w_2[n], ..., w_P[n])^T$ is additive noise. The SIMO LTI system arises in science and engineering areas where multiple sensors are needed such as time delay estimation [1] and seismic signal processing, etc. In communications, multiple antennas receiving signals and fractionally-spaced signal processing at receiver can also be modeled as SIMO LTI systems [2].

This work was supported by the National Science Council under Grant NSC 89-2213-E007-132.

2. BCE FOR SIMO LTI SYSTEMS

Let $\text{cum}\{y_1, y_2, ..., y_p\}$ denote the pth-order joint cumulant [3] of random variables $y_1, y_2, ..., y_p$ and

$$C_{p,q}\{y\} = \text{cum}\{y_1 = y_2 = \cdots = y_p = y, y_{p+1} = y_{p+2} = \cdots = y_{p+q} = y^*\} \qquad (2)$$

where y^* is the complex conjugate of y. Let $\mathcal{F}\{\bullet\}$ and $\mathcal{F}^{-1}\{\bullet\}$ denote the discrete-time Fourier transform and inverse Fourier transform operators, respectively. Assume that we are given a set of non-Gaussian measurements $\mathbf{x}[n]$, $n = 0, 1, ..., N-1$, modeled by (1) with the following assumptions:

- ($\mathcal{A}1$) $u[n]$ is zero-mean, independent identically distributed (i.i.d.), non-Gaussian and $C_{p,q}\{u[n]\} \neq 0$ for a chosen (p,q), where p and q are nonnegative integers and $p+q \geq 3$.
- ($\mathcal{A}2$) The SIMO system $\mathbf{h}[n]$ is stable.
- ($\mathcal{A}3$) The noise $\mathbf{w}[n]$ is zero-mean Gaussian (which can be spatially correlated and temporally colored) and statistically independent of $u[n]$.

Let $\mathbf{v}[n] = (v_1[n], v_2[n], ..., v_P[n])^T$ be a $P \times 1$ FIR inverse filter (deconvolution filter) for which $\mathbf{v}[n] = \mathbf{0}$ for $n < L_1$ and $n > L_2$, and let $e[n]$ be the inverse filter output, i.e.,

$$\begin{aligned} e[n] &= \sum_{l=L_1}^{L_2} \mathbf{v}^T[l] \cdot \mathbf{x}[n-l] \\ &= \sum_{k=-\infty}^{\infty} s[k] \cdot u[n-k] + \sum_{l=L_1}^{L_2} \mathbf{v}^T[l]\mathbf{w}[n-l] \end{aligned} \qquad (3)$$

where $s[n]$ is the overall system given by

$$s[n] = \sum_{l=L_1}^{L_2} \mathbf{v}^T[l]\mathbf{h}[n-l]. \qquad (4)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

Chi and Chen [4] design the inverse filter $\mathbf{v}[n]$ by maximizing the following multi-input multi-output inverse filter criteria (MIMO-IFC)

$$J_{p,q}(\mathbf{v}[n]) = \frac{|C_{p,q}\{e[n]\}|}{|C_{1,1}\{e[n]\}|^{(p+q)/2}} \tag{5}$$

where p and q are nonnegative integers and $p+q \geq 3$. They also proposed a fast iterative MIMO-IFC based algorithm [5] for obtaining the optimum inverse filter $\mathbf{v}[n]$ for $p+q \geq 3$ as $\mathbf{x}[n]$ is real and $p = q \geq 2$ as $\mathbf{x}[n]$ is complex. Based on the relation between the optimum $\mathbf{v}[n]$ and the MIMO linear minimum mean square error equalizer reported in [5], one can show the following fact on which the BCE algorithm for SIMO systems below is based:

Fact 1. Assume that $\boldsymbol{V}(\omega) = \mathcal{F}\{\mathbf{v}[n]\}$ is the optimum inverse filter associated with $J_{p,p}(\mathbf{v}[n])$ with $L_1 \to -\infty$ and $L_2 \to \infty$. Let

$$\begin{aligned} g_p[n] &= s^p[n](s^*[n])^{p-1} && (6)\\ G_p(\omega) &= \mathcal{F}\{g_p[n]\} && (7)\\ \mathcal{R}(\omega) &= \mathcal{F}\{\mathbf{R}[k]\} = \mathcal{F}\{E[\mathbf{x}[n]\mathbf{x}^H[n-k]]\}. && (8) \end{aligned}$$

Then

$$\boldsymbol{H}^*(\omega) = (\mathcal{F}\{\mathbf{h}[n]\})^* = \alpha \frac{\mathcal{R}^T(\omega)\boldsymbol{V}(\omega)}{G_p(\omega)} \tag{9}$$

where α is a non-zero constant.

SIMO BCE Algorithm:

Step 1. Blind Deconvolution.

With finite data $\mathbf{x}[n]$, obtain the inverse filter $\mathbf{v}[n]$ associated with $J_{p,p}(\mathbf{v}[n])$ using Chi and Chen's fast MIMO-IFC algorithm [5], and its $\mathcal{L}$-point FFT $\boldsymbol{V}(\omega_k)$, where $\omega_k = 2\pi k/\mathcal{L}$, $k = 0, 1, ..., \mathcal{L}-1$. Obtain $\mathcal{R}(\omega_k)$ using multichannel Levinson recursion algorithm [6].

Step 2. Channel Estimation.

(S1) Set $i = 0$. Set initial values $\boldsymbol{H}^{(0)}(\omega_k)$ and convergence tolerance $\epsilon_h > 0$.

(S2) Update i by $i+1$. Compute

$$S^{(i-1)}(\omega_k) = \left(\boldsymbol{H}^{(i-1)}(\omega_k)\right)^T \boldsymbol{V}(\omega_k) \tag{10}$$

by (4) and its $\mathcal{L}$-point inverse FFT $s^{(i-1)}[n]$.

(S3) Compute $g_p[n]$ using (6) with $s[n] = s^{(i-1)}[n]$ and its $\mathcal{L}$-point FFT $G_p(\omega_k)$.

(S4) Compute

$$\boldsymbol{H}^{(i)}(\omega_k) = \left(\frac{1}{G_p(\omega_k)} \cdot \mathcal{R}^T(\omega_k)\boldsymbol{V}(\omega_k)\right)^* \tag{11}$$

by (9) which is then normalized by $\sum_{k=0}^{\mathcal{L}-1} \|\boldsymbol{H}^{(i)}(\omega_k)\|^2 = 1$.

(S5) If

$$\sum_{k=0}^{\mathcal{L}-1} \| \boldsymbol{H}^{(i)}(\omega_k) - \boldsymbol{H}^{(i-1)}(\omega_k) \|^2 > \epsilon_h$$

then go to (S2), otherwise $\widehat{\boldsymbol{H}}(\omega_k) = \boldsymbol{H}^{(i)}(\omega_k)$ (except for a scale factor) and its $\mathcal{L}$-point inverse FFT $\widehat{\mathbf{h}}[n]$ are obtained.

Two worthy remarks regarding the proposed SIMO BCE algorithm are as follows.

(R1) The region of support associated with the estimate $\widehat{\mathbf{h}}[n]$ can be arbitrary as long as the FFT size $\mathcal{L}$ is chosen sufficiently large so that aliasing effects on the resultant $\widehat{\mathbf{h}}[n]$ are negligible.

(R2) The obtained estimate $\widehat{\boldsymbol{H}}(\omega)$ is robust against Gaussian noise because (9) is true regardless of the value of signal-to-noise ratio (SNR), although the inverse filter $\mathbf{v}[n]$ and the power spectrum $\mathcal{R}(\omega)$ depend on SNR.

3. TIME DELAY ESTIMATION (TDE)

In time delay estimation, a single source signal, denoted $\widetilde{s}[n]$, is received by P (≥ 2) spatially separate sensors. The received signal vector $\widetilde{\mathbf{x}}[n]$ can be modeled as

$$\begin{aligned} \widetilde{\mathbf{x}}[n] &= \widetilde{\mathbf{s}}[n] + \widetilde{\mathbf{w}}[n] \\ &= (\widetilde{s}[n], a_1\widetilde{s}[n-d_1], ..., a_{P-1}\widetilde{s}[n-d_{P-1}])^T + \widetilde{\mathbf{w}}[n] \quad (12) \end{aligned}$$

where a_i and d_i, $i = 1, 2, ..., P-1$ are amplitudes and time delays, respectively, $\widetilde{s}[n]$ is a wide-sense stationary, colored non-Gaussian signal modeled by

$$\widetilde{s}[n] = \sum_{k=-\infty}^{\infty} h[k]u[n-k] \tag{13}$$

in which $h[n]$ is a stable LTI system and $u[n]$ is zero-mean, i.i.d. non-Gaussian, and $\widetilde{\mathbf{w}}[n]$ is a $P \times 1$ additive Gaussian noise vector which can be spatially correlated and temporally colored.

From (12) and (13), one can easily see that $\widetilde{\mathbf{x}}[n]$ can also be expressed as an SIMO model as follows

$$\widetilde{\mathbf{x}}[n] = \sum_{k=-\infty}^{\infty} \widetilde{\mathbf{h}}[k]u[n-k] + \widetilde{\mathbf{w}}[n] \tag{14}$$

where

$$\widetilde{\mathbf{h}}[n] = (h[n], a_1 h[n-d_1], ..., a_{P-1}h[n-d_{P-1}])^T. \tag{15}$$

Let

$$\begin{aligned} \widetilde{\boldsymbol{H}}(\omega) &= \mathcal{F}\{\widetilde{\mathbf{h}}[n]\} && (16)\\ \widetilde{\boldsymbol{\phi}}(\omega) &= (\phi_1(\omega), ..., \phi_P(\omega))^T = \arg\{\widetilde{\boldsymbol{H}}(\omega)\} && (17)\\ \boldsymbol{B}(\omega) &= (1, e^{j(\phi_2(\omega)-\phi_1(\omega))}, ..., e^{j(\phi_P(\omega)-\phi_1(\omega))})^T && (18) \end{aligned}$$

It can be easily shown that

$$\begin{aligned}\mathbf{b}[n] &= (b_1[n], b_2[n], ..., b_P[n])^T = \mathcal{F}^{-1}\{\boldsymbol{B}(\omega)\} \\ &= (\delta[n], \delta[n-d_1], ..., \delta[n-d_{P-1}])^T. \end{aligned} \quad (19)$$

TDE Algorithm:

(T1) Process $\widetilde{\mathbf{x}}[n]$ using the proposed SIMO BCE algorithm to estimate $\widetilde{\boldsymbol{H}}(\omega_k)$, $k = 0, 1, ..., \mathcal{L}-1$, and then obtain its phase $\widetilde{\boldsymbol{\phi}}(\omega_k)$.

(T2) Obtain $\widehat{\boldsymbol{B}}(\omega_k)$ using (18) and its inverse $\mathcal{L}$-point FFT $\widehat{\mathbf{b}}[n]$. Then the estimate $\widehat{d}_i$ is obtained as

$$\widehat{d}_i = \arg\{\max_n\{|\widehat{b}_{i+1}[n]|\}\}, \quad i = 1, 2, ..., P-1 \quad (20)$$

by (19).

4. SIMULATION RESULTS

A. Simulation Results for the Proposed SIMO BCE Algorithm

Consider a 2-channel MA(6) system taken from [7] whose transfer function was

$$\boldsymbol{H}(z) = \begin{bmatrix} 0.6140 + 0.3684z^{-1} \\ \\ -0.2579z^{-1} - 0.6140z^{-2} + 0.8842z^{-3} \\ +0.4421z^{-4} + 0.2579z^{-6} \end{bmatrix} \quad (21)$$

The driving input $u[n]$ was a real zero-mean, exponentially distributed i.i.d. random sequence with unit variance. The noise vector $\mathbf{w}[n] = (w_1[n], w_2[n])^T$ was assumed to be spatially independent and temporally white Gaussian. The synthetic data $\mathbf{x}[n]$ were processed by the proposed SIMO BCE algorithm with $p = 2$, FFT length $\mathcal{L} = 64$, $L_1 = 0$ and $L_2 = 7$ for the inverse filter $\mathbf{v}[n]$ and the initial condition $\boldsymbol{H}^{(0)}(\omega_k) = 1$ for all k. Thirty independent realizations were performed for $N = 1024$, 2048 and 4096, and SNR = 10 dB, 5 dB, 0 dB and -5 dB, respectively, where SNR is defined as

$$\text{SNR} = \frac{\sum_{i=1}^{P} E[|x_i[n] - w_i[n]|^2]}{\sum_{i=1}^{P} E[|w_i[n]|^2]}. \quad (22)$$

For comparison, $\mathbf{h}[n]$ is also estimated by Tugnait's BCE method [7] as follows:

$$\hat{h}_i[n] = \frac{E\,[x_i[k]e^*[k-n]]}{E\,[|e[k]|^2]} \quad (23)$$

where $e[n]$ is the optimum inverse filter output associated with $J_{2,2}$.

Let $\widehat{\mathbf{h}}^{(l)}[n]$ denote the estimate of $\mathbf{h}[n]$ at the lth realization normalized by a constant energy, and the time delay between $\widehat{\mathbf{h}}^{(l)}[n]$ and the true $\mathbf{h}[n]$ was artificially removed. The normalized mean-square error (NMSE) for the ith channel estimate $\widehat{h}_i[n]$ is defined as

$$\text{NMSE}_i = \frac{1}{30} \cdot \frac{\sum_{l=1}^{30}\left[\sum_{n=0}^{20}(\hat{h}_i^{(l)}[n] - h_i[n])^2\right]}{\sum_n h_i^2[n]}. \quad (24)$$

Then the overall NMSE (ONMSE) [7] can be obtained by averaging NMSE_i over P channels as follows:

$$\text{ONMSE} = \frac{1}{P}\sum_{i=1}^{P}\text{NMSE}_i. \quad (25)$$

Table 1 shows the ONMSEs for different values of data length N and SNR associated with the proposed SIMO BCE algorithm and Tugnait's method, respectively. One can see from Table 1 that the proposed SIMO BCE algorithm performs much better than Tugnait's method (smaller ONMSE).

B. Simulation Results for the Proposed TDE Algorithm

Assume that there were 2 sensor elements ($P = 2$), the amplitude $a_1 = 1$, the true time delay $d_1 = 5$ and the driving input $u[n]$ was a real zero-mean, exponentially distributed i.i.d. random sequence with unit variance. The system $h[n]$ (see (13)) was a non-minimum phase ARMA(3,2) system taken from [1]

$$H(z) = \frac{1 - 2.95z^{-1} + 1.9z^{-2}}{1 - 1.3z^{-1} + 1.05z^{-2} - 0.32z^{-3}} \quad (26)$$

and noise $\widetilde{\mathbf{w}}[n]$ was coherent (i.e., $\widetilde{w}_1[n] = \widetilde{w}_2[n]$) and $\widetilde{w}_1[n]$ was generated as the output of a first-order MA model [1]

$$H_w(z) = 1 + 0.8z^{-1} \quad (27)$$

driven by white Gaussian noise. The synthetic data $\widetilde{\mathbf{x}}[n]$ were processed by the proposed TDE algorithm with $p = 2$, FFT length $\mathcal{L} = 32$, $L_1 = 0$ and $L_2 = 9$ for the inverse filter $\mathbf{v}[n]$ and the initial condition $\boldsymbol{H}^{(0)}(\omega_k) = 1$ for all k. Thirty independent runs were performed for $N = 2048$ and 4096, and SNR= 0 dB and -5 dB. For comparison, d_1 is also estimated by Tugnait's TDE methods [1] as follows:

$$\widehat{d}_1 = \arg\{\max_d\{T_i[d]\}\}, \quad i = 1 \text{ or } 2 \quad (28)$$

where

$$T_1[d] = \frac{|\text{cum}\{\widetilde{x}_1[n-d], \widetilde{x}_1[n-d], \widetilde{x}_2[n], \widetilde{x}_2[n]\}|}{\sqrt{|C_{4,0}\{\widetilde{x}_1[n]\}| \cdot |C_{4,0}\{\widetilde{x}_2[n]\}|}} \quad (29)$$

$$T_2[d] = \frac{|C_{4,0}\{\widetilde{x}_1[n-d] + \widetilde{x}_2[n]\}|}{16\sqrt{|C_{4,0}\{\widetilde{x}_1[n]\}| \cdot |C_{4,0}\{\widetilde{x}_2[n]\}|}}. \quad (30)$$

Table 1. ONMSE associated with the proposed SIMO BCE algorithm and Tugnait's method, respectively.

	Proposed algorithm				Tugnait's method			
	SNR (dB)							
N	10	5	0	-5	10	5	0	-5
1024	0.0358	0.0460	0.1568	0.7055	0.0438	0.0606	0.1846	0.8888
2048	0.0183	0.0228	0.0650	0.5481	0.0210	0.0291	0.0767	0.7979
4096	0.0109	0.0139	0.0354	0.2812	0.0110	0.0163	0.0400	0.4647

Table 2. Mean, standard deviation and RMSE for $\widehat{d}_1$ associated with Tugnait's methods and the proposed TDE algorithm, respectively.

True Time Delay $d_1 = 5$							
		$N = 2048$			$N = 4096$		
SNR (dB)	TDE Method	Mean	σ	RMSE	Mean	σ	RMSE
0	$T_1[d]$	4.8333	1.5992	1.5811	4.8667	0.9732	0.9661
	$T_2[d]$	5.0333	1.6078	1.5811	5.0000	0.0000	0.0000
	Proposed Algorithm	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000
-5	$T_1[d]$	6.5000	6.0272	6.1128	4.8667	5.5238	4.4497
	$T_2[d]$	4.9667	5.8101	5.7126	3.3667	4.2221	4.4609
	Proposed Algorithm	4.1667	1.8952	2.0412	4.6667	1.2685	1.2910

Table 2 shows mean, standard deviation (σ) and root-mean-square error (RMSE) for $\widehat{d}_1$ associated with Tugnait's methods and the proposed TDE algorithm, respectively. One can see from Table 2 that the proposed TDE algorithm performs much better than Tugnait's methods (smaller variance and RMSE).

5. CONCLUSIONS

We have presented an SIMO BCE algorithm using cumulant based MIMO-IFC (see (5)) which is robust against Gaussian noise, and a TDE algorithm that estimates $P-1$ time delays only using the phase information of the single-input P-output ($P \geq 2$) system estimated by the proposed SIMO BCE algorithm. Simulation results show that the proposed SIMO BCE algorithm and TDE algorithm outperform Tugnait's channel estimation method and TDE methods, respectively.

6. REFERENCES

[1] J. K. Tugnait, "Time delay estimation with unknown spatially correlated Gaussian noise," *IEEE Trans. Signal Processing*, vol. SP-41, pp. 549-558, Feb. 1993.

[2] L. Tong and S. Perreau, "Multichannel blind identification: From subspace to maximum likelihood methods," *Proc. IEEE*, vol. 86, no. 10, pp. 1951-1968, Oct. 1998.

[3] C. L. Nikias and A. P. Petropulu, *Higher Order Spectral Analysis: A Nonlinear Signal Processing Framework*, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[4] C.-Y. Chi and C.-H. Chen, "Blind MAI and ISI suppression for DS/CDMA systems using HOS based inverse filter criteria," *IEEE Trans. Signal Processing* (in revision).

[5] C.-Y. Chi and C.-H. Chen, "Cumulant based inverse filter criteria for blind deconvolution: properties, algorithms, and application to DS/CDMA systems," to appear in *IEEE Trans. Signal Processing*, July 2001.

[6] S. M. Kay, *Modern Spectral Estimation*, Prentice-Hall, 1988.

[7] J. K. Tugnait, "Identification and deconvolution of multichannel linear non-Gaussian processes using higher order statistics and inverse filter criteria," *IEEE Trans. Signal Processing*, vol. 45, No. 3, pp. 658-672, Mar. 1997.

Multichannel System Identification Methods for Sensor Array Calibration in Uncertain Multipath Environments

Vijay Varadarajan and Jeffrey L. Krolik,
Department of Electrical and Computer Engineering
Duke University, Box 90291, Durham, NC 27708

Abstract

This paper concerns the problem of estimating the array element relative gain and phase responses using sources of opportunity in an uncertain multipath environment. Unlike previous methods, which assume uncorrelated source wavefronts, we propose to perform calibration in a correlated multipath environment. We present two algorithms that apply traditional blind, single input multiple output (SIMO) methods to the array calibration problem. Calibration is performed by discriminating the received components which remain the same for all sources and are thus due to the receiver gains and phases. Simulation results demonstrate the effectiveness of both techniques in terms of reduced sidelobe levels.

1. INTRODUCTION

Performance of manyarray processing algorithms (cf. e.g. [1], [2]) are seriously limited by the knowledge of the array response. The objective of array calibration can be defined as the accurate characterization or estimation of the array manifold. Previous approaches to array calibration include maximum likelihood techniques [3],[4] and eigenstructure methods [5],[6],[7]. These methods either assume one or more uncorrelated source wavefronts are available which can be used to fit the sensor calibration factors, or assume specific array geometries. In complex correlated multipath, however, it is difficult to accurately model the source wavefronts and this has led to development of so called blind beamforming techniques which exploit alternative signal properties to estimate the array calibration factors. In the techniques proposed here, array calibration is performed by discriminating the received components which remain the same for all sources and are thus due to the receiver gains and phases. Thus to apply SIMO methods, for example, the "single corresponds to the spatial frequency domain element gain and phase response, while the "multiple outputs" correspond to the wavenumber spectra of the different multipath sources of opportunity. The observed data is assumed to consist of angularly separated sources of opportunity at the same frequency which are measured at different times at a fixed sensor array whose element locations are known *a priori*. A technique has been proposed by Leshem et al [9], which also identifies the isomorphism between the SIMO channel identification problem and array calibration. However, their method requires that data be collected along a fine grid in azimuth, which in practice may not be possible.

In this work, we present two computationally efficient non-iterative techniques using only the approximate knowledge of the source location and the range of angles that the multipath wavefronts from a given source can take. The first technique proposed is a least squares technique using a modified form of the well known cross relation (CR) technique used in blind identification of multipath wireless communication channels. In the second technique, known as the "Direct Method", a solution to the reciprocal calibration factors is obtained in a single step without recourse to least squares.

Simulation results from a 44 element ULA indicate that good performance is achieved even at low SNR levels in terms of reduced side-lobe levels and array gain degradation.

2. SIGNAL MODEL

Consider an array of N sensors located in a multipath environment consisting of L angularly separated sources. From each source, there exists several paths to the array. It is assumed that the multipaths due to a given source are correlated. Since the source is at a fixed location, the different multipaths arrive from approximately the same azimuth, but differ in the elevation angle of arrival. It is assumed that the range of elevation angles that the multipaths can take is known *a priori*. Note however, that we do not need to know the values of the elevation angles or the number of multipaths that may exist. Denoting θ_l as the location of the lth source, and $\phi_{l,i}$ as the elevation angle of the ith multipath, the received data from the lth source can be written as :

$$\mathbf{z}_l = \mathbf{G}\left[\mathbf{d}_{\theta_l,\phi_{l,1}} \vdots \ \mathbf{d}_{\theta_l,\phi_{l,2}} \vdots \cdots \ \mathbf{d}_{\theta_l,\phi_{l,P}}\right]\mathbf{s}_l + \mathbf{a}_l \quad (1)$$

where $diag(\mathbf{G}) = [g_1, \ g_2, \cdots \ g_N]^T$ is the $N \times 1$ complex gain vector that represents the gain and

0-7803-7011-2/01/$10.00 ©2001 IEEE

phase response of the array, l_p denotes the total number of multipaths from the lth source, with their individual complex amplitudes and phases contained in $\mathbf{s}_l$, and $\mathring{\mathbf{a}}_l$ is the additive noise vector with covariance $\sigma^2\mathbf{I}$. Temporarily ignoring the presence of noise, the above equation can be written as

$$z_l(n) = v_l(n) \odot g(n) \quad (2)$$

for which

$$\mathbf{v}_l = \mathbf{G}\left[\mathbf{d}_{\theta_l,\phi_{l,1}} \vdots \ \mathbf{d}_{\theta_l,\phi_{l,2}} \vdots \cdots \ \mathbf{d}_{\theta_l,\phi_{l,P}}\right]\mathbf{s}_l, \quad (3)$$

where n denotes the sensor index, $\odot$ denotes point-wise multiplication, and $\mathbf{v}_l$ represents the $N\times 1$ replica vector from the lth source. Due to propagation through the ionosphere, the multipaths arrive only in a small subset of angles in $[0 \ \ \pi]$. Therefore it is possible to represent the multipaths from a single source with fewer parameters than the number of sensors N. Therefore we have

$$\mathbf{v}_l = \boldsymbol{\varnothing}_l \hat{\mathbf{a}}_l, \quad (4)$$

where $\boldsymbol{\varnothing}_l$ is an $N\times K$ matrix $(K \ll N)$ whose columns correspond to the dominant eigenvectors of the correlation matrix $\mathbf{R}_{v_l} = E_\phi\left[\mathbf{v}_l\mathbf{v}_l^\dagger\right]$, where the averaging is done over the range of elevation angles of the multipaths from the lth source.

III. CROSS RELATION TECHNIQUE

We now incorporate the well known cross relation technique [8] to estimate the calibration factors. The multiplication of the sequences in equation (2) can be replaced by circular convolution in the frequency domain as

$$z_l(u) = g(u) \otimes v_l(u) \ , \quad (5)$$

where $g(u)$ denotes the DFT of the sequence of sensor gains, $v_l(u)$ represents the DFT of the received replica vector, and $\otimes$ denotes circular convolution. The above equation is analogous to the convolution of the source signal with the lth channel in the blind SIMO multichannel identification problem. However, it is important to note that in the SIMO identification problem, the source is "linearly" convolved with the channel. A cross relation between the lth and the source can be expressed as

$$\tilde{\mathbf{Z}}\tilde{\mathbf{v}} = \mathbf{0} \quad (6)$$

where $\tilde{\mathbf{v}}^\dagger = \left[\tilde{\mathbf{v}}_l^\dagger \ \vdots \ \tilde{\mathbf{v}}_{l'}^\dagger\right]^\dagger$ and $\left[\tilde{\mathbf{v}}_l\right]_u = v_l(u)$. The matrix $\tilde{\mathbf{Z}}$ is given by $\tilde{\mathbf{Z}} = \left[\tilde{\mathbf{Z}}_l \ \vdots \ \tilde{\mathbf{Z}}_{l'}\right]$, where $\tilde{\mathbf{Z}}_l$ is an $N\times N$ circulant matrix formed by $\tilde{\mathbf{z}}_l$, which represents the received data in the spatial frequency domain.

Using the reduced parameter expression for the replica vector, we have

$$\tilde{\mathbf{v}} = \tilde{\boldsymbol{\varnothing}}\,\tilde{\hat{\mathbf{a}}} \quad (7)$$

where the $2N\times 2K$ matrix ˜ can be written as

$$\tilde{\boldsymbol{\varnothing}} = \begin{bmatrix} \tilde{\boldsymbol{\varnothing}}_l & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{\varnothing}}_{l'} \end{bmatrix}. \quad (8)$$

The columns of $\tilde{\boldsymbol{\varnothing}}_l$ are the DFT's of the columns of $\boldsymbol{\varnothing}_l$. In order to estimate the vector $\mathbf{v}$ in the presence of noise, equation (X) can be reformulated by the minimization $\left\|\tilde{\mathbf{Z}}\tilde{\mathbf{v}}\right\|^2$ as

$$\tilde{\hat{\mathbf{a}}} = \arg\min_{\tilde{\mathbf{a}}} \tilde{\mathbf{a}} \left[\tilde{\boldsymbol{\varnothing}} \ \tilde{\mathbf{Z}} \ \tilde{\mathbf{Z}}\tilde{\boldsymbol{\varnothing}}\right]\tilde{\mathbf{a}} \quad (9)$$

subject to the constraint $\tilde{\mathbf{a}}^\dagger\tilde{\mathbf{a}} = 1$. The above minimization can be achieved by calculating the eigen vector with minimum eigen value of $\left[\tilde{\boldsymbol{\varnothing}} \ \tilde{\mathbf{Z}} \ \tilde{\mathbf{Z}}\tilde{\boldsymbol{\varnothing}}\right]$. Given $\tilde{\hat{\mathbf{a}}}$, we can estimate $\tilde{\mathbf{v}}$ using equation (X) as $\tilde{\hat{\mathbf{v}}} = \tilde{\boldsymbol{\varnothing}}\,\tilde{\hat{\mathbf{a}}}$. Consequently, a least squares estimate of the vector of the DFT of the complex sensor gains $\mathbf{g}$ can be calculated as

$$\hat{\tilde{\mathbf{g}}} = \left(\hat{\tilde{\mathbf{V}}}^\dagger\hat{\tilde{\mathbf{V}}}\right)^{-1}\hat{\tilde{\mathbf{V}}}^\dagger\tilde{\mathbf{z}} \quad (10)$$

where $\tilde{\mathbf{z}} = \left[\tilde{\mathbf{z}}_l^\dagger \ \vdots \ \tilde{\mathbf{z}}_{l'}^\dagger\right]^\dagger$, $\hat{\tilde{\mathbf{V}}} = \left[\hat{\tilde{\mathbf{V}}}_l \ \vdots \ \hat{\tilde{\mathbf{V}}}_{l'}\right]^\dagger$ and $\hat{\tilde{\mathbf{V}}}$ is the circulant form of $\tilde{\mathbf{v}}$.

IV DIRECT METHOD

In this method, the gains and phases are estimated in a single step without the use of the least squares technique used above. Using Equations (1) and (3) we have,

$$\mathbf{G}^{-1}\mathbf{z}_l = \mathbf{v}_l, \quad (11)$$

where $\mathbf{G}$ is assumed to be invertible. Since $\mathbf{G}$ is diagonal, the above condition which is equivalent to all the complex gains having their magnitude greater than 0, is satisfied in practice. Further, since $\mathbf{G}^{-1}$ is diagonal, the above equation can be expressed as

$$\mathbf{Z}_l \breve{\mathbf{g}} = \mathbf{v}_l \quad (12)$$

where $\mathbf{g} = diag\left(\mathbf{G}^{-1}\right)$ and $\mathbf{Z}_l = diag(\mathbf{z}_l)$. Using equations (11) and (12) we obtain

$$\mathbf{Zc} = \mathbf{0} \quad (13)$$

where

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_l & \mathbf{ø}_l & \mathbf{0} \\ \mathbf{Z}_{l'} & \mathbf{0} & \mathbf{ø}_{l'} \end{bmatrix} \quad (14)$$

while $\mathbf{c} = \left[\breve{\mathbf{g}} \;\vdots\; \hat{\mathbf{a}} \;\vdots\; \hat{\mathbf{a}}\right], \mathbf{Z}_l = diag(\mathbf{z}_l)$, $\left[\breve{\mathbf{g}}\right]_n = \breve{g}(n)$. Clearly, the complex gains can be estimated by estimating the null space of $\mathbf{Z}$. In presence of noise, we can estimate $\mathbf{c}$ by the minimization of $\|\mathbf{Zc}\|^2$ as

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \mathbf{c}^{\dagger}\mathbf{Z}^{\dagger}\mathbf{Zc} \quad (15)$$

subject to the constraint $\mathbf{c}^{\dagger}\mathbf{c} = 1$. This is achieved by calculating the minimum eigenvector of the matrix $\mathbf{Z}^{\dagger}\mathbf{Z}$.

Since the gains and phases are estimated in a single step, this method was found to be much faster than the cross relation based method from simulations performed.

V SIMULATION RESULTS

Simulations were performed using a 44 element uniformly spaced linear array (ULA) with inter-element spacing of $\lambda/2$. The frequency of the sources was chosen to be 10MHz. In all simulations, 4 multipaths were assumed to exist from each source to the sensor array. The complex sensor gains were chosen to be i.i.d complex gaussian random numbers. Since the gains and phases are estimated only up to a complex multiplicative constant, the performance metric that we use is the array gain degradation (AGD) which is defined as $AGD = -E\{20\log\cos(\theta)\}$, where

$$\cos(\theta) = \frac{\left\|\mathbf{g}^{\dagger}\hat{\mathbf{g}}\right\|}{\sqrt{\left\|\mathbf{g}^{\dagger}\mathbf{g}\right\| \left\|\hat{\mathbf{g}}^{\dagger}\hat{\mathbf{g}}\right\|}} \quad (16)$$

is the generalized cosine of the angle between the true gain $\mathbf{g}$ and the estimated gain $\mathbf{g}$. Under the ideal case wherein the estimated gain vector is a complex scalar multiple of the true gain, the AGD takes on the value 0. The calibration factors were estimated using the Direct Method and the beampatterns obtained using these coefficients was compared with the case when the sensor gains are known exactly and with an uncalibrated array. The beampattern for the uncalibrated, calibrated and the gain known exactly cases is shown in Figure 1.The SNR for this simulation was set at 20db. Figure 2 shows the results obtained with the SNR set at 40db. As can be seen, there is a significant improvement in the sidelobe levels with an increase in SNR, which is intuitively satisfying. The AGD was computed using both methods as a function of SNR, with 50 trials being conducted at each SNR value. Figures 3 and 4 show the variation of AGD with SNR for the cross relation method and the "Direct" method respectively, for different source locations. As can be seen, both methods perform slightly better when the sources are close to broadside. This is due to the fact that fewer parameters are required to characterize a source closer to broadside and therefore the replica vector can be more accurately represented.

VI CONCLUSIONS

In this paper we present two techniques methods for array gain and phase calibration using multipath sources of opportunity. Multichannel blind system identification techniques are applied to yield computationally efficient solutions. Simulation results demonstrate the potential of the proposed algorithms.

ACKNOWLEDGEMENT

This work was supported by ONR under grant number N00014-99-1-0532.

REFERENCES

[1] R. O. Schmidt, "Multiple emitter location and signal parameter estimation", IEEE Trans. Antennas. Propagat., vol 34, pp.276-280, March 1986.

[2] R. Roy and T. Kailath, "Esprit-estimation of signal parameters using rotational invariance techniques ",IEEE Trans. Acoust,. Speech, Signal Processing, vol. 37, pp.984-995, July 1989.

[3] B.C. Ng and C.M.S. See, "Sensor-array calibration using a maximum-likelihood approach," IEEE Trans. Antennas Propagat., vol. 44, pp. 827-835, June 1996.

[4] D. Fuhrmann, "Estimation of sensor gain and phase," IEEE Trans. Signal Processing, vol. 42, pp. 77-87, January 1994.

[5] B. Friedlander and A.J. Weiss, "Eigenstructure methods for direction finding and sensor gain and phase uncertainties," Proceedings IEEE Int. Conf. On Acoust Speech and Signal Processing, pp. 2681-2684, May 1988.

[6] G.C Brown, J.H. McClellan and E.J. Holder, "Eigenstructure approach for array processing and calibration with general gain

and perturbations," Proc. IEEE Int. Conf. On Acoust. Speech and Signal Processing pp. 1365 1368, 1991.

sensor gain and phase uncertainties," IEEE Trans. Antennas Propagat., vol 43, pp. 880-883, August 1995.

[8] G.Xu, H.Liu, L.Tong and T.Kailath, "A least- to blind channel identification," IEEE Trans. Signal Processubg, vol 43, pp. 2982-

[9] A. Leshem and M.Wax, "Array calibration in presence of multipath," IEEE Trans. Signal Processing, vol. 48, pp. 53- January 2000.

Figure : Beam pattern for Direct Method 20 dB SNR

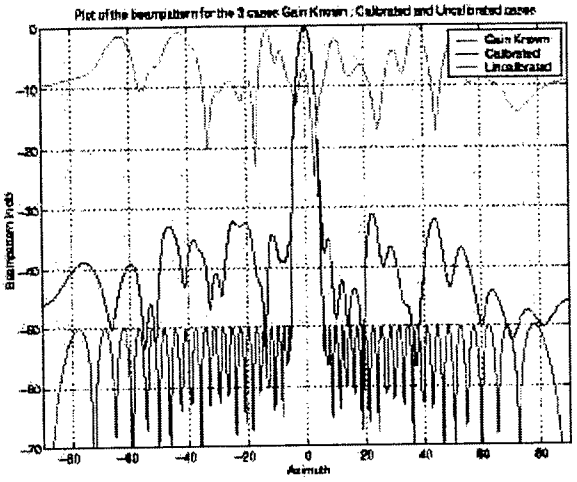
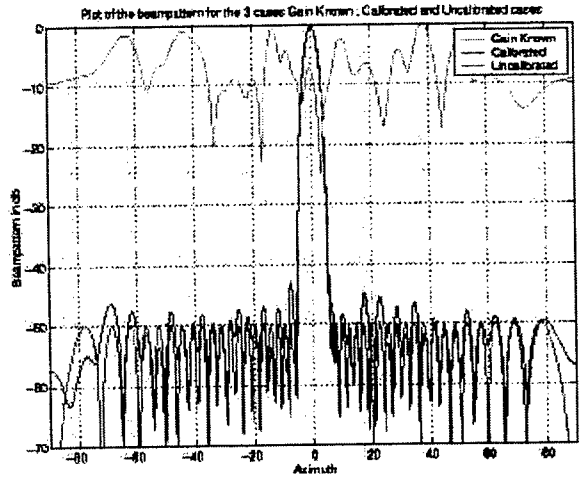


Figure 2



3: Array Gain Degradation

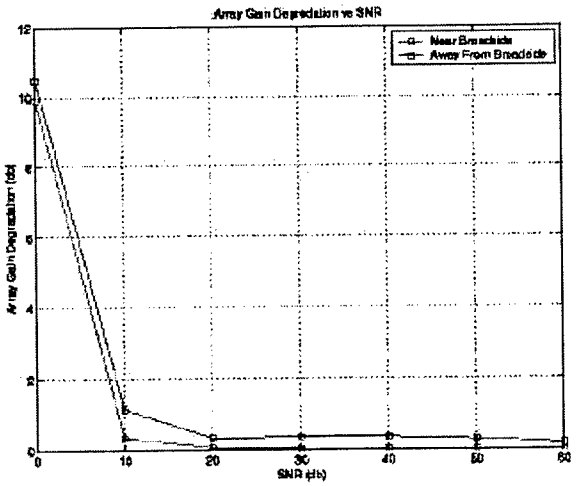
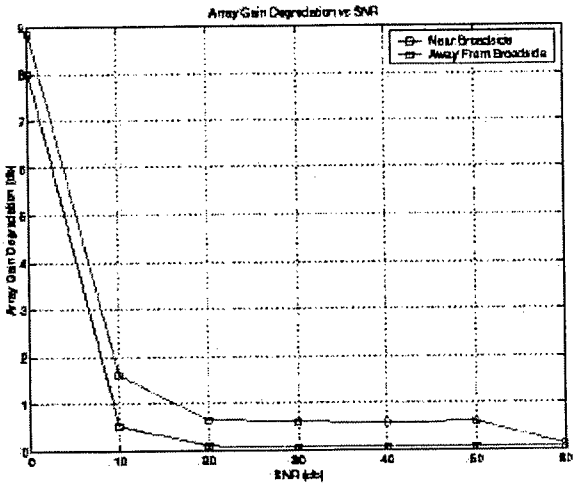


Figure 4: Array Gain Degradation vs. SNR for Direct Calibration Method



A COOPERATIVE MAXIMUM LIKELIHOOD MIMO CHANNEL ESTIMATOR

Marc Chenu-Tournier[1] *and Pascal Larzabal*[2]

[1]Thales-Communication, TTC/TSI/LTA, 66 Rue du fossé Blanc, 92231 Gennevilliers, FR
[2]LESiR/CNRS, 61 Av du Président Wilson, 94235 Cachan, FR
[1]marc.chenu@fr.thalesgroup.com, [2]larzabal@lesir.ens-cachan.fr

ABSTRACT

In future wireless telecommunication systems, the needs in terms of capacity will lead the operators to make simultaneous communications in the same time at the same frequency. When the system benefits from diversity, for example channel diversity (spatial separation of emitters) or code diversity (CDMA, space-time coding), the transmission may be assured with reasonable quality. Classical techniques of channel estimation for the single emission case do not offer sufficient performances for the multi-emission case.
We propose a cooperative maximum likelihood (ML) channel estimator adapted to the emitted signals in the multi-emitter (multi-user) context. This estimator relies on the hypothesis that the propagation channel is specular. The Cramer Rao bound (CRB) is derived and compared to the performances of the proposed ML estimator. The empirical performances from Monte-Carlo simulations show that this estimator is efficient at high SNR.

1. INTRODUCTION

The needs in terms of wireless communication capacity are increasing dramatically with video and music demand. On the other hand, the released frequencies are still limited. As it has recently been proved, the capacity of a MIMO (multiple input multiple output) system in increased compared to a classical single input, single output system [6]. To meet these capacities, accurate MIMO channel estimation techniques should be used for the demodulator to perform well. We propose a cooperative maximum likelihood (ML) channel estimator taking into account the knowledge of the emitted signals in the multi-emitter (multi-user) context. This estimator relies on the hypothesis that the propagation channel is specular. The Cramer Rao bound (CRB) is derived and compared to the performances of the proposed ML estimator. The empirical performances from Monte-Carlo simulations show that this estimator is efficient at high SNR. The proposed estimator uses multiple sensors at the reception. We suppose that the antenna array is not calibrated. If the antenna is calibrated, the directions of arrival may be estimated and the modified estimator may be found in [4]. A similar work in the radar context for the single emission case may be found in [2].

For CDMA systems, a typical simultaneous resource sharing system, the studies performed on the demodulators is much larger than the works that have been done concerning the channel estimation ([10], [7] and [1]).

The single user techniques [11] are applicable in the UMTS (CDMA) systems when the number of users is small but rapidly degrade when the number of users increase.
These single user techniques have the same weakness that the RAKE receiver has when multiple users transmit in the same cell.

The work proposed in [8] estimates the propagation delays using a bank of non coherent detectors. The results are proposed in multi-path propagation channels, with Doppler and inter-cell interference. In [9] the authors propose an optimal decision rule based on the output of these correlators. These techniques are single user based and thus are limited in the multi-user case.

Many proposed estimators based on the ML ([10], [7] and [1]) use rectangular pulse shape filters and are thus unfortunately inapplicable when filters longer than a chip are used as in real systems.

In the following, we first present the signal model for a multi-emission system with specular propagation channels. Then in section 3 the proposed estimator is developed and the theoretical bounds are derived. At last, in section 4, we will propose some simulated results and compare them to the theoretical bounds. These performances will be followed by the conclusion and some perspectives.

2. SIGNAL MODEL

We suppose that several emitters (users) are transmitting simultaneously at the same frequency. Each emitter u is supposed to transmit a known signal (pilot signal) $s^u(t)$ as it is done in the UMTS FDD norm. This pilot signal is generally used to estimate the propagation channel so that the demod-

0-7803-7011-2/01/$10.00 ©2001 IEEE

ulator may estimate the transmitted symbols. To simplify the presentation, we will consider that the received signal is only composed of the pilot signal, the data signals being considered as additive Gaussian noise. As supposed previously, the propagation channel is considered to be specular, meaning that it may be written as :

$$\mathbf{h}^u(t) = \sum_{p=1}^{P^u} \mathbf{h}_p^u \delta(t - \tau_p^u)$$

where P^u is the number of paths of the propagation channel of user u, $\mathbf{h}_p^u$ and τ_p^u are respectively the vector of the responses of the antenna and the associated delay to the path p of user u. $\delta(t)$ is the Dirac function.

The received signal $\mathbf{x}(t)$ is proposed in equation (1).

$$\mathbf{x}(t) = \sum_{u=1}^{U} \sum_{p=1}^{P^u} \mathbf{h}_p^u s^u(t - \tau_p^u) + \mathbf{b}(t) \qquad (1)$$

where : U is the number of users, $s^u(t)$ is the known signal of user u and $\mathbf{b}(t)$ is the noise vector at time t.

This signal is sampled every T_e on a period $t = [T_e, N_e T_e]$ during which the $\mathbf{h}_p^u$ are considered to be constant. These N_e samples of dimension $N \times 1$ are concatenated in a vector $\mathbf{Y}$ of dimension $N_e N \times 1$ verifying :

$$\mathbf{Y} = \left[\underbrace{x_1(T_e), \cdots, x_1(N_e T_e)}_{\text{sensor 1}}, \cdots, \underbrace{x_N(T_e), \cdots, x_N(N_e T_e)}_{\text{sensor N}} \right]^T$$

where $x_i(nT_e)$ is the sample n of the sensor i.

The vector $\mathbf{N}$ of dimension $N_e N \times 1$ contains the concatenation of the noise samples :

$$\mathbf{N} = \left[\underbrace{b_1(T_e), \cdots, b_1(N_e T_e)}_{\text{sensor 1}}, \cdots, \underbrace{b_N(T_e), \cdots, b_N(N_e T_e)}_{\text{sensor N}} \right]^T$$

where $b_i(nT_e)$ is the noise sample n on sensor i.
The signal vector $\mathbf{Y}$ may be written as :

$$\mathbf{Y} = \mathbf{\Psi}(\boldsymbol{\tau})\boldsymbol{\alpha} + \mathbf{N} \qquad (2)$$

where the matrix $(N_e N \times NP)$ $\mathbf{\Psi}$ contains the samples of the $\mathbf{s}^u(\tau_p^u)$ as follows :

$$\mathbf{\Psi}(\boldsymbol{\tau}) = \mathbf{I}_N \otimes \boldsymbol{\mathcal{S}}(\boldsymbol{\tau})$$

where $\boldsymbol{\mathcal{S}}(\boldsymbol{\tau})$ is the $(N_e \times P)$ matrix verifying :

$$\boldsymbol{\mathcal{S}}(\boldsymbol{\tau}) = \left[\mathbf{s}^1(\tau_1^1), \cdots, \mathbf{s}^1(\tau_{P_1}^1), \cdots, \mathbf{s}(\tau_{P^U}^U)\right]$$

with

$$\mathbf{s}^u(\tau_p^u) = \left[\; s^u(T_e - \tau_p^u) \quad \cdots \quad s^u(N_e T_e - \tau_p^u) \; \right]^T$$

$\boldsymbol{\alpha}$ contains the responses of the sensors to the paths of the users :

$$\begin{aligned} \boldsymbol{\alpha} &= \left[\underbrace{h_{1,1}^1, \cdots, h_{P^U,1}^U}_{\text{sensor 1}}, \cdots, h_{P^U,N}^U \right]^T \\ &= \left[\mathbf{h}_1^T, \cdots, \mathbf{h}_N^T\right]^T \end{aligned}$$

The modelisation of the received signals in equation (2) is linear in the nuisance parameters : $\boldsymbol{\alpha}$ and the noise $\mathbf{N}$. As the noise is considered as Gaussian, temporally and spatially white, the log-likelihood of the received signal can be easily deduced.

3. ML ESTIMATOR AND CRAMER RAO BOUND

3.1. ML Estimator

We consider here that the complex amplitudes $\boldsymbol{\alpha}$ are unknown but deterministic. The signals used in $\mathbf{\Psi}(\boldsymbol{\tau})$ are supposed to be known but parametrized by the variables $\boldsymbol{\tau}$ to be estimated. This leads to a model where only the noise $\mathbf{N}$ is random, with Gaussian components and thus the log-likelihood is given by :

$$L(\mathbf{Y}|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\tau}) = -N_e N \log(\pi\sigma^2) - \frac{1}{\sigma^2} \left\| \mathbf{Y} - \mathbf{\Psi}(\boldsymbol{\tau}\boldsymbol{\alpha}) \right\|^2 \qquad (3)$$

The following of this chapter is dedicated to the estimation of the parameters σ^2, $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$. Recall that the antenna is not calibrated and thus the DOAs are not estimated.

We may determine the analytical expressions of estimators for the complex amplitudes $\boldsymbol{\alpha}$ and for the power of the noise σ^2, parametrized by the vector of delays $\boldsymbol{\tau}$. These estimators are obtained by nulling the derives of the log-likelihood in $\boldsymbol{\alpha}$ and in σ^2. We get :

$$\widehat{\sigma}^2 = \frac{1}{N_e N} \left\| \mathbf{Y} - \mathbf{\Psi}(\boldsymbol{\tau})\boldsymbol{\alpha} \right\|^2$$

and

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{\Psi}^\dagger(\boldsymbol{\tau})\mathbf{\Psi}(\boldsymbol{\tau}) \right)^{-1} \mathbf{\Psi}^\dagger(\boldsymbol{\tau})\mathbf{Y} = \mathbf{\Psi}^\#(\boldsymbol{\tau})\mathbf{Y} \qquad (4)$$

By replacing $\boldsymbol{\alpha}$ and σ^2 by their estimators the log-likelihood simplifies and the estimator of $\boldsymbol{\tau}$ is given by :

$$\widehat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} \left\| \mathbf{Y} - \mathbf{\Psi}(\boldsymbol{\tau}) \left(\mathbf{\Psi}^\dagger(\boldsymbol{\tau})\mathbf{\Psi}(\boldsymbol{\tau}) \right)^{-1} \mathbf{\Psi}^\dagger(\boldsymbol{\tau})\mathbf{Y} \right\|^2$$

Let $\mathbf{\Pi}_{\mathbf{\Psi}}^{\perp}(\boldsymbol{\tau})$ be the projector on the image built by the rows of $\mathbf{\Psi}(\boldsymbol{\tau})$:

$$\mathbf{\Pi}_{\mathbf{\Psi}}^{\perp}(\boldsymbol{\tau}) = \mathbf{I} - \mathbf{\Psi}(\boldsymbol{\tau}) \left(\mathbf{\Psi}^\dagger(\boldsymbol{\tau})\mathbf{\Psi}(\boldsymbol{\tau}) \right)^{-1} \mathbf{\Psi}^\dagger(\boldsymbol{\tau})$$

The estimate of τ is given by :

$$\begin{aligned}\widehat{\tau} &= \arg\min_{\tau} \left\| \mathbf{\Pi}_{\Psi}^{\perp}(\tau)\mathbf{Y} \right\|^2 \\ &= \arg\min_{\tau} tr\left(\mathbf{Y}^{\dagger}\mathbf{\Pi}_{\Psi}^{\perp}(\tau)\mathbf{Y} \right)\end{aligned}$$

This estimator will be compared to the theoretical Cramer Rao bound. To implement the ML estimator, a Gauss newton algorithm has been used (for more details see [4]).

3.2. Cramer Rao Bound

In this section we will determine the statistical performances of the ML estimator in terms of estimation variance. The Cramer-Rao bounds are calculated and give us the minimum reachable variance for an un-biased estimator at high SNR. Once these bounds are given, they are compared to the Monte-Carlo simulations giving empirical estimations of the variance of the estimator.

The Cramer-Rao bounds (CRB) give the inferior variance limit for an un-biased estimator. These limits are given by :

$$var(\hat{\mathbf{q}}) \geq \mathbf{CRB} = \mathbf{J}_{\mathbf{q}}^{-1}$$

where $\mathbf{J_q}$ is the Fisher information matrix :

$$\mathbf{J_q} = E\left\{ \frac{\partial L(\mathbf{q})}{\partial \mathbf{q}} \cdot \left(\frac{\partial L(\mathbf{q})}{\partial \mathbf{q}} \right)^{\dagger} \right\}, \tag{5}$$

and where $\mathbf{q}$ is the complex vector of the wanted parameters :

$$\mathbf{q} = \begin{bmatrix} \sigma^2 \\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^* \\ \boldsymbol{\tau} \end{bmatrix}$$

$\boldsymbol{\alpha}, \boldsymbol{\alpha}^*$ and σ^2 are nuisance parameters and $\boldsymbol{\tau}$ are the useful (delay) parameters.

The choice of the complex notation in the Fisher matrix is motivated by the simplifications it introduces in the calculation of the inverse of the bloc $\left[\mathbf{J}_{\mathbf{q}}^{-1}\right]_{\boldsymbol{\tau}}$ on the delay parameters. Such an approach is much simpler than an approach isolating the real and imaginary part as is done in [3].

The matrix $\mathbf{J_q}$ has the following bloc structure :

$$\mathbf{J_q} = \begin{bmatrix} J_{\sigma^2} & \mathbf{J}_{\sigma^2\boldsymbol{\alpha}^T} & \mathbf{J}_{\sigma^2\boldsymbol{\alpha}^T} & \mathbf{J}_{\sigma^2\boldsymbol{\tau}^T} \\ [\mathbf{J}_{\sigma^2\boldsymbol{\alpha}^T}]^{\dagger} & \mathbf{J}_{\boldsymbol{\alpha}} & \mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}^T} & \mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\tau}^T} \\ [\mathbf{J}_{\sigma^2\boldsymbol{\alpha}^T}]^{\dagger} & [\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}^T}]^{\dagger} & \mathbf{J}_{\boldsymbol{\alpha}^*} & \mathbf{J}_{\boldsymbol{\alpha}^*\boldsymbol{\tau}^T} \\ [\mathbf{J}_{\sigma^2\boldsymbol{\tau}^T}]^{\dagger} & [\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\tau}^T}]^{\dagger} & [\mathbf{J}_{\boldsymbol{\alpha}^*\boldsymbol{\tau}^T}]^{\dagger} & \mathbf{J}_{\boldsymbol{\tau}} \end{bmatrix} \tag{6}$$

where the non null blocs are given by :

$$J_{\sigma^2} = \frac{N_e \cdot N}{\sigma^4}. \tag{7}$$

$$\mathbf{J}_{\boldsymbol{\alpha}} = \frac{1}{\sigma^2} \cdot \left(\mathbf{\Psi}^{\dagger}(\mathbf{p}) \cdot \mathbf{\Psi}(\mathbf{p}) \right)^*, \tag{8}$$

$$\begin{cases} [\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\tau}^T}]_{i,j} = \frac{1}{\sigma^2} \cdot \left(\boldsymbol{\alpha}^{\dagger} \cdot \frac{\partial \mathbf{\Psi}^{\dagger}(\mathbf{p})}{\partial \tau_j} \cdot \mathbf{\Psi}(\mathbf{p}) \cdot \mathbf{e}_i \right) \\ \mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\tau}^T} = \frac{1}{\sigma^2} \cdot \left(\mathbf{\Psi}^{\dagger} \cdot \mathbf{D}_{\mathbf{\Psi}} \cdot \mathbf{H_d} \right)^* \end{cases}, \tag{9}$$

with

$$\mathbf{e}_i = \begin{bmatrix} \underset{1}{0} & \cdots & \underset{i-1}{0} & \underset{i}{1} & \underset{i+1}{0} & \cdots & \underset{L}{0} \end{bmatrix}^T$$

and where

$$\mathbf{H_d} = [diag(\mathbf{h}_1), \cdots, diag(\mathbf{h}_N)]^T$$

$$\begin{aligned}\mathbf{D}_{\mathbf{\Psi}}(\boldsymbol{\tau}) &= (\mathbf{I}_N \otimes \mathbf{D}_{\boldsymbol{\mathcal{S}}}) \\ \mathbf{D}_{\boldsymbol{\mathcal{S}}}(\boldsymbol{\tau}) &= \left[\frac{\partial \mathbf{s}_1^1(\tau)}{\partial \tau}, \cdots, \frac{\partial \mathbf{s}_{PU}^U(\tau)}{\partial \tau} \right]\end{aligned}$$

For the terms relative to the delay parameters :

$$\mathbf{J}_{\boldsymbol{\tau}} = \frac{2}{\sigma^2} \cdot \Re\left(\mathbf{H}_{\mathbf{d}}^{\dagger} \cdot \mathbf{D}_{\mathbf{\Psi}}^{\dagger} \cdot \mathbf{D}_{\mathbf{\Psi}} \cdot \mathbf{H_d} \right) \tag{10}$$

Thus the Fisher information matrix becomes :

$$\mathbf{J_q} = \begin{bmatrix} J_{\sigma^2} & \mathbf{0}_{1\times N} & \mathbf{0}_{1\times N} & \mathbf{0}_{1\times N} \\ \mathbf{0}_{N\times 1} & \mathbf{J}_{\boldsymbol{\alpha}} & \mathbf{0}_{N\times N} & \mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\tau}^T} \\ \mathbf{0}_{N\times 1} & \mathbf{0}_{N\times N} & \mathbf{J}_{\boldsymbol{\alpha}^*} & \mathbf{J}_{\boldsymbol{\alpha}^*\boldsymbol{\tau}^T} \\ \mathbf{0}_{N\times 1} & [\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\tau}^T}]^{\dagger} & [\mathbf{J}_{\boldsymbol{\alpha}^*\boldsymbol{\tau}^T}]^{\dagger} & \mathbf{J}_{\boldsymbol{\tau}} \end{bmatrix}$$

Using the equalities $\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\tau}^T} = (\mathbf{J}_{\boldsymbol{\alpha}^*\boldsymbol{\tau}^T})^*$ and $\mathbf{J}_{\boldsymbol{\alpha}} = (\mathbf{J}_{\boldsymbol{\alpha}^*})^*$, and the bloc inversion lemma, we get :

$$\mathbf{BCR}(\boldsymbol{\tau}) = \left[\mathbf{J}_{\mathbf{q}}^{-1}\right]_{\boldsymbol{\tau}} = \left[\mathbf{J}_{\boldsymbol{\tau}} - 2\Re\left\{ \mathbf{J}_{\boldsymbol{\alpha}^*\boldsymbol{\tau}^T}^{\dagger} \mathbf{J}_{\boldsymbol{\alpha}^*}^{-1} \mathbf{J}_{\boldsymbol{\alpha}^*\boldsymbol{\tau}^T} \right\} \right]^{-1}$$

that simplifies by :

$$\mathbf{BCR}(\boldsymbol{\tau}) = \frac{\sigma^2}{2} \cdot \left[\Re\left\{ \mathbf{H}_{\mathbf{d}}^{\dagger} \cdot \mathbf{D}_{\mathbf{\Psi}}^{\dagger} \cdot \mathbf{\Pi}_{\mathbf{\Psi}}^{\perp} \cdot \mathbf{D}_{\mathbf{\Psi}} \cdot \mathbf{H_d} \right\} \right]^{-1}$$

This general expression of the CRB relative to the parameters $\boldsymbol{\tau}$ gives in the single path scenario the classical result. In deed, in this case, the matrix $\mathbf{J_q}$ becomes diagonal and $\left[\mathbf{J}_{\mathbf{q}}^{-1}\right]_{\boldsymbol{\tau}} = \mathbf{J}_{\boldsymbol{\tau}}^{-1}$.

4. PERFORMANCES AND FURTHER DEVELOPMENTS

We compare on figure 1 the performances of the ML to the CRB. The signals used for the simulations are UMTS-FDD like wave forms (CDMA) where only the pilot channel is generated and known for all the users. This means that the received signals do not contain any data signals. The

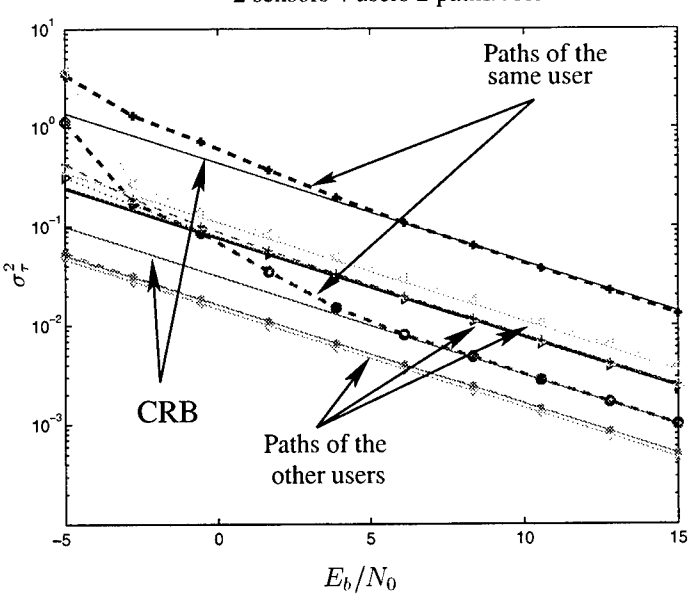


Figure 1: CRB and ML for the delays

spreading factor is set to 32 and the observations are of 320 samples long with a 2 sensor reception antenna in a cell containing 4 users. Each users signal propagates through a 2 path channel. The optimization of the ML has been done with a Gauss-Newton algorithm. Note that the simulation is made in a typical reduced UMTS-FDD scenario. The simulations are indexed in abscissa in E_b/N_0 representing the energy bit over the noise power.

We notice, on the figure, that when the performances (variance of the estimation on the delay parameter) of the estimation of one of the paths for one of the users is degraded, the performances of the other path is also degraded. On the other hand, the performances of the other users are not affected by the performance loss of the estimated delay of the other degraded user.

For high SNR scenario, the performances of the ML and the CRB are the same and thus, the ML estimator is efficient. A more detailed demonstration showing that the ML variance is equal to the CRB is proposed in [4] and in [5] for high SNR.

At low SNRs, the CRB and the ML algorithm do not match. The SNR for which the ML and the CRB separate is called the SNR threshold. In future work, the SNR threshold will be studied and tighter limits will be obtained using a Bayesian bound like the Ziv-Zakaï bound. These bounds offer more accurate information on the lower bounds for channel estimation schemes at low SNR.

5. REFERENCES

[1] S.E. Bensley and B. Aazhang. Maximum likelihood estimation of a single user's delay for code division multiple acess communication systems. Conf. Information Sciences ans Systems, 1994.

[2] N. Bertaux. *Contribution à l'utilisation des méthodes du Maximum de Vraisemblance en traitement radar actif.* PhD thesis, Ecole Normale Supérieure de Cachan, Janvier 2000.

[3] A.Van Bos. A cramer-rao lower bound for complex parameters. *IEEE Transaction on Signal Processing*, Vol: 42:pp: 2859, October 1994.

[4] M. Chenu-Tournier. *Contribution à l'utilisation des techniques de traitement d'antenne dans un système de radio-communication numérique : Application à l'UMTS et au GSM.* PhD thesis, ENS-Cachan, LESiR, France, 2000.

[5] M. Chenu-Tournier and P. Larzabal. Space-time channel estimation for multi-user communications. *to be submitted to IEEE Signal processing.*

[6] G.J. Foschini and M.J. Gans. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications, Kluwer Academic Publishers*, Vol: 6(No: 3):pp: 311–335, March 1998.

[7] S. Parkvall. *Near-Far Resistant DS-CDMA Systems : Parameter estimation and Data Detection.* PhD thesis, Royal Institute of Technology Stockholm, Sweden, 1996.

[8] R. Rick and L. Milsteil. Performance acquisition in mobile ds-cdma systems. *IEEE Trans on Communications*, Vol: 45(No: 11):pp: 1466–1476, November 1997.

[9] R. Rick and L. Milsteil. Optimal decision strategies for acquisition of spread spectrum signals in frequency selective fading channels. *IEEE Trans. on Communications*, Vol: 46(No: 5):pp: 686–694, May 1998.

[10] E. Strom, S. Parkvall, S. Miller, and B. Ottersen. Propagation delay estimation in asynchronous direct-sequence code-division multiple access systems. *IEEE Trans on Communications*, Vol: 44:pp: 84–93, January 1996.

[11] J.K. Tugnait, L. Tong, and Z. Ding. Single-user chanel estimation and equalization. *IEEE Signal Processing Magazine*, Vol: 17(No: 3):pp: 17–28, May 2000.

NONSTATIONARY SIGNAL CLASSIFICATION USING SUPPORT VECTOR MACHINES

Arthur Gretton[1], *Manuel Davy*[1], *Arnaud Doucet*[2], *Peter J. W. Rayner*[1]

[1] Signal Processing Group, University of Cambridge
Department of Engineering, Trumpington Street
CB2 1PZ, Cambridge, UK
{alg30,md283,pjwr}@eng.cam.ac.uk
[2] Department of Electrical and Electronic Engineering
The University of Melbourne
Victoria 3010, Australia
a.doucet@ee.mu.oz.au

ABSTRACT

In this paper, we demonstrate the use of support vector (SV) techniques for the binary classification of nonstationary sinusoidal signals with quadratic phase. We briefly describe the theory underpinning SV classification, and introduce the Cohen's group time-frequency representation, which is used to process the non-stationary signals so as to define the classifier input space. We show that the SV classifier outperforms alternative classification methods on this processed data.

1. INTRODUCTION

The classification of nonstationary signals is a difficult and much studied problem. On one hand, the nonstationarity precludes classification in the time or frequency domain; on the other hand, nonparametric representations such as time-frequency or time-scale representations, while suited to nonstationary signals, have high dimension. Time-Frequency Representations (TFRs) and distance measures adapted to their comparison have previously been used to classify nonstationary signals [1, 2, 6], however the decision rules chosen in these studies limit the performance of these classification algorithms.

Support vector machines (SVMs) [10] provide efficient and powerful classification algorithms, which are capable of dealing with high dimensional input features, and with theoretical bounds on the generalisation error and sparseness of the solution provided by statistical learning theory [12, 10]. Classifiers based on SVMs have few free parameters requiring tuning, are simple to implement, and are trained through optimisation of a convex, quadratic cost function, which ensures the uniqueness of the SVM solution. Furthermore, SVM based solutions are sparse in the training data, and are defined only by the most "informative" training points.

In this paper, we propose to use a support vector machine for binary classification of the TFRs of nonstationary signals. In Section 2, we review support vector classifiers. In Section 3, we propose a classifier implementation based on Cohen's group TFRs, and in Section 4 we compare the classification results obtained with the SVM-TFR approach to those found using other classification methods.

2. SUPPORT VECTOR CLASSIFICATION

We first describe how support vector machines may be used in binary classification, using the ν-SV procedure. The results in this section are derived in Schölkopf *et al.* [9], and are also described in detail in Schölkopf and Smola [10].

Assume a sample of N labeled training points,

$$z \triangleq ((\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N,$$

in which $\mathbf{x}_i \in \mathcal{X}$, where $\mathcal{X}$ is the input space, and $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ is the label space. For our purposes, we define $\mathcal{Y} \triangleq \{-1, 1\}$, which corresponds to a two class classification problem. We seek to determine a function

$$\begin{aligned} \psi : \mathcal{X} &\to \mathcal{Y} \\ \mathbf{x} &\mapsto \psi(\mathbf{x}), \end{aligned}$$

that best predicts the label y for a vector $\mathbf{x}$. Assuming that random variable pairs $(\mathbf{x}, \mathrm{y})$ are generated i.i.d according to a distribution $\mathbf{P}_{\mathbf{x},\mathrm{y}}$, the optimal predicted class label for an input $\mathbf{x}$ is

$$\psi(\mathbf{x}) = \arg\max_y \mathbf{P}_{\mathrm{y}}(\mathrm{y}|\mathbf{x} = \mathbf{x}).$$

Since we do not know the mapping $\psi(\cdot)$, we define a learn-

0-7803-7011-2/01/$10.00 ©2001 IEEE

ing algorithm $\mathcal{A}$,

$$\mathcal{A}: \bigcup_{N=1}^{\infty} (\mathcal{X}, \mathcal{Y})^N \to \mathcal{H}$$
$$z \mapsto \psi_z(\cdot),$$

within a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ (here $\mathcal{Y}^{\mathcal{X}}$ refers to the set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$), which we call the *hypothesis space*, that is flexible enough to model a wide range of decision boundaries. We next define a *feature space* $\mathcal{F}$, endowed with an inner product[1] $\langle \cdot, \cdot \rangle_{\mathcal{F}}$, and a mapping from $\mathcal{X}$ to $\mathcal{F}$,

$$\mathbf{\Phi}: \mathcal{X} \to \mathcal{F}$$
$$\mathbf{x} \mapsto \mathbf{\Phi}(\mathbf{x}).$$

Let us restrict $\mathcal{H}$ to functions of the form

$$\mathcal{H} := \{x \mapsto \operatorname{sign}(\langle \mathbf{\Phi}(\mathbf{x}), \mathbf{w} \rangle + b) \,|\, \mathbf{w} \in \mathcal{F}, b \in \mathbb{R}\}.$$

We can then define a function $f_z(x)$ in $\mathbb{R}^{\mathcal{X}}$, such that $\psi_z(\cdot) = \mathcal{A}(z) = \operatorname{sign}(f_z(\cdot))$; thus

$$f_z(x) = \langle \mathbf{\Phi}(\mathbf{x}), \mathbf{w} \rangle + b, \tag{1}$$

and the problem of finding a *nonlinear* decision boundary in $\mathcal{X}$ has been transformed into a problem of finding the optimal *hyperplane* in $\mathcal{F}$ separating the two classes, where this hyperplane is parametrised by $(\mathbf{w}, b)$.

The mapping $\mathbf{\Phi}(\cdot)$ need never be computed explicitly; instead, we use the fact that if $\mathcal{F}$ is the reproducing kernel Hilbert space induced by $k(\cdot, \cdot)$, then

$$\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j).$$

The latter requirement is met for kernels fulfilling the Mercer conditions [10]. These conditions are satisfied for a wide range of kernels, including Gaussian radial basis functions,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathcal{X}}}{2\sigma^2}\right). \tag{2}$$

An estimate $f_z(\cdot)$ associated with the *loss* $c(\mathbf{x}, y, f_z(\cdot))$ is attained by minimising the *risk* $R(g_z(\cdot))$, i.e.

$$f_z(\cdot) = \operatorname*{argmin}_{g_z(\cdot) \in \mathcal{F}} \left[R(g_z(\cdot)) \triangleq \mathbf{E}_{\mathbf{x},\mathbf{y}}[c(\mathbf{x}, \mathbf{y}, g_z(\mathbf{x}))]\right]. \tag{3}$$

Possible loss functions include the soft margin loss [3, 5],

$$c(\mathbf{x}, y, g_z(\mathbf{x})) = \begin{cases} 0 & \text{if } y\, g_z(\mathbf{x}) \geq \rho, \\ \rho - y\, g_z(\mathbf{x}) & \text{otherwise,} \end{cases} \tag{4}$$

and the logistic regression loss [8],

$$c(\mathbf{x}, y, g_z(\mathbf{x})) = \log(1 + \exp(-y\, g_z(\mathbf{x}))), \tag{5}$$

[1] We omit the inner product subscript in the subsequent discussion, unless the inner product is taken in a space other than $\mathcal{F}$.

among others. The present study is confined to the case of soft margin loss, which has been used successfully with support vector methods in a wide variety of classification problems [10].

In practice, equation (3) cannot readily be solved, as we do not usually know the distribution $\mathbf{P}_{\mathbf{x},\mathbf{y}}$. Minimising the empirical risk alone does not take into account other factors, such as the complexity of the classifying function, and can therefore result in overfitting [10, 12].

We now describe the optimisation problem to be undertaken in finding $f_z(\mathbf{x})$. All support vector classification methods involve the minimisation of a regularised risk functional, which represents a tradeoff between classifier complexity and training error (the latter is determined by the cost functional). In the case of the ν-SV method, the regularised risk $R_{\text{reg}}(f_z(\cdot), z)$ at the optimum is given by

$$\min_{f_z(\cdot) \in \mathcal{F}} [R_{\text{reg}}(f_z(\cdot), z)] = \min_{\mathbf{w}, b, \rho} \left[\frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + R^{\rho}_{\text{emp}}(f_z(\cdot), z)\right], \tag{6}$$

where we use the soft margin loss from equation (4) in the empirical risk;

$$R^{\rho}_{\text{emp}}(f_z(\cdot), z) = \frac{1}{N}\sum_{i=1}^{N} c(\mathbf{x}_i, y_i, f_z(\cdot)) = \frac{1}{N}\sum_{i=1}^{N} \xi_i,$$

in which

$$\xi_i = \max\{0,\ \rho - y\, f_z(\mathbf{x})\}.$$

All training points $(\mathbf{x}_i, y_i)$ for which $y_i f_z(\mathbf{x}_i) \leq \rho$ are known as *support vectors*; it is only these points that determine $f_z(\cdot)$. The rôle of the term ν in equation (6) is described in the following theorem, from Schölkopf *et al.* [9].

Theorem 1 *The following results hold* only *for solutions to the optimisation problem in equation (6) for which* $\rho > 0$.

1. *ν is an upper bound on the fraction of training points for which $y_i f_z(\mathbf{x}_i) < \rho$, which we call* margin errors.
2. *ν is a lower bound on the fraction of training points for which $y_i f_z(\mathbf{x}_i) \leq \rho$ (the support vectors).*
3. *Assume a data set z generated iid according to $\mathbf{P}_{\mathbf{x},\mathbf{y}}$, and that neither $\mathbf{P}_{\mathbf{x}}(\mathbf{x}|\mathbf{y} = 1)$ nor $\mathbf{P}_{\mathbf{x}}(\mathbf{x}|\mathbf{y} = -1)$ contains any discrete component. Then, given a kernel $k(\cdot, \cdot)$ that is analytic and non-constant, with probability 1, asymptotically, ν is equal to the fraction of support vectors and the fraction of margin errors.*

It can be shown [9] that the component $\mathbf{w}$ in equation (1) is a linear combination of the mapped training points,

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{\Phi}(\mathbf{x}_i).$$

and that solving equation (6) is equivalent to finding

$$\max_{\boldsymbol{\alpha}} \left(-\frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right)$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{N},$$
$$\sum_{i=1}^{N} y_i \alpha_i = 0,$$
$$\sum_{i=1}^{N} \alpha_i \geq \nu.$$

There exist a number of methods that can be used to solve this quadratic programming problem. Our results were obtained using the LOQO algorithm in Vanderbei [11]. In the case of large training sets, data decomposition methods exist to speed convergence; see e.g. Chang *et al.* [4]. The offset b and soft margin loss parameter ρ are found using

$$y_j \left(\langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x}_j) \rangle_{\mathcal{H}} + b \right) = \rho \quad \text{when } \alpha_j \in \left(0, \frac{1}{N}\right);$$

the set of equations thus obtained can be solved via linear least squares.

3. KERNEL DESIGN

The ν-SVM classification procedure relies on the choice of a kernel $k(\cdot,\cdot)$ suited to the problem at hand. In Davy *et al.* [6], a nonstationary signal classification algorithm was introduced, based on Cohen's group time-frequency representations. In this paper, we choose a ν-SVM kernel $k(\cdot,\cdot)$ based on a similar approach.

We write the Cohen's group time-frequency representation of $s(t)$ as $C_s^{\phi}(t,f)$ (parametrised by its TFR kernel[2] ϕ). Given two signals $s(t)$ and $s'(t)$, the Gaussian radial basis function kernel of equation (2) then becomes

$$k(\mathbf{x},\mathbf{x}') = \exp -\frac{1}{2\sigma^2} \left[\int\int \left| \mathrm{NC}_s^{\phi}(t,f) - \mathrm{NC}_{s'}^{\phi}(t,f) \right|^2 dtdf \right], \quad (7)$$

[2]In order to avoid confusion between the ν-SVM kernel and the TFR kernel, the latter will be referred to as the *TFR kernel* at all times.

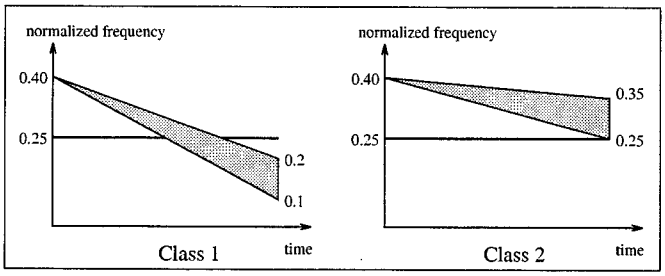


Fig. 1. Support of the noise-free TFRs (the gray areas represent the possible instantaneous frequencies for each class).

where the notation $\mathrm{NC}_s^{\phi}(t,f)$ is used to show that the TFR is normalised;

$$\mathrm{NC}_s^{\phi}(t,f) = \frac{\left| C_s^{\phi}(t,f) \right|}{\int\int \left| C_s^{\phi}(t,f) \right| dtdf}. \quad (8)$$

In this formulation, the input space $\mathcal{X}$ defined in previous section is the space of normalised TFRs (i.e., $\mathbf{x} = \mathrm{NC}_s^{\phi}(t,f)$), which depends on the choice of the TFR kernel ϕ.

4. RESULTS

We now apply the ν-SVR algorithm to the binary classification of chirp signals, and compare our results to those obtained previously by Davy *et al.* [6, 7]. The test signals are defined as the sum of two linear chirps:

$$\begin{aligned} x(k) &= A \sin\left[2\pi(a_0 + a_1 k)\right] \\ &+ B \sin\left[2\pi(b_0 + b_1 k + b_2 k^2)\right] \\ &+ \epsilon(k)\ ,\ k = 0, \ldots, K-1, \end{aligned} \quad (9)$$

where the $\epsilon(k)$ are iid, and are generated by a zero mean Gaussian process with variance σ_ϵ^2. Each test signal $x(k)$ is parametrized by $\theta = \left(A, B, a, b, \sigma_\epsilon^2\right)$, with $a = (a_0, a_1)$ and $b = (b_0, b_1, b_2)$. The problem consists of classifying a given signal $x(k)$ into one of the two following classes:

- Class $\omega_1 : p(b_2) \sim \mathcal{U}(\frac{-0.30}{2(K-1)}, \frac{-0.20}{2(K-1)})$, where $\mathcal{U}(a,b)$ is the uniform distribution on (a,b),
- Class $\omega_2 : p(b_2) \sim \mathcal{U}(\frac{-0.15}{2(K-1)}, \frac{-0.05}{2(K-1)})$.

The remaining signal parameters are identical in both classes, i.e. $A = B = 1$, $a_0, b_0 \sim \mathcal{U}(0,1)$, $a_1 = 0.25$ and $b_1 = 0.40$. The support of the noise-free time-frequency representation for signals in each class is plotted in figure 1.

The ν-SVM algorithm was trained using 100 signals, with an equal number of examples in each class. We specified a kernel width of $\sigma^2 = 0.1$ (see equation (7)), and set $\nu = 0.2$. A radially symmetric Gaussian TFR kernel ϕ was

selected, with parameters optimised to minimise the error rate observed on the test data.

To measure the performance of the algorithm, a total of 20000 randomly generated test signals were used, again divided equally between the two classes (note that the training signals did not form part of the test set). Table 1 shows the average error over these test signals, compared with the average obtained over the same number of test signals for alternative classification methods. We see that for this problem, the ν-SVM algorithm achieves the lowest error rate.

Classification method	Error rate
Wigner distribution [1]	22.30 %
Ambiguity plane [2]	4.56 %
Time-Frequency [6]	2.25 %
MCMC classification [7]	5.24 %
SVM classification [this paper]	1.51 %

Table 1. Error rates for the classification of chirps, using the proposed SVM implementation and other classifiers.

5. CONCLUSION

In this study, we show that the good performance of SV classifiers in high dimensions allows us to effectively classify chirp signals, when these are transformed using Cohen's group time-frequency kernels. Additional advantages of the SV classification method include simplicity of implementation, relatively low computational cost, and uniqueness of the SVM solution.

6. REFERENCES

[1] S. Abeysekera and B. Boashash. Methods of signal classification using the images produced by the wigner distribution. *Pattern Recognition Letters*, 12:717 – 729, November 1991.

[2] L. Atlas, J. Droppo, and J. McLaughlin. Optimizing time-frequency distributions for automatic classification. In *SPIE - The International Society for Optical Engineering*, 1997.

[3] K Bennett and O. Mangasarian. Robust linear programming discrimination of two linearly inseperable sets. *Optimisation methods and software*, 1:23–34, 1993.

[4] C.-C. Chang, C.-W. Hsu, and C.-J. Lin. The analysis of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 11(4):1003–1008, 2000.

[5] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, pages 273–297, 1995.

[6] M. Davy, C. Doncarli, and G. Faye Boudreaux-Bartels. Improved optimization of time-frequency based signal classifiers. *IEEE Signal processing letters*, 8(2):52–57, February 2001.

[7] M. Davy, C. Doncarli, and J.Y. Tourneret. Supervised bayesian learning using mcmc methods. application to the classification of chirps. Technical Report CUED/F-INFENG/TR.401, Engineering Department, University of Cambridge, UK, 2001.

[8] P. Huber. *Robust statistics*. John Wiley and Sons, New York, 1981.

[9] B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[10] A. Smola and B. Schölkopf. *Learning with Kernels*. MIT press, To appear.

[11] R. J. Vanderbei. LOQO: An interior point code for quadratic programming. Technical Report TR SOR-94-15, Department of Civil Engineering and Operations Research, Princeton University, 1995.

[12] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, N.Y., 1995.

IMPROVED AUXILIARY PARTICLE FILTERING: APPLICATIONS TO TIME-VARYING SPECTRAL ANALYSIS

Christophe Andrieu[1], *Manuel Davy*[2] *and Arnaud Doucet*[3]

[1] Department of Mathematics, Statistics Group, University of Bristol, University Walk, Bristol BS8 1TW, U.K. C.Andrieu@bristol.ac.uk
[2]Department of Engineering, University of Cambridge, Trumpington Street, CB2 1PZ Cambridge, U.K. md283@eng.cam.ac.uk
[3] Department of Electrical and Electronic Engineering, University of Melbourne, Victoria 3010 Australia. a.doucet@ee.mu.oz.au

ABSTRACT

This paper addresses optimal estimation for time varying autoregressive (TVAR) models. First, we propose a statistical model on the time evolution of the frequencies, moduli and real poles instead of a standard model on the AR coefficients as it makes more sense from a physical viewpoint. Second, optimal estimation involves solving a complex optimal filtering problem which does not admit any closed-form solution. We propose a new particle filtering scheme which is an improvement over the so-called auxiliary particle filter. The hyperparameters tuning the evolution of the model parameters are also estimated on-line so as to robustify the model. Simulations demonstrate the efficiency of both our model and algorithm.

1. INTRODUCTION

Many models in signal processing can be cast in a state space form. In most applications, prior knowledge of the system is also available. This knowledge allows us to adopt a Bayesian approach; that is, to combine a prior distribution for the unknown quantities with a likelihood function relating these quantities to the observations. Within this setting one performs inference on the unknown state via the posterior distribution. Often, the observations arrive sequentially in time and one is interested in *estimating recursively in time* the evolving posterior distribution. This problem is known as the Bayesian or *optimal filtering* problem [1]. In many realistic problems, state space models must include elements of non linearity and non Gaussianity that preclude a closed form expression for the optimal filter. For over thirty years, many approximation schemes, such as the extended Kalman filter, have been proposed to tackle this problem; see [1]. Unfortunately, in many cases, these suboptimal methods are unreliable.

Following the seminal paper by Gordon, Salmond and Smith introducing the *bootstrap filter/SIR* [5], there has been a surge of interest in particle filtering methods to solve the optimal filtering problem numerically; see [3], [4]. These methods are Sequential Monte Carlo (SMC) methods that utilize a large number, N, of random samples (or particles) to represent the posterior probability distributions. They are very flexible and can be easily applied to nonlinear and non Gaussian dynamic models.

We sum up here the contributions of our paper: we first propose an original model for TVAR. It relies on a pole type parameterization of the problem which is physically sound, versatile and robust. More precisely, our model takes into account model uncertainty: the order of the TVAR is assumed unknown and is estimated on line. The hyperparameters which might influence the results are also part of the inference process, therefore robustifying the model. The proposed model is complex and requires the use of state of the art particle filtering techniques. We introduce here a modification of the auxiliary particle filtering method relying on an approximate computation of the "one-step ahead likelihood". It is shown to lead to substantial improvements in simulation.

2. IMPROVED AUXILIARY PARTICLE FILTERING

2.1. Problem Statement

Let (Ω, F, P) be a probability space on which we have defined two real vector-valued stochastic processes $X = \{X_t,\ t \in \mathbb{N}\}$ and $Y = \{Y_t,\ t \in \mathbb{N}^*\}$. The process X is usually called the *signal* process and the process Y is called the *observation* process. Let $\mathbb{R}^{n_x}$ and $\mathbb{R}^{n_y}$ be the dimensions of the state space of X and Y. The *signal* process X is a Markov process with initial density $p(x_0)$ and probability transition density $p(x_t|x_{t-1})$. The*observations* are independent conditional upon X and have marginal density $g(y_t|x_t)$.

For $p < q$ and any sequence z_t, we denote $z_{p:q} = (z_p, z_{p+1}, \ldots, z_q)$. Bayes' theorem allows us to propagate over time the joint posterior distribution $p(x_{0:t}|y_{1:t})$

$$p(x_{0:t}|y_{1:t}) \propto g(y_t|x_t)p(x_t|x_{t-1})p(x_{0:t-1}|y_{1:t-1})$$

and the marginal filtering distribution

$$p(x_t|y_{1:t}) \propto g(y_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_t$$

where $\propto$ denotes "proportional to". Except in very special cases, these densities do not admit any closed-form expression and some numerical methods are required to approximate them.

2.2. Particle Filtering Method

Particle filtering methods are loosely speaking a set of sampling/ resampling methods. These are recursive algorithms which pro-

0-7803-7011-2/01/$10.00 ©2001 IEEE

duce, at each time t, a cloud of particles $\left\{x_{0:t}^{(i)}\right\}_{i=1}^{N}$ whose empirical measure

$$P^N(dx_{0:t}|y_{1:t}) = \sum w_t^{(i)} \delta_{x_{0:t}^{(i)}}(dx_{0:t}),\ w_t^{(i)} > 0,\ \sum_{i=1}^{N} w_t^{(i)} = 1$$

closely "follows" the distribution

$$P(dx_{0:t}|y_{1:t}) = p(x_{0:t}|y_{1:t})dx_{0:t}$$

2.2.1. Sequential Importance Sampling/Resampling

At time $t-1$, assume one has the following approximation of $P(dx_{0:t-1}|y_{1:t-1})$

$$P^N(dx_{0:t-1}|y_{1:t-1}) = \frac{1}{N}\sum \delta_{x_{0:t-1}^{(i)}}(dx_{0:t-1}).$$

We extend the current paths $x_{0:t-1}^{(i)}$ by sampling $x_t^{(i)} \sim q(x_t|y_{1:t}, x_{0:t-1})$. Then using the importance sampling identity,

$$p(x_{0:t}|y_{1:t}) \propto$$
$$p(x_{0:t-1}|y_{1:t-1})\frac{g(y_t|x_t)p(x_t|x_{t-1})}{q(x_t|y_{1:t}, x_{0:t-1})}q(x_t|y_{1:t}, x_{0:t-1})$$

one gets for the new weights

$$w_t^{(i)} \propto \frac{g(y_t|x_t^{(i)})p(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|y_{1:t}, x_{0:t-1}^{(i)})}.$$

Then one uses a resampling step: particles with high weights $w_t^{(i)}$ are copied several times whereas particles with low weights are discarded. After this resampling step, the weights are reset to N^{-1}.

In the case where one uses the "optimal" importance distribution [3]

$$q(x_t|y_{1:t}, x_{0:t-1}) = p(x_t|y_{1:t}, x_{t-1}) = \frac{g(y_t|x_t)p(x_t|x_{t-1})}{p(y_t|x_{t-1})},$$

then it is easy to see that

$$w_t^{(i)} \propto p(y_t|x_{t-1}^{(i)}).$$

That is the weight is independent of $x_t^{(i)}$. This suggests that resampling should be performed before sampling $\left\{x_t^{(i)}\right\}_{i=1}^{N}$ as it will obviously lead to an increased number of distinct particles. Unfortunately, this method cannot be used for most models as $p(y_t|x_{t-1})$ does not admit a closed-form expression.

2.2.2. An improved auxiliary particle filtering method

We present here an improved version of the Auxiliary Particle Filtering method [6]. In the APF, one approximates the "predictive likelihoods" $\left\{p(y_t|x_{t-1}^{(i)})\right\}_{i=1}^{N}$ by, say, $\left\{\widehat{p}(y_t|x_{t-1}^{(i)})\right\}_{i=1}^{N}$. Then one resamples the particles $\left\{x_{0:t-1}^{(i)}\right\}_{i=1}^{N}$ w. r. t. $w_{t-1}^{(i)}\widehat{p}(y_t|x_{t-1}^{(i)})$; the aim being to boost the number of particles in useful regions of the state space. Then, if we sample $\left\{x_t^{(i)}\right\}_{i=1}^{N}$ according to $q(x_t|y_{1:t}, x_{0:t-1})$. It is easy to check that the weights must then satisfy

$$w_t^{(i)} \propto \frac{g(y_t|x_t^{(i)})p(x_t^{(i)}|x_{t-1}^{(i)})}{\widehat{p}(y_t|x_{t-1}^{(i)})q(x_t^{(i)}|y_{1:t}, x_{0:t-1}^{(i)})}$$

for this procedure to be statistically consistent.

This method will only work well when the approximation of $\widehat{p}(y_t|x_{t-1}^{(i)})$ is correct. One has

$$p(y_t|x_{t-1}) = \int p(y_t|x_t)p(x_t|x_{t-1})dx_t.$$

In [6], it is suggested to approximate this integral by $p(y_t|x_t = \mu(x_{t-1}))$ where $\mu(x_{t-1})$ is the mode or median of $p(x_t|x_{t-1})$. This approximation might be very poor if $p(x_t|x_{t-1})$ is rather diffuse and $p(y_t|x_t)$ varies a lot over the prior $p(x_t|x_{t-1})$. It would be of course possible to approximate $p(y_t|x_{t-1})$ using a (second stage) Monte Carlo method but this would be highly computationally intensive.

It is possible to approximate this expression using numerical methods. We propose here a simple though efficient deterministic method known as the unscented transform. This has the advantage of computing both $\widehat{p}(y_t|x_{t-1})$ and $q(x_t|y_{1:t}, x_{0:t-1})$, see [7, 8] for details.

Improved Auxiliary particle filtering algorithm

At time $t = 0$, **Step 0**:
Initialization

- For $i = 1, ..., N$, sample $\widetilde{x}_0^{(i)} \sim p(x_0)$ and set $t = 1$.

At time $t \geq 1$, **Step 1**: *Auxiliary variable resampling step*

- For $i = 1, ..., N$, compute $\lambda_t^{(i)}$ as

$$\lambda_t^{(i)} \propto w_{t-1}^{(i)}\,\widehat{p}(y_t|\widetilde{x}_{t-1}^{(i)}),\ \sum_{i=1}^{N}\lambda_t^{(i)} = 1 \qquad (1)$$

  where $\widehat{p}(y_t|\widetilde{x}_{t-1}^{(i)})$ is computed using the unscented approximation.

- Multiply/Discard particles $\left\{\widetilde{x}_{t-1}^{(i)}\right\}_{i=1}^{N}$ with respect to high/low importance weights $\lambda_t^{(i)}$ to obtain N particles $\left\{x_{t-1}^{(i)}\right\}_{i=1}^{N}$.

Step 2: *Importance sampling step*

- For $i = 1, ..., N$, use the unscented Kalman filter to compute $\widehat{x}_t^{(i)}$ and $\widehat{P}_{t|t-1}^{(i)}$ that are respectively the estimates of $E(x_t|y_{1:t}, x_{t-1}^{(i)})$ and $\mathrm{Cov}(x_t|y_{1:t}, x_{t-1}^{(i)})$
- For $i = 1, ..., N$, sample $\widetilde{x}_t^{(i)} \sim q(x_t|y_{1:t}, x_{0:t-1}^{(i)})$ where the importance distribution is

$$q(x_t|y_{1:t}, x_{0:t-1}^{(i)}) = \mathcal{N}(x_t; \widehat{x}_t^{(i)}, \widehat{P}_{t|t-1}^{(i)})$$

- Update the importance weights as

$$w_t^{(i)} = \frac{p(y_t|\widetilde{x}_t^{(i)})\,p(\widetilde{x}_t^{(i)}|x_{t-1}^{(i)})}{\widehat{p}(y_t|\widetilde{x}_{t-1}^{(i)})\,q(\widetilde{x}_t^{(i)}|y_{1:t}, x_{0:t-1})}$$

3. APPLICATION TO BAYESIAN TIME-VARYING SPECTRAL ANALYSIS

Time-Varying models have received much attention as tools for nonstationary signals analysis [9, 10]. Time-Varying AR models consist in the following recursive process:

$$y_t = a_{1,t}y_{t-1} + a_{2,t}y_{t-2} \dots a_{p,t}y_{t-p} + v_t \quad (2)$$

where $\mathbf{a}_t = [a_{1,t}, a_{2,t}, \dots, a_{p,t}]^T$ is the vector of AR coefficients at time t, v_t is a zero-mean Gaussian white noise of variance R, and p is the AR model order. Denote $\mathbf{y}_t = [y_{t-1}\, y_{t-2}\, y_{t-p}]^T$, then Eq. (2) can be written as:

$$y_t = \mathbf{a}_t^T \mathbf{y}_t + v_t \quad (3)$$

The AR coefficients can be equivalently expressed in terms of frequencies $\boldsymbol{\nu}_t = [\nu_{1,t}\, \nu_{2,t}\, \dots\, \nu_{p^c,t}]^T$, moduli $\boldsymbol{\rho}_t = [\rho_{1,t}\, \rho_{2,t}\, \dots\, \rho_{p^c,t}]^T$, and real poles $\mathbf{r}_t = [r_{1,t}\, r_{2,t}\, \dots\, r_{p^r,t}]^T$, with $p = 2p^c + p^r$. The poles $r_{k,t}$, $z_{k,t} = \rho_{k,t}e^{j2\pi\nu_{k,t}}$ and its complex conjugate $z^\star_{k,t}$ are the roots of the polynomial:

$$1 - a_{1,t}X - a_{2,t}X^2 - \dots - a_{p,t}X^p$$

In the following, the transform $(\boldsymbol{\nu}_t, \boldsymbol{\rho}_t, \mathbf{r}_t) \rightarrow \mathbf{a}_t$ is denoted $\mathbf{a}_t = AR(\boldsymbol{\nu}_t, \boldsymbol{\rho}_t, \mathbf{r}_t)$. This latter formulation enables a physically sound model of the time evolution of the AR coefficients.

3.1. Bayesian model

A simple state space representation is given by:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}_t \quad (4)$$
$$y_t = \mathrm{AR}(\mathbf{x}_t)^T \mathbf{y}_t + v_t \quad (5)$$

where $\mathbf{u}_t$ is a zero-mean Gaussian white noise (referred to as *dynamic noise*) with diagonal covariance matrix $\mathbf{Q}$. Given an integer $M > 1$, the state vector $\mathbf{x}_t$ is $M \times p$-dimensional and consists of the frequencies, moduli and real poles from time $t-M+1$ to time t as:

$$\mathbf{x}_t = \left[\boldsymbol{\nu}_{t-M+1:t}^T\, \boldsymbol{\rho}_{t-M+1:t}^T\, \mathbf{r}_{t-M+1:t}^T\right]^T \quad (6)$$

the matrices $\mathbf{A}$ and $\mathbf{B}$ are such that, e.g.,

$$\nu_{1,t} = \frac{\nu_{1,t-1} + \dots, \nu_{1,t-M}}{M} + u_{1,t}$$

This is a simple smoothness prior. Rather than relying on the AR coefficients, the model defined in Eqs. (4) and (5) is based on a modulus-frequency representation, which is more convenient for modelling the signal evolution in time. This model is however highly non-linear due to the transform $\mathbf{a}_t = AR(\mathbf{x}_t)$. The estimate $\widehat{\mathbf{x}}_t$ of $\mathbf{x}_t$ given $y_{1:t}, \mathbf{x}_{0:t-1}$ requires filters adapted to nonlinear Gaussian models, such as those presented in previous section.

3.2. Extended Bayesian model

In most applications, however, the hyperparameters $\mathbf{Q}$, R and the model order p are not known a priori. It is still possible to tune these hyperparameters "by hand", but they may not be constant w.r.t. time: for example, p might change when a spectral trajectory appears or disappears. A possible solution consists of extending the model so as to define an extended Bayesian model, that includes the hyperparameters. The above model becomes:

$$p_t^c, p_t^r \sim h(p_t^c, p_t^r | p_{t-1}^c, p_{t-1}^r) \quad (7)$$
$$\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}, p_t^c, p_t^r, \mathbf{Q}_{t-1}) \quad (8)$$
$$\mathbf{Q}_t \sim p(\mathbf{Q}_t | \mathbf{Q}_{t-1}, p_t^c, p_t^r) \quad (9)$$
$$R_t \sim p(R_t | R_{t-1}) \quad (10)$$
$$y_t \sim g(y_t | \mathbf{x}_t, p_t^c, p_t^r, R_t, y_{1:t-1}) \quad (11)$$

where Eq. (8) is similar to Eq. (4), the likelihood given Eq. (11) is computed in the same way as in Eq. (5). Eq.'s (9) and (10) correspond to a random walk (written, e.g., for R_t):

$$\log(R_t) = \log(R_{t-1}) + \epsilon_t^R \quad (12)$$

where ϵ_t^R is a centered Normal noise with variance δ_R^2. Eq. (7) is a bivariate discrete distribution which enables five possible moves with equal probability: ($p_t^c = p_{t-1}^c$ and $p_t^r = p_{t-1}^r$), ($p_t^c = p_{t-1}^c$ and $p_t^r = p_{t-1}^r + 1$), ($p_t^c = p_{t-1}^c$ and $p_t^r = p_{t-1}^r - 1$), ($p_t^c = p_{t-1}^c + 1$ and $p_t^r = p_{t-1}^r$) and ($p_t^c = p_{t-1}^c - 1$ and $p_t^r = p_{t-1}^r$) corresponding to update, birth or death of a trajectory. This prior does not enable trajectory birth (resp. death) when the total number of trajectories – i.e., the TVAR model order – reaches an upper bound (resp. a lower bound).

3.3. Implementation

The algorithm presented in Section 2 is applied to the model described above (the actual state vector includes the frequencies, the moduli and real poles from time $t-M+1$ to time t, the logarithms of the diagonal terms of the matrix $\mathbf{Q}_t$, $\log(R_t)$ and the orders p_t^c and p_t^r). For the sake of simplicity, we consider constant orders p_t^c and p_t^r in this section. The proposal distribution for $[\boldsymbol{\nu}_t, \boldsymbol{\rho}_t, \mathbf{r}_t]$ is a multivariate normal density with parameters estimated using the unscented Kalman filter, as described in Section 2. The hyperparameters are sampled using an accept/reject method, according to, e.g., for R_t:

$$p(R_t | R_{t-1}, v_t) = \frac{p(v_t|R_t)p(R_t|R_{t-1})}{p(v_t|R_{t-1})}$$
$$\propto \exp -\frac{1}{2}\left[\frac{v_t^2}{R_t} + \log R_t + \frac{(\log R_t - \log R_{t-1})^2}{\delta_R^2}\right]$$

where the innovation is $v_t = y_t - \mathrm{AR}(\mathbf{x}_t)^T \mathbf{y}_t$, from Eq. (5). The algorithm is initialized using the frequencies, moduli and real poles computed from the AR coefficients estimated using the modified covariance AR estimator applied to the $2p$ first points of the signal.

3.4. Simulations

In this section, we present results obtained with a three frequency components signal (RSB=24 dB). The model orders are fixed such that $p^c = 3$ and $p^r = 1$. Figure 1 represents the spectrogram of the analysed signal. The MMSE estimates of the frequencies, moduli and real poles at each time instant are computed using the proposed particle filter, with $N = 500$ particles, and $M = 8$ (see Figure 2). These parameters are accurately tracked, in spite of the signal nonstationarity. Figure 3 displays the hyperparameters estimates over time: in particular, the excitation noise variance R_t is stable, which confirms the accuracy of the spectral trajectories tracking.

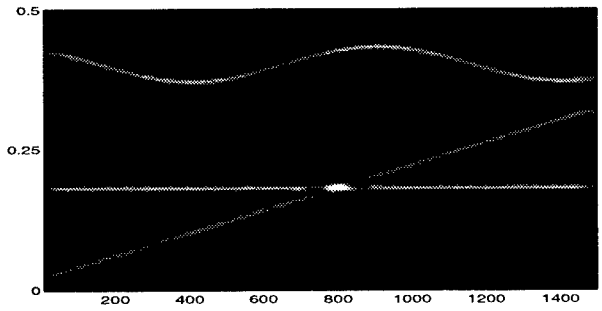


Fig. 1. Spectrogram of the processed signal.

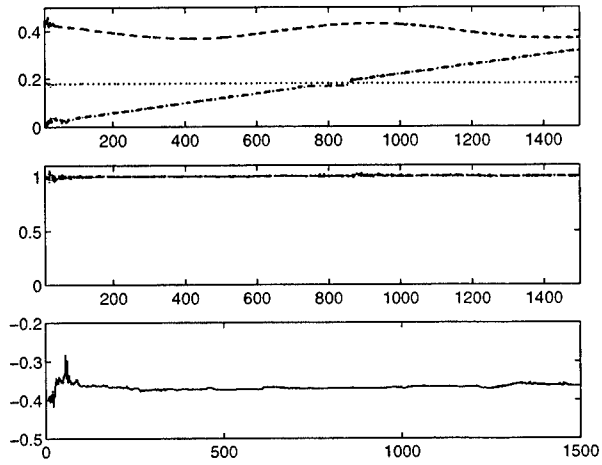


Fig. 2. TVAR estimation of the frequencies (top), moduli (middle) and real pole (bottom) of a three-component signal.

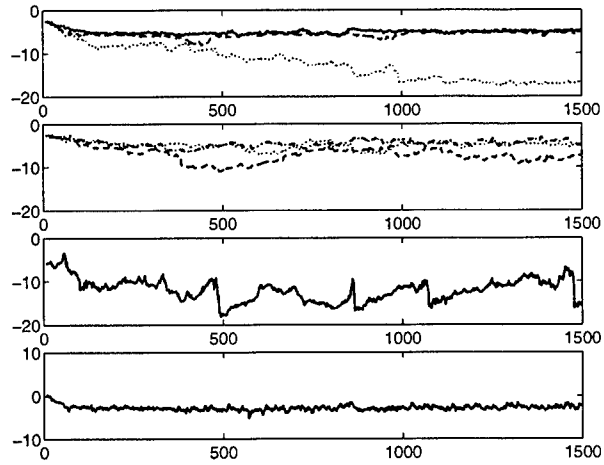


Fig. 3. Evolution of the hyperparameters over time. The $\log_{10}$ of the hyperparameters tuning the frequencies is plotted in the top row. The second row corresponds to the moduli hyperparameters, and the third row corresponds to the real pole. $\log_{10}(R_t)$ is plotted in the bottom row.

4. CONCLUSION

In this paper, we have introduced an improved auxiliary particle filtering method. Its implementation requires one resampling step, and the proposal distribution is computed using a numerical approximation. Simulations demonstrate the efficiency of both the model and the algorithm: the frequencies, magnitudes and hyperparameters are accurately tracked. In particular, we would like to underline the robustness of the method, as the input information from the user is minimum with our approach. The full implementation of the algorithm is still under development at the time of writing this paper, but results will be reported in [11].

5. REFERENCES

[1] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*, Englewood Cliffs, 1979.

[2] D. Crisan, P. Del Moral and T. Lyons, "Discrete filtering using branching and interacting particle systems", *Markov Proc. Rel. Fields*, vol. 5, 293-318, 1999.

[3] A. Doucet, S.J. Godsill and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering", *Statistics & Computing*, vol. 10, pp. 197-208, 2000.

[4] A. Doucet, J.F.G. de Freitas and N.J. Gordon (eds.), *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.

[5] N.J. Gordon, D.J. Salmond and A.F.M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation", *IEE-Proceedings-F*, vol. 140, no. 2, 1993, pp. 107-113.

[6] M.K. Pitt and N. Shephard, "Filtering via simulation:auxiliary particle filters", *J. Amer. Stat. Assoc.*, 1999.

[7] S.J. Julier, "The scaled unscented transformation", *Automatica*, 2000, to appear.

[8] E.A. Wan and R. van der Merwe, "The unscented Kalman filter for nonlinear estimation", in *Proc. Conf. Adaptive Systems for Signal Processing*, Communication and Control, Canada, 2000.

[9] N. Ikoma and H. Maeda, "Nonstationary spectral peak estimation by Monte Carlo Filter", in *Proc. Conf. Adaptive Systems for Signal Processing*, Communication and Control, Canada, 2000.

[10] R. Prado, G. Huerta and M. West, " Bayesian Time-Varying Autoregressions: Theory, Methods and Applications", To appear in *special issue on Time Series and Related Topics of "Resenhas", the Journal of the Institute of Mathematics and Statistics of the University of Sao Paolo.*

[11] C. Andrieu, M. Davy and A. Doucet, "Bayesian on-line estimation of TVAR model order", *Technical Report*, University of Cambridge, 2001.

SPATIAL AND TIME-FREQUENCY SIGNATURE ESTIMATION OF NONSTATIONARY SOURCES

Moeness G. Amin, Weifeng Mu, and Yimin Zhang

Department of Electrical and Computer Engineering,
Villanova University, Villanova, PA 19085, USA
E-mail: {moeness,weifeng,zhang}@ece.villanova.edu

ABSTRACT

Signal synthesis using time-frequency distributions can be improved using an antenna array receiver. The availability of the source signals at different array elements allows the implementation of time-frequency synthesis techniques that utilize the source spatial signatures for crossterm reduction and noise mitigation. In this paper, we introduce a new technique for signal synthesis based on array averaging of Wigner distributions. The source temporal waveforms are first synthesized and then used to estimate the source spatial signatures. Iterative process incorporating the source signal vector and array vector can be applied until desired results are reached.

1. INTRODUCTION

Synthesizing the signal from the Wigner-Ville distribution (WVD) is often impeded by the presence of high levels of noise and crossterms. These undesired terms not only obscure the true signal power localization in the time-frequency (t-f) domain, but also reduce the synthesized signal quality. Signal synthesis using time-frequency distributions (TFDs) can be improved using an antenna array receiver. The availability of the source signals at different array elements allows the implementation of t-f synthesis techniques that utilize the source spatial signatures for crossterm reduction and noise mitigation.

In [1], the WVDs of the data received at different antennas are averaged prior to synthesis. It is shown that spatial averaging of WVD decreases the noise levels, reduces the interactions of the source signals, and mitigates the crossterms. As such, it depicts enhanced t-f signatures of the sources incident on the multi-antenna receiver.

The procedures discussed in [1] is appropriate to synthesize the signal waveform whose t-f signatures are distinct. In this case, the masked t-f region always contains the autoterm of the desired source signal with the influence from other sources often negligible. However, if the source t-f signatures overlap, the mask is deemed to capture undesired autoterms. This problem cannot be mitigated by spatial averaging of TFDs and a modification of the proposed method is in order.

This paper discusses an iterative process that incorporates both the estimated source signal vector and array vector. The source temporal waveforms are first synthesized and then used to estimate the source spatial signatures.

The paper is organized as follows. A review of the technique proposed in [2, 3] for bilinear signal synthesis is given in Section 2. In Section 3, we introduce the array averaged WVD that reduces the effect of cross-terms and noise. Section 4 discusses the iterative synthesis process for signals with overlapping t-f signatures. Section 5 presents simulation results.

2. SIGNAL SYNTHESIS BASED ON WVD

The signal synthesis techniques based on WVDs can be found in [2, 3]. In this paper, we apply the method of extended discrete-time Wigner distribution (EDTWD), introduced in [4]. The EDTWD for a received data of $x(t)$ is defined as

$$W_{xx}(t,f) = \sum_{k:t+\frac{k}{2}\in Z} x(t+\frac{k}{2})x^*(t-\frac{k}{2})e^{-j2\pi kf}, \quad t = 0, \pm\frac{1}{2}, \pm 1, \cdots, \tag{1}$$

where * denotes complex conjugation, t and f represent the time index and the frequency index, respectively. Equation (1) is often referred to as the auto EDTWD of the signal $\mathrm{y}(t)$. Similarly, the cross EDTWD of any two signals $\mathrm{y}_1(t)$ and $\mathrm{y}_2(t)$ is defined as

$$W_{xx}(t,f) = \sum_{k:t+\frac{k}{2}\in Z} x(t+\frac{k}{2})x^*(t-\frac{k}{2})e^{-j2\pi kf}, \quad t = 0, \pm\frac{1}{2}, \pm 1, \cdots, \tag{2}$$

This work is supported by the Office of Naval Research under Grant N00014-98-1-0176, and the Air Force Research Laboratory under grant no. F30602-00-1-0515.

0-7803-7011-2/01/$10.00 ©2001 IEEE

The advantage of using the EDTWD in signal synthesis lies in the fact that it does not require *a priori* knowledge of the source waveform, and thereby avoids the problem of matching the two "uncoupled" even-indexed and odd-indexed vectors. In this paper, we refer to EDTWD as WVD for simplicity.

The overall procedure of WVD-based signal synthesis is summarized in the following steps.

1. Place an appropriate mask on $W_{xx}(t,f)$ such that only the desired signal autoterms are retained.

2. Take the inverse fast Fourier transform (IFFT) of $W_{xx}(t,f)$

$$p(t,\tau) = \int W_{xx}(t,f)e^{j2\pi\tau f}df. \tag{3}$$

3. Construct the matrix $\mathbf{Q} = [q_{i,j}]$ with

$$q_{i,j} = p\left(\frac{i+j}{2}, i-j\right). \tag{4}$$

4. Take the Hermitian component $\mathbf{Q}_H$ of $\mathbf{Q}$

$$\mathbf{Q}_H = \frac{1}{2}\left[\mathbf{Q} + \mathbf{Q}^H\right], \tag{5}$$

where the superscript H denotes transpose conjugation.

5. Apply eigen-decomposition to the matrix $\mathbf{Q}_H$ and obtain the maximum eigenvalue $\lambda_{\max}$ and the associated eigenvector $\mathbf{u}$. The synthesized signal is given by

$$\hat{x} = e^{j\phi}\sqrt{\lambda_{\max}}\mathbf{u}, \tag{6}$$

where ϕ is an unknown value representing the phase.

3. THE ARRAY AVERAGED WVD

Assume L source signals incident on an M-sensor array. The data received across the array is given by the narrowband model

$$\mathbf{x}(t) = \mathbf{y}(t) + \mathbf{n}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t), \quad t = 1, \cdots, N, \tag{7}$$

where $\mathbf{x}(t) = [\mathrm{x}_1(t), \cdots, \mathrm{x}_M(t)]^T$ and $\mathbf{s}(t) = [\mathrm{s}_1(t), \cdots, \mathrm{s}_L(t)]^T$ are the $M \times 1$ data snapshot vector and the $L \times 1$ source signal vector at time instant t, respectively, where we assume $\mathrm{s}_i(t), i = 1, \cdots, L$, are monocomponent signals. In (7), the superscript ^{T} denotes the vector/matrix transpose. The $M \times 1$ vector $\mathbf{n}(t)$ is the noise vector, whose elements are modeled as stationary, spatially and temporally white complex Gaussian processes with zero mean and variance of σ^2, i.e.,

$$E\left[\mathbf{n}(t+\tau)\mathbf{n}^H(t)\right] = \sigma^2\delta(\tau)\mathbf{I} \tag{8}$$

where $\delta(\tau)$ is the kronecker delta, $\mathbf{I}$ denotes the identity matrix, and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_L]$ denotes the $M \times L$ mixing matrix. The columns of matrix $\mathbf{A}$ are the source spatial signatures and are given by

$$\mathbf{a}_i = [\mathrm{a}_{i1}, \ldots, \mathrm{a}_{iM}]^T. \tag{9}$$

We assume that matrix $\mathbf{A}$ is of full column rank, which implies that the spatial signatures associated with the L sources are linearly independent. To simplify the discussion, we exchange any possible scalar factor embedded in $\mathbf{a}_i$ to the source signal and assume that $||\mathbf{a}_i||_2 = M$. It is obvious that this exchange does not affect the data observed from the antenna array.

It is evident that when $L > 1$, equation (7) represents a multi-component scenario due to the mixture of the signals at each sensor. Therefore, a quadratic TFD at the individual sensors would contain not only the autoterms of all source signals, but also the interactions of the source signals, causing undesirable crossterms.

For the purpose of subsequent derivation, we first rewrite the noise-free data vector in (7) as

$$\mathbf{y}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{L}\mathbf{a}_i\mathrm{s}_i(t), \tag{10}$$

and its kth element (i.e., the data received at sensor $k, k = 1, \cdots, M$) is given by

$$\mathrm{y}_k(t) = \sum_{i=1}^{L}\mathrm{a}_{ik}\mathrm{s}_i(t). \tag{11}$$

Substituting (11) into (1), we can express the auto-sensor WVD of the signal at the kth sensor, $\mathrm{y}_k(t)$, as

$$\mathrm{W}_{\mathrm{y}_k\mathrm{y}_k}(t,f) = \sum_{i=1}^{L}\sum_{j=1}^{L}\mathrm{a}_{ik}\mathrm{a}_{jk}^*\mathrm{W}_{\mathrm{s}_i\mathrm{s}_j}(t,f), \tag{12}$$

where $\mathrm{W}_{\mathrm{s}_i\mathrm{s}_j}(t,f)$ corresponds to the auto-source or cross-source WVD, depending on whether $i = j$, or $i \neq j$. Averaging the auto-sensor WVDs over the array yields

$$\begin{aligned}\bar{\mathrm{W}}(t,f) &= \frac{1}{M}\sum_{k=1}^{M}\mathrm{W}_{\mathrm{y}_k\mathrm{y}_k}(t,f) \\ &= \sum_{i=1}^{L}\sum_{j=1}^{L}\left(\frac{1}{M}\sum_{j=1}^{L}\mathrm{a}_{ik}\mathrm{a}_{jk}^*\right)\mathrm{W}_{\mathrm{s}_i\mathrm{s}_j}(t,f) \\ &= \sum_{i=1}^{L}\sum_{j=1}^{L}\beta_{ij}\mathrm{W}_{\mathrm{s}_i\mathrm{s}_j}(t,f),\end{aligned} \tag{13}$$

where

$$\beta_{ij} = \frac{1}{M}\sum_{k=1}^{M}\mathrm{a}_{ik}\mathrm{a}_{jk}^* = \frac{1}{M}\mathbf{a}_j^H\mathbf{a}_i \tag{14}$$

is defined as the spatial correlation coefficient.

Equation (13) shows that $\bar{\mathrm{W}}(t,f)$ is a linear combination of the auto-source and cross-source WVDs of all signal arrivals. Since

$$|\beta_{ij}| < 1, i \neq j \quad \text{and} \quad \beta_{ij} = 1, i = j, \tag{15}$$

the constant coefficients in (13) for the auto-source WVDs are always greater than those for the cross-source WVDs. For a large array or widely separated sources, $|\beta_{ij}| << 1$. This property is utilized by the array averaging process and is shown to improve the signal synthesis performance.

Specifically, when all spatial signatures are orthogonal, i.e., $\beta_{ij} = 0$ for any $i \neq j$,

$$\bar{\mathrm{W}}(t,f) = \sum_{k=1}^{L} \mathrm{W}_{\mathrm{s}_k\mathrm{s}_k}(t,f), \tag{16}$$

which is solely the summation of the source signal autoterms. The above equation highlights the fact that all source signal crossterms are entirely eliminated from $\bar{\mathrm{W}}(t,f)$ and only the autoterms are maintained, which is most desirable from the synthesis perspective.

4. ITERATIVE SYNTHESIS PROCESS

Assume that upon implementing the synthesis process described in section 2, we obtain the estimate of the mixing matrix $\hat{\mathbf{A}}$. Since there are interfering signal autoterms from other sources, $\hat{\mathbf{A}}$ is likely to be different from $\mathbf{A}$. We use $\hat{\mathbf{A}}$ to construct a beamformer applied to the data received across the array. Assuming a noise-free scenario,

$$\mathbf{z}(t) = \frac{1}{M}\hat{\mathbf{A}}^H\mathbf{x}(t) = \frac{1}{M}\hat{\mathbf{A}}^H\mathbf{A}\mathbf{s}(t). \tag{17}$$

where $\mathbf{z}(t) = [\mathrm{z}_1(t), \ldots, \mathrm{z}_L(t)]$ is a $L \times 1$ vector. Obviously,

$$\mathrm{z}_k(t) = \left(\frac{1}{M}\hat{\mathbf{a}}_k^H\mathbf{a}_k\right)\mathrm{s}_k(t) + \sum_{l\neq k}^{L}\left(\frac{1}{M}\hat{\mathbf{a}}_k^H\mathbf{a}_l\right)\mathrm{s}_l(t). \tag{18}$$

It is expected that $\hat{\mathbf{a}}_k$ would be a perturbed version of $\mathbf{a}_k$ with the approximations

$$\left|\frac{1}{M}\hat{\mathbf{a}}_k^H\mathbf{a}_k\right| \approx \beta_{kk} = 1 \tag{19}$$

and

$$\left|\frac{1}{M}\hat{\mathbf{a}}_k^H\mathbf{a}_l\right| \approx |\beta_{lk}| \ll 1, l \neq k. \tag{20}$$

From (18)–(20), the WVD of $\mathrm{z}_k(t)$ is given by

$$\begin{aligned}\mathrm{W}_{\mathrm{z}_k\mathrm{z}_k}(t,f) \;&\approx\; \mathrm{W}_{\mathrm{s}_k\mathrm{s}_k}(t,f) \\ &+ \sum_{i=1}^{L}\sum_{(j=1,j\neq i)}^{L}\beta_{ik}\beta_{jk}^*\mathrm{W}_{\mathrm{s}_i\mathrm{s}_j}(t,f).\end{aligned} \tag{21}$$

Clearly, when $j \neq i$, $\left|\beta_{ik}\beta_{jk}^*\right| \ll 1$. This shows that in equation (21), except for the kth auto-source term, all other terms, either auto- or cross-source terms, are significantly reduced in $\mathrm{W}_{\mathrm{z}_k\mathrm{z}_k}(t,f)$. In the case of ULA, the suppression of those terms are at least 13dB for large value of M. The suppression of the autoterms other than source k is $|\beta_{ik}|^2$, which is more than 26dB down from the kth source. Therefore, the effect of the overlapping autoterms from other sources becomes negligible. If we apply the steps (3)–(8) of the synthesis procedures of Section 2 using the improved WVD in (21), the synthesized signal will be significantly enhanced, as shown below.

5. SIMULATION RESULTS

In this section, computer simulations are provided to demonstrate the performance of the proposed technique. We consider two chirp signals with overlapping t-f signatures incident on an eight-sensor ULA ($M = 8$) with inter-element spacing of half-wavelength. The signals arrive at the array with AOAs of $-20°$ and $20°$, with the respective start and end frequencies given by $(0.7\pi, 0.3\pi)$ and $(0.3\pi, 0.7\pi)$, respectively. The length of the signal sequence is set to $N = 128$. There is no additive noise in this example.

Fig. 1 shows the WVD of data at the reference sensor #1. The two signal autoterms overlap, and their cross source terms could also be clearly noticed. The array averaged WVD is plotted in Fig. 2. Using the conclusions derived in Section 3, we expect that the cross-source terms would be suppressed by about 19dB after the array averaging process. Indeed, such suppression is supported by the plots in Fig. 2. To synthesize the signal, we place the mask along each t-f signature. Any reasonable selection of the mask inevitably includes components from the other source. Therefore, each signal synthesized following the procedures described in Section 2 is, in essence, corrupted by the other signal. Fig. 3(b) depicts the WVD of one synthesized but corrupted waveform, compared to the WVD from the original source, which is shown in Fig. 3(a). By implementing the beamformer and synthesis procedures proposed in Section 4, we obtain less noisy waveform. The WVD of the improved synthesized signal is shown in Fig. 3(c). The power leakage in Fig. 3(b) almost disappears in Fig. 3(c).

6. CONCLUSION

We have presented an iterative synthesis process to estimate the signals with overlapping t-f signatures based on the averaging of the Wigner-Ville distributions across an antenna array. By first synthesizing the source temporal waveforms and then using the results to estimate the source spatial signatures, the problem of power leakage that may occur in other conventional Wigner-Ville based synthesis techniques is solved. It is shown that the proposed method provides clear t-f signature and yields improved synthesis performance.

7. REFERENCES

[1] W. Mu, Y. Zhang, and M.G. Amin, "Bilinear signal synthesis in array processing," in *Proc. ICASSP*, Salt Lake City, UT, May 2001.

[2] G. Boudreaux-Bartels and T. Parks, "Time-varying filtering and signal estimation using Wigner distribution synthesis techniques," *IEEE Trans. ASSP*, vol. ASSP-34, pp. 442-451, June 1986.

[3] F. Hlawatsch and W. Krattenthaler, "Bilinear signal synthesis," *IEEE Trans. Signal Processing*, vol. 40, pp. 352-363, Feb. 1992.

[4] J. Jeong and W. Williams, "Time-varying filtering and signal synthesis," in B. Boashash ed., *Time-Frequency Signal Analysis — Methods and Applications*, Longman Cheshire, 1995.

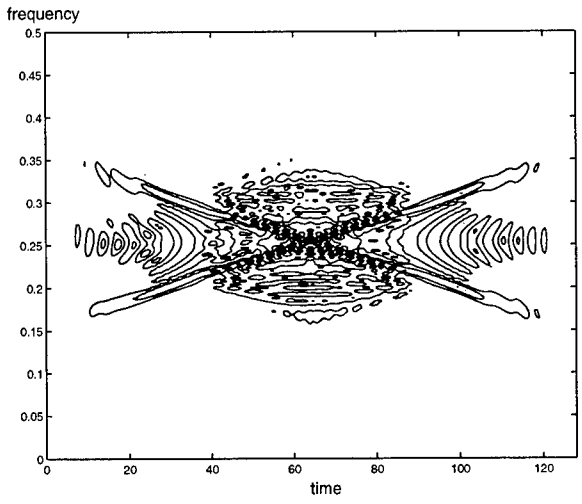


Fig. 1. WVD of the two overlapping signals at a reference sensor.

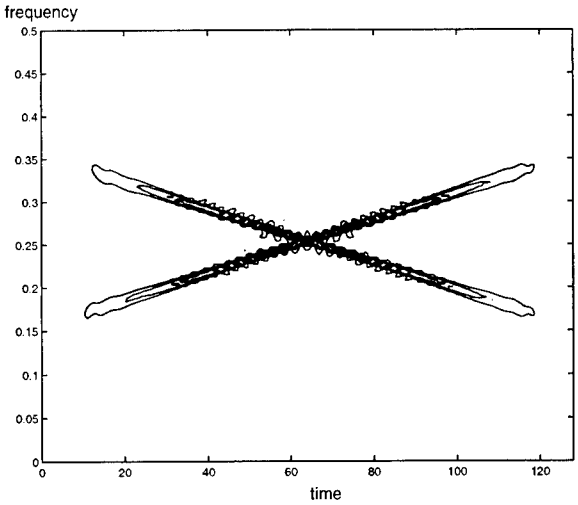


Fig. 2. Array averaged WVD.

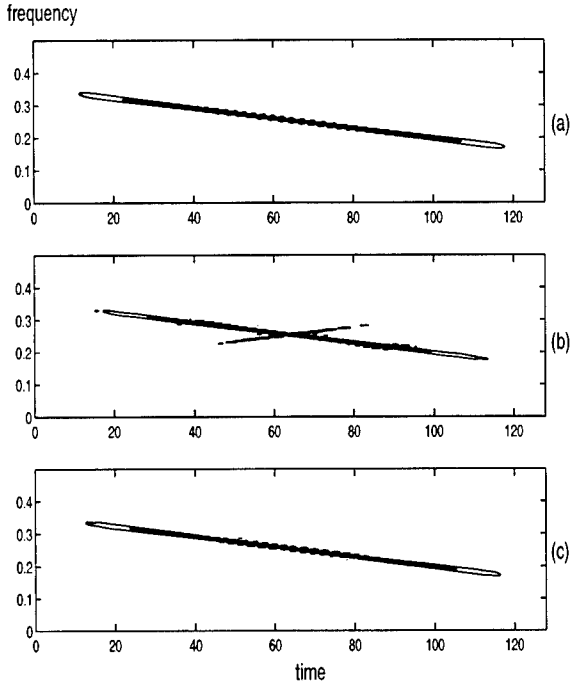


Fig. 3. WVD of: (a) original signal (top); (b) synthesized signal from array averaging (middle); (c) synthesized signal from iterative process (bottom).

FAST COMPUTATION OF DISCRETE SLTF TRANSFORM

Osama A. Ahmed

Center for Communications & Computer Research
Research Institute, KFUPM, Saudi Arabia
osamaa@kfupm.edu.sa

ABSTRACT

In this paper, a fast algorithm for SLTF analysis and SLTF synthesis computations is presented. The proposed algorithm exploits the special structure of the SLTF transformation matrix. The algorithm requires $K\left(2\log_2 M - \frac{1}{2}\right)$ multiplications and $K\left(4\log_2 M - \frac{3}{2}\right)$ additions for calculating the biorthogonal function and $K\left(\log_2 K - \frac{1}{2}\log_2 N\right)$ multiplications and $K\left(2\log_2 K - \log_2 N - 3\right)$ additions for both the analysis and the synthesis transform computations where K is the signal length and M and N are arbitrary numbers such that $MN = K$.

1. INTRODUCTION

Time-frequency (TF) transforms are of interest in many areas due to their natural decomposition of a signal into functions localized in both time and frequency. Specifically, Gabor transform has an optimal localization in the TF domain [1]. SLTF transform maintains the same optimality as Gabor transform [2]. Moreover, it overcomes the two main problems of the critically-sampled TF transforms: stability and localization of the window and its biorthogonal function [2]. Compared to other TF transforms, SLTF has several advantages that makes it suitable for many applications. First, it is a linear critically-sampled transform. This simplifies the synthesis transform procedure (from the TF domain to the original domain) after filtering. This is in contrast to bilinear transforms where difficulties are encountered in retrieving the signal from the TF domain, or the over-sampled transforms where iterative methods are needed for the synthesis transform [3]. Compared to other linear critically-sampled TF transforms, this transform has two major advantages: stability and localization of both the window and its biorthogonal function [2].

The direct computations of the SLTF transform, however, require $\mathcal{O}(K^3)$ operations for the biorthogonal function computations and $\mathcal{O}(K^2)$ for the analysis and the synthesis transform computations where K is the signal length.

The author would like to acknowledge the support of King Fahd University of Petroleum and Minerals.

For Gabor transform, several algorithms have been developed for fast computations with results in the range $\mathcal{O}(K^2)$ to $\mathcal{O}(K\log_2 K)$ [4–6]. In this paper, a fast algorithm for SLTF transform computations is derived which drastically reduces the computation requirements to $\mathcal{O}(K\log_2 M)$ where $M << K$ for both the analysis and synthesis transforms. The proposed algorithm is developed via a matrix approach by exploiting the special structure of the SLTF transformation matrix.

2. SLTF TRANSFORM

The SLTF analysis transform is defined for a finite extent discrete signal $x(k)$, for $0 \le k < K$, as [2]:

$$a_{m,n} = \sum_{k=0}^{K-1} x(k)\,\gamma_m^*(k)\,\mathrm{csin}\,\frac{\pi(k+\frac{1}{2})(n+\frac{1}{2})}{N} \qquad (1a)$$

and the synthesis transform is defined as:

$$x(k) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} a_{m,n}h_m(k)\,\mathrm{csin}\,\frac{\pi(k+\frac{1}{2})(n+\frac{1}{2})}{N} \qquad (1b)$$

where M and N are the number of analysis samples in time and frequency, respectively ($MN = K$), $a_{m,n}$ are the SLTF transform coefficients, csin stands for cos for even m and sin for odd m, and:

$$h_m(k) \;,\; \delta^{-\frac{1}{2}}\exp\left(-\frac{\pi}{2\delta^2}\left(k - mN - \frac{N-1}{2}\right)^2\right)$$

is the normalized discrete Gaussian window shifted to the center of the mth window with δ controlling the window width, and $\gamma_m(k) = \gamma(k-mN)$ where $\gamma(k)$ is the biorthogonal function to $h(k)$, *i.e.*, it satisfies the condition:

$$\sum_{k=0}^{K-1} h_m(k)\,\gamma(k)\,\mathrm{csin}\,\frac{\pi(k+\frac{1}{2})(n+\frac{1}{2})}{N} = \delta_n\delta_m$$

The analysis transform (1a) and the synthesis transform (1b) can be rewritten in matrix form as [2]:

$$\mathsf{a} = \mathsf{E}\mathsf{H}^{-1}\mathsf{x} \quad \text{(analysis equation)} \qquad (2a)$$

$$\mathsf{x} = \mathsf{H}\mathsf{E}\,\mathsf{a} \quad \text{(synthesis equation)} \qquad (2b)$$

where a , $[a_{0,0}, a_{0,1}, \ldots, a_{0,N-1}, a_{1,0}, \ldots, a_{M-1,N-1}]^T$, x , $[x(0), x(1), \ldots, x(K-1)]^T$, and E is an $K \times K$ block-diagonal matrix which is given by:

$$\mathsf{E} \;,\; \mathrm{diag}\,(C, S, -C, -S, \ldots, C, S) \tag{3}$$

where $\mathsf{C} = [c_{n,k}]_{N\times N}$ is the N-point DCT-IV transform matrix with:

$$c_{n,k} = \sqrt{\frac{2}{N}} \cos \frac{\pi\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)}{N}$$

and $\mathsf{S} = [s_{n,k}]_{N\times N}$ is the N-point DST-IV transform matrix with:

$$s_{n,k} = \sqrt{\frac{2}{N}} \sin \frac{\pi\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)}{N}$$

The $K\times K$ matrix H provides the desired windowing effect, i.e., localization in the time domain. H is given by:

$$\mathsf{H} = \begin{bmatrix} H_0 & H_{M-1} & \cdots & H_1 \\ H_1 J & H_0 J & \cdots & H_2 J \\ \vdots & \vdots & \ddots & \vdots \\ H_{M-1} J & H_{M-2} J & \cdots & H_0 J \end{bmatrix} \tag{4}$$

where J is the $N\times N$ row exchange matrix:

$$J = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{pmatrix}_{N\times N} \tag{5}$$

and H_m is an $N\times N$ diagonal matrix defined by:

$$H_m = (-1)^{\left\lfloor \frac{m+1}{2} \right\rfloor} \mathrm{diag}\,(h_m(0), \ldots, h_m(N-1)) \tag{6}$$

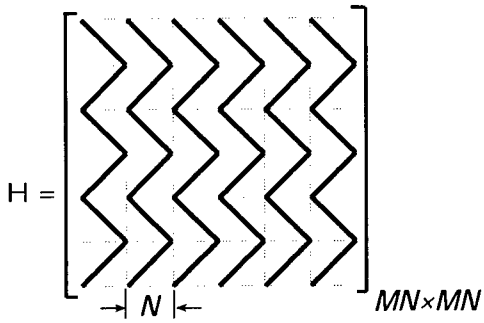
with $\lfloor x \rfloor$ being the integer part of x.

The direct computation of the analysis transform (2a) involves an inversion of the $K \times K$ matrix H which requires $\mathcal{O}(K^3)$ operations and a multiplication of the EH^{-1} by x which requires $\mathcal{O}(K^2)$ operations. Also, direct computation of the synthesis transform coefficients (2b) requires $\mathcal{O}(K^2)$ operations. In the following sections, it will be shown how to perform the transform computations with increased efficiency by exploiting the special structure of H and E.

3. CALCULATION OF BIORTHOGONAL FUNCTION

From (1a) and (2a), the diagonal or anti-diagonal elements from each block in the mth block-row of H^{-1} comprise the biorthogonal function $\gamma_m^*(k)$.

From (4), (5), and (6), H is block matrix with $N\times N$ blocks. These blocks are either diagonal blocks $[H_m]$ or anti-diagonal blocks $[H_m J]$. Therefore, H has non-zero elements only at the block's diagonal for even block-column indices or the block's anti-diagonal for odd block-column indices, *i.e.*, H has the structure:



where the solid lines indicate the only non-zero elements in H. The matrix H can be transformed to a block diagonal matrix with block circulant blocks using row and column permutations as follows. Define the permutation matrices[1] P_1 and P_2 whose encoding vectors p_1 and p_2 are given by:

$$p_1(k) = \left\lfloor \frac{k}{M} \right\rfloor + N\,(k \bmod M)$$

$$p_2(k) = l\,(N-1) + (-1)^l \left\lfloor \frac{k}{M} \right\rfloor + N\,(k \bmod M)$$

for $k = 0 : K-1$, where $l = k \bmod 2$ and $y \bmod x$ is the remainder of y/x.

The matrix $\mathsf{P}_2\mathsf{H}\mathsf{P}_1^T$ is a block diagonal matrix given by:

$$\mathsf{P}_2\mathsf{H}\mathsf{P}_1^T = \mathrm{diag}\,(\mathsf{D}_0, \mathsf{D}_1, \ldots, \mathsf{D}_{N-1}) \tag{7}$$

where D_n is an $M\times M$ block-circulant matrix with each block being a 2×2 matrix $D_{n,m}$

$$\mathsf{D}_n = \begin{pmatrix} D_{n,0} & D_{n,\frac{M}{2}-1} & \cdots & D_{n,1} \\ D_{n,1} & D_{n,0} & \cdots & D_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ D_{n,\frac{M}{2}-1} & D_{n,\frac{M}{2}-2} & \cdots & D_{n,0} \end{pmatrix} \tag{8}$$

Each D_n can be inverted separately. Therefore, H^{-1} is given by[2]:

$$\mathsf{H}^{-1} = \mathsf{P}_1^T\, \mathrm{diag}\,(\mathsf{D}_0^{-1}, \mathsf{D}_1^{-1}, \ldots, \mathsf{D}_{N-1}^{-1})\, \mathsf{P}_2 \tag{9}$$

Also, each D_n can be converted to block-diagonal matrix B_n by:

$$\begin{aligned} \mathsf{B}_n &= \left(E_{\frac{M}{2}} \otimes I_2\right) \cdot \mathsf{D}_n \cdot \left(E_{\frac{M}{2}} \otimes I_2\right)^T \\ &= \mathrm{diag}\left(B_{n,0}, B_{n,1}, \ldots, B_{n,\frac{M}{2}-1}\right) \end{aligned} \tag{10}$$

[1]The permutation matrix P is the identity matrix with rows reordered [7]. The permutation matrix is represented by the encoding vector p whose element p(k) is the column index of the sole "1" in the kth row of P.

[2]Permutation matrices are orthogonal, i.e., $\mathsf{P}^{-1} = \mathsf{P}^T$.

where $\otimes$ is the Kronecker tensor product, I_2 is the 2×2 identity matrix, and $E_{\frac{M}{2}} = [e_{m,k}]_{\frac{M}{2}\times\frac{M}{2}}$ is the $\frac{M}{2}$-point discrete Fourier transform matrix with $e_{m,k} = \frac{1}{\sqrt{M/2}} \exp \frac{-j2\pi nk}{M/2}$ and $B_{n,m}$ is a 2×2 matrix. Equation (10) is a simplified form of Theorem 5.6.4 of [8]. Since B_n is a block-diagonal matrix with 2×2 blocks, its inversion reduces to $\frac{M}{2}$ inversions of 2×2 matrices $B_{n,m}$. From (10), D_n^{-1} is given by:

$$\mathsf{D}_n^{-1} = \left(E_{\frac{M}{2}} \otimes I_2\right)^T \cdot \mathsf{B}_n^{-1} \cdot \left(E_{\frac{M}{2}} \otimes I_2\right) \tag{11}$$

Instead of using (11) to calculate D_n^{-1}, the 2×2 matrices $B_{n,m}$ can be directly calculated using:

$$\begin{bmatrix} B_{n,0} \\ B_{n,1} \\ \vdots \\ B_{n,\frac{M}{2}-1} \end{bmatrix} = \frac{M}{2}\left(E_{\frac{M}{2}} \otimes I_2\right) \begin{bmatrix} D_{n,0} \\ D_{n,1} \\ \vdots \\ D_{n,\frac{M}{2}-1} \end{bmatrix} \tag{12}$$

and D_n^{-1} can be directly obtained as:

$$\mathsf{D}_n^{-1} = \begin{pmatrix} \check{D}_{n,0} & \check{D}_{n,\frac{M}{2}-1} & \cdots & \check{D}_{n,1} \\ \check{D}_{n,1} & \check{D}_{n,0} & \cdots & \check{D}_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ \check{D}_{n,\frac{M}{2}-1} & \check{D}_{n,\frac{M}{2}-2} & \cdots & \check{D}_{n,0} \end{pmatrix} \tag{13}$$

where:

$$\begin{bmatrix} \check{D}_{n,0} \\ \check{D}_{n,1} \\ \vdots \\ \check{D}_{n,\frac{M}{2}-1} \end{bmatrix} = \left(E_{\frac{M}{2}} \otimes I_2\right)^T \begin{bmatrix} B_{n,0}^{-1} \\ B_{n,1}^{-1} \\ \vdots \\ B_{n,\frac{M}{2}-1}^{-1} \end{bmatrix} \tag{14}$$

The H^{-1} computation can be further simplified as follows:

Firstly, to save time and memory requirements, instead of establishing the $K\times K$ matrix $\mathsf{P}_2\mathsf{H}\mathsf{P}_1$, $D_{n,m}$ required in (12) can be obtained directly for any window function $h(k)$ by:

$$D_{n,m} = \begin{bmatrix} h_{2m}(n) & h_{(2m-1) \bmod M}(n) \\ -h_{2m+1}(n1) & h_{2m}(n1) \end{bmatrix} \tag{15}$$

for $m = 0, 1, \ldots, \frac{M}{2} - 1$ and $n = 0, 1, \ldots, N-1$ where $n1 = N - 1 - n$.

Secondly, the $M\times M$ matrix $\left(E_{\frac{M}{2}} \otimes I_2\right)$ can be converted to a block-diagonal matrix with $\frac{M}{2}\times\frac{M}{2}$ blocks using the permutation matrix P_3 whose encoding vector is given by:

$$p_3(m) = \left\lfloor \frac{m}{M/2} \right\rfloor + 2\left(m \bmod \frac{M}{2}\right) \tag{16}$$

for $m = 0, \ldots, M-1$. Using P_3 leads to:

$$\mathsf{P}_3\left(E_{\frac{M}{2}} \otimes I_2\right)\mathsf{P}_3^T = \begin{bmatrix} E_{\frac{M}{2}} & \\ & E_{\frac{M}{2}} \end{bmatrix} \tag{17}$$

Thus, (12) can be written as

$$\begin{bmatrix} B_{n,0} \\ B_{n,1} \\ \vdots \\ B_{n,\frac{M}{2}-1} \end{bmatrix} = \frac{M}{2}\mathsf{P}_3^T \begin{bmatrix} E_{\frac{M}{2}} & \\ & E_{\frac{M}{2}} \end{bmatrix} \mathsf{P}_3 \begin{bmatrix} D_{n,0} \\ D_{n,1} \\ \vdots \\ D_{n,\frac{M}{2}-1} \end{bmatrix} \tag{18}$$

Similarly, (14) can be written as

$$\begin{bmatrix} \check{D}'_{n,0} \\ \check{D}'_{n,1} \\ \vdots \\ \check{D}'_{n,\frac{M}{2}-1} \end{bmatrix} = \mathsf{P}_3 \begin{bmatrix} E_{\frac{M}{2}}^T & \\ & E_{\frac{M}{2}}^T \end{bmatrix} \mathsf{P}_3^T \begin{bmatrix} B_{n,0}^{-1} \\ B_{n,1}^{-1} \\ \vdots \\ B_{n,\frac{M}{2}-1}^{-1} \end{bmatrix} \tag{19}$$

Note that multiplication by the permutation matrices P_1, P_2 and P_3 represents only a change of row or column indices and multiplication by $E_{\frac{M}{2}}$ represents taking the $\frac{M}{2}$-point FFT. Thus, using (15), (18), (19), and (13), calculating D_n^{-1} reduces to four times the $\frac{M}{2}$-point FFT operation to calculate $B_{n,m}$, $\frac{M}{2}$ times the inversion of a 2×2 matrix $B_{n,m}$, and four times the $\frac{M}{2}$-point inverse FFT operation to calculate $\check{D}'_{n,m}$. The M-point FFT takes $\frac{M}{2}\log_2 M$ multiplications and $M\log_2 M$ additions and the inversion of a 2×2 matrix takes 12 multiplications and 5 additions, and the whole inversion process required to calculate D_n^{-1} takes: $2\times 4\left(\frac{M/2}{2}\log_2\frac{M}{2}\right) + 12\frac{M}{2} = M(2\log_2 M + 5)$ multiplications and $M\left(4\log_2 M - \frac{3}{2}\right)$ additions and the whole inversion process of H takes $K(2\log_2 M + 5)$ multiplications and $K\left(4\log_2 M - \frac{3}{2}\right)$ additions.

4. SLTF ANALYSIS TRANSFORM COMPUTATIONS

The computations of the SLTF transform coefficients $a_{m,n}$ can be reduced as follows:

Substituting (9) in (2a) leads to:

$$\mathsf{a} = \mathsf{E}\,\mathsf{P}_1^T \,\mathrm{diag}\left(\mathsf{D}_0^{-1}, \mathsf{D}_1^{-1}, \ldots, \mathsf{D}_{N-1}^{-1}\right)\mathsf{P}_2\,\mathsf{x} \tag{20}$$

From (11), each D_n^{-1} can be replaced by: $\left(E_{\frac{M}{2}} \otimes I_2\right)^T \mathsf{B}_n^{-1}\left(E_{\frac{M}{2}} \otimes I_2\right)$. Substituting (17) in (11) gives:

$$\mathsf{D}_n^{-1} = \mathsf{P}_3^T \begin{bmatrix} E_{\frac{M}{2}}^T & \\ & E_{\frac{M}{2}}^T \end{bmatrix} \mathsf{P}_3 \mathsf{B}_n^{-1} \mathsf{P}_3^T \begin{bmatrix} E_{\frac{M}{2}} & \\ & E_{\frac{M}{2}} \end{bmatrix} \mathsf{P}_3$$

where $\mathsf{B}_n^{-1} = \mathrm{diag}\left(B_{n,0}^{-1}, B_{n,1}^{-1}, \ldots, B_{n,\frac{M}{2}-1}^{-1}\right)$.

Therefore, (20) can be reduced to:

$$\begin{aligned} \mathbf{a} &= \mathbf{E}\,\mathbf{P}_1^T\mathbf{P}_4^T\mathbf{E}_1^T\mathbf{P}_4 \times \\ &\quad \mathrm{diag}\left(B_{0,0}^{-1},\ldots,B_{0,\frac{M}{2}-1}^{-1},B_{1,0}^{-1},\ldots,B_{N-1,\frac{M}{2}-1}^{-1}\right) \times \\ &\quad \mathbf{P}_4^T\mathbf{E}_1\mathbf{P}_4\mathbf{P}_2\,\mathbf{x} \end{aligned} \tag{21}$$

where $\mathbf{P}_4 = \mathrm{diag}\,(\mathbf{P}_3, \mathbf{P}_3, \ldots, \mathbf{P}_3)_{MN\times MN}$, and

$$\mathbf{E}_1 = \mathrm{diag}\left(E_{\frac{M}{2}}, E_{\frac{M}{2}}, \ldots, E_{\frac{M}{2}}\right)_{MN\times MN}$$

Thus equation (21) includes:

- $2N$ times taking the $\frac{M}{2}$-point FFT of the vector $\mathbf{P}_4\mathbf{P}_2\,\mathbf{x}$ which requires $\frac{NM}{2}\log_2 M - NM$ multiplications and $NM\log_2 M - 2NM$ additions;
- $N\frac{M}{2}$ times the multiplication of an 2×2 matrix $B_{n,m}^{-1}$ by an 2×1 vector which requires $2NM$ multiplications and NM additions;
- $2N$ times taking the $\frac{M}{2}$-point inverse FFT which requires $\frac{NM}{2}\log_2 M - NM$ multiplications and $NM\log_2 M - 2NM$ additions;
- M times taking the N-point DCT-IV or DST-IV transforms. Assuming that the N-point DCT-IV or DST-IV transform requires $\frac{N}{2}\log_2 N$ multiplications and $N\log_2 N$ additions, this operation requires $\frac{NM}{2}\log_2 N$ multiplications and $NM\log_2 N$ additions.

Thus, calculating the transform coefficients $a_{m,n}$ requires $NM\log_2 M + \frac{NM}{2}\log_2 N = K\left(\log_2 K - \frac{1}{2}\log_2 N\right)$ multiplications and $K\,(2\log_2 K - \log_2 N - 3)$ additions, *i.e.*, less than $\mathcal{O}(K\log_2 K)$.

5. SLTF SYNTHESIS TRANSFORM COMPUTATIONS

The computations of the synthesized signal $x\,(k)$ can be reduced as follows:

Substituting (7) in (2b) leads to:

$$\mathbf{x} = \mathbf{P}_2^T\,\mathrm{diag}\,(\mathbf{D}_0, \mathbf{D}_1, \ldots, \mathbf{D}_{N-1})\,\mathbf{P}_2\mathbf{E}\mathbf{a} \tag{22}$$

Each $\mathbf{D}_n$ can be replaced by:

$$\mathbf{D}_n = \mathbf{P}_3^T \begin{bmatrix} E_{\frac{M}{2}} & \\ & E_{\frac{M}{2}} \end{bmatrix} \mathbf{P}_3\mathbf{B}_n\mathbf{P}_3^T \begin{bmatrix} E_{\frac{M}{2}}^T & \\ & E_{\frac{M}{2}}^T \end{bmatrix} \mathbf{P}_3$$

where $\mathbf{B}_n = \mathrm{diag}\left(B_{n,0}, B_{n,1}, \ldots, B_{n,\frac{M}{2}-1}\right)$. Therefore, (22) can be reduced to

$$\begin{aligned} \mathbf{x} &= \mathbf{P}_1^T\mathbf{P}_4^T\mathbf{E}_1\mathbf{P}_4 \times \\ &\quad \mathrm{diag}\left(B_{0,0},\ldots,B_{0,\frac{M}{2}-1},B_{1,0},\ldots,B_{N-1,\frac{M}{2}-1}\right) \times \\ &\quad \mathbf{P}_4^T\mathbf{E}_1^T\mathbf{P}_4\mathbf{P}_2\,\mathbf{E}\mathbf{a} \end{aligned} \tag{23}$$

Thus equation (23) includes

- M times taking the N-point DCT-IV or DST-IV transforms;
- $2N$ times taking the $\frac{M}{2}$-point inverse FFT;
- $N\frac{M}{2}$ times multiplication of an 2×2 matrix by an 2×1 vector;
- $2N$ times taking the $\frac{M}{2}$-point FFT.

Thus, calculating the synthesized signal $x\,(k)$ requires: $K\left(\log_2 K - \frac{1}{2}\log_2 N\right)$ multiplications and $K\,(2\log_2 K - \log_2 N - 3)$ additions.

6. CONCLUSION

In conclusion, by exploiting the special structure of the SLTF transformation matrix $\mathbf{E}\mathbf{H}^{-1}$, a very fast algorithm for SLTF computations is derived. The proposed algorithm reduces the complexity to the of order $\mathcal{O}(K\log_2 M)$ instead of the order $\mathcal{O}(K^2)$ for the transform coefficients computations and to the order $\mathcal{O}(K\log_2 M)$ instead of the order $\mathcal{O}(K^3)$ for the biorthogonal function computations

7. REFERENCES

[1] D. Gabor, "Theory of communication", *J. Inst. Elect. Eng.*, vol. 93, pp. 429–459, Nov. 1946.

[2] O. Ahmed and M. Fahmy, "Stable critically sampled Gabor transform with localized biorthogonal function", in *IEEE Int. Sym. Time-Freq. Time-Scale Anal.*, 1998, pp. 37–40.

[3] S. Qian and D. Chen, *Joint Time-Frequency Analysis: Methods and Applications*, Prentice Hall PTR, 1996.

[4] R. S. Orr, "The order of computation for finite discrete Gabor transforms", *IEEE Trans. Signal Proc.*, vol. 41, no. 1, pp. 122–130, Jan. 1993.

[5] R. Balart, "Matrix reformulation of the Gabor transform", *Opt. Eng.*, vol. 31, no. 6, pp. 1235–1242, 1992.

[6] D. F. Stewart, L. C. Potter, and S. C. Ahalt, "Computationally attractive real Gabor transforms", *IEEE Trans. Signal Proc.*, vol. 43, no. 1, pp. 77–83, Jan. 1995.

[7] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, second edition, 1989.

[8] P. J. Davis, *Circulant Matrices*, John Wiley, 1979.

FRACTIONAL-FOURIER-DOMAIN WEIGHTED WIGNER DISTRIBUTION

LJubiša Stanković,[1] *Tatiana Alieva,*[2] *and Martin J. Bastiaans*[2]

[1]University of Montenegro, Podgorica, Montenegro. Email: l.stankovic@ieee.org
[2]Technische Universiteit Eindhoven, Eindhoven, Netherlands. Email: m.j.bastiaans@ieee.org

ABSTRACT

A fractional-Fourier-domain realization of the weighted Wigner distribution (or S-method), producing auto-terms close to the ones in the Wigner distribution itself, but with reduced cross-terms, is presented. The computational cost of this fractional-domain realization is the same as the computational cost of the realizations in the time or the frequency domain, since the short-time Fourier transform of the fractional Fourier transform of a signal corresponds to the short-time Fourier transform of the signal itself, with the window being the fractional Fourier transform of the initial one. The appropriate fractional domain is found from the analysis of the second-order fractional Fourier transform moments. Numerical simulations show a qualitative advantage in the time-frequency representation, when the calculation is done in the optimal fractional domain.

1. INTRODUCTION

Different types of joint time-frequency distributions (TFDs) are nowadays used in signal processing in order to extract the characteristic behavior of a signal. The advantages and disadvantages of most joint representations are well known. Thus, for example, the appropriate short-time Fourier transform (STFT) of multicomponent signals, although not stressing the auto-terms very well, is free from cross-terms, if the components do not overlap. On the other hand, the Wigner distribution (WD) of such signals is highly concentrated, but suffers from cross-terms, which may hide some of the auto-terms. Various distributions, belonging to Cohen's class of TFDs, are defined in order to try to find the optimal representation that will significantly reduce the cross-terms, without significantly changing the auto-terms. Since the commonly used Cohen class TFDs – such as, for example, the Choi-Willams, Bertrand, Butterworth, and Born-Jordan distributions – have been designed for a general signal, they do not correspond to the optimal representation of a particular signal. In order to construct an optimal TFD, the distribution kernel should be adapted to the given signal type [1]-[5]. Moreover, the adaptation has to be made fast and with minimum knowledge about the signal to be analyzed.

It was shown in [6]-[10] that the weighted pseudo WD, or the S-method (SWWD), of a multicomponent signal leads to a representation with significantly reduced cross-terms, while the auto-terms are close to or exactly the same as the ones in the pseudo WD. This signal representation is based on the STFT, and two different forms of it have been proposed [7, 10]. One of them combines the values of the STFT along the frequency axis for a given time instant, while the other is based on calculation in the time direction for a given frequency value. The cross-term reduction and the efficiency of convergence towards the WD auto-terms depend on the orientation of the auto-terms in the time-frequency plane. Thus, if the auto-terms are oriented in parallel to the time (or frequency) axis, then the STFT-based calculations have to be applied in the frequency (or time) domain, correspondingly. In the more general case, however, the auto-terms may lie in some region that might be oriented in a skew direction in the time-frequency plane.

In this paper we introduce the SWWD in the fractional (mixed time-frequency) domain. To this aim, the STFT of the fractional FT of the signal has to be calculated. We will see that the STFT of a signal's fractional FT with a given window, corresponds to the STFT of the signal itself with the window being the fractional FT of the original one, combined with a rotation of the coordinate system. The STFT in the most appropriate fractional domain can thus be performed without significantly more computational costs. After we get the STFT in the optimal fractional domain, the standard, very simple SWWD calculation is performed in that domain. As a result, we obtain a distribution that preserves the WD auto-terms and reduces the cross-terms at the same time. The standard time- or frequency-direction realizations [6]-[10] follow as special cases.

In order to find the fractional domain in which a signal is represented in the simplest way and which matches best to the chosen model, and to find the corresponding STFT window, the analysis of fractional FT moments is applied. In particular, we suppose that an optimal SWWD calculation direction corresponds to minimal signal width, i.e., minimal fractional second-order moment. Determination of this moment can be done analytically, based on three known moments for three different directions. The proposed approach is demonstrated on examples.

2. STFT IN THE FRACTIONAL FT DOMAIN

The STFT was originally introduced for better time localization of frequency components of a signal $f(x)$, by using a suitable, commonly real-valued, window $g(x)$:

$$STF_f^0(t,\omega) = \int_{-\infty}^{\infty} f(t+x)g^*(x)\exp(-j2\pi x\omega)dx. \qquad (1)$$

Clearly, for analyzing a signal with a constant frequency content, one needs a wide window, while for the analysis of pulse-like signals, a narrow window has to be applied. This rule also holds for the analysis of very wide-spread and very narrow signals, respectively. So we can adjust the window, if the signal width is known. Suppose now that the minimal

0-7803-7011-2/01/$10.00 ©2001 IEEE

signal width does not correspond to the time or the frequency domain. Then an affine transformation of the phase plane could lead to an optimal (for example, minimal-width) signal representation. In this paper we restrict ourselves to a rotation of the coordinate system.

To represent a signal in the new coordinate system, we use the fact that a rotation of the time-frequency plane corresponds to a fractional FT of $f(x)$. The fractional FT of a signal $f(x)$ can be defined as [11, 12]

$$R_f^\alpha(u) = \int_{-\infty}^{\infty} K(\alpha, x, u) f(x) dx, \tag{2}$$

where the kernel $K(\alpha, x, u)$ is given by

$$K(\alpha, x, u) = \frac{\exp(j\frac{\alpha}{2})}{\sqrt{j \sin\alpha}} \exp[j\pi \frac{(x^2+u^2)\cos\alpha - 2ux}{\sin\alpha}]. \tag{3}$$

Note that, in particular, $R_f^0(u) = f(u)$, $R_f^\pi(u) = f(-u)$, and that $R_f^{\pi/2}(u)$ corresponds to the normal FT of $f(x)$.

Let us consider the fractional STFT $ST_f^\alpha(t,\omega)$, defined as the STFT of the fractional FT $R_f^\alpha(x)$ of the signal $f(x)$

$$ST_f^\alpha(t,\omega) = \int_{-\infty}^{\infty} R_f^\alpha(t+x) g^*(x) \exp(-j2\pi x\omega) dx. \tag{4}$$

From the symmetry property $R_{g^*}^\alpha(u) = [R_g^{-\alpha}(u)]^*$ of the fractional FT, and from the relations between the fractional FT and the STFT as given in [11, Section IV], we have the relationship

$$\begin{aligned} & \exp(j\pi t\omega) \int_{-\infty}^{\infty} R_f^\alpha(x) g^*(t-x) \exp(-j2\pi x\omega) dx \qquad (5) \\ = \; & \exp(j\pi uv) \int_{-\infty}^{\infty} f(x) [R_g^{-\alpha}(u-x)]^* \exp(-j2\pi xv) dx, \end{aligned}$$

with $u = t\cos\alpha - \omega\sin\alpha$, $v = t\sin\alpha + \omega\cos\alpha$. We thus conclude that calculating the STFT of the signal's fractional FT $R_f^\alpha(x)$ with the window $g(x)$ [cf. (5)], is the same as calculating the STFT of the signal $f(x)$ itself with the window $R_g^{-\alpha}(x)$, combined with a rotation of the coordinate system. Since $R_g^{-\alpha}(x)$, which is the (inverse) fractional FT of the window $g(x)$, can be calculated for all possible angles and stored in a computer memory, this implies that calculation of the fractional STFT $ST_f^\alpha(t,\omega)$ will not be significantly more demanding in numerical complexity than calculation of the standard STFT $ST_f^0(t,\omega)$.

3. FRACTIONAL FT MOMENTS

It is known that the signal width in the time or the frequency domain can be estimated from its second-order central moments. Analogously, the signal width in a fractional domain is related to its second-order central fractional FT moments [13].

The normalized second-order central fractional FT moment p_α is defined by

$$p_\alpha = \frac{1}{E} \int_{-\infty}^{\infty} |R_f^\alpha(x)|^2 (x - m_\alpha)^2 dx = \frac{(w_\alpha - m_\alpha^2)}{E}, \tag{6}$$

where the zero-order moment $E = \int_{-\infty}^{\infty} |R_f^\alpha(x)|^2 dx$ represents the signal's energy (which, in accordance with Parseval's theorem for a unitary transformation, does not depend on α); where the first-order moment $m_\alpha = \int_{-\infty}^{\infty} |R_f^\alpha(x)|^2 x dx$ is related to the center of gravity of the fractional power spectrum; and where $w_\alpha = \int_{-\infty}^{\infty} |R_f^\alpha(x)|^2 x^2 dx$ is the second-order moment. The first-order moment m_α in the fractional α-domain can be calculated from the relationship

$$m_\alpha = m_0 \cos\alpha + m_{\pi/2} \sin\alpha, \tag{7}$$

where m_0 and $m_{\pi/2}$ are the first-order moments in the time and the frequency domain, respectively. Meanwhile, any second-order moment w_α can be obtained from three others w_β, w_γ, and w_μ, say, if the angles β, γ, and μ are different, and the difference between them is not equal to π [13]. Let us choose three second-order moments: w_0, $w_{\pi/2}$, and $w_{\pi/4}$. Then using the results from [13], we have:

$$w_\alpha = w_0 \cos^2\alpha + w_{\pi/2} \sin^2\alpha + [w_{\pi/4} - \tfrac{1}{2}(w_0 + w_{\pi/2})] \sin 2\alpha. \tag{8}$$

Taking into account Eqs. (6), (7), and (8), we conclude that three fractional FT power spectra define all normalized central second-order moments p_α, which characterize the signal widths in the corresponding fractional domains:

$$\begin{aligned} p_\alpha E \;=\; & p_0 \cos^2\alpha + p_{\pi/2} \sin^2\alpha \\ & + [w_{\pi/4} - m_0 m_{\pi/2} - \tfrac{1}{2}(w_0 + w_{\pi/2})] \sin 2\alpha. \end{aligned} \tag{9}$$

In order to find the fractional domain where the signal has an extremal (minimum or maximum) width, we study the behavior of the derivatives of p_α. It is easy to see from Eq. (9) that the first derivative of p_α equals zero for those angles α_e for which

$$\tan 2\alpha_e = \frac{2(w_{\pi/4} - m_0 m_{\pi/2}) - (w_0 + w_{\pi/2})}{p_0 - p_{\pi/2}}. \tag{10}$$

Since the fractional FT is periodic in α with period 2π and satisfies the half-period relation $R_f^{\alpha+\pi}(x) = R_f^\alpha(-x)$, the signal width takes a minimum and a maximum value once over the region $\alpha \in [0, \pi)$. From the behavior of the second derivative of p_α for $\alpha = \alpha_e$, $Ed^2p_\alpha/d\alpha^2\,|_{\alpha=\alpha_e} = 2(p_{\pi/2} - p_0)/\cos 2\alpha_e$, we conclude that the signal reaches its minimum width for that value α_e for which $\cos 2\alpha_e$ has the same sign as $p_{\pi/2} - p_0$; the other value of α_e in the interval $[0, \pi)$ then corresponds to the maximum width. Thus, the appropriate fractional domain where the signal is the best concentrated or most widely spread, can be found from the knowledge of only three fractional power spectra.

4. WEIGHTED WIGNER DISTRIBUTION (S-METHOD)

In the previous section we have discussed how to find the fractional angle corresponding to the minimal or the maximal spread of the signal. In this section we discuss the rotated version of the weighted WD and use the knowledge of the optimal fractional angle to find its optimal realization. The SWWD has been introduced for the analysis of multicomponent signals, with the aim to produce a representation close to the sum of the WDs of each component separately, but with reduced (or even without) cross-terms.

Consider a multicomponent signal $f(x) = \sum_{i=1}^{M} f_i(x)$. Its pseudo WD defined by

$$\begin{aligned} PWD_f(t,\omega) \;=\; & \int_{-\infty}^{\infty} f(t + \tfrac{1}{2}x) f^*(t - \tfrac{1}{2}x) \\ & \times g(\tfrac{1}{2}x) g^*\left(-\tfrac{1}{2}x\right) \exp(-j2\pi x\omega) dx \end{aligned} \tag{11}$$

has the following form: $PWD_f(t,\omega) = \sum_{i=1}^{M} PWD_{fi}(t,\omega) + \sum_{i=1}^{M}\sum_{k=1,k\neq i}^{M} PWD_{f_i}(t,\omega)PWD_{f_k}(t,\omega)$. In most applications, the aim is to get a distribution that contains only the auto-terms, $PWD_f^a(t,\omega) = \sum_{i=1}^{M} PWD_{f_i}(t,\omega)$. It is also known that the WD, among all other quadratic signal-independent time-frequency distributions, has the best auto-term concentration. In most cases the reduction of cross-terms is obtained at the cost of auto-terms degradation.

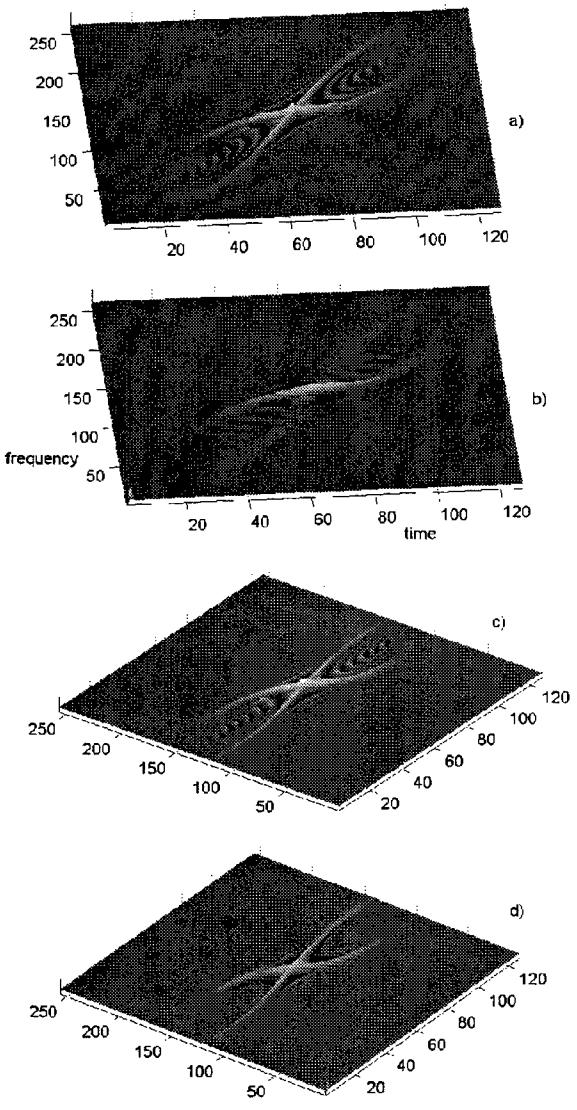


Figure 1: a) Wigner distribution of the signal, b) The SWWD calculated in the frequency direction, c) Rotated Wigner distribution (Wigner distribution of the fractional Fourier transform), d) The SWWD calculated in the "optimal" fractional frequency direction.

The weighted pseudo WD can be written as [6, 10]

$$P_f^0(t,\omega) = \int_{-\infty}^{\infty} ST_f^0(t,\omega+\theta)z(\theta)ST_f^{0*}(t,\omega-\theta)d\theta, \quad (12)$$

where $z(\theta)$ is a weighting function used to exclude the interference pattern between frequency-misaligned versions, while it should be wide enough to provide complete integration over the auto-terms of the STFT $ST_f^0(t,\omega)$. It is easy to see that if $z(\theta) = 1$ we get the pseudo WD (11), whereas for $z(\theta) = \delta(\theta)$ we obtain the time-varying spectrogram. If the width of $z(\theta)$ is somewhere in between, we can expect – as it was proved in [6] and [10] – that the corresponding distribution combines the nice properties of both the spectrogram and the WD. It is known that, unlike the WD, the spectrogram does not suffer from cross-terms. On the other hand, the spectrogram has a significant leakage due to a window usage, which is less exhibited in the case of the WD. By choosing an appropriate function $z(\theta)$, the sharpness of the WD can be preserved while the cross-terms will be reduced or even completely removed. In this case the time window has to be such that the components of the STFT are not far from the instantaneous frequencies of the signal components, in order to obtain fast convergence inside $z(\theta)$.

The SWWD can also be calculated based on time-direction combined STFTs. It is then of the form [7, 10]

$$\int_{-\infty}^{\infty} ST_f^0(t+\theta,\omega)z(\theta)\exp(j4\pi\omega\theta)ST_f^{0*}(t-\theta,\omega)d\theta. \quad (13)$$

Which one of the two forms [(12) or (13)] would produce better results, depends on the signal. If the auto-terms in the STFT are well concentrated along the frequency direction, then (12) would be a better choice, and vice versa.

We have already found that for a given signal there exists a fractional domain where the STFT can be performed in an optimal way, based on the fractional moments. Finding the domain where the signal is best concentrated (minimal second order moment), we can expect that the application of the SWWD in the domain where the direction of best concentration is "the frequency axis" ($\alpha = \alpha_e - \frac{1}{2}\pi$) will be the most efficient one. There, the FT of the signals' fractional FT occupies the narrowest range. The SWWD in this fractional domain, is defined as

$$P_f^\alpha(t,\omega) = \int_{-\infty}^{\infty} ST_f^\alpha(u,v+\theta)z(\theta)ST_f^{\alpha*}(u,v-\theta)d\theta, \quad (14)$$

where $ST_f^\alpha(u,v)$ and u,v are defined by Eqs. (4) and (5), respectively. Using the rotational properties of the STFT [cf. (5)], we could rewrite the definition (14) as

$$P_f^\alpha(t,\omega) = \int_{-\infty}^{\infty} ST_f^0(t+\theta\sin\alpha,\omega+\theta\cos\alpha)z(\theta)$$

$$\times \exp(j4\pi\omega\theta\sin\alpha)ST_f^{0*}(t-\theta\sin\alpha,\omega-\theta\cos\alpha)d\theta, \quad (15)$$

from which it is clear that the SWWD in the fractional domain corresponds to the SWWD calculated simultaneously in the time and the frequency direction. The two cases (12) and (13) follow as special cases from (15) with $\alpha = 0$ and $\alpha = \frac{1}{2}\pi$, respectively.

5. DISCRETE FORM AND EXAMPLE

The analog form (15) suggests that the discrete form of the SWWD in an arbitrary domain could be calculated based on the signal's normal STFT. However, the values of the STFT arguments do not correspond to the discretization

grid, and the STFT values should be calculated by using some interpolation for each time-frequency point and a given α. A much simpler calculation is based on (4) [using (5)] and (14). After the angle α_e – for which the second-order fractional moment is minimal, see Section 3, Eq. (10) – has been determined, the discrete fractional FT $R_f^\alpha(n)$ of the signal (or of the window) for the angle $\alpha = \alpha_e - \frac{1}{2}\pi$, that will result in the SWWD calculation along the "frequency axis" of α domain, is calculated. The discrete STFT reads

$$ST_f^\alpha(n,k) = \sum_{m=-N/2}^{N/2-1} R_f^\alpha(n+m)g^*(m)\exp(-j2\pi mk/N).$$

The discrete form of (14) reads

$$P_f^\alpha(n,k) = \left|ST_f^\alpha(n,k)\right|^2$$

$$+ 2\,\mathrm{Re}\{\sum_{m=1}^{N_z} ST_f^\alpha(n,k+m)ST_f^{\alpha *}(n,k-m)\} \qquad (16)$$

where a rectangular $z(m)$ of the width $2N_z+1$ is assumed. Therefore, the calculation of the SWWD can be understood as calculation of the fractional spectrogram in the domain defined by α, and its improving by terms $2\,\mathrm{Re}\{ST_f^\alpha(n,k+m)ST_f^{\alpha *}(n,k-m)\}$ towards the rotated WD quality of auto-terms, without or with reduced cross-terms. Taking just a few of these fractional spectrogram correcting terms, around the considered time-frequency point, we start immediately improving auto-terms concentration, while the cross-terms will start appearing only when we take the values from another auto-term. If we would take that the width of the window $z(m)$ were $N_z = \frac{1}{2}N$ we would get the rotated WD.

As an example, consider the signal

$$f(t) = \exp[-(3t)^8]\{\exp[j(192\pi t^2 - 8\cos(4\pi t)/\pi)]$$

$$+ \exp[j(64\pi t^2 + 8\cos(4\pi t)/\pi)]\}$$

sampled at $T = 1/256$. A Hanning lag window, with $N_w = 128$ samples, is used for the STFT calculation. The values of the normalized central moments are $p_0 = 1$, $p_{\pi/2} = 1.38$, $p_{\pi/4} = 0.07$. According to (10), and using the fact that $p_0 < p_{\pi/2}$, we get $\alpha_e = 41°$. The second-order moment in this direction is smaller than in any other direction: $p_{41°} = 0.057$. Now the fractional FT of the signal for the angle $\alpha = \alpha_e - \frac{1}{2}\pi = -49^0$ (with $p_{-49°} = 2.01$) can be calculated by using the discrete fractional FT algorithms, or just by using the inversion property of the rotated WD. The next step is to calculate the STFT of the fractional FT and to use it in (16).

The results of this analysis are presented in Fig. 1. The standard WD is shown in Fig. 1a. The SWWD calculated by the standard definition, i.e., along the frequency axis, with $N_z = 10$ correcting terms, is presented in Fig. 1b. We see that some cross-terms already appear, although the auto-terms are still very different from those in the WD in Fig. 1a. The reason lies in the very significant spread of one component along the frequency axis. Fig. 1c shows the WD of the fractional FT for $\alpha = -49°$, obtained as the optimal angle for this signal; note that it is just a rotated version of the original WD. The SWWD based on the fractional FT is presented in Fig. 1d. We can see that, as a consequence of the high concentration of the components along the optimal fractional angle, we almost achieved the goal of getting the auto-terms of the WD without any cross-terms.

Note that if the signal is already well concentrated in time or in frequency, then the proposed procedure will also produce the standard calculation directions, as special cases.

6. CONCLUSION

A generalized form of the weighted WD (or SWWD) is presented. The realization is done in the fractional FT domain with minimal signal width. This domain is optimal with respect to auto-terms convergence and cross-terms suppression in the SWWD. Further research could be directed toward the application of local optimization of the angle α as a time dependent one.

7. REFERENCES

[1] B. Ristic and B. Boashash, "Kernel design for time-frequency signal analysis using the Radon transform," *IEEE Trans. SP*, vol. **41**, pp. 1996 -2008, 1993.

[2] D. L. Jones and R. G. Baraniuk, "An adaptive optimal kernel for time-frequency representation," *IEEE Trans. SP*, vol. **43**, pp. 2361-2371, 1995.

[3] X.-G. Xia, Y. Owechko, B. H. Soffer, and R. M. Matic, "Generalized-marginal time-frequency distributions," *Proc. IEEE-SP Int. Symp. on TFTSA*, Paris, 1996, pp. 509-512.

[4] H. M. Ozaktas, M. A. Kutay, and D. Mendlovic, "Introduction to the fractional Fourier transform and its applications," in *Advances in Imaging and Electron Physics*, vol. **106**, ed. P. W. Hawkes, Academic Press, San Diego, CA, pp. 239-291, 1999.

[5] A. K. Ozdemir and O. Arikan, "A high resolution time frequency representation with significantly reduced cross-terms," *IEEE Proc. ICASSP '00*, vol. **2**, pp. II693-II696.

[6] LJ. Stankovic, "A method for time-frequency signal analysis," *IEEE Trans. SP*, vol. **42**, pp. 225-229, 1994.

[7] LJ. Stankovic, "An analysis of time-frequency representations", *Ann. Telecomm.*, vol. **49**, pp. 505-517, 1994.

[8] B. Boashash and B. Ristić, "Polynomial time-frequency and time-varying higher order spectra: Application to the analysis of multicomponent FM signals and to the treatment of multiplicative noise," *Signal Process.*, vol. **67**, pp. 1-23, 1998.

[9] P. Goncalves and R. G. Baraniuk, "Pseudo affine Wigner distributions: Definition and kernel formulation," *IEEE Trans. SP*, vol. **46**, pp. 1505-1517, 1998.

[10] L. L. Scharf and B. Friedlander, "Toeplitz and Hankel kernels for estimating time-varying spectra of discrete-time random processes," *IEEE Trans. SP*, vol. **49**, pp. 179-189, 2001.

[11] L. B. Almeida, "The fractional Fourier transform and time-frequency representations," *IEEE Trans. SP*, vol. **42**, pp. 3084-3091, 1994

[12] Y. Zhang, B.Y. Gu, B. Z. Dong, G.Z. Yang "A new kind of windowed fractional transforms," *Optics Commun.*, vol. 152, pp. 127-134, 1998.

[13] T. Alieva and M. J. Bastiaans, "On fractional Fourier transform moments," *IEEE SP Lett.*, vol. **7**, pp. 320-323, 2000.

WIGNER DISTRIBUTION RECONSTRUCTION FROM TWO PROJECTIONS

Tatiana Alieva,[1] *Martin J. Bastiaans,*[1] *and LJubiša Stanković*[2]

[1]Technische Universiteit Eindhoven, Eindhoven, Netherlands. Email: m.j.bastiaans@ieee.org
[2]University of Montenegro, Podgorica, Montenegro. Email: l.stankovic@ieee.org

ABSTRACT

The connection between the instantaneous frequency and the angle derivative of the fractional power spectra is established. It permits to solve the signal retrieval problem if only two close fractional power spectra are known. This fact is used in the reconstruction of the Wigner distribution or the pseudo Wigner distribution from two close projections.

1. INTRODUCTION

The reconstruction of a signal – and in particular its phase – from the distributions associated with the instantaneous power of the signal, its power spectrum or, more general, its fractional power spectra, is an important problem in signal processing, radio location, optics, quantum mechanics, etc. In spite of several successful iterative algorithms for phase reconstruction from the squared modulus of the signal and its power spectrum, or its Fresnel spectrum, which were proposed recently [1]-[3], the development of noniterative procedures remains an attractive research topic.

The fractional power spectra, which are the squared moduli of the fractional Fourier transform (FT) [4], are now a popular tool in optics and signal processing [4]-[11]. As it is known, they are equal to the projections of the Wigner distribution (WD) of the signal under consideration [11, 12]. Thus, by using the tomographic approach and the inverse Radon transform, the WD – and therefore the signal itself, up to a constant phase factor – can be reconstructed by knowing all its projections [5, 8]. The method is based on the rotation in the time-frequency plane of the WD under the fractional FT. It demands the measurements of the fractional FT spectra in the wide angular region $[0, \pi)$, which sometimes is impossible or very cost consuming [5].

In this paper we propose a new approach for the WD reconstruction from only two fractional FT spectra, i.e., only two WD projections. This approach significantly reduces the need for projections measurements and calculations. Its is also direct and does not use iterative procedures.

The paper is organized as follows. In Section 2 we present a review of the definition of the fractional FT, and the relationship between the fractional FT power spectra and the ambiguity function of a signal. In Section 3 we establish the connection between the instantaneous frequency in a fractional domain and the angular derivative of the fractional FT power spectra. We show that the instantaneous frequency is determined by the convolution of the angular derivative of the fractional power spectra and the signum function. In Section 4 we discuss the discrete version of the proposed phase retrieval method and demonstrate its efficiency on examples. The importance of the new algorithm and its possible applications are discussed in the Conclusions.

2. FRACTIONAL POWER SPECTRA AND AMBIGUITY FUNCTION

The fractional FT of function $x(t)$, can be written in the form [4]

$$R^{\alpha}[x(t)](u) = X_{\alpha}(u) = \int_{-\infty}^{\infty} K(\alpha, t, u) x(t) dt, \qquad (1)$$

where the kernel $K(\alpha, t, u)$ is given by

$$K(\alpha, t, u) = \frac{\exp(j\frac{1}{2}\alpha)}{\sqrt{j \sin\alpha}} \exp(j\pi \frac{(t^2+u^2)\cos\alpha - 2ut}{\sin\alpha}). \qquad (2)$$

Note that, in particular, $X_0(u) = x(u)$, $X_{\pi}(u) = x(-u)$, and that $X_{\pi/2}(u)$ corresponds to a normal FT. This transform is additive on the parameter α which corresponds to the rotation angle of the coordinate system.

It is known that the fractional power spectra $|X_{\alpha}(u)|^2$, i.e., the squared moduli of the fractional FT, are equal to the projections of the WD $W_x(t, f)$ of the signal $x(t)$,

$$|X_{\alpha}(u)|^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_x(t, f)\delta(t\cos\alpha + f\sin\alpha - u) df dt$$

$$= \int_{-\infty}^{\infty} W_x(u\cos\alpha - f\sin\alpha, u\sin\alpha + f\cos\alpha) df. \qquad (3)$$

The set of fractional power spectra in the angular region $[0, \pi)$ is also called the Radon-Wigner transform.

Since the ambiguity function $A_x(\tau, \nu)$ is the two-dimensional FT of the WD $W_x(t, f)$, the values of the ambiguity function along the line defined by α are – according to the Radon transform properties – equal to the FT of the WD projection for the same α [6, 8],

$$A_x(R\sin\alpha, -R\cos\alpha) = \int_{-\infty}^{\infty} |X_{\alpha}(u)|^2 e^{j2\pi Ru} du. \qquad (4)$$

We can also say that the fractional power spectrum $|X_{\alpha}(u)|^2$ is the FT of the ambiguity function. Note that this relationship is very important for the experimental determination of the ambiguity function in optics, where the fractional power spectra related to intensity distributions can be measured by a simple optical setup [6].

0-7803-7011-2/01/$10.00 ©2001 IEEE

3. WIGNER DISTRIBUTION PROJECTIONS AND INSTANTANEOUS FREQUENCIES

In this section we will derive that the well-known expression for the instantaneous frequency $f_0(t)$ at the position t [13]

$$f_0(t) = \frac{1}{2\pi j}\frac{1}{|x(t)|^2}\int_{-\infty}^{\infty}\left.\frac{\partial A_x(\tau,\nu)}{\partial\tau}\right|_{\tau=0} e^{j2\pi t\nu}d\nu, \quad (5)$$

can be written in terms of the local moments of the fractional power spectra. Indeed, using the relationship [14]

$$\left.\frac{\partial A_x(\tau,\nu)}{\partial\tau}\right|_{\tau=0} = -\frac{1}{\nu}\int_{-\infty}^{\infty}\left.\frac{\partial|X_\alpha(u)|^2}{\partial\alpha}\right|_{\alpha=0} e^{-j2\pi\nu u}du, \quad (6)$$

we get

$$f_0(t) = \frac{-1}{2\,|X_0(t)|^2}\int_{-\infty}^{\infty}\left.\frac{\partial|X_\alpha(u)|^2}{\partial\alpha}\right|_{\alpha=0}\mathrm{sgn}(t-u)du, \quad (7)$$

where $\mathrm{sgn}(t)$ is the signum function:

$$\mathrm{sgn}(t) = \frac{1}{\pi j}\int_{-\infty}^{\infty}\frac{1}{\nu}e^{j2\pi\nu t}d\nu = \begin{cases} 1 \text{ for } t>0, \\ -1 \text{ for } t<0. \end{cases} \quad (8)$$

We thus get for the signal $x(t) = |X_0(t)|\exp[j\varphi(t)]$, that its phase derivative $\varphi'(t) = d\varphi(t)/dt = 2\pi f_0(t)$ is determined by its intensity $|X_0(t)|^2$ and the convolution of the signum function with the angular derivative of the fractional power spectrum $\partial|X_\alpha(u)|^2/\partial\alpha$ at the angle $\alpha = 0$. Note that this relationship can easily be generalized for an arbitrary angle $\alpha \neq 0$, using $|X_\alpha(t)|^2$ and $f_\alpha(t)$ [14] .

In general, the complex-valued fractional FT $X_\alpha(t)$, and in particular the signal $x(t) = X_0(t)$, can be completely reconstructed (except for a constant phase shift) from its intensity distribution $|X_\alpha(t)|^2$ and its instantaneous frequency $f_\alpha(t)$. Since the instantaneous frequency is determined by the derivative of the fractional power spectra, see Eq. (7), this implies that only two fractional power spectra for close angles suffice to solve the signal retrieval problem, up to a constant phase factor. By reconstructing the signal, up to this constant phase factor, from two fractional power spectra (i.e., two WD projections), we can reconstruct the whole WD. Because $x(t)$ is related to $X_\alpha(t)$ through the inverse fractional FT, we can conclude that the signal phase can be reconstructed up to a constant term by a noniterative way from any two fractional power spectra taken for close angles.

Note that this result resembles the so-called transport of intensity equation, which deals with the Fresnel transformation [15]-[17]. This is not surprising since both the fractional FT and the Fresnel transform belong to the class of canonical integral transforms and the properties of any member of this class are related, too.

4. DISCRETIZATION AND EXAMPLES

4.1. Discretization

Here we will illustrate on some numerical examples how the signal, up to a constant phase factor, and its WD can be reconstructed from only two close fractional power spectra, i.e., two WD projections. Of course, instead of reconstructing the WD, we will actually reconstruct the *pseudo* WD, but for an appropriate window function and a small angle α, this will not have any noticeable effect on the final results.

After two WD projections have been obtained or in practice measured as fractional power spectra by using an appropriate optical setup, the instantaneous frequency is calculated as the output of the linear system, cf. Eq. (7),

$$f_0(nT) = -\frac{1}{2\alpha}\frac{|X_\alpha(nT)|^2 - |X_{-\alpha}(nT)|^2 *_n \mathrm{sgn}(nT)}{2\,|X_0(nT)|^2}T \quad (9)$$

where T is the discretization step, the angle α is small, and $*_n$ denotes the discrete-time convolution; moreover, in order to avoid a separate estimation of $|X_0(nT)|^2$, for small α the denominator $2|X_0(nT)|^2$ can be approximated by $|X_\alpha(nT)|^2 + |X_{-\alpha}(nT)|^2$. Note that instead of this symmetrical version of the system, we might as well have chosen an asymmetrical one with $-\alpha$ replaced by 0 and 2α by α.

The signal, up to the constant phase factor, is reconstructed as

$$\hat{x}(nT) = |X_0(nT)|\exp[j\sum_{m=-N}^{N}\varphi'(mT)T] \quad (10)$$

and the (pseudo) WD is calculated according to its definition

$$W_{\hat{x}}(n,k) = 2T\sum_{m=-N}^{N-1}\hat{x}[(n+m)T]\hat{x}^*[(n-m)T] \times w(mT)e^{-j2\pi mk/N}, \quad (11)$$

where $w(nT)$ is an appropriately chosen window function and where N is chosen such that $\hat{x}(nT) \approx 0$ for $|n| \geq N$.

The fractional spectra $|X_\alpha(nT)|^2$ and $|X_{-\alpha}(nT)|^2$ can be obtained in different ways: (i) measured in experiments (a simple optical set up for the measurements of the fractional power spectra was described in [18]); (ii) calculated as squared moduli of the corresponding fractional FT of $x(t)$; (iii) calculated as the Radon transform of the WD of $x(t)$ for two angles $\pm\alpha$.

4.2. Examples

Example 1: We start with the reconstruction of a **monocomponent signal**, whose instantaneous frequency is a monotonic function. Thus a signal of the form

$$x(t) = \exp[-(2.25t)^8] \times \exp\{j\int_{-\infty}^{t}[40\pi\sinh^{-1}(100t) + 256\pi t]dt\} \quad (12)$$

is considered, with $T = 1/1024$. Its (pseudo) WD is calculated, by using a Hanning window $w(t)$ with width $T_w = 1/8$. After the WD has been obtained, we assume that only two of its projections are known, for angles $\alpha = -1°$ and $\alpha = 1°$. The projections are calculated by using the MATLAB radon function, taking the WD matrix as the argument. This corresponds to the case where two fractional power spectra $|X_\alpha(nT)|^2$ and $|X_{-\alpha}(nT)|^2$ are obtained by measurements in an optical system. The procedure described before [cf. Eq. (9)] is then used for the reconstruction

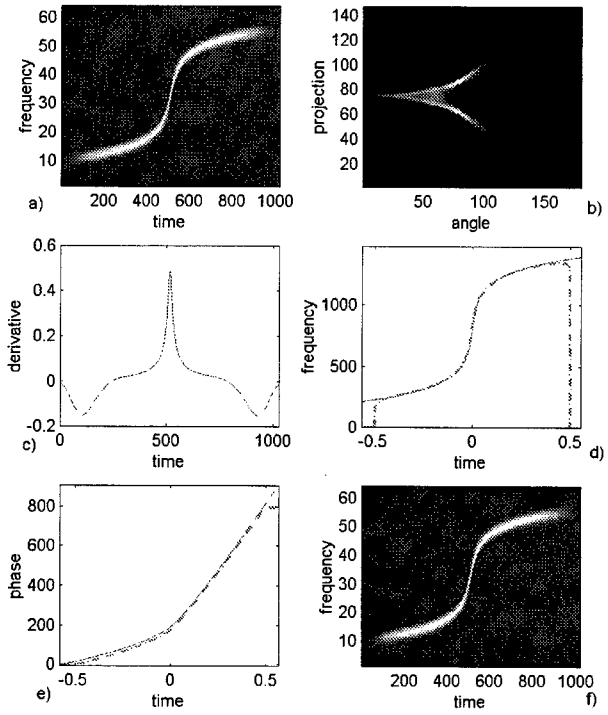


Figure 1: Monocomponent signal and its WD reconstruction from two close fractional power spectra (two WD projections): a) Original WD, b) Projections of the WD (Radon-Wigner transform), c) Derivative approximation: difference of two close projections calculated at $1°$ and $-1°$, and divided by the angle step, d) Reconstructed (dash-dot) and original (solid line) instantaneous frequency of the signal, e) Reconstructed (dash-dot) and original (solid line) phase of the signal, f) Reconstructed WD.

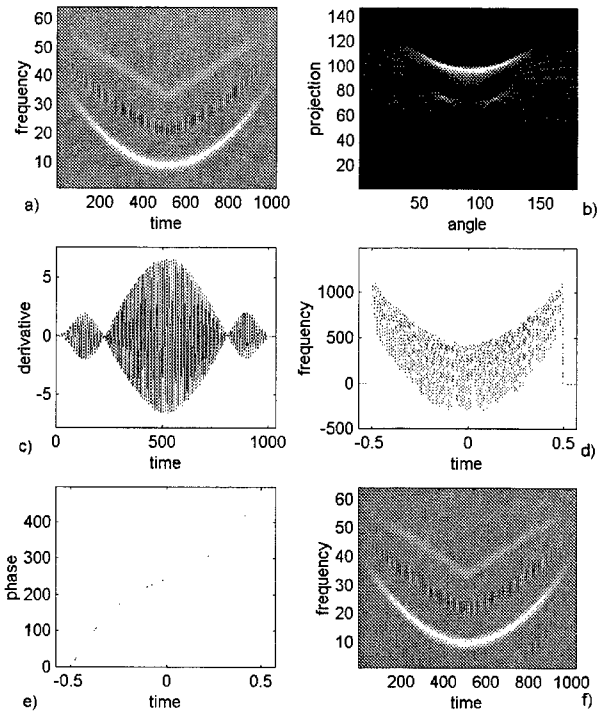


Figure 2: Multicomponent signal and its WD reconstruction from two close fractional power spectra (two WD projections): a) Original WD, b) Projections of the WD (Radon-Wigner transform), c) Derivative approximation: difference of two close projections calculated at $1°$ and $-1°$, and divided by the angle step, d) Reconstructed instantaneous frequency of the signal, e) Reconstructed phase of the signal, f) Reconstructed WD.

of the signal's instantaneous frequency, its phase, and the signal itself [Eq. (10)], from these two projections only.

The original WD is given in Fig. 1a. Its Radon-Wigner transform $|X_\alpha(nT)|^2$ [cf. Eq. (3)] is presented in Fig. 1b, for angles $\alpha \in [0°, 180°)$. Only two projections, for $\alpha = \pm 1°$, are used for further calculations. The difference of these projections, $(|X_\alpha(nT)|^2 - |X_{-\alpha}(nT)|^2)/2\alpha$ for $\alpha = 1°$, is shown in Fig. 1c. The reconstructed instantaneous frequency and the reconstructed phase are given in Fig. 1d and Fig. 1e, respectively, by a dash-dot line, while the original, exact values are represented by solid lines. We can see that the agreement between the reconstructed and the original instantaneous frequency is very high. The phase has a constant shift, as can be expected. The reconstructed WD according to (11) is given in Fig. 1f.

Example 2: The reconstruction of a **multicomponent signal**, having the same amplitude variation as the signal in Example 1, but with a different phase variation,

$$x(t) = \exp[-(2.25t)^8]$$
$$\times\{\exp[j\int_{-\infty}^{t}\omega_1(t)dt] + 0.5\exp[j\int_{-\infty}^{t}\omega_2(t)dt]\}, \quad (13)$$

$$\omega_1(t) = 384\pi\,|t| + 256\pi,\ \omega_2(t) = 1024\pi t^2 + 64\pi,$$

is considered in this example. Note that the instantaneous frequency of this signal is not a continuous function. Nevertheless, we still obtain a satisfactory reconstruction of the phase and the WD, using only two fractional power spectra (see Fig. 2).

Example 3: The reconstruction algorithm is tested for **noisy signals**, as well. The signal from Example 1, contaminated by Gaussian, complex-valued, white noise $\nu(t)$

$$x(t) = \exp[-(2.25t)^8]$$
$$\times\{A\exp[j\int_{-\infty}^{t}(40\pi\sinh^{-1}(100t) + 256\pi t)dt] + \nu(t)\} \quad (14)$$

is considered. Various values of the local signal-to-noise ratio $SNR = 20\log(A/\sigma_\nu)$ have been used in simulations. Figure 3 presents the reconstruction result for a SNR of 9 dB. Small deviations of the reconstructed distribution can be seen in this case. From numerous calculations, we have concluded that the reconstruction threshold is at about $SNR = 3$ dB. Below this value, the degradation of the reconstructed distribution is significant. Nevertheless, it

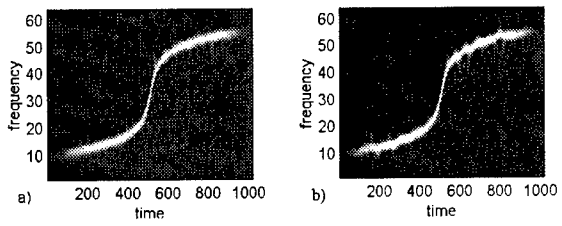


Figure 3: Noisy signal and its WD reconstruction from two close fractional power spectra (two WD projections): a) Original WD, b) Reconstructed WD. Signal to noise ratio is 9 dB.

seems that for signal reconstruction in very high noise, the knowledge of several pairs of close projections would improve the results. In that case we can calculate the differences of the fractional power spectra for different small angles and average them. Furthermore, using other discrete differentiators, different from the simple one given by a mere difference, would also improve noisy case results. However, since the original algorithm produces satisfactory reconstruction even for as low a SNR as a few dB, we have not implemented this variation of the algorithm, for now.

5. CONCLUSIONS

In this paper we have established the connection between the angular derivative of the fractional power spectra and the instantaneous frequency, and we have proposed a method of phase reconstruction from only two close fractional projections of the WD. The numerical simulations show that the discussed phase retrieval algorithm produces good results for different types of signals. The reconstruction technique works well for a signal-to-noise ratio as low as about 3 dB. The main advantages of the proposed method are that it is noniterative and that it demands a minimum number of initial data – only two fractional FT power spectra – which are related to easily measurable intensity distributions. Thus in optics and quantum mechanics, the fractional FT spectrum corresponds to the intensity distribution and probability distribution, respectively.

We have briefly discussed the possible applications of the angular derivatives of the fractional FT power spectra for signal processing, which becomes especially attractive if only the fractional projections of a signal are known.

6. REFERENCES

[1] Z. Zalevsky, D. Mendlovic, and R. G. Dorsch, "Gerchberg-Saxton algorithm applied in the fractional Fourier or the Fresnel domain," *Opt. Lett.*, vol. **21**, pp. 842-844, 1996.

[2] W. X. Cong, N. X. Chen, and B. Y. Gu, "Recursive algorithm for phase retrieval in the fractional Fourier-transform domain," *Appl. Opt.*, vol. **37**, pp. 6906-6910, 1998.

[3] W. X. Cong, N. X. Chen, and B. Y. Gu, "Phase retrieval in the Fresnel transform system - a recursive algorithm," *J. Opt. Soc. Am. A*, vol. **16**, pp. 1827-1830, 1999.

[4] L. B. Almeida, "The fractional Fourier transform and time-frequency representations," *IEEE Trans. Signal Process.*, vol. **42**, pp. 3084-3091, 1994.

[5] M. G. Raymer, M. Beck, and D. F. McAlister, "Complex wave-field reconstruction using phase-space tomography," *Phys. Rev. Lett.*, vol. **72**, pp. 1137-1140, 1994.

[6] J. Tu and S. Tamura, "Analytic relation for recovering the mutual intensity by means of intensity information," *J. Opt. Soc. Am. A*, vol. **15**, pp. 202-206, 1998.

[7] H. M. Ozaktas, N. Erkaya, and M. A. Kutay, "Effect of fractional Fourier transformation on time-frequency distributions belonging to the Cohen class," *IEEE Signal Process. Lett.*, vol. **3**, pp. 40-41, 1996.

[8] X.-G. Xia, Y. Owechko, B. H. Soffer, and R. M. Matic, "Generalized-marginal time-frequency distributions," *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pp. 509-512, 1996.

[9] O. Akay and G. F. Boudreaux-Bartels, "Joint fractional representations," *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pp. 417-420, 1998.

[10] B. Ristic and B. Boashash, "Kernel design for time-frequency signal analysis using the Radon transform," *IEEE Trans. Signal Process.*, vol. **41**, pp. 1996-2008, 1993.

[11] J. C. Wood and D. T. Barry, "Radon transformation of time-frequency distributions for analysis of multicomponent signals," *IEEE Trans. Signal Process.*, vol. **42**, pp. 3166 -3177, 1994.

[12] A. W. Lohmann and B. H. Soffer, "Relationships between the Radon-Wigner and fractional Fourier transforms," *J. Opt. Soc.Am. A*, vol. **11**, pp. 1798-1801, 1994.

[13] F. Boudreaux-Bartels, "Mixed time-frequency signal transformations," in *The Transforms and Applications Handbook*, ed. A. D. Poularikas, CRC Press, Alabama, pp. 887-962, 1996.

[14] T. Alieva and M. J. Bastiaans, "On fractional Fourier transform moments," *IEEE Signal Process. Lett.*, vol. **7**, pp. 320-323, 2000.

[15] M. R. Teague, "Deterministic phase retrieval: a Green function solution," *J. Opt. Soc.Am.*, vol. **73**, pp. 1434-1441, 1983.

[16] N. Streibl, "Phase imaging by the transport equation of intensity," *Opt. Commun.*, vol. **49**, pp. 6-10, 1984.

[17] K. Ichikawa, A. W. Lohmann, and M. Takeda, "Phase retrieval based on the Fourier transport method: experiments," *Appl. Opt.*, vol. **27**, pp. 3433-3436, 1988.

[18] A. W. Lohmann, "Image rotation, Wigner rotation, and the fractional Fourier transform," *J. Opt. Soc. Am. A*, vol. **10**, pp. 2181-2186, 1993.

PREDICTION OF TIME VARYING COMPOSITE SOURCES BY TEMPORAL FUZZY CLUSTERING

S. Policker, A.B. Geva
Electrical and Computer Engineering Department
Ben-Gurion University of the Negev, P.O.B. 653 Beer-Sheva 84105, Israel

ABSTRACT

We present a method for predicting non-stationary signals generated by a time varying composite source. The method is based on the concept of temporal fuzzy clustering. A fuzzy clustering algorithm is applied to the given part (past+present) of the time series and the calculated clusters and membership matrix are then used to estimate a mixture probability distribution function (PDF) underlying the series. In this way a continuous drift in the series distribution expressed as a drift in the clusters' appearance rate can be estimated. A future PDF can then be predicted by fitting a specific model to the estimated past and future PDF values. This also enables the generation of a minimal-mean-squared-error prediction for a future time series element using the estimated mean value of the predicted PDF.

1. INTRODUCTION

Many physical, biological and economical systems produce measurable time series. Analysis of the measured series may enable better understanding of the system and the processes underlying it, prediction of future behavior and detection of meaningful temporal patterns. Tools for time series analysis and prediction were developed for a wide range of applications but most of them share the assumption of stationarity [1]. Methods assuming semi-stationarity such as hidden markov models – HMM may be used in cases where the series can be segmented into stationary periods [2][3]. In many cases, however, there is a continuous change in the probability distribution function (PDF) of the series in which case the semi-stationarity assumption may impose a large error on the prediction result. Also, the series can be composed of stationary segments with long drift periods between them. Detection and prediction in these drift periods can be very important (for example in medical application where early alarm signs or short abnormal periods can have vital clinical importance). Attempts were made to merge two consecutive stationary states to achieve the ability to model such a drift and to use methods based on artificial neural networks for PDF estimaion [4].

We suggest modeling the generator of non-stationary time series by a time varying mixture of stationary or semi-stationary sources. We then combine fuzzy clustering in the observation space and an analysis of the membership matrix on the time scale in order to estimate the model parameters. The given part (past+present) series is clustered as an unindexed set of observations and then we project the resulting membership matrix back to the time scale. The membership matrix is given the interpretation of a continuous temporal change in the weights of a mixture probability distribution function. By estimating the current values of the weights from the membership matrix we can derive an optimal next-step prediction (in the minimal squared error sense) of the series.

In real time applications we also apply the clustering algorithm to a set of given observations to receive an initial condition. In this stage we may also estimate the number of sources by using an unsupervised clustering algorithm and not only their parameters. After receiving this initial state we use a fixed-length moving window of observations for continuous update of the membership matrix and also to recalculate the cluster parameters if they are assumed time varying as well.

2. METHODS

2.1 Model Definition

The output of a composite source [6][7] is a discrete time series generated from N, D dimensional, continuous sub-sources which are sampled by a *random switching function* $\{F(\boldsymbol{\theta}(t) \mid \boldsymbol{\theta}(t) \in [0,1]^N\}$. The input sources are represented by the temporal matrix $\mathbf{X}(t) \in \Re^{N \times D}$ where each one of its N columns containing a random vector process with dimension D originated in a different sub-source $\mathbf{x}_i(t)$. The dimension D can represent a multi-channel feature vector, a temporal sliding window or a combination of both. $\boldsymbol{\theta}(t)$ is a vector of probabilities for selecting a single sub-source in each time sample k_t, to be transmitted to the output:

$$\mathbf{y}(k_t) = \mathbf{x}_i(k_t) \text{ with} \tag{1}$$

and

$$\sum_{i=1}^{N} \theta_i(t) = 1 \quad \text{for all} \quad \infty - < t < \infty \tag{2}$$

The random switch takes a new position each time according to the probability vector $\boldsymbol{\theta}(t)$ and outputs the vector $\mathbf{y}(t) \in \Re^D$ which equals one of the columns of $\mathbf{X}(t)$.

A variety of models for the temporal behavior of $\boldsymbol{\theta}(t)$ can be assumed depending on the specific application or physical phenomenon and thus enabling the description of a wide range of time series.

2.2 Prediction Method

In [8] we presented an algorithm for the estimation of $\boldsymbol{\theta}(t)$ using temporal clustering which will be described briefly.

The clustering space is composed of L sampled points from the time series $\{\mathbf{y}(n) \in \Re^D \mid 1 \le n \le L\}$. Fuzzy partition is defined by a set of N cluster means (prototypical elements) $\{\hat{\mu}_i \in \Re^D \mid 1 \le i \le N\}$ and the membership matrix $\{\mathbf{U} \in [0,1]^{L \times N}\}$ of each element $\mathbf{y}_n$ in each cluster $\mathbf{c}_i$ with prototype μ_i. The clustering procedure that we

0-7803-7011-2/01/$10.00 ©2001 IEEE

are currently using is the hierarchical unsupervised fuzzy clustering (HUFC) algorithm recently presented in [5], which seems to better fit the non-stationary and transient nature of the given time series. After clustering we return to the time axis and divide the series into a set of L/K segments, each including K samples. A moving average of the membership matrices is then used to estimate a sampled version of $\theta(t)$ for each partition:

$$\left\{\hat{\theta}_i^N(k)=\frac{1}{K}\sum_{j=K(k-1)+1}^{K-k} u_{i,j}^N\right\} \quad (3)$$

$$\text{for } 1\le k\le \frac{L}{K};1\le i\le N;N_{min}\le N\le N_{max}$$

Where $u^N_{i,j}\in \mathbf{U}^N$ is the estimated membership of the j-th sample in cluster i of the N-th partition. The result for each partition is a sampled version of the estimated $\theta(k)$ in a sampling rate of f_s/K where f_s is the sampling rate of the given time series. We consider two tasks regarding the prediction of non-stationary time series generated by a time varying composite source. The first task, predicting future values of the series PDF can be performed by using a model for the behavior of $\theta(t)$. The type of the model can be selected using additional information related to the application at hand or by analyzing a long baseline period. The model may describe a deterministic process (for example a linear trend or a periodic one that generates a cyclo-stationary series) or a random one (for example a markov chain that will generate an HMM-like series). This issue is highly dependent on the specific time series or application at hand and will not be addressed here in detail. The second task on which we will focus is predicting the next future element of the series by using the results of temporal clustering i.e. cluster prototypes and temporal regime $\theta(t)$.

Given the time series:

$$\{\mathbf{y}_n\in\Re^D \mid L\ge n\ge 1\} \quad (4)$$

$\boldsymbol{\theta}_n$ can be viewed as a sampled version of the continuous time varying signal $\theta(t)$

If we can forecast a future value for the vector $\boldsymbol{\theta}_{n+\Delta n}$ then, given all prototypical elements for each of the sub-sources, a prediction for the element $\mathbf{y}_{n+\Delta n}$ can be formulated. The optimal predictor in the Minimal Mean Square Error (MMSE) sense is given by:

$$\hat{\mathbf{y}}_{n+\Delta n}=E\left\{\mathbf{y}_{n+\Delta n}\mid \mathbf{y}_1^n\right\}. \quad (5)$$

where

$$\mathbf{y}_1^n\equiv\{\mathbf{y}_1,\ldots,\mathbf{y}_n\} \quad (6)$$

Given that all sub-sources are i.i.d we can calculate all optimal predictors from the estimated means.

$$\left\{\hat{\mathbf{x}}_{1,n+\Delta n},\ldots,\hat{\mathbf{x}}_{N,n+\Delta n}\right\}\equiv \quad (7)$$

$$\equiv\left\{E\{\mathbf{x}_{1,n+\Delta n}\mid \mathbf{y}_1^n\},\ldots,E\{\mathbf{x}_{N,n+\Delta n}\mid \mathbf{y}_1^n\}\right\}\equiv$$

$$\left\{\hat{\boldsymbol{\mu}}_1,\ldots,\hat{\boldsymbol{\mu}}_N\right\}$$

where N is the number of sub-sources in the composite source.

By estimating the future *a priori* probabilities vector (APV) for each sub-source $\boldsymbol{\theta}_{n+\Delta n}$ we can estimate the future value of the time series $\mathbf{y}_{n+\Delta n}$.

We conclude that the MMSE predictor for our model is given by:

$$\hat{\mathbf{y}}_{n+\Delta n}=E\left\{\mathbf{y}_{n+\Delta n}\mid \mathbf{y}_1^n\right\}= \quad (8)$$

$$=\sum_{i=1}^{N} p(\mathbf{y}_{n+\Delta n}=\mathbf{x}_{i,n+\Delta n}\mid \mathbf{y}_1^n)\hat{\boldsymbol{\mu}}_i$$

We recall that

$$\theta_{i,n}\equiv p(\mathbf{y}_n=\mathbf{x}_{i,n}) \quad (9)$$

$$\Theta=\{\theta_{i,n}\mid 1\le i\le N,1\le n\le L\}$$

Where N is the number of sub-sources and L is the number of samples in the series.

The estimation of the APV is given by:

$$\hat{\theta}_{i,k|n}\equiv p(\mathbf{y}_k=\mathbf{x}_{i,k}\mid \mathbf{y}_1^n) \quad (10)$$

For simplicity of presentation we shall use the notation:

$p(\mathbf{x}_{i,n})\equiv p(\mathbf{y}_n=\mathbf{x}_{i,n})$

The sub-sources expectations estimation:

$$\hat{\boldsymbol{\mu}}_i\equiv E\left\{\mathbf{x}_i\mid \mathbf{y}_1^n\right\}=\sum_{j=1}^{n}\hat{\theta}_{i,j}\mathbf{y}_j \text{ and} \quad (11)$$

$$\hat{\Omega}=\{\hat{\boldsymbol{\mu}}_i\mid 1\le i\le N\}$$

We get:

$$\hat{\mathbf{y}}_{n+\Delta n}=E\left\{\mathbf{y}_{n+\Delta n}\mid \hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\Omega}}\right\}=\sum_{i=1}^{N}\hat{\theta}_{i,n+\Delta n|n}\hat{\boldsymbol{\mu}}_i \quad (12)$$

Temporal clustering offers the ability to calculate both $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\theta}}_n$ simultaneously and allows for a flexible trade-off between time and frequency resolution in estimating the temporal change of the APV $\boldsymbol{\theta}_n$

3. RESULTS

We will now present two examples of time series prediction tasks performed by temporal clustering. First, a non-stationary time series with 5000 elements was generated by a random source that was composed of a continuous time varying mixture of 3 Gaussian stationary sub-sources. All sub-sources variances were 10 and the means were 10, 50 and 90. The time varying mixture

that was used for the source was a combination of exponentially and linearly changing expressions with step functions. The 3 mixture weights are presented in the upper 3 traces of figure 1. The resulting time series is presented in the bottom trace. The segments used for train and test are marked on the figure. The train period was used to extract the sources parameters namely the source means by unsupervised fuzzy clustering (results were 88.8, 49.5 and 10.0). The estimated means were used to classify all the time series elements during the test period and produce a prediction for each element from its past window of 30 elements (i.e. for predicting element i we used elements: i-31, i-30, .., i-1). The resulting time series prediction is drawn in the middle trace of figure 2a against the original test period presneted in the upper trace. The prediction error is presneted in the lower trace. Figure 2b presents a small portion of the time series (connected circles) with the corresponding predicted values (connected x).

In figure 3 the predictions for theta are drawn against the original mixture temporal probabilities (3 upper traces) together with the overall classification mean squared error (bottom trace) defined as the weighted average of error in estimating θ_t .

We can see that when there is a low error in the estimation of θ_t then the prediction error is mainly related to the error in estimating the sub-source parameters. See for example the points before n=1000 in figure 2a that have very low classification error. Around n=500 (in figure 2a) when the distance between the confused cluster means is relatively small there is a low prediction error even though there is a substantial classification. When there is a large error in estimation of θ_t then the prediction error term depends on the distribution of the sub-sources means. Around n=750 there is a peak in classification error also resulting in a large variance in the prediction error. In figure 3 we can also see the tracking capability of the algorithm when there are changes in the time series distribution. The tracking speed depends on the length of the segment used to avergae the membership and the sensitivity of the membership to the distance from the cluster mean. In figure 2b we can see the change in the prediction value due to a change in the estimated mixture in a small window of observations.

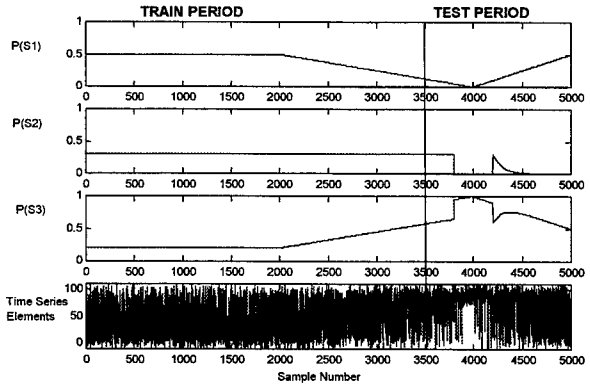


Figure 1 – A simulated example for a time varying composite source produced by 3 gaussian sources with means: 10, 50, 90 and all with variance 10. The three upper figures show temporal mixture a-priori probability. The bottom figure shows the resulting time series. Vertical line separates training and test periods as used for the simulation.

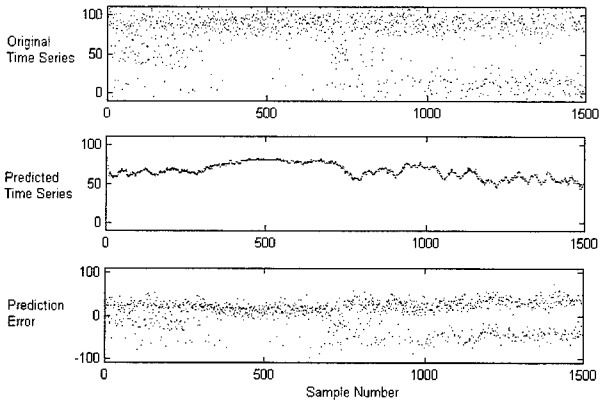


Figure 2a – Original time series (upper figure), predicted time series (middle figure) and prediction error (bottom figure) during the test period.

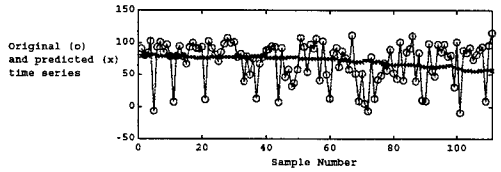


Figure 2b – Original time series (circle trace) and predicted time series (X trace) of a portion of the simulated time series

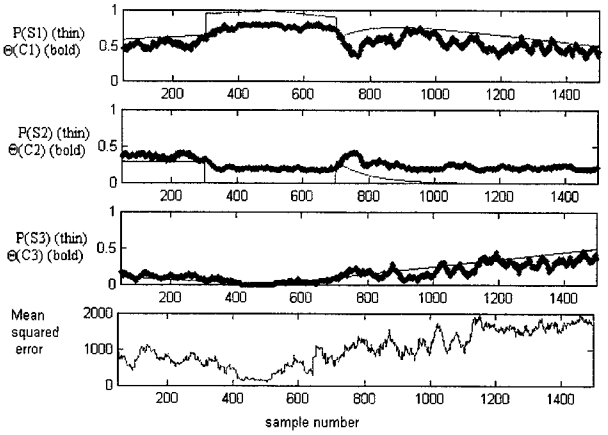


Figure 3 – Original (thin line) and predicted (thick line) temporal mixture probabilities in the 3 upper figures (corresponding to sub-sources with means 10, 50 and 90). Mean squared classification error (bottom figure).

Next, we applied the prediction algorithm to a time series representing the RR intervals (Time intervals between consecutive heart beats) of rats with hyperbaric-oxygen-induced generalized seizures. The rats were implanted with chronic surface cortical electrodes and sub-coetaneous ECG electrodes and exposed to pure oxygen in a pressure chamber. Selected sections of the ECG were digitized at a sampling rate of 1000 Hz. All digitized sections were then analyzed by software that was

developed by the Israeli Naval Medical Institute. All irregular and pathological beats were left to be included in the analyzed series. The final output of the software was an indexed list of consecutive RR-intervals (RRi). The list is converted into a point array in an N-dimensional space, the axes being the durations: $Rri_{(n)}$, $RRi_{(n+1)}$, ..., $RRi_{(n+N-1)}$ (lag plots)

The RR-interval time series extracted from minutes 18-25 at pressure, in a rat that seized after 25 min, is shown in the upper part of figure 5. The middle trace of figure 5 presents the prediction results obtained using the algorithm and the lower trace presents the prediction error. The input for the algorithm was a 3 point window of previous elements (i.e. the prediction for x(i) was obtaind using x(i-3),x(i-2), x(i-1)). The series was fuzzy partitioned to 5 clusters and the resulting membership matrix was averaged for each consecutive 4 points.

The first 18 minuets of the RR series were defined as a train phase and were used to select the best partition of the data to clusters. In the prediction stage we used the number of clusters and cluster prototypes obtained from the train stage and produced the prediction presented in figure 4.

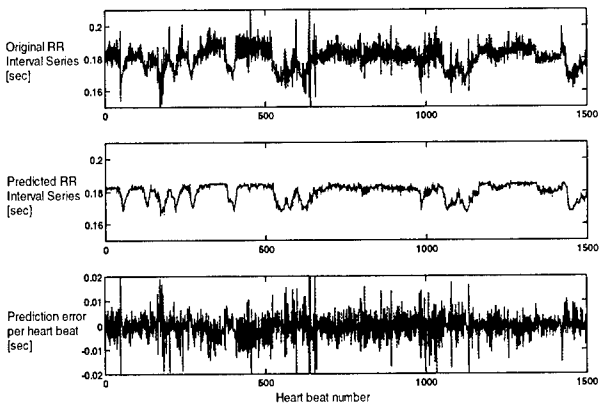


Figure 4 - Measured (upper trace) and predicted (middle trace) RR series of a rat exposed to high oxygen presure. prediction error is drawn in the bottom trace.

4. SUMMARY

The presented method for prediction of non-stationary time series can be used for time series with continuous change in regime that can be described by a time varying mixture PDF. When the non-stationarity of the series is mainly caused by changes in the PDF of the sub-sources and not in the mixture (for example a single gaussian noise source with a drift in its mean value) then the method is expected to be less adequate for prediction.

The suggested prediction procedure can also be used as a non-linear filtering scheme that may be especially useful in applications where clusters of outliers are expected to appear with some probability and add a non-stationary and non-linear aspect to an otherwise stationary linear baseline as in the RR example.

The prediction of elements farther than the adjacent future demands more assumptions regarding the process underlying $\hat{\theta}_n$ and should be treated with a specific application in mind.

The algorithm can be enhanced by performing re-estimation of cluster prototypes in a moving window (longer than the one used to calculate $\hat{\theta}_n$). This may be unavoidable for a long series and can also be used to compensate for a drift in the sub-sources means.

We conclude by stating that our future goal is to investigate temporal clustering as a tool for other signal processing tasks with comparison to common signal processing methods..

5. REFERENCES

[1] J. D. Hamilton. *Time Series Analysis*, Princeton University Press, 1994

[2] A. S. Weigend and N. A. Gershenfeld Ed. *Time series prediction, forcasting the future and understanding the past* . Perseus books 1994

[3] J. R. Deller, J. G. Proakis and J. H. L. Hansen. *Discrete-time processing of speech signals*. Prentice-Hall, 1987.

[4] J. Kohlmorgen, K.R. Muller, J. Rittweger, K. pawelzik, "Identification of nonstationary dynamics in physiological recordings" *Biological Cybernetics* 83 (1), 73-84, Springer Berlin Heidelberg

[5] I. Gath and A.B. Geva, Unsupervised optimal fuzzy clustering. *IEEE Trans. On Pattern Anal. Machine Intell.*, Vol 7, pp. 773-781, 1989.

[6] T. Berger, *Rate Distortion Theory*. Prentice-Hall, 1971.

[7] Y. Ephraim and N. Merhav, "Lower and Upper Bounds on the Minimum Mean-Square Error in Composite Source Signal Estimation", IEEE trans. on information theory, Vol. 38, No. 6, 1992

[8] S. Policker and A. B. Geva, "Non-Stationary Time Series Analysis by Temporal Clustering", IEEE trans. on system, man and cybernetics B. 2000

[9] A.B. Geva and D.H. Kerem,. "Forecasting Generalized Epileptic Seizures from the EEG Signal by Wavelet Analysis and Dynamic Unsupervised Fuzzy Clustering", *IEEE Trans. on Biomedical Engineering*, Vol. 45, No. 10, pp. 1205-1216, October 1998..

ON THE USE OF A NEW COMPACT SUPPORT KERNEL IN TIME FREQUENCY ANALYSIS

Adel Belouchrani and Mohamed Cheriet ***

* Electrical Engineering Department
Ecole Nationale Polytechnique
10 Avenue Hassan Badi, B.O. 182, 16200 El-Harrach, Algiers, Algeria.
E-mail: belouchrani@hotmail.com
** Imagery, Vision and Artificial Intelligence Laboratory
Ecole de technologie supérieure
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3
Email: cheriet@gpa.etsmtl.ca

ABSTRACT

This contribution introduces a new Time Frequency Distribution based on a kernel with compact support. The properties of the new distribution are emphasized. Through a parameter that controls the kernel width, the new representation allows a tradeoff between a good autoterm resolution and a high crossterm rejection. A signal representation example is provided and compared with respect to the Wigner distribution, the Choi-Williams distribution, the spectrogram and the Born-Jordan distribution as a member of the Reduced Interference Distribution class.

1. INTRODUCTION

Time Frequency distributions are more and more widely used for nonstationary signal analysis. They perform a mapping of one-dimensional signal $x(t)$ into a two dimensional function of time and frequency $TFD_x(t,f)$. Herein, we are interested by the Cohen's Class Distributions. The general expression of a member of this class is given by [1],

$$TFD_x(t,f) = \int\int\int \phi(\eta,\tau)x(t'+\frac{\tau}{2})x^H(t'-\frac{\tau}{2}) e^{-j2\pi\eta t}e^{-j2\pi\tau f}e^{j2\pi\eta t'}dt'd\tau d\eta \quad (1)$$

where t and f represent time and frequency, respectively, and ^{H} the transpose conjugate operator. The kernel $\phi(\eta,\tau)$ determines the main properties of the resulting Time Frequency Distribution (TFD). Many authors [2, 1] start from the Cohen's class of distributions to define kernels whose main property is to reduce the interference patterns induced by the distribution itself. In this contribution [3], we propose to use a new kernel derived from the Gaussian kernel [4]. Unlike the Gaussian kernel, the new kernel has the compact support analytical property, i.e., it vanishes itself outside a given compact set. Hence, It recovers the information loss that occurs for the Gaussian kernel due to truncating and improves the processing time. Moreover, the Compact support kernel keeps the most important properties of the Gaussian kernel. This compact support property is different from the finite support property of time frequency representations. It turns out that through a control parameter the new time frequency representation allows a trade of between a good auto term resolution and a high cross term rejection. In the next section, the expression of this new kernel is given together with the properties of the induced time frequency distribution. Finally a signal representation example of two crossing chirps is given and compared with respect to the Wigner distribution [1], Choi-Williams distribution [2], the spectrogram and the Born-Jordan distribution as a member of the Reduced Interference Distribution class [5].

2. THE NEW KERNEL

The new kernel is derived from the Gaussian kernel by transforming the $\mathbb{R}^2$ space into a unit ball through a change of variables. This transformation packs all the information in the unit ball. With the new variables, the Gaussian is defined on the unit ball and vanishes on the unit sphere. Then, it is extended over all the $\mathbb{R}^2$ space by taking zero values outside the unit ball. The obtained kernel still belongs to the space of functions with derivatives of any order. The new kernel refereed

0-7803-7011-2/01/$10.00 ©2001 IEEE

to as Compact Support Kernel (CSK) has the following expression [4],

$$\phi(\eta,\tau) = \begin{cases} \epsilon^{\frac{1}{2}(\frac{\gamma}{\eta^2+\tau^2-1}+\gamma)} & \eta^2+\tau^2 < 1 \\ 0 & \text{elsewhere} \end{cases} \quad (2)$$

where γ is a parameter that controls the kernel width. Figure 1 shows the Compact Support Kernel (CSK) with $\gamma = 5.5$.

2.1. Features of the new kernel

Let us first recall two practical limitations of the Gaussian kernel: information loss due to diminished accuracy when the Gaussian is cut off to compute the time frequency distribution, and the prohibitive processing time due to the mask's width which is increased to minimize the accuracy loss. The main features of the new kernel is that it recover the above information loss and improves processing time and retains the most important properties of the Gaussian kernel [4]. These features are achieved thanks to the compact support analytical property of the new kernel. This compact support property means that the kernel vanishes itself outside a given compact set.

3. THE NEW TIME FREQUENCY DISTRIBUTION

The resultant distribution from the above compact support kernel (2) does not satisfy the marginal property just like the spectrogram. It is consistent with the energy conservation ($\phi(0,0) = 1$) and verifies both the reality and the time and frequency shift properties.

The waveform of any kernel $\phi(\eta,\tau)$ determines the autoterm resolution and the cross term reduction of a time frequency distribution. Note that there is a tradeoff between the autoterm resolution and the interference suppression [5]. More the kernel width is wide, more the resultant distribution suffers from interference while maintaining good auto term resolution. On the other hand, more the kernel width is narrow, better is the interference term suppression at the expense of the autoterm resolution.

The new kernel (2) has by definition a limited width extend since it has a compact support. In this extend, its width can be controlled through the parameter γ to allow a tradeoff between a good autoterm resolution and a sufficient cross term suppression.

Compared to the Reduced Interference Distribution (RID), the resultant distribution from the new kernel does not satisfy all the distribution properties [5]. However, in contrast to these RIDs, it provides a good tradeoff between autoterm resolution and cross term suppression.

Note that a new RID kernel can be derived from the kernel of (2) following the design procedure proposed in [5].

4. EXPERIMENTAL RESULTS

In order to compare the performances of the new distribution, we consider Four typical time frequency distributions (Wigner Distribution, Choi-Williams Distribution, Spectrogram and Born-Jordan Distribution as member of the RID class [5]). In this section, an example of two crossing chirps is shown which clearly reveals the differences in performance among the five distributions.

In figures 2, 3, 4, 5 and 6, the time frequency representation of the two crossing chirps is plotted using the CSK TFD, the Wigner TFD, the Choi-Williams TFD, the spectrogram and the Born-Jordan TFD (as a member of the RID class), respectively. The new time frequency representation shows its ability to remove the cross terms and presents cute curves in contrast to the other representations.

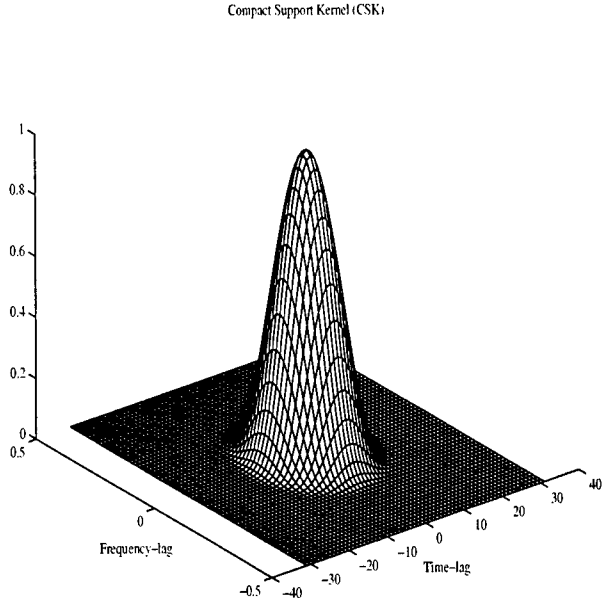


Figure 1: The Compact Support Kernel.

5. REFERENCES

[1] L. Cohen, *Time-frequency analysis.* Prentice Hall, 1995.

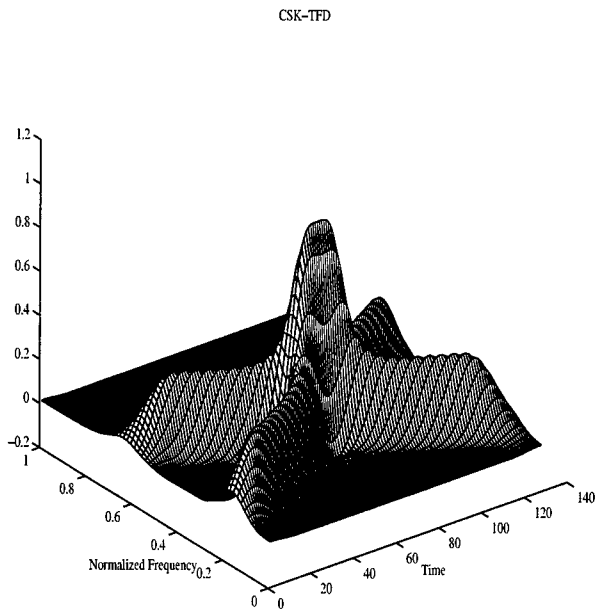


Figure 2: The CSK TFD of the two crossing chirps.

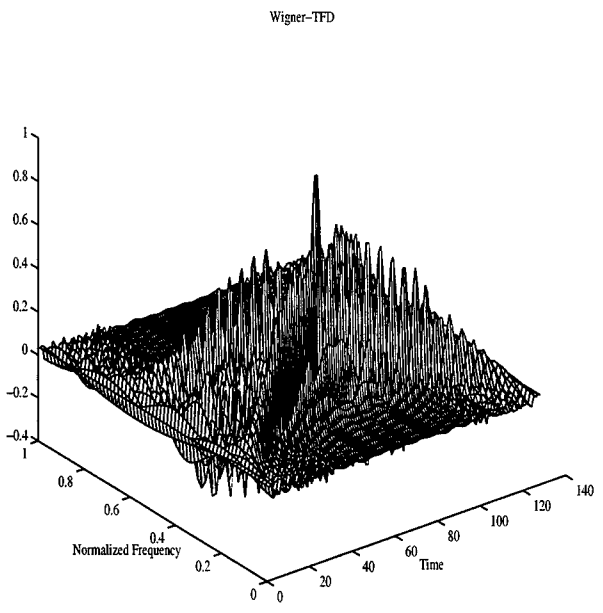


Figure 3: The Wigner TFD of the two crossing chirps.

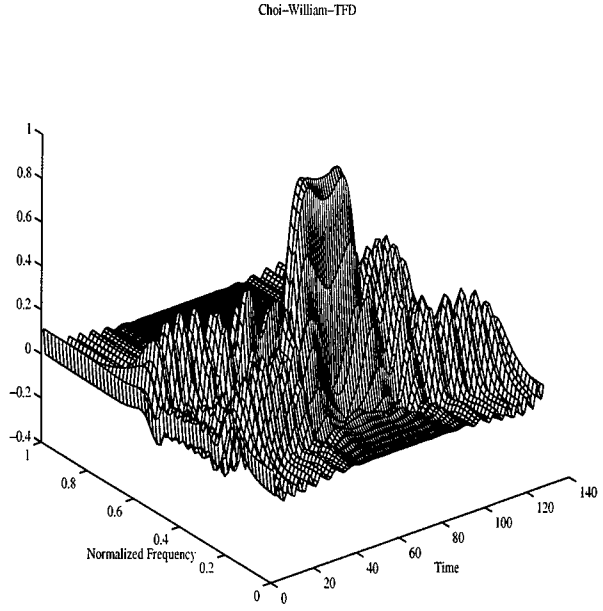


Figure 4: The Choi-Williams TFD of the two crossing chirps.

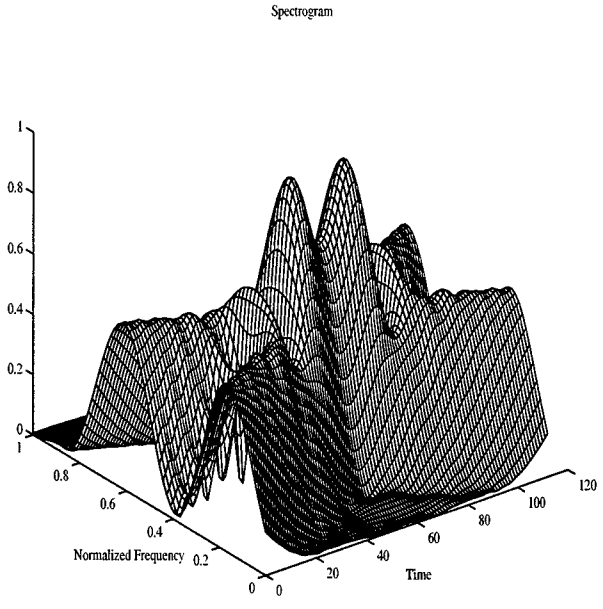


Figure 5: The spectrogram of the two crossing chirps.

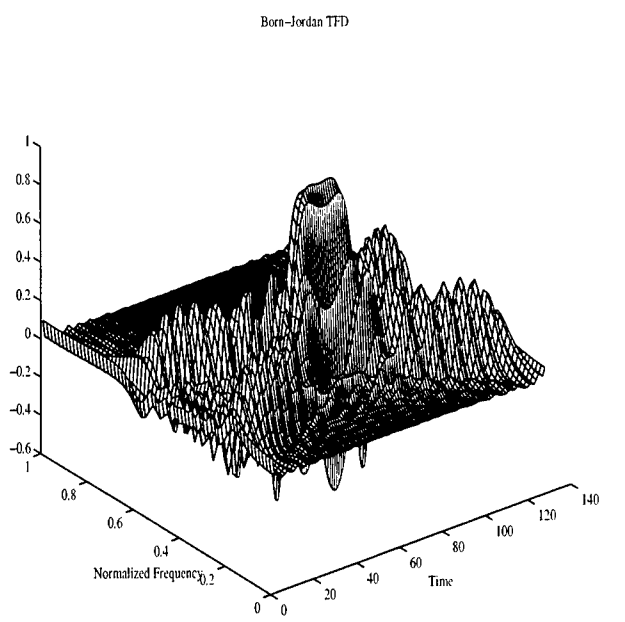


Figure 6: The Born-Jordan TFD (as member of RID) of the two crossing chirps.

[2] H. Choi and W. J. Williams, "Improved time-frequency representation of multicomponent signal kernels," *IEEE Trans. Acoust. Speech. Signal. Proc.*, 7, pp. 862–871, 1989.

[3] M. Cheriet and A. Belouchrani, "Méthode de mesure et d'analyse de l'energie d'un signal en fonction du temps et de la fréquence avec une tres bonne réduction des fausses manifestations d'énergie et une bonne résolution temps fréquence utilisant un filtre à support compact" *Patent pending*, ETS VAL-11, May 2001.

[4] L. Remaki and M. Cheriet, "KCS - New Kernel Family with Compact Support in Scale Space: Formulation and Impact," *IEEE T-PAMI*, 9(6), pp. 970–982, June 2000.

[5] J. Jeong and W. J. Williams, "Kernel Design for reduced interference distributions," *IEEE Trans. on Signal Processing*, vol. 40, No. 2, pp. 402–412, Feb. 1992.

A WEIGHTED DECOMPOSITION OF THE WIGNER DISTRIBUTION

Junfeng Wang, Xiang Yan, Antonio H. Costa, and Dayalan Kasilingam

Department of Electrical and Computer Engineering
University of Massachusetts Dartmouth
North Dartmouth, Massachusetts 02747-2300, USA
acosta@umassd.edu, dkasilingam@umassd.edu

ABSTRACT

The Wigner distribution (WD) can be decomposed into a linear combination of elementary WDs. Slow-oscillatory elementary WDs and fast-oscillatory elementary WDs mainly contribute to auto-terms and cross-terms, respectively. Using a weight function to keep slow-oscillatory elementary WDs and attenuate fast-oscillatory elementary WDs, one can balance auto-term resolution and cross-term suppression and obtain a weighted Wigner distribution (WWD).

1. INTRODUCTION

Time-frequency representations (TFRs) describe the variation of signals simultaneously in time and frequency. The short-time Fourier transform (STFT) [1,2], a typical linear TFR, finds wide use in practice. Its performance, however, is restricted by the tradeoff between its time resolution and frequency resolution. The 2-norm of the STFT is known as the spectrogram [2].

Quadratic TFRs, also called time-frequency distributions, are loosely interpreted as two-dimensional (2-D) signal energy densities in the time-frequency domain. Cohen's class of shift covariant TFRs includes all quadratic TFRs that satisfy the time-shift and frequency-shift covariance properties [1,2]. The spectrogram and the Wigner distribution (WD), prominent members of Cohen's class, can be viewed as "opposites." The WD portrays optimal resolution of auto-terms and no cross-term attenuation whereas the spectrogram portrays poor auto-term resolution and massive reduction of cross-terms [1,2]. The cross-terms in the WD greatly restrict its practical use [3].

In the WD, cross-terms oscillate and auto-terms vary slowly. Hence, cross-terms can be suppressed by convolving the WD with a 2-D low-pass, fixed kernel (or filter) [1,2]. The resulting TFR is a member of Cohen's class and corresponds to a smoothed WD [2]. The cost of attenuating cross-terms, however, usually comes at the expense of auto-term resolution [3].

The time-frequency distribution series (TFDS) of [4] attenuates cross-terms differently. The WD is decomposed into a linear combination of elementary WDs. Slow-oscillatory elementary WDs and fast-oscillatory elementary WDs mainly contribute to auto-terms and cross-terms, respectively. In the TFDS, auto-term resolution and cross-term suppression is balanced by keeping slow-oscillatory elementary WDs and discarding fast-oscillatory elementary WDs.

Our work, inspired by the TFDS idea, makes progress in: (1) deriving a continuous-parameter decomposition of the WD; (2) generalizing the weight function of the TFDS; and (3) presenting an approach for choosing the sampling intervals of the parameters.

2. A WEIGHTED DECOMPOSITION OF THE WD

The STFT of a signal x(t) is defined as[1]

$$X^{(h)}(t,f)=\int x(\alpha)h^{*}(\alpha-t)\exp(-j2\pi f\alpha)\,d\alpha \qquad (1)$$

where h(t) is an analysis window [2]. x(t) can be recovered from its STFT by

$$x(t)=\int\int X^{(h)}(\alpha,\theta)g(t-\alpha)\exp(j2\pi\theta t)\,d\alpha d\theta \qquad (2)$$

where g(t) is a synthesis window satisfying [2]

$$\int g(t)h^{*}(t)\,dt=1. \qquad (3)$$

The cross-WD of two signals x(t) and y(t) is defined by [2]

$$W_{x,y}(t,f)=\int x\left(t+\frac{\tau}{2}\right)y^{*}\left(t-\frac{\tau}{2}\right)\exp(-j2\pi f\tau)\,d\tau. \qquad (4)$$

Taking the WD of (2), we get

$$W_x(t,f)=\int\int\int\int X^{(h)}(\alpha,\theta)X^{(h)*}(\beta,\upsilon)\times \times W_{\tilde{g},\hat{g}}(t,f,\alpha,\theta,\beta,\upsilon)\,d\alpha d\theta d\beta d\upsilon \qquad (5)$$

where $W_{\tilde{g},\hat{g}}(t,f,\alpha,\theta,\beta,\upsilon)$ is the cross-WD of

$$\tilde{g}(t)=g(t-\alpha)\exp(j2\pi\theta t) \qquad (6)$$

and

$$\hat{g}(t)=g(t-\beta)\exp(j2\pi\upsilon t). \qquad (7)$$

Eq. (5) shows that the WD can be decomposed into a linear combination of elementary WDs. Approximating the integrations by summations in (5), we obtain

$$W_x(t,f)=T^2F^2\sum_n\sum_k\sum_p\sum_m X^{(h)}(nT,kF)X^{(h)*}(pT,mF)\times \times W_{\tilde{g},\hat{g}}(t,f,nT,kF,pT,mF). \qquad (8)$$

Eq. (8) is essentially the decomposition of the WD in [4].

$W_{\tilde{g},\hat{g}}(t,f,\alpha,\theta,\beta,\upsilon)$ can be expressed as

$$W_{\tilde{g},\hat{g}}(t,f,\alpha,\theta,\beta,\upsilon)=W_g\left(t-\frac{\alpha+\beta}{2},f-\frac{\theta+\upsilon}{2}\right)\times \times\exp\{j\pi[(\theta+\upsilon)(\alpha+\beta)]+j2\pi[(\theta-\upsilon)t+(\beta-\alpha)f]\}. \qquad (9)$$

Eq. (9) shows that an elementary WD is halfway between the corresponding elementary signals, has the same envelope as $W_g(t,f)$, and oscillates at the "frequency" of $\theta-\upsilon$ in time and at the "frequency" of $\alpha-\beta$ in frequency.

In the WD, auto-terms vary slowly and cross-terms oscillate. Hence, in (5) slow-oscillatory elementary WDs and fast-oscillatory elementary WDs mainly contribute to auto-terms and cross-terms, respectively. Using a weight function to preserve

[1] Unless otherwise noted, all integrations are from $-\infty$ to ∞.

0-7803-7011-2/01/$10.00 ©2001 IEEE

slow-oscillatory elementary WDs and attenuate fast-oscillatory elementary WDs in (5), auto-term resolution and cross-term suppression can be balanced, i.e.,

$$\tilde{W}_x(t,f) = \int\int\int\int X^{(h)}(\alpha,\theta)X^{(h)*}(\beta,\upsilon)\times \\ \times\lambda(\alpha,\theta,\beta,\upsilon)W_{\tilde{g},\tilde{g}}(t,f,\alpha,\theta,\beta,\upsilon)d\alpha d\theta d\beta d\upsilon . \quad (10)$$

We call $\tilde{W}_x(t,f)$ in (10) the weighted Wigner distribution (WWD). The weight function $\lambda(\alpha,\theta,\beta,\upsilon)$ is generally even in both α-β and θ-υ. It usually tends to decrease with increasing $|\alpha-\beta|$ and $|\theta-\upsilon|$. Slower-decreasing $\lambda(\alpha,\theta,\beta,\upsilon)$ usually means higher auto-term resolution and lower cross-term reduction. The WD can be obtained from (10) by letting $\lambda(\alpha,\theta,\beta,\upsilon)=1$.

If the integrations in (10) are approximated by summations and the weight function is chosen as

$$\lambda(n,k,p,m) = 1 \quad ,|n-p|+|k-m| \le d , \quad (11)$$

we get

$$\tilde{W}_x(t,f) = T^2F^2\sum\sum_{|n-p|+|k-m|\le d}\sum\sum X^{(h)}(nT,kF)X^{(h)*}(pT,mF)\times \\ \times W_{\tilde{g},\tilde{g}}(t,f,nT,kF,pT,mF) . \quad (12)$$

Eq. (12) is essentially the TFDS with order d of [4].

3. THE WWD AND COHEN'S CLASS

The TFR of a signal x(t) is said to belong to Cohen's fixed kernel shift-covariant class if and only if the TFR is a 2-D filtered WD, i.e., the TFR can be expressed by

$$T_x(t,f) = \iint\varphi(t-\tilde{t},f-\tilde{f})W_x(\tilde{t},\tilde{f})d\tilde{t}d\tilde{f} \quad (13)$$

where $\varphi(t,f)$ is a fixed (signal independent) kernel [2]. Eq. (13) can be equivalently written as [2]

$$T_x(t,f) = \iint\Psi(\tau,\nu)A_x(\tau,\nu)\exp[j2\pi(\nu t-f\tau)]d\tau d\nu \quad (14)$$

where the kernel function $\Psi(\tau,\nu)$ and the ambiguity function $A_x(\tau,\nu)$ are the 2-D Fourier transforms (a Fourier transform with respect to t and an inverse Fourier transform with respect to f) of $\varphi(t,f)$ and $W_x(t,f)$, respectively.

Substitution of (5) into (13) yields

$$T_x(t,f) = \int\int\int\int X^{(h)}(\alpha,\theta)X^{(h)*}(\beta,\upsilon)\times \\ \times\left\{\iint\varphi(t-\tilde{t},f-\tilde{f})W_{\tilde{g},\tilde{g}}(\tilde{t},\tilde{f},\alpha,\theta,\beta,\upsilon)d\tilde{t}d\tilde{f}\right\}d\alpha d\theta d\beta d\upsilon \quad (15)$$

where $\{\cdot\}$ denotes a filtered elementary WD. Comparing (10) to (15), we see that the WWDs and the TFRs of Cohen's class have similar forms. However, the WWD of (10) results from weighted elementary WDs whereas the TFRs given by (15) result from filtered elementary WDs.

4. WWD ALGORITHM

4.1. Algorithm

The algorithm for the WWD (10) proceeds as follows.

(a) Find the STFT $X^{(h)}(t,f)$.

(b) Find $W_g(t,f)$.
$W_g(t,f)$ is used in the computation of every elementary WD and, thus, its computation is treated as a separate step. Note that if g(t) is chosen in such a way that $W_g(t,f)$ has a closed-form expression, the computation can be carried out from this expression and, consequently, the full Nyquist bandwidth can be obtained without over-sampling [5].

(c) Find $\tilde{W}_x(t,f)$. For each pair of elementary signals whose weight is non-zero, the corresponding pair of elementary WDs is computed and weighted. All the weighted elementary WDs are summed to obtain $\tilde{W}_x(t,f)$.

4.2. Sampling of the STFT

The computation of the STFT involves its sampling in time and in frequency. The sampling intervals used affect the precision and the speed of the algorithm considerably. Smaller sampling intervals cause higher precision but slower speed. An empirical method for choosing the sampling intervals is given in [4,6]. Here, we solve the problem theoretically.

The discrete version of $\tilde{W}_x(t,f)$ can be obtained by discretizing (10). Also, it can be derived from the inverse sampled STFT, also known as Gabor expansion [4,6,7]. The sampled STFT is defined as

$$X^{(h)}(nT,kF) = \int x(t)h^*(t-nT)\exp(-j2\pi kFt)dt \quad (16)$$

where T and F are the sampling intervals of time and frequency, respectively. The inverse sampled STFT reconstructs the signal from the sampled STFT, i.e.,

$$x(t) = TF\sum_n\sum_k X^{(h)}(nT,kF)g(t-nT)\exp(j2\pi kFt) . \quad (17)$$

Eq. (17) holds if

$$\int g(t)h\left(t-\frac{m}{F}\right)\exp\left(-j2\pi\frac{n}{T}t\right)dt = \delta(m)\delta(n) . \quad (18)$$

Taking the WD of (17), we get

$$W_x(t,f) = T^2F^2\sum_n\sum_k\sum_p\sum_m X^{(h)}(nT,kF)X^{(h)*}(pT,mF)\times \\ \times W_{\tilde{g},\tilde{g}}(t,f,nT,kF,pT,mF) . \quad (19)$$

Using a weight function and discretizing t and f, we obtain the discrete version of the WWD of (10), i.e.,

$$\tilde{W}_x(\tilde{n}T_s,\tilde{k}F_s) = \\ = T^2F^2\sum_n\sum_k\sum_p\sum_m X^{(h)}(nT,kF)X^{(h)*}(pT,mF)\times \\ \times\lambda(nT,kF,pT,mF)W_{\tilde{g},\tilde{g}}(\tilde{n}T_s,\tilde{k}F_s,nT,kF,pT,mF) . \quad (20)$$

The above development shows that if h(t), g(t), T, and F are chosen such that (18) is satisfied, (17) and thus (20) will hold. Hence, since the following conditions are sufficient for the validity of (18), they can be used as the criteria for choosing h(t), g(t), T, and F:

A. h(t) and g(t) are real even window functions.

B. h(t) and g(t) satisfy (3).

C. If F_h and F_g denote the bandwidths of h(t) and g(t), respectively, then

$$T \le \frac{2}{F_h+F_g} . \quad (21)$$

D. If T_h and T_g are the width of h(t) and g(t), respectively, then

$$F \le \frac{2}{T_h+T_g} . \quad (22)$$

Conditions A and D guarantee that (18) holds when $m \neq 0$. Conditions A and C guarantee that (18) holds when m = 0 and

$n \neq 0$. When $m = 0$ and $n \neq 0$, (18) becomes

$$\int g(t)h(t)\exp\left(-j2\pi\frac{n}{T}t\right)dt = 0, \tag{23}$$

which is equivalent to

$$\int G\left(\frac{n}{T}-f\right)H(f)\,df = 0. \tag{24}$$

Eq. (24) holds if conditions A and C are satisfied. Condition B guarantees that (18) holds when $m = n = 0$.

5. EXPERIMENTAL RESULTS

Using (10), a WWD is given below. We call this WWD the Gauss-Hamming distribution (GHD). We choose

$$h(t) = g(t) = \sqrt[4]{\frac{2}{\pi t_0^2}}\exp\left(-\frac{t^2}{t_0^2}\right). \tag{25}$$

Note that h(t) and g(t) satisfy conditions A and B of section 4 above. Taking the Fourier transform of (25), we get

$$H(f) = G(f) = \sqrt[4]{2\pi t_0^2}\exp\left(-\pi^2 t_0^2 f^2\right). \tag{26}$$

Defining the width of the function $\exp(-x^2/x_0^2)$ as $4x_0$, we obtain

$$T_h = T_g = 4t_0 \tag{27}$$

and

$$F_h = F_g = \frac{4}{\pi t_0}. \tag{28}$$

Substituting (28) and (27) into conditions C and D of section 4, we obtain

$$T \leq \frac{\pi t_0}{4} \quad \text{and} \quad F \leq \frac{1}{4t_0}, \tag{29}$$

respectively. $W_g(t, f)$ has the closed-form expression

$$W_g(t,f) = 2\exp\left[-\left(\frac{2t^2}{t_0^2}+2\pi^2 t_0^2 f^2\right)\right]. \tag{30}$$

The weight function is chosen as

$$\lambda(\alpha,\theta,\beta,\upsilon) = 0.54 + 0.46\cos\left[\pi\sqrt{\frac{(\alpha-\beta)^2}{a^2}+\frac{(\theta-\upsilon)^2}{b^2}}\right],$$

$$\frac{(\alpha-\beta)^2}{a^2}+\frac{(\theta-\upsilon)^2}{b^2} \leq 1. \tag{31}$$

The effect of the weight function (31) on a signal composed of two Gaussian components is shown[2] in Fig. 1. A faster decreasing $\lambda(\alpha,\theta,\beta,\upsilon)$, i.e., smaller a and b in (31), results in poorer auto-term resolution and more cross-term suppression (see Fig. 1(a)). The reverse situation is shown in Fig. 1(b).

The kernel $\Psi(\tau,\nu)$ in (14) uniquely specifies a given Cohen's class TFR. For example, we used a Hamming window to obtain

$$\Psi(\tau,\nu) = 0.54 + 0.46\cos\left(\pi\sqrt{\frac{\nu^2}{a^2}+\frac{\tau^2}{b^2}}\right). \tag{32}$$

We call the resulting TFR the Hamming distribution (HD).

We first consider a simulation involving a 128-point signal composed of three Hamming-windowed parallel complex linear chirps. The results are shown in Fig. 2. The WD is shown in Fig. 2(a) and a spectrogram using a Hamming window is given in Fig. 2(b). Fig. 2(c) shows the Hamming distribution (HD) and the GHD is given in Fig. 2(d). Note that the HD and the GHD portray better auto-term resolution than the spectrogram and no visible cross-terms.

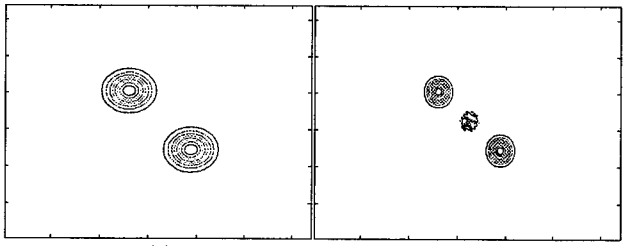


Fig. 1. GHD contour plots for the simulation involving two Gaussian signals. (a) Faster decreasing weight function; and (b) slower decreasing weight function.

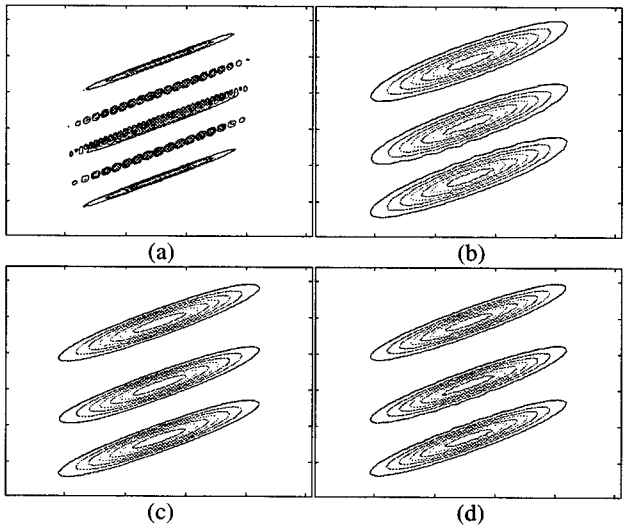


Fig. 2. Contour plots for the simulation involving three parallel complex linear chirps. (a) WD; (b) spectrogram; (c) HD; and (d) GHD.

A simulation involving a 256-point signal consisting of a complex sinusoid, a linear chirp, and a hyperbolic chirp is depicted in Fig. 3. Fig. 3(a) displays the WD of the composite signal whereas Fig. 3(b) shows the spectrogram when a 32-point Hamming window is used. The spectrogram exhibits essentially no cross terms and a major resolution loss in the auto-terms. On the other hand, the WD exhibits perfect resolution in the auto-terms and the presence of all cross-terms. The HD and the GHD are given in Figs. 3(c) and 3(d), respectively. The HD and the GHD exhibit similar performance. In particular, both preserve auto-term resolution better than the spectrogram and exhibit fewer and much reduced cross-terms than the WD.

Fig. 4 shows the simulation involving a 128-point signal composed of one Gaussian component, one linear chirp, and one component having parabolic instantaneous frequency. Fig. 4(a) displays the WD with its undesirable inner and outer cross-terms.

[2] All TFRs are displayed graphically by employing 7 linearly spaced contours. All contour plots have time running horizontally and increasing to the right; frequency runs vertically and increases to the top.

Fig. 4(b) displays a spectrogram with its undesirable low-resolution auto-terms. As Figs. 4(c)-(d) show, both the HD and the GHD are good compromises between the WD and the spectrogram.

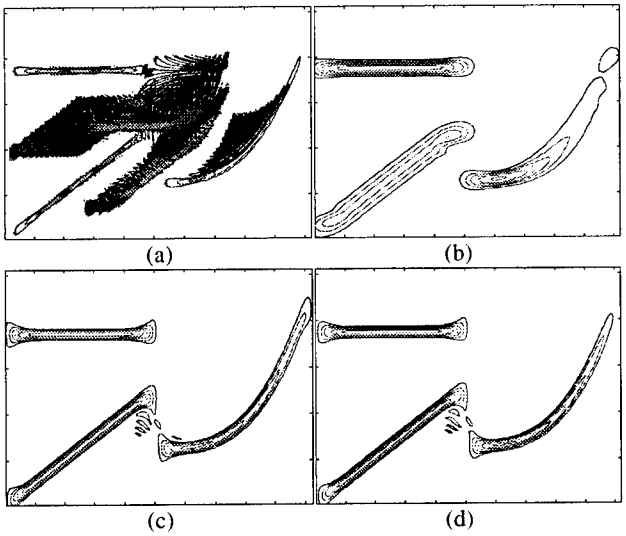


Fig. 3. Contour plots for the simulation involving a signal composed of one complex sinusoid, one linear chirp, and one hyperbolic chirp. (a) WD; (b) spectrogram; (c) HD; and (d) GHD.

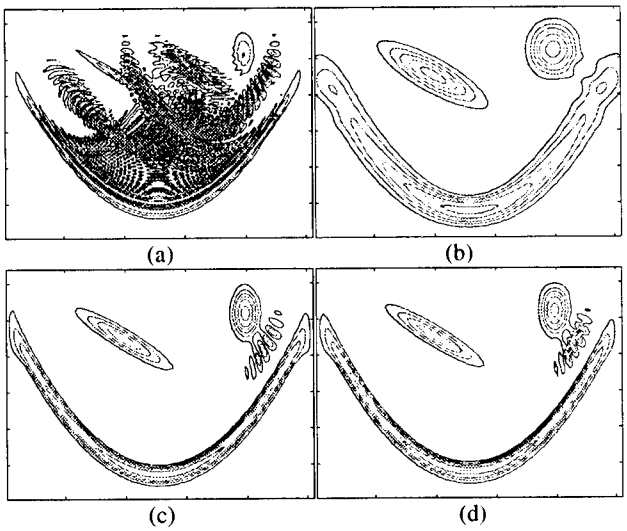


Fig. 4. Contour plots for the simulation involving a signal composed of one Gaussian component, one linear chirp, and one parabolic chirp. (a) WD; (b) spectrogram; (c) HD; and (d) GHD.

Fig. 5 shows the simulation involving a bat chirp signal emitted by the Large Brown Bat (Eptesicus Fuscus)[3]. Both the HD and the GHD in Figs. 5(c)-(d) exhibit much better auto-term resolution than the spectrogram in Fig. 5(b) and much less cross-terms than the WD in Fig. 5(a).

[3] The authors wish to thank Curtis Condon, Ken White, and Al Feng of the Beckman Institute of the University of Illinois for the bat data and for permission to use it in this paper.

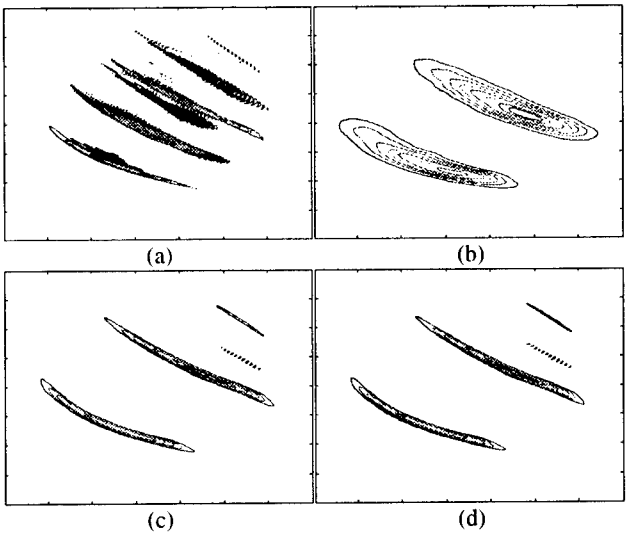


Fig. 5. Contour plots for the simulation involving a bat echolocation pulse. (a) WD; (b) spectrogram; (c) HD; and (d) GHD.

6. SUMMARY

The WD can be decomposed into a linear combination of elementary WDs. Slow-oscillatory elementary WDs and fast-oscillatory elementary WDs mainly contribute to auto-terms and cross-terms, respectively. Using a weight function to keep slow-oscillatory elementary WDs and attenuate fast-oscillatory elementary WDs, auto-term resolution and cross-term suppression can be effectively balanced.

7. REFERENCES

[1] L. Cohen, "Time-frequency distributions – a review," *Proceedings of the IEEE*, vol. 77, pp. 941-981, July 1989.

[2] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency representations," *IEEE Signal Processing Magazine*, vol. 9, pp. 21-67, April 1992.

[3] A. H. Costa and G. F. Boudreaux-Bartels, "Design of time-frequency representations using a multiform, tiltable exponential kernel," *IEEE Transactions on Signal Processing*, vol. 43, pp. 2283-2301, October 1995.

[4] S. Qian and D. Chen, "Decomposition of the Wigner-Ville distribution and time-frequency distribution series," *IEEE Transactions on Signal Processing*, vol. 42, pp. 2836-2842, October 1994.

[5] A. H. Costa and G. F. Boudreaux-Bartels, "An overview of aliasing errors in discrete-time formulations of time-frequency representations," *IEEE Transactions on Signal Processing*, vol. 47, pp. 1463-1474, May 1999.

[6] S. Qian and D. Chen, "Discrete Gabor transform," *IEEE Transactions on Signal Processing*, vol. 41, pp. 2429-2438, July 1993.

[7] J. Wexler and S. Raz, "Discrete Gabor expansions," *Signal Processing*, vol. 21, pp. 207-220, November 1990.

SUPPORT VECTOR REGRESSION FOR BLACK-BOX SYSTEM IDENTIFICATION

Arthur Gretton[1], *Arnaud Doucet*[2], *Ralf Herbrich*[3], *Peter J. W. Rayner*[1] *and Bernhard Schölkopf*[3]

[1] Signal Processing Group, University of Cambridge
Department of Engineering, Trumpington Street
CB2 1PZ, Cambridge, UK
{alg30,pjwr}@eng.cam.ac.uk
[2] Department of Electrical and Electronic Engineering
The University of Melbourne
Victoria 3010, Australia
a.doucet@ee.mu.oz.au
[3] Microsoft Research, Cambridge
St. George House, 1 Guildhall Street
Cambridge, CB2 3NH, UK
{rherb,bsc}@microsoft.com

ABSTRACT

In this paper, we demonstrate the use of support vector regression (SVR) techniques for black-box system identification. These methods derive from statistical learning theory, and are of great theoretical and practical interest. We briefly describe the theory underpinning SVR, and compare support vector methods with other approaches using radial basis networks. Finally, we apply SVR to modeling the behaviour of a hydraulic robot arm, and show that SVR improves on previously published results.

1. INTRODUCTION

System identification of nonlinear black-box models is a crucial but complex problem. There have been numerous recent papers in the area based on neural networks, wavelet networks, hinging hyperplanes, etc. Roughly speaking, one selects a set of regressors/basis functions, and tries to determine the number of basis/regressors and their parameters according to a given statistical criterion. Many methods are based on a penalised maximum likelihood criterion. Performing model selection and estimation is usually a difficult task, however, as it involves solving complex integration and/or optimisation problems. Gradient methods are often used, but are only guaranteed to converge toward local optima. Recently, in a Bayesian framework, Markov chain Monte Carlo algorithms have also been developed. These methods are computationally intensive, however.

We propose here an alternative approach based on support vector machines. These comprise a set of powerful tools to perform classification and regression [8], and have become very popular recently in the machine learning community. This approach, motivated by Statistical Learning Theory [10], is systematic and principled. One can list its main advantages:

- There are very few free parameters to adjust.
- Estimating the unknown parameters only involves optimisation of a convex cost function. This can be achieved using standard quadratic programming algorithms. This is *fast* and there are *no local minima*.
- The model constructed depends explicitly on the most "informative" data (the support vectors).
- It is possible to obtain theoretical bounds on the generalisation error and the sparseness of the solution (see [8]). These bounds are *independent* of the distribution generating the training and test data.

To the best of our knowledge, support vector regression (SVR) has never been used in the context of system identification, although it has been used in estimating time series by Müller *et al.* [4], and Mattera and Haykin [3]. This work differs from these previous studies in that it investigates the ν-SVR method [5], which does not require us to specify an a priori level of accuracy. We demonstrate the application of this algorithm to modeling a standard data set, and show that it is possible to obtain results that improve on current state-of-the-art methods [6], [7], with very little tuning.

2. BLACK-BOX SYSTEM IDENTIFICATION

The problem of nonlinear black-box system identification consists of conducting non-parametric regression, as described in Sjöberg *et al.* [6], [7], among others. This means that random variables $(\mathbf{x}, \mathrm{y})$, which take values in $\mathcal{X} \times \mathcal{Y}$, are generated according to a distribution $P_{\mathbf{x},\mathrm{y}}$, and we are required to estimate the *regression function* y on $\mathbf{x}$, or

$$\mathbf{E}_{\mathrm{y}}\left(\mathrm{y} | \mathbf{x} = \mathbf{x}\right) = f\left(\mathbf{x}\right).$$

We call $\mathbf{x}$ the regressor, and y the output. We further define $\mathcal{X} \triangleq \mathbb{R}^d$ and $\mathcal{Y} \triangleq \mathbb{R}$. We want to estimate $f\left(\cdot\right)$ from the training sample

$$z_N = \left(\left(\mathbf{x}_1, y_1\right), \ldots, \left(\mathbf{x}_N, y_N\right)\right) \in \left(\mathcal{X} \times \mathcal{Y}\right)^N,$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

each element of which is drawn from $P_{\mathbf{x},y}$. Since we do not know the mapping $f(\cdot)$, we define a learning algorithm $\mathcal{A}$, which gives us an estimate $f_{\mathbf{z}}(\cdot)$ of $f(\cdot)$,

$$\mathcal{A}: \bigcup_{N=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^N \rightarrow \mathcal{H}$$
$$\mathbf{z} \mapsto f_{\mathbf{z}}(\cdot),$$

within a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ (here $\mathcal{Y}^{\mathcal{X}}$ refers to the set of functions mapping $\mathcal{X}$ to $\mathcal{Y}$), called the *hypothesis space*, which is flexible enough to model a wide range of functions. An estimate $f_{\mathbf{z}}(\cdot)$ associated with the *loss* $c(\mathbf{x}, y, f_{\mathbf{z}}(\cdot))$ is attained by minimising the *risk*,

$$f_{\mathbf{z}}(\cdot) = \underset{g_{\mathbf{z}}(\cdot)\in\mathcal{H}}{\operatorname{argmin}} \left[R(g_{\mathbf{z}}(\cdot)) \triangleq \mathbf{E}_{\mathbf{x},y}[c(\mathbf{x}, \mathrm{y}, g_{\mathbf{z}}(\mathbf{x}))]\right]. \quad (1)$$

Possible loss functions include quadratic loss,

$$c(\mathbf{x}, y, g_{\mathbf{z}}(\cdot)) = |y - g_{\mathbf{z}}(\mathbf{x})|^2,$$

Vapnik's ϵ-insensitive loss [10],

$$c(\mathbf{x}, y, g_{\mathbf{z}}(\cdot)) = \max\{0,\ |g_{\mathbf{z}}(\mathbf{x}) - y| - \epsilon\},$$

and Huber loss,

$$c(\mathbf{x}, y, g_{\mathbf{z}}(\cdot)) = \begin{cases} \epsilon|g_{\mathbf{z}}(\mathbf{x}) - y| - \frac{\epsilon^2}{2} & \text{for } |g_{\mathbf{z}}(\mathbf{x}) - y| \geq \epsilon, \\ \frac{1}{2}|g_{\mathbf{z}}(\mathbf{x}) - y|^2 & \text{otherwise,} \end{cases}$$

among others.

In practice, the regression function $f_{\mathbf{z}}(\cdot)$ cannot readily be obtained from equation (1), since we do not usually know the distribution $\mathbf{P}_{\mathbf{x},y}$. Minimising the empirical risk alone does not take into account other requirements that we would like to satisfy, such as smoothness, and can therefore result in overfitting [8], [10].

Classes of system identification problems falling within the nonlinear black-box identification framework are described in [6], [7]. These include nonlinear finite impulse response models, nonlinear autoregressive models with external input, nonlinear output error models, nonlinear autoregressive moving average, nonlinear Box-Jenkins models, etc.

3. SUPPORT VECTOR REGRESSION

We now describe how support vector machines may be used to solve the system identification problem described in the previous section. The results in this section are derived in Schölkopf *et al.* [5].

To describe the ν-SVR procedure, we must first define a mapping from the space $\mathcal{X}$ of regressors to the possibly infinite dimensional hypothesis space $\mathcal{H}$, in which an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is defined. We formally describe this map as

$$\mathbf{\Phi}: \mathcal{X} \rightarrow \mathcal{H}$$
$$\mathbf{x} \mapsto \mathbf{\Phi}(\mathbf{x}).$$

We choose to limit our choice of regression function $f_{\mathbf{z}}(\cdot)$ to the class of functions which can be expressed as inner products in $\mathcal{H}$,

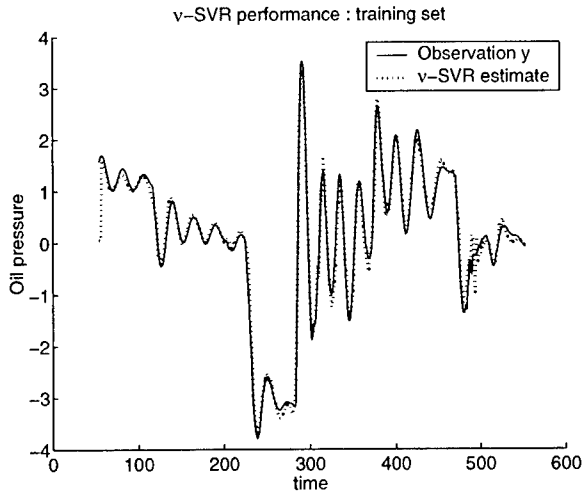


Fig. 1. Robot arm data and ν-SVR model : Training set and model approximation.

taken between some *weight vector* $\mathbf{w}$ and the mapped regressor $\mathbf{\Phi}(\mathbf{x})$;

$$f_{\mathbf{z}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x})\rangle_{\mathcal{H}} + b. \quad (2)$$

The regression function in the hypothesis space is consequently linear, and thus the *nonlinear* regression problem of estimating $f_{\mathbf{z}}(\mathbf{x})$ has become a *linear* regression problem in the hypothesis space $\mathcal{H}$. Note that the mapping $\mathbf{\Phi}(\cdot)$ need never be computed explicitly; instead, we use the fact that if $\mathcal{H}$ is the reproducing kernel Hilbert space induced by $k(\cdot,\cdot)$, then writing $\mathbf{\Phi}(\mathbf{x}) = k(\mathbf{x},\cdot)$, we get

$$\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j)\rangle_{\mathcal{H}} = k(\mathbf{x}_i, \mathbf{x}_j).$$

The latter requirement is met for kernels fulfilling the Mercer conditions [8]. These conditions are satisfied for a wide range of kernels, including Gaussian radial basis functions (see equation (6)). We emphasise that the feature space need never be defined explicitly, since only the kernel is used in SVR algorithms. Indeed, it is possible for multiple feature spaces to be induced by a single kernel.

We now describe the optimisation problem to be undertaken in finding $f_{\mathbf{z}}(\cdot)$. All support vector regression methods involve the minimisation of a regularised risk functional, which represents a tradeoff between smoothness and training error (the latter is determined by the cost functional). In the case of the ν-SVR method, the regularised risk $R^{\epsilon}_{\mathrm{emp}}(f_{\mathbf{z}}, \mathbf{z})$ at the optimum is given by

$$\min_{\mathbf{w},b,\epsilon} \left[R^{\epsilon}_{\mathrm{reg}}(f_{\mathbf{z}}(\cdot), \mathbf{z})\right] = \min_{\mathbf{w},b,\epsilon} \left[\frac{1}{2}\|\mathbf{w}\|^2_{\mathcal{H}} + C\left(\nu\epsilon + R^{\epsilon}_{\mathrm{emp}}(f_{\mathbf{z}}(\cdot), \mathbf{z})\right)\right], \quad (3)$$

where we use the Vapnik ϵ-insensitive loss in the empirical risk;

$$R^{\epsilon}_{\mathrm{emp}}(f_{\mathbf{z}}(\cdot), \mathbf{z}) = \frac{1}{N}\sum_{i=1}^{N} c(\mathbf{x}_i, y_i, f_{\mathbf{z}}(\cdot)) = \frac{1}{N}\sum_{i=1}^{N} \xi_i + \xi_i^*,$$

in which

$$\xi_i = \max\{0,\ f_{\mathbf{z}}(\mathbf{x}) - y - \epsilon\} \quad \text{and}$$
$$\xi_i^* = \max\{0,\ -f_{\mathbf{z}}(\mathbf{x}) + y - \epsilon\}.$$

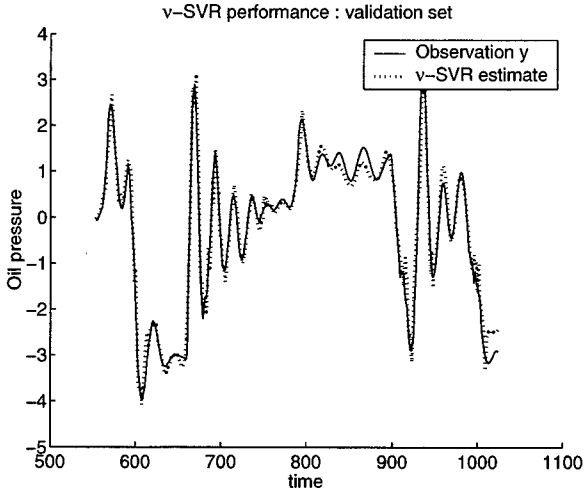


Fig. 2. Robot arm data and ν-SVR model : Validation set and model approximation.

All training points $(\mathbf{x}_i, y_i)$ for which $|f_{\mathbf{z}}(\mathbf{x}_i) - y_i| \geq \epsilon$ are known as *support vectors*; it is only these points that determine $f_{\mathbf{z}}(\cdot)$. Note that other loss functions, such as the Huber loss, can also be used in support vector regression, although not all loss functions result in a sparse representation. The terms C and ν in equation (3) specify the tradeoff between model simplicity, the size of the parameter ϵ below which the loss is zero, and the total empirical loss over the training set, $R^{\epsilon}_{\text{emp}}(\cdot)$. Schölkopf *et al.* [5] describe the theoretical behaviour of ν and C in more detail.

It can be shown [5] that the component $\mathbf{w}$ in equation (2) is a linear combination of the mapped training points,

$$\mathbf{w} = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) \mathbf{\Phi}(\mathbf{x}_i), \tag{4}$$

and that solving equation (3) is equivalent to finding

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*} \; -\frac{1}{2} \sum_{i,j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{N} y_i (\alpha_i^* - \alpha_i), \tag{5}$$

subject to

$$\sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0, \qquad \alpha_i, \alpha_i^* \in \left[0, \frac{C}{N}\right],$$

$$\sum_{i=1}^{N} (\alpha_i + \alpha_i^*) \leq C\nu.$$

There exist a number of methods that can be used to solve this quadratic programming problem. Our results were obtained using the LOQO algorithm in Vanderbei [9]. In the case of large training sets, data decomposition methods exist to speed convergence; see

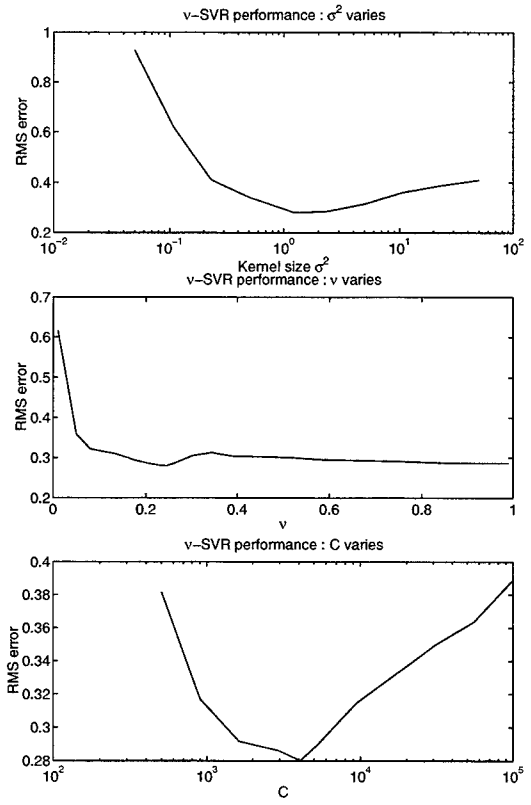


Fig. 3. RMS error variation with ν, σ^2 and C. In each case, the fixed parameters take their optimal values.

e.g. Chang *et al.* [2]. The offset b is found using

$$\langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x}_j) \rangle_{\mathcal{H}} + b - y_j = \epsilon \quad \text{when } \alpha_j \in \left(0, \frac{C}{m}\right),$$

$$y_j - \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x}_j) \rangle_{\mathcal{H}} - b = \epsilon \quad \text{when } \alpha_j^* \in \left(0, \frac{C}{m}\right);$$

the set of equations thus obtained can be solved via linear least squares.

4. COMPARISON WITH STANDARD RBF APPROACHES

A popular set of regression functions are the radial basis functions. The radial basis function expansion is

$$f_{\mathbf{z}}(\mathbf{x}) = \sum_{i=0}^{M} w_j k_j(\boldsymbol{\mu}_j, \mathbf{x}),$$

where M is the number of radial basis functions used (this need not be the same as the number N of training points), $w_j \in \mathbb{R}$ scale the various basis functions $k_j(\boldsymbol{\mu}_j, \cdot)$, w_0 scales the constant offset term $k_0(\boldsymbol{\mu}_0, \cdot) \triangleq 1$, and each basis function $k_j(\boldsymbol{\mu}_j, \cdot)$ has an individual centre parameter $\boldsymbol{\mu}_j$ and width parameter σ_j. For instance, in the case of Gaussian radial basis networks,

$$k_j(\boldsymbol{\mu}_j, \mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2\sigma_j^2}\right). \tag{6}$$

It is also possible to use a more general covariance matrix, rather than σ_j^2; this results in a greater number of parameters that require adjustment.

It is clear that SVR methods in fact produce radial basis function networks, with all width parameters σ_j^2 set at the same value, and centres $\boldsymbol{\mu}_j$ corresponding to support vectors $\mathbf{x}_j$ (thus M is the number of support vectors). As discussed previously, the SVR training procedure selects the training points to be used in this expansion so as to avoid overfitting, and to achieve sparseness with regards to the training data. Furthermore, the attendant optimisation process is convex, and has a single optimum.

There is a great deal of literature, past and present, on methods for training radial basis function networks; see for instance Bishop [1]. Without going into detail, it is fairly common practice to centre the basis functions on the training data and fix the basis width a priori, as in SVR. Model selection (determining the number of non-null weight vector components $\mathbf{w}_i$) and parameter estimation (estimating the values of the $\mathbf{w}_i$) in traditional radial basis function network methods, however, are usually based on Bayesian/penalized maximum likelihood approaches; the associated optimisation problems are often non-convex and possess multiple local minima, which can lead to greater computational complexity.

5. EXPERIMENTAL RESULTS

In the following experiments, we make use of a Gaussian radial basis function kernel, as described in equation (6), with kernel width σ^2 (note that other kernel options, such as polynomial kernels or sigmoid kernels, could also be used). As we perform SV regression, the kernel centres are set at the training point locations $\mathbf{x}_j$. We apply the ν-SVR algorithm to modeling behaviour of a hydraulic robot arm; our result will be compared with the neural network NARX and wavelet network NARX models in Sjöberg *et al* [6]. The input u_t represents the size of the valve through which oil flows into the actuator, and the output y_t is a measure of oil pressure (the latter determines the arm position). For the purpose of comparison, we used the regressor

$$\mathbf{x}_t = \begin{bmatrix} y_{t-1} & y_{t-2} & y_{t-3} & u_{t-1} & u_{t-2} \end{bmatrix}^T,$$

since this is also used by Sjöberg *et al*. We also used half the data set for training, and half as validation data, again following the procedure of Sjöberg *et al*. The kernel width was set at $\sigma^2 = 1.2242$, and we used the ν-SVR parameters $\nu = 0.2444$ and $C = 4.07 \times 10^3$. It must be emphasised that the experimental outcome varies little for a wide range of parameter values; see figure 3. Note also that prior knowledge of the observation noise would allow us to select a value of ν that is asymptotically optimal in the number of data [8].

The ν-SVR model output on the training data is given in figure 1, and the model output on the validation data in figure 2. The RMS error of this prediction on the validation set is 0.280, which is lower than both the wavelet network RMS error (0.579), and the prediction made by a one-hidden-layer sigmoid neural network with ten hidden units (0.467). Although Sjöberg *et al.* were able to further reduce the RMS error to 0.328 on this data set, this required assumptions regarding the model structure not made in our algorithm. Further advantages of the ν-SVR solution include simplicity, computational efficiency, robustness in the face of decreased training set size, and ease of tuning, due to the low sensitivity of the solution to changes in σ^2, ν and C. Our implementation of the ν-SVR algorithm required 56 lines of Matlab code (excluding the standard quadratic programming component), and took 193 seconds to train on the data set in figure 1, using a Pentium III processor running at 500MHz.

6. CONCLUSION

In this study, we describe the important theoretical and practical advantages of support vector regression for back box system identification. The simplicity of implementation, coupled with good performance in both this and other studies on time series prediction, make SVR methods an attractive alternative to standard system identification techniques.

7. REFERENCES

[1] C. Bishop. *Neural Networks for Pattern Recognition*, chapter 5. Oxford University Press, Oxford, 1995.

[2] C.-C. Chang, C.-W. Hsu, and C.-J. Lin. The analysis of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 11(4):1003–1008, 2000.

[3] D. Mattera and S. Haykin. Support vector machines for dynamic reconstruction of a chaotic system. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*. MIT Press, 1999.

[4] K.R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Using support vector machines for time series prediction. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*. MIT Press, 1999.

[5] B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[6] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691–1724, 1995.

[7] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: Mathematical foundations. *Automatica*, 31(12):1725–1750, 1995.

[8] A. Smola and B. Schölkopf. *Learning with Kernels*. MIT press, To appear.

[9] R. J. Vanderbei. LOQO: An interior point code for quadratic programming. Technical Report TR SOR-94-15, Department of Civil Engineering and Operations Research, Princeton University, 1995.

[10] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, N.Y., 1995.

TIME-VARYING QUADRATIC MODEL SELECTION USING WAVELET PACKETS

Matthew Green

Australian Telecommunications Cooperative Research Center
Curtin University of Technology, GPO Box U1987, Perth WA 6845, Australia
Email: matthewgreen@ieee.org

ABSTRACT

Model selection and system identification for cases where the model is required to have both characteristics of time-variance and nonlinearity is considered. To enable identification from a single input/output observation record, the time-variation is approximated by a weighted sum of orthogonal sequences. Wavelet packets are chosen for these sequences and an adapted basis for each time-varying coefficient is selected via the Best Basis algorithm [1]. Individual wavelet packets are then selected via a multiple hypothesis test which determines those packets that are significant to each approximation, and which may be discarded from the model.

1. IDENTIFICATION OF TIME-VARYING NONLINEAR SYSTEMS

All physical systems exhibit some degree of time-varying nonlinear behavior. Despite the linear time-invariant model being adequate for many systems, there is a growing requirement for more accurate models to achieve higher performance. Of particular interest at present is communications where channel characterisation is required for such tasks as analysis and equalisation. In mobile communications, multipath propagation and user motion result in a dispersive time-varying communication channel [2]. Nonlinear communication channels are also widely studied [3, 4, 5], typically modeled by a Volterra model.

A time-varying quadratic Volterra model with memory M may be written as

$$y(n) = \sum_{m_1=0}^{M-1} h_1(n,m_1)x(n-m_1) + \sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1} h_2(n,m_1,m_2)x(n-m_1)x(n-m_2)$$

for the observed time record of $n = 0, 1, \ldots, N-1$. The time-varying kernels $h_1(n,m_1)$ and $h_2(n,m_1,m_2)$ respectively represent the linear and quadratic time-varying dynamics of the system. Taking into account the symmetry in arguments of the quadratic kernel (i.e. $h_2(n,m_1,m_2) = h_2(n,m_2,m_1)$), the model contains $(M^2+3M)/2$ time-varying coefficients and thus $N(M^2+3M)/2$ parameters. To enable their estimation from a single input/output record, a subset of sequences from an orthogonal basis may be used to approximate the time-variation. The basis from which the sequences are taken may be determined using *a priori* knowledge of the characteristics of the system's time-variation, or from a general basis when this information is lacking. An approximation for the linear kernel may be written as

$$h_1(n,m_1) \approx \sum_{sfp\in\mathcal{Q}_{m_1}} \beta_{1sfp}(m_1)\psi_{sfp}(n)$$

with a similar expression for $h_2(n,m_1,m_2)$. The summation is over $\mathcal{Q}_{m_1}$, a set of Q triples $\{sfp\}$ indexing the coefficients $\beta_{1sfp}(m_1)$, and their corresponding orthogonal wavelet packets $\psi_{sfp}(n)$. For example, say

$$\begin{aligned} h_1(n,0) \approx & \beta_{1200}(0)\psi_{200}(n) + \beta_{1210}(0)\psi_{210}(n) \\ & + \beta_{1110}(0)\psi_{110}(n) + \beta_{1111}(0)\psi_{111}(n), \end{aligned}$$

then $\mathcal{Q}_0 = \{\{200\},\{210\},\{110\},\{111\}\}$, for $Q = 4$.

Since all $\psi_{sfp}(n)$ are known, we need only estimate the time-invariant kernels $\beta_{1sfp}(m_1)$ and $\beta_{2sfp}(m_1,m_2)$. If Q wavelet packets are used for each approximation there are $Q(M^2+3M)/2$ unknown time-invariant coefficients — a great reduction in unknowns if $Q \ll N$. Being *linear-in-the-parameters*, the Volterra model may be written as a linear regression in matrix form

$$\mathbf{y} = \mathbf{X}_\psi \mathbf{b} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of N output observations, $\mathbf{X}_\psi$ is a matrix incorporating the input observations $x(n)$ with the wavelet packets $\psi_{sfp}(n)$, $\mathbf{b}$ is a vector containing coefficients from $\beta_{1sfp}(m_1)$ and $\beta_{2sfp}(m_1,m_2)$ for $sfp \in \mathcal{Q}_{m1}$ and $sfp \in \mathcal{Q}_{m2}$ respectively, and $m_1 = 0, 1, \ldots, M-1$ and $m_2 = m_1, \ldots, M-1$. The error term $\mathbf{e}$ denotes noise and modeling mismatch. The least squares solution is

$$\hat{\mathbf{b}} = (\mathbf{X}_\psi^T \mathbf{X}_\psi)^{-1} \mathbf{X}_\psi^T \mathbf{y}.$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

2. BEST WAVELET PACKET BASIS

Wavelets have been advocated for their flexibility and applicability to real-life signal characteristics [6]. All wavelets in particular basis are scaled and translated versions of a single analysing wavelet $\psi(t)$,

$$\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$$

where $j, k \in \mathbb{Z}$. The analysing wavelet is defined in terms of a *scaling function*, $\phi(t)$,

$$\psi(t) = \sum_k (-1)^k h_{1-k}\sqrt{2}\phi(2t-k).$$

The scaling function satisfies

$$\phi(t) = \sum_k h_k \sqrt{2}\phi(2t-k),$$

where $h_k = \langle \phi(t), \sqrt{2}\phi(2t-k)\rangle$ are refinement coefficients. The set of functions $\phi(t)$ and $\psi_{jk}(t)$, $j = 0, 1, \ldots$, $k = 0, 1, \ldots, 2^j - 1$, forms an orthogonal basis in $L^2(\mathbb{R})$.

Generalising this basis, wavelet packets are particular combinations or superpositions of wavelets. Defining the following sequence of functions

$$\psi_{2r}(t) = \sqrt{2}\sum_k h_k \psi_r(2t-k)$$
$$\psi_{2r+1}(t) = \sqrt{2}\sum_k g_k \psi_r(2t-k)$$

where $\psi_0(t) = \phi(t)$ and $\psi_1(t) = \psi(t)$, the wavelet packets are dilated/translated versions of these functions:

$$\psi_{sfp}(t) = 2^{s/2}\psi_f(2^s t - p)$$

for $0 \le s \le L$, $0 \le f < 2^s$, $0 \le p < 2^{L-s}$ and $Q = 2^L$. The indices s, f and p relate to scale, frequency and position, respectively. Using these indices the packets may be organized into a rectangular binary tree of nodes corresponding to packets of equivalent sf, as illustrated in Figure 1. Each node (rectangle) is the parent of the two nodes

000	001	002	003	004	005	006	007
100	101	102	103	110	111	112	113
200	201	210	211	220	221	230	231
300	310	320	330	340	350	360	370

Figure 1: Organisation of wavelet packets into nodes showing the indices sfp.

directly below it, and the child of the one above. A wavelet packet in any one node is orthogonal to the other packets in that node as well as packets in other nodes with which its node does not vertically overlap.

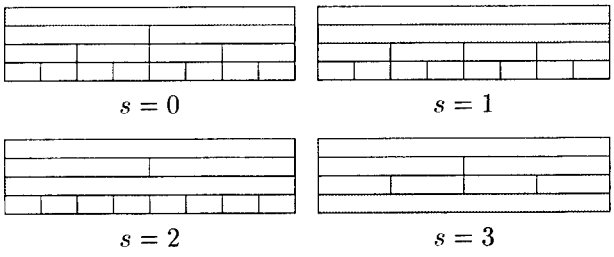


Figure 2: Wavelet packet bases used in identification.

Orthogonality properties of wavelet packets allows the formation of many bases. Furthermore, these properties enable the wavelet packet bases to be efficiently searched through to find the most appropriate basis for use in approximating a signal. The Best Basis algorithm [1] is used for this, determining the minimum entropy basis. Minimising the entropy results in a basis whose set of coefficients has a relatively small number that are non-negligible and the sum of magnitudes of the negligible coefficients is negligible. Entropy of a sequence $c = \{c_i : i = 1, 2, \ldots, N\}$ is defined as

$$M(\mathbf{c}) = -\sum_i p_i \log(p_i) \quad \text{where } p_i = \frac{|c_i|^2}{\|c\|^2}.$$

The algorithm begins by calculating the entropies for the coefficients of the wavelet packets in each node. Then, starting from the bottom of the tree, the entropies of the child nodes are compared to their parents. If the child nodes have lower entropy, they replace the parent node and become a child of the next parent node.

To find the best basis for each of the system's time-varying coefficients, a time-varying model is identified for several bases of wavelet packets to get values for all the coefficients. These bases are illustrated in Figure 2 for $Q = 8$. Four distinct bases cover the entire library of wavelet packets which consists of 32 possible bases. Best Basis is then run on both set of coefficients $\beta_{1sfp}(m_1)$, $sfp \in \mathcal{Q}_{m_1}$ and $\beta_{2sfp}(m_1, m_2)$, $sfp \in \mathcal{Q}_{m_1 m_2}$, for $m_1 = 0, 1, \ldots, M-1$, $m_2 = m_1, \ldots, M-1$. Now that a set of Q sequences has been chosen for each time-varying coefficient, it is desired to discard those sequences that do not contribute significantly to the approximation. This may be done several ways including thresholding or by hypothesis testing.

3. MULTIPLE HYPOTHESIS TESTING FOR BASIS SEQUENCE SELECTION

To test the significance a particular wavelet packet has on the regression, rewrite the regression as

$$\mathbf{y} = \mathbf{x}_{\psi_i}\beta_i + \mathbf{X}_{\bar{i}}\mathbf{b}_{\bar{i}} + \mathbf{e}$$

where β_i is the coefficient under test, $\mathbf{x}_{\psi_i}$ is its corresponding vector from $\mathbf{X}_\psi$ and $\mathbf{X}_{\bar{i}}$ and $\mathbf{b}_{\bar{i}}$ is the rest of $\mathbf{X}_\psi$ and $\mathbf{b}$ from the regression respectively. There are $Q(M^2 + 3M)/2 = QP$ coefficients to test so $i = 1, 2, \ldots, QP$. The hypothesis

$$H_i : \beta_i = 0, \quad \mathbf{b}_{\bar{i}} \text{ unspecified}$$

is tested against the two-sided alternative

$$K_i : \beta_i \neq 0$$

for $i = 1, 2, \ldots, QP$. Performing an α-level test on this hypothesis, if we reject H_i, the probability that the sequence is significant is $(1 - \alpha)$. Strictly speaking, if we do not reject H_i we cannot say whether the sequence is significant or not. However, we risk not rejecting a false hypothesis and remove the sequence from the model.

A suitable test statistic for testing H_i measures the reduction in the residual sum of squares due to adding the parameter β_i to the regression:

$$T_i = (N - QP)\frac{||\mathbf{y} - \mathbf{X}_{\bar{i}}\hat{\mathbf{b}}_{\bar{i}}||^2 - ||\mathbf{y} - \mathbf{X}_\psi\hat{\mathbf{b}}||^2}{||\mathbf{y} - \mathbf{X}_\psi\hat{\mathbf{b}}||^2},$$

where $\hat{\mathbf{b}}_{\bar{i}} = (\mathbf{X}_{\bar{i}}^T\mathbf{X}_{\bar{i}})^{-1}\mathbf{X}_{\bar{i}}^T\mathbf{y}$. Under the null, T_i is F-distributed with degrees of freedom 1 and $N - QP$. The $\mathcal{P}$-value, obtained from $\hat{T}_i$,

$$\mathcal{P}_i = \Pr\left\{F_{1,(N-QP)} \geq \hat{T}_i \mid H_i\right\},$$

determines the significance of the sequence.

Since many parameters are to be tested, a multiple hypothesis test is employed. The most widely applied is the *Bonferroni* test [7]. The Bonferroni test is simple to use and enables individual hypotheses to be identified. We reject H_i if

$$\mathcal{P}_i \leq \frac{\alpha}{QP}, \quad i = 1, 2, \ldots, QP.$$

This is seen as quite a conservative test though.

To increase the power of the Bonferroni test a sequentially rejective Bonferroni (*Holm's*) procedure [8] may be used. Sorting the hypotheses, $H_{(1)}, H_{(2)}, \ldots, H_{(QP)}$, so their corresponding $\mathcal{P}$-values satisfy $P_{(1)} \leq \mathcal{P}_{(2)} \leq \cdots \leq \mathcal{P}_{(QP)}$, $H_{(i)}$ is then rejected if

$$\mathcal{P}_{(l)} \leq \frac{\alpha}{QP - l + 1}, \text{ for all } l = 1, 2, \ldots, i.$$

Other modifications to the Bonferroni procedure aiming to make it less conservative, provide a test for H_0 only, not individual hypotheses. Such tests may be extended, enabling rejection of individual hypotheses [9]. This was considered in [10].

4. SIMULATION RESULTS

The simulated system was a quadratic Volterra model with memory $M = 2$. Each of the 5 time-varying coefficients used summed the first $Q = 8$ Haar-Walsh wavelet packets with random coefficients uniformly distributed on [-1,1]. The input was Gaussian, $\mathcal{N}(0, 1)$, and of length $N = 128$.

Four models were identified, each using a basis of 8 wavelet packets (as in Figure 2). The basis for the 4th model corresponded to a Walsh basis. The entropies of the basis expansions for this model were 1.2074, 1.2552, 1.0993, 1.5547 and 1.5102. Using Best Basis to find the most efficient basis expansion for each time-varying coefficient led to the entropies 1.2055, 1.2177, 1.0132, 0.9884 and 1.2571, which are lower. The wavelet packets chosen were different from that used to simulate the system, as illustrated in Figure 3.

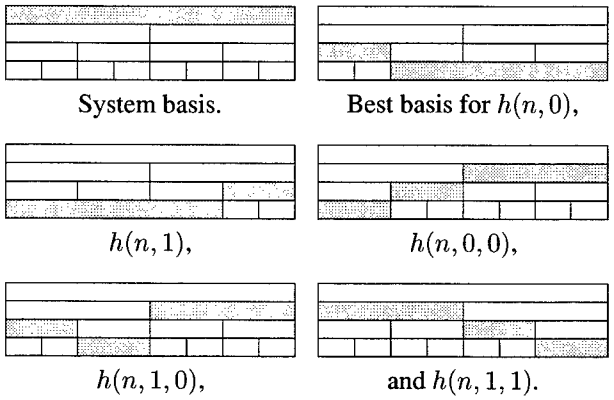


Figure 3: Basis used to simulate the system and best bases chosen for model.

Next, the contribution of each wavelet packet is tested. It was found that the $\mathcal{P}$-values were better indicators of the contribution rather than the packets coefficient magnitude. Starting with a model containing no wavelet packets at all, the packets were added one at a time according to $\mathcal{P}$-values and the model's output mean squared error (MSE) calculated. This was also done according to coefficient magnitudes. The results are shown in Figure 4. Including coefficients in the model according to their $\mathcal{P}$-value leads to a more accurate model. Thus if a model with only 30 coefficients is desired, the 30 coefficients with the smallest $\mathcal{P}$-values would be chosen, not those with the largest magnitude. This demonstrates that selecting coefficients via hypothesis testing is more effective than by simple thresholding.

Both Bonferroni and Holms' multiple test procedures were applied to the selection of sequences over a range of signal to noise ratios (SNRs). The number of hypotheses rejected by each test was averaged over 1000 runs (models). Figure 5(a) shows that Holms' proceedure usually rejects

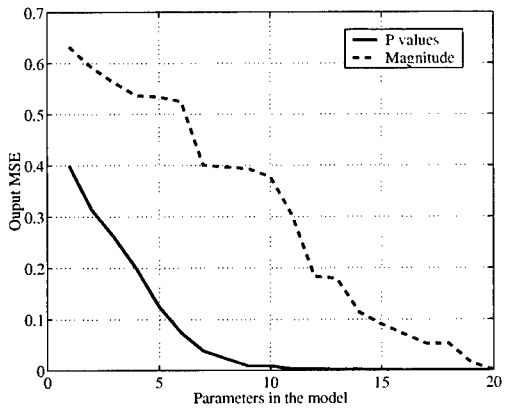
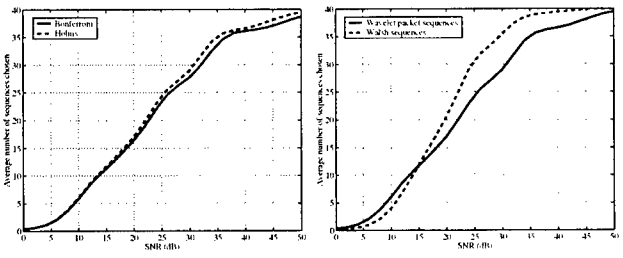


Figure 4: Effect of increasing model size on ouput MSE.

more hypotheses than the Bonferroni test. Figure 5(b) compares applying Holm's proceedure to the best basis model and the Walsh basis model. At higher SNR, less wavelet packets are significant than Walsh sequences indicating a more parsimonious model. However, at lower SNR more wavelet packets are selected. This may be due to the more flexible model being able to model the noise more and thus a number of falsely significant wavelet packets arising.



(a) (b)

Figure 5: Multiple hypothesis test results.

5. CONCLUSION

Choosing to use wavelet packets as the sequences to approximate the system's time-varying coefficients makes the model more flexible since a library of $Q(\log_2(Q)+1)$ sequences becomes available, from which Q orthogonal sequences may be chosen. Using another (non-wavelet) basis restricts the range of consideration to just Q specific sequences. Thus, the wavelet packets may lead to a better characterisation. To select the wavelet packets that most efficiently approximate the time-varying coefficients, the Best Basis algorithm may be employed.

To then determine which wavelet packets to keep in the model a multiple hypothesis test was applied. The $\mathcal{P}$-values were found to be a better indicator of the significance of a packet than the magnitude of its coefficient. Holms' test was slightly more powerful than Bonferroni at higher SNR, usually selecting 1 or 2 extra wavelet packets. Less wavelet packets were selected, at high SNR, when compared to using a model with Q walsh sequences. At low SNR, more wavelet packets were usually selected.

6. REFERENCES

[1] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for Best Basis selection," *IEEE Trans. on Information Theory*, vol. 38, pp. 713–718, March 1992.

[2] B. Sklar, "Rayleigh fading channels in mobile digital communication systems Part 1: Characterization," *IEEE Comm. Magazine*, pp. 90–100, July 1997.

[3] S. Benedetto, E. Biglieri, and R. Daffara, "Modeling and performance evaluation of nonlinear satellite links—A Volterra series approach," *IEEE trans. on Aerospace and Electronic Systems*, vol. 15, pp. 494–507, July 1979.

[4] H. Shichun and H. Zhenya, "Blind equalization of nonlinear communication channels using recurrent wavelet neural networks," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 4, (Los Alamitos, CA, USA), pp. 3305–3308, 1997.

[5] C. H. Cheng and E. J. Powers, "Optimal Volterra kernel estimation algorithms for a nonlinear communication system for PSK and QAM inputs," *IEEE Trans. on Signal Processing*, pp. 147–163, 2001.

[6] M. K. Tsatsanis and G. B. Giannakis, "Time-varying system identification and model validation using wavelets," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3512–3523, December 1993.

[7] Y. Hochberg and A. C. Tamhane, *Multiple comparison procedures*. New York: John Wiley, 1987.

[8] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, no. 6, pp. 65–70, 1979.

[9] G. Hommel, "A stagewise rejective multiple test procedure based on a modified Bonferroni test," *Biometrika*, vol. 75, no. 2, pp. 383–386, 1988.

[10] M. Green and A.M. Zoubir, "Selection of time-varying volterra model using multiple hypothesis testing.," in *34th Asilomar conf. on signals, systems and computers*, (Pacific Grove, U.S.A.), pp. 1782–1785, October 2000.

INTERNET TRANSPORT LAYER SYSTEM IDENTIFICATION

Langford B. White

Department of Electrical and Electronic Engineering
Adelaide University, Adelaide, SA 5005, Australia.
email : Lang.White@adelaide.edu.au

ABSTRACT

This paper addresses the problem of building appropriate statistical models of the way the Internet appears from the point of view of congestion, to a Transmission Control Protocol (TCP) sender. TCP is a mechanism for implementing full duplex, acknowledged, end-to-end transmission over an Internet Protocol (IP) network. This work has been motivated by the most recent TCP variant, the so-called Vegas implementation. TCP Vegas is really the first implementation to be based loosely on system theoretic ideas in the sense that it measures the segment round-trip times across the network to adjust its transmission rate. This paper develops a new linear system framework for TCP, and applies Recursive Prediction Error identification techniques to specify statistical models which may be used to develop alternative control strategies. Network simulations are used to illustrate behaviour.

1. INTRODUCTION

The rapid increase in Internet utilisation has led to high levels of congestion in some parts of the network, and has provided a driving force to improve efficiencies and throughput. End-to-end congestion control is included as part of the implementation of the Transmission Control Protocol (TCP) which provides a full duplex connection across the network. Reliability is achieved in TCP, by use of acknowledgements (ACKs) returned from the receiver to the sender. The vast majority ($> 90\%$) of applications use TCP as the transport mechanism, so improvements in the operation of TCP could well lead to significant improvements in Internet performance. Indeed, variants of the original TCP implementation have been proposed in an attempt to achieve just this. One significant recent proposal is that of TCP Vegas [1]. While earlier variants, such as TCP Reno (the most common in use currently) utilised a very coarse method of congestion control, Vegas adjusts its transmission rate in response to the round trip times (RTTs) of segments, ie the time between a segment being sent and the reception of the corresponding ACK. Thus there is at least in principle, the presence of a feedback control mechanism. Vegas however does not attempt to use any model of the relationship between transmission rate and RTTs, except to recognise that large RTTs are indicative of congestion and thus force reductions in transmission rate, albeit in a rather crude and inefficient manner. Vegas relies on a very simple control law precisely because it does not utilise any modelling information. The purpose of this work is to (i) derive a linear modelling representation consistent with the operation of Vegas, and (ii) investigate closed loop system identification schemes which are suitable for operation in this framework, including appropriate disturbance modelling.

2. OPERATION OF TCP VEGAS

The network will be characterised by a set of Q TCP sources $S = \{s_i : i = 1, \ldots, Q\}$, and associated receivers. Each source i sets its transmission rate by maintaining a congestion window of length $w_i(t)$, where t denotes a discrete time index. The source also measures the round trip time (RTT) $D_i(t)$ associated with the time of each received ACK. The standard TCP Vegas control law for adjusting $w_i(t)$ is

$$w_i(t+1) = w_i(t) + \begin{cases} D_i(t)^{-1} & \tilde{\epsilon}_i(t) > \alpha \\ -D_i(t)^{-1} & \tilde{\epsilon}_i(t) < \beta \\ 0 & \text{else}, \end{cases} \qquad (1)$$

where

$$\tilde{\epsilon}_i(t) = w_i(t)\left(\frac{1}{d_i} - \frac{1}{D_i(t)}\right). \qquad (2)$$

Here d_i denotes the round trip *propagation delay* for source i, and α and β are design parameters. We address their selection subsequently. We initially assume that $\alpha = \beta$ for ease of analysis. Thus in this case, (1) may be written

$$w_i(t+1) = w_i(t) + \frac{\mathrm{sgn}(e_i(t))}{D_i(t)}, \qquad (3)$$

where the error term is given by

$$e_i(t) = \alpha d_i - w_i(t)\left(1 - \frac{d_i}{D_i(t)}\right). \qquad (4)$$

In the following, we use the transformed measurements

$$y_i(t) = 1 - \frac{d_i}{D_i(t)}, \qquad (5)$$

which may be interpreted as the bandwidth efficiency for source i. So (3) and (4) may be written

$$\begin{aligned} w_i(t+1) &= w_i(t) + \frac{\mathrm{sgn}(e_i(t))}{d_i(1-y_i(t))} \\ e_i(t) &= r_i - w_i(t)\, y_i(t). \end{aligned} \qquad (6)$$

This error term can be easily interpreted as the difference between the term $r_i = \alpha\, d_i$ and the number of packets queued in the system for source i at time t. Thus r_i can be regarded as the desired

0-7803-7011-2/01/$10.00 ©2001 IEEE

number of packets from source i to be queued in the system at any given time. These parameters can be regarded as resource allocation parameters [2]. The equations (3) and (4) thus define a control system model for TCP Vegas. The measurements $D_i(t)$, and thus $y_i(t)$ are related to all users window sizes ie $w_1(s), \ldots, w_Q(s)$ for $s \leq t$, and other factors such as the network topology and queueing disciplines etc.

3. A LINEAR SYSTEM MODEL

In this section, we define a linear system model consistent with the operation of Vegas. The key idea we shall use is to work with *logarithmic* values $v_i(t)$ of the congestion window $w_i(t)$, and to use a linear update rule for these quantities. This corresponds to a proportional update rule for the window length itself. The error term $\epsilon_i(t)$ is chosen to be of the form

$$\begin{aligned} \epsilon_i(t) &= \log\left(\frac{r_i}{y_i(t)\, w_i(t)}\right) \\ &= s_i - z_i(t) - v_i(t), \end{aligned} \tag{7}$$

where $s_i = \log r_i$ and $z_i(t) = \log y_i(t)$. Clearly ϵ_i has the same qualitative type of behaviour as e_i in that it is positive if the number of queued packets for user i is less than the desired number, and negative otherwise. We argue however that such an error function is more appropriate for positive quantities than a difference of terms.

The key challenge addressed here is the derivation of a decentralised model for the observation process

$$z_i(t) = \log\left(1 - \frac{d_i}{D_i(t)}\right), \tag{8}$$

in terms of the inputs $v_i(t)$, where we recall that $D_i(t)$ is the RTT seen by user i. Depending on the state of the network, which is determined by the other users as well as user i, this quantity (which is always negative) will be some function of $v_i(t)$. We postulate a linear model of the form

$$z_i(t) = G_i(z,\theta)\, v_i(t) + H_i(z,\theta)\, u_i(t)\,, \tag{9}$$

for a particular value of parameter vector θ_i, where $G_i(z,\theta_i)$ and $H_i(z,\theta_i)$ are respectively strictly proper and proper rational transfer functions with $H_i(z,\theta_i)$ and $H_i^{-1}(z,\theta_i)$ having all poles inside the unit circle. Here $u_i(t)$ denotes a white noise process. The role of H_i here is to model the effect *in closed loop* of other users on congestion seen by user i, while G_i models the effect of user i transmission rate on the congestion as seen by user i. Since $z_i(t)$ can take the value of $-\infty$ when there are no packets queued in the network, we need to include a hard limit on $z_i(t)$. This is consistent with the operation of TCP, where the receiver sets an upper bound on the congestion window size and thus $v_i(t)$ in our model.

Looking forward to the potential application of our identified model to the derivation of alternative congestion control strategies, we follow the methodology of [3], and choose the Recursive Prediction Error (RPE) identification procedure for its robustness properties. The prediction error is given by

$$\begin{aligned} \mu_i(t,\theta_i) &= z_i(t) - \hat{z}_i(t|t-1,\theta_i) \\ &= -H_i(\infty)\, H_i^{-1}(z,\theta_i)\, G_i(z,\theta_i)\, v_i(t) \\ &\quad + H_i(\infty)\, H_i^{-1}(z,\theta_i)\, z_i(t)\,. \end{aligned}$$

We recursively minimise the variance of the prediction error using the familiar Recursive Least Squares (RLS) approach, yielding estimates $\hat{\theta}_i(t)$ for θ_i. Figure 1 depicts the structure of the closed-loop system with the identification process included.

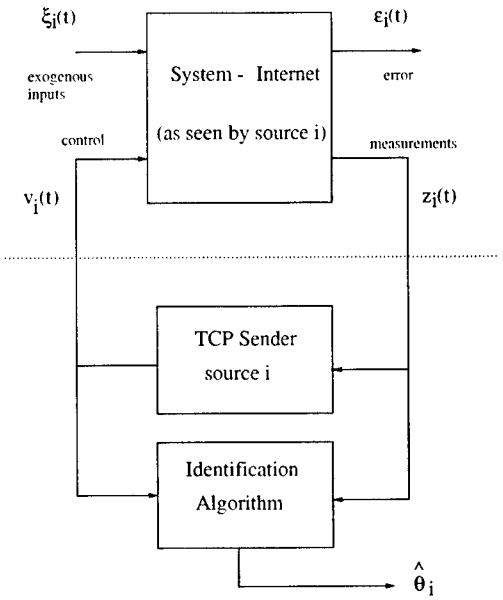


Fig. 1. TCP closed loop identification system model

4. NETWORK SIMULATIONS

We conducted a network simulation with 2 TCP senders and 2 receivers, transmitting over a link of length 1000 km and capacity 2.5 Mbyte/sec. Thus the round trip propagation delay = 667 μsec. A third (Poisson) source simulates background traffic on the link. For the purposes of this simulation, slow start and duplicate acknowledgements arising from timeouts were switched off, so that we can concentrate on the essential features of the congestion management function. A Vegas type control process was used for each TCP sender. The sample rate was 250 samples/sec and the simulation time was 50 s. The mean background Poissonian traffic was set at 1.75 Mbyte/sec with a jump to 2.25 MByte/sec at $t = 25$s. TCP sender 1 has a set point $r_1 = 0.1$, while for TCP sender 2, the set point was $r_2 = 1$.

Figures 3 - 5 show respectively, the instantaneous background traffic rate, and the sending rates for each TCP sender.The mean total combined rate in this example is 2.49 MByte/sec, a 99.5% utilisation. The data has been smoothed with a single pole filter with pole at 0.97, in order to aid readability. Our class of models has the form

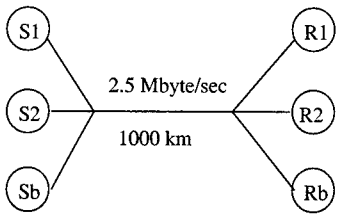


Fig. 2. TCP Network Simulation

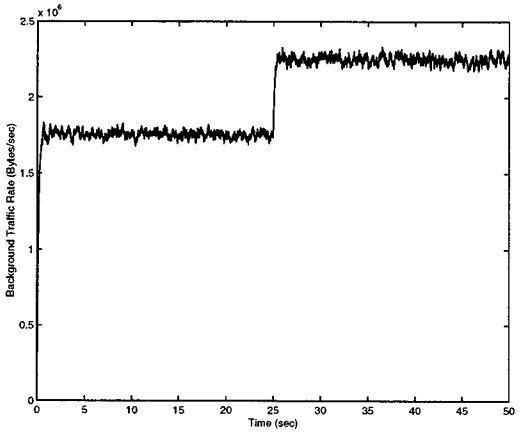


Fig. 3. Background Traffic Rate

$$G_i(z) = \frac{\sum_{j=1}^{M} b_{i,j}\, z^{-j}}{1 + \sum_{j=1}^{N} a_{i,j}\, z^{-j}} \qquad H_i(z) = \delta_i \, \frac{1 + \sum_{j=1}^{M} d_{i,j}\, z^{-j}}{1 + \sum_{j=1}^{N} c_{i,j}\, z^{-j}}, \tag{10}$$

where $M < N$. We chose (after some experimentation) $N = 10$, $M = 9$ and $P = 6$. Figure 6 shows an estimated power spectrum of the signal $z_1(t)$. The dynamic range is approximately 22 dB. Figure 7 shows the estimated power spectrum of the prediction error $\mu_1(t)$. The dynamic range has been reduced to about 7.5 dB indicating that the model has captured significant information about the system. It has both suppressed the low frequency variability by about 6 dB and significantly whitened the spectrum at mid to high frequencies. Figures 8 and 9 show the filter transfer function magnitudes $|G_1(z)|$ and $|H_1(z)|$ for user 1 at the end of the first 25 sec, and then at the end of 50 sec. Thus we can observe the effect of the change in background traffic intensity at $t = 25$ sec. It should be pointed out that the RPE technique does not explicitly yield the power of the exogenous disturbance white noise process $u_i(t)$, which is encompassed in the parameter δ_i in (10). Thus we can only estimate $H_i(\infty)^{-1} H_i(z)$. Thus figure 9 shows little variation in the colouring of the background traffic. However figure 8 illustrates both the drop in transmission rate for sender 1 together with its lower frequency content as evident from figure 4.

5. CONCLUSION

This paper has presented a linearised model for congestion at the transport layer of the internet as observed by a particular TCP sender in closed loop. The model incorporates both the effect of that user's transmission rate, and the aggregated effect on that user's observed congestion from all other users. Some preliminary simulations for a bottleneck queue suggest that such a modelling strategy can encompass some of the relevant information about observed congestion. These studies are very preliminary : there are many issues to investigate such as choice of model order and tracking capability. More extensive results are presented in [4]. Our main objective in formulating the TCP congestion control problem as we have here, is to replace the existing TCP control which employs no underlying model of the observed congestion (through RTTs), with a model based control approach. This approach is depicted in figure 10. The design of adaptive robust controllers as in [3] will build on the system identification studies presented here.

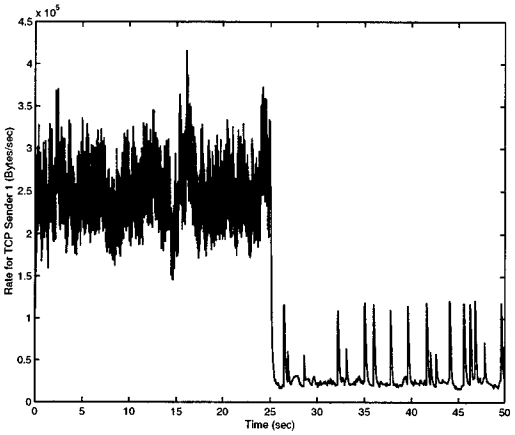


Fig. 4. Transmission Rate - TCP Sender 1

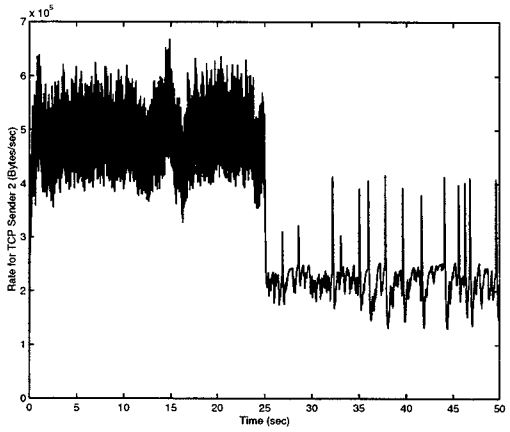


Fig. 5. Transmission Rate - TCP Sender 2

6. ACKNOWLEDGEMENT

The author acknowledges the support of the Co-operative Research Centre for Sensor Signal and Information Processing (CSSIP).

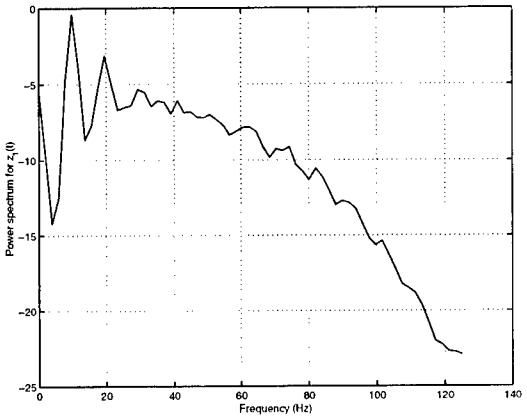


Fig. 6. Estimated PSD for user 1 data

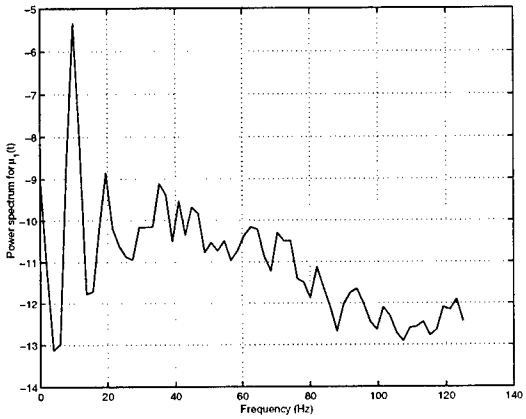


Fig. 7. Estimated PSD for user 1 prediction error

7. REFERENCES

[1] L. S. Brakmo and L. L. Peterson, "TCP Vegas : End to end congestion avoidance on a global Internet", *IEEE J. Sel. Areas Comms.*, v. 13, no. 8, October 1995.

[2] S. Low, L. L. Peterson and L. Wang, "Understanding TCP Vegas : Theory and Practice", preprint.

[3] R. R. Bitmead, M. Gevers and V. Wertz, *Adaptive Optimal Control*, New York : Prentice-Hall, 1990.

[4] L. B. White, "Internet Transport Layer System Identification", *IEEE Trans. Automatic Control*, in preparation.

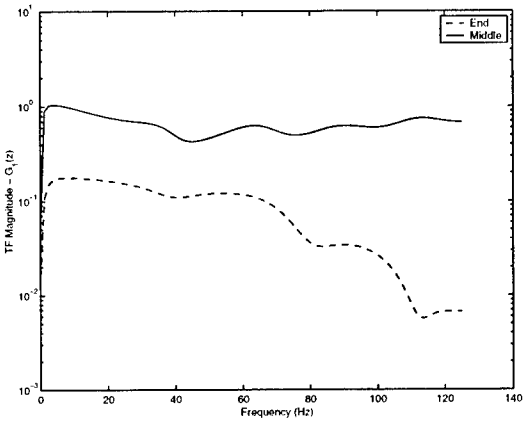


Fig. 8. Transfer Functions G_1

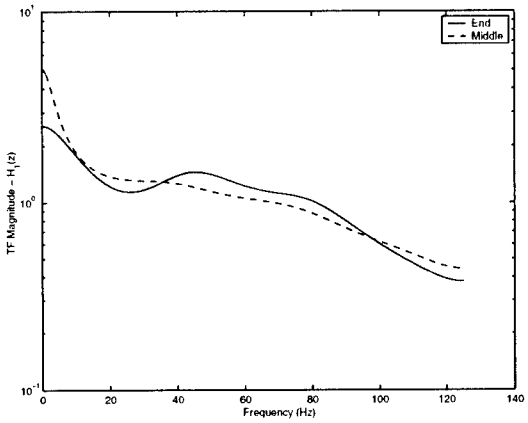


Fig. 9. Transfer Functions H_1

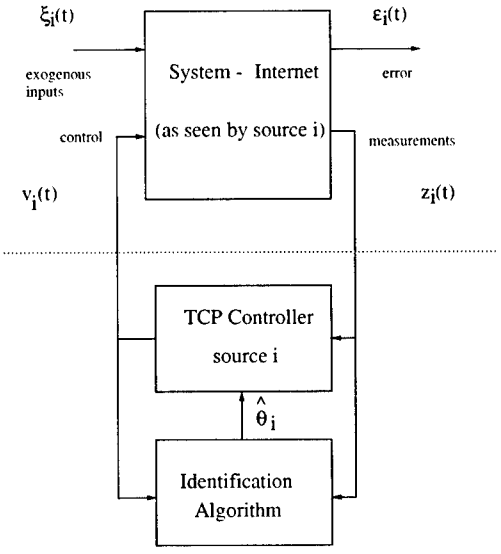


Fig. 10. TCP closed loop adaptive control system model

OPTIMAL DESIGN OF VARIABLE FRACTIONAL-DELAY DIGITAL FILTERS

Tian-Bo Deng

Department of Information Science
Faculty of Science, Toho University
Miyama 2-2-1, Funabashi, Chiba 274-8510, Japan

E-mail: deng@is.sci.toho-u.ac.jp

ABSTRACT

This paper presents a closed-form solution for obtaining the optimal coefficients of variable FIR filters with continuously adjustable fractional-delay (FD) response. The design is formulated as a weighted-least-squares (WLS) approximation problem without discretizing the frequency and fractional-delay parameters. Compared with the existing WLS method, the *discretization-free* one can yield a closed-form optimal solution with considerably reduced computational complexity.

1. INTRODUCTION

In many digital signal processing (DSP) applications and telecommunications, the frequency characteristics of digital filters are required to be continuously variable (adjustable). Such digital filters are referred to as *variable digital filters* [1, 2, 3, 4, 5, 6, 7]. Recently, the variable filters with adjustable fractional-delay (FD) response have been found useful in various applications [8]. The survey paper [8] provides a comprehensive review and comparison of most of the existing methods for designing and implementing such variable FD filters. Among those methods, the Lagrange interpolation method is considered to be the most attractive due to its simplicity [8], but the frequency response of the resulting Lagrange interpolator cannot be uniformly balanced in the entire frequency band, i.e., the frequency response in the low frequency region is better than high frequency response as demonstrated in [9]. Therefore, it is difficult to achieve a satisfactory design with low filter orders in the whole frequency band by using the Lagrange interpolation method. To solve this problem, paper [9] proposed a general technique using the weighted-least-squares (WLS) method, which can yield a more satisfactory design with lower filter order than the Lagrange interpolator. However, the WLS method requires sampling the frequency ω and fractional-delay p. That is, the parameter discretizations are necessary. Usually, the sampling grids of the parameter discretizations must be dense enough to guarantee high design accuracy, which in turn increases the computational complexity needed in filter design process. Therefore, it is desirable to derive a closed-form optimal solution without discretizations.

This paper presents a new method for obtaining the closed-form optimal solution of variable FD filter coefficients by formulating the FD filter design as an integral WLS approximation problem without parameter discretizations [10]. Compared with the existing WLS method in [9], the *discretization-free* method can achieve higher design accuracy with considerably reduced computational complexity. A design example is given to demonstrate its effectiveness.

2. CLOSED-FORM DESIGN

In this paper, we consider the polynomial-based variable FD filter

$$H(z,p) = \sum_{n=-N_1}^{N_2} a_n(p) z^{-n} \tag{1}$$

where N_1 and N_2 are positive integers determining the filter order

$$N = N_1 + N_2$$

whose values are selected as

$$N_1 = \frac{N-1}{2}, \quad N_2 = \frac{N+1}{2}$$

for odd N, and

$$N_1 = N_2 = \frac{N}{2}$$

for even N, respectively. The filter coefficients $a_n(p)$ are expressed as the polynomials of the fractional-delay p as

$$a_n(p) = \sum_{k=0}^{K} a(n,k) p^k \tag{2}$$

where $p \in [-0.5, 0.5]$. Thus the transfer function $H(z,p)$ can be rewritten as

$$H(z,p) = \sum_{n=-N_1}^{N_2} \sum_{k=0}^{K} a(n,k) p^k z^{-n} = \boldsymbol{a}^T (\boldsymbol{p} \otimes \boldsymbol{z}) \tag{3}$$

where the notation $\otimes$ denotes the Kronecker product

$$\begin{aligned} \boldsymbol{p} &= \begin{bmatrix} 1 & p & p^2 & \cdots & p^K \end{bmatrix}^T \\ \boldsymbol{z} &= \begin{bmatrix} z^{N_1} & \cdots & z & 1 & z^{-1} & \cdots & z^{-N_2} \end{bmatrix}^T \end{aligned} \tag{4}$$

and vector $\boldsymbol{a}$ is the column string of the matrix $\boldsymbol{A} = [a(n,k)]$.

The actual variable frequency response is

$$H(\omega,p) = \boldsymbol{a}^T (\boldsymbol{p} \otimes \boldsymbol{\omega}) \tag{5}$$

where the complex vector ω is

$$\boldsymbol{\omega} = \begin{bmatrix} e^{jN_1\omega} & \cdots & e^{j\omega} & 1 & e^{-j\omega} & \cdots & e^{-jN_2\omega} \end{bmatrix}^T.$$

Also, the desired variable frequency response is given by

$$H_d(\omega,p) = e^{-j\omega p} \tag{6}$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

where the normalized frequency ω is in the range $\omega \in [0, \pi]$, and the fractional-delay p is continuously variable in the range $p \in [-0.5, 0.5]$. Our objective here is to find the optimal coefficient vector $\boldsymbol{a}$ such that the weighted squared error of the variable frequency response

$$J(\boldsymbol{a}) = \int_0^{\pi} \int_{-0.5}^{0.5} W(\omega, p)|H(\omega, p) - H_d(\omega, p)|^2 d\omega dp \quad (7)$$

is minimized. For deriving a closed-form solution, we assume that the 2-D weighting function $W(\omega, p)$ is separable as

$$W(\omega, p) = W_1(\omega) W_2(p) \quad (8)$$

which is the product of 1-D stepwise functions

$$\begin{array}{ll} W_1(\omega) = \alpha_l & \text{for } \omega \in [\omega_{l-1}, \omega_l), \quad l = 1, 2, \cdots, L \\ W_2(p) = \beta_m & \text{for } p \in [p_{m-1}, p_m), \quad m = 1, 2, \cdots, M \end{array} \quad (9)$$

where α_l and β_m are constants. Since frequency response error is

$$e(\omega, p) = H(\omega, p) - H_d(\omega, p) = \boldsymbol{a}^T(\boldsymbol{p} \odot \boldsymbol{\omega}) - e^{-j\omega p}$$

thus

$$\begin{aligned} & |e(\omega, p)|^2 \\ &= \left[\boldsymbol{a}^T(\boldsymbol{p} \odot \boldsymbol{\omega}) - e^{-j\omega p}\right]\left[(\boldsymbol{p} \odot \boldsymbol{\omega})^* \boldsymbol{a} - e^{j\omega p}\right] \\ &= \boldsymbol{a}^T(\boldsymbol{p} \odot \boldsymbol{\omega})(\boldsymbol{p}^T \odot \boldsymbol{\omega}^*)\boldsymbol{a} - 2\mathrm{Re}\left[\boldsymbol{a}^T(\boldsymbol{p} \odot \boldsymbol{\omega}) e^{j\omega p}\right] + 1 \\ &= e_1(\omega, p) - 2e_2(\omega, p) + 1 \end{aligned} \quad (10)$$

where $[\ \cdot\ ]^*$ means the Hermitian adjoint, and $\mathrm{Re}[\ \cdot\ ]$ denotes the real part of $[\ \cdot\ ]$. By using the property of Kronecker product that

$$(\boldsymbol{A} \odot \boldsymbol{B})(\boldsymbol{C} \odot \boldsymbol{D}) = (\boldsymbol{AC}) \odot (\boldsymbol{BD}) \quad (11)$$

we obtain the first term in (10) as

$$\begin{aligned} e_1(\omega, p) &= \boldsymbol{a}^T(\boldsymbol{p} \odot \boldsymbol{\omega})(\boldsymbol{p}^T \odot \boldsymbol{\omega}^*)\boldsymbol{a} \\ &= \boldsymbol{a}^T\left[(\boldsymbol{pp}^T) \odot (\boldsymbol{\omega\omega}^*)\right]\boldsymbol{a} \end{aligned} \quad (12)$$

where

$$\boldsymbol{pp}^T = \begin{bmatrix} 1 & p & p^2 & \cdots & p^K \\ p & p^2 & p^3 & \cdots & p^{K+1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ p^K & p^{K+1} & p^{K+2} & \cdots & p^{2K} \end{bmatrix} \quad (13)$$

is a Hankel matrix, and $\boldsymbol{\omega\omega}^*$ is a complex Hermitian Toeplitz matrix. Furthermore, the second term in (10) is

$$e_2(\omega, p) = \mathrm{Re}\left[\boldsymbol{a}^T(\boldsymbol{p} \odot \boldsymbol{\omega}) e^{j\omega p}\right] = \boldsymbol{a}^T(\boldsymbol{p} \odot \boldsymbol{q}) \quad (14)$$

where the real vector $\boldsymbol{q}$ is defined as

$$\boldsymbol{q} = [\cos\omega(p + N_1) \ \cdots \ \cos p\omega \ \cdots \ \cos\omega(p - N_2)]^T. \quad (15)$$

Consequently, substituting (8) and (10) into (7) obtains

$$J(\boldsymbol{a}) = J_1(\boldsymbol{a}) - 2J_2(\boldsymbol{a}) + J_3 \quad (16)$$

where

$$\begin{aligned} J_1(\boldsymbol{a}) &= \int_0^{\pi} \int_{-0.5}^{0.5} e_1(\omega, p) d\omega dp \\ &= \boldsymbol{a}^T \left\{ \left[\int_{-0.5}^{0.5} W_2(p) \boldsymbol{pp}^T dp \right] \odot \left[\int_0^{\pi} W_1(\omega) \boldsymbol{\omega\omega}^* d\omega \right] \right\} \boldsymbol{a} \\ &= \boldsymbol{a}^T(\boldsymbol{P} \odot \boldsymbol{\Omega}_c)\boldsymbol{a}. \end{aligned} \quad (17)$$

The above matrix $\boldsymbol{P}$ can be computed by

$$\boldsymbol{P} = \int_{-0.5}^{0.5} W_2(p) \boldsymbol{pp}^T dp = \sum_{m=1}^{M} \beta_m \int_{p_{m-1}}^{p_m} \boldsymbol{pp}^T dp. \quad (18)$$

It should be noted that since $\boldsymbol{P}_m$ is a Hankel matrix, thus only its first column and last row are needed for generating the $\boldsymbol{P}_m$. Moreover, it is clear that the resulting matrix $\boldsymbol{P}$ in (18) is also a Hankel matrix whose size is (K+1)-by-(K+1). On the other hand, although the $\boldsymbol{\Omega}_c$ in (17) is a complex Hermitian matrix, we only need to determine its real part for evaluating $J_1(\boldsymbol{a})$ since $J_1(\boldsymbol{a})$ is real-valued. The real part of $\boldsymbol{\Omega}_c$ is calculated by

$$\boldsymbol{\Omega} = \mathrm{Re}[\boldsymbol{\Omega}_c] = \sum_{l=1}^{L} \alpha_l \boldsymbol{\Omega}_l \quad (19)$$

whose elements are

$$\Omega_l(i, j) = \begin{cases} \omega_l - \omega_{l-1} & \text{if } i = j \\ \dfrac{\sin[(i-j)\omega_l] - \sin[(i-j)\omega_{l-1}]}{i - j} & \text{for } i \neq j \end{cases} \quad (20)$$

and $i, j = 1, 2, \cdots, (N + 1)$.

It should also be pointed out that since $\boldsymbol{\Omega}_l$ is a symmetric Toeplitz matrix, thus only its first row needs to be computed for generating $\boldsymbol{\Omega}_l$. As a result, the resulting matrix $\boldsymbol{\Omega}$ is also a symmetric Toeplitz matrix whose size is (N+1)-by-(N+1).

Based on the above derivations of matrices $\boldsymbol{P}$ and $\boldsymbol{\Omega}$, the $J_1(\boldsymbol{a})$ in (16) can be obtained as

$$J_1(\boldsymbol{a}) = \boldsymbol{a}^T(\boldsymbol{P} \odot \boldsymbol{\Omega})\boldsymbol{a}. \quad (21)$$

In addition, the second term $J_2(\boldsymbol{a})$ in (16) is evaluated by

$$\begin{aligned} J_2(\boldsymbol{a}) &= \int_0^{\pi} \int_{-0.5}^{0.5} W_1(\omega) W_2(p) \left[\boldsymbol{a}^T(\boldsymbol{p} \odot \boldsymbol{q})\right] d\omega dp \\ &= \boldsymbol{a}^T \left\{ \int_{-0.5}^{0.5} W_2(p) \left[\boldsymbol{p} \odot \int_0^{\pi} W_1(\omega) \boldsymbol{q} d\omega \right] dp \right\} \\ &= \boldsymbol{a}^T \left[\int_{-0.5}^{0.5} W_2(p)(\boldsymbol{p} \odot \boldsymbol{u}) dp \right] \\ &= \boldsymbol{a}^T \boldsymbol{v} \end{aligned} \quad (22)$$

where

$$\boldsymbol{u} = \int_0^{\pi} W_1(\omega) \boldsymbol{q} d\omega = \sum_{l=1}^{L} \alpha_l \boldsymbol{u}_l \quad (23)$$

and the elements of vector $\boldsymbol{u}_l$ are determined by

$$\begin{aligned} u_l(n) &= \int_{\omega_{l-1}}^{\omega_l} q(n) d\omega \\ &= \begin{cases} \omega_l - \omega_{l-1} & \text{if } \gamma = 0 \\ \dfrac{\sin(\gamma\omega_l) - \sin(\gamma\omega_{l-1})}{\gamma} & \text{if } \gamma \neq 0 \end{cases} \end{aligned} \quad (24)$$

with $\gamma = p + N_1 - n + 1$, and $n = 1, 2, \cdots, (N + 1)$.

Furthermore, the vector $\boldsymbol{v}$ in (22) can be calculated as

$$\boldsymbol{v} = \int_{-0.5}^{0.5} W_2(p)(\boldsymbol{p} \odot \boldsymbol{u}) dp = \sum_{l=1}^{L} \sum_{m=1}^{M} \alpha_l \beta_m \boldsymbol{v}_{lm} \quad (25)$$

where

$$\boldsymbol{v}_{lm} = \int_{p_{m-1}}^{p_m} (\boldsymbol{p} \odot \boldsymbol{u}_l) dp$$

whose elements can be computed by using numerical integration as

$$v_{lm}(i) = \int_{p_{m-1}}^{p_m} p^k u_l(n) dp \tag{26}$$

with

$$\begin{array}{l} k = 0, 1, \cdots, K, \quad n = 1, 2, \cdots, (N+1) \\ i = k(N+1) + n. \end{array} \tag{27}$$

Lastly, the third term in (16) is

$$J_3 = \int_0^{\pi} \int_{-0.5}^{0.5} W_1(\omega) W_2(p) d\omega dp = \text{constant}. \tag{28}$$

Thus, the final error function (16) can be formed by combining the derived $J_1(\boldsymbol{a})$, $J_2(\boldsymbol{a})$, and J_3 as

$$J(\boldsymbol{a}) = \boldsymbol{a}^T (\boldsymbol{P} \otimes \boldsymbol{\Omega}) \boldsymbol{a} - 2\boldsymbol{a}^T \boldsymbol{v} + \text{constant}. \tag{29}$$

The optimal solution for $\boldsymbol{a}$ can be found by setting the derivative of $J(\boldsymbol{a})$ with respect to $\boldsymbol{a}$ to zero, i.e.,

$$\frac{\partial J(\boldsymbol{a})}{\partial \boldsymbol{a}} = \left[(\boldsymbol{P} \otimes \boldsymbol{\Omega}) + (\boldsymbol{P} \odot \boldsymbol{\Omega})^T \right] \boldsymbol{a} - 2\boldsymbol{v} = 0. \tag{30}$$

Since $\boldsymbol{P}$ is a Hankel matrix, and $\boldsymbol{\Omega}$ is a symmetric Toeplitz matrix, thus

$$\boldsymbol{P}^T = \boldsymbol{P}, \quad \boldsymbol{\Omega}^T = \boldsymbol{\Omega}$$

which leads to

$$(\boldsymbol{P} \otimes \boldsymbol{\Omega})^T = \boldsymbol{P}^T \otimes \boldsymbol{\Omega}^T = \boldsymbol{P} \otimes \boldsymbol{\Omega}.$$

As a result, we obtain

$$\boldsymbol{a} = (\boldsymbol{P} \otimes \boldsymbol{\Omega})^{-1} \boldsymbol{v} = (\boldsymbol{P}^{-1} \otimes \boldsymbol{\Omega}^{-1}) \boldsymbol{v}. \tag{31}$$

Since the direct inversion of matrices $\boldsymbol{P}$ and $\boldsymbol{\Omega}$ may suffer from ill-conditioning problem due to their large condition numbers, some efficient measure must be taken to avoid the numerical problem. In considering that both $\boldsymbol{P}$ and $\boldsymbol{\Omega}$ are positive definite matrices, and they can be decomposed by using the Cholesky factorization as

$$\boldsymbol{P} = \boldsymbol{R}^T \boldsymbol{R}, \quad \boldsymbol{\Omega} = \boldsymbol{S}^T \boldsymbol{S} \tag{32}$$

where $\boldsymbol{R}$ and $\boldsymbol{S}$ are upper triangular matrices, thus we can indirectly compute the inverses of $\boldsymbol{P}$ and $\boldsymbol{\Omega}$ as

$$\boldsymbol{P}^{-1} = \boldsymbol{R}^{-1} \boldsymbol{R}^{-T}, \quad \boldsymbol{\Omega}^{-1} = \boldsymbol{S}^{-1} \boldsymbol{S}^{-T}. \tag{33}$$

Substituting (33) into (31) and applying the property (11) yields the closed-form optimal solution

$$\begin{aligned} \boldsymbol{a} &= \left[(\boldsymbol{R}^{-1} \boldsymbol{R}^{-T}) \otimes (\boldsymbol{S}^{-1} \boldsymbol{S}^{-T}) \right] \boldsymbol{v} \\ &= (\boldsymbol{R}^{-1} \otimes \boldsymbol{S}^{-1})(\boldsymbol{R}^{-T} \otimes \boldsymbol{S}^{-T}) \boldsymbol{v} \\ &= (\boldsymbol{R}^{-1} \otimes \boldsymbol{S}^{-1}) \left[(\boldsymbol{R}^{-T} \otimes \boldsymbol{S}^{-T}) \boldsymbol{v} \right]. \end{aligned} \tag{34}$$

It should be noticed that the re-grouping in the last expression of (34) is very important for ensuring a numerically stabilized optimal solution. Finally, some additional remarks should be given to clarify the following points.

(1). Once the optimal coefficient vector $\boldsymbol{a}$ in (34) is obtained, we can yield the coefficients $a(n,k)$ in (3). Substituting different values of p into (3) results in a variable filter $H(z,p)$ with different fractional-delays, which can be implemented by using Farrow structure [11].

(2). Since the range of variable fractional-delay p is $p \in [-0.5, 0.5]$, all non-integer delays with integer and fractional parts can be covered by just cascading $H(z,p)$ with D delay elements (z^{-D}), where D is a positive integer. Evidently, if $D \geq N_1$, $z^{-D} H(z,p)$ becomes causal.

3. DESIGN EXAMPLE

In this section, we illustrate the proposed discretization-free method by designing a variable FD filter with the same design specification as that in [9]: the maximum absolute error of variable frequency response defined by

$$e_{max} = \max \{ \epsilon(\omega, p) | \omega \in [0, 0.9\pi], p \in [-0.5, 0.5] \} \tag{35}$$

where

$$\epsilon(\omega, p) = 20 \log_{10} |H(\omega, p) - H_d(\omega, p)| \tag{36}$$

should not exceed the level of -100dB for any fractional-delay p and any frequency ω in the above range of our interest, i.e., $0 \leq \omega \leq 0.9\pi$, and $-0.5 \leq p \leq 0.5$.

We have tackled this design problem by using the method [9] and the proposed discretization-free method, and found that if the design parameters

$$\begin{array}{l} N = 65, \quad K = 7 \\ W_1(\omega) = \begin{cases} 0.64 & \text{for } \omega \in [0, 0.55\pi) \\ 4.9 & \text{for } \omega \in [0.55\pi, 0.85\pi) \\ 37 & \text{for } \omega \in [0.85\pi, 0.8996\pi) \\ 0 & \text{for } \omega \in [0.8996\pi, \pi] \end{cases} \\ W_2(p) = \begin{cases} 53 & \text{for } p \in [-0.5, -0.4) \\ 0.2 & \text{for } p \in [-0.4, 0.4) \\ 8 & \text{for } p \in [0.4, 0.5] \end{cases} \end{array} \tag{37}$$

are selected, the above design specification can be satisfied by the proposed method. To compare the discretization-free method with the method in [9], the maximum absolute deviation of the frequency response defined in (35) and the root-mean-squared error defined by

$$e_2 = \left[\int_0^{0.9\pi} \int_{-0.5}^{0.5} |H(\omega, p) - H_d(\omega, p)|^2 d\omega dp \right]^{1/2} \tag{38}$$

are used to evaluate the design accuracy, and the number of floating-point-operations (Flops) required for determining the optimal coefficient vector $\boldsymbol{a}$ is used to compare the computational complexity. Table 1 shows the design results by the two methods. It is observed from Table 1 that the proposed discretization-free method can meet the design requirement since the maximum deviation of the frequency response is below -100dB, but the maximum deviation by the method [9] is -99.9208dB. Moreover, the proposed method requires only 5.58% of the Flops required by method in [9]. Therefore, a significant reduction of computational complexity has been achieved. Fig. 1 depicts the variable fractional-delay response in the range $\omega \in [0, 0.9\pi]$ and $p \in [-0.5, 0.5]$. It

is observed that the variable fractional-delay response of the designed variable FD filter is considerably flat in the entire range. Also, the absolute error of variable frequency response is illustrated in Fig. 2, which shows significantly small error in our interested range.

4. CONCLUSION

A weighted-least-squares (WLS) method without parameter discretizations has been proposed for deriving the optimal closed-form solution of variable FD filter coefficients. Since the proposed method does not require sampling the frequency ω and fractional-delay p in obtaining the final closed-form solution, thus a considerable reduction of computational complexity and higher design accuracy can be achieved. An illustrative example has been presented to demonstrate the effectiveness of the proposed method.

5. REFERENCES

[1] P. Jarske, Y. Neuvo, and S. K. Mitra, "A simple approach to the design of linear phase FIR digital filters with variable characteristics", *Signal Processing*, vol. 14, no. 4, pp. 313-326, June 1988.

[2] R. Zarour and M. M. Fahmy, "A design technique for variable digital filters", *IEEE Trans. Circuits Syst.*, vol. 36, no. 11, pp. 1473-1478, Nov. 1989.

[3] T.-B. Deng, "Design of recursive 1-D variable filters with guaranteed stability", *IEEE Trans. Circuits Syst. II*, vol. 44, no. 9, pp. 689-695, Sept. 1997.

[4] R. Zarour and M. M. Fahmy, "A design technique for variable two-dimensional recursive digital filters", *Signal Processing*, vol. 17, no. 2, pp. 175-182, June 1989.

[5] T.-B. Deng, "Design of variable 2-D linear phase recursive digital filters with guaranteed stability", *IEEE Trans. Circuits Syst. I*, vol. 45, no. 8, pp. 859-863, Aug. 1998.

[6] T.-B. Deng, "Design of linear phase variable 2-D digital filters using real-complex decomposition", *IEEE Trans. Circuits Syst. II*, vol. 45, no. 3, pp. 330-339, Mar. 1998.

[7] G. Stoyanov and M. Kawamata, "Variable digital filters", *J. Signal Processing*, vol. 1, no. 4, pp. 275-289, July 1997.

[8] T. I. Laakso, V. Valimaki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay: Tools for fractional delay filter design", *IEEE Signal Processing Mag.*, vol. 13, no. 1, pp. 30-60, Jan. 1996.

[9] A. Tarczynski, G. D. Cain, E. Hermanowicz, and M. Rojewski, "WLS design of variable frequency response FIR filters", *Proc. 1997 IEEE Int. Symp. Circuits and Systems*, pp. 2244-2247, Hong Kong, June 9-12, 1997.

[10] T.-B. Deng, "Discretization-free design of variable fractional-delay FIR digital filters", to appear in *IEEE Trans. Circuits Syst. II.*

[11] C. W. Farrow, "A continuously variable digital delay elements", *Proc. 1988 IEEE Int. Symp. Ciruits and Systems*, vol. 3, pp. 2641-2645, Espoo, Finland, June 6-9, 1988.

Table 1: Design Errors and Computational Complexity

	Method in [9]	Proposed Method
ϵ_{max} (dB)	-99.9208	-100.3683
ϵ_2	4.4931×10^{-4}	4.4147×10^{-4}
Flops Used	84057162	4694097

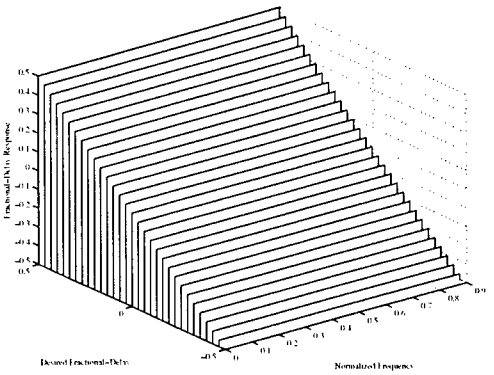


Fig. 1. Variable Fractional-Delay Response.

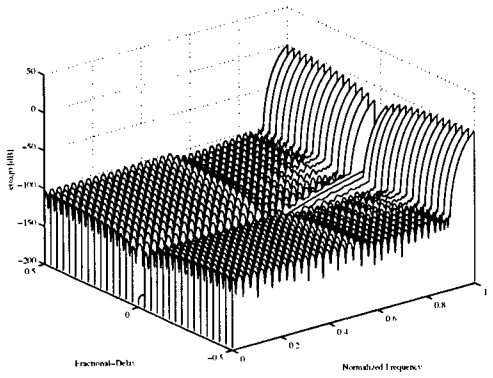


Fig. 2. Absolute Error of Variable Frequency Response.

DESIGN OF DIGITAL FILTERS WITH AMPLITUDE AND GROUP DELAY SPECIFICATIONS*

Zhuquan Zang† *Sven Nordholm*† *Sven Nordebo*‡ *Antonio Cantoni*†

† Australian Telecommunications Research Institute
Curtin University of Technology, Perth, WA 6102, Australia
‡ Dept of Telecommunications and Signal Processing
Blekinge Institute of Technology, S-372 25 Ronneby, Sweden

ABSTRACT

This paper considers the design of a digital filter with prescribed magnitude and group delay specifications. First, we outline the derivation of the phase and group delay functions of an Nth order digital Laguerre filter and show that the group delay function of the filter can be written as a ratio of quadratic functions in the filter coefficients. Then, we formulate our filter design problem as a constrained L_2 space minimization problem in which the performance requirement on the group delay and magnitude in the passband are treated as constraints while minimizing the L_2 norm of the error function between the designed and the desired filters. Methods for solving the proposed nonlinear optimization problem are outlined. Numerical results are presented to illustrate the usefulness of the proposed method. As a special case, corresponding results for general FIR filters are also derived.

1. Introduction

In filter design literature, the problem of designing linear phase FIR filters with desired magnitude characteristics has been well studied [7]. Algorithms and efficient softwares are now readily available for the design of such filters. Linear-phase restriction imposed on the whole frequency range converts the filter design problem into a real approximation problem for which efficient classical methods for real approximation such as the Remez exchange algorithms can be utilised to find the optimal solution. However, linear phase filter with short transition bands introduce large delays – an undesirable feature in many applications such as adaptive filtering in acoustic echo cancellation. Moreover, the linear-phase restriction is not needed in the stopband. In practice, filters with a well-specified phase or group delay functions in the passband together with desired magnitude characteristics are highly desirable in a variety of areas such as communication channel equalisation, chirp processing, and optimal beamforming with unequally spaced sensor arrays.

Imposing phase or group delay requirement only in the passband results in a complex approximation problem. In recent years, such problems have attracted the attention of many researchers, see [2], [5], [6] and the references therein. In this paper, using a filter structure based on the set of orthonormal Laguerre functions, we shall investigate the design of digital IIR filters which satisfy prescribed magnitude and group delay requirements. First, we show that the group delay function of an Nth order digital Laguerre filter can be written as a ratio of quadratic functions with respect to the filter coefficient vector. Then, we utilise the expression of the group delay function to pose our filter design problem as a constrained L_2 space optimization problem. Performance requirements on the group delay function and magnitude spectrum are treated as constraints while minimising the L_2 norm of the error function between the designed and the desired filters. Methods for solving the nonlinear optimization problem are briefly outlined. Numerical results are presented to illustrate the usefulness of the proposed method. The results obtained are also applicable to the design of FIR filters which is a special case of the digital Laguerre filters.

*This project was partially supported by a research grant from the Australian Research Council

2. Laguerre Filter and Its Phase and Group Delay Functions

Consider the following Nth order Laguerre filter

$$H_L(z) = \sum_{k=0}^{N-1} x_k \phi_k(z) \tag{1}$$

where $\phi_k(z)$'s are the Laguerre functions defined as

$$\phi_k(z) = \frac{\sqrt{1-a^2}}{1-az^{-1}} \left(\frac{z^{-1}-a}{1-az^{-1}} \right)^k, \quad k = 0, 1, 2, ...$$

Define

$$A_0(z) = \frac{\sqrt{1-a^2}}{1-az^{-1}}, \quad A_{ap}(z) = \frac{z^{-1}-a}{1-az^{-1}} \tag{2}$$

where $A_0(z) = \phi_0(z)$ is a lowpass filter for $0 < a < 1$ and a highpass filter for $-1 < a < 0$, $A_{ap}(z)$ is an allpass filter. Note that $H_L(z)$ is an N-tap FIR filter for $a = 0$.

2.1. The Phase Function of $H_L(z)$

Let us outline the derivation of the phase function of $H_L(z)$. Detailed derivation can be found in [10].

First, it is easy to verify that $A_0(e^{j\omega})$ can be written as

$$A_0(e^{j\omega}) = \frac{\sqrt{1-a^2}}{\sqrt{1-2a\cos\omega+a^2}} e^{-j \arctan\left[\frac{a\sin\omega}{1-a\cos\omega}\right]}$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

Therefore, the phase function of $A_0(z)$, denoted as $\theta_0(\omega)$, can be expressed as

$$\theta_0(\omega) \triangleq -\arg(A_0(e^{j\omega})) = \arctan\left[\frac{a\sin\omega}{1-a\cos\omega}\right] \tag{3}$$

Similarly, the phase function of the allpass filter $A_{ap}(e^{j\omega})$ can be written as

$$\theta_{ap} = \omega + 2\arctan\left[\frac{a\sin\omega}{1-a\cos\omega}\right] \tag{4}$$

Combining (4) with (3), we can conclude that the phase function, denoted as $\theta_k(\omega)$, of $\phi_k(z)$ is given by

$$\theta_k(\omega) \triangleq k\omega + (2k+1)\arctan\left[\frac{a\sin\omega}{1-a\cos\omega}\right] \tag{5}$$

As a result, $H_L(e^{j\omega})$ can be written as

$$\begin{aligned} H_L(e^{j\omega}) &= \sum_{k=0}^{N-1} x_k\phi_k(e^{j\omega}) \\ &= \gamma(\omega)\sum_{k=0}^{N-1} x_k e^{-j\theta_k(\omega)} \end{aligned}$$

where $\gamma(\omega)$ denotes the magnitude spectrum of $A_0(z)$

$$\gamma(\omega) = \frac{\sqrt{1-a^2}}{\sqrt{1-2a\cos\omega+a^2}}$$

Hence, the phase function of $H_L(z)$ can be expressed as

$$\theta_{H_L}(\omega) = \arctan\left[\frac{\sum_{k=0}^{N-1} x_k\sin\theta_k(\omega)}{\sum_{k=0}^{N-1} x_k\cos\theta_k(\omega)}\right] \tag{6}$$

2.2. The Group Delay Function of $H_L(z)$

Using (6), the group delay function of $H_L(z)$ is given by,

$$\tau_g(\omega) = \frac{d}{d\omega}\arctan\left[\frac{\sum_{k=0}^{N-1} x_k\sin\theta_k(\omega)}{\sum_{k=0}^{N-1} x_k\cos\theta_k(\omega)}\right]$$

It can be proved [10], through some algebra, that τ_g can be expressed as

$$\tau_g(\omega) = \frac{x^T P_\Gamma(\omega)x}{x^T P(\omega)x} \tag{7}$$

where $x = [x_0,\ x_1,\ \cdots,\ x_{N-1}]^T$ and

$$\begin{aligned} P(\omega) &= C(\omega)C^T(\omega) + S(\omega)S^T(\omega) \\ P_\Gamma(\omega) &= (C(\omega)C^T(\omega) + S(\omega)S^T(\omega))\Gamma(\omega) \\ \Gamma(\omega) &= \mathrm{diag}\{\Gamma_0(\omega), \Gamma_1(\omega), ..., \Gamma_{N-1}(\omega)\} \\ \Gamma_k(\omega) &= k + (2k+1)\frac{a(\cos\omega - a)}{1-2a\cos\omega+a^2} \end{aligned}$$

$$\begin{aligned} C(\omega) &= [\cos\theta_0(\omega),\ \cos\theta_1(\omega),\ ...,\ \cos\theta_{N-1}(\omega)]^T \\ S(\omega) &= [\sin\theta_0(\omega),\ \sin\theta_1(\omega),\ ...,\ \sin\theta_{N-1}(\omega)]^T \end{aligned}$$

$\theta_k(\omega)$ is given by (5). From (7) we know that the group delay function of an Nth order Laguerre filter is a ratio of two quadratic functions with respect to the filter coefficient vector x.

Note that, in general, P_Γ is not a symmetric matrix. However, since for any real x, $x^T P_\Gamma(\omega)x$ is always real for any ω, it can be concluded [?] that

$$x^T P_\Gamma(\omega)x = \frac{1}{2}x^T(P_\Gamma(\omega) + P_\Gamma^T(\omega)x$$

Obviously, $Q_\Gamma \triangleq \frac{1}{2}(P_\Gamma(\omega) + P_\Gamma^T(\omega))$ is a symmetric matrix. As a result, we have

$$\tau_g(\omega) = \frac{x^T Q_\Gamma(\omega)x}{x^T P(\omega)x} \tag{8}$$

2.3. Phase and Group Delay Functions of an FIR Filter

For a given FIR fiter $H_F(z) = \sum_{k=0}^{N-1} x_k z^{-k}$, its phase and group delay functions can be readily calculated by setting $a = 0$ in the expression (6) and (8). It is easy to verify that the phase function of $H_F(z)$ is

$$\theta_F(\omega) = \arctan\left(\frac{\sum_{k=0}^{N-1} x_k\sin k\omega}{\sum_{k=0}^{N-1} x_k\cos k\omega}\right) \tag{9}$$

The corresponding group delay function is

$$\tau_{g_F}(\omega) = \frac{x^T Q_{\Gamma_F(\omega)}x}{x^T P_F(\omega)x} \tag{10}$$

where

$$\begin{aligned} Q_{\Gamma_F}(\omega) &= \frac{1}{2}(P_{\Gamma_F}(\omega) + P_{\Gamma_F}^T(\omega)) \\ P_{\Gamma_F}(\omega) &= P_F(\omega)\Gamma_F(\omega) \\ P_F(\omega) &= C_F(\omega)C_F^T(\omega) + S_F(\omega)S_F^T(\omega)) \\ \Gamma_F(\omega) &= \mathrm{diag}\{0, 1, 2, ..., N-1\} \\ C_F(\omega) &= [1,\ \cos\omega,\ ...,\ \cos(N-1)\omega]^T \\ S_F(\omega) &= [0,\ \sin\omega,\ ...,\ \sin(N-1)\omega]^T \end{aligned}$$

As in the digital Laguerre filter case, the group delay function of an N tap FIR filter can be expressed as a ratio of two quadratic functions in the filter coefficient vector x.

Remark: Note that an IIR filter is the ratio of two FIR filters. Therefore, the group delay function of a given IIR filter can be expressed as the difference between the group delay function of the numerator and that of the denominator.

3. Frequency Domain Digital Filter Design with Magnitude and Group Delay Constraints

It has been advocated [1], [8] (at least in the context of linear phase FIR filter design) that minimizing the L_2 norm of a suitably chosen error function subject to relevant peak constraints is one of the most meaningful approaches for filter design. In the following, we adopt this philosophical point of view in the formulation and solution of our filter design problem as a constrained L_2 space nonconvex optimization problem.

3.1. Problem formulation and conversion

To start with, let us consider those filters whose frequency responses can be expressed as a linear combination of a finite set of the orthonormal Laguerre functions ϕ_k's. That is

$$H_L(z) = \sum_{k=0}^{N-1} x_k \phi_k(z) \tag{11}$$

Without loss of generality, We consider the following constrained filter design problem, denoted as problem (**P**)

$$\min_x \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi(\omega) |H_L(\omega) - H_d(\omega)|^2 d\omega \tag{12}$$

subject to

$$|H_L(e^{j\omega}) - H_d(\omega)| \leq \delta_m, \quad |\tau_g(\omega) - \tau_d(\omega)| \leq \delta_g, \quad \omega \in \Omega_p$$

where $\Phi(\omega)$ is a weighting function, $H_d(\omega)$ the desired frequency response to be modelled/approximated, τ_d the desired group delay function, and Ω_p and Ω_s are the sets of passband and stopband frequency points. Note that in time domain (12) is equivalent to minimizing the mean square (MMS) error between the outputs of the desired and the designed systems driven by the same random input sequence with $\Phi(\omega)$ as its power density spetral function. To simplify problem (**P**), we note that the objective function can be written as

$$\int_{-\pi}^{\pi} |H_L - H_d|^2 d\omega = \frac{1}{2} x^T H_f x - b_f^T x + c_f \tag{13}$$

where

$$\begin{aligned} H_f &= \frac{1}{\pi} \int_{-\pi}^{\pi} \Phi(\omega) |H(e^{j\omega})|^2 d\omega \\ b_f &= \frac{1}{\pi} \int_{-\pi}^{\pi} \Re\{\Phi(\omega) H_L(e^{j\omega}) H_d(\omega)\} d\omega \\ c_f &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi(\omega) |H_d(\omega)|^2 d\omega \end{aligned}$$

Let us now consider the linearisation of the magnitude constraint. For any complex number $z = \xi + j\eta$, from the rotation theorem [7] we know that

$$|z| = \max_{0 \leq \theta < 2\pi} \Re\{z e^{j\theta}\}, \tag{14}$$

Hence, the magnitude inequality constraint can be written as

$$\Re\{(H_L(e^{j\omega}) - H_d(\omega)) e^{j\theta}\} \leq \delta_m, \quad \forall \theta \in [0, 2\pi) \tag{15}$$

It is easy to verify that (15) can be written as

$$\mathbf{a}^T(\omega, \theta) x \leq c(\omega, \theta) \tag{16}$$

where $\mathbf{a}(\omega, \theta) = \Re\{\phi(\omega) e^{j\theta}\}$, $\phi = [\phi_0, \phi_1, ..., \phi_{N-1}]^T$ and $c(\omega, \theta) = \Re\{H_d(\omega) e^{j\theta}\} + \delta_m$. Constraint (16) is linear and continuous.

To summarise, using (8), (13) and (16) the optimization problem (**P**) can be written as

$$\min_x \{\frac{1}{2} x^T H_f x - b_f^T x\} \tag{17}$$

subject to

$$\mathbf{a}^T(\omega, \theta) x \leq c(\omega, \theta), \quad \forall (\omega, \theta) \in \Omega_p \times [0, 2\pi) \tag{18}$$

$$\left| \frac{x^T Q_\Gamma(\omega) x}{x^T P(\omega) x} - \tau_d \right| \leq \delta_g, \quad \forall \omega \in \Omega_p \tag{19}$$

Remark: Due to the group delay constraint, problem (**P**) is a general nonlinear optimization problem. This means that only local optimal solutions (if it exists) can be expected. However, our numerical simulation experience indicate that good choice of starting point can lead to very satisfactory solution. For instance, since the objective function is convex and the magnitude constraints are linear, one way to obtain a good solution is to choose a starting point as the solution to the optimization problem (17)-(18) which is a quadratic programming problem. Alternatively, the linearization approach in [3] can also be utilized to provide such a starting point for solving the general nonlinear optimization problem.

3.2. Suboptimal Approach

To simplify the group delay constraints, we consider the case when $|H(e^{j\omega})|^2 \approx |H_d(\omega)|^2$ in the passband (note that this can be guaranteed by the magnitude constraint with sufficiently small δ_m). It can be proved [10] that

$$|H_L(e^{j\omega})|^2 = \gamma^2(\omega)(x^T P(\omega) x)$$

This means that in the passband we have,

$$x^T P(\omega) x \approx \frac{1}{\gamma^2(\omega)} |H_d(\omega)|^2$$

Therefore, the group delay function can be written as

$$\tau_g(\omega) \approx x^T Q_a(\omega) x$$

where $Q_a(\omega) = \frac{\gamma(\omega) Q_\Gamma(\omega) \gamma(\omega)}{|H_d(\omega)|^2}$. This approximation converts the group delay constraint into a quadratic constraint which, according to our experience, can be handled much more easily using existing optimization subroutines such as *constr.m* in Matlab's Optimization Toolbox [4]. This suggests that, as an alternative, we shall consider the following simplified version of the nonlinear optimization problem (**P**).

Problem (P_b): Solve the following optimization problem for a discrete set of $\{\omega_k\}$.

$$\min_x \{\frac{1}{2} x^T H_f x - b_f^T x\} \tag{20}$$

subject to

$$|x^T \gamma^2(\omega) P(\omega) x - |H_d(\omega)|^2| \leq \epsilon_p$$

$$\left| x^T Q_a(\omega) x - \tau_d \right| \leq \delta_g$$

This is a quadratic problem with quadratic constraints. Specific methods exist [9] for solving such problems.

4. Numerical Examples

For illustration, consider the design of a 36th order Laguerre filter to approximate the desired frequency response $H_d(\omega) = H_r(\omega)e^{-j\tau_d\omega}$, where $H_r(\omega)$ is the frequency response of the noncausal and zero phase raised cosine filter defined by $H_r(\omega) = T$ for $\omega < (1-\alpha)\pi/T$, $H_r(\omega) = \frac{T}{2}\left(1-\sin(\frac{T}{2\alpha}(\omega-\frac{\pi}{T}))\right)$ for $(1-\alpha)\pi/T \leq \omega \leq (1+\alpha)\pi/T$, and $H_r(\omega) = 0$, elsewhere. τ_d is the desired group delay constant. Choosing $\alpha = 0.35$, N=36 and a=0.25, $\tau_d = 14$, $\delta_m = \delta_g = 0.2$, we solve the filter design optimization problem (P_b). The magnitude, group delay and impulse response are depicted, respectively, in Figs. 1–3. The figures indicate very satisfactory design results.

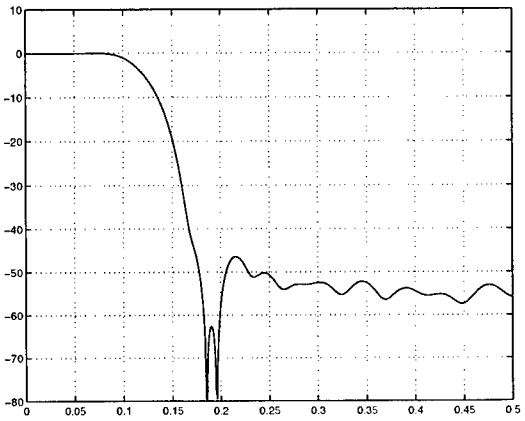


Figure 1. Plot of magnitude response of the designed Laguerre filter.

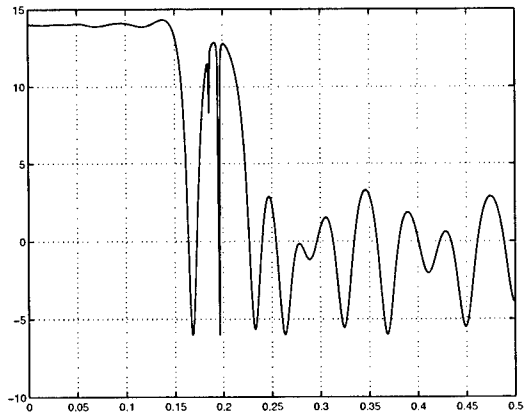


Figure 2. Plot of group delay function of the designed Laguerre filter.

5. Concluding Remarks

We have shown that the group delay function of a given Nth order digital Laguerre filter can be written as a ratio of quadratic functions with respect to the filter coefficients. We have posed a filter design problem as a constrained L_2 space minimization problem with both magnitude and group delay constraints. The approach is practically relevant and simulation results demonstrated the usefulness of our theoretical results and proposed solution methods.

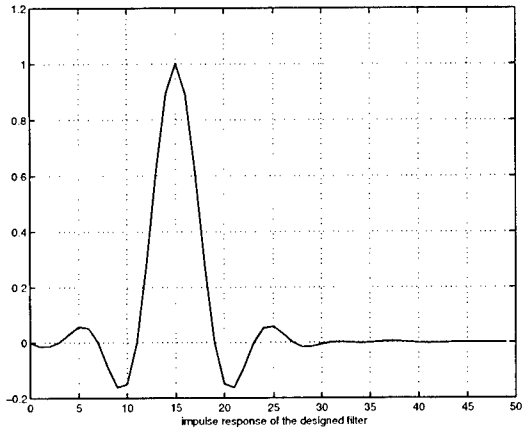


Figure 3. Plot of impulse response of the designed Laguerre filter.

REFERENCES

[1] J. W. Adams and J.L. Sullivan, "Peak-constrained least-squares optimization," *IEEE Transactions on Signal Processing*, vol. 46, no. 2, pp. 306-321, 1998.

[2] D. Burnside, T. W. Parks, "Optimal Design of FIR Filters with the Complex Chebyshev Error Criteria", *IEEE Transactions on Signal Processing*, vol. 43, no. 3, pp. 605–616, March 1995.

[3] X. Chen, T. W. Parks, "Design of FIR Filters in the Complex Domain", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP–35, no. 2, pp. 144–153, February 1987.

[4] A. Grace, *Optimization Toolbox: for use with MATLAB*, The Math Works Inc., South Natick, MA, 1993.

[5] L.J. Karam and J.H. McClellan, "Complex Chebyshev approximation for FIR filter design," *IEEE Trans. on Circuits and Systems-II*, vol. 42, no. 3, pp.207-216, 1995

[6] S. Nordebo and Z. Zang, "Semi-Infinite linear programming: a unified approach to digital filter design with time and frequency domain specifications." *IEEE Transactions on Circuit and Systems, Part II*, vol. 46, no. 6, pp. 765-775, 1999.

[7] T.W. Parks and C.S. Burrus, *Digital Filter Design*. New York: Wiley, 1987.

[8] I.W. Selesnick, M. Lang, and C.S. Burrus, "Constrained least square design of FIR filters without specified transition bands," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp.1897-1892, 1996.

[9] H. Tuy, *Convex Analysis and Global Optimization*. Dordrecht: Kluwer Academic Publishers, 1998.

[10] Z. Zang, S. Nordholm, and S. Nordebo, "Design of digital filter design with amplitude, phase and group delay specifications." Technical Report, Australian Telecomm. Research Institute, Curtin University of Technology, Australia, August, 2000.

PERFORMANCE ANALYSIS OF SUBSPACE PROJECTION TECHNIQUES FOR ANTI-JAMMING GPS USING SPATIO-TEMPORAL INTERFERENCE SIGNATURES

Moeness G. Amin[†], *Liang Zhao* [†], *and Alan R. Lindsey*[††]

[†] Department of Electrical and Computer Engineering
Villanova University, Villanova PA 19085, USA

[††] Air Force Research Laboratory / IFGC
525 Brooks Road, Rome, NY 13441, USA

ABSTRACT

Combined spatial and time-frequency signatures of signal arrivals at multi-antenna receivers has recently been used for effective nonstationary interference suppression in broadband communication platforms. This paper presents performance analysis of subspace projection array processing techniques for suppression of frequency modulated (FM) jammers in GPS receivers. The FM jammers are instantaneous narrowband and have clear time-frequency (t-f) signatures that are distinct from the GPS C/A spread spectrum code. The paper assumes that the spatial signature of the jammer is accurately estimated, but its instantaneous frequency (IF) estimate, which provides the basis for construction of the jammer subspace exhibits zero-mean independent Gaussian errors. Simulation results comparing the effects of IF errors in a single and multi-antenna GPS receivers are provided.

1. INTRODUCTION

Recently, subspace projection techniques based on time-frequency distributions and bilinear transforms have been devised for non-stationary FM interference excision in direct-sequence spread-spectrum (DSSS) communications using a single and multi-antenna receivers [1][2][3][4]. These techniques exploit the jammer time-frequency signatures and rely on the distinct differences in the time-frequency localization properties between the jammer and the spread spectrum signals. The jammer instantaneous frequency, whether provided by the time-frequency distributions or any other IF estimator, is used to define the temporal signature of the interference, which is in turn used to construct the interference subspace. The respective projection matrix is used to excise the jammer power in the incoming signal prior to correlation with the receiver pseudorandom noise (PN) sequence. The result is improved receiver signal-to-interference-plus-noise ratio (SINR). The use of multi-antenna array further improves the performance by increasing the dimension of the available signal subspace. It allows both the distinctions in the spatial and temporal signatures of the GPS signals from those of the interferers to play equal roles in suppressing the jammer with a minimum distortion of the desired signal[1][5].

As the jammer subspace is solely determined by the jammer IF, reliable estimation of the IF is important for FM interference mitigation. With perturbations in IF, the GPS receiver anti-jamming performance is degraded, lowering the receiver SINR. The single antenna receiver case was discussed in [2], where it was shown that small IF estimation errors may lead to a significant decrease in the receiver SINR. This paper analyzes the multi-antenna GPS receiver performance in the presence of zero-mean identical and independent Gaussian IF estimation errors. Accurate estimates of the jammer spatial signatures are assumed, and as such, exact values of the cross-correlation coefficients between the signal and the jammers are used. In comparing the single and multi-antenna cases, the paper shows that the employment of antenna arrays can improve the receiver SINR effectively by exploiting the difference in spatial signatures.

2. SUBSPACE PROJECTION ARRAY PROCESSING

Once the IF of the jammer is estimated from the t-f domain, or by using any other appropriate IF estimator, the jammer vector can be constructed up to an ambiguity in a constant complex amplitude. Subspace projection techniques perform jammer suppression by projecting the received data onto the jammer's orthogonal subspace. Herein, the focus is on jamming of GPS receivers.

In GPS, the PN sequence of length P (1023) repeats itself Q (20) times within one symbol of the 50 bps navigation data [6][7]. Discrete-time form is used, where all the signals are sampled at the chip-rate of the C/A code. Now, consider an antenna array of N sensors, and a communication channel restricted to flat-fading. In the proposed interference excision approach, the PNQ sensor output samples are partitioned into Q blocks, each of P chips and

This work is supported by the Air Force Research Laboratory, Information Directorate, Rome, NY, grant no. F30602-00-1-0515.

0-7803-7011-2/01/$10.00 ©2001 IEEE

PN samples. The jammer can be consecutively removed from the 20 blocks. This is achieved by projecting the received data in each block on the corresponding orthogonal subspace of the jammer. The jammer-free signal is then correlated with the replica PN sequence on a symbol-by-symbol basis. Subspace projection within each block is first considered. The array output vector at the kth sample is given by

$$\begin{aligned}\mathbf{x}(k) &= \mathbf{x}_s(k) + \mathbf{x}_u(k) + \mathbf{w}(k) \\ &= p(k)\mathbf{h} + A\sqrt{P}u(k)\mathbf{a} + \mathbf{w}(k)\end{aligned} \quad (1)$$

where $\mathbf{x}_s$, $\mathbf{x}_u$, and $\mathbf{w}$ are the signal spreading code, the jammer and the white Gaussian noise contributions, respectively. $\mathbf{h}$ is the signal spatial signature, and $p(k)$ is the spreading PN sequence. The jammer is considered as instantaneously narrowband FM signal with constant amplitude $u(k) = \frac{1}{\sqrt{P}}\exp[j\phi(k)]$ (jammer is normalized within each block). A and $\mathbf{a}$ are the jammer amplitude and spatial signature, respectively. Furthermore, the spatial channels are normalized and it is presumed that $\|\mathbf{h}\|_F^2 = N$ and $\|\mathbf{a}\|_F^2 = N$, where $\|\cdot\|_F^2$ is the Frobenius norm of a vector. The noise vector w(k) is zero-mean, temporally and spatially white with

$$E[\mathbf{w}(k)\mathbf{w}^T(k+l)] = 0, \quad E[\mathbf{w}(k)\mathbf{w}^H(k+l)] = \sigma^2\delta(l)\mathbf{I}_N \quad (2)$$

where σ^2 is the noise power, and $\mathbf{I}_N$ is the $N\times N$ identity matrix. Using P sequential array vector samples within the block, the following PN×1 data vector for one block is obtained:

$$\mathbf{X} = [\mathbf{x}^T(1)\ \ \mathbf{x}^T(2)\ \ \ldots\ \ \mathbf{x}^T(P)]^T = \mathbf{X}_s + \mathbf{X}_u + \mathbf{W}. \quad (3)$$

The spatial-temporal signature $\mathbf{X}_s$ can be rewritten as

$$\mathbf{X}_s = \mathbf{p} \otimes \mathbf{h} \quad (4)$$

where $\mathbf{p}=[p(1),\ (2),\ldots\ p(P)]$ is the signal in one block, and $\otimes$ denotes the Kronecker product. In the same way, the jammer vector $\mathbf{X}_u$ is expressed as

$$\mathbf{X}_u = A\sqrt{P}\mathbf{u} \otimes \mathbf{a}_i \triangleq A\sqrt{P}\mathbf{U} \quad (5)$$

where $\mathbf{u}=[u(1),\ u(2),\ldots\ u(\mathrm{P})]$ is the jammer normalized vector, and $\mathbf{U}$ is the spatial-temporal signature. Its orthogonal subspace projection matrix is given by

$$\mathbf{V} = \mathbf{I}_{LN} - \mathbf{U}(\mathbf{U}^H\mathbf{U})^{-1}\mathbf{U}^H = \mathbf{I}_{LN} - \frac{1}{N}\mathbf{U}\mathbf{U}^H \quad (6)$$

With the exact knowledge of the jammer IF, the projection of the received signal vector onto the orthogonal subspace yields

$$\mathbf{X}_\perp = \mathbf{V}\mathbf{X} = \mathbf{V}\mathbf{X}_s + \mathbf{V}\mathbf{W} \quad (7)$$

which excises the jammers. The result of despreading over one block is

$$y = \mathbf{X}_s^H\mathbf{X}_\perp = \mathbf{X}_s^H\mathbf{V}\mathbf{X}_s + \mathbf{X}_s^H\mathbf{V}\mathbf{W} \triangleq y_1 + y_2 \quad (8)$$

where y_1 and y_2 are the contributions of the PN and the noise sequences to the correlator output, respectively. For simplification, the jammers are assumed to share the same period as the GPS C/A code. y_1 is deterministic due to the fact that the spreading code for each satellite signal is fixed and periodic. The decision variable is the real part of the sum of the correlation output over Q blocks.

3. EFFECTS OF IF ERRORS ON THE PROJECTION OPERATION

Errors in IF may occur in many situations, where it becomes difficult to determine the IF due to a drop in the jammer power, presence of amplitude modulations, or high levels of cross-terms in the t-f domain. When IF estimation errors exist, the subspace projection operation will not entirely remove the jammer. The un-excised residual jammer at the projection filter output is often significant, specifically for high JSR. In this paper, the phase error model is a zero-mean Gaussian white noise process, motivated by the fact that phase errors, directly obtained from the analytic signal of FM in complex Gaussian additive noise, have wrapped normal distributions [8]. For high jammer power, the distribution variance becomes very small and the phase errors assume a Gaussian distribution.

The estimated unit jammer vector can be represented as

$$\hat{\mathbf{u}}^T = \frac{1}{\sqrt{P}}\left[e^{j(\phi(1)+\Delta(1))}\quad e^{j(\phi(2)+\Delta(2))}\quad \ldots\quad e^{j(\phi(P)+\Delta(P))}\right] \quad (9)$$

The phase estimation errors $\Delta(i)$ at different chips are assumed to be i.i.d random variables with a zero mean Gaussian distribution and variance σ_ε^2. The variance σ_ε^2 is assumed to be sufficiently small such that most errors lie inside the interval $[-\pi,\pi]$. The projection matrix, constructed from the inaccurate jammer vector, is

$$\hat{\mathbf{V}} = \mathbf{I}_{LN} - \frac{1}{N}\hat{\mathbf{U}}\hat{\mathbf{U}}^H \quad (10)$$

where $\hat{\mathbf{U}} = \hat{\mathbf{u}} \otimes \mathbf{a}$. In this case, the output of the correlator in one block is

$$y = \mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{X}_s + \mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{W} + \mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{X}_u \triangleq y_1 + y_2 + y_3 \quad (11)$$

where y_1, y_2 and y_3 represent, respectively, the contributions of the spreading code, the noise sequence, and the interfering signal. Due to phase estimation errors, these three variables are random which renders equation (11) different from its deterministic counterpart equation (8). Since the projection matrix $\hat{\mathbf{V}}_k$ is Hermitian, y_1 is always real. The mean value of y_1 is

$$E[y_1] = E[\mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{X}_s] = PN - \frac{1}{N}E[\mathbf{X}_s^H\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{X}_s] \quad (12)$$

Define

$$\alpha = \frac{\mathbf{h}^H\mathbf{a}}{N} \quad (13)$$

as the spatial cross-correlation coefficient between the signal and the jammer, and

$$\beta = \frac{\mathbf{p}^T\mathbf{u}}{\sqrt{P}},\quad \hat{\beta} = \frac{\mathbf{p}^T\hat{\mathbf{u}}}{\sqrt{P}} \tag{14}$$

as the exact and estimated temporal cross-correlation coefficient between the PN sequence and the jammer vector, respectively. Therefore,

$$\mathbf{X}_s^H\hat{\mathbf{U}} = (\mathbf{p}\otimes\mathbf{h})^H(\mathbf{u}\otimes\mathbf{a}) = (\mathbf{p}^T\mathbf{u})(\mathbf{h}^H\mathbf{a}) = \sqrt{P}N\alpha\hat{\beta} \tag{15}$$

It can be readily shown that

$$E[\hat{\mathbf{u}}\hat{\mathbf{u}}^H] = e^{-\sigma_\varepsilon^2}\mathbf{u}\,\mathbf{u}^H + \frac{1}{P}(1-e^{-\sigma_\varepsilon^2})\mathbf{I}_P \tag{16}$$

Hence,

$$\begin{aligned} E[y_1] &= PN(1-|\alpha|^2\,E[|\hat{\beta}|^2]) = PN(1-|\alpha|^2\,E[\frac{\mathbf{p}^T\hat{\mathbf{u}}\hat{\mathbf{u}}^H\mathbf{p}}{P}]) \\ &= PN\left[1-|\alpha|^2\left(e^{-\sigma_\varepsilon^2}|\beta|^2+\frac{1-e^{-\sigma_\varepsilon^2}}{P}\right)\right] \end{aligned} \tag{17}$$

With the noise assumption in (2),

$$E[y_2] = 0 \tag{18}$$

Similar to y_1, the mean value of y_3 is obtained as

$$\begin{aligned} E[y_3] &= E[\mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{X}_j] = A\sqrt{P}\left(\mathbf{X}_s^H\mathbf{U}-\frac{1}{N}E[\mathbf{X}_s^H\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{U}]\right) \\ &= A\sqrt{P}\left(\mathbf{X}_s^H\mathbf{U}-\frac{1}{N}N^2\alpha E[\mathbf{p}^T\hat{\mathbf{u}}\hat{\mathbf{u}}^H\mathbf{u}]\right) \\ &= AN\alpha\beta(1-e^{-\sigma_\varepsilon^2})(P-1) \end{aligned} \tag{19}$$

From (17)-(19), the mean value of y can be calculated by the sum

$$E[y] = E[y_1]+E[y_2]+E[y_3] \tag{20}$$

The mean square value due to the signal is

$$\begin{aligned} E[|y_1|^2] &= E[\mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{X}_s\,\mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{X}_s] \\ &= E[P^2N^2 - 2P\mathbf{X}_s^H\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{X}_s + \frac{1}{N^2}\mathbf{X}_s^H\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{X}_s\,\mathbf{X}_s^H\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{X}_s] \\ &= P^2N^2 - 2N^2P|\alpha|^2\,E[\mathbf{p}^T\hat{\mathbf{u}}\hat{\mathbf{u}}^H\mathbf{p}] + N^2|\alpha|^4\,E[\mathbf{p}^T\hat{\mathbf{u}}\hat{\mathbf{u}}^H\mathbf{p}\mathbf{p}^T\hat{\mathbf{u}}\hat{\mathbf{u}}^H\mathbf{p}] \\ &= P^2N^2 - 2N^2P|\alpha|^2(e^{-\sigma_\varepsilon^2}P|\beta|^2+1-e^{-\sigma_\varepsilon^2}) + \\ &\quad N^2|\alpha|^4\left[(2+4P|\beta|^2\,e^{-\sigma_\varepsilon^2})(1-e^{-\sigma_\varepsilon^2}) + P^2|\beta|^4\,e^{-2\sigma_\varepsilon^2}\right] \end{aligned} \tag{21}$$

In deriving the above expression, $\mathbf{p}^T\hat{\mathbf{u}}$ is approximated by a complex Gaussian random variable using the Central Limit Theorem. From (17) and (21), the variance of y_1 can be computed as

$$\sigma_{y_1}^2 = E[y_1^2] - E^2[y_1] \tag{22}$$

The mean square values of y_2 and y_3 are given by

$$\begin{aligned} E[|y_2|^2] = \sigma_{y_2}^2 &= E[\mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{W}\mathbf{W}^H\hat{\mathbf{V}}\mathbf{X}_s] \\ &= E[E[\mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{W}\mathbf{W}^H\hat{\mathbf{V}}\mathbf{X}_s\,|\,\hat{\mathbf{V}}]] \\ &= \sigma^2E[\mathbf{X}_s^H\hat{\mathbf{V}}\hat{\mathbf{V}}\mathbf{X}_s] \\ &= \sigma^2E[\mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{X}_s] = \sigma^2E[y_1] \\ &= \sigma^2PN\left[1-|\alpha|^2\left(e^{-\sigma_\varepsilon^2}|\beta|^2+\frac{1-e^{-\sigma_\varepsilon^2}}{P}\right)\right] \end{aligned} \tag{23}$$

$$\begin{aligned} E[|y_3|^2] &= E[\mathbf{X}_u^H\hat{\mathbf{V}}\,\mathbf{X}_s\mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{X}_u] = A^2PE[\mathbf{U}^H\hat{\mathbf{V}}\mathbf{X}_s\,\mathbf{X}_s^H\hat{\mathbf{V}}\mathbf{U}] \\ &= A^2PE[\mathbf{U}^H\mathbf{X}_s\,\mathbf{X}_s^H\mathbf{U} - \frac{1}{N}\mathbf{U}^H\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{X}_s\,\mathbf{X}_s^H\mathbf{U} \\ &\quad -\frac{1}{N}\mathbf{U}^H\mathbf{X}_s\,\mathbf{X}_s^H\,\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{U} + \frac{1}{N^2}\mathbf{U}^H\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{X}_s\,\mathbf{X}_s^H\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{U}] \\ &= A^2P\{N^2P|\alpha|^2\,|\beta|^2 - N^2|\alpha|^2\sqrt{P}\beta E[\mathbf{u}^H\hat{\mathbf{u}}\hat{\mathbf{u}}^H\mathbf{p}] \\ &\quad -N^2|\alpha|^2\sqrt{P}\beta^*E[\mathbf{p}^H\hat{\mathbf{u}}\hat{\mathbf{u}}^H\mathbf{u}] + N^2|\alpha|^2\,E[\mathbf{u}^H\hat{\mathbf{u}}\hat{\mathbf{u}}^H\mathbf{p}\mathbf{p}^H\hat{\mathbf{u}}\hat{\mathbf{u}}^H\mathbf{u}]\} \\ &= A^2N^2P|\alpha|^2\{P|\beta|^2 - 2P|\beta|^2[e^{-\sigma_\varepsilon^2}+(1-e^{-\sigma_\varepsilon^2})/P] \\ &\quad +[e^{-\sigma_\varepsilon^2}+(1-e^{-\sigma_\varepsilon^2})/P](P|\beta|^2e^{-\sigma_\varepsilon^2}+1-e^{-\sigma_\varepsilon^2})\} \end{aligned} \tag{24}$$

Therefore, the variance of y_3 is given by

$$\sigma_{y_3}^2 = E[|y_3|^2] - |E[y_3]|^2 \tag{25}$$

It can be shown that the covariance between y_i and y_j ($i\neq j$), $i,j=1,2,3$) assume small values relative to the respective variance values. The variance of the decision variable can be approximated by

$$\sigma_{y_r}^2 = \sigma_{y_1}^2 + \frac{1}{2}(\sigma_{y_2}^2+\sigma_{y_3}^2) \tag{26}$$

The above equations are derived for only one block of the signal symbol. Below, the subscript m is added to identify y with block m (m=1, 2, ..., Q), and should not be confused with those used in (11). The decision variable y_r can be expressed as

$$y_r = \mathrm{Re}[\sum_{m=1}^{Q} y_m] \tag{27}$$

Since the white Gaussian and estimation errors are independent for different blocks, the expected value and variance of y_r become

$$E[y_r] = \sum_{m=1}^{Q}\mathrm{Re}[E[y_m]],\quad Var[y_r] = \sum_{m=1}^{Q}\mathrm{var}[y_{rm}] \tag{28}$$

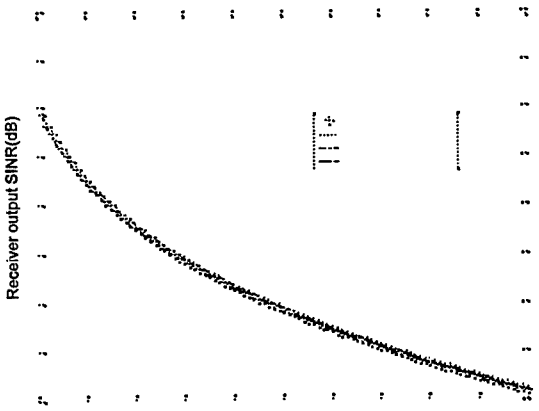
The above expressions can now be used to generate the desired receiver SINR,

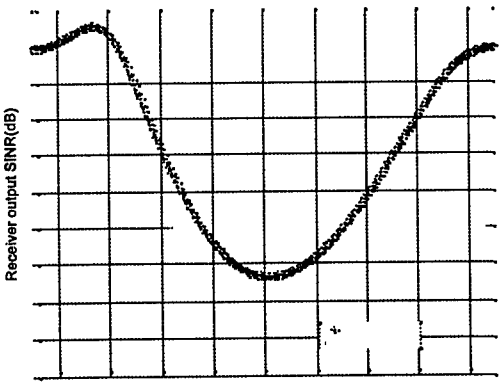
$$SINR = \frac{E^2[y_r]}{Var[y_r]} \tag{29}$$

4. SIMULATIONS

Computer simulations of the receiver performance are presented based on the following parameters: the PN code is the Gold code of satellite SV #1, and the jammer is a periodic linear FM signal with normalized frequency range [0, π]. The jammer has a period equal to the C/A code block length. The angle of arrival (AOA) of the satellite signal is 5^0. A two-element array is considered with half-wavelength spacing. The Jammer-to-Signal-Ratio (JSR) is set to 50 dB and SNR equal to –20 dB.

Figure 1 depicts the simulated values of the receiver SINR vs. the phase error variance σ_ε^2, which changes in the range [0, 0.01] for all blocks. The AOA of the jammer signals are set to be 5^0, 35^0, and 65^0, respectively. It is clear from the figure that, as the error variance increases, the output SINR decreases. The SINR of the single sensor case is also plotted for comparison. Unlike the result of exact IF estimation, where antenna arrays bring a constant 3 dB array gain [1], the receiver SINR in the presence of those errors is dependent on the spatial signatures of the signal and jammer. For small spatial cross-correlation coefficients, the use of antenna array allows the receiver to be more robust to the IF estimation errors. The relation between the receiver SINR and the jammer AOA is shown in Fig. 2. In this case, phase error variance σ_ε^2 in Fig. 2 kept constant at 0.01.





5. CONCLUSION

Subspace projection is a pre-correlation technique that can reject the instantaneous narrowband interference effectively. Using antenna arrays can provide further improvement in the receiver SINR performance by exploiting both temporal and spatial signatures. However, as the subspace projection matrix is solely dependent on the IF estimation, IF estimation errors will perturb the projection matrix and allow part of the jammer power escape the projection operation.

In this paper, the SINR performance of GPS receiver using array subspace projection in the presence of IF estimation errors has been analyzed. The phase errors are modeled as zero-mean white Gaussian, and independent over different chips. The spatial signature is assumed to be accurately estimated. The analysis and simulation shows that, although the IF estimation errors can affect the receiver SINR significantly, the combination of temporal and spatial signature can provide more robustness in the presence of IF estimation errors, and as such, render better performance than the single antenna scheme.

REFERENCES

[1] L. Zhao, M. G. Amin, and A. R. Lindsey, "Subspace Array Processing for the Suppression of FM Jamming in GPS Receivers," Proceedings of the 34th Asilomar Conference on Signals, Systems, and Computers, Monterey, CA, Oct. 2000.

[2] L. Zhao, M. G. Amin, and A. R. Lindsey, "Performance Analysis of Subspace Projection Techniques for FM Interference Rejection in GPS Receivers," Proceedings of SPIE, Digital Wireless Communications, Orlando, FL, Apr. 2001.

[3] L. Zhao, M. G. Amin, and A. R. Lindsey, "Mitigation of Periodic Interferers in GPS Receivers Using Subspace Projection Techniques," IEEE International Symposium on Signal Processing and Applications, Malaysia, Aug., 2001.

[4] S. Ramineni, M. G. Amin, and A. R. Lindsey, "Performance Analysis of Subspace Projection Techniques in DSSS Communications," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, May 2000.

[5] Y. Zhang and M. G. Amin, "Array processing for nonstationary interference suppression in DS/SS communications using subspace projection techniques," submitted to IEEE Transactions on Signal Processing, Sept. 2000.

[6] B. W. Parkinson, J. J. Spilker Jr. eds, Global Positioning System: Theory and Applications, American Institute of Aeronautics and Astronautics, 1996.

[7] E. D. Kaplan eds, Understanding GPS: Principles and Applications, Artech House Publishers, 1996.

[8] B. C. Lovell and R. C. Williamson, "The statistical performance of some instantaneous frequency estimators," IEEE Transactions on Signal Processing, vol. 40, no. 7, pp. 1708-1723, July 1992.

GAIN DECOMPOSITION METHODS FOR RADIO TELESCOPE ARRAYS

A.J. Boonstra[1,2] *and A.J. van der Veen*[2]

[1]ASTRON, Netherlands Foundation for Research in Astronomy
Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands
Tel: +31(0)521 595100, fax: +31(0)521 597332, email: boonstra@astron.nl

[2] Department of Electrical Engineering, Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
Tel: +31(0)15 2781372, fax: +31(0)15 2786190, email: allejan@cas.et.tudelft.nl

ABSTRACT

In radio telescope arrays, the complex receiver gains and sensor noise powers are initially unknown and have to be calibrated. Gain calibration enhances the quality of astronomical sky images and moreover, improve the effectiveness of certain radio telescope phased-array data processing techniques, such as radio interference (RFI) mitigation and beamforming. In this paper we present several closed form and iterative complex gain estimation methods. These methods are analyzed and compared to the Cramer-Rao lower bound for the variance of the estimated gain. The models are tested both on simulated data and on observed telescope data.

Keywords: applications in radio astronomy, sensor array processing.

1. INTRODUCTION

Gain calibration techniques for radio telescope systems exist already for a long time [1][2]. However, since studies started for a next generation of radio telescopes (the Square Kilometer Array radio telescope or SKA [3]), phased array beamforming issues and radio frequency interference (RFI) suppression techniques received renewed interest [4] in radio astronomy. For RFI suppression, and for phased array beamforming, gain calibration of the telescope array is an important factor. Maximum likelihood techniques exist for estimation of the gain and phase of signals impinging on the telescope array [5] and for estimation of the direction of arrival of the impinging signals [6]. For computational reasons (SKA will have many sensor elements) and for robustness reasons (iterative maximum likelihood techniques depend on a good initial point) we investigated several closed form and iterative complex gain estimation methods and found that these techniques perform well.

The complex gains and noise powers of individual telescopes of a telescope array (figure 1) can be estimated by observing a strong astronomical source in the centre of the field of view of the telescopes. In most cases, single point sources can be found which dominate the field of view of a radio telescope. A telescope output signal is the sum of the telescope system noise (uncorrelated among the telescopes) and the astronomical source flux, which is correlated, multiplied by the telescope gain. The source flux is the same for each of the telescopes, but the telescope gains and noise powers usually are not. The gains consist of the combined effect of atmospheric disturbances, telescope geometry, receiver characteristics, and electronic (amplifier) gains, whereas the system noise powers can differ by several dB's.

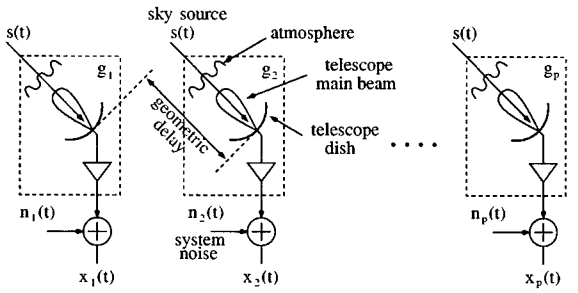


Fig. 1. Radio telescope array

The output of the backend processing is a sequence of covariance matrices formed by cross correlation of all the telescope outputs x_i. The aim in this paper is to estimate the complex gain factors and the system noise powers from an observed covariance matrix, assuming that the astronomical source flux is known from tables. We present three algorithms to extract these parameters.

2. DATA MODEL

Assume that during the calibration observations the telescopes are pointed at a single radio source in the sky. For a telescope array (figure 1) the output x_i of element i at a certain time t can be modeled (using the narrow band assumption) as

$$x_i(t) = g_i a_i s(t) + n_i(t) \qquad (1)$$

where g_i is the complex gain of the sensor, n_i is the system noise of channel i, a_i is the narrow band phase offset due to the geometric delay, and $s(t)$ is the flux of the impinging external source. For the gain calibration observation, the sky source is located in the centre of the field of view. The geometry and look direction of a telescope is known, so the narrow band phase offset due to the

A.J. Boonstra was supported by the NOEMI project of the STW under contract no. DEL77-4476.

geometric delay is known as well and can be compensated for. Hence without loss of generality we may assume in our model that the phase offsets are $a_i = 1$.

In radio astronomy, the sensor array output

$$\mathbf{x}(t) = [x_1(t), \cdots, x_p(t)]^t \tag{2}$$

is usually correlated with itself to form a covariance matrix. Here the superscript t means the transpose operator, and p is the number of telescopes. The true covariance matrix $\mathbf{R}$ and estimate $\widehat{\mathbf{R}}$ based on N samples, assuming stationarity over this interval, are given by

$$\mathbf{R} = E\{\mathbf{x}(t)\mathbf{x}(t)^H\} \tag{3}$$

$$\widehat{\mathbf{R}} = \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{x}(t+nT)\mathbf{x}(t+nT)^H \tag{4}$$

where superscript H denotes the complex conjugate transpose.

The signal power $\sigma_s^2 = E\{|s(t)|^2\}$ is known from tables, hence without loss of generality we may model it as $\sigma_s^2 = 1$. The covariance matrix can now be written as

$$\mathbf{R} = \mathbf{g}\mathbf{g}^H + \mathbf{D} \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix containing the system noise contributions, $d_i = E\{|n_i(t)|^2\} \geq 0$. The gain vector $\mathbf{g}$ can be written as a product of a gain magnitude $\boldsymbol{\gamma} = [\gamma_i, \cdots, \gamma_p]^t$ ($\gamma_i > 0$) and a phasor $\mathbf{z} = [e^{\jmath\phi_1}, \cdots, e^{\jmath\phi_p}]^t$; i.e. $\mathbf{g} = \boldsymbol{\gamma} \odot \mathbf{z}$, where $\odot$ is the Schur-Hadamard (elementwise) matrix product. The ij-th element of $\mathbf{R}$ is thus given by

$$r_{ij} = \gamma_i\gamma_j e^{\jmath(\phi_i - \phi_j)} + d_i\delta_{ij} \tag{6}$$

Since the phases are underdetermined, we define without loss of generality the phase of the first sensor to be zero: $\phi_1 = 0$. The objective at this point is, given $\widehat{\mathbf{R}}$, estimate $\mathbf{g}$ and $\mathbf{D}$ according to the model (5).

3. GAIN DECOMPOSITION ALGORITHMS

3.1. Alternating least squares gain estimation (ALS)

The covariance matrix in equation (5) is composed of a rank-one matrix $\mathbf{g}\mathbf{g}^H$ and a diagonal matrix $\mathbf{D}$. The gain extraction procedure is based on minimizing the model error:

$$\{\widehat{\mathbf{g}}, \widehat{\mathbf{D}}\} = \arg\min_{\mathbf{g},\mathbf{D}\geq 0} \|\widehat{\mathbf{R}} - \mathbf{D} - \mathbf{g}\mathbf{g}^H\|_F^2 \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. In the ALS technique, we alternatingly minimize over one component, keeping the other component fixed. In particular, assume at the k-th iteration that we have an estimate $\widehat{\mathbf{D}}^{(k)}$. The next step is to minimize equation (7) with respect to the gain vector only:

$$\widehat{\mathbf{g}}^{(k)} = \arg\min_{\mathbf{g}} \|\widehat{\mathbf{R}} - \mathbf{g}\mathbf{g}^H - \widehat{\mathbf{D}}^{(k)}\|_F^2 \tag{8}$$

The minimum is found from the eigenvalue decomposition of $\widehat{\mathbf{R}} - \widehat{\mathbf{D}}^{(k)}$,

$$\widehat{\mathbf{R}} - \widehat{\mathbf{D}}^{(k)} = \mathbf{U}^{(k)}\boldsymbol{\Lambda}^{(k)}\mathbf{U}^{(k)H} \tag{9}$$

where the matrix $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_p]$ contains the eigenvectors $\mathbf{u}_i$, and $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues λ_i. The gain estimate minimizing (8) is given by

$$\widehat{\mathbf{g}}^{(k)} = \mathbf{u}_1^{(k)}\sqrt{\lambda_1^{(k)}} \tag{10}$$

where λ_1 is the largest eigenvalue, and $\mathbf{u}_1$ is the corresponding eigenvector. The second step is to minimize (7) with respect to the system noise matrix $\mathbf{D}$, keeping the gain vector fixed. The minimum is obtained by subtracting $\widehat{\mathbf{g}}^{(k)}\widehat{\mathbf{g}}^{(k)H}$ from $\widehat{\mathbf{R}}$ and discarding all off-diagonal elements. The condition that the diagonal elements of $\widehat{\mathbf{D}}^{(k+1)}$ should be positive is implemented by subsequently setting $\widehat{d}_i^{(k+1)} = \max(\widehat{d}_i^{(k+1)}, 0)$. The two minimizations steps are repeated until the model error (7) converges. Since each of the minimizing steps in the iteration loop reduces the model error, we obtain monotonic convergence to a local minimum. Although the iteration is very simple to implement, simulations indicate that convergence usually is very slow, especially in the absence of a reasonable initial point.

3.2. Column ratio gain estimation (COL)

We now set out to find a closed form estimate of $\mathbf{g}$, which recovers $\mathbf{g}$ exactly when applied to $\mathbf{R}$ (hence asymptotic for $\widehat{\mathbf{R}}$). The crux of this method is the observation that the off-diagonal entries of $\mathbf{g}\mathbf{g}^H$ are equal to those of $\mathbf{R}$, hence known, so that we only need to reconstruct the diagonal entries of $\mathbf{g}\mathbf{g}^H$. This can be done in closed form by estimating the column ratios of $\mathbf{R}$ away from the diagonal as discussed below. The diagonal of the covariance matrix $\mathbf{R}$ is then replaced with the estimate producing a matrix of the form $\mathbf{R}' = \mathbf{g}\mathbf{g}^H$. The gain vector $\mathbf{g}$ can then be extracted by an eigenvalue decomposition of $\mathbf{R}'$.

The ratio α_{ij} of two of elements g_i and g_j of the complex gain vector $\mathbf{g}$ can be estimated from the data $\mathbf{R}$ by solving

$$\mathbf{c}_i = \alpha_{ij}\mathbf{c}_j \tag{11}$$

where $\mathbf{c}_i$ and $\mathbf{c}_j$ are the i-th and j-th column of the matrix $\mathbf{R}$, not including the entries r_{ii}, r_{ij}, r_{ji} and r_{jj} because r_{ii} and r_{jj} contain also the unknown system noise contributions d_i. Solving for α_{ij} in the Least Squares sense gives

$$\alpha_{ij} = (\mathbf{c}_i^H\mathbf{c}_i)^{-1}\mathbf{c}_i^H\mathbf{c}_j = \frac{\sum_{k\neq i,j} r_{ki}^* r_{kj}}{\sum_{k\neq i,j} r_{ki}^* r_{ki}} \tag{12}$$

We can subsequently estimate $|g_i|^2$ as $|g_i|^2 = \Re(\alpha_{ij}r_{ij}^*)$, for any choice of j. This estimate can be improved if all $(p-1)$ available column ratios are used. The next step is to form $\mathbf{R}'$ equal to $\mathbf{R}$ but with the diagonal entries replaced by the estimates of $|g_i|^2$ obtained above. The resulting matrix $\mathbf{R}'$ is an estimate of $\mathbf{g}\mathbf{g}^H$, and $\mathbf{g}$ is found from an eigenvalue decomposition of $\mathbf{R}' = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H$, similarly as in (9), (10) before. In the actual algorithm, we follow the same procedure but replace $\mathbf{R}$ by the sample estimate $\widehat{\mathbf{R}}$.

3.3. Logarithmic least square gain estimation (LOGLS)

An alternative closed form estimate (as in use at the Westerbork Sythesis Radio Telescope, WSRT [1] since 1980) is obtained by minimizing the mean square error of the *logarithms* of (6). Taking the logarithm has the effect that the equations become linear as the product of gains become sums. We start by taking the logarithm

of the off-diagonal elements ($\forall\, i \neq j$) of equation (6) and define the logarithmic model errors Δ_{ij} ($\forall\, i \neq j$) as

$$\Delta_{ij} \equiv \ln(\widehat{r}_{ij}) - \ln(\gamma_i) - \ln(\gamma_j) - \jmath(\phi_i - \phi_j) \bmod 2\pi\jmath \quad (13)$$

Minimization in the least square sense of the sum-squared error $\sum |\Delta_{ij}|^2$ over the real gains and and phases is obtained by setting

$$\frac{\partial}{\partial \ln(\gamma_k)} \sum_{\substack{i,j=1 \\ i\neq j}}^{p} |\Delta_{ij}|^2 = 0\,, \quad \frac{\partial}{\partial \ln(e^{\jmath\phi_k})} \sum_{\substack{i,j=1 \\ i\neq j}}^{p} |\Delta_{ij}|^2 = 0 \quad (14)$$

After some manipulations the equation for the gain magnitude (14) becomes:

$$\sum_{\substack{i=1 \\ i\neq k}}^{p} \Big(\Re\{\ln(\widehat{r}_{ik})\} - \ln(\gamma_i) - \ln(\gamma_k) \Big) = 0 \quad (15)$$

for $k = 1, \cdots, p$. This equation can easily be written in matrix form and solved in closed form using Woodbury's identity. The same procedure leads to a closed form solution for the phase. In this method, phase unwrapping is necessary. This is done by using a simple phase quadrant estimation procedure.

4. GAIN ESTIMATION SIMULATIONS

4.1. Method

The aim of the simulations is to evaluate the gain estimation accuracy as a function of signal to noise ratio ($\text{SNR}_i = \mathbf{g}^H \mathbf{g}/d_i$), i.e. the ratio of the astronomical source power (normalized here to unity) and array gain to the noise power in the i-th channel.

In the simulations we use eight telescope channels. The gain magnitude was kept fixed during the simulations, and was taken as a nominal value plus a (uniformly selected) random deviation of 10% of the nominal value . The gain phase was randomly distributed in the interval $[0, 2\pi]$ and also kept fixed during the simulations. In the presentation of the results, we split the gain estimates in a magnitude and a phase, since they might have different accuracies, and since the Cramer-Rao bounds are based on these (real) parameters.

4.2. Cramer-Rao lower bound of the gain estimates

The Cramer-Rao Bound (CRB) gives a lower bound on the variance of any unbiased estimator [7]. In our situation, we assume that the source signal and the channel noise are independent Gaussian distributed with zero mean, and satisfy the model in equation (5). Define the parameter vector

$\boldsymbol{\theta} \equiv [\gamma_1, \cdots, \gamma_p, \phi_2, \cdots, \phi_p, d_1, \cdots, d_p]^t$ (Note that the phase ϕ_1 of the first sensor is not a parameter.) The CRB is then given by [7]

$$\text{var}\,(\widehat{\theta}_i(\mathbf{X})|\boldsymbol{\theta}) \geq \left[\mathbf{I}_F^{-1}\right]_{ii} \quad (16)$$

where $\mathbf{I}_F$ is the Fisher information matrix, where $\mathbf{X}$ is defined as $\mathbf{X} \equiv (\mathbf{x}[1] \cdots \mathbf{x}[N])$, and N is the number of samples. Following standard techniques [7], the Fisher information matrix can be written as

$$\mathbf{I}_{F,ij}(\boldsymbol{\theta}) = N\text{tr}\Big(\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_i}\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_j}\Big) \quad (17)$$

Inserting the model (5), the components in the Fisher information matrix can easily be found as

$$\mathbf{R}^{-1} = \mathbf{D}^{-1}\left(\mathbf{I} - \frac{\mathbf{g}\mathbf{g}^H\mathbf{D}^{-1}}{1+\mathbf{g}^H\mathbf{D}^{-1}\mathbf{g}}\right) \quad (18)$$

$$\partial\mathbf{R}/\partial\gamma_i = (\mathbf{e}_i \odot \mathbf{z})\,\mathbf{g}^H + \mathbf{g}\,(\mathbf{e}_i \odot \mathbf{z}^H) \quad (19)$$

$$\partial\mathbf{R}/\partial\phi_i = \jmath(\,(\mathbf{g}\odot\mathbf{e}_i)\mathbf{g}^H - \mathbf{g}(\mathbf{g}\odot\mathbf{e}_i)^H) \quad (20)$$

$$\partial\mathbf{R}/\partial d_i = \mathbf{e}_i\mathbf{e}_i^t \quad (21)$$

where $\mathbf{e}_i$ denotes the i-th unit vector. The estimation variance of the model parameters is calculated by evaluating equation (16).

4.3. Comparison of the gain decomposition methods: simulation results

For a typical online gain calibration measurement at a radio observatory, astronomical sources are used with noise powers in the range of 0.1 to 10% of the telescope system noise power. The integration time of the correlation data can be several seconds to a few minutes.

Figure 2 shows the results of a gain estimation simulation in which the gain estimation variance is plotted versus SNR for a fixed number of samples. The three models are plotted together with the Cramer-Rao lower bound. In the -10 to 0 dB SNR range, the gain estimation errors lie very close to the CRB (for 16 k samples) and the estimators are unbiased. Below an SNR of -15 dB the gain estimation starts to deviate from the bound. The ALS method breaks down at higher SNR than the other two methods.

The phase estimation tends to break down earlier than the gain magnitude estimation. The phase breakdown point is observed around -16 dB, and is a bit lower for the LOGLS method. At low SNR some of the curves drop below the Cramer-Rao bound. Here, the estimators are biased.

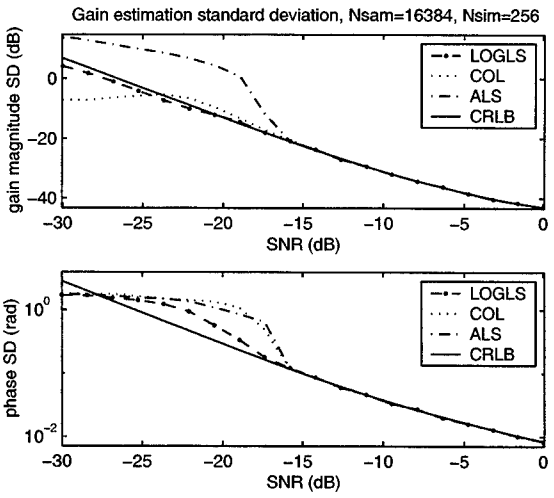


Fig. 2. Gain estimation standard deviation versus SNR

In figure 3 the gain estimates are plotted as a function of the number of observed time samples for a fixed SNR. Note that also here, the phase estimators break down earlier than the gain magnitude estimators.

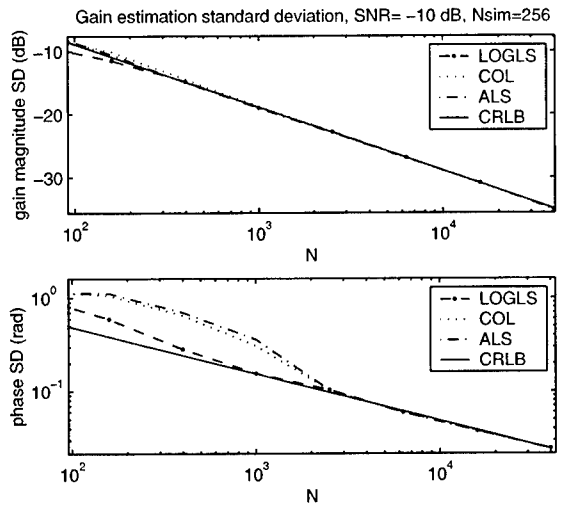


Fig. 3. Gain estimation standard deviation vs. number of samples

5. EXPERIMENTAL RESULTS

The gain estimation methods were tested on real telescope data. An eight channel datarecorder was connected to the Westerbork Synthesis Radio Telescope, which was pointing at the astronomical source 3C48. Baseband signals were recorded corresponding to a sky frequency of 1420 MHz. The SNR of the source relative to the system noise is -13 dB. A narrow band is selected (by means of an FFT) and covariance matrices are derived by cross correlation of the input sequences. The observed correlation coefficient is 0.055 with a spread of about 10% due to the different gains of the telescopes. The gain decomposition algorithms are applied to the covariance matrices. Figure 4 shows the observed gain magnitude estimation standard deviation and the CRLB. The entire dataset is

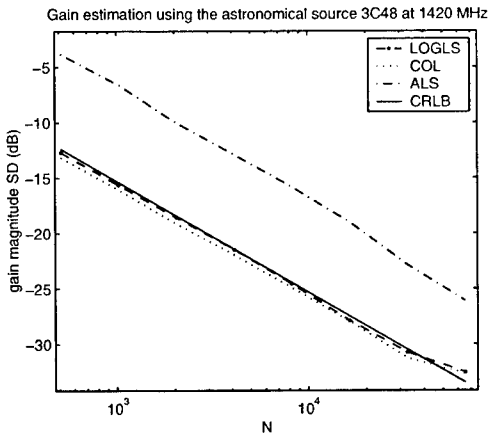


Fig. 4. Gain estimation standard deviation from observations with the astronomical source 3C48.

used to obtain a fair estimate of the complex gains, these numbers are used for the calculation of the CRLB. The curves for the LOGLS and COLS gain estimations lie very close to the CRLB, just as is the case with the simulations.The ALS however, performs not too well for SNR's in the range below about -15 dB; the ALS estimates are biased. The small deviation of the LOGLS and COLS curves from the CRLB curve could be caused by the fact that for the CRLB calculations not the true gains were used (as they are not known) but the estimated gains.

6. CONCLUSIONS

In our simulations the three gain estimation methods do not differ much in performance. The main difference is that the ALS method for gain magnitude estimation breaks down a bit earlier than the two other methods. Also, the phase estimation seems to break down earlier than the gain magnitude estimation. For 16 k samples and for SNRs higher than -15 dB, the estimators are unbiased (for the gain distribution used).

In general the measurement results support the conclusions from the simulations. However, the ALS gain magnitude estimates deviate 8 dB from the CRLB. The ALS estimator is biased for the SNR of the observation.

Further research will focus on other methods, like the Gauss-Newton iterative method [8], on processing efficiency, and the methods will be extended to multiple sources.

7. REFERENCES

[1] H. van Someren-Greve. Logarithmic least square gain decomposition algorithm for the WSRT. IWOS software documentation, ASTRON internal document, 1980.

[2] R.A. Perley F.R. Schwab and A.H.Bridle. Synthesis imaging in radio astronomy. *Astronomical Society of the Pacific Conference Series*, 6, 1994.

[3] A.B.Smolders and M.P. van Haarlem, editors. *Perspectives on Radio Astronomy: Technologies for Large Antenna Arrays*. ASTRON, conference proceedings edition, April 1999.

[4] A. Leshem A.J. van der Veen and A.J. Boonstra. Multichannel interference mitigation techniques in radio astronomy. *Astrophysical Journal Supplements*, 131(1):355–374, November 2000.

[5] Q. Cheng Y.Hua and Petre Stoica. Asymptotic performance of optimal gain and phase estimators of sensor arrays. *IEEE Transactions on Signal Processing*, 48(12):3587–3590, December 2000.

[6] M. Pesavento and A.B. Gershman. Array processing in the presence of unknown nonuniform sensor noise: a maximum likelihood direction finding algorithm and cramer-rao bounds. *IEEE Wokshop on Statistical Signal Processing*, 2000.

[7] S.M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, volume 1. Prentice Hall, Inc., New Jersey, 1993.

[8] D.N. Lawley and A.E. Maxwell. *Factor Analysis as a Statistical Method*. Butterworth & Co, London, 1971.

IDENTIFICATION OF GEAR MESH SIGNALS BY KURTOSIS MAXIMISATION AND ITS APPLICATION TO CH46 HELICOPTER GEARBOX DATA

Wenyi Wang
Aeronautical and Maritime Research Laboratory
Defence Science and Technology Organisation
PO Box 4331, Melbourne, VIC 3001, Australia
(e-mail: wenyi.wang@dsto.defence.gov.au*)*

ABSTRACT

The detection and diagnosis of gearbox faults is of vital importance for the safe operation of helicopters. This paper presents a new approach in identifying gear mesh signals for early and effective detection of localised gear faults. Using this approach, the gear mesh signal is identified using a non-minimum phase autoregressive (AR) model by maximising the kurtosis of the inverse filter error signal of the model. Sudden changes in the error signal are usually indications of the existence of localised gear faults in the monitored gear. It is demonstrated using well-regarded CH46 helicopter aft transmission test data that the approach shows great promise for detecting faults in complex gearboxes.

INTRODUCTION

To reduce life cycle cost and to improve the reliability and safety of helicopters, fault diagnosis schemes for their power transmission systems are essential. In general, vibration diagnosis has been proven effective in the detection and diagnosis of gear faults. However, vibration signals generated by helicopter transmissions are heavily corrupted by noise. Therefore, it is essential to employ advanced signal enhancement techniques to extract useful diagnostic information from the measured vibration signals. The synchronous signal averaging technique is a powerful tool in extracting the tooth mesh signal of the gear of interest from the measured vibration signal. Using the averaged signal, various analysis techniques can be used to further identify the gear mesh signal in determining the gear's health condition.

For healthy gears, the averaged signal is normally dominated by the gear meshing harmonics (sinusoids with the gear meshing frequency – the tooth number times the shaft rotational frequency, and its harmonics) modulated by the rotation of gear shaft [1]. In most cases, the modulation waveforms are also sinusoids with lower shaft orders, i.e., 1× and/or 2× the shaft frequency. In cases where the shaft has multiple gears, the signal will be more complex due to the cross gear modulation interaction [2]. When a localised tooth defect such as a tooth crack is present, the engagement of the cracked tooth will induce an impulsive change with comparatively low energy to the gear mesh signal. This can produce some higher shaft-order modulations and may excite structural resonances. For the purpose of fault diagnosis, this additive impulse must be extracted. A common approach, i.e., the conventional residual signal method, is to use a multiple band-stop filter to remove the gear meshing harmonics and sometimes their lower shaft-order modulation sidebands from the gear mesh signal, resulting in a residual signal that may expose the low energy impulses. Other approaches focus on the higher shaft-order modulation sidebands or a structural resonance to extract the information relevant to local gear faults. However, when dealing with incipient fault detection and cases involving multiple gears on the same shaft, or strong unknown (e.g., ghost [1]) components, the current techniques may become ineffective.

The minimum-phase autoregressive (MPAR) modelling approach was studied for gear fault diagnosis at the Aeronautical and Maritime Research Laboratory (AMRL) [3]-[5]. Under the assumption that gear mesh signals were derived from an AR system driven by Gaussian noise, an AR model was established for signals from the monitored gear under healthy conditions, and then used as a linear prediction error (LPE) filter. The future signals from the same gear, under healthy or faulty conditions, were processed by the LPE filter. The output prediction error at the LPE filter would resemble random noise if the monitored gear remains in a healthy condition. However, when a local fault (e.g., a tooth crack) is developed in the gear, the fault-affected region would not be well predicted by the AR model that was established under healthy conditions. Thus the LPE signal would reflect the changes caused by the fault. It was shown that the AR modelling approach outperforms current gear fault diagnostic techniques, such as the residual signal method. Using this method, there is no need to know the number of teeth on the monitored gear, the characteristics of the modulation waveforms or the number of gears on the same shaft. Moreover, it is *not* important whether structural resonances and additional modulations are presented in the signal, whereas this information is essential to some of the current techniques.

However, in terms of making an early detection of incipient local gear faults, the MPAR modelling method can still perform unsatisfactorily. Hence, a more general linear prediction modelling method – non minimum-phase AR (NMPAR) – has been studied. Here, we assume that gear mesh signals are derived from a non minimum-phase AR system driven by non-Gaussian noise. The AR coefficients are estimated by kurtosis maximisation using the inverse filter criteria. This technique has been applied to the analysis of the CH46 helicopter aft transmission seeded fault testing data [6], with encouraging results.

GEAR MESH SIGNAL

A signal model generated by the mesh of a faulty gear is presented here. The model consists of 1) the gear meshing vibration with some amplitude and phase modulation effects caused by geometric and assembly errors, together with speed

0-7803-7011-2/01/$10.00 ©2001 IEEE

and load fluctuations of the gears, and 2) additional modulation and structure resonant vibration caused by local gear fault. The model is expressed as:

$$y(t)=\sum_{m=0}^{M}A_m[1+a_m(t)]\cos\{2\pi f_m t+\beta_m+b_m(t)\} + d(t)\cos(2\pi f_r t+\theta_r)+v(t) \quad (1)$$

where A_m, f_m and β_m are the amplitude, frequency and initial phase of the m-th mesh harmonic, respectively. Functions $a_m(t)$ and $b_m(t)$ are the amplitude and phase modulation functions at the m-th mesh harmonic respectively, in which both the normal modulation effects created by shaft rotation and those produced by the fault-induced impact are considered. Function $d(t)$ is the envelope function of the resonant vibration: f_r is the resonance frequency (the carrier frequency) and θ_r the corresponding initial phase. The last term, $v(t)$, is considered the remaining noise after signal averaging. In most cases, Eq. (1) represents a sub-Gaussian signal with normalised kurtosis (see Eq. 4) below 2.80 (3.00 for a Gaussian signal).

Using the conventional residual signal method, the removal of gear meshing harmonics and lower shaft-order modulation will produce a residual signal in which the higher shaft-order modulation and resonance terms remain. If these terms are significant, the residual signal may become a super-Gaussian signal. Therefore, the kurtosis as a measure of Gaussianity of the residual signal can be used as the index for fault detection.

NON MINIMUM-PHASE AR MODELLING

Based on Eq. (1), it is assumed that the gear mesh signal is generated by a NMPAR system driven by non-Gaussian noise. Hence, the discrete signal of Eq. (1) can be expressed by

$$y[n]=-\sum_{k=1}^{p}a[k]\cdot y[n-k]+w[n]\cdot \quad (2)$$

The input random signal $w[n]$, n=1,2, ..., N, is assumed non-Gaussian, zero-mean and higher-order white (i.e., independently identically distributed – i.i.d.), which contains the fault-induced part in Eq. (1). For stability of non-causal systems, a constraint for Eq. (2) is

$$A(z)=\sum_{k=1}^{p}a[k]z^{-k}\neq 0, \quad \text{for } |z|=1, \quad (3)$$

which means that the Z-transform polynomial $A[z]$ parameterised by $a[k]$ has no roots on the unit circle. In identifying the system in Eq. (2), we denote $\varepsilon[n;\alpha]$ the estimate of $w[n]$ obtained by applying inverse signal model with parameter $\alpha[k]$ (estimate of $a[k]$) to the signal $y[n]$. In analogy to channel equalisation problems, the identification process is shown in Fig. 1.

Theorems developed by Tugnait [7] and Shalvi *et al* [8] show that, under energy constraint, the kurtosis of the inverse filter error signal $\varepsilon[n]$ is upper bounded by the kurtosis of true random input signal $w[n]$ to the model shown in Eq. (2), that is $K(\varepsilon[n])$ ≤ $K(w[n])$ with equality if and only if the parameter estimate $\alpha[k]$ equals to the true parameter of the system $a[k]$, k=1, 2, ..., p. Based on the theorems, the AR parameters $a[k]$ can be estimated by maximising the normalised kurtosis of $\varepsilon[n;\alpha]$, with some constraints on the variance of $\varepsilon[n;\alpha]$. In gear diagnosis, the kurtosis of a zero-mean signal $\varepsilon[n]$ is defined by:

$$K_{\varepsilon[n]}=N\cdot\sum_{n=1}^{N}\varepsilon^4[n]\Big/\left(\sum_{n=1}^{N}\varepsilon^2[n]\right)^2. \quad (4)$$

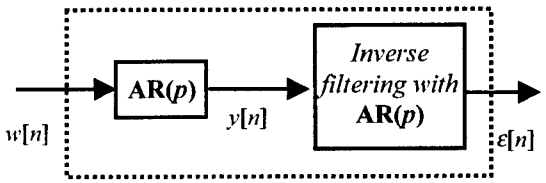


Fig. 1. Blind identification of an AR system

As shown in Eq. (1), the gear mesh signal is essentially amplitude/phase modulated sinusoidals. Maximisation of the above kurtosis alone could lead to an $\varepsilon[n]$ with repetitive spikes at a period identical to that of the dominant sinusoid. However, the system to be identified is assumed to have a non-Gaussian random noise input, where autocorrelation coefficients should be minimal. Therefore, we need to put some restrictions on the autocorrelation coefficients of the inverse filter error $\varepsilon[n]$, i.e., the estimate of $w[n]$. Furthermore, constraints on the power of $\varepsilon[n]$ also need to be considered to ensure convergence. These constraints can be incorporated into the cost function (refer to [9] for detailed necessary and sufficient conditions for a blind signal recovery). In our study, the cost function is as follows:

$$\max_{\alpha}\{J(\varepsilon[n;\alpha])\}=K_\varepsilon-\lambda\log_{10}(C_\varepsilon)-\mu(R_\varepsilon), \quad (5)$$

where C_ε and R_ε are the variance and autocorrelation coefficients of the error signal $\varepsilon[n]$, respectively. The logarithm for C_ε is to transform C to a similar range to K. The constants λ and μ can be chosen such that the decrease of K and the increases of C and R are to be equally penalised during optimisation.

Obviously, Eq. (5) represents a non-linear cost function. A number of unconstrained optimisation algorithms, such as simplex search, gradient-type methods and genetic algorithms, can be used for this maximisation process. In our study, we employed a quasi-Newton gradient-type algorithm with the BFGS Hessian updating, developed by Broyden, Fletcher, Goldfarb and Shanno, due to its wide availability; for example, in Microsoft Excel™ and Matlab™ etc. It is believed that the BFGS method is the most effective gradient-type method for general purpose non-linear optimisations. The initial parameters for the AR model can be obtained using a MPAR parameter estimation method, such as the Yule-Walker method. The MPAR parameters are then replaced by the mixed-phase parameters during iteration. The model order can be obtained by the Akaike Information Criteria (AIC), which is thought to be sufficient for gear fault detection.

APPLICATION TO THE CH46 HELICOPTER TRANSMISSION SEEDED FAULT TEST DATA

The proposed NMPAR technique has been applied to the well-known CH46 helicopter aft transmission seeded fault testing

data [6] (available at http://wisdom.arl.psu.edu/Westland/data). The U.S. Navy sponsored the test that was carried out in 1993 at Westland Helicopters Ltd. in the U.K. using its Universal Transmission Test Rig. The data used for this paper were acquired in Test #6 and Test #7, where gear crack propagation tests were conducted on the double helical idler and the collector gear, shown in Fig. 2 as Components #8 and #6 (both shaded), respectively.

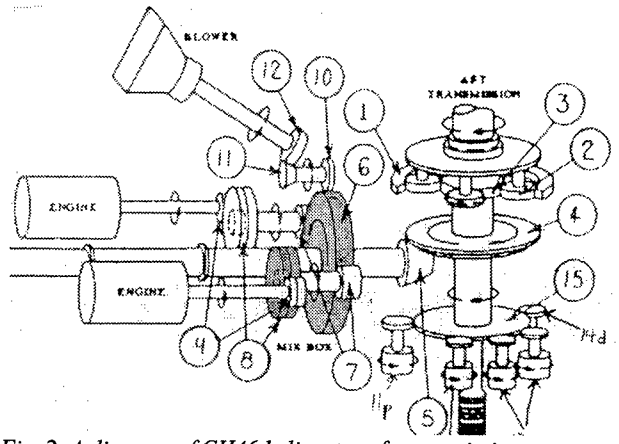


Fig. 2. A diagram of CH46 helicopter aft transmission.

IN TEST #6, the gearbox was tested with the port (i.e., the shaded) double helical or herringbone idler gear containing a spark-eroded notch in the roots of two teeth on each helical gear at an identical shaft angle. In the CH46 mix gearbox, there were two herringbone idler gears (Component #8 with 72 teeth) connecting to two input pinions (Component #9) on the turbine engine shafts, and two spur idler pinions (Component #7 with 25 teeth) on the same shafts connecting to the collector gear. Sensor #2 (at the port-engine-input location) was chosen for the purpose of this paper because of its close vicinity to the faulty gear. The data were acquired when the cracks had grown to 3~19mm and 2~22mm including the initial notch, and the gearbox was running at 100% power (Record #446 in [6]).

The averaged signal was normalised to zero-mean and unit variance and is shown in Fig. 3(a). The spectral analysis of this signal showed that the gear meshing harmonics of the 25-tooth spur pinion were dominant. By removing the mesh harmonics from both the 25-tooth pinion and the 72-tooth herringbone gear, together with their lower shaft-order sidebands, the conventional residual signal (with kurtosis of 3.07) was obtained and is shown in Fig. 3(b). Obviously, no convincing diagnostic information can be found from the figure because the cross gear interactive components [2] (i.e., $m\times72 \pm n\times25$; m, n = 1, 2, ...) still dominated the residual signal.

Fig. 3(c, d) respectively show the error signals of a minimum phase AR(76) model obtained using the Yule-Walker method and AIC, and of a non-minimum phase AR(76) model by maximising the cost function shown in Eq. (5) with λ=3 and μ=6. In Fig. 3(c), the error signal was reduced significantly compared to the conventional residual signal shown in Fig. 3(b), but, still no evidence of fault could be found. As can be seen in Fig. 3(d), a distinct spike at about 315° of shaft angle was detected by the NMP-AR error signal with the kurtosis of 9.32, which strongly indicated the existence of a localised gear fault in either the 25-tooth spur pinion or the 72-tooth herringbone gear.

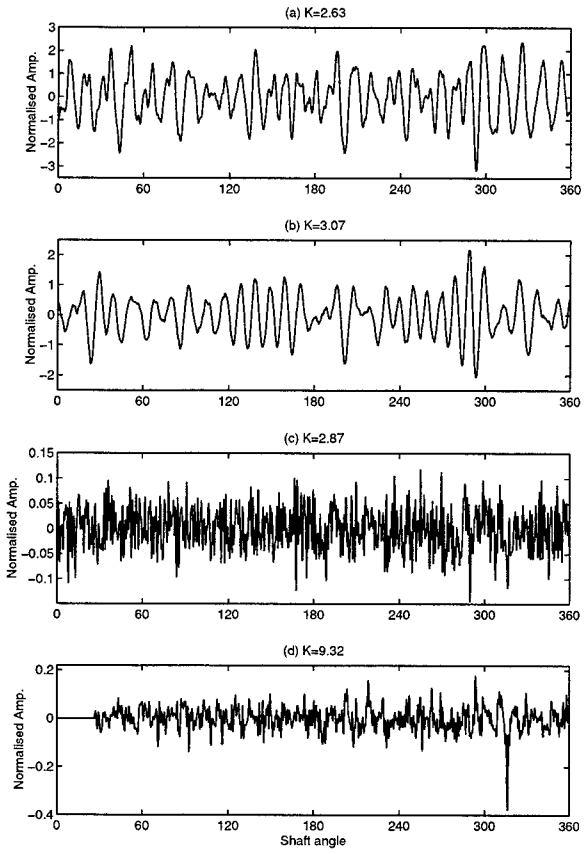


Fig. 3. Analysis result for Test #6 (helical idler gear crack propagation) with Sensor #2. (a) The normalised signal of the idler gear, (b) the residual signal with mesh harmonics and lower shaft-order sidebands (from both 25T and 72T gears) removed, (c) the AR(76) model error signal and (d) the error signal of a non-minimum phase AR(76) model.

IN TEST #7, a fatigue crack was initiated using a spark-eroded notch in the root of one tooth of the collector gear. After a total of 63 hours of testing, the crack propagated across the root of the tooth and brought the test to conclusion when the tooth detached from the gear [6]. The 74-tooth collector gear combined the drive from port and starboard engine inputs (180° apart) to the two main rotor transmission gearboxes. At the ends of the collector gear shaft, there were two spiral bevel pinions (Component #5 with 26 teeth) driving the front and aft main rotor gearboxes. Hence, the multiple gears on the same shaft would produce cross-gear interactions [2] as their tooth meshing components modulate with each other.

Fig. 4(a) shows the normalised signal acquired at 75% power by Sensor #3 (Record #1543 in [6] at 50.3 test-hours), where the initial notch (length×depth = 36×5mm, rectangular shape) had grown by 7~10mm on the aft end of the notch and 3~6mm on the forward end. Using the conventional residual signal method, it was found extremely difficult, if not impossible, to remove all the interactive components (i.e., $m\times74 \pm n\times26$; m, n = 1, 2, ...). Fig. 4(b) shows the residual signal with the removal of the meshing harmonics from both 74-tooth and 26-tooth gears and their lower-order sidebands. As can be seen, the residual signal was still dominated by some periodic

components, which is most likely due to the cross-gear interaction. Again, the residual signal is of little value for fault detection.

By applying a minimum-phase AR(45) model to the data, the inverse filter error signal was obtained and is shown in Fig. 4(c), where no evidence of sudden changes can be found. The kurtosis value of the error signal was found to be 3.20. Fig. 4(d) presents the inverse filter error signal produced by the proposed approach using a non minimum-phase AR(45) model. From this signal, we can easily identify two distinct spikes at about 150° and 330° of shaft angles respectively. The corresponding kurtosis was found to be 7.21, which strongly indicated the existence of the tooth cracking. By zooming into the spikes, we found they were exactly 180° apart. This is because the monitored collector gear engaged with two pinions, i.e., Component #7 as shown in Fig. 2, so that a local impact would be produced by the mesh of the cracked tooth with each of the two pinions.

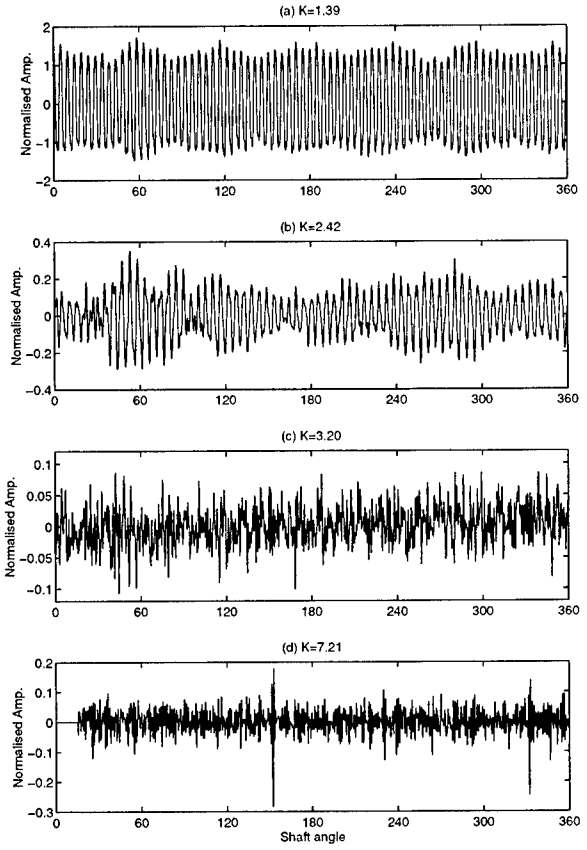


Fig. 4. Analysis results for Test #7 (collector gear crack propagation) with Sensor #3. (a) The normalised signal of the collector gear, (b) the residual signal with mesh harmonics and lower shaft-order sidebands (from both 74T and 26T gears) removed, (c) the MP AR(45) model error signal and (d) the NMP-AR(45) model error signal.

CONCLUSION

The analysis results presented in this paper have demonstrated the effectiveness of the proposed non minimum-phase AR modelling method by kurtosis maximisation for gear fault detection. The method is potentially superior to the current gear fault detection methods from the following perspectives: 1) earlier and more convincing detection of gear tooth cracking; 2) capability of detecting faults in complex gearboxes, such as helicopter transmission gearboxes, where cross-gear interactions are common.

The proposed technique requires the choice of some coefficients, such as λ and μ, in the cost function and the selection of non-linear optimisation algorithms. In our study, we mainly concentrated on the gradient-type optimisation algorithms and we chose λ and μ by putting all three terms in the cost function into a similar scale. For future work, we need to develop more systematic approaches for choosing λ and μ and to test other popular optimisation methods, such as neural network based non-linear optimisation and genetic algorithms, for suitability in gear fault detection.

Acknowledgment: The author would like to thank Dr. A.K. Wong and Mr. S.A. Fisher of AMRL for their comments on the paper, and Mr. Bill Hardman and Mr. Mark Hollins of the U.S. Navy Air Warfare Center for their support to this work.

REFERENCES

[1] B. Randall, "A New Method of Modeling Gear Faults," ASME Journal of Mechanical Design, Vol. 104, April 1982, pp.259-267.

[2] M. Zacksenhouse, S. Braun, M. Feldman and M. Sidahmed, "Toward Helicopter Gearbox Diagnostics from a Small Number of Examples," Mechanical System and Signal Processing, Vol. 14, No. 4, July 2000. p523-543.

[3] W. Wang and A.K. Wong, "A Model-based Gear Diagnostic Technique," DSTO Technical Report: DSTO-TR-1079, Dec. 2000, Australia.

[4] W. Wang and A.K. Wong, "Linear Prediction and Gear Fault Diagnosis," the Proceedings of the 13th International Congress on Condition Monitoring and Diagnostic Engineering Management, Dec. 3-8, 2000, Houston, Texas, USA. p797-807.

[5] W. Wang and A.K. Wong, " Autoregressive Model Based Gear Fault Diagnosis," submitted to ASME journal of Vibration and Acoustics.

[6] "Final Report on CH-46 Aft Transmission Seeded Fault Testing," Westland Research Paper RP907 (available at: http://wisdom.arl.psu.edu/Westland/report).

[7] J.K. Tugnait, "Estimation of Linear Parametric Models Using Inverse Filter Criteria and Higher Order Statistics," IEEE Trans. on Signal Processing, Vol. 41, No. 11, Nov. 1993. p3196-99.

[8] O. Shalvi and E. Weintein, "New Criteria for Blind Deconvolution of Nonminimum Phase Systems (Channels)," IEEE Trans. on Information Theory, Vol.36, No.2, Mach 1990. p312-321.

[9] C.B. Papadias, "Globally Convergent Blind Source Separation Based on a Multiuser Kurtosis Maximisation Criterion," IEEE Trans. on Signal Processing, Vol. 48, No. 12, Dec. 2000. p3508-3519.

A Hyperbolic LMS Algorithm for CORDIC Based Realization

Mrityunjoy Chakraborty, Suraiya Pervin and T. S. Lamba
Dept. of Electronics and Electrical Communication Engineering,
Indian Institute of Technology, Kharagpur-721302, W.B., India.
email: {mrityun,pervin}@ece.iitkgp.ernet.in

ABSTRACT

An alternate formulation of the LMS algorithm is presented by expressing the mean square error as a convex function of a set of hyperbolic variables that are monotonically related to the filter tap weights. The proposed algorithm is ideally suitable for CORDIC based realization and possesses very good convergence characteristics as revealed via extensive simulation studies.

Key words: LMS algorithm, CORDIC arithmetic, Time varying system.

1. INTRODUCTION

The CORDIC algorithm originally proposed by Volder [1], shot to prominence in last two decades, as it offered scopes for efficient implementation of a large class of signal processing algorithms. The popularity of the CORDIC method can be traced to its numerical stability, efficiency in evaluating trigonometric and hyperbolic functions and inherent pipelinability at the microlevel [2]. For the case of the LMS-based adaptive filters, however, the CORDIC based approach has so far remained confined largely to lattice filters ([3]-[4]) and seemingly has not been extended to the transversal form. This is because, in the case of the former, the computations in each stage can be related easily to a set of hyperbolic operations, while no such direct hyperbolic, or, trigonometric interpretation exists for the computations present in the latter. In this paper, we propose an alternate formulation of the LMS algorithm using a set of hyperbolic variables which are monotonically related to the transversal filter coefficients. The proposed hyperbolic LMS (HLMS) algorithm leads to a class of pipelined architectures and possesses satisfactory convergence characteristics as demonstrated via simulation studies.

2. A HYPERBOLIC FORMULATION OF THE LMS ALGORITHM

We begin by first considering the steepest descent search procedure that arises in the optimal FIR filtering problem. Given an input sequence $x(n)$, desired response $d(n)$ and an N-tap filter coefficient vector $\mathbf{w}=[w_0,w_1,\ldots,w_{N-1}]^t$, the optimal filter $\hat{\mathbf{w}}=[\hat{w}_0,\hat{w}_1,\ldots,w_{N-1}]^t$ is obtained by minimizing the mean square error (MSE) function $\varepsilon^2=E[e^2(n)]$, where $e(n)$ is the error signal at the filter output and is given by $e(n)=d(n)-\mathbf{w}^t\mathbf{x}(n)$, with $\mathbf{x}(n)=[x(n),x(n-1),\ldots,x(n-N+1)]^t$. The MSE ε is a convex function of the filter coefficients w_k, $k=0,\ 1,\ldots,N-1$ and defines the so called error performance surface in an N+1 dimensional space. In the proposed alternative, we select a set of N hyperbolic angles θ_k, $k=0,1,\ldots,N-1$ and express each tap weight as $w_k=\sinh\theta_k$, which defines a one-to-one mapping of an N-dimensional **w** space to an N-dimensional θ space. Clearly, the MSE ε, when expressed as a function of θ_k, has a unique minimum at $\hat{\theta}_k=\sinh^{-1}\hat{w}_k$, $k=0,1,\ldots,N-1$. Further, the function $\sinh\theta_k$ is a monotonically increasing, continuous function of θ_k, as θ_k varies from -∞ to +∞, meaning that, $\frac{\partial\varepsilon^2}{\partial\theta_k}$ has the same sign as that of $\frac{\partial\varepsilon^2}{\partial w_k}$ everywhere within the respective parameter spaces. In other words, the MSE does not exhibit any local minimum in the θ space as well. We illustrate this in Fig. 1, where a 2-tap filter is considered with chosen MSE $\varepsilon=1-2w_0-w_1+2w_0^2+2w_1^2+2w_0w_1$. Fig.1(a) shows the constant MSE contours vs. w_0,w_1 while Fig. 1(b) shows the contours as functions of θ_0 and θ_1. Note that as we move from Fig. 1(a) to Fig. 1(b), the general

0-7803-7011-2/01/$10.00 ©2001 IEEE

nature of the MSE doesn't change and the minimum is mapped from $[\frac{1}{2},0]^t$ to $[0.482,0]^t$.

In the proposed scheme, a steepest descent search is taken up in the θ space in order to reach $\hat{\theta}$. The gradient $\nabla_\theta \varepsilon^2$ is easily seen to be given by $\nabla_\theta \varepsilon^2 = -2\,\Delta(\mathbf{p} - \mathbf{R}\mathbf{w})$, where $\mathbf{w} = [\sinh\theta_0, \sinh\theta_1, \ldots, \sinh\theta_{N-1}]^t$, $\mathbf{p} = E[\mathbf{x}(n)d(n)]$, $\mathbf{R} = E[\mathbf{x}(n)\mathbf{x}^t(n)]$ and Δ is a diagonal matrix with j-th diagonal entry given by $\Delta_{j,j} = \cosh\theta_j$, $j = 0, 1, \ldots, N-1$. The iterate $\theta(i)$ arising in the i-th step of iteration is then updated as :

$$\theta(i+1) = \theta(i) - \mu \nabla_\theta \varepsilon^2 \Big|_{\theta = \theta(i)}$$

where μ is some appropriate step size. To move from the steepest descent to the LMS form, we simply replace $\mathbf{R}$ and $\mathbf{p}$ by $\mathbf{x}(n)\mathbf{x}^t(n)$ and $\mathbf{x}(n)d(n)$ respectively in the expression for $\nabla_\theta \varepsilon^2$ in order to obtain an estimate of the gradient at index n. This leads to the so called HLMS algorithm as follows:

$$\theta(n+1) = \theta(n) + \mu\, \Delta(n)\mathbf{x}(n)e(n), \qquad (1)$$

$$e(n) = d(n) - \sum_{k=0}^{N-1} \sinh\theta_k(n)x(n-k) \qquad (2)$$

The HLMS algorithm is particularly suitable for CORDIC based realization, since the two quantities: $\sinh\theta_k(n)x(n-k)$ and $\cosh\theta_k(n)x(n-k)$, $k = 0,1, \ldots, N-1$, required for filtering by and updatation of the k-th coefficient respectively *can be computed simultaneously by engaging only one CORDIC processor*. For pipelined realization, it may, however, be more appropriate to consider the hyperbolic analog of an approximate version of the LMS algorithm, popularly known as the "Delayed LMS"(DLMS) algorithm [5], where the filter coefficients at the n-th index are updated using a past estimate of the gradient, say, for index $(n-L)$, where L is an integer. The correction term in the weight update formula then gets modified to $\mu\mathbf{x}(n-L)e(n-L)$ and the resulting L cycle delay in the error feedback path is used for retiming purpose. The hyperbolic analog of the DLMS algorithm can be easily worked out by substituting $\mathbf{R}$, $\mathbf{p}$ and $\mathbf{w}$ in the gradient expression by $\mathbf{x}(n-L)\mathbf{x}^t(n-L)$, $\mathbf{x}(n-L)d(n-L)$ and $[\sinh\theta_0(n-L), \sinh\theta_1(n-L), \ldots, \sinh\theta_{N-1}(n-L)]^t$ respectively and is given by

$$\theta(n+1) = \theta(n) + \mu\Delta(n-L)\mathbf{x}(n-L)e(n-L) \qquad (3)$$

The CORDIC algorithm provides an efficient way of implementing (2) and (3). This algorithm essentially rotates a two dimensional vector by running the iterations: $x_{i+1} = x_i - \delta_i 2^{-i} y_i$, $y_{i+1} = \delta_i 2^{-i} x_i + y_i$ and $\varepsilon_{i+1} = \varepsilon_i - \delta_i \arctan h\,(2^{-i})$, where $\delta_i = \mathrm{sgn}(\varepsilon_i)$, $i = 0,1,\ldots, M-1$, M is the required number of iterations to perform a hyperbolic operation. After M iterations, $x_M \to k(x_0\cosh\theta + y_0\sinh\theta)$, $y_M \to k(x_0\sinh\theta + y_0$ and $\varepsilon_M \to 0$ where $k = 1/\prod_{i=0}^{M-1}\cosh(\arctan h(2^{-i}))$ is the so called scale factor having a constant value for a particular machine (x_0, y_0) is the initial two dimensional vector, $\varepsilon_0 = \theta$ being the desired angle of rotation. Fig. 2 shows a CORDIC realization of a N tap DHLMS-based adaptive filter which achieves microlevel pipelining by using pipelined CORDIC processor units (PCU) [6] and pipelined multipliers (PM). Since the critical path delay arising from the PCU as well as from the PM amounts to that of a single adder/subtractor, this architecture can indeed process very high throughput data, typically of the order of hundreds of megahertz.

3. SIMULATION STUDIES AND DISCUSSION

Unlike the conventional LMS, it is very difficult to prove convergence of the HLMS and the delayed HLMS (DHLMS) algorithms analytically owing to the presence of nonlinearities in the form of hyperbolic quantities in (1), (2) and (3). Both the HLMS and the DHLMS algorithms, however, have been simulated extensively in the context of a wide class of applications and promising convergence results observed in each case. In this paper, we present simulation results for equalizing an AWGN channel with transfer function $H(z) = (1+2z^{-1})(1-0.5z^{-1})(1+1.1z^{-1})(1-0.6z^{-1})$ and noise variance 0.077. The transmitted symbols were chosen from an alphabet of 16 equispaced, equiprobable discrete amplitude levels with transmitted signal power of 10 dB. A 15 tap equalizer with centre placed at the 8-th tap position was used for equalizing the channel and a step size of $\mu = .0004$ (.00005 for DHLMS) was adopted for weight updatation. The resulting output error characteristics displayed in Fig. 3(a) (for HLMS) and in Fig. 3(b) (for DHLMS) by plotting MSE vs. n, confirms satisfactory convergence properties of the proposed method. These two Figures also represent comparative studies of the convergence performance between LMS and HLMS (Fig.3(a)) and also between DLMS and DHLMS (Fig.3(b)) algorithms. For this, we have also plotted the variable $\eta(n)$, gives by the ratio of the MSE under LMS (DLMS) to the MSE under HLMS (DHLMS) algorithm.

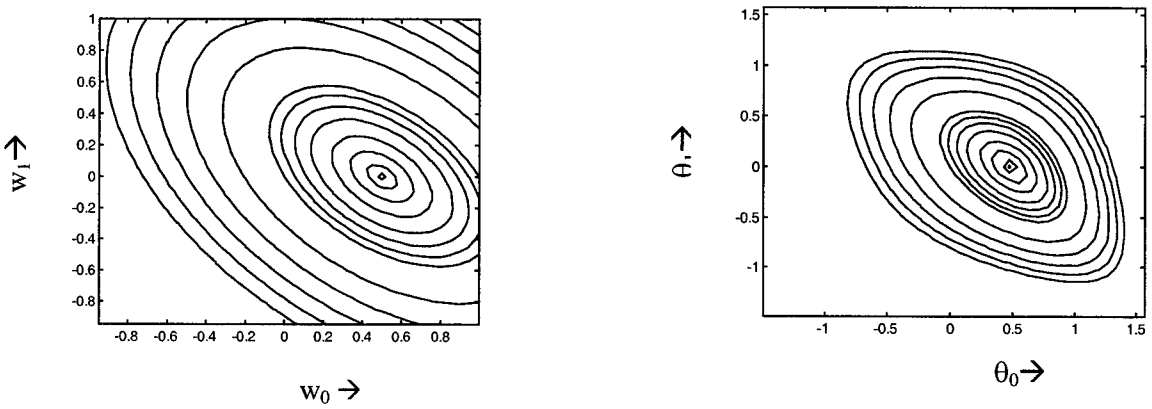


Fig. 1 Error performance contours as functions of (a) w_0 and w_1 and (b) θ_0 and θ_1.

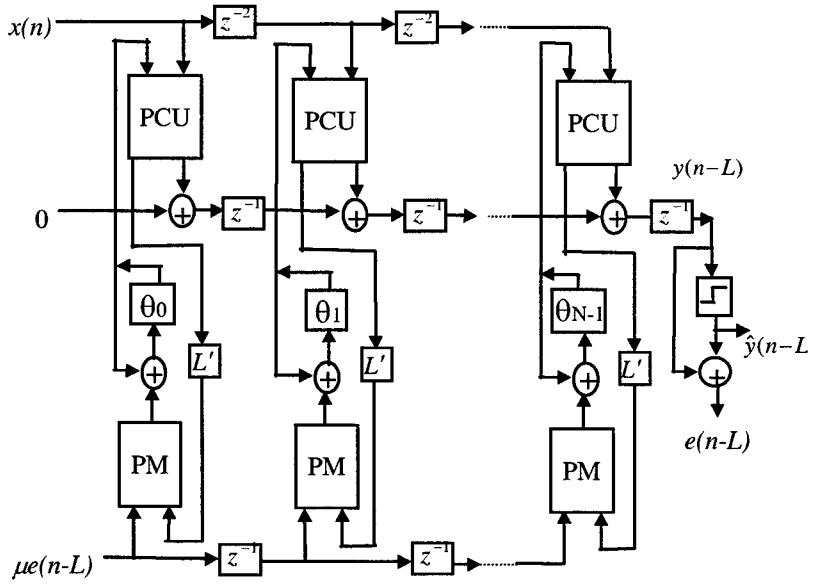


Fig 2. DHLMS based adaptive filter

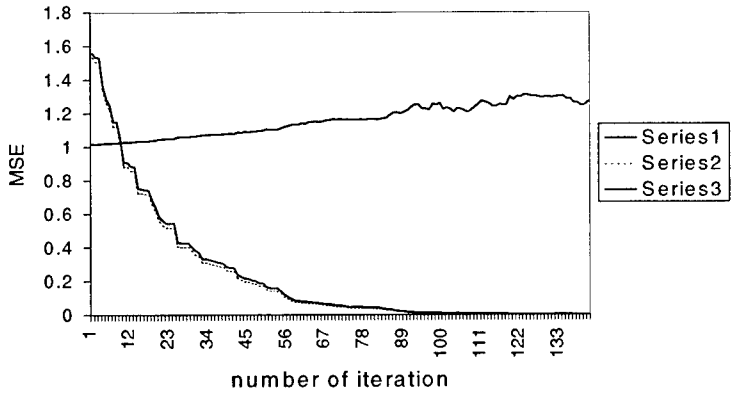


Fig. 3(a). Convergence performance of the LMS (series 1) and HLMS (series 2) algorithms, series 3 represents η.

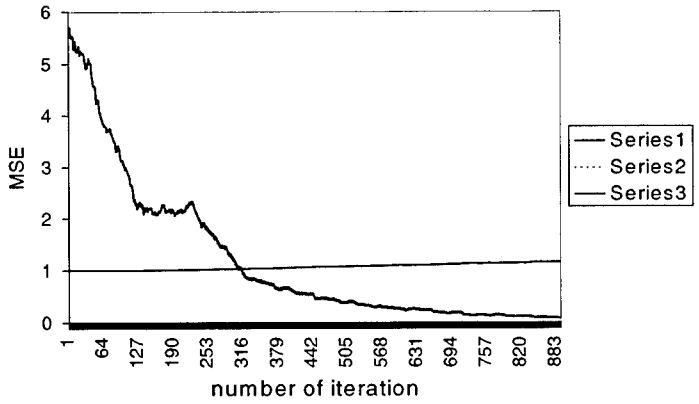


Fig. 3(b). Convergence performance of the DLMS (series 1) and DHLMS (series 2) algorithms, series 3 represents η.

5. REFERENCES

[1] J. E. Volder, "The CORDIC trigonometric computing technique," *IRE Trans. Electron. Comput.*, vol. EC-8, no. 3, pp. 330-334, Sept. 1959.

[2] Y. H. Hu, "CORDIC-Based VLSI architecture for digital signal processing", *IEEE Signal Processing Magazine*, vol. 9, no 3, pp 16-3, July 1992..

[3] Y. H. Hu and H. E. Liao, "CALF: a CORDIC adaptive lattice filter", *IEEE Trans. Signal Processing*", vol. 40, no. 4, pp. 990-993, April 1992.

[4] Yu Hen Hu, "On the convergence of the CORDIC adaptive lattice filtering (CALF) algorithm", *IEEE Trans. Signal Processing*, vol. 46, no. 7, pp. 1861-1871, July 1998.

[5] G.Long, F. Ling, J. G. Proakis, "The LMS algorithm with delayed coefficient adaptation", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1397-1405, September 1989.

[6] P. Dewilde, E. F. Deprettere and R. Nouta, " Parallel and pipelined VLSI implementation of signal processing algorithms", in VLSI and modern signal processing, S. Y. Kung et al, Eds., Prentice Hall series, 1985.

ON EXTERNAL CALIBRATION OF ANALOG-TO-DIGITAL CONVERTERS

Henrik Lundin, Mikael Skoglund and Peter Händel

Department of Signals, Sensors and Systems
Royal Institute of Technology
SE-100 44 Stockholm, Sweden

ABSTRACT

External calibration and compensation of analog-to-digital converters is considered. Two novel methods are presented. Both methods employ multiple calibration frequencies in order to improve the wide-band performance of the converter. Also, a dynamic table indexing is introduced to further improve the performance. A recursive sine-wave reconstruction filter algorithm is developed for calibration purposes.The proposed methods are evaluated using experimental converter data. Results indicate that the dynamic indexing provides good correction performance, also at frequencies not used during calibration. Thus, wide-band calibration can be achieved.

1. INTRODUCTION

The need for broad-band analog-to-digital converters (ADCs) is increasing rapidly. In third-generation mobile communications, for instance, the broad-band linearity of the radio receiver ADC is a crucial property. It is a well-known fact that practical AD converters suffer from various errors, e.g., gain errors, offset errors and linearity errors. These errors stem from numerous sources such as non-ideal spacing of transition levels and timing jitter, to mention a few, and they contribute to deterioration of the broad-band performance of the converter. Several methods have been proposed to *externally* compensate for such errors, e.g.,[5, 8]. External in this case implies that digital signal processing methods are used in the calibration and compensation schemes, which operate outside of the actual converter.

The methods presented in this paper extend that of [4], which is briefly recapitulated in this section. The scheme operates in two different modes: calibration and compensation, where the latter is engaged during normal ADC operation and the former is the process of analyzing the errors of the ADC. During compensation a compensation table of size 2^N, where N is the number of bits in the ADC output, is used to map each possible converter output x_i into a corresponding corrected value s_i. Thus, the ADC operation becomes $s(t) \rightarrow x_i \rightarrow s_i$. This is illustrated in Fig. 1. The corrected values $\{s_j\}$ are chosen to minimize the MSE criterion in accordance with [6], where it is understood that the corrected value s_i should equal the mean of all input values $s(t)$ that yield the ADC output x_i.

Calibration is performed with a sinusoidalreference signal. By using optimal sinusoid-reconstruction filtering [3], the input reference signal is reconstructed in the digital domain to form an estimate $\hat{s}(k)$, as shown in Fig. 2. Hence, the ADC error characteristics can be analyzed using $\hat{s}(k)$ and the sampled signal $x(k)$. It

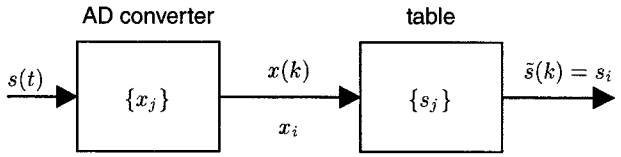


Fig. 1. Basic compensation system utilizing table mapping. The output of the ADC is used as address into the compensation table to produce a compensated value.

should be noted that the calibration is implemented without any reference device, such as a "better" ADC or a digitally generated reference signal (e.g., [10]). In [4] only one single frequency is used as reference signal, and the results indicate that the error correction performance deteriorates when evaluated at off-calibration frequencies.

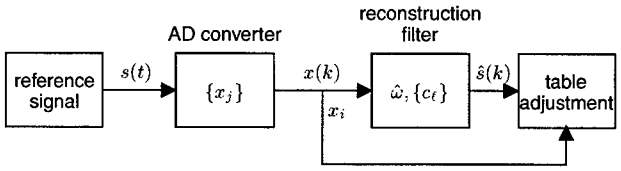


Fig. 2. Calibration system utilizing a reference signal reconstruction filter.

The first extension of the proceduredescribed above is the *sequential multi-tone calibrator* described in Section 2. In Section 3 the table indexing is altered to take signal dynamics into account. Finally, in Section 4 the methods introduced are evaluated using experimental ADC data.

2. SEQUENTIAL MULTI-TONE CALIBRATION

An improvement to the methods of [4] is a recursive sine-wave reconstruction filter algorithm based on [3]. The filter algorithm is applied in the *sequential multi-tone calibrator* presented in this section, as well as in the *dynamic calibrator* in Section 3. The adaptive behavior of the filter makes the calibration procedure robust and enables the employment of a sequential multi-tone calibration signal.

Single-tone calibration of ADCs usually results in peak performance at or near the selected calibration frequency. In the sequential multi-tone calibrator, the same correction table building proce-

0-7803-7011-2/01/$10.00 ©2001 IEEE

dure as in [4] is utilized but with several sinusoids with different frequencies applied subsequently. In other words, the reference signal $s(t)$ input to the ADC is described by

$$s(t) = A\sin\big(2\pi f(t)t + \phi\big), \tag{1a}$$

$$f(t) = \begin{cases} f_0 & 0 \le t < t_1, \\ f_1 & t_1 \le t < t_2, \\ \vdots & \\ f_{F-1} & t \ge t_{F-1}, \end{cases} \tag{1b}$$

where $\{f_j\}$ are the different calibration frequencies and $\{t_j\}$ are the time instants where the frequency changes. It should be understood that neither $\{f_j\}$ nor $\{t_j\}$ have to be uniformly spaced sequences, and $\{f_j\}$ does not have to be monotonically increasing or decreasing. Obviously, $\{t_j\}$ must be an increasing sequence.

The recursive reconstruction filter comprises two parts. The first part is an LMS-based frequency estimation, producing a rough estimate of the reference signal frequency. The second part is a sine-wave reconstruction filter, based on [3], which adapts to the reference signal through an LMS recursion. The reconstruction filter utilizes the frequency estimate to ensure rapid convergence to the global minimum. See [7] for a thorough description of the algorithms. By using this adaptive reconstruction filter the calibration frequencies $\{f_j\}$ and the time instants $\{t_j\}$ in (1) can be unknown to the calibration algorithms. Thus, the communication between the reference signal generator and the calibration algorithms is reduced to a binary control signal (*calibrate* or *compensate*).

The calibration of the ADC is summarized in Table 1, where the averaging of reconstructed input samples is implemented as a running average. The outcome of using several calibration frequencies in this manner is that the correction table will compensate for the *average* error over all calibration frequencies. During compensation the same compensated value s_i will be returned for a certain ADC output x_i regardless of the signal frequency. Thus, true dynamic compensation has not been achieved, although the performance has been improved compared with that of [4] with a small increase in complexity. The next section will discuss an extension to true dynamic compensation.

Table 1. Summary of the sequential multi-tone calibration procedure.

1. Initialize the table, e.g., $s_i = x_i$.
2. Apply the reference signal $s(t)$ in (1) to the ADC input.
3. Sample the reference signal to produce one sample $x(k)$.
4. Calculate the reference signal estimate $\hat{s}(k)$ using the recursive reconstruction filter.
5. If the filter has converged and $x(k) = x_i$, update table entry s_i as $s_i \rightarrow \frac{A_i(k)\, s_i + \hat{s}(k)}{A_i(k)+1}$ where $A_i(k)$ is the number of times s_i has been updated before.
6. Repeat from 3 with the next sample.

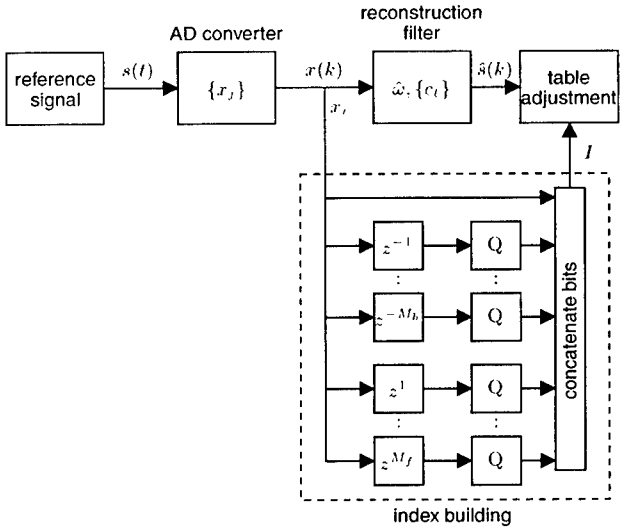


Fig. 3. Calibration system with dynamic table indexing. Q denotes a quantizer, i.e., a reduction of the number of bits.

3. DYNAMIC CALIBRATION

Since the errors sought to compensate for are in most cases frequency dependent, it would be desirable if the addressing of the compensation table involved information about the signal dynamics. Such an approach is now introduced.

The scheme utilizes a compound index I to address the table. For each sample $s(k)$ of the input signal, the ADC maps the value into an output $x_i(k)$ where the index $i = i(k)$ at time-instant k is in $\{0, 1, \ldots, N-1\}$. The index $i(k)$ can be represented as a binary number using N bits,

$$\begin{aligned} i(k) &= b_{N-1}(k)2^{N-1} + b_{N-2}(k)2^{N-2} + \ldots + b_0(k)2^0 \\ &= [b_{N-1}b_{N-2}\cdots b_0]_2\,(k), \end{aligned} \tag{2}$$

where b_{N-1} is the most significant bit (MSB), and where $[\cdot]_2$ denotes binary representation.

Now, let $\tilde{i}(k)$ be a binary number consisting of the K most significant bits in $i(k)$, such as

$$\tilde{i}(k) = [b_{N-1}b_{N-2}\cdots b_{N-K}]_2\,(k). \tag{3}$$

Let the compound index $I(k)$ at time-instant k consist of all N bits of $i(k)$ followed by the M_b previous $\tilde{i}$-s and the M_f future $\tilde{i}$-s,

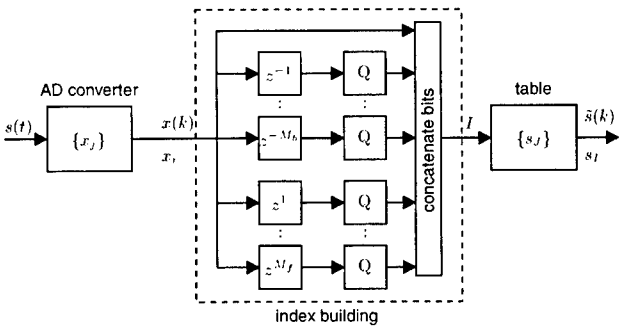
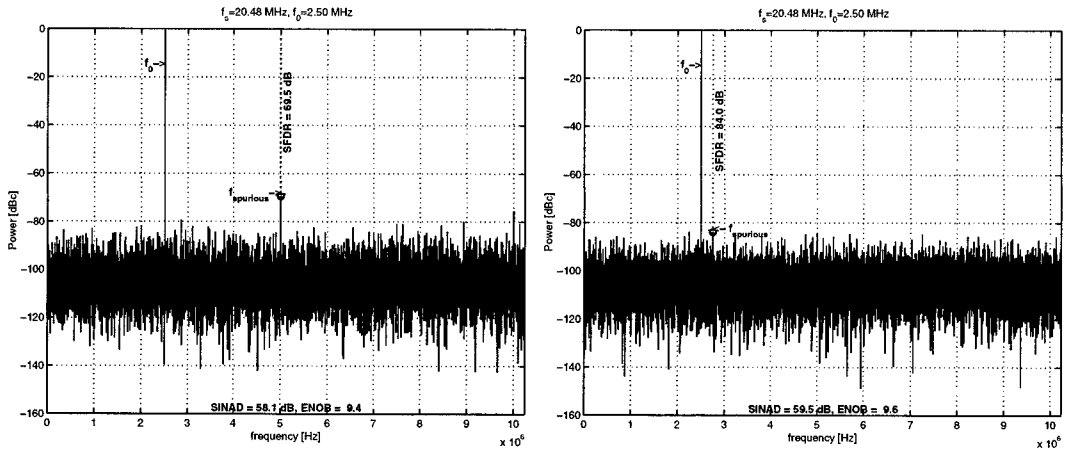


Fig. 4. Compensation system with dynamic table indexing.



(a) (b)

Fig. 5. Typical output spectrum of uncalibrated (left) and calibrated (right) ADC, in both cases employed with a full-scale single sine-wave. The improvement is given in terms of SFDR and SINAD.

according to

$$I(k) = [i(k)\,\tilde{i}(k-1)\cdots\tilde{i}(k-M_b)\,\tilde{i}(k+1)\cdots\ \tilde{i}(k+M_f)]_2 \tag{4}$$

Note that K in (3) can be different for the various lags in (4), i.e., the different $\tilde{i}$-s need not be of the same number of bits.

Calibration is performed in the same manner as in the sequential multi-tone method (see Table 1)using the compound index $I(k)$ for table addressing instead of the ADC output $x(k)$, as depicted in Fig. 3. Correspondingly, the compensation scheme is altered to include the index $I(k)$, so that the complete ADC operation becomes $s(t) \rightarrow x_i \rightarrow I \rightarrow s_I$, which is shown in Fig. 4. The compensation table is expanded to include the 2^M elements S_J, where M is the total number of bits in I.

4. PERFORMANCE

Both methods introduced in this paper have been evaluated using experimental ADC data from an Analog Devices AD 876 10-bit 20 MSPS converter. Fig. 5(a) shows a typical single sine-wave spectrum of the ADC output without compensation and Fig. 5(b) shows typical results after compensation. Performance improvement can bemeasured in terms of signal-to-noise and distortion ratio (SINAD)[1] and spurious free dynamic range (SFDR) [2, 9].

4.1. Sequential multi-tone calibration

The sequential multi-tone calibration method is evaluated for performance at off-calibration frequencies, i.e., frequencies *not* used for calibration. Calibration is done at seven frequencies dispersed over the Nyquist band. The results are presented in Fig. 6. The results are also compared with single-tone calibration with evaluation at off-calibration frequencies, and the comparison is presented in Table 2.

[1] SINAD is introduced in IEEE Std 1241 [2, 9] and is equivalent to SNR in IEEE Std 1057 [1].

Table 2. Comparison between sequential multi-tone calibration and single-tone calibration with evaluation at other frequencies than the calibration frequency.

	SFDR improvement [dB]	
	Multi-tone	Single-tone
Average	5.0	3.3
Minimum	-0.8	-6.4

4.2. Dynamic calibration

The performance of the dynamic compensation system was tested for different index configurations. Let the notation

$$\langle K_{M_b}\ \ldots\ K_1\,\mathbf{N}\,K_{-1}\ \ldots\ K_{-M_f}\rangle$$

implythat the index is built starting with all N bits of $i(k)$ (indicated by the bold N), followed by the K_1 most significant bits of $i(k-1)$ up to the K_{M_b} most significant bits of $i(k-M_b)$, and finally the K_{-1} most significant bits of $i(k+1)$ up to the K_{-M_f} most significant bits of $i(k+M_f)$. For example, $\langle 4\,2\,\mathbf{10}\,4\rangle$ means that the index is the concatenation of all 10 bits of $i(k)$, the 2 MSBs of $i(k-1)$, the 4 MSBs of $i(k-2)$ and the 4 MSBs of $i(k+1)$.

The tests were conducted in the following manner for every index configuration:

1. Reset and initialize the compensation table.
2. Calibrate the table using n_{fcal} different frequencies, chosen at random in the Nyquist band.
3. Evaluate the compensation performance at n_{feval} different frequencies, chosen at random but *not* coinciding with any of the frequencies used for calibration. Calculate the SFDR improvement for each evaluation frequency.
4. Repeat from 1. Test each configuration n_{test} times.

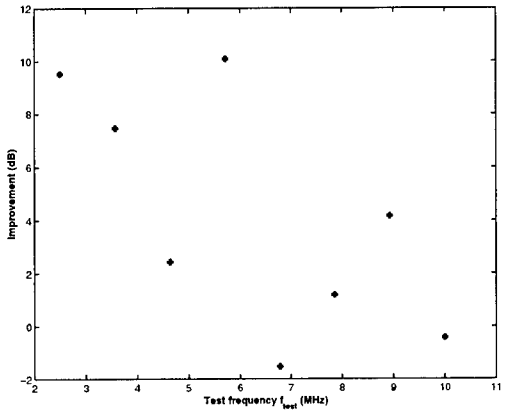


Fig. 6. SFDR improvement when calibrating at seven different frequencies and evaluating at a frequency f_{test} not used for calibration.

In the presentation below, the median improvement over all tests for every configuration is presented. Also, as an indication of how well the calibration signals managed to excite the compensation table, the 'fill rate' is presented. This is nothing but the percentage of the table entries that were actually altered during calibration.

The test illustrates the impact of different index configurations. In this example, the number of calibration frequencies $n_{\text{fcal}} = 50$, the number of evaluation frequencies $n_{\text{feval}} = 20$ and the number of tests per configuration $n_{\text{test}} = 10$. Each calibration frequency was maintained for 16 384 samples. Fig. 7 shows the outcome of the test and the different configurations used.

5. CONCLUSIONS

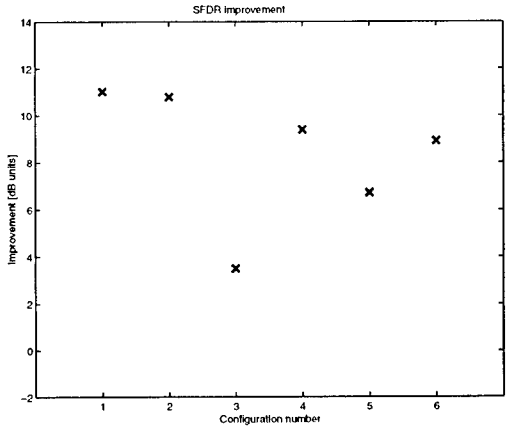
Two methods to make ADC calibration work well in a wider frequency range have been introduced. Both are based on look-up table compensation with reference signal reconstruction in the digital domain. The first method, referred to as sequential multi-tone calibration, subsequently applies several sinusoids with different frequencies as reference signal to the ADC during calibration. This results in a better performance in a wider frequency range. In the second method, referred to as dynamic calibration, the look-up table indexing is altered to a compound index consisting of the present sample together with (quantized versions of) past and future samples. Through this scheme the compensation depends on the signal dynamics, as do the ADC errors.

The proposed schemes have been evaluated with experimental AD converter data. The results indicate that the wide-band performance, in terms of SFDR, of the calibration schemes is superior to that of [4].

6. REFERENCES

[1] *IEEE Standard for Digitizing Waveform recorders*. IEEE Std. 1057-1994.

[2] *IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters*, draft edition, Mar. 2000. IEEE Std. 1241.



Number	Configuration	Table fill rate
1	⟨4 4 **10**⟩	19%
2	⟨**10** 4 4⟩	19%
3	⟨8 **10**⟩	36%
4	⟨1 1 1 1 1 1 1 1 **10**⟩	15%
5	⟨4 **10**⟩	93%
6	⟨2 2 2 **10**⟩	39%

Fig. 7. SFDR improvement using different index configurations.

[3] P. Händel. Predictive digital filtering of sinusoidal signals. *IEEE Transactions on Signal Processing*, 46(2):364–374, Feb. 1998.

[4] P. Händel, M. Skoglund, and M. Pettersson. A calibration procedure for imperfect quantizers. *IEEE Transactions on Instrumentation and Measurement*, 49:1063–1068, Oct. 2000.

[5] F. H. Irons, D. M. Hummels, and S. P. Kennedy. Improved compensation for analog-to-digital converters. *IEEE Transactions on Circuits and Systems*, 38(8):958–961, Aug. 1991.

[6] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, IT-28(2):129–137, Mar. 1982.

[7] H. Lundin. Dynamic compensation of analogue-to-digital converters. Technical Report IR-SB-EX-0023, Dept. of Signals, Sensors and Systems, Royal Institute of Technology, Stockholm, Sweden, Dec. 2000. ftp://ftp.s3.kth.se/pub/signal/reports/exjobb/00/IR-SB-EX-0023.pdf.

[8] D. Moulin. Real-time equalization of A/D converter nonlinearities. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, volume 1, pages 262–267. IEEE, 1989.

[9] S. J. Tilden, T. E. Linnenbrink, and P. J. Green. Overview of the IEEE-STD-1241 "Standard for Terminology and Test Methods for Analog-to-Digital Converters". In *Proceedings of the 16th IEEE Instrumentation and Measurements Technology Conference, 1999. IMTC/99, vol. 3*, pages 1498–1503. IEEE, 1999.

[10] J. Tsimbinos and K. V. Lever. Improved error-table compensation of A/D converters. *IEE Proceedings - Circuits, Devices and Systems*, 144(6):343–349, Dec. 1997.

SEMI-BLIND CHANNEL ESTIMATION FOR BLOCK PRECODED SPACE-TIME OFDM TRANSMISSIONS

Shengli Zhou[1*], *Bertrand Muquet*[2] and *Georgios B. Giannakis*[1*]

[1]Dept. of ECE, University of Minnesota, 200 Union Str. SE, Minneapolis, MN 55455
[2]Motorola Labs Paris, Espace Technologique, Saint-Aubin, 91193 Gif-sur-Yvette, France

ABSTRACT

We develop in this paper a (semi-) blind channel estimation algorithm for space time (ST) block precoded OFDM transmissions over frequency-selective channels. We establish that multi-channel identifiability is guaranteed up to one or two scalar ambiguities, when distinct or identical precoders are employed for even and odd indexed symbol blocks. With known pilots inserted before precoding, we resolve the residual scalar ambiguities and show that distinct precoders require less pilots than identical precoders to achieve the same channel estimation accuracy. Simulation results confirm our theoretical analysis and illustrate that the proposed semi-blind algorithm is capable of tracking slow channel variations and improving the overall system performance relative to competing differential ST alternatives.

1. INTRODUCTION

New applications such as high speed Internet access and wireless digital television call for high data rate transmissions. Usage of multiple transmit- and receive-antennas has the potential to increase the channel capacity, and thus the maximum achievable rate. Equipped with Space-Time Coding (STC) at the transmitter and intelligent signal processing at the receiver, multi-antenna transceivers offer also diversity and coding advantages over single antenna systems (see [4,6] for tutorial treatments). But all these enhancements in capacity, diversity and coding gains can be realized if the underlying channels can be acquired at the receiver.

Conventionally, training symbols are transmitted periodically to assist the receiver in acquiring channel state information (CSI), see e.g., [2] for ST-OFDM systems. However, training sequences consume bandwidth and thereby incur spectral efficiency loss especially in rapidly varying environments. For this reason, blind channel estimators receive growing attention. Relying on non-redundant and non-constant modulus precoding, [1] proposed blind channel estimation and equalization for OFDM-based multi-antenna systems using cyclostationary statistics. For ST-OFDM, a deterministic blind channel estimator was derived in [3] when the channel transfer functions are coprime (no common zeros) and the transmitted signals have constant-modulus (CM).

In this paper, we deal with a linearly precoded ST-OFDM system with two transmit antennas and derive (semi-) blind channel identification algorithms for frequency-selective FIR channels. With properly designed redundant precoders, the proposed subspace-based blind channel estimator possesses the following three attractive features: i) it can be applied to arbitrary signal constellations; ii) it guarantees channel identifiability regardless of the underlying channel zero locations; iii) it can estimate multiple channels simultaneously up to one or two scalar ambiguities.

To enable channel equalization, we also show how to resolve the residual scalar ambiguities using a minimal number of pilots that we insert before precoding.

Notation: Bold upper (lower) letters denote matrices (column vectors); $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^{\mathcal{H}}$ denote conjugate, transpose and Hermitian transpose; $\mathcal{R}(\cdot)$ stands for range space; $\mathbf{I}_K$ denotes the identity matrix of size K and $\mathbf{0}$ denotes an all-zero matrix or vector; $\mathrm{diag}(\mathbf{x})$ will stand for a diagonal matrix with $\mathbf{x}$ on its diagonal; $[\cdot]_p$ denotes the pth entry of a vector, and $[\cdot]_{p,q}$ denotes the (p,q)th entry of a matrix.

2. SYSTEM DESCRIPTION

Figure 1 depicts the wireless system considered in this paper, where the ST transceiver is equipped with two transmit antennas and one receive antenna as in [4]. Prior to transmission, the information bearing symbols are first grouped into blocks $\mathbf{s}(n)$ of size $K \times 1$. Two different linear block precoders denoted by the tall $J \times K$ matrices $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$, one for the even block indices $2n$ and one for the odd indices $2n+1$, are used to introduce redundancy ($J > K$). The corresponding $J \times 1$ precoded blocks

$$\tilde{\mathbf{s}}(2n) := \mathbf{\Theta}_1 \mathbf{s}(2n) \quad \text{and} \quad \tilde{\mathbf{s}}(2n+1) := \mathbf{\Theta}_2 \mathbf{s}(2n+1), \quad (1)$$

are fed to the ST encoder $\mathcal{M}(\cdot)$. The ST encoder takes as input two consecutive precoded blocks, $\tilde{\mathbf{s}}(2n)$ and $\tilde{\mathbf{s}}(2n+1)$, to output the following $2J \times 2$ code matrix:

$$\begin{bmatrix} \bar{\mathbf{s}}_1(2n) & \bar{\mathbf{s}}_1(2n+1) \\ \bar{\mathbf{s}}_2(2n) & \bar{\mathbf{s}}_2(2n+1) \end{bmatrix} := \begin{bmatrix} \tilde{\mathbf{s}}(2n) & -\tilde{\mathbf{s}}^*(2n+1) \\ \tilde{\mathbf{s}}(2n+1) & \tilde{\mathbf{s}}^*(2n) \end{bmatrix}.$$

Each block column of this matrix is transmitted over successive time intervals with the blocks $\bar{\mathbf{s}}_1(n)$ and $\bar{\mathbf{s}}_2(n)$ sent through transmit-antennas 1 and 2, respectively.

The frequency-selective channels between the two transmit antennas and the receive antenna can be modeled as FIR linear time-invariant filters with impulse responses $\mathbf{h}_i := [h_i(0), \ldots, h_i(L)]$, $i = 1,2$, where L is an upper bound on the channel orders of $\mathbf{h}_1$ and $\mathbf{h}_2$. Moreover, we assume that OFDM modulation has been deployed to convert the FIR channels into a set of parallel flat faded subchannels (see e.g., [8] for detailed derivations). Let $\mathcal{D}_1$ and $\mathcal{D}_2$

*Supported by the NSF Wireless Initiative grant no. 99-79443.

0-7803-7011-2/01/$10.00 ©2001 IEEE

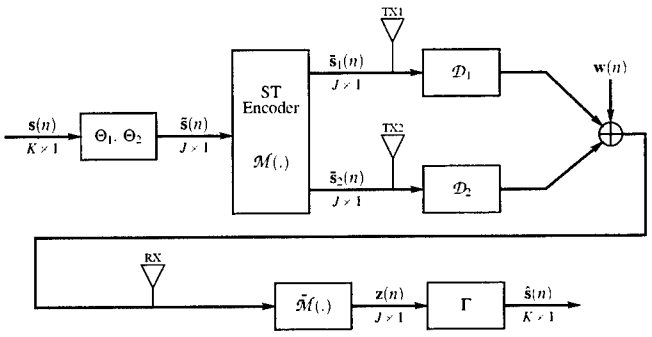


Fig. 1. Block precoded ST-OFDM transceiver model

be the diagonal matrices corresponding to these subchannels: $\mathcal{D}_i := \text{diag}[H_i(0)\dots H_i(J-1)]$, where $H_i(k) := \sum_{l=0}^{L} h_i(l)e^{-j\frac{2\pi}{J}lk}$. Considering two successive received blocks: $\bar{\mathbf{y}}(2n)$ and $\bar{\mathbf{y}}(2n+1)$, let us define the super blocks $\check{\mathbf{y}}(n)$ and $\check{\mathbf{s}}(n)$ as: $\check{\mathbf{y}}(n) := [\bar{\mathbf{y}}^T(2n), \bar{\mathbf{y}}^{\mathcal{H}}(2n+1)]^T$ and $\check{\mathbf{s}}(n) := [\mathbf{s}^T(2n), \mathbf{s}^T(2n+1)]^T$. Letting $\check{\mathbf{w}}(n)$ be the additive noise, the received block $\check{\mathbf{y}}(n)$ can then be expressed as (see also [4] for further details):

$$\check{\mathbf{y}}(n) = \mathcal{D}\Phi_{12}\check{\mathbf{s}}(n) + \check{\mathbf{w}}(n) := \mathcal{H}\check{\mathbf{s}}(n) + \check{\mathbf{w}}(n), \quad (2)$$

where $\mathcal{D}, \Phi_{12}, \mathcal{H}$ are defined respectively as:

$$\mathcal{D} := \begin{bmatrix} \mathcal{D}_1 & \mathcal{D}_2 \\ \mathcal{D}_2^* & -\mathcal{D}_1^* \end{bmatrix}, \ \Phi_{12} := \begin{bmatrix} \Theta_1 & 0 \\ 0 & \Theta_2 \end{bmatrix}, \ \mathcal{H} := \mathcal{D}\Phi_{12}.$$

When the channel matrices $\mathcal{D}_1$ and $\mathcal{D}_2$ become available at the receiver, it is possible to demodulate $\check{\mathbf{y}}(n)$ with diversity gains by a simple matrix multiplication:

$$\check{\mathbf{z}}(n) = \mathcal{D}^{\mathcal{H}}\check{\mathbf{y}}(n) = \begin{bmatrix} \bar{\mathcal{D}}_{12}\Theta_1 & 0 \\ 0 & \bar{\mathcal{D}}_{12}\Theta_2 \end{bmatrix} \check{\mathbf{s}}(n) + \mathcal{D}^{\mathcal{H}}\check{\mathbf{w}}(n), \quad (3)$$

where the diagonal matrix $\bar{\mathcal{D}}_{12} := \mathcal{D}_1^*\mathcal{D}_1 + \mathcal{D}_2^*\mathcal{D}_2$ equals $\text{diag}[\sum_{i=1}^{2}|H_i(e^{j0})|^2, \dots, \sum_{i=1}^{2}|H_i(e^{j\frac{2\pi}{J}(J-1)})|^2]$. Eq. (3) reveals that zero-forcing recovery of $\check{\mathbf{s}}(n)$ from $\check{\mathbf{z}}(n)$ requires the matrices $\bar{\mathcal{D}}_{12}\Theta_i, i \in [1,2]$, to have full column rank. Because the channels have maximum order L, $\bar{\mathcal{D}}_{12}$ has at most L zero diagonal entries. Hence, the full rank of $\bar{\mathcal{D}}_{12}\Theta_i$ can be always assured if we adopt the following design conditions on the block lengths and the linear precoders:

a1) $J > K + L$;

a2) $\Theta_i, i \in \{1,2\}$, *is designed so that any* K *rows of* Θ_i *are linearly independent.*

Based on a1) and a2), our objective in this paper is to develop a subspace-based (semi-) blind multichannel estimation algorithm.

3. (SEMI-) BLIND MULTICHANNEL ESTIMATION

At the receiver, we collect N received blocks $\check{\mathbf{y}}(n)$, with:

a3) N *large enough* ($\geq 2K$) *so that* $\mathbf{S}_N\mathbf{S}_N^{\mathcal{H}}$ *has full rank* $2K$, *where* $\mathbf{S}_N := [\check{\mathbf{s}}(0), \dots, \check{\mathbf{s}}(N-1)]$.

Under a1), a2) and a3), a consistent blind channel estimator has been developed in [4, 9]. We summarize the resulting algorithm in the following steps:

S1. Collect the received data blocks $\check{\mathbf{y}}(n)$ and compute $\mathbf{R}_{\check{\mathbf{y}}\check{\mathbf{y}}}^{(N)} = (1/N)\sum_{n=0}^{N-1}\check{\mathbf{y}}(n)\check{\mathbf{y}}^{\mathcal{H}}(n)$;

S2. Determine the eigenvectors $\tilde{\mathbf{u}}_k$, $k = 1, \dots, 2J-2K$ corresponding to the smallest $2J-2K$ eigenvalues of matrix $\mathbf{R}_{\check{\mathbf{y}}\check{\mathbf{y}}}^{(N)}$; split each vector $\tilde{\mathbf{u}}_k$ into its upper and lower parts as: $\tilde{\mathbf{u}}_k = [\hat{\mathbf{u}}_k^T, \check{\mathbf{u}}_k^T]^T$ and form the matrix

$$\mathcal{D}(\tilde{\mathbf{u}}_k) := \begin{bmatrix} \text{diag}(\hat{\mathbf{u}}_k^*) & -\text{diag}(\check{\mathbf{u}}_k) \\ \text{diag}(\check{\mathbf{u}}_k^*) & \text{diag}(\hat{\mathbf{u}}_k) \end{bmatrix}.$$

S3. From these eigenvectors, estimate $[\mathbf{h}_1^T, \mathbf{h}_2^{\mathcal{H}}]^T$ as the left eigenvector corresponding to the smallest eigenvalue of $\mathbf{Q}$, where $\mathbf{Q}$ is defined as:

$$\mathbf{Q} := \mathcal{F}[\mathcal{D}(\tilde{\mathbf{u}}_1)\Psi, \dots, \mathcal{D}(\tilde{\mathbf{u}}_{2J-2K})\Psi], \quad (4)$$

with $\mathcal{F} := \begin{bmatrix} \mathbf{V}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^{\mathcal{H}} \end{bmatrix}$, $\Psi := \begin{bmatrix} \Theta_1 & \mathbf{0} \\ \mathbf{0} & \Theta_2^* \end{bmatrix}$, and $\mathbf{V}$ a tall Vandermonde matrix with $[\mathbf{V}]_{p+1,q+1} = e^{-j\frac{2\pi}{J}pq}$.

An inherent problem to all subspace based estimators is their relatively slow convergence with respect to the number of data required. To facilitate data efficiency and also enable tracking of slow channel variations, a semi-blind implementation of the subspace based method can be devised by capitalizing on training sequences, which are anyways present for synchronization and quick channel acquisition in practical systems. Proceeding as in [5], the *semi-blind implementation* of our algorithm is outlined next:

1. Obtain initial channel estimates $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$ (and thus $\hat{\mathcal{H}}$) through training (using e.g., [2]); and estimate the autocorrelation matrix $\mathbf{R}_{\check{\mathbf{y}}\check{\mathbf{y}}}$ as (σ_s^2 denotes symbol energy): $\mathbf{R}_{\check{\mathbf{y}}\check{\mathbf{y}}}^{(0)} = \sigma_s^2\hat{\mathcal{H}}\hat{\mathcal{H}}^{\mathcal{H}}$.
2. Refine iteratively the autocorrelation matrix each time a new symbol block $\check{\mathbf{y}}(N)$ becomes available using a rectangular sliding window of length W:

$$\mathbf{R}_{\check{\mathbf{y}}\check{\mathbf{y}}}^{(N)} = \mathbf{R}_{\check{\mathbf{y}}\check{\mathbf{y}}}^{(N-1)} + \frac{1}{W}[\check{\mathbf{y}}(N)\check{\mathbf{y}}^{\mathcal{H}}(N) - \check{\mathbf{y}}(N-W)\check{\mathbf{y}}^{\mathcal{H}}(N-W)]. \quad (5)$$

3. Perform the subspace algorithm based on $\mathbf{R}_{\check{\mathbf{y}}\check{\mathbf{y}}}^{(N)}$.

4. CHANNEL IDENTIFIABILITY

The key question here is whether the solution of S3 is unique. With the proof provided in [9], we present channel identifiability results for two precoder choices: identical precoders and distinct precoders.

Theorem 1 (identical precoders): *Suppose a1), a2) and a3) hold true; if* $\Theta_1 = \Theta_2 = \Theta$, *the matrix* $\mathbf{Q}$ *in (4) loses row rank by two and the resulting estimate* $[\mathbf{h}_3^T, \mathbf{h}_4^{\mathcal{H}}]^T$ *belongs to a two-dimensional vector space that is spanned by* $\mathbf{h}_{12} = [\mathbf{h}_1^T, \mathbf{h}_2^{\mathcal{H}}]^T$ *and* $\mathbf{h}_{21} = [\mathbf{h}_2^T, -\mathbf{h}_1^{\mathcal{H}}]^T$. *The underlying channels are identified up to two scalar ambiguities as:*

$$\begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} = \frac{1}{|\alpha_1|^2 + |\alpha_2|^2} \begin{bmatrix} \alpha_1^*\mathbf{I} & -\alpha_2\mathbf{I} \\ \alpha_2^*\mathbf{I} & \alpha_1\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{h}_3 \\ \mathbf{h}_4 \end{bmatrix}. \quad (6)$$

Theorem 2 (distinct precoders): *Suppose a1), a2) and a3) hold true; let* $\bar{\mathbf{D}}$ *denote any diagonal matrix with unit amplitude diagonal entries, and* $\bar{\mathbf{\Theta}}_1$, $\bar{\mathbf{\Theta}}_2$ *be formed by any* $J-L$ *rows of* $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$, *respectively. If* $\bar{\mathbf{\Theta}}_1$ *and* $\bar{\mathbf{\Theta}}_2$ *satisfy:* $\bar{\mathbf{D}}\bar{\mathbf{\Theta}}_1 \notin \mathcal{R}(\bar{\mathbf{\Theta}}_2)$, *the resulting estimate* $[\mathbf{h}_3^T, \mathbf{h}_4^{\mathcal{H}}]^T$ *is unique up to a constant and thus channel identifiability within one scalar is guaranteed:*

$$\begin{bmatrix}\mathbf{h}_1\\ \mathbf{h}_2\end{bmatrix} = \alpha \begin{bmatrix}\mathbf{h}_3\\ \mathbf{h}_4\end{bmatrix}. \tag{7}$$

Therefore, from the received data only, multiple channels can be estimated simultaneously up to one or two scalar ambiguities with linearly block precoded ST-OFDM transmissions. To enable channel equalization, we show next how to resolve these scalar ambiguities by inserting known symbols in the transmitted sequence.

5. RESOLVING SCALAR AMBIGUITIES

To resolve the scalar ambiguities inherent to all blind channel estimators, known symbols are needed in the transmitted sequence. We here focus on pre-precoding pilots where the known symbols are inserted before precoding by $\mathbf{\Theta}$; alternatively, known symbols can be inserted after precoding and we will call them post-precoding pilots [9].

We pursue the scalar determination with identical precoders first. Notice that the estimated channels of (6) satisfy:

$$\mathcal{D}_{34} = \begin{bmatrix}\mathcal{D}_3 & \mathcal{D}_4\\ \mathcal{D}_4^* & -\mathcal{D}_3^*\end{bmatrix} = \begin{bmatrix}\mathcal{D}_1 & \mathcal{D}_2\\ \mathcal{D}_2^* & -\mathcal{D}_1^*\end{bmatrix}\begin{bmatrix}\alpha_1\mathbf{I}_J & -\alpha_2^*\mathbf{I}_J\\ \alpha_2\mathbf{I}_J & \alpha_1^*\mathbf{I}_J\end{bmatrix}, \tag{8}$$

and thus $\bar{\mathcal{D}}_{34} := \mathcal{D}_3^*\mathcal{D}_3 + \mathcal{D}_4^*\mathcal{D}_4$ equals:

$$\bar{\mathcal{D}}_{34} = (|\alpha_1|^2+|\alpha_2|^2)(\mathcal{D}_1^*\mathcal{D}_1+\mathcal{D}_2^*\mathcal{D}_2) = (|\alpha_1|^2+|\alpha_2|^2)\bar{\mathcal{D}}_{12}.$$

Multiplying $\check{\mathbf{y}}(n)$ by $\mathcal{D}_{34}^{\mathcal{H}}$ yields [c.f. (3)]:

$$\begin{aligned}\check{\mathbf{z}}(n) &= \begin{bmatrix}\mathbf{z}(2n)\\ \mathbf{z}(2n+1)\end{bmatrix} := \mathcal{D}_{34}^{\mathcal{H}}\check{\mathbf{y}}(n) \qquad (9)\\ &= \frac{1}{|\alpha_1|^2+|\alpha_2|^2}\begin{bmatrix}\alpha_1^*\mathbf{I}_J & \alpha_2^*\mathbf{I}_J\\ -\alpha_2\mathbf{I}_J & \alpha_1\mathbf{I}_J\end{bmatrix}\begin{bmatrix}\bar{\mathcal{D}}_{34}\mathbf{\Theta}\mathbf{s}(2n)\\ \bar{\mathcal{D}}_{34}\mathbf{\Theta}\mathbf{s}(2n+1)\end{bmatrix},\end{aligned}$$

where, for brevity, we omitted the noise.

Because the known symbols are inserted in the data stream before precoding, we need to equalize the channel and compensate for the precoding first, before resolving the residual scalar ambiguities. With identical precoders, a zero-forcing (ZF) equalizer can be applied to $\mathbf{z}(2n)$ and $\mathbf{z}(2n+1)$ in (9) by pre-multiplying with $(\bar{\mathcal{D}}_{34}\mathbf{\Theta})^\dagger$, where † stands for matrix pseudo-inverse. Based on (9), the equalizer outputs $\grave{\mathbf{s}}(2n) := (\bar{\mathcal{D}}_{34}\mathbf{\Theta})^\dagger\mathbf{z}(2n)$ and $\grave{\mathbf{s}}(2n+1) := (\bar{\mathcal{D}}_{34}\mathbf{\Theta})^\dagger\mathbf{z}(2n+1)$ can be written as:

$$\begin{bmatrix}\grave{\mathbf{s}}(2n)\\ \grave{\mathbf{s}}(2n+1)\end{bmatrix} = \frac{1}{|\alpha_1|^2+|\alpha_2|^2}\begin{bmatrix}\alpha_1^*\mathbf{I}_J & \alpha_2^*\mathbf{I}_J\\ -\alpha_2\mathbf{I}_J & \alpha_1\mathbf{I}_J\end{bmatrix}\begin{bmatrix}\mathbf{s}(2n)\\ \mathbf{s}(2n+1)\end{bmatrix}. \tag{10}$$

Suppose that two known symbols p_1 and p_2 are placed inside two consecutive blocks $\mathbf{s}(2m)$ and $\mathbf{s}(2m+1)$ at position k. Letting $\grave{s}_1 := [\grave{\mathbf{s}}(2m)]_k$ and $\grave{s}_2 := [\grave{\mathbf{s}}(2m+1)]_k$; we obtain:

$$\begin{bmatrix}\grave{s}_1\\ \grave{s}_2\end{bmatrix} = \frac{1}{|\alpha_1|^2+|\alpha_2|^2}\begin{bmatrix}\alpha_1^* & \alpha_2^*\\ -\alpha_2 & \alpha_1\end{bmatrix}\begin{bmatrix}p_1\\ p_2\end{bmatrix}, \tag{11}$$

from which α_1 and α_2 can be solved as:

$$\begin{bmatrix}\alpha_1\\ \alpha_2^*\end{bmatrix} = \frac{1}{|\grave{s}_1|^2+|\grave{s}_2|^2}\begin{bmatrix}\grave{s}_1^* & \grave{s}_2\\ -\grave{s}_2^* & \grave{s}_1\end{bmatrix}\begin{bmatrix}p_1\\ p_2^*\end{bmatrix}. \tag{12}$$

With α_1 and α_2 resolved, the true channels can then be found from (6). However, this step is not necessary since the symbol estimates can be obtained directly from (10) by:

$$\begin{bmatrix}\hat{\mathbf{s}}(2n)\\ \hat{\mathbf{s}}(2n+1)\end{bmatrix} = \begin{bmatrix}\alpha_1\mathbf{I}_J & -\alpha_2^*\mathbf{I}_J\\ \alpha_2\mathbf{I}_J & \alpha_1^*\mathbf{I}_J\end{bmatrix}\begin{bmatrix}\grave{\mathbf{s}}(2n)\\ \grave{\mathbf{s}}(2n+1)\end{bmatrix}. \tag{13}$$

With distinct precoders, we should equalize $\mathbf{z}(2n)$ by $(\bar{\mathcal{D}}_{34}\mathbf{\Theta}_1)^\dagger$ and $\mathbf{z}(2n+1)$ by $(\bar{\mathcal{D}}_{34}\mathbf{\Theta}_2)^\dagger$. Substituting $\alpha = \alpha_1$ and $\alpha_2 = 0$ in (12), the scalar α can be figured out as:

$$\alpha = (\grave{s}_1^* p_1 + \grave{s}_2 p_2^*)/(|\grave{s}_1|^2 + |\grave{s}_2|^2). \tag{14}$$

Similarly, if $|p_1| = |p_2|$ and thus $|\grave{s}_1| = |\grave{s}_2|$, eq. (14) can be further simplified to $\alpha = (1/2)\,(p_1/\grave{s}_1 + p_2^*/\grave{s}_2^*)$.

Remark (advantage of distinct over identical precoders): As indicated by Theorems 1 and 2, with distinct precoders $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ the channels can be identified up to one scalar α instead of two scalars (α_1, α_2) that must be determined with identical precoders $\mathbf{\Theta}_1 = \mathbf{\Theta}_2 = \mathbf{\Theta}$. With one pair of known symbols (p_1, p_2), the residual scalar ambiguities can be resolved by (12) and (14) for pre-precoding pilots. Therefore, the advantage of distinct precoders over identical precoders is not clearly justified since two scalars are also not difficult to resolve for identical precoders as in (12). However, the noise analysis we detail in [9] for the scalar ambiguity determination of (12) and (14) reveals that distinct precoders lead to a 3dB gain over identical precoders for suppressing the channel estimation error caused by the imperfectly resolved scalar ambiguities. To achieve the same channel estimation accuracy, identical precoders need to employ twice the number of pilots relative to distinct precoders.

In a nutshell, designing distinct precoders instead of identical precoders pays off either in terms of increasing the system efficiency by using half the number of pilots, or, in terms of improving the system performance with the same number of pilots, a feature that we also verified by simulations.

6. SIMULATIONS

To test the proposed channel estimation algorithm we use as figures of merit the averaged Normalized Mean Square Error (NMSE) of the channels defined as: $(1/2)\sum_{i=1}^{2}\|\hat{\tilde{\mathbf{h}}}_i - \tilde{\mathbf{h}}_i\|^2/\|\tilde{\mathbf{h}}_i\|^2$, and the Bit Error Rate (BER). We set the system parameters as: $L = 8$, $K = 3L$, $J = K + L = 32$; and generate the channels according to the Channel Model

A specified by ETSI. We assume here that each data burst has $N = 400$ symbol blocks, in which the first one $\check{\mathbf{s}}(0)$ is not precoded and serves as a training block. The semi-blind channel estimator is implemented using (5) with $W = 100$ and is initialized using the training based method of [2]. The channel estimates are updated every 50 blocks in order to render the complexity reasonable. To resolve the residual scalar ambiguities, $N_p = 1, 2, 4$ pairs of pre-precoding pilots are distributed inside each set of 50 symbol blocks.

To illustrate the advantage of distinct over identical precoders, we depict in Fig. 2 the NMSE averaged over the entire data burst with different number of pilots employed. From Fig. 2, we infer that indeed at high SNR identical precoders need to double the number of pilots to be able to catch up with the performance of distinct precoders, which is consistent with our noise analysis in [9]. To check the overall performance of channel estimation, equalization and ST decoding, we plot in Fig. 3 the BERs averaged over the entire data burst with ZF equalizers constructed from different channel estimates. Compared to the benchmark BER performance obtained with perfect channel knowledge at the receiver, our semi-blind channel estimator only incurs less than 2 dB SNR loss, while a high error floor is observed for the training based approach since the channels are time varying and no tracking mechanism has been invoked.

To illustrate the advantage of channel acquisition and coherent detection at the receiver, we also plot in Fig. 3 the BER performance of a competing differential ST-OFDM alternative, where the differential encoding of [7] is applied on each subcarrier to dispense with channel estimation. To make up for the same information rate, convolutional coding with rate 3/4 ($= K/J$) is also tested for differential ST-OFDM. Since the differential decoder output takes binary values [7], the Viterbi decoding algorithm with hard decision is applied here. Without assuming any side channel information, the path metric for Viterbi decoding is set to be the Hamming distance between the received bit stream at the output of differential decoder (denoted by $\hat{c}_1, \ldots, \hat{c}_n$, where $\hat{c}_i \in [0, 1]$) and the possible codeword candidates (denoted by $c_1, \ldots, c_n$). If side information on the channel fading coefficients $\{f_i^2\}_{i=1}^n$, where $f_i^2 := |H_1(\rho_i)|^2 + |H_2(\rho_i)|^2$, can be acquired at the receiver, the path metric could be modified using the weighted Hamming distance: $\sum_{i=1}^n f_i^2(\hat{c}_i - c_i)^2$. Fig. 3 demonstrates that precoded ST-OFDM equipped with our semi-blind channel estimator outperforms the differential ST-OFDM considerably, for both uncoded and coded transmissions at the considered SNR range.

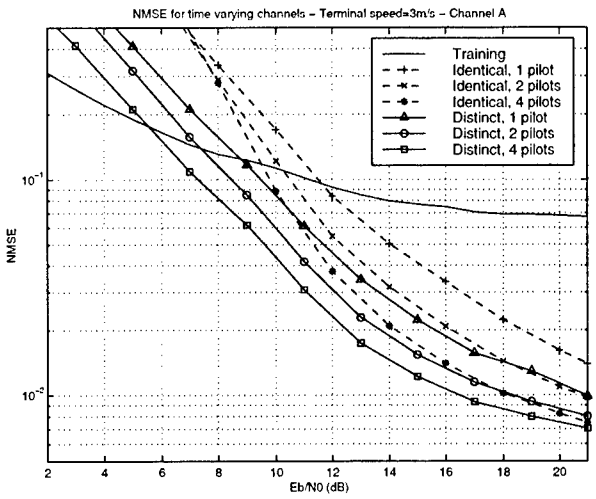


Fig. 2. Channel NMSE versus SNR

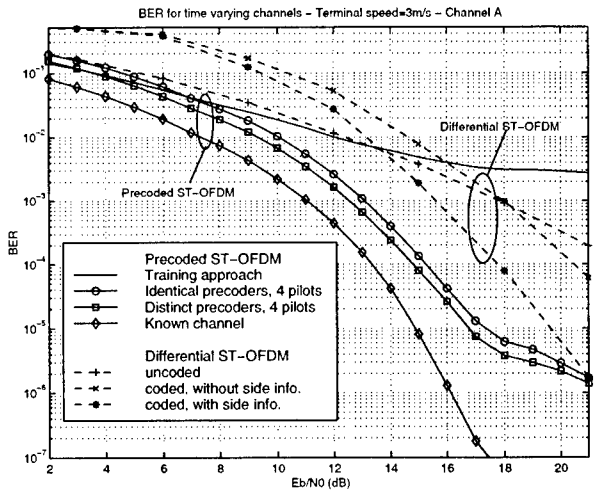


Fig. 3. BER versus SNR comparisons

7. REFERENCES

[1] H. Bölcskei, R. W. Heath Jr. and A. J. Paulraj, "Blind channel identification and equalization in OFDM-based multi-antenna systems," *IEEE Trans. Signal Processing*, 2001 (to appear).

[2] Y. Li, N. Seshadri and S. Ariyavisitakul, "Channel estimation for OFDM systems with transmitter diversity in mobile wireless channels," *IEEE JSAC*, pp. 461–471, March 1999.

[3] Z. Liu, G. B. Giannakis, S. Barbarossa and A. Scaglione, "Transmit-antennae space-time block coding for generalized OFDM in the presence of unknown multipath," *IEEE JSAC*, July 2001.

[4] Z. Liu, G. B. Giannakis, B. Muquet and S. Zhou, "Space-Time Coding for Broadband Wireless Communications," *Wireless Comm. and Mobile Computing*, vol. 1, no.1, pp. 33-53, 2001.

[5] B. Muquet, M. de Courville, P. Duhamel and V. Buzenac "A subspace based blind and semi-blind channel identification method for OFDM systems," in *Proc. of IEEE SPAWC*, 1999, pp. 170–173.

[6] A. F. Naguib, N. Seshadri and R. Calderbank, "Space-time coding and signal processing for high data rate wireless communications," *IEEE SP Magazine*, pp. 76–92, May 2000.

[7] V. Tarokh and H. Jafarkhani, "A differential detection scheme for transmit diversity", *IEEE JSAC*, pp. 1169–1174, July 2000.

[8] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications: Where Fourier meets Shannon," *IEEE SP Magazine*, pp. 29–48, May 2000.

[9] S. Zhou, B. Muquet and G. B. Giannakis, "Subspace-based (Semi-) Blind Estimation of Frequency-Selective Channels for Space-Time Block Precoded Transmissions," *IEEE Trans. on Signal Processing*, submitted February 2001; see also *Proc. of 34th Asilomar Conf.*, 2000, pp. 975-979.

COMPARING DS-CDMA AND MULTICARRIER CDMA WITH IMPERFECT CHANNEL ESTIMATION

Lucy L. Chong and Laurence B. Milstein

Department of Electrical and Computer Engineering
University of California, San Diego, La Jolla CA 92093
Email: lchong@ucsd.edu, milstein@ece.ucsd.edu

ABSTRACT

Probability of bit error expressions are derived for direct sequence CDMA (DS-CDMA) and multicarrier CDMA (MC-CDMA) with imperfect diversity combining. Pilot and data channels are transmitted through a Rayleigh fading channel with an exponential multipath intensity profile. Channel statistics are estimated using simple integrators. Then the multipath in the DS system and the multiple subcarriers in the MC system are weighted by the imperfect channel estimates and combined. Keeping the data rate, the transmit power, and the fading power constant, as the bandwidth increases, the number of multipaths increases in the DS system, and the number of subcarriers increases in the MC system. At the same time, the signal strength in each path/subcarrier decreases, and results in larger errors in the channel estimates. We show that there is a tradeoff between diversity order and SNR available for channel estimation in both DS-CDMA and MC-CDMA. Moreover, we also show that MC-CDMA performs better than DS-CDMA.

1. INTRODUCTION

Future personal communication systems are proposed to be wide-band to support high rate applications, such as video and data. High bandwidth brings the possibility of diversity gain. Among the various diversity combining techniques, it has been shown that maximal ratio combining (MRC) maximizes output signal-to-noise ratio (SNR). However, ideal MRC requires perfect knowledge of the channel fading statistics of each diversity branch. Of course, channel estimates will not be perfect when the received signal is corrupted by fading and noise. This error in the channel estimates will degrade performance. In a diversity system with a fixed total energy, as the number of diversity branches increases, the energy-per-branch decreases, and the weaker signals result in larger errors in the channel estimates. On one hand, the diversity gain improves with more diversity branches, on the other hand, greater degradation is caused by the errors in the channel estimates. Thus, there is a point when extra order of diversity actually degrades overall receiver performance, due to the decreasing SNR available for estimation.

In this paper, we examine the tradeoff between diversity order and channel estimation errors in direct sequence CDMA (DS-CDMA) [1–4] and multicarrier CDMA (MC-CDMA) [5–9]. In a wideband DS-CDMA system with an exponential multipath intensity profile (MIP), each resolved multipath has a different average power. Thus, there will be less accurate estimates for the weaker paths. In this same environment, if we trade path diversity for frequency diversity, we can design a MC-CDMA system with L subcarriers, where L is the number of resolvable paths in the direct sequence system. The processing gain in each subcarrier is L times smaller than the processing gain in the direct sequence system. The CDMA signal with lower processing gain is repeated L times, once in each of the L subcarriers. Each subcarrier now experiences flat Rayleigh fading, and the average fade power is the same for each of the subcarriers. We show that MC-CDMA performs better than DS-CDMA, with the extent of the improvement depending on the rate of decay of the exponential delay profile [10].

2. DIRECT SEQUENCE CDMA

2.1. Signal and Channel Model

The direct sequence BPSK signal transmitted by the kth user is

$$s_k(t) = \Re\left\{\sum_{n=-\infty}^{\infty} u_k[n]h(t-nT_c-\tau_k)e^{j(\omega_c t-\psi_k)}\right\},$$

where $u_k[n] = AC_{pk}[n]+BC_{dk}[n]d_k[n]$, $h(t)$ is the Nyquist chip-shaping filter, τ_k is the relative time delay of user k, ω_c is the carrier frequency, and ψ_k is the carrier phase. Also, $C_{pk}[n]$ and $C_{dk}[n]$ are binary orthogonal spreading sequences for the pilot channel and the data channel, respectively, A

Acknowledgment: This work was partially supported by the Army Research Office under Grant No. DAAG55-98-1-0473, by Airtouch, and by the MICRO program of the State of California.

0-7803-7011-2/01/$10.00 ©2001 IEEE

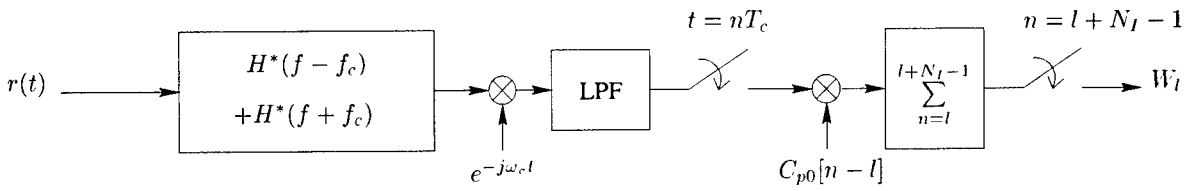


Fig. 1. The Direct Sequence Complex Channel Estimator Block Diagram.

and B are their corresponding transmit amplitudes and $d_k[n]$ is binary data with bit interval T. The spreading sequences chip rate is $1/T_c$, where $T_c = T/N$, and N is the processing gain. Assuming perfect average power control, the received signal is $r(t) = Re\{\tilde{r}(t)e^{j\omega_c t}\}$, where

$$\tilde{r}(t) = \sum_{k=0}^{K-1}\sum_{l=0}^{L-1}\sum_{n=-\infty}^{\infty} \alpha_{k,l}e^{j\theta_{k,l}}u_k[n]h(t-(n+l)T_c-\tau_k) + \tilde{n}_w(t),$$

K is the total number of users, L is the number of resolvable paths, $\alpha_{k,l}(t)$ and $\theta_{k,l}(t)$ are the amplitude and phase of the kth user's l^{th} path, respectively, and $\tilde{n}_w(t)$ is complex white Gaussian noise, with two-sided spectral density η_0. The fading amplitudes $\{\alpha_{k,l}\}$ are independent Rayleigh processes. Each user has an i.i.d. exponential multipath intensity profile with normalized decay factor δ, *i.e.*, $E[\alpha_{k,l}^2] = \Omega_0 e^{-\delta l/L}$. The phases $\{\theta_{k,l}\}$ are i.i.d. uniform random processes in $[0, 2\pi]$; the delays $\{\tau_k\}$ are independent uniform random variables in $[0, T_c]$. We assume the channel to be slowly varying such that the channel fading parameters can be assumed constant during the estimation interval, T_I. The delay spread is assumed to be sufficiently less than a bit interval T such that there is no significant intersymbol interference.

2.2. Receiver Model

The receiver estimates the fading parameters on each path. The estimator block diagram is shown in Figure 1. Assume chip timing, bit timing, and local code generator synchronization have been established. The channel estimate for the lth path of user 0 consists of a self-interference term, a multiple access interference term, and a noise term:

$$W_l = AN_I\alpha_{0,l}e^{j\theta_{0,l}} + S_l + M_{el} + N_{el}.$$

Their definitions can be found in [10].

Each estimate is updated at the end of the estimation interval T_I. Then the estimates are used to form the decision statistic $Z = \Re\{Y\}$, where $Y = \sum_{l=0}^{L-1} W_l^* Y_l$, and the complex conjugate is denoted by $*$. Demodulation is similar to the channel estimation operation, except despreading is done with the data spreading. The lth demodulator output is given by

$$Y_l = BN\alpha_{0,l}e^{j\theta_{0,l}}d_0[(v-1)N_I] + I_l + M_{dl} + N_{dl}.$$

Self interference, multiple access interference, and noise terms are defined in [10].

2.3. Probability of Bit Error

For well-chosen spreading sequences with large processing gain, we can approximate the set $\{W_l^* Y_l\}_{l=0,\ldots,L-1}$ as being independent. With the help of [3, Appendix 4, pp.345], the probability of bit error is given by

$$P_e \approx \frac{-1}{2\pi j}\int_{-\infty+j\epsilon}^{\infty+j\epsilon}\frac{1}{v}\prod_{l=0}^{L-1}\frac{v_{1l}v_{2l}}{(v+jv_{1l})(v-jv_{2l})}dv \quad (1)$$

$$\approx -\left(\prod_{i=0}^{L-1}v_{1i}v_{2i}\right)\sum_{l=0}^{L-1}\lim_{v\to jv_{2l}} \quad (2)$$

$$\left[\frac{v-jv_{2l}}{v}\prod_{d=0}^{L-1}\frac{1}{(v+jv_{1d})(v-jv_{2d})}\right], \quad (3)$$

assuming the decay factor δ of the exponential MIP is nonzero, and that the integrand in (1) has distinct roots. The variables v_{1l} and v_{2l} in (3) are defined in [3, equation 4B.6], where they involve the following second moments of W_l and Y_l [10],

$$\mu_{W_lW_l} \approx \frac{1}{2}N_I^2A^2\Omega_0e^{-\delta l/L} + \frac{\eta_0N_I}{2} + \frac{1}{2}N_I(K-1)(A^2+B^2)\Omega_0\frac{e^{-\delta}-1}{e^{-\delta/L}-1},$$

$$\mu_{Y_lY_l} \approx \frac{1}{2}N^2B^2\Omega_0e^{-\delta l/L} + \frac{\eta_0N}{2} + \frac{1}{2}N(K-1)(A^2+B^2)\Omega_0\frac{e^{-\delta}-1}{e^{-\delta/L}-1},$$

$$\mu_{W_lY_l} \approx \frac{1}{2}ABN_INE[\alpha_{0,l}^2] = \frac{1}{2}ABN_IN\Omega_0e^{-\delta l/L}.$$

In the special case that the estimation interval is one bit long (*i.e.*, $N_I = N$), and the pilot channel has the same power as the data channel (*i.e.*, $A = B$), the integrand in (1) has an Lth order repeated root at $v = jv_2$, where $v_2 = v_{2l}$,

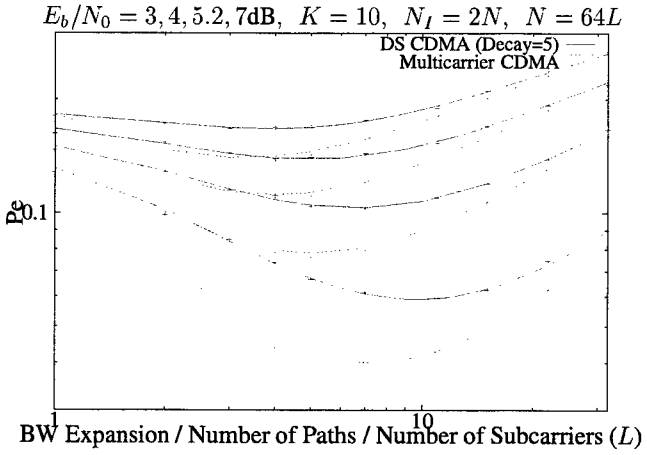


Fig. 2. Probability of error versus bandwidth for varying E_b/η_0.

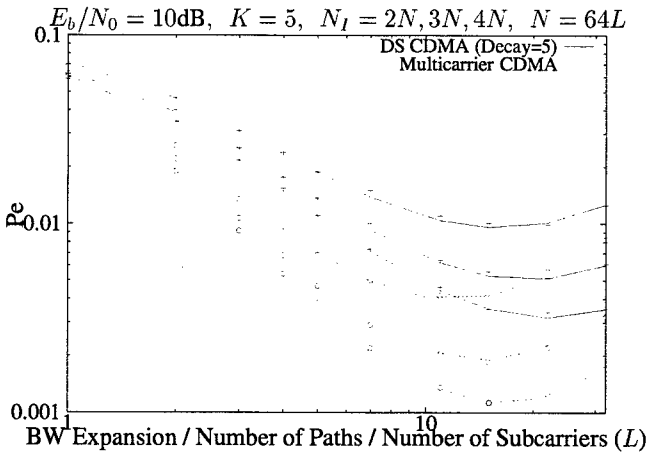


Fig. 3. Probability of error versus bandwidth for different estimation intervals.

$\forall l$. Then the probability of bit error is given by

$$P_e \approx -v_2^L \left(\prod_{i=0}^{L-1} v_{1i} \right) \lim_{v \to jv_2} \frac{1}{(L-1)!} \cdot \frac{\mathrm{d}^{L-1}}{\mathrm{d}v^{L-1}} \left[\frac{1}{v} \prod_{d=0}^{L-1} \frac{1}{(v + jv_{1d})} \right].$$

3. MULTICARRIER CDMA

In a multicarrier CDMA system, the same user signal as in the direct sequence section, with the same data rate but proportionally smaller processing gain and therefore with lower chip rate, is transmitted in L non-overlapping sub-bands. The bandwidth of the subcarriers is selected to be the coherence bandwidth of the channel such that each subcarrier experiences independent, slowly varying, flat fading. The signal model and receiver structure are similar to the previous direct sequence section and are described in [11]. The probability of bit error expression is also found in [11].

4. NUMERICAL RESULTS

We compare the performance of DS-CDMA and MC-CDMA with increasing bandwidth. We fix the data rate of both systems and, as the bandwidth increases, the processing gain (N) of the DS system increases, and the MC signal is repeated in more subcarriers. When L (number of paths/ number of subcarriers) equals unity, the two systems are identical, with the same processing gain (64 in our results). In the numerical results to follow, we use equal power in the pilot and the data channel, *i.e.*, $A = B$.

To make the comparison between different bandwidths, we keep the total transmit power constant and the total fading power constant, that is, decreasing the transmit power per subcarrier as the number of subcarriers increases, and renormalizing the power of each path as the number of paths increases. Traditionally, assuming perfect channel estimation, probability of error improves monotonically with the bandwidth [6]. However, when there is estimation error, the situation is different.

The probability of error is plotted against bandwidth in Figure 2 with 10 total users, the estimation interval N_I equaling 2 bits, and E_b/η_0 of 3, 4, 5.2, and 7 dB. The normalized decay factor is five for the DS-CDMA system. We have used a processing gain of 64 for each subcarrier of the MC system, and a processing gain of $64L$ for the DS system. As the bandwidth increases, the bit error rate first improves and then degrades. The increasing L helps performance by introducing diversity gain. At the same time, as L goes up, the SNR available for each estimate goes down; this causes more estimation error, and in turn, results in performance degradation. Thus, an optimal value of L exists. When we increase the E_b/η_0, the optimal L becomes larger, because the higher E_b/η_0 reduces the degradation due to the estimation error.

The MC system performs better than the DS system. All the subcarriers in the MC system have equal SNR. In the DS system, the multipaths follow an exponential profile. Some of the paths have higher SNR than the MC system, and others have lower SNR. However, the paths with lower SNR are hurting performance more than the paths with higher SNR are helping. Therefore, the overall performance of the DS system is worse than the MC system.

In Figure 3, we have plotted the probability of error for

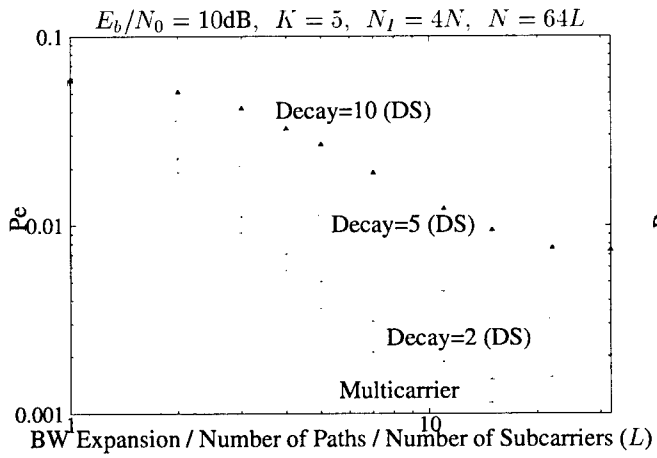


Fig. 4. Probability of error versus bandwidth for different normalized decay factors.

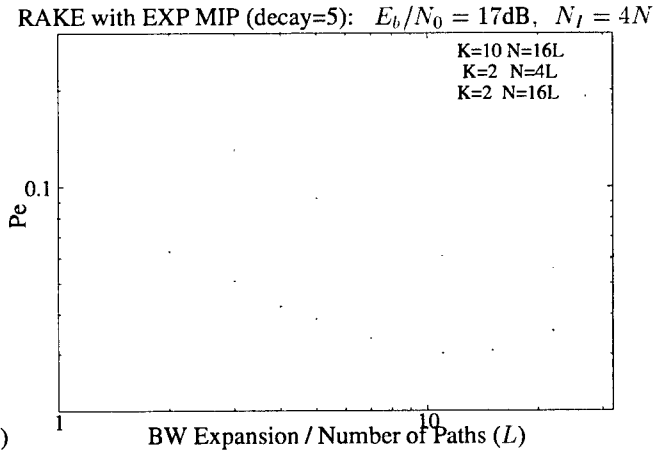


Fig. 5. Effects of approximations.

estimation intervals of 2, 3, and 4 bits. The increasing estimation intervals reduce the channel estimation error; this allows the optimal L to be higher.

In Figure 4, we have plotted the probability of error for different normalized decay factors of the DS system. For a small decay factor, the DS system performs similar to the MC system, but as the decay factor becomes larger, the DS system becomes much worst than the MC system.

We have approximated the performance of our DS-CDMA system by a system with uncorrelated estimates. This approximation is valid for large processing gain. We have also neglected self-interference in our DS-CDMA system, assuming a large number of users. We examine the effects of these approximations in Figure 5. For a moderate number of users ($K = 10$) and moderate processing gain ($N = 16L$), the simulation results match our theoretical expression very well. For a very small number of users ($K = 2$) and moderate processing gain ($N = 16L$), the simulation results still match quite well with the theoretical expression. When the number of users and the processing gain are both very small ($K = 2, N = 4L$), the effects of both approximations become apparent, and there are discrepancies between simulation and theoretical results.

5. REFERENCES

[1] G. L. Turin, "Introduction to spread-spectrum antimultipath techniques and their application to urban digital radio," *Proc. IEEE*, Mar. 1980, vol. 68, pp.328-53.

[2] J. S. Lehnert and M. B. Pursley, "Multipath diversity reception of spread-spectrum multiple-access communications," *IEEE Trans. Commun.*, Nov. 1987, vol. 35, pp.1189-98.

[3] J. G. Proakis, *Digital Communications*, McGraw-Hill, second ed., 1989.

[4] F. Adachi, "Rake combining effect on link capacity and peak transmit power of power-controlled reverse link of DS-CDMA cellular mobile radio," *IEICE Trans. Commun.*, Oct. 1997, vol. E80-B, no. 10, pp. 1547-55.

[5] S. Kondo and L. B. Milstein, "Multicarrier CDMA System with Cochannel Interference Cancellation," *Proc. VTC '94*, June 1994, pp. 1640-44.

[6] S. Kondo and L. B. Milstein, "Performance of Multicarrier DS CDMA Systems," *IEEE Trans. Commun.*, Feb. 1996, vol. 44, pp. 238-46.

[7] E. A. Sourour and M. Nakagawa, "Performance of orthogonal multicarrier CDMA in a multipath fading channel," *IEEE Trans. Commun.*, Mar. 1996, vol. 44, pp.356-67

[8] D. N. Rowitch and L. B. Milstein, "Coded Multicarrier DS-CDMA in the Presence of Partial Band Interference," *Proc. MILCOM '96*, Oct 1996, vol. 1, pp. 204-9.

[9] T. B. Welch and R. E. Ziemer, "DSSS/MCM System Performance in a Doppler Spread Channel," *Proc. MILCOM '97*, Nov. 1997, vol. 2, pp.587-91.

[10] L. L. Chong, *The effects of channel estimation errors on wideband CDMA systems*, Ph.D. thesis, University of California, San Diego, La Jolla, CA 92093 USA.

[11] L. L. Chong and L. B. Milstein, "Error rate of a multicarrier CDMA system with imperfect channel estimates," *Proc. ICC 2000*, June 2000, vol. 2, pp. 934-938.

ASYMPTOTIC PERFORMANCE ANALYSIS FOR REDUNDANT BLOCK PRECODED OFDM SYSTEMS WITH MMSE EQUALIZATION

Mérouane Debbah[1,2], *Walid Hachem*[3], *Philippe Loubaton*[2] *and Marc de Courville*[1]

[1]Motorola Labs – Paris, Espace Technologique Saint-Aubin 91193 Gif-sur-Yvette, France

[2]Laboratoire système de communication, Université de Marne la Vallée, France

[3]Service de radioélectricité, Supelec, France

ABSTRACT

Linear Precoding consists in multiplying by a $N \times K$ matrix a K-dimensional vector obtained by serial to parallel conversion of a symbol sequence to be transmitted. In this paper, the performance of MMSE receivers for certain large random isometric precoded OFDM systems on fading channels is analyzed. Using new tools, borrowed from *Free Probability Theory*, it can be shown that the Signal to Interference plus Noise Ratio at the equalizer output converges almost surely to a deterministic value depending on the probability distribution of the channel coefficients when $N \to +\infty$ and $K/N \to \alpha \leq 1$. These asymptotic results are used to answer the trade-off Convolutional Coding versus Linear Precoding issue while preserving a simple MMSE equalization scheme at the receiver.

1. INTRODUCTION.

A multi-carrier OFDM system [1] using a Cyclic Prefix for preventing inter-block interference is known to be equivalent to multiple flat fading parallel transmission channels in the frequency domain. In such a system, the information sent on some carriers might be subject to strong attenuations and could be unrecoverable at the receiver. This has motivated the proposal of more robust transmission schemes combining the advantages of CDMA with the strength of OFDM known as OFDM-CDMA [2], in which the information is precoded across all the carriers by a pre-coding matrix. This combination increases the overall frequency diversity of the modulator, so that unreliable carriers can still be recovered by taking advantage of the subbands enjoying a high Signal to Noise Ratio (SNR). Although originally proposed for a multiuser access scheme, this concept is extended to all single user OFDM systems and is referred in the sequel as Linear Precoded OFDM (LP OFDM)[3]. The LP OFDM equivalent frequency model scheme is depicted in figure 1, in which the input symbol stream is serial to parallel converted, then the resulting K-dimensional symbol vector $\mathbf{s}(n)$ (a white vector process with $E\left(\mathbf{s}(n)\mathbf{s}^H(n)\right) = \mathbf{I}_K$) is multiplied by an isometric $N \times K$ matrix $\mathbf{W}_N$ (i.e. $\mathbf{W}_N^H \mathbf{W}_N = \mathbf{I}_K$) where $N \geq K$. This N-dimensional vector $\mathbf{W}_N\mathbf{s}(n)$ is parallel to serial converted, and the corresponding generated data stream is sent across a frequency non selective Rayleigh fading channel. After serial to parallel conversion, the N–dimensional received vector $\mathbf{y}(n)$ can be written as:

$$\mathbf{y}(n) = \mathbf{H}_N(n)\,\mathbf{W}_N\,\mathbf{s}(n) + \mathbf{n}(n) \qquad (1)$$

where $\mathbf{n}(n)$ is a white additive Gaussian noise such that $E\left(\mathbf{n}(n)\mathbf{n}(n)^H\right) = \lambda \mathbf{I}_N$, and where $\mathbf{H}_N(n) = \mathrm{diag}([h_1(n), \ldots, h_N(n)])$ is the $N \times N$ diagonal complex matrix bearing on its diagonal the channel gains.

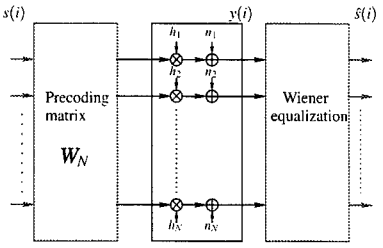


Fig. 1. System Model

Note that Giraud and Belfiore [4], and then Boutros and Viterbo [5] already introduced such a scheme called *signal space diversity*. An important problem lies in the choice of the amount of redundancy introduced by Linear Precoding, i.e. the ratio K/N, and also in the choice of matrix $\mathbf{W}_N$. [4] and [5] considered the case where $K/N = 1$, i.e. $\mathbf{W}_N$ is unitary. They assumed the entries of $\mathbf{H}_N(n)$ independent and identically distributed, and proposed to derive an upper bound of the error probability for the Maximum Likelihood (ML) detector of $\mathbf{s}(n)$. They discovered that, at least for high signal to noise ratios, $\mathbf{W}_N$ has to be chosen in such a way that the minimum L–distance product of the constellation $\{\mathbf{W}_N\mathbf{s}_i, i \in I\}$, where $\{\mathbf{s}_i, i \in I\}$ denotes the set of possible values taken by vector $\mathbf{s}(n)$, be maximum. More recently, Wang and Giannakis ([6], see also [3]) generalized these results to the case $K < N$ when the covariance matrix of random vector $\mathbf{h}(n) = (h_1(n), \ldots, h_N(n))^T$ is rank deficient.

The high computational cost of the ML detector prevents its use in practical contexts. Actually, due to its lower complexity, MMSE detection is often preferred. Therefore, in this paper, we study the impact of the choice of $\mathbf{W}_N$ and of parameter K/N on the asymptotic performance of the MMSE receiver when N and K converge toward $+\infty$ in such a way that $K/N \to \alpha \leq 1$. Several papers [7][8] have recently analyzed the behaviour of the SINR at the output of the MMSE detector when the entries of $\mathbf{W}_N$ are independent and identically distributed random variables (to be referred to in the sequel as the i.i.d. case) and in the case where $\mathbf{H}_N(n)$ is reduced to $\mathbf{I}_N$. The originality of our contribution lies in the fact that the linear random precoder $\mathbf{W}_N$ is isometric instead of being i.i.d. Such a choice is justified by the fact that isometric precoders provide much better results than i.i.d. ones as will be seen below.

From a technical stand point, the i.i.d. case study of [8] is based on mathematical results that concern the limiting distribution of

0-7803-7011-2/01/$10.00 ©2001 IEEE

eigenvalues of some large random matrices with independent and identically distributed entries (see e.g. [9]). The results given here rely on the so-called *Free Probability Theory* initially developed by D. Voiculescu. Due to the lack of space, the corresponding tools and derivations are not introduced here. The interested reader may consult [10] (this paper can be downloaded at http://www-syscom.univ-mlv.fr/~loubaton/index.html). Note that Evans and Tse already introduced free probability theory in [7], but for solving quite different problems.

2. ASYMPTOTICAL SINR PERFORMANCE

In this section, the SINR of Linear Precoded OFDM systems designed with random matrices is derived. Since the time index n is irrelevant in the following, we simply discard it from now on. We assume that $\mathbf{H}_N = \mathrm{diag}([h_1, \ldots, h_N])$ has identically distributed centered random diagonal entries. $|h_i|^2$ is supposed to have a probability density $p(t)$ with finite moments of all orders. We set $E(|h_i|^2) = 1$, so that λ represents the inverse of the SNR at the receiver input. Notice that random variables $\{h_i\}_{i=1,N}$ are not assumed to be independent. However, we assume that for each $l \geq 1$,

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{k=1}^{N} |h_k|^{2l} = E(|h_i|^{2l}) \text{ almost surely.} \qquad (2)$$

which implies some kind of asymptotic independence between the random variables h_i and h_j if $|i-j| \to \infty$. This hypothesis is quite realistic in the context of LP OFDM schemes if large time interleavers/deinterleavers are inserted in the scheme represented in figure 1. We stress on the fact that the ergodicity relation expressed by (2) influences deeply our results.

We now explain how the random matrix $\mathbf{W}_N$ is generated. A random unitary matrix is said *Haar distributed* if its probability distribution is invariant by left multiplication by constant unitary matrices (this invariance condition specifies the distribution). Such a matrix can be generated by the following way: let $\mathbf{X} = [x_{i,j}]_{1\leq i,j\leq N}$ be a $N \times N$ random matrix with independent complex Gaussian centered unit variance entries. Then the unitary matrix $\mathbf{X}(\mathbf{X}^H\mathbf{X})^{-1/2}$ is Haar distributed (see [10]). $\mathbf{W}_N$ is generated by extracting any K columns from a $N \times N$ Haar distributed unitary matrix independent of $\mathbf{H}_N$.

Before going further, let recall the expression of the SINR at one of the K outputs of the MMSE detector. The SINR $\beta_{\mathbf{w}_N}$ is easily shown to express as $\beta_{\mathbf{w}_N} = \frac{\eta_{\mathbf{w}_N}}{1-\eta_{\mathbf{w}_N}}$ where:

$$\eta_{\mathbf{w}_N} = \mathbf{w}_N^H \mathbf{H}_N^H \left(\mathbf{H}_N \mathbf{W}_N \mathbf{W}_N^H \mathbf{H}_N^H + \lambda \mathbf{I}_N\right)^{-1} \mathbf{H}_N \mathbf{w}_N \, .$$

We are now in position to state the main results of this contribution:

Theorem 1 (Isometric Case) *Assume that matrices $\mathbf{W}_N$ and $\mathbf{H}_N$ are chosen as above and moreover, that the probability density $p(t)$ of the random variables $(|h_i|^2)_{i\in\mathbb{N}}$ has a compact support included in the interval $[0,c]$ (which implies that $\sup_{i\in\mathbb{N}} |h_i|^2 \leq c < \infty$ almost surely).*
When N grows towards infinity and $K/N \to \alpha \leq 1$, the SINR $\beta_{\mathbf{w}_N}$ at the output of a MMSE equalizer converges almost surely to a value $\bar{\beta}$ that is the unique solution of the equation

$$\int_0^\infty \frac{t}{\alpha t + \lambda(1-\alpha)\bar{\beta} + \lambda} \, p(t)\, dt = \frac{\bar{\beta}}{\bar{\beta}+1} \, . \qquad (3)$$

The first important conclusion provided by this result is that the SINR $\beta_{\mathbf{w}_N}$ converges to a deterministic value depending only on the channel coefficients distribution, but not on the particular channel realization. Relation (2) plays a fundamental role in this respect. In some sense, the precoded system equipped with a MMSE receiver allows to transform a flat fading Rayleigh channel into a Gaussian channel with signal to noise ratio $\bar{\beta}$. This confirm the results of [5] stated in the case of the maximum likelihood detector, and the observations made in [11] and [12] in the context of MC-CDMA systems.
The second conclusion is that the SINR limit does not depend on the particular realization of $\mathbf{W}_N$ as well. This suggests that for large blocks, it is irrelevant to optimize the performance of a MMSE receiver with respect to $\mathbf{W}_N$. In the statement of theorem 1, $p(t)$ is assumed to be compactly supported. This technical hypothesis is needed because the most powerful results of free probability theory applied to random matrices require compactly supported measures. Although the usual channel probability distributions like the Rayleigh or the Rice distributions are not compactly supported, in practice, formula (3) predicts quite well the performance of our precoded modulation scheme using MMSE detection.

Since the proof of (3) is non trivial and needs the use of sophisticated arguments, only an outline is provided. We have shown in [10] that $\eta_{\mathbf{w}_N}$ converges almost surely to a value $\bar{\eta}$ which does not depend on the choice of $\mathbf{w}_N$. For a given N, there are K quantities $\eta_{\mathbf{w}_N}$ each corresponding to the choice of a particular column code in $\mathbf{W}_N$. Their sum over all the columns of this matrix is $\mathrm{trace}\left((\mathbf{H}_N\mathbf{W}_N\mathbf{W}_N^H\mathbf{H}_N^H + \lambda\mathbf{I}_N)^{-1}\mathbf{H}_N\mathbf{W}_N\mathbf{W}_N^H\mathbf{H}_N^H\right)$. Hence, the limit value $\bar{\eta}$ is given by :

$$\bar{\eta} = \lim_{N\to\infty} \frac{1}{\alpha N} \mathrm{trace}\left((\mathbf{H}_N\mathbf{W}_N\mathbf{W}_N^H\mathbf{H}_N^H + \lambda\mathbf{I}_N)^{-1}\mathbf{H}_N\mathbf{W}_N\mathbf{W}_N^H\mathbf{H}_N^H\right)$$

By applying free probability results, we can show that the empirical eigenvalue distribution of $\mathbf{H}_N\mathbf{W}_N\mathbf{W}_N^H\mathbf{H}_N^H$ converges almost surely to a compactly supported measure θ, which can be derived explicitly. Therefore, $\bar{\eta}$ converges almost surely to: $\frac{1}{\alpha}\int \frac{t}{t+\lambda}\, d\theta(t)$. This also shows that $\beta_{\mathbf{w}_N}$ converges to a deterministic value $\frac{\bar{\eta}}{1-\bar{\eta}}$ denoted $\bar{\beta}$ and solution of (3).

For comparison sake, it is useful to give the expression of the asymptotic SINR when $\mathbf{W}_N$ is i.i.d. :

Theorem 2 (i.i.d. Case) *Assume that the entries of $\mathbf{W}_N$ are centered i.i.d. random variables with variance $1/N$, that the elements $\{h_i\}$ are identically distributed random variables such that $|h_i|^2$ has a probability density $p(t)$ with compact support, and that for each bounded continuous function φ,*

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{n=1}^{N} \varphi\left(\frac{1}{|h_i|^2}\right) = E\left(\varphi\left(\frac{1}{|h|^2}\right)\right) = \int \varphi\left(\frac{1}{t}\right) p(t)\, dt$$

almost surely.
When N grows towards infinity and $K/N \to \alpha \leq 1$, the SINR $\beta_{\mathbf{w}_N}$ at the output of a MMSE equalizer converges almost surely to a value $\bar{\beta}_1$ that is the unique solution of the equation

$$\int_0^\infty \frac{t}{\alpha t + \lambda\bar{\beta}_1 + \lambda} \, p(t)\, dt = \frac{\bar{\beta}_1}{\bar{\beta}_1 + 1} \, . \qquad (4)$$

This result is a direct consequence of corollary 1 in [7] and the main result of [9]. The two main conclusions stated in the isometric case remain valid in the i.i.d. case. However, we observe that

for a fixed value of α, $\bar{\beta}_1 \leq \bar{\beta}$ because for each $\beta > 0$,

$$\int_0^\infty \frac{t}{\alpha t+\lambda\beta+\lambda}\, p(t)\, dt \leq \int_0^\infty \frac{t}{\alpha t+\lambda(1-\alpha)\beta+\lambda}\, p(t)\, dt \ .$$

Moreover, the asymptotic performance of the MMSE receiver in the isometric case is all the more better with respect to the i.i.d. case that α is close to 1. Conversely, $\bar{\beta}_1 \simeq \bar{\beta}$ if α is close to 0.

3. INFLUENCE OF CHANNEL AND PRECODER

In this section, we provide better insights of the effects of the fading channel and of the nature of the precoder (i.i.d. versus isometric) on the SINR performance.

Effect of the fading on the performance in the i.i.d. case. We first study the effect of the fading in the i.i.d. case, and compare (4) with the results of [8] obtained in the Gaussian channel case. For that purpose, rewrite (4) as

$$\bar{\beta}_1 = \mathbf{E}_{|h|^2}\left(\frac{1}{\frac{\alpha}{\bar{\beta}_1+1}+\frac{\lambda}{|h|^2}}\right) \tag{5}$$

where $\mathbf{E}_{|h|^2}$ denotes the mathematical expectation with respect to $|h|^2$ distribution. If N is large enough, the righthandside of (5) can be approximated by the empirical mean $\frac{1}{N}\sum_{k=0}^{N-1}\frac{1}{\frac{\alpha}{\bar{\beta}_1+1}+\frac{\lambda}{|h_k|^2}}$ In the Gaussian channel case, the distribution of $|h|^2$ is reduced to $\delta(1)$, and equation (5) coincides with the result given in [8]. Recall that Tse and Hanly interpreted in [8] the factor $\frac{1}{\bar{\beta}_1+1}$ as the effective interference of user i on the desired user k at the desired target SINR $\bar{\beta}_1$. The term $\frac{\alpha}{\bar{\beta}_1+1} = \frac{1}{N}\frac{K}{\bar{\beta}_1+1}$ thus represents the total amount of multiuser interference at the output of the MMSE receiver (the term $\frac{1}{\bar{\beta}_1+1}$ is multiplied by K the number of users, while the coefficient $\frac{1}{N}$ is due to the spreading gain provided by the precoder). It is interesting to note that in the fading channel case, the righthandside of (5) coincides with an averaged version (on the amplitude of the channel coefficients) of the inverse of the sum of the *same interference term* and of the term $\frac{\lambda}{|h|^2}\cdot\frac{\lambda}{|h|^2}$ represents the contribution of a thermal noise of variance $\frac{\lambda}{|h|^2}$ in a Gaussian channel. The diversity provided by the precoder is of course due to the averaging on the values taken by $|h|^2$ in (5).

Comparison between i.i.d. and isometric precoders. In order to compare the two kind of precoders, rewrite (3) as

$$\bar{\beta} = \mathbf{E}_{|h|^2}\left(\frac{1}{\frac{\alpha}{\bar{\beta}+1}+\frac{\lambda}{|h|^2}(1-\alpha\frac{\bar{\beta}}{(1+\bar{\beta})})}\right) \tag{6}$$

It is interesting to note that the second term of the righthandside of (5) and (6) are similar. The multiuser interference term $\frac{\alpha}{\bar{\beta}+1}$ appears in both formulas, while the term $\frac{\lambda}{|h|^2}$, representing the effect of the thermal noise in the i.i.d. case, is multiplied in the isometric case by $1-\alpha\frac{\bar{\beta}}{1+\bar{\beta}}$, which is of course smaller than 1. In other words, for a given target SINR of $\bar{\beta}$, an isometric precoded system corrupted by a thermal noise of variance λ provides the same performance than an i.i.d. precoded one corrupted by a thermal noise of variance $(1-\alpha\frac{\bar{\beta}}{1+\bar{\beta}})\lambda$. We note in particular that the attenuation factor is all the more favorable that α is close to 1.

4. SYSTEM DESIGN IMPLICATIONS

Simulations have been performed assuming a QPSK constellation, independent Rayleigh channel attenuations and perfect channel knowledge at the receiver. Figure 2 shows the BER in the isometric case for various spectral efficiencies $\alpha = 1, \frac{1}{2}$, and $\frac{1}{4}$. The curves closely match the simulation results using a realistic number of subchannels ($N = 256$). The "Gaussian Channel" curve is provided as a reference and corresponds to $\mathbf{H}_N = \mathbf{I}_N$. In this situation, the receiver output SINR is easily shown to be $1/\lambda$.

In order to determine the optimal amount of redundancy that should be spent on Linear Precoding, the throughput of an LP OFDM system equipped with a MMSE receiver is analyzed. The throughput $\gamma(\alpha,\lambda)$ is the total number of bit/s/Hz that can be reliably transmitted with this system. It is defined by $\gamma(\alpha,\lambda) = \alpha C(\alpha,\lambda)$ where the capacity $C(\alpha,\lambda)$ is given by $C(\alpha,\lambda) = \log_2(1+\mathrm{SINR}(\alpha,\lambda))$ and $\mathrm{SINR}(\alpha,\lambda)$ is $\bar{\beta}$ or $\bar{\beta}_1$. Note that $Eb/N0 = (C\lambda)^{-1}$ (see [13] for more details). Figure 3 shows the behaviour of the optimum value of α (i.e. for which the throughput is maximum) with respect to Eb/N_0 for both isometric and i.i.d. cases. For maximizing the throughput, nearly no redundancy should be spent on the Linear Precoder in the isometric case. In contrast, in the i.i.d case, a significant amount of redundancy is required when $\frac{Eb}{N_0} > 4$dB. Figure 4 shows the maximum throughput vs Eb/N_0 for isometric and i.i.d. linear precoders. The throughput for a Gaussian channel is also provided. Isometric precoding increases the throughput with respect to i.i.d. precoding.

This throughput analysis is now used to study the performance of a system where Linear Precoding of rate α is combined with a classical Convolutional Coding of rate R. Assuming an overall coding rate αR of $1/2$, the purpose is to determine the optimum balance between α and R. Figure 3 suggests that when isometric precoding is used, $\alpha_{opt} \approx 1$ and with i.i.d. precoding, $\alpha_{opt} \approx 2/3$ for the most common values of Eb/N_0. The optimum values of R are thus close to $1/2$ and to $3/4$ respectively. These claims are sustained by figures 5 and 6.

5. CONCLUSION

This contribution extends the pioneering work of [8] devoted to asymptotic performance analysis of DS-CDMA systems employing i.i.d. signatures. Here, the theoretical asymptotic SINR at the output of a MMSE receiver with isometric Linear Precoding is derived using new mathematical tools, borrowed from the so-called Free Probability theory. It is shown that in a system where isometric Linear Precoding is combined with Convolutional Coding, nearly no redundancy should be spent on LP. However, for Linear i.i.d Precoders, redundancy is required at the emitter side. Finally, in all the cases, isometric Linear Precoders always outperform i.i.d ones. We finally remark that these results do not contradict those of [6] devoted to the study of maximum likelihood receivers of precoded OFDM systems. [6] found useful the use of redundant precoders ($\alpha < 1$) in the isometric case because [6] did not assumed the presence of an interleaver/deinterleaver structure. The ergodicity assumptions (2) are therefore not valid in [6], and the signal to interference plus noise ratio converges toward a value depending on the particular realizations of random variables $(h_k)_{k\geq 0}$.

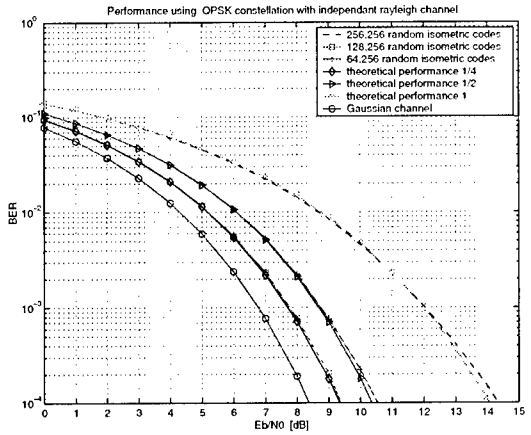


Fig. 2. Probability of error

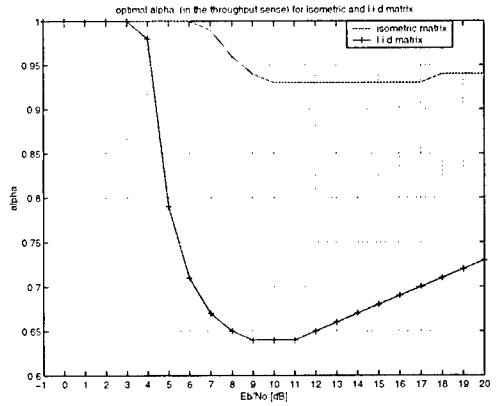


Fig. 3. Optimum α

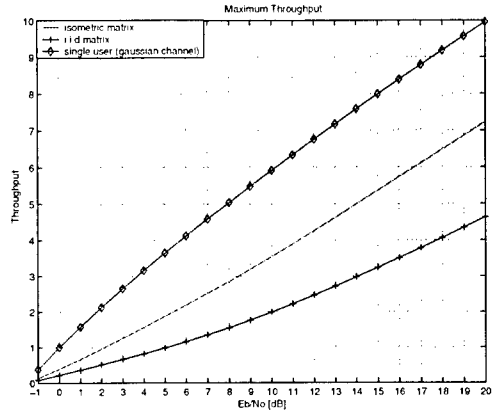


Fig. 4. Optimum Throughput

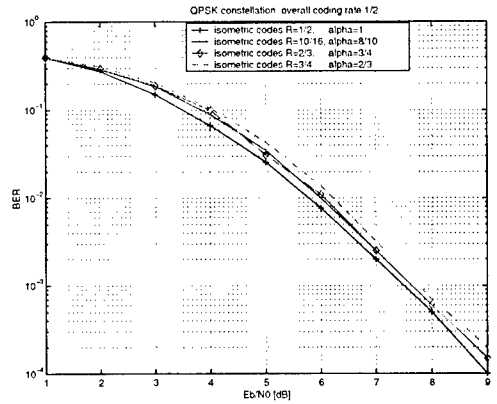


Fig. 5. Isometric precoding matrix

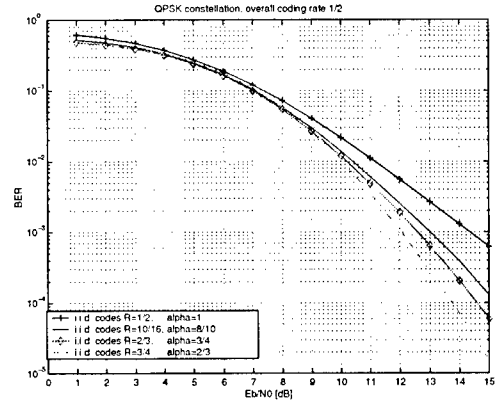


Fig. 6. i.i.d precoding matrix

6. REFERENCES

[1] J.A.C. Bingham, "Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come," *IEEE Comm. Mag.*, vol. 28, no. 5, pp. 5–14, May 1990.

[2] K. Fazel, "Performance of CDMA/OFDM for mobile communication system," in *Proc. ICUPC*, 1993, vol. 2, pp. 975–979.

[3] Z. Wang and G.B. Giannkis, "Linearly Precoded OFDM for Fading Wireless Channels," *IEEE Trans. on IT*, to appear.

[4] X. Giraud and J.C Belfiore, "Constellations Matched to the Rayleigh Fading Channel ," *IEEE Trans. on IT*, pp. 106–115, Jan. 1996.

[5] J. Boutros and E. Viterbo, "Signal space diversity: a power and bandwidth efficient diversity technique for the Rayleigh fading channel," *IEEE Trans. on IT*, pp. 1453–1467, July 1998.

[6] Z. Wang and G.B. Giannakis, "Linearly Precoded or Coded OFDM against Wireless Channel Fades ?," in *Proceedings of the 3rd Workshop on Signal Processing Advances in Wireless Communications*, Mar. 2001, pp. 267–270.

[7] J. Evans and D.N.C Tse, "Large System Performance of Linear Multiuser Receivers in Multipath Fading Channels," *IEEE Trans. on IT*, pp. 2059–2078, Sept. 2000.

[8] D.N.C Tse and S. Hanly, "Linear Multi-user Receiver: Effective Interference, Effective bandwidth and User Capacity," *IEEE Trans. on IT*, pp. 641–657, Mar. 1999.

[9] J.W. Silverstein and Z.D. Bai, "On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices," *J. Multivariate Anal.*, vol. 54, no. 2, pp. 175–192, 1995.

[10] M. Debbah, W. Hachem, Ph. Loubaton, and M. de Courville, "MMSE Analysis of Certain Large Isometric Random Precoded Systems," *IEEE Trans. on IT*, submitted may 2001.

[11] S. Kaiser, "Trade-Off between Channel Coding and Spreading in Multi-Carrier CDMA," in *IEEE Spread Spectrum Techniques And Applications Proceedings*, 4th International Symposium, 1996, vol. 3, pp. 1366–1370.

[12] Jürgen Lindner, "MC-CDMA and its Relation to General Multiuser/Multisubchannel Transmission Systems," in *International Symposium on Spread Spectrum Techniques & Applications*, Mainz, Germany, Sept. 1996, pp. 115–121.

[13] E. Biglieri, G. Caire, G. Taricco, and E. Viterbo, "How fading affects CDMA: an asymptotic analysis," *JSAC Wireless Series*, 2001, to appear.

AN ALGORITHM FOR JOINT SYMBOL TIMING AND CHANNEL ESTIMATION FOR OFDM SYSTEMS

Erik G. Larsson†, *Guoqing Liu*†, *Jian Li*† *and Georgios B. Giannakis*⋆

†Dept. of Electrical and Computer Engineering, University of Florida
P.O. Box 116130, Gainesville, FL 32611. Email: larsson@dsp.ufl.edu
⋆Dept. of Electrical and Computer Engineering, University of Minnesota
200 Union Street SE, Minneapolis, MN 55455.

ABSTRACT

In this paper, we consider the problem of joint synchronization and channel estimation for orthogonal frequency division multiplexing (OFDM) systems. A new algorithm is proposed that estimates the channel and symbol timing simultaneously by using a technique based on maximum-likelihood (ML) theory and the generalized Akaike information criterion (GAIC). Finally, we demonstrate the performance of our algorithm by simulation results.

1. OFDM

The OFDM access technology is based on the transmission of data packets, each of which consists of a number of consecutive *OFDM symbols*. Each OFDM symbol $\boldsymbol{x}$ has a length of N samples and carries a certain number of information bits or training data (that is, known data that are used to assist the demodulator). An OFDM symbol is created by taking the discrete Fourier transform (DFT) of N data symbols (taken from a finite constellation $\mathcal{A}$, such as BPSK, QPSK or QAM). Furthermore, each OFDM symbol is preceded by a *cyclic prefix* (CP) (also called *guard interval* (GI)) of length M that is an exact replica of the M last samples of the OFDM symbol. The reason for this (as will become apparent below) is that demodulation in the presence of frequency-selective fading can be carried out very easily. Before proceeding, let us remark on the fact that in the case that two (or more) *identical* OFDM symbols are transmitted directly subsequent to each other, the tail of the first symbol can serve as the CP for the second.

In the WLAN standard recently adopted by the IEEE 802.11 standardization group [1], each data packet consists of a *preamble* and a data carrying part. The preamble consists of 10 "short" identical known OFDM symbols of length $N_s = 16$ concatenated with 2 "long" identical and known OFDM symbols of length $N_l = 64$ which are all utilized for carrier frequency offset (CFO) correction, channel estimation and synchronization. The data carrying part consists of a variable number of OFDM symbols of length $N_d = 64$, where each OFDM symbol contains useful information plus some known *pilot* bits, which are typically used for updating the phase of the channel estimates. An OFDM packet for the IEEE 802.11 standard is depicted in Figure 1. Note that t_1 serves as a CP for t_2, t_2 is the CP for t_3, and so on. For the long symbols in the preamble, GI2 is the CP for T_1 and it contains the 32 last samples of T_1.

Let $\boldsymbol{y} = [y_1 \ \cdots \ y_N]^T$ be a vector of N data symbols taken from $\mathcal{A}$ (the elements of $\boldsymbol{y}$ are sometimes referred to as *sub-carriers*) and let $\boldsymbol{W}$ be a DFT matrix of size $N \times N$, that is, element k,l of $\boldsymbol{W}$ is equal to $\boldsymbol{W}_{k,l} = e^{-j2\pi\frac{(k-1)(l-1)}{N}}$. Then the OFDM symbol $\boldsymbol{x}$ corresponding to the data $\boldsymbol{y}$ is computed by taking the inverse DFT of $\boldsymbol{y}$, viz. $\boldsymbol{x} = \boldsymbol{W}^*\boldsymbol{y}$, where $(\cdot)^*$ denotes the conjugate transpose. The CP $\tilde{\boldsymbol{x}}$ corresponding to $\boldsymbol{x}$ contains the $M = 16$ last samples of $\boldsymbol{x}$, a relation that can be expressed as $\tilde{\boldsymbol{x}} = \boldsymbol{T}_M\boldsymbol{W}^*\boldsymbol{y}$ where $\boldsymbol{T}_M$ consists of the last M rows of the $N \times N$ identity matrix.

Assume that the effect of the propagation channel can be described by a finite impulse response (FIR) filter with an effective length $L \leq M + 1$ and impulse response $\{h_0, \ldots, h_{L-1}\}$. For reasons that will be apparent later, we augment the channel impulse response with $M - L$ zeros and define

$$\boldsymbol{h} = [h_0 \ \cdots \ h_{M-1}]^T = [h_0 \ \cdots \ h_{L-1} \ 0 \ \cdots \ 0]^T$$

To illustrate the demodulation procedure, we write the

THIS WORK WAS PARTLY SUPPORTED BY THE SWEDISH FOUNDATION FOR STRATEGIC RESEARCH (SSF) AND THE NATIONAL SCIENCE FOUNDATION GRANT MIP-9457388.

0-7803-7011-2/01/$10.00 ©2001 IEEE

received signal as (neglecting receiver noise)

$$\boldsymbol{r} = \begin{bmatrix} r_0 \\ \vdots \\ r_{N-1} \end{bmatrix} = \begin{bmatrix} h_{M-1} & \cdots & h_0 & & & \\ & h_{M-1} & \cdots & h_0 & & \\ & & \ddots & \cdots & \ddots & \\ & & & h_{M-1} & \cdots & h_0 \end{bmatrix} \begin{bmatrix} \boldsymbol{T}_{M-1}\boldsymbol{W}^*\boldsymbol{y} \\ \boldsymbol{W}^*\boldsymbol{y} \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} h_0 & & & h_{M-1} & & h_1 \\ h_1 & h_0 & & & h_{M-1} & h_2 \\ \vdots & \ddots & \ddots & & & \vdots \\ & \ddots & \ddots & & & h_{M-1} \\ h_{M-1} & & \ddots & \ddots & & \\ & \ddots & & \ddots & \ddots & \\ & & h_{M-1} & \cdots & h_1 & h_0 \end{bmatrix}}_{\boldsymbol{H}} \boldsymbol{W}^*\boldsymbol{y}$$

The matrix $\boldsymbol{H}$ is of dimension $N \times N$ and *circulant*, so the DFT of the received data vector $\boldsymbol{r}$ can be written [9]

$$\boldsymbol{W}\boldsymbol{r} = \boldsymbol{W}\boldsymbol{H}\boldsymbol{W}^*\boldsymbol{y} + \text{noise} = \boldsymbol{\Delta}\boldsymbol{y} + \text{noise} \qquad (1)$$

where $\boldsymbol{\Delta} = \text{diag}\{\delta_1, \ldots, \delta_N\}$ is a *diagonal* matrix containing the DFT of the channel impulse response $\boldsymbol{h}$, that is, $[\delta_1 \ \cdots \ \delta_N]^T = N\boldsymbol{W}\boldsymbol{h}$. We remark on the fact that another way to see this is that the cyclic prefix gives the effect of the propagation channel an interpretation in terms of circular convolution; however we prefer to remain in the matrix algebra framework. From (1) we see that, provided the channel $\boldsymbol{\Delta}$ is known, each data bit y_n in the OFDM symbol under consideration can be estimated as $\hat{y}_n = P_{\mathcal{A}}\left(\frac{[\boldsymbol{W}\boldsymbol{r}]_n}{\delta_n}\right)$ where $P_{\mathcal{A}}(\cdot)$ denotes projection onto the alphabet $\mathcal{A}$, and $[\cdot]_n$ denotes the nth element of a vector. This common and simple demodulator can be implemented by one single FFT.

2. ML CHANNEL ESTIMATION

Channel estimation for OFDM is discussed in some detail in [2, 3, 6], so we merely summarize some results using our notation and framework. Assume that the received data $\boldsymbol{r}$ has been adjusted to compensate for a possible CFO [4], that a proper timing T is obtained and that the effective channel length L is known. Consider first the estimation of the channel $\boldsymbol{h}(L)$ (we use the index L to emphasize that the last $M-L$ elements of $\boldsymbol{h}$ are zero) based on a least-squares (LS) criterion using received data corresponding to the first (known) long OFDM symbol in the preamble. Denote the $N \times 1$ vector of the known data symbols in the long OFDM preamble symbol with $\boldsymbol{p}$. Then LS channel estimation (conditioned on the timing T) amounts to (cf. (1))

$$\min \|\boldsymbol{W}\boldsymbol{r} - \boldsymbol{\Delta}(L)\boldsymbol{p}\|^2$$

subject to the constraint that the effective channel length is L, i.e., $\boldsymbol{T}_{M-L}\boldsymbol{h} = \boldsymbol{0}$. This is equivalent to

$$\min \|\boldsymbol{W}\boldsymbol{r} - \text{diag}\{\boldsymbol{p}\}\boldsymbol{W}\boldsymbol{h}(L)\|^2 \qquad (2)$$

subject to $\boldsymbol{T}_{M-L}\boldsymbol{h} = \boldsymbol{0}$. For the symbols in the IEEE 802.11 WLAN standard, 12 of the elements of $\boldsymbol{p}$ are equal to zero, and the rest belong to the (unitary) BPSK constellation. Using this fact it is not difficult to see that (2) has the solution (see, e.g., [7])

$$[h_1 \ \cdots \ h_L]^T = (\tilde{\boldsymbol{W}}^*\tilde{\boldsymbol{W}})^{-1}\tilde{\boldsymbol{W}}^*\bar{\boldsymbol{W}}\boldsymbol{r} \qquad (3)$$

where $\bar{\boldsymbol{W}}$ equals the matrix $\boldsymbol{W}$ with all rows removed for which the corresponding element of $\boldsymbol{p}$ is zero, and $\tilde{\boldsymbol{W}}$ equals the first L columns of $\bar{\boldsymbol{W}}$. Note that $(\tilde{\boldsymbol{W}}^*\tilde{\boldsymbol{W}})^{-1}\tilde{\boldsymbol{W}}^*\bar{\boldsymbol{W}}$ can be precomputed (for different L) and further that in case the noise in (1) is Gaussian and white, (3) gives the ML estimate of the channel. Having established this, it is straightforward to show that the LS (or ML) channel estimate based on *both* long OFDM symbols in the preamble is nothing but the average of the estimate based on the first and the second symbol, respectively.

3. JOINT TIMING AND CHANNEL ESTIMATION

Once an initial timing T_1 is obtained that is less (earlier) than the true timing, the channel impulse response will contain leading zeros (due to the too early timing) and tailing zeros (provided that the effective channel length plus the synchronization error is less than M). If the number of leading and tailing zeros (or equivalently, the correct channel length and timing) can be estimated, the number of unknown channel coefficients will decrease. Hence a more accurate channel estimate can be obtained, which will reduce the bit error rate (BER) in the system (this is known as the *parsimonious principle* in the system identification literature [7]). This is exactly the idea behind our joint timing, channel length and channel coefficient vector estimation algorithm.

To obtain the initial timing estimate, we use a simple correlation approach (see, e.g., [2]) that exploits the fact that the two long OFDM symbols in the preamble are identical. The initial timing estimate T_1 is determined such that it is (with a very large probability) less than the true timing (unless this is ensured, the channel impulse response will not contain leading zeros). Following this, we refine the timing estimate at

the same time as the channel estimation is performed. The details of the procedure are as follows:

1. Let T_1 denote the sample number corresponding to the initial timing (based on a correlation approach [2]).

2. Fix $L = 16$ and increment the timing T starting from $T = T_1$ until the criterion in (2) is minimized. Let the so-obtained T be denoted by T_2.

3. Decrease L starting from $L = 16$ until the following generalized Akaike information criterion (GAIC) is minimized:

$$\ln \|\boldsymbol{W}\boldsymbol{r} - \mathrm{diag}\{\boldsymbol{p}\}\boldsymbol{W}\boldsymbol{h}(L)\|^2 + \gamma L \qquad (4)$$

where $\gamma = 0.08$ (the rationale behind GAIC are discussed in some detail in, e.g., [7]). Denote by L_1 the L that minimizes (4).

4. Increment T (starting from $T = T_2$) and simultaneously decrease L (starting from $L = L_1$) until (4) is minimized. Let the so-obtained final timing and channel length estimates be denoted by $\hat{T}$ and $\hat{L}$, respectively.

Note that the algorithm is iterative but terminates within a finite number of steps.

4. PHASE CORRECTION BASED ON PILOT SYMBOLS

The received signal will inevitably suffer from a CFO, which can be estimated and corrected for using methods such as those in [3, 2, 4]. These methods estimate the CFO based on the received data in the preamble only, and despite being statistically sound, they will never be perfectly accurate. The remaining CFO error results in a phase error that increases linearly with time. As a remedy to this problem, we perform an additional phase correction for each OFDM symbol to compensate for the (small) remaining CFO error.

Each OFDM symbol contains 4 known pilot symbols. Let $\boldsymbol{q}$ be a 4×1 vector of these pilot symbols, and let $\boldsymbol{z}$ be the corresponding 4 elements of the DFT of the received data, i.e., of $\boldsymbol{W}\boldsymbol{r}$. For each OFDM symbol, we estimate a channel phase correction ϕ by minimizing the LS criterion $\|\boldsymbol{z} - \boldsymbol{q}e^{j\phi}\|^2$ which has the solution $\phi = \arg(\boldsymbol{q}^*\boldsymbol{z})$. This phase correction is used to obtain a compensated received signal $\hat{\boldsymbol{r}} = e^{-j\phi}\boldsymbol{r}$, upon which the detection of the data symbols is based. As we illustrate below, this phase correction can have a significant influence on the performance.

5. NUMERICAL EXAMPLES

We provide a few Monte-Carlo simulation results to illustrate the effectiveness of our new algorithm. In all simulations, we consider a Rayleigh fading channel according to [8], with $L = 6$ Gaussian distributed coefficients h_l having a mean power of $\sigma_l^2 = E[|h_l|^2] = \sigma_0^2 e^{-\alpha l}$ for $l = 1, \ldots, L$ and where σ_0 is such that $\sum_{l=1}^{L} \sigma_l^2 = 1$ and $\alpha = 5/3$. The channel is fixed during the transmission of one packet but independent from one packet to another. A CFO of 0.025 Hz is introduced in the simulation and a simple algorithm based on the phase of the correlation of two subsequent OFDM symbols in the preamble is applied to estimate and remove the CFO error (see, e.g., [4]). White Gaussian noise is added to the data to simulate a received signal with a certain ratio of energy per information bit to the spectral density of the noise (E_b/N_0).

Example 1: Timing estimation. Figure 2 shows the distribution of the different timing estimates T_1 (initial coarse timing), T_2 (refined timing estimate from Step 2) and $\hat{T}$ (final timing estimate). The true timing is $T = 194$ and E_b/N_0 is 14 dB. It is clear from the figure that our algorithm succeeded to recover the true timing exactly in more than 90% of the realization, and to within a few sample intervals in virtually all test cases.

Example 2: Estimation of the effective channel length. In Figure 3 we show the distribution of the channel length estimates L_1 (after Step 3) and $\hat{L}$ (the final channel length estimate). Note that the channel length is underestimated in most realizations since the last elements of the impulse response are usually very small.

Example 3: Bit error rate (BER) for QPSK data symbols. We illustrate the BER obtained by simulation of an IEEE 801.11 OFDM system using (a) using our algorithm without the additional channel phase correction; (b) using our algorithm together with the additional phase correction based on the pilot symbols; and (c) perfect knowledge of the timing, channel and CFO. The results are shown in Figure 4. We observe from the figure that our synchronization and channel estimation algorithm achieves a performance close to the bound provided by the exact knowledge of the timing and the transmission channel. Furthermore, it is evident that the usage of the pilot symbols is necessary to fully compensate for the CFO.

6. CONCLUDING REMARK

We have presented a novel and conceptually simple algorithm for joint synchronization and channel estimation for the IEEE 801.11 WLAN standard. The algorithm is based on ML estimation and the GAIC information theoretic criterion. Numerical examples show that as far as the BER is concerned, our algorithm achieves a performance close to the ultimate bound

provided by the exact knowledge of the transmission channel; and therefore eliminates the need for more complicated approaches to the CFO, timing and channel estimation.

7. REFERENCES

[1] R. van Nee, G. Awater, M. Morikura, H. Takanashi, M. Webster, and K. Halford, "New high-rate wireless LAN standards," *IEEE Communications Magazine*, vol. 37, pp. 82–88, December 1999.

[2] T. M. Schmidl and D. C. Cox, "Robust frequency and timing synchronization for OFDM," *IEEE Transactions on Communications*, vol. 45, pp. 1613–1621, December 1997.

[3] P. H. Moose, "A technique for orthogonal frequency division multiplexing frequency offset correction," *IEEE Transactions on Communications*, vol. 42, pp. 2908–2914, October 1994.

[4] J. Li, G. Liu, and G. B. Giannakis, "Carrier frequency offset estimation for OFDM based WLANs," *IEEE Signal Processing Letters*, vol. 8, no. 3, pp. 80–82, March 2001.

[5] J. van de Beek, O. Edfors, M. Sandell, S. Wilson, and P. Börjesson, "On channel estimation in OFDM systems," *Proc. of the Vehicular Technology Conf.*, Chicago, USA, vol. 2, pp. 815–819, July 1995.

[6] R. Negi, and J. Coiffi, "Pilot tone selection for channel estimation in a mobile OFDM system," *IEEE Transactions on Consumer Electronics*, vol. 44, pp. 1122–1128, August 1998.

[7] T. Söderström and P. Stoica, *System Identification*. London, U.K.: Prentice-Hall International, 1989.

[8] N. Chayat, "Tentative criteria for comparison of modulation methods," *doc: IEEE P802.11-97/96*, September 1997.

[9] Z. Wang and G. B. Giannakis, "Wireless multicarrier communication," *IEEE Signal Processing Magazine*, vol. 17, pp. 29–48, May 2000.

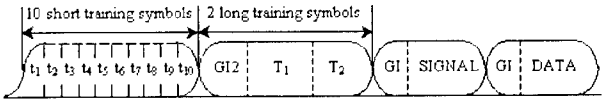


Figure 1: The structure of an OFDM packet.

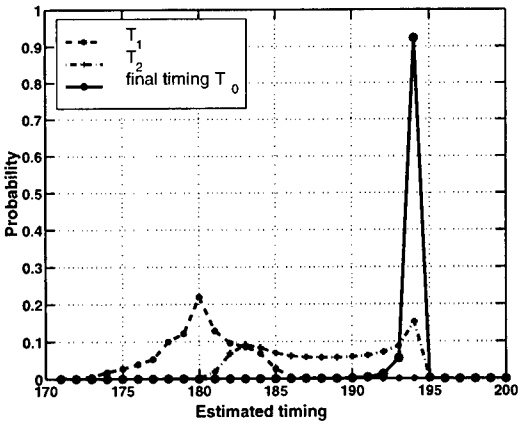


Figure 2: Distribution of the timing estimates.

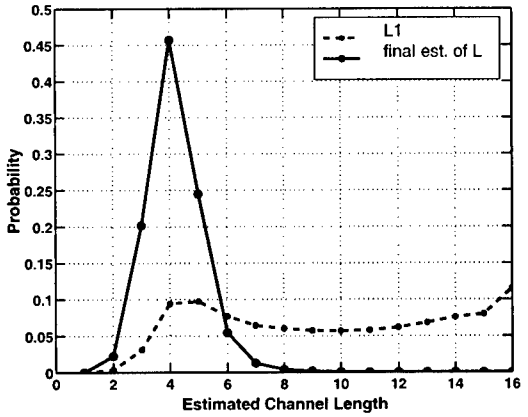


Figure 3: Distribution of the channel length estimates.

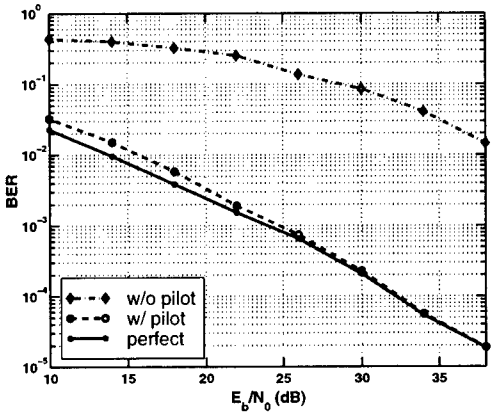


Figure 4: Simulated BER for QPSK data.

A DISTRIBUTED ALGORITHM FOR DYNAMIC SUB-CHANNEL ASSIGNMENT IN A MULTI USER OFDM COMMUNICATION SYSTEM

Alireza Seyedi and Gary J. Saulnier

Department of ECSE, Rensselaer Polytechnic Institute
Room JEC6003, 110, 8th St., Troy, NY, 12180-3590, USA
email: seyeds@rpi.edu

ABSTRACT

In this paper we propose and study a Sub-Channel Selective OFDM spread spectrum system and a distributed algorithm for sub-channel allocation. The proposed system is a combination of OFDMA and OFDM-CDMA systems. For each user the maximal ratio combining weights of the sub-channels are used as channel information. The algorithm implements a simple but sub-optimal version of the water filling power allocation, thus increasing the transmitter power efficiency and decreasing the interference generated for other users. This algorithm also provides the ability to offer different Quality of Service (QoS) levels for different users. The convergence of the algorithm and its performance have been studied through simulation. The proposed system has significant advantage in BER performance, over the conventional OFDM-SS system.

1. INTRODUCTION

Among various techniques used for communication over a wireless channel, Orthogonal Division Frequency Multiplexing (OFDM) is one of the most promising. In an OFDM system the data stream is divided into N parallel streams, which are transmitted over N orthogonal sub-carriers (sub-channels). In the spread spectrum version of OFDM (OFDM-SS, also known as Multi Carrier Spread Spectrum), the same data bit is transmitted over all N sub-channels. OFDM-SS achieves frequency diversity as well as processing gain. Furthermore, OFDM-SS is robust against jamming and interference and enables us to use Code Division Multiple Access (CDMA) as an efficient Radio Resource Allocation (RRA) scheme, for a multi user system. In an OFDM-CDMA system, orthogonal codes of length N are assigned to each user. These codes (signatures) are applied to the sub-channels, to reduce the interference generated by other users.

At the receiver, the output of the sub-channels are combined to obtain the decision variable. For linear sub-channels with additive Gaussian noise and/or interference, Maximal Ratio Combining (MRC) provides the optimum Signal-to-Noise and Interference Ratio (SNIR) [1]. MRC combines the output of the sub-channels by giving less weight to sub-channels with low SNIR and more weight to those with high SNIR. i.e. if sub-channel i is modeled by

$$y_i = \alpha_i x + n_i, \tag{1}$$

where y_i is the output of the sub-channel, x is the data symbol, α_i is the sub-channel gain and $n_i \sim \mathcal{N}(0, \sigma^2)$ is the additive Gaussian noise, the MRC decision variable is given by

$$y = \sum_{i=1}^{N} w_i y_i, \tag{2}$$

where $w_i = \alpha_i^*$. In practice, the sub-channel gains are unknown and a Least Mean Square (LMS) or a Recursive Least Square (RLS) algorithm with decision feedback is employed to estimate these gains [2].

While MRC performs optimum detection, it does not consider the transmitted power efficiency. We observe that a sub-channel with smaller SNIR has smaller contribution to the decision variable compared to one with high SNIR, thus less benefit is gained from the transmit power spent on the sub-channel with low SNIR. The solution known as the water filling power distribution [3] gives the distribution of the power over N sub-channels which results in maximum channel capacity for given total transmission power. The water filling solution allocates more power to *good* (high SNIR) sub-channels and little or no power to the *bad* ones.

Several algorithms have been proposed to implement the water filling solution or a sub-optimal version of it for the multi carrier system. Optimal power and symbol size allocation, assuming that different sub-channels experience different fading and other channel effects, is derived in [4]. [5] proposes an adaptive sub-channel and bit allocation for the down-link of a multi user OFDM system. A dynamic sub-channel allocation for the down-link of OFDM system is proposed in [6] which assumes quasi-static channels. These algorithms are computationally complex, and assume that

0-7803-7011-2/01/$10.00 ©2001 IEEE

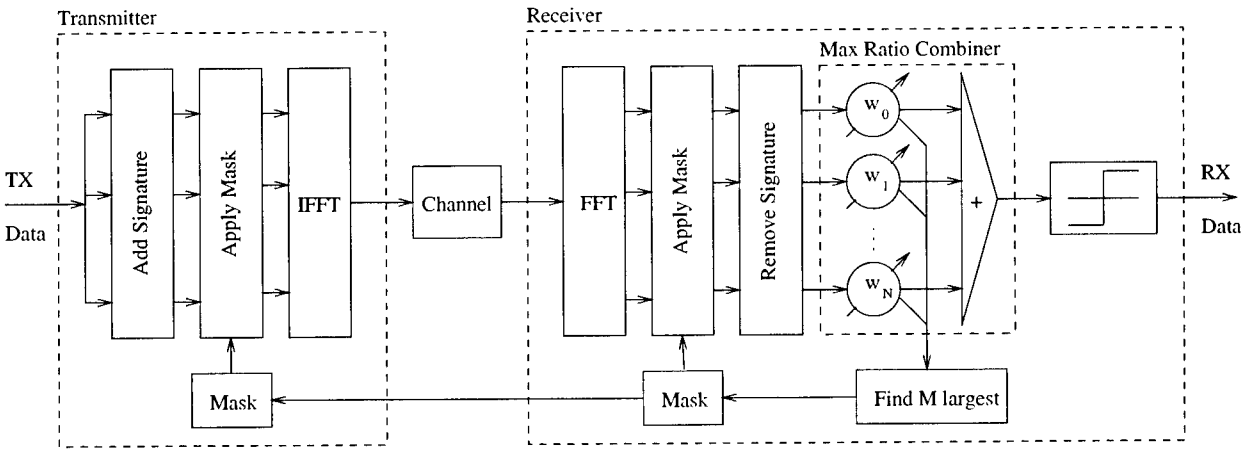


Figure 1: Sub-Channel Selective OFDM-SS system

the perfect channel information of all the users is known by the base station at all times.

In this paper we propose a Sub-Channel Selective OFDM-SS (SCS OFDM-SS) scheme that implements a sub-optimal but simple distributed algorithm for the sub-channel power allocation in a multi user OFDM system. The result is a hybrid of OFDM-CDMA and OFDMA systems [7][8]. An Orthogonal Frequency Division Multiple Access (OFDMA) system is one in which each sub-channel is assigned to one user.

The system is described in section 2. The proposed distributed algorithm is detailed in section 3, and in section 4 the convergence of this algorithm for specific cases is studied. Section 5 presents the simulation results for the performance of the proposed system and compares them with the performance of the conventional system. Finally in section 6 conclusions are drawn.

2. SYSTEM DESCRIPTION

In the SCS OFDM-SS system, the magnitude of the MRC channel weights, $|w_i|$, are used to distinguish *good* sub-channels from *bad* ones. At the receiver, M best sub-channels are selected and a set of binary mask variables are defined as $m_i = 1$ if the sub-channel i is selected and $m_i = 0$ otherwise. These variables are then reported to the transmitter via a feedback channel. The transmitter in turn allocates equal power to the selected sub-channels and no power to the unselected ones (Figure 1).

Although not using the *bad* sub-channels in the combining slightly deteriorates the resulting Bit Error Rate (BER), redirecting the power spent on these sub-channels to the *good* ones improves the over all BER performance. This is similar to the water filling algorithm if the value of the assigned power to each sub-channel is limited to 0 or a constant value.

The more important benefit of this algorithm, however, appears in a multi user system. In this case, a *bad* channel for user k is not necessarily *bad* for user l, thus turning off such a sub-channel will reduce the interference that user l receives. In other words, the users try to avoid the sub-channels with high level of interference. This will result in less average interference for the users, and improves the over all performance.

Furthermore, by selecting different number of selected sub-channels for different users, different Quality of Service (QoS) levels can be provided for different users.

3. DROP AND ADD (DA) ALGORITHM

When the channel is time variable, or when the users enter and leave the system, the selected sub-channels must be updated periodically. The period of the updating of the *good* sub-channels must be short enough to allow the system to follow the changes in the channel and long enough to allow the channel weight estimator to obtain a good estimate.

Since the channel weight estimator uses the output of the sub-channels to estimate the weights, channel weights will not be calculated for the off sub-channels and these sub-channels should be probed before the selected sub-channels are updated. Here we propose the Drop and Add (DA) algorithm, in which each user drops the worst selected sub-channel (minimum $|w_i|$) and adds a random off sub-channel to the selected sub-channels with a period KT, where K is the total number of users in the system, . If the new added sub-channel happens to be a *bad* one it will be dropped in a later iteration, otherwise the user has found a *good* sub-channel.

For faster convergence, we assume that all users are synchronized in such a way that user k performs the updating iterations in times $t_n = nKT + kT, n = 0, 1, ...$, such that the next user has at least T seconds to adapt to the change.

4. CONVERGENCE OF THE DA ALGORITHM

The system can be modeled as a Markov Chain, where each state will describe a specific sub-channel selection for all the users. Thus the number of states in this Markov Chain will be

$$n_{MC} = \sum_{k=1}^{K} \begin{pmatrix} N \\ M_k \end{pmatrix}, \quad (3)$$

where M_k is the number of selected channels for user k. Furthermore, since different users have different channel gains, the convergence analysis for such a system is complicated even for small number of users in the network.

In this paper we have studied the convergence of the DA algorithm through simulation. A Gaussian model has been used for the interference. Over time, the average SNIR of the users has been compared to the optimum case.

Figure 2 shows this comparison for a system with $K = 4$ users with flat channels (equal sub-channel gains), $N = 64$ sub-channels and $M_k = 16, k = 1, ..., 4$ and Figure 3 shows the comparison with $K = 2$ users with different frequency selective channels, with $N = 64$ sub-channels and $M_k = 32, k = 1, 2$. For simplicity, these values have been chosen to satisfy $N = \sum M_k$, which means that sub-channel distributions exist which allow all users to communicate without any interference. In general $\sum M_k$ can be larger than N which will result in some overlap between the selected sub-channels of different users. Perfect synchronization and exact channel weight estimation has been assumed. Since in the down link the original and the interference signal have the same path, all sub-channels have the same SNIR and will be equally preferred by the selection process. In other words the down link can be modeled considering same (flat) channels for all users. On the other hand, for the up-link, the original and the interference signals have different paths, thus the channels used for the original and the interference signals can be frequency selective and different.

In both cases it can be seen that although the average SNIR resulting from the DA algorithm does not converge to the optimum value, the resulting average SNIR is close to optimum after a few iterations.

5. BER PERFORMANCE RESULTS

The BER performance of the SCS OFDM-SS system has been studied using Monte Carlo simulations of the system.

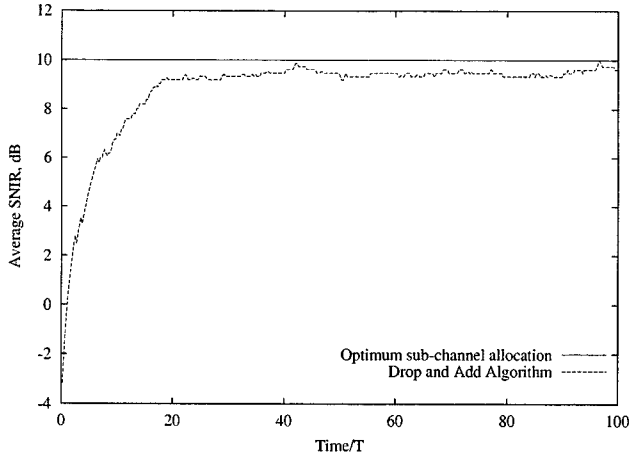


Figure 2: Convergence of the DA algorithm for $K = 4$ users with flat channels (down-link).

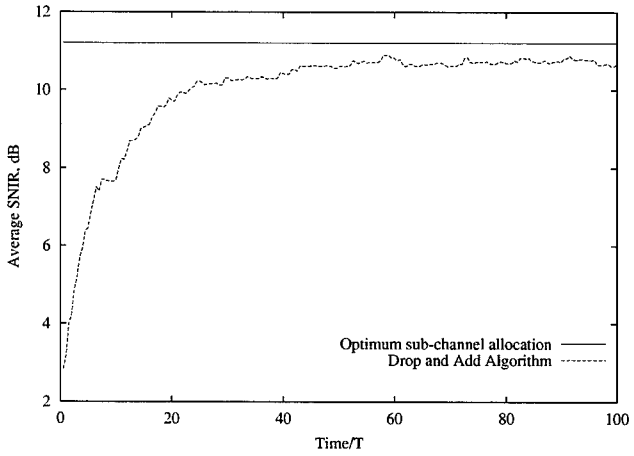


Figure 3: Convergence of the DA algorithm for $K = 2$ users with different frequency selective channels (up-link).

For these simulations it has been assumed that the synchronization and the channel weight estimates are perfect. Also it has been assumed that perfect power control exists, therefore the original and interference signals have equal power. A value of $N = 64$ sub-channels has been used, and a fair distribution of the sub-channels is considered, i.e. $M_k = N/K, k = 1, ..., K$. Also random spreading codes (signatures) have been used.

Figure 4 shows the BER performance of each user for the down-link (flat channels) of a SCS OFDM-SS system with $K = 4$ users. The result has been compared with the BER performance in an equivalent conventional OFDM-SS system. Figure 5 compares the BER performance of the SCS and conventional systems for the uplink (frequency selective channels) for $K = 2$ user case. The frequency

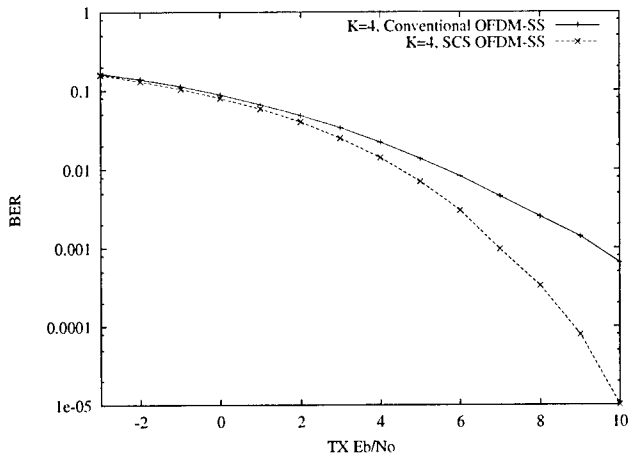


Figure 4: Comparison of BER performance of the down link of the SCS and conventional OFDM-SS systems for $K = 4$ users.

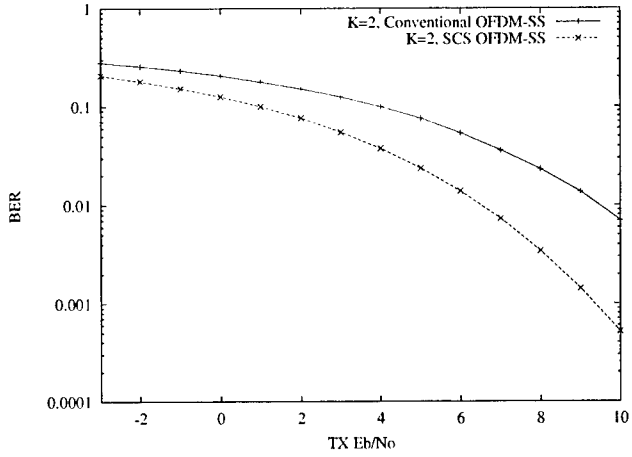


Figure 5: Comparison of BER performance of the up link of the SCS and conventional OFDM-SS systems for $K = 2$ users.

selective channels have been modeled by two rays, with equal power and different delays. The second ray delays of $D_1 = 16T_b/N$ and $D_2 = 4T_b/N$ has been used for users 1 and 2 respectively.

In both cases we can see that the SCS OFDM-SS system has a strong advantage in BER performance compared to the conventional OFDM-SS system.

6. CONCLUSIONS

A simple, sub-optimal distributed algorithm for sub-channel allocation in a OFDM-SS system has been proposed and simulated. The convergence of the algorithm has been studied. Also the overall BER performance of the system has been obtained and compared to that of the conventional system. The results show significant improvement in the BER performance of the users.

7. REFERENCES

[1] J. G. Proakis, *Digital Communications*, McGraw Hill, New York, NY, 1995.

[2] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, 1991.

[4] B. S. Krongold, K. Ramachandran and D. L. Jones, "Computationally Efficient Optimal Power Allocation Algorithms for Multi Carrier Communication Systems", *IEEE Transactions on Communications*, Vol. 48, pp. 23-27, 2000.

[5] C. Y. Wong, R. S. Cheng, K. B. Letaief and R. D. Murch, "Multi User Sub-Carrier Allocation for OFDM Transmission using Adaptive Modulation", *Proceedings of IEEE 49th Vehicular Technology Conference*, Houston, TX, Jul. 1999.

[6] W. Rhee and J. M. Cioffi, "Increase in Capacity of Multi User OFDM System Using Dynamic Sub-Channel Allocation", *Proceedings of IEEE 51st Vehicular Technology Conference*, Tokyo, Japan, Spring 2000.

[7] J. Lindner, M. Nold, W. G. Teich and M. Schreiner, "MC-CDMA and OFDMA for Indoor Communications: the Influence of Multiple Receiving Antennas", *Proceedings of IEEE International Symposium on Spread Spectrum Techniques and Applications*, Sun City, South Africa, 31st Aug. - 1st Sep., 1998.

[8] R. Nogueroles, M. Bossert and V. Zyablov, "Estimation of User Capacity in Mobile Radio Multiple Access Systems Based on Multi Carrier Modulation", *Proceedings of International Symposium on Communication Theory and Applications*, U.K., Jul. 1997.

A SOS SUBSPACE METHOD FOR BLIND CHANNEL IDENTIFICATION AND EQUALIZATION IN BANDWIDTH EFFICIENT OFDM SYSTEMS BASED ON RECEIVE ANTENNA DIVERSITY

Hassan Ali, Jonathan H. Manton and Yingbo Hua

ARC Special Research Centre for Ultra Broadband Information Networks (CUBIN)
Department of Electrical and Electronic Engineering
The University of Melbourne, Victoria 3010, Australia

ABSTRACT

A blind second order statistical (SOS) subspace based channel identification and equalization technique is introduced and investigated for bandwidth efficient Orthogonal Frequency Division Multiplexing (OFDM) systems. A suitable zero-forcing linear equalizer (ZF-LE) is also proposed. Simulations show that identification and equalization is possible with only a small number of short length OFDM symbols.

1. INTRODUCTION

In OFDM systems, a cyclic prefix (CP) is used to combat inter-block-interference (IBI) caused by the multi-path channel. Additionally, the CP allows for a simple one tap equalization scheme. However, due to the extra symbols required by the CP, the OFDM spectrum is underutilized. This overhead can be significant for channels with long impulse responses and short block transmission formats. One may omit the CP in an attempt to raise the spectral efficiency but at the expense of increased receiver complexity.

A number of techniques [9, 4, 8] have been proposed for spectrally efficient OFDM systems which do not use a CP. However, it appears that the only technique that has so far been proposed to estimate and equalize the channel characteristics in a spectrally efficient OFDM system is the blind iterative block technique in [8]. This paper proposes a new channel estimation and equalization scheme which has a number of advantages over the scheme in [8].

The contributions of this paper are as follows:

(i) A new subspace based blind channel estimator for CP-free OFDM is developed. The main idea is based upon oversampling the channel output in the spatial domain by using multiple antennas and then estimating the channel based on the second order statistics (SOS) of the received signal. (By using spatial diversity, one can achieve enhanced performance without additional power or bandwidth consumption.)

(ii) It is shown how the proposed estimator takes into account known zeros in the input stream for channel identification. (These known zeros in the input stream are referred to as virtual carriers, and are sometimes used in practical OFDM systems [7].) The performance of the new method is compared with the channel subspace (CS) method of Moulines et al. [6] and the scheme in [8] through computer simulations. The method is shown to outperform existing methods if the number of short length OFDM symbols is small. The technique appears to achieve a good trade off between estimation accuracy and receiver complexity.

(iii) A matching zero-forcing linear equalizer (ZF-LE) is developed.

It is noted that the channel estimation and equalization scheme in this paper requires (i) $N > 2L$ and (ii) the sub-channels to have no common zeros. Here, N denotes the OFDM symbol length and L is an upper bound on the order of the sub-channels.

2. BANDWIDTH EFFICIENT OFDM BASEBAND MODEL BASED ON SPATIAL DIVERSITY AND BLIND CHANNEL ESTIMATION

We consider the baseband discrete time OFDM system with multiple FIR model arrangement as shown in Fig.1(a). We assume that this multi-channel FIR system arises from oversampling the channel output in the spatial domain using a multiple receiver system (shown in Fig.1(b)) and consists of Z sub-channels of length $L+1$ at the most. The input signal $s(n)$ and the additive white Gaussian noise (AWGN) $v(n)$ are mutually uncorrelated and stationary. The $v(n)$ is also assumed to be uncorrelated among channels. Also, we assume perfect synchronization of carriers and symbols.

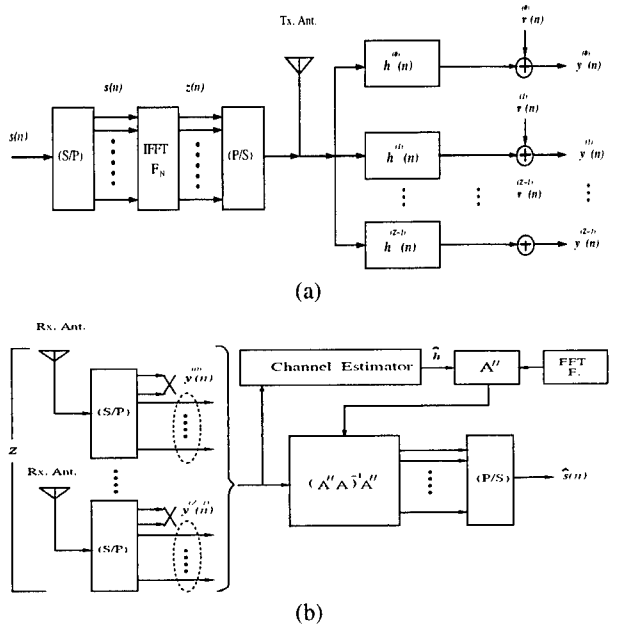
Let us consider a block of N complex valued source symbols at the OFDM transmitter which is given by: $\mathbf{s}(n) = [s(nN), \ldots, s(nN-N+1)]^T$. The elements of $\mathbf{s}(n)$ are assumed to be i.i.d and taken from the complex alphabet $V = \{v_1, v_2, \ldots, v_u\}$ of size u. Considering $\mathbf{s}(n)$ to be in the frequency domain, the time domain signal is generated by the N point inverse fast Fourier transform (IFFT) of the source symbol block $\mathbf{s}(n)$ expressed as: $\mathbf{z}(n) = \mathbf{F}_N\mathbf{s}(n) = [z(nN), z(nN-1), \ldots, z(nN-N+1)]^T$, where $\mathbf{F}_N = [\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_{N-1}]$ is an $N \times N$ IFFT matrix. After parallel-to-serial (P/S) conversion and modulation, the transmitted symbols propagate through multiple channels with impulse response vectors $\mathbf{h}^{(r)} := [h^{(r)}(0), \ldots, h^{(r)}(L)]^T, r = 0, \ldots, Z-1$. At each sensor the received signal is band pass filtered and down-converted to baseband OFDM. The nth received signal at each sensor can be written in block form as: $\mathbf{y}^{(r)}(n) = [y(nN), y(nN-1), \ldots, y(nN-N+1)]^T$. We can relate transmit with receive blocks as:

$$\mathbf{y}^{(r)}(n) = \mathbf{H}_0^{(r)}\mathbf{F}_N\mathbf{s}(n) + \mathbf{H}_1^{(r)}\mathbf{F}_N\mathbf{s}(n-1) + \mathbf{v}^{(r)}(n) \quad (1)$$

where (i) $\mathbf{v}^{(r)}(n)$ is the AWGN vector at rth sensor, (ii) $\mathbf{H}_0^{(r)}$ is the $N \times N$ Toeplitz matrix with first column $[h^{(r)}(0), \ldots, h^{(r)}(L),$

The 3rd author is with The Department of Electrical Engineering, University of California, Riverside, CA, 92521, USA.

0-7803-7011-2/01/$10.00 ©2001 IEEE



(a)

(b)

Fig. 1. Discrete time baseband: (a) transmitter model with multiple FIR channels, (b) receiver model with multiple antennas

$0, \ldots, 0]^T$ and first row $[h^{(r)}(0), 0, \ldots, 0]$ and (iii) $\mathbf{H}_1^{(r)}$ is the $N \times N$ Toeplitz matrix with first column $[0, \ldots, 0]^T$ and first row $[0, \ldots, h^{(r)}(L), \ldots, h^{(r)}(1)]$. Due to the dispersive nature of the rth channel, IBI arises between successive blocks and renders $\mathbf{y}^{(r)}(n)$ in (1) dependent on both $\mathbf{s}(n)$ and $\mathbf{s}(n-1)$. This IBI causes loss of orthogonality and the distinct sub-carriers are no longer orthogonal. To avoid the IBI, we consider a truncated version of $\mathbf{y}^{(r)}(n)$. Suppose that the channel order L is known and L is smaller than N. The truncated version of $\mathbf{y}^{(r)}(n)$ can be written as: $\overline{\mathbf{y}}^{(r)}(n) = [y^{(r)}(nN-L), y^{(r)}(nN-L-1), \ldots, y^{(r)}(nN-N+1)]^T$. Note that the length of $\overline{\mathbf{y}}^{(r)}(n)$ is Q, where $Q = N-L$. Then, $\overline{\mathbf{y}}^{(r)}(n)$ depends only on $\mathbf{s}(n)$, not on $\mathbf{s}(n-1)$, that is

$$\overline{\mathbf{y}}^{(r)}(n) = \mathcal{H}^{(r)}\mathbf{z}(n) + \overline{\mathbf{v}}^{(r)}(n) = \mathcal{H}^{(r)}\mathbf{F}_N\mathbf{s}(n) + \overline{\mathbf{v}}^{(r)}(n) \quad (2)$$

where $\mathcal{H}^{(r)}$ is the Toeplitz channel filtering matrix of size $(N-L) \times N$, with first row $[h^{(r)}(L), \ldots, h^{(r)}(0), 0, \ldots, 0]$ and first column $[h^{(r)}(L), 0, \ldots, 0]^T$. Stacking the outputs of the Z channels gives:

$$\overline{\mathbf{y}}(n) = \mathcal{H}\mathbf{F}_N\mathbf{s}(n) + \overline{\mathbf{v}}(n) \quad (3)$$

where $\overline{\mathbf{y}}(n) = [\overline{\mathbf{y}}^{(0)}(n), \ldots, \overline{\mathbf{y}}^{(Z-1)}(n)]^T, \mathcal{H} = [\mathcal{H}^{(0)} \ldots, \mathcal{H}^{(Z-1)}]^T$ and $\overline{\mathbf{v}}(n) = [\overline{\mathbf{v}}^{(0)}(n), \ldots, \overline{\mathbf{v}}^{(Z-1)}(n)]^T$.

The matrix $\mathcal{H}$ is known as a Generalized Sylvester Matrix, which has full column rank N under the conditions [6]: $(\mathbf{a0})$ the polynomials $H^{(i)}(z) = \sum_{j=0}^{L} h_j^{(i)} z^j$ have no common zero, $(\mathbf{a1})$ Q is greater than the maximum degree L of the polynomials $H^{(i)}(z)$, i.e., $Q \geq L$, and $(\mathbf{a2})$ at least one polynomial $H^{(i)}(z)$ has degree L. The conditions $(\mathbf{a0}) - (\mathbf{a2})$ are assumed to hold throughout this paper, that is $\mathcal{H}$ is assumed to have full rank.

Given a block of data $\{\overline{\mathbf{y}}(n)\}_{n=0}^{K-1}$, the objective here is to estimate the $Z(L+1) \times 1$ vector $\mathbf{h} = [\mathbf{h}^{(0)T}, \ldots, \mathbf{h}^{(Z-1)T}]^T$.

We choose to collect K consecutive data vectors $\{\mathbf{y}(n)\}_{n=0}^{K-1}$ in a matrix: $\overline{\mathbf{Y}}_K := [\overline{\mathbf{y}}(0), \ldots, \overline{\mathbf{y}}(K-1)] = \mathcal{H}\mathbf{F}_N\mathbf{S}_K + \overline{\mathbf{V}}_K$. The covariance matrix of the received data is thus

$$\mathbf{R}_{yy} = E(\overline{\mathbf{Y}}_K\overline{\mathbf{Y}}_K^H) = \mathcal{H}\mathbf{F}_N\mathbf{R}_{ss}\mathbf{F}_N^H\mathcal{H}^H + \mathbf{R}_{vv} \quad (4)$$

where $\mathbf{R}_{ss} = E(\mathbf{S}_K\mathbf{S}_K^H)$ and $\mathbf{R}_{vv} = E(\overline{\mathbf{V}}_K\overline{\mathbf{V}}_K^H)$. It is assumed that the noise is white ($\mathbf{R}_{vv} = \sigma^2\mathbf{I}$) and the input signal is rich enough that $\mathbf{R}_{ss}$ has full rank, i.e., $rank(\mathbf{R}_{ss}) = N$. As in [6], the EVD of $\mathbf{R}_{yy}$ is expressed as

$$\mathbf{R}_{yy} = \mathbf{S}\mathrm{diag}(\lambda_0, \ldots, \lambda_{N-1})\mathbf{S}^H + \sigma_v^2\mathbf{G}\mathbf{G}^H \quad (5)$$

where $\mathbf{S} = [\mathbf{S}_0, \ldots, \mathbf{S}_{N-1}]$ and $\mathbf{G} = [\mathbf{G}_0, \ldots, \mathbf{G}_{Z(N-L)-N-1}]$. The columns of $\mathbf{S}$ span the signal subspace, while those of $\mathbf{G}$, the noise subspace. The columns of $\mathcal{H}$ also span the signal subspace and thus by orthogonality, we have:

$$\mathbf{G}_i^H\mathcal{H} = \mathbf{0}, \quad 0 \leq i \leq Z(N-L) - N - 1. \quad (6)$$

This expression is different from that in the CS method because we calculate $\mathbf{R}_{yy}$ based on $\overline{\mathbf{y}}(n)$ and not on $\mathbf{y}(n)$. However, it can be solved in the least squares sense, as in the CS method to uniquely identify $\mathbf{h}$.

In practical OFDM systems, some sub-carriers are not modulated [7]. These virtual carriers are not used for data transmission but are usually introduced inside the roll-off region (to create a null guard interval) to avoid aliasing effects on data symbols when the system operates over multi-path propagation channels [2].

We assume the presence of $N-P$ virtual carriers at the tail end of each OFDM symbol. Thus each OFDM symbol consists of P modulated source symbols and $N-P$ non-modulated symbols. The IFFT matrix $\mathbf{F}_N$ thus reduces to a partial $N \times P$ matrix $\tilde{\mathbf{F}}_N = [\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_{P-1}]$. The removal of $N-P$ columns of $\mathbf{F}_N$ correspond to the $N-P$ virtual carriers in $\mathbf{s}(n)$. The received data model is then given by:

$$\tilde{\mathbf{y}}(n) = \mathcal{H}\tilde{\mathbf{F}}_N\tilde{\mathbf{s}}(n) + \overline{\mathbf{v}}(n) \quad (7)$$

where $\tilde{\mathbf{s}}(n)$ denotes the new data vector of reduced length P. Under the data model (7), the resulting equation (6) no longer has a unique solution because $rank(\mathbf{R}_{ss}) = P$. (This is in contrast to $rank(\mathbf{R}_{ss}) = N$ in case of no virtual carriers.) In the following section, the corresponding adjustments which take into account the virtual carriers are detailed.

Following the analogous steps (4)-(5) for the data model (7), the corresponding basis for the noise subspace is given by: $\tilde{\mathbf{G}} = [\tilde{\mathbf{G}}_0, \tilde{\mathbf{G}}_1, \ldots, \tilde{\mathbf{G}}_{Z(N-L)-N-1}]$. Also, we observe that the columns of $\mathcal{H}\tilde{\mathbf{F}}_N$ also span the signal subspace and thus by orthogonality we have: $\tilde{\mathbf{G}}_i^H\mathcal{H}\tilde{\mathbf{F}}_N = \mathbf{0}$. In practice, since the output data vectors are noisy, this equation is solved by minimizing the quadratic form:

$$q(\mathbf{h}) = \sum_{i=0}^{Z(N-L)-N-1} |\ \tilde{\mathbf{G}}_i^H\mathcal{H}\tilde{\mathbf{F}}_N|^2. \quad (8)$$

Let $\tilde{\mathbf{G}}_i^H\mathcal{H}\tilde{\mathbf{F}}_N = \mathbf{h}^T\tilde{\mathcal{G}}_i\tilde{\mathbf{F}}_N$ where $\tilde{\mathcal{G}}_i$ is the $Z(L+1) \times N$ filtering matrix associated with $\tilde{\mathbf{G}}_i$ and can be obtained by back substitution. Therefore $|\ \tilde{\mathbf{G}}_i^H\mathcal{H}_P\tilde{\mathbf{F}}_N\ |^2 = \mathbf{h}^H\tilde{\mathcal{G}}_i\tilde{\mathbf{F}}_N\tilde{\mathbf{F}}_N^H\tilde{\mathcal{G}}_i^H\mathbf{h}$ and equation (8) can thus be expressed as: $q(\mathbf{h}) = \mathbf{h}^H\tilde{\mathbf{Q}}\mathbf{h}$, where $\tilde{\mathbf{Q}} = \sum_{i=0}^{Z(N-L)-N-1}\tilde{\mathcal{G}}_i\tilde{\mathbf{F}}_N\tilde{\mathbf{F}}_N^H\tilde{\mathcal{G}}_i^H$ and the channel estimate can thus be formulated as

$$\hat{\mathbf{h}} = \arg\min_{\|\mathbf{h}\|=1} \|\ \mathbf{h}^H\tilde{\mathbf{Q}}\mathbf{h}\ \|^2 . \quad (9)$$

This quadratic optimization criterion allows unique estimation of $\mathbf{h}$ up to a scale factor and $\hat{\mathbf{h}}$ is thus obtained as the eigenvector associated with the minimum eigenvalue of $\tilde{\mathbf{Q}}$.

3. ZERO FORCING-LINEAR EQUALIZER (ZF-LE)

In this section we present a ZF-LE which gives linear estimates of the input symbols $\mathbf{s}(n)$ based on the received data and channel state information according to the ZF criteria.

Let $\hat{\mathcal{H}}$ is the estimate of matrix $\mathcal{H}$. With $\hat{\mathcal{H}}$ and $\mathbf{F}_N$ (or $\tilde{\mathbf{F}}_N$ in case of virtual carriers) known at the receiver, the received signal matrix in (3) can be written as: $\overline{\mathbf{y}}(n) = \mathbf{A}\mathbf{s}(n) + \overline{\mathbf{v}}(n)$ where $\mathbf{A} = \hat{\mathcal{H}}\mathbf{F}_N$.

From the estimation theory, the continuous valued unbiased maximum-likelihood estimate $\hat{\mathbf{s}}(n)$ of the vector $\mathbf{s}(n)$ is given by: $\hat{\mathbf{s}}(n) = \mathbf{G}_{zf}\overline{\mathbf{y}}(n)$ where $\mathbf{G}_{zf}$, is the ZF-LE given by $\mathbf{G}_{zf} = \mathbf{A}^{\dagger}$, and $\mathbf{A}^{\dagger}$ denotes the Moore Penrose Pseudo-inverse of $\mathbf{A}$. (We observe that the ZF-LE is the same as the MMSE-ZF-LE in [3].)

Note that the source symbols can be recovered provided $\mathbf{A}$ has full rank, which it does if assumptions $(\mathbf{a0}) - (\mathbf{a2})$ hold. The ZF-LE can be implemented as shown in Fig.1(b).

4. NUMERICAL RESULTS

We provide simulation results which compare the performance of our algorithm to the schemes in [6] and [8]. The input symbols were generated using randomly drawn BPSK symbols. We simulated the output of $Z = 4$ receivers and all evaluations are made for a $N = 25$ carrier OFDM system (unless specified). The channel coefficients are chosen as in [6]. Fig.2(a) shows the zero locations of the channel set. To evaluate the channel estimation error, we employed the normalized-root-mean-square- error (NRMSE) (as defined in [1]) with 100 Monte Carlo runs.

Simulation Example 1. Fig.2(b) shows the estimator performance at an SNR=35 dB as a function of the number of OFDM symbols (from 40-160). We can see the performance of the estimator improves with increasing the number of symbols and large number of symbols are required to obtain good channel estimates.

Simulation Example 2. In this simulation study, we fixed the number of symbols to be 100 and varied the SNR from 10-40 dB. Fig.2(c) shows NRMSE as a function of SNR. We observe that the estimator performance improves with increasing SNR.

Simulation Example 3. We investigated the influence of the OFDM symbol length N on the channel estimator performance. The SNR is fixed at 35 dB. Fig.2(d) shows that the performance of the estimator improves by increasing N, however, performance does not improve much beyond $N = 25$ for increasing number of OFDM symbols. It is observed that in contrast to $N = 25$, the estimator performance degrades with $N = 30$ and $N = 35$ for 40 OFDM symbols. This indicates that the proposed channel estimator is quite suitable for sufficiently small number of short length OFDM symbols.

Simulation Example 4. Fig.3(a) illustrates the effect of increasing the number of virtual carriers at the tail end of OFDM symbols. We observe that with the small number of OFDM symbols, virtual carriers degrade the estimator performance. It is also clear that this effect can be suppressed by increasing the number of OFDM symbols.

Simulation Example 5. We also implemented CS method for SNR=35 dB, 2 virtual carriers and 80 OFDM symbols. As can be seen from Fig.3(b), the NRMSE of CS method increases with the increase in window length.

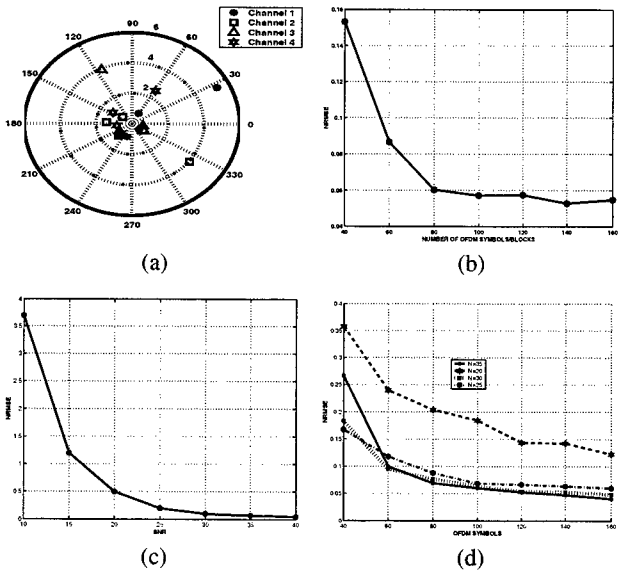


Fig. 2. (a) zero locations (b) Channel error vs OFDM symbols (c) Channel error vs SNR; and (d) Channel error vs OFDM symbols of different lengths

Simulation Example 6. Fig.3(c) shows the performances of CS and the new estimator for a fixed window length Q=22, 2 virtual carriers and SNR=35 dB with increasing the number of OFDM symbols. The new technique is seen to be closely approaching the performance of CS method for number of OFDM symbols greater than 60. It is also seen that, the new technique outperforms CS method for number of OFDM symbols less than 60. We thus have the possibility of complexity trade-off: An increase in performance as well as identification with much smaller number of symbols can be achieved with the new technique by providing additional receiver antennas.

Simulation Example 7. For the channel set, the phase pattern of the receiver outputs is plotted in Fig.3(d). Using the channel estimates via the proposed method for 40 OFDM symbols at 35 dB ZF equalizer was implemented. The equalized phase pattern is shown in Fig.4(a).

Simulation Example 8. An interesting point to compare the performance of the proposed estimator with [8] for short length OFDM symbols. Allowing for 4 virtual carriers, 1000 OFDM symbols and 100 iterations, Fig.4(d) gives an insight of channel estimates by [8] where phase pattern is obtained by using corresponding channel estimates. It is clear that channel estimates by [8] are unacceptable and equalization is impossible. In contrast, as shown in previous simulation results the proposed method rapidly converges with much smaller number of short length OFDM symbols.

Remark 1: The processing of large length OFDM symbols causes significantly large time delays which is an important factor in time delay sensitive services. Since our proposed technique is feasible for short length OFDM symbols, it is suitable for use in wireless local area networks (WLANs) [5].

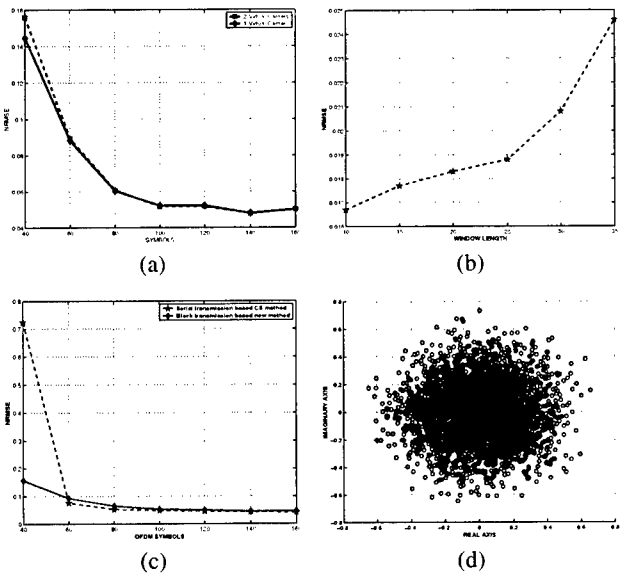


Fig. 3. (a) Channel error vs OFDM symbols with different number of virtual carriers (b) CS method: channel error vs window length (c) CS vs new method: channel error vs OFDM symbols and (d) phase pattern before equalization

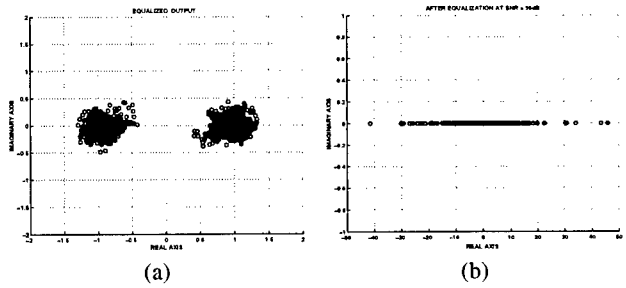


Fig. 4. Phase pattern after ZF equalization: (a) using proposed method channel estimates (b) using estimates of [8]

5. CONCLUSION

Conclusion is that although Moulines' method works even if there are virtual carriers, this method can be modified to improve the performance if virtual carriers are present and the number OFDM symbols is small.

6. REFERENCES

[1] H. Ali, J. H. Manton, and Y. Hua. Modified channel subspace method for identification of SIMO FIR channels driven by a trailing zero filter bank precoder. In *IEEE International Conference on Acoustics, Speech and Signal Processing 2001, ICASSP'2001*, Salt Lake City, Utah, May 2001.

[2] D. Dari, V. Tralli, and A. Vaccari. A novel complexity technique to reduce non-linear distortion effects in OFDM systems. In *The ninth IEEE international symposium on personal, indoor and mobile radio communications*, volume 2, pages 795–800, 1998.

[3] E. de Carvalho and D. T. M. Slock. Burst mode equalization: optimal approach and suboptimal continuous-processing approximation. *Signal Processing*, 80(10):1999–2015, 2000.

[4] M. de Courville, P. Duhamel, P. Madec, and J. Palicot. Blind equalization of OFDM systems based on the minimization of a quadratic criterion. In *Proc. International Conference on Communications 1996, ICC'96*, pages 1318–1322, 1996.

[5] J. Mikkonen, J. Aldis, G. Awater, A. Lunn, and D. Hutchison. The Magic WAND-Functional Overview. *IEEE Journal on Selected Areas in Communications*, 16(6):953–972, August 1998.

[6] E. Moulines, P. Duhamel, J. Cardoso, and S. Mayrargue. Subspace methods for the blind identification of multichannel FIR filters. *IEEE Transactions on Signal Processing*, 43(2):516–525, February 1995.

[7] H. Sari, G. Karam, and I. Jeanclaude. Transmission techniques for digital terrestrial TV broadcasting. *IEEE Commun. Mag.*, 33:100–109, February 1995.

[8] Y. Song, S. Roy, and L. Akers. Joint blind estimation of channel and data symbols in OFDM. In *IEEE 51st Vehicular Technology Conference, VTC'2000*, volume 1, pages 46–50, Tokyo, 2000.

[9] Y. Sun and L. Tong. Channel equalization for wireless OFDM systems with ICI and ISI. In *Proc. IEEE International Conference on Communications, ICC'99*, volume 1, pages 182–186, 1999.

A CHANNEL CODED CP-OFDM INTERPRETATION OF TZ-OFDM SYSTEMS

Jonathan H. Manton

ARC Special Research Centre for Ultra-Broadband Information Networks
Department of Electrical and Electronic Engineering
The University of Melbourne, Victoria 3010, Australia.
j.manton@ee.mu.oz.au

ABSTRACT

Trailing zero OFDM systems replace the cyclic prefix in OFDM systems by a sequence of zeros. However, this paper shows that a cyclic prefix is still present in TZ-OFDM systems. Indeed, it is shown that a TZ-OFDM system implicitly works by first adding redundancy to the symbols to be transmitted (channel coding) and then adding a cyclic prefix. This paper also proves that the channel coding introduced by the TZ-OFDM system is spectrally balanced, meaning that it maps white noise to (essentially) white noise. This is a desirable property because it is known that any coding scheme which achieves the channel capacity over an unknown multipath channel must be white-like. By introducing the Cramer-Rao Bound as a figure of merit, it is shown that there exist channels over which a TZ-OFDM system performs worse than an uncoded OFDM system. The Cramer-Rao Bound is also used to explain why using a cyclic prefix is desirable; it allows channels with otherwise unstable inverses to be inverted accurately.

1. INTRODUCTION

Orthogonal Frequency Division Multiplex (OFDM) systems [2, 8] transmit data in blocks, with a cyclic prefix added to the start of each block. Recently, it was proposed in [5, 7] to replace the cyclic prefix by a sequence of zeros. The resulting system is known as a Trailing Zero OFDM (TZ-OFDM) system. The main advantage [4] of TZ-OFDM over OFDM is that the source symbols can always be recovered regardless of the location of the channel zeros.

On the surface, it may appear that TZ-OFDM and OFDM (henceforth referred to as Cyclic Prefixed OFDM, or CP-OFDM for short) have little in common. The main contribution of this paper is to show that TZ-OFDM, with one proviso[1], is actually a special case of channel coded CP-OFDM. Moreover, it is proved that the channel coding is such that the transmitted signal has a flat power spectrum provided the source symbols are white. This is a desirable property of a channel coder for unknown multipath channels; if the location of the channel spectral nulls are unknown, then it is best to spread the transmitter power evenly over all subchannels [1, 3, 6].

This work was performed while the author was a Tan Chin Tuan Exchange Fellow at Nanyang Technological University, Singapore.

[1]The channel coded CP-OFDM interpretation results in 2L-2 zeros between each block whereas TZ-OFDM uses only L-1 zeros. However, the performance of these two systems, as measured by the MSE of the equalised source symbols, is identical.

2. CP-OFDM INTERPRETATION OF TZ-OFDM

The channel coded CP-OFDM interpretation of TZ-OFDM is illustrated below by way of example. Consider transmitting two blocks of three symbols each over a length $L = 2$ FIR channel using a TZ-OFDM framework. For convenience, the following two blocks of symbols are chosen:

$$[6, -1.5 + \jmath 0.866, -1.5 - \jmath 0.866]^T, \quad [15, -1.5 + \jmath 0.866, -1.5 - \jmath 0.866]^T. \quad (1)$$

First, the blocks are IDFT'ed, to obtain:

$$[1, 2, 3]^T, \qquad [4, 5, 6]^T. \quad (2)$$

A trailing zero (just one, since $L - 1 = 1$) is now added to each block. However, it is also necessary to add a zero at the very start for initialisation purposes. Therefore, the transmitted sequence is

$$\{0, 1, 2, 3, 0, 4, 5, 6, 0\}. \quad (3)$$

It is now shown how the same result can be obtained in a linearly precoded CP-OFDM framework.

Define the linear precoder matrix

$$P = \begin{bmatrix} 1 & 0 & 0 \\ -0.333\jmath & 0.789 + 0.455\jmath & 0.211 - 0.122\jmath \\ 0.333 & 0.333 - 0.577\jmath & 0.333 + 0.577\jmath \\ 0.333\jmath & 0.211 + 0.122\jmath & 0.789 - 0.455\jmath \end{bmatrix}. \quad (4)$$

Then, apply P to the two blocks in (1) to obtain:

$$P[6, -1.5 + \jmath 0.866, -1.5 - \jmath 0.866]^T = [6, -2 - 2\jmath, 2, -2 + 2\jmath], \quad (5)$$

$$P[15, -1.5 + \jmath 0.866, -1.5 - \jmath 0.866]^T = [15, -2 - 5\jmath, 5, -2 + 5\jmath]. \quad (6)$$

Now, encode each block as in a conventional CP-OFDM system, first by taking the IDFT of each block to obtain:

$$[1, 2, 3, 0]^T, \qquad [4, 5, 6, 0]^T \quad (7)$$

and then by adding a cyclic prefix:

$$[0, 1, 2, 3, 0]^T, \qquad [0, 4, 5, 6, 0]^T. \quad (8)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

Now, a CP-OFDM system would transmit

$$\{0, 1, 2, 3, 0, 0, 4, 5, 6, 0\}. \tag{9}$$

However, since the memory length of the channel is one, there is no reason for transmitting two consecutive zeros. (Precisely, no extra information is gained at the receiver by transmitting two zeros rather than a single zero.) A TZ-OFDM system exploits this fact by replacing the two zeros by a single zero, as in (3). In effect, it is using the zero suffix of the first block (introduced by the channel coding operation) to also serve as a cyclic prefix of the second block.

The fact that a TZ-OFDM receiver does indeed use the zero suffix of the previous block as a cyclic prefix for the current block can be verified by observing that, unlike a CP-OFDM receiver which discards the guard interval, a TZ-OFDM receiver does not. Continuing the above example, a CP-OFDM transmitter encodes the blocks in (1) as $\{3, 1, 2, 3, 6, 4, 5, 6\}$ yet a CP-OFDM receiver can use only 3 received symbols (those corresponding to $1, 2, 3$) to recover the data in the first block. This is due to the problem of inter-block interference (IBI). A TZ-OFDM receiver, on the other hand, can use all 4 symbols (corresponding to $1, 2, 3, 0$ and $4, 5, 6, 0$ in (3)) to recover each block. This is because the zero suffix of the previous block eliminates IBI. That is, the receiver does indeed depend on the previous block having a zero suffix.

Key Point: A CP-OFDM receiver applied to (9) and modified to take into account the presence of the precoder P works identically to a TZ-OFDM receiver applied to (3). This is because the CP-OFDM system uses a transmit block size of 4 (as measured before the cyclic prefix is added) and hence the receiver uses 4 symbols per received block (corresponding to $1, 2, 3, 0$ and $4, 5, 6, 0$ in (9) since it discards the guard interval) to recover the data. The TZ-OFDM system uses a transmit block size of 3 (as measured before the trailing zero is added) and hence the receiver uses 4 symbols per received block (corresponding to $1, 2, 3, 0$ and $4, 5, 6, 0$ in (3)) to recover the data. Therefore, *the performance of the TZ-OFDM system can be analysed by studying the performance of the channel coded CP-OFDM interpretation of it.*

3. CHANNEL CODING

The previous section showed that the performance of a TZ-OFDM system is equivalent to the performance of a channel coded CP-OFDM system, where P in (4) performs the channel coding. Since a CP-OFDM system transmits each symbol over an independent sub-channel, the effect of P in (5) is to spread the 3 source symbols out over 4 sub-channels, that is, *it spreads the spectrum of the source symbols.* This helps explain why a TZ-OFDM system can recover the source symbols regardless of the location of the channel zeros.

This section proves that the precoder P spreads the spectrum in a very special way; it maps white noise to (essentially) white noise. This is a desirable property because it is known that any coding scheme which achieves the channel capacity over an unknown multipath channel must be white-like.

Consider a general TZ-OFDM system operating over a channel of length L and using a block size of p. In order to make (9) with the $2L-2$ zeros between each block reduced to only $L-1$ zeros identical to (3), it is necessary to choose the channel coder P in (5) to be

$$P = \frac{1}{p} D_{p+L-1} \begin{bmatrix} D_p^H \\ 0_{(L-1)\times p} \end{bmatrix} \tag{10}$$

where D_n denotes the $n \times n$ DFT matrix. (Substituting $p = 3$ and $L = 2$ into (10) yields (4).) Such a P has the following property, the proof of which is omitted.

Theorem 1 *Define P as in (10). Then PP^H is a circulant matrix with ones along the diagonal.*

If the source symbols $\boldsymbol{s} \in \mathbb{C}^p$ are white ($\boldsymbol{E}[\boldsymbol{s}\boldsymbol{s}^H] = I$) then the covariance of the precoded symbols $P\boldsymbol{s}$ is $\boldsymbol{E}[(P\boldsymbol{s})(P\boldsymbol{s})^H] = PP^H$. It follows from Theorem 1 that, on average, power is distributed equally on the $p + L - 1$ sub-channels in a TZ-OFDM system. (Recall that the channel coded CP-OFDM interpretation of TZ-OFDM systems makes it possible to speak of independent sub-channels in a TZ-OFDM system.)

Remark: Since P is a tall matrix, it is not possible for the transmitted symbols to be white (that is, for $PP^H = I$). However, Theorem 1 shows that PP^H has ones along the diagonal, which is interpreted here as being "almost white", or white-like. The important point is that power is distributed equally on the sub-channels, a pleasing property if the location of the channel spectral nulls is not known.

4. A PERFORMANCE MEASURE

OFDM receivers operate on the received blocks independently of each other. Therefore, the performance of an OFDM system is fully determined once the performance of recovering a single block is known. This section shows how the Cramer-Rao Bound can be used to measure the performance of a CP-OFDM or TZ-OFDM system. Also, the channel coded CP-OFDM interpretation of a TZ-OFDM system is used to construct a channel $\boldsymbol{h}$ over which, somewhat surprisingly, a TZ-OFDM system performs worse than an uncoded CP-OFDM system.

Consider using an arbitrary linear precoder matrix $\tilde{P} \in \mathbb{C}^{n\times p}$ to encode a single block of unknown source symbols $\boldsymbol{s} \in \mathbb{C}^p$ prior to transmission through a finite impulse response (FIR) channel $\boldsymbol{h} = [h_0, \cdots, h_{L-1}]^T \in \mathbb{C}^L$. The received vector $\boldsymbol{y} \in \mathbb{C}^{n-L+1}$ is given by

$$\boldsymbol{y} = \mathcal{H}\tilde{P}\boldsymbol{s} + \boldsymbol{n}, \qquad \boldsymbol{n} \sim N(0, I) \tag{11}$$

where $\mathcal{H}$ is the upper triangular $(n - L + 1) \times n$ Toeplitz channel matrix with first row equal to $[h_{L-1}, \cdots, h_0, 0, \cdots, 0]$ and $\boldsymbol{n}$ denotes additive white Gaussian noise (AWGN) with unit variance ($\boldsymbol{E}[\boldsymbol{n}\boldsymbol{n}^H] = I$).

Remark 1: The fact that $\mathcal{H}$ has fewer rows than columns is due to the memory of the channel and is related to the problem of IBI; without knowing something about the symbols transmitted just prior to $\tilde{P}\boldsymbol{s}$ being transmitted, the received symbols corresponding to the first $L-1$ symbols of $\tilde{P}\boldsymbol{s}$ provide no information to the receiver about $\boldsymbol{s}$. Similarly, if nothing is known about the symbols transmitted just after $\tilde{P}\boldsymbol{s}$ is transmitted, then no information about $\boldsymbol{s}$ is gained by observing symbols received after $\tilde{P}\boldsymbol{s}$ is transmitted.

It is assumed for simplicity that the receiver has perfect knowledge of the channel parameters $\boldsymbol{h}$. Then, for any unbiased estimate

$\widehat{s}$ of s, the error covariance matrix $\boldsymbol{E}\left[(s-\widehat{s})(s-\widehat{s})^H\right]$ is lower bounded by the Cramer-Rao Bound (CRB)

$$R = \left(\tilde{P}^H \mathcal{H}^H \mathcal{H} \tilde{P}\right)^{-1}. \tag{12}$$

In fact, this lower bound is met with equality if the maximum-likelihood (ML) decoder

$$\widehat{s} = \left(\tilde{P}^H \mathcal{H}^H \mathcal{H} \tilde{P}\right)^{-1} \tilde{P}^H \mathcal{H}^H \boldsymbol{y} \tag{13}$$

is used.

Remark 2: Since (13) is an unbiased estimate, it is also referred to as a Zero Forcing (ZF) equaliser in the literature. Since it achieves the CRB, it is the best possible ZF equaliser for recovering s from $\boldsymbol{y}$.

It is proposed to use R as a figure of merit for the precoder $\tilde{P}$. Indeed, for any given channel $\boldsymbol{h}$, the diagonal element R_{ii} is the mean-square error (MSE) of the estimate of the ith element of s if the ML-decoder is used, while the off-diagonal element R_{ij} is the correlation between the estimates of the ith and jth elements of s.

Remark 3: The CRB ignores the fact that practical communication systems transmit symbols coming from a finite alphabet. However, it is clear that if the elements of R are large then the bit error rate of a practical system using the linear precoder $\tilde{P}$ will be adversely affected.

The relevance of (11) and (12) to practical CP-OFDM and TZ-OFDM systems is now explained. A CP-OFDM system first IDFT's the block and then adds a cyclic prefix of length $L-1$. Since a CP-OFDM receiver operates on blocks separately, it does not keep any information about the previously decoded block. Therefore, the first $L-1$ received symbols of the current block provide no information about the transmitted symbols (see Remark 1 above) and are discarded by the receiver. Thus, for a CP-OFDM system, if $\tilde{P} = CD^H$ where

$$C = \begin{bmatrix} 0_{(L-1)\times(n-2L+2)} & I_{L-1} \\ & I_{n-L+1} \end{bmatrix} \tag{14}$$

adds a cyclic prefix and D is a DFT matrix, then (11) correctly models all the information available to the receiver about the current block s, and hence (12) is the best achievable performance of any unbiased equaliser in a CP-OFDM system.

A TZ-OFDM system first IDFT's the block and then appends $L-1$ trailing zeros. Unlike in a CP-OFDM system though, the first $L-1$ received symbols of the current block do provide information about the transmitted symbols. This is because the receiver knows that $L-1$ zeros (corresponding to the trailing zeros of the previous block) were transmitted just prior to the current block! In order to incorporate this extra information into (11), it is necessary to use a $\tilde{P}$ different from expected. Specifically, the correct $\tilde{P}$ is one which first IDFT's the block and then adds both $L-1$ leading zeros and $L-1$ trailing zeros to the block. Then, (11) correctly models all the information available to the receiver about the current block s, and thus (12) is the best achievable performance of any unbiased equaliser in a TZ-OFDM system.

Remark 4: The simple rule is that $\tilde{P}$ in (11) is chosen so that $\boldsymbol{y}$ contains all the information available to the equaliser in a single received block.

Remark 5: An alternative way of deriving the correct $\tilde{P}$ to use in (11) to model a TZ-OFDM system is to use the channel coded CP-OFDM interpretation. Then, it is clear that $\tilde{P} = CD^H P$ for appropriately sized cyclic prefix matrix C and DFT matrix D. Here, P is defined as in (10).

Although a TZ-OFDM system performs better than an uncoded CP-OFDM system over most channels $\boldsymbol{h}$, the proof of the following theorem uses the channel coded CP-OFDM interpretation of a TZ-OFDM system to construct a channel $\boldsymbol{h}$ over which a TZ-OFDM system performs worse than an uncoded CP-OFDM system.

Theorem 2 *Consider sending a single block of p symbols over an FIR channel of length L, using either a CP-OFDM precoder $\tilde{P}_1$ of size $(p+L-1)\times p$ or a TZ-OFDM precoder $\tilde{P}_2$ of size $(p+2L-2)\times p$. Here, $\tilde{P}_1 = CD^H$ where C is a $(p+L-1)\times p$ cyclic prefix matrix and D a DFT matrix, while $\tilde{P}_2 = C[D\ 0_{p\times(L-1)}]^H$ where C is now a $(p+2L-2)\times(p+L-1)$ cyclic prefix matrix. Let R_1 and R_2 be the associated CRB matrices, defined in (12), for a given channel vector $\boldsymbol{h}$. Then, there exist values of p, L and $\boldsymbol{h}$ for which $\mathrm{tr}\{R_1\} < \mathrm{tr}\{R_2\}$, meaning that a CP-OFDM system can sometimes perform better than a TZ-OFDM system.*

PROOF. Choose $p = 3$ and $L = 3$. The TZ-OFDM system spreads the $p = 3$ symbols over $p+L-1 = 5$ sub-channels. Choose $\boldsymbol{h}$ to have spectral nulls on the 2nd and 5th sub-channels, that is, $\boldsymbol{h} = [1\ -0.618\ 1]^T$. Then $\mathrm{tr}\{R_1\} = 1.29$ and $\mathrm{tr}\{R_2\} = 1.88$. □

It is emphasised that $\tilde{P}_1$ and $\tilde{P}_2$ are chosen in Theorem 2 so that (13) correctly models the best CP-OFDM equaliser and the best TZ-OFDM equaliser respectively, where best means the minimum variance unbiased equaliser based on all the available information at the receiver. Since the CP-OFDM equaliser discards the guard interval whereas the TZ-OFDM equaliser does not, it is necessary for $\tilde{P}_2$ to have more rows than $\tilde{P}_1$.

5. IMPORTANCE OF CYCLIC PREFIX

For completeness, this section mentions that the cyclic prefix, besides allowing individual data symbols to be transmitted over independent sub-channels, serves another important purpose. A cyclic prefix *allows channels with otherwise unstable inverses to be inverted accurately*. This is readily demonstrated with the aid of the CRB (12).

Consider sending the symbols $1, 2, 3$ over a length $L = 2$ FIR channel by first precoding the symbols to form one of: (A) $0, 1, 2, 3$, (B) $1, 2, 3, 0$, or (C) $3, 1, 2, 3$. Consider the following four test channels:

$$\boldsymbol{h}_1 = [1\ 0]^T,\ \boldsymbol{h}_2 = [0\ 1]^T,\ \boldsymbol{h}_3 = [1\ -5]^T,\ \boldsymbol{h}_4 = [-5\ 1]^T. \tag{15}$$

The resulting CRB, given by R in (12), can be calculated for any combination of precoder and channel. Of most interest are the following combinations:

$$R_{A3} = \begin{bmatrix} 1 & 5 & 25 \\ 5 & 26 & 130 \\ 25 & 130 & 651 \end{bmatrix},\quad R_{B4} = \begin{bmatrix} 651 & 130 & 25 \\ 130 & 26 & 5 \\ 25 & 5 & 1 \end{bmatrix} \tag{16}$$

where R_{A3} denotes the combination of precoder A and channel $\boldsymbol{h}_3$, and similarly for R_{B4}. Furthermore, the CRB is not defined (implying that not all the symbols can be recovered) if precoder A is used over channel 2, or if precoder B is used over channel

1. It can be verified that all other combinations lead to reasonable values of R. For example,

$$R_{C3} = R_{C4} = \begin{bmatrix} 0.04 & 0.01 & 0.01 \\ 0.01 & 0.04 & 0.01 \\ 0.01 & 0.01 & 0.04 \end{bmatrix} \quad (17)$$

There is a simple explanation for (16). The channel $\boldsymbol{h}_3$ is non-minimum phase, and hence has an unstable inverse. This is reflected by the exponential growth in the diagonal elements of R_{A3}. Precoder B puts a known symbol at the end, and hence performs poorly over channels which have an unstable inverse when run backwards (such as any non-maximum phase channel). Since a channel generated at random has a reasonable chance of being non-minimum phase, it is clear that Precoder A is unsuitable for use in practice, and similarly for Precoder B.

Remark: Note that using Precoder B in (11) does *not* model a TZ-OFDM system. The correct precoder to use in (11) so as to model a TZ-OFDM system is $\tilde{P}_2$ in Theorem 2, and in particular, it will become clear that the performance of TZ-OFDM systems is not affected by non-minimum phase channels.

This observation can be generalised as follows. Any precoder which does not make the last $L-1$ transmitted symbols in a block[2] a known function of the first $L-1$ transmitted symbols will perform badly if the channel is non-minimum phase. This is because errors introduced by noise near the start of the block build up exponentially in magnitude and go unchecked unless the receiver is able to reconcile the last $L-1$ symbols with their true values (both Precoders B and C allow this reconciliation, for instance). Similarly, if the first $L-1$ transmitted symbols are not a known function of the last $L-1$ transmitted symbols in a block, then channels which are non-maximum phase will lead to an exponential growth of errors in the reverse direction (as shown by R_{B4}). A cyclic prefix precoder is distinguished by the fact that it satisfies both properties; the first $L-1$ symbols of each block are a function of the last $L-1$ symbols, and the last $L-1$ symbols of each block are a function of the first $L-1$ symbols. Therefore, a cyclic prefix prevents an exponential growth of errors regardless of the phase of the channel.

In fact, the key property of the cyclic prefix is that its performance is invariant to the phase of the channel spectrum (see Theorem 3 below, whose straightforward proof is omitted), and moreover, it keeps this property regardless of what other linear operations are performed prior to adding the cyclic prefix. Note too that a cyclic prefix is the most efficient way of attaining the phase invariance property because a necessary condition for the inverse in (12) to exist is for the precoder to introduce at least $L-1$ redundant symbols.

Theorem 3 *In (11), assume that $\tilde{P}$ factorises as $\tilde{P} = CA$ where $A \in \mathbb{C}^{(n-L+1)\times p}$ is an arbitrary linear precoder and C is the cyclic prefix matrix defined in (14). Then, the CRB R, defined in (12), depends on $\mathbf{h}$ only through the magnitude of the channel spectrum, and in particular, is invariant to the phase of the channel spectrum.*

[2]Here, "block" must be interpreted with care. In the notation of Section 4, it refers to $\tilde{P}\boldsymbol{s}$. In this sense then, a TZ-OFDM block has $L-1$ zeros at the start and at the end.

6. CONCLUSION

The standard description of a TZ-OFDM system as a CP-OFDM system with the cyclic prefix replaced by a null guard provides little insight into the performance of TZ-OFDM systems. This paper showed that the performance of TZ-OFDM systems can be understood by considering an equivalent channel coded CP-OFDM system. Since CP-OFDM systems are particularly simple to understand — they transmit each data symbol over an independent subchannel — this interpretation is believed to be an attractive way of understanding TZ-OFDM systems. Indeed, using this interpretation, this paper proved that TZ-OFDM systems spread the power of the source symbols equally across the independent sub-channels. Furthermore, this intuition led to the discovery of a channel for which a TZ-OFDM system performs worse than an uncoded CP-OFDM system. Finally, for completeness, the need for a cyclic prefix (or leading and trailing zeros) was explained in terms of being able to invert accurately non-minimum and non-maximum phase channels.

7. REFERENCES

[1] E. Biglieri, J. Proakis, and S. Shamai. Fading channels: Information-theoretic and communications aspects. *IEEE Transactions on Information Theory*, 44(6):2619–2692, October 1998.

[2] B. Le Floch, M. Alard, and C. Berrou. Coded orthogonal frequency division multiplex. *Proceedings of the IEEE*, 83(6):982–996, 1995.

[3] T. L. Marzetta and B. M. Hochwald. Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading. *IEEE Transactions on Information Theory*, 45(1):139–157, January 1999.

[4] B. Muquet, Z. Wang, G. B. Giannakis, M. de Courville, and P. Duhamel. Cyclic-prefixed or zero-padded multicarrier transmissions? *IEEE Transactions on Communications*, 2000. Submitted.

[5] A. Scaglione, G. B. Giannakis, and S. Barbarossa. Redundant filterbank precoders and equalizers: Parts I and II. *IEEE Transactions on Signal Processing*, 47:1988–2022, July 1999.

[6] I. E. Telatar and D. N. C. Tse. Capacity and mutual information of wideband multipath fading channels. *IEEE Transactions on Information Theory*, 46(4):1384–1400, July 2000.

[7] Z. Wang and G. B. Giannakis. Wireless multicarrier communications: Where Fourier meets Shannon. *IEEE Signal Processing Magazine*, 17, May 2000.

[8] S. B. Weinstein and P. M. Ebert. Data transmission by frequency division multiplexing using the discrete Fourier transform. *IEEE Transactions on Communications*, 19(5):628–634, October 1971.

FREQUENCY ESTIMATION UTILIZING THE HADAMARD TRANSFORM

Tomas Andersson, Mikael Skoglund and Peter Händel

Department of Signals, Sensors and Systems
Royal Institute of Technology
SE–100 44 Stockholm, Sweden

Abstract—Fast analog to digital conversion with only one bit per sample does not only make high sampling rates possible but also reduces the required hardware complexity. For short data buffers or block lengths, it has been shown that tone frequency estimators can be implemented by a simple table look-up. In this paper we present an analysis of such tables using the Hadamard transform. As an outcome of the analysis, we propose a class of nonlinear estimators of low complexity. Their performance is evaluated using numerical simulations. Comparisons are made with the proper Cramér–Rao bound and with the table look-up approach.

1. INTRODUCTION

Tone frequency estimation from an N-sequence

$$\{x[0],\ldots,x[N-1]\} \tag{1}$$

of noise corrupted data is a well-established research area and several estimators have been proposed during the past decades. In this paper, we consider the signal model

$$x[n] = s[n] + e[n], \qquad s[n] = A\sin(2\pi f n + \phi) \tag{2}$$

where $A > 0$ is the amplitude, ϕ the initial phase, and f is the normalized frequency, $0 < f < 1/2$, i.e. $f = F/f_s$ where F is the signal frequency and f_s is the sampling frequency. The frequency f is an unknown parameter and the phase ϕ is assumed to be uniformly distributed over the interval $[0, 2\pi]$ (and independent of other signal parameters). The noise is assumed white Gaussian with variance σ^2.

We make the assumption that the observed data $y[n]$ is a quantized version of $x[n]$ forming a *binary sequence*

$$\{y[0],\ldots,y[N-1]\} \tag{3}$$

according to

$$y[n] = \mathrm{sign}(x[n]) \tag{4}$$

This work was supported in part by the Junior Individual Grant Program of the Swedish Foundation for Strategic Research.

where $\mathrm{sign}(x) = 1$ for $x \geq 0$ and $\mathrm{sign}(x) = -1$ for $x < 0$. In an electronic circuit we would represent such binary data by ones and zeros.

We are interested in estimators that strive to estimate the true value, say f_0 (a deterministic constant), of the unknown frequency f, based on a *binary sequence* of the observed data according to (3). Our goal is to find an estimator $\hat{f} : \{\pm 1\}^N \to \mathbb{R}$, operating on the observed and quantized data and optimal in the sense of minimum mean square error (MMSE). That is, we strive to find the estimator that minimizes $\mathrm{E}[(\hat{f} - f)^2]$ subject to an assumed *a priori* distribution for the unknown frequency f. That is, the *a priori* distribution for the frequency is a design parameter of the estimator.

Because of the quantization, the number of possible different sequences (3) is finite. Hence, a particular observed sequence, of length N, can always be mapped to an index $i \in \{0,\ldots,M-1\}$, with $M = 2^N$, where we chose the mapping from an observed sequence to the index i as

$$i = \sum_{n=0}^{N-1} \frac{1 - y[n]}{2}\, 2^n. \tag{5}$$

Since there is only a finite number of possible observed sequences, there is also a finite number of possible estimator outputs. Thus any estimator can be implemented in two steps: (a) determine the index i that corresponds to the observed sequence according to (5), and (b) use this index as a pointer to an entry in a *table*

$$\{\hat{f}(0), \hat{f}(1),\ldots,\hat{f}(M-1)\} \tag{6}$$

containing all possible frequency estimates. Under the MMSE criterion we have that the table entries should be chosen as

$$\hat{f}(i) = \mathrm{E}[f \,|\, i]. \tag{7}$$

where the expectation is with respect to the assumed *a priori* distribution for f, the phase and the noise, conditioned on the observed sequence (as represented by the index i). In [1] we studied methods for computing estimator tables (6) based on (7). We also investigated the performance

0-7803-7011-2/01/$10.00 ©2001 IEEE

of the resulting MMSE estimator. As demonstrated in [1], table based frequency estimation performs well compared, e.g., with the Cramér–Rao bound for one-bit quantized data [2]. However, the size of the table grows exponentially with the block-length N, and the method is hence not feasible for block-lengths larger than, say, 24–26 samples. Our aim in the present study is therefore to investigate methods to *compress* the table, that is, characterizing the set of possible estimates $\hat{f}$ using (much) less than 2^N table entries. Our main tool in achieving such compression is the Hadamard transform, as explained next.

2. THE HADAMARD TRANSFORM

We note that any function $\gamma : \{0, \ldots, M-1\} \to \mathbb{R}$, where $M = 2^N$ and with a finite domain represented by the integers $\{0, \ldots, M-1\}$, can be expanded as

$$\gamma(i) = \mathbf{t}^{\mathsf{T}}\,\mathbf{h}(i), \quad \text{with}$$

$$\mathbf{h}(i) \triangleq \begin{bmatrix} 1 \\ y[N-1] \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 \\ y[0] \end{bmatrix} = \begin{bmatrix} 1 \\ y[0] \\ y[1] \\ y[0]y[1] \\ \vdots \\ \prod_{n=0}^{N-1} y[n] \end{bmatrix} \tag{8}$$

where $\otimes$ denotes the Kronecker matrix product and with the relation between the index i and the binary variables $\{y[n]\}$ defined as in (5). The vector $\mathbf{t}$, with elements $\{t_m\}$, is then the *Hadamard transform* of $\mathbf{g} = [\,\gamma(0) \cdots \gamma(M-1)\,]^{\mathsf{T}}$, computed as

$$\mathbf{t} = 2^{-N}\,\mathbf{H}\,\mathbf{g} \tag{9}$$

where $\mathbf{H}$ is the size $M \times M$ *Hadamard matrix*, with rows $\mathbf{h}(0)$, ..., $\mathbf{h}(M-1)$. Computing $\mathbf{t}$, as in (9), requires $\mathcal{O}(NM)$ operations [3]. We see that the representation $\gamma(i) = \mathbf{t}^T\mathbf{h}(i)$ gives the value $\gamma(i)$ in terms of the "bits" $\{y[n]\}$ of the index i. This property has proven to be of great use in synthesis and analysis of quantizers [4]. In the application studied in this paper, the finite-domain function of interest is the estimator $\hat{f}(i)$, and the binary variables $\{y[n]\}$ are the one-bit quantized data samples (4). Using (8) we conclude that the Hadamard transform can hence be employed to represent this estimator as

$$\hat{f}(i) = \mathbf{t}^{\mathsf{T}}\mathbf{h}(i) = \sum_{m=0}^{M-1} t_m h_m(i) = t_0 + t_1\,y[0] + t_2\,y[1] + t_3\,y[0]y[1] + \cdots + t_{M-1}\prod_{n=0}^{N-1} y[n]. \tag{10}$$

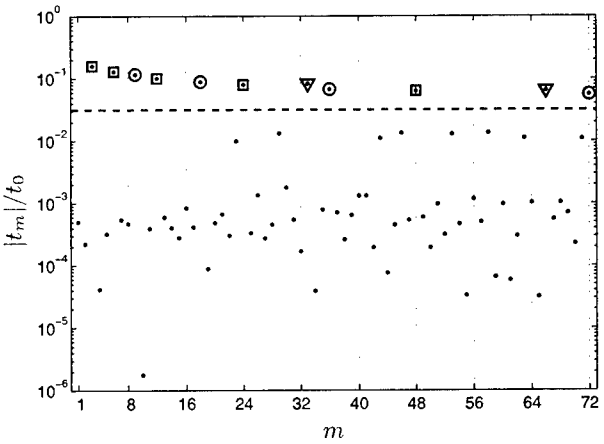


Fig. 1. The normalized magnitudes of the first 72 t-coefficients in (10) for a fixed table of size $M = 2^{16}$. The coefficients above the dashed line correspond to weight two binary products of neighboring samples (□), samples at distance 3 (○) and 5 (▽), respectively.

That is, $\hat{f}$ can be represented in terms of the transform coefficients $\{t_m\}$ and *all possible different products that can be formed using the variables* $\{y[n]\}$. For a given $\hat{f}(i)$ the coefficients $\{t_m\}$ (the t-coefficients, for short) are calculated via the Hadamard transform. It is important to note that the representation (10) is exact.

We aim to use (10) as a basis for reducing the number of parameters needed in implementing the estimator $\hat{f}$, noting that the t-coefficients completely define $\hat{f}$. However, since there are M different t_m nothing is gained by using (10) to implement the estimator (on the contrary there is a loss in computational complexity since the sum in (10) needs to be calculated, while a table look-up implementation based on (6) basically requires no computation at all). It is reasonable, however, to assume that not all of the t-coefficients are significant (in the sense that some of them are zero or close-to zero). Hence, if we can identify the t-coefficients that are most significant we need only to store these and then use (10) (setting "insignificant" coefficients to zero) to compute an approximate estimate. Compared with using a table look-up implementation we can hence use such an approach to trade storage complexity for computations.

3. TABLE ANALYSIS

Consider a known table used in a table look-up frequency estimator (6), say $\mathbf{g}$. That is

$$\mathbf{g} = \left[\hat{f}(0), \hat{f}(1), \ldots, \hat{f}(M-1)\right]^{\mathsf{T}}. \tag{11}$$

The table entries in $\mathbf{g}$ can be expressed as a function of the corresponding t-coefficients and a binary representation of the entry index, as in (10). To illustrate the structure of

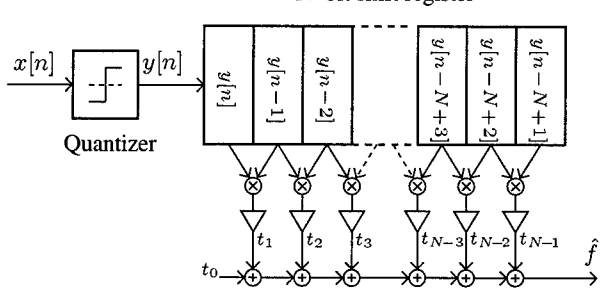


Fig. 2. A proposed estimator where neighboring binary products of weight two are used (type-$\mathcal{A}$).

the t-coefficients we use a table (11) trained at SNR $= A^2/(2\sigma^2) = 20$ dB and for a block-length $N = 16$, according to [1]. The t-coefficients for this table are computed, as in (9), and their normalized magnitudes $|t_i|/t_0$ are displayed in Figure 1. We see that there exist coefficients that are significantly larger in magnitude than the rest (marked in Figure 1 above the dashed line). In further analyzing the t-coefficients we note that all the dominant t-coefficients correspond to a weight two product in the sum (10), i.e t_3 is multiplied with the product $y[0]y[1]$ and t_6 is multiplied with $y[1]y[2]$ and so forth. Further we can divide the dominant t-coefficients into two sets:

- $\mathcal{A}$) t-coefficients that correspond to a weight two product of neighboring samples, for example t_{12} corresponding to the product $y[2]y[3]$, or t_{24} corresponding to the product $y[3]y[4]$.
- $\mathcal{B}$) t-coefficients that correspond to a weight two product of samples separated by a distance of an even number of samples. The set $\mathcal{B}$ is exemplified by t_9 corresponding to the binary product $y[0]y[3]$, or t_{33} corresponding to $y[0]y[5]$.

The coefficient t_0 is included in both sets. Neighboring samples are separated by a zero distance, hence set $\mathcal{A}$ is a subset of $\mathcal{B}$. Using one of the sets $\mathcal{A}$ or $\mathcal{B}$ we can form an approximation of each entry in the true g and build a table estimate $\hat{\mathbf{g}}$. By calculating an entry estimate only when needed, fewer coefficients need to be stored. Accordingly, the memory complexity is reduced from storing the entire table with 2^N coefficients to N or $N^2/4+1$ using set $\mathcal{A}$ or $\mathcal{B}$, respectively. That is, a reduction from an exponential to a polynomial relation between the block length and the number of coefficients. A block diagram of a type-$\mathcal{A}$ estimator is given in Figure 2.

4. ESTIMATOR DESIGN

In was shown above how to form an approximation of each table entry using a reduced set of t-coefficients. Calculating the entire set of t-coefficients requires storage of the full table g. This is not feasible for, say, $N > 26$. The structure of the approximate estimator is, however, independent of N. Here, we use the structure of such an estimator and calculate the corresponding reduced set of coefficients under the MMSE criterion.

Let $\tilde{\mathbf{h}}_{\mathcal{A}}(i)$ and $\tilde{\mathbf{h}}_{\mathcal{B}}(i)$ denote vectors containing the signal products in (10) corresponding to the t-coefficients in the sets $\mathcal{A}$ and $\mathcal{B}$, respectively. That is,

$$\tilde{\mathbf{h}}_{\mathcal{A}}(i) = \begin{bmatrix} 1 \\ y[0]y[1] \\ y[1]y[2] \\ \vdots \end{bmatrix}, \qquad \tilde{\mathbf{h}}_{\mathcal{B}}(i) = \begin{bmatrix} 1 \\ y[0]y[1] \\ y[1]y[2] \\ y[0]y[3] \\ \vdots \end{bmatrix} \tag{12}$$

where the relation between the index i and the sequence $y[n]$ is given by (5). We denote the corresponding vectors with t-coefficients by $\tilde{\mathbf{t}}_{\mathcal{A}}$ and $\tilde{\mathbf{t}}_{\mathcal{B}}$, respectively. We can now formulate two corresponding frequency estimators as

$$\hat{f}_{\mathcal{A}}(i) = \tilde{\mathbf{t}}_{\mathcal{A}}^{\mathrm{T}}\tilde{\mathbf{h}}_{\mathcal{A}}(i), \qquad \hat{f}_{\mathcal{B}}(i) = \tilde{\mathbf{t}}_{\mathcal{B}}^{\mathrm{T}}\tilde{\mathbf{h}}_{\mathcal{B}}(i). \tag{13}$$

In order to optimize the performance of the estimators in (13) let $\tilde{\mathbf{t}}_k$ be a design parameter to be chosen optimally. Using the MMSE criterion $\tilde{\mathbf{t}}_k$ is given by

$$\begin{aligned} \tilde{\mathbf{t}}_k &= \arg\min_{\mathbf{a}} \mathrm{E}(f - \mathbf{a}^{\mathrm{T}}\tilde{\mathbf{h}}_k(i))^2 \\ &= (\mathrm{E}[\tilde{\mathbf{h}}_k(i)\tilde{\mathbf{h}}_k(i)^{\mathrm{T}}])^{-1}\,\mathrm{E}[\tilde{\mathbf{h}}_k(i)f] \quad k = \mathcal{A}, \mathcal{B} \end{aligned} \tag{14}$$

where the expectation is with respect to frequency f, phase ϕ and noise $e[n]$.

A feasible approach to calculate the expectations needed in (14) is by aid of Monte Carlo integration. Such a training procedure for the problem at hand is discussed in [1].

5. NUMERICAL EVALUATION

In Figure 3, the empirical mean square error (MSE) is shown as function of SNR for a data record of length $N = 16$. The performance of the estimator using the full table g in (11) is compared with using subsets of parameters, that is type-$\mathcal{A}$ and type-$\mathcal{B}$ in (13), respectively. As reference, the asymptotic ($N \to \infty$) CRB for the given signal model is included [2]. The table g in (11) is obtained using a training approach discussed in [1]. The t-coefficients $\tilde{\mathbf{t}}_{\mathcal{A}}$ and $\tilde{\mathbf{t}}_{\mathcal{B}}$ for $\hat{f}_{\mathcal{A}}(i)$ and $\hat{f}_{\mathcal{B}}(i)$ are calculated according to (14) using Monte Carlo integration at SNR $= 20$ dB. The *a priori* distribution of f is chosen as a uniform distribution on the interval $[\varepsilon, 0.5-\varepsilon]$ where ε is a design parameter and has been set to $\varepsilon = 0.04$. Our experience indicates that a smaller value of ε typically results in a significant performance reduction while a larger value does not appear to influence the performance negatively. In Figure 3 (as well as in Figure 4), the performance is evaluated for the signal in (2) with the true frequency $f_0 = 0.1$. Further the MSE figures are averaged

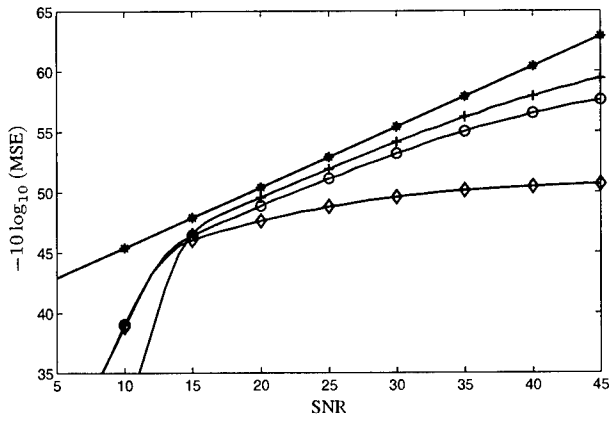


Fig. 3. Performance of the proposed estimators for $N=16$. Displayed are: CRB ($*$), table look-up estimator (6) (+), estimator $\hat{f}_{\mathcal{A}}(i)$ (o) and $\hat{f}_{\mathcal{B}}(i)$ ($\Diamond$).

over 100.000 independent trials. From Figure 3, we note a decreased performance when the complexity of the estimator is reduced. We observe further that for high SNRs the performance of (11) starts to deviate from the CRB due to a non-negligible bias term in the MSE. For $\tilde{f}_{\mathcal{A}}(i)$ and $\tilde{f}_{\mathcal{B}}(i)$ the bias is even more significant.

The experiment is repeated in Figure 4, but now for $N=64$. In this case, it is not feasible to implement (11) and it is therefore excluded from the comparison. From the figure, we note that the performance of $\hat{f}_{\mathcal{B}}(i)$ almost coincides with the asymptotic CRB for all SNRs above a threshold at about 15 dB. We also note that the difference in performance between $\hat{f}_{\mathcal{A}}(i)$ and $\hat{f}_{\mathcal{B}}(i)$ is negligible for low SNRs. At high SNRs the difference is more significant.

In Figure 5, the empirical MSE is shown as a function of the unknown signal frequency f_0 at a fixed SNR $=20$ dB and block length $N=64$. As a reference the asymptotic CRB is displayed. We observe that both the estimators, $\hat{f}_{\mathcal{A}}(i)$ and $\hat{f}_{\mathcal{B}}(i)$ performs well, except at frequencies near 0 or 0.5, and that the difference in performance between them is negligible. However, the performance is dependent on the unknown signal frequency f_0 and for some isolated frequencies the performance is significantly deteriorated.

6. SUMMARY AND CONCLUSIONS

We have shown that the table based approach used as a frequency estimator in [1] can be transformed using the Hadamard transform to an equivalent representation based on a sum over binary products and a set of coefficients. We have also investigated how the set of coefficients can be reduced, and how such reduction makes it possible to handle large blocks of data. We furthermore showed how the remaining coefficients can be optimized to increase the performance of the estimator with reduced complexity. The performance of the new estimators was then evaluated by aid of simulations and their performance was compared with the appropriate Cramér–Rao bound. The simulations indicated that the considered methods are able to produce nearly statistically efficient estimates.

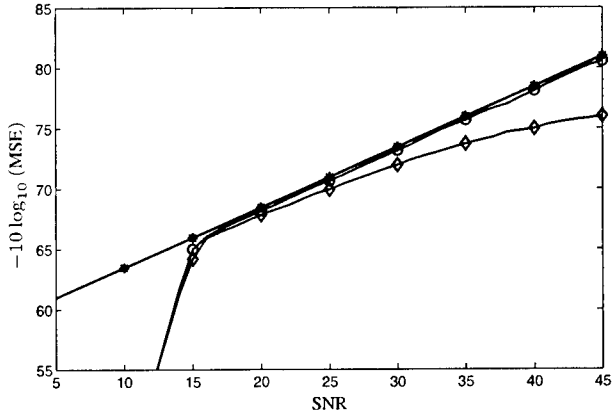


Fig. 4. Performance of the proposed estimators for $N=64$. Displayed are: CRB ($*$), estimator $\hat{f}_{\mathcal{A}}(i)$ (o) and $\hat{f}_{\mathcal{A}}(i)$ ($\Diamond$).

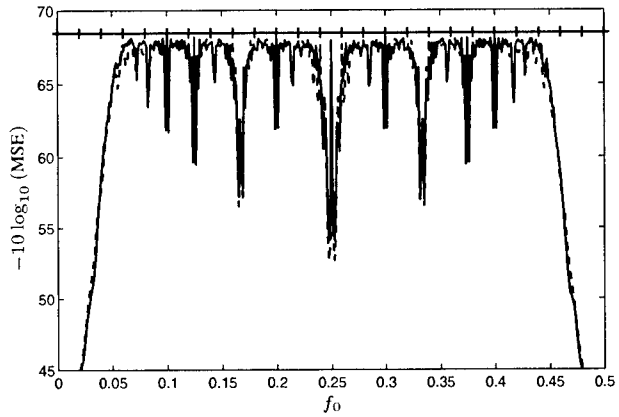


Fig. 5. Emperical MSE as a function of the frequency f for $N=64$ and SNR $=20$ dB. Displayed are: CRB (+), $\hat{f}_{\mathcal{A}}(i)$ (- -) and $\hat{f}_{\mathcal{B}}(i)$ (-).

REFERENCES

[1] T. Andersson, M. Skoglund, and P. Händel, "Frequency estimation by 1-bit quantization and table look-up processing," in *Proc. EUSIPCO*, Finland, pp. 1807–1810, Sep. 2000.

[2] A. Høst-Madsen and P. Händel, "Effects of sampling and quantization on single tone frequency estimation," *IEEE Trans. Signal Processing*, Vol. 48, No. 3, pp. 650-662, 2000.

[3] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[4] P. Hedelin, P. Knagenhjelm, and M. Skoglund, "Theory for transmission of vector quantization data," in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 10, pp. 347–396. Elsevier Science, 1995.

BEST QUADRATIC UNBIASED ESTIMATOR (BQUE) FOR TIMING AND FREQUENCY SYNCHRONIZATION

Javier Villares and Gregori Vázquez

Department of Signal Theory and Communications, Polytechnic University of Catalonia
UPC Campus Nord - Mòdul D5, c/Jordi Girona 1-3, 08034 Barcelona (Spain)
e-mail:{javi,gregori}@gps.tsc.upc.es

ABSTRACT

This paper deals with the optimal design of quadratic Non-Data-Aided (NDA) open- and closed-loop estimators. The new approach supplies the minimum variance, unbiased NDA quadratic estimators, without the need of assuming a given statistics for the nuisance parameters, that is, avoiding the common adoption of the gaussian assumption, which does not apply in digital communications. Alternatively, if the unbiased constraint is relaxed, a bayesian open-loop estimator is presented and its performance compared with the open-loop $BQUE$ solution. On the other hand, the closed-loop $BQUE$ is developed, showing that it outperforms the well-known 'ad hoc' Gaussian Stochastic Maximum Likelihood ($GSML$) scheme for short observation windows, that is, for low-complexity implementations, only converging to the $UCRBG$ asymptotically. Finally, the quadratic analysis is naturally extended to higher-order techniques which exhibit a better performance for high $SNRs$.

1. INTRODUCTION

Non-Data-Aided (NDA) synchronization has received lately a lot of attention because it simplifies the system protocols and makes unnecessary the transmission of training sequences (preambles) that reduce significantly the spectral efficiency.

So far, most algorithms have been devised by heuristic reasoning. In [1] [2] the authors presented a general framework that allowed the formulation of any NDA synchronizer based on second order moments under a Maximum Likelihood (ML) perspective. There are two basic reasons for limiting the analysis to quadratic synchronizers. It is shown [1] that the stochastic ML solution becomes quadratic for low $SNRs$, being still unknown for moderate or high SNRs because the difficult treatment of the unknown transmitted symbols and, moreover, they allow efficient digital implementations.

In this paper, we propose a totally different approach to the design of NDA discriminators that does not have to cope with the symbols extraction problem as in [1]. Making use of basic concepts from the estimation theory [3], we have deduced the Best Quadratic Unbiased Estimator ($BQUE$), that is, an estimator of the desired parameter that is quadratic, unbiased and has minimum variance. Its name has been chosen by analogy with the classical Best Linear Unbiased Estimator ($BLUE$) [3]. The $BQUE$ approach allows to unify the design of open- and closed-loop synchronizers by constraining the value and/or slope of certain points of the S-curve and optimizing its performance within the expected operation range.

The structure of the paper is the following. The signal model and the problem statement are presented in Section 2. Section 3 introduces the algebraic notation used through the paper. Section 4 deduces the exact solution to the open-loop $BQUE$ and closed expressions are obtained with the help of a discrete approximation. In Section 5 the non-bias constraint is lifted and an alternative open-loop bayesian discriminator having minimum mean squared error is presented. Section 6 deduces the closed-loop $BQUE$ for tracking. Section 7 compares the $BQUE$ and Gaussian Stochastic Maximum-Likelihood ($GSML$) feedback detectors with the Gaussian Unconditional Cramer-Rao Bound ($UCRBG$). Section 8 extends the results to higher-order estimators. Simulations results and their comments can be found in Section 9 and, finally, conclusions are drawn in Section 10.

2. DISCRETE-TIME SIGNAL MODEL

A lot of problems in the signal processing field can be unified using the following signal model:

$$\mathbf{r} = \mathbf{A}_\lambda \cdot \mathbf{x} + \mathbf{w} \tag{1}$$

where $\mathbf{r}$ is a vector containing N samples of the received signal (N_{ss} samples per symbol), λ is the parameter to estimate (for instance, the timing error or frequency error) embedded in the transfer matrix $\mathbf{A}_\lambda$, $\mathbf{x}$ is the vector of transmitted symbols (unknown in a NDA scheme), $\mathbf{w}$ is the vector of noise samples with covariance matrix $\mathbf{R_w} = E\{\mathbf{w}\mathbf{w}^H\}$.

3. NOTATION

The following notation is introduced here to facilitate the deduction of the proposed BQU estimators:

$$\begin{aligned}
&\widehat{\mathbf{R}}_\lambda = \mathbf{r}\mathbf{r}^H \\
&\mathbf{R}_\lambda = E\left\{\widehat{\mathbf{R}}_\lambda\right\} = \mathbf{A}_\lambda \mathbf{A}_\lambda{}^H + \mathbf{R_w} \\
&\mathbf{S}_\lambda = \tfrac{\partial}{\partial\lambda}\mathbf{R}_\lambda = \mathbf{D}_\lambda \mathbf{A}_\lambda{}^H + \mathbf{A}_\lambda \mathbf{D}_\lambda{}^H \;\; ; \;\; \mathbf{D}_\lambda = \tfrac{\partial}{\partial\lambda}\mathbf{A}_\lambda \\
&\widehat{\mathcal{R}}_\lambda = vec(\widehat{\mathbf{R}}_\lambda) \;\; ; \;\; \mathcal{R}_\lambda = vec(\mathbf{R}_\lambda) \;\; ; \;\; \mathcal{S}_\lambda = vec(\mathbf{S}_\lambda)
\end{aligned} \tag{2}$$

This work has been supported by: TIC98-0703, TIC99-0849 (CICYT) and CIRIT/Generalitat de Catalunya 2000SGR 00083.

0-7803-7011-2/01/$10.00 ©2001 IEEE

where $\widehat{\mathbf{R}}_\lambda$ is the sample covariance matrix and the operator $vec(\mathbf{M})$ stacks the columns of $\mathbf{M}$ in the column vector $\mathcal{M}$. Using the notation defined in the previous section, the generic equation of a quadratic discriminator is:

$$\hat{\lambda} = \mathbf{r}^H \mathbf{M} \mathbf{r} = Tr(\mathbf{M}\widehat{\mathbf{R}}_\lambda) = \mathcal{M}^H \widehat{R}_\lambda \tag{3}$$

where $Tr(\cdot)$ is the trace operator, $\mathbf{M}$ is the matrix containing the discriminator coefficients (complex) and $\mathcal{M}$ is defined as:

$$\mathcal{M} = vec(\mathbf{M}^H) \tag{4}$$

4. OPEN-LOOP (OL) BEST QUADRATIC UNBIASED ESTIMATOR (OL-BQUE)

The minimum variance unbiased (MVU) estimator $\mathcal{M}$ if the received parameter is λ has this expression:

$$\mathcal{M} = argmin_{\mathcal{M}} E\left\{|\hat{\lambda}|^2\right\} \tag{5}$$

subject to $\mathcal{M}^H \widehat{\mathcal{R}}_\lambda = \lambda$ and, thus, the cost function to minimize is the following:

$$C(\lambda) = \mathcal{M}^H \mathbf{Q}_\lambda \mathcal{M} + \left(\mathcal{M}^H \mathcal{R}_\lambda - \lambda\right) \widetilde{\alpha}_\lambda \tag{6}$$

where $\mathbf{Q}_\lambda = E\left\{\widehat{\mathcal{R}}_\lambda \widehat{\mathcal{R}}_\lambda^H\right\}$ and the Lagrange multiplier $\widetilde{\alpha}_\lambda$ impose the non-bias constraint $\mathcal{M}^H \mathcal{R}_\lambda = \lambda$. The fourth-order moments contained in $\mathbf{Q}_\lambda$ can be computed as follows for any symmetric constellation ($E\{x_i^n\} = 0\ \forall i$ if n is odd) with uncorrelated and identically distributed symbols ($E\{x_i x_j^*\} = E\{x_i x_j\} = 0\ \forall i \neq j$):

$$\begin{aligned} &Q_\lambda(Ni+j, Nk+l) = E\left\{r_i r_j^* r_k r_l^*\right\} = R_{ij}R_{kl} + R_{il}R_{kj} + \\ &+ (\mu_4 - 2\mu_2) \cdot (\mathbf{a_i} \odot \mathbf{a_j^*})(\mathbf{a_k^*} \odot \mathbf{a_l})^H \\ &+ (\mathbf{a_i} \odot \mathbf{p})\,\mathbf{a_k^T}\,(\mathbf{a_j^*} \odot \mathbf{p})\,\mathbf{a_l^H} - (\mathbf{a_i} \odot \mathbf{a_j^*} \odot \mathbf{p} \odot \mathbf{p})(\mathbf{a_k^*} \odot \mathbf{a_l})^H \end{aligned} \tag{7}$$

where R_{ij} is the element (i,j) of $\mathbf{R}_\lambda$, $\mathbf{a_n}$ the n-th row of $\mathbf{A}_\lambda$, $\mu_4 = E\{|x_n|^4\}$ and $\mu_2 = E\{|x_n|^2\}$ $(\forall n)$, $p_n = E\{x_n^2\} = E\{(x_n^*)^2\}$ the n-th element of the row vector $\mathbf{p}$ and $\odot$ stands for the Hadamard product of matrices.

Any statistical a priori knowledge of the parameter of interest $\lambda \in \Lambda = \{|\lambda| \leq \Delta_\lambda\}$ can be introduced in the optimization process by means of a bayesian approach, that is, by averaging the cost function in (6) with respect the adopted prior:

$$\begin{aligned} C_{ol} &= E_\lambda\{C(\lambda)\} = \textstyle\int_\Lambda C(\lambda) W_\lambda\, d\lambda \equiv \\ &\equiv \mathcal{M}^H \left(\textstyle\int_\Lambda \mathbf{Q}_\lambda W_\lambda\, d\lambda\right) \mathcal{M} + \mathcal{M}^H \textstyle\int_\Lambda \mathcal{R}_\lambda \alpha_\lambda d\lambda \end{aligned} \tag{8}$$

where $\alpha_\lambda = W_\lambda \widetilde{\alpha}_\lambda$ and all the irrelevant terms have been wiped out from the last equation. The weighting function $W_\lambda = f_\lambda(\lambda)$ is the prior of the parameter of interest. If no a priori knowledge is available, then, a uniform prior shall be adopted within the operative range Λ, that is, $W_\lambda = \frac{1}{2\Delta_\lambda}$.

The solution of (8) has the following expression:

$$\mathcal{M} = \mathbf{Q}^{-1} \overline{\mathcal{R}} \tag{9}$$

with $\mathbf{Q} = \int_\Lambda \mathbf{Q}_\lambda W_\lambda\, d\lambda$ and $\overline{\mathcal{R}} = \int_\Lambda \mathcal{R}_\lambda \alpha_\lambda d\lambda$. The value of the multipliers α_λ $(\forall \lambda \in \Lambda)$ that force the unbiased solution within the whole interval Λ, requires the solution to the following system of integral equations:

$$\mathcal{M}^H \mathcal{R}_\lambda = \overline{\mathcal{R}}^H \mathbf{Q}^{-1} \mathcal{R}_\lambda = \int_\Lambda \mathcal{R}_\nu^H \alpha_\nu^* d\nu \cdot \mathbf{Q}^{-1} \mathcal{R}_\lambda = \lambda \tag{10}$$

At that point, we have opted for a discrete approximation of the integral in (10) considering only a finite set of constraints $\Lambda_s = [\lambda_1, \cdots, \lambda_L]^T$. Thus, we obtain that $\overline{\mathcal{R}} \approx \mathbf{R}_s \alpha$ and (10) becomes $\mathbf{R}_s^H \mathbf{Q}^{-1} \mathbf{R}_s \alpha = \Lambda_s$ (after some manipulations) incorporating the definitions below:

$$\mathbf{R}_s = [\mathcal{R}_{\lambda_1}, \cdots, \mathcal{R}_{\lambda_L}] \quad \alpha = [\alpha_{\lambda_1}, \cdots, \alpha_{\lambda_L}]^T \tag{11}$$

The discrete approximation of the solution is therefore:

$$\mathcal{M} = \mathbf{Q}^{-1}(\mathbf{R}_s \alpha) = \mathbf{Q}^{-1} \mathbf{R}_s \left(\mathbf{R}_s^H \mathbf{Q}^{-1} \mathbf{R}_s\right)^{\#} \Lambda_s \tag{12}$$

It turns out that the matrix $\mathbf{R_s}^H \mathbf{Q}^{-1} \mathbf{R_s}$ in (12) can be singular if the oversampling (N_{ss}) and the length of $\mathbf{r}$ are not large enough. In that case, the set of constraints α cannot be exactly fulfilled and the pseudo-inverse operator $(\cdot)^{\#}$ will supply the least-squares fitting.

5. UNCONSTRAINED OPEN-LOOP BAYESIAN ESTIMATOR (OL-BAYES)

In this section we present an alternative criterion to design open-loop estimators from a bayesian approach when relaxing the unbiased constraint. The discriminator coefficients $\mathcal{M}$ will be those that minimize the following cost function:

$$C_{uol} = E_\lambda\left\{E\|\mathcal{M}^H \widehat{\mathcal{R}} - \lambda\|^2\right\} \equiv \mathcal{M}^H \mathbf{Q} \mathcal{M} - \mathcal{M}^H \widetilde{\mathcal{R}} \tag{13}$$

where the last equivalence only conserves the terms dependent on $\mathcal{M}^H$ and where $\widetilde{\mathcal{R}} = \int_\Lambda \mathcal{R}_\lambda W_\lambda \lambda d\lambda$ and $\mathbf{Q}$ was introduced in (8).

The expression of the discriminator $\mathcal{M}$ minimizing (14) is therefore:

$$\mathcal{M} = \mathbf{Q}^{-1} \widetilde{\mathcal{R}} \tag{14}$$

and both $\widetilde{\mathcal{R}}$ and $\mathbf{Q}$ admit analytical solutions for uniform priors, i.e., $W_\lambda = \frac{1}{2\Delta_\lambda}$.

6. CLOSED-LOOP (CL) BEST QUADRATIC UNBIASED ESTIMATOR (CL-BQUE)

In this section the estimator is required to track the parameter fluctuations around λ=0 with minimum variance for a given loop bandwidth, i.e., for a given value of the S-curve slope at λ=0. The pretended BQU discriminator of the parameter errors will be that minimizing the following cost function:

$$C_{cl} = \mathcal{M}^H \mathbf{Q}_0 \mathcal{M} + \mathcal{M}^H \mathcal{R}_0 \cdot \alpha_0 + \left(\mathcal{M}^H \mathcal{S}_0 - 1\right) \cdot \beta_0 \tag{15}$$

where the Lagrange multipliers α_0 and β_0 impose the non-bias constraints $\mathcal{M}^H \mathcal{R}_0 = 0$ and $\mathcal{M}^H \mathcal{S}_0 = 1$ at λ=0.

It can be shown that the constraint α_0 is always fitted due to the discriminator symmetry and the $CL-BQUE$ solution reduces to:

$$\mathcal{M} = \beta_0 \cdot \mathbf{Q}_0^{-1} \mathcal{S}_0 = \frac{\mathbf{Q}_0^{-1} \mathcal{S}_0}{\mathcal{S}_0^H \mathbf{Q}_0^{-1} \mathcal{S}_0} \tag{16}$$

and its tracking error variance is therefore:

$$\sigma_{\hat{\lambda}}^2 = \frac{1}{\mathcal{S}_0^H \mathbf{Q}_0^{-1} \mathcal{S}_0} \tag{17}$$

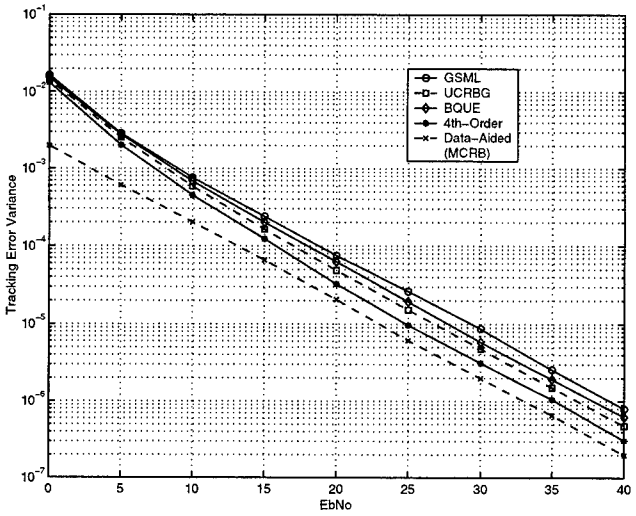


Figure 1: Normalized timing variance (σ_λ^2/T^2) for the $GSML$ and $CL-BQUE$ discriminators N=4. The fourth-order discriminator proposed in reference [4] is also plotted and compared with the $MCRB$ [5].

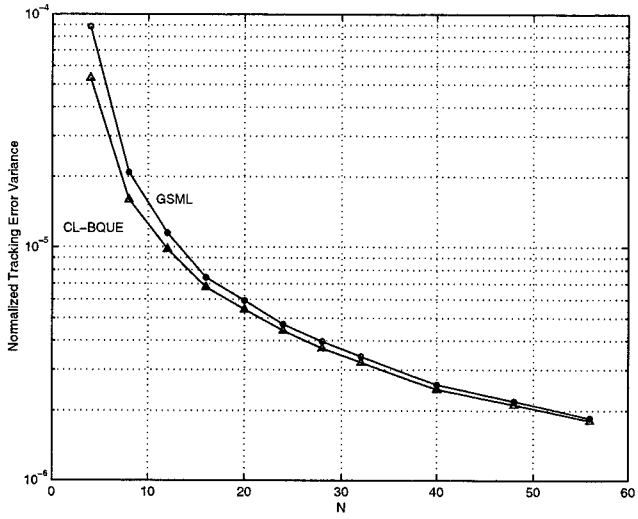


Figure 2: Normalized timing variance as a function of N for the $GSML$ and $CL-BQUE$ discriminators ($EbNo$=40dB).

7. CL-BQUE VS. MAXIMUM LIKELIHOOD APPROACH

Other suitable approach to the problem is to treat the received data ($\mathbf{r}$) as if they were gaussian. The resulting discriminator, called Gaussian Stochastic Maximum Likelihood ($GSML$), has the following structure in the uniparametric case [2]:

$$\hat{\lambda} = \frac{Tr(\mathbf{R}_0^{-1}\mathbf{S}_0\mathbf{R}_0^{-1}\widehat{\mathbf{R}}_\lambda)}{Tr\left[(\mathbf{R}_0^{-1}\mathbf{S}_0)^2\right]} \tag{18}$$

This discriminator is known to attain the Gaussian Unconditional Cramer-Rao Bound ($UCRBG$) if $N \to \infty$ [1].

In Section 9 simulations have shown that the $CL-BQUE$ (Section 6) becomes asymptotically equivalent to the $GSML$ and, hence, both attain the $UCRBG$ (asymptotically). However, when the length of $\mathbf{r}$ is short, the variance of the $CL-BQUE$ is below that of the $GSML$ and the $UCRBG$ is not attained. This fact confirms the $UCRBG$ as a suitable bound for quadratic (unbiased) NDA discriminators but it also proves that it can not be attained if the observation window is not large enough. In that case, the performance of the $CL-BQUE$ (17) can be used as a tighter bound valid for any quadratic unbiased NDA discriminator.

8. EXTENSION TO HIGHER-ORDER DISCRIMINATORS

In this section the procedure for designing optimal open- and closed-loop estimators is extended to n-order discriminators (with n>2).

For the n-order case, equation (3) can be rewritten as follows:

$$\hat{\lambda} = \mathcal{M}^H \widehat{\mathcal{R}}^{(n)} \tag{19}$$

where

$$\widehat{\mathcal{R}}^{(n)} = \widehat{\mathcal{R}} \otimes \widehat{\mathcal{R}}^{(n-2)} \quad n > 2 \text{ (even)} \tag{20}$$

and $\otimes$ stands for the Kronecker product of matrices.

Odd values of n are not considered because all modulation schemes in practice are symmetric with respect to the origin and so the odd moments are always null.

All the expressions included in previous sections are then usable if the following substitutions are done previously:

$$\begin{aligned}
&\widehat{\mathcal{R}} \longrightarrow \widehat{\mathcal{R}}^{(n)} \\
&\mathcal{R}_\lambda \longrightarrow \mathcal{R}_\lambda^{(n)} = E\left\{\widehat{\mathcal{R}}^{(n)}\right\} \\
&\mathcal{S}_\lambda \longrightarrow \mathcal{S}_\lambda^{(n)} = \tfrac{\partial}{\partial\lambda} R_\lambda^{(n)} \\
&Q_\lambda \longrightarrow Q_\lambda^{(n)} = E\left\{\mathcal{R}_\lambda^{(n)} \left(\mathcal{R}_\lambda^{(n)}\right)^H\right\}
\end{aligned} \tag{21}$$

Generally, the complexity and the minor improvement reflected in the system BER with respect to quadratic algorithms, do not justify the utilization of higher-order synchronizers in communications systems. However, when the purpose is not strictly synchronization, but the exact estimation and/or tracking of the time of arrival and/or the Doppler offset of the incoming signal, which is the case of advanced navigation and positioning systems ($DGPS$, $GNSS$, etc.), they could be taken into account despite their complexity. In any case, the higher-order study herein is valuable because it supplies new bounds that give information on the potential improvement these techniques can yield (Figure 1).

9. SIMULATION RESULTS

The simulations have been carried out for the MSK (Minimum Shift Keying) modulation as a particular case of the binary Continuous Phase Modulations CPM. This transmission scheme is adopted because it allows a simple extension to any linear digital modulation and to any multiple access modulation, as well [6]. Recall that the Laurent expansion [7] [5] allows the formulation of binary CPM signals in terms of the model presented in Section 2. The simulations have been done for additive white gaussian noise (AWGN) and two samples per symbol ($N_{ss} = 2$).

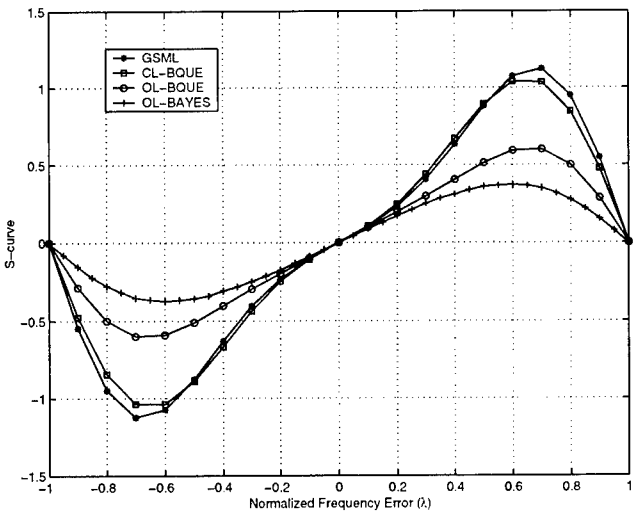


Figure 3: Frequency S-curves for the $CL-BQUE$, $GSML$, $OL-BQUE$ and $OL-BAYES$ for $\Delta_\lambda = 0.5$ and N=4.

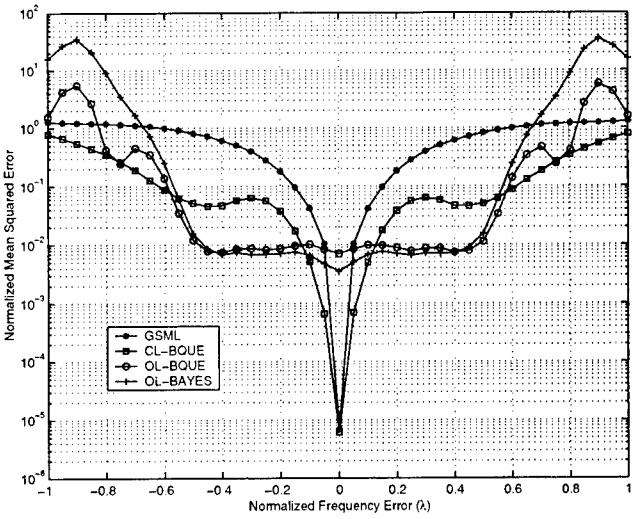


Figure 4: Normalized mean square error as a function of λ for the $CL-BQUE$, $GSML$, $OL-BQUE$ and $OL-BAYES$ with Δ_λ=0.5. N=16 and N_{ss}=4

- Figure 1: the $GSML$ and $CL-BQUE$ are compared with the $UCRBG$ when the vector of samples $\mathbf{r}$ is short (N=4). It is clear that in this case the variance of both discriminators is above the $UCRBG$. The discrepancies are more important for high $EbNo$ because the gaussian assumption becomes more exact as the $EbNo$ is reduced. Figure 1 also shows how fourth-order detectors [4] are capable to be below the $UCRBG$ and nearer the DA performance. It is also remarkable that for low $SNRs$ the $UCRBG$ becomes a valid bound for the variance of any NDA detector irrespective of its order.

- Figure 2: the asymptotic convergence of the $CL-BQUE$ and the $GSML$ is shown. The variance depicted in the figure is for an open-loop configuration when $\lambda \simeq 0$ (also Fig. 4). The corresponding closed-loop tracking variance is approximately $L_0 = 0.5/B_nT$ times lower if the normalized loop bandwidth (B_nT) is very small and $L_0 \gg N$.

- Figure 3 and 4: the two open-loop schemes proposed in the paper (Sections 4 and 5) are compared. Figure 4 shows how the $OL-BAYES$ can reduce the mean squared error within the designed interval Δ_λ (even for a extremely high $EbNo$=40 dB) because it is not forced to be unbiased (see figure 3).

The behaviour of the studied closed-loop discriminators ($GSML$ and $CL-BQUE$, Sec. 6) for $\lambda \neq 0$ is also depicted. Figures 3 and 4 show their specialization for $\lambda = 0$ (tracking). It is also notorious that the $CL-BQUE$ has a better behaviour than the $GSML$ outside the steady-state situation ($\lambda \neq 0$).

10. CONCLUSIONS

This paper presented a new, versatile approach for designing both open- and closed-loop optimal synchronizers with constraints on the S-curve shape (non-bias restrictions). If a little amount of bias is tolerated, a very simple, elegant bayesian estimator was formulated in Sec. 5 which is found to reduce the mean squared error within the designed range of the parameter.

For the closed-loop case, an optimal (unbiased) parameter error detector is obtained whose tracking variance is a lower bound for any NDA quadratic unbiased discriminator with independence of the number of samples it processes. A comparison with the classical $GSML$ is carried out and their asymptotic convergence proved empirically. However, the $BQUE$ is observed to outperform the $GSML$ for short data vectors and high $SNRs$.

Finally, the formulation of the paper is extended to higher-order synchronizers and their utilization discussed.

11. REFERENCES

[1] G. Vázquez, J.Riba. *Non-Data-Aided Digital Syncronization.* In G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, editors. *Signal Processing Advances in Wireless Communications*, volume. II: Trends in Single and Multi-User Systems, chapter. 9, pp. 357-402. Prentice-Hall, 2000.

[2] J. Riba and G. Vázquez, "Parameter Estimation of Binary CPM Signals", Proc. of ICASSP 2001, Salt Lake City (USA)

[3] Louis. L. Scharf, *Statistical Signal Processing. Detection, Estimation, and Time Analysis*, Addison Wesley, 1991.

[4] X. Villares and G. Vázquez, "Fourth-order Non-Data-Aided Synchronization", Proc. of ICASSP 2001, Salt Lake City (USA)

[5] Umberto Mengali, Aldo N. D'Andrea, *Synchronization Techniques for Digital Receivers*, Plenum Press,1997.

[6] F.Rey, G. Vázquez, J. Riba, "Joint Synchronization and Symbol Detection in Asynchronous DS-CDMA Systems". Proc. of ICASSP 2000, Istambul (Turkey)

[7] P.A. Laurent, "Exact and Approximate Construction of Digital Phase Modulations by Superposition of Amplitude Modulated Pulses", IEEE Trans. on Comm., vol. 34, Feb. 1986

AN EFFICIENT HILBERT TRANSFORM INTERPOLATION ALGORITHM FOR PEAK POSITION ESTIMATION

Saman S. Abeysekera

School of Electrical and Electronic Engineering,
Nanyang Technological University, Nanyang Avenue, SINGAPORE 639798.
E-mail: esabeysekera@ntu.edu.sg

ABSTRACT

An efficient algorithm for estimating the peak position of a sampled function is presented. The algorithm uses the Hilbert transform of the function for peak detection via interpolation. The accuracy of the proposed method is demonstrated using an example where the frequency of a sinusoid is determined by detecting the peak of the FFT of the signal. It is shown that the algorithm has computational advantage when the positions of many peaks of the sampled function are required to be estimated, e.g. as in the fundamental and harmonic frequency estimation of an audio signal. Spectral characteristics of the Hilbert transform amplitude and phase functions and the rationale for the use of Hilbert transform for interpolation are also discussed in detail.

1. INTRODUCTION

Accurate peak position estimation of a sampled function is necessary in many Digital Signal Processing applications. For example, precise time delay estimation in radar/sonar applications [1], accurate signal frequency estimation via the FFT algorithm [2], detection of R wave in ECG signals for time alignment [3] and the estimation of the position of many peaks in a time-frequency distribution [4] are some of the applications where peak detection is required in the processing. When the function is continuous in time, the peak position can be simply estimated using any existing gradient finding algorithm. In sampled signals, however, accurate peak detection becomes a computationally intensive procedure as most of the time the exact peak lies in between sample values of the function. In such cases the peak position is determined by various signal interpolation techniques which are computationally intensive. For example, frequency estimation using FFT algorithm requires many DFT calculations before sufficiently accurate frequency estimate could be obtained [2].

This paper proposes a computationally efficient algorithm for the peak detection and position estimation of a sampled function. The algorithm is based on a novel signal interpolation technique: The technique relies on the Hilbert Transform (HT) of the sampled signal which can be efficiently used to interpolate the signal in between samples. (The HT interpolation technique has been successfully used in a fractional sampling application in an array processing example in reference [5]). The accuracy of the HT based signal interpolation technique as well as the performance of the peak position estimation algorithm are discussed in the following sections.

2. HILBERT TRANSFORM INTERPOLATION

Let the sequence $x(k)$, $(k \in Z)$, has been obtained by uniformly sampling a real function $x(t)$, at sampling intervals of T, i.e. $x(k) = x(t)\big|_{t=kT}$. Consider the problem of estimating the signal value $x(t)$ at some time t given by $t = kT + \varepsilon T$ where $0 \le \varepsilon \le 1$ using the sequence $x(k)$.

Suppose z(k) is the analytic signal associated with $x(k)$, i.e.

$$z(k) = x(k) + jH\{x(k)\}\ , \qquad (1)$$

where $H\{.\}$ denotes the Hilbert transform (HT). Using equation (1) the amplitude and phase of the analytic signal can be respectively obtained as,

$$A(k) = |z(k)| \quad ; \quad \phi(k) = \arg(z(k)) \qquad (2)$$

The following points are noted:

1. In some applications, e.g. in radar/sonar and also in digital communications, the Hilbert transform of the signal is available at the receiver without the need of additional processing. This is because of the quadrature demodulation at the receiver.
2. In the absence of noise, the functions $A(k)$ and $\phi(k)$ obtained via the Hilbert transform operation, are both slowly varying. (See Appendix for a detailed discussion.)

As, $A(k)$ and $\phi(k)$ are slow varying it is possible to linearly interpolate them to obtain an estimate of the analytic signal at time $t = kT + \varepsilon T$. That is $z(k+\varepsilon)$ can be derived using the following relations:

$$|z(k+\varepsilon)| = \varepsilon A(k+1) + (1-\varepsilon)A(k) \quad ; \qquad (3)$$

$$\arg(z(k+\varepsilon)) = \varepsilon\phi(k+1) + (1-\varepsilon)\phi(k)\ . \qquad (4)$$

Once $z(k+\varepsilon)$ is known, $x(kT+\varepsilon T)$ then results from the real part of $z(k+\varepsilon)$.

Table 1 shows the results of an experiment performed to determine the accuracy of the Hilbert transform interpolation technique. The following Linear Frequency modulated signal having a Gaussian shaped envelope was used in the experiment.

$$x(t) = e^{-10t^2}\cos(2\pi\alpha t + \pi\beta t^2)\ , \qquad (5)$$

with $\alpha = 240Hz$ and $\beta = 120Hz/s$. The signal duration was selected as $-1 < t < +1$ seconds. At first a sequence $x(k)$ was obtained by sampling the signal in equation (5) by a *1 kHz* sampling frequency. Note that the signal in equation (5) occupies

0-7803-7011-2/01/$10.00 ©2001 IEEE

the full Nyquist bandwidth $\pm 500Hz$. Suppose another sequence is defined as the values obtained by sampling the signal $x(t)$ at a frequency of $d\ kHz$ ($d \in \Re$), i.e.

$$x_d(k) = x(t)\big|_{t=d} \quad . \tag{6}$$

Value of d	Maximum Error from H.T. Interpolation	Maximum Error from Linear Interpolation
0.092	7.2403×10^{-6}	4.3975
0.320	7.2402×10^{-6}	4.3975
0.900	7.2413×10^{-6}	4.3989
1.100	7.2370×10^{-6}	4.3931
3.900	10.454×10^{-6}	3.9211
13.70	11.990×10^{-6}	4.9335

TABLE 1: Comparison of Hilbert Transform Interpolation with Linear Interpolation.

(Note that the above corresponds to a sampling rate conversion of the original signal.) We can estimate the sequence $x_d(k)$ from the sequence $x(k)$ via the HT interpolation technique. The exact value of $x_d(k)$ can also be obtained using equation (5). Therefore, it is possible to calculate estimation error, and thus evaluate the performance of the HT interpolation algorithm. Table 1 shows the performance results of the Hilbert transform interpolation technique in obtaining the sequence $x_d(k)$ from sequence $x(k)$. For comparison purposes results from a Linear Interpolation algorithm is also shown in Table 1. Results from Table 1 demonstrate that sequence values $x(k)$ could be accurately estimated (within an error of 10^{-5}) using the described Hilbert transform interpolation technique.

3. PEAK DETECTION VIA THE HILBERT TARNSFORM

What is required here is to estimate the peak position of the function $x(t)$ using the sampled sequence $x(k)$. The first step is a coarse estimate; to determine the sampled interval where the peak of the function $x(t)$ would be located. This could easily be performed by detecting the peak value of the sequence $x(k)$ and then investigating the right and left neighbor samples of the peak sample.

Suppose m and $(m+1)$ denote the interval resulting from the coarse estimate. Using the relations (3) and (4), the HT interpolated function $x(t)$ within this interval can be obtained as,

$$\begin{aligned} x(mT+\varepsilon T) = &\{\varepsilon A(m+1)+(1-\varepsilon)A(m)\} \\ &\times\cos(\varepsilon\phi(m+1)+(1-\varepsilon)\phi(m)) \end{aligned} \tag{7}$$

The value of ε, which maximizes $x(mT+\varepsilon T)$ can be determined by differentiating the right hand side of equation (7) with respect to ε and equating it to zero. That is to obtain

$$\begin{aligned} &\{\varepsilon A(m+1)+(1-\varepsilon)A(m)\}\{\phi(m+1)-\phi(m)\} \\ &\quad\times\sin(\varepsilon\phi(m+1)+(1-\varepsilon)\phi(m)) = \\ &\{A(m)-A(m+1)\}\cos(\varepsilon\phi(m+1)+(1-\varepsilon)\phi(m)) \end{aligned} \tag{8}$$

Suppose γ is defined as

$$\gamma = \big(A(m+1)-A(m)\big)\big/A(m) \tag{9}$$

then equation (8) can be expressed as

$$\begin{aligned} \varepsilon = &\frac{1}{(\phi(m+1)-\phi(m))}\tan^{-1}\left(\frac{-\gamma}{1+\varepsilon\gamma}\right) \\ &-\frac{\phi(m)}{(\phi(m+1)-\phi(m))} \end{aligned} \tag{10}$$

It is then possible to numerically solve equation (10) for the value of ε. However, as $0\le\varepsilon\le1$ and $\gamma << 1$ (see appendix), $1+\varepsilon\gamma \cong 1$. Therefore, equation (8) simplifies to yield a direct solution for ε as

$$\varepsilon_0 \cong \frac{\phi(m)+\tan^{-1}(\gamma)}{\phi(m)-\phi(m+1)} \tag{11}$$

And thus the peak position of the function $x(t)$ can be obtained as $t = mT+\varepsilon_0 T$ using the amplitude $A(k)$ and phase $\phi(k)$ of the analytic signal.

4. AN EXAMPLE OF FREQUENCY ESTIMATION USING THE FFT

To demonstrate the accuracy of the above peak position estimation algorithm, consider the following example. Suppose it is required to estimate the frequency f_0 of a noisy sinusoid signal given by,

$$S(p) = e^{j2\pi f_0 p}+v(p)\big/\sigma^2 \quad 0\le p\le N-1 \tag{12}$$

where $v(p)$ is an independent identical distributed complex white Gaussian noise sequence with unit variance; σ^2 is the signal to ratio (SNR) associated with the signal. The maximum likelihood (ML) method of frequency estimation is to compute the Discrete Fourier Transform (DFT) of the signal and determine the frequency where the absolute value of the DFT is a maximum. As this is an extremely computationally intensive procedure, conventional algorithms works on the following method [2].

STEP 1: Use an N point FFT of the sampled signal as

$$x(k) = \left|\sum_{p=0}^{N-1} S(p)e^{-\frac{j2\pi pk}{N}}\right| \qquad 0\le k\le N-1 \tag{13}$$

to determine a coarse frequency estimate, i.e. determine the peak position interval. An indication of the computational load in this step can be obtained by looking into number of required multiplication operations, which is $N\log_2 N$.

STEP 2: Once the coarse estimate is obtained, a fine frequency estimate is obtained by evaluating a large number of DFTs with in the estimated coarse interval. Suppose the required frequency

estimation accuracy is f_{acc} the multiplication operations necessary to achieve this can be obtained as $N\log_2(1/Nf_{acc})$. Note that STEP 2 is computationally intensive in comparison to STEP 1. The computation speed can be greatly increased by using the HT interpolation on the sequence $x(k)$ defined in equation (13), followed by the detection of the peak position.

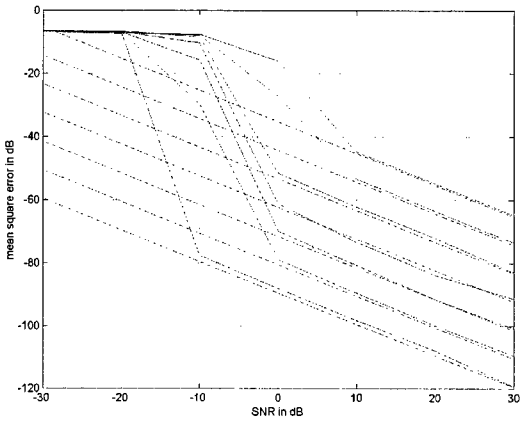


Figure 1: The mean square frequency estimation error versus the SNR, for different values of N. The value of N increases from top plot to bottom plot as *8 ,16, 32, 64, 128, 256, 512*. The HT interpolation method is used for estimating the Peak. (Dashed lines are the Cramer-Rao lower bounds.)

Figure 1 shows the frequency estimation results obtained from such an HT interpolation method. Results of Figure 1 are identical to the results reported in [2]. The results of frequency estimation in Figure 1 also achieves the Cramer-Rao bound given by,

$$CRB(snr;N)=\sqrt{\frac{3}{2\pi^2N(N+1)(N-1)snr}} \qquad (14)$$

which is the theoretically possible achievable minimum error [2]. Note that in equation (3) it was assumed that $A(k)$ is slow varying. In the presence of noise, to satisfy this condition, $x(k)$ in equation (13) was obtained using an FFT (a real valued) of length $4N$. Two Hilbert transforms at the end points of the coarse interval was then evaluated. The number of multiplication operations required for the two HT operation is $8N$. (It is assumed here that the HT can be calculated with $4N$ multiplications.)

5. A COMPARISON OF COMPUTATIONAL EFFICIENCY

From the discussion in the previous section, the number of multiplication operations required for the peak position using the HT interpolation method can be deduced as :

$$Q_{HT}=N\log_2(4N)+8N \qquad (15)$$

Using the Cramer-Rao bound in equation (14) as the required frequency accuracy, f_{acc}, the number of multiplication operations required in the method using DFTs can be determined to be:

$$Q_{DFT}=N\log_2(N)+N\log_2\left(\sqrt{\frac{2\pi^2(N+1)(N-1)snr}{3N}}\right) \qquad (16)$$

Calculating the number of multiplication operations in equations (15) and (16) for various values of N and *SNR*, it can be shown that the computational load of the proposed HT method is comparable to the DFT method in estimating the frequency of the sinusoid. However, the proposed HT method does not require the evaluation of many DFTs and is extremely efficient when a large number of peaks are required to be estimated, such as in a periodogram (time-frequency) analysis [4]. This is because the computational load in the proposed method is independent of the number of peaks in the estimate: The amount of computations in the DFT method is proportional to the number of peaks that are necessary to be estimated.

6. CONCLUSION

A technique to estimate the peak position of a sampled signal is proposed. The technique is based on the HT of the sampled signal, which can be efficiently used to interpolate the signal within sampled points. The proposed technique can be used in many engineering applications to reduce the computational load of algorithms. In a frequency estimation example it has been shown that the proposed method can reduce the computational load by a significant factor, especially when the number of peak positions that are required to be estimated is large.

7. APPENDIX: SPECTRAL CHARACTERISTICS OF FUNCTIONS $A(k)$ AND $\phi(k)$ ASSOCIATED WITH THE ANALYTIC SIGNAL

In section 2, the signal $x(k)$ has been represented using an analytic signal derived via the Hilbert transform. Spectral characteristics of the amplitude $A(k)$ and phase $\phi(k)$ of the analytic signal, in such a representation is provided in this appendix. The following discussion also provides a rationale for the selection of Hilbert transform for the interpolation.

A1. Frequency Support of $A(k)$ and $\phi(k)$:

From equations (1) and (2) we get

$$x(k)+jH\{x(k)\}=A(k)\cos(\phi(k))+jA(k)\sin(\phi(k)) \qquad (a1)$$

This requires that

$$H\{A(k)\cos(\phi(k))\}=A(k)\sin(\phi(k)) \qquad (a2)$$

To determine the necessary conditions for equation (a2) to be satisfied, consider $H\{A(k)B(k)\}$, where $B(k)=\cos(\phi(k))$. Using the convolution relation of the Discrete Time Fourier Transform (DTFT) we get

$$H\{A(k)B(k)\}=\int_{-1/2T}^{1/2T}\int_{-1/2T}^{1/2T} jS_A(\phi)S_B(f-\phi)d\phi\,\mathrm{sgn}(f)e^{j2\pi fkT}df \qquad (a3)$$

where $S_A(\phi)$ and $S_B(f)$ denote the DTFT of the signals $A(k)$ and $B(k)$, respectively; *sgn()* is the sign function. Suppose $S_A(\phi)$ and $S_B(f)$ are such that, in the frequency support regions of the product $S_A(\phi)S_B(f)$ and $\mathrm{sgn}(f+\phi)$,

$\text{sgn}(f+\phi)$ can be expressed independent of the argument ϕ, i.e.,

$$S_A(\phi)S_B(f)\text{sgn}(f+\phi) = S_A(\phi)S_B(f)I(f) \quad \text{(a4)}$$

then it can be shown that

$$H\{A(k)B(k)\} = A(k)H\{B(k)\} \quad \text{(a5)}$$

That is,

$$H\{A(k)\cos(\phi(k))\} = A(k)H\{\cos(\phi(k))\} \quad \text{(a6)}$$

The condition in equation (a4) requires that functions $S_A(f)$ and $S_B(f)$ are of low-pass and high-pass (but limited to the half sampling frequency) type, respectively, and that their spectra do not over-lap [6][7]. In other words, $S_A(f)$ and $S_B(f)$ are such that,

$$\begin{aligned} S_A(f) &= 0 \quad for \quad |f| > u \\ S_B(f) &= 0 \quad for \quad v^- < |f| < v^+ \end{aligned} \quad \text{(a7)}$$

where $0 < u \le v^-$ and $v^+ < f_s/2 = 1/2T$.

A2. Conditions on the Variation of the Amplitude Function $A(k)$:

Since $A(k)$ is band-limited to $\pm u$, as shown in equation (a7), via the use of DTFT the following expression can be obtained.

$$A(k) - A(k-1) = \int_{-u}^{u} S_A(f)[1 - e^{-j2\pi fT}]e^{j2\pi fkT}df \quad \text{(a8)}$$

Using Schwartz equality the following results:

$$\gamma = \left|\frac{A(k) - A(k-1)}{A(k)}\right| \le \left|\int_{-u}^{u} 2\sin(\pi fT)df\right| \le 4\pi\left(\frac{u}{f_s}\right)^2 \quad \text{(a9)}$$

As $u << f_s/2$, $A(k)$, is a slowly varying function and therefore a linear interpolation can be performed to estimate the signal $A(k)$ between sampling points.

A3. Conditions on the Variation of Phase Function $\phi(k)$:

From equations (a2) and (a6) we get,

$$H\{\cos(\phi(k))\} = \sin(\phi(k)) \quad \text{(a10)}$$

If the condition in equation (a10) is satisfied then the signal $B(k) = \cos(\phi(k))$ can be obtained as the real part of an analytic signal $c(k) = e^{j\phi(k)}$. Note that the instantaneous frequency of the signal $c(k)$ is given by [8],

$$(\phi(k+1) - \phi(k))/2\pi T \quad \text{(a11)}$$

Combining equations (a7) and (a11) the following can be obtained.

$$2\pi v^-/f_s < \phi(k+1) - \phi(k) < 2\pi v^+/f_s \quad \text{(a12)}$$

The phase function $\phi(k)$, therefore, is monotonic and has a positive slope. Using (a12) a condition for the variation of the second difference of $\phi(k)$ can be obtained as,

$$|\phi(k+1) - 2\phi(k) + \phi(k-1)| < 2\pi(v^+ - v^-)/f_s \quad \text{(a13)}$$

As the accuracy of linear interpolation depends on the deviation of the function from linearity, it can be noted from (a13) that if $(v^+ - v^-) << f_s/2$, the phase $\phi(k)$ can be accurately interpolated linearly.

A4. Total Signal Bandwidth and Rationale for Interpolating using Functions $A(k)$ and $\phi(k)$:

As $x(k) = A(k)B(k)$, the high frequency extent of the signal $x(k)$ is given by $f_H = v^+ + 2u$. (A more rigorous proof of this via the concepts of instantaneous frequency and instantaneous bandwidth is provided in reference [9].) Therefore, to avoid aliasing in the sampling process, it is also necessary that $v^+ + 2u \le f_s/2$. Note that the signal $x(k)$ can be directly linearly interpolated provided that $f_H = v^+ + 2u << f_s/2$. Where as the conditions for HT interpolation are such that (i) $u << f_s/2$ (for interpolating $A(k)$) and (ii) $(v^+ - v^-) << f_s/2$ (for interpolating $\phi(k)$). As noted in equation (a7) since $0 < u \le v^- < v^+ < f_H < f_s/2$, it is clear from the discussion that the conditions for HT interpolation are far less stringent than the conditions for a linear interpolation.

8. REFERENCES

[1] G. C. Carter, "Coherence and Time Delay Estimation", Proceedings of the IEEE, vol. 75, pp. 236-255, 1987.

[2] D.C. Rife and R. R. Boorstyn, "Single Tone Parameter Estimation from Discrete Time Observations", IEEE Transactions on Information Theory, IT-20, pp. 591-589, 1974.

[3] S. S. Abeysekera, "Detection and Classification of ECG Signals in the Time-Frequency Domain", Applied Signal Processing, vol. 1, pp. 35-51, Springer-Verlag London 1994.

[4] J. Laroche and Mark Dobson, "Improved Phase Vocoder Time-Scale Modification of Audio", IEEE Transactions on Speech and Audio Processing, vol. 7, no. 3, pp. 323-332, 1999.

[5] S. S. Abeysekera, "A Non-Linear Algorithm for Digital Beamforming of a Wideband Active Sonar Array", Proceedings of the IEEE- EURASIP Workshop on Non-Linear Signal and Image Processing NSIP-99, Antalya, Turkey, June 1999.

[6] E. Bedrosian, "A Product Theorem for Hilbert Transforms", Proceedings of the IEEE, vol. 51, pp. 686-689, 1963.

[7] A. H. Nuttall, "On the Quadrature Approximation to the Hilbert Transform of Modulated Signals", Proceedings of the IEEE, vol. 54, p. 1458, 1966.

[8] L. Cohen, "Time-Frequency Distributions", Proceedings of the IEEE, vol. 77, pp. 941-981, 1989.

[9] L. Cohen and C. Lee, "Instantaneous Frequency, Its Standard Deviation and Multicomponent Signals," Proceedings of the SPIE, vol. 975, pp. 186-208, 1988.

COMPUTATIONALLY EFFICIENT INTERATIVE REFINEMENT TECHNIQUES FOR POLYNOMIAL PHASE SIGNALS

Simon Sando, Dawei Huang, Tony Pettitt

Centre in Statistical Science and Industrial Mathematics,
Queensland University of Technology, GPO Box 2434, Brisbane, Australia, 4001.
E-mail: sando@fsc.qut.edu.au.

ABSTRACT

Recursive, and efficient estimation of polynomial-phase is considered here, with alternatives to the standard Gauss-Newton approach presented. We consider approximations of the likelihood and phase noise distribution to derive recursive approximate maximum likelihood and Bayesian estimators. Monte Carlo simulations indicate that these methods compare favourably with the Gauss-Newton scheme both in terms of computational expense and efficiency thresholds.

1. INTRODUCTION

Estimating the coefficients of a polynomial-phase signal in the presence of noise has arisen from many applications in signal processing [3][4][9]. Assume that we have observations

$$z_t = Ae^{i\phi_t} + \eta_t, \qquad (1)$$

where $\phi_t = \sum_{k=0}^{K} \theta_k t^k, t = 0, 1, \ldots, T-1, A$ is a real-valued constant, $\Theta = [\theta_0, ..., \theta_K]'$ is the unknown parameter vector to be estimated, and $\{\eta_t\}$ is a complex white normal sequence with mean zero and variance σ^2.

Maximum likelihood estimation of these parameters is difficult due to the many local maxima in the likelihood, and the subsequent computational expense. Recently, [1] has proposed a nonlinear least squares estimation scheme that improves the numerical properties significantly, however the computational burden remains high.

Fast methods for polynomial-phase estimation are concerned with phase unwrapping followed by regression, as in [7][11], or differencing in phase, as in [3][9]. A method to obtain efficient and direct estimation of polynomial phase signals in real time is still an open problem. However, if we have initial estimates of the parameters that are within a certain accuracy, but do not attain the Cramér-Rao Bound (CRB), we could obtain efficient estimation based on these initial values. It has long been recognised that inefficient estimation can be improved by a single step of an iterative process that leads to fully efficient estimates; see [2], section 9.2. Such procedures have been reported in frequency estimation [5][10], and a similar idea has been introduced for polynomial phase estimation [4].

The standard approach to iterative refinement is the use of the Gauss-Newton method. The main advantage is the locally quadratic convergence to the solution, however if the initialisation is not sufficiently accurate, convergence may be towards a local minimum or saddle point. Techniques exist to improve the estimation accuracy, however the majority of these involve line searching. The computational expense relating to the inversion of the Hessian matrix at each step is also a disadvantage.

In this paper, we consider alternative iterative refinement techniques for polynomial phase estimation. Firstly, a 2nd−order Taylor approximation of the likelihood equations is used to derive a recursive scheme in section 2. We then consider Bayesian approaches to recursive estimation, where we propose a 2nd−order approximation of the likelihood function, considered in

0-7803-7011-2/01/$10.00 ©2001 IEEE

terms of the phase angle, to produce a Gaussian density. This is derived in section 3. Monte Carlo simulations were then computed, and the results shown in section 4. Here it is shown that the Bayesian approach yields the best performance as far as attaining the Cramér-Rao bound is concerned, while the approximate maximum likelihood scheme is the most computationally efficient.

2. LINEARISING THE LIKELIHOOD EQUATIONS

For the signal model given in (1), the negative log-likelihood (up to additive and multiplicative constants) is

$$\begin{aligned} J &= \sum_{t=0}^{T-1} |z_t - A\exp(i\phi_t)|^2 \\ &= \sum_{t=0}^{T-1} \left\{\rho_t^2 + A^2 - 2A\rho_t\cos(y_t - \phi_t)\right\}, \end{aligned}$$

where y_t is the wrapped phase of z_t, and $\rho_t = |z_t|$. This is of course also the cost function that would be minimised in the nonlinear least squares approach to estimation. Using this representation, it is clear that we need only minimise the expression

$$J = J(\theta) = \sum_{t=0}^{T-1} \rho_t \cos(y_t - \phi_t) \tag{2}$$

with respect to ϕ_t, and hence Θ. The partial derivatives of (2) are, for each $j = 0, 1, ..., K$,

$$\begin{aligned} \frac{\partial J}{\partial \theta_j} &= \sum_{t=0}^{T-1} \rho_t t^j \sin(y_t - \phi_t) && (3) \\ &= \sum_{t=0}^{T-1} \rho_t t^j \sin(y_t + 2\pi x_t - \phi_t) && (4) \\ &\approx \sum_{t=0}^{T-1} \rho_t t^j (y_t + 2\pi x_t - \phi_t). && (5) \end{aligned}$$

where the final line is a 2nd–order Taylor approximation about zero, and we have introduced the integer process $\{x_t\}$ to account for the 2π phase ambiguity. This approximation is accurate only if we have estimated x_t correctly for each t. Setting (5) to zero for $j = 0, 1, \ldots, K$ yields a set of linear equations in the parameter vector Θ, which can be solved to yield the improved estimate. We of course do not know the sequence $\{x_t\}$ precisely, so we use an algorithm that uses this solution method but proceeds recursively through the sample. At each sample step t, if we have a reasonable estimate $\hat{\Theta}_t$ of Θ, with $\hat{\phi}_t = \sum_{k=0}^{K} \hat{\theta}_t t^k$, then

$$p\left(\left|x_t - \left[\frac{\hat{\phi}_t - y_t}{2\pi}\right]\right| > 1 \,\middle|\, y_t; \Theta^{(n)}\right) \approx 0 \tag{6}$$

when $\frac{3\sigma_v}{\sqrt{2}A} < 2\pi$, and where $[a]$ is the integer such that $|a - [a]|$ achieves the minimum. The criterion mentioned above corresponds to a signal-to-noise ratio of approximately -10dB. We then estimate x_t via

$$\hat{x}_t = \hat{x}_t(y_t, \hat{\Theta}_t) = \left[\frac{\hat{\phi}_t - y_t}{2\pi}\right], \tag{7}$$

and hence solve the equations $P_t\Theta = \mathbf{b_t}$ where,

$$P_t = \begin{bmatrix} \sum \rho_n & \sum n\rho_n & \cdots & \sum n^K\rho_n \\ \sum n\rho_n & \sum n^2\rho_n & \cdots & \sum n^{K+1}\rho_n \\ \vdots & \vdots & \ddots & \vdots \\ \sum n^K\rho_n & \sum n^{K+1}\rho_n & \cdots & \sum n^{2K}\rho_n \end{bmatrix}$$

$$\mathbf{b_t} = \begin{bmatrix} \sum \rho_n (y_n + 2\pi x_n) \\ \sum n\rho_n (y_n + 2\pi x_n) \\ \vdots \\ \sum n^K \rho_n (y_n + 2\pi x_n) \end{bmatrix}$$

where the summations are for $n = 0, 1, \ldots, t$. Using $h_t = [0, t, \ldots, t^K]'$, we obtain

$$\begin{aligned} P_t &= \sum_{n=0}^{t} \rho_n h_n h_n' && (8) \\ &= P_{t-1} + \rho_t h_t h_t' && (9) \end{aligned}$$

which, after inverting using the standard identity [8] yields

$$P_t^{-1} = P_{t-1}^{-1} - \frac{\rho_t P_{t-1}^{-1} h_t h_t' P_{t-1}^{-1}}{1 + \rho_t h_t' P_{t-1}^{-1} h_t}. \tag{10}$$

We then have the following algorithm to implement the recursive approximate maximum likelihood estimator.

a Initialise Θ and calculate P_K^{-1} and b_K.

b for t=K+1,...,T

(i) calculate P_t^{-1} from (10) and
$b_t = b_{t-1} + \rho_t(y_t + 2\pi x_t)h_t$

(ii) calculate $\Theta_t = P_t^{-1} b_t$

c end

Steps (i) and (ii) above can be calculated in parallel (with (ii) delayed) for greater speed.

3. GAUSSIAN APPROXIMATION BY 2ND-ORDER TAYLOR EXPANSION

A 2nd–order Taylor approximation of the likelihood for the tth observation, similarly to [6], yields

$$p(y_t|\Theta) \propto \exp\left(-\frac{1}{\sigma^2}\rho_t A(y_t + 2\pi x_t - \phi_t)^2\right).$$

We again estimate x_t by (7), however to robustify this estimation, and unlike the scheme proposed in section 2, we search the integers in the neighborhood of the best estimate. Define

$$G(X;\mu,\Sigma) = \frac{\exp\left(-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right)}{\sqrt{2\pi\det(\Sigma)}}.$$

Using $m = \hat{x}_t - 1, \hat{x}_t, \hat{x}_t + 1$, Bayesian formulae yield the recursion

$$\begin{aligned} p(\Theta|y_t,...,y_0) &= \frac{p(y_t|\Theta)p(\Theta|y_{t-1},...,y_0)}{\int p(y_t|\Theta)p(\Theta|y_{t-1},..,y_0)\mathrm{d}\Theta} \\ &\propto \sum_m G(h_t'\Theta; y_t + 2\pi m, S_t)G(\Theta;\hat{\Theta}_{t-1},\Sigma_{t-1}) \\ &= \sum_m W_m G(\Theta;\hat{\Theta}_{m,t},\hat{\Sigma}_t) \qquad (11) \end{aligned}$$

where, using the lemma in [6],

$$\begin{aligned} S_t &= \frac{\sigma^2}{A\rho_t} \\ \hat{\Sigma}_t &= \Sigma_{t-1} - \frac{\Sigma_{t-1}h_t h_t'\Sigma_{t-1}}{(S_t + h_t'\Sigma_{t-1}h_t)} \\ \hat{\Theta}_{m,t} &= \hat{\Theta}_{t-1} + \frac{\hat{\Sigma}_t h_t}{S_t}(y_t + 2\pi m - h_t'\hat{\Theta}_{t-1}) \\ W_m &\propto \exp\left(-\frac{(y_t + 2\pi m - h_t'\hat{\Theta}_{t-1})^2}{2(S_t + h_t'\Sigma_{t-1}h_t)}\right) \end{aligned}$$

and $\sum W_m = 1$. The number of Gaussian components in (11) will increase exponentially. To overcome this, the maximum entropy criterion [6] is used to combine the Gaussian sum into a single Gaussian pdf. The updated mean and variance estimates can then be shown to be

$$\hat{\Theta}_t = \sum_m W_m \hat{\Theta}_{m,t}, \qquad (12)$$

$$\Sigma_t = \hat{\Sigma}_t + \sum_m W_m \left(\hat{\Theta}_{m,t} - \hat{\Theta}_t\right)\left(\hat{\Theta}_{m,t} - \hat{\Theta}_t\right)' \qquad (13)$$

4. SIMULATIONS

We considered the computational and statistical efficiency of the above estimators and compared them with the Gauss-Newton method. We consider a constant amplitude chirp signal for this problem, with parameter vector $\Theta = [1.4, -0.4, 0.03]$. Figure 1 gives an indication of the performance of these estimators for varying initial accuracy and signal-to-noise ratios. The initial values were chosen as Gaussian random variables with mean equal to the true values and variance chosen such that the mean square error of the estimators was the relative efficiency prescribed.

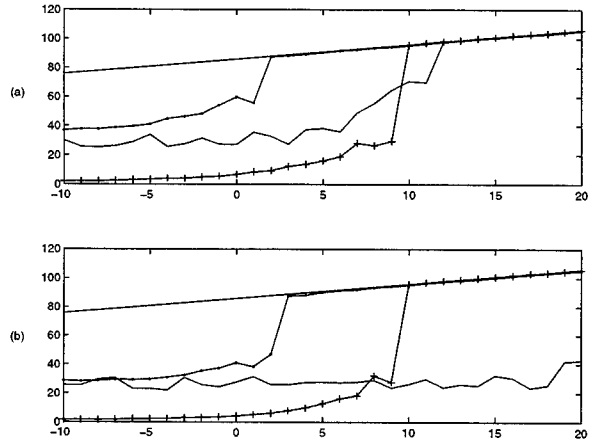


Figure 1: Performance of the Gauss-Newton (-), approximate maximum likelihood (+-) and Bayesian filtering (·-) schemes compared with the Cramér-Rao bound (solid). 1000 simulations were run, with $T = 128$ and initial values having accuracy 50% and $\frac{100}{T}$% for parts (a) and (b) respectively.

From these plots, we can clearly see the superior performance of the Bayesian scheme when compared with the approximate maximum likelihood scheme and

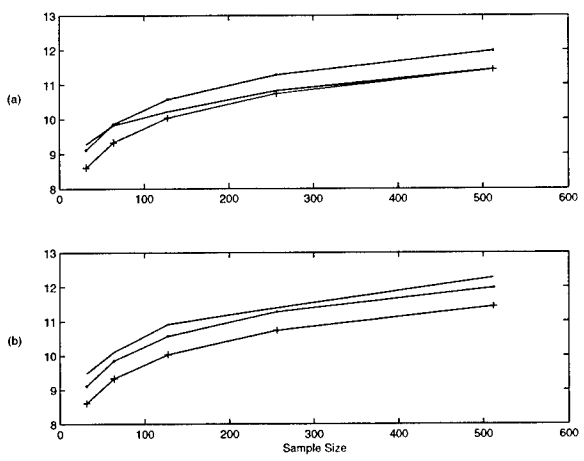


Figure 2: Number of floating point operations using Gauss-Newton (-), the approximate maximum likelihood method (+-) and the Bayesian filtering scheme (·-). 50 simulations were run, with $T = \{32, 64, 128, 256, 512\}$ and initial values having relative efficiency of 50% and $\frac{100}{T}\%$ for parts (a) and (b) respectively. The signal to noise ratio was fixed at 10dB.

the Gauss-Newton, especially when initialisation is poor. The approximate maximum likelihood approach performs slightly better than the Gauss-Newton scheme utilised, and is more robust to poor initialisation with no additional computation.

This improved performance comes at the cost of slightly more computation, as seen in figure 2. No exact theoretic analysis on computational complexity has been provided; this plot merely provides flop counts as calculated in Matlab. From these, we can see the linear performance of the approximate maximum likelihood and Bayesian methods, and that the computational requirements of the Gauss-Newton method are not significantly different. It should be noted that the Gauss-Newton approach we have taken is the fastest converging approach; the estimation performance may be improved at the expense of greater computation.

5. REFERENCES

[1] Angeby, J. Estimating Signal Parameters Using the Nonlinear Instantaneous Least Squares Approach. *IEEE Trans. on Signal Processing*, 48(10):2721–2732, 2000.

[2] Cox, D.R. and Hinkley, D.V. *Theoretical Statistics*. Chapman and Hall, London, 1974.

[3] Djuric, P. and Kay, S.M. Parameter Estimation of Chirp Signals. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 38:2118–2126, 1990.

[4] Golden, S. and Friedlander, B. Maximum Likelihood Estimation, Analysis, and Applications of Exponential Polynomial Signals. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 47(6):1493–1501, June 1999.

[5] Hannan, E.J. and Huang, D. On-Line Frequency Estimation. *Journal of Time Series Analysis*, 14(2):147–161, 1993.

[6] Huang, D. Efficient Estimation for Non-linear and Non-Gaussian State Space Models. In *Proc. of the 36th Conference on Decision and Control*, pages 5036–5041. IEEE, December 1997.

[7] Huang, D., Sando, S., and Wen, L. Least Squares Estimation of Polynomial Phase Signals via Stochastic Tree-Search. In *Proc. of ICASSP*, pages 1569–1572. IEEE, March 1999.

[8] Kay, S.M. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Upper Saddle River, NJ 07458, 1993.

[9] Peleg, S. and Porat, B. Estimation and Classification of Polynomial-Phase Signals. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37:422–428, 1991.

[10] Quinn, B.G. and Fernandes, J.M. A Fast Efficient Technique for the Estimation of Frequency. *Biometrika*, 78:489–497, 1991.

[11] Tretter, S.A. Estimating the Frequency of a Noisy Sinusoid by Linear Regression. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 31:832–835, 1985.

ON DESIGN OF CORRELATION BASED FREQUENCY ESTIMATORS

Björn Völcker and Peter Händel

Department of Signals, Sensors and Systems
Royal Institute of Technology, SE-100 44 Stockholm, Sweden

ABSTRACT

As a complement to the Periodogram, low complexity frequency estimators are of interest. Different designs of these estimators may affect the performance significantly. In this paper we consider correlation based estimators and present a design strategy that outperforms most estimators in the same class. We give a closed form expression for the asymptotic performance together with a new method of phase unwrapping to resolve an introduced frequency ambiguity. Finally, we illustrate the performance through a design example.

1. INTRODUCTION

Estimation of the parameters of a noise corrupted sinusoidal model is a frequently addressed problem in the signal processing literature. The signal model is of interest in different application areas, such as communications, radar, measurements and geophysical exploration, among others. Starting with an observed sample $\{y(0),\ldots,y(N-1)\}$, where N is the number of data points, there exist numerous methods which can be used to estimate the sought parameters.

Often, the estimation of the frequencies is of particular interest. It is well known that in most applications, excellent estimates of the sought frequencies are easily obtained by peak-picking the Periodogram of data, i.e., the magnitude square of the discrete Fourier transform. Besides the fact that the Periodogram is an excellent frequency estimator, it can be efficiently implemented using the fast Fourier transform of the observations followed by a search, or interpolation, for the spectral maxima. Thus, there are basically only two scenarios where there is a need for alternative methods, that is *i)* when the frequencies are so closely spaced that they cannot be resolved by the Periodogram, and *ii)* when the real-time constraints on numerical complexity requires low-complexity methods.

We focus on the second scenario only, i.e., low complexity methods. Therefore, consider the single-tone model

$$y(n) = ae^{\mathrm{i}\omega n} + e(n), \qquad n = 0,\ldots,N-1 \qquad (1)$$

where $a = |a|e^{\mathrm{i}\phi}$ is a complex-valued amplitude, and $\omega \in [-\pi,\pi)$ is the normalized (angular) frequency. The noise $e(n)$ is zero mean complex-valued circular white Gaussian with variance σ^2. The parameters $(|a|,\phi,\omega,\sigma^2)$ are all unknown, but the frequency ω is the parameter of main interest.

For the signal model in (1) it is well known that the maximum likelihood estimate (MLE) of the frequency is given by the location at which the Periodogram attains its maximum [1]. As a complement to the Periodogram, there have been a large amount of papers on low-complexity estimators. Basically, they can be divided into two classes, that is data based and correlation based. The well known weighted phase averager [2] belongs to the former class. In this paper, we consider correlation based estimators, i.e., an estimate of the frequency is obtained from one or several estimated entries of the autocorrelation sequence $\{r(m)\}$ of $y(n)$

$$r(m) = \mathrm{E}[y(n)y^*(n-m)] = |a|^2e^{\mathrm{i}\omega m} + \sigma^2\delta_{m,0}. \qquad (2)$$

Here, $\delta_{m,0}$ is the Kronecker delta and $(^*)$ denote complex conjugate. From the data we can form the sample correlation sequence $\{\hat{r}(m)\}$ where $\hat{r}(m)$ is, for example, the unbiased estimator

$$\hat{r}(m) = \frac{1}{N-m}\sum_{n=m}^{N-1} y(n)y^*(n-m). \qquad (3)$$

Considering computational complexity, one may note from the above that correlation based methods are only an alternative to the Periodogram based methods when the number K of correlation lags is fixed, and $K \ll \log(N)$. Due to the averaging in $\hat{r}(m)$, this class of methods have SNR thresholds between the threshold of MLE and the thresholds of data based methods. The correlation based estimators may, however, not be statistically efficient at high SNR.

Clearly, one can use the truncated sequence $\hat{r}(1),\ldots,\hat{r}(K)$ and the observation that the sequence itself is a noise corrupted sinusoidal signal with the same frequency as the raw data and fit the unwrapped phase to a straight line [3]. By considering estimators formed from an arbitrary set of correlations, we show that it is possible to improve their accuracy (in terms of a lower error variance, or a lower SNR threshold) while retaining a low numerical complexity. *We present a methodology to find correlation based estimators with minimal error variance performance, subject to an arbitrary set of correlations* $\{\hat{r}(L_1),\ldots,\hat{r}(L_K)\}$. By a simple example, we illustrate that using the given methodology, we are able to outperform many of the previously published tone frequency estimators in the trade-off between accuracy/threshold and complexity.

Estimation of phase parameters by linear regression requires an unwrapped phase. This is often done on the entire data set, for which the process is straightforward. In [4] a frequency estimator based on two correlations was proposed. It was further shown how the frequency ambiguity can be resolved if the correlation lags are relatively prime. We take this approach a step further and show that the phase unwrapping, from an arbitrary set of phases, is an integer assignment problem related to frequency estimation. By invoking the Chinese remainder theorem (CRT) we propose an efficient implementation of the phase unwrapping.

Corresponding author: email: bjorn.volcker@s3.kth.se, Voice: +46-8-790 7749, Fax: +46-8-790 7260.

0-7803-7011-2/01/$10.00 ©2001 IEEE

2. FREQUENCY ESTIMATION FROM SETS OF CORRELATIONS

From (2) it is evident that information about the frequency is gathered in the phase angle of $r(m)$, that is, for $m \neq 0$,

$$m\omega = \angle[r(m)] + 2\pi\ell \tag{4}$$

for some integer ℓ satisfying $0 \leq \ell < m$. Here, $\angle[\cdot]$ denotes the phase angle in $[0, 2\pi)$. For notational brevity and without loss of generality, m is restricted to be positive, and ω is mapped to the interval $[0, 2\pi)$ instead of $[-\pi, \pi)$. For $m = 1$ the frequency can be unambiguously estimated, i.e. $\hat{\omega} = \angle[\hat{r}(1)]$, but it is known to have poor performance [5]. With prior knowledge of the frequency interval of interest, ℓ, it is shown that the error variance can be significantly reduced by increasing m [5], i.e., an estimator

$$\hat{\omega}_m = \frac{\angle[\hat{r}(m)] + 2\pi\ell}{m}.$$

If ℓ is not known a priori the frequency cannot be uniquely determined from one correlation only. In [4] a method, based on two correlations with relatively prime correlation lags ($m = L_1, m = L_2$), was introduced to resolve the ambiguity. Here we extend this to the general case of K correlations, and further suggest a simple implementational design using shift registers.

Starting with the problem of estimating the frequency from phase information of the correlations, a system of K equations and $K + 1$ unknowns $(\omega, \ell_1, \ldots, \ell_K)$ follows from (4), i.e.,

$$\mathbf{L}\omega = \boldsymbol{\varphi} + 2\pi\boldsymbol{\ell}.$$

Here, $\mathbf{L} = [L_1, \ldots, L_K]^{\mathrm{T}}$ and $\boldsymbol{\varphi}, \boldsymbol{\ell}$ are defined accordingly and further, $\varphi_k = \angle[r(L_k)]$. In an ideal case (no noise) only one ω satisfies all the K equations if there is no common divisor among $\{L_k\}$. With noisy measurements we can solve for ω in a least squares sense for every combination of $\{\ell_k\}$ and pick the best one. The weighted least squares solution is

$$\hat{\boldsymbol{\ell}} = \arg\min_{\boldsymbol{\ell} \in \mathcal{L}_\ell} (\hat{\boldsymbol{\varphi}} + 2\pi\boldsymbol{\ell})^{\mathrm{T}} \boldsymbol{\Pi}_{\mathbf{L}}^{\perp} (\hat{\boldsymbol{\varphi}} + 2\pi\boldsymbol{\ell}) \tag{5}$$

$$\hat{\omega}(\hat{\boldsymbol{\ell}}) = \frac{\mathbf{L}^{\mathrm{T}}\mathbf{W}(\hat{\boldsymbol{\varphi}} + 2\pi\hat{\boldsymbol{\ell}})}{\mathbf{L}^{\mathrm{T}}\mathbf{W}\mathbf{L}} \tag{6}$$

$$\boldsymbol{\Pi}_{\mathbf{L}}^{\perp} = \mathbf{W} - \frac{\mathbf{W}\mathbf{L}\mathbf{L}^{\mathrm{T}}\mathbf{W}}{\mathbf{L}^{\mathrm{T}}\mathbf{W}\mathbf{L}}$$

where $(^{\mathrm{T}})$ denotes transpose, $\mathcal{L}_\ell$ is the set of feasible combinations of $\boldsymbol{\ell}$ and $\mathbf{W}$ is a weighting matrix. The original frequency estimation problem is now separated into two subproblems. First, phase unwrapping, i.e., determining the unknown set $\{\ell_k\}$, (5). Secondly, frequency estimation from the unwrapped phase, i.e., (6). Despite the joint nature of the problem, phase unwrapping and frequency estimation are often treated separately in the literature. This simplifies analyses significantly. A proper analysis requires a careful treatment of errors in the phase unwrapping. An optimal weighting will then be frequency dependent, hence not very applicable in practice. In a high SNR scenario though, the probability of an incorrect phase unwrapping is negligible. Therefore, the assumption of a correct phase unwrapping, used in the performance analyses, is justified.

In Sect. 2.2 we introduce a new alternative approach to the phase unwrapping in (5). But first, we assume that $\boldsymbol{\ell}$ is known or estimated, and consider the frequency estimation problem given a set of K correlations.

2.1. Frequency Estimator

Let $\hat{r}(L_1), \ldots, \hat{r}(L_K)$ (such that $L_1 < \ldots < L_K$) be K sample correlations. Any frequency estimator based on phase information can be formed as a weighted average of the unwrapped phase, i.e.,

$$\hat{\omega}_{\boldsymbol{\alpha}}(\mathbf{L}) = \boldsymbol{\alpha}^{\mathrm{T}}(\hat{\boldsymbol{\varphi}} + 2\pi\boldsymbol{\ell}) \tag{7}$$

where $\boldsymbol{\alpha}$ is a weighting vector with $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{L} = 1$ for unbiased estimates. For clarification the dependence of $\mathbf{L}$ is stated. The asymptotic error variance of (7) as well as the optimal $\boldsymbol{\alpha}$ and correlation lag constellation $\mathbf{L}$ are studied in detail in Sect. 3.

Note that the WLSE as well as the estimators in [3–5] are special cases of (7). For [3] the correlation lag constellation is $\mathbf{L} = [1, \ldots, K]^{\mathrm{T}}$ and the weighting vector is $[\boldsymbol{\alpha}]_k = 6k^2/[K(K+1)(2K+1)]$. In [4], $K = 2$. Here, the correlation lags are $\mathbf{L} = [2N/3,\ 2N/3+1]^{\mathrm{T}}$ and the weighting vector is $\boldsymbol{\alpha} = [1,\ 0]^{\mathrm{T}}$.

2.2. Phase Unwrapping

The optimization problem in (5) requires numerous computations, which has to be kept low by complexity reasons. We therefore introduce another approach to this problem, which is less complex.

Define the dummy variable P as

$$P \triangleq -\frac{1}{2\pi}\sum_{k=1}^{K} \beta_k L^{(k)} \angle[r(k)]$$

where $L^{(k)} = (\prod_{q=1}^{K} L_q)/L_k$ and $\{\beta_k\}$ are integers that satisfy $\sum_{k=1}^{K} \beta_k = 0$. We now show that the set $\{\ell_k\}$ can be uniquely determined from P if and only if $\{L_k\}$ are all relatively prime and $\{\beta_k\}$ properly chosen. It is straightforward to verify that

$$P = \sum_{k=1}^{K} \beta_k L^{(k)} \ell_k = \text{integer}$$

$$\ell_k = (b_k P \mod L_k)$$

where b_k is the modulo L_k inverse of $\beta_k L^{(k)}$, i.e., the integer b_k satisfies $(b_k \beta_k L^{(k)} \mod L_k) = 1$. This is an example of the CRT (see e.g. [6]) and a direct consequence of this theorem is that $\{\ell_k\}$ are identifiable if and only if $\{\beta_k, L_k\}$ are all relatively prime.

For noisy measurements, P will not likely be an integer and we have to round towards the closest one. This introduces an error of course, but the error probability can be reduced by choosing β_k small in magnitude. As K and/or L_k increases, the variance of $\hat{P}$ increases and can be quite large due to large values of the product $\beta_k L^{(k)}$. This increases the probability of an incorrect phase unwrapping, which is the main contribution to the threshold effect occurring in non-linear estimation. In practice the algorithm is applied on subsets with two correlations at a time, rendering a lower error probability. For a (sub)set of $K = 2$, choose $\beta_k = \pm 1$. This special case gives the setup in [4].

An alternative to the modulo operator is tabulation. We can generate a table of all possible P and store the values of $\{\ell_k\}$. Despite that a table look up can be very efficient, the modulo approach has its advantages. It can for example be implemented with shift registers. Finally, the proposed phase unwrapping algorithm is summarized in Table 1.

1. Let $\angle[\hat{r}(L_k)] \in [0, 2\pi)$ denote the phase of the sample correlation $\hat{r}(L_k)$, where $\{L_k\}$ are K relatively prime integers. Choose the set $\{\beta_k\}$ properly, i.e., integers satisfying $\sum_{k=1}^{K} \beta_k = 0$ and relatively prime to $\{L_k\}$.
2. With $L^{(k)} = (\prod_{q=1}^{K} L_q)/L_k$, calculate

$$\hat{P} = -\text{round}\left[\frac{1}{2\pi}\sum_{k=1}^{K}\beta_k L^{(k)}\angle[\hat{r}(L_k)]\right]. \qquad (8)$$

3. Find the integers $\{\hat{\ell}_k\}$ that satisfy $\hat{P} = \sum_{k=1}^{K}\beta_k L^{(k)}\hat{\ell}_k$. The solution is unique and can for example be found by tabulation, or

$$\hat{\ell}_k = (b_k\hat{P} \mod L_k)$$

where b_k is the modulo L_k inverse of $\beta_k L^{(k)}$, i.e., satisfies $(b_k\beta_k L^{(k)} \mod L_k) = 1$.

Table 1. A method to resolve the ambiguity in correlation based tone frequency estimation.

3. PERFORMANCE ANALYSIS

In this section the performance of the weighted average estimator in (7) is analyzed. We derive an expression for the asymptotic error variance as well as a lower bound tighter than the Cramér-Rao bound (CRB). The performance is a function of $\boldsymbol{\alpha}$ and $\mathbf{L}$, and we further investigate the choice of them.

Let $\mathbf{R}$ be the covariance matrix of $\hat{\boldsymbol{\varphi}}$. Then the variance of the weighted estimator $\hat{\omega}_{\boldsymbol{\alpha}}(\mathbf{L})$, as given in (7), is

$$\text{var}[\hat{\omega}_{\boldsymbol{\alpha}}] = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{R}\boldsymbol{\alpha}.$$

In Lemma 3.1 the asymptotic covariance matrix of $\hat{\boldsymbol{\varphi}}$ (as SNR $\to \infty$) is given explicitly.

Lemma 3.1 (Asymptotic Covariance Matrix) *For fixed N and $L_l \geq L_k$, let $\{\hat{r}(L_k)\}$ be estimates according to (3). If the phase is correctly unwrapped ($\boldsymbol{\ell}$ is known), then element (k,l) of the asymptotic covariance matrix $\mathbf{R}$ of the phases $\hat{\boldsymbol{\varphi}}$, as* SNR $\to \infty$, *is*

$$[\mathbf{R}]_{k,l} = \frac{\min(L_k, N - L_l)}{\text{SNR}(N - L_k)(N - L_l)}.$$

Proof: The proof is given in [7]. ■

3.1. Optimal Weighting

With use of the Gauss-Markov Theorem the optimal (minimal variance) weighting scheme, for a given SNR is

$$\boldsymbol{\alpha}_{\text{opt}}(\mathbf{L}) = \frac{\mathbf{R}^{-1}\mathbf{L}}{\mathbf{L}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{L}} \qquad (9)$$

with the corresponding estimator $\hat{\omega}_{\text{opt}}(\mathbf{L}) = \boldsymbol{\alpha}_{\text{opt}}^{\mathrm{T}}(\mathbf{L})(\hat{\boldsymbol{\varphi}} + 2\pi\boldsymbol{\ell})$. Note that this coincides with the WLSE when the weighting matrix $\mathbf{W} = \mathbf{R}^{-1}$. The variance of this estimator is

$$\text{var}[\hat{\omega}_{\text{opt}}(\mathbf{L})] = \frac{1}{\mathbf{L}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{L}}$$

and may serve as a lower bound on the performance of this class of frequency estimators, given the correlation lags $\mathbf{L}$. This bound is tighter than the CRB given by [1]

$$\text{CRB}[\hat{\omega}] = \frac{6}{\text{SNR}\,N(N^2-1)}.$$

With use of Lemma 3.1, an explicit expression of the asymptotic (as SNR $\to \infty$) performance is given. This case is studied in detail in Section 3.2.

The weighting $\boldsymbol{\alpha}_{\text{opt}}$ in (9) is SNR dependent, which without prior knowledge of the SNR is of little practical use. In addition, it is difficult to derive an explicit expression of the covariance matrix for an arbitrary SNR. This can be overcome by considering a high or low SNR case, for a suboptimal weighting scheme:

$$\boldsymbol{\alpha}_\infty = \lim_{\text{SNR}\to\infty}\boldsymbol{\alpha}_{\text{opt}}, \qquad \boldsymbol{\alpha}_0 = \lim_{\text{SNR}\to 0}\boldsymbol{\alpha}_{\text{opt}}.$$

The estimator in (7) with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_\infty$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ results in the estimator with lowest variance in the limit of high SNR and low SNR, respectively.

3.2. The High SNR Case

The (sub)optimal weighting schemes are subject to a given correlation lag constellation $\mathbf{L}$. By choosing the constellation properly we can increase the performance further. In Lemma 3.2 the proper constellation for high SNR together with the weights and the resulting variance are stated.

Lemma 3.2 (Proper Choice of Correlations) *In the limit (as* SNR $\to \infty$*) the proper correlation lag constellation is given by*

$$L_k = \frac{k}{2K+1}N, \qquad k = 1, \ldots, K < \frac{N}{2}. \qquad (10)$$

If L_k is a non-integer value, it is rounded to the closest one. Further, every lag L_k has a mirror point $N - L_k$ with the same performance asymptotically in SNR. It follows that the suboptimal weighting is

$$[\boldsymbol{\alpha}_\infty]_k = \frac{3k(2K+1-k)}{K(K+1)(2K+1)}$$

resulting in an asymptotic variance of $\hat{\omega}_\infty = \boldsymbol{\alpha}_\infty^T(\hat{\boldsymbol{\varphi}} + 2\pi\boldsymbol{\ell})$ as

$$\text{var}[\hat{\omega}_\infty] = \frac{6(2K+1)^2}{\text{SNR}\,N^3((2K+1)^2-1)}.$$

Proof: The proof is given in [7]. Strictly, an optimization with respect to L_k is subject to the condition that it is an integer. If it is not, we choose the closest one. If N is large this quantization effect is negligible. ■

From Lemma 3.2 we see that the resulting efficiency tends to

$$\text{efficiency} = \frac{\text{var}[\hat{\omega}_\infty]}{\text{CRB}} = \frac{(2K+1)^2(N^2-1)}{((2K+1)^2-1)N^2}$$

as SNR $\to \infty$. Consider the special case when $2K + 1 = N$. Then the variance becomes $\text{var}[\hat{\omega}_\infty] = \text{CRB}[\hat{\omega}]$, and the method is asymptotically (as SNR $\to \infty$) efficient for any fixed N. We make the conclusion that we do not need all $N - 1$ correlations to make a correlation based estimator asymptotically efficient. Only half of the set is needed. Note that for a sequence of correlations conventional phase unwrapping applies. For this special case the constellation equals that in [3], but the weights differ. Hence, Fitz' estimator is not efficient.

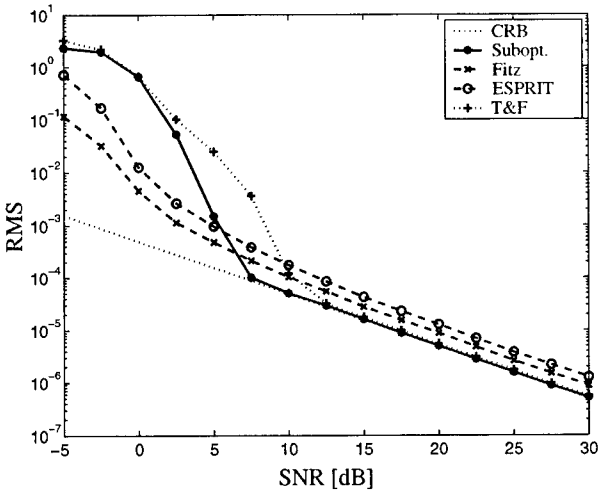


Fig. 1. The RMS plotted versus the SNR for the case $K = 3$.

3.3. The Low SNR Case and Threshold Effects

It is well known that frequency estimation suffers from threshold effects. Algorithms that rely on phase data have a higher threshold due to their use of phase unwrapping. An incorrect phase unwrapping gives a dramatic error in the frequency estimate, which is the main contribution to the threshold effect.

The error probability of $\hat{P}$ in (8) increases with the values in $\mathbf{L}$. Thus, the mirror correlation lag constellations of (10) suffer from higher probability of an incorrect phase unwrapping than the constellation in (10). This contribute to a high SNR threshold, whereas the variance are asymptotically equal. Hence, in practice the constellation in (10) is chosen.

As the SNR decreases a correct analysis must incorporate both the probability of an incorrect phase unwrapping as well as a variance expression. For SNR values below the threshold the phases tend to be uncorrelated and uniformly distributed over $[0, 2\pi)$ and the frequency cannot be determined. In this case the weighting has no effect on the performance. Thus, if the frequency estimator is to operate in a low SNR environment one has to choose a correlation lag constellation $\mathbf{L}$ that gives a low SNR threshold. This is in general achieved for small L_k.

4. DESIGN EXAMPLE

To illustrate the performance of the proposed design strategy we consider the cases $K = 3$, i.e., all estimators use three correlations, except Tufts & Fiore (T&F) which uses two correlations by construction [4]. The experimental setup is a single complex-valued sinusoid with $\omega = 0.71$ and $N = 24$ sample points. In Fig. 1 the root mean square error (RMS) is plotted versus the SNR. The RMS is calculated over 5000 trials. It is easily verified that the proposed design outperforms most of the previous estimators (Fitz [3], ESPRIT and Tufts & Fiore (T&F) [4]) for high SNR scenarios. We use the high SNR suboptimal method with the constellation in (10). From Fig. 1 it seems like both the suboptimal estimator and the T&F method are efficient at high SNR, but this is not true. In fact, their efficiencies are $49/48$ and $9/8$ respectively at high SNR.

The main contribution to the complexity is the calculation of the sample correlations, which is compared in Table 2. Small correlation lags are used for the estimator with the constellation in (10), which render a low SNR threshold. This to a cost in complexity. We have a trade-off between complexity and accuracy/threshold that has to be treated from case to case. In [7] a design strategy given a numerical complexity is introduced. The strategy determines a good constellation in a trade off between asymptotic performance and low SNR threshold. The reference also includes a more detailed complexity analysis.

	PROPOSED		FITZ	T&F	KAY
	L in (10)	Mirror			[2]
Adds/Mults	$\sim 6KN$	$\sim 2KN$	$\sim 8KN$	$\sim 5.3N$	$\sim 7N$
Phases	K	K	K	2	$N-1$

Table 2. Number of real valued multiplications/additions as well as the number of phase calculations, for the different estimators is given. In addition, a comparison with Kays estimator is included.

5. CONCLUSIONS

In this paper we proposed a design strategy for correlation based frequency estimators. From an arbitrary set of sample correlations we formed an estimator by weighting the unwrapped phase. An optimal weighting scheme was derived and as a complement, one suboptimal strategy (high SNR case) was analyzed. For good performance we showed how to choose the correlation lag constellation properly. We also proposed a new method of phase unwrapping, based on an integer assignment problem and the CRT. For easy reference, see Table 1.

We compared the performance of our design with other estimators in the same class. These estimators are special cases of the proposed design, and we can outperform them as well as many similar estimator.

The analysis assumes a correct phase unwrapping, i.e., $\hat{P} = P$. For reasonably high SNR the error probability is negligible, but is a main error source for SNR < 0 dB [4, 7].

6. REFERENCES

[1] D. C. Rife and R. R. Boorstyn, "Single tone parameter estimation from discrete-time observations," *IEEE Trans. on Info. Theory*, vol. IT-20, no. 5, pp. 591–598, 1974.

[2] Steven M. Kay, "A fast and accurate single frequency estimator," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 37, no. 12, pp. 1987–1990, 1989.

[3] M. P. Fitz, "Further results in the fast estimation of a single frequency," *IEEE Trans. on Communications*, vol. 42, no. 2/3/4, pp. 862–864, 1994.

[4] D. W. Tufts and P. D. Fiore, "Simple, effective estimation of frequency based on Prony's method," in *ICASSP*, 1996, vol. 5, pp. 2801–2804.

[5] G. W. Lank, I. S. Reed, and G. E. Pollon, "A semicoherent detection and Doppler estimation statistic," *IEEE Trans. on Aero. and Elect. Syst.*, vol. AES-9, pp. 151–165, 1973.

[6] N. K. Bose, *Digital filters: Theory and applications*, North-Holland, 1985.

[7] B. Völcker and P. Händel, "Frequency estimation from proper sets of correlations," *IEEE Trans. on Signal Proc.*, To appear.

GAUSSIAN PARTICLE FILTERING

Jayesh H. Kotecha and Petar M. Djurić

Department of Electrical and Computer Engineering
State University of New York at Stony Brook, Stony Brook, NY 11794
jkotecha@ece.sunysb.edu, djuric@ece.sunysb.edu

ABSTRACT

Sequential Bayesian estimation for dynamic state space models involves recursive estimation of hidden states based on noisy observations. The update of filtering and predictive densities for nonlinear models with non-Gaussian noise using Monte Carlo particle filtering methods is considered. The Gaussian particle filter (GPF) is introduced, where densities are approximated as a single Gaussian, an assumption which is also made in the extended Kalman filter (EKF). It is analytically shown that, if the Gaussian approximations hold true, the GPF minimizes the mean square error of the estimates asymptotically. The simulations results indicate that the filter has improved performance compared to the EKF, especially for highly nonlinear models where the EKF can diverge.

1. INTRODUCTION

Nonlinear filtering problems arise in many fields including statistical signal processing, economics, statistics, biostatistics and engineering such as communications, radar tracking, sonar ranging, target tracking, and satellite navigation. Many of these problems can be written in the form of the so called Dynamic State Space (DSS) model [1]. The signal of interest $\{\mathbf{x}_n; n \in \mathbb{N}\}, \mathbf{x} \in \mathbb{R}^{m_x}$, is an unobserved (hidden) Markov process of initial distribution $p(\mathbf{x}_0)$ represented by the distribution $p(\mathbf{x}_n|\mathbf{x}_{n-1})$. The observations $\{\mathbf{y}_n; n \in \mathbb{N}\}, \mathbf{y} \in \mathbb{R}^{m_y}$, are conditionally independent given the state process $\{\mathbf{x}_n; n \in \mathbb{N}\}$ and represented by the distribution $p(\mathbf{y}_n|\mathbf{x}_n)$. Alternatively, the model can be written as

$$\begin{aligned} \mathbf{x}_n &= \mathbf{f}(\mathbf{x}_{n-1}) + \mathbf{u}_n \quad \text{(process equation)} \\ \mathbf{y}_n &= \mathbf{h}(\mathbf{x}_n) + \mathbf{v}_n \quad \text{(observation equation)} \end{aligned} \tag{1}$$

where $\mathbf{u}_n$ and $\mathbf{v}_n$ are *additive*, random noise vectors of given distributions.

In a Bayesian context, our aim is to estimate *recursively in time*, the marginal posterior distribution referred to as the filtering distribution $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ and the predictive distribution $p(\mathbf{x}_{n+1}|\mathbf{y}_{0:n})$, where $\mathbf{y}_{0:n} \equiv \{\mathbf{y}_0, \ldots, \mathbf{y}_n\}$. Given these densities, an estimate of the state can be determined for any performance criterion suggested for the problem. The filtering density or the marginal posterior of the state at time n can be written as

$$p(\mathbf{x}_n|\mathbf{y}_{0:n}) = C_n p(\mathbf{x}_n|\mathbf{y}_{0:n-1}) p(\mathbf{y}_n|\mathbf{x}_n) \tag{2}$$

This work was supported by the National Science Foundation under Award Nos. CCR-9903120 and CCR-0082607.

where $C_n = (\int p(\mathbf{x}_n|\mathbf{y}_{0:n-1}) p(\mathbf{y}_n|\mathbf{x}_n) d\mathbf{x}_n)^{-1}$ is the normalizing constant. Furthermore, the predictive density can be expressed as

$$p(\mathbf{x}_{n+1}|\mathbf{y}_{0:n}) = \int p(\mathbf{x}_{n+1}|\mathbf{x}_n) p(\mathbf{x}_n|\mathbf{y}_{0:n}) d\mathbf{x}_n. \tag{3}$$

When the model is linear with additive Gaussian noise, and $p(\mathbf{x}_0)$ is Gaussian, the filtering and predictive densities are Gaussian and the Kalman filter provides the mean and covariance sequentially, which is the optimal Bayesian solution [2]. However, for most nonlinear models and non-Gaussian noise problems, closed form analytic expression for the posterior densities do not exist in general. Numerical solutions often require high dimensional integrations which are not practical to implement. As a result, several approximations which are more tractable have been proposed.

A class of filters called *Gaussian filters* provide Gaussian approximations to the filtering and predictive densities. For example, the EKF linearizes the nonlinearities around the current state and provides Gaussian approximations to the densities. Although the EKF has been successfully implemented in some problems, in others it diverges or provides very poor approximations. This is especially emphasized when the model is highly nonlinear or when the posterior densities are multimodal. In such cases however, significant improvements are possible. Efforts to improve upon the EKF have led to new filters by Julier et al. [3] and Ito et al. [4], which use deterministic sets of points in the space of the state variable to obtain more accurate approximations to the mean and covariance than the EKF.

Recently, particle based sampling filters have been used to update the posterior distributions [5],[6],[7], [8]. A density is represented by a weighted set of samples from the density, which are propagated through the dynamic system to sequentially update the posterior densities. These methods are collectively called sequential importance sampling (SIS) filters.

In this paper, we present the GPF for nonlinear DSS models in Section 2. Similar to the above mentioned Gaussian filters, the GPF approximates (2) and (3) as Gaussians. The justifications are that under this assumption only the mean and covariance need to be tracked and given just the mean and covariance, the Gaussian maximizes entropy of the random variable or it is the least informative distribution. The GPF updates the Gaussian approximations using a particle based approach, wherein random samples are generated and Monte Carlo estimates of mean and covariance are provided. In fact, all moments can be calculated similarly. It is shown analytically, that as the number of particles used $\rightarrow \infty$, the estimates converges almost surely to

0-7803-7011-2/01/$10.00 ©2001 IEEE

the minimum mean square estimates (given that the Gaussian assumption holds true). It is important to note that unlike the EKF, the assumption of additive Gaussian noise can be relaxed for the GPF. **The noises can in general be non-Gaussian and non-additive**, as long as the Gaussian approximation is valid. The GPF has improved performance compared to the EKF as demonstrated by the simulations in Section 3. Finally, we conclude the paper in section 4.

2. GAUSSIAN PARTICLE FILTERING

The GPF applies particle filtering methodology [5],[7],[9] to update the mean and covariance based on the Bayesian update equations (2) and (3). The basic idea in Monte Carlo methods is to represent a distribution $p(\mathbf{x}_n)$ of a random variable $\mathbf{x}_n$ by a collection of samples (particles) from that distribution. M particles, $\mathcal{X} = \{\mathbf{x}_n^{(1)}, \ldots, \mathbf{x}_n^{(M)}\}$, from a so called importance sampling (IS) distribution $\pi(\mathbf{x}_n)$ (which satisfies certain conditions; see [9] for details) are generated. The particles are then weighted as $w^{(j)} = \frac{p(\mathbf{x}_n^{(j)})}{\pi(\mathbf{x}_n^{(j)})}$. If $W = \{w^{(1)}, \ldots, w^{(M)}\}$, then the set $\{\mathcal{X}, W\}$ represents samples from the posterior distribution $p(\mathbf{x}_n)$. Monte Carlo integration suggests that the estimate of

$$E_p(g(\mathbf{x}_n)) = \int g(\mathbf{x}_n)p(\mathbf{x}_n)d\mathbf{x}_n \tag{4}$$

can be computed as

$$\hat{E}_p(g(\mathbf{x}_n)) = \frac{\sum_j w^{(j)} g(\mathbf{x}_n^{(j)})}{\sum_j w^{(j)}}. \tag{5}$$

Using the Strong Law of Large Numbers it can be shown that

$$\hat{E}_p(g(\mathbf{x}_n)) \longrightarrow E_p(g(\mathbf{x}_n)) \tag{6}$$

almost surely as $M \to \infty$; see for example [9]. The posterior density can be approximated as

$$p(\mathbf{x}_n)d\mathbf{x}_n = P(d\mathbf{x}_n) \approx \frac{\sum_{j=1}^{M} w^{(j)} \delta_{x_n^{(j)}}(d\mathbf{x}_n)}{\sum_{j=1}^{M} w^{(j)}} \tag{7}$$

where $\delta_{\mathbf{x}_n}(d\mathbf{x}_n)$, is the Dirac delta function. For the DSS models, SIS filters have been developed, which essentially obtain particles and their weights from the posterior densities in a recursive manner. However, a phenomenon called *sample degeneration* occurs where only a few particles representing the distribution have significant weights. A procedure called *resampling* [7] is applied to mitigate this problem, but it can give limited results and can be computationally expensive.

Since the GPF approximates posterior densities as Gaussians, particle resampling is not required, as long as the Gaussian approximations are valid. This results in an advantage of the GPF over SIS methods. Using the underlying ideas, the update mechanism for GPF is explained below.

The density of Gaussian random variable $\mathbf{x}$ is written as $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ where the m dimensional vector $\boldsymbol{\mu}$ is the mean, and the covariance is the positive definite matrix $\boldsymbol{\Sigma}$. Assume that at time $n = 1$, we have $p(\mathbf{x}_1|\mathbf{y}_0) = \mathcal{N}(\mathbf{x}_1; \bar{\boldsymbol{\mu}}_0, \bar{\boldsymbol{\Sigma}}_0)$, where $\bar{\boldsymbol{\mu}}_0$ and $\bar{\boldsymbol{\Sigma}}_0$ are chosen based on prior information. As new measurements are received, measurement and time updates are performed to obtain the filtering and predictive densities as discussed in the following sections.

2.1. Measurement Update

After receiving the n-th observation $\mathbf{y}_n$, from (2) the filtering density can be approximated as

$$p(\mathbf{x}_n|\mathbf{y}_{0:n}) \approx C_n p(\mathbf{y}_n|\mathbf{x}_n)\mathcal{N}(\mathbf{x}_n; \bar{\boldsymbol{\mu}}_n, \bar{\boldsymbol{\Sigma}}_n). \tag{8}$$

The GPF measurement update approximates the above density as a Gaussian, so that the mean and covariance of $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ are preserved, i.e.,

$$\hat{p}(\mathbf{x}_n|\mathbf{y}_{0:n}) = \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n). \tag{9}$$

In general, analytical expressions for the mean $\boldsymbol{\mu}_n$ and covariance $\boldsymbol{\Sigma}_n$ of $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ are not available. However, for the GPF update, Monte Carlo estimates of $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ can be computed from (8), where samples $\mathbf{x}_n^{(i)}$ are obtained from an importance sampling function $\pi(\mathbf{x}_n|\mathbf{y}_{0:n})$. The measurement update algorithm is given in Chart 1.

GPF - Measurement update algorithm.

1. Obtain samples from the density $\pi(\mathbf{x}_n|\mathbf{y}_{0:n})$ and denote them as $\{\mathbf{x}_n^{(j)}\}_{j=1}^{M}$.
2. Obtain the respective weights by
$$\bar{w}_n^{(j)} = \frac{p(\mathbf{y}_n|\mathbf{x}_n^{(j)})\mathcal{N}(\mathbf{x}_n = \mathbf{x}_n^{(j)}; \bar{\boldsymbol{\mu}}_n, \bar{\boldsymbol{\Sigma}}_n)}{\pi(\mathbf{x}_n^{(j)}|\mathbf{y}_{0:n})}. \tag{10}$$
3. Normalize the weights as
$$w_n^{(j)} = \bar{w}_n^{(j)} / \sum_{j=1}^{M} \bar{w}_n^{(j)}. \tag{11}$$
4. Estimate the mean and covariance by
$$\begin{aligned} \boldsymbol{\mu}_n &= \textstyle\sum_{j=1}^{M} w_n^{(j)} \mathbf{x}_n^{(j)} \\ \boldsymbol{\Sigma}_n &= \textstyle\sum_{j=1}^{M} w_n^{(j)} (\mathbf{x}_n^{(j)} - \boldsymbol{\mu}_n)(\mathbf{x}_n^{(j)} - \boldsymbol{\mu}_n)^T. \end{aligned} \tag{12}$$

GPF - Time update algorithm.

1. Draw samples from $\mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ and denote them as $\{\mathbf{x}_n^{(j)}\}_{j=1}^{M}$.
2. For $j = 1, \ldots, M$, sample from $p(\mathbf{x}_{n+1}|\mathbf{x}_n = \mathbf{x}_n^{(j)})$ to obtain $\{\mathbf{x}_{(n+1)}^{(j)}\}_{j=1}^{M}$.
3. Compute the mean $\bar{\boldsymbol{\mu}}_{n+1}$ and covariance $\bar{\boldsymbol{\Sigma}}_{n+1}$ by taking sample means and covariances.

Chart 1.

Theorem 1 *Assume $p(\mathbf{x}_n|\mathbf{y}_{0:n-1}) = \mathcal{N}(\mathbf{x}_n; \bar{\boldsymbol{\mu}}_n, \bar{\boldsymbol{\Sigma}}_n)$ at time n. Upon receiving the n-th observation $\mathbf{y}_n$, the GPF measurement updates the filtering density as shown in Chart 1. Then $\boldsymbol{\mu}_n$ computed in (12) converges almost surely as*

$M \to \infty$ to the minimum mean square error (MMSE) estimate of $\mathbf{x}_n$. In addition, the estimate of the MMSE given by $\bar{\mathbf{\Sigma}}_n$ in (12) converges almost surely as $M \to \infty$ to the true MMSE.

For a proof, see [10]. The same is true for all central and non-central moments.

The above corollary shows that given that the Gaussian approximation is valid, the GPF provides the MMSE estimate asymptotically during the measurement update, which is clearly not true for the EKF. Hence, the GPF is expected to perform better than the EKF.

2.1.1. Choice of $\pi(\cdot)$

The choice of IS density $\pi(\cdot)$ depends on the problem, [8],[9]. For the GPF, a simple choice for $\pi(\cdot)$ is $p(\mathbf{x}_n|\mathbf{y}_{0:n-1}) = \mathcal{N}(\mathbf{x}_n; \bar{\boldsymbol{\mu}}_n, \bar{\mathbf{\Sigma}}_n)$. Alternatively, samples obtained in the time update step (presented in the next section) in step 2 can be used. However, this choice can be inadequate in some applications. Another choice is $\mathcal{N}(\mathbf{x}_n; \bar{\boldsymbol{\mu}}_{n|n}, \bar{\mathbf{\Sigma}}_{n|n})$, where $\bar{\boldsymbol{\mu}}_{n|n}$ and $\bar{\mathbf{\Sigma}}_{n|n}$ are obtained from the measurement update step of the EKF or from the unscented Kalman filter [3].

2.2. Time update

Assume that at time n, it is possible to obtain samples from $p(\mathbf{x}_{n+1}|\mathbf{x}_n)$. From (3) and (9)

$$p(\mathbf{x}_{n+1}|\mathbf{y}_{0:n}) \approx \int \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_n, \mathbf{\Sigma}_n) p(\mathbf{x}_{n+1}|\mathbf{x}_n) d\mathbf{x}_n. \quad (13)$$

A Monte Carlo approximation for (13) is

$$p(\mathbf{x}_{n+1}|\mathbf{y}_{0:n}) \approx \frac{1}{M} \sum_{i=1}^{M} p(\mathbf{x}_{n+1}|\mathbf{x}_n^{(i)}) \quad (14)$$

where $\mathbf{x}_n^{(i)}$ are particles from $\mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_n, \mathbf{\Sigma}_n)$. The GPF time update approximates $p(\mathbf{x}_{n+1}|\mathbf{y}_{0:n})$ as a Gaussian, such that its mean and covariance are preserved, i.e.,

$$\hat{p}(\mathbf{x}_{n+1}|\mathbf{y}_{0:n}) = \mathcal{N}(\mathbf{x}_n; \bar{\boldsymbol{\mu}}_n, \bar{\mathbf{\Sigma}}_n). \quad (15)$$

However, since closed form analytical expressions of $\bar{\boldsymbol{\mu}}_n$ and $\bar{\mathbf{\Sigma}}_n$ may not be available, we compute Monte Carlo estimates from (14). The Monte Carlo time update steps are shown in Chart 1.

Similar to Theorem 1, it can be shown that $\bar{\boldsymbol{\mu}}_{n+1}$ converges almost surely as $M \to \infty$ to the MMSE estimate of $\mathbf{x}_{n+1}$ given the observations until time n.

3. SIMULATION RESULTS

The GPF was applied to some numerical examples, and here we present results for the univariate non-stationary growth model (UNGM), which has been used previously in [5],[11]. We choose this model because it is highly nonlinear and is bimodal in nature. The DSS equations are

$$\begin{aligned} x_n &= \alpha x_{n-1} + \beta \frac{x_{n-1}}{1+x_{n-1}^2} + \gamma \cos(1.2(n-1)) + u_n \\ y_n &= x_n^2/20 + v_n, \quad n = 1, \ldots, N \end{aligned} \quad (16)$$

where $v_n \sim \mathcal{N}(v_n; 0, \sigma_v^2)$ and $u_n \sim \mathcal{N}(u_n; 0, \sigma_u^2)$ This model is highly nonlinear in both the process and observation equations. Notice the term in the process equation which is independent of x_n but varies with time n, this can be interpreted as time varying noise. The likelihood $p(y_n|x_n)$ has bimodal nature when $y_n > 0$, but when $y_n < 0$ it is unimodal. The bimodality makes the problem more difficult to address using conventional methods.

		M=20		M=100		M=1000	
	EKF	GPF	SIS	GPF	SIS	GPF	SIS
1	175.7	26.3	28.6	12.7	14.5	11.2	11.8
2	164.7	25.7	29.4	12.9	14.0	11.2	11.6
3	176.1	25.1	30.6	12.2	13.4	10.5	10.9
4	160.7	29.9	27.4	13.6	14.6	11.8	12.0
5	199.4	24.6	26.5	11.6	14.6	11.2	11.4
6	182.3	30.8	30.3	15.2	15.5	12.8	13.2
7	185.9	27.1	24.9	13.3	15.0	11.3	12.0
8	175.3	27.5	28.8	11.9	14.6	10.9	11.7
9	171.1	25.6	28.6	12.1	14.0	10.6	11.7
10	168.2	26.7	27.8	12.6	14.0	10.6	11.7

Table 1: MSEx_f for 10 random simulation runs for the EKF, GPF and SIS. M is the number of particles for GPF and SIS.

We compare performance of the EKF, GPF and SIS filters based on the following metrics. MSEx_f is defined by $\frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{x}_n)^2$ where $\hat{x}_n = E(\mathbf{x}_n|\mathbf{y}_{0:n})$, which is obtained from the filtering density. When the ratio $\frac{x^2/20}{\sigma_v^2}$ is small, then the bimodality of the problem is more severe and we expect to see improved performance of the GPF in the presence of this high nonlinearity over that of the EKF.

Data were generated using $x_0 = 0.1$, $\sigma_v^2 = 1$, $\sigma_u^2 = 1$, $\alpha = 0.5, \beta = 25$, $\gamma = 8$, and $N = 5000$ in each simulation. The initial distribution was $p(\mathbf{x}_0) \sim \mathcal{N}(0, 1)$. For both GPF and SIS, the IS density chosen is the prior given by $p(x_n|y_{0:n-1})$ and $p(x_n|x_{n-1})$ respectively. For the GPF and SIS, since we draw particles from $p(x_n|y_{0:n})$ in the measurement update, we obtain a Monte Carlo estimate for $\hat{y}_n$.

A large number of simulations were performed where all the three filters were used for state estimation. Results are shown for different choices of the number of particles $M = 20, 100, 1000$. In Table 1, we show MSEx_f for 10 random simulations, with M varied for GPF and SIS filters. The GPF has marginally better performance than the SIS for each choice of M. It is noted that even for $M = 20$, the GPF and SIS have better performance than the EKF. Increasing M to 100 gave significant improvement in performance, however increasing M to 1000 did not change the performance much. Note the significant improvement of the GPF over the EKF in terms of the MSEs for this model. The MSEs for the EKF were large due to its tendency to diverge at high nonlinearities.

In Figures 1 and 2, we show a plot for the first 100 states and the estimates obtained using the EKF and GPF respectively. Note the tendency of the EKF to track the opposite mode of the bimodality, especially when $\frac{x^2/20}{\sigma_v^2}$ is small. This behavior was observed in general for most simulation runs. In Figures 3 and 4, we plot the error $x_n - \hat{x}_n$ and the $3\hat{\sigma}_{err}$ intervals, where $\hat{\sigma}_{err}$ was the estimated standard deviation of the prediction error. Note that as expected, the errors lie mostly within this interval for the GPF, however not so for the EKF. Also, the values of $\hat{\sigma}_{err}$ for the EKF are much higher, pointing to the occurrence of divergence of

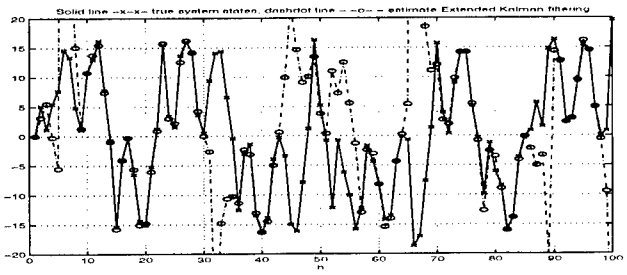


Figure 1: Plot of the true state and estimate of the EKF

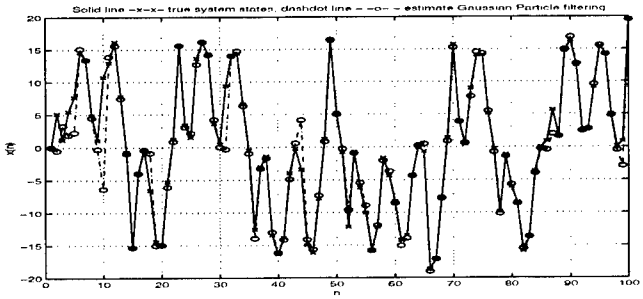


Figure 2: Plot of the true state and estimate of the GPF

the filter. All of the above observations, were made in most of the simulation runs. Clearly, the GPF outperforms the EKF significantly for this highly nonlinear example. The GPF had marginally better performance than the SIS for this model, but the computational complexity of GPF is much lower than SIS, since resampling is required for the SIS. In general, however, we cannot expect the GPF to work better than the SIS since the Gaussian assumption is not present in the SIS.

4. CONCLUSION

The Gaussian particle filter provides much better performance than the EKF. Moreover, the additive Gaussian noise assumption can be relaxed without any modification to the filter algorithm. Updating the filtering and predictive densities as Gaussians using particle based approaches has the advantages of easy implementation and better performance. The parallelizibility of the filter makes it convenient for

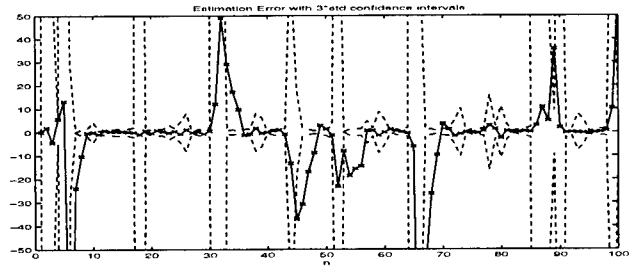


Figure 3: Plot of the prediction error and $3\hat{\sigma}_{err}$ interval for the EKF

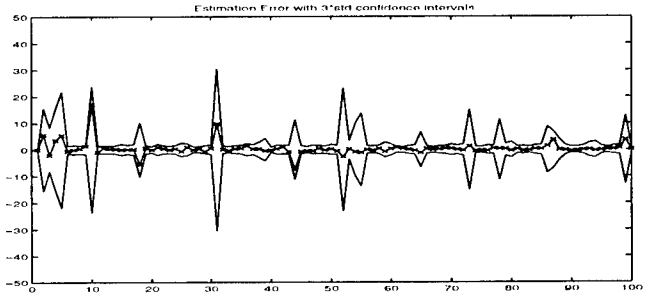


Figure 4: Plot of the prediction error and $3\hat{\sigma}_{err}$ interval for the GPF

VLSI implementation and hence more feasible for practical real time applications. For extensions to this work, see [10].

5. REFERENCES

[1] P.J. Harrison and C.F. Stevens, "Bayesian forecasting (with discussion)," *Journal of the Royal Statistical Society , Ser B*, vol. 38, pp. 205–247, 1976.

[2] B.D.O Anderson and J.B.Moore, *Optimal Filtering*, Prentice Hall Inc., Englewood Cliffs, New Jersey 07632, 1979.

[3] S.J.Julier, J.K.Uhlmann, and H.F.Durrant-Whyte, "A new method for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Transactions on Automatic Control*, vol. 45, no. 3, pp. 477–482, March 2000.

[4] K. Ito and K. Xiong, "Gaussian Filters for Nonlinear Filtering Problems," *IEEE Transactions on Automatic Control*, vol. 45, no. 5, pp. 910–927, May 2000.

[5] N. Gordon, D. Salmond, and C. Ewing, "Bayesian state estimation for tracking and guidance using the bootstrap filter," *Journal of Guidance, Control and Dynamics*, vol. 18, no. 6, pp. 1434–1443, Nov-Dec. 1995.

[6] J. H. Kotecha and P.M. Djurić, "Sequential Monte Carlo sampling detector for Rayleigh fast-fading channels," *International Conference on Acoustics, Speech and Signal Processing*, 2000.

[7] J.S. Liu and R. Chen, "Monte Carlo methods for dynamic systems," *Journal of American Statistical Association*, vol. 93, no. 443, pp. 1032–1044, 1998.

[8] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[9] J. Geweke, "Bayesian inference in econometrics models using Monte Carlo integration," *Econometrica*, vol. 33, no. 1, pp. 1317–1339, 1989.

[10] J. H. Kotecha and P.M. Djurić, " Gaussian Sum Particle Filtering - Part I," *submitted to IEEE Transactions on Signal Processing*, 2001.

[11] G. Kitagawa, "Non-Gaussian state-space modelling of nonstationary time series," *Journal of American Statistical Association*, vol. 82, no. 400, pp. 1032–1063, 1987.

Bayesian Learning using Gaussian Process for time series prediction

Sofiane Brahim-Belhouari and Jean-Marc Vesin
Signal Processing Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
e-mail: Sofiane.Brahim@epfl.ch, Jean-Marc.Vesin@epfl.ch

Abstract

In this paper, the problem of time series prediction is studied. A Bayesian procedure based on Gaussian process models is proposed and compared to the radial basis function networks. In our experiments, Gaussian process models show an excellent prediction. The conceptual simplicity, and good performance of Gaussian process models should make them very attractive for a wide range of problems.

1 Introduction

In the Bayesian approach to the regression problem a prior distribution over the model parameters induces a prior over functions. This prior is combined with a noise model to yield a posterior distribution over functions which can then be used for predictions. In general the prior over functions has a complex form. The idea of Gaussian Process (GP) modeling is, without parameterizing the model function, to place a prior directly on the functions space. The simplest type of prior over functions is called a Gaussian process.
It has been known for many years that such priors over functions can be defined using Gaussian process [7]. Neal has shown that many Bayesian regression models based on neural networks converge to Gaussian processes in the limit of an infinite network [6]. This has motivated the application of Gaussian process models for modeling noisy data [4] [9], noise free data [3] and also for classification problems [2] [4].
In this paper we use Gaussian process for forecasting problem, and compare its performance with other method, Radial Basis Function (RBF) neural network. The advantage of the Gaussian process formulation is that the combination of the prior and noise models can be carried out exactly using matrix operations. We also show how the hyperparameters of the covariance function which control the form of the Gaussian process can be estimated from the data using a maximum likelihood approach.

2 Forecasting Problem

The outcomes of a phenomenon over time form a time series. Time series are encountered in science as well as in real life. Most commonly time series are the result of unknown or incomplete understood systems. A time serie $x(t)$ is defined as a function x of an independent variable t, generating from an unknown system. Its main characteristic is that its evolution can not be described exactly. The observation of past values of a phenomenon in order to anticipate its futur behavior represents the essence of forecasting. A typical approach is to try to predict by constructing a prediction model which take into account previous outcomes of the phenomenon. We can take a set of d such values $x_{t-d+1}, ..., x_t$ to be the model input and use the next value x_{t+1} as the target.

2.1 Parametric approaches to the problem

In a parametric approach to forecasting we express the predictor in terms of nonlinear function $y(\boldsymbol{x}, \boldsymbol{\theta})$ parameterized by parameters $\boldsymbol{\theta}$. It implements a non linear mapping from input vector $\boldsymbol{x} = [x_{t-d+1}, ..., x_t]^\top$

0-7803-7011-2/01/$10.00 ©2001 IEEE

to the real value :

$$t^i = y(\boldsymbol{x}^i, \boldsymbol{\theta}) + \epsilon^i \qquad i = 1, ..., n \tag{1}$$

where ϵ is a noise corrupting the data points.
Time series processing is an important application area of neural networks. In fact, $\boldsymbol{y}$ can be given by a specified network. The output of the Radial Basis Function (RBF) network is computed as a linear superposition [1] :

$$y(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k g_k(\boldsymbol{x}) \tag{2}$$

where $w_k (k = 1, ..., K)$ denotes the weights of the output layer. The Gaussian basis functions g_k are defined as :

$$g_k(\boldsymbol{x}) = \exp(-\frac{||\boldsymbol{x} - \mu_k||^2}{2\sigma_k^2}) \tag{3}$$

where μ_k and σ_k^2 denotes means and variances. Thus we define the parameters as $\theta = [w_k, \mu_k, \sigma_k]^\top (k = 1, ..., K)$.

2.2 Nonparametric approaches

In nonparametric methods, predictions are obtained without representing the unknown system as an explicit parameterized function. A new method for regression was inspired by Neal's work [6] on Bayesian learning for neural networks. It is an attractive method for modelling noisy data, based on priors over function using Gaussian Processes.

3 Gaussian Process models

The Bayesian analysis of interesting forecasting models is difficult because a simple prior over parameters implies a complex prior distribution over functions. Rather than expressing our prior knowledge in terms of a prior for the parameters, we can instead integrate over the parameters to obtain a prior distribution for the model outputs in any set of cases. The prediction operation is most easily carried out if all the distributions are Gaussian. Fortunately, Gaussian process are flexible enough to represent a wide variety of interesting model structure, many of which would have a large number of parameters if formulated in more classical fashion.
A Gaussian process is a collection of random variables, any finite set of which have a joint Gaussian distribution [4]. For a finite collection of inputs, $\boldsymbol{x} = [\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}]^\top$, we consider a set of random variables $\boldsymbol{y} = [y^{(1)}, ..., y^{(n)}]^\top$, to represent the corresponding function values. A Gaussian process is used to define the joint distribution between the y's :

$$\wp(\boldsymbol{y}|\boldsymbol{x}) \sim \exp(-\frac{1}{2}\boldsymbol{y}^\top \Sigma^{-1} \boldsymbol{y}) \tag{4}$$

where the covariance matrix Σ is given by the covariance function :

$$\Sigma_{pq} = cov(y^{(p)}, y^{(q)}) = C(\boldsymbol{x}^{(p)}, \boldsymbol{x}^{(q)})$$

3.1 Predicting with Gaussian Process

The goal of Bayesian forecasting is to compute the distribution $\wp(y^{(n+1)}|D, \boldsymbol{x}^{(n+1)})$ of output $y^{(n+1)}$ given a test input $\boldsymbol{x}^{(n+1)}$ and a set of n training points $D = \{\boldsymbol{x}^{(i)}, t^{(i)} | i = 1, ..., n\}$.
Using Baye's rule, we obtain the posterior distribution for the $(n + 1)$ Gaussian process outputs. By conditioning on the observed targets in the training set, the predictive distribution is Gaussian with mean and variance [8] :

$$\wp(y^{(n+1)}|D, \boldsymbol{x}^{(n+1)}) \sim N(\mu_{y^{(n+1)}}, \sigma^2_{y^{(n+1)}}) \tag{5}$$

where :

$$\begin{aligned}
\mu_{y^{(n+1)}} &= \boldsymbol{a}^\top Q^{-1} \boldsymbol{t} \\
\sigma^2_{y^{(n+1)}} &= b - \boldsymbol{a}^\top Q^{-1} \boldsymbol{a} \\
Q_{pq} &= C(\boldsymbol{x}^{(p)}, \boldsymbol{x}^{(q)}) + r^2 \delta_{pq} \\
a_p &= C(\boldsymbol{x}^{(n+1)}, \boldsymbol{x}^{(p)}), \qquad p = 1, ..., n \\
b &= C(\boldsymbol{x}^{(n+1)}, \boldsymbol{x}^{(n+1)})
\end{aligned}$$

r^2 is the unknown variance of the Gaussian noise. We get a predictive distribution, not just a point prediction. This advantage can be used to obtain the prediction intervals that describe a degree of belief of the predictions.

3.2 Training a Gaussian Process

There are many possible choices of prior covariance functions. From a modeling point of view, we wish to

specify prior covariances which contain our prior beliefs about the structure of the function we are modeling. Formally, we are required to specify a function which will generate a non-negative definite covariance matrix for any set of inputs points. We find that the following covariance function works well [9] :

$$C(\boldsymbol{x}^{(p)}, \boldsymbol{x}^{(q)}) = v_0 \exp\{-\frac{1}{2}\sum_{l=1}^{d} w_l (x_l^{(p)} - x_l^{(q)})^2\} + a_0 + a_1 \sum_{l=1}^{d} x_l^{(p)} x_l^{(q)} \quad (6)$$

where $\boldsymbol{\theta} = (a_0, a_1, w_1, ..., w_d, v_0, r^2)$ plays the role of hyperparameters.

Let us assume that a form of covariance function has been chosen, but that it depends on undertermined hyperparameters $\boldsymbol{\theta}$. We would like to learn these hyperparameters from the training data. In a maximum likelihood framework, we adjust the hyperparameters so as to maximize the log likelihood of the hyperparameters :

$$\begin{aligned} \log \wp(D|\boldsymbol{\theta}) &= log\wp(t^{(1)}, ..., t^{(n)}|\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}, \boldsymbol{\theta}) \\ &= -\frac{1}{2}\log\det Q - \frac{1}{2}\boldsymbol{t}^\top Q^{-1}\boldsymbol{t} - \frac{n}{2}\log 2\pi \end{aligned}$$

It is possible to express analytically the partial derivatives of the log likelihood, which can form the basis of an efficient learning scheme. These derivatives are :

$$\frac{\partial}{\partial\theta_i}\log\wp(D|\boldsymbol{\theta}) = -\frac{1}{2}tr(Q^{-1}\frac{\partial Q}{\partial\theta_i}) + \frac{1}{2}\boldsymbol{t}^\top Q^{-1}\frac{\partial Q}{\partial\theta_i}Q^{-1}\boldsymbol{t}$$

We initialize the hyperparameters to random values (in a reasonable range) and then use an iterative method, for example conjuate gradient, to search for optimal values of the hyperparameters. We have found that this approach is somethimes susceptible to local minima, so it is advisable to try a number of random starting positions in the hyperparameters space.

4 Experimental results

In order to compare Gaussian process performances with RBF ones, we consider a high chaotic system

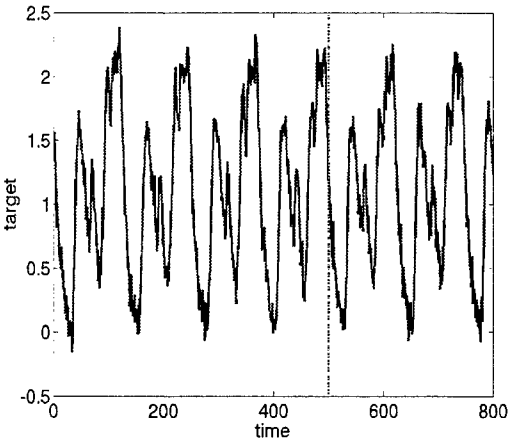


Figure 1: *Temporal patterns for the noisy Mackey Glass equation.*

generated by the Mackey-Glass equation :

$$\frac{dx(t)}{dt} = \frac{0.2x(t-\lambda)}{1+x(t-\lambda)^{10}} - 0.1x(t) \quad (7)$$

with delay $\lambda = 30$. The Mackey-Glass equation was originally developed for modeling white blood cells production [5], and became quite common as an artificial forecasting benchmark. The difficulty associated with this data set is the high nonlinearity. After integrating (7), we added noise to time series. We obtained 500 patterns for training and 300 for testing candidates models, the data set consisted of 800 samples is shown in Figure 1. Patterns were generated windowing 6 inputs and 1 output. We conducted experiments for different signal to noise ratios (SNR) using a Gaussian noise. We define the SNR as the ratio between the variance of the respective noise and the underlying time series.

The RBF network uses 30 centers chosen according to the validation set. The hayperparameters were adapted to the training data using conjugate gradient search algorithm (the linear term in the covariance function (6) involving a_1 was not present). Results of prediction errors for different SNR, using a GP models and RBF networks, are given in Figure 3. This shows that the Bayesian learning using GPs performs better

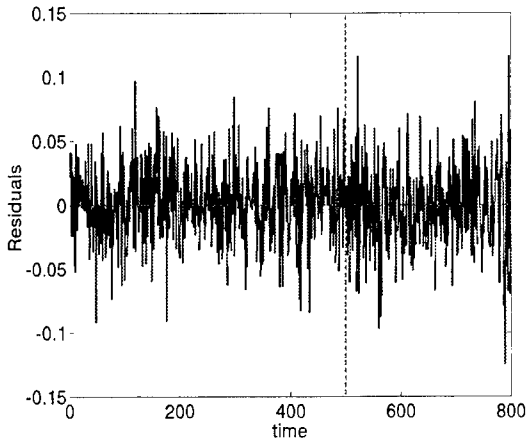


Figure 2: *Residuals given by the GP model.*

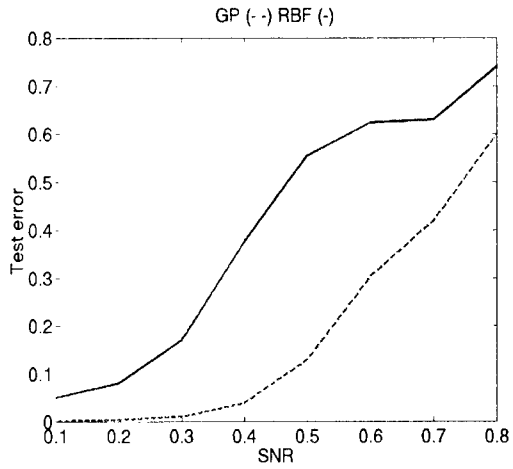


Figure 3: *Prediction errors on the test set given by GP model (- -) and RBF one (-).*

than RBF networks, specially in the noisy data set.

5 Conclusions

We have presented the method of forecasting with Gaussian process models, and shown how Bayesian learning performs better than RBF neural networks. Gaussian process models are a simple, practical and powerful Bayesian tool for data analysis.

Real life time series are often non-stationary, meaning that the data distribution is changing over time. It is why, we have also conducted some experiments on the use of the Gaussian process with a simple non-stationary covariance function for real temporal patterns. Although space limitations do not allow these to be described here.

The examination of a more complicated parameterized covariance function is currently under investigation to improve the predictors tracking in non-stationary cases.

References

[1] C. M. Bishop: "Neural networks for pattern recognition," *Clarendon press* , Oxford, 1995.

[2] M. N. Gibbs and D. J. Mackay: "Variational Gaussian process classifiers," *draft manuscript*, 1997.

[3] R. M. Neal: "Monte Carlo implementation of Gaussian process models for Bayesian regression and classification," *Technical Report*, No. 9702, University of Toronto.

[4] D. J. C. MacKay: "Gaussian Processes - A Replacement for Supervised Neural Networks," *Lecture notes for a tutorial at NIPS '97* , 1997.

[5] M. C. Mackey and L. Glass: "Oscillations and chaos in physiological control systems," *Science*, No. 197, pp. 287-289, 1977.

[6] R. M. Neal: " Bayesian learning for neural networks," *Springer-Verlag*, New York, 1996.

[7] A. O'Hagan: "Curve fitting and optimal design for prediction," (with discussion), *Journal of the Royal Statistical Society B*, Vol. 40, pp. 1-42, 1978.

[8] C. E. Rasmussen: " Evaluation of Gaussian processes and other methods for non-linear regression," *PhD. thesis*, University of Toronto. 1996.

[9] C. K. I. Williams and C. E. Rasmussen: "Gaussian process for regression," *In Advances in information processing systems 8 eds.*, pp. 111-116, 1996.

ON THE ESTIMATION OF COMMON NON-LINEARITY AMONG REPEATED TIME SERIES

Adrian G. Barnett, Rodney C. Wolff

Centre in Statistical Science and Industrial Mathematics
Queensland University of Technology
Brisbane 4001 Qld Australia
fax: 61 7 3864 2310
a.barnett@fsc.qut.edu.au

ABSTRACT

The bispectrum is a higher-order statistic and is known to be a useful tool for detecting non-linearity. A recent succinct example of its power to identify non-linear sound waves from broken bridge struts was given by [5]. As well as detecting non-linearity it has the further advantage that its magnitude and shape can be used to estimate the third order non-linear structure [1]. When a time series is repeated (such as sound waves from a collection of bridge struts) [2] showed how to produce a common spectrum and to estimate individual departures from this global quantity. The purpose of this paper is to extend this method to the bispectrum and give a summary of common non-linearity among repeated time series. We evaluate our method using data from a group of people speaking the letter 'A' and from one person repeatedly speaking this letter.

1. INTRODUCTION

The modulus-squared bispectrum is estimated as

$$|\hat{b}(\omega_j,\omega_k)|^2 = \frac{1}{n^2}|H(\omega_j)H(\omega_k)H^*(\omega_j+\omega_k)|^2, \quad j,k = 1,\ldots,\frac{n}{2}-1$$

where $\omega_j = 2\pi j/n$, $H(\omega) = \sum_{t=1}^{n} X_t e^{-i\omega t}$ is the Fourier transform of a time series $X_1,\ldots,X_n$, and * indicates the complex conjugate.

For r repeated time series each coordinate (j,k) of the estimated bispectrum is modelled as

$$|\hat{b}_i(\omega_j,\omega_k)|^2 = |b_c(\omega_j,\omega_k)|^2 Z_i(\omega_j,\omega_k)U_{ijk}, \quad i = 1,\ldots,r,\ j,k = 1,\ldots,\frac{n}{2}-1 \qquad (1)$$

where $b_c(.)$ is the common bispectrum, $Z_i(.)$ is the perturbation of the common spectrum for the ith replicate, and U_{ijk} are independently distributed error terms. Modelling this is a two-stage process. In the first stage the individual departure from the common bispectrum $Z_i(.)$ is estimated, then the actual realisation of the series for a replicate is modelled through the U_{ijk}. This is a logical basis as one might expect individual readings to vary in a reasonably consistent manner from a common quantity, and to give different readings on any one occasion. The modelling is done on the log scale as this transforms (1) to an additive function and improves the behaviour of the estimated bispectrum.

The key is then to summarise the degree of heterogeneity between repeated responses and to look for an overall non-linear structure using the common bispectrum.

2. COMMON BISPECTRUM

[2] used a parametric method to find the common spectrum but this rather restricts its shape as well as introducing the chance of making a wrong decision. Ordinary non-parametric kernel smoothing can be used to estimate the common bispectrum, which allows the data govern to its shape. To estimate a common bispectrum we need a two-dimensional smoothing process and an ideal method was proposed by [4]. As noted by the author the method does not work well at the borders, and this is overcome by reflecting the bispectrum data in the boundaries (so we now have an area of size $\tilde{n} = \frac{n}{2} - 1 + 2\gamma$). Working on the log scale allows the use of a state space model with $Y_{ijk} = \log \hat{b}_i(\omega_j,\omega_k)$, $S_{ijk} = \log Z_i(\omega_j,\omega_k)$ and $\epsilon_{ijk} = \log U_{ijk}$. The Kalman filter requires a vector so we set $\mathbf{Y_{ik}} = (Y_{i1k}, Y_{i2k},\ldots,Y_{ink})'$, and the filter equations become,

$$\mathbf{Y_{ik}} = F\alpha_{\mathbf{k}} + \mathbf{S_{ik}} + \epsilon_{\mathbf{ik}}, \qquad \epsilon_{\mathbf{ik}} \sim N(0_{r\times\tilde{n}}, \sigma^2)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

$$\alpha_{\mathbf{k}} = G\alpha_{\mathbf{k-1}} + \mathbf{u_k}, \qquad \mathbf{u_k} \sim N(0_{2\tilde{n}}, \frac{\sigma^2}{\lambda}H)$$

$$F = [I_{\tilde{n}}, \emptyset], \quad H = \begin{bmatrix} I_{\tilde{n}} & \emptyset \\ \emptyset & \emptyset \end{bmatrix},$$

$$G = \left[\begin{array}{cccc|c} 3 & -1 & & 0 & \\ -1 & 4 & -1 & & \\ & & \ddots & & -I_{\tilde{n}} \\ & -1 & 4 & -1 & \\ 0 & & -1 & 3 & \\ \hline \multicolumn{4}{c|}{I_{\tilde{n}}} & \emptyset \end{array}\right]$$

where $\emptyset$ is the $\tilde{n} \times \tilde{n}$ zero matrix, $\mathbf{S_{ik}}$ are the individual specific effects, σ and λ control the degree of noise in the smoothing and observation equation; G is a matrix that smooths the data according to the discrete thin plate method; $\alpha_{\mathbf{k}} = (\mathbf{y'_{ik}}, \mathbf{y'_{ik-1}})$, where $\mathbf{y_{ik}} = (y_{i1k}, y_{i2k}, \ldots, y_{i\tilde{n}k})'$ is the smoothed surface.

We propose estimating the parameters λ and σ and the shapes of $\mathbf{S_{ik}}$ using a Bayesian MCMC method with the following steps. Initial values for $\lambda_{(0)}$ and $\sigma_{(0)}$ are taken from vague Gamma(0.5,0.5) priors. The initial subject and error effects are assumed to be zero, $\mathbf{S_{ik(0)}} = 0_{r\times\tilde{n}}$, $\hat{\epsilon}_{\mathbf{ik(0)}} = 0_{r\times\tilde{n}}$.

Step 1 - Forward sweep of Kalman filter

To get a smooth estimate of the common bispectrum we first remove the subject and error effects from the observed data to give $\mathbf{Y^*_{ik}} = \mathbf{Y_{ik}} - \mathbf{S_{ik}} - \hat{\epsilon}_{\mathbf{ik}}$.

In a forward sweep of the Kalman filter we calculate the mean and variance of the innovation equation

$$\mathbf{a_{k+1}} = G\mathbf{p_k}, \qquad \mathbf{R_{k+1}} = G\mathbf{C_k}G' + \frac{\sigma^2}{\lambda}H, \quad k = 0, \ldots, \tilde{n}-1$$

with $\mathbf{p_0} \sim N(\overline{Y}^*, Var(Y^*))$ and $\mathbf{C_0} = I_{2\tilde{n}}$, so that the initial estimates of each column are not null and neither are the variances.

The one-step forecast mean and variance are then

$$F\mathbf{a_{k+1}}, \qquad \mathbf{Q_{i,k+1}} = F\mathbf{R_{k+1}}F' + \sigma^2 I_{\tilde{n}}, \quad i = 1, \ldots, r, \; k = 0, \ldots, \tilde{n}-1$$

We can then predict the error $\mathbf{e_{i,k+1}} = \mathbf{Y^*_{i,k+1}} - F\mathbf{a_{k+1}}$ for each subject. The filtering formula which runs for $k = 1, \ldots, \tilde{n}-1$ is then

$$\mathbf{p_{k+1}} = \mathbf{a_{k+1}} + \frac{1}{r}\sum_{i=1}^{r} \mathbf{R_{k+1}}F'\mathbf{Q_{i,k+1}}^{-1}\mathbf{e_{i,k+1}}$$

$$\mathbf{C_{k+1}} = \mathbf{R_{k+1}} - \frac{1}{r}\sum_{i=1}^{r} \mathbf{R_{k+1}}F'\mathbf{Q_{i,k+1}}^{-1}F\mathbf{R_{k+1}}$$

So the effect of the error $\mathbf{e_{ik}}$ is averaged over the subjects.

Step 2 - Smoothing

The smoothing backward step is then run across $\tilde{n}-1, \ldots, 1$.

$$\mathbf{h_k} = \mathbf{p_k} + \mathbf{A_k}(\mathbf{h_{k+1}} - \mathbf{a_{k+1}})$$
$$\mathbf{H_k} = \mathbf{C_k} + \mathbf{A_k}(\mathbf{H_{k+1}} - \mathbf{R_{k+1}})\mathbf{A'_k}$$
$$\alpha_{\mathbf{k}} \sim N(\mathbf{h_k}, diag\mathbf{H_k})$$

where $\mathbf{A_k} = \mathbf{C_k}G'\mathbf{R_{k+1}}^{-1}$. The initial values for the vector and matrix $\mathbf{h_{\tilde{n}}}$ and $\mathbf{H_{\tilde{n}}}$ are $\mathbf{p_{\tilde{n}}}$ and $\mathbf{C_{\tilde{n}}}$ respectively.

We estimate $\hat{\epsilon}_{\mathbf{ik}} = \mathbf{Y_{ik}} - \mathbf{S_{ik}} - F\mathbf{h_k}$.

Step 3 - Update σ

We update σ and λ using the Metropolis-Hastings algorithm [3]. At the m-th MCMC progression generate $\sigma_* = |\sigma_{(m-1)} + \Phi|$, where $\Phi \sim U[-1, 1]$. The joint likelihood for the two variance parameters is

$$p(\lambda, \sigma_* | \mathbf{Y_{ik}})$$
$$\propto \prod_{i=1}^{r}\prod_{k=1}^{\tilde{n}} p\left(\mathbf{Y_{ik}} | \alpha_{*\mathbf{k}}, \mathbf{S_{ik}}, \sigma^2_*\right) \prod_{k=1}^{\tilde{n}} p\left(\alpha_{*\mathbf{k}} | \lambda, \sigma^2_*\right) p\left(\lambda, \sigma^2_*\right)$$
$$= \frac{\lambda^{\tilde{n}^2/2}}{\sigma_*^{(r+1)\tilde{n}^2}} \exp\left\{-\frac{1}{2\sigma^2_*}\left[\sum_{i=1}^{r}\sum_{k=1}^{\tilde{n}} \mathbf{c_{ik}}'\mathbf{c_{ik}} + \lambda\sum_{k=1}^{\tilde{n}} \mathbf{d_k}'\mathbf{d_k}\right]\right\}$$

where

$$\mathbf{c_{ik}} = \mathbf{Y_{ik}} - \mathbf{S_{ik}} - F\alpha_{*\mathbf{k}}, \qquad \mathbf{d_k} = \alpha_{*\mathbf{k}} - G\alpha_{*\mathbf{(k-1)}}$$

We can safely assume our prior probabilities are independent so that $p(\lambda, \sigma^2_*) = p(\lambda)p(\sigma^2_*)$ and proportional to 1. To generate the $\alpha_{*\mathbf{k}}$ we need to repeat steps 1 and 2 with the updated σ_*. We then accept σ_* with probability $\min(1, r_\sigma)$ where

$$r_\sigma = \exp\left\{\frac{L(\lambda_{(m-1)}, \sigma_*)}{L(\lambda_{(m-1)}, \sigma_{(m-1)})}\right\}$$

and $L(\lambda_{(m-1)}, \sigma_*) = \log p\left(\lambda_{(m-1)}, \sigma_* | \mathbf{Y_{ik}}\right)$. Otherwise $\sigma_{(m)} = \sigma_{(m-1)}$. If σ_* is accepted then so are the $\alpha_{*\mathbf{k}}$ and $\hat{\epsilon}_{*\mathbf{ik}}$.

Step 4 - Update λ

Using the same logic as the previous step we now accept λ_* with probability $\min(1, r_\lambda)$ where

$$r_\lambda = \exp\left\{\frac{L(\lambda_*, \sigma_{(m)})}{L(\lambda_{(m-1)}, \sigma_{(m)})}\right\}$$

Step 5 - Update subject effects

We again use rejection sampling to estimate the subject effects.

A function that respects the symmetries of the bispectrum as well as providing a range three-dimensional shapes is

$$S_{ijk} = \phi_0(B_{i,0}, \omega_j, \omega_k) + \phi_1(B_{i,1}, \omega_j, \omega_k) \qquad (2)$$

where

$$\begin{aligned}\phi_0(B_{i,0},\omega_j,\omega_k) &= B_{i,0}\\ \phi_1(B_{i,1},\omega_j,\omega_k) &= \sin(B_{i,1}\omega_j)\sin(B_{i,1}\omega_k)\end{aligned}$$

We can see that the first term controls the position on the z-axis and the other term control the shape. Note that setting $B_{i,s}=0,\ s=0,1$, gives $S_{ijk(0)}=0$.

Starting with the first subject we generate $B_{*1,0} = B_{1,0(m-1)} + \Phi$. And then calculate the new surface using (2) to give an updated set of surfaces

$$\mathbf{S_{*jk}} = [S_{*1jk}, S_{2jk(m)}, \ldots, S_{rjk(m)}]$$

Terms yet to be updated revert to their previous values so $S_{ijk(m)} = S_{ijk(m-1)},\ i=2,\ldots r$. The required likelihood is

$$\begin{aligned}p(\mathbf{S_{*ik}}|\mathbf{Y_{ik}}) &\propto p(\mathbf{Y_{ik}}|\mathbf{S_{*ik}})\,p(\mathbf{S_{*ik}})\\ &= \frac{1}{\sigma^{r\tilde{n}^2}}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{r}\sum_{k=1}^{\tilde{n}}\mathbf{c_{*ik}}'\mathbf{c_{*ik}}\right]\right\}\end{aligned}$$

where

$$\mathbf{c_{*ik}} = \mathbf{Y_{ik}} - \mathbf{S_{*ik}} - F\alpha_{\mathbf{k}}$$

and $p(\mathbf{S_{*ik}}) = 1$. We accept $\mathbf{S_{*ik}}$ with probability $\min(1, r_S)$ where

$$r_S = \exp\left\{\frac{L^S(\mathbf{S_{*ik}})}{L^S(\mathbf{S_{ik(m)}})}\right\}$$

and $L^S(\mathbf{S_{*ik}}) = \log p(\mathbf{S_{*ik}}|\mathbf{Y_{ik}})$. If the new value is accepted then $S_{ijk(m)} = [S_{*1jk}, S_{2jk(m)}, \ldots, S_{rjk(m)}]$, otherwise $S_{ijk(m)} = [S_{1jk(m)}, S_{2jk(m)}, \ldots, S_{rjk(m)}]$. The step is then repeated for $B_{1,1}$ and then the procedure repeated for the next subject. Again if a new $B_{i,s}$ is accepted then so are the associated $\alpha_{*\mathbf{k}}$ and $\hat{\epsilon}_{*\mathbf{ik}}$.

We repeat steps 1 to 5 M times. We then assess whether the estimates have converged, disregard the initial burn-in of the chain and give estimates for λ and σ and plot their marginal densities. The plot of the common bispectrum is used to identify the type of non-linearity whilst the degree of subject heterogeneity is a measure of the deviation from this overall norm.

3. RESULTS

A group of four people were recorded speaking the letter 'A', and one person repeated the letter four times. These signals were then resampled at 1/20 of the original sample rate to give a shorter series. To make all signals the same length they were tapered with zeros. We used $n=250$, $r=4, \gamma=10$ and $M=200$. The common bispectrum for the two data sets are shown

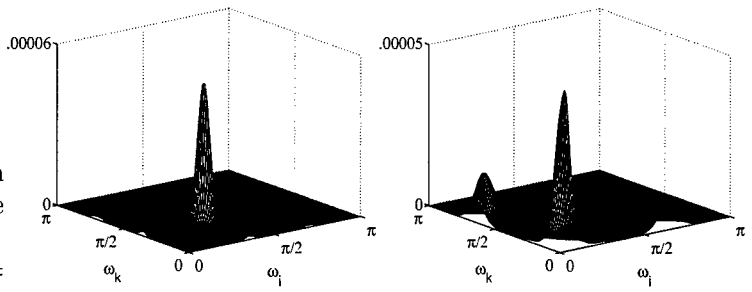


Figure 1: Common bispectrum for repeated and group data

in figure 1. For the repeated speaker $\lambda=0.87$, $\sigma=2.2$, $B_{i,1}=[2.1,-0.97,-0.86,-0.24]$, for the group data $\lambda=0.83$, $\sigma=3.2$, $B_{i,1}=[6.2,-4.0,1.9,8.9]$. For the single speaker the subject effects are generally smaller than the grouped data indicating that the bispectrum (and hence the third order non-linearity) is similar for all samples. For the grouped data the subject effects are much larger indicating that they do not conform to an overall common bispectrum. The shape of the population bispectrum for the single speaker is consistent with a Bilinear model with the non-linear term at lags 2 and 4 ($X_t = \beta_1 X_t X_{t-2} + \beta_2 X_t X_{t-4}$). For the grouped data the important lag appears to be at 1.

Future work will look for a common bispectrum in the Mel frequency Cepstral Coefficients.

4. REFERENCES

[1] Adrian G. Barnett and Rodney C. Wolff. A bootstrap test to detect non-linearity in time series based on the bispectrum. 2001. Submitted to Journal of Time Series Analysis.

[2] P.J Diggle and I Al-Wasel. Spectral analysis of replicated biomedical time series. *Journal of The Royal Statistical Society Series C (Applied Statistics)*, 46(1):31–71, 1997.

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall texts in Statistical Science series. Chapman & Hall, London, 1995.

[4] Nobuhisa Kashiwagi. On use of the kalman filter for spatial smoothing. *Annals of the Institute of Statistical Mathematics*, 45(1):21–34, 1993.

[5] A Rivola and P.R White. Bispectral analysis of the bilinear oscillator with application to the detection of fatigue cracks. *Journal of Sound and Vibration*, 216(5):889–910, 1998.

TIME DELAY ESTIMATION FOR MULTIPATH CDMA-SYSTEMS BASED ON A FAST MINIMIZATION TECHNIQUE FOR SUBSPACE FITTING

Patrik Bohlin, Anders Ranheim and Per Pelin

Department of Signals and Systems,
Chalmers University of Technology
E-mail: pbohlin@s2.chalmers.se

ABSTRACT

A low complexity algorithm for parameter estimation in a block fading multipath DS-CDMA system is presented. The main contribution in this paper is a novel technique for minimizing a subspace fitting criterion which is obtained as a large sample approximation of the Maximum Likelihood (ML) estimator. The minimization procedure is based on an approximation of the exact criterion function, allowing a direct analytic solution.

For the acquisition phase, an initialization procedure similar to alternating projections is employed which exhibits remarkable global convergence properties. The efficacy of the proposed method is demonstrated by means of numerical simulations.

1. INTRODUCTION

One of the main concerns in CDMA-systems is the near-far problem, i.e. that the signal received from different users have very dissimilar power levels. If the signals are nonorthogonal, as is often the case in real systems, conventional detectors such as the matched filter are known to deteriorate rapidly as the ratio between the power levels increase. To overcome this problem, a number of near-far resistant multiuser detectors have been proposed, see e.g. [5]. These detectors often assume various degrees of knowledge regarding channel parameters, such as time-delays, complex amplitudes and noise variances. It has also been observed that the performance of multiuser detectors are highly sensitive to the quality of channel parameter estimates [2], in particular to errors in the time-delays. This has led to the development of a number of near-far resistant time-delay estimators in the DS-CDMA context (see e.g. [7] and references therein).

In this paper, we are considering a single user approach, where the interfering users and the background noise are treated as temporally white Gaussian noise with an unknown spatial covariance. By deriving a large sample approximation of the Maximum Likelihood (ML) estimator as in [4, 7], the resulting criterion function has the structure of a subspace fitting problem. The novel idea presented here is how to search for the minimum of this function. In short, it is based on a linearization of the criterion function [3] around a prespecified number of points in the parameter-space (typically equal to the number of chips/symbol). In the case of a multidimensional parameter-space (multipath), the search can be decoupled into several one-dimensional minimization problems, in a similar fashion as in the Alternating Projection (AP) approach described in [8]. This point obviously has important implications with regards to the computational complexity. Since the proposed method is based on a linearization of a quadratic error-criterion, it is naturally interpreted as a Gauss-Newton step, being performed in a number of grid points.

The following section contains a brief description of the signal model being used, as well as relevant assumptions. Section 3 outlines the derivation of the criterion function to be minimized, and describes how the linearization leads to a closed form expression for the parameter estimates. Finally, the results of numerical simulations are presented in Section 4, together with concluding remarks.

2. SYSTEM MODEL

Consider a K-user asynchronous DS-CDMA system operating in a slowly fading multipath environment i.e. the fading is constant during the observation interval. All transmitted symbols are members of some complex symbol alphabet $\mathbf{\Omega}$ and have duration T. The code waveforms are assumed to be of unit energy and have zero support outside $[0, T\rangle$. Each chip has duration $T_c = T/L$, where L is the processing gain. After down conversion, IQ-demodulation and an integrate and dump stage with integration time T_c, the discrete baseband formulation of L consecutive samples (viz. one symbol interval) of the received signal form user k can at symbol interval n be written as

$$\mathbf{r}_k(n) = \mathbf{H}_k \mathbf{B}_k \mathbf{z}_k(n) \qquad (1)$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

where

$$\mathbf{H}_k \triangleq \begin{bmatrix}\mathbf{h}_{kl}(\tau_{k,1}) & \mathbf{h}_{kr}(\tau_{k,1}) & \dots & \mathbf{h}_{kl}(\tau_{k,R_k}) & \mathbf{h}_{kr}(\tau_{k,R_k})\end{bmatrix}$$

$$\mathbf{B}_k \triangleq \begin{bmatrix}\beta_{k,1} & 0 \\ 0 & \beta_{k,1} \\ \vdots & \vdots \\ \beta_{k,R_k} & 0 \\ 0 & \beta_{k,R_k}\end{bmatrix} \qquad \mathbf{z}_k(n) \triangleq \begin{bmatrix}d_k(n-1) \\ d_k(n)\end{bmatrix}.$$

Here, R_k denotes the number of multipath components from user k, $\beta_{k,1}$ is the complex path gain (for the first path) and d_k represents the transmitted bits. Furthermore, $\mathbf{h}_{kl}$ and $\mathbf{h}_{kr}$ are functions of the path delays τ and code waveforms $\mathbf{c}_k$ [4, 7];

$$\mathbf{h}_{kl}(\tau) = (1-\delta)\mathcal{T}_L^{L-p}\mathbf{c}_k + \delta\mathcal{T}_L^{L-p-1}\mathbf{c}_k \tag{2}$$

$$\mathbf{h}_{kr}(\tau) = (1-\delta)\mathcal{T}_R^{p+1}\mathbf{c}_k + \delta\mathcal{T}_R^{p}\mathbf{c}_k \tag{3}$$

$$\mathbf{c}_k = \begin{bmatrix}c_k(1) & c_k(2) & \cdots & c_k(L)\end{bmatrix} \tag{4}$$

where $p = \left\lfloor \frac{\tau}{T_c} \right\rfloor$, $\delta = \frac{\tau}{T_c} - p$ and $\mathcal{T}_L^p$ and $\mathcal{T}_R^p$ are the p-step left- and right acyclic shift operators defined as

$$\mathcal{T}_L^p[x_1,\dots,x_N] \triangleq [x_{p+1},\dots,x_N,0,\dots,0] \tag{5}$$

$$\mathcal{T}_R^p[x_1,\dots,x_N] \triangleq [0,\dots,0,x_1,\dots,x_{N-p}]. \tag{6}$$

Collecting the contributions from all K users, the total received vector will be

$$\begin{aligned}\mathbf{r}(n) &= \sum_{k=1}^{K}\mathbf{r}_k(n) + \mathbf{n}(n) \\ &= \mathbf{H}_1\mathbf{B}_1\mathbf{z}_1(n) + \sum_{k=2}^{K}\mathbf{H}_k\mathbf{B}_k\mathbf{z}_k(n) + \mathbf{n}(n) \\ &= \mathbf{H}_1\mathbf{B}_1\mathbf{z}_1(n) + \mathbf{j}(n).\end{aligned} \tag{7}$$

The superposition of multiuser interference and background noise, is modelled as a zero-mean complex Gaussian random process with second-order moments

$$E\{\mathbf{j}(n_1)\mathbf{j}^*(n_2)\} = \mathcal{R}_{jj}\delta(n_1-n_2) \tag{8}$$

$$E\{\mathbf{j}(n_1)\mathbf{j}^T(n_2)\} = \mathbf{0}, \tag{9}$$

where $\mathcal{R}_{jj}$ is an unknown positive definite matrix.

3. ALGORITHM

3.1. Criterion Function

Invoking the assumptions stated above, the negative log-likelihood function of the received data $\{\mathbf{r}(n)\}_{n=1}^N$ is proportional to

$$l(\boldsymbol{\tau},\boldsymbol{\beta},\mathcal{R}_{jj}) = \log|\mathcal{R}_{jj}| + \mathrm{Tr}\left\{\mathcal{R}_{jj}^{-1}\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{r}(n)-\mathbf{D}\mathbf{z}(n)\}\{\mathbf{r}(n)-\mathbf{D}\mathbf{z}(n)\}^*\right\}. \tag{10}$$

where $|\cdot|$ denotes the determinant of a matrix and $\mathbf{D} \triangleq \mathbf{H}\mathbf{B}$. The user index 1 has been dropped for notational convenience, since we are considering an arbitrary user. Elimination of $\mathcal{R}_{jj}$ gives the following criterion function

$$l(\boldsymbol{\tau},\boldsymbol{\beta}) = \left|\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{r}(n)-\mathbf{D}\mathbf{z}(n)\}\{\mathbf{r}(n)-\mathbf{D}\mathbf{z}(n)\}^*\right| \tag{11}$$

from which an unstructured estimate of $\mathbf{D}$ can be obtained as [6]

$$\hat{\mathbf{D}} = \hat{\mathcal{R}}_{zr}^*\hat{\mathcal{R}}_{zz}^{-1} \tag{12}$$

and a consistent estimate of $\hat{\mathcal{R}}_{jj}$ as

$$\hat{\mathcal{R}}_{jj} = \hat{\mathcal{R}}_{rr} - \hat{\mathcal{R}}_{zr}^*\hat{\mathcal{R}}_{zz}^{-1}\hat{\mathcal{R}}_{zr}. \tag{13}$$

Here, $\hat{\mathcal{R}}_{rr} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{r}(n)\mathbf{r}^*(n)$, and $\hat{\mathcal{R}}_{zz}$ and $\hat{\mathcal{R}}_{zr}$ are defined similarly.

Minimizing (11) can be shown [4] to be asymptotically equivalent to minimizing

$$l(\boldsymbol{\tau},\boldsymbol{\beta}) = ||\tilde{\mathbf{D}} - \tilde{\mathbf{H}}\mathbf{B}||_F^2 \tag{14}$$

where $\tilde{\mathbf{D}} \triangleq \hat{\mathcal{R}}_{jj}^{-\frac{1}{2}}\hat{\mathbf{D}}$ and $\tilde{\mathbf{H}} \triangleq \hat{\mathcal{R}}_{jj}^{-\frac{1}{2}}\hat{\mathbf{H}}$ are the prewhitened estimates. By exploiting the stacked diagonal structure of the matrix $\mathbf{B}$, the cost function can be reformulated as

$$l(\boldsymbol{\tau},\boldsymbol{\beta}) = ||\mathrm{vec}\,(\tilde{\mathbf{D}}) - (\mathbf{I}\otimes\tilde{\mathbf{H}})\mathbf{T}\boldsymbol{\beta}||^2 \tag{15}$$

where

$$\mathbf{T} = \begin{pmatrix}1 & 0 & \cdots \\ 0 & 0 & \cdots \\ 0 & 1 & \cdots \\ 0 & 0 & \cdots \\ & \cdots & \\ 0 & 0 & \cdots \\ 1 & 0 & \cdots \\ 0 & 0 & \cdots \\ 0 & 1 & \cdots \\ 0 & 0 & \cdots \\ & \cdots & \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix}\beta_1 \\ \vdots \\ \beta_{R_k}\end{pmatrix}. \tag{16}$$

If $\boldsymbol{\Upsilon}(\boldsymbol{\tau}) \triangleq (\mathbf{I}\otimes\tilde{\mathbf{H}})\mathbf{T}$, the criterion function becomes

$$l(\boldsymbol{\tau},\boldsymbol{\beta}) = ||\mathrm{vec}(\tilde{\mathbf{D}}) - \boldsymbol{\Upsilon}(\boldsymbol{\tau})\boldsymbol{\beta}||^2. \tag{17}$$

Furthermore, if the estimate of the complex path gains[1]

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Upsilon}^\dagger(\boldsymbol{\tau})\,\mathrm{vec}(\tilde{\mathbf{D}}) \tag{18}$$

is substituted back into (17), the desired criterion, as a function of the time-delay parameters, will finally be

$$\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}}||\boldsymbol{\Pi}_{\boldsymbol{\Upsilon}}^{\perp}(\boldsymbol{\tau})\,\mathrm{vec}(\tilde{\mathbf{D}})||^2. \tag{19}$$

Here, $\boldsymbol{\Pi}_{\boldsymbol{\Upsilon}}^{\perp}(\boldsymbol{\tau}) = \mathbf{I} - \boldsymbol{\Upsilon}(\boldsymbol{\tau})\boldsymbol{\Upsilon}^\dagger(\boldsymbol{\tau})$ is the projection matrix projecting onto the orthogonal complement of span $(\boldsymbol{\Upsilon}(\boldsymbol{\tau}))$.

[1] $\dagger$ denotes the Moore-Penrose pseudoinverse.

3.2. Minimization Procedure

In what follows, we will describe a novel approach for the minimizing the criterion in (19) that leads to a closed form expression for the minimizing argument, $\hat{\boldsymbol{\tau}}$.

Given an estimate $\boldsymbol{\tau}^q$ in the close vicinity of the global minimizer $\boldsymbol{\tau}^*$, consider the first order Taylor-series expansion of the projection matrix around $\boldsymbol{\tau}^q$ (such that $\boldsymbol{\tau}^* \cong \boldsymbol{\tau}^q + \tilde{\boldsymbol{\tau}}$)

$$\mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp}(\boldsymbol{\tau}^*) \cong \mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp}(\boldsymbol{\tau}^q) + \sum_i \tilde{\tau}_i \frac{\partial}{\partial \tau_i} \mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp}(\boldsymbol{\tau})\Big|_{\boldsymbol{\tau}^q} \tag{20}$$

In order to evaluate (20) we need to find the derivatives of $\boldsymbol{\Upsilon}(\boldsymbol{\tau})$. So, by noting that

$$\boldsymbol{\Upsilon}(\boldsymbol{\tau}) = \begin{bmatrix} \hat{\mathcal{R}}_{jj}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathcal{R}}_{jj}^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{h}_l(\tau_1) & \dots & \mathbf{h}_l(\tau_R) \\ \mathbf{h}_r(\tau_1) & \dots & \mathbf{h}_r(\tau_R), \end{bmatrix} \tag{21}$$

the derivative of $\boldsymbol{\Upsilon}(\boldsymbol{\tau})$ with respect to τ_i will be

$$\mathbf{G}_i = \frac{\partial}{\partial \tau_i} \boldsymbol{\Upsilon}(\boldsymbol{\tau}) = \begin{bmatrix} \mathbf{0}, \dots, \hat{\mathcal{R}}_{jj}^{-\frac{1}{2}} \mathbf{g}_l(\tau_i), \dots, \mathbf{0} \\ \mathbf{0}, \dots, \hat{\mathcal{R}}_{jj}^{-\frac{1}{2}} \mathbf{g}_r(\tau_i), \dots, \mathbf{0} \end{bmatrix}. \tag{22}$$

The derivatives of $\mathbf{h}_l$ and $\mathbf{h}_r$ with respect to τ_i on each chip interval are easily found, using (2) and (3), as

$$\frac{\partial}{\partial \tau} \mathbf{h}_l(\tau) = \mathbf{g}_l(\tau) = -\mathcal{T}_L^{L-p}\mathbf{c} + \mathcal{T}_L^{L-p-1}\mathbf{c} \tag{23}$$

$$\frac{\partial}{\partial \tau} \mathbf{h}_r(\tau) = \mathbf{g}_r(\tau) = -\mathcal{T}_R^{p+1}\mathbf{c} + \mathcal{T}_R^{p}\mathbf{c} \tag{24}$$

where $p = \left\lfloor \frac{\tau}{T_c} \right\rfloor$. Note that the derivative is undefined on the chip borders so we will have a piecewise linear derivative. With this notation, the projection matrix at the global minimum can be written as

$$\mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp}(\boldsymbol{\tau}^q + \tilde{\boldsymbol{\tau}}) \cong \mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp} + \boldsymbol{\Upsilon}^{\dagger *} \mathbf{\Delta} \mathbf{G}^* \mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp} + \mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp} \mathbf{G} \mathbf{\Delta} \boldsymbol{\Upsilon}^{\dagger}, \tag{25}$$

where $\mathbf{G} = \sum_{i=1}^{R_k} \mathbf{G}_i$ and $\mathbf{\Delta} = \mathrm{diag}(\tilde{\boldsymbol{\tau}})$. The notational dependence of $\boldsymbol{\tau}^q$ has also been dropped in $\mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp}$, $\boldsymbol{\Upsilon}$ and $\mathbf{G}$. Hence an approximate criterion function in the variable $\tilde{\boldsymbol{\tau}}$ is obtained as

$$V(\tilde{\boldsymbol{\tau}}) = ||\mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp}(\boldsymbol{\tau}^q + \tilde{\boldsymbol{\tau}}) \,\mathrm{vec}(\tilde{\mathbf{D}})||_F^2. \tag{26}$$

This criterium can be minimized with respect to $\tilde{\boldsymbol{\tau}}$ to find the increments to $\boldsymbol{\tau}^q$. Therefore we define

$$\mathbf{f} = \mathrm{vec}\left\{ \mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp} \,\mathrm{vec}(\tilde{\mathbf{D}}) \right\} \tag{27}$$

$$\mathbf{A} = \left(\mathbf{G}^* \mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp} \,\mathrm{vec}(\tilde{\mathbf{D}}) \right)^T \Diamond \boldsymbol{\Upsilon}^{\dagger *} + \left(\boldsymbol{\Upsilon}^{\dagger} \,\mathrm{vec}(\tilde{\mathbf{D}}) \right)^T \Diamond (\mathbf{G} \mathbf{\Pi}_{\boldsymbol{\Upsilon}}^{\perp}), \tag{28}$$

where $\Diamond$ is the Khatri-Rao product, i.e. columnwise Kronecker product [1]. Using these definitions it is possible to rewrite (26) as

$$V(\tilde{\boldsymbol{\tau}}) = ||\mathbf{f} - \mathbf{A}\tilde{\boldsymbol{\tau}}||_F^2 \tag{29}$$

from which the increments $\tilde{\boldsymbol{\tau}}$ easily are computed as

$$\hat{\tilde{\boldsymbol{\tau}}} = \mathbf{A}^{\dagger}\mathbf{f}. \tag{30}$$

Note that to enforce a real solution the real and imaginary parts of $\mathbf{f}$ and $\mathbf{A}$ can be stacked. The new estimate of $\boldsymbol{\tau}$ will then be

$$\boldsymbol{\tau}^{q+1} = \boldsymbol{\tau}^q + \hat{\tilde{\boldsymbol{\tau}}} \tag{31}$$

This procedure can then be iterated with the last estimate as the new starting value.

Acquisition: In order to find the global minimum of the criterion function (19), we propose to work on the linearized version derived above as follows;

1) Since the derivative of the projection matrix is only piecewise continuous with discontinuities on multiples of chip intervals i.e. $\tau = pT_c$, we have to evaluate the criterion function on each chip-transition as well as solving (30) on a grid of at least L points, i.e one starting-point per chip interval. A closer spacing will improve performance at the expense of increased complexity. Of all the candidate solutions obtained in this way, we select the time-delay corresponding to the smallest value of the criterion function. If desirable, evaluate (30) with the last estimate as the new starting point as long as a significant improvement occurs.

2) For multipath, i.e. $R > 1$, the previous estimates can be used as starting point(s), while carrying out the same procedure as outlined above.

This is akin to the initialization procedure used in the Alternating Projection (AP) algorithm [8], with one very important distinction; Whereas AP turns a D-dimensional grid search into D 1-dimensional grid searches, we instead propose to carry out what amounts to a Gauss-Newton (GN) step in each grid point. The results of the numerical simulations presented in the next section will clearly demonstrate the advantage of this approach.

4. NUMERICAL SIMULATIONS

In the simulations presented here we consider a $K = 5$ user scenario where each user has an $R_k = 2$ path channel and is assigned an $N = 15$ chips per bit Gold-like code sequence. Both paths for the *user of interest* have the same strength, i.e. $|\beta_{1,1}| = |\beta_{1,2}|$. Further, all interfering signals have the same amplitude, defined by the near-far ratio $|\beta_{2,1}|/|\beta_{1,1}|$, which is chosen to be either 0 or 20 dB. The proposed method, labelled GN, is compared to the method proposed in [4], here labelled AP. In Figure 1 the Root Mean-Square Error (RMSE) of the estimated time-delays is plotted versus the Signal-to-Noise Ratio SNR$\triangleq 1/\sigma^2 \sum_{r=1}^{R_k} |\beta_{1,r}|$. The number of training bits is set to 50. For clarity, only the RMSE for the first multipath component is shown, but the same

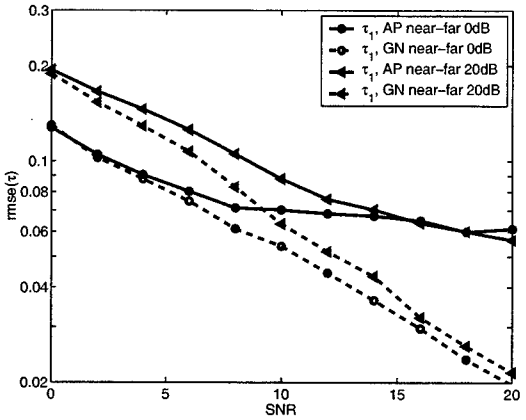


Figure 1: The RMSE as a function of the SNR

trends can be observed also for the estimates of the second component. From the figure we note that the difference in performance increases with the SNR, and one might conclude that the algorithms will have comparable performance at low SNR. This is not entirely correct as all outliers have been removed in these plots. The outliers correspond to those events where the estimate is so poor, here defined as $|\hat{\tau}_1 - \tau_1| > 0.5T_c$, that the estimate is useless and the acquisition fails. The probability of acquisition failure is given in Table 1 and 2. It can be seen that failure to acquire the paths is only a problem in scenarios with high levels of interference combined with low SNR.

The performance of the estimators also depends on the number of known symbols. In Figure 2 this is shown for the same scenario as above. The SNR was set to 10 dB. Again, these results support the conclusions that

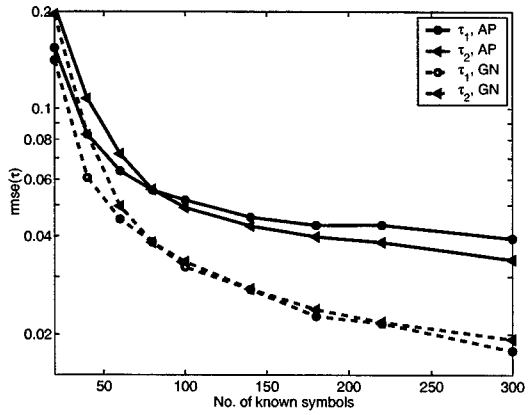


Figure 2: The RMSE as a function of the number of known symbols

the proposed method for minimizing the subspace fitting criterion function in (19) clearly outperforms the alternating projection approach, while maintaining the moderate complexity.

$\frac{\lvert\beta_{2,1}\rvert}{\lvert\beta_{1,1}\rvert}$	0 dB		20 dB	
path	$\hat{\tau}_1(1)$	$\hat{\tau}_1(2)$	$\hat{\tau}_1(1)$	$\hat{\tau}_1(2)$
SNR				
2	2%	1%	26 %	9 %
8	~ 0 %	0.5 %	2 %	1 %
14	~ 0 %	~ 0%	~ 0 %	~ 0%

Table 1: Probability of acquisition failure, AP.

$\frac{\lvert\beta_{2,1}\rvert}{\lvert\beta_{1,1}\rvert}$	0 dB		20 dB	
path	$\hat{\tau}_1(1)$	$\hat{\tau}_1(2)$	$\hat{\tau}_1(1)$	$\hat{\tau}_1(2)$
SNR				
2	1%	1%	10 %	4 %
8	~ 0 %	~ 0%	1 %	~ 0%
14	~ 0 %	~ 0%	~ 0 %	~ 0%

Table 2: Probability of acquisition failure, GN.

5. REFERENCES

[1] J. Brewer. "Kronecker Products and Matrix Calculus in System Theory". *IEEE Transactions on Circuits and Systems*, Vol.CAS-25(9):pp.772–781, September 1978.

[2] S. Parkvall, E. Ström, and B. Ottersten. "The Impact of Timing Errors on the Performance of Linear DS-CDMA Receivers". *IEEE Journal on Selected Areas in Communications*, Vol.14(8):pp.1660–1668, October 1996.

[3] P. Pelin. "A Fast Minimization Technique for Subspace Fitting with Arbitrary Array Manifolds". *Preprint. To appear in IEEE Transactions on Signal Processing.*

[4] A. Ranheim and P. Pelin. "Joint Symbol Detection and Parameter Estimation in Asynchronous DS-CDMA Systems". *IEEE Transactions on Signal Processing*, Vol.48 (2):pp.545–550, February 2000.

[5] Sergio Verdú. *Multiuser Detection.* Cambridge University Press, 1998.

[6] M. Viberg, P. Stoica, and B. Ottersten. "Maximum Likelihood Array Processing in Spatially Correlated Noise Fields Using Parametrized Signals". *IEEE Transactions on Signal Processing*, April 1997.

[7] D. Zheng, J. Li, S.L. Miller, and E.G. Ström. "An Efficient Code-Timing Estimator for DS-CDMA Signals". *IEEE Transactions on Signal Processing*, Vol.45(1):pp.82–89, January 1997.

[8] I. Ziskind and M. Wax. "Maximum Likelihood Localization of Multiple Sources by Alternating Projection". *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol.36(10):pp.1553–1560, October 1988.

BLIND SEPARATION OF SECOND-ORDER NONSTATIONARY AND TEMPORALLY COLORED SOURCES

Seungjin Choi [§†], *Andrzej Cichocki* [†], *Adel Belouchrani* [‡]

[§] Department of Computer Science & Engineering, POSTECH, Korea
seungjin@postech.ac.kr
[†] Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, Japan
cia@brain.riken.go.jp
[‡] Department of Electrical Engineering, Ecole Nationale Polytechnique, Algeria
belouchrani@hotmail.com

ABSTRACT

This paper presents a method of blind source separation that jointly exploits the nonstationarity and temporal structure of sources . The method needs only multiple time-delayed correlation matrices of the observation data, each of which is evaluated at different time-windowed data frame, to estimate the demixing matrix. We show that the method is quite robust with respect to the spatially correlated but temporally white noise. We also discuss the extension of some existing second-order blind source separation methods. Extensive numerical experiments confirm the validity of the proposed method.

1. INTRODUCTION

Blind source separation (BSS) is a fundamental problem that is encountered in many practical applications such as telecommunications, image/speech processing, and biomedical signal analysis where multiple sensors are involved. In its simplest form, the m-dimensional observation vector $\boldsymbol{x}(t) \in \mathbb{R}^m$ is assumed to be generated by

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{v}(t), \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is the unknown mixing matrix, $\boldsymbol{s}(t)$ is the n-dimensional source vector (which is also unknown and $n \leq m$), and $\boldsymbol{v}(t)$ is the additive noise vector that is statistically independent of $\boldsymbol{s}(t)$.

A variety of methods/algorithms for BSS have been developed for last decade (for example, see [1] and references therein). Although many different BSS algorithms are available, their principles can be categorized by three distinctive methods which are based on (1) the non-Gaussianity of source [2], (2) the temporal structure of source [3], and (3) the nonstationarity of source [4].

In this paper we present methods that jointly exploits the nonstationarity and temporal structure of sources to estimate the mixing matrix (or the demixing matrix) in the presence of spatially correlated but temporally white noise (not necessarily Gaussian). Thus our methods works even for the case where multiple Gaussian sources with no temporal correlations exit as long as their variances are slowly time-varying. Moreover, we show that if we use just time-delayed correlations of the observation data, we can find a robust estimate of the demixing matrix. To this end, we introduce a method of robust whitening and present the Second-Order Nonstationary source Separation (SEONS) method. We also present the extension of some existing second-order BSS methods which are (1) the extended matrix pencil method and (2) the extended Pham-Cardoso method.

Throughout this paper, the following assumptions are made:

(AS1) The mixing matrix $\boldsymbol{A}$ is of full column rank.

(AS2) Sources are spatially uncorrelated but are temporally correlated (colored) stochastic signals with zero mean.

(AS3) Sources are second-order nonstationary signals in the sense that their variances are time varying.

(AS4) Additive noises $\{v_i(t)\}$ are spatially correlated but temporally white, i.e.,

$$E\{\boldsymbol{v}(t)\boldsymbol{v}^T(t-\tau)\} = \delta_\tau \boldsymbol{\Gamma}, \tag{2}$$

where δ_τ is the Kronecker symbol and $\boldsymbol{\Gamma}$ is an arbitrary $m \times m$ matrix.

2. ROBUST WHITENING

The whitening (or data sphering) is an important pre-processing step in a variety of BSS methods. The conventional whitening exploits the equal-time correlation matrix of the data $\boldsymbol{x}(t)$, so that the effect of additive noise can not be removed. The idea of robust whitening lies in utilizing the time-delayed correlation matrices that are not sensitive to the additive white noise. The robust whitening method is explained for the case of stationary signals.

It follows from the assumptions (AS2) and (AS4) that the time-delayed correlation matrix of the observation data $\boldsymbol{x}(t)$ has the form

$$\begin{aligned} \boldsymbol{R}_x(\tau) &= E\{\boldsymbol{x}(t)\boldsymbol{x}^T(t-\tau)\} \\ &= \boldsymbol{A}\boldsymbol{R}_s(\tau)\boldsymbol{A}^T, \end{aligned} \tag{3}$$

for $\tau \neq 0$. One can easily see that the transformation $\boldsymbol{R}_x^{-\frac{1}{2}}(\tau)$ whiten the data $\boldsymbol{x}(t)$ without the effect of the noise vector $\boldsymbol{v}(t)$. It reduces the noise effect and project the data onto the signal subspace, in contrast to the conventional whitening transformation $\boldsymbol{R}_x^{-\frac{1}{2}}(0)$. Some source separation methods already employ this robust whitening transformation [5, 6, 7, 8].

0-7803-7011-2/01/$10.00 ©2001 IEEE

In general, however, the matrix $\boldsymbol{R}_x(\tau)$ is not always positive definite, so the whitening transformation $\boldsymbol{R}_x^{-\frac{1}{2}}(\tau)$ may not be valid for some time-lag τ. The idea of the robust whitening is to consider a linear combination of several time-delayed correlation matrices, i.e.,

$$\boldsymbol{C}_x = \sum_{i=1}^{K} \alpha_i \boldsymbol{M}_x(\tau_i), \tag{4}$$

where

$$\boldsymbol{M}_x(\tau_i) = \frac{1}{2}\left\{\boldsymbol{R}_x(\tau_i) + \boldsymbol{R}_x^T(\tau_i)\right\}. \tag{5}$$

A proper choice of $\{\alpha_i\}$ may result in a positive definite matrix $\boldsymbol{C}_x$. For example, the FSGC method [9] can be used to find a set of coefficients $\{\alpha_i\}$ such that the matrix $\boldsymbol{C}_x$ is positive definite.

The matrix $\boldsymbol{C}_x$ has the eigen-decomposition

$$\boldsymbol{C} = [\boldsymbol{U}_1, \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{D}_1 & \\ & \boldsymbol{0} \end{bmatrix} [\boldsymbol{U}_1, \boldsymbol{U}_2]^T, \tag{6}$$

where $\boldsymbol{U}_1 \in \mathbb{R}^{m\times n}$ and $\boldsymbol{D}_1 \in \mathbb{R}^{n\times n}$. Then the robust whitening transformation matrix is given by $\boldsymbol{Q} = \boldsymbol{D}_1^{-\frac{1}{2}}\boldsymbol{U}_1^T$. The transformation $\boldsymbol{Q}$ project the data onto n-dimensional signal subspace as well as whitening.

Let us denote the whitened n-dimensional data by $\boldsymbol{z}(t)$

$$\begin{aligned} \boldsymbol{z}(t) &= \boldsymbol{Q}\boldsymbol{x}(t) \\ &= \boldsymbol{B}\boldsymbol{s}(t) + \boldsymbol{Q}\boldsymbol{v}(t), \end{aligned} \tag{7}$$

where $\boldsymbol{B} \in \mathbb{R}^{n\times n}$. The whitened data $\boldsymbol{z}(t)$ is a unitary mixture of sources with additive noise, i.e., $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{I}$.

3. SECOND-ORDER NONSTATIONARY SOURCE SEPARATION

This section describes our main method, SEONS, as well as some extensions such as the extended matrix pencil method and the extended Pham-Cardoso method.

Now we consider the case where sources are second-order nonstationary and have non-vanishing temporal correlations. It follows from the assumptions (AS1)-(AS4) that we have

$$\boldsymbol{M}_x(t_k, \tau_i) = \boldsymbol{A}\boldsymbol{M}_s(t_k, \tau_i)\boldsymbol{A}^T, \tag{8}$$

for $\tau_i \neq 0$ and the index t_k is for time since we deal with nonstationary sources. In practice $\boldsymbol{M}_x(t_k, \tau_i)$ is computed using the samples in the kth time-windowed data frame, i.e.,

$$\boldsymbol{R}_x(t_k, \tau_i) = \frac{1}{N_k}\sum_{t\in\mathcal{N}_k} \boldsymbol{x}(t)\boldsymbol{x}^T(t-\tau_i), \tag{9}$$

$$\boldsymbol{M}_x(t_k, \tau_i) = \frac{1}{2}\left\{\boldsymbol{R}_x(t_k, \tau_i) + \boldsymbol{R}_x^T(t_k, \tau_i)\right\}, \tag{10}$$

where $\mathcal{N}_k$ contains the data points in the kth time-windowed frame and N_k is the number of data points in $\mathcal{N}_k$.

The matrix pencil method [4] was applied to the blind separation of temporally colored sources. In general, however, the pencil that consists of two time-delayed correlation matrices is not symmetric definite pencil, which may cause some numerical problems in calculating generalized eigenvectors. The extended matrix pencil method (which is described below) employs a symmetric definite pencil.

Algorithm Outline: Extended Matrix Pencil Method (nonstationary case)

1. We partition the observation data into two non-overlapping blocks, $\{\mathcal{N}_1, \mathcal{N}_2\}$.
2. Compute $\boldsymbol{M}_x(t_2, \tau_2)$ for some time-lag $\tau_2 \neq 0$ using the data points in $\mathcal{N}_2$.
3. Calculate the matrix $\boldsymbol{C}_1(t_1) = \sum_{i=1}^{J} \alpha_i \boldsymbol{M}_x(t_1, \tau_i)$ by the FSGC method using the data points in $\mathcal{N}_1$.
4. Find the generalized eigenvector matrix $\boldsymbol{V}$ of the pencil $\boldsymbol{M}_x(t_2, \tau_2) - \lambda\boldsymbol{C}_1(t_1)$ which satisfies
$$\boldsymbol{M}_x(t_2, \tau_2)\boldsymbol{V} = \boldsymbol{C}_1(t_1)\boldsymbol{V}\boldsymbol{\Lambda}. \tag{11}$$
5. The demixing matrix is given by $\boldsymbol{W} = \boldsymbol{V}^T$.

In order to improve the statistical efficiency, we can employ a joint approximate diagonalization method [10], as in the JADE [11] and SOBI [3]. The joint approximate diagonalization method in [10] finds an unitary transformation that jointly diagonalizes several matrices (which do not have to be symmetric nor positive definite). The method SEONS is based on this joint approximate diagonalization. In this sense the SEONS includes the SOBI as its special case (if sources are stationary). The algorithm is summarized below.

Algorithm Outline: SEONS

1. The robust whitening method (described in Section 2) is applied to obtain the whitened vector $\boldsymbol{z}(t) = \boldsymbol{Q}\boldsymbol{x}(t)$. In the robust whitening step, we used the whole available data points.
2. Divide the whitened data $\{\boldsymbol{z}(t)\}$ into K non-overlapping blocks and calculate $\boldsymbol{M}_z(t_k, \tau_j)$ for $k = 1, \ldots, K$ and $j = 1, \ldots, J$. In other words, at each time-windowed data frame, we compute J different time-delayed correlation matrices of $\boldsymbol{z}(t)$.
3. Find a unitary joint diagonalizer $\boldsymbol{V}$ of $\{\boldsymbol{M}_z(t_k, \tau_j)\}$ using the joint approximate diagonalization method in [10], which satisfies
$$\boldsymbol{V}^T\boldsymbol{M}_z(t_k, \tau_j)\boldsymbol{V} = \boldsymbol{\Lambda}_{k,j}, \tag{12}$$
where $\{\boldsymbol{\Lambda}_{k,j}\}$ is a set of diagonal matrices.
4. The demixing matrix is computed as $\boldsymbol{W} = \boldsymbol{V}^T\boldsymbol{Q}$.

Recently Pham [12] developed a joint approximate diagonalization method where non-unitary joint diagonalizer of several Hermitian positive matrices is computed by a way similar to the classical Jacobi method. Second-order nonstationarity was also exploited in [13], but only noise-free data was considered. In order to extend the Pham-Cardoso algorithm into the case of noisy data, we employ a linear combination of multiple time-delayed correlation matrices which is ensured to be positive definite, at each data block. The method is referred to as the extended Pham-Cardoso (which is summarized below). One advantage of the extended Pham-Cardoso is that it does not require the whitening step because the joint approximate diagonalization method in [13] finds a non-unitary joint diagonalizer. However, its drawback lies in the fact that it requires the set of matrices to be Hermitian and positive definite, so we need to find a linear combination of time-delayed

correlation matrices that is positive definite at each data frame, which increase the computational complexity.

Algorithm Outline: Extended Pham-Cardoso

1. Divide the data $\{\boldsymbol{x}(t)\}$ into K non-overlapping blocks and calculate $\boldsymbol{M}_x(t_k, \tau_j)$ for $k = 1, \ldots, K$ and $j = 1, \ldots, J$.
2. At each data frame, we compute

$$\boldsymbol{C}_k = \sum_{i=1}^{J} \alpha_i^{(k)} \boldsymbol{M}_x(t_k, \tau_i) \tag{13}$$

   by the FSGC method for $k = 1, \ldots, K$. Note that $\{\boldsymbol{C}_k\}$ is symmetric and positive definite.
3. Find a non-unitary joint diagonalizer $\boldsymbol{V}$ of $\{\boldsymbol{C}_k\}$ using the joint approximate diagonalization method in [12], which satisfies

$$\boldsymbol{V}\boldsymbol{C}_k\boldsymbol{V}^T = \boldsymbol{\Lambda}_k, \tag{14}$$

   where $\{\boldsymbol{\Lambda}_k\}$ is a set of diagonal matrices.
4. The demixing matrix is computed as $\boldsymbol{W} = \boldsymbol{V}$.

4. NUMERICAL EXPERIMENTS

Several numerical experimental results are presented to evaluate the performance of our method (SEONS) and to compare it with some existing methods such as JADE [11], SOBI [3], matrix pencil methods [4], and Pham-Cardoso [13]. Through numerical experiments, we confirm the useful behavior of the proposed method, SEONS, in two cases: (1) the case where several nonstationary Gaussian sources exist and each Gaussian source has no temporal correlation; (2) the case where additive noises are spatially correlated but temporally white Gaussian processes.

In order to measure the performance of algorithms, we use the performance index (PI) defined by

$$\begin{aligned} \text{PI} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left\{ \left(\sum_{k=1}^{n} \frac{|g_{ik}|}{\max_j |g_{ij}|} - 1 \right) \right. \\ &\quad \left. + \left(\sum_{k=1}^{n} \frac{|g_{ki}|}{\max_j |g_{ji}|} - 1 \right) \right\}, \end{aligned} \tag{15}$$

where g_{ij} is the (i,j)-element of the global system matrix $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{A}$ and $\max_j g_{ij}$ represents the maximum value among the elements in the ith row vector of $\boldsymbol{G}$, $\max_j g_{ji}$ does the maximum value among the elements in the ith column vector of $\boldsymbol{G}$. When the perfect separation is achieved, the performance index is zero. In practice, the value of performance index around 10^{-3} gives quite a good performance.

4.1. Experiment 1

The first experiment was designed to evaluate the effectiveness of the proposed method in the presence of several Gaussian signals. In this experiment, we used three speech signals that are sampled at 8 kHz and two Gaussian signals (with no temporal correlations) whose variances are slowly varying. These 5 sources were mixed using a randomly generated 5×5 mixing matrix to generate 5-dimensional observation vector with 10000 data points. No measurement noise was added.

In this experiment, we compared the SEONS with JADE, SOBI, and Pham-Cardoso [13]. It is expected that the performance of JADE and SOBI is degraded because of the presence of two white Gaussian sources. The result is shown in Fig. 1 in which the Hinton diagram of the global system matrix $\boldsymbol{G}$ is plotted. In Hinton diagram, each square's area represents the magnitude of the element of the matrix and each square's color represents the sign of the element (red for negative value and green for positive value). For successful separation, each row and column has only one dominant square (regardless of its color). Small squares contribute performance degradation. One can observe that SEONS and Pham-Cardoso work well even in the presence of nonstationary Gaussian sources (see (a) and (b) in Fig. 1), compared to JADE and SOBI (see (c) and (d) in Fig. 1). For the case of JADE, the first and last row of $\boldsymbol{G}$ has a relatively big square besides the dominant square, which verifies that the two white Gaussian sources are difficult to be separated out. The SOBI gives slightly better performance than JADE, but its performance is not comparable to SEONS (see the first and fourth row of $\boldsymbol{G}$, (d) in Fig. 1).

The following parameters were used in this experiment:

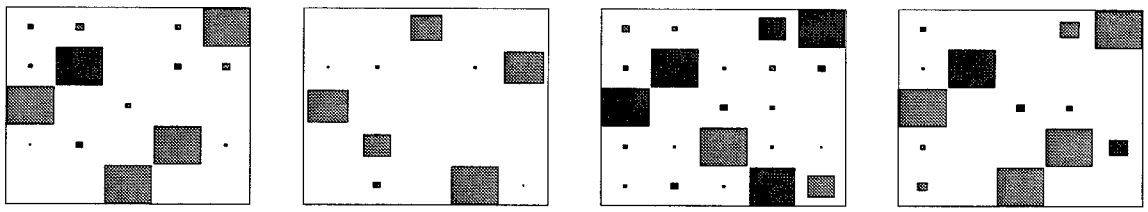
- In SEONS and Pham-Cardoso, we partitioned the whole data (10000 data points) into 100 different frames of data (each frame contains 100 data points) to calculate 100 different equal-time correlation matrices. These matrices were used to estimate the demixing matrix.
- In SOBI, we used 20 different time-delayed correlation matrices to estimate the demixing matrix.

4.2. Experiment 2

The second experiment was designed to show the robustness of the SEONS in the presence of spatially correlated but temporally white noise. We used 3 digitized voice signals and 2 music signals, all of which were sampled at 8 kHz. The mixing matrix $\boldsymbol{A} \in \mathbb{R}^{5\times 5}$, all the elements of which were drawn from standardized Gaussian distribution (i.e., zero mean and unit variance). As in the experiment 1, the whole data has 10000 samples.

The algorithms that are tested in this experiment, include the extended matrix pencil method (Extended MP), SEONS, extended Pham-Cardoso, JADE, SOBI, and SOBI with robust whitening method [8] (see Fig. 2). In SEONS, we partitioned the data into 50 no overlapping blocks (each frame has 200 data points). The robust whitening was performed using a combination of 5 time-delayed correlation matrices (with time-lags $\{1, 2, \ldots, 5\}$). In each data frame, we computed 5 time-delayed correlation matrices. The joint approximate diagonalizer of 250 correlation matrices (5 of each blocks = 5×50) was computed to estimate the demixing matrix.

At high SNR, most of algorithms worked very well, except for the extended MP method since it uses only two matrices. At low SNR, one can observe that the SOBI with robust whitening outperforms the SOBI without whitening. The SEONS gives slightly better performance than the SOBI with robust whitening in most of ranges of SNR. In the range between 0 and 6 dB, the SEONS is worse than the SOBI with robust whitening. It might result from the fact that the SEONS takes only 200 data points to calculate the time-delayed correlation matrices, so the temporal whiteness of the noise vector is not really satisfied. One can use less number of blocks (so more data points for each block) to reduce this drawback. The advantage of SEONS over SOBI with robust whitening



(a) (b) (c) (d)

Figure 1: Hinton diagrams of global system matrices: (a) SEONS; (b) Pham-Cardoso; (c) JADE; (d) SOBI with PI .001, .001, .05, .01, respectively.

lies in the fact that the first method works even for the case of nonstationary sources with identical spectra shape, whereas the latter does not (see the result of Experiment 1).

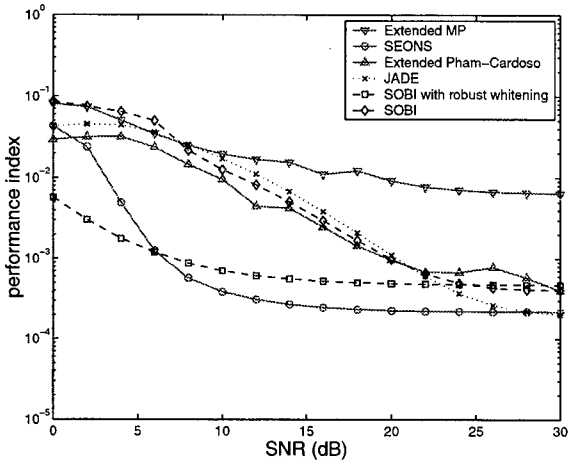


Figure 2: The performance comparison for SEONS, SOBI, SOBI with robust whitening, extended MP, JADE, and extended Pham-Cardoso.

5. CONCLUSION

In this paper we have presented a BSS method that jointly exploits the nonstationarity and temporal structure of sources. We have shown that our method, SEONS, was robust with respect to the temporally white noise and worked well even for the case of several nonstationary Gaussian sources (with no temporal correlations).

6. ACKNOWLEDGMENT

This work was supported by KOSEF 2000-2-20500-009-5 and by Korea Ministry of Science and Technology under an International Cooperative Research Project and by ETRI.

7. REFERENCES

[1] S. Haykin, *Unsupervised Adaptive Filtering: Blind Source Separation*. Prentice-Hall, 2000.

[2] S. Amari and A. Cichocki, "Adaptive blind signal processing - neural network approaches," *Proc. of the IEEE, Special Issue on Blind Identification and Estimation*, vol. 86, no. 10, pp. 2026–2048, Oct. 1998.

[3] A. Belouchrani, K. Abed-Merain, J. F. Cardoso, and E. Moulines, "A blind source separation technique using second order statistics," *IEEE Trans. Signal Processing*, vol. 45, pp. 434–444, Feb. 1997.

[4] C. Chang, Z. Ding, S. F. Yau, and F. H. Y. Chan, "A matrix-pencil approach to blind separation of colored nonstationary signals," *IEEE Trans. Signal Processing*, vol. 48, pp. 900–907, Mar. 2000.

[5] K. R. Müller, P. Philips, and A. Ziehe, "JADE$_{TD}$: Combining higher-order statistics and temporal information for blind source separation (with noise)," in *Proc. ICA'99*, (Aussois, France), pp. 87–92, 1999.

[6] S. Choi and A. Cichocki, "Blind separation of nonstationary sources in noisy mixtures," *Electronics Letters*, vol. 36, pp. 848–849, Apr. 2000.

[7] S. Choi and A. Cichocki, "Blind separation of nonstationary and temporally correlated sources from noisy mixtures," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, (Sidney, Austrailia), pp. 405–414, 2000.

[8] A. Belouchrani and A. Cichocki, "Robust whitening procedure in blind source separation context," *Electronics Letters*, vol. 36, pp. 2050–2051, Nov. 2000.

[9] L. Tong, Y. Inouye, and R. Liu, "A finite-step global convergence algorithm for the parameter estimation of multichannel MA processes," *IEEE Trans. Signal Processing*, vol. 40, pp. 2547–2558, Oct. 1992.

[10] J. F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM J. Mat. Anal. Appl.*, vol. 17, pp. 161–164, Jan. 1996.

[11] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, 1993.

[12] D. T. Pham, "Joint approximate diagonalization of positive definite hermitian matrices," Tech. Rep. LMC/IMAG, University of Grenoble, France, 2000.

[13] D. T. Pham and J. F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," in *Proc. ICA*, (Helsinki, Finland), pp. 187–192, 2000.

BLIND SEPARATION OF NON STATIONARY SOURCES USING JOINT BLOCK DIAGONALIZATION

Hicham Bousbia-Salah, Adel Belouchrani*, and Karim Abed-Meraim***

* Electrical Engineering Department,
Ecole Nationale Polytechnique,
P.O. Box182 El-Harrach, Algiers 16200, Algeria.
E-mail: hichambs@yahoo.fr, belouchrani@hotmail.com
** Signal and Image processing Department,
Télécom Paris 46 rue Barrault, 75013 Paris, France.
E-mail: abed@tsi.enst.fr

ABSTRACT

Recovering independent source signals from their convolutive mixtures without any a priori knowledge on their structure represents a great challenge in signal processing. In this paper, we present an efficient solution that is based on the joint block-diagonalization of positive spatio-temporal covariance matrices. In the case of instantaneous mixtures, robust solutions have been proposed previously. Taking advantage of possible non-stationarity of the sources, this new technique uses only second order statistics. The new approach has been successfully applied to the separation of speech signals.

1. INTRODUCTION

If we consider a set of received signals that are convolutive mixtures of independent source signals, the objective of blind separation is to recover the source signals from the set of received signals without any knowledge of the linear mixtures or the Linear Time Invariant (LTI) systems. For instantaneous mixtures, a Second Order Blind Identification (SOBI) algorithm has been presented [1] and showed to be very robust for temporally correlated sources. An analog technique based on Block Gaussian likelihood, presented by Pham [2] uses a joint diagonalization of positive correlation matrices of the received data. An extension of the SOBI technique to the convolutive mixtures has been considered in [3, 4].

When dealing with convolutive mixtures, classical blind separation can be achieved in two ways. One way is to first identify the channel system from the output mixtures and then to design an equalizer accordingly [5]. The other way consists of directly designing an equalizer from the output mixtures. The latter bypasses the problem of blind system identification and is computation less expensive. Herein, we consider the separation of the source signals up to a scalar filter instead of a full deconvolution. For this purpose, we propose to extend the Block Gaussian likelihood technique [2] to the convolutive mixture case. It is based on the joint block-diagonalization of positive spatio-temporal covariance matrices of the received data. In this contribution, the measure of block-diagonality is directly related to the likelihood objective function and is optimized without any orthogonality constraint which bypasses any prior whitening of the observations. The proposed method has been successfully applied to the separation of speech signals up to a scalar filter. In the next sections, we will present the data model and describe the proposed algorithm. And finally, some simulation results are provided in section 5.

2. DATA MODEL

For simplicity, we shall restrict ourselves to the simplest discrete time multiple input multiple output (MIMO) linear time invariant model given by,

$$x_i(n) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} h_{ij}(l) s_j(n-l), \quad \text{for } i = 1, \cdots, N. \quad (1)$$

where $s_j(n)$, $j = 1, \cdots, M$ are the M source signals (model inputs), $x_i(n)$, $i = 1, \cdots, N$, are the N sensor signals (model outputs), h_{ij} is the transfer function with an overall duration L between the j-th source and the i-th sensor.

The assumptions made about the data model are as follows:

A1) The source signals $s_j(n)$, $j = 1, \cdots, M$, are mutually decorrelated.

A2) Each source signal is non stationary.

A3) The channel matrix $\tilde{\mathbf{H}}$ defined in (3) is full column rank.

The purpose of blind source separation is to recover the

0-7803-7011-2/01/$10.00 ©2001 IEEE

source signals based only on the sensor signals. In some applications as in speech processing, the separation of the sources up to a scalar filter is sufficient. In this paper, we consider the problem of the source separation up to a scalar filter instead of the full MIMO deconvolution procedure.
We can rewrite equation (1) in the following matrix form,

$$\mathbf{x}(n) = \tilde{\mathbf{H}}\mathbf{s}(n) \tag{2}$$

where

$$\begin{aligned}\mathbf{s}(n) &= [s_1(n), \cdots, s_1(n-(L+L'-1)+1), \\ &\quad \cdots, s_M(n), \cdots, s_M(n-(L+L'-1)+1)]^T \\ \mathbf{x}(n) &= [x_1(n), \cdots, x_1(n-L'+1), \\ &\quad \cdots, x_N(n), \cdots, x_N(n-L'+1)]^T\end{aligned}$$

Subscript ^{T} denotes the transpose of a vector, and:

$$\tilde{\mathbf{H}} = \begin{bmatrix} \tilde{\mathbf{H}}_{11} & \cdots & \tilde{\mathbf{H}}_{1M} \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{H}}_{N1} & \cdots & \tilde{\mathbf{H}}_{NM} \end{bmatrix} \tag{3}$$

with

$$\tilde{\mathbf{H}}_{ij} = \begin{bmatrix} h_{ij}(0) & \cdots & h_{ij}(L-1) & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \cdots & h_{ij}(0) & \cdots & h_{ij}(L-1) \end{bmatrix}$$

Note that $\tilde{\mathbf{H}}$ is a $[NL' \times M(L+L'-1)]$ matrix and $\tilde{\mathbf{H}}_{ij}$ are $[L' \times (L+L'-1)]$ matrices. L' is chosen such that $NL' \geq M(L+L'-1)$.
We assume that $\tilde{\mathbf{H}}$ is a square matrix, i.e., $NL' = M(L+L'-1)$, if not, it can be made square by projecting the sensors data $\mathbf{x}(n)$ into the sources subspace.

3. THE PROPOSED ALGORITHM

In this section, we extend the Block Gaussian likelihood technique [2] to the convolutive mixture case. It is based on the joint block-diagonalization of positive spatio-temporal covariance matrices of the received data.
The interval $[0, T]$ may be divided into K consecutive subintervals $T_1, \cdots, T_K$ such that the approximate covariance matrice of the received data in the sub-interval T_k is given by:

$$\tilde{\mathbf{R}}_x(k) = \frac{1}{n_{T_k}} \sum_{n \in T_k} \mathbf{x}(n)\mathbf{x}(n)^* \tag{4}$$

where n_{T_k} is the number of elements (samples) in the subinterval T_k and subscript * denotes the conjugate transpose of a vector. Implicitly, we assume approximate local stationarity in each data sub-block.

Under the linear model (2), the above equation can be put in the following form:

$$\tilde{\mathbf{R}}_x(k) = \tilde{\mathbf{H}}\tilde{\mathbf{R}}_s(k)\tilde{\mathbf{H}}^H \tag{5}$$

where, $\tilde{\mathbf{R}}_s(k)$ are the approximate covariance matrices of the source signals and subscript ^{H} denotes the conjugate transpose of a matrix. Taking advantage of the mutual decorrelation of the source signals, $\tilde{\mathbf{R}}_s(k)$ is approximately block diagonal, with M diagonal blocks of dimension $(L+L'-1) \times (L+L'-1)$ each, i.e.

$$\tilde{\mathbf{R}}_s(k) \approx \begin{bmatrix} \tilde{\mathbf{R}}_{s_1}(k) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{R}}_{s_2}(k) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{R}}_{s_M}(k) \end{bmatrix} \tag{6}$$

$\tilde{\mathbf{R}}_{s_1}(k), \tilde{\mathbf{R}}_{s_2}(k), \cdots, \tilde{\mathbf{R}}_{s_M}(k)$ are the 'local' covariance matrices of the M sources, k being the data sub-block index. Equations (5) and (6) just mean that any data covariance matrix is block-diagonal in the basis of the column vectors of matrix $\tilde{\mathbf{H}}$, which can be retrieved by computing the joint block-diagonalization of a set of K covariance matrices $\tilde{\mathbf{R}}_x(k), k = 1, \cdots, K$.

4. A JOINT BLOCK-DIAGONALIZATION CRITERION

In this section, we derive a joint block-diagonalization criterion inspired from the joint diagonalization criterion of [2]. Using the Kullback-Leiber divergence between two zero mean K-variate normal densities with covariance matrices $\mathbf{R}_a$ and $\mathbf{R}_b$ respectively, the deviation between $\mathbf{R}_a$ and $\mathbf{R}_b$ is defined as:

$$D(\mathbf{R}_a, \mathbf{R}_b) > 0 \tag{7}$$

with equality if and only if $\mathbf{R}_a = \mathbf{R}_b$ and thus is a legitimate measure of deviation between positive definite matrices.
Therefore, a measure of deviation from block-diagona lization could be derived from:

$$D(\tilde{\mathbf{H}}^{-1}\tilde{\mathbf{R}}_x(k)\tilde{\mathbf{H}}^{-H}, \tilde{\mathbf{R}}_s(k)) \tag{8}$$

Following the same steps as in [2], the above measure of deviation is equivalent to,

$$\sum_{k=1}^{K} [\log\det(bdiag(\mathbf{M}_k)) - \log\det(\mathbf{M}_k)] \tag{9}$$

with

$$\mathbf{M}_k = \mathbf{B}\tilde{\mathbf{R}}_x(k)\mathbf{B}^H \tag{10}$$

over the set of matrices $\mathbf{B}$, where $bdiag(\mathbf{M}_k)$ denotes the block-diagonal matrix with the same diagonal blocks of size

$(L+L'-1)\times(L+L'-1)$ as $\mathbf{M}_k$.
From the generalized Hadamard inequality [6] and for Hermitian positive definite matrices:

$$\det(\mathbf{M}_k) < \det(bdiag(\mathbf{M}_k)) \tag{11}$$

with equality if and only if $\mathbf{M}_k$ is block-diagonal.
It follows that criterion (9) is a measure of the global deviation of the matrices from block-diagonal structure. Hence, minimization of (9) leads to,

$$\mathbf{B} \approx \mathbf{D}\tilde{\mathbf{H}}^{-1} \tag{12}$$

where the matrix $\mathbf{D}$ is an arbitrary block-diagonal matrix coming from the inherent indeterminacy of the joint block-diagonalization problem.

Once the matrix $\mathbf{B}$ is determined, the recovered signal are obtained up to a filter by,

$$\hat{\mathbf{s}}(n) = \mathbf{B}\mathbf{x}(n) \tag{13}$$

Accordingly, the recovered signals will verify,

$$\hat{\mathbf{s}}(n) = \mathbf{D}\mathbf{s}(n) \tag{14}$$

5. SIMULATIONS

We present here a simulation to illustrate the effectiveness of our algorithm in separating speech signals. The parameter settings are :

1. $M = 3, N = 2, L = 3$ and $L' = 4$.
2. The two speech signals are sampled at 8kHz.
3. The transfer function matrix of the simulated multi channel is given by,

$$\mathbf{H}(z) = \begin{bmatrix} 1+0.5z^{-1}+0.7z^{-2} & 0.1z^{-1}+0.85z^{-2} \\ 0.8+0.7z^{-1}+0.4z^{-2} & 1+0.9z^{-1} \\ 1+0.5z^{-1}+0.3z^{-2} & 0.7+0.85z^{-1}+0.1z^{-2} \end{bmatrix}$$

Figures 1, 2 and 3 show a sample run of the proposed algorithm. Note that only two speech signals among twelve recovered ones are displayed. These two signals lead to the smallest correlation coefficients.

6. CONCLUSION

In this contribution, we considered the problem of the blind separation of convolutive mixtures of non-stationary source signals. We proposed a solution based on the joint block-diagonalization of positive spatio-temporal covariance matrices. This technique uses only second order statistics and unlike [3, 4] has no orthogonality constraint which bypasses any prior whitening of the data. This method is well suited when applied to the deconvolution of speech signals, which is of great importance in practical applications [7].

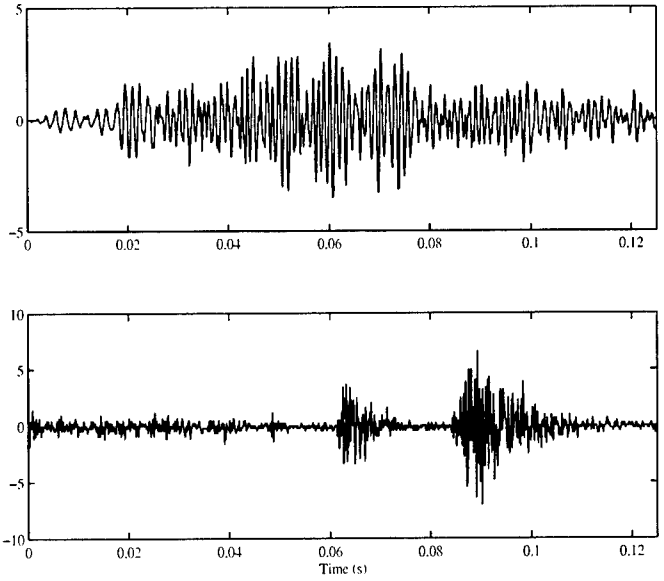


Fig. 1. Original speech signals.

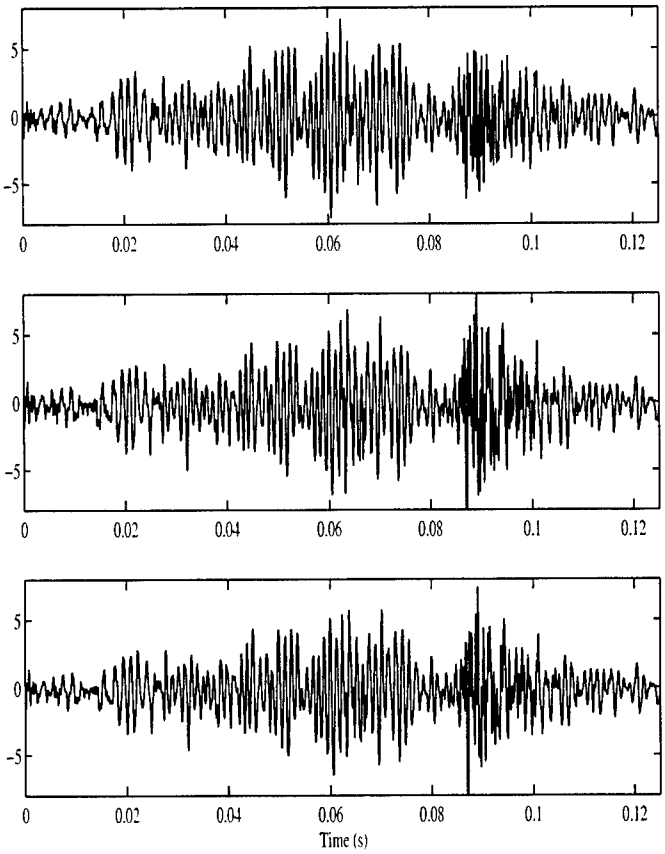


Fig. 2. Mixed speech signals.

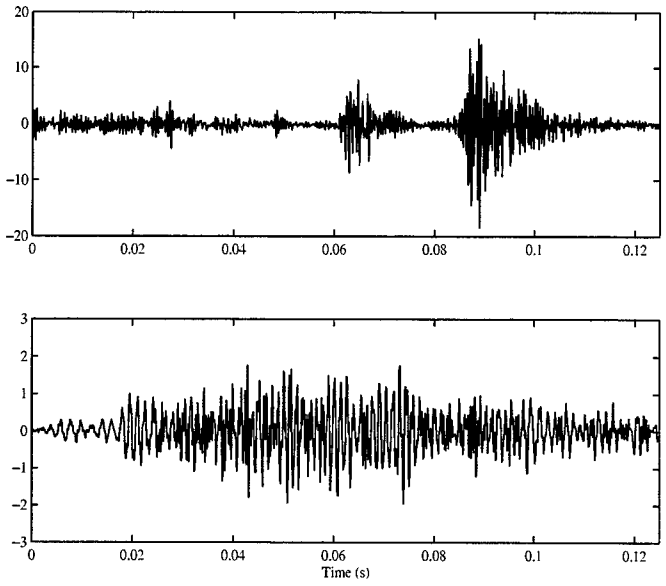


Fig. 3. Two of the twelve recovered speech signals.

7. REFERENCES

[1] A. Belouchrani and K. Abed Meraim and J.-F Cardoso and E. Moulines, "A blind source separation technique using second order statistics," *IEEE Trans. on SP*, vol. 45, pp. 434–444, Feb. 1997.

[2] D. Pham and J. F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. on SP*, vol. 45, pp. 2608–2612, Oct. 1997.

[3] H. Bousbia-Salah and A. Belouchrani, "A Second Order Multi Output Deconvolution (SOMOD) Technique," *In the tenth IEEE Workshop on Statistical Signal and Array Processing*, pp.306-310, Poconos,Pennsylvania,USA, Aug. 2000.

[4] H. Bousbia-Salah, A. Belouchrani and K. Abed-Meraim, "Blind source separation of convolutive mixtures using joint block diagonalization," *Proc. ISSPA'2001*, Kuala-Lumpur, Malaysia, Aug. 2001.

[5] G. B. Giannakis, Y. Inouye, and J. M. Mendel, "Cumulant-based identification of multichannel moving-average processes," *IEEE Trans. on Automat. Contr.*, vol. 34, pp. 783–787, Jul. 1989.

[6] B. Flury and B.E. Neuenschwander, "Principal component models for patterned covariance matrices, with applications to canonical correlation analysis of several sets of variables," *Descriptive multivariate analysis, Oxford University Press*, 1994.

[7] F. Ehlers and H. G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment," *IEEE Trans. on SP*, vol. 45, pp. 2608–2612, Oct. 1997.

BLIND SOURCE SEPARATION OF AUDIO SIGNALS USING IMPROVED ICA METHOD

F. Sattar, Member, IEEE, *M. Y. Siyal,* Member, IEEE, *L. C. Wee and L. C. Yen*

School of EEE,
Nanyang Technological University,
Nanyang Avenue, Singapore 639798.

ABSTRACT

Blind source separation (BSS) of independent sources from their convolutive mixtures is a problem in many real-world multi-sensor applications. In this paper, we propose an improved BSS method for audio signals based on ICA (Independent Component Analysis) technique. It is performed by implementing non-causal filters instead of causal filters within the feedback network of the ICA based BSS method. It reduces the required length of the unmixing filters considerably as well as provides better results and faster convergence compared to the case with the conventional causal filters. The proposed method has been simulated and compared for real world audio signals.

1. INTRODUCTION

Blind signal separation refers to performing inverse channel (or unmixing filter) estimation despite having no knowledge about the true channel (or mixing filter). The word "blind" refers that the independent original source signals and the mixing process are unknown.

A typical scenario would be to record two people talking at the same time using two microphones. The recorded signals would then of-course consist of a mixture of the two speech signals. The applied algorithm then tries to estimate the inverse channel and force the recored signals to be independent of each other (in order to separate the signals).

BSS method based on ICA (independent component analysis) technique has been found effective in signal separation comparing other BSS methods. The serious limitation of this technique is the requirement of long unmixing filters in order to estimate inverse channels[1]

The objective of this paper is thus to improve an ICA based BSS method by reducing the length of the unmixing filters. This can be achieved by implementing non-causal filters instead of conventional causal filters within the feedback network of the ICA based BSS method. This non-causal filters within the feedback loop is able to reduce the length of the unmixing/separation filters, while improve the results of the source separation by reducing the whitening effect (i.e. not sensitive to whitening in the inversion of *non-minimum* phase system). The feedback network within the non-causal filters is then able to invert the mixing even if the direct paths are not "good", i.e. when the direct channels filters are not gauranteed to have stable inverse. Moreover, for adaptation of the learning process, a variable step-size parameter is adopted providing the stable convergence.

2. BACKGROUND OF THE BSS ALGORITHM

2.1. "Infomax" or Entropy Maximization Criterion

BSS is the main application of independent component analysis (ICA), which reduces redundancy between source signals and make them "as independent as possible". In BSS, second order statistics are inadequate to reduce redundancy between the input signals. Higher-order statistics are required for redundancy reduction and these are determined mainly in two ways. The first is the explicit estimation of the cumulants and polyspectra[2]. The second is by obtaining higher-order statistics through the use of static nonlinear functions[3].

Bell and Sejnowski[4] proposed an information-theoretic approach for blind source separation (BSS), which is referred to as the "Infomax algorithm". Information theory can be used to unify several lines of research[5] and different theories recently proposed for independent component analysis (ICA), leading to the same iterative learning algorithm for BSS.

2.2. Separation of Convolutive Mixture

The initial algorithm of Bell and Sejnowski[4] deals with the instantaneous mixture problem. The algorithm was further extended by Torkkola for the convolutive mixture problem. Given measured signals, which are combinations of independent sources, the aim of blind separation is to produce outputs, which recreate the source signals, i.e., $y_1(k) = s_1(k)$, $y_2(k) = s_2(k), \cdots, y_n(k) = s_n(k)$. Nothing can be assumed about the sources except that they are statistically independent. Torkkola[6] suggested the feedback structure for the separation of convolutive mixture (see also[5]). The nonlinear function, f, must be a monotonically increasing or decreasing function. In this paper the nonlinear function used is defined as $y = f(u) = 1/(1+e^{-u})$. The learning rule for the convolutive mixture can follow the same steps as the instantaneous case[4]. Minimizing the mutual information between outputs y_1 and y_2 can be achieved by maximizing the entropy at the output[5]. Assuming causal FIR filters for w^{ij}, the network performs the following operations in

0-7803-7011-2/01/$10.00 ©2001 IEEE

the time domain:

$$u_1(t) = \sum_{k=0}^{L_{11}} w_k^{11} x_1(t-k) + \sum_{k=1}^{L_{12}} w_k^{12} u_2(t-k)$$
$$u_2(t) = \sum_{k=0}^{L_{22}} w_k^{22} x_2(t-k) + \sum_{k=1}^{L_{21}} w_k^{21} u_1(t-k) \quad (1)$$

where w_k^{ij} is the kth tap of the filter from source j to sensor i and L_{ij} is the filter length for the respective filter.

The relationships between the mixing filter and the separation filter can be expressed in z-transform[6]:

$$W_{11}(z) = A_{11}(z)^{-1}, \quad W_{12}(z) = -A_{12}(z)A_{11}(z)^{-1}$$
$$W_{22}(z) = A_{22}(z)^{-1}, \quad W_{21}(z) = -A_{21}(z)A_{22}(z)^{-1} \quad (2)$$

This is a network which combines the separation and deconvolution problem. Maximizing the entropy at the output will result in W_{11} and W_{22} not only inverting A_{11} and A_{22}, but also whitening the sources. This can be avoided by forcing W_{11} and W_{22} to mere scaling coefficients. In the ideal case, W_{11} and W_{22} will have the following solutions:

$$W_{11}(z) = 1, \quad W_{12}(z) = -A_{12}(z)A_{22}(z)^{-1}$$
$$W_{22}(z) = 1, \quad W_{21}(z) = -A_{21}(z)A_{11}(z)^{-1} \quad (3)$$

Further, when the feedback network is used, we have to consider the relations: $U_1(z) = A_{11}(z)S_1(z)$ and $U_2(z) = A_{22}(z)S_2(z)$ which are related to what each sensor would observe in the absence of interference from the other source.

The learning rules for the separation matrix are:

$$\Delta w_0^{ii} \propto (1-2y_i)x_i + 1/w_0^{ii},$$
$$\Delta w_k^{ii} \propto (1-2y_i)x_i(t-k)$$
$$\Delta w_k^{ij} \propto (1-2y_i)u_j(t-k) \quad (4)$$

where $k = 0, 1, 2, \cdots, L_{ij}$.

3. THE IMPROVED ICA BASED BSS METHOD

Torkkola's algorithm[6] works only when the stable inverse of the direct channel filters (A_{11} and A_{22}) exist. This is not always guaranteed in real world systems. In the separation of audio signals, the direct channel is the path from the source to the ipsi microphone. The corresponding transfer function would come from a very complex process, for which it is not guaranteed that there will a stable inverse for this transfer function.

However, even if a filter does not have a stable causal inverse, there still exists a stable non-causal inverse. Therefore, the algorithm of Torkkola can be modified and used even though there is no stable (causal) inverse filter for the direct channel.

The relationships between the signals are now changed to:

$$u_1(t) = \sum_{k=-M}^{M} w_k^{11} x_1(t-k) + \sum_{k=-M}^{M} w_k^{12} u_2(t-k)$$
$$u_2(t) = \sum_{k=-M}^{M} w_k^{22} x_2(t-k) + \sum_{k=-M}^{M} w_k^{21} u_1(t-k) \quad (5)$$

where M(even) is half of the (total filter length-1) and the zero lag of the filter is at $(M+1)$. In (5) there exist an initialization problem regarding filtering. To calculate the value of $u_1(t)$, the values of $u_2(t), u_2(t+1), \cdots, u_2(t+M)$ are required which are not initially available. Since learning is an iteration process, we have used some pre-assigned values to solve this filter initialization problem or padded signals with zeros of length M. For example, the input value of $x_2(t)$ is used for the output $u_2(t)$ at the first iteration. The new values generated at the first iteration are then used for the second iteration. This process is repeated until its convergence to certain values.

The derivative of the learning rule can follow the same procedure as in Torkkola[6]. According to (5), only the coefficients of W_{12} and W_{21} have to be learned. The learning rule is the same in notation but different in nature because the values of k have changed:

$$\Delta w_k^{ij} \propto (1-2y_i)u_j(t-k) \quad (6)$$

where $k = -M, -M+1, \cdots, M$.

The step-size is considered to be an exponentially time-varying step-size and the initial step-size is calculated as $1/(2\lambda_{max})$, where λ_{max} is the maximum singular value of the correlation matrix $\mathbf{R}$ for the initial input (mixed signal) block of length $(2M+1)$.

4. RESULTS AND PERFORMANCES

Separations of audio signals have been performed for various real mixed data, e.g. for two musics, for two speech of the same languages and different languages, for a music and a speech. In the following we present two illustrative results for the real audio-files, which are available in *http://www.cnl.salk.edu/~tewon/blind.html*[1].

Example 1: In this example, two different music signals are separated when sampling frequencies of the signals are 22 kHz. Fig. 1 shows small portions of the separation results. The original signals are shown in Figs. 1(a)–(b), whereas mixed signals and separated signals are presented in Figs. 1(c)–(d) and Figs. 1(e)–(f). In the above, the unmixing filters length used is 161, which is the minimum filter length needed for the successful separation. The stopping criterion for the learning process is when the change of weights (Δw_k^{12} and Δw_k^{21}) are less than a threshold value, which is set to be 0.0001. Then the number of iterations required is about 100. The audio signals can be listened in our newly developed web-page *(http://members.tripod.com/zen76/index.htm)*. Satisfactory separation results are obtained from both subjective (listening) and objective (cross-correlation) performance testings. Fig. 2 shows a very low cross-correlation values between the separated signals compared to that of the mixed signals indicating efficiency of the present method.

Example 2: In Fig. 3 the separation results are illustrated for the two recorded speech data having sampling frequencies of 16 kHz. The experiments have been performed with two speakers speaking simultaneously in a normal office room[7]. Figs. 3(a)–(b) show the small portion of the mixed signals recorded from two microphones. The corresponding separated signals are shown in Figs. 3(c)–(d). The listening

[1]The details of the experimental setup for the audio sound recording can be found in http://www.cnl.salk.edu/~tewon/blind.html

test shows speech separation is almost perfect. In this example, the filter length used is 321 being the minimum filter length, which provides good results. The stopping learning threshold is chosen to be same as in *Example 1* and needed 120 iterations to reach the threshold. It is found that extreme value of the cross-correlation for the mixed signals in *Example 2* is much higher than that of mixed signals in *Example 1*. This could be the reason for requiring larger filter length and more iterations for the former case compared to the latter case.

Also note that this method can be extended for more than two-source two-sensor case. An illustrative result for the three-source three-sensor case is shown in *http://members/tripod.com/zen76/index.htm*. It is found that the three sources from their three mixtures can be successfully separated.

5. COMPARISON

The results are compared with the Te-Wons results shown in[1, 7] (see also in *http://www.cnl.salk.edu/~tewon/blind.html*). The limitation of the Te-Wons method is that it requires large filter length (e.g. 1024 samples), which is significantly reduced by the proposed method (e.g., for the case in our simulation examples the reduction is 5 to 6 times). Moreover, according to Fig. 4, the cross-correlation values are found less for the presented method compared to Te-Won's results. From the listening test it is also found that the separation quality is better for the proposed method (see in *http://members.tripod.com/zen76/index.htm* to compare results as well as other examples). Here we do not compare our results with that of other more conventional methods, since it is found that Te-Won's method works much better than the other existing methods when real-world audio signals are used.

6. DISCUSSION

Separations of audio signals have been performed for various real-world signals using an efficient ICA based BSS method. Using feedback network within non-causal filters it is successful to reduce the length of the unmixing filters. Satisfactory results are obtained from both the subjective (listening) and objective (cross-correlation) performance tests, which overcome the results shown in[7]. The length of separation filters and the required number of iterations may depend on the amount of cross-correlation between the recorded signals.

REFERENCES

[1] T-W Lee, "Independent Component Analysis - Theory and Applications", Kluwer Academic Publishers, 1998.

[2] R.M. Gray, "Entropy and Information Theory", New York: Springer-Verlag, 1990.

[3] P. Comon, "Independent component analysis, a new concept?", *Signal Processing*, vol. 36, 1994, pp.287-314.

[4] A.J. Bell and T.J. Sejnowski, "An information maximisation approach to blind separation", *Neural Computation*, vol. 7, 1995, pp. 1129-1159.

[5] H.H. Szu, I. Kopriva, A. Persin, "Independent component analysis to resolve the multi-source limitation of the nutating rising-sun reticle based optical trackers", *Optics Communication*, vol. 176, March 2000, pp. 77–89.

[6] K. Torkkola, "Blind separation of convolved sources based on information maximization", *IEEE Workshop Neural Networks for Signal Processing*, Kyoto, Japan, Sept 4-6,1996.

[7] T-W Lee, A.J. Bell and R. Orglmeister, "Blind source separation of real world signals", *Proc. IEEE Int. Conf. Neural Networks*, June 97, Houston, pp. 2129–2135.

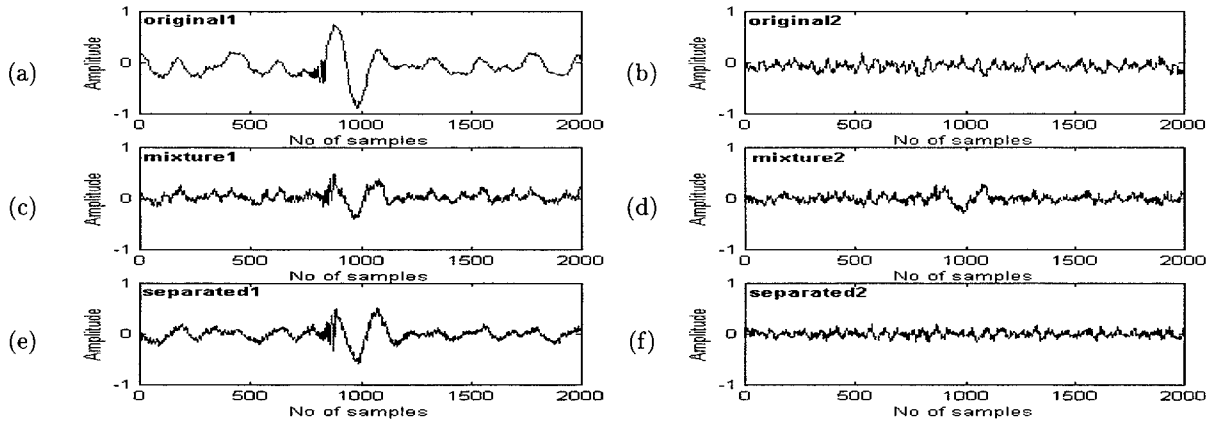


Figure 1: Separation of two music signals (between 0s–.091s); (a)–(b) Original source signals, (c)–(d) Mixed signals, (e)–(f) Separated signals.

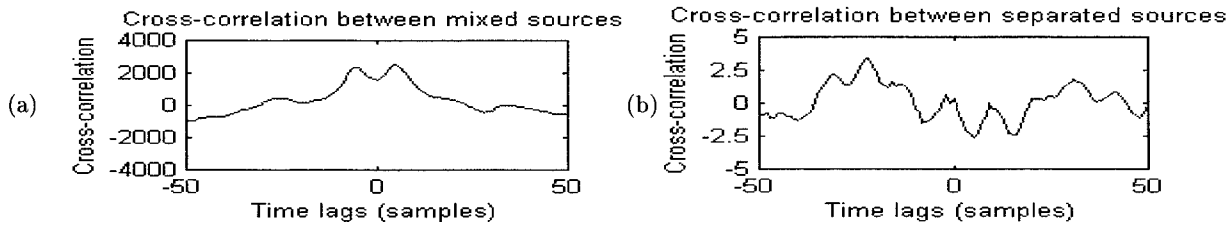


Figure 2: Performance measure using cross-correlation; (a) Cross-correlation between mixed signals in Figs. 1(c)–(d), (b) Cross-correlation between separated signals in Figs. 1(e)–(f).

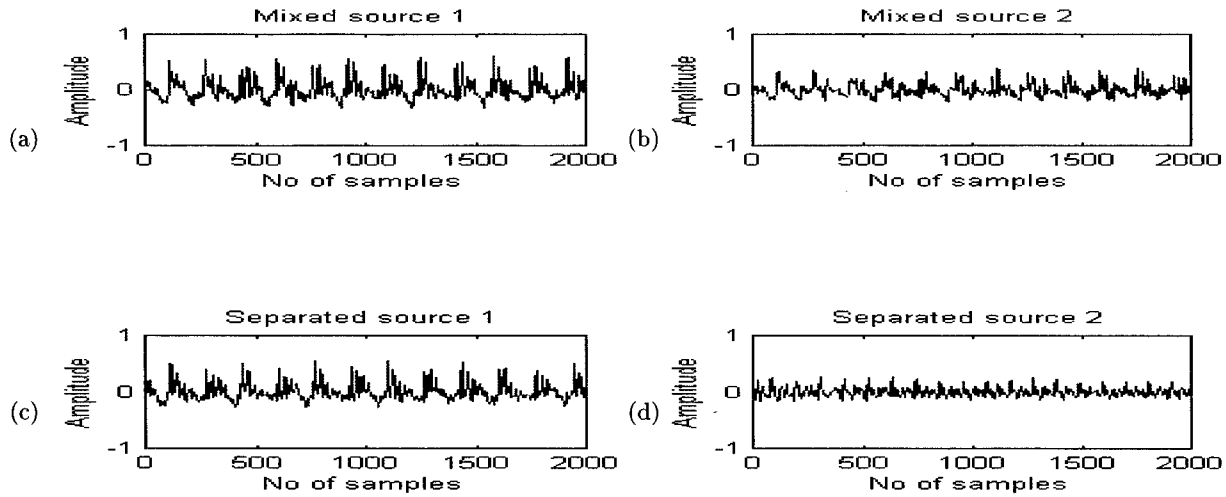


Figure 3: Separation of two speech signals (between 0.3125s–0.4375s); (a)–(b) Mixed signals, (c)–(d) Separated signals.

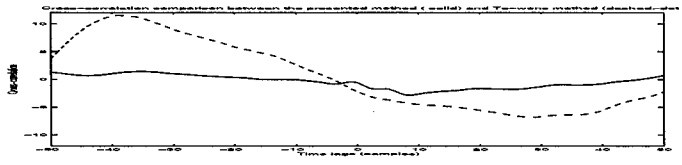


Figure 4: Cross-correlation between the separated signals for the presented method (solid line) and Te-Won's method [7] (dashed-dot line) for the results shown in Figs. 1(e)–(f) and http://www.cnl.salk.edu/~tewon/blind.html.

WEIGHTED CLOSED-FORM ESTIMATORS FOR BLIND SOURCE SEPARATION

Vicente Zarzoso, Frank Herrmann and Asoke K. Nandi

Signal Processing and Communications Group, Department of Electrical Engineering and Electronics,
The University of Liverpool, Brownlow Hill, Liverpool L69 3GJ, UK
Tel/Fax: +44 151 794 4525/4540, e-mail: {vicente, fherrm, aknandi}@liv.ac.uk
http://www.liv.ac.uk/~{vicente, fherrm, aknandi}

ABSTRACT

This paper investigates a novel closed-form estimation class, so-called weighted estimator (WE), for blind source separation in the basic two-signal problem. Proper combination of previously proposed estimators yields consistent estimates of the separation parameters under general conditions. In the real-mixture case, we determine analytic expressions for the WE asymptotic (large-sample) variance and the source-dependent weight value of the most efficient estimator in the class. By means of the bicomplex-number formalism, the WE is extended to the complex-mixture scenario, for which Cramér-Rao bounds are also derived. Simulations compare the WE with other methods, demonstrating its potential.

Keywords: blind source separation, estimation theory, higher-order statistics, non-Gaussian signal processing, sensor array processing.

1. INTRODUCTION

The problem of blind source separation (BSS) arises in a great variety of applications, in fields as diverse as wireless communications, seismic exploration and biomedical signal processing. BSS aims to reconstruct an unknown set of q mutually independent source signals $\mathbf{x} \in \mathbb{C}^q$ which appear mixed at the output of a p-sensor array $\mathbf{y} \in \mathbb{C}^p$, $p \geqslant q$. In the noiseless instantaneous linear case, sources and observations are linked through an unknown mixing transformation $M \in \mathbb{C}^{p \times q}$:

$$\mathbf{y} = M\mathbf{x}. \tag{1}$$

The problem consists of estimating the source vector $\mathbf{x}$ and the mixing matrix M from the exclusive knowledge of sensor vector $\mathbf{y}$. Neither the ordering nor the power and phase-shift of the sources can be identified in the model above, so we may assume, with no loss of generality, an identity source covariance matrix.

When the time structure of the signals cannot be exploited (e.g., due to the source spectral whiteness), one needs to resort to higher-order statistics (HOS) [1]. The success of the separation then relies on the non-Gaussian nature of the sources. A previous spatial whitening process (entailing second-order decorrelation and power normalization) helps to reduce the number of unknowns, resulting in a set of normalized uncorrelated components $\mathbf{z} \in \mathbb{C}^q$:

$$\mathbf{z} = Q\mathbf{x}, \tag{2}$$

with $Q \in \mathbb{C}^{q \times q}$ unitary. As the general scenario $p > 2$ can be tackled through an iterative approach over the signal pairs [2], the two-signal case, $p = q = 2$, is of fundamental importance. The unitary transformation Q is then a complex elementary Givens rotation matrix:

$$Q = \begin{bmatrix} \cos\theta & -\mathrm{e}^{-j\alpha}\sin\theta \\ \mathrm{e}^{j\alpha}\sin\theta & \cos\theta \end{bmatrix}. \tag{3}$$

Hence, the source-signal extraction and mixing-matrix identification reduce to the estimation of angular parameters θ, $\alpha \in \mathbb{R}$.

In the real-valued mixture case, $\alpha = 0$ and only θ is unknown. The performance of the first closed-form solution for the estimation of θ, based on the output 4th-order cross-cumulant nulling [3], was later shown to depend on θ itself [4, 5]. The maximum-likelihood (ML) approach on the Gram-Charlier expansion of the source probability density function (pdf) produced the solution of [6], whose validity was broadened through the extended ML (EML) and the alternative EML (AEML) estimators [4, 7, 8]. Such estimators lose their consistency for zero source kurtosis sum (sks) and source kurtosis difference (skd), respectively. This deficiency was overcome in [8] and [9]. In the latter, adopting the framework of [6] the two estimators were joined into a single analytic expression, the approximate ML (AML). The MaSSFOC estimator [10], derived from the approximate maximization of a contrast function made up of the sum of output squared kurtosis [2], exhibits a strikingly resembling form. The notion of linearly combining estimation expressions using arbitrary weights was originally put forward in [9], giving rise to the so-called weighted AML (WAML) estimator. It was suggested that the weight parameter could be adjusted by taking advantage of a priori information on the source pdfs, although no specific guidelines were given on how the actual choice should be made.

The present contribution fills this gap by studying in finer detail this weighted estimator (WE) for BSS and emphasizing its potential benefits. In the real-mixture case, we capitalize on the complex-centroid notation used in the EML and AEML estimators in order to provide an analytic formula for the WE large-sample variance. From this formula, the weight parameter of the asymptotically most efficient WE is obtained as a function of the source statistics. In addition, the WE is neatly extended to the complex-valued mixture case with the bicomplex number formalism developed in [4, 11]. We deduce Cramér-Rao lower bounds (CRLBs) for the pertinent parameters, and show in simulations that the WE is able to follow the CRLB trend of an objective separation-quality performance index. The connections between the WE and other analytic solutions are also highlighted throughout the paper.

First, we summarize a few mathematical notations. Symbol $\mu^x_{mn} = \mathrm{E}[x_1^m x_2^n]$, where $\mathrm{E}[\cdot]$ denotes the mathematical expectation, stands for the $(m+n)$th-order moment of the source signals $\mathbf{x} = (x_1, x_2)$. For convenience, the cumulants of complex vector $\mathbf{z} = (z_1, \ldots, z_q)$ are defined as $\mathrm{Cum}^z_{i_1 i_2 i_3 \ldots} =$

Vicente Zarzoso would like to thank the Royal Academy of Engineering for supporting this work through the award of a Post-doctoral Research Fellowship.

$\mathrm{Cum}[z_{i_1}^*, z_{i_2}, z_{i_3}^*, \ldots]$, $1 \leqslant i_k \leqslant q$, with the convention, in the two-component case, $\kappa^z_{n-r,\,r} = \mathrm{Cum}^z_{\underbrace{1\ldots1}_{n-r}\,\underbrace{2\ldots2}_{r}}$. We also define $\gamma = \kappa^x_{40} + \kappa^x_{04}$ (sks) and $\eta = \kappa^x_{40} - \kappa^x_{04}$ (skd). Symbol $\angle a$ represents the principal value of the argument of $a \in \mathbb{C}$.

2. REAL-MIXTURE CASE

2.1. Fourth-Order Weighted Estimator

The WAML estimator [9] accepts a more convenient formulation when adopting the EML/AEML approach [4, 5, 7, 8], which is based on the polar representation of real-valued bivariate random vector $\mathbf{z} = (z_1, z_2)$ as $\rho e^{j\phi} = z_1 + jz_2$, $j = \sqrt{-1}$. Higher-order expectations then generate complex-valued linear combinations (*centroids*) of the whitened-sensor statistics which lead to explicit estimation expressions for the parameter of interest. Accordingly, the EML is expressed as

$$\hat{\theta}_{\mathrm{EML}} = \tfrac{1}{4}\angle(\gamma\xi_4), \tag{4}$$

where ξ_4 is the 4th-order complex centroid:

$$\xi_4 = \mathrm{E}[\rho^4 e^{j4\phi}] = (\kappa^z_{40} + \kappa^z_{04} - 6\kappa^z_{22}) + j4(\kappa^z_{31} - \kappa^z_{13}), \tag{5}$$

and the sks can be estimated from the array output through $\gamma = \mathrm{E}[\rho^4] - 8 = \kappa^z_{40} + \kappa^z_{04} + 2\kappa^z_{22}$. Similarly, the AEML [4, 8] reads:

$$\hat{\theta}_{\mathrm{AEML}} = \tfrac{1}{2}\angle\xi_2, \tag{6}$$

$$\xi_2 = \mathrm{E}[\rho^4 e^{j2\phi}] = (\kappa^z_{40} - \kappa^z_{04}) + j2(\kappa^z_{31} + \kappa^z_{13}). \tag{7}$$

Under mild conditions [4, 7], centroids ξ_4 and ξ_2 are consistent estimators of $\gamma e^{j4\theta}$ and $\eta e^{j2\theta}$, respectively, so that $\hat{\theta}_{\mathrm{EML}}$ and $\hat{\theta}_{\mathrm{AEML}}$ consistently estimate θ as long as $\gamma \neq 0$ and $\eta \neq 0$, respectively. It follows that

$$\hat{\theta}_{\mathrm{WE}} = \tfrac{1}{4}\angle\xi_{\mathrm{WE}}, \qquad \text{with} \tag{8}$$

$$\xi_{\mathrm{WE}} = w\gamma\xi_4 + (1-w)\xi_2^2, \qquad 0 < w < 1. \tag{9}$$

is a consistent estimator of θ for *any* source distribution (besides when the sources are both Gaussian). Eqn. (8) is essentially the WAML estimator [9] written in centroid form. Nonetheless, we adhere to the more general denomination of *weighted estimator (WE)*, since its ML nature becomes unclear when extended to the complex-signal domain (Section 3).

Some special cases of the WE are:

(i) $w = 0$: AEML estimator of [4, 8].
(ii) $w = 1/3$: AML estimator of [9].
(iii) $w = 1/2$: MaSSFOC estimator of [10].
(iv) $w = 1$: EML estimator of [4, 7].

2.2. Performance Analysis

Along the lines of [4, 5], and omitting tedious algebraic details, the asymptotic (large-sample) variance of the WE (8) is determined as:

$$\sigma^2_{\hat{\theta}_{\mathrm{WE}}} = \frac{\mathrm{E}\Big\{\big[w\gamma(x_1^3x_2 - x_1x_2^3) + (1-w)\eta(x_1^3x_2 + x_1x_2^3)\big]^2\Big\}}{T\big[w\gamma^2 + (1-w)\eta^2\big]^2}, \tag{10}$$

where T is the number of samples. Remark that:

(i) $\sigma^2_{\hat{\theta}_{\mathrm{WE}}}$ reduces to the asymptotic variance of the AEML and EML estimators [4, 5] for $w = 0$ and $w = 1$, respectively.

(ii) When $\gamma = 0$ (resp. $\eta = 0$), WE performance reduces to that of the AEML (resp. EML) estimator, for any $0 < w < 1$.

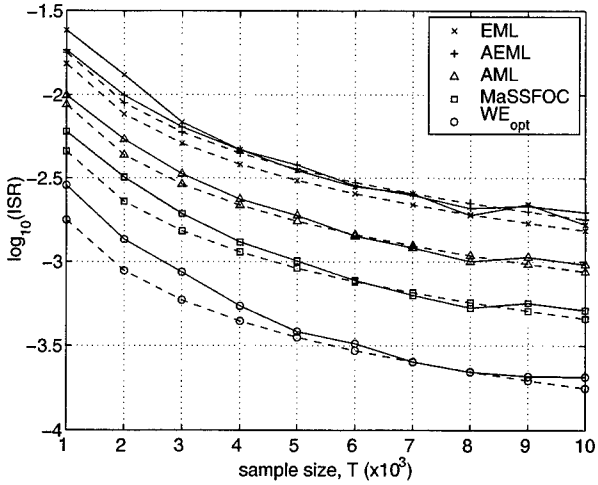


Fig. 1. ISR vs. sample size. Uniform–Rayleigh sources, $\theta = 15°$, ν independent Monte Carlo runs, with $\nu T = 5 \times 10^6$. Solid lines: average empirical values. Dashed lines: asymptotic variances (10).

2.3. Optimal Large-Sample Performance

If $|\kappa^x_{40}| \neq |\kappa^x_{04}|$, the derivative of eqn. (10) with respect to w cancels at:

$$w_{\mathrm{opt}} = \frac{1}{2} + \frac{\mu^x_{40}\mu^x_{04}\big[(\kappa^x_{40})^2 - (\kappa^x_{04})^2\big] + \kappa^x_{40}\kappa^x_{04}(\mu^x_{60} - \mu^x_{06})}{2\big[(\kappa^x_{40})^2\mu^x_{06} - (\kappa^x_{04})^2\mu^x_{60}\big]}. \tag{11}$$

Since $\partial^2(\sigma^2_{\hat{\theta}_{\mathrm{WE}}})/\partial w^2\big|_{w_{\mathrm{opt}}} > 0$, w_{opt} corresponds to the minimum variance estimator of the WE family. Hence, given the source statistics, one can select the WE with optimal asymptotic performance. If $w_{\mathrm{opt}} \notin [0, 1]$, we choose between $w_{\mathrm{opt}} = 0$ (AEML) and $w_{\mathrm{opt}} = 1$ (EML) the value that gives the lowest $\sigma^2_{\hat{\theta}_{\mathrm{WE}}}$ in (10).

2.4. Simulation Results

A few simulations illustrate the benefits of the WE and show the goodness of asymptotic approximation (10). First, observe that any angle estimate of the form $\hat{\theta} = \theta + n\pi/2$, $n \in \mathbb{Z}$, provides a valid separation solution up to the indeterminacies mentioned in Sec. 1. The interference-to-signal ratio (ISR) performance index [1] approximates the variance of $\hat{\theta}$, $\sigma^2_{\hat{\theta}}$, around any valid separation solution [4]. The ISR is an objective measure of separation performace, for it is method independent.

Fig. 1 shows the ISR results obtained by the EML, AEML, AML, MaSSFOC and optimal WE, together with the expected asymptotic variances, for varying sample size and i.i.d. sources with uniform and Rayleigh distributions [$w_{\mathrm{opt}} = 0.7141$, from eqn. (11)]. Centroids are computed from their polar forms. The optimal WE substantially outperforms the other estimators, being, e.g., five and ten times as efficient [12] as the AML and the AEML, respectively. The fitness of asymptotic approximation (10) is very precise in all cases.

The generalized Gaussian distribution (GGD) with shape parameter λ, $p(x) \propto \exp(-|x|^\lambda)$, is used as source pdf in the simulation of Fig. 2. We fix $\kappa^x_{04} = 0.5$ and smoothly vary κ^x_{40} to generate

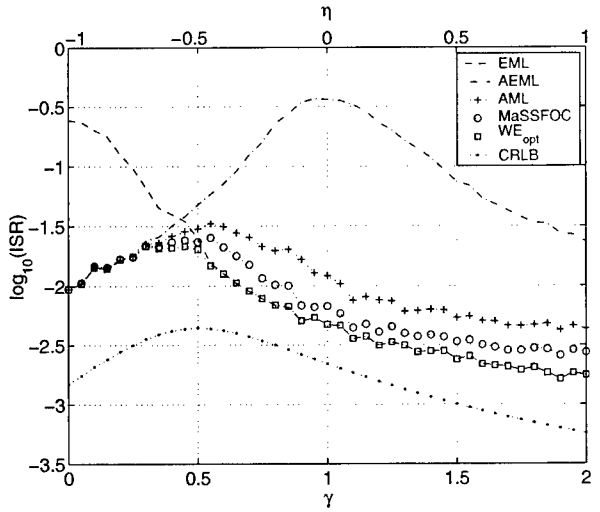


Fig. 2. ISR vs. sks γ and skd η. GGD sources, $\kappa_{04}^{x} = 0.5$, $\theta = 15°$, $T = 5 \times 10^3$ samples, 10^3 Monte Carlo runs.

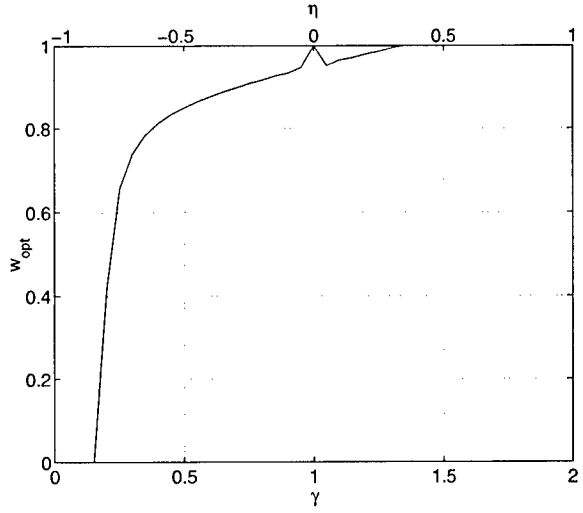


Fig. 3. Optimal value of the WE weight parameter in the separation scenario of Fig. 2.

a range of sks and skd values. The optimal WE, with w_{opt} calculated as in Sec. 2.3 and shown in Fig. 3, is compared with other analytic solutions and the CRLB obtained in [9] for the real case. The optimal WE follows the CRLB more closely than any of the other methods.

3. COMPLEX-MIXTURE CASE

3.1. Bicomplex Numbers

In [4, 11], the so-called bicomplex numbers prove useful in simplifying the development of closed-form estimators in the complex-mixture scenario. Given a unitary matrix $Q = \left[\begin{smallmatrix} a & -b^* \\ b & a^* \end{smallmatrix}\right]$, $a, b \in \mathbb{C}$, where * denotes complex conjugation, the associated bicomplex number is defined as $\bar{x} = a + \jmath b$. Though analogous to j, the *bimaginary unit* $\jmath$ is actually a distinct algebraic element. Terms $a = \mathbb{R}\text{e}(\bar{x})$ and $b = \mathbb{I}\text{m}(\bar{x})$ are the *breal* and *bimaginary* parts of $\bar{x}$, respectively. The product of two bicomplex numbers $\bar{x}_1 = a_1 + \jmath b_1$ and $\bar{x}_2 = a_2 + \jmath b_2$ is defined in accordance with the product of unitary transformations:

$$\bar{x}_1 \bar{x}_2 = (a_1 a_2 - b_1^* b_2) + \jmath(b_1 a_2 + a_1^* b_2). \tag{12}$$

In this manner, an isomorphism is created between the set of unitary matrices under usual matrix product and the set of bicomplex numbers under the above product operation. Note that, as with j, $\jmath^2 = -1$. A special class of bicomplex numbers arises when the associated unitary transformation shows the shape of (3):

$$\text{e}_{\alpha}^{\jmath\theta} = \cos\theta + \jmath \text{e}^{j\alpha} \sin\theta, \tag{13}$$

which we call bicomplex exponential.

3.2. Fourth-order Weighted Estimator

By means of the bicomplex formalism, one can easily generalize centroids (5) and (7) to the complex-mixture case. Effectively,

$$\bar{\xi}_4 = (\kappa_{40}^{\bar{z}} + \kappa_{04}^{\bar{z}} - 6\kappa_{22}^{\bar{z}}) + \jmath 4(\kappa_{31}^{\bar{z}} - \kappa_{13}^{\bar{z}}) \tag{14}$$

and

$$\bar{\xi}_2 = (\kappa_{40}^{\bar{z}} - \kappa_{04}^{\bar{z}}) + \jmath 2(\kappa_{31}^{\bar{z}} + \kappa_{13}^{\bar{z}}) \tag{15}$$

are consistent estimators of $\gamma \text{e}_{\alpha}^{\jmath 4\theta}$ and $\eta \text{e}_{\alpha}^{\jmath 2\theta}$, respectively, under the same general conditions as in the real case. Centroid (14) gives rise to the complex EML (CEML) estimator [4, 11], whereas (15) yields the complex AEML (CAEML) estimator [4]. Bearing in mind the bicomplex product (12), it follows immediately that the linear combination

$$\bar{\xi}_{\text{CWE}} = w\gamma\bar{\xi}_4 + (1-w)\bar{\xi}_2^2 \tag{16}$$

consistently estimates $(w\gamma^2 + (1-w)\eta^2)\text{e}_{\alpha}^{\jmath 4\theta}$. The sks γ may be obtained from the available data just as in the real case. For $w \in [0,\ 1]$, parameters $(\theta,\ \alpha)$ are estimated through

$$\begin{cases} 4\hat{\theta}_{\text{CWE}} = \angle\big(\mathbb{R}\text{e}(\bar{\xi}_{\text{CWE}}) + j|\mathbb{I}\text{m}(\bar{\xi}_{\text{CWE}})|\big) \\ \hat{\alpha}_{\text{CWE}} = \angle\, \mathbb{I}\text{m}(\bar{\xi}_{\text{CWE}}), \end{cases} \tag{17}$$

which is the *complex WE (CWE)*.

3.3. Cramér-Rao Lower Bounds

Assuming circularly distributed source signals composed of T independent samples, the Fisher information matrix (FIM) for the estimation of parameters $(\theta,\ \alpha)$ in model (2)–(3) reads:

$$\text{FIM}_{(\theta,\,\alpha)} = T \begin{bmatrix} I & 0 \\ 0 & \frac{1}{4} I \sin^2 2\theta \end{bmatrix}, \tag{18}$$

where

$$I = I_1 + I_2 - 4,$$
$$I_k = \frac{1}{2} \iint_{\mathsf{D}_k} \frac{1}{p_k} \left[\left(\frac{\partial p_k}{\partial u} \right)^2 + \left(\frac{\partial p_k}{\partial v} \right)^2 \right] \text{d}u\text{d}v, \tag{19}$$

and $p_k(u,\ v)$ is the pdf of the kth source signal $x_k = u_k + jv_k$, $u_k,\ v_k \in \mathbb{R}$, $k = 1,\ 2$. Integration extends over the definition domain D_k of the corresponding random variable.

It is interesting to note that:
(i) The CRLBs of θ and α are decoupled, and therefore:

$$\mathrm{CRLB}_\theta = (TI)^{-1} \tag{20}$$

$$\mathrm{CRLB}_\alpha = 4(TI\sin^2 2\theta)^{-1} \tag{21}$$

(ii) For sources with complex generalized Gaussian distribution (CGGD) of shape parameter λ, given by

$$p(u,\, v) \propto \exp\{-(u^2+v^2)^{\frac{\lambda}{2}}\}, \quad \lambda > 0, \tag{22}$$

we have

$$I_k = \tfrac{1}{2}\lambda_k^2 \Gamma(4/\lambda_k)/\Gamma^2(2/\lambda_k). \tag{23}$$

Then, the FIM is zero, and hence the model unidentifiable, iff $\lambda_1 = \lambda_2 = 2$, i.e., both sources are Gaussian.

(iii) When $\theta = n\pi/2$, $\forall n \in \mathbb{Z}$, estimation of α becomes unfeasible. However, in such cases the correct estimation of α does not affect the source extraction, e.g., if $\theta = 0$, Q in (3) is just an identity matrix; if $\theta = \pi/2$, Q only contains off-diagonal phase factors which are 'absorbed' by the source signals.

(iv) Endorsing the previous point we have that, for accurate estimates of $(\theta,\, \alpha)$, ISR $\approx \sigma_{\hat\theta}^2 + \frac{1}{4}\sigma_{\hat\alpha}^2 \sin^2 2\theta$, so that ISR is lower bounded by $2 \times \mathrm{CRLB}_\theta$. When $\theta = n\pi/2$, $n \in \mathbb{Z}$, and if $\hat\theta$ is still precise enough, this bound decreases to CRLB_θ. That is, the lower bound of separation-performance objective measure ISR is independent of θ and is (asymptotically) determined by the source statistics only [via I in (19)].

3.4. Simulation Results

A simple simulation experiment compares the behaviour of the CEML, CAEML and CWE (with $w = 1/3$ and $w = 1/2$, which would correspond to the complex extensions of AML and MaSS-FOC, resp.). Two independent CGGDs are used as sources. Average ISR results as a function of sks and skd are displayed in Fig. 4. As expected, the CEML and CAEML worsen near $\gamma = 0$ and $\eta = 0$, respectively. By contrast, the CWE maintains a satisfactory separation in both tested cases over all γ and η range, and, as occurred in the real case (Fig. 2), its performance follows closely the CRLB trend.

4. CONCLUSIONS AND OUTLOOK

A new class of closed-form estimators of the separation parameters in the fundamental two-signal instantaneous linear mixture BSS problem has been investigated. A weighted estimator (WE) arises from the linear combination of the EML and AEML centroids, and produces consistent estimates under rather general conditions (essentially, if at most one source is Gaussian). For real-valued mixtures, prior knowledge on the source statistics can be exploited by selecting the WE with optimal large-sample performance (minimum asymptotic variance). With the aid of the bicomplex numbers the WE has also been extended to the complex-mixture case, where it has shown a performance variation similar to the CRLB, that we have derived for circular sources.

Paths of further research include the asymptotic performance analysis of the WE in the complex environment, which is of relevance in areas as important as digital communications. Also, in order to enable a fully blind operation, it is necessary to develop the optimal weight coefficient as a function of the array-output statistics. The estimator's behaviour in the presence of additive noise and impulsive interference needs to be explored as well.

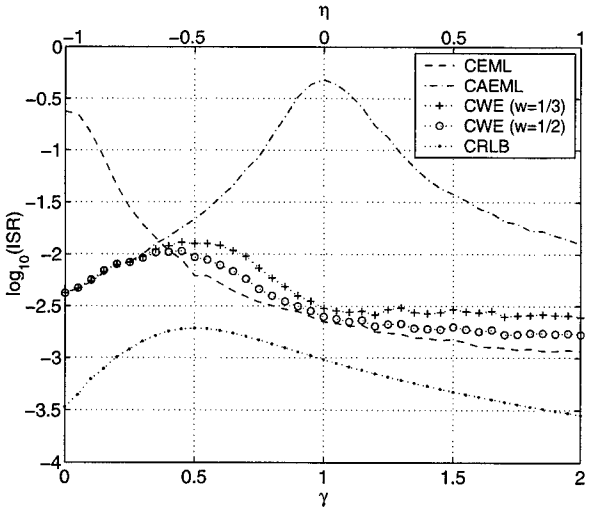


Fig. 4. ISR vs. sks γ and skd η. CGGD sources, $\kappa_{04}^x = 0.5$, $\theta = 15°$, $\alpha = 65°$, $T = 5 \times 10^3$ samples, 10^3 independent Monte Carlo iterations.

5. REFERENCES

[1] V. Zarzoso and A. K. Nandi, "Blind Source Separation," in *Blind Estimation Using Higher-Order Statistics*, A. K. Nandi (Ed.), pp. 167–252. Kluwer Academic Publishers, Boston, 1999.

[2] P. Comon, "Independent Component Analysis, A New Concept?," *Signal Processing*, Vol. 36, No. 3, pp. 287–314, Apr. 1994.

[3] P. Comon, "Separation of Sources Using Higher-Order Cumulants," in *Proc. SPIE*, San Diego, CA, 1989, Vol. 1152, pp. 170–181.

[4] V. Zarzoso, *Closed-Form Higher-Order Estimators for Blind Separation of Independent Source Signals in Instantaneous Linear Mixtures*, Ph.D. thesis, The University of Liverpool, UK, Oct. 1999.

[5] V. Zarzoso and A. K. Nandi, "Unified Formulation of Closed-Form Estimators for Blind Source Separation in Real Instantaneous Linear Mixtures," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, Vol. V, pp. 3160–3163.

[6] F. Harroy and J.-L. Lacoume, "Maximum Likelihood Estimators and Cramer-Rao Bounds in Source Separation," *Signal Processing*, Vol. 55, No. 2, pp. 167–177, Dec. 1996.

[7] V. Zarzoso and A. K. Nandi, "Blind Separation of Independent Sources for Virtually Any Source Probability Density Function," *IEEE Transactions on Signal Processing*, Vol. 47, No. 9, pp. 2419–2432, Sept. 1999.

[8] V. Zarzoso, A. K. Nandi, F. Herrmann, and J. Millet-Roig, "Combined Estimation Scheme for Blind Source Separation with Arbitrary Source PDFs," *IEE Electronics Letters*, Vol. 37, No. 2, pp. 132–133, Jan. 18, 2001.

[9] M. Ghogho, A. Swami, and T. Durrani, "Approximate Maximum Likelihood Blind Source Separation with Arbitrary Source Pdfs," in *Proc. IEEE SSAP Workshop*, Pocono Manor Inn, PA, Aug. 2000.

[10] F. Herrmann, *Independent Component Analysis with Applications to Blind Source Separation*, Ph.D. thesis, The University of Liverpool, UK, Sept. 2000.

[11] V. Zarzoso and A. K. Nandi, "Unified Formulation of Closed-Form Estimators for Blind Source Separation in Complex Instantaneous Linear Mixtures," in *Proc. EUSIPCO*, Tampere, Finland, Sept. 2000, Vol. I, pp. 597–601.

[12] E. L. Lehmann, *Theory of Point Estimation*, Wadsworth, Inc., Pacific Grove, CA, 1991.

LARGE SAMPLE PERFORMANCE ANALYSIS OF ACMA

Alle-Jan van der Veen

Delft University of Technology, Department of Electrical Engineering/DIMES
Mekelweg 4, 2628 CD Delft, The Netherlands

The "Algebraic Constant Modulus Algorithm" (ACMA) is a non-iterative block algorithm for blind separation of constant modulus sources. We previously showed that, unlike CMA, it asymptotically converges to the (non-blind) Wiener receiver. In this paper, we present a finite sample statistical performance analysis. This can be used to predict the SINR performance, as well as the deviation from the Wiener receivers. The theoretical performance is illustrated by numerical simulations and shows a good match.

1. INTRODUCTION

In this paper we study the performance of ACMA ("Analytical Constant Modulus Algorithm"), proposed in [1]. ACMA is a non-recursive blind source separation algorithm for constant modulus signals. It is a batch algorithm that under noise-free conditions can compute exact separating beamformers for all sources at the same time, using only a small number of samples. Although it has been derived as a deterministic method, it is closely related to JADE and other fourth-order statistics based source separation techniques.

We could recently show that (unlike CMA), ACMA beamformers converge asymptotically in the number of samples to the (non-blind) Wiener receivers [2]. Here, we will extend the analysis by deriving the large *finite* sample performance of a block of N samples. For this we need the statistics of the eigenvectors of a fourth order covariance matrix with non-Gaussian sources.

2. DATA MODEL

We consider a linear data model of the form

$$\mathbf{x}_k = \mathbf{A}\mathbf{s}_k + \mathbf{n}_k\,, \tag{1}$$

where $\mathbf{x}_k \in \mathbb{C}^M$ is the data vector received by an array of M sensors at time k, $\mathbf{s}_k \in \mathbb{C}^d$ is the source vector at time k, and $\mathbf{n}_k \in \mathbb{C}^M$ an additive noise vector. $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_d]$ represents an $M\times d$ complex-valued instantaneous mixing matrix (or array response matrix). The sources are constant modulus (CM), i.e. each entry s_i of $\mathbf{s}$ satisfies $|s_i| = 1$.

We collect N samples in a matrix $\mathbf{X} = [\mathbf{x}_1\,,\,\cdots\,,\,\mathbf{x}_N] : M\times N$. Similarly defining $\mathbf{S} : d\times N$ and $\mathbf{N} : M\times N$, we obtain

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}. \tag{2}$$

$\mathbf{A}$, $\mathbf{S}$ and $\mathbf{N}$ are unknown. The objective is to reconstruct $\mathbf{S}$ using linear beamforming, i.e., to find a beamforming matrix $\mathbf{W} = [\mathbf{w}_1\,,\cdots,\mathbf{w}_d] \in \mathbb{C}^{M\times d}$ of full row rank d such that $\hat{\mathbf{S}} = \mathbf{W}^{\mathrm{H}}\mathbf{X}$ approximates $\mathbf{S}$. Since $\mathbf{S}$ is unknown, the criterion for this is that $\hat{\mathbf{S}}$ should be as close to a CM matrix as possible, i.e., we aim to make $|\hat{\mathbf{S}}_{ik}| = |\mathbf{w}_i^{\mathrm{H}}\mathbf{x}_k| = 1\ \forall i,k$. If this is the case, then $\hat{\mathbf{S}}$ is equal to $\mathbf{S}$ up to unknown permutations and unit-norm scalings of its rows. With noise, we can obviously recover the sources only approximatively.

We work under the following assumptions:

1. $N \geq d^2$. $\mathbf{A}$ has full rank d, and $M \geq d$. To avoid complications in the analysis, we assume $M = d$.
2. The sources are statistically independent constant modulus sources, circularly symmetric, with covariance $\mathbf{R}_s := \mathrm{E}(\mathbf{s}\mathbf{s}^{\mathrm{H}}) = \mathbf{I}$.
3. The noise is additive white Gaussian, zero mean, circularly symmetric, independent from the sources, with covariance $\mathbf{R}_n := \mathrm{E}(\mathbf{n}\mathbf{n}^{\mathrm{H}}) = \sigma^2\mathbf{I}$.

Notation Overbar ($\bar{}$) denotes complex conjugation, ^{T} is the matrix transpose, ^{H} the matrix complex conjugate transpose, † the matrix pseudo-inverse (Moore-Penrose inverse). $\mathbf{I}$ (or $\mathbf{I}_p$) is the $(p\times p)$ identity matrix; $\mathbf{e}_i$ is its i-th column. $\mathbf{0}$ and $\mathbf{1}$ are vectors with all entries equal to 0 and 1, respectively. $\mathrm{vec}(\mathbf{A})$ is a stacking of the columns of a matrix $\mathbf{A}$ into a vector. For a vector, $\mathrm{diag}(\mathbf{v})$ is a diagonal matrix with the entries of $\mathbf{v}$ on the diagonal. $\odot$ is the Schur-Hadamart (entry-wise) matrix product, $\otimes$ is the Kronecker product, $\circ$ is the Khatri-Rao product, which is a column-wise Kronecker product. $\mathrm{E}(\cdot)$ denotes the expectation operator.

For a matrix-valued stochastic variable $\hat{\mathbf{R}}$, define its covariance matrix $\mathrm{cov}\{\hat{\mathbf{R}}\} = \mathrm{E}\{[\mathrm{vec}(\hat{\mathbf{R}}-\mathrm{E}(\hat{\mathbf{R}}))][\mathrm{vec}(\hat{\mathbf{R}}-\mathrm{E}(\hat{\mathbf{R}}))]^{\mathrm{H}}\}$.

For a zero mean random vector $\mathbf{x} = [x_i]$, define the fourth order cumulant matrix

$$\begin{aligned}\mathbf{K}_{\mathbf{x}} \;=\; & \mathrm{E}(\bar{\mathbf{x}}\otimes\mathbf{x})(\bar{\mathbf{x}}\otimes\mathbf{x})^{\mathrm{H}} - \mathrm{E}(\bar{\mathbf{x}}\otimes\mathbf{x})\mathrm{E}(\bar{\mathbf{x}}\otimes\mathbf{x})^{\mathrm{H}} - \mathrm{E}(\bar{\mathbf{x}}\mathbf{x}^{\mathrm{H}})\otimes\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{H}}) \\ & - \mathrm{E}(\bar{\mathbf{x}}\otimes\mathbf{1})(\mathbf{1}\otimes\mathbf{x})^{\mathrm{H}} \odot \mathrm{E}(\mathbf{1}\otimes\mathbf{x})(\bar{\mathbf{x}}\otimes\mathbf{1})^{\mathrm{H}}.\end{aligned}$$

For circularly symmetric variables, the last term vanishes.

3. FORMULATION OF THE ALGORITHM

In brief outline, ACMA consists of two main steps: a prewhitening operation, and the algorithm proper. Define the data covariance matrix and its sample estimate

$$\mathbf{R}_{\mathbf{x}} := \mathrm{E}\{\mathbf{x}\mathbf{x}^{\mathrm{H}}\}, \qquad \hat{\mathbf{R}}_{\mathbf{x}} := \frac{1}{N}\sum \mathbf{x}_k\mathbf{x}_k^{\mathrm{H}}.$$

Assuming that $M = d$ for simplicity of the analysis, the prewhitening filter transforms the data to

$$\underline{\mathbf{X}} := \hat{\mathbf{R}}_{\mathbf{x}}^{-1/2}\mathbf{X} \quad =: \quad \underline{\mathbf{A}}\mathbf{S} + \underline{\mathbf{N}}$$

where the underscore indicates the prewhitening. Note that $\hat{\underline{\mathbf{R}}}_{\mathbf{x}} = \mathbf{I}$.

Given the N data samples $[\mathbf{x}_k]$, the purpose of a beamforming vector $\mathbf{w}$ is to recover one of the sources as $\hat{s}_k = \mathbf{w}^{\mathrm{H}}\mathbf{x}_k$. One technique for estimating such a beamformer is by minimizing the deterministic CMA(2,2) cost function, $\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}}\ \frac{1}{N}\sum(|\mathbf{w}^{\mathrm{H}}\mathbf{x}_k|^2-1)^2$. Define

$$\hat{\underline{\mathbf{C}}}_{\mathbf{x}} = \tfrac{1}{N}\sum(\bar{\underline{\mathbf{x}}}_k\otimes\underline{\mathbf{x}}_k)(\bar{\underline{\mathbf{x}}}_k\otimes\underline{\mathbf{x}}_k)^{\mathrm{H}} - \left[\tfrac{1}{N}\sum\bar{\underline{\mathbf{x}}}_k\otimes\underline{\mathbf{x}}_k\right]\left[\tfrac{1}{N}\sum\bar{\underline{\mathbf{x}}}_k\otimes\underline{\mathbf{x}}_k\right]^{\mathrm{H}}.$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

In [2], we have derived that CMA(2,2) is equivalent to (up to a scaling of $\mathbf{w}$ which is not of interest to its performance)

$$\hat{\mathbf{w}} = \hat{\mathbf{R}}_{\mathbf{x}}^{-1/2}\hat{\mathbf{t}}, \qquad \hat{\mathbf{t}} = \underset{\substack{\mathbf{y}=\bar{\mathbf{t}}\otimes\mathbf{t}\\ \|\mathbf{y}\|=1}}{\operatorname{argmin}} \mathbf{y}^{\mathrm{H}}\hat{\underline{\mathbf{C}}}_{\mathbf{x}}\mathbf{y}, \tag{3}$$

ACMA is obtained as a two-step approach to the latter minimization problem [2]:

1. Find an orthonormal basis $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_d]$ of independent minimizers of $\mathbf{y}^{\mathrm{H}}\hat{\underline{\mathbf{C}}}_{\mathbf{x}}\mathbf{y}$, i.e., the eigenvectors corresponding to the d smallest eigenvalues of $\hat{\underline{\mathbf{C}}}_{\mathbf{x}}$.
2. Find a basis $\{\bar{\hat{\mathbf{t}}}_1 \otimes \hat{\mathbf{t}}_1, \cdots, \bar{\hat{\mathbf{t}}}_d \otimes \hat{\mathbf{t}}_d\}$ that spans the same linear subspace as $\{\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_d\}$, and with $\|\hat{\mathbf{t}}_i\| = 1$, i.e., solve

$$\hat{\mathbf{T}} = \min_{\mathbf{T},\mathbf{M}} \|\hat{\mathbf{Y}} - (\bar{\mathbf{T}} \circ \mathbf{T})\mathbf{M}\|_{\mathrm{F}}^2, \tag{4}$$

subject to the constraint $\operatorname{diag}(\mathbf{T}^{\mathrm{H}}\mathbf{T}) = \mathbf{I}$.

It was shown in [2] that $\hat{\mathbf{T}}$ converges asymptotically in N to a matrix $\mathbf{T} = \underline{\mathbf{A}}_0$, where $\underline{\mathbf{A}}_0$ is equal to $\underline{\mathbf{A}}$ except for a scaling and permutation of its columns. In the non-whitened domain, $\hat{\mathbf{W}} = \hat{\mathbf{R}}_{\mathbf{x}}^{-1/2}\hat{\mathbf{T}}$ converges asymptotically to $\mathbf{W} = \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{A}_0$, the Wiener receiver (except for the scaling and the permutation).

A performance analysis is now possible, and follows in outline the analysis of the MUSIC and WSF DOA estimators [3], but extended to fourth order statistics of non-Gaussian sources. The following limitations are introduced to keep the derivations tractable.

1. N is sufficiently large, and we neglect terms of order N^{-2} over terms of order N^{-1}. The noise power σ^2 is sufficiently small and we neglect σ^4 over σ^2.
2. We assume that the prewhitening step is based on the *true* covariance matrix $\mathbf{R}_{\mathbf{x}}$. (This is accurate for $M = d$.)
3. We assume that the exact solution to (4) is computed.

4. COVARIANCE OF $\hat{\mathbf{C}}_{\mathbf{x}}$

In this and the next sections, we drop for convenience the underscore from the notation since all variables are based on whitened data. Our objective in this section is to find a compact approximative expression for the covariance of $\hat{\mathbf{C}}_{\mathbf{x}}$, denoted by $\mathbf{\Omega}_{\mathbf{x}}$. Define

$$\mathbf{C}_{\mathbf{x}} = \mathrm{E}\{(\bar{\mathbf{x}}_k \otimes \mathbf{x}_k)(\bar{\mathbf{x}}_k \otimes \mathbf{x}_k)^{\mathrm{H}}\} - \mathrm{E}\{\bar{\mathbf{x}}_k \otimes \mathbf{x}_k\}\mathrm{E}\{\bar{\mathbf{x}}_k \otimes \mathbf{x}_k\}^{\mathrm{H}}.$$

Using properties of cumulants, we can show that [2]

$$\mathbf{C}_{\mathbf{x}} = -[\bar{\mathbf{A}} \circ \mathbf{A}][\bar{\mathbf{A}} \circ \mathbf{A}]^{\mathrm{H}} + \bar{\mathbf{R}}_{\mathbf{x}} \otimes \mathbf{R}_{\mathbf{x}} = -[\bar{\mathbf{A}} \circ \mathbf{A}][\bar{\mathbf{A}} \circ \mathbf{A}]^{\mathrm{H}} + \mathbf{I}. \tag{5}$$

Furthermore, a straightforward derivation shows that

$$\operatorname{cov}\{\hat{\mathbf{R}}_{\mathbf{x}}\} = \tfrac{1}{N}\mathbf{C}_{\mathbf{x}}. \tag{6}$$

Thus, $\mathbf{C}_{\mathbf{x}}$ is the covariance of $\hat{\mathbf{R}}_{\mathbf{x}}$, and $\hat{\mathbf{C}}_{\mathbf{x}}$ is a (biased) sample estimate of it. A second interpretation of $\mathbf{C}_{\mathbf{x}}$ is obtained by defining a "data" sequence

$$\mathbf{g}_k := \bar{\mathbf{x}}_k \otimes \mathbf{x}_k - \mathrm{E}\{\bar{\mathbf{x}}_k \otimes \mathbf{x}_k\}, \qquad k = 1, \cdots, N, \tag{7}$$

and considering its covariance and sample covariance

$$\mathbf{R}_{\mathbf{g}} := \mathrm{E}\{\mathbf{g}_k\mathbf{g}_k^{\mathrm{H}}\}, \qquad \hat{\mathbf{R}}_{\mathbf{g}} := \tfrac{1}{N}\textstyle\sum \mathbf{g}_k\mathbf{g}_k^{\mathrm{H}}.$$

It is straightforward to show that

$$\mathrm{E}\{\hat{\mathbf{R}}_{\mathbf{g}}\} = \mathbf{R}_{\mathbf{g}} = \mathbf{C}_{\mathbf{x}}, \quad \hat{\mathbf{R}}_{\mathbf{g}} = \hat{\mathbf{C}}_{\mathbf{x}}(1 + \mathcal{O}(\tfrac{1}{N})).$$

Thus, $\mathbf{C}_{\mathbf{x}}$ is the covariance of $\mathbf{g}_k$, and $\hat{\mathbf{R}}_{\mathbf{g}}$ is an unbiased sample estimate of it; in first order approximation it has the same properties as the biased estimate $\hat{\mathbf{C}}_{\mathbf{x}}$. Similar to (6), it follows that $\operatorname{cov}\{\hat{\mathbf{R}}_{\mathbf{g}}\} = \frac{1}{N}\mathbf{C}_{\mathbf{g}}$ where

$$\mathbf{C}_{\mathbf{g}} := \mathrm{E}\{(\bar{\mathbf{g}} \otimes \mathbf{g})(\bar{\mathbf{g}} \otimes \mathbf{g})^{\mathrm{H}}\} - \mathrm{E}\{\bar{\mathbf{g}} \otimes \mathbf{g}\}\mathrm{E}\{\bar{\mathbf{g}} \otimes \mathbf{g}\}^{\mathrm{H}}. \tag{8}$$

In summary, we can prove

Theorem 1. $\mathbf{\Omega}_{\mathbf{x}} := \operatorname{cov}\{\hat{\mathbf{C}}_{\mathbf{x}}\} = \frac{1}{N}\mathbf{C}_{\mathbf{g}} + \mathcal{O}(\frac{1}{N^2})$.

It remains to find a compact description of $\mathbf{C}_{\mathbf{g}}$ in terms of our data model. Inserting the model $\mathbf{x}_k = \mathbf{A}\mathbf{s}_k + \mathbf{n}_k$ in the definition of $\mathbf{g}_k$, we obtain

$$\mathbf{g}_k = \mathbf{A}_{\mathbf{c}}\mathbf{c}_k + \underset{\sim}{\mathbf{n}}_k$$

where

$$\begin{aligned}
\mathbf{c} &:= \sum_{i \neq j} \mathbf{e}'_{ij}\bar{s}_j s_i = [\bar{s}_1 s_2, \cdots, \bar{s}_1 s_d,\ \bar{s}_2 s_1,\ \bar{s}_2 s_3, \cdots]^{\mathrm{T}} \\
\mathbf{A}_{\mathbf{c}} &:= [\bar{\mathbf{a}}_j \otimes \mathbf{a}_i]_{i \neq j} \\
\underset{\sim}{\mathbf{n}} &:= \bar{\mathbf{n}} \otimes \mathbf{n} - \mathbf{R}_{\mathbf{n}} + \bar{\mathbf{A}}\bar{\mathbf{s}} \otimes \mathbf{n} + \bar{\mathbf{n}} \otimes \mathbf{A}\mathbf{s}
\end{aligned}$$

where $\mathbf{e}'_{ij} = \operatorname{vec}'(\mathbf{e}_i\mathbf{e}_j^{\mathrm{H}})$, and $\operatorname{vec}'(\cdot)$ is a vectoring operator which skips the main diagonal. The vector $\mathbf{c}$ is CM (with certain dependencies among its entries). Likewise, the matrix $\mathbf{A}_{\mathbf{c}}$ skips the $\bar{\mathbf{a}}_i \otimes \mathbf{a}_i$ columns of $\bar{\mathbf{A}} \otimes \mathbf{A}$.

The model $\mathbf{g}_k = \mathbf{A}_{\mathbf{c}}\mathbf{c}_k + \underset{\sim}{\mathbf{n}}_k$ has several properties that are similar to that of $\mathbf{x}_k = \mathbf{A}\mathbf{s}_k + \mathbf{n}_k$. However, $\mathbf{c}$ and $\underset{\sim}{\mathbf{n}}$ are not independent (only uncorrelated), not circularly symmetric, and $\mathbf{K}_{\underset{\sim}{\mathbf{n}}} \neq \mathbf{0}$. A good approximation for $\mathbf{C}_{\mathbf{g}}$ taking into account all terms up to $\mathcal{O}(\sigma^2)$, is given as

Theorem 2. $\mathbf{C}_{\mathbf{g}} \approx [\bar{\mathbf{A}}_{\mathbf{c}} \otimes \mathbf{A}_{\mathbf{c}}]\mathbf{K}'_{\mathbf{c}}[\bar{\mathbf{A}}_{\mathbf{c}} \otimes \mathbf{A}_{\mathbf{c}}]^{\mathrm{H}} + \bar{\mathbf{R}}_{\mathbf{g}} \otimes \mathbf{R}_{\mathbf{g}} + \mathbf{E} + \mathbf{E}^{\mathrm{H}}$ *where*

$$\begin{aligned}
\mathbf{E} &= [\mathbf{A} \otimes \bar{\mathbf{R}}_{\mathbf{n}}^{1/2} \otimes \mathbf{A}_{\mathbf{c}}]\mathbf{E}_1[\bar{\mathbf{A}}_{\mathbf{c}} \otimes \bar{\mathbf{R}}_{\mathbf{n}}^{1/2} \otimes \mathbf{A}]^{\mathrm{H}} \\
&\quad + [\mathbf{R}_{\mathbf{n}}^{1/2} \otimes \bar{\mathbf{A}} \otimes \mathbf{A}_{\mathbf{c}}]\mathbf{E}_2[\bar{\mathbf{A}}_{\mathbf{c}} \otimes \bar{\mathbf{A}} \otimes \mathbf{R}_{\mathbf{n}}^{1/2}]^{\mathrm{H}} \\
\mathbf{K}'_{\mathbf{c}} &= \mathbf{K}_{\mathbf{c}} + \sum_{i \neq j} \sum_{k \neq l} (\mathbf{e}'_{ij} \otimes \mathbf{e}'_{kl})(\mathbf{e}'_{lk} \otimes \mathbf{e}'_{ji})^{\mathrm{H}} \\
\mathbf{K}_{\mathbf{c}} &= -[\sum_{i \neq j} (\mathbf{e}'_{ij} \otimes \mathbf{e}'_{ij})(\mathbf{e}'_{ij} \otimes \mathbf{e}'_{ij})^{\mathrm{H}} + (\mathbf{e}'_{ji} \otimes \mathbf{e}'_{ij})(\mathbf{e}'_{ji} \otimes \mathbf{e}'_{ij})^{\mathrm{H}} \\
&\qquad + (\mathbf{e}'_{ij} \otimes \mathbf{e}'_{ij})(\mathbf{e}'_{ji} \otimes \mathbf{e}'_{ji})^{\mathrm{H}}] \\
\mathbf{E}_1 &= \sum_i \sum_{j \neq i} \sum_{k \neq i} \mathbf{e}'^{\mathrm{H}}_{ji} \otimes \mathbf{e}_k \otimes \mathbf{I}_d \otimes \mathbf{e}_j^{\mathrm{H}} \otimes \mathbf{e}'_{ik} \\
&\quad + \mathbf{e}'^{\mathrm{H}}_{ij} \otimes \mathbf{e}_j \otimes \mathbf{I}_d \otimes \mathbf{e}_k^{\mathrm{H}} \otimes \mathbf{e}'_{ki} + \mathbf{e}'^{\mathrm{H}}_{ij} \otimes \mathbf{e}_k \otimes \mathbf{I}_d \otimes \mathbf{e}_k^{\mathrm{H}} \otimes \mathbf{e}'_{ji}(1 - \delta_k^j) \\
\mathbf{E}_2 &= \sum_i \sum_{j \neq i} \sum_{k \neq i} \mathbf{e}'^{\mathrm{H}}_{ji} \otimes \mathbf{e}_k^{\mathrm{H}} \otimes \mathbf{I}_d \otimes \mathbf{e}_j \otimes \mathbf{e}'_{ik} \\
&\quad + \mathbf{e}'^{\mathrm{H}}_{ij} \otimes \mathbf{e}_j^{\mathrm{H}} \otimes \mathbf{I}_d \otimes \mathbf{e}_k \otimes \mathbf{e}'_{ki} + \mathbf{e}'^{\mathrm{H}}_{ij} \otimes \mathbf{e}_k^{\mathrm{H}} \otimes \mathbf{I}_d \otimes \mathbf{e}_k \otimes \mathbf{e}'_{ji}(1 - \delta_k^j).
\end{aligned}$$

(All indices range over $1, \cdots, d$. Note, the latter matrices are data independent and simply collections of '1' entries.)

PROOF Omitted.

It can be shown experimentally that the term $\bar{\mathbf{R}}_{\mathbf{g}} \otimes \mathbf{R}_{\mathbf{g}}$ is the dominant term, so that

$$\mathbf{C}_{\mathbf{g}} \approx \bar{\mathbf{C}}_{\mathbf{x}} \otimes \mathbf{C}_{\mathbf{x}} \tag{9}$$

is a good approximation. This is the same as regarding $\mathbf{c}$ and $\underset{\sim}{\mathbf{n}}$ as Gaussian vectors with independent entries. Making this approximation would lead to particularly simple results in the eigenvector perturbation study and subsequent steps, as we basically can apply the theory in Viberg [3].

5. EIGENVECTOR PERTURBATION

In this section we consider the statistical properties of the eigenvectors of $\hat{\mathbf{C}}_{\mathbf{x}}$, a fourth order sample covariance matrix based on nonGaussian signals. We first give a general derivation and then specialize to the case at hand. The generalization is needed because most existing derivations consider Gaussian sources.

For a covariance matrix $\mathbf{R}$ with unbiased sample estimate $\hat{\mathbf{R}}$ based on N samples of a (not necessarily Gaussian) vector process, consider the eigenvalue decompositions $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{H}}$, $\hat{\mathbf{R}} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^{\mathrm{H}}$. If we elaborate on the equality

$$\hat{\mathbf{R}}-\mathbf{R} = (\hat{\mathbf{U}}-\mathbf{U})\mathbf{\Lambda}\mathbf{U}^{\mathrm{H}}-\hat{\mathbf{R}}(\hat{\mathbf{U}}-\mathbf{U})\mathbf{U}^{\mathrm{H}} + \hat{\mathbf{U}}(\hat{\mathbf{\Lambda}}-\mathbf{\Lambda})\mathbf{U}^{\mathrm{H}}$$

and assume that we partition the eigenvalue decomposition of $\mathbf{R}$ as

$$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{H}} = \mathbf{U}_s\mathbf{\Lambda}_s\mathbf{U}_s^{\mathrm{H}}+\mathbf{U}_n\mathbf{\Lambda}_n\mathbf{U}_n^{\mathrm{H}}, \tag{10}$$

where the eigenvalues in $\mathbf{\Lambda}_s$ are distinct and unequal to any eigenvalue in $\mathbf{\Lambda}_n$, then we can derive directly that in first order

$$\mathrm{vec}(\mathbf{P}_n\hat{\mathbf{U}}_s) \approx [\mathbf{I}\otimes\mathbf{U}_n][\mathbf{\Lambda}_s\otimes\mathbf{I}-\mathbf{I}\otimes\mathbf{\Lambda}_n]^{-1}[\bar{\mathbf{U}}_s\otimes\mathbf{U}_n]^{\mathrm{H}}\mathrm{vec}(\hat{\mathbf{R}}-\mathbf{R})$$

where $\mathbf{P}_n = \mathbf{U}_n\mathbf{U}_n^{\mathrm{H}}$. From the latter we can immediately find an expression for the covariance of the "signal" eigenvectors projected into the "noise" subspace:

Lemma 3. *Let $\hat{\mathbf{R}}$ be a sample covariance matrix converging to $\mathbf{R}$, and assume that $\mathbf{R}$ has eigenvalue decomposition (10) where the entries in $\mathbf{\Lambda}_s$ are distinct and unequal to any entry in $\mathbf{\Lambda}_n$. Then*

$$\begin{aligned}\mathrm{cov}\{\mathbf{P}_n\hat{\mathbf{U}}_s\} = {}& [\mathbf{I}\otimes\mathbf{U}_n][\mathbf{\Lambda}_s\otimes\mathbf{I}-\mathbf{I}\otimes\mathbf{\Lambda}_n]^{-1}[\bar{\mathbf{U}}_s\otimes\mathbf{U}_n]^{\mathrm{H}}\cdot\mathrm{cov}\{\hat{\mathbf{R}}\}\cdot\\ &\cdot[\bar{\mathbf{U}}_s\otimes\mathbf{U}_n][\mathbf{\Lambda}_s\otimes\mathbf{I}-\mathbf{I}\otimes\mathbf{\Lambda}_n]^{-1}[\mathbf{I}\otimes\mathbf{U}_n]^{\mathrm{H}} + o(N^{-1}).\end{aligned} \tag{11}$$

Essentially the same result appears in [4], but written as summations and with a more indirect proof.

We now specialize to our situation. We have

$$\begin{aligned}\mathbf{R} \leftrightarrow \mathbf{R}_{\mathbf{g}} &= \mathbf{C}_{\mathbf{x}}\\ \mathrm{cov}\{\hat{\mathbf{R}}\} \leftrightarrow \mathbf{\Omega}_{\mathbf{x}} &= \mathrm{cov}\{\hat{\mathbf{C}}_{\mathbf{x}}\} = \tfrac{1}{N}\mathbf{C}_{\mathbf{g}} + O(N^{-2}).\end{aligned}$$

Introduce the eigenvalue decomposition of $\mathbf{C}_{\mathbf{x}}$ as

$$\mathbf{C}_{\mathbf{x}} = \underset{\sim}{\mathbf{U}}\underset{\sim}{\mathbf{\Lambda}}\underset{\sim}{\mathbf{U}}^{\mathrm{H}} = \underset{\sim}{\mathbf{U}}_s\underset{\sim}{\mathbf{\Lambda}}_s\underset{\sim}{\mathbf{U}}_s^{\mathrm{H}}+\underset{\sim}{\mathbf{U}}_n\underset{\sim}{\mathbf{\Lambda}}_n\underset{\sim}{\mathbf{U}}_n^{\mathrm{H}} \tag{12}$$

where $\underset{\sim}{\mathbf{\Lambda}}_s$ collects the d *smallest* eigenvalues of $\mathbf{C}_{\mathbf{x}}$. Likewise, $\underset{\sim}{\mathbf{U}}_s$ is a basis for the approximate null space of $\mathbf{C}_{\mathbf{x}}$. Also introduce the singular value decomposition

$$\underset{\sim}{\mathbf{A}} := \bar{\mathbf{A}}\circ\mathbf{A} = \underset{\sim}{\mathbf{U}}_{\mathbf{A}}\underset{\sim}{\mathbf{\Sigma}}_{\mathbf{A}}\underset{\sim}{\mathbf{V}}_{\mathbf{A}}, \tag{13}$$

where $\underset{\sim}{\mathbf{U}}_{\mathbf{A}}$ has d orthonormal columns, $\underset{\sim}{\mathbf{\Sigma}}_{\mathbf{A}} = \mathrm{diag}[\varsigma_k]$ is a $d\times d$ diagonal matrix, and $\underset{\sim}{\mathbf{V}}_{\mathbf{A}}$ is $d\times d$ unitary. Let $\underset{\sim}{\mathbf{U}}_{\mathbf{A}}^{\perp}$ be the orthogonal complement of $\underset{\sim}{\mathbf{U}}_{\mathbf{A}}$. It follows from (5) that the eigenvalue decomposition of $\mathbf{C}_{\mathbf{x}}$ is given by

$$\mathbf{C}_{\mathbf{x}} = [\underset{\sim}{\mathbf{U}}_{\mathbf{A}} \quad \underset{\sim}{\mathbf{U}}_{\mathbf{A}}^{\perp}]\begin{bmatrix}\mathbf{I}-\underset{\sim}{\mathbf{\Sigma}}_{\mathbf{A}}^2 & \\ & \mathbf{I}\end{bmatrix}[\underset{\sim}{\mathbf{U}}_{\mathbf{A}} \quad \underset{\sim}{\mathbf{U}}_{\mathbf{A}}^{\perp}]^{\mathrm{H}}. \tag{14}$$

In view of the partitioning in (12) we set $\underset{\sim}{\mathbf{U}}_s = \underset{\sim}{\mathbf{U}}_{\mathbf{A}}$, $\underset{\sim}{\mathbf{\Lambda}}_s = \mathbf{I}-\underset{\sim}{\mathbf{\Sigma}}_{\mathbf{A}}^2$, and $\underset{\sim}{\mathbf{\Lambda}}_n = \mathbf{I}$. Inserting this in (11), we obtain

Theorem 4. $\mathrm{cov}\{\underset{\sim}{\mathbf{P}}_{\mathbf{A}}^{\perp}\hat{\mathbf{U}}_s\} = \frac{1}{N}\mathbf{C}_{\mathbf{u}} + o(N^{-1})$, *where*

$$\mathbf{C}_{\mathbf{u}} := [\underset{\sim}{\mathbf{\Sigma}}_{\mathbf{A}}^{-2}\bar{\underset{\sim}{\mathbf{U}}}_{\mathbf{A}}^{\mathrm{H}}\otimes\underset{\sim}{\mathbf{P}}_{\mathbf{A}}^{\perp}]\,\mathbf{C}_{\mathbf{g}}\,[\bar{\underset{\sim}{\mathbf{U}}}_{\mathbf{A}}\underset{\sim}{\mathbf{\Sigma}}_{\mathbf{A}}^{-2}\otimes\underset{\sim}{\mathbf{P}}_{\mathbf{A}}^{\perp}].$$

Significant simplifications are possible if we allow the approximation of $\mathbf{C}_{\mathbf{g}}$ in (9).

6. SUBSPACE FITTING

6.1. Cost function

The next item in the analysis is the subspace fitting problem in (4). We can follow in outline the performance analysis technique described in [3]. Some notational changes are necessary.

In equation (4), we computed a $d\times d$ separating beamforming matrix $\hat{\mathbf{T}}$ (in the whitened domain), with columns constrained to have unit norm. W.l.o.g., we can further constrain the first nonzero entry of each column to be positive real. Let $\mathbf{A}(\boldsymbol{\theta})$ be a minimal parametrization of such matrices. The true mixing matrix can then be written as $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta}_0)\mathbf{B}$, where $\mathbf{B}$ is a diagonal scaling matrix which is unidentifiable by the subspace fitting. We assume that the true parameter vector $\boldsymbol{\theta}_0$ is uniquely identifiable and that $\mathbf{A}(\boldsymbol{\theta})$ is continuously differentiable around $\boldsymbol{\theta}_0$. We proved in [2] that as $N\to\infty$, $\hat{\mathbf{T}}$ converges to $\mathbf{A}_0 \equiv \mathbf{A}(\boldsymbol{\theta}_0)$, and thus we can write $\hat{\mathbf{T}} = \mathbf{A}(\hat{\boldsymbol{\theta}})$. In this notation, equation (4) becomes

$$\mathbf{A}(\hat{\boldsymbol{\theta}}) = \underset{\mathbf{A}(\boldsymbol{\theta}),\mathbf{M}}{\mathrm{argmin}}\,\|\hat{\underset{\sim}{\mathbf{U}}}_s-\underset{\sim}{\mathbf{A}}(\boldsymbol{\theta})\mathbf{M}\|_{\mathrm{F}}^2, \qquad \underset{\sim}{\mathbf{A}}(\boldsymbol{\theta}) := \bar{\mathbf{A}}(\boldsymbol{\theta})\circ\mathbf{A}(\boldsymbol{\theta}).$$

As usual, the problem is separable, and the optimum for $\mathbf{M}$ given $\underset{\sim}{\mathbf{A}}(\boldsymbol{\theta})$ is $\underset{\sim}{\mathbf{A}}(\boldsymbol{\theta})^{\dagger}\hat{\underset{\sim}{\mathbf{U}}}_s$. Eliminating $\mathbf{M}$, we obtain

$$\mathbf{A}(\boldsymbol{\theta}) = \underset{\mathbf{A}(\boldsymbol{\theta})}{\mathrm{argmin}}\,\|\underset{\sim}{\mathbf{P}}_{\mathbf{A}(\boldsymbol{\theta})}^{\perp}\hat{\underset{\sim}{\mathbf{U}}}_s\|_{\mathrm{F}}^2$$

where $\underset{\sim}{\mathbf{P}}_{\mathbf{A}(\boldsymbol{\theta})}^{\perp} = \mathbf{I}-\underset{\sim}{\mathbf{A}}(\boldsymbol{\theta})\underset{\sim}{\mathbf{A}}(\boldsymbol{\theta})^{\dagger}$. Hence we will consider the minimization of the cost function

$$J(\boldsymbol{\theta}) = \|\underset{\sim}{\mathbf{P}}_{\mathbf{A}(\boldsymbol{\theta})}^{\perp}\hat{\underset{\sim}{\mathbf{U}}}_s\|_{\mathrm{F}}^2 = \mathrm{vec}(\underset{\sim}{\mathbf{P}}_{\mathbf{A}(\boldsymbol{\theta})}^{\perp}\hat{\underset{\sim}{\mathbf{U}}}_s)^{\mathrm{H}}\mathrm{vec}(\underset{\sim}{\mathbf{P}}_{\mathbf{A}(\boldsymbol{\theta})}^{\perp}\hat{\underset{\sim}{\mathbf{U}}}_s) \tag{15}$$

(This can be generalized to a *weighted* norm as usual.)

6.2. Covariance of $\hat{\boldsymbol{\theta}}$

Choose a specific parametrization of $\mathbf{A}(\boldsymbol{\theta})$. Since the columns of $\mathbf{A}(\boldsymbol{\theta})$ are not coupled, we can write $\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\boldsymbol{\theta}_1), \cdots, \mathbf{a}(\boldsymbol{\theta}_d)]$, where $\mathbf{a}(\boldsymbol{\theta}_i)$ is a parametrization of a unit-norm vector with real non-negative first entry, which requires $p := 2(d-1)$ real-valued parameters per vector. Denote θ_{ij} the i-th parameter of $\boldsymbol{\theta}_j$, and define the derivative matrix

$$\mathbf{D} = [\frac{\partial\mathbf{a}_1}{\partial\theta_{11}}, \frac{\partial\mathbf{a}_1}{\partial\theta_{21}}, \cdots, \frac{\partial\mathbf{a}_2}{\partial\theta_{12}}, \cdots](\boldsymbol{\theta}_0). \tag{16}$$

Theorem 5. *Let $\underset{\sim}{\mathbf{A}}_0 := \bar{\mathbf{A}}(\boldsymbol{\theta}_0)\circ\mathbf{A}(\boldsymbol{\theta}_0)$, $\mathbf{A}_e := \mathbf{A}(\boldsymbol{\theta}_0)\otimes\mathbf{1}_p^{\mathrm{T}}$,*

$$\begin{aligned}\underset{\sim}{\mathbf{D}} &= \bar{\mathbf{A}}_e\circ\mathbf{D}+\bar{\mathbf{D}}\circ\mathbf{A}_e\\ \mathbf{M} &:= (\underset{\sim}{\mathbf{A}}_0^{\dagger}\underset{\sim}{\mathbf{U}}_{\mathbf{A}})^{\mathrm{H}}\otimes\mathbf{1}_p^{\mathrm{T}}\\ \mathbf{C}_{\mathbf{u}} &:= [\underset{\sim}{\mathbf{\Sigma}}_{\mathbf{A}}^{-2}\bar{\underset{\sim}{\mathbf{U}}}_{\mathbf{A}}^{\mathrm{H}}\otimes\underset{\sim}{\mathbf{P}}_{\mathbf{A}}^{\perp}]\,\mathbf{C}_{\mathbf{g}}\,[\bar{\underset{\sim}{\mathbf{U}}}_{\mathbf{A}}\underset{\sim}{\mathbf{\Sigma}}_{\mathbf{A}}^{-2}\otimes\underset{\sim}{\mathbf{P}}_{\mathbf{A}}^{\perp}]\\ \mathbf{Q} &:= 4[\mathbf{M}\circ\underset{\sim}{\mathbf{P}}_{\mathbf{A}}^{\perp}\underset{\sim}{\mathbf{D}}]^{\mathrm{H}}\mathbf{C}_{\mathbf{u}}[\mathbf{M}\circ\underset{\sim}{\mathbf{P}}_{\mathbf{A}}^{\perp}\underset{\sim}{\mathbf{D}}]\\ \mathbf{H} &:= 2[\mathbf{M}\circ\underset{\sim}{\mathbf{P}}_{\mathbf{A}}^{\perp}\underset{\sim}{\mathbf{D}}]^{\mathrm{H}}[\mathbf{M}\circ\underset{\sim}{\mathbf{P}}_{\mathbf{A}}^{\perp}\underset{\sim}{\mathbf{D}}],\end{aligned}$$

where $\underset{\sim}{\mathbf{U}}_{\mathbf{A}}$ and $\underset{\sim}{\mathbf{\Sigma}}_{\mathbf{A}}$ are defined in (13). For large N, the covariance of $\hat{\boldsymbol{\theta}}$ that minimizes the subspace fitting problem (15) is in first order approximation

$$\mathbf{R}_{\boldsymbol{\theta}} := \mathrm{cov}\{\hat{\boldsymbol{\theta}}\} = \tfrac{1}{N}\mathbf{H}^{-1}\mathbf{Q}\mathbf{H}^{-1}.$$

PROOF Omitted; along the lines of [3].

6.3. Covariance of T

It remains to map the previous result to an expression for the covariance of the beamforming vectors. With some abuse of notation, let $\mathbf{t} = \mathrm{vec}(\mathbf{T})$, where $\mathbf{T} = \mathbf{A}(\boldsymbol{\theta}_0)$, and let $\hat{\mathbf{t}} = \mathrm{vec}(\hat{\mathbf{T}}) = \mathrm{vec}(\mathbf{A}(\hat{\boldsymbol{\theta}}))$. Then, for small perturbations, $\hat{\mathbf{t}} = \mathbf{t} + \sum_\eta \frac{\partial \mathbf{t}}{\partial \theta_\eta}(\hat{\theta}_\eta - \theta_\eta)$, so that $\hat{\mathbf{t}}$ has covariance

$$\begin{aligned}\mathbf{R_t} &= [\frac{\partial \mathbf{t}}{\partial \theta_{11}}, \frac{\partial \mathbf{t}}{\partial \theta_{21}}, \cdots]\mathbf{R_\theta}[\frac{\partial \mathbf{t}}{\partial \theta_{11}}, \frac{\partial \mathbf{t}}{\partial \theta_{21}}, \cdots]^{\mathrm{H}} \\ &= [(\mathbf{I}_d \otimes \mathbf{1}_p^{\mathrm{T}}) \circ \mathbf{D}]\mathbf{R_\theta}[(\mathbf{I}_d \otimes \mathbf{1}_p^{\mathrm{T}}) \circ \mathbf{D}]^{\mathrm{H}},\end{aligned} \tag{17}$$

where $\mathbf{D}$ was defined in (16). The covariance of a beamformer $\mathbf{t}_j$ is the jj-th subblock of size $p \times p$ of $\mathbf{R_t}$.

6.4. SINR performance

To allow a better interpretation of the performance of the beamformers, we derive a mapping of $\mathbf{R_t}$ to the inverse SINR, or the INSR (interference plus noise to signal ratio), defined for a beamforming vector $\mathbf{t}$ and array response vector $\mathbf{a}$ of the corresponding source as (recall that $\mathbf{R_x} = \mathbf{I}$)

$$\mathrm{INSR}(\mathbf{t}) := \frac{\mathbf{t}^{\mathrm{H}}(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathrm{H}})\mathbf{t}}{\mathbf{t}^{\mathrm{H}}\mathbf{a}\mathbf{a}^{\mathrm{H}}\mathbf{t}}.$$

The optimal solution that minimizes the INSR is $\mathbf{t} = \alpha\mathbf{a}$ (for an arbitrary nonzero scaling α). Consider a perturbation: $\hat{\mathbf{t}} = \mathbf{t} + \mathbf{d}$ where $\mathbf{t} = \alpha\mathbf{a}$. Then

$$\mathrm{INSR}(\hat{\mathbf{t}}) \approx \frac{1}{\mathbf{a}^{\mathrm{H}}\mathbf{a}}(1 - \mathbf{a}^{\mathrm{H}}\mathbf{a} + \frac{\mathbf{d}^{\mathrm{H}}\mathbf{P}_{\mathbf{a}}^{\perp}\mathbf{d}}{\mathbf{t}^{\mathrm{H}}\mathbf{t}}), \tag{18}$$

where the approximation is good if $\mathbf{d}^{\mathrm{H}}\mathbf{P_a}\mathbf{d} \ll \mathbf{t}^{\mathrm{H}}\mathbf{t}$. Let $\boldsymbol{\Delta} := \frac{\mathrm{E}(\mathbf{d}\mathbf{d}^{\mathrm{H}})}{\mathbf{t}^{\mathrm{H}}\mathbf{t}}$ be a normalized (scale-invariant) definition of the covariance of $\hat{\mathbf{t}}$. Then in the above approximation

$$\mathrm{E}\{\mathrm{INSR}(\hat{\mathbf{t}})\} = \frac{1 - \mathbf{a}^{\mathrm{H}}\mathbf{a}}{\mathbf{a}^{\mathrm{H}}\mathbf{a}} + \frac{\mathrm{tr}(\mathbf{P}_{\mathbf{a}}^{\perp}\boldsymbol{\Delta})}{\mathbf{a}^{\mathrm{H}}\mathbf{a}}. \tag{19}$$

The first term represents the asymptotic performance of the Wiener beamformer ($\hat{\mathbf{t}} = \mathbf{a}$ with $\boldsymbol{\Delta} = 0$). The second term is the excess INSR due to the deviation of $\hat{\mathbf{t}}$ from the optimum. We can simply plug in the estimates of $\mathbf{R}_{\mathbf{t}_j}$ from equation (17) in place of $\boldsymbol{\Delta}$ to obtain the INSR corresponding to the ACMA beamformers.

For comparison, we consider the Wiener beamformer estimated from finite samples and known $\mathbf{S}$, or $\hat{\mathbf{T}}_W = (\mathbf{X}\mathbf{X}^{\mathrm{H}})^{-1}\mathbf{X}\mathbf{S}^{\mathrm{H}}$. Let $\hat{\mathbf{t}}_W$ be one of the columns of $\hat{\mathbf{T}}_W$, and $\mathbf{a}$ the corresponding column of $\mathbf{A}$. The normalized covariance of $\hat{\mathbf{t}}_W$ is derived as

$$\boldsymbol{\Delta}_W = \frac{\mathrm{cov}(\hat{\mathbf{t}}_W - \mathbf{a})}{\mathbf{a}^{\mathrm{H}}\mathbf{a}} = \frac{1}{N}\frac{1 - \mathbf{a}^{\mathrm{H}}\mathbf{a}}{\mathbf{a}^{\mathrm{H}}\mathbf{a}}\mathbf{I} + \mathcal{O}(\frac{1}{N^2}),$$

so that for the expected INSR of the finite-sample Wiener we find in first order approximation

$$\mathrm{E}\{\mathrm{INSR}(\hat{\mathbf{t}}_W)\} = \frac{1 - \mathbf{a}^{\mathrm{H}}\mathbf{a}}{\mathbf{a}^{\mathrm{H}}\mathbf{a}} + \frac{d-1}{N} \cdot \frac{1 - \mathbf{a}^{\mathrm{H}}\mathbf{a}}{(\mathbf{a}^{\mathrm{H}}\mathbf{a})^2}. \tag{20}$$

7. SIMULATIONS

Figure 1 shows performance plots of the first source for a simulation with $d = 3$ sources, $M = 3$ antennas in a uniform linear array, source powers $\mathbf{B} = \mathrm{diag}(1, 1.2, 0.9)$, and source angles $\boldsymbol{\alpha} = [0, \alpha, -\alpha]$, for varying N and SNR. The figure shows the excess INSR relative to the INSR of the asymptotic Wiener beamformer,

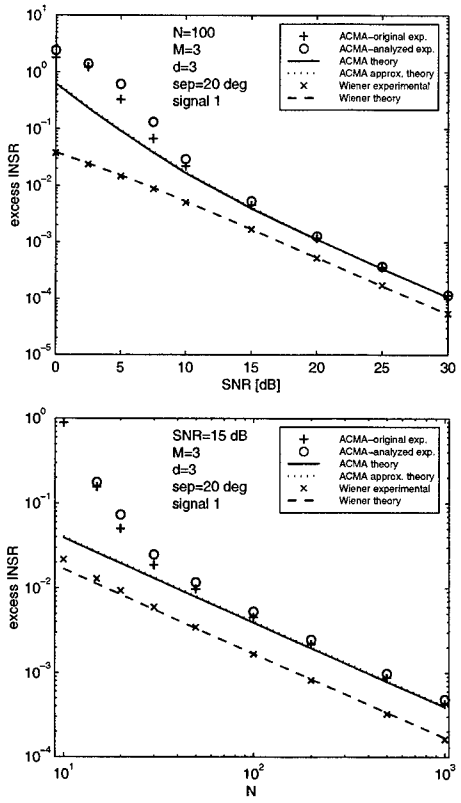


Figure 1. Finite sample INSR in excess of the asymptotic INSR of the Wiener beamformer.

evaluated for source 1 (i.e. the second terms in (19) and (20)). The experimental results show with '+' the outcome of the original ACMA algorithm of [1], and with '∘' the algorithm as analyzed here, i.e., with prewhitening based on the true covariance matrix $\mathbf{R_x}$, and using Gauss-Newton optimization to solve the subspace fitting step. The dotted line is the approximation resulting from (9), which is indeed very good. As is seen from the figures, the theoretical curves are a good prediction of the actual performance once $N > 30$, SNR > 5 dB. The small difference in performance between the original algorithm and the analyzed algorithm is caused by the different prewhitening. Not shown in the figures are the results for *weighted* subspace fitting: these turned out to be virtually identical to the unweighted results.

References

[1] A.J. van der Veen and A. Paulraj, "An analytical constant modulus algorithm," *IEEE Trans. Signal Processing*, vol. 44, pp. 1136–1155, May 1996.

[2] A.J. van der Veen, "Asymptotic properties of the Algebraic Constant Modulus Algorithm," IEEE Trans. Signal Processing, vol. 49, Aug. 2001.

[3] M. Viberg and B. Ottersten, "Sensor array processing based on subspace fitting," *IEEE Trans. Signal Proc.*, vol. 39, pp. 1110–1121, May 1991.

[4] N. Yuen and B. Friedlander, "Asymptotic performance analysis of ESPRIT, Higher-Order ESPRIT, and Virtual ESPRIT algorithms," *IEEE Trans. Signal Processing*, vol. 44, pp. 2537–2550, Oct. 1996.

FAST-CONVERGENCE ALGORITHM FOR ICA-BASED BLIND SOURCE SEPARATION USING ARRAY SIGNAL PROCESSING

Hiroshi SARUWATARI, Toshiya KAWAMURA, and Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN
Phone: +81-743-72-5281, Fax: +81-743-72-5289
E-mail: sawatari@is.aist-nara.ac.jp

ABSTRACT

We propose a new algorithm for blind source separation (BSS), in which independent component analysis (ICA) and beamforming are combined to resolve the low-convergence problem through optimization in ICA. The proposed method consists of the following two parts: frequency-domain ICA with direction-of-arrival (DOA) estimation, and null beamforming based on the estimated DOA. The alternation of learning between ICA and beamforming can realize fast- and high-convergence optimization. The results of the signal separation experiments reveal that the signal separation performance of the proposed algorithm is superior to that of the conventional ICA-based BSS method.

1. INTRODUCTION

Blind source separation (BSS) is the approach taken to estimate original source signals using only the information of the mixed signals observed in each input channel. This technique is applicable to the realization of noise-robust speech recognition and high-quality hands-free telecommunication systems. In the recent works for the BSS based on the independent component analysis (ICA) [1, 2], several methods, in which the inverse of the complex mixing matrices are calculated in the frequency domain, have been proposed to deal with the arrival lags among each of the elements of the microphone array system [3, 4, 5]. However, this ICA-based approach has the disadvantage that there is difficulty with the low convergence of nonlinear optimization [6].

In this paper, we describe a new algorithm for BSS in which ICA and beamforming are combined. The proposed method consists of the following two parts: (1) frequency-domain ICA with estimation of the direction of arrival (DOA) of the sound source, and (2) null beamforming based on the estimated DOA. The alternation of learning between ICA and null beamforming can realize fast- and high-convergence optimization. The following sections describe the proposed method in detail, and it is shown that the signal separation performance of the proposed algorithm is superior to that of the conventional ICA-based BSS method.

2. DATA MODEL AND CONVENTIONAL BSS METHOD

In this study, a straight-line array is assumed. The coordinates of the elements are designated as d_k ($k = 1, \cdots, K$), and the directions of arrival of multiple sound sources are designated as θ_l ($l = 1, \cdots, L$) (see Fig. 1), where we deal with the case of $K = L = 2$.

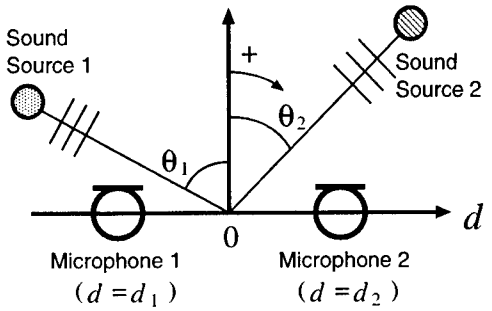


Figure 1: Configuration of a microphone array and signals.

In general, the observed signals in which multiple source signals are mixed linearly are given by the following equation in the frequency domain:

$$\boldsymbol{X}(f) = \boldsymbol{A}(f)\boldsymbol{S}(f), \tag{1}$$

where $\boldsymbol{X}(f)$ is the observed signal vector, $\boldsymbol{S}(f)$ is the source signal vector, and $\boldsymbol{A}(f)$ is the mixing matrix; these are given as

$$\boldsymbol{X}(f) = [X_1(f), \cdots, X_K(f)]^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{S}(f) = [S_1(f), \cdots, S_L(f)]^{\mathrm{T}}, \tag{3}$$

$$\boldsymbol{A}(f) = \begin{bmatrix} A_{11}(f) & \cdots & A_{1L}(f) \\ \vdots & & \vdots \\ A_{K1}(f) & \cdots & A_{KL}(f) \end{bmatrix}. \tag{4}$$

$\boldsymbol{A}(f)$ is the mixing matrix which is assumed to be complex-valued because we introduce a model to deal with the arrival lags among each of the elements of the microphone array and room reverberations.

In the frequency-domain ICA, first, the short-time analysis of observed signals is conducted by frame-by-frame discrete Fourier transform (DFT). By plotting the spectral values in a frequency bin of each microphone input frame by frame, we consider them as s time series. Hereafter, we designate the time series as

$$\boldsymbol{X}(f,t) = [X_1(f,t), \cdots, X_K(f,t)]^{\mathrm{T}}. \tag{5}$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

Next, we perform signal separation using the complex-valued inverse of the mixing matrix, $\boldsymbol{W}(f)$, so that the L time-series output $\boldsymbol{Y}(f,t)$ becomes mutually independent; this procedure can be given as

$$\boldsymbol{Y}(f,t) = \boldsymbol{W}(f)\boldsymbol{X}(f,t), \tag{6}$$

where

$$\boldsymbol{Y}(f,t) = [Y_1(f,t),\cdots,Y_L(f,t)]^{\mathrm{T}}, \tag{7}$$

$$\boldsymbol{W}(f) = \begin{bmatrix} W_{11}(f) & \cdots & W_{1K}(f) \\ \vdots & & \vdots \\ W_{L1}(f) & \cdots & W_{LK}(f) \end{bmatrix}. \tag{8}$$

We perform this procedure with respect to all frequency bins. Finally, by applying the inverse DFT and the overlap-add technique to the separated time series $\boldsymbol{Y}(f,t)$, we reconstruct the resultant source signals in the time domain.

In the conventional ICA-based BSS method, the optimal $\boldsymbol{W}(f)$ is obtained by the following iterative equation [3, 7]:

$$\boldsymbol{W}_{i+1}(f) = \eta\Big[\mathrm{diag}\Big(\big\langle\boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\big\rangle_t\Big) - \big\langle\boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\big\rangle_t\Big]\boldsymbol{W}_i(f)+\boldsymbol{W}_i(f), \tag{9}$$

where $\langle\cdot\rangle_t$ denotes the time-averaging operator, i is used to express the value of the i th step in the iterations, and η is the step-size parameter. Also, we define the nonlinear vector function $\boldsymbol{\Phi}(\cdot)$ as

$$\boldsymbol{\Phi}(\boldsymbol{Y}(f,t)) \equiv [\Phi(Y_1(f,t)),\cdots,\Phi(Y_L(f,t))]^{\mathrm{T}}, \tag{10}$$

$$\Phi(Y_l(f,t)) \equiv \big[1+\exp(-Y_l^{(\mathrm{R})}(f,t))\big]^{-1} + j\cdot\big[1+\exp(-Y_l^{(\mathrm{I})}(f,t))\big]^{-1}, \tag{11}$$

where $Y_l^{(\mathrm{R})}(f,t)$ and $Y_l^{(\mathrm{I})}(f,t)$ are the real and imaginary parts of $Y_l(f,t)$, respectively.

3. PROPOSED ALGORITHM

The conventional ICA method inherently has a significant disadvantage which is due to low convergence through nonlinear optimization in ICA. In order to resolve the problem, we propose an algorithm based on the alternation of learning between ICA and beamforming; the inverse of the mixing matrix, $\boldsymbol{W}(f)$, obtained through ICA is periodically substituted by the matrix based on null beamforming for a temporal initialization. The proposed algorithm is conducted by the following steps with respect to all frequency bins in parallel (see Fig. 2).

[Step 1: Initialization] Set the initial $\boldsymbol{W}_{jP+i}(f)$, i.e., $\boldsymbol{W}_0(f)$, to an arbitrary value, where the subscripts i and j are set to be 0.

[Step 2: P-time ICA iteration] Optimize $\boldsymbol{W}_{jP+i}(f)$ using the following P-time ICA iteration:

$$\boldsymbol{W}_{jP+i+1}(f) = \eta\Big[\mathrm{diag}\Big(\big\langle\boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\big\rangle_t\Big) - \big\langle\boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\big\rangle_t\Big]\boldsymbol{W}_{jP+i}(f) + \boldsymbol{W}_{jP+i}(f), \tag{12}$$

where $i\ (=0,\cdots,P-1)$ is increased by one every iteration.

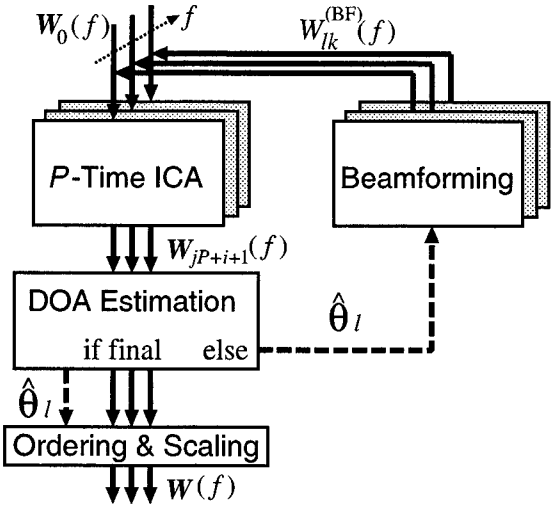


Figure 2: Proposed algorithm combining frequency-domain ICA and beamforming.

[Step 3: DOA estimation] Estimate DOAs of the sound sources by utilizing the directivity pattern of the array system, $F_l(f,\theta)$, which is given by

$$F_l(f,\theta) = \sum_{k=1}^{K} W_{lk}(f)\ \exp\left[j2\pi f d_k \sin\theta/c\right], \tag{13}$$

where $W_{lk}(f)$ is the element of $\boldsymbol{W}_{jP+i+1}(f)$, and c is the velocity of sound. In the directivity patterns, directional nulls exist in only two particular directions. Accordingly, by obtaining statistics with respect to the directions of nulls at all frequency bins, we can estimate the DOAs of the sound sources. The DOA of the l th sound source, $\hat{\theta}_l$, can be estimated as

$$\hat{\theta}_l = \frac{2}{N}\sum_{m=1}^{N/2}\theta_l(f_m), \tag{14}$$

where N is a total point of DFT, and $\theta_l(f_m)$ represents the DOA of the l th sound source at the m th frequency bin. These are given by

$$\theta_1(f_m) = \min\Big[\operatorname*{argmin}_{\theta}|F_1(f_m,\theta)|, \operatorname*{argmin}_{\theta}|F_2(f_m,\theta)|\Big], \tag{15}$$

$$\theta_2(f_m) = \max\Big[\operatorname*{argmin}_{\theta}|F_1(f_m,\theta)|, \operatorname*{argmin}_{\theta}|F_2(f_m,\theta)|\Big], \tag{16}$$

where $\min[x,y]$ ($\max[x,y]$) is defined as a function in order to obtain the smaller (larger) value among x and y.

[Step 4] If the $(jP+i+1)$th iteration was the final iteration, go to **step 6**; otherwise go to **step 5** with an increment of j.

[Step 5: Beamforming] Construct an alternative matrix for signal separation based on the null-beamforming technique where the DOA information obtained in the ICA section is used. In the case that the look direction is $\hat{\theta}_1$ and the directional null is steered to $\hat{\theta}_2$ (see solid line in Fig. 3), the elements of the matrix for signal separation are given as

$$W_{11}^{(\mathrm{BF})}(f_m) = \exp\left[-j2\pi f_m d_1 \sin\hat{\theta}_1/c\right]$$

$$\times \left\{\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c\right] - \exp\left[j2\pi f_m d_2(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c\right]\right\}^{-1}, \quad (17)$$

$$W_{12}^{(\mathrm{BF})}(f_m) = -\exp\left[-j2\pi f_m d_2 \sin\hat{\theta}_1/c\right] \times \left\{\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c\right] - \exp\left[j2\pi f_m d_2(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c\right]\right\}^{-1}. \quad (18)$$

Also, in the case that the look direction is $\hat{\theta}_2$ and the directional null is steered to $\hat{\theta}_1$ (see broken line in Fig. 3), the elements of the matrix are given as

$$W_{21}^{(\mathrm{BF})}(f_m) = -\exp\left[-j2\pi f_m d_1 \sin\hat{\theta}_2/c\right] \times \left\{-\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c\right] + \exp\left[j2\pi f_m d_2(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c\right]\right\}^{-1}, \quad (19)$$

$$W_{22}^{(\mathrm{BF})}(f_m) = \exp\left[-j2\pi f_m d_2 \sin\hat{\theta}_2/c\right] \times \left\{-\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c\right] + \exp\left[j2\pi f_m d_2(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c\right]\right\}^{-1}. \quad (20)$$

The elements given by Eqs. (17)–(20) are inserted into $\boldsymbol{W}_{jP}(f)$, where the subscript i is reset to be 0. Then we go back to **step 2** and repeat the ICA iteration using the $\boldsymbol{W}_{jP}(f)$ as an initial value. **[Step 6: Ordering and scaling]** Using the DOA information obtained in **step 3**, we detect and correct the source permutation and the gain inconsistency [8]. By applying the above-mentioned modifications, we can finally obtain the optimal $\boldsymbol{W}(f)$ as follows:

$$\boldsymbol{W}(f) = \begin{cases} \begin{bmatrix} 1/F_1(f,\hat{\theta}_1) & 0 \\ 0 & 1/F_2(f,\hat{\theta}_2) \end{bmatrix} \cdot \boldsymbol{W}_{jP+i+1}(f), \\ \text{(without permutation)} \\ \begin{bmatrix} 0 & 1/F_2(f,\hat{\theta}_1) \\ 1/F_1(f,\hat{\theta}_2) & 0 \end{bmatrix} \cdot \boldsymbol{W}_{jP+i+1}(f), \\ \text{(with permutation).} \end{cases} \quad (21)$$

4. EXPERIMENTS AND RESULTS

4.1. Conditions for Experiments

A two-element array with the interelement spacing of 4 cm is assumed. The speech signals are assumed to arrive from two directions, $-30°$ and $40°$. Two kinds of sentences, those spoken by two male and two female speakers selected from the ASJ continuous speech corpus for research, are used as the original speech samples. Using these sentences, we obtain 12 combinations with respect to speakers and source directions. In these experiments, we use the following signals as the source signals: the original speech convolved with the impulse responses specified by different reverberation times (RTs) of 0 msec, 150 msec and 300 msec. The impulse responses are recorded in a variable reverberation time room as shown in Fig. 4. The analytical conditions of these experiments are as follows: the sampling frequency is 8 kHz, the frame length is 32 msec, the frame shift is 16 msec, the window function is a Hamming window, the parameter P is set to be 100, and the step-size parameter η for iterations is set to be 1.0×10^{-5}.

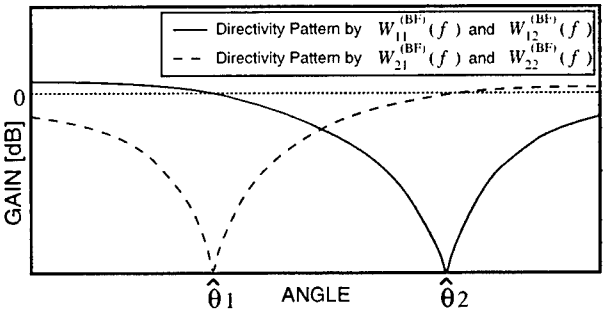


Figure 3: Example of directivity patterns in beamforming.

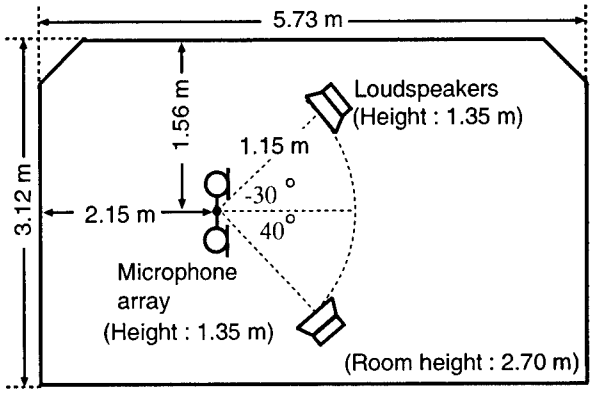


Figure 4: Layout of reverberant room used in experiments.

4.2. Objective Evaluation of Separated Signal

In order to compare the performance of the proposed algorithm with that of the conventional BSS described in Sect. 2 for different iteration points in ICA, the *noise reduction rate* (NRR), defined as the output signal-to-noise ratio (SNR) in dB minus input SNR in dB, is shown in Figs. 5(a)–(c). These values were averages of all of the combinations with respect to speakers and source directions.

In Fig. 5(a), for the nonreverberant test, it is evident that the separation performance of the proposed algorithm is superior to that of the conventional ICA-based BSS method at every iteration after 100 iterations. For example, the proposed method can improve the NRR of about 6.4 dB at the 200-iteration point. As for the results of DOA estimation, Fig. 6 shows the average and deviation of the estimated DOA at each frequency corresponding to $-30°$. As shown in Fig. 6, the proposed algorithm can update $\boldsymbol{W}(f)$ properly with a more accurate estimation of DOA compared with the conventional method (the same tendency was shown at $40°$). This contributes to the realization of fast and high convergence through the optimization of $\boldsymbol{W}(f)$ in the proposed algorithm under the nonreverberant condition.

As shown in Figs. 5(b) and (c), by the reverberant tests, it is shown that the performance of the proposed algorithm is superior to those of the conventional ICA-based BSS method at every iteration after 100 iterations. For example, the proposed method can improve the NRRs of about 2.4 dB (RT=150 msec) and 0.7 dB (RT=300 msec) at the 200-iteration point. Although null beamforming is not suitable for signal separation under the condition that the direct sounds and their reflections exist, we can confirm that the utilization of null beamforming for temporal initialization

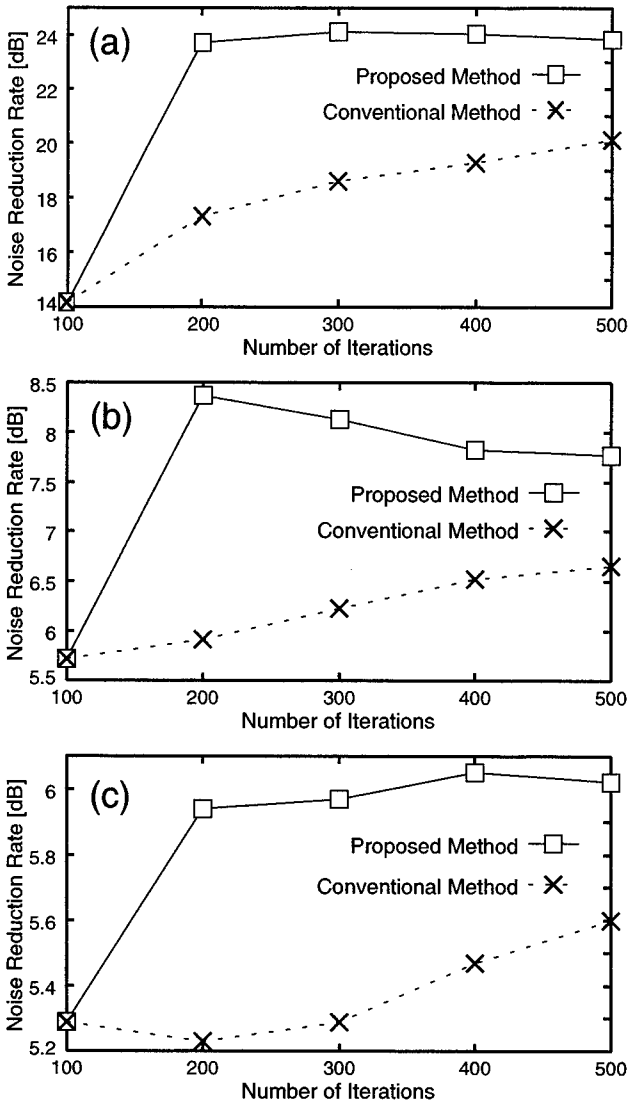


Figure 5: Noise reduction rates for different iterations in ICA in the case that the RT is (a) 0 msec, (b) 150 msec, and (c) 300 msec.

through ICA iterations is effective for improving the separation performance, even under reverberant conditions.

5. CONCLUSION

In this paper, we described a fast- and high-convergence algorithm for BSS where null beamforming is used for temporal initialization through ICA iterations. The results of the signal separation experiments reveal that the signal separation performance of the proposed algorithm is superior to that of the conventional ICA-based BSS method, and the utilization of null beamforming in ICA is effective for improving the separation performance and convergence, even under reverberant conditions. In future, further investigations regarding the adjustment of the periodical-alternation parameter , e.g., P, will be required, and we will apply the proposed method to a noise-robust speech recognition system.

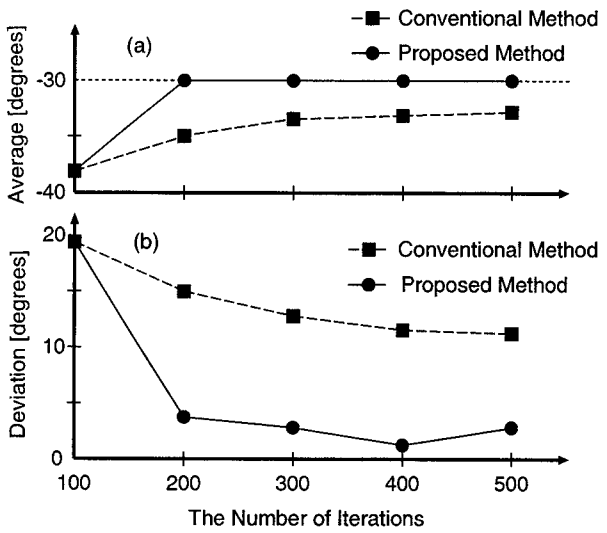


Figure 6: Average and deviation of estimated DOA at each frequency corresponding to $-30°$ under the nonreverberant condition.

6. ACKNOWLEDGEMENT

This work was partly supported by CREST (Core Research for Evolutional Science and Technology) in Japan.

7. REFERENCES

[1] P. Common, "Independent component analysis, a new concept?," *Signal Processing*, vol.36, pp.287–314, 1994.

[2] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol.7, pp.1129–1159, 1995.

[3] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *Proceedings of 1998 International Symposium on Nonlinear Theory and Its Application (NOLTA '98)*, vol.3, pp.923–926, Sep. 1998.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol.22, pp.21–34, 1998.

[5] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech & Audio Process.*, vol.8, pp.320–327, 2000.

[6] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, and K. Shikano, "Blind source separation based on subband ICA and beamforming," *Proc. ICSLP2000*, vol.3, pp.94–97, Oct. 2000.

[7] A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. Circuits and Systems I*, vol.43, no.11, pp.894–906, 1996.

[8] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *Proc. ICASSP2000*, vol.5, pp.3140–3143, June 2000.

RECOGNITION OF FACIAL IMAGES USING SUPPORT VECTOR MACHINES

K. I. Kim, J. Kim,

A. I. Lab, CS. Dept. Korea Advanced Institute of Science and Technology, Taejon, 305-701, Korea

K. Jung

School of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, 440-746, Korea

ABSTRACT

A novel support vector machine (SVM)-based method for appearance-based face recognition is presented. The proposed method does not use any external feature extraction process. Accordingly the intensities of the raw pixels that make up the face pattern are fed directly to the SVM. However, it takes account of prior knowledge about facial structures in the form of a kernel embedded in the SVM architecture. The new kernel efficiently explores spatial relationships among potential eye, nose, and mouth objects and is compared with existing kernels. Experiments with ORL database show a recognition rate of 98% and speed of 0.22 seconds per face with 40 classes.

1. INTRODUCTION

It is reported that sales of identity verification products exceed $100 million [1]. Accordingly, many methods have been developed for easy and reliable identification. Among them, face recognition has the benefit of being a passive, nonintrusive system for verifying personal identity. This paper presents a face recognition method designed for the use of applications such as security monitoring and location tracking. In these applications, multiple images per person are often available for training and real-time recognition is required [2]. To allow the system being real time, the proposed method excludes any of time-consuming feature extraction or pre-processing stage. Instead the gray values of raw pixels that make up the face pattern are directly feed to recognizer. In order to absorb the resulting high-dimensionality of input space, support vector machines (SVMs), which are known to work well even in high-dimensional space, are used as face recognizer.

This idea is somewhat similar to recent applications of SVMs [3][4]. However, the method proposed here differs in that it takes account of prior knowledge about facial structures and uses this in the form of a kernel (called a local correlation kernel) that is embedded in the SVM architecture. A brief introduction to SVMs and the use of prior knowledge for face recognition are given in Section 2. Section 3 presents the performance results of the proposed method when using the ORL database [5]. It was found that the proposed method correctly recognized 98.0% of the face patterns with a speed of 0.22 seconds per face with 40 classes. The conclusions and directions for future research are given in Section 4.

2. SUPPORT VECTOR MACHINES FOR FACE RECOGNITION

A SVM constructs a binary classifier from a set of patterns called training examples, which are available prior to classification. Let $(\mathbf{x}_i, y_i) \in \mathbf{R}^N \times \{\pm 1\}$, $i = 1,\ldots,l$ be such a set of training examples. The classifer constructs a linear decision surface (hyperplane) of the form:

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l^*} y_i \alpha_i \mathbf{x}_i^* \cdot \mathbf{x} + b\right). \quad (1)$$

where $\{\mathbf{x}_i^*\}_{i=1}^{l^*}$ is a subset of the training data set. These are called *support vectors* (SVs) and are the points from the data set that fall closest to the separating hyperplane. The coefficients α_i and b are determined by solving the large-scale quadratic programming problem [6]. This hyperplane is known to minimize the bound on its VC-dimension and accordingly, has shown to provide high generalization performance even in high-dimensional spaces [6]. However, since it is unlikely that a general pattern classification problem can actually be solved by a linear classifier, the SVM needs to be augmented in order to allow for non-linear decision surfaces. The basic idea is to map the data into another dot product space (called the *feature space*) F via a nonlinear map

$$\Phi : \mathbf{R}^N \to F. \quad (2)$$

and perform the above linear algorithm in F.
Since the solution has the form

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l^*} y_i \alpha_i \Phi(\mathbf{x})_i^* \cdot \Phi(\mathbf{x}) + b\right). \quad (3)$$

it is nonlinear in the original input variables.

In SVMs, the mapping Φ is usually performed by the kernel function as defined by:

$$k(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}). \quad (4)$$

Then, by selecting the proper kernels k, various mappings (or feature extractions) Φ can be indirectly induced [6]. One of these mappings can be achieved by taking the p-order correlations between the entries, x_i, of the input vector $\mathbf{x}$. It should be noted that these features cannot be extracted by simply computing all the correlations, since the required computation is prohibitive when p is not small (p>2): for N-dimensional input patterns, the dimensionality of the feature space F is $(N+p-1)!/p!(N-1)!$. However, this is facilitated by the

0-7803-7011-2/01/$10.00 ©2001 IEEE

introduction of a polynomial kernel, as a polynomial kernel with degree p ($k(\mathbf{x},\mathbf{y})=(\mathbf{x}\cdot\mathbf{y})^p$) corresponds to the dot product of two monomial mappings, Φ_p [6]:

$$\begin{aligned}(\Phi_p(\mathbf{x})\cdot\Phi_p(\mathbf{y}))&=\sum_{i_1,\ldots,i_d=1}^{N}x_{i_1}\cdot\ldots\cdot x_{i_d}\cdot y_{i_1}\cdot\ldots\cdot y_{i_d}\\&=\left(\sum_{i=1}^{N}x_i\cdot y_i\right)^p=(\mathbf{x}\cdot\mathbf{y})^p\end{aligned}\tag{5}$$

When $\mathbf{x}$ represents an image pattern, the use of this kernel allows all possible correlations of p pixels in the image to be taken into account.

From the feature extraction viewpoint, however, the mapping Φ_p induced by a polynomial kernel has an important shortcoming—it does not utilize any prior knowledge while it gets to be common to use it for improving the system performance [7][8]. With this observation, it is reasonable to expect that polynomial kernel can be improved by incorporating available prior knowledge. The following set of intuitive knowledge is considered: 1. It is usually the case that images have a local structure in that not all the correlations between image regions carry equal amounts of information [8]. 2. The human face is a complex and meaningful pattern that contains most of its information in its structure. A human face is therefore expressed as a composition of its components (or objects) such as eyes, nose, mouth, etc. and can be well represented by exhibiting the features of such individual objects and the context between them [7].

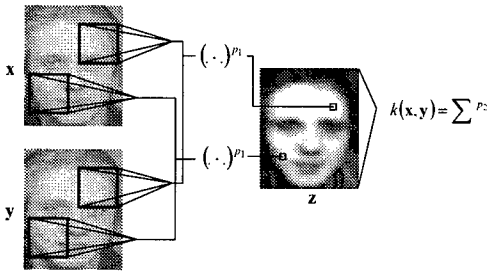


Figure 1. Architecture of local correlation kernel.

The local correlation kernel presented by Schölkopf, et al. gives a way of utilizing the first knowledge [8]. The basic idea is that the local correlations between each adjacent pixel are computed first, and then the long-range correlations are only computed based on the local correlations. The resulting kernel corresponds to a dot product in a polynomial space spanned mainly by localized correlations between pixels. Fig. 1 shows the architecture of a kernel utilizing local correlations in face images. To compute $k(\mathbf{x},\mathbf{y})$ for two patterns $\mathbf{x}$ and $\mathbf{y}$, the products between the corresponding pixels of the localized regions in the two images are summed (indicated by dot products $(\cdot\cdot)$), as weighed by the pyramidal receptive fields. The first nonlinearity, in the form of the exponent p_1, is then applied to the output. The resulting values are summed, and the p_2 -th power of the result is taken as the value $k(\mathbf{x},\mathbf{y})$. The resulting kernel will be of the order up to $p_1\cdot p_2$, however, this does not contain all the possible pixel correlations but mainly just the local ones. In the rest of this paper, we call this kernel as pure local correlation kernel in order to distinguish this from the new local correlation kernel that will be described later.

The second knowledge is the basis of feature-based methods. While this knowledge has been effectively adopted in feature-based methods [2][7], it was not well established in appearance-based methods. The basic idea of pure local correlational kernel can be extended to accommodate this: The knowledge suggests that a face image should be characterized using a two-level hierarchy of within and between the object features. In the case of correlations as the feature, the hierarchy is realized based on correlations between object features, which are defined as the correlations between the pixels constituting these objects. With a local correlational kernel, this is achieved by simply removing the pixel-level inter-object correlations (for example, the correlations between two pixels, which are located in the left eye and mouth, respectively) from all the possible correlations.

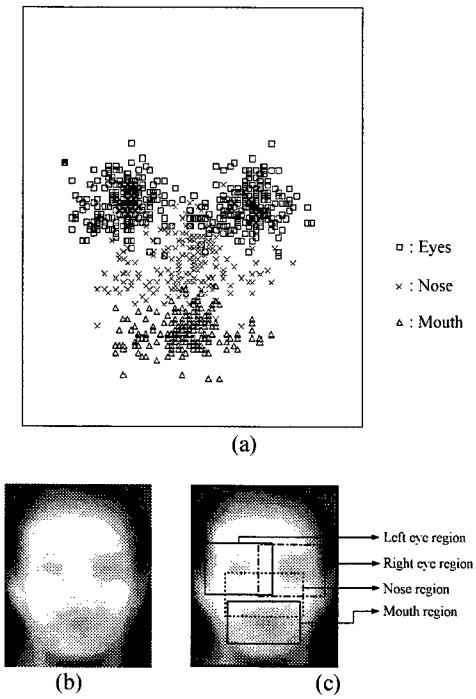


Figure 2. Object region configuration: (a) coordinates of objects, (b) average face image, and (c) object configuration overlaid in (b).

As mentioned in Section 1, the problem with this approach is that this requires the facial image to be analyzed into individual facial objects, while object location task itself is not trivial and usually requires dense computation. Accordingly, an alternative is adopted instead of directly utilizing this approach. When the target is a single frontal face image (with rather controlled zoom and pose), rough locations of some objects, as a priori

information. are available even without structural analysis. Fig. 2a shows the coordinates of objects in consideration (eyes. nose. and mouth) obtained from 200 facial images of the ORL database [5]. It can be observed that the locations of these objects do not significantly intercept with each other nor very significantly. and accordingly can be estimated in the rough. This is supported by the fact that the average of 200 frontal images still retains the shape of human face (Fig. 2b) (eyes and mouth regions are observed). Furthermore. from this. we can decompose a face image into a set of overlapping regions. which probably contain only one object. respectively (Fig. 2c). Then. the following strategy is adopted to improve the correlation kernel:

Restrict inter-region correlations. which may be pixel-level inter-object correlations, while retaining intra region correlations or hopefully intra-object correlations.

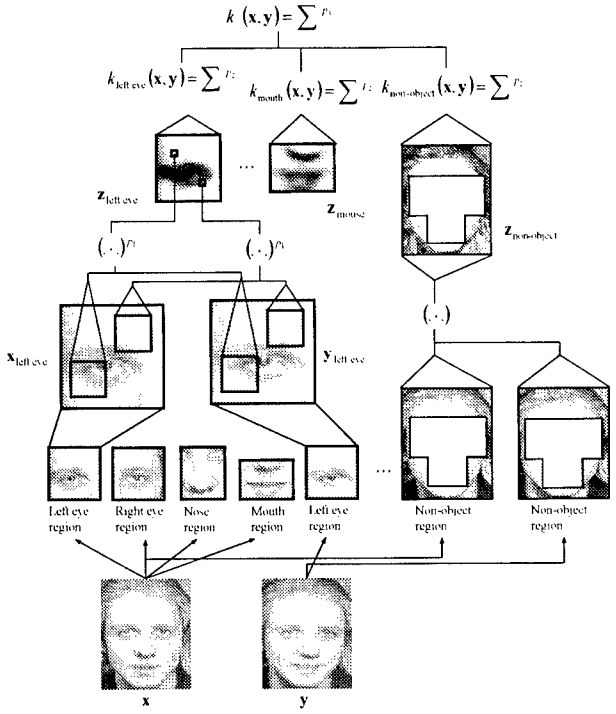


Figure 3. Architecture of modified local correlation kernel for facial feature extraction.

This strategy is implemented in a modified form of local correlational kernel in Fig. 3. To compute $k(\mathbf{x},\mathbf{y})$ for two facial images $\mathbf{x}$ and $\mathbf{y}$, it first decomposes them into a set of object regions and computes local correlation kernels of order $p_1 \cdot p_2$ for each region. The resulting kernel $\mathbf{z}_{\text{object}}$ then represents localized characteristics of object regions. Then the global correlation is computed from only the summed products (of order p_3) of these kernel outputs and non-object regions. Non-object region is included to take into account information contained in objects with irregular shape (such as hair). The resulting kernel will be of the order $p_1 \cdot p_2 \cdot p_3$ polynomial kernel which differes from a standard polynomial in that it does not utilize all products of $p_1 \cdot p_2 \cdot p_3$ pixels. but mainly inter object ones.

Since SVMs were originally developed for two-class classification. their basic scheme for multi-face recognition is extended by adopting a *one-against-others* decomposition method. In this strategy. R different SVMs are constructed. one for each class. Here the r-th SVM ω_r is trained on the whole training data set in order to classify the members of class r against the rest. Then. in the recognition phase. the index of SVM with largest output for a given pattern is regarded as the recognition result.

3. EXPERIMENTAL RESULTS

The system has been tested with ORL face database [5]. This set of faces includes ten different images of 40 distinct subjects. The images are grayscale with a resolution of 92×112. For the training and testing of the recognizer. the grayscale was linearly normalized to lie within [-1. 1]. All experiments were performed using 5 training images and 5 test images per person for a total of 200 training images and 200 test images. There was no overlap between the training and test sets. Since the recognition performance will be affected by the selection of training images. the reported results were obtained by training 20 recognizers[1] for each dichotomy with different training examples (random selection of 5 images from 10 per each subject. resulting in 5 positive and 35 negative for each SVM) and selecting the average error over all the results. The system was implemented using Visual C++ language on a Pentium III compatible CPU. The average recognition time was 0.22 seconds for a face pattern with 40 classes. This speed is sufficient for tasks such as security monitoring and location tracking.

Table 1. Error rates with different kernel degrees.

p_3	p_1	p_2			
		1	2	3	4
1	1	4.0	3.2	2.3	2.5
	2	2.3	3.0	2.1	3.1
	3	2.2	2.0	2.5	3.2
	4	2.6	2.4	3.2	4.7
2	1	2.1	2.0	2.8	2.9
	2	2.5	2.1	2.2	4.2
	3	2.7	2.3	3.8	4.7
	4	4.1	2.7	6.3	8.4

Table 1 shows the error rates with different kernel degrees p_1. p_2 . and p_3 . The best performances was obtained with $(p_1 = 3, p_2 = 2, p_3 = 1)$ and $(p_1 = 1, p_2 = 2, p_3 = 2)$ (shaded entries in table) which yields degree $6(= 3 \cdot 2 \cdot 1)$ and $4(= 1 \cdot 2 \cdot 2)$ correlations.

[1] Out of a total of 10!/5! = 30240 combinations.

To gain a better understanding of the relevance of the results obtained using local correlation kernels, benchmark comparisons with other kernels were carried out. A set of experiments was performed using SVMs with different kernels. Table 2 summarizes the type of kernels and their parameter settings used in the experiments. These parameters were set empirically i.e. those parameters which yielded the best performances from several experiments. Table 3 shows the recognition results. For comparison, the result obtained from the local correlation kernel is also presented. It should be noted that linear SVMs ranked as the third to the pure and proposed local correlation kernels. This is because the problem was linearly separable as the face space was high-dimensional (92×112) and very sparse (with few training and testing examples). The zero training error for the linear SVMs supported this observation. In this case, making the classification space larger than the input space is not preferable as in other possible linear non-separable applications [4]. In contrast, the superior performance of the local correlation kernel confirms the usefulness of prior knowledge for constructing the classification space and verifies its appropriateness for face recognition.

Table 2. Different kernels their parameter settings used in experiments

Kernels	Parameters
None (linear SVM)	
$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y})^p$	$p = 3$
$k(\mathbf{x},\mathbf{y}) = \exp(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|^2)$	$\sigma = 0.5$
$k(\mathbf{x},\mathbf{y}) = \tanh(\mathbf{x}\cdot\mathbf{y}-\Theta)$	$\Theta = 1.5$
Pure local correlation kernel	$(p_1 = 3, p_2 = 2)$

Table 3. Error rates of SVMs using different kernels.

Kernels	Error rates (%)
None (Linear SVM)	3.2
Polynomial	3.4
Gaussian	4.2
Tangent hyperbolic	5.3
Pure local correlation kernel	2.7
Local correlation kernel	2.0

Table 4. Error rates of various systems.

System	Error rates (%)
Eigenfaces [9]	10.0
Psudo-2DHMM [9]	5.0
Convolutional neural network [10]	3.8
Linear SVMs [3]	3.0

Table 4 shows a summary of the performance of various systems for which results using the ORL database are available [3][9][10]. The proposed method showed the best performance and significant reduction of error rate (33.3%) from the second best performing system–linear SVMs [3].

4. CONCLUSIONS

A novel SVM-based method is proposed for appearance-based face recognition. The proposed method takes account of prior knowledge about facial structures in the form of a kernel embedded in the SVM architecture. The new kernel explores spatial relationships among potential eye, nose, and mouth objects and showed better performance than other kernels.

The application domain of the proposed method is not limited in the problem of face recognition. It can also be applied to the problem of face authentification and face detection. By shifting the detection window to all locations within an image, a face detection problem can be reduced to a problem of binary classification (i.e. face class or background class). Accordingly, further experiments are required for these two-class face classification applications. The proposed method is insensitive to color, which is often present in single face images, although color is often unreliable because of the difficulty of accurate camera calibration. However, it would also be interesting to explore the utility of color information for face recognition.

5. REFERENCES

[1] B. Miller, "Vital signs of identity", *IEEE Spectrum*, pp. 22-30, Feb, 1994.

[2] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey", *Proc. IEEE*, vol. 83, pp. 705-740, 1995.

[3] G. D. Guo, S. Z. Li, and K. L. Chan, "Face Recognition by Support Vector Machines," in *Proc. International Conference on Automatic Face and Gesture Recognition*, pp. 196-201, 2000.

[4] K. I. Kim, K. Jung, S. H. Park, and H. J. Kim, "Supervised texture segmentation using support vector machines," *IEE Electronics Letters*, vol. 35, pp. 1935-1937, 1999.

[5] AT&T Laboratories, Cambridge, *http://www.cam-orl.co.uk/facedatabase.html*.

[6] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, New York, 1995.

[7] K. C. Yow and R. Cipolla, "Feature-based human face detection," *Image and Vision Computing*, vol. 15, pp. 713-735, 1997.

[8] B. Schölkopf, P. Simard, A. Smola, and V. Vapnik, "Prior knowledge in support vector kernels," *Advances in Neural information processing systems*, M. Jordan, M. Kearns, and S. Solla, Eds., vol. 10, MIT Press, Cambridge, MA, 1998, pp. 640-646.

[9] F. S. Samaria, *Face Recognition Using Hidden Markov Models*, Ph. D. dissertation, Univ. Cambridge, Cambridge, 1994.

[10] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," *IEEE Trans. on Neural Networks*, vol. 8, pp. 98-113, 1997.

HIGHER ORDER CONDITIONAL ENTROPY-CONSTRAINED TRELLIS-CODED RVQ WITH APPLICATION TO PYRAMID IMAGE CODING

Mohammad Asmat Ullah Khan

Department of Electrical Engineering
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia.
maukhan@kfupm.edu.sa

ABSTRACT

This paper introduces an extension of conditional entropy-constrained RVQ (CEC-RVQ) to include quantization cell shape gain. The method is referred to as conditional entropy-constrained trellis-coded RVQ (CEC-TCRVQ). The new design is based on coding image vectors by taking into account their 2-D correlation and employing a higher order entropy model with a trellis structure. We employed CEC-TCRVQ to code image subbands at low bit rate. The CEC-TCRVQ coded images do well in term of preserving low-magnitude textures present in some images

1. INTRODUCTION

For ergodic stationary sources Vector quantization (VQ) is optimal in a rate-distortion sense for a given vector size. However, it has not been successfully applied to image coding in spatial domain. One of the primary reason is the fact that due to high inter-pixel correlation of real world imagery, to get good performance, a fairly large vector sizes are needed. Since for VQ implementation, the codebook size, and hence the complexity, memory and the needed training data size, all grow exponentially with the vector size and the encoding rate, large vector sizes become prohibitive.

Relief can be obtained by employing a multi-stage VQ (MSVQ), also known as residual VQ (RVQ), for image coding purposes. Entropy-constrained residual vector quantization (EC-RVQ) [6], is a high-performance, computationally efficient implementation over conventional VQ for image coding. It was shown in [5]that improved rate-distortion performance of an EC-RVQ for image coding can be realized by exploiting adjacent vector dependencies. The improved image coding design is called conditional entropy-constrained RVQ (CEC-RVQ). The CEC-RVQ employed a higher-order conditional entropy model with multistage structure of RVQ, to achieve a reduction of as much as 40% for the same image quality as for EC-RVQ.

In order to incorporate quantization cell shape gain, a trellis-based coding was employed in CEC-RVQ design. The method was called conditional entropy-constrained trellis-coded RVQ (CEC-TCRVQ) [3].The approach taken in CEC-TCRVQ is to employ adjacent and stage-conditioning symbols and select conditioning symbols for higher-order model jointly over the long term.

The direct application of CEC-TCRVQ to image coding leads to a blocky appearance of the reconstructed image. This problem becomes more apparent at low bit rates. It was found that this problem does not occur when we code subbands. Another advantage in coding image subbands is that the vector dimensions need not be large. Further more coding of the various subbands can be done in parallel and thus is suitable for real-time implementation. Pyramid image coding is a form of subband coding and differs from conventional subband coding in that it involves an octave-step division of the frequency axes, whereas conventional subband coding splits the frequency axes uniformly. The motivation behind the research presented in this paper is to show the application of CEC-TCRVQ to pyramid image coding.

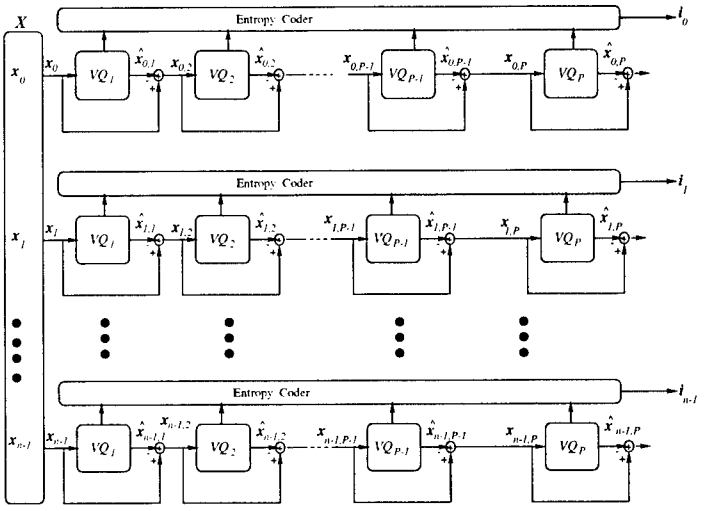


Figure 1: Quantizing supervector using residual vector quantization

The paper is organized as follows. Section 2 provides the review of conditional entropy-constrained RVQ. The CEC-RVQ is then extended to include trellis-coding in Section 3. Section 4 discusses subband quantization scheme employed. Bit allocation problem is described in Section 5. Simulation results and comparison with other subband coding techniques are presented in Section 6.

2. CONDITIONAL ENTROPY-CONSTRAINED RVQ

Let

$$\boldsymbol{X} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\}$$

be a supervector of n consecutive vectors. Each component of the supervector is quantized by a P-stage EC-RVQ encoder as shown

0-7803-7011-2/01/$10.00 ©2001 IEEE

in Figure 1. It is noteworthy that the cascade of stage VQs shown in the figure enacts a direct sum codebook. That is, the direct sum codebook associated with $\boldsymbol{X}$ is defined by summing all combinations of the stage codevectors in $VQ_1, VQ_2, \cdots, VQ_P$. After quantizing the supervector, the codebook indices are fed to an entropy coder which outputs a variable-length bit sequence for each input component of the supervector. The variable-length index sequence for the supervector is denoted as

$$\begin{aligned} \boldsymbol{I} &= \{i_0, i_1, \ldots, i_{n-1}\} \\ &= \{(i_{0,1}, i_{0,2}, \ldots, i_{0,P}), (i_{1,1}, i_{1,2}, \ldots, i_{1,P}), \ldots, \\ &\quad (i_{n-1,1}, i_{n-1,2}, \ldots, i_{n-1,P})\} \end{aligned} \quad (1)$$

Let $P(\boldsymbol{C})$ be the probability that a supervector $\boldsymbol{X}$ is represented by the codevector sequence $\boldsymbol{C} = \{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_{n-1}\}$ according to index sequence $\boldsymbol{I}$. Here each component $\boldsymbol{c}_j$ represents a direct sum codevector. The distortion associated with the supervector is given by

$$d(\boldsymbol{X}, \boldsymbol{C}) = \sum_{i=0}^{n-1} d_i(\boldsymbol{x}_i, \boldsymbol{c}_i)$$

The design goal for the Conditional Entropy-constrained RVQ is to minimize the Lagrangian

$$J_\lambda = E\{d(\boldsymbol{X}, \boldsymbol{C})\} + \lambda E\{l(\boldsymbol{I})\}$$

where $d(\boldsymbol{X}, \boldsymbol{C})$ is the distortion between the supervector $\boldsymbol{X}$ and the codevector sequence $\boldsymbol{C}$, and $l(\boldsymbol{I})$ is the length of the index sequence $\boldsymbol{I}$. Ideally, we choose the length of the codevector sequence to be

$$l(\boldsymbol{I}) = -\log P(\boldsymbol{C}).$$

In order to minimize J_λ, we compute the Lagrangian for all possible combinations of codevector sequences, which can grow extremely large as n increases. Large supervectors will require a large number of additions, a large amount of storage for $P(\boldsymbol{C})$ and a large variable length code.

The solution we adopt is to use first or second-order conditioning models to approximate the probability of occurrence of a particular codevector sequence. Assuming a first-order conditional model, the probability of a specific codevector sequence has the form

$$P(\boldsymbol{C}) = P(\boldsymbol{c}_0)P(\boldsymbol{c}_1|\boldsymbol{c}_0)P(\boldsymbol{c}_2|\boldsymbol{c}_1)\cdots, P(\boldsymbol{c}_{n-1}|\boldsymbol{c}_{n-2}). \quad (2)$$

The Lagrangian associated with this model is given by

$$\begin{aligned} J_\lambda &= d(\boldsymbol{x}_0, \boldsymbol{c}_0) - \lambda \log P(\boldsymbol{c}_0) \\ &+ \sum_{i=1}^{n-1} \{d(\boldsymbol{x}_i, \boldsymbol{c}_i) - \lambda \log P(\boldsymbol{c}_i|\boldsymbol{c}_{i-1})\}. \end{aligned} \quad (3)$$

The Lagrangian J_λ in equation (3) is the sum of Lagrangians from each component of the supervector, where the Lagrangian component vector is given by

$$d(\boldsymbol{x}_i, \boldsymbol{c}_i) - \lambda \log P(\boldsymbol{c}_i|\boldsymbol{c}_{i-1}). \quad (4)$$

The above equations dictate that we need to find conditional probabilities and then find the best codevector sequence to represent the supervector that minimizes the Lagrangian in equation (3). For

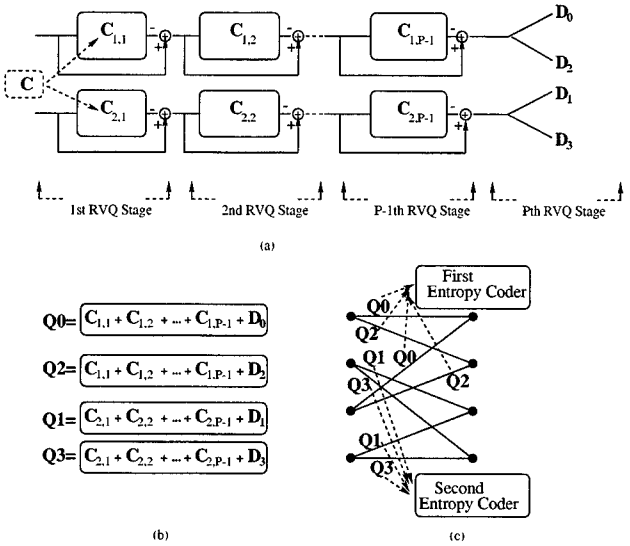


Figure 2: Conditional Entropy-constrained Trellis Coded Residual Vector Quantizer

residual vector quantizers, conditioning codevectors (or symbols) may come from the previous spatial location (intra-stage) or from a previous residual stage location (inter-stage). The procedure proposed in [5] is to search a small neighboring region in both intra- and inter-stage space to find the optimal symbol for conditioning. The candidate symbols are then arranged in a tree structure. Subject to conditioning complexity, the BFOS algorithm [8] may be used to determine conditioning symbols for every residual stage.

Once the best conditioning symbols and order are determined for each residual stage, the next task is to compute the Lagrangian for all the possible codevector sequences under the above conditioning model. In [5], the authors adopted an algorithm to find the best codevector for each component of the supervector in isolation by minimizing the Lagrangian in equation (4).

3. CONDITIONAL ENTROPY-CONSTRAINED TCRVQ

Let R be the encoding rate (in bits per sample) and n the source vector dimension. The conditional entropy-constrained trellis-coded residual vector quantizer (CEC-TCRVQ) proposed here uses an N-state trellis with two branches entering and leaving each state. Figure 2(c) shows a 4-state trellis with two branchess entering and leaving. The trellis branches are labeled with codebooks obtained as follows. The encoding rate R can be decomposed into stage component rates given by $R = R_1 + R_2 + \ldots + R_P$. Let $\boldsymbol{C}$ be the first stage expanded codebook with 2^{nR_1+1} code vectors. Then $\boldsymbol{C}$ is partitioned, in the sense of increasing intra-codebook distance, to form two first stage codebooks $\boldsymbol{C}_{1,1}$ and $\boldsymbol{C}_{2,1}$ as shown in Figure 2(a). Two RVQs are designed next to match their first stage codebooks respectively. The last stage of each RVQ is partitioned again to form four sub-codebooks i.e $\boldsymbol{D}_0, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3$. The codebook labeling for the trellis, $\boldsymbol{Q0}, \boldsymbol{Q1}, \boldsymbol{Q2}, \boldsymbol{Q3}$ are obtained by joining stage codebooks and last stage subcodebooks as shown in Figure 2 (b). Selection of one or the other RVQ structure at any given time instant depends on which trellis state we are in at that instant. The indices from the two RVQs are subsequently entropy coded

by using their respective entropy coders.

The entropy-coders employed in CEC-TCRVQ make use of conditioning symbols, which come from the previous adjacent vectors as well as previous residual vectors. Like CEC-RVQ described in [5], we search a smaller region in both intra-residual stage and inter-residual stage space for finding conditioning symbols for a given residual stage in CEC-TCRVQ design. Then a complexity-entropy trade off tree is constructed with the number of branches equal to the number of residual stages present in the underlying RVQ. The tree is searched using a BFOS algorithm [8] to find the best conditioning symbols along with conditioning model order for each residual stage. Then Lagrangian of the form of equation (3) is found by using the Viterbi algorithm along a trellis structure with component Lagrangians of equation (4) as a branch metric.

4. SUBBAND QUANTIZATION SCHEME

In our proposed scheme, the image is split into a pyramid, and each pyramid is coded independently. The pyramid construction process begins with the splitting of the image into four subbands, and then continues with the division of the lowpass band recursively up to the required level of decomposition. Here we used three levels of decomposition to get ten subbnads.

We used trellis-coded residual vector quantization (TCRVQ) [2] for designing codebooks of the image pyramids by employing a training set of 14 (512×512) images. The image subbands usually differed in their spectral contents [7], therefore normalized codebooks were designed for each subband by dividing all of the training data by their respective standard deviations. The mean of the baseband (LL3 band) was also subtracted. Therefore, the mean of the baseband and the standard deviations of its ten bands needed to be sent to the decoder. This overhead information corresponds to a negligible increase in the overall bit rate.

Figure 3 shows various trellis-coded residual vector quantization schemes used to quantize the subbands. The LL3 band which contains the texture, also contains strong two-dimensional correlation. In order to effectively exploit the correlation to reduce the bit rate, we coded the LL3 band using three-stage conditional entropy-constrained trellis-coded residual scalar quantization (CEC-TCRSQ). The reason for using a scalar quantizer lies in the fact that it is difficult to code textures using vector quantization without producing visual artifacts. The bands HL3, HH3 and LH3, also contain some vertical and horizontal correlation so we employed three-stage two-dimensional conditional entropy-constrained trellis-coded residual vector quantization. There is little correlation present in the HL2, HH2, and LH2 bands. Therefore, we used four-dimensional trellis-coded residual vector quantization for these bands. The HL1, HH1 and LH1 bands contained very small correlation and also a small amount of image energy. Hence we needed to code them at low bit rates. In our scheme, we coded these bands using 16-dimensional trellis-coded residual vector quantization.

5. BIT ALLOCATION

Once the quantization scheme is specified for the image pyramids, the next issue is how to distribute the bit budget among the subbands. Westerink, Biemond, and Boekee [11] developed an optimal bit allocation algorithm based on the subband variance. Riskin

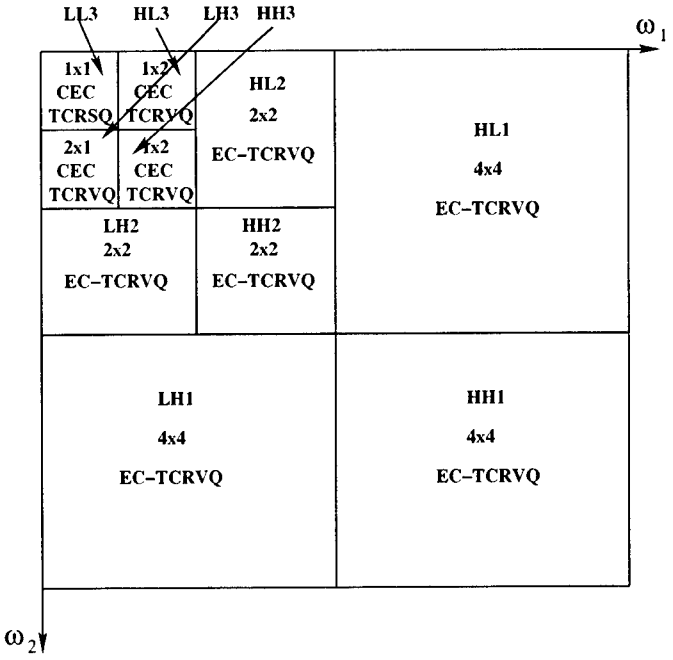


Figure 3: Trellis-coded quantization scheme for image subbands.

[8] restated their algorithm using the generalized BFOS algorithm for both cases of convex and non-convex operational distortion-rate functions.

We employed bit allocation using the generalized BFOS algorithm.The BFOS algorithm can be used as follows: construct a tree T with l subtrees where each subtree is a unary tree and represents a subband. In each subtree we have k nodes where each node is represented by an (R, D) point found during the quantization design. If we denote the initial tree by T_I, the generalized BFOS algorithm will prune off the branches of the initial tree in order to form the final pruned tree T_F. In this pruning operation, the algorithm obtains a sequence of trees where each intermediate tree T_{i+1} is obtained by pruning off the node having the smallest slope s in the tree T_i. The pruned leaf node belongs to a certain subtree, and therefore this iteration provides a new leaf node in the previous tree. After this procedure, the s ratio must be re-calculated in this new tree T_{i+1}. The algorithm ends when the sum of the leaf node rates drops below the target rate. The codebook used to encode each subband corresponds to the codebook specified by the leaf nodes of the final pruned tree T_F.

6. SIMULATION RESULTS

In this section, we present results for 512×512 Lena at low bit rates. For the bit allocation tree, we obtained thirty rate-distortion pairs for each subband. The tree has ten branches with thirty points on each branch.Figure 4 shows Lena image coded at 0.125 bits per sample using our scheme. We observe that the image coded by our scheme is slightly blurred in nature. We also noticed the presence of small magnitude texture on the Lena hat in our coded image.

Figure 5 compares our TCRVQ-based subband coder (TCRVQ-SBC) with other results in the literature for the test image Lena. Kim and Modestino [4] report PSNR's of 34.04 dB, 35.28 dB, 35.98 dB, 37.23 dB for bit rates of 0.31, 0.41,



Figure 4: Image Lena coded at 0.125 bits per pixel using our proposed image coder, PSNR = 30.43 dB.

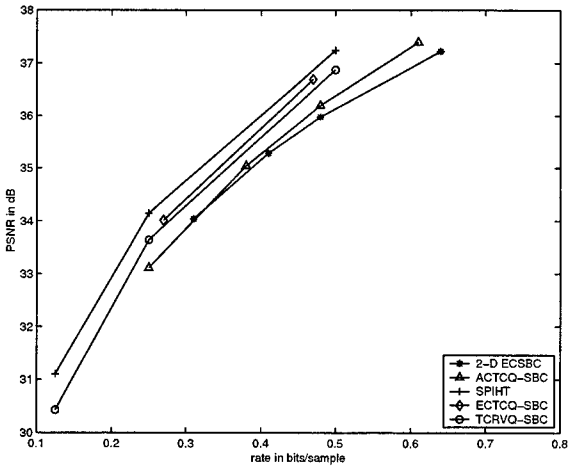


Figure 5: Comparison of subband coding performance for "Lena".

0.48, and 0.64 bpp, respectively, for their entropy-constrained subband coder (2-D ECSBC). Joshi, Crump and Fischer [1] devloped arithmetic-coded trellis-coded subband image coder (ACTCQ-SBC) and is shown to provide about 0.25 dB improvement over the 2-D ECSBC design. Sriram and Marcellin [10] report PSNR's of 34.01, 36.70, and 40.06 dB for bit rates of 0.27, 0.47, and 0.95 bits per pixel, respectively, for their entropy-constrained trellis-coded quantization based subband image coder (ECTCQ-SBC). SPIHT [9] results are also displayed in the figure. The figure shows that our coder does better than the ACTCQ-SBC and the 2-D ECSBC. Comparing the performance of our coder with that of ECTCQ-SBC shows that TCRVQ-SBC performance is worse by about 0.5 dB at 0.5 bits per pixel and is about 0.15 dB worse at 0.25 bits per pixel. This may be due to the reason that ECTCQ is a single stage system as compared to ECTCRVQ. The TCRVQ-SBC performs worse in comparison to SPIHT by about 0.6 dB. We believe that this gap is due to the reason that SPIHT coder exploits inter-band dependence while our coder does not.

7. REFERENCES

[1] R.L. Joshi, V.J. Crump, and T.R. Fischer. Image subband coding using arithmetic-coded trellis-coded quantization. *IEEE Trans. on circuits and systems for video technology*, 5(6):515–523, Dec. 1995.

[2] M.A. Khan, M.J.T. Smith, and S.W. Mclaughlin. Trellis-coded residual vector quantization with application to image coding. In *proceedings of IEEE International symposium on circuits and systems*, Orlando, Florida, Jun 1999.

[3] M.A.U. Khan, M.J.T. Smith, and S.W. McLaughlin. Conditional entropy-constrained trellis-coded rvq with application to image coding. *IEEE transactions on signal processing letters*, 7(3):49–51, March 2000.

[4] Y.H. Kim and J.W. Modestino. Adaptive entropy-coded subband coding of images. *IEEE trans. on image processing*, 1:31–48, Jan 1992.

[5] F. Kossentini, W.C. Chung, and M.J.T. Smith. Conditional entropy-constrained residual vq with application to image coding. *IEEE transactions on image processing*, 5(2):311–320, Feb 1996.

[6] F. Kossentini, M.J.T. Smith, and C.F. Barnes. Image coding using entropy-constarined residual vector quantization. *IEEE transactions on image processing*, 4(10):1349–1356, October 1995.

[7] B. Mahesh and W.A. Pearlman. Multiple-rate structured vector quantization of image pyramids. *Journal of visual communications and image representation*, 2:103–113, June 1991.

[8] E.A. Riskin. Optimal bit allocation via the generalized bfos algorithm. *IEEE Trans. on Information Theory*, 37:400–402, Mar 1991.

[9] A. Said and W. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. on circuits and systems for video technology*, 6:243–250, June 1996.

[10] P. Sriram and M.W. Marcellin. Image coding using wavelet transforms and entropy-constrained trellis-coded quantization. *IEEE Trans. on Image Processing*, 4:725–733, June 1995.

[11] P.H. Wasterink, J. Biemond, and D.E. Boekee. An optimal bit allocatiohn algorithm for sub-band coding. In *proceedings of ICASSP*, pages 757–760, 1988.

TEXTURE SEGMENTATION BY CLUSTRING THE PHASE OF HOS CEPSTRA

[1]S. Sanei,[2]J. Li, [2]S. H. Ong

[1]EEE department, Singapore Polytechnic, Singapore
[2]EE department, National University of Singapore, Singapore

ABSTRACT

Many applications require identification, segmentation and deconvolution of textures and detection of the objects of combined patterns. Frequency-based analysis of patterns [1] does not avoid the redundancy which highly deteriorates the process. On the other hand, spatial quantifiers [2] rely on the rough estimation of the model parameters, which are not robust enough to details and alteration of the regions size. By using HOS [3], classification based on variation in minimum and maximum phase of cepstra, through a spatial-based sliding window incorporates both space and frequency differences. A small number of minimum and maximum phase coefficients are then evaluated for an sliding window of fixed size. The results show an attractive implementation of HOS estimation in texture segmentation.

1. INTRODUCTION

Image segmentation process separates the regions of different statistics. Grey-level, colour and texture segmentations are the main topics in this field. Texture segmentation has been widely demanded due to its application in pattern and object recognition. Deconvolution of mixed textures in image segmentation has been the subject of research recently. Higher order statistics (HOS) in signal separation and deconvolution has also been under research by many researchers [3][4][5]. Traditional methods in texture segmentation such as application of Gabor filtering, Auto-regressive modelling, etc fails in deconvolution of mixtures and detection of the boundary between the true texture and contaminated one. For example detection of the objects partially covered by nets and stains or partitioning the human tissue into normal and slightly malignant can be mentioned as suitable patterns for our experiments. The malignant tissue is a combination of normal cells pattern and a non-uniform granular texture. However, most of our experiments are on Brodatz textures and their combinations.

The proposed method requires measurement of 3rd order statistics and their spectrum. In bispectrum domain a zero-mean quasi-Gaussian noise will be suppressed or highly abated. This enhances accuracy of estimation of the signals parameters in that domain. Application of accurate measurement criteria and near-optimal estimation of the pattern statistics enhance the outcome of the process. In next part the theoretical approach will be explained. The implementation result comes next.

2. PRELIMINARIES

Some images can be viewed as an original texture partially contaminated by one or more other textures. The textures may also be polluted by Gaussian noise. In this case recovering the actual texture from the mixed pattern is required so we can convert the question of texture segmentation into a question of signal reconstruction.

For a minimum phase sequence, the log magnitude of its Fourier transform and its Fourier phase form a Hilbert transform pair. Hence, we can compute the signal's Fourier magnitude from its Fourier phase and vice versa. Consequently, the knowledge of only the Fourier phase or magnitude of minimum phase signal can lead to the unique reconstruction of the signal. However the reconstruction is subject to fulfilling certain requirements. The conditions under which a general FIR sequence can be reconstructed form its bispectral phase only can be stated as [6]:

Let $x(k)$ and $y(k)$ be two FIR sequences which are zero outside the interval $[0,N-1]$, and their Z transform have no zeros on the unit circle, nor its reciprocal pairs. Let $\varphi_3^x(\omega_1,\omega_2), \varphi_3^y(\omega_1,\omega_2)$ be the bispectral phase of $x(k)$ and $y(k)$ respectively. Also suppose we sample the bispectral phases at $L = 2^v > 2N-1$ equal-space frequency points. If $\varphi_3^x(\omega_1,\omega_2) = \varphi_3^y(\omega_1,\omega_2)$ at discrete frequency pairs within the non-redundant bispectrum region $\{0 \le \omega_1 + \omega_2 \le \pi,\ \omega_2 \le \omega_1, \omega_1 \ge 0\}$, then we

0-7803-7011-2/01/$10.00 ©2001 IEEE

have $x(k) = ay(k-k_0)$, for some positive constant α. and some integer k_0.

In this case let $x(k)$ be a FIR sequence which is zero outside the interval $[0,N-1]$, and its Z transform has no zeros on the unit circle, nor its reciprocal pairs. Let $\varphi_3^x(\omega_1,\omega_2)$ be the bispectral phase of $x(k)$. Suppose we sample the bispectral phases at $L = 2^v > 2N - 1$ equispaced frequency points. The BIRA algorithm [3] can recover a scaled and shifted version of the original signal form its bispectral phase only. This algorithm proceeds as follows:

Estimate the bicepstrum of $x(k)$ and then compute the values of $A(m) - B(m)$, we have:

$$D(m) = A(m) - B(m),\ m = 1,2,\ldots r \qquad (1)$$

Where $r = max\ (p, q)$, p and q are the lengths of $A(m)$ and $B(m)$ respectively. When the Fourier magnitude is corrupted, only the differences of the computed cepstral coefficients contain undistorted information. Note that:

$$D(m) = -2m \cdot bic_x^0(m) \qquad (2)$$

where $bic_x^0(m)$ is the initial bispectrum, i.e. at iteration $i = 0$. Initially we set each sum of the cepstral coefficients to some arbitrary value such as zero;

$$A^0(m) + B^0(m) = 0,\ m = 1,2,\ldots,r \qquad (3)$$

Where $A^0(m)$ and $B^0(m)$ denote the values of the cepstral coefficients at iteration $i = 0$. Thus we have:

$$A^0(m) + B^0(m) = -mp_x^0(m) = 0,\ m = 1,2,\ldots,r \qquad (4)$$

Where $p_x^0(m)$ denotes the value of the cepstrum of $x(k)$ at iteration $i = 0$. The reconstructed signal will be achieved after following the iterations below.

Step 1: Combine (2) and (3) for any iteration i, we have:

$$A^i(m) = \frac{D(m) - m \cdot p_x^i(m)}{2} \qquad (5)$$

$$B^i(m) = \frac{-D(m) - m \cdot p_x^i(m)}{2} \qquad (6)$$

Where $m=1,2,\ldots r$ and $\{i\}$ is the iteration index.

Step 2: Compute $x^i(k)$ using the following relationship:

$$x^i(k) = F_1^{-1}\left\{e^{F_1\{c_x^i(m)\}}\right\} k = 0,\ldots M-1 \qquad (7)$$

$$c_x^i(m) = \frac{1}{m} \times \begin{cases} -A^i(m) & m > 0 \\ 0 & m = 0 \\ B^i(m) & m < 0 \end{cases} \qquad (8)$$

Where $x^i(k)$ is the computed sequence at each iteration and M ($M > 2r$) is the length of the Fourier transform used in equation (7).

Step 3: Generate the sequence $y^i(k)$ as follows:

$$y^i(k) = x^i(k) w_N(k), \quad k = 0,\ldots M-1 \qquad (9)$$

Where

$$w_N(k) = \begin{cases} 0 & M - k_0 \le k \le N - k_0 - 1 \\ 1 & otherwise \end{cases} \qquad (10)$$

Where k_0 is the time shift introduced to the signal due to its reconstruction from its cepstrum coefficients and W_N shows the size of the window. Due to this shift, $x^i(k)$ will appear in the interval $[-k_0,N-k_0-1]$, and is computed from (7). k_0 is identified by using an iteration algorithm. This will be discussed later in this part.

Step 4: Calculate the power cepstrum of $x^i(k)$,

$$p_y^i(m) = F_1^{-1}\left\{\ln\left|Y^i(\omega)\right|^2\right\} \qquad (11)$$

and set

$$p_x^{i+1}(m) = p_y^i(m) \qquad (12)$$

Repeat Steps1-4 until the reconstructed sequence $x^i(k)$ remains unchanged. In other words, if we define

$$E_i = \sum_{k=0}^{M-1}\left[x^i(k) - x^{i-1}(k)\right]^2 \qquad (13)$$

The algorithm stops at $i = I$, *when* $E_I < \delta$ where δ is a very small constant.

Then we can get the result:

$$x^I(k) = \alpha x(k - k_0) \qquad (14)$$

For the algorithm to converge k_0 has to be accurately identified. In order to determine the time shift k_0 we guess an initial value for k_0 within $[0,N-1]$ (we can start from 0). Apply Step1-4 while checking E_0 for the successive iterations. A second loop is used to decide about the value

of k_0 . The value of k_0 is incremented one by one and the outcome will be tested.

However, since $x(k)$ must be a minimum phase FIR sequences which is zero outside the interval $[0,N-1]$ and estimation of the HOS parameters involves error, for some natural data, the performance of the algorithm is not satisfactory. Conditioning of the signal without its deterioration is a solution. However, in our experiment we tried to use long enough signal (one dimensional scan of the image) and lowpass filter the signal before processing.

3. TEXTURE SEPARATION

In our specific implementation, let one signal $x(n)$ to pass through two LTI channels $h_1(n)$, $h_2(n)$ and result in $x_1(n)$, $x_2(n)$. We then use above algorithm to restore $x(n)$. There are some prerequisites:

i. the channels $h_1(n)$ and $h_2(n)$ are finite-duration impulse/response sequences;
ii. there are no zero-pole cancellation between $X(Z)$ and the channels $H_1(Z)$ and $H_2(Z)$;
iii. $H_1(Z)$ and $H_2(Z)$ have no common zeros.

Petropulu extended above arguments to non-linear signals too [7][8]. The reconstruction process operates on row-by-row of the image and restores the original texture from the overlapping ones. Obviously the difference between the original mixed texture and the final reconstructed one is expected to be mainly in the overlapping region.

$$d_j(n)= x^I_j(n) - x^I_j(n-1) \qquad (15)$$

where j denotes the data segment which can be a row of the image. $d_j(n)$ varies smoothly and with low amplitude if there is no change in the texture. At the overlapping section it introduces a remarkable change. Obviously, $d_j(n)$ is not sensitive to Gaussian white noise since HOS of noise tends to zero.

Different combination of patterns yield different measures for $d_j(n)$. a limited number of regions introduce a number of distinct clusters for $d_j(n)$. A differential competitive learning (CL) neural network has been built up to cluster above $d_j(n)s$. The network is similar to the traditional Kohonen unsupervised NN, except, the winner neurons are defined as those whose current and one level previous values are above a threshold level. The weights to the winner and its two adjacent neurons are updated. This highly avoids the effect of non-Gaussian noise in the texture and idle spikes in $d_j(n)$.

4. EXPERIMENTAL RESULTS

Combination of various Brodatz textures and their mixtures has been used to show the performance of the proposed algorithm. Figure 1 represents a combination of two Brodatz textures in which one is partly overlapped by another. The values of $d_j(n)$ have been measured for the 256×256 image of Figure 1. The image is scanned line by line and $d_j(n)s$ are measured for each sliding window at J

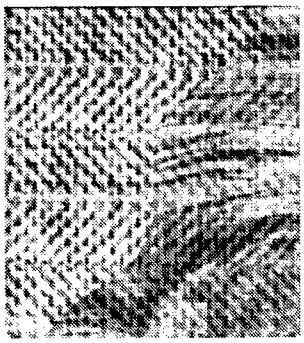


Figure 1. A Brodatz texture overlapped with another pattern

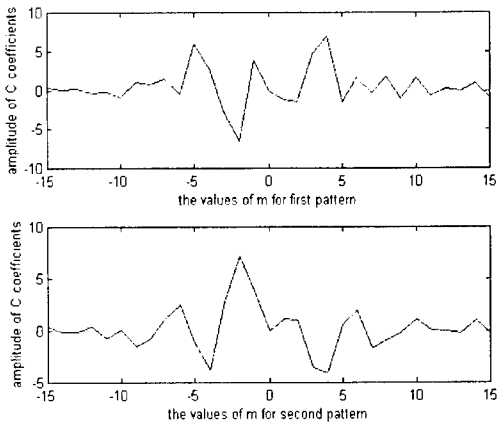
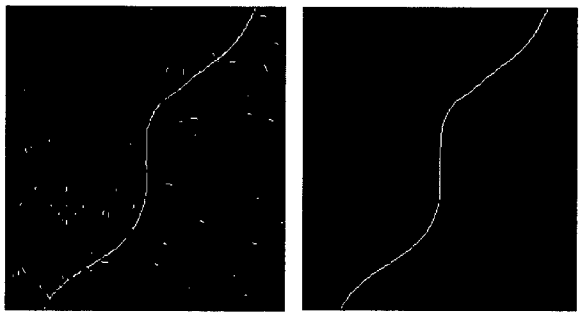


Figure 2. $C_k(m)$ for two different regions (mixed) partly by the other texture

positions. The value of p has been considered to be constant equal to 8. Therefore only 16values for each $d_j(n)$ (i.e. n = *-8 to 8*, $n \neq 0$) have been used. The number of clusters is initially set based on the desired number of regions. If there is no prior knowledge about the number of regions, the number of peaks in $d_j(n)s$ can be considered as the maximum number of clusters. However in majority of cases where there is only one background texture, the number of outputs can be tentatively set to 2. Figure 2 shows $C_k(m)$ for two different regions of the image. Finally in Figure 3 *a* and *b*

The segmented regions before and after post-processing are depicted.

Figure 4 represents another pattern. In this figure, part of the image has been covered with another texture. Figure 5 illustrates the texture after reconstruction process. Figure 6.a and b show the boundary of the contaminated area before and after post-processing respectively.



(a) (b)

Figure 3. The boundary detected before (a) and after (b) post-processing

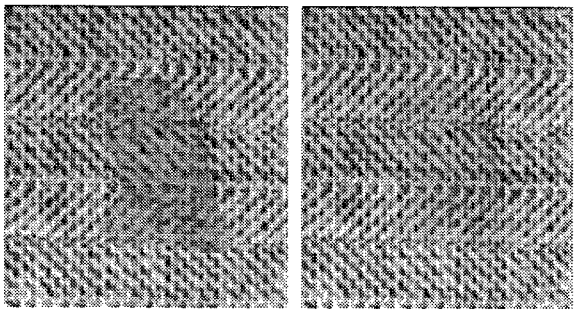
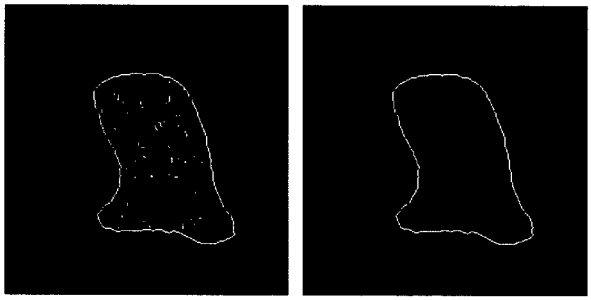


Figure 4. The mixed pattern

Figure 5. The reconstructed Texture



a b

Figure 6. The boundary of the contaminated region a. before and b. after post-processing

4. CONCLUSIONS

Unlike frequency-based methods which are sensitive to noise components the proposed method avoids the noise effect in two different stages. In the first stage the components of WGN are eliminated in HOS of the signal. In he second stage a differential updating the weights in the modified CL-NN suppresses the remaining noise and also enhances the performance of the system. The method overcomes some shortcoming of the time-based systems that mainly handle distinct uniform patterns. In application of wavelet in texture segmentation definition of the frequency of the basic function is one of the major issues. This requires a rough prior knowledge about the image variations. Application of HOS in image segmentation is superior to above established methods especially when a uniform pattern follows by a region of the mixed pattern. The mixture includes the original image. In fact the original pattern has been somehow used in extraction of the required prior knowledge.

REFERENCES

[1] T.N. Tan and A.G. Constantinides, " Texture analysis based on human visual model", *Proceeding of IEEE, ICASSP*, 1990, pages 2137-2140

[2] S. Sanei, R. I. Kitney and R. Benjamin, "adapting transform coding to texturally segmented images", *proceeding of IEEE, ACSSC*, 1991, CA, USA, pages 252-255

[3] C. L. Nikias and A. P. petropulu, "Higher-order spectra analysis", 1993, Prentice Hall

[4] K. J. Pope and R. E. Bonger, "Blind Signal Separation, 1. Linear, Instantaneous Combinations, *Journal of Digital Signal Processing*, No. 6, 1996, pages 5-16

[5] K. J. Pope and R. E. Bonger, "Blind Signal Separation, 1. Linear, Convolutive Combinations, *Journal of Digital Signal Processing*, No. 6, 1996, pages 17-28

[6] A. P. Petropulu and C.L. Nikias, "Signal reconstruction from the phase of bispectrum", *IEEE trans. on signal processing*, Vol. 40, No. 3, March 1992, pages 601-610

[7] A. P. Petropulu and C.L. Nikias, "Blind deconvolution of coloured signals based on higher-order cepstra and data fusion", *IEE Proceeding -F*, Vol. 140, No. 6, Dec. 1993, pages 352-361

[8] A. P. Petropulu, "Blind deconvolution of non-linear random signals", *IEEE Proceeding*, 1993, pages 205-209

R-D QUANTISATION OF COMPLEX COEFFICIENTS IN ZEROTREE CODING

T.H. Reeves and N.G. Kingsbury

Signal Processing Group
Department of Engineering
University of Cambridge, U.K.
thr20@eng.cam.ac.uk, ngk@eng.cam.ac.uk

ABSTRACT

This paper describes a rate-distortion (R-D) optimal scheme for bit-plane based quantisation of complex coefficients, which is suitable for zerotree image coding systems. Most zerotree-type image codecs operate on real-valued wavelet coefficients. The Dual-Tree Complex Wavelet Transform, which has several advantages over the discrete wavelet transform, produces complex coefficients. Our scheme offers progressive bit-by-bit refinement of coefficient magnitude and phase values. It ensures that refinement decisions always maximise the expected distortion decrease.

1. INTRODUCTION

Zerotree-type (ZT-T) coding systems, *e.g.* [1] [2], are well-known for providing efficient and effective image compression. They belong to a larger class of bit-plane based systems that implicitly quantise data values as a consequence of the encoder's execution path. The coded bitstreams produced by these systems are often progressive and embedded.

ZT-T codecs usually operate on a discrete wavelet transform (DWT) of an image. They exploit the multiscale observation that, when a small (insignificant) wavelet coefficient appears in a coarser level, the coefficients in the same spatial locations of the finer scales are likely also to be insignificant. These codecs are more efficient when coefficients in a local neighbourhood all have similar magnitudes. The shift invariance of a transform's response to image features increases the likelihood that coefficient magnitudes will be locally correlated between and within scales.

ZT-T codecs are considered close to optimal in a rate-distortion (R-D) sense for scalar-quantised (SQ) real values. Because the overwhelming majority of wavelet transforms produce real-valued coefficients, most literature analysing the rate-distortion performance of ZT-T codecs, and proposing improvements, tends to focus on real-valued data.

Some notable complex wavelet transforms are Daubechies' complex wavelets [3] and the Dual-Tree Complex Wavelet Transform (DT-CWT) [4] [5], both of which are redundant for real image data. The DT-CWT is a perfect reconstruction transform with Gabor-like filters. It uses two trees per dimension, each with short, linear phase real lowpass and highpass filters, to simulate a single complex lowpass/-highpass filter pair. The filters in the two trees of [5] are just the time-reverse of each other, as are the analysis and reconstruction filters. For 2 dimensional signals, the DT-CWT has 4:1 redundancy.

The DT-CWT has several advantages over the conventional critically-sampled DWT. It has good directional selectivity in multiple dimensions, and can distinguish between positive and negative signal frequencies. Significantly, the magnitude response of the DT-CWT is approximately shift invariant. This is a very beneficial property for ZT-T coding systems which rely on local interscale and intrascale correlations of wavelet coefficient magnitudes.

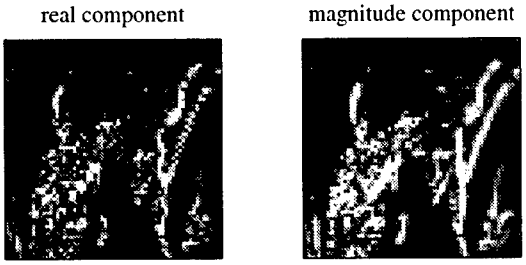


Fig. 1. Comparison of complex components

Generally, complex-valued data is quantised using scalar quantisers independently on the data's real and imaginary components, or using some form of vector quantisation (VQ), such as trellis coded quantisation. The problem when quantising complex-valued wavelet coefficients within a ZT-T codec is the lack of correlation between related coefficients when expressed as real and imaginary components. ('Related' refers to the family tree structure imposed on coefficients by ZT-T systems.) For example, the real and imaginary components of DT-CWT coefficients exhibit shift dependence like DWT coefficients.

0-7803-7011-2/01/$10.00 ©2001 IEEE

Figure 1 illustrates the enhanced local correlations among the magnitudes of complex coefficients, compared to the magnitudes of the real components only. From the 2nd level, vertical subband of the DT-CWT transform of the 256×256 'Lena' image, outlines of Lena's hair, hat, and mirror are visible. The enhanced local correlations are evident from the smoother, more continuous lines and edges in the magnitude component image.

The purpose of this paper is to develop a bit-plane based quantisation scheme for complex coefficients, suitable for ZT-T codecs, that is optimal in an R-D sense.

2. OPTIMAL R-D QUANTISATION

The magnitudes of related DT-CWT coefficients are well correlated. They are therefore more appropriate for ZT-T codecs than the individual magnitudes of the real and imaginary components. To extend a ZT-T codec for complex coefficients, the coefficients' magnitudes could be the basis for the codec's significance threshold tests (*i.e.* the 'Sorting Pass' of the SPIHT algorithm [2]). The individual magnitudes of the real and imaginary components could then be refined one bit each for each subsequent level of the algorithm (*i.e.* SPIHT's 'Refinement Pass'). The obvious inefficiency here is that, if one component is much larger than the other, many 0 bits are used to describe the insignificant component.

Instead, we propose refining coefficients' magnitude and phase components individually. The coefficients' magnitudes are used for the ZT-T algorithm's significance-based tests and decisions. As with real-valued coefficients, once a coefficient is found to be significant compared to the current threshold, its magnitude is refined by one bit at each subsequent threshold. We assume that the usual thresholding system applies; *i.e.* all thresholds can be expressed as powers of 2, and if the threshold at level k is t_k, then the threshold during the next pass is $t_{k-1} = t_k/2$.

Obviously, the phases of significant coefficients must be refined concurrently with the magnitude refinement. The issue here is the determination of how many phase refinement bits to process at each level. We use an R-D approach.

Let $\mathbf{x} = re^{j\theta}$ be the true value of a complex coefficient, and $\hat{\mathbf{x}}_{k,l} = \hat{r}_k e^{j\hat{\theta}_l}$ be its quantised (*i.e.* estimated or reconstructed) value at level k of the algorithm. The l subscript denotes that fact that phase, unlike magnitude, is not necessarily refined 1 bit/level. We choose the squared error (square of the l_2 norm) as our distortion measure:

$$D_{\mathbf{x}}(r,\theta;\hat{r}_k,\hat{\theta}_l) = (\hat{r}_k\cos\hat{\theta}_l - r\cos\theta)^2 + (\hat{r}_k\sin\hat{\theta}_l - r\sin\theta)^2$$

The change in distortion if the next refinement bit processed for $\mathbf{x}$ is a magnitude refinement bit, or phase bit respectively, is:

$$\begin{aligned}\Delta D_{k-1,l} &= D_{\mathbf{x}}(r,\theta;\hat{r}_k,\hat{\theta}_l) - D_{\mathbf{x}}(r,\theta;\hat{r}_{k-1},\hat{\theta}_l)\\ \Delta D_{k,l-1} &= D_{\mathbf{x}}(r,\theta;\hat{r}_k,\hat{\theta}_l) - D_{\mathbf{x}}(r,\theta;\hat{r}_k,\hat{\theta}_{l-1})\end{aligned}$$

Let the rate changes associated with the above distortion changes be $\Delta R_{k-1,l}$ and $\Delta R_{k,l-1}$. The codec should process phase refinement bits for $\mathbf{x}$, before the next refinement bit, while:

$$\frac{E[\Delta D_{k,l-1}]}{E[\Delta R_{k,l-1}]} > \frac{E[\Delta D_{k-1,l}]}{E[\Delta R_{k-1,l}]} \qquad (1)$$

Without entropy compression of the coded bitstream, the rate change due to phase or magnitude refinement is exactly 1 bit. Equation (1) reduces to the question of whether increasing the phase quantisation precision by 1 bit is expected to result in a larger distortion decrease than increasing the magnitude quantisation precision by 1 bit. This is the strategy of our R-D based complex quantiser. Before the next algorithm level – when the next magnitude refinement bit will be (de)coded – *process phase refinement bits while they give greater expected distortion decreases than the next magnitude refinement bit.*

2.1. Quantisation cells

Before proceeding to calculation of the expected distortions, let us briefly consider the geometry of the 2-D quantiser we propose. Assume that, at level k, coefficient $\mathbf{x}$ is newly significant, *i.e.* $t_k \leq \mathbf{x} < t_{k+1}$. Without any phase information, the decoder knows only that $\mathbf{x}$ lies in the ring with inner radius is t_k and outer radius t_{k+1}. With 1 bit of phase information, the range of possible values of $\mathbf{x}$ is half of the ring; with 2 bits, the range is a quarter of the ring (figure 2), *etc.*.

The next refinement bit for $\mathbf{x}$ received by the decoder reduces the ring range segment (*i.e.* quantisation cell) to one of the four overlapping segments shown in figure 2. The encoder knows the true distortion change associated with each segment, and can therefore decide to send the refinement bit that is R-D optimal. However, unless the decoder can follow the same decision paths as the encoder, the encoder must include decision overhead information in the coded bitstream. Any gains from using the R-D optimal bits are easily offset by the cost of the overhead [6].

The decoder can calculate the expected distortion over each of the four segments, and decide which one offers the greatest expectation of distortion reduction. The encoder must use the same decision rule to determine which type of refinement bit to code.

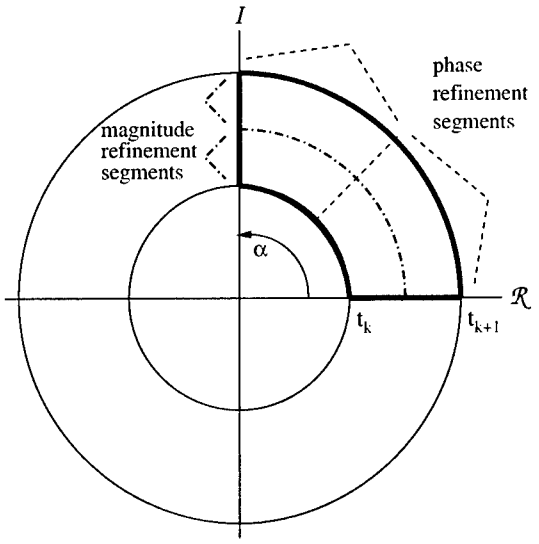


Fig. 2. Quantisation cells in complex plane

2.2. Expected distortion

Calculation of expected distortions over a quantisation cell requires the magnitude and phase joint probability distribution $p(r,\theta)$. Actually, because we are concerned with the pdf only *after* the coefficient has become significant, when we know the most significant bit (MSB) of its magnitude, we want the joint conditional distribution:

$$p(r,\theta \mid 2^{\lfloor \log_2 r \rfloor} \leq r < 2^{\lceil \log_2 r \rceil})$$

where $\lfloor\rfloor$ and $\lceil\rceil$ denote the floor and ceiling, respectively.

For the moment we shall use the simple and not unrealistic assumptions that r and θ are independent, and uniformly distributed. Our magnitude model therefore assumes that while there is high correlation amongst the MSBs of related coefficients, all lesser bits are uncorrelated – in fact, independent – and equiprobable.

The radial width of the ring is the decoder's uncertainty in the magnitude of $\mathbf{x}$. The angular width α of the ring range segment is the decoder's uncertainty in the phase of $\mathbf{x}$. Note that, when $p(\theta)$ is assumed uniform, the phase uncertainty is independent of the true phase value. Consequently, to simplify expected distortion calculations, we can treat all ring range segments as being bounded between angles 0 and α.

The expected distortion for any ring segment quantisation cell, such as the cells shown in figure 2, is:

$$\begin{aligned} E[D_{\mathbf{x}}(r,\theta;\hat{r},\hat{\theta})]_{t_1,0}^{t_2,\alpha} &= \int_0^{\alpha}\int_{t_1}^{t_2} p(r,\theta) D_{\mathbf{x}}(r,\theta;\hat{r},\hat{\theta})\, dr\, d\theta \\ &= \hat{r}^2 + \frac{(t_2^3 - t_1^3)}{3\Delta t} - \frac{(t_2+t_1)}{\alpha}\hat{r} \cdot \\ &\quad (\sin(\alpha-\hat{\theta}) + \sin\hat{\theta}) \qquad (2) \\ \Delta t &= t_2 - t_1 \end{aligned}$$

The reconstruction values $\hat{r}$ and $\hat{\theta}$ should minimise the expected distortion; *i.e.* $\hat{\mathbf{x}}$ should be the centroid of the ring segment bounded by $r \in [t_1, t2)$ and $\theta \in [0, \alpha)$. For the squared error distortion measure, the centroid is simply the expectation of $\mathbf{x}$, given that $\mathbf{x}$ lies in the ring segment above [7]. Therefore, the optimal reconstruction estimates are:

$$\begin{aligned} \hat{\theta} &= \alpha/2 \\ \hat{r} &= \frac{1}{2\alpha}(t_2+t_1)(\sin(\alpha-\hat{\theta}) + \sin\hat{\theta}) \\ &= \frac{1}{2}(t_2+t_1)\,\mathrm{sinc}(\alpha/2) \end{aligned}$$

Using the optimal reconstruction estimates above, equation (2) simplifies to:

$$E[D_{\mathbf{x}}(r,\theta;\hat{r},\hat{\theta})]_{t_1,0}^{t_2,\alpha} = \frac{(t_2^3 - t_1^3)}{3\Delta t} - \left[\frac{1}{2}(t_2+t_1)\,\mathrm{sinc}(\alpha/2)\right]^2 \qquad (3)$$

The encoder and decoder can use (3) to calculate the expected distortions of the four new quantisation cells that result from processing another refinement bit for $\mathbf{x}$. Since the two possible cells that result from a phase refinement bit have the same distortions, the $E[\Delta D_{k,l-1}]$ term from (1) can be re-written:

$$E[\Delta D_{k,l-1}] = E[D_{\mathbf{x}}(r,\theta;\hat{r},\hat{\theta})]_{t_1,0}^{t_2,\alpha} - E[D_{\mathbf{x}}(r,\theta;\hat{r},\hat{\theta})]_{t_1,0}^{t_2,\alpha/2}$$

The two possible cells that result from processing a magnitude refinement bit do not have the same distortions. Since the next magnitude bit is 0 or 1 with equal probability, the $E[\Delta D_{k-1,l}]$ term from (1) can be re-written:

$$\begin{aligned} E[\Delta D_{k-1,l}] &= E[D_{\mathbf{x}}(r,\theta;\hat{r},\hat{\theta})]_{t_1,0}^{t_2,\alpha} - \\ &\quad \frac{1}{2}(E[D_{\mathbf{x}}(r,\theta;\hat{r},\hat{\theta})]_{t_1,0}^{\frac{3}{2}t_1,\alpha} + E[D_{\mathbf{x}}(r,\theta;\hat{r},\hat{\theta})]_{\frac{3}{2}t_1,0}^{t_2,\alpha}) \end{aligned}$$

3. DISCUSSION AND RESULTS

3.1. SQ zerotree coding of complex coefficients

We implemented two complex-coefficient extensions of SPIHT. The coefficients are generated by applying the 2D DT-CWT to an input image to ensure good correlation between the magnitudes of related coefficients. In one sys-

tem, the coefficients are separated into their real and imaginary components, and are treated as two separate pixels in SPIHT's lists of insignificant and significant pixels. Each 2×2 neighbourhood of coefficients, as defined by the SPIHT family tree structure, contains 8 pixels, and has 2 parents (one real and one imaginary). The magnitudes of the complex coefficients are used for pixel and set significance tests.

The second system separates coefficients into magnitude and phase components. It implements the R-D based phase refinement bit versus magnitude refinement bit decision rule described in this paper. All of the lists and tree structures are the same as in SPIHT, with an extra list to manage phase refinement information. For low bit rates, initial tests show the second system provides up to 0.5 dB PSNR improvement over the first system at the same bit rates.

Figure 2 shows the coding performance of the two systems described above when applied to 8-bit 512×512 'Lena' and 'Peppers' images. Because of the DT-CWT's 4:1 redundancy, the performance curves in fig. 2 lie a few dB below those achievable using critically-sampled DWTs. We are currently investigating methods to realise fully the coding gains the DT-CWT's properties should provide.

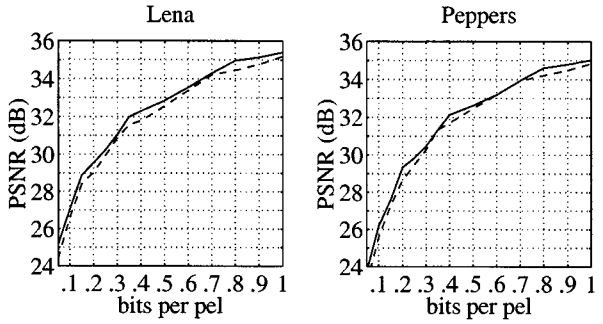


Fig. 3. Comparison of quantisers' performance
- - - - - - - real - imaginary based quantiser
————magnitude - phase based quantiser

3.2. VQ zerotree coding of complex coefficients

Several successful codecs use vector quantisers within a ZT-T framework. Because tree-structured VQ and multistage VQ offer progressive refinement of codewords they are natural choices, although the bitstreams they produce are not fully embedded and progressive. The most significant problem with VQ ZT-T systems is that they are very difficult to optimise in an R-D sense. Within a given significance level (and even between levels), bits are often spent refining vectors with little reduction in overall distortion when those bits would be better spent elsewhere.

With energy-normalised wavelet transforms, a few large magnitude coefficients possess much of the energy of the transformed data. These coefficients are difficult to code efficiently with a vector quantiser. (The space of possible vectors is too large, and the number of realised vectors in a given data set is too small.) We are developing a hybrid SQ/VQ SPIHT-like codec which combines the the system described in this paper with regular VQ coding. The smaller coefficients are gathered into multidimensional vectors and quantised with a tree-structured vector quantiser. The largest coefficients are quantised using the progressive refinement, R-D based system described in this paper.

4. CONCLUSIONS

Most ZT-T codecs deal with real-valued wavelet coefficients. Transforms that produce complex coefficients, such as the DT-CWT, can offer desirable properties, such as shift-invariance. However, extension of bit-plane based coding to complex coefficients is not straightforward. We developed an R-D optimal strategy for progressive bit-by-bit refinement of magnitude and phase values. By calculating expected distortion changes, the encoder and decoder can make the same decisions without the need for overhead bits. The decision to code a magnitude or phase refinement bit is determined by which type of bit maximises the expected distortion decrease. We are investigating more sophisticated magnitude pdf models, for instance pdfs conditioned on the magnitudes of a coefficients' neighbours and parent.

5. REFERENCES

[1] J.M. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients", *IEEE Transactions on Signal Processing*, Vol. 41, No. 12, pp. 3445–3462, December 1993.

[2] A. Said, W.A. Pearlman, "A New, Fast, and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 6, No. 3, pp. 243–249, June 1996.

[3] J.M. Lina, M. Mayrand, "Complex Daubechies wavelets", *J. of Applied and Computational Harmonic Analysis*, 2:2219–229, 1995.

[4] N.G. Kingsbury, "Shift invariant properties of the Dual-Tree Complex Wavelet Transform" *Proc. IEEE ICASSP 1999*, March 1999, paper SPTM 3.6.

[5] N.G. Kingsbury, "A Dual-Tree Complex Wavelet Transform with improved orthogonality and symmetry properties", *Proc. IEEE ICIP 2000*, September 2000, paper 1429.

[6] J. Li, S. Lei, "An Embedded Still Image Coder with Rate-Distortion Optimization", *IEEE Trans. Image Processing*, Vol. 8, No. 7, July 1999, pp. 913–924.

[7] A. Gersho, R.M. Gray, *Vector Quantization and Signal Compresssion*, Kluwer Academic Publishers, 1992.

IRREGULAR SAMPLING PROBLEMS AND SELECTIVE RECONSTRUCTIONS ASSOCIATED WITH MOTION TRANSFORMATIONS. *

Jean-Pierre Leduc

University of Maryland
Department of Mathematics
1301 Mathematics Bldg
College Park, MD 20742
Email: jleduc@math.umd.edu

ABSTRACT

This paper introduces the irregular sampling problem associated with motion transformations embedded in image sequences. Moving patterns in image sequences undergo a sampling which is function of the relative position of the object and the sampling grid. To solve this problem, it is effective to consider motion as a smooth invertible time-warping transformation. Important applications are related to this topic. Let us mention the focalization on selected moving areas characterized by a specific scale and a specific kinematic. Focalization and selective reconstruction can be performed either for analysis purpose with interpolation, prediction, and de-noising or for coding purpose with transmission of limited areas of interest. The Shannon sampling theorem and its generalizations as Kramer and Parzen theorems apply in this context with Clark's theorem. Clark's theorem shows that signals formed by warping band-limited signals admit formulae for reconstruction from samples. Furthermore, in this paper, the warping operators that lift the pattern up to a trajectory are chosen as unitary irreducible and square-integrable group representations. These operators bring important tools to motion-selective analysis and reconstruction, namely continuous wavelets, frames, discrete wavelet transforms, and reproducing kernel subspaces. In this paper, two examples are treated with motion at constant translational velocity and angular velocity. It is shown that the analysis and reconstruction structures directly derived from motion-based groups are equivalent to warping the same structures from the usual affine multidimensional group defined for space-time transformations.

Key Words: wavelets, motion detection, classification and reconstruction, signal and system modeling.

1. INTRODUCTION

In this paper, the motion transformations that occur in space-time signals like image sequences $\mathbb{R}^2 \times \mathbb{R}$ are characterized as a smooth (i.e. differentiable) warping of the spatio-temporal space which lifts a still signal into a moving signals on a trajectory. As the object is moving from one frame to the next, its sampling is irregular except in the particular case where the displacements correspond to an integer number of samples in the grid. The approach developed in this paper will refer to the Shannon sampling theorem and its generalization to Parzen's theorem for multidimensional signals and to Kramer's theorem as a general integral transform [1, 2, 3, 4]. Clark's theorem is also relevant. Indeed, Clark's theorem states that signals obtained by warping band-limited signals admit formulae for reconstruction from samples [5, 6].

In this paper, the warping models rely on the physical structure of motion which involves Lie algebras or Lie groups. The warping operators are in fact constructed as a Lie group representations i.e. operators in the Hilbert space $H = L^2(\mathbb{R}^2 \times \mathbb{R}, d\vec{k}d\omega)$ of the signals. $\vec{k}$ and ω stand respectively for the spatial and temporal frequency. This technique relates these operators to important analysis tools. One of these tools consist in multi-dimensional affine wavelets which correspond to the affine group of dilations in space and translations in space-time. The deformation of the affine group into a group of motion induces a warping of the continuous wavelets, frame and discrete wavelets. These warping transformations require to be generated by invertible operators, to compose one with the other and to preserve the band-limitedness of the still signals. How to build such warping operator? The answer stays in the following choice. The Lie group Representations [11, 12, 13] are Unitary Irreducible (UIR) operators in $H = L^2(\mathbb{R}^2 \times \mathbb{R}, d\vec{k}d\omega)$ defined from a group homomorphism i.e. a on-to-one mapping from the group element $g \in G$ to operator Π_g in the Hilbert space $H = L^2(\mathbb{R}^2 \times \mathbb{R}, d\vec{k}d\omega)$: $g \in G \to \Pi_g$ such that $\Pi_{g_1}\Pi_{g_2} = \Pi_{g_1 \circ g_2}$ and $\Pi_{g^{-1}} = T_g^{-1}$; then $\Pi_e = I_H$. Moreover, when this warping operator is square-integrable, it provides a strong structure for signal analysis, decomposition and reconstruction. This structure is made of Continuous Wavelet Transform (CWT) and Frames Operator (FO) along with Reproducing Kernel Spaces (RKS), Discrete Wavelet Transforms (DWT) or Orthonormal Bases (ONB), and in a weaker sense, Riesz bases. Square-integrable warping operators preserve the signal band-limitedness. Two examples of motion warping are considered in this paper. The first concerns translational velocity $\vec{v}$ (called Galilean transformation) [12, 13] and the second introduces the angular velocity θ_1 [11]. For the time-warpings defined as above, this paper shows that diagrams like the following commute.

$$
\begin{array}{ccc}
 & \Pi_{\{\vec{v}\}} & \\
L^2(\mathbb{R}^2 \times \mathbb{R}) & \longleftrightarrow & L^2_{\Pi}(\mathbb{R}^2 \times \mathbb{R}) \\
\downarrow & & \downarrow \\
\text{CWT} & & \text{CWT} \\
\downarrow & & \downarrow \\
\text{affine-wavelets} & \longleftrightarrow & \text{affine-Galilei wavelets} \\
\downarrow & & \downarrow \\
\text{frame dec.} & & \text{frame dec.} \\
\downarrow & & \downarrow \\
\text{affine frame} & \longleftrightarrow & \text{affine-Galilei frame} \\
\downarrow & & \downarrow \\
\text{DWT} & & \text{exact FO} \\
\downarrow & & \downarrow \\
\text{ONB} & \longrightarrow & \text{Galilei-Riesz basis}
\end{array}
$$

*This research work is supported by the AFOSR grant No. F49620-99-1-0068.

This diagram means that the analysis and reconstruction structures (CWT, FO, bases) directly derived from motion-based groups are equivalent to warping the same structures defined from the usual affine multidimensional group for space-time transformations. This scheme generalizes up to general forms of motion (deformational motion and motion on manifolds) as long as the warping satisfies unitary, irreducibility and square-integrability. This paper also shows how signals formed by the warping of band-limited signals admit different kinds of reconstruction formulas from samples. This theory extends beyond the limited scope of this presentation to consider singular self-adjoint boundary-value problems also known as Sturm-Liouville boundary-value problem related to generalized special functions characterizing motion transformations.

2. WARPING AND SAMPLING THEOREMS

This section first defines time-warping operators and proceeds to three related sampling theorems [5, 6]. The warping transformation is defined as a space-time mapping γ: $D = (\mathbb{R}^2 \times \mathbb{R}) \rightarrow D_\gamma = (\mathbb{R}^2 \times \mathbb{R})$. This mapping acts on band-limited functions f defined in image sequences. The appropriate space for these finite-energy functions is the Hilbert space $H = L^2(\mathbb{R}^2 \times \mathbb{R}, d\vec{x}dt)$. The composition $(f \circ \gamma)$ defines a warping operator $[\Pi_\gamma f](\mathbf{x}) = f(\gamma(t))$ with $f \in H$ and $\mathbf{x} \in (\mathbb{R}^2 \times \mathbb{R})$. In the following, the notation $\mathbf{x}$ stands for space-time variables, $\vec{x}$ for space vectors, and t for time i.e. $\mathbf{x} = (\vec{x}, t)$. $\mathbf{k}$ stands for $\mathbf{k} = (\vec{k}, \omega)$.

The *Shannon sampling theorem* has been generalized by Parzen [4] for multi-dimensional signals, then in space-time.
Theorem 1 If $\mathbf{I}$ is a bounded interval symmetric to origin defined as a the spatio-temporal frequency torus $\mathbf{I} = I_s^2 \times I_t$ and $f(\mathbf{x})$ is band-limited in $\mathbf{I}$ i.e. $\int_{\mathbf{I}} |f(\mathbf{k})|^2 d\mathbf{k} < \infty$, ten

$$f(\mathbf{x}) = \sum_{\mathbf{K}} f(\mathbf{x_K}) \prod_{i=1,2,3} \frac{sin[W_i(x_i - n_i T]}{W_i(x_i - n_i T)} \quad (1)$$

where $\mathbf{K} = (n_1, n_2, n_3)$, $n_i \in \mathbb{Z}$, $W_i = \frac{\pi}{\Pi_i}$, $I_s = [-W_i, +W_i]$, $i = 1, 2$, and $I_t = [-W_3, +W_3]$. Parzen theorem establishes that a band-limited function $f(\mathbf{k}) \in \mathbf{I}$ can be completely determined by giving its ordinates on a grid of points. *Clark's theorem* states the following.
Theorem 2 If D admits a sampling formula as

$$f(\mathbf{x}) = \sum_{\mathbf{K}} f(\mathbf{x_K}) \Psi_{\mathbf{K}}(\mathbf{x}) \quad (2)$$

then, D_γ admits a sampling formula for $h = f \circ \gamma$

$$h(\mathbf{x}) = \sum_{\mathbf{K}} h(\rho_{\mathbf{K}}) \Phi_{\mathbf{K}}(\mathbf{x}) \quad (3)$$

where $\rho_{\mathbf{K}} = \gamma^{-1}(\mathbf{x_K})$ and $\Phi_{\mathbf{K}} = \Psi_{\mathbf{K}} \circ \gamma$. When γ is an affine transformation, then the band-limitedness of $f(t)$ is preserved. When γ is not an affine transformation, we need additional conditions on the warping operator to preserve band-limitedness. In the following, time-warping operators are derived from unitary and irreducible square-integrable representations of groups for motion transformation.

The *Kramer's generalized sampling theorem* states:
Theorem 3 Let us suppose a bounded interval $\mathbf{I}$ defined as above, and the space $L^2(\mathbf{I}, d\mathbf{k})$ of functions $f(x)$ for which $\int_{\mathbf{I}} |f(\mathbf{k})|^2 d\mathbf{k} < \infty$ Let us further suppose the existence of a kernel $K(\mathbf{x}, \mathbf{k}) \in L^2(\mathbf{I}, d\mathbf{k})$ for all $\mathbf{x} \in \mathbb{R}^2 \times \mathbb{R} : [\mathbb{R}^2 \times \mathbb{R}] \times \mathbf{I} \rightarrow \mathbb{C}$. Then, the Kramer space associated with $\mathbf{I}$ and K consists of all the signals of the form

$$f(\mathbf{x}) = \int_{\mathbf{I}} K(\mathbf{x}, \mathbf{k})\ f(\mathbf{k}) d\mathbf{k} \quad (4)$$

where $f(\mathbf{k}) \in L^2(\mathbf{I}, d\mathbf{k})$. If there exists a countable set $E = \{\mathbf{x_n}\}$ such that $\{K(\mathbf{x}, \mathbf{k})\}$ is a complete orthogonal set on $L^2(\mathbf{I}, d\mathbf{k})$, then we have the following reconstruction formula

$$f(\mathbf{x}) = \lim_{\mathbf{n} \rightarrow \infty} \sum_{-\mathbf{n}}^{+\mathbf{n}} f(\mathbf{x_n})\ S_{\mathbf{n}}(\mathbf{x}) \quad (5)$$

where

$$S_{\mathbf{n}}(\mathbf{x}) = S(\mathbf{x_n}, \mathbf{x}) = \frac{\int_{\mathbf{I}} K(\mathbf{x}, \mathbf{k})\ \overline{K(\mathbf{x_n}, \mathbf{k})}\ d\mathbf{k}}{\int_{\mathbf{I}} |K(\mathbf{x_n}, \mathbf{k})|^2\ d\mathbf{k}} \quad (6)$$

If the kernel K is chosen as a Fourier kernel, Kramer's theorem retrieves Shannon's theorem. Let us proceed further on RKS.

A basis $\{\Psi_{\mathbf{n}}\}$ is a sampling basis for a Reproducing Kernel Hilbert Space (RKHS) H [6] with sampling set $\{x_{\mathbf{n}} \in D\}$ yields a reconstruction formula

$$f(\mathbf{x}) = \sum_{\mathbf{n}} f(x_{\mathbf{n}}) \Psi_{\mathbf{n}}(\mathbf{x}) \quad \forall\ f \in H \quad (7)$$

if and only if its bi-orthogonal basis $\{\Upsilon_{\mathbf{n}}\}$ is given by

$$\Upsilon_{\mathbf{n}}(\mathbf{x}) = \langle \Upsilon_{\mathbf{n}}\ ,\ \Psi_{\mathbf{n}} \rangle\ K(x_{\mathbf{n}}, \mathbf{x}) \quad (8)$$

where $K(x_{\mathbf{n}}, \mathbf{x})$ is a reproducing kernel for the functions $f \in H$. $\langle\ ,\ \rangle$ defines the inner product. A reproducing kernel K for H is such that $K : D \times D \rightarrow \mathbb{C}$ with $K(\mathbf{x}_1, \mathbf{x}_2) \in H$ for all $\mathbf{x}_1, \mathbf{x}_2 \in D$ and $f(\mathbf{x}_1) = \int_D K(\mathbf{x}_1, \mathbf{x}_2)\ f(\mathbf{x}_2) d\mathbf{x}_2$ for all $f \in H$.

3. CWT, DWT FOR MOTION PATHS

This section restarts from the definition of the CWT and warps this structure along motion transformations. The condition of square-integrability imposed on the group representations implies the existence of CWT and frames along with RKHS. A time-warping is applied in forms of a velocity-based transformation: it generates continuous Galilean wavelets, frames and RKHS. This defines new CWTs and frames along the path of a constant velocity transformation i.e. $\vec{x} = \vec{b}_0 + \vec{v}\tau$. Similar constructions apply for other kinds of motion like rotation at constant the angular velocity.

Let us recall the definition of the Continuous Wavelet Transform (CWT). Let us denote $S(\vec{x}, t)$ the signal in the Hilbert space $L^2(\mathbb{R}^n \times \mathbb{R}, d^n\vec{x}dt)$. The CWT $[W_\Psi S](g)$ is defined as a linear map $W_\Psi : L^2(\mathbb{R}^n \times \mathbb{R}, d^n\vec{x}dt) \rightarrow L^2(G, dg)$

$$[W_\Psi S](g) = < \widehat{\Psi}_g\ ,\ \widehat{S} > = \int_{\mathbb{R}^n \times \mathbb{R}} d\vec{k}d\omega\ \widehat{\Psi}_g(\vec{k}, \omega) \overline{\widehat{S}}(\vec{k}, \omega) \quad (9)$$

The overbar $\bar{}$ and $\widehat{}$ symbols denote complex conjugate and Fourier domain. As a CWT, this linear map (9) is an inner product endowed with more properties than an usual cross-correlation function since it enables perfect reconstruction from the inverse CWT. The CWT is in fact an isometry from the space of observation $H = L^2(\mathbb{R}^2 \times \mathbb{R}, d\vec{x}dt)$ to a subspace $H_\eta \subset L^2(G, dg)$. The space H_η is a space of

complex-valued functions on G; it is is a reproducing kernel space. This means the existence of a reproducing kernel i.e. the autocorrelation of Ψ. The reproducing kernel K is such that $G \times G \rightarrow \mathbb{C}$ with $K(g_1, g_2) = \langle \Psi_{g_1}, \Psi_{g_2} \rangle$ ansd $f(g_1) = \langle K(g_1, g_2), f(g_2) \rangle$ for all $g \in G$. The link between reproducing kernel spaces and sampling theorem has already been defined; it will be warped.

The n-dimensional spatio-temporal affine group [12], denoted here G_1), is an ordered 4-tuple of elements $g = \{\vec{b}, \tau, a, R\}$ where the parameters $\vec{b} \in \mathbb{R}^n$, $\tau \in \mathbb{R}$, $a \in \mathbb{R}^+ \backslash \{0\}$ and $R \in SO(n)$ stand respectively for spatial translation (the Cartesian position), temporal translation, dilation (the scale) and rotation in n-dimensional space (the angular orientation). The group elements in G_1 have the following matrix representation

$$g = \begin{bmatrix} a\,R(\Lambda) & \vec{v} & \vec{b} \\ 0 \quad 0 & \tau \\ 0 \quad 0 & 1 \end{bmatrix} \qquad \Lambda \in [0, 2\pi) \qquad \vec{v} = \vec{0} \tag{10}$$

This group is a subgroup of $GL(n+2, \mathbb{R})$. The UIRs of G_1 are eventually given as operators Π_g: $\widehat{\Psi}(\vec{k}, \omega) \rightarrow [\Pi_g \widehat{\Psi}](\vec{k}, \omega)$ as

$$[\Pi_g \widehat{\Psi}](\vec{k}, \omega) = a^{\frac{n}{2}}\, e^{i(\vec{b}\cdot\vec{k} + \omega\tau)}\, \widehat{\Psi}(\vec{k}\,', \omega) \tag{11}$$

with

$$\vec{k}\,' = aR^{-1}\vec{k} \tag{12}$$

From now on, we consider $n = 2$.

Theorem 4 The UIRs of G_1 in the space $L^2(\mathbb{R}^2 \times \mathbb{R}, d^2\vec{k}d\omega)$ are square-integrable. The condition of square-integrability requires that $\widehat{\Psi} \in L^2(\mathbb{R}^n, d^n\vec{x})$ be such that

$$\int_{\mathbb{R}^2 \times \mathbb{R}} |\widehat{\Psi}(\vec{\xi}, \eta)|^2 \, \frac{d^2\vec{\xi}d\eta}{|\vec{\xi}|^2} = C_\psi < +\infty \tag{13}$$

and the representation $\Pi_g \widehat{\Psi}$ is bounded for all g. The variable $\vec{\xi} \in \mathbb{R}^n$ is a Fourier variable. Let us warp the group G_1 with a warping parameter $\vec{v} \in \mathbb{R}^n$. This deformation defines a group G_2 called the Galilei group [13, 12] made of ordered 5-tuple of elements $g = \{\vec{b}, \tau, \vec{v}, a, R\}$. The parameter $\vec{v} \in \mathbb{R}^n$ is in fact the velocity. This group is still a subgroup of $GL(n+2, \mathbb{R})$ but the UIRs read now

$$[\Pi_g \widehat{\Psi}_{m_0}](\vec{k}, \omega) = a^{\frac{n}{2}}\, e^{i(\vec{b}\cdot\vec{k} + \tau\omega)} \widehat{\Psi}(\vec{k}\,', \omega') \tag{14}$$

with

$$\begin{aligned} \vec{k}\,' &= aR^{-1}(\vec{k} + m_0\vec{v}) \\ \omega' &= (\omega - \tfrac{m_0||\vec{v}||^2}{2} - \vec{v}\cdot\vec{k}) \end{aligned} \tag{15}$$

When $\vec{v}$ tends to $\vec{0}$, these UIRs tend to Equation (11).

Theorem 5 The UIRs of G_2 in the space $L^2(\mathbb{R}^2 \times \mathbb{R}, d^2\vec{k}d\omega)$ are square-integrable. The condition of square-integrability requires that the following integral be finite

$$c_\Psi = \int_{\mathbb{R}^2 \times \mathbb{R}} |\widehat{\Psi}_{m_0}(\vec{k}, \omega)|^2 \, I_{m_0}(\vec{k}, \omega)\, d\vec{k}\, d\omega < \infty \tag{16}$$

where $I_{m_0}(\vec{k}, \omega)$ is equal to

$$\int_{\mathbb{R}^2 \times \mathbb{R}} |\widehat{\Psi}_{m_0}(\vec{k}\,', \omega')|^2 \left[\frac{|\vec{k}|^2 + 2m_0(\omega - \omega')}{|\vec{k}|^4\, m_0}\right] d\vec{k}\,'\, d\omega' \tag{17}$$

It is clear that for $m_0 \neq 0$, $\vec{v} = \vec{0}$, and $\omega = \omega'$ the condition of admissibility for G_2 (17) is equivalent to G_1 in (13).

Let us apply a second warping with Λ as a warping parameter $\Lambda = [\theta_0 + \theta_1\tau] \mathrm{mod}\ 2\pi$ with $\theta_0 \in [0, 2\pi)$ and $\theta_1 \in \mathbb{R}$. θ_1 is the angular velocity [11]. We have defined a new group G_3 composed of ordered 6-tuple of elements $g = \{\vec{b}, \tau, \vec{v}, \theta_0, \theta_1, a\}$. The UIRs of G_3 in $L^2(\mathbb{R}^2 \times \mathbb{R}, d\vec{k}d\omega)$ are expressed as

$$\left[T(g)\widehat{\Psi}_{m_0}\right](\vec{k}, \omega) = a^{n/2}\, e^{i[R^{-1}[\theta_1\tau]\vec{b}\cdot\vec{k} + \omega\tau]}\, \widehat{\Psi}_{m_0}\left(\vec{k}\,', \omega'\right) \tag{18}$$

where $\vec{k}\,'$ and ω' are as in Equation 15. The character $e^{i[R(\theta_1\tau)\vec{b}\cdot\vec{k} + \omega\tau]}$ of the UIRs in Equation 18 introduces a special function derived by integration on τ. This yields with $\Omega = \frac{\omega}{\theta_1}$, and polar coordinates $\vec{k} = (k, \alpha)$, $\vec{b} = (r, \beta)$

$$J_\Omega(kr) = \frac{1}{2\pi} \int_0^{2\pi} e^{i[\Omega u + kr\sin u]} du \tag{19}$$

which is not a Bessel function except for $\Omega \in \mathbb{Z}$.

Let us recall that the definition of a frame. A sequence of functions $\{\phi_j\}$ in a separable Hilbert space is called a frame if there exist two constants $A, B > 0$ and $B < \infty$ so that, for all $f \in H$, we have

$$A\, ||f||^2 \leq \sum_j |\langle f\ ,\ \phi_j \rangle|^2 \leq B\, ||f||^2 \tag{20}$$

where the sequence of functions $\{\phi_j\}$ is computed on a discrete lattice j derived from discretizing the group parameters $\vec{b}$, τ, a, θ_0, $\vec{v}$. Square-integrable UIRs imply the existence of associated frames. Such frames have an associated invertible bounded operator $F : H \rightarrow H$; and $F(f) = \sum_n \langle f, \phi_n \rangle \phi_n$. This frame allows a perfect reconstruction and a sampling theorem

$$f(\mathbf{x}) = \sum_{\mathbf{n}} \langle f\ ,\ F^{-1}(\phi_{\mathbf{n}}) \rangle\, \phi_{\mathbf{n}}(\mathbf{x}) = \sum_{\mathbf{n}} \langle f\ ,\ \phi_{\mathbf{n}} \rangle\, F^{-1}[\phi_{\mathbf{n}}(\mathbf{x})] \tag{21}$$

where F^{-1} is the inverse or dual frame operator for F.

Discrete wavelet transforms are well defined as dyadic Multi-Resolution Analysis (MRA) wavelets Ψ [7]. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be the continuous scaling function. V_s is a RKHS with $K_s(x_1, x_2) = 2^s \sum_n \phi(2^s x_1 - n)\phi(2^s x_2 - n)$. The basis $\{\Psi(x - n)\}$ of V_0 is bi-orthogonal to $\{K_0(x, n)\}$.

4. SAMPLING THEOREM FOR MOTION TRANSFORMATIONS

This section shows that the UIRs deduced in Equations (14) and (18) lead naturally to a Kramer theorem for motion transformations. Let $\Psi \in L^2(\mathbf{I}, d\mathbf{k})$ and I be the spatio-temporal frequency torus $\mathbf{I} = I_s^2 \times I_t$, and $h(t) = \Psi \circ \Pi_g$. If we integrate these UIRs (14, 18) on the spatio-temporal frequency torus I, we get in each case a generalized Fourier transform of the form (a=1)

$$h(\mathbf{x}) = \frac{1}{(2\pi)^3} \int_{\mathbf{I}} e^{i\gamma[\mathbf{x}]\cdot\mathbf{k}}\, \widehat{\Psi}(\mathbf{k})\, d\mathbf{k} \tag{22}$$

which can be restated as a Kramer's sampling theorem

$$h(\mathbf{x}) = \int_{\mathbf{I}} K(\mathbf{x}, \mathbf{k})\, \widehat{\Psi}(\mathbf{k})\, d\vec{\mathbf{k}} \tag{23}$$

Therefore, motion-based warped spaces admit a reconstruction formula with $K(\mathbf{x}, \mathbf{k}) = \frac{e^{i\gamma[\mathbf{x}]\cdot\mathbf{k}}}{(2\pi)^3}$ and $\vec{\mathbf{x}}_{\mathbf{n}} = \gamma^{-1}(\mathbf{n})$ where the basis $\{K(\mathbf{n}, \vec{k})\} = e^{i\mathbf{n}\cdot\mathbf{k}}$ is complete on the spatio-temporal torus I. The interpolating functions $s_{\mathbf{n}}$ can be derived from Kramer's theorem.

For the Galilei group of constant velocity, the transformation $\gamma[\mathbf{x}]$ is a matrix $A\mathbf{x}$ where A is of the form

$$A = \begin{pmatrix} I_n & \vec{v} \\ \vec{0}^T & 1 \end{pmatrix} \quad \text{with} \quad \mathbf{k}' = A^T\mathbf{k} \tag{24}$$

where I_n is the $n \times n$ unit matrix. For the rotational motion, Equation (19) leads to an integral transform of the form

$$[H_\Omega f](k) = \int_0^\infty f(r)\, J_\Omega(k\, r)\, r dr \tag{25}$$

on which Kramer's theorem applies on the zeroes of the Bessel functions defined for $\Omega = \frac{\omega}{\theta_1} \in \mathbb{Z}$ i.e. when ω is a multiple of the angular velocity. To avoid any aliasing, we need $\theta_1 < 1$. Hence, for signal processing $\Omega = -1, 0, 1$ do only matter. From Section 2, if $\Omega = m \in \mathbb{Z}$, then Kramer's theorem has the sampling function

$$S_{m,n}(k) = \frac{\int_0^\infty f(r)\, J_m(k\, r)\, \overline{J_m(k_{m,n}\, r)}\, r dr}{\int_0^\infty f(r)\, |J_m(k\, r)|^2\, r dr} \tag{26}$$

$$= \frac{2\, k_{m,n}\, J_\Omega(k)}{(k_{m,n}^2 - k^2) J_{m+1}(k_{m,n})} \tag{27}$$

where $k_{m,n}$ are the zeroes of the Bessel function J_m i.e. $J_m(k_{m,n}) = 0$ and usual properties of Bessel functions have enabled the evaluation of the integrals in (26).

5. WARPING CWT, FRAMES AND DWT

In Section 3, an invertible motion-based warping transformation has been constructed from group G_1 to G_2 and then to G_3. In fact, we have much more properties to state on the CWT, frames and DWT.

Unitary irreducible and square-integrable warping preserves the inner product and then the CWT. Indeed, if $s, \Psi \in H$, $f_\gamma = \Pi_\gamma(f)$ and $\Psi_\gamma = \Pi_\gamma(\Psi_{g_1}) = \Psi_{g_i}; i = 2, 3$ the inner product is preserved by the warping such that

$$\langle s_\gamma, \Psi_\gamma \rangle = \langle s, \Psi_{g_1} \rangle \tag{28}$$

This means that the motion-based CWT computed on a moving signal (rigid pattern) is equivalent to computing the multi-dimensional affine CWT on the still/frozen version of the same signal (pattern).

The warping of the multi-dimensional affine frame computed from the UIRs (11) gives rise to the same frames as computed directly from the CWT of the correspond motion-based UIRs i.e. to motion-compensated frame, and convolutional filters. Motion-compensated structures are defined as structures applied on the assumed trajectory of motion. For the frame operator, the same conclusions as for the CWT apply. If $E = \{\phi_j\}$ is a frame for G_1, then $E_\gamma = \{\Pi(\phi_j)\}$ is a frame for any G_i with rescaled bounds. Since the inner product is preserved, the reconstruction process delivers a still version of the moving signal (pattern).

The warping of the multi-dimensional affine DWT and its MRA do NOT give rise in whole generality to DWTs (or ONBs) on the corresponding motion group but instead gives Riesz bases or exact frames. To have DWT in the Galilei group, we need velocity vectors whose components provide integer translations. For integer velocities and discrete group parameters g_* as defined in [13], the Galilean case mimics the affine group as follows. Let $a_* = 2$ and $\Pi_{g_*}\Psi(\vec{x}, t) = a_*^{-m/2}\Psi\left(a_*^{-m}\vec{x} - n_b\vec{b}_* - n_v\vec{v}_*(t - n_\tau\tau_*)\ ,\ t - n_*\tau_*\right)$ where we retrieve the ONBs $\Psi_{m,\vec{p},q}(\vec{x}, t) = 2^{-m/2}\Psi\left(2^{-m}\vec{x} - \vec{p}, t - q\right)$ in $L^2(\mathbb{R}^2 \times \mathbb{R})$ at $\vec{p} = n_b\vec{b}_* + n_v\vec{v}_*n_\tau\tau_*$, $q = n_\tau\tau_*$ with $\vec{p} \in \mathbb{Z}^2$, and $q \in \mathbb{Z}$.

6. CONCLUSIONS

This paper has shed some introductory light on the irregular sampling problem associated with motion embedded in image sequences. Related sampling theorems have been stated as selective reconstruction formulae. This theory extends to more general statements involving involving deformational motion, motion on manifolds, Sturm-Liouville boundary problems and special functions out of which interesting and practical derivations will be presented. The applications of motion-compensated wavelets in de-noising, interpolating, coding and transmitting digital image sequences have been presented as efficient schemes in [8, 10, 11, 12, 13]. Results on motion-selective reconstructions from digital image sequences will also be presented to demonstrate the effectiveness of this approach.

REFERENCES

[1.] A. Jerri. "The Shannon Sampling Theorem - Its Various Extensions and Applications: A tutorial Review", *Proceedings of the IEEE, Vol. 65, No. 11, November 1977, pp. 1565-1596.*

[2.] H. Feichtinger and K. Gröchenig. "Theory and Practice of Irregular Sampling", *Wavelets: Mathematics and Applications, Stud. Adv. Math., CRC, Boca Raton, pp. 305-363, 1994.*

[3.] J. Benedetto. "Irregular Sampling and Frames", *Wavelets: A Tutorial in Theory and Applications, C. Chui Editor, Boston: Academic Press, pp. 445-508, 1992.*

[4.] A. Zayed. "Advances in Shannon's Sampling Theory", *CRC Press, Boca Raton, 1993.*

[5.] Y. Zeevi and E. Shlomot. "Nonuniform Sampling and Antialiasing in Image Representation", *IEEE Transactions on Signal Processing, Vol. 41, No. 3, March 1993, pp. 1223-1236.*

[6.] S. Azizi and D. Cochran. "Reproducing Kernel Structure and Sampling on Time-Warped Kramer Spaces", *Proceedings of ICASSP-99, Phoenix, Vol.3, May 15-19 1999, pp. 1649-1652.*

[7.] G. Walter. "A Sampling Theorem for Wavelet Subspaces", *IEEE Transactions on Information Theory, Vol. 38, No. 2, part 2, pp. 881-884, 1992.*

[8.] E. Dubois. "Motion-Compensated Filtering of Time-Varying Images", *Multidimensional Systems and Signal Processing, Vol. 3, pp. 211-239, 1992.*

[9.] M. Nashed and G. Walter. "General Sampling Theorems for Functions in Reproducing Kernel Hilbert Spaces", *Mathematics of Control, Signals, and Systems, Vol. 4, pp. 363-390, 1991*

[10.] J.-P. Leduc, J.-M. Odobez and C. Labit. "Adaptive Motion-Compensated Wavelet Filtering for Image Sequence Coding", *IEEE Transactions on Image processing, Vol. 6, No. 6, pp. 862-878, June 1997.*

[11.] M. Kong, J.-P. Leduc, B. Ghosh, J. Corbett, V. Wickerhauser. "Wavelet based Analysis of Rotational Motion in Digital Image Sequences", *Proceedings of ICASSP-98, Seattle, May 12-15, 1998, pp. 2781-2784.*

[12.] J.-P. Leduc, F. Mujica, R. Murenzi, and M. Smith. "Spatio-Temporal Wavelets: a Group-Theoretic Construction for Motion Estimation and Tracking", to appear in SIAM Journal of Applied Mathematics in April 2001, electronic version available on SIAM server.

[13.] J.-P. Leduc. "Spatio-Temporal Wavelet Transforms for Digital Signal Analysis", Signal Processing, Elsevier, Vol. 60 (1), pp. 23-41, July 1997.

NONLINEAR PERCEPTUAL AUDIO FILTERING USING SUPPORT VECTOR MACHINES

Simon I. Hill, Patrick J. Wolfe, and Peter J. W. Rayner*

Signal Processing Group, University of Cambridge
Department of Engineering, Trumpington Street
CB2 1PZ, Cambridge, UK
{sih22, pjw47, pjwr}@eng.cam.ac.uk
http://www-sigproc.eng.cam.ac.uk

ABSTRACT

In this paper, the perceptually based loss functions for audio filtering used by Wolfe and Godsill [1] are shown to fit well within a complex-valued Support Vector Machine (SVM) framework. SVM regression is extended to estimation of complex-valued functions, including the derivation of a variant of the Sequential Minimal Optimisation (SMO) algorithm. Audio filters are derived using this based on an autoregressive (AR) model used for audio and two different Hermitian kernel functions. Results are found to be promising, and further improvements are discussed.

1. INTRODUCTION

Recent attempts to design audio filters based on perceptual considerations (see, for example, [1–3]), have in general assumed independence between Discrete Fourier Transform (DFT) components. Indeed, this simplifying assumption has also been made in standard approaches to audio signal enhancement [4].

While mathematically convenient, such an assumption is unrealistic. This can be shown by considering a common standard speech model: the autoregression (AR). Assume that the signal of interest is generated by

$$y_n = \sum_{p=1}^{P} a_p y_{n-p} + e_n, \tag{1}$$

where $e_n \sim \mathcal{N}(0, \sigma_e^2)$. With reference to Box and Jenkins [5] and Hopgood [6],

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_N \Lambda \mathbf{W}_N^T), \tag{2}$$

where $\mathbf{Y}$ is an N-length vector of DFT elements, Λ is related to σ_e^2 and is generally *not* diagonal, and the AR coefficients $\mathbf{a}$, and $\mathbf{W}_N$ is a matrix with elements $\mathbf{W}_N(k+1, n+1) = W_N^{kn} = e^{-j\frac{2\pi kn}{N}}$, $k, n \in \{0, \ldots, N-1\}$.

Wolfe and Godsill [1] consider a loss function based on masked thresholds, ϵ_k, below which additive noise is assumed to be imperceptible to human listeners:

$$C(\hat{Y}_k, Y_k) = \begin{cases} 0 & \left|\,|\hat{Y}_k| - |Y_k|\,\right| < \epsilon_k \\ \left(|\hat{Y}_k| - |Y_k|\right)^2 - \epsilon_k^2 & \text{otherwise.} \end{cases} \tag{3}$$

*Material by the second author is based upon work supported under a U.S. National Science Foundation Graduate Fellowship.

This results in an estimation of the magnitude of the DFT components; in the absence of a quantitative perceptual motivation the observed phase is retained. Masked thresholds are calculated at each frequency bin for a given short-time block via the masking model proposed in [7], which takes into account both simultaneous masking and absolute hearing thresholds, and has been used in other recent perceptually motivated noise reduction systems [2,3].

This paper uses the perceptually based loss function of (3) in a Support Vector Machine (SVM) framework, as described in Section 2. A variant of the Sequential Minimal Optimisation (SMO) algorithm for the complex problem is presented in Section 3. Experimental results are presented in Section 4.

2. SVM FRAMEWORK

As demonstrated in (2), the ubiquitous AR model of audio is incompatible with the assumption of independence of frequency components in the DFT. Intuitively, and from informal observations, it appears reasonable to favour the idea that there is some correlation. This being the case, it seems logical to consider some form of kernel-based regression in order to capture this correlation while at the same time allowing the freedom of nonlinear estimation. The most readily identifiable problem with this approach is the introduction of the underlying assumption that the audio statistics remain constant over time, which, in itself, is not necessarily accurate. This reservation is, of course, a standard one for such audio filtering and should be borne in mind when considering the final results.

2.1. Kernel based Regression

Consider the problem of estimating a latent function relating some input, $\mathbf{x}$, with a corresponding output, y,

$$y = f(\mathbf{x}),$$

using some training data, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Assume that the output data is drawn from a probability density function

$$y|f(\cdot), \mathbf{x} \sim p_{Y|F,\mathbf{X}}\left(y|f(\cdot), \mathbf{x}\right).$$

Using this a posterior probability density function for $f(\cdot)$ can be found,

$$p_{F|Y,\mathbf{X}}\left(f(\cdot)|y, \mathbf{x}\right) \propto p_{Y|F,\mathbf{X}}\left(y|f(\cdot), \mathbf{x}\right) p_F\left(f(\cdot)\right), \tag{4}$$

where $p_F(f(\cdot))$ is the prior probability density function. It is the aim of many authors [8–10] to relate this formulation to ideas of regularization, also known as stabilization, or prior smoothing. In the case at hand, $p_F(f(\cdot))$ would be related to the result in (2).

Taking negative logs of (4) yields

$$H[f(\cdot)] = V[\mathbf{x}, y; f(\cdot)] + \Omega[f(\cdot)], \tag{5}$$

where $V[\mathbf{x}, y; f(\cdot))]$ can be interpreted as some error cost function or loss function, used to measure the interpolation error and $\Omega[f(\cdot)]$ as a smoothness functional, stabilizer, or regularization term. A standard situation is one which sees

$$y_i = f(\mathbf{x}_i) + v_i,$$

where $v_i \sim \mathcal{N}(0, \sigma^2)$. In this case,

$$V[\mathbf{x}, y; f(\cdot)] \propto \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2.$$

In any case for which this has, in combination with $\Omega[f(\cdot)]$, a single minimum, finding $f(\cdot)$ at this minimum is equivalent to finding the maximum *a-posteriori* (MAP) solution for $f(\cdot)$.

The regularization term in kernel-based regression is proportional to the square of the norm of the projection of the function into some Reproducing Kernel Hilbert space (RKHS) $\mathcal{H}_K$, a subspace of some Hilbert Space, $\mathcal{H}$, in which $f(\cdot)$ is taken to exist,

$$\Omega[f(\cdot)] = \lambda \|P_K f(\cdot)\|^2_{\mathcal{H}_K}.$$

It is beyond the scope of this summary to go into detail; related references are [9, 10], among others.

Traditional, real-valued support vector regression is a subset of kernel-based regression in which the Hilbert space containing the function is constrained to the span of some kernel, $K(\cdot, \cdot)$, and a constant. This has the result that the functional approximation takes the form

$$f(\cdot) = \sum_{i=1}^{N} \beta_i K(\cdot, \mathbf{x}_i) + b. \tag{6}$$

Appropriate kernels which satisfy Mercer's conditions (see, for example, Smola and Schölkopf [9]), can also be considered to be the inner product of some mapping to a feature space $\mathcal{F}$, $\Phi(\mathbf{x})$ that is, $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{F}}$. From this perspective, there exists some vector, $\mathbf{w}$, in the feature space such that

$$\sum_{i=1}^{N} \beta_i K(\mathbf{x}_i, \cdot) = \langle \mathbf{w}, \Phi(\cdot) \rangle_{\mathcal{F}},$$

and it can be shown that,

$$\Omega[f(\cdot)] = \lambda \|\mathbf{w}\|^2_{\mathcal{F}}.$$

Typically, when using an SVM approach for regression, a threshold-based loss function is employed. This has the advantage that the vector $\boldsymbol{\beta}$ may well be sparse, thereby reducing computational requirements.

2.2. Complex-Valued SVM

Returning to the original problem, with the consideration of correlation between the frequency components, a more applicable starting point (in that it includes phase considerations) is

$$\begin{array}{ll} \text{Minimise} & \sum_k \eta_k^2 + \lambda \|\mathbf{w}\|^2 \\ \text{Subject to} & |f(\mathbf{X}_k) - Y_k| \leq \epsilon_k + \eta_k. \end{array} \tag{7}$$

This constraint, although appearing to be linear, is in fact quadratic, making the problem much harder to solve. Linear constraints are preferable since the problem then reduces to that of quadratic programming; separating the above constraints such that there is individual consideration of the real and imaginary components can reduce the problem to the more desirable linear form.

Before doing this, consider the representation in the complex feature space where

$$\begin{aligned} \langle \mathbf{w}, \Phi(\mathbf{X}_k) \rangle_{\mathcal{F}} &= \mathbf{w}^H \Phi(\mathbf{X}_k) \\ &= \mathbf{w}_R^T \Phi_R(\mathbf{X}_k) + \mathbf{w}_I^T \Phi_I(\mathbf{X}_k) \\ &\quad + j \left(\mathbf{w}_R^T \Phi_I(\mathbf{X}_k) - \mathbf{w}_I^T \Phi_R(\mathbf{X}_k) \right), \end{aligned}$$

with subscripts R and I denoting real and imaginary components respectively. In this paper, looking to the simpler case of linear loss outside the threshold, the problem in equation (7) can be then rewritten as

$$\begin{array}{ll} \text{Minimise} & \frac{1}{\lambda} \sum_k (\eta_k + \hat{\eta}_k + \eta_k^* + \hat{\eta}_k^*) + \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{Subject to} & \left\{ \begin{array}{l} Y_{R,k} - \mathbf{w}_R^T \Phi_R(\mathbf{X}_k) - \mathbf{w}_I^T \Phi_I(\mathbf{X}_k) - b_R \\ \leq \epsilon_{R,k} + \eta_k \\ \mathbf{w}_R^T \Phi_R(\mathbf{X}_k) + \mathbf{w}_I^T \Phi_I(\mathbf{X}_k) + b_R - Y_{R,k} \\ \leq \epsilon_{R,k} + \hat{\eta}_k \\ Y_{I,k} - \mathbf{w}_R^T \Phi_I(\mathbf{X}_k) + \mathbf{w}_I^T \Phi_R(\mathbf{X}_k) - b_I \\ \leq \epsilon_{I,k} + \eta_k^* \\ \mathbf{w}_R^T \Phi_I(\mathbf{X}_k) - \mathbf{w}_I^T \Phi_R(\mathbf{X}_k) + b_I - Y_{I,k} \\ \leq \epsilon_{I,k} + \hat{\eta}_k^* \\ \eta_k, \hat{\eta}_k, \eta_k^*, \hat{\eta}_k^* \geq 0. \end{array} \right. \end{array}$$

This can be used to form the Lagrangian,

$$\begin{aligned} L \;=\; & \frac{1}{\lambda} \sum_k (\eta_k + \hat{\eta}_k + \eta_k^* + \hat{\eta}_k^*) + \frac{1}{2} \mathbf{w}^H \mathbf{w} \\ & - \sum_k \Big[\alpha_k (\mathbf{w}_R^T \Phi_R(\mathbf{X}_k) + \mathbf{w}_I^T \Phi_I(\mathbf{X}_k) + b_R - Y_{R,k} \\ & + \epsilon_{R,k} + \eta_k) + \hat{\alpha}_k (Y_{R,k} - \mathbf{w}_R^T \Phi_R(\mathbf{X}_k) - \mathbf{w}_I^T \Phi_I(\mathbf{X}_k) \\ & - b_R + \epsilon_{R,k} + \hat{\eta}_k) + \alpha_k^* (\mathbf{w}_R^T \Phi_I(\mathbf{X}_k) - \mathbf{w}_I^T \Phi_R(\mathbf{X}_k) \\ & + b_I - Y_{I,k} + \epsilon_{I,k} + \eta_k^*) + \hat{\alpha}_k^* (Y_{I,k} - \mathbf{w}_R^T \Phi_I(\mathbf{X}_k) \\ & + \mathbf{w}_I^T \Phi_R(\mathbf{X}_k) - b_I + \epsilon_{I,k} + \hat{\eta}_k^*) \Big] \\ & - \sum_k \left[r_k \eta_k + \hat{r}_k \hat{\eta}_k + r_k^* \eta_k^* + \hat{r}_k^* \hat{\eta}_k^* \right], \end{aligned}$$

where $\{\alpha_k, \hat{\alpha}_k, \alpha_k^*, \hat{\alpha}_k^*\}$ and $\{r_k, \hat{r}_k, r_k^*, \hat{r}_k^*\}$ are Kuhn-Tucker multipliers. Applying an extension of the standard SVM steps leads to a dual expression to maximise,

$$\begin{aligned} L \;=\; & \sum_k \left([Y_{R,k} \beta_{R,k} - Y_{I,k} \beta_{I,k}] \right. \\ & \left. - [\epsilon_{R,k} |\beta_{R,k}| + \epsilon_{I,k} |\beta_{I,k}|] \right) - \frac{1}{2} \boldsymbol{\beta}^H \mathbf{K} \boldsymbol{\beta}, \end{aligned}$$

where $\beta_k = (\alpha_k - \hat{\alpha}_k) - j(\alpha_k^* - \hat{\alpha}_k^*)$, $\sum_k \beta_k = 0$ and $\mathbf{K}$ is the matrix with (i,j)th entry $K(\mathbf{x}_i, \mathbf{x}_j)$. With this formulation,

$$f(\cdot) = \sum_k \tilde{\beta}_k K(\cdot, \mathbf{X}_k) + b,$$

where $\tilde{\beta}_k$ is the complex conjugate of β_k.

3. COMPLEX SMO

The Sequential Minimal Optimisation (SMO) algorithm was developed for SVM classification by Platt [11], and presented for regression by Flake and Lawrence [12]. The central idea of the SMO algorithm is that, when the Lagrangian is maximised with respect to two points (generically labelled β_1 and β_2), the maximisation becomes analytically tractable.

The complex-valued problem can be similarly structured and effectively decomposes into real and imaginary cases: an overview of it will be given here. As for the studied cases the two points are constrained to add to a constant. First the new value β_2 is found and, if outside the possible region, it is clipped to the nearest extremum. With this result the corresponding value for β_1 is found by subtracting the new β_2 from the old sum.

The unclipped updates for the real and imaginary points are,

$$\beta_{R,2}^{new} = \beta_{R,2}^{old} + \frac{E_{R,1} - E_{R,2} - (\epsilon_{R,2}\mathrm{sgn}(\beta_{R,2}) - \epsilon_{R,1}\mathrm{sgn}(\beta_{R,1}))}{K_{11} + K_{22} - K_{12} - K_{21}} \quad (8)$$

$$\beta_{I,2}^{new} = \beta_{I,2}^{old} + \frac{E_{I,2} - E_{I,1} - (\epsilon_{I,2}\mathrm{sgn}(\beta_{I,2}) - \epsilon_{I,1}\mathrm{sgn}(\beta_{I,1}))}{K_{11} + K_{22} - K_{12} - K_{21}}, \quad (9)$$

where $E_j = f(\mathbf{X}_j) - Y_j$. Note that both terms are real and, importantly, that $K_{12} = \tilde{K}_{21}$.

4. RESULTS AND DISCUSSION

Preliminary trials have been conducted using both linear and Gaussian kernels, with promising results. In these experiments a 30-second male voice recording was used for training as well as estimation, keeping the sound source consistent. Of this approximately six seconds were used as training data, in the form of 200 DFTs of time interval length 512, with an overlap in time of 50%. The speech signal was degraded artificially with additive white Gaussian noise to yield a signal-to-noise ratio of 15 dB; audio examples typical of results obtained are available at `http://www-sigproc.eng.cam.ac.uk/~sih22`.

Training was conducted for each frequency bin; in every case the 129 nearest noisy frequency bins were used as input to the filter. For each instance a corresponding set of 200 β-values were found. The final filters were then implemented on all data and the original time signal reconstructed.

Initial results indicate that, graphically and on a local basis, the algorithm is performing as desired. This is illustrated by a representative example in Figure 1. Here it is clear that, in the main, the filter output lies with in the threshold regions, as intended. Frequency bins have been chosen to illustrate different phenomena of the same DFT, and as such, note should be taken of the amplitude scales.

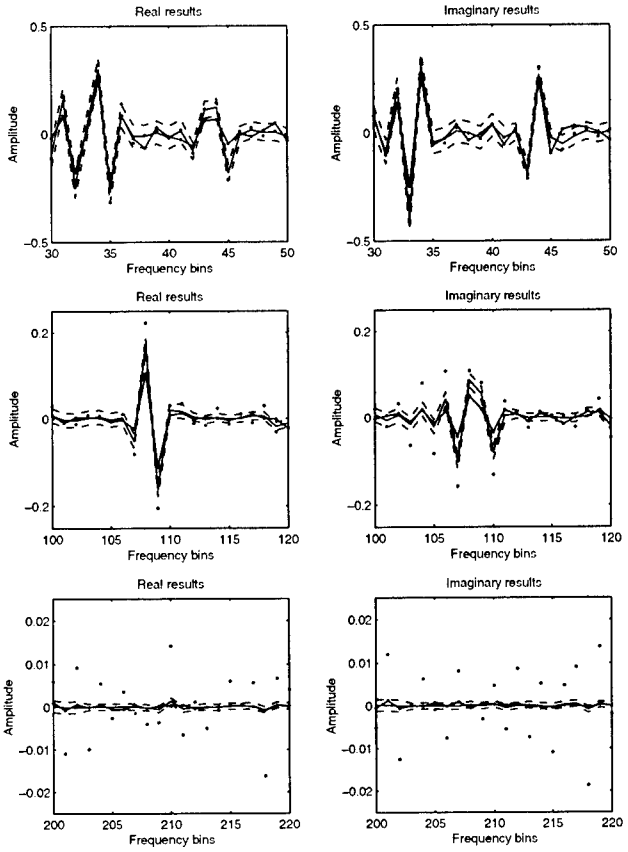


Fig. 1. Successful filtering in the frequency domain. *The original signal is shown as a continuous line, the thresholds as dashed lines, the filter output as a dotted continuous line and the received noisy signal as points. Note that the filtered output closely follows the original signal.*

While visibly good, audibly the results achieved are not yet superior to current perceptually motivated techniques, e.g., [1–3]. On closer inspection some reasons for this become apparent. Two readily observed problem cases are shown in Figure 2. The first of these demonstrates an extreme case of amplitude underestimation. It has been observed that the frequency estimations tend to be lower than the original signal more often than higher; this is even the case when the estimation is within the thresholds. It may be possible to overcome this problem through a more intelligent choice of kernel, see, for example, [13]. Here the linear kernel has been used for initial investigations, which is perhaps more appropriate in the estimation of smooth curves, as indeed underestimation in this case appears due to smoothing.

The lower plots in Figure 2 illustrate what may be a slight failing of the *a-priori* assumptions. This is namely that regions with broad thresholds allow significant deviation from the original signal. While it has been claimed that the thresholds are derived such that this variation is not audible, this is for individual variation with respect to the global status quo. In the case that a large number of variations are occurring and, in addition, the smoothing discussed previously is flattening the overall DFT (albeit all within thresholds), then clearly a significant global deviation from the original

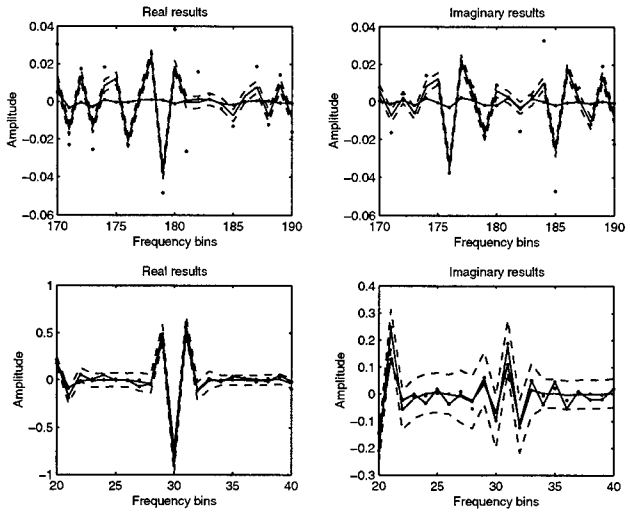


Fig. 2. Problematic filtering in the frequency domain. *The original signal is shown as a continuous line, the thresholds as dashed lines, the filter output as a dotted continuous line and the received noisy signal as points.*

is taking place. Obviously there is no simple solution to this, it being an essential part of the overall framework. One approach may be to identify regions of excessively large margin, such as that illustrated, and to reduce them. Ideas similar to this appear in [1].

In addition to the extensions mentioned, to date no rigorous parameter determinations have been made. Note that this is a problem faced in all SVM applications; methods for such determinations appear in [14], among others. As well, a linear loss function has been used, which is not at all perceptually derived. As the loss function determines the tradeoff in the optimisation process, it appears sensible to attempt to determine in, some manner, one which better reflects the sensitivities of the ear.

5. CONCLUSION

It has been seen that the perceptually based loss functions of [1] fit well in a complex-valued SVM framework. In addition this framework allows a considerable extension of the audio model, incorporating more realistic prior belief about the correlation between frequency components. This includes the previously unconsidered (in the context of perceptually based filtering) aspect of prior belief about phase. In this sense the algorithm takes a more holistic approach to estimating the spectrum of the audio signal.

In the course of applying SVMs to the problem of perceptual audio filtering, new results have been presented on the application of SVMs to the estimation of a complex-valued function. These include the derivation of the Lagrangian formulation and of the complex SMO algorithm. These are results which should prove more widely applicable for such problems.

While a large proportion of Section 4 dwelt on potential improvements regarding the audio results, it is important to emphasise that the results obtained did improve the quality of the signal, albeit not to state-of-the-art levels. This is no surprise given that present results are the result of initial investigations, and several choices (e.g., parameters, kernel, and loss function) have yet to be optimised. However, the preliminary results presented herein indicate that, with the refinements discussed, further improvements will likely be possible.

6. REFERENCES

[1] Wolfe, P.J., Godsill S.J., "Towards a Perceptually Optimal Spectral Amplitude Estimator for Audio Signal Enhancement," in *Proc. IEEE ICASSP*, Istanbul, 2000, vol. 2, pp. 821–824.

[2] Tsoukalas, D.E., Mourjopoulos, J., Kokkinakis, G., "Speech Enhancement Based on Audible Noise Suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 497–514, November 1997.

[3] Virag, N., "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[4] Ephraim, Y., Malah, D., "Speech Enhancement using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[5] Box, G.E.P., Jenkins, G.M., *Time Series Analysis Forecasting and Control*, Prentice Hall, first edition, 1976.

[6] Hopgood, J.R., *Nonstationary Signal Processing with Application to Reverberation Cancellation in Acoustic Environments*, Ph.D. thesis, Cambridge University Engineering Department, September 2000.

[7] Johnston, J.D., "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[8] Girosi, F., Jones, M., Poggio, T., "Priors, Stabilizers and Basis Functions: from regularization to radial, tensor and additive splines," Memo 1430, Massachusetts Institute of Technology Artificial Intelligence Laboratory, June 1993.

[9] Smola, A.J., Schölkopf, B., "A Tutorial on Support Vector Regression," Neurocolt technical report, Royal Holloway College, University of London, 1998.

[10] Wahba, G., *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, 1990.

[11] Platt, J.C., "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in *Advances in Kernel Methods - Support Vector Learning*, Schölkopf, B., Burges, C.J.C., Smola, A.J., Ed., pp. 185–208. MIT Press, Cambridge, MA, 1999.

[12] Flake, G.W., Lawrence, S., "Efficient SVM Regression Training with SMO," *Machine Learning*, 2001, submitted; available from http://www.neci.nj.nec.com/homepages/flake.

[13] Jaakkola, T.S., Haussler, D., "Probabilistic Kernel Regression Models," in *Proceedings of the 1999 Conference on AI and Statistics*, 1999.

[14] Wahba, G., "Support Vector Machines, Reproducing Kernel Hilbert Spaces and Randomized GACV," in *Advances in Kernel Methods: Support Vector Learning*, Schölkopf, B., Burges, C.J.C., Smola, A.J., Ed., pp. 69–88. The MIT Press, 1999.

A PROBABILISTIC FRAMEWORK FOR SUBBAND AUTOREGRESSIVE MODELS APPLIED TO ROOM ACOUSTICS

James R. Hopgood and Peter J. W. Rayner

Signal Processing Laboratory, Department of Engineering,
University of Cambridge, England CB2 1PZ
jrh1008@eng.cam.ac.uk, pjwr@eng.cam.ac.uk

ABSTRACT

Real room acoustic impulse responses (AIRs) modelled by infinite impulse response (IIR) filters require high model orders. Many problems involving the estimation of AIRs reduce to high dimensional optimisation problems. Subband autoregressive (AR) modelling techniques reduce this difficult optimisation problem to a number of simpler low dimensional optimisations. This paper introduces a formulation for subband AR modelling in a probablistic framework which facilitates robust Bayesian parameter estimation. The paper also provides new results to show that the subband AR representation accurately models typical AIRs and, therefore, is suitable for modelling room reverberation.

1. INTRODUCTION

The transfer function due to the acoustics of a room generally do not change considerably with time, but do vary with the spatial locations of the sound source and observer. Assuming both are spatially stationary, a linear time-invariant (LTI) model is appropriate. The *all-pole model* can parsimoniously approximate rational transfer functions, and typical all-pole model orders required for approximating room transfer functions (RTFs) are in the range $50 \leq P \leq 500$ – around a factor of 40 lower than all-zero model orders [1]. A room acoustic impulse response (AIR), $h(t)$, may be modelled by a LTI all-pole filter of order P, as given by:

$$h(t) = -\sum_{p\in\mathcal{P}} a(p)\, h(t-p) + \delta(t), \quad t \in \mathbb{Z} \tag{1}$$

where $\mathbf{a} = \{a(p),\, p \in \mathcal{P} \triangleq \{1,\ldots,P\}\}$ are the model parameters, P is the number of poles, and $\delta(t)$ is the Kronecker delta.

In many applications, such as single channel blind dereverberation [2], an estimate of the AIR is required and, in general, this reduces to a high-dimensional optimisation problem. This is difficult to solve because attempts to model the entire acoustic spectrum by a single IIR filter leads to a large computational load, as well as numerical problems resulting from the size of the parameter space. The problem is that the all-pole model must simultaneously fit the entire frequency range, even though the model may fit some regions in this frequency space better than others. Thus, it is better to model a particular frequency band of the filter's spectrum by an all-pole model, resulting in a lower model order *for that frequency band* and, therefore, improved parameter estimation. Effectively, the modelling of different frequency bands has been *decoupled*, leading to a better model fit and also reducing a high-dimensional optimisation problems to a number of low-dimensional ones.

Subband methods have previously been used to model acoustic environments with much success [3–6]. Subband linear prediction has been considered in [7–9]. This paper introduces a probabilistic formulation for subband AR modelling which leads to Bayesian parameter estimation. The paper also demonstrates that subband AR models are suitable for modelling room acoustics and, therefore, are suitable for modelling room reverberation.

Supported by the Schiff Foundation, University of Cambridge.

2. FREQUENCY DOMAIN FORMULATION

If the room is excited by white Gaussian noise (WGN), the parameter vector of (1), $\mathbf{a}$, can be estimated by considering (1) as an AR process. In the time-domain formulation of the method of least-squares, it is sought to find $\mathbf{a}$ which minimises the expected value of the square of the excitation sequence for the AR sequence:

$$s(t) = -\sum_{p\in\mathcal{P}} a(p)\, s(t-p) + e(t), \quad \forall t \in \mathbb{Z} \tag{2}$$

where $e(t) \sim \mathcal{N}\left(e(t) \,|\, 0,\, \sigma^2\right)$ and $s(t)$ are the input excitation and output, respectively. However, estimators for AR models can also be formulated in the frequency domain [10].

2.1. Likelihood Function

The data sequence $\{s(t),\, t \in \mathcal{T} \triangleq \{0,\ldots,T-1\}\}$ denotes a segment of the infinite sequence introduced in (2) and, for simplicity, is assumed to be periodic; as $T \to \infty$, this approximation becomes more accurate. Application of the DFT to (2), gives:

$$\mathcal{E}(k) = \mathcal{S}(k) + \sum_{p\in\mathcal{P}} a(p)\, \exp\left\{-\frac{2\pi\jmath kp}{T}\right\} \mathcal{S}(k) \tag{3}$$

Denoting $\boldsymbol{\mathcal{E}} = \{\mathcal{E}(k),\, k \in \mathcal{K} \equiv \mathcal{T}\}$, (3) may be written as:

$$\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{S}} + \mathbf{S}\,\mathbf{a} \tag{4}$$

where $\mathbf{S} = [\boldsymbol{\mathcal{S}}_1 \;\ldots\; \boldsymbol{\mathcal{S}}_P]$, and $[\boldsymbol{\mathcal{S}}_p]_k = \exp\{-\frac{2\pi\jmath kp}{T}\}\mathcal{S}(k), k \in \mathcal{K}$. Define $[\mathbf{W}_T]_{k+1,t+1} = \exp\{-\frac{2\pi\jmath kt}{T}\}$, $\forall k \in \mathcal{K}$, $\forall t \in \mathcal{T} \Rightarrow \boldsymbol{\mathcal{E}} = \mathbf{W}_T\,\mathbf{e}$. Noting $\mathbf{e}$ is WGN and $\mathbf{W}_T\,\mathbf{W}_T^\dagger \equiv \mathrm{I}_T \Rightarrow |\mathbf{W}_T| = 1$, where $\mathrm{I}_T \in \mathbb{R}^{T\times T}$ is the identity matrix, then using the probability transformation:

$$p_{\mathcal{E}}\left(\boldsymbol{\mathcal{E}}\right) = \frac{1}{|\mathbf{W}_T| \times |\mathbf{W}_T|} p_{\mathbf{e}}\left(\mathbf{W}_T^{-1}\,\boldsymbol{\mathcal{E}}\right)$$

it follows: $$p_{\mathcal{E}}(\mathcal{E}) = \mathcal{N}(\mathcal{E} \mid \mathbf{0}, \sigma^2 \mathrm{I}_T) \quad (5)$$

Since the Jacobian $\mathcal{J}(\mathcal{S}, \mathcal{E})$ is unity, the likelihood function is:

$$p_{\mathcal{S}}(\mathcal{S} \mid \mathbf{a}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}} \exp\left\{-\frac{\|\mathcal{S} + \mathbf{S}\,\mathbf{a}\|^2}{2\sigma^2}\right\}$$

The maximum-likelihood estimate (MLE) is, therefore,

$$\hat{\mathbf{a}} = -(\mathbf{S}^\dagger \mathbf{S})^{-1} \mathbf{S}^\dagger \mathcal{S} \quad (6)$$

Hence, with suitable priors for the unknown parameters, subband AR modelling can be formulated in the Bayesian framework, with the likelihood function given above.

3. SELECTIVE SUBBAND MODELLING

Consider modelling the power spectrum, $\mathcal{P}(e^{\jmath\omega}) = |\mathcal{S}(e^{\jmath\omega})|^2$, of $s(t)$, where $s(t) \rightleftharpoons \mathcal{S}(e^{\jmath\omega})$ is a Fourier transform pair, in the region $\Omega_k = (\omega_k, \omega_{k+1})$. Consider a signal $\tilde{s}(t)$ whose power spectrum, $\tilde{\mathcal{P}}(e^{\jmath\omega'})$, is related to $\mathcal{P}(e^{\jmath\omega})$ by the mapping

$$\tilde{\mathcal{P}}(e^{\jmath\omega'}) \triangleq \mathcal{P}(e^{\jmath\omega}),\ \omega = \left\{\frac{\omega_{k+1} - \omega_k}{\pi}\right\}\omega' + \omega_k,\ \omega' \in (0, \pi)$$

It is seen that the region $\omega \in \Omega_k$ is mapped onto $\omega' \in (0, \pi)$, and the new process, $\tilde{s}(t)$, can be modelled as an all-pole filter across the entire spectrum, with approximate power spectrum:

$$\tilde{\mathcal{P}}(e^{\jmath\omega}) = \frac{\tilde{G}_k^2}{\left|1 + \sum_{p \in \mathcal{P}} a_k(p)\, e^{-\jmath p\omega}\right|^2},\quad \omega \in (0, \pi) \quad (7)$$

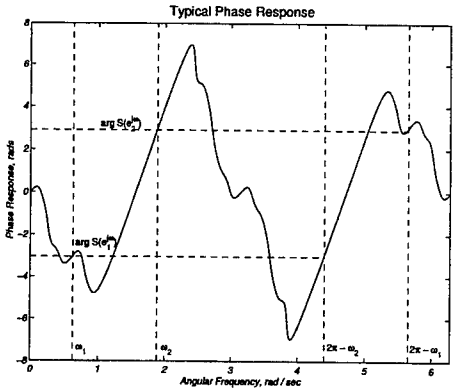
Hence, the estimated power spectrum for $\mathcal{P}(e^{\jmath\omega})$, over the complete frequency range, $(0, \pi)$, can be represented by a series of subband models, as given by:

$$|\mathcal{S}(e^{\jmath\omega})|^2 = \sum_{k=0}^{K-1} \frac{\hat{G}_k^2\, \mathbb{I}_{(\omega_k, \omega_{k+1})}(\omega)}{\left|1 + \sum_{p \in \mathcal{P}_k} a_k(p)\, e^{-\jmath p\pi \frac{\omega - \omega_k}{\omega_{k+1} - \omega_k}}\right|^2} \quad (8)$$

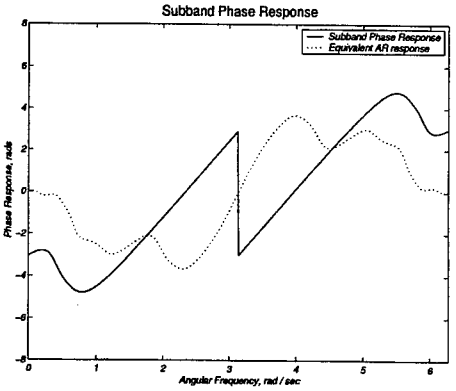
where $\mathcal{P}_k \triangleq \{1, \ldots, P_k\}$, the spectrum of the excitation sequence is given by $e(t) \rightleftharpoons \mathcal{E}(e^{\jmath\omega})$, $\mathbb{I}_\Omega(\omega) = 1$ if $\omega \in \Omega$ and zero otherwise, $\omega_0 = 0$, $\omega_{K+1} = \pi$, and K is the number of subbands. The excitation variance must be scaled proportionally, $\pi\,\hat{G}_k^2 = \tilde{G}_k^2(\omega_{k+1} - \omega_k)$, since energy must be conserved in the transformation. Although, in the frequency range Ω_k, (8) models the power spectrum of the process $s(t)$ and, therefore, the magnitude of the spectrum $\mathcal{S}(e^{\jmath\omega})$, it does not accurately model the phase of $s(t)$, *i.e.* $\arg \mathcal{S}(e^{\jmath\omega})$ since phase information is lost. Hence, (8) suggests that $\{s(t)\}$ is related to its excitation sequence $\{e(t)\}$ by:

$$\mathcal{S}(e^{\jmath\omega}) = \sum_{k=0}^{K-1} \frac{G_k\, \mathcal{E}(e^{\jmath\omega})\, \mathbb{I}_{(\omega_k, \omega_{k+1})}(\omega)}{1 + \sum_{p \in \mathcal{P}_k} a_k(p)\, e^{-\jmath p\pi \frac{\omega - \omega_k}{\omega_{k+1} - \omega_k}}} \quad (9)$$

The phase response of a true AR process is always minimum phase. Consider the typical phase response of an AR process shown in Figure 1(a). The phase of the subband $\Omega_1 \cap \Omega_1' = \{\omega_1 \le \omega_2\} \cap \{2\pi - \omega_2, 2\pi - \omega_1\}$ is shown in Figure 1(b), where $\arg \mathcal{S}(e^{\jmath\omega_1}) \neq 0$ or π. The model in (9) cannot model this phase response; the best it can do is shown in Figure 1(b). A more accurate model is:



(a) Typical Phase Response



(b) Subband Phase Response

Fig. 1. The Phase Ambiguity.

$$\mathcal{S}(e^{\jmath\omega}) = \sum_{k=0}^{K-1} \frac{G_k\, e^{\jmath\phi_k(\omega)\, \mathbf{I}_{(\omega_k, \omega_{k+1})}(\omega)}\, \mathcal{E}(e^{\jmath\omega})}{1 + \sum_{p \in \mathcal{P}_k} a_k(p)\, e^{-\jmath p\pi \frac{\omega - \omega_k}{\omega_{k+1} - \omega_k}}} \quad (10)$$

where $e^{\jmath\phi_k(\omega)}$ corresponds to an additional phase term to compensate for the difference between the actual phase response, and the phase response of an AR process with identical magnitude response. Estimation of this phase term is considered in §5.2.

Given the model in (10), the analysis in §2.1 can be applied to each subband, $k \in \{0, \ldots, K-1\}$, to obtain estimates of the parameters $\mathbf{a}_k$, provided it is reformulated so that the optimisation is over the frequency range $\omega \in \Omega_k$: *i.e.* apply (4) to the spectral error sequence $\mathcal{E}(e^{\jmath\omega_m})$, $\omega_m \in \Omega_k$, where $\omega_m = \frac{2\pi m}{T}$ and T is the number of error samples corresponding to the sequence $s(t)$, and use (6) to obtain an estimate of $\mathbf{a}_k$. The temporal subband AR modelling method implicitly uses a filter bank network and, therefore, care must be taken to ensure that the filter bank possesses perfect reconstruction properties. Details of such techniques are discussed in [12] but, for brevity, are not taken into account here.

4. SUBBAND MODELLING EXAMPLE

As an example of subband AR modelling, a *true* 8th-order AR spectrum is modelled using three subbands; since the number of subbands and the model order in each band are fixed, the location of the spectral changepoint is determined using Bayesian changepoint estimation: see [11]. Figure 2 shows the estimated spectra in each subband for a particular choice of model order.

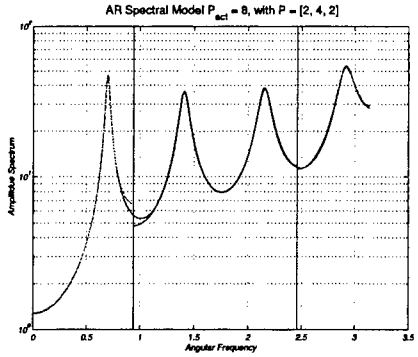


Fig. 2. Subband modelling an AR(8) process. The original and estimated spectra in each of the three subbands are shown, and the vertical line denotes the boundary of the subbands.

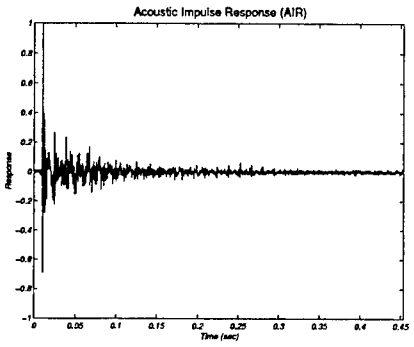
5. SUBBAND MODELLING OF ROOM ACOUSTICS

The subband model in (10) is used to represent a *known* AIR and, thus, could be inverted directly. Naturally, in practice the AIR will be unknown and this will not be possible. This section investigates the ability of the subband model to equalise the RTF.

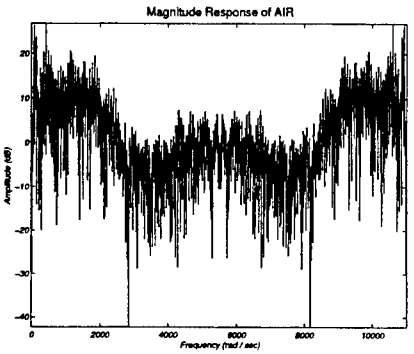
5.1. Reconstructing the Magnitude Frequency Response

A typical AIR, measured in a stairwell, with no direct path from source to observer, is shown in Figure 3(a), with magnitude frequency response shown in Figure 3(b). The length of this impulse response is $T = 4000$ samples. The *minimum-phase equivalent* of this impulse response [13] is modelled using (10). Then, a MLE is calculated using (6), where the phase response is modified as discussed in §5.2; $P_k = 50$, $\forall k \in \{0, \ldots, K-1\}$, $K = 500$, and the AIR is zero-padded by a factor of 200 to improve numerical stability [11]. The *equalised* impulse response is calculated by inverting the frequency response of the model in each subband, multiplying by the *original* frequency response, and taking the inverse Fourier transform. Figure 4 shows the equalised impulse response, and the magnitude response of the equalised RTF shown in Figure 5 indicates that the spectral coloration is significantly reduced.

However, as demonstrated in Figure 6, a closer inspection reveals why the magnitude response contains many sharp spectral components: since the model in each subband is completely decoupled from the other subbands, there are discontinuities in the spectrum at the subband boundaries. The model in (10) does not enforce any continuity between blocks, although this can be ensured by modifying the prior distributions for the AR parameters such that the end point at the lower subband boundary is constrained to match the estimated spectrum in the previous subband.



(a) Impulse Response



(b) Magnitude Frequency Response

Fig. 3. Typical acoustic impulse response.

Moreover, the resonant spikes at the subband boundaries in Figure 6 are due to the implicit use of a filter bank that does not possess perfect reconstruction properties [12]. The modelling of the magnitude frequency response of the nonminimum-phase AIR gives similar results to the minimum-phase system, since the magnitude responses are identical. For nonminimum-phase systems, the phase response must be modelled as discussed below.

5.2. Reconstructing the Phase Frequency Response

In modelling the AIR shown in Figure 3(a), it is sought to minimise the phase discrepancy, $\phi_k(\omega)$, introduced when using a parameter estimate based on the spectral error function. Estimating $\phi_k(\omega)$ is difficult, and is modelled using a polynomial such that a least-squares fit may be obtained. For room acoustics, observational experiments suggest a good approximation for $\phi_k(\omega)$ is:

$$\hat{\phi}_k(\omega) \approx \psi_0 + \psi_1 \omega + \psi_2 \omega^2 \qquad (11)$$

The coefficients $\{\psi_i\}$ can be estimated using least-squares. The equalised impulse response when $\phi_k(\omega)$ is not accounted for is shown in Figure 7. Compared with Figure 4, where $\phi_k(\omega)$ has been accounted for, the equalised response is much longer, and thus no longer accurately reflects an impulse. Acoustic listening tests, in which a clean speech signal is filtered by each of the

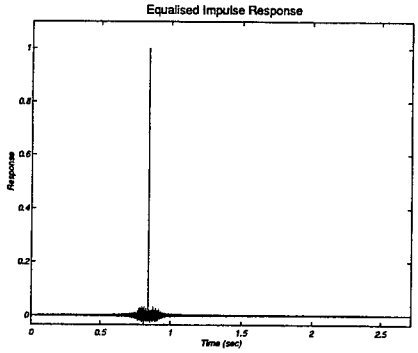


Fig. 4. Equalised impulse response.

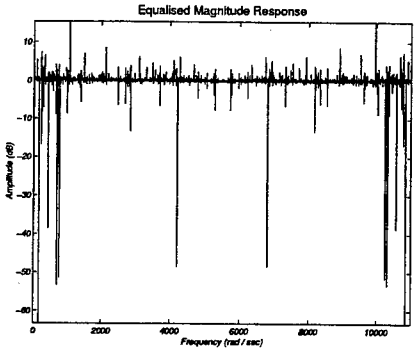


Fig. 5. Equalised magnitude response.

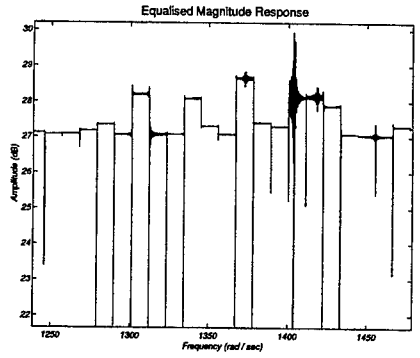


Fig. 6. Discontinuities between the subbands.

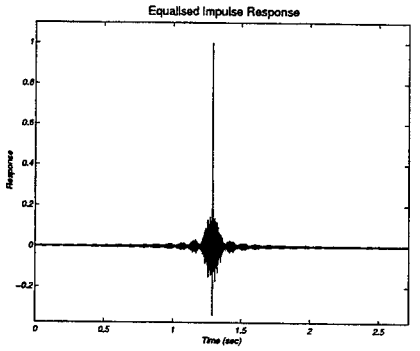


Fig. 7. Effect of ignoring additional phase term $\phi_k(\omega)$.

equalised responses in Figures 4 and 7, indicate that the speech is heavily distorted in the latter case when $\phi_k(\omega)$ is not modelled.

6. CONCLUSIONS

In this paper, a subband AR model has been shown to represent a typical AIR reasonably accurately. The likelihood-function for this spectral model is identical in form that obtained for a time series and, therefore, the subband model elegantly fits into the Bayesian framework [2, 11]. The model produced in this paper means that a difficult high-dimensional optimisation problem reduces to a number of simpler low-dimensional optimisation problems. For nonminimum-phase AIRs, where a causal inverse does not exist, only subbands which possess minimum-phase characteristics can be inverted; hence, the method for detecting the minimum-phase subbands in [5] should be applied to this subband model.

7. REFERENCES

[1] J. N. Mourjopoulos and M. A. Paraskevas, "Pole and zero modeling of room transfer functions," *Journal of Sound and Vibration*, vol. 146, no. 2, pp. 281 – 302, Apr. 1991.

[2] J. R. Hopgood and P. J. W. Rayner, "Bayesian single channel blind deconvolution using parametric signal and channel models," in *Proc. IEEE WASPAA*, Mohonk, New York, Oct. 1999, pp. 151 – 154.

[3] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," in *Proc. IEEE ICASSP*, Paris, France, May 1982, vol. 1, pp. 156–159.

[4] J. N. Mourjopoulos and J. K. Hammond, "Modelling and enhancement of reverberant speech using an envelope convolution method," in *Proc. IEEE ICASSP*, Boston, Apr. 1983, vol. 3, pp. 1144–1147.

[5] H. Wang and F. Itakura, "Dereverberation of speech signals based on sub-band envelope estimation," *IEICE Transactions on*, vol. E74, no. 11, pp. 3576–3583, Nov. 1991.

[6] H. Yamada, H. Wang, and F. Itakura, "Recovering of broadband reverberant speech signal by sub-band MINT method," in *Proc. IEEE ICASSP*, Toronto, Ont., Canada, May 1991, vol. 2, pp. 967–972.

[7] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[8] S. L. Tan and T. R. Fischer, "Linear prediction of subband signals," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 9, pp. 1576–1583, Dec. 1994.

[9] S. Rao and W. A. Pearlman, "Analysis of linear prediction, coding, and spectral estimation from subbands," *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1160–1178, July 1996.

[10] J. Makhoul, "Spectral analysis of speech by linear prediction," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-21, no. 3, pp. 140–148, June 1973.

[11] J. R. Hopgood, *Nonstationary Signal Processing with Application to Reverberation Cancellation in Acoustic Environments*, Ph. D. Thesis, University of Cambridge, UK, Nov. 2000.

[12] P. P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial," *Proc. IEEE*, vol. 78, no. 1, pp. 56–93, Jan. 1990.

[13] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, July 1979.

SIMPLE ALTERNATIVES TO THE EPHRAIM AND MALAH SUPPRESSION RULE FOR SPEECH ENHANCEMENT

Patrick J. Wolfe and Simon J. Godsill*

Signal Processing Group, University of Cambridge
Department of Engineering, Trumpington Street
CB2 1PZ, Cambridge, UK
{pjw47, sjg}@eng.cam.ac.uk
http://www-sigproc.eng.cam.ac.uk

ABSTRACT

Short-time spectral attenuation is a common form of audio signal enhancement in which a time-varying filter, or suppression rule, is applied to the frequency-domain transform of a corrupted signal. The Ephraim and Malah suppression rule for speech enhancement is both optimal in the minimum mean-square error sense and well-known for its associated colourless residual noise; however, it requires the computation of exponential and Bessel functions. In this paper we show that, under the same modelling assumptions, alternative Bayesian approaches lead to suppression rules exhibiting almost identical behaviour. We derive three such rules and show that they are efficient to implement and yield a more intuitive interpretation.

1. INTRODUCTION

1.1. Short-Time Spectral Attenuation

Short-time spectral attenuation is a popular method of broadband noise reduction in which a time-varying filter is applied to the frequency-domain transform of a corrupted audio signal. Often such a signal is modelled as follows: let $\{x_n\} \triangleq \{x(nT)\}$ in general represent a set of values from a finite-duration analogue signal sampled regularly at intervals of T, so that at time n one has the additive observation model $y_n = x_n + d_n$, where y_n is the observed signal, x_n is the original signal, and d_n is random noise.

In many implementations the set of observations $\{y_n\}$ is analysed using the discrete Fourier transform (DFT), via the overlap-add method of short-time Fourier analysis and synthesis. Noise reduction in this manner may be viewed as the application of a suppression rule, or nonnegative real-valued gain H_k, to each bin k of the observed signal spectrum $\mathbf{Y}_k$, in order to form an estimate $\widehat{\mathbf{X}}_k$ of the original signal spectrum.

In the ensuing discussion of such suppression rules we consider, for simplicity of notation and without loss of generality, the case of a single (windowed) short-time block. To facilitate a comparison our notation follows that of Ephraim and Malah [1], except that complex quantities appear in bold throughout.

*Material by the first author is based upon work supported under a U.S. National Science Foundation Graduate Fellowship. The authors also wish to acknowledge the contribution of Shyue Ping Ong to this paper.

1.2. The Ephraim and Malah Suppression Rule

Ephraim and Malah [1] derive a minimum mean-square error (MMSE) short-time spectral amplitude estimator for speech enhancement under the assumption that the Fourier expansion coefficients of the original signal x_n and the noise d_n may be modelled as independent, zero-mean, Gaussian random variables. Thus the observed spectral component in DFT bin k, $\mathbf{Y}_k \triangleq R_k \exp(j\vartheta_k)$, is equal to the sum of the spectral components of the signal, $\mathbf{X}_k \triangleq A_k \exp(j\alpha_k)$, and the noise, $\mathbf{D}_k$. This model leads to the following marginal, joint, and conditional distributions:

$$p(a_k) = \begin{cases} \frac{2a_k}{\lambda_x(k)} \exp\left(-\frac{a_k^2}{\lambda_x(k)}\right) & \text{if } a_k \in [0, \infty), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

$$p(\alpha_k) = \begin{cases} \frac{1}{2\pi} & \text{if } \alpha_k \in [-\pi, \pi), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$$p(a_k, \alpha_k) = \frac{a_k}{\pi\lambda_x(k)} \exp\left(-\frac{a_k^2}{\lambda_x(k)}\right) \tag{3}$$

$$p(\mathbf{Y}_k | a_k, \alpha_k) = \frac{1}{\pi\lambda_d(k)} \exp\left(-\frac{\left|\mathbf{Y}_k - a_k e^{j\alpha_k}\right|^2}{\lambda_d(k)}\right) \tag{4}$$

where it is understood that (3) and (4) are defined over the range of a_k in (1) and α_k in (2); $\lambda_x(k) \triangleq E[|\mathbf{X}_k|^2]$ and $\lambda_d(k) \triangleq E[|\mathbf{D}_k|^2]$ denote the respective variances of the kth short-time spectral component of the signal and noise. The MMSE spectral amplitude estimator derived by Ephraim and Malah, when combined with their derived optimal phase estimator (the observed phase ϑ_k [1]), takes the form of a suppression rule:

$$H_k = \frac{\sqrt{\pi v_k}}{2\gamma_k} \left[(1 + v_k) I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] \exp\left(\frac{-v_k}{2}\right), \tag{5}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of order zero and one, respectively. Additionally,

$$\frac{1}{\lambda(k)} \triangleq \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)}$$

and

$$v_k \triangleq \frac{\xi_k}{1 + \xi_k} \gamma_k; \qquad \xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)}, \quad \gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)},$$

where ξ_k and γ_k are interpreted after [2] as the *a priori* and *a posteriori* SNR, respectively.

0-7803-7011-2/01/$10.00 ©2001 IEEE

2. DERIVATION OF EFFICIENT APPROXIMATIONS

2.1. Joint Maximum *A Posteriori* Spectral Amplitude and Phase Estimator

Joint estimation of the real and imaginary components of $\mathbf{X}_k$ under either the maximum *a posteriori* (MAP) or MMSE criterion leads to the Wiener estimator (due to symmetry of the resultant posterior distribution, which is Gaussian). However, one may reformulate the problem in terms of spectral amplitude A_k and phase α_k, and then obtain a joint MAP estimate by maximising the posterior distribution $p(a_k, \alpha_k|\mathbf{Y}_k)$:

$$\begin{aligned} p(a_k, \alpha_k|\mathbf{Y}_k) &\propto p(\mathbf{Y}_k|a_k, \alpha_k)p(a_k, \alpha_k) \\ &\propto \frac{a_k}{\pi^2\lambda_x(k)\lambda_d(k)} \exp\left(-\frac{\left|\mathbf{Y}_k - a_k e^{j\alpha_k}\right|^2}{\lambda_d(k)} - \frac{a_k^2}{\lambda_x(k)}\right). \end{aligned}$$

Since $\ln(\cdot)$ is a monotonically increasing function, one may equivalently maximise the natural logarithm of $p(a_k, \alpha_k|\mathbf{Y}_k)$. Define

$$J_1 = -\frac{\left|\mathbf{Y}_k - a_k e^{j\alpha_k}\right|^2}{\lambda_d(k)} - \frac{a_k^2}{\lambda_x(k)} + \ln a_k + \text{constant}.$$

Differentiating J_1 w.r.t. α_k yields

$$\frac{\partial}{\partial \alpha_k} J_1 = -\frac{1}{\lambda_d(k)}\left[(\mathbf{Y}_k^* - a_k e^{-j\alpha_k})(-ja_k e^{j\alpha_k}) + (\mathbf{Y}_k - a_k e^{j\alpha_k})(ja_k e^{-j\alpha_k})\right].$$

Setting to zero and substituting $\mathbf{Y}_k = R_k \exp(j\vartheta_k)$, we get

$$\begin{aligned} 0 &= j\hat{a}_k R_k e^{j(\vartheta_k - \hat{\alpha}_k)} - j\hat{a}_k R_k e^{-j(\vartheta_k - \hat{\alpha}_k)} \\ &= 2j \sin(\vartheta_k - \hat{\alpha}_k), \end{aligned}$$

and therefore

$$\hat{\alpha}_k = \vartheta_k, \tag{6}$$

i.e., the joint MAP phase estimate is simply the noise phase. Differentiating J_1 w.r.t. a_k yields

$$\frac{\partial}{\partial a_k} J_1 = -\frac{1}{\lambda_d(k)}\left[(\mathbf{Y}_k^* - a_k e^{-j\alpha_k})(-e^{j\alpha_k}) + (\mathbf{Y}_k - a_k e^{j\alpha_k})(-e^{-j\alpha_k})\right] - \frac{2a_k}{\lambda_x(k)} + \frac{1}{a_k}.$$

Setting the above to zero implies

$$\begin{aligned} 2\hat{a}_k^2 &= \lambda_x(k) - \frac{\lambda_x(k)}{\lambda_d(k)}\hat{a}_k[2\hat{a}_k - R_k e^{-j(\vartheta_k - \hat{\alpha}_k)} - R_k e^{j(\vartheta_k - \hat{\alpha}_k)}] \\ &= \lambda_x(k) - \xi_k\hat{a}_k[2\hat{a}_k - 2R_k\cos(\vartheta_k - \hat{\alpha}_k)]. \end{aligned}$$

From (6), we have $\cos(\vartheta_k - \hat{\alpha}_k) = 1$; therefore

$$0 = 2(1+\xi_k)\hat{a}_k^2 - 2R_k\xi_k\hat{a}_k - \lambda_x(k).$$

Solving the above quadratic equation, and substituting

$$\lambda_x(k) = \frac{\xi_k}{\gamma_k} {R_k}^2, \tag{7}$$

we have

$$\widehat{A}_k = \frac{\xi_k + \sqrt{{\xi_k}^2 + 2(1+\xi_k)\frac{\xi_k}{\gamma_k}}}{2(1+\xi_k)} R_k. \tag{8}$$

Together (8) and (6) define the following suppression rule:

$$H_k = \frac{\xi_k + \sqrt{{\xi_k}^2 + 2(1+\xi_k)\frac{\xi_k}{\gamma_k}}}{2(1+\xi_k)}.$$

2.2. Maximum *A Posteriori* Spectral Amplitude Estimator

First we note that the posterior density $p(a_k|\mathbf{Y}_k)$ arising from integration over the phase term α_k is Rician with parameters (σ_k^2, s_k^2):

$$p(a_k|\mathbf{Y}_k) = \frac{a_k}{\sigma_k^2}\exp\left(-\frac{a_k^2 + s_k^2}{2\sigma_k^2}\right) I_0\left(\frac{a_k s_k}{\sigma_k^2}\right) \tag{9}$$

$$\sigma_k^2 \triangleq \frac{\lambda(k)}{2}, \quad s_k^2 \triangleq \upsilon_k\lambda(k). \tag{10}$$

for large arguments of $I_0(\cdot)$ we may substitute the approximation

$$I_0(|x|) \approx \frac{1}{\sqrt{2\pi\,|x|}}\exp(|x|)$$

into (9), yielding

$$p(a_k|\mathbf{Y}_k) \approx \frac{1}{\sqrt{2\pi\sigma_k^2}}\left(\frac{a_k}{s_k}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\left[\frac{a_k - s_k}{\sigma_k}\right]^2\right), \tag{11}$$

which is almost Gaussian. Considering (11), and maximising its natural logarithm w.r.t. a_k, we obtain

$$J_2 = -\frac{1}{2}\left[\frac{a_k - s_k}{\sigma_k}\right]^2 + \frac{1}{2}\ln a_k + \text{constant}$$

$$\begin{aligned} \frac{d}{da_k} J_2 &= \frac{s_k - a_k}{\sigma_k^2} + \frac{1}{2a_k} \\ 0 &= \hat{a}_k^2 - s_k\hat{a}_k - \frac{\sigma_k^2}{2}. \end{aligned} \tag{12}$$

Substituting (10) and (7) into (12) and solving, we arrive at an estimator differing from that of the joint MAP solution only by a factor of two under the square root (owing to the factor $\sqrt{a_k}$ in (11); replacement with a_k would yield the joint MAP spectral amplitude solution):

$$\widehat{A}_k = \frac{\xi_k + \sqrt{{\xi_k}^2 + (1+\xi_k)\frac{\xi_k}{\gamma_k}}}{2(1+\xi_k)} R_k. \tag{13}$$

Combining (13) with the Ephraim and Malah optimal phase estimator (i.e., the observed phase ϑ_k; *cf.* (6) also) yields the following suppression rule:

$$H_k = \frac{\xi_k + \sqrt{{\xi_k}^2 + (1+\xi_k)\frac{\xi_k}{\gamma_k}}}{2(1+\xi_k)}.$$

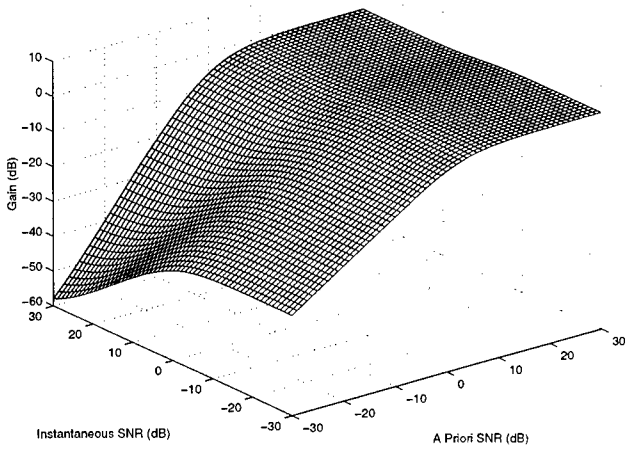


Fig. 1. Ephraim and Malah MMSE suppression rule

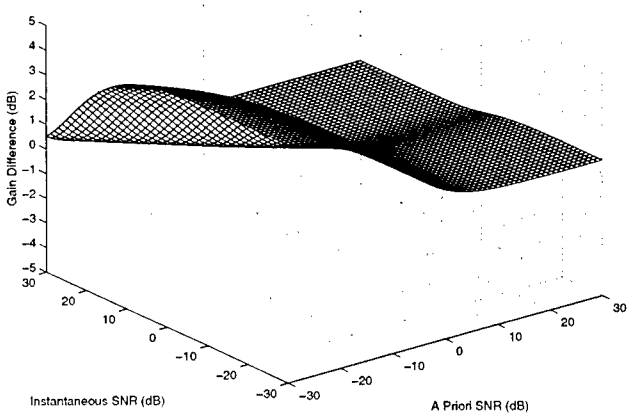


Fig. 3. MAP approximation suppression rule gain difference

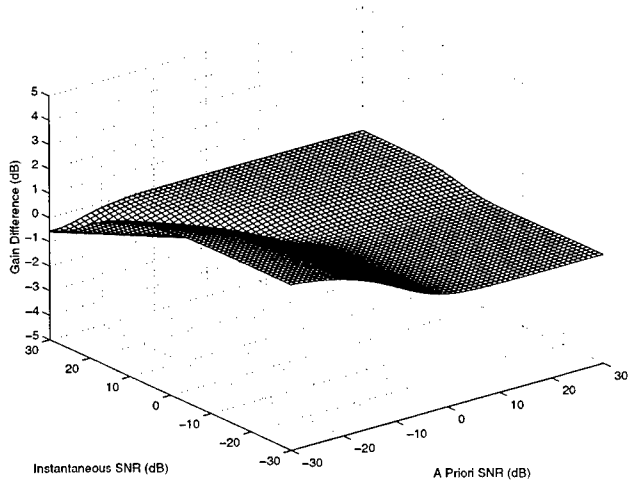


Fig. 2. Joint MAP suppression rule gain difference

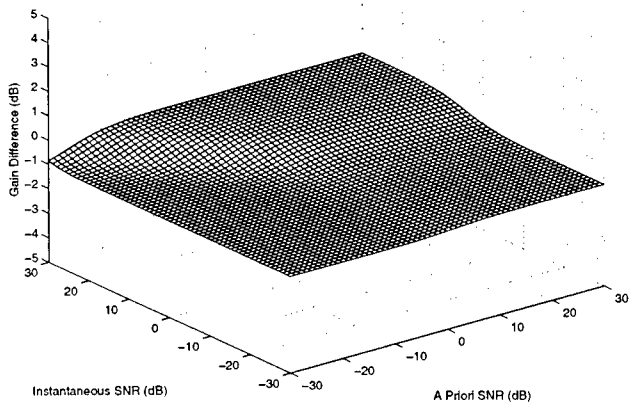


Fig. 4. MMSE power suppression rule gain difference

2.3. Minimum Mean-Square Error Spectral Power Estimator

Recall that Ephraim and Malah formulate the first moment of a Rician posterior distribution, $E[A_k|\mathbf{Y}_k]$, as a suppression rule. The second moment $E[A_k^2|\mathbf{Y}_k]$ of that distribution is given by a much simpler formula (see, e.g., [3]):

$$E[A_k^2|\mathbf{Y}_k] = 2\sigma_k^2 + s_k^2, \tag{14}$$

where σ_k^2 and s_k^2 are as defined previously in (10). Letting $B_k = A_k^2$ and substituting for σ_k^2 and s_k^2 in (14) yields

$$\widehat{B}_k = \frac{\xi_k}{1+\xi_k}\left(\frac{1+\upsilon_k}{\gamma_k}\right)R_k^2,$$

where $\widehat{B}_k$ is the optimal spectral power estimator in the MMSE sense, as it is also the first moment of a new posterior distribution $p(b_k|\mathbf{Y}_k)$ having a noncentral chi-square probability density function with two degrees of freedom and parameters (σ_k^2, s_k^2).

When combined with the observed phase ϑ_k, this estimator also takes the form of a suppression rule:

$$H_k = \sqrt{\frac{\xi_k}{1+\xi_k}\left(\frac{1+\upsilon_k}{\gamma_k}\right)}. \tag{15}$$

3. COMPARISON OF APPROXIMATIONS

Figure 1 shows the Ephraim and Malah suppression rule as a function of instantaneous SNR (defined in [1] as $\gamma_k - 1$) and *a priori* SNR ξ_k. Figures 2, 3, and 4 show the gain difference (in dB) between it and each of the three derived suppression rules (note the difference in scale). Table 1 on the following page shows a comparison of the magnitude of gain differences for the three approximations. The MMSE spectral power suppression rule provides the best and most consistent approximation to the Ephraim and Malah

Suppression Rule	$(\gamma_k - 1, \xi_k) \in [-30, 30]$ dB			$(\gamma_k - 1, \xi_k) \in [-100, 100]$ dB		
	Mean	Maximum	Range	Mean	Maximum	Range
MMSE Spectral Power	0.68473	−1.0491	1.0469	0.63092	−1.0491	1.0491
Joint MAP Spectral Amplitude and Phase	0.52192	+1.7713	2.3352	0.74507	+1.9611	2.5250
MAP Spectral Amplitude Approximation	1.2612	+4.7012	4.7012	1.7423	+4.9714	4.9714

Table 1. Magnitude of deviation from Ephraim and Malah MMSE suppression rule gain

rule, with only slightly less suppression. The MAP spectral amplitude approximation, although still within 5 dB of the optimal value over a wide range of SNR, is the poorest. While the sign of the deviation of each of these two approximations is constant, that of the joint MAP suppression rule depends on the instantaneous and *a priori* SNR.

4. DISCUSSION

Ephraim and Malah [1] show that at high SNR, their derived suppression rule approaches the Wiener suppression rule:

$$H_k = \frac{\xi_k}{1 + \xi_k}. \tag{16}$$

Although not immediately obvious upon inspection of (5), this relationship is easily seen in the MMSE spectral power suppression rule given by (15), expanded slightly to the following:

$$H_k = \sqrt{\frac{\xi_k}{1 + \xi_k}\left(\frac{1}{\gamma_k} + \frac{\xi_k}{1 + \xi_k}\right)}. \tag{17}$$

As the instantaneous SNR γ_k becomes large, (17) may be seen to approach the Wiener suppression rule given by (16). As it becomes small, the $1/\gamma_k$ term in (17) lessens the severity of the attenuation. Cappé [4] makes the same qualitative observation concerning the behaviour of the Ephraim and Malah suppression rule, although the simpler form of the MMSE spectral power estimator shows the influence of the *a priori* and *a posteriori* SNR more explicitly.

Lastly, we note that the success of the Ephraim and Malah suppression rule is largely due to the so-called 'decision-directed approach' for estimating the *a priori* SNR ξ_k [4]. For a given short-time block l, the decision-directed *a priori* SNR estimate $\widehat{\xi}_k$ is given by a geometric weighting of the SNR in the previous and current blocks:

$$\widehat{\xi}_k = \alpha \frac{|\widehat{\mathbf{X}}_k(l-1)|^2}{\lambda_d(l-1,\, k)} + (1-\alpha)\max\left[0, \gamma_k(l) - 1\right],\ \alpha \in [0, 1). \tag{18}$$

It is instructive to consider the case in which $\xi_k = \gamma_k - 1$; i.e., $\alpha = 0$ in (18) so that the estimate of the *a priori* SNR is based only on the current block. In this case the MMSE spectral power suppression rule given by (17) reduces to the method of power spectral subtraction (see, e.g., [2]). Figure 5 shows a comparison of the derived suppression rules under this constraint.

5. CONCLUSION

Herein we have presented a derivation and comparison of three simple alternatives to the Ephraim and Malah MMSE spectral amplitude estimator. These may be implemented where increased efficiency is desired, and each may be coupled with hypotheses concerning uncertainty of speech presence, as in [1, 2]. Moreover, the form of the MMSE spectral power suppression rule given by (17) provides a clear insight into the behaviour of the Ephraim and Malah solution, and in particular its connection to simpler suppression rules.

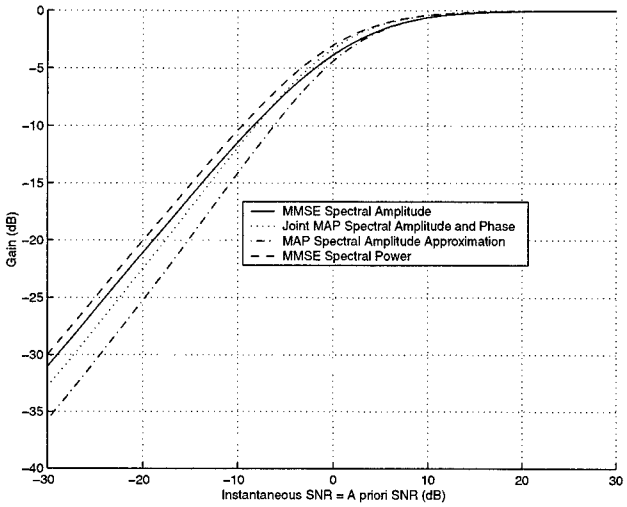


Fig. 5. Optimal and derived suppression rules

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[3] S. O. Rice, "Statistical properties of a sine wave plus random noise," *Bell System Technical Journal*, vol. 27, pp. 109–157, Jan. 1948.

[4] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

A SPECTRAL DISTANCE MEASURE FOR SPEECH DETECTION IN NOISE AND SPEECH SEGMENTATION

K. Drouiche[1], P. Gómez[2], A. Alvarez[2], R. Martínez[2], V. Rodellar[2], V. Nieto[2]

[1]Université de Cergy-Pontoise
2 Avenue Adolphe Chauvin
95302 Cergy Pontoise Cédex, France
e-mail: kd@u-cergy.fr

[2]Universidad Politécnica de Madrid
Campus de Montegancedo, s/n
Boadilla del Monte, 28660 Madrid, Spain
e-mail: pedro@pino.datsi.fi.upm.es

ABSTRACT

Through this paper a new *Spectral Distance Measure* is introduced and its properties explained. This measure is especially designed to evaluate distances between spectral densities, and presents important properties, as the invariance to scaling factors, or shifts in amplitude. The measure may be used as a *Test for Whiteness*, to determine the similarity between independent processes, or to check the *Quasi-stationarity* condition in a single process. Its special ability to detect spectral similarities may be exploited for *Speech Segmentation* and in the *detection of Speech* under strong noise levels, and may be used in *End-point Detection* applications. The fundamentals of the measure are given, some case studies are described and the results discussed.

1. INTRODUCTION

One of the main problems in *Noise-Robust Speech Recognition* is the detection of speech boundaries when the Signal-to-Noise Ratio is rather low, such that energy-based criteria can not be used. The inaccurate detection of speech gives incorrect information to Speech Recognizing Engines, thus increasing recognition errors. Good algorithms for end-point detection and speech segmentation have been designed and developed throughout the last years, using energy-based thresholding, stochastic spectral detection, neural networks, autocorrelation techniques, etc. [2][1][5][8][9]. In a previous work *Adaptive Lattice-Ladder Filters* have also been successfully used to determine speech boundaries [6]. Through this work a different method based on the application of a *Spectral Distance Measure (SDM)* supported by *Statistical Test Theory* will be exposed [3].

For the formulation of the *SDM* the following considerations will be made. Assume two auto-regressive processes $(x_n)_{n\in Z}$ and $(y_n)_{n\in Z}$ defined as:

$$x_n + \sum_{i=1}^{p_x} a_{x,i} x_{n-i} = \varepsilon_{x,n} \qquad (1)$$

$$y_n + \sum_{i=1}^{p_y} a_{y,i} y_{n-i} = \varepsilon_{y,n} \qquad (2)$$

where $\varepsilon_{x,n}$ and $\varepsilon_{y,n}$ are two independent and identically distributed random stochastic processes with zero mean and variances given by σ_x^2 and σ_y^2 respectively. Let $\varphi_x(\omega)$ and $\varphi_y(\omega)$ be their respective positive spectral densities on $0\leq\omega\leq 2\pi$, where ω is the angular frequency:

$$2\pi\varphi_x(\omega) = \frac{\sigma_x^2}{|P_x(\omega)|^2} \qquad (3)$$

$$2\pi\varphi_y(\omega) = \frac{\sigma_y^2}{|P_y(\omega)|^2} \qquad (4)$$

$P_x(\omega)$ and $P_y(\omega)$ being respectively the transfer functions in the domain of z associated with both processes evaluated on the unity circle, i.e.:

$$P_x(z=e^{j\omega}) = 1 + \sum_{k=1}^{p} a_{x,k} e^{-jk\omega} \qquad (5)$$

$$P_y(z=e^{j\omega}) = 1 + \sum_{k=1}^{p} a_{y,k} e^{-jk\omega} \qquad (6)$$

p being the maximum order of both processes, setting the corresponding non-existing coefficients of the lowest-order process to zero: $p=max\{p_x,p_y\}$.

Let $\rho(\omega)$ denote the positive ratio between both spectral densities:

$$\rho(\omega) = \frac{\varphi_x(\omega)}{\varphi_y(\omega)} \qquad (7)$$

The *SDM* test is based thence in checking whether this ratio is constant or not over the frequency span. In other words, if the ratio is near to constant for all the frequencies considered, one may conclude that both autoregressive processes are identical except for a factor explainable as the *SNR* of the noise input processes (given by σ_x^2/σ_y^2).

The proposed measure may be then formulated as:

$$D(\rho) = D\{\varphi_x(\omega),\varphi_x(\omega)\} = log\{(1/2\pi)\int_{-\pi}^{\pi}\rho(\omega)d\omega\} - $$

$$- (1/2\pi)\int_{-\pi}^{\pi} log\,\rho(\omega)d\omega \qquad (8)$$

It may be shown that D fulfils the following properties:

- *(P1)*: $D(\rho)\geq 0$ for all $\rho(\omega)$; $0\leq\omega\leq 2\pi$.

- *(P2): $D(\lambda\rho) = D(\rho)$* for any real positive number λ.
- *(P3): $D(\rho) = 0$* iff *$\rho(\omega)$ = constant almost everywhere.*

Property *(P1)* follows directly from *Jenssen's* inequality. Property *(P2)* says that D is invariant to a change in the scale factor between the power distribution (variance) of both processes. Property *(P3)* is by far the most important one, as it establishes that if the ratio between the power distribution of two processes is almost constant they may be considered as generated by the same system fed with two white noise processes with different variances, thus separating the contribution of the systems from that of their inputs in the overall behavior of both power distributions.

From property *(P3)* an *absolute* measure of the distance from a given process x_n with respect to a hypothetical white noise process ε_n with spectral density given by $\sigma^2/2\pi$=*constant* may be implemented just imposing φ_y=*1* in *(7)*, this constituting a *Test for Whiteness* of process x_n:

$$D_w(\varphi_x) = \log\{(1/2\pi)\int_{-\pi}^{\pi}\varphi_x(\omega)d\omega\} -$$

$$-(1/2\pi)\int_{-\pi}^{\pi}\log\varphi_x(\omega)d\omega \qquad (9)$$

where D_w establishes the distance of the considered process x_n with respect to randomness. This is the essential theory underlying the methodology proposed. Through the rest of the paper this methodology will be detailed (Section 2), and its application for *Speech Segmentation* (Section 3) and *End-Point Detection* (Section 4) will be commented. A brief discussion (Section 5) will extract the most relevant conclusions on the applicability of the referred technique.

2. TEST METHODOLOGY

Through the present paper the applicability of the *SDM* for the detection of speech in noise is sought. Without a lose of generality it will be considered that the processes to be compared are extracted from the same process, namely a signal frame as given in Figure 1. To implement the test two sliding windows will be used, each of one being N samples long, separated by an interval of M samples. There may be two possible techniques used to obtain an estimation of the spectral densities of the traces in the sliding windows:

a) To use the *squared absolute value of the FFT* applied on the *Hamming windowed* N-sample trace.

b) To use the *spectral envelope* defined by the *p-th order Auto-regressive Model* of the N-sample trace.

This last possibility is being implied in the example given in Figure 1. In the examples given throughout the paper an *N-point FFT* was used instead.

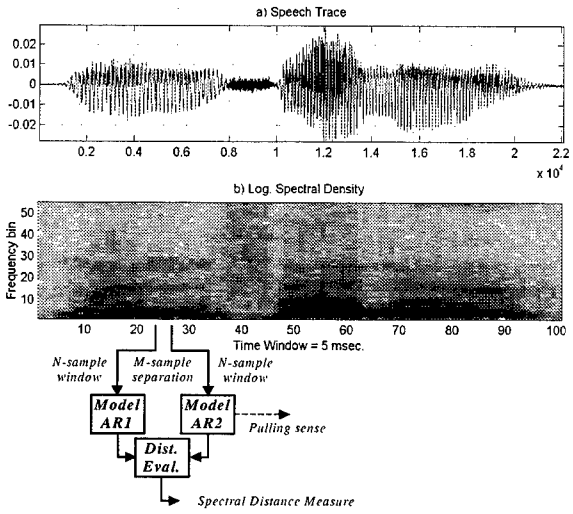


Figure 1. Two sliding and non-overlaping windows N samples long separated by M samples to grant independency are used to define processes x_n and y_n.

A trade-off must be established regarding the size N and separation M of both windows to preserve the quasi-stationarity conditions on one side, and to grant independency on the other. It is well established in Speech Processing and Recognition that quasi-stationarity is well preserved with time windows not exceeding *10 msec.* This compromise could be used both to establish M and N, in the sense that they should not be large enough to exceed the number of samples allowed by the sampling frequency to keep the time intervals below the mentioned value. In the examples shown the time windows were *N=220* samples long for a sampling frequency of *f=22050*, and the separation interval *M=0*, therefore the sliding windows are contiguous and non-overlaping. Another important aspect to be considered is the possibility of carrying out the test sample by sample or by blocks of samples. This aspect will have a direct expression on the computational costs of the method. Block processing will be used in the examples presented. In principle the tests are devised to establish the similarity between processes, or the whiteness of a given process. Therefore, if it may be assumed that noise contaminating speech is purely white. It will be quite simple to detect where there is speech present just checking a given segment of a record against itself, as described in Section 1. For non-white noise the comparison of near-neighbor segments of a signal supposedly containing speech would also give interesting hints to end-point segmentation, assuming that contaminating noise is quasi-stationary. In the case that this condition can not be granted, the measure may be used to detect the degree in which a given process is far apart from the quasi-stationary condition, adding a new feature to the capabilities of the *SDM*.

3. SPEECH SEGMENTATION

The first experiment to be presented is oriented to measure the spectral variations inside a clean speech frame. For the analysis an utterance of the word */manθana/* (apple) was used. Figure 2.a shows the time-domain speech trace. Its spectrogram using a

220-point FFT is being shown in Figure 2.b. The so called *Global Distance Measure* summarizing the most important spectral changes is given in Figure 2.c.

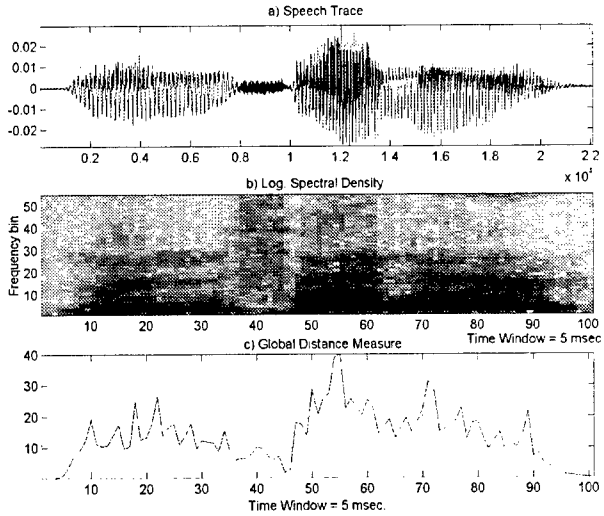


Figure 2. a) Speech Trace for */manθana/*. b) Associated Spectral Density from *200-point FFT*. c) Global Distance Measure.

It may be observed that each time an important change in the spectral density takes place the *Global Distance Measure* increases, to decrease again immediately after. The peaks determine quite reasonably the spectral changes. Therefore the speech frame could be divided into different sectors of *quasi-stationary behavior*. The most dramatic change takes place around templates *54* and *55*, and corresponds apparently to the transition inside a given vowel *[a]* from a non-nasal character in *[θa]* to a nasalized coloring in the articulation of *[ān]*. Another important transition takes place around window *71-72* when the nasalization ends.

4. END-POINT DETECTION

Another important problem to be treated using the technique described is *End-point Detection* for speech traces in the presence of strong noise levels. Under this assumption three different cases may be studied:

a) The contaminating noise is stationary and white. In this case the *Spectral Distance Measure* may be very efficient, as it may detect the presence of speech using two combined hints: where the spectral density changes from a white case to a non-white case using (9) on a single sliding window, or where there is a spectral change when comparing two sliding windows using (8).

b) Noise is colored and quasi-stationary. In this case spectral changes will be spotted using two sliding windows and (8).

c) Noise is colored and non-stationary. This is the worst case, as there will be no means to infer if the spectral changes are due to hidden speech or to noise spectral changes. Even in this case it will be useful to detect changes in the spectral density function, as they may be hints to the possible presence of speech. One such example is given in Figure 3.

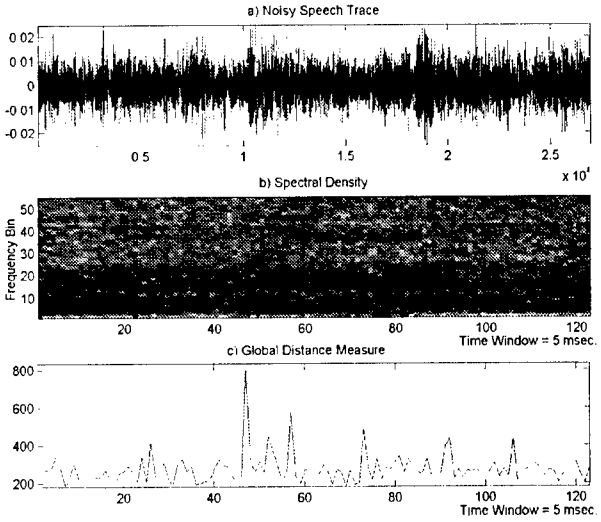


Figure 3. a) Noisy Speech Trace. Actual SNR is under – *5 dB*. The underlying speech trace corresponding to an utterance of the isolated word */eight/* can not be seen. b) Associated Spectral Density showing the presence of strong noise components. c) Global Distance measure pinpointing the possible presence of speech.

The Noisy Speech Trace was recorded under strong noise levels (racing car noise above 95 dB SPL), corresponding to isolated words. Under these conditions neither the time signal nor the associated spectrogram allow to infer the presence of speech. Nevertheless the *Global Distance Measure* in Figure 3.c gives specific points where important spectral changes take place, these being around windows *46* and *57*. Other smaller changes are spotted for templates around *73*, *82* or *106*. To see whether this detection corresponds with actual speech segments, a *speech enhancement technique* previously developed [7] and [4] was used. The results may be seen in Figure 4 to be contrasted against the data in Figure 3.

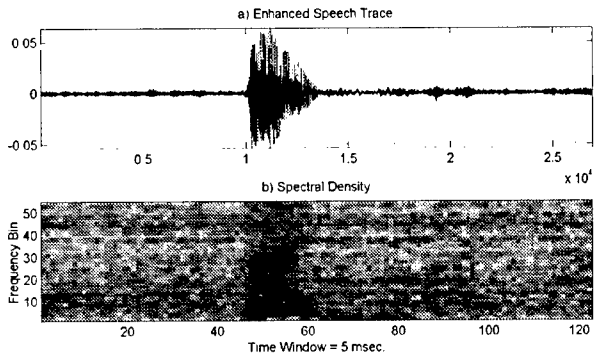


Figure 4. a) Enhanced Speech Trace using the technique described in [7] and [4]. b) Associated Spectral Density showing the spectrogram of the word */eight/*.

In fact, it may be seen that within window frames *46* and *58* there is a speech frame present, corresponding to the isolated word embedded in noise.

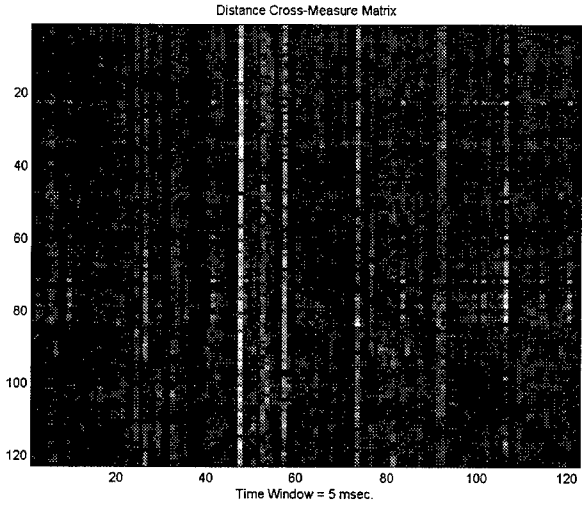


Figure 5. Matrix corresponding to the measure of each window with respect to other windows in the speech trace, showing clearly the boundaries in the spectrogram corresponding to the spotting factor given in Figure 3.c.

5. DISCUSSION

To study the consistency of the results an exhaustive test was carried out cross-contrasting the spectral density estimates corresponding to the whole set of windows in the speech frame of Figure 3 among themselves, the corresponding results being given as a matrix D_{ij}:

$$D_{ij} = D(\rho_{ij}); \qquad 0 \le i \le N_w\text{-}1;\ 0 \le j \le N_w\text{-}1 \qquad (10)$$

where N_w is the total number of time windows used in the analysis described. As it would be reasonable, the main diagonal of the matrix D is $D_{ii}=0$, as this distance measure is zero when using (8). It is important to remark that the matrix is non-symmetric respect to the main diagonal, because from (7):

$$\rho_{ij} = 1/\rho_{ji} \qquad (11)$$

therefore the condition that:

$$D_{ij} = D_{ji} \qquad (12)$$

will signal the areas where the spectral density is *white*. The most important property of the *Cross-Distance Matrix* is that its columns will spot the spectral changes, as they will show a correlation among a given window W_i and all the other windows W_j; $0 \le j \le N_w\text{-}1$. The lighter columns (those keeping higher distances with respect to the other windows) will signal the spectral boundaries. It may be seen that there is a perfect correspondence between these boundaries and the ones given in Figure 3.c.

There are many other fields of application for the *Spectral Distance Measure* presented, as for example in random signal whitening. Most random generators do not produce true white traces, although they are very much used in noise-staining experiments. Using the measure introduced here a pre-produced trace may be whitened to approach a flat spectral density with a rippling error of less than *0.01 dB*. Long white series may be produced this way with important practical applications, such as in sound equipment calibrations and others similar.

6. ACKNOWLEDGMENTS

This research is being supported by Project TIC99-0960 (Programa Nacional de las Tecnologías de la Información y las Comunicaciones), Project 07T/0001/2000 (Plan Regional de Investigación de la Comunidad Autónoma de Madrid), and the support of the *University of Cergy Pontoise*.

7. REFERENCES

[1] Acero, A., Crespo, C., Torre, C. and Torrecilla, J., "Robust HMM-Based Endpoint Detector", *Eurospeech'93*, Berlin, Germany, 1993, pp. 1551-1554.

[2] Dermatas, E., Fakotakis, N. and Kokkinakis, G., "Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment", *ICASSP'91*, Toronto, Canada, 1991, pp. 733-736.

[3] Drouiche, K., "A New Test for Whiteness", *IEEE Trans. on Signal Proc.*, Vol. 48, July 2000, pp. 1864-1871.

[4] Gómez, P., Alvarez, A., Martínez, R., Nieto, V. and Rodellar, V., "A Hybrid Signal Enhancement Method for Robust Speech Recognition", *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Robust'99*, Tampere, Finland, May 25-26, 1999, pp. 203-206.

[5] Mak, B., Junqua, J. C. and Reaves, B., "A Robust Speech/Non-Speech Detection Algorithm Using Time and Frequency-based Features", *ICASSP'92*, San Francisco, CA, 1992, pp. 269-272.

[6] Martínez, R., Alvarez, A., Gómez, P., Pérez, M., Nieto, V. and Rodellar, V. "A Speech Pre-Processing Technique for End-Point Detection in Highly Non-Stationary Environments", *Proc. of EUROSPEECH'97*, Rhodes, Greece, 22-25 September, 1997, pp. 1111-1114.

[7] Martínez, R., Gómez, P., Alvarez, A., Nieto, V., Rodellar, V., Rubio, M. and Pérez, M., "Dynamic Adjustment of the Forgetting Factor in Adaptive Filters for Non-Stationary Noise Cancellation in Speech", *1998 Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'98*, Seattle, Washington, USA, May 12-15, 1998, pp. 1009-1012.

[8] Rangoussi, M., Bakamidis, S. and Carayannis, G., "Robust Endpoint Detection of Speech in the Presence of Noise", *EUROSPEECH'93*, Berlin, Germany, 1993, pp. 649-652.

[9] Zhu, J., and Chen, F. L., "The analysis and application of a new endpoint detection method based on distance of autocorrelated similarity", *Proc. of Eurospeech'99*, Budapest, Hungary, September 5-9, 1999, paper S1.PO1.12

A NEW ALGORITHM FOR JOINT DOA AND MULTIPATH DELAY ESTIMATION: SEPARABLE DIMENSION SUBSPACE METHOD

Jian Mao
Sigprowireless Inc.
Ottawa, Ont., K2C 0R4, Canada
mao@sigprowireless.com

Benoit Champagne
Dept. of ECE, McGill University,
Montreal, QC, H3A 2A7, Canada
champagne@tsp.ece.mcgill.ca

Mairtin O'Droma
Dept.of ECE, University of Limerick,
Limerick, Ireland
mairtin.odroma@ul.ie

Lijia Ge
Chongqin University,
Chongqin, China
lj4500629@sina.com

ABSTRACT

With the growing capacity demand in wireless communication systems, space division multiplexing and space-time processing by means of antenna arrays are becoming ever more attractive as a technology to improve the system performance, especially for reduction of multipath effects. This paper presents a new low complexity and high accuracy algorithm to estimate the multipath delays and direction of arrivals (DOAs) simultaneously in wireless communication systems. By using separable dimension correlation processing, the temporal and spatial signal subspaces are formed and the joint two dimensional delay/DOA estimation problem is separated into two simpler one dimensional estimations.

1. INTRODUCTION

Three major performance and capacity limiting impairments in current mobile communication systems are: multipath fading, intersymbol interference (ISI) and co-channel interference (CCI). Especially, the ISI impairment resulting from delay spread constrains the maximum data rate. Current mobile communication systems, using temporal processing alone, cannot effectively address these impairments. By means of an antenna array, a combination of temporal and spatial processing can potentially yield good performance improvements over existing systems. Several joint delay and direction estimation algorithms for signals in multipath environments have thus been developed recently [1, 2, 3].

This paper proposes a new low complexity and high accuracy algorithm based on a separable dimension subspace method [7] to estimate the multipath delays and direction of arrivals (DOAs) simultaneously. With separable dimension processing, a joint spatial and temporal estimation problem is separated, i.e., the delays are first estimated by using a one-dimensional subspace method and then the DOAs are estimated for each estimated delay. In this way, the computational complexity of the proposed method is reduced while its performance for the joint delay/DOA estimation is still satisfied as supported by computer simulations.

2. PROBLEM FORMULATION

Consider a base station receiving array composed of M antennas and assume that the single user signal of interest arrives at the base station via D paths, with the DOA of the ith path denoted as θ_i $(i = 1, 2, \cdots, D)$. Then, the received complex baseband signal vector at the antenna array can be described as:

$$\mathbf{x}(t) = \sum_{i=1}^{D} \mathbf{a}(\theta_i)\, \beta_i\, r(t - \tau_i) \; + \; \mathbf{n}(t) \qquad (1)$$

where $\mathbf{a}(\theta_i)$ is an $M \times 1$ spatial steering vector for the ith path, β_i is the complex fading factor of the ith ray, $r(t)$ is a transmitted complex baseband signal, τ_i is the ith path propagation delay and $\mathbf{n}(t)$ is a spatially and temporally white additive Gaussian noise with zero mean and equal covariance σ_n^2.

In a linear time-invariant system, the transmitted signal $r(t)$ can be represented as a convolution of the data bits and a pulse shaping function $g(t)$. i.e. $r(t) = \sum_l s_l \cdot g(t - lT_s)$. Therefore, by passing the signal vector $\mathbf{x}(t)$ through a set of tapped-delay lines (TDL) of length Q and delay T_0, as shown in Figure 1, and sampling the resulting outputs, a data matrix $\mathbf{X}[n]$ is formed as:

$$\begin{aligned} \mathbf{X}[n] &= [\boldsymbol{x}_1{}^T[n] \;\; \boldsymbol{x}_2{}^T[n] \;\; \cdots \;\; \boldsymbol{x}_M{}^T[n]]^T \\ &= \sum_{i=1}^{D} \mathbf{a}(\theta_i)\, \beta_i \otimes [\mathbf{G}(\tau_i) \cdots \mathbf{G}(\tau_i + LT_s)]\mathbf{s}[n] + \mathbf{N}[n] \end{aligned} \qquad (2)$$

where symbol $\otimes$ denotes the Kronecker product, $\mathbf{G}(\tau_i) = [g(t_0 - \tau_i),\, g(t_0 - T_0 - \tau_i),\, \cdots,\, g(t_0 - (Q-1)T_0 - \tau_i)]^T$ is referred to as the temporal manifold, $g(\cdot)$ is the pulse shaping function, which models the total impulse response of the filters used in the system, t_0 is the sampling reference time of the nth data, $\mathbf{s}[n] = [s(n)s(n-1)\cdots s(n-L)]^T$ is a vector consisting of $L+1$ consecutive symbols, T_s is symbol duration and L is the length of the channel, which covers the range of delays $\{\tau_i\}_{i=1}^{D}$. We assume that

a training sequence is embedded in the transmitted signal; the training portion, represented here by $\mathbf{s}[n]$, can be extracted by receiver and is assumed to be known. Typically, in TDMA systems, $L = L_g + \tau_{max}/T_s$, where $L_g T_s$ is the duration of the pulse shaping function $g(t)$ and τ_{max} is the maximum integer delay [4]. Likewise, in DS-CDMA systems, $L = 2\mathcal{N}_c$, where $\mathcal{N}_c = T_{cs}/T_c$, T_{cs} is the data symbol period and T_c is the chip duration [5].

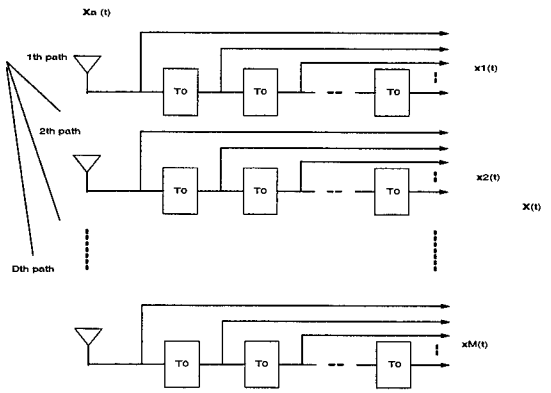


Fig. 1. Array front-end with TDLs

The equation (2) can be rewritten as:

$$\mathbf{X}[n] = \sum_{i=1}^{D} \mathbf{a}(\theta_i)\,\beta_i \otimes \begin{bmatrix} \tilde{g}(t_0 - \tau_i) \\ \vdots \\ \tilde{g}(t_0 - (Q-1)T_0 - \tau_i) \end{bmatrix} + \mathbf{N}[n]$$

$$= [\mathbf{a}(\theta_1)\ \cdots\ \mathbf{a}(\theta_D)] \diamond [\tilde{\mathbf{G}}(\tau_1)\ \cdots\ \tilde{\mathbf{G}}(\tau_D)] \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_D \end{pmatrix} + \mathbf{N}[n]$$

$$= \mathbf{D}(\boldsymbol{\theta}, \boldsymbol{\tau})\ \mathbf{B}\ +\ \mathbf{N}[n] \tag{3}$$

where $\diamond$ denotes the Khtari-Rao product (see [6]), which represents column-Kronecker product;

$$\tilde{\mathbf{G}}(\tau_i) = [\mathbf{G}(\tau_i)\ \cdots\ \mathbf{G}(\tau_i + LT_s)]\,\mathbf{s}[n]$$

is a modified temporal manifold, which is the convolution between the training sequence and delayed shaping function;

$$\mathbf{D}(\boldsymbol{\theta}, \boldsymbol{\tau})\ =\ [\mathbf{a}(\theta_1)\ \cdots\ \mathbf{a}(\theta_D)] \diamond [\tilde{\mathbf{G}}(\tau_1)\ \cdots\ \tilde{\mathbf{G}}(\tau_D)]$$

is a spatio-temporal manifold and $\mathbf{B} = [\beta_1\ \beta_2\ \cdots\ \beta_D]^T$.

Thus, based on the available samples $\{\mathbf{X}[n]\}_{n=1}^{N}$, the problem of interest here is to estimate the DOAs θ_i and the multipath delays τ_i $(i = 1, 2, \cdots, D)$ simultaneously with subspace methods.

3. SUBSPACE PARTITION

Assume that the number of paths D, the maximum path delay τ_{max}, the array response $\mathbf{a}(\cdot)$ and the pulse shaping function $g(\cdot)$ are known. Also assume that the complex path fading factor $\{\beta_i\}_{i=1}^{D}$ remain constant during a data symbol period.

Define

$$\mathbf{r}'_h = \tfrac{1}{M} \sum_{l=1}^{M} E[x_l(nT_s - h \cdot T_0)\mathbf{x}_l^H(n)], \qquad h = 0, 1, \cdots, D-1 \tag{4}$$

$$\boldsymbol{\eta}'_l = \tfrac{1}{Q} \sum_{h=0}^{Q-1} \mathrm{E}[x_l(nT_s - h \cdot T_0)\mathbf{Y}^H(nT_s - h \cdot T_0)], \qquad l = 1, 2, \cdots, D \tag{5}$$

where

$$\mathbf{Y}(t) = [x_1(t) \cdots x_M(t)]^T$$

and $E(\cdot)$ denotes mathematical expectation. We refer to $\{\mathbf{r}'_h\}_{h=0}^{D-1}$ and $\{\boldsymbol{\eta}'_l\}_{l=1}^{D}$ as the set of temporal vectors and spatial vectors, respectively. It can be shown that the linear space spanned by these sets of vectors are equal to the range space of $\tilde{\mathbf{G}}(\boldsymbol{\tau})$ and $\mathbf{a}(\boldsymbol{\theta})$, respectively. That is, let $\mathbf{r} = [\mathbf{r}'_1, \mathbf{r}'_2, \cdots, \mathbf{r}'_D]$ and $\boldsymbol{\eta} = [\boldsymbol{\eta}'_1, \boldsymbol{\eta}'_2, \cdots, \boldsymbol{\eta}'_D]$, then we have

$$\mathcal{R}(\mathbf{r}) = \mathcal{R}(\tilde{\mathbf{G}}(\boldsymbol{\tau})) \qquad \mathcal{R}(\boldsymbol{\eta}) = \mathcal{R}(\mathbf{a}(\boldsymbol{\theta})) \tag{6}$$

where $\mathcal{R}(\cdot)$ denotes the range space of its matrix argument, $\tilde{\mathbf{G}}(\boldsymbol{\tau}) = [\tilde{\mathbf{G}}(\tau_1)\ \cdots\ \tilde{\mathbf{G}}(\tau_D)]$ and $\mathbf{a}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1)\ \cdots\ \mathbf{a}(\theta_D)]$.

4. SEPARABLE DIMENSIONAL ALGORITHM

Equation (6) suggests that a subspace method may be used independently to estimate DOA and delay parameters. Specifically, we may use correlation processing in spatial and temporal dimension respectively to get the estimates of $\{\mathbf{r}'_h\}_{h=0}^{D-1}$ and $\{\boldsymbol{\eta}'_l\}_{l=1}^{D}$; These estimates are then used separately to generate null subspace projections. Finally, the path delay and DOAs can be estimated with subspace methods by two one-dimensional searches. This leads to the following algorithm:

Step 1. Formation of temporal and spatial projection matrices

(1) Estimation of temporal vectors:

$$\widehat{\mathbf{r}}'_h = \frac{1}{M} \sum_{l=1}^{M} \frac{1}{N} \sum_{n=1}^{N} x_l(nT_s - h \cdot T_0)\mathbf{X}_l^H(n) \tag{7}$$

$$\widehat{\mathbf{r}}_h = (\widehat{\mathbf{r}}'_h - \widehat{\sigma}^2 \mathbf{e}_h)^H, \quad h = 0, 1, \cdots, D-1 \tag{8}$$

where $\mathbf{e}_h = [\underbrace{0\cdots0}_{h}10\cdots0]$ and $\hat{\sigma}^2$ is an estimate of the noise variance.

(2) Gram-Schmidt (GS) orthogonalization and formation of temporal projection matrix $\mathbf{P}_\tau$: From the vectors $\{\hat{\mathbf{r}}_h\}_{h=0}^{D-1}$, we can get D orthogonal vectors, $\{\mathbf{q}_k\}_{k=1}^{D}$ via GS orthogonalization. Let $\mathbf{Q}_\tau = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_D]$, then compute the temporal projection matrix $\mathbf{P}_\tau = \mathbf{I} - \mathbf{Q}_\tau\mathbf{Q}_\tau^H$, which spans the null space of $\{\tilde{\mathbf{G}}(\tau_k)\}_{k=1}^{D}$.

(3) Estimation of spatial vectors:

$$\hat{\eta}_l' = \frac{1}{Q}\sum_{h=0}^{Q-1}\frac{1}{N}\sum_{n=1}^{N} x_l(nT_s - h\cdot T_0)\mathbf{Y}^H(nT_s - h\cdot T_0) \tag{9}$$

$$\hat{\eta}_l = (\hat{\eta}_l' - \hat{\sigma}^2\mathbf{e}_l)^H, \qquad l = 1, 2, \cdots, D \tag{10}$$

(4) Gram-Schmidt (GS) orthogonalization and formation of spatial projection matrix $\mathbf{P}_\theta$: Via Gram-Schmidt orthogonalization of $\{\hat{\eta}_l\}_{l=1}^{D}$, the D orthogonal vectors, $\{\zeta_l\}_{l=1}^{D}$ and spatial orthogonal projection matrix $\mathbf{P}_\theta = \mathbf{I} - \mathbf{Q}_\theta\mathbf{Q}_\theta^H$ are obtained, where $\mathbf{Q}_\theta$=[ζ_1 ζ_2 $\cdots$ ζ_D].

Step 2. Multipath delays and direction of arrivals estimation

The path delays $\{\tau_k\}_{k=1}^{D}$ are estimated as the D largest peaks of the function $P(\tau) = (\tilde{\mathbf{G}}^H(\tau)\mathbf{P}_\tau\tilde{\mathbf{G}}(\tau))^{-1}$, searching over the delay sector of interest, measured by symbol period T. Likewise, the DOAs $\{\theta_k\}_{k=1}^{D}$ are estimated by searching over the direction sector of interest to get the D largest peaks of the function $P(\theta) = (\mathbf{a}^H(\theta)\mathbf{P}_\theta\mathbf{a}(\theta))^{-1}$.

Step 3. Delay and DOA pairing

With estimated delays $\hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_D$, if we select the estimated DOAs $\hat{\theta}_i$ $(i = 1, 2, \cdots, D)$ to minimize the cost function $\mathcal{L}$, then the delays $\hat{\tau}_i$ and the DOAs $\hat{\theta}_i$ can be paired. The cost function $\mathcal{L}$ is:

$$\mathcal{L} = \mathbf{D}^H(\hat{\theta}, \hat{\tau})\mathbf{E}_n\mathbf{E}_n^H\mathbf{D}(\hat{\theta}, \hat{\tau}) \tag{11}$$

where $\mathbf{D}(\hat{\theta}, \hat{\tau}) = \mathbf{a}(\hat{\theta}) \diamond \tilde{\mathbf{G}}(\hat{\tau})$ is joint spatial-temporal vector and $\mathbf{E}_n$ is a matrix whose columns are the eigenvectors corresponding to the smallest eigenvalues of covariance matrix $\mathbf{R}_x = E\,[\,\mathbf{X}[n]\,\mathbf{X}^H[n]\,]$.

5. ESTIMATION OF UNKNOWN NOISE COVARIANCE

In the case of unknown noise covariance, the performance of separable dimension subspace estimation method will be degraded. In equation (7) and equation (8), we can see that the vector $\mathbf{r}_h$ will be noise free when the TDL delay T_0 is greater than the correlation time of noise. Therefore, to improve the estimation performance in the case of unknown noise covariance, we can change the length of tapped-delay-lines (TDL) from Q to $2Q$ ($Q \geq D$) to estimate unknown noise covariance σ^2 and then remove it from temporal/spatial vectors.

Let the first Q TDL outputs of the lth sensor be represented by the vector $\mathbf{x}_l = [x_l(nT_s), x_l(nT_s - T_0), \cdots, x_l(nT_s - (Q-1)T_0)]^T$ and the later Q TDL outputs by the vector $\bar{\mathbf{x}}_l = [x_l(nT_s - QT_0), x_l(nT_s - (Q+1)T_0), \cdots, x_l(nT_s - (2Q-1)T_0)]^T$. Then, the sample covariance matrix for cross $\mathbf{x}_l$ and $\bar{\mathbf{x}}_l$ is

$$\hat{\mathbf{R}}_l = \frac{1}{N}\sum_{n=1}^{N}[\mathbf{x}_l(n)\,\bar{\mathbf{x}}_l(n)^H] \tag{12}$$

We can estimate the temporal vector $\mathbf{r}_h$

$$\hat{\mathbf{r}}_h = \frac{1}{M}\sum_{l=1}^{M}\frac{1}{N}\sum_{n=1}^{N} x_l(nT_s - (h-1)T_0)\bar{\mathbf{x}}_l^H(n) \tag{13}$$

With the method described in Section 4 and the estimated vector $\hat{\mathbf{r}}_h$, we can estimate the multipath delay $\hat{\tau}_i$, $i = 1, 2, \cdots, D$ with equation $P(\tau) = (\tilde{\mathbf{G}}^H(\tau)\mathbf{P}_\tau\tilde{\mathbf{G}}(\tau))^{-1}$.

To estimate the signal covariance matrix $\mathbf{R}_s$, we can reconstruct the modified temporal manifold $\tilde{\mathbf{G}}(\hat{\tau}_i)$ and $\tilde{\mathbf{G}}'(\hat{\tau}_i)$ with the estimated delay, $\hat{\tau}_i$, herein $\tilde{\mathbf{G}}'(\hat{\tau}_i) = [\tilde{g}(t_0 - QT_0 - \hat{\tau}_i), \cdots, \tilde{g}(t_0 - 2QT_0 - \hat{\tau}_i)]^T$, and have

$$\hat{\mathbf{R}}_s = [\,\tilde{\mathbf{G}}^H(\hat{\tau})\tilde{\mathbf{G}}(\hat{\tau})\,]^{-1}\tilde{\mathbf{G}}^H(\hat{\tau})\hat{\mathbf{R}}_l\tilde{\mathbf{G}}'(\hat{\tau})[\,\tilde{\mathbf{G}}'^H(\hat{\tau})\tilde{\mathbf{G}}'(\hat{\tau})\,]^{-1} \tag{14}$$

With estimated signal covariance matrix $\hat{\mathbf{R}}_s$ and $\tilde{\mathbf{G}}(\hat{\tau})$, the noise covariance matrix can be estimated by

$$\sigma^2\mathbf{I} = \mathbf{R}_l' - \tilde{\mathbf{G}}(\hat{\tau})\,\mathbf{R}_s\,\tilde{\mathbf{G}}^H(\hat{\tau}) \tag{15}$$

where $\mathbf{R}_l' = E[\mathbf{x}_l(n)\mathbf{x}_l(n)^H]$.

6. COMPUTER SIMULATIONS

We assume that a transmitted signal with $D = 3$ paths arrives at a linear array of $M = 6$ sensors with half-wavelength spacing. The multipath delays are $[0,\ 0.5,\ 1.2]\,T$, $T = 1$, and the direction of arrivals are $[15°,\ 40°,\ 70°]$. The path fadings are $[1, 0.85, 0.8]$. Additive white Gaussian noise is added, the corresponding SNR = 10 dB. 100 samples are accumulated. The pulse shape function is a raised cosine with 0.35 excess bandwidth, the TDL length is $Q = 6$ and delay $T_0 = 0.5$.

Fig. 2 and Fig 3 show the estimation results of multipath delays and DOAs with 30 trials by using the proposed separable dimension subspace method. Simulations and performance comparison with other proposed algorithms such as JADE [2] are presented in Fig. 4 and Fig. 5.

7. CONCLUSIONS

A new algorithm based on separable dimension subspace method is proposed for joint estimation of DOAs and multipath delays in the wireless communication systems. In this paper, spatial and temporal separable dimension correlation processing are used to replace EVD (Eigen Value Decomposition) or SVD (Singluar Value Decomposition). Therefore, compared to other joint DOA and delay estimation methods, the computational complexity of the proposed method is relatively small. The presented algorithm has been tested by computer simulation studies and has been found to perform satisfactorily.

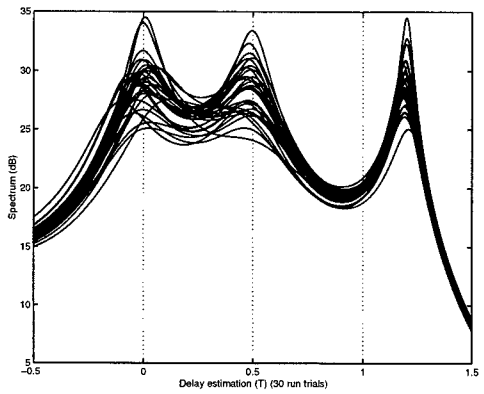


Fig. 2. Delay estimation with 30 trials

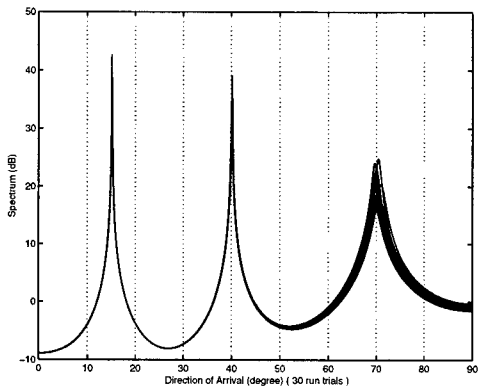


Fig. 3. DOA estimation with 30 trials

8. REFERENCES

[1] A. J. van der Veen, M. C. Vanderveen, and A. Paulraj "Joint angle and delay estimation using shift invariance techniques ". *IEEE Trans. on Signal Processing*, **46**(2):405 - 418, Feb. 1998.

[2] M. C. Vanderveen, B. C. Ng, C. B. Papadias and A. Paulraj, "Joint angle and delay estimation (JADE) for signals in multipath environments". *30th Asilomar Conf. on Circuit, Systems and Computer*, pp. 1250 - 1254, Pacific Grove, USA, November 1996.

[3] P. Pelin, "Space-time algorithms for mobile communications". *Ph.D thesis*, Chalmers University of Technology, Gothenburg, Sweden, 1999.

[4] M. C. Vanderveen, A. J. van der veen, and A. Paulraj, "Estimation of multipath parameters in wireless communications". *IEEE Trans. Signal Processing*, **46** (3):682-690, March 1998.

[5] L. Huang and A. Manikas, "Blind single-user array receiver for MAI cancellation in multipath fading DS-CDMA channels ". *European Signal Processing Conference (EUSIPCO'00)*, Tampere, Finland, Sept. 2000.

[6] A. J. van der veen, "Algeberic methods for deterministic blind beamforming". *Proceedings of The IEEE* , **86** (10):1987-2008, October 1998.

[7] J. Mao, B. Champagne, and M. O'Droma, "Separable dimension subspace method for joint signal frequencies, DOAs, and sensor mutual coupling estimation ". *34th Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, USA, Oct. 2000.

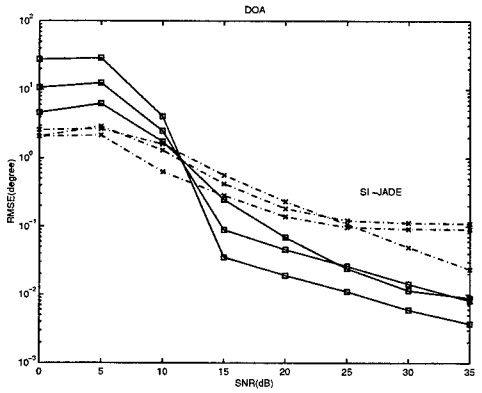


Fig. 4. DOA estimation performance

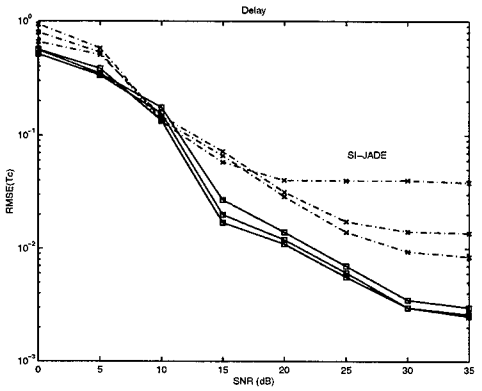


Fig. 5. Delay estimation performance

HIGH-RESOLUTION DIRECTION FINDING USING A SWITCHED PARASITIC ANTENNA

Thomas Svantesson[1]

Department of Signals and Systems,
Chalmers University of Technology,
tomaso@s2.chalmers.se

Mattias Wennström[2]

Signals and Systems Group
Uppsala University
mw@signal.uu.se

ABSTRACT

Direction finding by exploiting the directional radiation patterns of a Switched Parasitic Antenna (SPA) is considered. By employing passive elements (parasites), which can be shorted to ground using pin diodes, directional radiation patterns can be obtained. The direction finding performance of the SPA is examined by calculating a lower bound on the direction finding accuracy, the Cramér-Rao lower Bound (CRB). It is found that the SPA offers a compact implementation with high-resolution direction finding performance using only a single radio receiver. Thus, exploiting SPAs for direction finding is an interesting alternative to traditional antenna arrays offering compact and low-cost antenna implementations.

1. INTRODUCTION

Direction finding is of great importance in a variety of applications, such as radar, sonar, communications, and recently also personal locating services. In the last two decades, direction finding and sensor array processing has attracted considerable interest in the signal processing community. The focus of this work has been on high resolution, i.e. a resolution higher than the width of the main lobe, Direction Of Arrival (DOA) estimation algorithms [3]. These algorithms exploit the fact that an electromagnetic wave that is received by an array of antenna elements reaches each element at different time instants. Although the performance of these systems is excellent, an unfortunate aspect is the high costs of employing a radio receiver for each antenna element. Furthermore, it is expensive to calibrate and maintain antenna arrays with many antenna elements.

Recently, it was proposed to employ an SPA for direction finding [6, 7] that only uses a single active radio receiver, thereby significantly reducing the cost. The SPA offers characteristics similar to an array antenna with several beams by using passive antenna elements that serve as reflectors when shorted to ground. Different directional patterns can be achieved by switching the short-circuits of the passive elements using pin diodes. The possibilities of exploiting these patterns for high-resolution DOA estimation will be examined in this paper, since no attempt to employ high-resolution DOA methods was undertaken in [6, 7].

2. SWITCHED PARASITIC ANTENNA

Switched Parasitic Antennas offering directional patterns dates back to the early work of Yagi and Uda in the 1930's [1]. The concept is to use a single active antenna element, connected to a radio transceiver, in a structure with one or several passive antenna elements, operating near resonance. These passive elements are called Parasitic Element (PE)s and act together with the active element to form an array, as in the well known Yagi-Uda array [1]. To alter the radiation pattern, the termination impedances of the PEs are switchable, to change the current flowing in those elements. The PEs become reflectors when shorted to the ground plane using pin diodes [8] and when not shorted, the PEs have little effect on the antenna characteristics. The receiver is always connected to the center antenna element so there are no switches in the RF direct signal path.

An interesting possibility to obtain directional information is to sample the received signal with several different radiation patterns, since the switching time of a pin diode is only of the order of a few nanoseconds. This technique of oversampling the received signal is common in many communication systems, but here the oversampling is performed in both time and space, i.e. spatio-temporal oversampling. If the increased sampling rate (or bandwidth) poses a problem, a bandpass sampling strategy could also be employed. In this paper, the potential in using the different radiation patterns of an SPA for direction finding will

[1]This work was supported in part by the Swedish Foundation for Strategic Research, under the Personal Computing and Communications Program.

[2]This work was supported in part by NUTEK.

0-7803-7011-2/01/$10.00 ©2001 IEEE

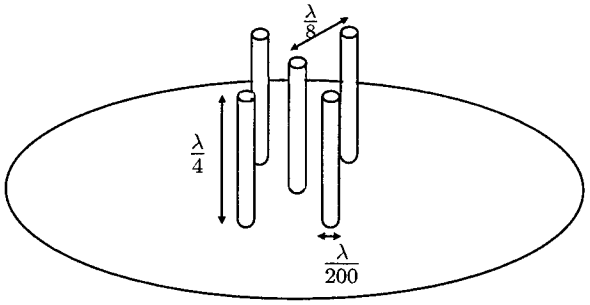


Figure 1: A five element monopole SPA. The center element is active and connected to the transceiver. The four passive antenna elements can be switched in or out of resonance using appropriately biased pin diodes.

be examined. However, further work is needed on the practical aspects of the antenna design as well as sampling strategies.

In the literature, it has been proposed to use monopoles on a ground plane [8] or patch antennas [5] as SPAs. In this paper, a monopole on a ground plane is used because of its omnidirectional properties. A four Direction Symmetry (4-DS) monopole parasitic antenna is shown in Figure 1 and a three Direction Symmetric (3-DS) antenna is shown in Figure 2. The antenna in Figure 2 has an additional circle of parasitic elements that always are shorted to ground. The effect of this arrangement is an increased directivity as their length are shorter than the corresponding resonant length ($\approx \lambda/4$) and will lead the induced Electro-Motive Force (EMF) [1]. The lengths and distances displayed in Figure 1 and 2 are not optimal in any way. Note that the resulting antennas are very compact ($\lambda/4$ & $\lambda/2$) compared to corresponding linear arrays with $\lambda/2$ separation distance (2λ & $3\lambda/2$).

The antennas in Figures 1 and 2 were simulated using HFSS (High Frequency Structure Simulator) from Agilent Technologies Inc. which is a 3D simulator using the Finite Element Method (FEM) to solve for the electromagnetic field. The software was used to calculate the far-field radiation pattern of the antenna for different settings of the switched parasitics. The monopole elements were cylindrical with a length to radius ratio $l/r = 100$ which yields a first resonance at approximately 0.24λ [1].

The far-field power radiation pattern in the azimuth plane $F(\phi)$ for the 4-DS SPA is shown in Figure 3. The corresponding pattern for the 3-DS SPA is similar and not shown. The directivity of the two antennas are 9.9 dB and 10.0 dB respectively. Thereby, only a small gain in directivity was achieved by the extra ring of shorted parasites.

Once the far-field radiation properties are found, it

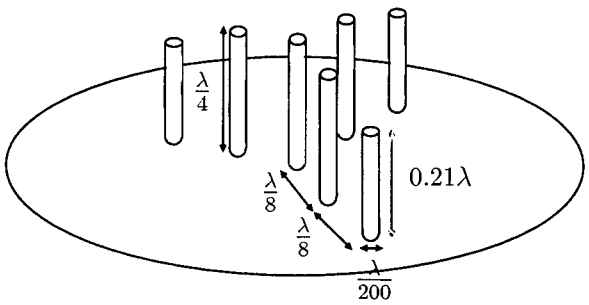


Figure 2: A seven element monopole SPA. The center element is active and connected to the transceiver. The three passive antenna elements closest to the active can be switched in or out of resonance using appropriately biased pin diodes. The three outermost monopoles are hardwired to ground.

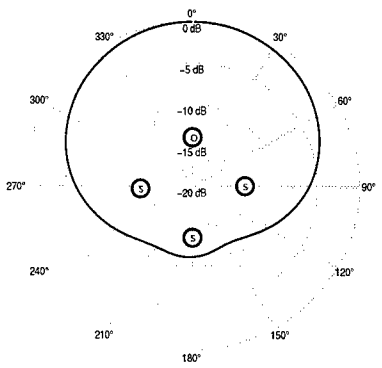


Figure 3: Power radiation pattern of the five element monopole antenna shown in Figure 1 with three parasitics shorted (S) to ground and one open (O).

is straightforward to derive a model for the received voltages [2]. If p waves are incident upon an antenna with M symmetry directions, the received voltages can be written in matrix form as

$$\mathbf{x}(t) = \mathbf{A}(\boldsymbol{\phi})\mathbf{s}(t) + \mathbf{e}(t), \qquad (1)$$

where the vector of measured voltages $\mathbf{x}(t)$ is $M \times 1$. The matrix $\mathbf{A}(\boldsymbol{\phi})$ $(M \times p)$ corresponds to the response of the different symmetry directions and has elements $[\mathbf{A}(\boldsymbol{\phi})]_{qr} = F(\phi_r + 2q\pi/M)$. This matrix is often called steering matrix in the sensor array processing literature. The signal vector $\mathbf{s}(t)$ is $p \times 1$ and contains the strength of the received fields. Finally, the noise vector $\mathbf{e}(t)$ is $M \times 1$.

In order for the analysis in the following sections to be valid, some additional assumptions are needed:

- the steering matrix has full rank, i.e., $\mathrm{rk}(\mathbf{A}) = p$
- $\mathbf{e}(t)$ is temporally white and circularly Gaussian distributed: $\mathbf{e}(t) \in \mathcal{N}(0, \sigma^2\mathbf{I})$

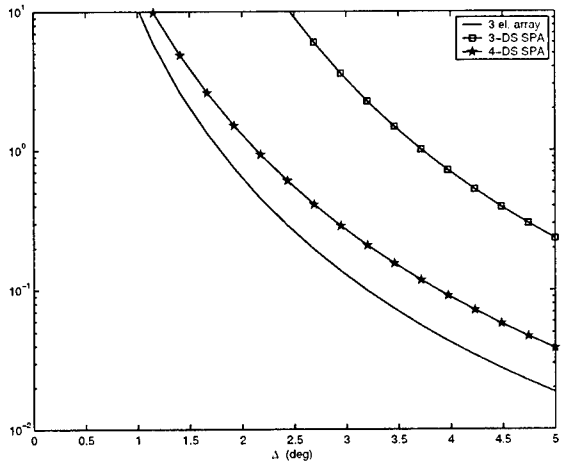


Figure 4: The square root of the CRB for the configurations in Figure 1 and 2 when two waves are incident from $(30°, 30°+\Delta)$ with SNR=10dB and 1000 samples.

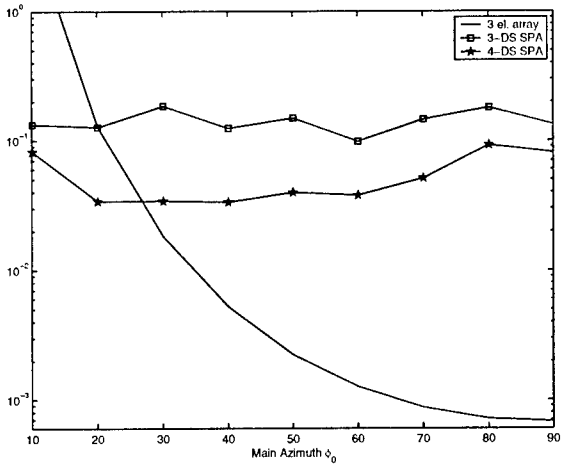


Figure 5: The square root of the CRB for the configurations in Figure 1 and 2 when two waves are incident from $(\phi_0, \phi_0 + 5°)$ with SNR=10dB and 1000 samples.

- $\mathbf{s}(t)$ is also temporally white and circularly Gaussian distributed: $\mathbf{s}(t) \in \mathcal{N}(0, \mathbf{S})$

The noise is both spatially and temporally white, while the signal is only assumed to be temporally white. Furthermore, the signal is assumed to be uncorrelated with the noise.

3. DIRECTION FINDING PERFORMANCE

The data model (1) is identical to the usual data model used in sensor array processing [3], except for a new steering matrix. This will of course change the direction finding properties. Before the properties of a specific DOA estimation scheme is studied, a lower bound, the Cramér-Rao lower Bound (CRB), on the variance of the DOA estimates will be analyzed. Note that it is possible to asymptotically achieve this bound with many methods in the literature [3].

Expressions for the CRB was derived for an array of antenna elements in [4]; and can also be applied to the parasitic antenna by changing the steering matrix.

$$E\{(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)^T\} \geq \mathbf{B} \qquad (2)$$

$$\mathbf{B} = \frac{\sigma^2}{2N}\left[Re\{(\mathbf{D}^H\mathbf{P}_A^{\perp}\mathbf{D}) \odot (\mathbf{S}\mathbf{A}^H\mathbf{R}^{-1}\mathbf{A}\mathbf{S})^T\}\right]^{-1},$$

where the elements of $\mathbf{D}_{qr} = \frac{\partial F(\phi + 2q\pi/M)}{\partial \phi}\Big|_{\phi=\phi_r}$. Furthermore, $\odot$ denotes the Hadamard (or Schur) product, i.e., element-wise multiplication and $\mathbf{P}_A^{\perp} = \mathbf{I} - \mathbf{P}_A = \mathbf{I} - \mathbf{A}\mathbf{A}^{\dagger}$ [1] is the orthogonal projector onto the null space of $\mathbf{A}^H$. The matrix $\mathbf{R} = \mathbf{A}\mathbf{S}\mathbf{A}^H + \sigma^2\mathbf{I}$ is the covariance matrix of the measured voltages $\mathbf{x}(t)$ and N denotes the number of time samples.

The square root of the CRB, i.e. the standard deviation, is shown in Figure 4 for the antenna configurations in Figure 1 and 2 as two waves are incident from $(30°, 30° + \Delta)$. Only the CRB for the first DOA, i.e. the wave arriving from 30°, is shown since the CRB for the second DOA will behave similarly. The standard deviation for a uniform linear array of three elements spaced $\lambda/2$ apart is compared to the 4-DS and 3-DS SPAs. As expected, the performance is better when using four rather than the three symmetry directions. Also, note that the three element array performs slightly better the 4-DS SPA. However, these results depend on the incidence angles, since the array will work best for broadside and worst for end-fire incidence.

In Figure 5, the standard deviation is shown for the same antenna configurations as in Figure 4 when two waves are incident from $(\phi_0, \phi_0 + 5°)$. The parasitic antenna, due to its symmetrical properties, offers similar direction finding performance properties for all incidence angles. The linear array performs worse than the parasitic antenna at end-fire incidence, while performing much better at broad-side incidence. However, for many direction finding applications, the direction finding performance of the parasitic antenna is sufficient and the cost reduction of using only a single radio receiver outweighs the loss in performance for broad-side angles. It should also be stressed that the antenna designs in Figure 1 and 2 are by no means optimal and better DOA properties may be obtained by a proper optimization.

[1] $\mathbf{M}^{\dagger}$ is the Moore-Penrose pseudo inverse of M.

4. ESTIMATION METHODS

The analysis in the previous section was based on the CRB on the estimation error. In this section, algorithms that approximately achieve this lower bound will be discussed. In principle, all DOA estimation schemes derived for a general antenna array can also be applied to a parasitic antenna by inserting a new steering matrix. For an overview of DOA estimation methods, see [3].

In [9], a popular high resolution DOA estimation method, MUltiple SIgnal Classification (MUSIC), was introduced where the DOA estimates are taken as those ϕ that maximizes the MUSIC criterion function

$$\hat{\phi} = \arg\max_{\phi} \frac{\mathbf{a}^H(\phi)\mathbf{a}(\phi)}{\mathbf{a}^H(\phi)\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^H\mathbf{a}(\phi)}, \qquad (3)$$

where the steering vector $\mathbf{a}_q(\phi) = F(\phi + 2q\pi/M)$. Usually this is formulated as finding the p largest peaks in the "MUSIC spectrum". Here, $\hat{\mathbf{E}}_n$ denotes the $M-p$ eigenvectors corresponding to the $M-p$ smallest eigenvalues of the estimated covariance matrix $\hat{\mathbf{R}}$. A typical example of a MUSIC spectrum is shown in Figure 6, where two waves are incident from 25° and 45° upon a 4-DS SPA and a three element array with SNR=10dB and 1000 samples. This figure indicates that the SPA, in this case, offers a high-resolution direction finding performance similar to that of an antenna array without the cost of many radio receivers. Most other DOA estimation schemes [3] can also be applied to SPAs with similar results. For instance, the Stochastic Maximum Likelihood (SML) algorithm [4] for this type of antenna was implemented. The RMSE of the ML estimator achieved the CRB bound from Section 3, as expected.

5. CONCLUSIONS

The potential use of a Switched Parasitic Antenna for high-resolution direction finding was investigated. By employing passive elements, which can be shorted to ground using pin diodes, directional radiation patterns are obtained that can be used successfully to estimate DOAs. The main advantage with this concept is that only one radio receiver is needed, thereby reducing the costs significantly compared to traditional antenna arrays where one radio receiver per element typically is employed. Another advantage of the SPA is that a very compact implementation of the antenna is possible.

A data model for the SPA was presented and the direction finding performance was examined by calculating the CRB and the MUSIC estimator. It was found that the SPA offers a compact implementation

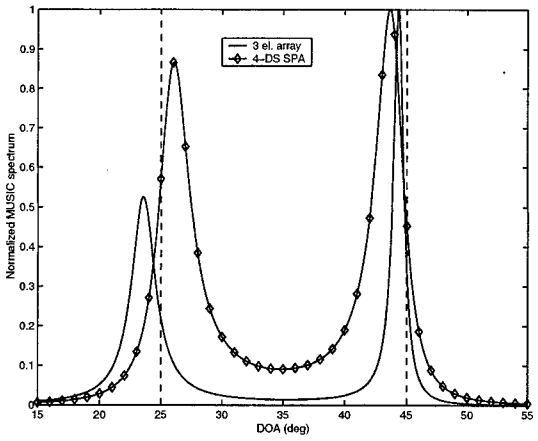


Figure 6: The normalized MUSIC spectrum when two waves are incident from 25° and 45° upon a 4-DS parasitic antenna and a three element array with SNR=10dB and 1000 samples.

with high-resolution direction finding performance using only a single radio receiver. Thus, exploiting SPAs for direction finding is an interesting alternative that offers several advantages over traditional arrays.

6. REFERENCES

[1] C. Balanis. *Antenna Theory: Analysis and Design.* John Wiley and sons, Inc., New York, 1982.

[2] P-S. Kildal. "Equivalent Circuits of Receive Antennas in Signal Processing Arrays". *Microwave and Optical Technology Letters*, 21(4):244–246, May 1999.

[3] H. Krim and M. Viberg. "Two Decades of Array Signal Processing Research: The Parametric Approach". *IEEE Signal Processing Magazine*, 13(4):67–94, July 1996.

[4] B. Ottersten, B. Wahlberg, M. Viberg, and T. Kailath. "Stochastic Maximum Likelihood Estimation in Sensor Arrays by Weighted Subspace Fitting". In *Proc. 23rd Asilomar Conf. Sig., Syst.,Comput.*, pages 599–603, Monterey, CA, November 1989.

[5] S.L. Preston and D.V. Thiel. "Direction Finding Using a Switched Parasitic Antenna Array". In *Proc. IEEE 1997 AP-S*, pages 1024–1027, Montreal, July 1997.

[6] S.L. Preston, D.V. Thiel, J.W. Lu, S.G. O'Keefe, and T.S. Bird. "Electronic Beam Steering Using Switched Parasitic Patch Elements". *Elect. Lett.*, 33:7–8, 1997.

[7] S.L. Preston, D.V. Thiel, T.A. Smith, S.G. O'Keefe, and J.W. Lu. "Base-Station Tracking in Mobile Communications Using a Switched Parasitic Antenna Array". *IEEE Trans. Antennas Propagat.*, 46(6):841–844, 1998.

[8] R. Schlub, D.V. Thiel, J.W. Lu, and S.G. O'Keefe. "Dual-Band Six-Element Switched Parasitic Array For Smart Antenna Cellular Communications Systems". *Electronic Letters*, 36:1342–1343, 2000.

[9] R.O. Schmidt. "Multiple Emitter Location and Signal Parameter Estimation". In *Proc. RADC Spectrum Estimation Workshop*, pages 243–258, Rome, NY, 1979.

TWO-STEP MUSIC ALGORITHM FOR IMPROVED ARRAY RESOLUTION

R. Chavanne

Office National d'Études et
de Recherches Aérospatiales,
DEMR/TSI BP72,
F92322 Châtillon Cedex, France
Email : chavanne@onera.fr

K. Abed-Meraim

Ecole Nationale Supérieure
des Télécommunications
de Paris TSI 46, rue Barrault
75013 Paris France
Email : abed@tsi.enst.fr

D. Médynski

Office National d'Études et
de Recherches Aérospatiales,
DEMR/TSI BP72,
F92322 Châtillon Cedex, France
Email : medynski@onera.fr

ABSTRACT

High resolution methods such as MUSIC fail to separate closely spaced sources in difficult contexts (low SNR, short sample size,...). Halder *et* al.(1997) have applied an interleaving technique to improve the resolution as well as the performances in the case of frequencies estimation. Here we extend this work and deal with the application of this technique to array processing. We aim to estimate closely spaced DOAs. After a first estimation with MUSIC, a second step of the algorithm consists in refining the angle resolution using downsampled covariance matrices together with a Joint Estimation Strategy (JES) similar to that proposed by Gershman *et* al. (1996). This method improves MUSIC performances especially for low SNRs. Simulations examples are provided to illustrate the performance of the proposed method referred to as Two Step-MUSIC (TS-MUSIC).

1. INTRODUCTION

Direction of Arrival (DOA) estimation is a recurrent problem in array processing that can be treated with high resolution methods such as MUSIC (MUltiple Signal Classification) [4]. Unfortunately, these methods are less efficient as the DOAs come closer. Recently, Halder *et* al. have suggested a temporal downsampling technique that improves the frequencies estimation of subspace-based methods [3]. The effect of the downsampling is to artificially increase the separation between the sources. Here, we apply this technique in the case of DOA estimation by replacing temporal downsampling by spatial downsampling. More precisely, the proposed technique combines the effects of spatial downsampling and JES to further improve the resolution of MUSIC for closely spaced DOAs. Section II presents the data model and problem formulation. In section III, the TS-MUSIC algorithm is introduced and discussed. Section IV provides simulation results to assess the performances of TS-MUSIC comparatively to that of MUSIC algorithm. Concluding remarks are given in section V.

2. PROBLEM FORMULATION

In the following we consider a uniform linear array (ULA) of N antennas separated by half a wavelength[1]. T samples are collected on each antenna. We assume d plane waves sources impinging on the array from angles θ_1, ..., θ_d. The received signal is corrupted by additive white gaussian noise and is expressed as [4] :

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \quad t = 1....T$$

where $\mathbf{A} = [\mathbf{a}(\theta_1)....\mathbf{a}(\theta_d)]$ is the (N x d) steering matrix, $\mathbf{a}(\theta) = [1\, e^{j\pi \sin(\theta)} \ldots e^{j\pi \sin(\theta)(N-1)}]^T$ is the (N x 1) vector steering vector toward the direction θ, $\mathbf{s}(t)$ the (d x 1) vector of zero-mean random source waveforms, $\mathbf{n}(t)$ is the (N x 1) vector of white zero-mean sensor noise and $(.)^T$ denotes the transposition operator. The following assumptions on the model are considered to hold throughout this work:

- A_1: the number of sensors is at least L times (L>1 a positive integer) larger than the number of sources.
- A_2: the sources number is unknown (but small in comparison with N) and two or more sources are closely located.

Our objective is to improve the sources detection and resolution performances of MUSIC using both spatial downsampling and joint estimation strategy.

3. TS-MUSIC

The first step of our algorithm consists in the application of the standard MUSIC method : i.e. singular value decomposition (SVD) of the covariance matrix of received signal is processed, the noise subspace is built and the sources angles are estimated by minimizing the projection of the steering

[1]The results in this paper do not require a specific geometry of the array. We assumed ULA model just to simplify the notation.

0-7803-7011-2/01/$10.00 ©2001 IEEE

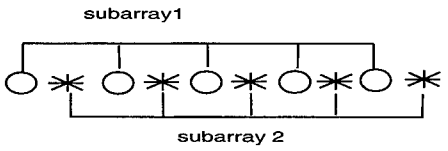


Figure 1: example of downsampling with L=2

vector $\mathbf{a}(\theta)$ on the noise subspace [5]. Now, in the situation where two or more sources are closely located, MUSIC fails in separating them and estimates less DOAs than effective sources. The following is proposed to improve the resolution.

3.1. Spatial Downsampling

To improve the detection performances, we propose to use a spatial downsampling and re-apply MUSIC in a second step to the downsampled data.

Let L be the downsampling factor (L>1) and M>d be the size of each subarray (see figure1 for an illustration of spatial downsampling with L=2).

The steering vector corresponding to the kth subarray is:

$$\mathbf{a}_k(\theta) = e^{j\pi\,(k-1)\sin(\theta)} \begin{bmatrix} 1 \\ e^{j\pi L \sin(\theta)} \\ \dots \\ e^{j\pi(M-1)L\sin(\theta)} \end{bmatrix}$$

It is clear that the downsampling artificially increases the separation between the sources by a factor L. This results to improved performances of the method.

Let illustrate our algorithm with the following example : consider the case where we have two sources located at $\theta_1 = \theta + \delta\theta_1$ and $\theta_2 = \theta + \delta\theta_2$ with $|\delta\theta_m| \ll 1 \;\; m = 1, 2$. We can then write

$$\mathbf{a}(\theta_m) = \begin{bmatrix} 1 \\ e^{j2\pi \sin(\theta+\delta\theta_m)} \\ \dots \\ e^{j(N-1)\pi \sin(\theta+\delta\theta_m)} \end{bmatrix} \quad m = 1, 2$$

Applying MUSIC in this context will result in one angle estimate $\hat{\theta}$ (i.e. MUSIC fails to distinguish between the two sources).

In a second step, we use a downsampling factor $L > 1$ and the previous estimate of θ (i.e. $\hat{\theta}$) to separate the sources and improve the estimation of their respective DOAs. More precisely, we re-apply MUSIC algorithm by restricting our search in the vicinity of $\hat{\theta}$ and using the following expression for the steering vector :

$$\mathbf{a}(\delta\theta_i) = \begin{bmatrix} 1 \\ e^{j\pi L \sin(\hat{\theta}+\delta\theta_i)} \\ \dots \\ e^{j(M-1)L\pi \sin(\hat{\theta}+\delta\theta_i)} \end{bmatrix}$$

This has the advantage of "virtually" increasing the angle difference between θ_1 and θ_2 by a factor of L which leads to a better resolution of the two angles.

3.2. Joint Estimation Strategy

One of the major issue of high resolution methods for DOAs estimation is the determination of the number of sources. Here, we propose to use a technique based on a joint estimation strategy following the same spirit as the one given by Gershman and Böhme in [1]. In their contribution resampling techniques are used to build "artificially" several trials of the observations. Then different methods are used to estimate DOAs from each trials. The major idea of their algorithm is that these methods show different local behavior[2]. The reason is that DOA's estimation methods are noise sensitive and thus their local behavior is different in each estimation trial. For example, considering two estimators with comparable performance, one can always find some trials where the first one resolves the sources while the second does not. Using this JES, the number of sources is determined as the maximum number (best case) obtained from the different estimation trials. We have adapted this technique in the following manner:

We first make a coarse estimation of sources number and positions using standard MUSIC applied to the global array outputs. If $\hat{q}$ peaks appear, this leads to the determination of a set of $\hat{q}$ angular intervals in which the real angles lie. For example, each interval can be defined by the -3dB points around those peaks. Now, to refine the search we place us in one of the angular sector determined before . As we work on subarrays in the second step of our algorithm, we can expect that each subset of antennas can lead to various estimation behavior because noise trials are different on each antenna. So we can perform a selection on the subarrays to get the best estimates.

In our case, if L is the interleaving factor value, we can obtain L different DOAs estimates. Then among these L sets of estimates, we can keep those that give the highest number of peaks in the angular sector being scanned.

The algorithm can be summarized as follows:
First step:

1. Coarse estimation of the number $\hat{q}$ of sources by a conventional method like beamforming or MUSIC.

2. Definition of the set of intervals in which the refined search will be performed:

$$\bigcup_{i=1}^{\hat{q}} [\theta^i_{left}, \theta^i_{right}]$$

[2]Local behavior refers to the instantaneous performance of any estimation algorithm achieved in a single trial without any statistical averaging.

where θ^i_{left} and θ^i_{right} are the left and right bounds of the angular sector corresponding to the ith DOA.

Second step:

3. Application of the interleaving method leading to L subarrays outputs (L trials).

4. On each trial apply MUSIC algorithm where the number of sources in each angular sector is selected as the number of peaks of MUSIC spectrum[3] observed in this sector. Each peak argument corresponds to an estimate of a source angle.

5. The number of sources in each angular sector is chosen as the maximum number of sources detected from the L subarrays.

6. In each angular sector, select the p subarrays (trials) that lead to the maximum number of sources in this interval and after sorting the angles, we compute the final angle estimates as their averaged values, i.e.

$$\hat{\theta}_j = \frac{1}{p}\sum_{l=1}^{p}\hat{\theta}_j^{(l)}$$

where $\hat{\theta}_j^{(l)}$, l=1...p represent the p estimates of $\hat{\theta}_j$ from p different subarrays.

3.3. Discussion

We discuss here the possible extensions of the algorithm and give some observations on our method:

- We have applied array downsampling plus JES to improve MUSIC resolution. Other standard localization techniques [4] can be used as well and improved in the same way. Also, it is always possible, as in [1, 2], to use different estimation methods on each subarray and combine the results according to the JES.

- The proposed method does not necessarily improve the estimation accuracy of the DOA but only their resolution when they are closely spaced. The reason is that in the second step of TS-MUSIC, subarrays (L time smaller than the global array) are used to estimate DOAs which may deteriorate the estimation accuracy (see figure 6 for illustration).

- The method can be further improved by applying both spatial downsampling and temporal resampling, e.g. bootstrap, on each subarray to increase the number of estimation trials.

[3]We limit our angle search in the second step of the algorithm to the angular sectors previously computed.

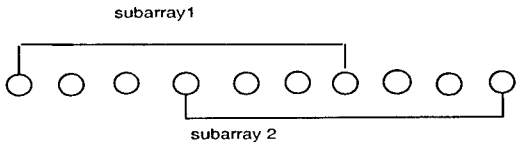


Figure 2: two subarrays spaced by L=3 sensors inter-spaces.

- A major limitation of TS-MUSIC is the restrictive condition $N > Ld$ with $L > 1$. An alternative solution would be to use in the second step of the method ESPRIT algorithm [4] with 2 subarrays spaced by L sensors inter-spaces (see figure 2). In this case, condition A_1 (§ 1) becomes $N > d + L$ which is much less restrictive than the previous one. Furthermore, a resampling technique can be used to apply the JES to further improve the resolution.

4. SIMULATION EXAMPLE

To illustrate the performance improvement achieved by our method, we consider a simple example of d=2 equipower sources impinging on the array from θ_1 =22° and $\theta_2 = \theta_1 + \delta\theta$, $\delta\theta$ being a small angle difference. The ULA is constituted of N=10 antennas. The sample size is set to T=100. The results that we obtained are based on 1000 independent Monte Carlo experiments. We have applied both MUSIC and TS-MUSIC for comparison. In figure 3, we plot the resolution probability (i.e. the percentage of successful detection of the exact sources number) versus the angle difference for a SNR of 5 dB and a downsampling factor L=3. We can see that for low SNRs, TS-MUSIC achieves a higher rate of successful source separation in comparison with MUSIC. For example, for an angle difference of 3° the resolution probability of MUSIC was less than 10% while it is of 50% for TS-MUSIC. In figure 4, we plot the resolution probability versus the SNR for an angle difference $\delta\theta = 2^o$ and a downsampling factor L=3. A significant improvement is obtained in terms of resolution probability for low and moderate SNRs. In figure 5, we represent the histograms of the number of sources detected by MUSIC and TS-MUSIC for a SNR of 0 dB and different angular distances. In figure 6, we plot the DOAs MSE (mean square error) against SNR for L=2 and an angle difference of $\delta\theta = 8^o$ (we chose here a situation where both MUSIC and TS-MUSIC achieve a correct source separation). We note that accuracies of the estimation with MUSIC and TS-MUSIC are very close. This illustrates the fact, in such context, that the proposed method only improves the resolution but not necessarily the estimation accuracy.

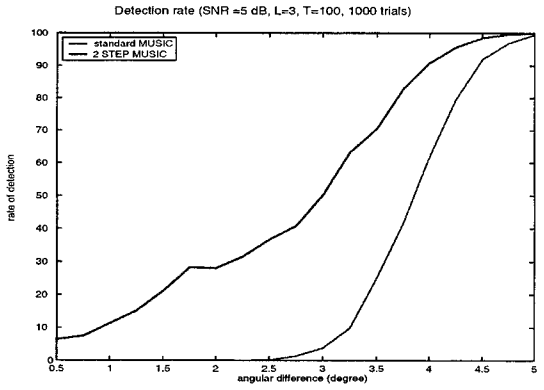


Figure 3: Detection rate (in %) vs. $\delta\theta$

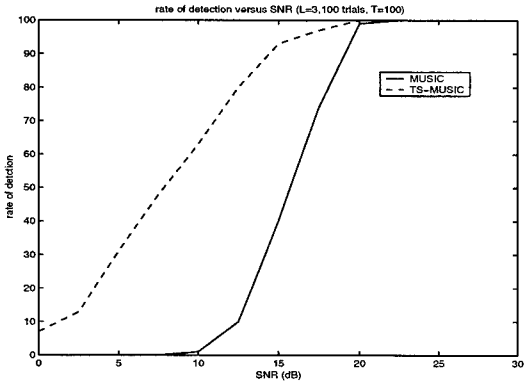


Figure 4: Detection rate (in %) vs. SNR

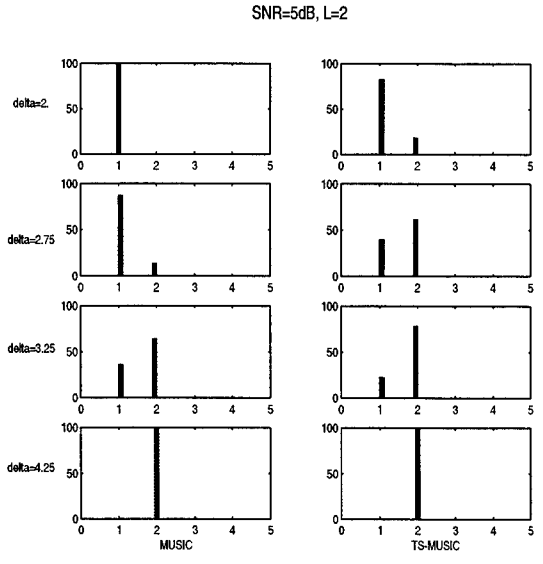


Figure 5: Estimated sources number histograms

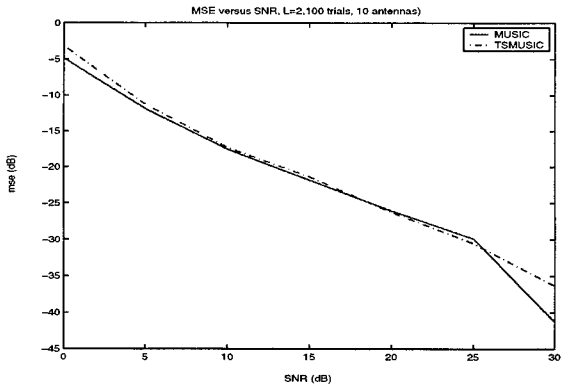


Figure 6: MSE vs. SNR

5. CONCLUSION

We have introduced a new method TS-MUSIC based on the combination of spatial downsampling and a joint estimation strategy to improve array resolution. Spatial downsampling artificially increases sources separation and provides several subarrays (i.e. trials) whereas JES enables a selection of the best estimates that are obtained with the different subarrays.

6. REFERENCES

[1] Alex B. Gershman and Johann F. Böhme. Joint estimation strategy with application to eigenstructure methods. *8th IEEE Signal Processing Workshop on Stastical Signal and Array Processing*, June 24-26 1996.

[2] Alex B. Gershman and Johann F. Böhme. Improved doa estimation via pseudo-random resampling of spatial spectrum. *IEEE Signal Processing Letters*, 4(2):54–57, february 1997.

[3] Bijit Halder and Thomas Kailath. Efficient estimation of closely spaced sinusoidal frequencies using subspace-based methods. *IEEE Signal Processing Letters*, Vol. 4(num. 2):p. 49–51, 1997.

[4] Don H. Johnson and Dan E. Dudgeon. *Array Signal Processing : Concepts and Techniques*. Prentice Hall, 1993.

[5] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *Proceedings of the RADC Spectral Estimation Workshop*, Rome(N-Y):243–258, 1979.

OPTIMIZATION OF ELEMENT POSITIONS FOR DIRECTION FINDING WITH SPARSE ARRAYS*

Fredrik Athley

Department of Signals and Systems
Chalmers University of Technology
SE-412 96 Göteborg, Sweden
athley@s2.chalmers.se

ABSTRACT

Sparse arrays are attractive for Direction-Of-Arrival (DOA) estimation since they can provide accurate estimates at a low cost. A problem of great interest in this matter is to determine the element positions that yield the best DOA estimation performance. A major difficulty with this problem is to define a suitable performance measure to optimize. In this paper, a novel criterion is proposed for optimizing element positions. The ambiguity threshold of the Weiss-Weinstein Bound (WWB) is used to optimize the element positions of a sparse linear array. The array obtained from the optimization is compared with some other sparse array structures that have been proposed in the literature.

1. INTRODUCTION

Direction-Of-Arrival (DOA) estimation using an array of sensors finds application in many fields, such as radar, sonar, communications etc. Over the past decades, there has been intense research in this area, see e.g. [1] and the references therein. The DOA estimation accuracy is critically dependent on the array size. Large arrays can thus provide very accurate estimates. DOA estimation with arrays with many elements are, however, expensive to implement, both in terms of receiver hardware and computational complexity.

For non-ambiguous DOA estimation with Uniform Linear Arrays (ULAs), the inter-element spacings should not exceed half a wavelength of the impinging wavefronts. In sparse arrays, elements are spaced further apart in order to obtain a large aperture with few elements. Sparse arrays thus have the potential of very accurate DOA estimation at a low cost. The price paid is the risk of obtaining ambiguous estimates, caused by grating lobes in the array beam pattern. To reduce such grating lobes, non-uniform element spacing is employed. An important problem is then to determine which element positions yield the most accurate DOA estimates.

Different approaches in optimizing the element positions with respect to DOA estimation accuracy have been taken in the literature. In [2, 3], the element positions of Non-Uniform Linear Arrays (NULAs) were optimized by minimization of the Cramér-Rao Bound (CRB). A problem with this approach is that the CRB is a local bound that does not take into account large estimation errors caused by near ambiguities. For the single signal problem this means that only the curvature of the mainlobe is considered; high sidelobes have no effect. At low Signal-to-Noise Ratios (SNRs) these sidelobes may cause large estimation errors, rendering the CRB a far too optimistic bound in this case.

Various approaches have been proposed to account for near ambiguities. In [4] the mainlobe area was minimized subject to a peak sidelobe constraint and in [5] competitive criteria involving maximum aperture and identifiability were considered. Although these approaches are intuitively appealing, there is no explicit connection between these ambiguity/aperture trade-offs and the resulting mean square estimation error.

In this paper, another approach is taken. A lower bound on the mean square estimation error that takes ambiguity errors into account is used to optimize the element positions of a NULA with fixed aperture. The bound used is the Weiss-Weinstein Bound (WWB), which was first presented in [6] and subsequently applied to DOA estimation in e.g. [7, 8, 9, 10].

2. PROBLEM FORMULATION

Consider a linear array of K sensors receiving a single planar wavefront from the DOA θ measured relative to the array boresight. For mathematical convenience, the estimation of $u \triangleq \sin\theta$ is considered. The element positions, denoted by $d_k, k = 1, \ldots, K$ are normalized by the standard spacing $\lambda/2$ where λ is the wavelength, i.e. $d_k = 2\tilde{d}_k/\lambda$ where $\tilde{d}_k$ is the physical distance. In the sequel, different linear array geometries, keeping the array length D (normalized by $\lambda/2$) fixed, will be studied. Without loss of generality, the end elements d_1 and d_K are fixed at 0 and D respectively. Assuming an ideal array with omnidirectional elements, the array output at time t can be modeled by the $K \times 1$ complex vector

$$\mathbf{x}(t) = \mathbf{a}(u)s(t) + \mathbf{n}(t), \qquad t = 1, \ldots, N \tag{1}$$

where

$$\mathbf{a}(u) = \begin{bmatrix} 1 & e^{-j\pi d_2 u} & \ldots & e^{-j\pi d_{K-1} u} & e^{-j\pi D u} \end{bmatrix}^T$$

is the $K \times 1$ array steering vector. Furthermore, $s(t)$ denotes the impinging signal at baseband, $\mathbf{n}(t)$ is an additive noise term and N denotes the number of temporal snapshots. The signal $s(t)$ and noises $\mathbf{n}(t)$ are assumed independent and are modeled as white (spatially and temporally), zero mean, circular complex Gaussian random variables with second order moments

$$\mathrm{E}\left[|s(t)|^2\right] = \mathrm{SNR} \quad \text{and} \quad \mathrm{E}\left[\mathbf{n}(t)\mathbf{n}^H(t)\right] = \mathbf{I}, \tag{2}$$

The signal variance is thus equal to the Signal-to-Noise-Ratio per space-time sample since the noise variance has been normalized to unity.

*This work was supported in part by *Ericsson Microwave Systems AB*

The problem considered in this paper is, given the noisy observations $\mathbf{x}(t)$, $t = 1, \ldots, N$, to determine the element positions $d_2, \ldots, d_{K-1}$ that maximize the DOA estimation performance. A crucial issue that is addressed herein is how to define DOA estimation performance when the estimation is prone to ambiguities.

3. ESTIMATION PERFORMANCE MEASURES

Often, DOA estimation performance is evaluated by means of the CRB. This bound is relatively easy to compute but is a local bound that does not take into account large errors that may be caused by near ambiguities. Various global bounds have been proposed in the literature. These are more tedious to compute but, on the other hand, they provide insights into how ambiguity errors affect the overall estimation error. One such bound is the Weiss-Weinstein Bound (WWB) [6]. This is a lower bound on the Mean Square Error (MSE) that rests on the Bayesian framework of estimation. This means that the parameter of interest is considered to be a random variable with known prior distribution. Throughout the paper, a uniform distribution on $[-1, 1]$ is assigned to u. For details concerning the computation of the WWB for DOA estimation, see [7].

To illustrate the difference between the WWB and the CRB, Figure 1 shows the CRB and WWB as a function of SNR for a particular NULA with 8 elements[1]. At high SNR, the WWB and CRB coincide since, in this region, ambiguous estimates do not occur. Below a certain SNR threshold the WWB increases rapidly. At this threshold, ambiguous estimates from grating lobes begin to yield contribution to the total MSE which is comparable to that of the mainlobe. This threshold effect is not captured by the CRB. The performance measure that is used in this paper for finding optimal element positions is this SNR threshold. An array with a low threshold is likely to provide very accurate estimates and still be robust to ambiguity errors. The SNR threshold can be defined in different ways, e.g. where the WWB exceeds the CRB by a certain amount or the maximum of the second derivative of the WWB curve. Since no analytical expressions for the SNR threshold are derived in this paper, the liberty is taken to identify the SNR threshold simply by ocular inspection of the produced graphs. It is likely that the conclusions of the present paper is independent of the precise definition of SNR threshold.

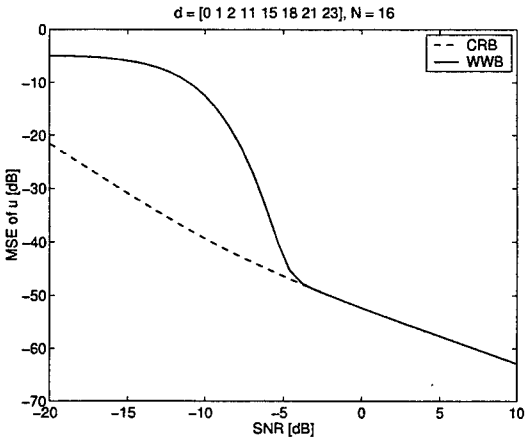


Fig. 1. CRB and WWB as a function of SNR for a NULA.

[1]Since the WWB is a Bayesian bound, a direct comparison with the CRB is not meaningful in general. However, arguments similar to those in [10] can be used to justify such a comparison.

4. OPTIMIZATION OF ELEMENT POSITIONS

The basic ideas behind the optimization procedure are as follows:

1. Generate a large number of different arrays with random element positions.
2. Compute the WWB as a function of SNR for each array.
3. Identify a reduced set of arrays with the lowest SNR thresholds.
4. The element positions of these arrays are used as starting points in a numerical optimization routine to improve the best arrays from the previous step.
5. The optimal element positions are then taken from the best array after the numerical optimization.

Importance was attached to analyzing as many random arrays as possible within a limited computing time. Therefore, a somewhat simplified procedure was implemented:

- The WWB as a function of SNR was computed for 10^3 different arrays. The array with the lowest SNR threshold of these arrays was identified. This array had a threshold at about SNR = -5 dB. Then, the WWB at SNR = - 5 dB was computed for 10^6 random arrays.
- The 10 arrays with the lowest WWB at SNR = - 5 dB were selected for numerical optimization.
- The element positions of these arrays were used as starting points when minimizing the WWB with respect to element positions at SNR = -5 dB. The "fminsearch" routine in Matlab's Optimization Toolbox was used for this purpose.
- The array with the lowest WWB after the numerical optimization was then considered to be the optimal array.

There is no guarantee that the global optimum is found with this procedure. If a very large number of random arrays are generated, however, it is likely that the obtained solution is "sufficiently optimal" in any practical application.

The optimization procedure was evaluated by generating 10^6 eight-element linear arrays with random element positions and $D = 23, N = 16$. The element positions were generated according to a uniform distribution on $[0, D]$. Figure 2 shows the WWB as a function of SNR for the arrays which had the lowest and highest WWB at SNR = -5 dB. In order to show the statistical nature of the WWB of the randomly generated arrays, there is also a histogram of the WWB at SNR = -5 dB for all the arrays in the plot. The histogram has been rotated 90° compared to the standard orientation of a histogram. It can be seen from the figure that the difference between the WWB for the best and the worst array is quite large. The positions of the array elements thus have a great influence on the attainable estimation performance. The element positions of the 10 best arrays were then used as starting points in a numerical optimization routine to minimize the MSE at SNR = -5 dB. Finally, the optimal element positions are taken from the array with the lowest MSE at this SNR. The numerical optimization reduced the minimum MSE from -44.8 dB to -45.2 dB.

Hitherto, the element positions were considered as continuous variables. This implies an infinite number of possible arrays with different element positions. On the other hand, constraining the element positions to a discrete grid leads to a finite number of

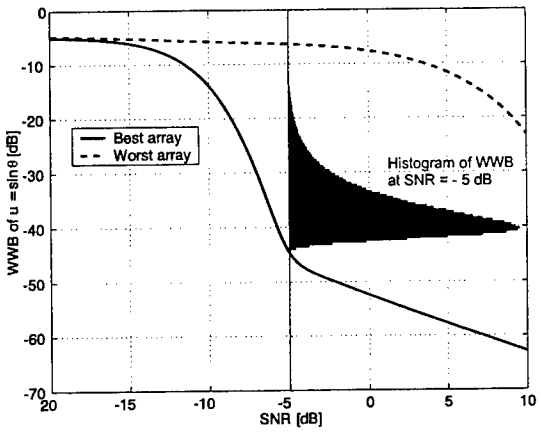


Fig. 2. WWB for the best and worst array and a histogram of WWB at SNR = -5 dB.

possible different arrays. Therefore, it should be possible to compute the WWB for all these possible arrays if the number of grid points and array elements are not too large. A common approach is to start with a ULA with $\lambda/2$ element spacing and the required length. Then, a given number of elements are removed from the full array in order to produce the sparse array. These arrays are often called *thinned arrays*. In the present example, the two end elements are fixed. Thus, there are 22 element positions to choose 6 positions from. The number of different ways to pick 6 elements out of 22 is equal to $\binom{22}{6} = 74613$. This is a reasonable number of arrays for being able to compute the WWB for all of these arrays on a standard PC. The WWB at SNR = -5 dB was computed for all these 74613 arrays and the result is illustrated in Figure 3. The

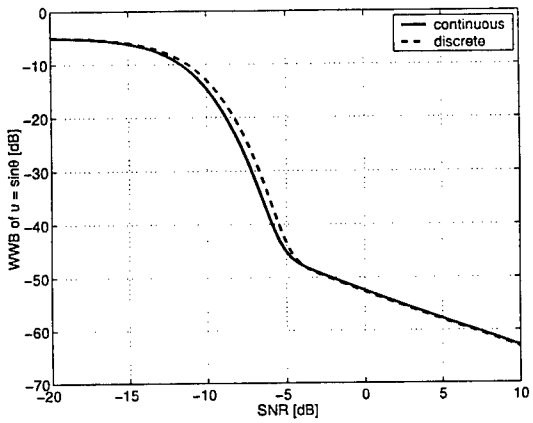


Fig. 3. WWB for the best array using continuous and discrete element positions respectively.

WWB vs SNR for the arrays with the lowest WWB at SNR = -5 dB using continuous and discrete positions respectively is shown. Clearly, the difference between the two is negligible.

5. COMPARISON WITH OTHER ARRAYS

The arrays obtained from the optimization procedure described in the previous section were compared with a few other array configurations that have been studied in the literature. A type of thinned array that has been widely studied is the so called minimum-redundancy array [11]. Another array configuration that also has been studied is two separated subarrays where each subarray is a ULA with $\lambda/2$ inter-element spacing, see e.g. [9, 10]. The array geometry that minimizes the CRB for NULAs with fixed length is given by two point clusters at the array end points [3]. Due to mutual coupling effects and mechanical considerations, the element spacing cannot be too small. The separated subarrays configuration can thus be viewed as a realizable approximation of the CRB-optimal geometry. Figure 4 shows the element positions for the arrays under consideration. In Figure 5, the WWB of the best thinned ar-

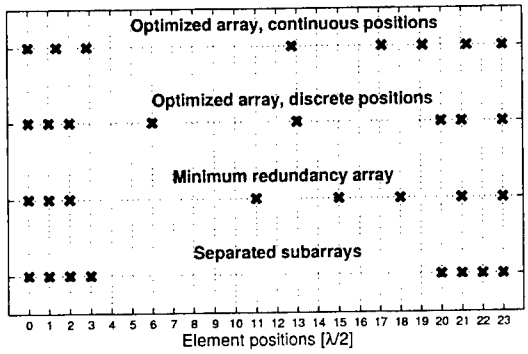


Fig. 4. Element positions

ray obtained from the numerical optimization is compared with the WWB of a minimum redundancy array. There is practically no

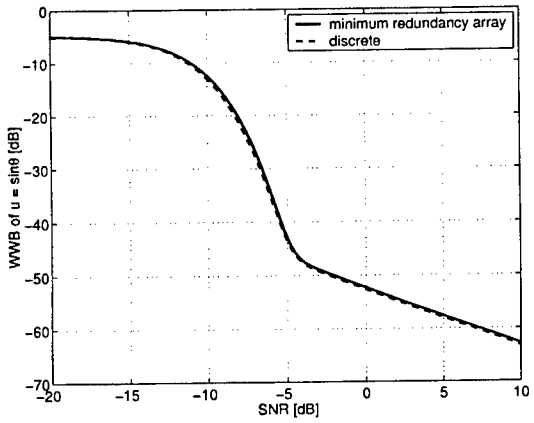


Fig. 5. WWB for minimum redundancy array and the best thinned array.

difference between the WWB for the two arrays. Therefore, in this example, the minimum redundancy array can be considered to be the optimal thinned array with respect to robustness to ambiguity errors. Recall that the difference between the WWB for the arrays obtained from minimization over continuous and discrete element positions respectively was very small. Therefore, it is concluded that the minimum redundancy array is near optimal in the present example.

Figure 6 shows WWB vs SNR for the array obtained from optimization over continuous element positions and the separated subarray structure. The separated subarray structure has a consid-

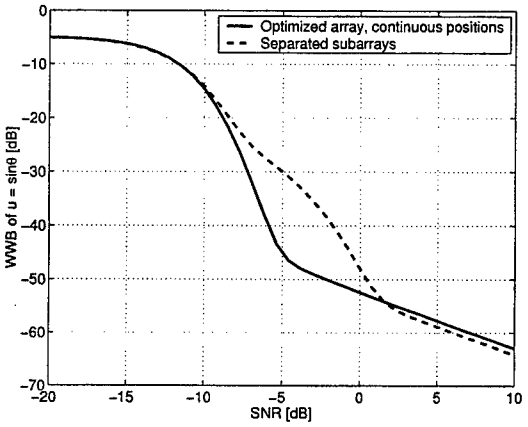


Fig. 6. WWB for optimized array, continuous positions and separated subarrays respectively.

erably higher SNR threshold and somewhat lower WWB at high SNR as compared to the optimal array. This is expected, since the separated subarray structure has a narrower mainlobe but higher sidelobes, due to concentration of the elements near the array endpoints.

Common engineering practice suggests that low sidelobes are important for ambiguity-free DOA estimation. In order to investigate the adequacy of this, Figure 7 displays a 2-D histogram of the WWB at SNR = -5 dB and the peak sidelobe in the beam pattern for each of the 10^6 arrays as a contour plot. Some interesting con-

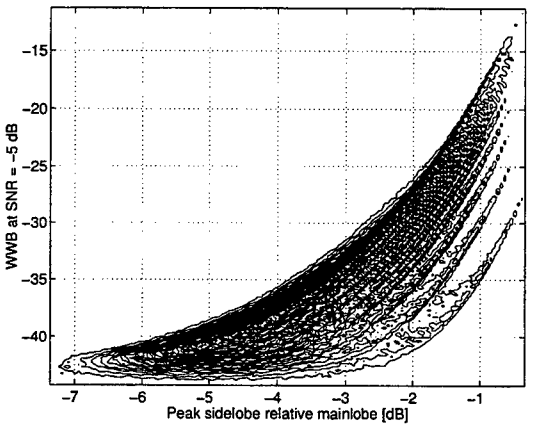


Fig. 7. 2-D histogram of WWB at SNR = -5 dB and peak sidelobe, displayed as a contour plot.

clusions can be drawn from Figure 7. For most arrays, high peak sidelobe means high WWB. A few ridges can be discerned in the contour plot. These are probably due to the peak sidelobe being at different distances from the mainlobe. An ambiguous estimate from a sidelobe far from the mainlobe gives a greater contribution to the MSE than from a sidelobe close to the mainlobe. A remarkable property appears, however, if the lowest contour of the histogram is scrutinized. Apparently, if only the best arrays are considered, there seems to be no relationship between the peak sidelobe and the WWB, at least as long as the peak sidelobe does not exceed -3 dB (relative to the mainlobe). It is concluded that high peak sidelobe does not necessarily give high mean square estimation error, if the element positions are determined judiciously. Inspection of the corresponding beam patterns revealed that for the arrays with low WWB and high peak sidelobe, the peak sidelobe is relatively close to the mainlobe.

6. CONCLUSIONS

A novel criterion for optimizing the element positions of sparse linear arrays has been presented. The criterion used was the ambiguity threshold of the Weiss-Weinstein Bound (WWB). This is a lower bound on the mean square DOA estimation error that takes into account large errors caused by ambiguities. An optimization procedure was implemented in order to find the array with lowest ambiguity threshold. The WWB for this array was compared with a minimum-redundancy array and a separated subarrays structure. It was found that the optimal array and the minimum-redundancy array had similar performance. Furthermore, it was found that low peak sidelobe is not necessary for obtaining lowest possible WWB.

7. REFERENCES

[1] H. Krim and M. Viberg, "Two Decades of Array Signal Processing Research: The Parametric Approach", *IEEE Signal Processing Magazine*, **13**(4):67–94, July 1996.

[2] Y. I. Abramovich, D. A. Gray, A. Y. Gorokhov, and N. K. Spencer, "Comparison of DOA Estimation Performance for Various Types of Sparse Antenna Array Geometries", In *Proceedings of EUSIPCO-96*, 1996.

[3] A. B. Gershman and J. F. Böhme, "A Note on Most Favorable Array Geometries for DOA Estimation and Array Interpolation", *IEEE Signal Processing Letters*, **4**(8):232–5, August 1997.

[4] M. Viberg and C. Engdahl, "Element Position Considerations for Robust Direction Finding Using Sparse Arrays", In *Proc.* 33rd *Asilomar Conf. on Signals, Systems, and Computers*, 1999.

[5] Y. I. Abramovich and N. K. Spencer, "Design of Nonuniform Linear Antenna Array Geometry and Signal Processing Algorithm for DOA Estimation of Gaussian Sources", *Digital Signal Processing*, **10**(4):340–54, October 2000.

[6] A. J. Weiss and E. Weinstein, "A Lower Bound on the Mean-Square Error in Random Parameter Estimation", *IEEE Trans. on Information Theory*, **31**(5):680–682, Sep. 1985.

[7] D. F. DeLong, "Use of the Weiss-Weinstein Bound to Compare the Direction-Finding Performance of Sparse Arrays", Technical Report TR-982, MIT Lincoln Lab., Aug. 1993.

[8] H. Nguyen and H. L. Van Trees, "Comparison of Performance Bounds for DOA Estimation", In *Seventh SP Workshop on Statistical Signal & Array Processing*, 1994.

[9] M. Zatman and S. T. Smith, "Resolution and Ambiguity Bounds for Interferometric-Like Systems", In *Proc.* 32nd *Asilomar Conf. on Signals, Systems, and Computers*, 1998.

[10] F. Athley and C. Engdahl, "Direction-of-Arrival Estimation Using Separated Subarrays", In *Proc.* 34th *Asilomar Conf. on Signals, Systems, and Computers*, 2000.

[11] A. T. Moffet, "Minimum-Redundancy Linear Arrays", *IEEE Trans. on Antennas and Propagation*, **AP-16**(2):172–175, March 1968.

HIGH RESOLUTION DF WITH A SINGLE CHANNEL RECEIVER

Chong Meng Samson See *

Abstract— **This paper presents a method for high resolution multiple source direction finding with an antenna array. Unlike previously developed algorithms, the proposed approach can achieve high resolution direction finding with only ONE receiver, thereby, offering significant hardware savings. The proposed approach requires the signal received by the antenna array to be pre-processed by a beamformer network where each of the beamformer output ports are sequentially sampled by a RF switch. As the power of each beamformer output port is a function of the array covariance matrix, we derive a Kronecker form that leads to an unique least squares estimates of the array covariance matrix using the power measured from all the beamformer output ports. With the array covariance matrix estimated, conventional high resolution DF algorithms can be applied to determine the direction of arrival estimation of the multiple sources impinging the antenna array.**

I. INTRODUCTION

High resolution direction finding algorithms enable the antenna array system to achieve accurate direction of arrival estimation in the presence of co-channel and multipaths. Most high resolution direction finding (DF) algorithms, such as MUSIC, ESPRIT, WSF etc, require the number of receivers to the number of antennas to be matched. However, these antenna array processing system can be costly to implement in applications that require to achieve wide instantaneous frequency coverage and where weight, size and volume are subjected to tight constraints. In [1], a method for high resolution direction finding is proposed which allows the number of antenna to be larger than the number of receivers. While the number of receivers needed is significantly reduced (minimum of two), the proposed method requires multi-dimensional search algorithms to determine the DOA of multiple sources which generally are computationally demanding and do not guarantee global convergence. In [2], the application of computationally efficient MUSIC and Capon's beamformer is made possible by estimating the DOA from a restricted number of antenna outputs (sub-array). Later in [3], a similar approach was proposed that combined the cost-function of each sub-array incoherently. Apart from poorer estimation performance, the number of signal sources that can be detected and resolved by these approaches are limited by the number of elements in the sub-array. Recently, an approach based on reconstructing of the array covariance from the sub-array data was proposed in [4]. The significance of this approach is that it allows the direct application of the MUSIC estimator.

Direction of arrival estimation processing architectures using only one receiver channel to sample the antenna array element, such as [5], have been proposed. With the sampling and antenna switching synchronized, the basic tenet of these approaches is to sequentially sample the antenna elements as fast as possible such that the effect of sequential sampling of the antenna can be approximated by phase shifts. As a result,

*C.M.S. See is with DSO National Laboratories, 20 Science Park Drive, Singapore 118230. Tel: 065-8712423. Fax: 065-8724366. Email:schongme@dso.org.sg

0-7803-7011-2/01/$10.00 ©2001 IEEE

the received signal vector, as in the case of a full channel antenna array system, can be approximated and, therefore, allowing the application of computationally efficient algorithms like MUSIC estimator. However, such approach requires the antennas to be sampled at very high rates and will increase proportionally with receiver bandwidth and number of antennas. It was recommended in [5] that the sampling rate should be in excess of 1GHz. Unfortunately, the need for very high speed RF switches and ADC as well as the ability to handle the large amount of data (due to high sampling rate) will increase the cost and complexity of the DF system, hence, diminishing the gain due to the reduced number of receivers.

In this paper, we present a method for high resolution DF with an antenna array. Unlike previously the approach proposed in [5], it can achieve high resolution DF of multiple signal sources with only one receiver without the need for high speed ADC and RF switches. As shown in Figure 1, in our proposed approach, the signals received by a N element antenna array is pre-processed by an analog beamformer network with M output ports. The M channel signals are sequentially sampled by a M to 1 RF switch. The single channel output from the RF switch is down-converted by a single channel receiver and sampled at Nyquist rate by an ADC for digital signal processing.

II. Proposed Approach

The received signal power associated with each beamformer output, $z_i(t)$, is estimated from the sampled data and is given by

$$z_i(t) = \boldsymbol{\alpha}_i^H \mathbf{r}(t)\,\mathbf{r}(t)^H \boldsymbol{\alpha}_i + n_i(t)$$

where $\boldsymbol{\alpha}_i$ is the vector of beamformer weights and $n_i(t)$ is the receiver noise. Without any loss of generality, we assume that the dominant noise comes from the receiver. The signal received by the antenna array, $\mathbf{r}(t)$, is given by

$$\mathbf{r}(t) = \mathbf{A}(\boldsymbol{\Theta})\,\mathbf{s}(t)$$

where $\mathbf{A}(\boldsymbol{\Theta}) = \left[\begin{array}{ccc} \mathbf{a}(\theta_1) & \cdots & \mathbf{a}(\theta_d) \end{array}\right]$ with $\mathbf{a}(\theta_i)$ being the steering vector associated with angle of arrival θ_i and $\mathbf{s}(t)$ is the source waveform. The measure power of the beamformer outputs is function of the array covariance matrix and the beamformer weight vector:

$$z_i = \mathrm{E}\langle z_i(t)\rangle = \boldsymbol{\alpha}_i^H \mathbf{R}(\boldsymbol{\Theta})\,\boldsymbol{\alpha}_i$$

where

$$\mathbf{R}(\boldsymbol{\Theta}) = \mathbf{A}(\boldsymbol{\Theta})\,\mathbf{P}\mathbf{A}(\boldsymbol{\Theta})^H \; and \mathbf{P} = E\left\langle \mathbf{s}(t)\,\mathbf{s}(t)^H \right\rangle.$$

By writing $\mathbf{z} = \left[\begin{array}{ccc} z_1 & \cdots & z_M \end{array}\right]^T$, we have

$$\mathbf{z} = \mathrm{diag}\left(\boldsymbol{\Gamma}^H \mathbf{R}(\boldsymbol{\Theta})\,\boldsymbol{\Gamma}\right) \qquad (1)$$

where $\boldsymbol{\Gamma} = \left[\begin{array}{ccc} \boldsymbol{\alpha}_1 & \cdots & \boldsymbol{\alpha}_M \end{array}\right]$ and the operator $\mathrm{diag}(\mathbf{C})$ returns a vector from diagonal terms extracted from $\mathbf{C}$. Given (1), we relate the power of all the beamformer outputs to the array covariance by

$$\begin{aligned} \mathbf{z} &= \mathbf{Q}\mathrm{vec}\left(\boldsymbol{\Gamma}^H \mathbf{R}(\boldsymbol{\Theta})\,\boldsymbol{\Gamma}\right) \\ &= \mathbf{Q}\left(\boldsymbol{\Gamma}^T \otimes \boldsymbol{\Gamma}^H\right)\mathbf{G}\boldsymbol{\rho} \\ &= \left(\tilde{\boldsymbol{\Omega}}_{\mathrm{Re}} + j\tilde{\boldsymbol{\Omega}}_{\mathrm{Im}}\right)\boldsymbol{\rho} \end{aligned}$$

where $\mathbf{Q}$ is a selection matrix such that $\mathrm{diag}(\mathbf{C}) = \mathbf{Q}\mathrm{vec}(\mathbf{C})$, $\boldsymbol{\rho}$ is a real valued parameter vector and $\mathbf{G}$ is another selection matrix (noting that $\mathbf{R}(\boldsymbol{\Theta})$ is a Hermitian matrix) such that $\mathbf{G}\boldsymbol{\rho} = \mathrm{vec}(\mathbf{R}(\boldsymbol{\Theta}))$. The operator $\otimes$ denotes Kronecker product and,

$$\tilde{\boldsymbol{\Omega}}_{\mathrm{Re}} = \mathrm{Re}\left(\mathbf{Q}\left(\boldsymbol{\Gamma}^T \otimes \boldsymbol{\Gamma}^H\right)\mathbf{G}\right)$$

and

$$\tilde{\boldsymbol{\Omega}}_{\mathrm{Im}} = \mathrm{Im}\left(\mathbf{Q}\left(\boldsymbol{\Gamma}^T \otimes \boldsymbol{\Gamma}^H\right)\mathbf{G}\right).$$

With sufficiently large number of beamformer outputs $M > N$, the least squares estimates of the array covariance matrix, $\mathbf{R}(\boldsymbol{\Theta})$ or $\boldsymbol{\rho}$, can be obtained by evaluating

$$\hat{\boldsymbol{\rho}} = \left(\boldsymbol{\Omega}^H\boldsymbol{\Omega}\right)^{-1}\boldsymbol{\Omega}^H \left[\begin{array}{c} \mathrm{Re}(\mathbf{z}) \\ \mathrm{Im}(\mathbf{z}) \end{array}\right]$$

where $\boldsymbol{\Omega} = \left[\begin{array}{cc} \tilde{\boldsymbol{\Omega}}_{\mathrm{Re}}^T & \tilde{\boldsymbol{\Omega}}_{\mathrm{Im}}^T \end{array}\right]^T$. Once the array covariance matrix is reconstructed from $\hat{\boldsymbol{\rho}}$, high resolution DF algorithms, such as MUSIC, can be used to estimate the DOA of the signal sources. It is important to point out that the signal power of each beamformer port can be estimated in frequency or in time domain. When wideband receiver is used to achieve wide instantaneous spectrum coverage, the signal power at the frequency of interest can be estimated in the frequency domain using Fast Fourier Transform.

A. A Numerical Example

In this example, we consider a 4 element antenna array uniformly spaced circular array with radius of 0.8λ, where λ is the wavelength of the signals impinging the antenna array. The analogy beamformer network has 16 output ports and the weights of each beamformer port are randomly generated. Figure 2 depicts a MUSIC spectrum computed using the array covariance matrix estimated by the proposed method. Two uncorrelated signal impinge the antenna array with SNR of 5dB from 70 and 180 as indicated by the red dotted lines. The number of snapshots used to estimate the power at each beamformer output z_i is 1000. As shown in Figure 2, the proposed method is able to resolve and estimate the source location accurately.

III. Concluding Remarks

Apart from achieving significant hardware savings by enabling high resolution DF with only ONE receiver, the proposed method also offers the following advantages:

1. As it derives the array covariance matrix from the beamformer output power, the sampling of the beamformer outputs can be done at very low rate. Hence, only low speed RF switches will be needed here.
2. It does not require the received signals to be oversampled and the ADC sampling rate is only low bounded by Nyquist rate.
3. When only ONE receiver is used, the DF processor based on the proposed method does not require receiver calibration. This will further simplify and reduce the cost of hardware.

As seen from these advantages, the proposed approach can offer low cost and parsimonious DF architecture for high resolution direction of arrival estimation.

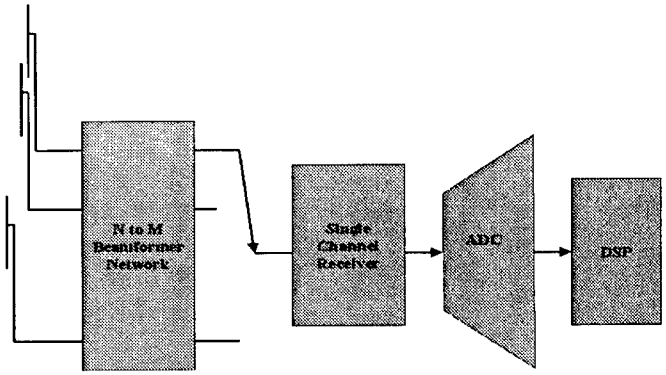


Figure 1 – Proposed Architecture.

References

[1] J. SHEINVALD and M. WAX. "Detection and Localisation of Multiple Signals Using Subarrays Data". S. HAYKIN. Advances in Spectrum Analysis and Array Processing, vol. 3, pp. 324-351: Prentice Hall, 1995.

[2] K.M. Buckley, X.L. Xu, "Recent Advances in High Resolution Spatial Spectrum Estimation", *Proc. of EUSIPCO-90*, Barcelona, Spain, Sep. 1990, pp. 17-25.

[3] J. G. Worms, "RF Direction Finding with a Reduced Number of Receivers by Sequential Sampling", *IEEE Conference on Phased Array Systems and Technology*, May 2000.

[4] Fishler, E.; Messer, H., "Multiple source direction finding with an array of M sensors using two receivers ", *Statistical Signal and Array Processing. 2000. Proceedings of the Tenth IEEE Workshop on , 2000*, pp. 86 -89.

[5] US Patent 5,497,161, "Angle of Arrival Solution using a single receiver".

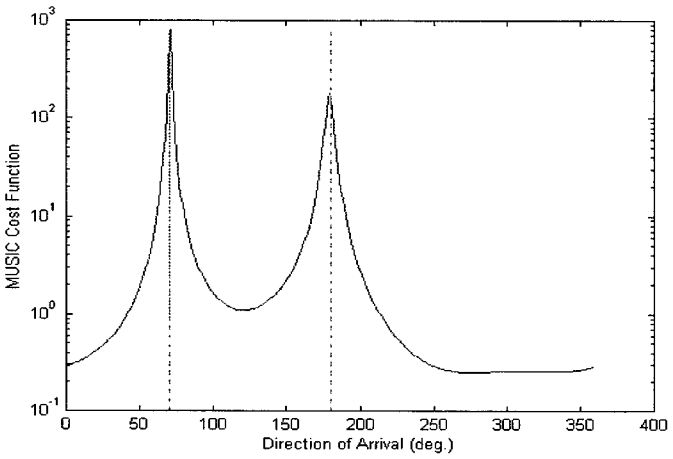


Figure 2 – MUSIC Spectrum. - - - : True DOA. $M = 20$, $N = 4$, Number of snapshots : 1000. Array Geometry: Uniform circular array, radius = 0.8 wavelength.

DIRECTIONS-OF-ARRIVAL ESTIMATION OF CYCLOSTATIONARY SIGNALS IN MULTIPATH PROPAGATION ENVIRONMENT

Jingmin Xin [†] and *Akira Sano* [§]

[†] YRP Mobile Telecoms. Key Tech. Res. Lab., 3-4 Hikari-no-oka, Yokosuka 239-0847, Japan
[§] Department of System Design Engineering, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan

ABSTRACT

By exploiting the spatial and temporal properties of most communication signals, we propose a forward-backward linear prediction (FBLP) approach for estimating the directions-of-arrival (DOA) of coherent signals impinging on a uniform linear array. In the proposed method, the evaluation of the cyclic array covariance matrix is avoided and the difficulty of choosing the optimal lag parameter is alleviated. As a result, the proposed method has two advantages: the computational load is relatively reduced and the robustness of estimation is significantly improved. It is shown through numerical examples that this approach is superior in resolving closely spaced coherent signals with small length of array data and at relatively low signal-to-noise ratio (SNR).

1. INTRODUCTION

For estimating the directions-of-arrival (DOA) of multiple narrow-band signals from the noisy array data, maximum likelihood (ML) methods and subspace-based methods are well known. Subspace-based methods such as MUSIC [1] and MODE (in a uniform linear array (ULA)) [2] are more computationally efficient than the ML methods [3], but all of them except MODE are unsuitable for coherent signals. To tackle the problem of coherent signals, several modifications to the subspace-based methods have been proposed. Among them, spatial smoothing (SS) [4] is a popular preprocessing scheme. However, in array processing of wireless communication systems, there are some practical situations where the overall number of incident signals is greater than the number of sensors even though the number of desired signals is smaller, and multipath propagation due to various reflections is often encountered. Furthermore, the number of snapshots is usually limited. In these scenarios, the performance of most subspace-based methods and their variants will degrade. Moreover, the subspace-based methods basically rely on the spatial information contained in the received data, whereas the temporal properties of the desired incident signals are ignored.

Most communication signals exhibit cyclostationarity for a given cycle frequency because of the underlying periodicity arising from carrier frequencies or baud rates [7]. Many direction estimation methods exploiting this inherently temporal property have been developed recently (e.g. [8]), in which the stationary noise and the interfering signals that do not share a cycle frequency common to the desired signals are suppressed. For estimating the directions of coherent cyclostationary signals, a cyclic ML method [9] and an SS-based cyclic MUSIC method [10] were proposed. However, the former is computationally expensive because it involves a multidimensional optimization, while the latter is still not computationally efficient enough since the cyclic correlation matrices of subarrays must be evaluated.

In this paper, by utilizing the spatial and temporal properties of the incoming signals impinging on a ULA, we investigate an efficient method for estimating the directions of narrow-band cyclostationary signals in a multipath propagation environment. In the proposed cyclic method, the forward-backward linear prediction (FBLP) model is incorporated with a subarray scheme, and the directions of the desired coherent signals can be estimated from the corresponding prediction polynomial. In this paper, we use multiple lags to exploit the cyclic statistical information efficiently and to alleviate the difficulty of choosing the optimal time lag. For achieving the best performance of DOA estimation, the choice of the subarray size (*i.e.* the order of the LP model plus one) is considered. We derive an analytical expression of error variance of spectral peak position by using linear approximation for sufficiently high signal-to-noise ratio (SNR) and clarify the optimal subarray size for minimizing the peak position variance. As a result, the proposed method has two advantages: the computational load is relatively reduced and the robustness of estimation is significantly improved. The performance of the proposed approach is verified through numerical examples.

2. PROBLEM FORMULATION

2.1 Data Model and Assumptions

We consider a ULA of M identical and omnidirectional sensors with spacing d, and assume that p narrow-band signals $\{s_k(n)\}$ with zero-mean and center frequency f_c are far enough away and come from distinct directions $\{\theta_k\}$. The received signal $y_i(n)$ at the ith sensor can be expressed by

$$y_i(n) = x_i(n) + w_i(n) \tag{1}$$

$$x_i(n) = \sum_{k=1}^{p} s_k(n) e^{j\omega_0(i-1)\tau_k(\theta)} \tag{2}$$

where $x_i(n)$ and $w_i(n)$ are the noiseless received signal and additive noise, $\omega_0 = 2\pi f_c$, $\tau_k(\theta) = (d/c)\sin\theta_k$, and c is the speed of propagation.

The received signals can be rewritten in a compact form as

$$\boldsymbol{y}(n) = \boldsymbol{A}(\theta)\boldsymbol{s}(n) + \boldsymbol{w}(n) \tag{3}$$

where $\boldsymbol{y}(n)$ and $\boldsymbol{w}(n)$ are the $M\times 1$ vectors of the received signals and noise, $\boldsymbol{s}(n)$ is the $p\times 1$ vector of the incident signals, and the array matrix $\boldsymbol{A}(\theta)$ is given by $\boldsymbol{A}(\theta) = [\boldsymbol{a}(\theta_1), \boldsymbol{a}(\theta_2), \cdots, \boldsymbol{a}(\theta_p)]$ with $\boldsymbol{a}(\theta_k) = [1, e^{j\omega_0\tau_k(\theta)}, \cdots, e^{j\omega_0(M-1)\tau_k(\theta)}]^T$.

In this paper, the array is assumed to be unambiguous. Without loss of generality, under a frequency-flat multipath propagation [4], the first q ($1 \le q \le p$ and $q < 2M/3$) signals are coherent ones from the desired source expressed by $s_k(n) = \beta_k s_1(n)$, where β_k is the multipath coefficient which represents the complex attenuation of the kth signal with respect to the first one $s_1(n)$ with $\beta_k \neq 0$ and $\beta_1 = 1$. The desired source exhibits the second-order cyclostationarity with the cycle frequency α, and it is cyclically uncorrelated with the other signals at this cycle frequency. The noise $\{w_i(n)\}$ are cyclically uncorrelated with themselves and with the incident signals at the considered cycle frequency α. The number of coherent signals q and the cycle frequency α are known or estimated *a priori*.

2.2 Forward-Backward Linear Prediction with Subarrays

Here we consider the case that the interfering signals are absent. The noiseless received signals $\{x_i(n)\}$ in (2) differ only by a phase factor $\omega_0\tau_k(\theta)$, so from Prony's method [11], we can find that the noiseless signals $\{x_i(n)\}$ obey a linear difference equation [5]. By dividing the array into L overlapping subarrays of size m, where $L=M-m+1$ and $m\geq q+1$, *i.e.* the lth forward subarray comprises sensors $\{l,l+1,\cdots,l+m-1\}$, the signal $x_{l+m-1}(n)$ can be exactly predicted as follows [4]

$$x_{l+m-1}(n)=\boldsymbol{x}_{f,l}^T(n)\boldsymbol{a} \tag{4}$$

where $\boldsymbol{x}_{f,l}(n)=[x_l(n),x_{l+1}(n),\cdots,x_{l+m-2}(n)]^T$, $\boldsymbol{a}=[a_{m-1},a_{m-2},\cdots,a_1]^T$, and $\{a_i\}$ are the LP coefficients. Similarly by partitioning the full array into L subarrays with m sensors in the backward direction, we obtain the backward LP equation for the lth backward subarray as

$$x_{L-l+1}^*(n)=\boldsymbol{x}_{b,l}^T(n)\boldsymbol{a} \tag{5}$$

where $\boldsymbol{x}_{b,l}(n)=[x_{M-l+1}(n),x_{M-l}(n),\cdots,x_{L-l+2}(n)]^H$, $(\cdot)^*$ and $(\cdot)^H$ denote the complex conjugate and the Hermitian transpose. Then we get the following FLP and BLP models for the received data

$$y_{l+m-1}(n)=\boldsymbol{y}_{f,l}^T(n)\boldsymbol{a}+\varepsilon_{f,l}(n) \tag{6}$$

$$y_{L-l+1}^*(n)=\boldsymbol{y}_{b,l}^T(n)\boldsymbol{a}+\varepsilon_{b,l}(n) \tag{7}$$

where $\boldsymbol{y}_{f,l}(n)=[y_l(n),y_{l+1}(n),\cdots,y_{l+m-2}(n)]^T$, $\boldsymbol{y}_{b,l}(n)=[y_{M-l+1}(n),y_{M-l}(n),\cdots,y_{L-l+2}(n)]^H$, $\varepsilon_{f,l}(n)$ and $\varepsilon_{b,l}(n)$ are the forward and backward prediction errors given by $\varepsilon_{f,l}(n)=w_{l+m-1}(n)-\boldsymbol{w}_{f,l}^T(n)\boldsymbol{a}$ and $\varepsilon_{b,l}(n)=w_{L-l+1}^*(n)-\boldsymbol{w}_{b,l}^T(n)\boldsymbol{a}$, $\boldsymbol{w}_{f,l}(n)=[w_l(n),w_{l+1}(n),\cdots,w_{l+m-2}(n)]^T$, and $\boldsymbol{w}_{b,l}(n)=[w_{M-l+1}(n),w_{M-l}(n),\cdots,w_{L-l+2}(n)]^H$.

The accumulation of the additive noise in $y_{l+m-1}(n)$, $y_{L-l+1}^*(n)$, $\boldsymbol{y}_{f,l}(n)$ and $\boldsymbol{y}_{b,l}(n)$ will cause the ordinary least squares (LS) or minimum-norm estimate from (6) and (7) to become biased and inconsistent [13], and this estimate will make the DOA estimation unreliable. In the paper, we thus exploit the inherent cyclostationarity of most communication signals to suppress the interfering signals and noise.

3. FBLP-BASED CYCLIC DOA ESTIMATION

3.1 Cyclic Correlation of Noisy Data

First the noiseless signal $x_i(n)$ can be rewritten compactly as

$$x_i(n)=\boldsymbol{b}_i^T(\theta)\boldsymbol{s}(n)=\boldsymbol{s}^T(n)\boldsymbol{b}_i(\theta) \tag{8}$$

where $\boldsymbol{b}_i(\theta)=[e^{j\omega_0(i-1)\tau_1(\theta)},e^{j\omega_0(i-1)\tau_2(\theta)},\cdots,e^{j\omega_0(i-1)\tau_p(\theta)}]^T$. Then from the definition of the cyclic correlation [7], and under the model assumptions, we obtain the cyclic correlation function $r_{y_i,y_k}^{\alpha}(\tau)$ between the noisy signals $y_i(n)$ and $y_k(n)$ as

$$r_{y_i,y_k}^{\alpha}(\tau)=\left\langle y_i(n)y_k^*(n+\tau)e^{-j2\pi\alpha n}\right\rangle=\boldsymbol{b}_i^T(\theta)\boldsymbol{R}_s^{\alpha}(\tau)\boldsymbol{b}_k^*(\theta) \tag{9}$$

where $\langle z(n)\rangle=\lim_{N\to\infty}(1/N)\sum_{n=0}^{N-1}z(n)$ denotes the time average of $z(n)$, τ is the lag parameter, and $\boldsymbol{R}_s^{\alpha}(\tau)$ is the cyclic covariance matrix of the source signals given by

$$\begin{aligned}\boldsymbol{R}_s^{\alpha}(\tau)&=\left\langle \boldsymbol{s}(n)\boldsymbol{s}^H(n+\tau)e^{-j2\pi\alpha n}\right\rangle\\&=\left\langle \boldsymbol{\beta}\,s_1(n)\boldsymbol{\beta}^H s_1^*(n+\tau)e^{-j2\pi\alpha n}\right\rangle=r_s^{\alpha}(\tau)\boldsymbol{\beta}\boldsymbol{\beta}^H\end{aligned} \tag{10}$$

where $\boldsymbol{\beta}$ is the vector of multipath coefficients given by $\boldsymbol{\beta}=[\beta_1,\cdots,\beta_q,\beta_{q+1},\cdots,\beta_p]^T$ with $\beta_{q+1}=\cdots=\beta_p=0$, and $r_s^{\alpha}(\tau)$ is the cyclic autocorrelation function of the signal $s_1(n)$ given by $r_s^{\alpha}(\tau)=\left\langle s_1(n)s_1^*(n+\tau)e^{-j2\pi\alpha n}\right\rangle$.

Clearly the influence of the arbitrary (not necessarily stationary and/or spatially white) noise and interference vanish if the cycle frequency α is appropriately selected, so the signal detection capability can be improved. However, because of the coherency of the q signals from the desired source, we can easily find that the cyclic matrix $\boldsymbol{R}_s^{\alpha}(\tau)$ is singular, and it degrades to the performance of the ordinary cyclic methods.

3.2 Linear Prediction Based DOA Estimation

In the absence of interfering signals, from (1), (6) and (9), we obtain the cyclic correlation $r_{y_{l+m-1},y_M}^{\alpha}(\tau)$ between $y_{l+m-1}(n)$ in the lth forward subarray and $y_M(n)$ as

$$\begin{aligned}r_{y_{l+m-1},y_M}^{\alpha}(\tau)&=\left\langle y_{l+m-1}(n)y_M^*(n+\tau)e^{-j2\pi\alpha n}\right\rangle\\&=\left\langle \boldsymbol{y}_{f,l}^T(n)y_M^*(n+\tau)e^{-j2\pi\alpha n}\right\rangle\boldsymbol{a}=\boldsymbol{\varphi}_{f,l}^T(\tau)\boldsymbol{a}\end{aligned} \tag{11}$$

where $\boldsymbol{\varphi}_{f,l}(\tau)=[r_{y_l,y_M}^{\alpha}(\tau),r_{y_{l+1},y_M}^{\alpha}(\tau),\cdots,r_{y_{l+m-2},y_M}^{\alpha}(\tau)]^T$. Equivalently, we can obtain the cyclic correlation $r_{y_1,y_{L-l+1}}^{\alpha}(\tau)$ between $y_1(n)$ and $y_{L-l+1}(n)$ in the lth backward subarray as

$$\begin{aligned}r_{y_1,y_{L-l+1}}^{\alpha}(\tau)&=\left\langle y_1(n)y_{L-l+1}^*(n+\tau)e^{-j2\pi\alpha n}\right\rangle\\&=\left\langle y_1(n)\boldsymbol{y}_{b,l}^H(n+\tau)e^{-j2\pi\alpha n}\right\rangle\boldsymbol{a}=\boldsymbol{\varphi}_{b,l}^T(\tau)\boldsymbol{a}\end{aligned} \tag{12}$$

where $\boldsymbol{\varphi}_{b,l}(\tau)=[r_{y_1,y_{M-l+1}}^{\alpha}(\tau),r_{y_1,y_{M-l}}^{\alpha}(\tau),\cdots,r_{y_1,y_{L-l+2}}^{\alpha}(\tau)]^T$.

As shown in (9) and (10), even in the presence of interfering signals, the influence of the interfering signals and noise are eliminated by exploiting the cyclostationarity, we can find that the prediction relations (11) and (12) in the cyclic domain are valid when the interfering signals are present. Now we consider the DOA estimation of the desired coherent cyclostationary signals by utilizing the LP technique. By letting $l=1$ to L, from (11) and (12), we can obtain the following FBLP equation

$$\boldsymbol{z}(\tau)=\boldsymbol{\Phi}(\tau)\boldsymbol{a} \tag{13}$$

where $\boldsymbol{z}(\tau)=[\boldsymbol{z}_f^T(\tau),\boldsymbol{z}_b^T(\tau)]^T$, $\boldsymbol{\Phi}(\tau)=[\boldsymbol{\Phi}_f^T(\tau),\boldsymbol{\Phi}_b^T(\tau)]^T$, $\boldsymbol{z}_f(\tau)=[r_{y_m,y_M}^{\alpha}(\tau),r_{y_{m+1},y_M}^{\alpha}(\tau),\cdots,r_{y_M,y_M}^{\alpha}(\tau)]^T$, $\boldsymbol{z}_b(\tau)=[r_{y_1,y_{M-m+1}}^{\alpha}(\tau),r_{y_1,y_{M-m}}^{\alpha}(\tau),\cdots,r_{y_1,y_1}^{\alpha}(\tau)]^T$, $\boldsymbol{\Phi}_f(\tau)=[\boldsymbol{\varphi}_{f,1}(\tau),\boldsymbol{\varphi}_{f,2}(\tau),\cdots,\boldsymbol{\varphi}_{f,L}(\tau)]^T$, and $\boldsymbol{\Phi}_b(\tau)=[\boldsymbol{\varphi}_{b,1}(\tau),\boldsymbol{\varphi}_{b,2}(\tau),\cdots,\boldsymbol{\varphi}_{b,L}(\tau)]^T$.

To combat the rank deficiency resulting from signal coherency, we have the following proposition.

Proposition: If the array is partitioned properly to ensure $2L\geq q$, the rank of the cyclic matrix $\boldsymbol{\Phi}(\tau)$ in (13) equals the number of the desired coherent signals.

Proof: By defining $\boldsymbol{A}_1(\theta)$ and $\boldsymbol{A}_2(\theta)$ as the submatrices of the array steering matrix $\boldsymbol{A}(\theta)$ consisting of the first $m-1$ and L rows respectively, after some manipulations, we can obtain [14]

$$\begin{aligned}\boldsymbol{\Phi}(\tau)&=r_s^{\alpha}(\tau)\rho_M^*\begin{bmatrix}\boldsymbol{A}_2(\theta)\boldsymbol{B}\\ \rho_1\boldsymbol{A}_2(\theta)\boldsymbol{B}^*\boldsymbol{D}^{-(M-1)}/\rho_M^*\end{bmatrix}\boldsymbol{A}_1^T(\theta)\\&=r_s^{\alpha}(\tau)\rho_M^*\boldsymbol{C}\boldsymbol{B}\boldsymbol{A}_1^T(\theta)\end{aligned} \tag{14}$$

where $\rho_i=\boldsymbol{b}_i^T(\theta)\boldsymbol{\beta}$, $\boldsymbol{B}=\mathrm{diag}(\beta_1,\beta_2,\cdots,\beta_p)$, $\boldsymbol{D}=\mathrm{diag}(e^{j\omega_0\tau_1(\theta)},e^{j\omega_0\tau_2(\theta)},\cdots,e^{j\omega_0\tau_p(\theta)})$, $\boldsymbol{C}=[\boldsymbol{A}_2^T(\theta),(\boldsymbol{A}_2(\theta)\boldsymbol{\Gamma})^T]^T$, $\boldsymbol{\Gamma}=\mathrm{diag}(\gamma_1,\cdots,\gamma_q,\gamma_{q+1},\cdots,\gamma_p)$, and $\gamma_i=(\rho_1\beta_i^*/\rho_M^*\beta_i)e^{-j\omega_0(M-1)\tau_i(\theta)}$ for $i=1,2,\cdots,q$ while $\gamma_i=0$ for $i=q+1,\cdots,p$.

From the model assumptions, we have $\mathrm{rank}(\boldsymbol{B})=\mathrm{rank}(\boldsymbol{\Gamma})=q$ and $\mathrm{rank}(\boldsymbol{A}_1(\theta))=\min(m-1,p)$ and $\mathrm{rank}(\boldsymbol{A}_2(\theta))=\min(L,p)$. Consequently, from the fact that $m\geq q+1$ and $q\leq p$, we can obtain that $\mathrm{rank}(\boldsymbol{A}_1(\theta))\geq q$. Additionally, the rank of the matrix $\boldsymbol{C}$ is given by $\mathrm{rank}(\boldsymbol{C})=\min(2L,p)$, so $\mathrm{rank}(\boldsymbol{C})=q$ iff $2L\geq q$. Thus if $2L\geq p$, the rank of the cyclic matrix $\boldsymbol{\Phi}(\tau)$ is equal to the number of the desired signals q regardless of the coherence of these signals. Here the fact $\rho_M\neq 0$ and the assumption $r_s^{\alpha}(\tau)\neq 0$ are used implicitly. ■

However, the matrix $\boldsymbol{\Phi}(\tau)$ is usually rank-deficient because $q\leq 2L$ and $q\leq m-1$, so we use the truncated singular value decomposition (SVD) to obtain a numerically reliable estimation, where the SVD of the matrix $\boldsymbol{\Phi}(\tau)$ is given by

$$\boldsymbol{\Phi}(\tau)=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^H \tag{15}$$

where $\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_{2L}]$, $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{m-1}]$, and $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_{\min(2L,m-1)})$. Then from (13), the minimum-norm estimate

of the LP parameter $\boldsymbol{a}$ is obtained [5], [14]

$$\hat{\boldsymbol{a}} = \sum_{i=1}^{q} \lambda_i^{-1} \boldsymbol{v}_i \boldsymbol{u}_i^H \boldsymbol{z}(\tau) \tag{16}$$

Finally by finding the phase of the q zeros of the polynomial $D(z) = 1 - \hat{a}_1 z^{-1} - \hat{a}_2 z^{-2} - \cdots - \hat{a}_{m-1} z^{-(m-1)}$ closest to the unit circle in the z-plane, or by searching for the q highest peaks of the spectrum $1/|D(e^{j\omega_0(d/c)\sin\theta})|^2$, the directions of the desired coherent signals can be estimated.

3.3 Cyclic DOA Estimation Algorithm

As the cyclic correlation function is dependent on the lag parameter τ [7], if the cyclic correlation of one source is zero or insignificant for a given τ, then this signal will not be resolved. The choice of the optimal lag parameter is important in cyclic methods [8], but it is rarely available. For alleviating the difficulty in choosing the optimal lag and to exploit the cyclic statistics effectively, we use multiple lags to obtain a robust estimate of the LP parameter $\boldsymbol{a}$. By concatenating (13) for $\tau = -Q, \cdots, -1, 0, 1, \cdots, Q$, we can obtain a modified cyclic vector-matrix form as

$$\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{a} \tag{17}$$

where $\boldsymbol{z} = [\boldsymbol{z}^T(-Q), \cdots, \boldsymbol{z}^T(-1), \boldsymbol{z}^T(0), \boldsymbol{z}^T(1), \cdots, \boldsymbol{z}^T(Q)]^T$, and $\boldsymbol{\Phi} = [\boldsymbol{\Phi}^T(-Q), \cdots, \boldsymbol{\Phi}^T(-1), \boldsymbol{\Phi}^T(0), \boldsymbol{\Phi}^T(1), \cdots, \boldsymbol{\Phi}^T(Q)]^T$. Here we can choose Q large enough so that $\hat{r}^{\alpha}_{y_i,y_i}(\tau)$ is non-zero and significantly varying for $|\tau| > Q$ [14]. Then we can estimate the directions of the desired coherent signals with the cycle frequency α from (17).

In summary, the proposed FBLP-based DOA estimation algorithm from the finite array data $\{y_1(n), y_2(n), \cdots, y_M(n)\}_{n=0}^{N-1}$ is as follows.

a) Set the subarray size m to satisfy $m \geq q+1$ and $2L \geq q$, where $L = M - m + 1$.

b) Calculate the estimates of the cyclic correlations $r^{\alpha}_{y_i,y_M}(\tau)$ and $r^{\alpha}_{y_i,y_i}(\tau)$ for $\tau = -Q, \cdots, -1, 0, 1, \cdots, Q$ as

$$\hat{r}^{\alpha}_{y_i,y_k}(\tau) = (1/N) \sum_{n=0}^{N-1-\tau} y_i(n) y_k^*(n+\tau) e^{-j2\pi\alpha n}, \quad \text{for } \tau \geq 0 \tag{18}$$

$$\hat{r}^{\alpha}_{y_i,y_k}(\tau) = (1/N) \sum_{n=-\tau}^{N-1} y_i(n) y_k^*(n+\tau) e^{-j2\pi\alpha n}, \quad \text{for } \tau < 0 \tag{19}$$

where $i = 1,2,\cdots,M$ and $k = M$ for $r^{\alpha}_{y_i,y_M}(\tau)$, while $k = 1,2 \cdots, M$ and $i = 1$ for $r^{\alpha}_{y_1,y_k}(\tau)$.

c) Form the estimated cyclic vector $\hat{\boldsymbol{z}}$ and matrix $\hat{\boldsymbol{\Phi}}$ as (17) by using (18), (19) and (13).

d) Perform the SVD on the estimated matrix $\hat{\boldsymbol{\Phi}}$ as (15), where L is replaced by $\bar{L} = (2Q+1)L$.

e) Calculate the estimate of the LP parameter $\boldsymbol{a}$ as

$$\hat{\boldsymbol{a}} = \sum_{i=1}^{q} \hat{\lambda}_i^{-1} \hat{\boldsymbol{v}}_i \hat{\boldsymbol{u}}_i^H \hat{\boldsymbol{z}} \tag{20}$$

f) Estimate the DOA of the signals from the q highest peak locations of the spectrum given by $1/|D(e^{j\omega_0(d/c)\sin\theta})|^2$.

Remark: Calculating the cyclic correlations for multiple lags takes approximately $52N_\tau NM$ flops, where a flop is defined as a floating-point addition or multiplication operation as adopted by MATLAB. The number of flops needed by the SVD of matrix $\hat{\boldsymbol{\Phi}}$ is of the order $O((2LN_\tau)^2(m-1))$, while the computation of $\hat{\boldsymbol{a}}$ requires $8(m-1)(q^2 + 2LN_\tau q + 2LN_\tau) + q$ flops. Thus a rough estimate of the number of MATLAB flops required by the dominant steps in the implementation of proposed approach is $52N_\tau NM$ when $N >> M$, where the computations needed by the remaining steps are negligible.

3.4 Optimal Subarray Size

For estimating the directions of the q coherent signals, from Proposition, it follows that the subarray size m (*i.e.* the order of the prediction model plus one) must be chosen to satisfy the inequality $q+1 \leq m \leq M - q/2 + 1$ [14]. The choice of optimal value of m is crucial to achieve the best performance of direction estimation [5], but it generally depends on the number of desired coherent signals, the SNR and the angle separation of incident signals.

Now we investigate the choice of the subarray size to minimize the variance of peak position error of the spectrum $1/|D(e^{j\omega_0(d/c)\sin\theta})|^2$. The derivation of the error variance of spectral peak position for direction estimation is tedious, so here we only give the result for the sufficiently high SNR. As the interfering signals are suppressed in the proposed cyclic approach, for notational simplicity, we assume that the interfering signals are absent and that the noise is temporally and spatially uncorrelated white complex Gaussian noise, *i.e.* $p = q$, $\beta_k \neq 0$ for $k = 1,2, \cdots, p$, and $E\{w_i(n)w_k^*(n)\} = \sigma^2\delta_{i,k}$ and $E\{w_i(n)w_k(n)\} = 0$, where $E\{\cdot\}$ and $\delta_{i,k}$ denote the expectation and Kronecker delta. As the true parameters $\{a_i\}$ with order $m-1$ can be determined exactly by using the method of undetermined coefficients [11], by adopting the linear approximation as used in [12], we can obtain the variance for the peak position error in terms of noise variance, signal power and subarray size as follows [14]

$$\mathrm{var}(\hat{\omega}_k) \approx \begin{cases} \dfrac{\sigma^2}{3(m-1)L^2|\beta_k|^2 r_s}, & \text{for } m \leq M/2+1 \\ \dfrac{(3m(m-2) - 2L^2 + 2)\sigma^2}{3m^2(m-1)^2 L|\beta_k|^2 r_s}, & \text{for } m > M/2+1 \end{cases} \tag{21}$$

where ω_k denotes the "spatial frequency" $\omega_0\tau_k(\theta)$ for convenience, and $r_s = E\{s_1(n)s_1^*(n)\}$. Therefore we can find that $\mathrm{var}(\hat{\omega}_k)$ increases with subarray size m for $m > M/2+1$ while $\mathrm{var}(\hat{\omega}_k)$ has the minimum m at about $M/3+1$ for $m > M/2+1$. It is straightforward to show that the minimum variance of ω_k (and hence θ_k) can be obtained when $m \approx M/3+1$.

4. NUMERICAL EXAMPLES

The effectiveness of the proposed cyclic FBLP-based direction estimation method is illustrated through numerical examples, in which the desired coherent binary phase-shift keying (BPSK) signals can be distinguished from the interfering BPSK signals with different cycle frequencies. In the simulations, the sensor separation of the ULA with $M = 8$ is half-wavelength, where $f_c = 8$ MHz, $c = 3\times10^6$ m/s, the sensor outputs are collected at the rate $f_s = 8$ MHz, and the lag parameter Q is chosen as $Q = 10$. The BPSK signals have a raised-cosine pulse shape with 50% excess bandwidth. The additive noise is temporally and spatially uncorrelated white complex Gaussian noise with zero-mean and variance σ^2. The SNR is defined as the ratio of the power of the source signals to that of the noise at each sensor. The results shown below are all based on 100 independent trials.

Example 1: Performance versus SNR

The direct-path signal from the BPSK 1 source impinges on the array from angle $\theta_1 = -10°$ with 1.6 MHz baud rate ($\alpha = 0.2$ normalized to the sampling rate [8]), while one coherent arrival comes from $\theta_2 = 4°$ with multipath coefficient $\beta_2 = 1$. There is one interfering BPSK 2 signal that arrives from $\theta_3 = 0°$ with 2.0 MHz baud rate ($\alpha = 0.25$). The number of snapshots and the subarray size are $N = 512$ and $m = 5$. The SNR of the desired is varied, while that of the interference is fixed at 10 dB. The root mean-squared-errors (RMSEs) of the estimates and Cramer-Rao lower bound (CRLB) [3] versus SNR are shown in Fig. 1. Because SS-based MUSIC [4] and smoothed LP method [6] do not exploit the temporal properties of the incoming signals, they

are unable to distinguish the desired signals from the interference correctly even when the dimension of signal subspace is assumed to be the number of coherent signals. Although the RMSE of estimate θ_2 obtained by MODE [2] decreases as the SNR increases, the performance of MODE degrades severely at low SNR, and the estimate θ_1 has a rather large RMSE. Except at very low SNR, the proposed approach performs better, and it is more accurate than SS-based cyclic MUSIC [10] with its RMSE very close to the CRLB at higher SNR.

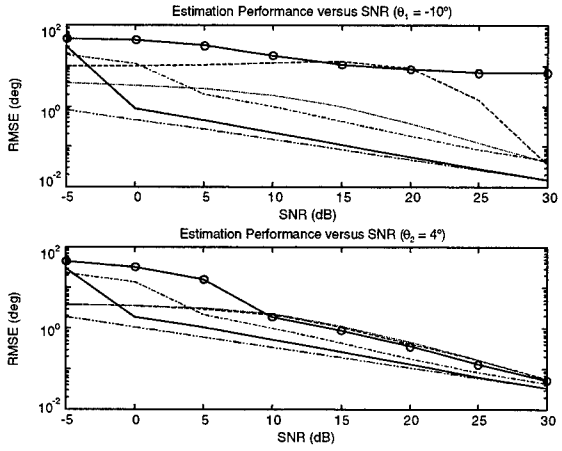


Fig. 1 RMSEs of the estimates versus SNR (dotted: SS-based MUSIC; dashed: smoothed LP; dash-dot: SS-based cyclic MUSIC; solid with "o": MODE; solid: the proposed approach; and dash-dots: CRLB).

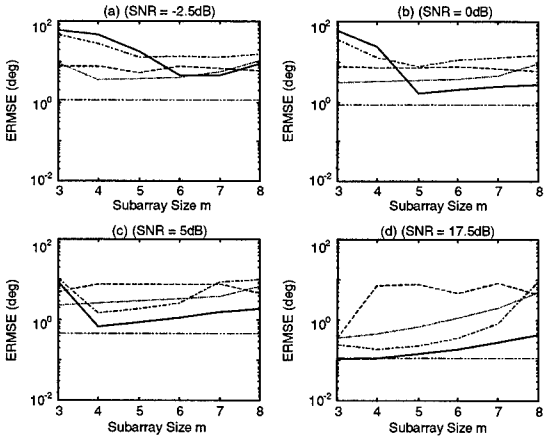


Fig. 2 ERMSEs of the estimates versus subarray size (dotted: SS-based MUSIC; dashed: smoothed LP; dash-dot: SS-based cyclic MUSIC; solid: the proposed approach; and dash-dots: empirical CRLB).

Example 2: Performance versus Subarray Size

The simulation parameters are the same as that in Example 1, except that the subarray size m is varied from 3 to 8. For measuring the overall estimation performance, we define an "empirical RMSE (ERMSE)" of the estimated directions as

$$\text{ERMSE} = \sqrt{(1/qK)\sum_{i=1}^{q}\sum_{k=1}^{K}(\hat{\theta}_i^{(k)} - \theta_i)^2} \qquad (22)$$

where K is the number of trials, and $\theta_i^{(k)}$ is the estimate obtained in the kth trial. Under the SNRs of the desired signal of –2.5 dB, 0 dB, 5 dB and 17.5 dB, the ERMSEs of the estimates against subarray size are shown in Fig. 2, where the "empirical CRLB" is calculated by averaging the corresponding CRLBs over the number of coherent signals. We find that the best estimation can usually be attained when m is about $M/3+1$ for medium and high SNR, while a reasonable estimation can be obtained with a larger value of m for low SNR.

5. CONCLUSIONS

For estimating the directions of coherent cyclostationary signals impinging on a ULA, we proposed a new cyclic FBLP method. In order to improve the estimation performance, multiple lag parameters are used to exploit the cyclic statistics sufficiently and effectively. The optimal subarray size that minimizes the peak position variance was derived using linear approximation for sufficiently high SNR. As a result, the proposed method has two advantages: the computational load is relatively reduced and the robustness of estimation is significantly improved. The performance of the proposed method was verified through numerical examples.

REFERENCES

[1] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 276-280, 1986.

[2] P. Stoica and K.C. Sharman, "Novel eigenanalysis method for direction estimation," *IEE Proc.*, Part F, vol. 137, no. 1, pp. 19-26, 1990.

[3] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Trans. ASSP*, vol. 37, no. 5, pp. 720-741, 1989.

[4] T.J. Shan, M. Wax and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," *IEEE Trans. ASSP*, vol. 33, no. 4, pp. 806-811, 1985.

[5] S. Haykin, "Radar array processing for angle of arrival estimation," in *Array Signal Processing* (S. Haykin, ed.), Englewood Cliffs, NJ: Prentice-Hall, pp. 194-292, 1985.

[6] H. Krim and J.G. Proakis, "Smoothed eigenspace-based parameter estimation," *Automatica*, vol. 30, no. 4, pp. 27-38, 1994.

[7] W.A. Gardner, "Exploitation of spectral redundancy in cyclostationary signals," *IEEE Signal Process. Mag.*, vol. 8, pp. 14-37, April 1991.

[8] G. Xu and T. Kailath, "Direction-of-arrival estimation via exploitation of cyclostationarity – A combination of temporal and spatial processing," *IEEE Trans. Signal Process.*, vol. 40, no. 7, pp. 1775-1786, 1992.

[9] S.V. Schell and W.A. Gardner, "Signal-selective high-resolution direction finding in multipath," *Proc. IEEE ICASSP*, pp. 2667-2670, Albuquerque, NM, April 1990.

[10] J. Xin, H. Tsuji, Y. Hase and A. Sano, "Directions-of-arrival estimation of cyclostationary coherent signals in array processing," *IEICE Trans. Fundamentals*, vol. E81-A, no. 8, pp. 1560-1569, 1998.

[11] S.M. Kay, *Modern Spectral Estimation: Theory and Application.* Englewood Cliffs, NJ: Prentice-Hall, 1987.

[12] S.W. Lang and J.H. McClellan, "Frequency estimation with maximum entropy spectral estimators," *IEEE Trans. ASSP*, vol. 28, no. 6, pp. 716-724, 1980.

[13] T. Söderström, "Identification of stochastic linear systems in presence of input noise," *Automatica*, vol. 17, no. 5, pp. 713-725, 1981.

[14] J. Xin and A. Sano, "Linear prediction approach to direction estimation of cyclostationary signals in multipath environment," *IEEE Trans. Signal Process.*, vol. 49, no. 4, pp. 710-720, 2001.

ITERATIVE ALGORITHM FOR THE ESTIMATION OF DISTRIBUTED SOURCES LOCALIZATION PARAMETERS

Antonio Pascual Iserte[1], *Ana I. Pérez-Neira*[1] *and Miguel Ángel Lagunas Hernández*[1,2]

[1]Department of Signal Theory and Communications
Polytechnic University of Catalonia (UPC)

[2]CTTC-Centre Tecnològic de Telecomunicacions de Catalunya
Edifici NEXUS I
C/ Jordi Girona 1-3 (Campus Nord UPC - mòdul D5), 08034 Barcelona (SPAIN)

e-mail:{tonip,anuska,miguel}@gps.tsc.upc.es

ABSTRACT

We present a novel algorithm for the estimation of the direction of arrival and angular distribution parameters of sources that, as a result from the scattering effects, cannot be considered as punctual.

The algorithm is iterative and is based on the maximization of the likelihood function associated to the received snapshots at the antenna array (ML). It is proposed a computationally efficient method for estimating the localization and angular distribution parameters of more than one source transmitting at the same frequency in a noisy environment. This algorithm solves the problem of the joint maximization in the case of two sources by formulating two new problems of single-source ML.

Key Words- Array signal processing, DOA estimation, distributed sources, statistical parameter estimation.

1. INTRODUCTION

Classically, the methods for the estimation of the direction of arrival (DOA) have considered punctual sources and spatio-temporal thermal white noise. This problem can be assumed equivalent to those of frequency detection based on temporal diversity. However, whereas in spectral analysis, two different frequencies are always totally uncorrelated, in spatial diversity two signals impinging from different angles can be partially correlated.

The multipath propagation implies an increase in the temporal correlation between signals from different directions, making the performance of the classical spectral analysis techniques worse. Besides the spatial smoothing techniques, the most effective solution to this problem is represented by the spreading systems, such as radar "pulse compression techniques" and spread spectrum communications (DSSS), which are based on an increase of the signal bandwidth. By means of these techniques, the classical non-parametric spectral analysis algorithms can be applied, converting the DOA detection in a multipath environment in multiple uncorrelated punctual sources DOA detection problems, even with minimum time shift differences between echoes.

However, the presence of scatterers near the transmitter with no relative delays between different DOAs, makes the performance of the spreading systems worse. The source must be considered as distributed, and therefore the classical methods may fail because of the high spatio-temporal correlation. As in the case of specular multipath, this problem cannot be solved by manipulating the transmitted signal. The distributed source signs with a unique temporal waveform and a spatial signature. Although the scattering may change over the time, it can be considered time-invariant within the frame duration in most of the cases. That means that in large periods of time, the correlation matrix of the received snapshots, considering a free-noise environment, is full-rank [1] [2], whereas in short periods, due to the local analysis of the scenario, each source is contributing as a rank one covariance matrix, so it is completely coherent during a frame. Our interest is to characterize this quasi-static behaviour of a source, and not the inter-frame or inter-scan changes [1] [2] [3].

The proposed technique consists in maximizing the likelihood function associated to the parameters of the angular distribution of the sources. This represents a multi-variable maximization problem with a very high computational cost. Solutions based on EM (Estimate & Maximize) [4] [5] or AP (Alternating Projection) [6] and RCAP (Reduced Complexity Array Processing) [7] [8] convert this problem into multiple one-dimensional problems.

This paper presents an algorithm for the estimation of the spatial signature of multiple distributed sources with rank one contributions, that is, estimation within the time duration of a frame. In the case of a more prolongated observation period, the classical spectral analysis methods can be applied. In general terms, this work presents the generalization of the AP and RCAP techniques to the case of distributed sources.

2. SIGNAL MODEL

In the case of a single source scenario and an array of antennas, a known angle distribution can be assumed $f_0(\theta, n)$, whose mean is θ_0 and n is the temporal index of the received snapshot. The snapshot model is as follows [9]:

$$\begin{aligned} \mathbf{x}_n &= a(n)\int_{-\pi/2}^{\pi/2} f_0(\theta,n)\mathbf{s}(\theta)d\theta + \mathbf{w}_n = a(n)\mathbf{b}_n + \mathbf{w}_n \\ \mathbf{b}_n &= \int_{-\pi/2}^{\pi/2} f_0(\theta,n)\mathbf{s}(\theta)d\theta \end{aligned} \quad (1)$$

where $a(n)$ is the complex envelope of the transmitted signal, $\mathbf{s}(\theta)$ is the steering vector for a punctual source in the elevation angle θ and $\mathbf{w}_n$ is the noise contribution at the front-end. In this model the complex envelope is the same for all the angles of arrival of the source, so it is totally correlated. The goal is to estimate the spatial signature $\mathbf{b}_n$ of the source, which is already defined in (1).

In the case of long or inter-frame observation periods, the spatial signature can change, and so its temporal correlation can be

This work was partially supported by the European Commission under project IST-1999-10322 SATURN; the Spanish Government (CICYT) TIC98-0703, TIC99-0849, TIC2000-1025, FIT-070000-2000-649; and the Catalan Government (CIRIT) 2001FI 00714, 2000SGR 00083.

exploited by the system to update the estimate of the source movement or position, in the case of low-mobility environments. The inter-frame spatial signature tracking system can be based on a Kalman filter, although this is outside the scope of this work [8].

Our interest is centered in the spatial signature estimate for the case of non-varying sources. This is the general situation in a space-time diversity scheme at the receiver and/or the transmitter side of the communication system, in which an estimate of $\mathbf{b}$ is required within a frame duration. Although this signature can change in time periods longer than the frame duration, it can be assumed constant within one frame and the N received snapshots, so the signal model is as follows:

$$\mathbf{x}_n = a(n)\mathbf{b} + \mathbf{w}_n \tag{2}$$

where now it is remarked that the spatial signature $\mathbf{b}$ does not depend on the temporal index n. This is the same situation as the one described in [2] for the case of *Coherently Distributed Sources*. In that case, the angular distribution $f_0(\theta)$, called *deterministic angular signal density*, is assumed to be known.

In order to estimate the spatial signature, the preamble and reference symbols in the frame may be used. In the problem we assume, no reference signal is used. In this case, and taking equation (2) as a basis, the estimated covariance matrix for a scenario with NS independent sources is as follows:

$$\widehat{\mathbf{R}} = \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{x}_n\mathbf{x}_n^H \cong \sum_{s=1}^{NS}\alpha_s\mathbf{b}_s\mathbf{b}_s^H + \sigma^2\mathbf{I} \tag{3}$$

$$\sigma^2\mathbf{I} = E\left\{\mathbf{w}_n\mathbf{w}_n^H\right\} \quad \alpha_s = E\left\{|a_s(n)|^2\right\}$$

It will be assumed in the theoretical description of the algorithm, that there is no error in the estimation of the covariance matrix. In the simulations, the impact of the variation of the number of snapshots will be shown.

Our goal is the estimation of the parameters of the distribution $f_0(\theta)$. It is practical to consider a discrete model of the spatial signature instead of its integral model (1). We consider the contribution of the same signal impinging from M different angles, where M is much larger than the total number of sensors Q [9]:

$$\mathbf{b} = \sqrt{Q}\frac{\mathbf{S}\,\mathbf{f}_0}{\|\mathbf{S}\,\mathbf{f}_0\|} \quad \Longrightarrow \quad \frac{\mathbf{b}^H\mathbf{b}}{Q} = 1 \tag{4}$$

where $\mathbf{S} = \left[\begin{array}{cccc}\mathbf{s}(\theta(0)) & \mathbf{s}(\theta(1)) & \cdots & \mathbf{s}(\theta(M-1))\end{array}\right]$ and we have normalized the spatial signature $\mathbf{b}$ so that the mean signal power measured at the sensors of the array is equal to the signal power α_s. $\theta(k)$ represents the discretized angular axis.

At this point, it is only necessary to parametrize the distribution. The experimental results and measurements at 2 GHz indicate that the best-fitting model is the exponential one, with its mean value situated at the real position of the source. The discretization of the angular distribution is expressed in the vector $\mathbf{f}_0 = \left[\begin{array}{cccc} f_0(0) & f_0(1) & \cdots & f_0(M-1)\end{array}\right]^T$, where $f_0(m) = \exp(-c_0\,|\theta(m) - \theta_0|)\ \ 0 \le m \le M-1$. In the simulations, not only a Laplacian, but also Gaussian and rectangular profiles have been proved. In all the cases, the performance of the algorithms described throughout the paper is good. In the simulations section, the results for the case of a Laplacian distribution are shown.

In section 3 our goal is to estimate the spread parameter c_0 and the mean angle of arrival θ_0 based on the estimate of the covariance matrix in the case of a single-source scenario. In the next sections, the generalization to the case of a two sources environment is discussed through algorithms based on AP and RCAP.

3. SINGLE SOURCE ESTIMATION

It is well known that the maximum likelihood (ML) estimator of the spatial signature $\mathbf{b}$ in an AWGN environment is as follows:

$$\widehat{\mathbf{b}} = \arg\max\frac{\widehat{\mathbf{b}}^H\widehat{\mathbf{R}}\widehat{\mathbf{b}}}{\widehat{\mathbf{b}}^H\widehat{\mathbf{b}}} \Rightarrow \widehat{\mathbf{b}} = k\mathbf{e},\ \widehat{\mathbf{R}}\mathbf{e} = \lambda_{\max}\mathbf{e},\ \|\mathbf{e}\| = 1 \tag{5}$$

so, it is an eigenvector problem, where the maximum eigenvalue must be chosen. Due to errors in the estimation of the the covariance matrix, it is possible that the eigenvector $\mathbf{e}$ does not exactly fit the parametric definition of the spatial signature $\mathbf{b}$ as expressed in (4). A MSE (Minimum Square Error) criterion is proposed so as to fit the distribution parameters to the eigenvector $\mathbf{e}$:

$$\left[\widehat{\theta}_0, \widehat{c}_0, \widehat{\beta}_0\right] = \arg\min\left\|\sqrt{\lambda_{\max}}\mathbf{e} - \widehat{\beta}_0\mathbf{b}(\widehat{\theta}_0, \widehat{c}_0)\right\|^2 \tag{6}$$

where $|\beta_0|$ is an estimate of the RMS value of the source and $\mathbf{b}$ is defined as shown in (4). The validity of the expression is based on the idea that the maximum eigenvalue is an approximated measurement of the source power in a single source and typical SNR environment. In the case of extremely low SNR conditions, a noise calibration should be carried out.

The mean value of the angular distribution can be easily estimated through the following expression, where the spatial response of the eigenvector $\mathbf{e}$ is calculated:

$$\widehat{\theta}_0 = \arg\max\left|\mathbf{s}^H(\widehat{\theta}_0)\mathbf{e}\right|^2 \tag{7}$$

We admit that this estimator has a bias. However, for sources situated within the angle view [-40°,40°] and typical spreadings in mobile communications, the deviation is minimum. The great advantage is that, making use of this estimator, a two dimensional search (mean angle and spreading parameter) is avoided in (5).

The estimate of the parameter $|\beta_0|$ can be expressed in function of the estimate of the spatial signature, where now, only an unidimensional search on the spreading parameter c_0 is necessary based on the MSE expression (6).

$$\widehat{\mathbf{b}} = \sqrt{Q}\frac{\mathbf{S}\,\widehat{\mathbf{f}}_0}{\left\|\mathbf{S}\,\widehat{\mathbf{f}}_0\right\|} \quad \widehat{\mathbf{f}}_0 = \mathbf{f}_0(\widehat{\theta}_0, \widehat{c}_0) \quad \widehat{\beta}_0 = \sqrt{\lambda_{\max}}\frac{\widehat{\mathbf{b}}^H\mathbf{e}}{\widehat{\mathbf{b}}^H\widehat{\mathbf{b}}} \tag{8}$$

At this point, it is important to highlight that the traditional methods suffer important degradations when trying to estimate parameters of distributed sources using the classical punctual source model. It is the same degradation as that produced by a system with an uncalibrated array.

4. TWO SOURCES ESTIMATION

Now the previous method is extended to the case of a scenario with two distributed sources. The extrapolation of the algorithm to the case of more sources is direct, although the computational cost grows importantly. In most of the real communication systems, it is not exaggerated to assume that only two independent sources with no negligible power level are radiating, where one of them can be considered as the desired signal and the other as the interference.

In the following presentation, the spatial signatures $\mathbf{b}_1$ and $\mathbf{b}_2$ will be used. The algorithm is iterative, and in each step the

parameters of one of the sources are estimated, based on the previous estimates of the other source parameters. It can be shown in the simulations that the algorithm converges and the quality of the estimates improves as the number of iterations grows. In this paper, one of the steps is presented, where a previous estimate of the second source is assumed:

$$\widehat{\mathbf{b}}_2 = \sqrt{Q}\frac{\mathbf{S}\,\widehat{\mathbf{f}}_2}{\left\|\mathbf{S}\,\widehat{\mathbf{f}}_2\right\|} \quad \widehat{\mathbf{f}}_2 = \mathbf{f}_0(\widehat{\theta}_2, \widehat{c}_2) \quad \widehat{\beta}_2 = \sqrt{\lambda_{\max}}\frac{\widehat{\mathbf{b}}_2^H \mathbf{e}}{\widehat{\mathbf{b}}_2^H \widehat{\mathbf{b}}_2} \quad (9)$$

Based on this estimate, it is possible to estimate the covariance matrix $\mathbf{R}_1$ of the snapshots without the contribution of the second source. The algorithm presented in section 3 is now applied to $\widehat{\mathbf{R}}_1$ so as to estimate the parameters of the first source. The mechanism must be applied iteratively to obtain admissible estimates.

$$\widehat{\mathbf{R}}_1 = \widehat{\mathbf{R}} - \left|\widehat{\beta}_2\right|^2 \widehat{\mathbf{b}}_2 \widehat{\mathbf{b}}_2^H \quad (10)$$

It is interesting to comment a conflictive point in this algorithm, which deals with the second source RMS power estimate $\widehat{\beta}_2$ (9), similar to the one deduced by Li and Stoica [10]. Equation (10) does not guarantee that $\widehat{\mathbf{R}}_1$ is positive defined. The maximum source power estimate that guarantees it is $\left(\widehat{\mathbf{b}}_2^H \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{b}}_2\right)^{-1}$ while the MSE based $\left|\widehat{\beta}_2\right|^2$ estimate (6) (9) may be greater. However, in the simulations it will be proved that the algorithm works well, although some of the eigenvalues may be negative in the first steps of the iterative mechanism.

The algorithm presented up to this point and based in (10) is called *"substraction method"*. Now, we present another method based on a blocking matrix as the AP [6] or RCAP [8] algorithms, whose name is *"blocking method"*. In this case we consider the second source as Gaussian noise with a known covariance matrix. Our goal is to estimate the parameters of the first source based on the maximization of the log-likelihood associated to the received snapshots. Taking into account all the snapshots, and assuming that the symbols emitted from the second source are independent and the noise is white, the estimated spatial signature for the first source is calculated as follows (see Appendix):

$$\widehat{\mathbf{b}}_1 = \arg\max \frac{\widehat{\mathbf{b}}_1^H \mathbf{P} \widehat{\mathbf{R}} \mathbf{P} \widehat{\mathbf{b}}_1}{\widehat{\mathbf{b}}_1^H \mathbf{P} \widehat{\mathbf{b}}_1} \Rightarrow \widehat{\mathbf{b}}_1 = k\mathbf{e}, \quad (11)$$

$$\widehat{\mathbf{R}}\mathbf{P}\mathbf{e} = \lambda_{\max}\mathbf{e}, \quad \|\mathbf{e}\| = 1$$

where $\mathbf{P}$, called blocking matrix, is defined as follows:

$$\mathbf{P} = \mathbf{I} - \phi_2 \widehat{\mathbf{b}}_2 \widehat{\mathbf{b}}_2^H \quad (12)$$

$$\phi_2 = \frac{SNR_2}{1 + \widehat{\mathbf{b}}_2^H \widehat{\mathbf{b}}_2 SNR_2} \qquad SNR_2 = \frac{\left|\widehat{\beta}_2\right|^2}{\sigma^2}$$

Equation (11) is a modified eigenvector problem very similar to the one described in the case of AP or RCAP algorithms. As explained in section 3, when the eigenvector $\mathbf{e}$ is calculated, the spatial signature $\mathbf{b}_1$ and the distribution parameters of the first source (mean DOA and spreading parameter) must be fitted as expressed in the next equation, based on a MSE criterion:

$$\left[\widehat{\theta}_1, \widehat{c}_1, \widehat{\beta}_1\right] = \arg\min \left\| \sqrt{\frac{\lambda_{\max}}{1 - \phi_2 \left|\mathbf{e}^H \widehat{\mathbf{b}}_2\right|^2}}\mathbf{e} - \widehat{\beta}_1 \mathbf{b}_1(\widehat{\theta}_1, \widehat{c}_1) \right\|^2 \quad (13)$$

The constant multiplying the eigenvector is found by calculating the maximum eigenvalue of the matrix $\mathbf{RP}$ which is the one that solves equation (11): $\lambda_{\max} \cong |\beta_1|^2 \|\mathbf{b}_1\|^2 - |\beta_1|^2 \phi_2 \left|\mathbf{b}_1^H \mathbf{b}_2\right|^2$, assuming a sufficiently high SNR environment.

The blocking matrix is defined by the parameter ϕ_2, which depends on the level of the source to be blocked. In the case of AP or RCAP, this blocking is independent of the source level. In this sense, the performance of the algorithm presented in this paper copes better with variations of the source levels. Only for high SNR conditions, the blocking matrix is equal to the case of AP and RCAP:

$$\begin{aligned} SNR_2 &\to \infty \quad \phi_2 \to \left\|\widehat{\mathbf{b}}_2\right\|^{-2} \\ SNR_2 &\to 0 \quad \phi_2 \to SNR_2 \end{aligned} \quad (14)$$

If more than two sources are considered, the matrix $\mathbf{P}$ is defined as the inverse of the correlation matrix of the noise plus the $NS-1$ source signals different from the one whose parameters are being estimated. This can be easily deduced from the Appendix.

5. SIMULATIONS AND RESULTS

Some simulations and results about the performance of the algorithms are now presented. We consider the presence of two independent distributed sources at 8° and -10°, with spreading parameters 0.5 and 0.1 respectively. The smaller the spreading parameter is, the more distributed the source is.

The next figure shows the RMS error in the mean angle estimation of the two sources as a function of the number of antennas and the algorithm. The SNR for each source is 10 dB, and the number of snapshots used to estimate the covariance matrix is 100. 100 simulations have been carried out per point in the curves.

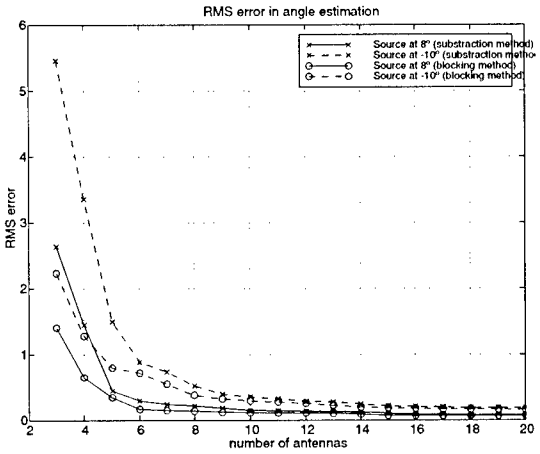


Fig.1. RMS error vs. number of antennas.

It can be observed that the angle estimation of the source at -10°,which is the most distributed, presents higher error. It can be also seen that the blocking method performs better than the substraction method.

The impact of the SNR variation on the estimation performance of the spreading parameters is shown in Fig. 2. The simulation parameters are the same as in the previous figure, where now 8 antennas are used and both sources have the same level. SNR refers to the signal-to-noise ratio of each source. For high SNR

conditions, both methods perform similar.

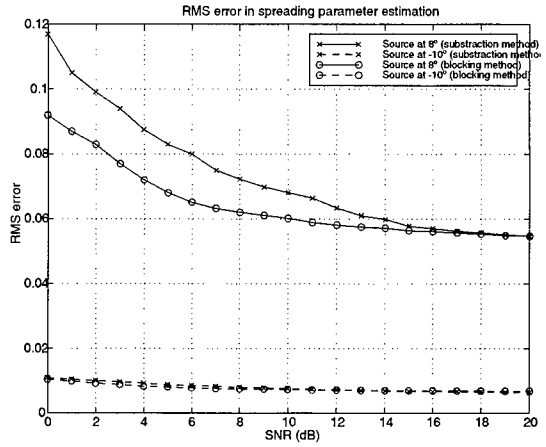


Fig.2. RMS error vs. SNR.

In Fig. 3, the dependence on the number of snapshots used to estimate the covariance matrix is analysed. Both signal sources have a SNR equal to 10 dB and the number of antennas is 8.

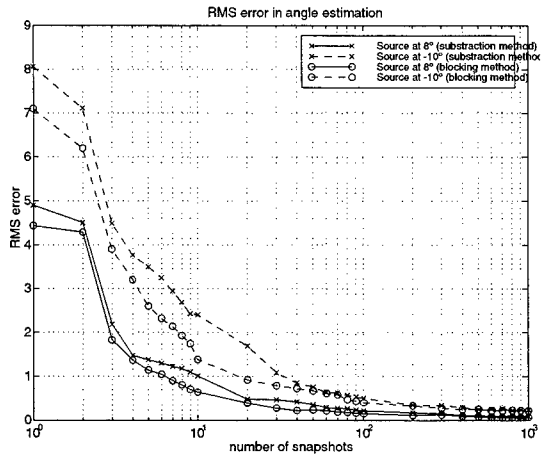


Fig.3. RMS error vs. number of snapshots.

It is concluded that the number of snapshots can impact seriously on the performance, specially when this number is low. For a high number of snapshots, the performance of both methods is equivalent. However, the duration of the frame and the number of snapshots for considering a non-varying source are generally limited.

In all the cases, the initial values in the simulations were based on the assumption that only one source was present in the scenario. The simulations prove that this initial value does not affect negatively the performance of the algorithm.

6. APPENDIX

We present here, the mathematical formulation of the ML estimator described in section 4 for the case of the "blocking method", where in each step, one of the sources is considered as Gaussian noise. The PDF is $f_X(\mathbf{x};\beta_1,\mathbf{b}_1,\sigma^2;\widehat{\beta}_2,\widehat{\mathbf{b}}_2)$:

$$f_X(\mathbf{x}) = \prod_{n=0}^{N-1} \frac{1}{\pi^Q \det\left(\widehat{\mathbf{R}}_2\right)} e^{\left\{-(\mathbf{x}_n - a_1(n)\mathbf{b}_1)^H \widehat{\mathbf{R}}_2^{-1}(\mathbf{x}_n - a_1(n)\mathbf{b}_1)\right\}}$$

$$\widehat{\mathbf{R}}_2 = \sigma^2\mathbf{I} + \left|\widehat{\beta}_2\right|^2 \widehat{\mathbf{b}}_2\widehat{\mathbf{b}}_2^H \qquad \mathbf{P} = K_2\widehat{\mathbf{R}}_2^{-1} = \mathbf{I} - \phi_2\widehat{\mathbf{b}}_2\widehat{\mathbf{b}}_2^H$$

The log-likelihood is expressed as follows $\Lambda_{\mathbf{x}} = \log\left(f_X(\mathbf{x})\right)$:

$$\Lambda_{\mathbf{x}} = K - \sum_{n=0}^{N-1} (\mathbf{x}_n - a_1(n)\mathbf{b}_1)^H \widehat{\mathbf{R}}_2^{-1} (\mathbf{x}_n - a_1(n)\mathbf{b}_1)$$

Maximizing with respect to the information symbols, they can be estimated as expressed in the next equations:

$$\widehat{a}_1(n) = \frac{\mathbf{b}_1^H\widehat{\mathbf{R}}_2^{-1}\mathbf{x}_n}{\mathbf{b}_1^H\widehat{\mathbf{R}}_2^{-1}\mathbf{b}_1} \qquad \mathbf{P}_1 = \mathbf{I} - \frac{\mathbf{b}_1\mathbf{b}_1^H\mathbf{P}}{\mathbf{b}_1^H\mathbf{P}\mathbf{b}_1}$$

$$\begin{aligned}
\Lambda_{\mathbf{x}} &= K - \sum_{n=0}^{N-1} \mathbf{x}_n^H\mathbf{P}_1^H\widehat{\mathbf{R}}_2^{-1}\mathbf{P}_1\mathbf{x}_n = K - \frac{N}{K_2} tr\left(\mathbf{P}_1^H\mathbf{P}\mathbf{P}_1\widehat{\mathbf{R}}\right) \\
&= K - \frac{N}{K_2} tr\left(\mathbf{P}\mathbf{P}_1\widehat{\mathbf{R}}\right) = K - \frac{N}{K_2} tr\left(\mathbf{P}_1\widehat{\mathbf{R}}\mathbf{P}\right) \\
&= K - \frac{N}{K_2}\left\{tr\left(\widehat{\mathbf{R}}\mathbf{P}\right) - tr\left(\frac{\mathbf{b}_1\mathbf{b}_1^H\mathbf{P}\widehat{\mathbf{R}}\mathbf{P}}{\mathbf{b}_1^H\mathbf{P}\mathbf{b}_1}\right)\right\}
\end{aligned}$$

The estimate of the spatial signature for the first source is obtained by maximizing $\Lambda_{\mathbf{x}}$:

$$\widehat{\mathbf{b}}_1 = \arg\max \frac{\widehat{\mathbf{b}}_1^H\mathbf{P}\widehat{\mathbf{R}}\mathbf{P}\widehat{\mathbf{b}}_1}{\widehat{\mathbf{b}}_1^H\mathbf{P}\widehat{\mathbf{b}}_1} \quad \Rightarrow \quad \widehat{\mathbf{R}}\mathbf{P}\mathbf{e} = \lambda_{\max}\mathbf{e}, \quad \|\mathbf{e}\| = 1$$

7. REFERENCES

[1] R. Raich, J. Goldberg, H. Messer, "Bearing Estimation for a Distributed Source: Modeling, Inherent Accuracy Limitations and Algorithms," *IEEE Trans. Signal Proc.*, vol. 48, no. 2, pp. 429-441, February 2000.

[2] S. Valaee, B. Champagne, P. Kabal, "Parametric Localization of Distributed Sources," *IEEE Trans. Signal Proc.*, vol. 43, no. 9, pp. 2144-2153, September 1995.

[3] M. Bengtsson, B. Ottersten, "Low Complexity Estimators for Distributed Sources," *IEEE Trans. Signal Proc.*, vol. 48, no. 8, pp. 2185-2194, August 2000.

[4] D. Kraus, D. Maiwald, J. Bohme, "Maximum Likelihood Source Localization Estimation via EM Algorithm," *Signal Proc.*, pp. 649-652, 1992.

[5] M. Feder, E. Weinstein, "Parameter Estimation of Superimposed Signals Using the EM Algorithm," *IEEE Trans. ASSP*, vol. 36, no. 4, pp. 477-489, April 1988.

[6] I. Ziskind, M. Wax, "Maximum Likelihood Localization of Multiple Sources by Alternating Projection," *IEEE Trans. ASSP*, vol. 36, no. 10, pp. 1553-1560, October 1988.

[7] A. I. Pérez, M. A. Lagunas, "High Performance DOA Trackers derived from Parallel Low Resolution Detectors," *Proc. IEEE-SSAP Workshop*, pp. 558-561, Corfu, Greece 1996.

[8] A. I. Pérez, R. Villarino, M. A. Lagunas, "The RCAP: a Concept for Robust Beamforming and High Resolution DOA Tracking". Submitted to *IEEE Trans. Ant. Prop.*

[9] V. Veen, R. Roberts, "Partially Adaptive Beamformer Design via Output Power Minimization," *IEEE Trans. ASSP*, vol. 35, pp. 1524-1532, November 1987.

[10] P. Stoica, H. Li, J. Li, "Amplitude Estimation of Sinusoidal Signals: Survey, new Results, and an Application," *IEEE Trans. Signal Proc.*, vol. 48, pp. 338-352, February 2000.

ARRAY SIGNAL PROCESSING FOR RECURSIVE TRACKING OF MULTIPLE MOVING SOURCES BASED ON LPA BEAMFORMING

Vladimir Katkovnik, Yonghoon Kim

Kwangju Institute of Science and Technology* (K-JIST), Kwangju, 500-712, Korea

ABSTRACT

The windowed linear local polynomial approximation (LPA) of the time-varying direction-of-arrival (DOA) is developed for nonparametric high-resolution estimation of multiple moving sources. The method gives the estimates of instantaneous values of the directions as well as their first derivatives. The asymptotic variance and bias of these estimates are derived and used for the optimal window size selection. Marginal beamformers are proposed for estimation and sources visualization. These marginal beamformers are able to localize and track every source individually nulling signals from all other moving sources. Recursive implementation of estimation algorithms are developed for two different tasks: estimation of $DOAs$ with varying number of sources and multiple source tracking in time.

1. INTRODUCTION

Localization and tracking multiple narrow band moving sources by a passive array is one of the fundamental problems in radar, communication, sonar, seismology and in other arrears. In recent years a significant progress was achieved on the base of development and application of source movement models for DOA estimation. These techniques are mainly based on the maximum likelihood (ML) which follows Kalman-style recursive algorithms [7] or the expectation-maximization algorithms [2]. Our approach mainly is in line with the ideas of the ML and model identification approach those are known to give high-resolution estimates of $DOAs$ [1].

A recently developed local polynomial approximation (LPA) beamforming is originated as an adjustment of the conventional beamformer to nonstationary environments with moving a single [3]-[5] and multiple sources . It is shown that the LPA beamforming is able to yield a very useful visualization of the DOA of rapidly moving sources as well as improved estimation and high resolution of tight sources. In this paper we use the LPA beamformers in order to obtain efficient Gauss-Newton recursive tracking algorithms.

2. PROBLEM FORMULATION

Let the uniform linear array of n sensors receive q narrowband signals impinging from far-field sources with unknown time-varying directions $\theta_r(t)$, $r = 1, ..., q$. Assume that the $n \times 1$ array observation vector $x(t)$ can be expressed as

$$x(t) = \sum_{r=1}^{q} a_r s_r(t) = A(\Theta(t))\, s(t) + \varepsilon(t), \quad (1)$$

$$A(\Theta(t)) = [a_1, \ldots, a_q],\ a_r = a(\theta_r(t)),\ r = 1, ..., q,$$

where $A(\Theta)$ is the $n \times q$ direction matrix, $a(\theta)$ is the $n \times 1$ steering vector, $\Theta = (\theta_1, \ldots, \theta_q)^T$ is the $q \times 1$ vector of $DOAs$, $s(t)$ is the $q \times 1$ vector of source waveforms, and $\varepsilon(t)$ is the $n \times 1$ vector of a sensor noise.

Let us define the steering vector as $a(\theta) = \left(1, q, \ldots, q^{n-1}\right)^T$, where $q = \exp\{-j\,\frac{2\pi}{\lambda}\, d \sin\theta\}$, d is the interelement spacing, λ is the wavelength, and $(\cdot)^T$ stands for transpose. Assume that the sensor noise is a white zero-mean circular process with $E\{\varepsilon^H \varepsilon\} = \sigma^2$. The problem is to find estimates of $DOAs$ $\hat{\theta}_r(t)$ of $\theta_r(t)$ from observations (1). It is assumed that the directions $\theta_r(t)$ are arbitrary functions of time belonging to a nonparametric class of piece-wise continuous differentiable functions.

3. LPA ESTIMATION OF DOA

Let $C = (c_0, c_1)^T$ be a generic notation for the $2D$ vector with c_0 and c_1 giving estimates of $\theta(t)$ and $\theta^{(1)}(t)$ respectively; C with a superscript "k" in square brackets designates similar estimates for the kth source, $C^{[k]} = (c_{0,k}, c_{1,k})^T$; $\Theta(t) = (\theta_1(t), ..., \theta_q(t))^T$ and $\Theta^{(1)}(t) = (\theta_1^{(1)}(t), ..., \theta_q^{(1)}(t))^T$ be vectors of the true values of the DOA and their first derivatives at the time-instant t. We use also $q \times 1$ vectors $\mathbf{C}_0 = (c_{0,1}, ..., c_{0,q})^T$, $\mathbf{C}_1 = (c_{1,1}, ..., c_{1,q})^T$ and $q \times 2$ matrix $\mathbf{C} = (\mathbf{C}_0\ \mathbf{C}_1)$.

Multiple source nonparametric LPA estimates of $DOAs$ are defined as a solution of the optimization problem [5]:

$$\hat{\mathbf{C}}(t) = \arg\Big(\min_{\mathbf{C},\, s(t+u)} J_h(\mathbf{C}, t)\Big), \quad (2)$$

$$J(\mathbf{C}, t) = \sum_u w_h(u) ||e(t+u)||^2, \quad (3)$$

where $e(t+u) = x(t+u) - A(\mathbf{C}_0 + \mathbf{C}_1 u) s(t+u)$. The criteria function $J_h(\mathbf{C}, t)$ is a measure of the quality-of-fit of the observations $x(t+u)$ by the model $A(\mathbf{C}_0 + \mathbf{C}_1 u) s(t+u)$ in the neighborhood of the "centre" t. The window $w_h(u) = w(u/h)/h$ formalizes a localization of fitting. u denotes a shift of an observation snapshot with respect to the center

0-7803-7011-2/01/$10.00 ©2001 IEEE

t, while the scale parameter $h > 0$ determines a length of the window. It is shown that $A(\mathbf{C}_0 + \mathbf{C}_1 u)s(t+u)$ fits the output of the array $x(t+u)$ with $\mathbf{C}_0$ and $\mathbf{C}_1$ as estimates of $\Theta(t)$ and $\Theta^{(1)}(t)$. It follows from (2) that

$$\hat{s}(t+u) = [A^H A]^{-1} A^H x(t+u),\ A \triangleq A(\mathbf{C}_0 + \mathbf{C}_1 u). \tag{4}$$

Inserting $\hat{s}(t+u)$ into (2)-(3) results in

$$\hat{\mathbf{C}}(t) = \arg(\max_{\mathbf{C}} P),$$

$$P = \sum_u w_h(u) x^H(t+u) Q_A x(t+u), \tag{5}$$

$$Q_A = A[A^H A]^{-1} A^H, \tag{6}$$

where Q_A is the projection matrix onto the column space of $A \triangleq A(\mathbf{C}_0 + \mathbf{C}_1 u)$.
For $q = 1$ (5) gives the LPA beamformer power [3] as

$$P_{LPA}(C) = \frac{1}{n} \sum_u w_h(u) |a^H(c_0 + c_1 u) x(t+u)|^2. \tag{7}$$

Let us consider P as a $2D$ conditional function of the parameters of the kth source, $C \triangleq C^{[k]}$, provided that all other parameters $C^{[r]}$, $r \neq k$, are fixed. Rewrite A in (6) as a structured matrix $A = [a_k\ A_k]$, where $n \times (q-1)$ matrix A_k is A, where kth column a_k is omitted. Then $Q_k = A_k[A_k^H A_k]^{-1} A_k^H$ is a projector on the subspace spanned by the columns of A_k, $Q_k^{\perp} = I - Q_k$, I is the 2×2 identity matrix. The only term of P depending on C can be represented in the form [6]:

$$P_{MARG}^{(k)}(C,t,q) = \sum_u \frac{1}{a_k^H Q_k^{\perp} a_k} w_h(u) |a_k^H Q_k^{\perp} x(t+u)|^2. \tag{8}$$

In this notation q is a total number of sources, k indicates a desirable source to which this partial power function is designated while all other variable $C^{[r]}$, $r = 1, ..., q$, $q \neq k$, are fixed. We call $P_{MARG}^{(k)}(C,t,q)$ a marginal LPA beamformer. Thus, on the definition, the marginal beamformer is a varying part of the array output power depending on $C^{[k]}$. It works as a conventional beamformer for the kth source with a generalized sidelobe canceller nulling the signals from all other sources.
The marginal power $P_{MARG}^{(k)}(C,t,q)$ contains all information needed for optimization on $C^{[k]}$ and the $2q$ dimensional optimization problem (5) can be replaced by q two dimensional optimization problems (with motivation similar to given for unmoving sources in [6])

$$(\hat{\theta}_k(t), \hat{\theta}_k^{(1)}(t)) = \arg(\max_{\mathbf{C}} P_{MARG}^{(k)}(C,t,q)). \tag{9}$$

4. ACCURACY ANALYSIS

Let the estimation errors be defined as vectors $\Delta C^{[k]} = (\Delta\theta_k, \Delta\theta_k^{(1)})^T$, $\Delta\theta_k = \theta_k(t) - \hat{c}_{0,k}$, $\Delta\theta^{(1)} = \theta_k^{(1)}(t) - \hat{c}_{1,k}$, the sampling be periodical with a sampling period T. In the above formulas we assume that $u = kT$ and that the summation is produced over the observations into the window $w_h(kT) = \frac{1}{h} w(\frac{kT}{h})$. Let the angle $\theta_k(t)$ be continuous differentiable function of time, $\theta_k(t+u) = \theta_k(t) + \theta_k^{(1)}(t)u + \theta_k^{(2)}u^2/2 + ...$ and the waveform $s_k(t)$ be constant in a small neighborhood of the center t. Then the accuracy of the estimates (5) can be presented in the following structured form [5].

Proposition 1 *Let $h \to 0$, $T \to 0$, and $h/T \to \infty$, then:*
(1) The bias of the estimates is given as

$$D_h E\{\Delta C^{[k]}\} \simeq$$
$$-\Phi_1^{-1} \int (h^2 \theta_k^{(2)}(t) \frac{1}{2} u^2 + h^3 \theta_k^{(3)}(t) \frac{1}{6} u^3) U du,$$
$$\Phi_1 = \int w(u) U U^T dv,\ U^T = (1, u).$$

(2) The covariance matrix of the estimates is

$$D_h cov\{\Delta C^{[k]}\} D_h \simeq \frac{T}{h} \frac{L_k}{\cos^2 \theta_k(t)} \Phi_1^{-1} \Phi_2 \Phi_1^{-1}, \tag{10}$$
$$\Phi_2 = \int w^2(u) U U^T dv,\ D_h = diag\{1, h\},$$

where

$$L_k = \frac{\lambda^2}{2\eta_k(t)(2\pi l)^2} \frac{1}{SNR_k},\ SNR_k = |s_k(t)|^2/\sigma^2,$$

and

$$\eta_k(t) = a^H(\theta_k(t)) B Q_k^{\perp} B a(\theta_k(t)) - \frac{|a^H(\theta_k(t)) Q_k^{\perp} B a(\theta_k(t))|^2}{a^H(\theta_k(t)) Q_k^{\perp} a(\theta_k(t))},$$
$$B = diag(0, 1, 2, ..., n-1).$$

The proof is obtained by the technique mainly based on the Taylor series assuming that the estimation errors as well as all disturbances are small.
Comments to the proposition: (1). The one source signal case can be considered as a particular case of the derived results provided that $Q_k = I_{n\times n}$.
Then $\eta_k(t) = n(n^2-1)/12$ and

$$L_k = \frac{6\lambda^2}{n(n^2-1)(2\pi d)^2} \frac{1}{SNR}. \tag{11}$$

(2). The parameter $\eta_k(t)$ is the only term in the above formulas depending on the fact that the multiple source case is considered.
(3). Assume that the window is symmetric, $w(u) = w(-u)$. Then the formulas of Proposition 1 are simplified and the results can be presented into the following explicit form:

(a) For the bias

$$E\{\Delta\theta_k\} = -\frac{h^2}{2}\theta_k^{(2)}(t)\int w(v)v^2dv, \qquad (12)$$

(b) For the covariance

$$var\{\Delta\theta_k\} = \frac{TL_k}{h\cos^2\theta_k(t)}\int w^2(v)dv, \qquad (13)$$

$$L_k = \frac{\lambda^2}{2\eta_k(t)(2\pi d)^2}\frac{1}{SNR_k}.$$

(4). Consider the mean squared errors (MSE) of estimation using the formulas (12)-(13). It is clear that the

$$E\{(\Delta\theta_k)^2\} = \frac{T}{h\cos^2\theta_k(t)}L_k\int w^2(v)dv + \qquad (14)$$
$$+ (\frac{h^2}{2}\theta_k^{(2)}(t)\int w(v)v^2dv)^2.$$

has a minimum on h which defines the optimal window size as

$$h_{opt} = \left(\frac{TL_k\int w^2(v)dv}{(\theta_k^{(2)}(t)\cos\theta_k(t)\int w(v)v^2dv)^2}\right)^{1/5}. \qquad (15)$$

Note that the optimal window size h_{opt} depends both on the value of the angle $\theta_k(t)$ and its second derivatives $\theta_k^{(2)}(t)$. For slowly varying $\theta_k(t)$, the second derivative $\theta_k^{(2)}(t)$ becomes small and, according to (15), the optimal window length tends to increase.

5. ALGORITHMS AND RESULTS

In our study we use the Gauss-Newton recursive procedures in order to design two types of recursive tracking algorithms for the problems (5) and (9). The first, corresponding to the problem (5), produces recursive steps on all $2q$ variables simultaneously. The second implements a cyclic source-wise optimization. In the latter the Gauss-Newton recursive procedure has a deal with a $2D$ optimization problems (9) only. The gradient and approximate Hessian matrices of the algorithms are obtained in an analytical form. In particular, the gradient and the Hessian of $P_{LPA}^{(k)}(C,t,q)$ are as follows

$$\frac{\partial P_{LPA}^{(k)}(C,t,q)}{\partial C} = \frac{-j2\pi d}{\lambda}\sum_u w_h(u)\frac{\cos(c_0+c_1u)}{a^HQ_k^{\perp}a}\times$$
$$\{x^H(t+u)Q_k^{\perp}\times$$
$$[Baa^H - aa^HB + a\cdot aa^H]Q_k^{\perp}x(t+u)\}U,$$

$$\frac{\partial^2 P_{LPA}^{(k)}(C,t,q)}{\partial C\partial C^T} \approx \left(\frac{2\pi d}{\lambda}\right)^2\times$$
$$\sum_u w_h(u)\frac{\cos^2(c_0+c_1u)}{a^HQ_k^{\perp}a}\times$$
$$\{x^H(t+u)Q_k^{\perp}RR^HQ_k^{\perp}x(t+u)\}UU^T,$$
$$R = Ba + a\cdot a.$$

where $a = a(c_0+c_1u)$, $B = diag\{0,1,...,n-1\}$, $a = (a^HBQ_k^{\perp}a - a^HQ_k^{\perp}Ba)/a^HQ_k^{\perp}a$, and the vector a as a function of c_0+c_1u is defined by (1).
Comparison of the mentioned two types of the algorithms is definitely in a favor of the latter, which is much simpler in implementation and provide a similar accuracy of tracking.

6. ACKNOWLEDGMENT

This work was supported by the Brain Korea 21 Project.

7. REFERENCES

[1] Y. Bresler and Macovski A. "Exact maximum likelihood parameter estimation of superimposed exponential signals in noise", *IEEE Trans. on Acoustic, Speech, and Signal Processing*, vol. 34, N^{0}5, pp. 1081-1089, 1986.

[2] L.Frenkel and Feder M. "Recursive expectation-minimization (EM) algorithms for time-varying parameters with application to multiple target tracking", *IEEE Trans. on Signal Processing*, vol. 47, N^{0}2, pp. 306-320, 1999.

[3] V. Katkovnik and A. Gershman, "A local polynomial approximation based beamforming for source localization and tracking in nonstationary environments", *IEEE Signal Processing Letters*, vol. 7, N 1, pp. 3-5, 2000.

[4] V. Katkovnik, "A new concept of adaptive beamforming for moving sources and impulse noise environment," *Signal Processing*, vol. 80, N^0 9, pp. 1863-1882, 2000.

[5] V. Katkovnik, *Adaptive Robust Array Signal Processing for Moving Sources and Impulse Noise Environment (Nonparametric M-estimation Approach).* Tampere International Center for Signal Processing, TICSP Series, N 11, pp. 215, Tampere, TTKK, Monistamo, 2000.

[6] I. Ziskind and Wax M. "Maximum likelihood localization of multiple sources by alternating projection", , *IEEE Trans. on Acoustic, Speech, and Signal Processing*, vol. ASSP-36, N^{0}10, pp. 1553-1560, 1988.

[7] Y. Zhou, P. C. Yip, and H. Leung, "Tracking the direction-of arrival of multiple moving targets by passive arrays algorithms", *IEEE Trans. on Signal Processing*, vol. 47, N^{0}10, pp. 2635-2666, 1999.

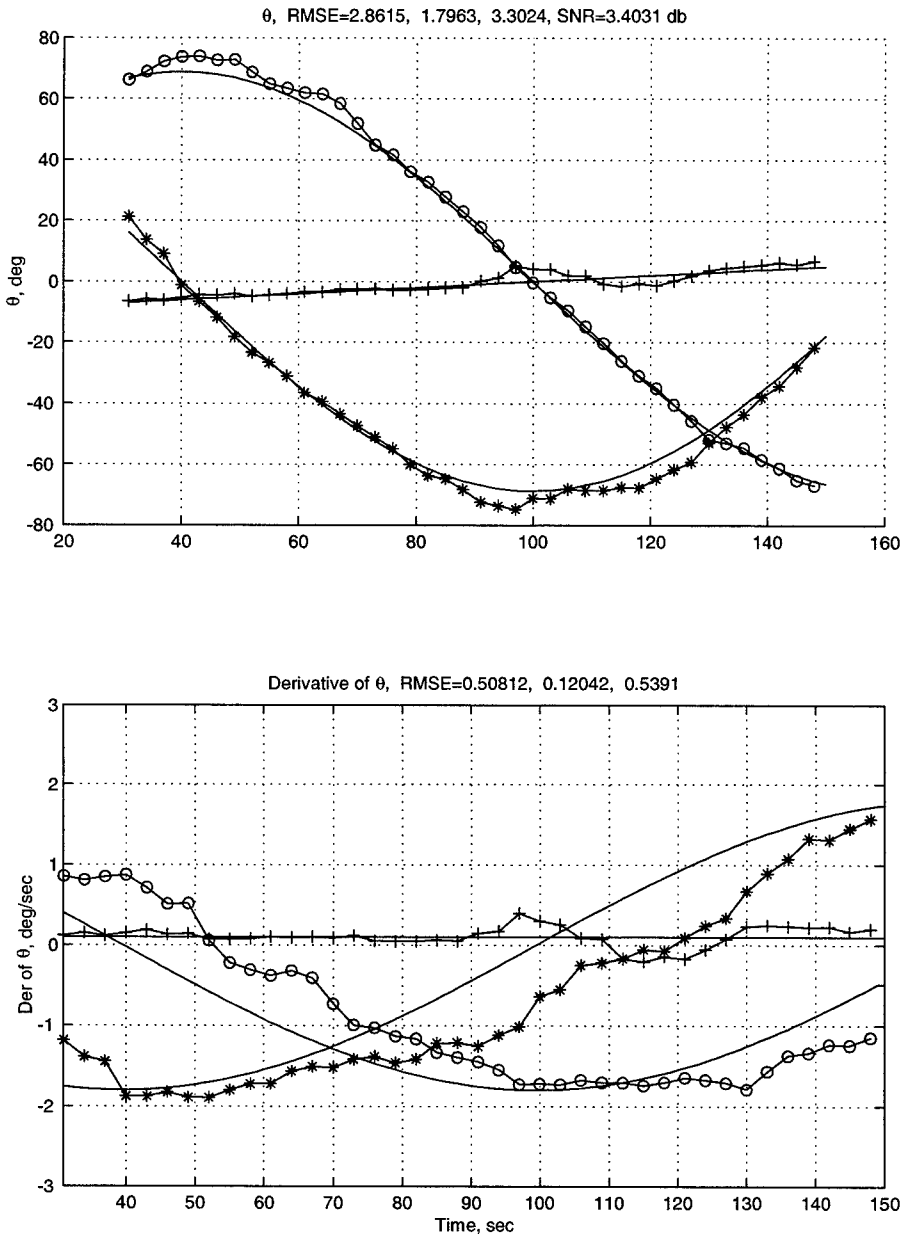

θ, RMSE=2.8615, 1.7963, 3.3024, SNR=3.4031 db
θ, deg
80
60
40
20
0
-20
-40
-60
-80
20
40
60
80
100
120
140
160
Derivative of θ, RMSE=0.50812, 0.12042, 0.5391
Der of θ, deg/sec
3
2
1
0
-1
-2
-3
40
50
60
70
80
90
100
110
120
130
140
150
Time, sec

Figure 1:

DIRECTION FINDING IN PARTLY CALIBRATED ARRAYS COMPOSED OF NONIDENTICAL SUBARRAYS: A COMPUTATIONALLY EFFICIENT ALGORITHM FOR THE RANK REDUCTION (RARE) ESTIMATOR

Marius Pesavento† *Alex B. Gershman** *Kon Max Wong** *Johann F. Böhme*†

†Signal Theory Group, Ruhr University, Bochum, D-44780, Germany
*Department of ECE, McMaster University, Hamilton, Ontario, L8S 4K1 Canada

ABSTRACT

We consider the direction finding problem in partly calibrated arrays composed of nonidentical subarrays which are displaced by unknown vector translations. A new computationally efficient algorithm is developed for the recently proposed RAnk REduction (RARE) estimator [1].

1. INTRODUCTION

Direction-Of-Arrival (DOA) estimation of narrowband sources in large sensor arrays composed of multiple subarrays has recently attracted a significant attention of specialists because using subarrays on a sparse grid extends the array aperture without a corresponding increase in hardware and software costs [2]. Also, exploiting some particular subarray structure may enable simple formulations of the DOA estimation problem. For example, a search-free polynomial rooting-based formulation of the MUSIC algorithm has been obtained in [3] for sensor arrays composed of identical (and identically oriented) subarrays displaced by arbitrary but known translations. An essential shortcoming of this approach is that the subarrays are restricted to be identical ULA's and the exact knowledge of all sensor positions is required. Obviously, such knowledge may be unavailable in large array systems where calibration of the whole array usually represents much more challenging task than calibration of each subarray and, additionally, subarray positions may change with time.

Recently, a new search-free eigenstructure-based approach to DOA estimation has been proposed [1], which overcomes the aforementioned shortcomings of [3] and other self-calibration techniques. This approach is referred to as the RAnk REduction (RARE) estimator and is applicable to partly calibrated arrays which may involve several nonidentical but identically oriented subarrays displaced by arbitrary unknown vector translations. In [1], it has been shown that RARE approaches the corresponding Cramér-Rao Bound (CRB) and enjoys simple implementation, which entails computing the eigendecomposition of the sample array covariance matrix and polynomial rooting. However, the computation of the coefficients of the RARE polynomial represents quite a complicated problem.

In this paper, we obtain a new condition for the number of subarrays which guarantees the uniqueness of the RARE DOA estimates, present the procedure to determine the degree of the RARE polynomial, and develop an efficient technique to compute its coefficients.

2. RARE ESTIMATOR

Consider an array of M omnidirectional sensors which receives $L < M$ narrowband signals impinging from the unknown DOA's $\{\theta_1, \ldots, \theta_L\}$. Let this array consist of K identically oriented linear subarrays whose interelement spacings are integer multiples of the known *shortest baseline* d. The geometry of each subarray is assumed to be known, whereas the *inter-subarray displacements* are assumed to be unknown. An example of such array (composed of four subarrays) is shown in Fig. 1. Note that unlike [2] and [3], the subarrays are allowed to be nonidentical to each other and some of them even may consist of a single sensor[1] (as the fourth subarray in Fig. 1).

Let $M_k \geq 1$ be the number of sensors of the kth subarray, so that $M = \sum_{k=1}^{K} M_k$. Note that M_k may take different values for various subarrays. For the sake of simplicity, it is convenient to define each subarray by means of a certain translation of a part of sensors of an $\tilde{M}$-element *nominal* (virtual) uniform linear array (ULA), where $\tilde{M} \geq M$. This representation is illustrated in Fig. 2 for the specific case[2] $\tilde{M} = M$ where the second, third, and fourth subarrays of Fig. 1 are interpreted as a result of three unknown vector

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

[1]However, note that a certain condition formulated below must be fulfilled for the number of subarrays.

[2]Note that in what follows, mostly the case $\tilde{M} = M$ will be considered, where each sensor of the virtual ULA becomes the part of some subarray.

0-7803-7011-2/01/$10.00 ©2001 IEEE

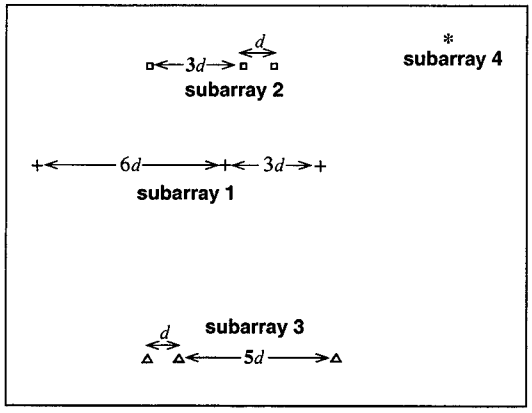


Fig. 1. A particular example of the considered type of sensor array: first subarray (+), second subarray (□), third subarray (△), fourth subarray (∗).

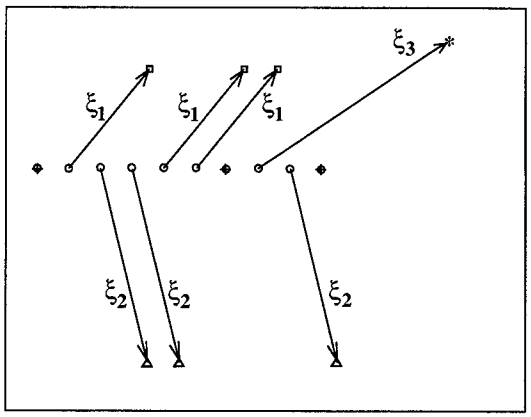


Fig. 2. Representation of the array structure of Fig. 1 by means of the virtual ULA (∘) and three vector displacements $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$.

translations $\{\boldsymbol{\xi}_k,\ k=1,2,3\}$. In the general case of K subarrays, the $K-1$ translation vectors $\{\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\ldots,\boldsymbol{\xi}_{K-1}\}$ are required to determine the array geometry (we assume without loss of generality that $\boldsymbol{\xi}_0=\mathbf{0}$). The problem is to estimate the DOA vector $\boldsymbol{\theta}=[\theta_1,\theta_2,\ldots,\theta_L]^T$ from array observations.

It can be readily shown that the narrowband model for the $M\times 1$ steering vector can be written as [1]

$$\boldsymbol{a}(\theta,\boldsymbol{\alpha})=\boldsymbol{Q}(\theta)\boldsymbol{T}\boldsymbol{h}(\theta,\boldsymbol{\alpha}) \tag{1}$$

where the $2(K-1)\times 1$ vector $\boldsymbol{\alpha}=\mathrm{vec}\{\boldsymbol{\Omega}\}$ and the $(K-1)\times 2$ matrix $\boldsymbol{\Omega}=\left[\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\ldots,\boldsymbol{\xi}_{K-1}\right]^T$. The vector $\boldsymbol{\alpha}$ combines all unknown inter-subarray displacement parameters,

$$\boldsymbol{h}(\theta,\boldsymbol{\alpha})=\left[1,e^{j(2\pi/\lambda)\boldsymbol{\xi}_1^T\boldsymbol{\phi}},\ldots,e^{j(2\pi/\lambda)\boldsymbol{\xi}_{K-1}^T\boldsymbol{\phi}}\right]^T$$

$$\boldsymbol{Q}(\theta)=\mathrm{diag}\left\{1,e^{j(2\pi/\lambda)d\sin\theta},\ldots,e^{j(M-1)(2\pi/\lambda)d\sin\theta}\right\}$$

$\boldsymbol{\phi}=[\sin\theta,\cos\theta]^T$, $\boldsymbol{\xi}_k=[\xi_{x,k},\xi_{y,k}]^T$, and λ is the wavelength. The $M\times K$ full column rank *selection matrix* $\boldsymbol{T}$ consists of zeros and ones and shows how the sensors of the nominal ULA are distributed among the subarrays, so that the (m,k)th element of $\boldsymbol{T}$ is equal to one if, after the translation by $\boldsymbol{\xi}_{k-1}$, the mth virtual ULA sensor becomes a part of the kth subarray, and equal to zero otherwise. For example, for the particular array configuration shown in Fig. 2, the selection matrix is given by

$$\boldsymbol{T}=\begin{bmatrix}1&0&0&0&0&0&1&0&0&1\\0&1&0&0&1&1&0&0&0&0\\0&0&1&1&0&0&0&0&1&0\\0&0&0&0&0&0&0&1&0&0\end{bmatrix}^T$$

According to (1), the array snapshots can be modeled as

$$\boldsymbol{x}(t)=\boldsymbol{A}(\boldsymbol{\theta},\boldsymbol{\alpha})\boldsymbol{s}(t)+\boldsymbol{n}(t),\quad t=1,2,\ldots,N \tag{2}$$

where $\boldsymbol{A}(\boldsymbol{\theta},\boldsymbol{\alpha})=[\boldsymbol{a}(\theta_1,\boldsymbol{\alpha}),\ldots,\boldsymbol{a}(\theta_L,\boldsymbol{\alpha})]$ is the $M\times L$ direction matrix which is composed of the signal direction vectors $\{\boldsymbol{a}(\theta_l,\boldsymbol{\alpha})\}_{l=1}^L$, $\boldsymbol{s}(t)$ is the $L\times 1$ vector of the signal waveforms, $\boldsymbol{n}(t)$ is the $M\times 1$ vector of white sensor noise, and N is the number of snapshots. The sample estimate of the covariance matrix

$$\boldsymbol{R}=\mathrm{E}\left\{\boldsymbol{x}(t)\boldsymbol{x}^H(t)\right\}=\boldsymbol{A}(\boldsymbol{\theta},\boldsymbol{\alpha})\boldsymbol{S}\boldsymbol{A}^H(\boldsymbol{\theta},\boldsymbol{\alpha})+\sigma^2\boldsymbol{I} \tag{3}$$

is given by

$$\hat{\boldsymbol{R}}=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}(t)\boldsymbol{x}^H(t) \tag{4}$$

where $\boldsymbol{S}=\mathrm{E}\left\{\boldsymbol{s}(t)\boldsymbol{s}^H(t)\right\}$ is the $L\times L$ source covariance matrix, $\boldsymbol{I}$ is the identity matrix, $(\cdot)^H$ is the Hermitian transpose, and $\mathrm{E}\{\cdot\}$ denotes the statistical expectation. The noise is assumed to be a zero-mean spatially white process with the identical variance σ^2 in each sensor.

Write the eigendecompositions of the matrices (3) and (4) as

$$\boldsymbol{R} = \boldsymbol{E}_S\boldsymbol{\Lambda}_S\boldsymbol{E}_S^H+\boldsymbol{E}_N\boldsymbol{\Lambda}_N\boldsymbol{E}_N^H \tag{5}$$

$$\hat{\boldsymbol{R}} = \hat{\boldsymbol{E}}_S\hat{\boldsymbol{\Lambda}}_S\hat{\boldsymbol{E}}_S^H+\hat{\boldsymbol{E}}_N\hat{\boldsymbol{\Lambda}}_N\hat{\boldsymbol{E}}_N^H \tag{6}$$

where the $L\times L$ diagonal matrices $\boldsymbol{\Lambda}_S$ and $\hat{\boldsymbol{\Lambda}}_S$ contain the L signal-subspace eigenvalues of $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$, respectively, and the $(M-L)\times(M-L)$ diagonal matrices $\boldsymbol{\Lambda}_N$ and $\hat{\boldsymbol{\Lambda}}_N$ contain the $M-L$ noise-subspace eigenvalues of $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$, respectively. In turn, the columns of the $M\times L$ matrices $\boldsymbol{E}_S$ and $\hat{\boldsymbol{E}}_S$ contain the signal-subspace eigenvectors of $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$, respectively, whereas the $M\times(M-L)$ matrices $\boldsymbol{E}_N$ and $\hat{\boldsymbol{E}}_N$ are composed of the noise-subspace eigenvectors of $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$, respectively.

Let us consider the conventional spectral MUSIC algorithm which estimates the signal DOA's from the L deepest

minima of the function

$$f_{\mathrm{MUSIC}}(\theta) = \boldsymbol{a}^H(\theta,\boldsymbol{\alpha})\hat{\boldsymbol{E}}_N\hat{\boldsymbol{E}}_N^H\boldsymbol{a}(\theta,\boldsymbol{\alpha}) \quad (7)$$

In the ideal case of exactly known $\boldsymbol{R}$, the DOA's can be found from the equation

$$\boldsymbol{a}^H(\theta,\boldsymbol{\alpha})\boldsymbol{E}_N\boldsymbol{E}_N^H\boldsymbol{a}(\theta,\boldsymbol{\alpha}) = 0 \quad (8)$$

Since the vector parameter $\boldsymbol{\alpha}$ is unknown, the minimization of (7) requires an exhaustive $(2(K-1)+1)$-dimensional search which becomes totally impractical for $K > 1$. Inserting (1), we can rewrite (8) as

$$\boldsymbol{h}^H(\theta,\boldsymbol{\alpha})\boldsymbol{B}(z)\boldsymbol{h}(\theta,\boldsymbol{\alpha}) = 0 \quad (9)$$

where

$$\boldsymbol{B}(z) \triangleq \boldsymbol{T}^T\boldsymbol{Q}(1/z)\hat{\boldsymbol{E}}_N\hat{\boldsymbol{E}}_N^H\boldsymbol{Q}(z)\boldsymbol{T} \quad (10)$$

is the $K \times K$ Hermitian matrix, $\boldsymbol{Q}$ is reformulated in terms of $z = e^{j(2\pi/\lambda)d\sin\theta}$ as $\boldsymbol{Q}(z) = \mathrm{diag}\{1, z, \ldots, z^{M-1}\}$ and the obvious property $\boldsymbol{Q}^H(z) = \boldsymbol{Q}(1/z)$ is used. A very important observation here is that the vector parameter $\boldsymbol{\alpha}$ is contained in $\boldsymbol{h}(\theta,\boldsymbol{\alpha})$ only, so that the matrix $\boldsymbol{B}(z)$ is independent of $\boldsymbol{\alpha}$.

It is worth noting that $\boldsymbol{B}(z)$ in the general case is full rank. Note that (9) may hold true only if the matrix $\boldsymbol{B}(z)$ drops rank, so that $\mathrm{rank}\{\boldsymbol{B}(z)\} < K$ or, equivalently, when the polynomial

$$P(z) = \det\{\boldsymbol{B}(z)\} = 0 \quad (11)$$

An important question now is whether $\boldsymbol{B}(z)$ may drop rank for some values of z which lie on the unit circle but do not nullify the MUSIC polynomial $f_{\mathrm{MUSIC}}(z)$. The following theorem answers this question.

Theorem: Let

$$\boldsymbol{v}(z) = [1, z, \ldots, z^{M-1}]^T \quad (12)$$

be the $M \times 1$ virtual ULA steering vector written in terms of z and

$$\boldsymbol{d}_k(z) = [d_{k,1}(z), d_{k,2}(z), \ldots, d_{k,M_k}(z)]^T \quad (13)$$

be the $M_k \times 1$ vector composed of $\boldsymbol{v}(z)$ by taking into account only the sensors of the kth subarray. Let

$$\breve{\boldsymbol{d}}_k(z) = \boldsymbol{d}_k(z)/d_{k,1}(z) \quad (14)$$

be the steering vector of the kth subarray with the phase origin in the first sensor of this subarray, and

$$\beta_k = \begin{cases} 1, & \breve{\boldsymbol{d}}_k(z) = \breve{\boldsymbol{d}}_k(z') \text{ if and only if } \theta = \theta' \\ 0, & \text{otherwise} \end{cases}$$

be the parameter determining whether the kth subarray manifold is unambiguous. Then, provided that the following condition is satisfied

$$L < \sum_{k=1}^{K} \beta_k(M_k - 1) \quad (15)$$

equation (9) holds true if and only if $P(z)\Big|_{|z|=1} = 0$. □

An important corollary following from this theorem is that *the signal DOA's can be found by rooting the polynomial $P(z)$ without exploiting any knowledge of the inter-subarray displacement parameters $\boldsymbol{\alpha}$.* From (15) we obtain that $L \le \sum_{k=1}^{K}\beta_k(M_k-1) \le \sum_{k=1}^{K}(M_k-1) = M-K$. Therefore, condition (15) can be interpreted as a strengthened version of the necessary condition $K \le M - L$ discussed in [1]. Furthermore, (15) simplifies to the latter inequality $K \le M - L$ in the case when all subarrays an unambiguous, i.e. $\beta_k = 1,\ k = 1, 2, \ldots, K$.

3. IMPLEMENTATION

To apply RARE to the finite observation case (where only the sample covariance matrix (4) is available), we should root the sample polynomial

$$\hat{P}(z) = \det\{\hat{\boldsymbol{B}}(z)\} \quad (16)$$

where

$$\hat{\boldsymbol{B}}(z) = \boldsymbol{T}^T\boldsymbol{Q}^T(1/z)\hat{\boldsymbol{E}}_N\hat{\boldsymbol{E}}_N^H\boldsymbol{Q}(z)\boldsymbol{T} \quad (17)$$

and then find the signal DOA's from the L roots of (16) which are closest to the unit circle (in the similar way as in root-MUSIC). Note that, similar to root-MUSIC, the RARE roots enjoy *conjugate reciprocity* property, i.e. if $\hat{z}$ is the root of $\hat{P}(z)$ then $1/\hat{z}^*$ is also the root of $\hat{P}(z)$. Therefore, to obtain the signal DOA's, it is sufficient to examine the roots of $\hat{P}(z)$ inside the unit circle.

Another interesting observation that $\hat{P}(z)$ reduces to the root-MUSIC polynomial in the case $K = 1$ [1].

Two important questions arise when implementing the RARE estimator. First of all, it is important to determine the degree of the RARE polynomial. Second, a low-complexity algorithm to compute the coefficients of this polynomial is required. These issues are addressed below.

3.1. Degree of the RARE Polynomial

It can be readily shown that

$$D_{\mathrm{RARE}} = \sum_{k=1}^{K} D_{k,k} \quad (18)$$

where D_{RARE} is the degree of $\hat{P}(z)$,

$$D_{i,k} \triangleq \text{degree}\left\{\boldsymbol{d}_i^T(1/z)\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_k^H\boldsymbol{d}_k(z)\right\} \qquad (19)$$

and the matrix $\hat{\boldsymbol{U}}_l$ is composed of $\hat{\boldsymbol{E}}_N$ by taking only the rows of the latter matrix which correspond to the lth subarray. From (18)-(19), it follows that each particular degree $D_{k,k}$ and, therefore, the total degree D_{RARE} essentially depend on how the subarrays have been chosen.

3.2. Computing the RARE Polynomial Coefficients

Let us consider $b_{k,l}$, the (k,l)th element of the polynomial matrix $\hat{\boldsymbol{B}}(z)$. It can be written as the polynomial

$$b_{k,l}(z) = \boldsymbol{d}_k^T(1/z)\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_l^H\boldsymbol{d}_l(z) \qquad (20)$$

Let $p_{k,l}(n)$ denote the polynomial coefficient of $b_{k,l}(z)$ that corresponds to z^n. It can be readily verified that $p_{k,l}(n)$ is zero if z^n is not representable as a product of any two elements of vectors $\boldsymbol{d}_k(1/z)$ and $\boldsymbol{d}_l(z)$. However, if z^n can be represented as such product, i.e.

$$z^n = d_{k,p}(1/z)d_{l,m}(z) \qquad (21)$$

for some $p \in \{1,2\ldots,M_k\}$ and $m \in \{1,2\ldots,M_l\}$ then the coefficient $p_{k,l}(n)$ is nonzero and can be computed as a sum of (p,m)th elements of the matrix $\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_l^H$ over all pairs $\{p,m\}$ of indices which satisfy (21). Making use of the well-known recursive formula for computing determinants, we can write

$$\hat{P}(z) = \sum_{k=1}^{K}(-1)^{k+1}b_{k,1}(z)\det\left\{\hat{\boldsymbol{B}}_{k,1}(z)\right\} \qquad (22)$$

where $\boldsymbol{B}_{k,m}(z)$ is the $(K-1)\times(K-1)$ matrix obtained from $\hat{\boldsymbol{B}}(z)$ by deleting its kth row and mth column.

Let $p_{\hat{\boldsymbol{B}}}(n)$ denote the coefficient of the polynomial $\hat{P}(z)$ which corresponds to z^n. Similarly, let $p_{\hat{\boldsymbol{B}}_{k,m}}(n)$ denote the polynomial coefficient of $\det\{\boldsymbol{B}_{k,m}(z)\}$ which corresponds to z^n. Note that the sequences of polynomial coefficients and the polynomials themselves form the following z-transform pairs: $b_{k,m}(z) = \mathcal{Z}\{p_{k,m}(n)\}$, $\hat{P}(z) = \mathcal{Z}\{p_{\hat{\boldsymbol{B}}}(n)\}$, and $\det\{\boldsymbol{B}_{k,m}(z)\} = \mathcal{Z}\{p_{\hat{\boldsymbol{B}}_{k,m}}(n)\}$. Using (22) and the convolution property of z-transform, we obtain

$$p_{\hat{\boldsymbol{B}}}(n) = \sum_{k=1}^{K}(-1)^{k+1}p_{k,1}(n) * p_{\hat{\boldsymbol{B}}_{k,1}}(n) \qquad (23)$$

Equation (23) is the core formula for computing the RARE polynomial coefficients. It relates the coefficients of the polynomial $\det\{\hat{\boldsymbol{B}}(z)\}$ with the coefficients of the polynomials $\det\{\hat{\boldsymbol{B}}_{k,1}\}$. Obviously, (23) can be applied recursively to compute the coefficients of each polynomial $\det\{\hat{\boldsymbol{B}}_{k,1}(z)\}$ via the coefficients of the polynomials formed by the determinants of submatrices of $\hat{\boldsymbol{B}}_{k,1}$ and so on.

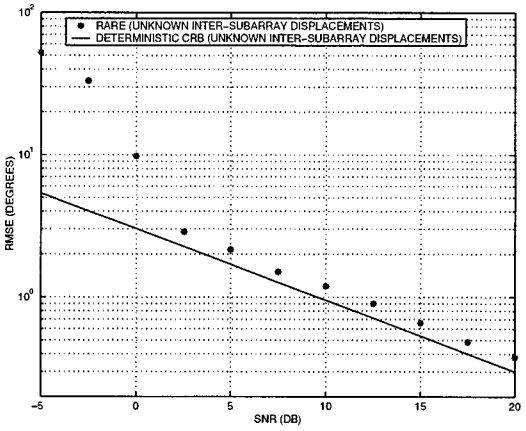


Fig. 3. The RMSE of RARE and the CRB [1] versus the SNR.

4. SIMULATIONS

We assume an array of $M = 6$ sensors which is composed of $K = 3$ two-element subarrays ($M_1 = M_2 = M_3 = 2$). The interelement spacings of the first, second, and third subarrays are $d = \lambda/6$, $2d = \lambda/3$, and $3d = \lambda/2$ respectively. The inter-subarray displacements are unknown and equal to $\boldsymbol{\xi}_1 = [7.56d, 25.43d]$ and $\boldsymbol{\xi}_2 = [0.93d, -12.27d]$, respectively.

Two uncorrelated equi-powered sources are assumed to impinge on the array from $\theta_1 = 5°$ and $\theta_2 = 11°$. All results are averaged over 100 simulation runs and $N = 100$ is taken. In Fig. 3, the RMSE of RARE is displayed versus the SNR along with the deterministic CRB derived in [1] for the case of unknown inter-subarray displacements. Our simulations show that the performance of RARE is very close to the corresponding CRB. It is worth noting that RARE appears to be the only method applicable in this scenario.

5. REFERENCES

[1] M. Pesavento, A.B. Gershman, and K.M. Wong, "Direction of arrival estimation in partly calibrated time-varying sensor arrays," in *Proc. ICASSP'01*, Salt Lake City, UT, May 2001.

[2] M.D. Zoltowski and K.T. Wong, "Closed-form eigenstructure-based direction finding using arbitrary but identical subarrays on a sparse uniform Cartesian array grid," *IEEE Trans. Signal Processing*, vol. 48, pp. 2205-2210, Aug. 2000.

[3] A.L. Swindlehurst, P. Stoica, and M. Jansson, "Application of MUSIC to arrays with multiple invariances," in *Proc. ICASSP'00*, Istanbul, Turkey, pp. 3057-3060, June 2000.

RECURSIVE EM AND SAGE ALGORITHMS

Pei Jung Chung and Johann F. Böhme

Dept. of Electrical Engineering and Information Science
Ruhr-Universität Bochum, 44780 Bochum, Germany
e-mail: {pjc,boehme}@sth.ruhr-uni-bochum.de

ABSTRACT

This work is concerned with recursive procedures in which the data run through sequentially. Two stochastic approximation recursions derived from EM and SAGE algorithms are proposed. We show that under regularity conditions, these recursions lead to strong consistency and asymptotic normality. Although the recursive EM and SAGE algorithm do not have the optimal convergence rate, they are usually easy to implement. As an example, we derive recursive procedures for direction of arrival (DOA) estimation. In numerical experiments both algorithms provide good results for low computational cost.

1. INTRODUCTION

This work is concerned with recursive parameter estimation using augmented data. The EM algorithm [1] is a well known numerical procedure to locate modes of a likelihood function which is characterized by its simple implementation and stability. One of its variants, the SAGE algorithm [3] generalizes the idea of data augmentation to facilitate more flexible choices of parameter sets and faster convergence rates in some settings.

Both algorithms provide iterative estimates based on the same batch of data. If the data sets are large, these procedures could become expensive. To solve this problem we propose two stochastic approximation recursions derived from EM and SAGE algorithms in which the data arrive sequentially.

The first recursive EM algorithm was suggested by Titterington [6] where the step size is limited to be $a_n = n^{-\alpha}$. The consistency and asymptotic normality was only presented for the univariate version. Here we will consider a more general case where $a_n = an^{-\alpha}$ and generalize the asymptotic properties to the multivariate case.

Furthermore, we propose a new recursion derived from the SAGE algorithm in which a more flexible choice of parameter sets is allowed. Under regularity conditions the sequence of estimates generated by the recursive SAGE algorithm enjoys strong consistency and asymptotic normality as well.

Compared to the stochastic approximation procedure with optimal convergence rate where the inversion of the observed information matrix is necessary [2], the inverse of the augmented information matrices used in recursive EM and SAGE algorithms are usually much easier to compute.

As an illustrative example, we apply the proposed algorithms to direction of arrival (DOA) estimation. Because of the diagonal structure of the augmented information matrices, the recursive procedures have very simple implementations. In numerical experiments, we consider a critical scenario in which two sources are closely located. Results show that both algorithms converge to the true parameters and the mean squared errors decrease with time.

This paper is organized as follows. The recursive EM and SAGE algorithms are developed in section 2 and 3. Consistency and asymptotic normality of both algorithms are proved in section 4. In section 5, we derive recursive procedures for DOA estimation. Numerical results are presented in section 6.

This work has been supported by German Science Foundation.

2. RECURSIVE EM ALGORITHM

Suppose $\underline{x}_1, \underline{x}_2, \ldots$ are independent observations, each with underlying probability density function (p.d.f.) $f_{\underline{X}}(\underline{x}, \underline{\theta})$, where $\underline{\theta} \in \mathbb{R}^M$ denotes an unknown parameter vector. $\underline{Y}$ is the augmented data used by the EM algorithm with p.d.f. $f_{\underline{Y}}(\underline{y}, \underline{\theta})$. Let $\underline{\theta}^n$ denote the estimate after n observations altogether. The following recursion is aimed at finding the maximizing parameter $\underline{\theta} = \underline{\theta}^*$ of $\log f_{\underline{X}}(\underline{x}, \underline{\theta})$:

$$\underline{\theta}^{n+1} = \underline{\theta}^n - an^{-\alpha}\, \mathcal{I}_{\mathrm{EM}}(\underline{\theta}^n)^{-1}\, \underline{\gamma}(\underline{x}_n, \underline{\theta}^n), \tag{1}$$

where $a > 0$ is a constant and

$$\underline{\gamma}(\underline{x}_n, \underline{\theta}^n) = -\underline{\nabla}_{\underline{\theta}} \log f_{\underline{X}}(\underline{x}_n, \underline{\theta})|_{\underline{\theta}=\underline{\theta}^n}, \tag{2}$$

$$\mathcal{I}_{\mathrm{EM}}(\underline{\theta}^n) = \mathsf{E}\left[-\underline{\nabla}_{\underline{\theta}}\underline{\nabla}_{\underline{\theta}}^T \log f_{\underline{Y}}(\underline{y}, \underline{\theta})|\, \underline{x}_n, \underline{\theta}\right]|_{\underline{\theta}=\underline{\theta}^n} \tag{3}$$

represent the gradient vector of log–likelihood of the observation at time n and the Fisher information matrix corresponding to the augmented data $\underline{Y}$, respectively. $\underline{\nabla}_{\underline{\theta}}$ is a column gradient operator with respect to $\underline{\theta}$. The choice of α depends on the matrix

$$\mathbf{D}_{\mathrm{EM}}(\underline{\theta}) = \frac{1}{2}\mathbf{I} - a\mathcal{I}_{\mathrm{EM}}(\underline{\theta})^{-1}\mathcal{I}(\underline{\theta}) \tag{4}$$

where $\mathcal{I}(\underline{\theta})$ and $\mathbf{I}$ denote the Fisher information matrix corresponding to one observation and the identity matrix, respectively. Use $\alpha = 1$ if $\mathbf{D}_{\mathrm{EM}}(\underline{\theta})$ is a stable matrix and otherwise $1/2 < \alpha < 1$. A matrix is called stable if all eigenvalues have negative real parts [5].

As pointed out by Titterington [6], there is a strong relationship between recursion (1) and the EM algorithm. The consistency and asymptotic normality for the univariate version of (1) with $a = 1$ was also presented in [6], [7]. These results will be generalized later in section 4.

3. RECURSIVE SAGE ALGORITHM

The space alternating generalized EM (SAGE) algorithm [3] generalizes the idea of data augmentation to simplify computations of the EM algorithm. It preserves the stability of EM and can improve the convergence rate significantly in some settings. Instead of estimating all parameters at once, each iteration of SAGE consists of C cycles. The parameter subset associated with the c–th cycle $\underline{\theta}_c$ is updated by maximizing the conditional expectation of the log–likelihood of the augmented data $\underline{Z}_c$ corresponding to this cycle.

To avoid mathematical difficulties, we will only consider the case $\underline{\theta} = (\underline{\theta}_1, \ldots, \underline{\theta}_C)$ where the parameter subsets are disjoint. Let $\underline{Z}_c$ denote the augmented data of the c–th cycle with p.d.f. $f_{\underline{Z}_c}(\underline{z}_c, \underline{\theta}_c)$. The recursive version of SAGE is based on the following procedure.

At time instant $n+1$ with the current estimate $\underline{\theta}^n$, define

$$L_{n+1}(\underline{\theta}) = \sum_{c=1}^{C} \mathsf{E}\left[\log f_{\underline{Z}_c}(\underline{\theta}_c) | \, \underline{x}_n, \underline{\theta}^n\right] + L_n(\underline{\theta}). \qquad (5)$$

Choose $\underline{\theta}^{n+1}$ by maximizing $L_{n+1}(\underline{\theta})$.

Applying Taylor expansion to the right hand side of (5), the recursion can be approximated by

$$\underline{\theta}^{n+1} = \underline{\theta}^n - (n+1)^{-1} \, \mathcal{I}_{\mathrm{SAGE}}(\underline{\theta}^n)^{-1} \, \underline{\gamma}(\underline{x}_n, \underline{\theta}^n), \qquad (6)$$

where $\mathcal{I}_{\mathrm{SAGE}}$ is a block diagonal matrix with the c–th block

$$\mathcal{I}^{[c]}_{\mathrm{SAGE}}(\underline{\theta}^n) = \mathsf{E}\left[-\nabla_{\underline{\theta}_c} \nabla^T_{\underline{\theta}_c} \, \log f_{\underline{Z}_c}(\underline{z}_c, \underline{\theta}_c) | \, \underline{x}_n, \underline{\theta}\right] |_{\underline{\theta}=\underline{\theta}^n}. \qquad (7)$$

Without losing the asymptotic properties of $\underline{\theta}^n$, (6) can be used with a more flexible step size, i.e.

$$\underline{\theta}^{n+1} = \underline{\theta}^n - a n^{-\alpha} \, \mathcal{I}_{\mathrm{SAGE}}(\underline{\theta}^n)^{-1} \, \underline{\gamma}(\underline{x}_n, \underline{\theta}^n). \qquad (8)$$

Use $\alpha = 1$ if

$$\mathbf{D}_{\mathrm{SAGE}}(\underline{\theta}) = \frac{1}{2}\mathbf{I} - a\mathcal{I}_{\mathrm{SAGE}}(\underline{\theta})^{-1}\mathcal{I}(\underline{\theta}). \qquad (9)$$

is a stable matrix and otherwise $1/2 < \alpha < 1$.

4. CONVERGENCE AND ASYMPTOTIC DISTRIBUTION

In this section we study the asymptotic behaviors of $\{\underline{\theta}^n\}$ generated by the recursive EM algorithm (1) and the recursive SAGE algorithm (8), respectively. Based on convergence results from stochastic approximation [4][5], it will be shown that $\underline{\theta}_n$ converges with probability one to $\underline{\theta}^*$ and is asymptotically normal distributed.

To begin with, define

$$\underline{g}(\underline{\theta}) = \mathsf{E}\left[-\nabla_{\underline{\theta}} \log f_{\underline{X}}(\underline{x}, \underline{\theta})\right] = \nabla_{\underline{\theta}} J(\underline{\theta}, \underline{\theta}^*) \qquad (10)$$

where

$$J(\underline{\theta}, \underline{\theta}^*) = \int \log\left[f_{\underline{X}}(\underline{x}, \underline{\theta}^*)/f_{\underline{X}}(\underline{x}, \underline{\theta})\right] f_{\underline{X}}(\underline{x}, \underline{\theta}^*) \mathrm{d}\underline{x} \qquad (11)$$

is the Kullback–Leibler divergence between $f_{\underline{X}}(\underline{x}, \underline{\theta}^*)$ and $f_{\underline{X}}(\underline{x}, \underline{\theta})$. It is well known that under regularity conditions $J(\underline{\theta}, \underline{\theta}^*) \geq 0$ and with equality if and only if $\underline{\theta} = \underline{\theta}^*$. Therefore $\underline{g}(\underline{\theta}^*) = \underline{0}$, $\mathcal{I}_{\mathrm{EM}}(\underline{\theta}^*)^{-1}\underline{g}(\underline{\theta}^*) = \underline{0}$ and $\mathcal{I}_{\mathrm{SAGE}}(\underline{\theta}^*)^{-1}\underline{g}(\underline{\theta}^*) = \underline{0}$. Clearly, (1) and (8) are recursive procedures to find the roots of $\underline{U}_1(\underline{\theta}) = \mathcal{I}_{\mathrm{EM}}(\underline{\theta})^{-1}\underline{g}(\underline{\theta})$ and $\underline{U}_2(\underline{\theta}) = \mathcal{I}_{\mathrm{SAGE}}(\underline{\theta})^{-1}\underline{g}(\underline{\theta})$, respectively.

In the following, we will consider the strong consistency and asymptotic normality of $\underline{\theta}^n$ generated by (1). These properties hold also for $\underline{\theta}^n$ generated by (8). Making use of results from stochastic approximation [4], we obtain the following.

Theorem 1 *Suppose (a)* $\mathsf{E}\left[\underline{\gamma}(\underline{x}_n, \underline{\theta}^n)\,\underline{\gamma}(\underline{x}_n, \underline{\theta}^n)^T\right] < \infty$ *and (b)* $\mathcal{I}_{\mathrm{EM}}(\underline{\theta}) > 0$. *hold for recursion (1). Then* $\underline{\theta}^n$ *converges with probability one to* $\underline{\theta}^*$.

Proof

(1) $\sum_{n=1}^{\infty} an^{-\alpha} = \infty$, $an^{-\alpha} > 0$, $an^{-\alpha} \to 0, \forall n \geq 0$.

(2) Under the assumption $\mathsf{E}\left[\underline{\gamma}(\underline{x}_n, \underline{\theta}^n)\underline{\gamma}(\underline{x}_n, \underline{\theta}^n)^T\right] < \infty$, we have $\mathsf{E}\left[\|\mathcal{I}_{\mathrm{EM}}(\underline{\theta}^n)^{-1}\underline{\gamma}(\underline{x}_n, \underline{\theta}^n)\|^2\right] < \infty$, $\forall n > 0$

(3) Since $\underline{x}_1, \underline{x}_2, \ldots$ are mutually independent,

$$\mathcal{I}_{\mathrm{EM}}(\underline{\theta}^n)^{-1}\underline{\gamma}(\underline{x}_n, \underline{\theta}^n) = \mathcal{I}_{\mathrm{EM}}(\underline{\theta}^n)^{-1}\underline{g}(\underline{\theta}^n) + \delta M_n \qquad (12)$$

where δM_n is a martingale difference.

With (1),(2),(3) and **Theorem 2.1** in [4], $\underline{\theta}^n$ converges to $\underline{\theta}^*$ with probability one. □

Remark: The result of **Theorem 1** holds for the recursive SAGE algorithm if $\mathcal{I}_{\mathrm{EM}}(\underline{\theta})$ in (b) is replaced by $\mathcal{I}_{\mathrm{SAGE}}(\underline{\theta})$.

The next theorem is concerned with the asymptotic properties of normalized errors about the limit point $\underline{\theta}^*$. Let $\mathbf{U}_i'(\underline{\theta})$, $(i = 1, 2)$ denote the Jacobi matrix of $\underline{U}_i(\underline{\theta})$ and $\mathcal{N}(.,.)$ the normal distribution.

Theorem 2 *Suppose (a)* $\mathsf{E}\left[\underline{\gamma}(\underline{x}_n, \underline{\theta}^n)\,\underline{\gamma}(\underline{x}_n, \underline{\theta}^n)^T\right] < \infty$ *and (b) finite* $\mathbf{U}_1'(\underline{\theta})$ *hold for (1). Then (i) if* $\alpha = 1$ *and* $\mathbf{D}_{\mathrm{EM}}$ *is a stable matrix,* $n^{1/2}(\underline{\theta}^n - \underline{\theta}^*)$ *has asymptotic distribution* $\mathcal{N}(\underline{0}, \mathbf{V})$ *where* $\mathbf{V}$ *is the solution of*

$$(a\mathbf{A} - \frac{1}{2}\mathbf{I})\mathbf{V} + \mathbf{V}(a\mathbf{A} - \frac{1}{2}\mathbf{I})^T = a^2\mathbf{C} \qquad (13)$$

with $\mathbf{A} = \mathcal{I}_{\mathrm{EM}}(\underline{\theta}^*)^{-1}\mathcal{I}(\underline{\theta}^*)$, $\mathbf{C} = \mathcal{I}_{\mathrm{EM}}(\underline{\theta}^*)^{-1}\mathcal{I}(\underline{\theta}^*)\mathcal{I}_{\mathrm{EM}}(\underline{\theta}^*)^{-1}$, *(ii) if* $1/2 < \alpha < 1$, $n^{\alpha/2}(\underline{\theta}^n - \underline{\theta}^*)$ *has asymptotic distribution* $\mathcal{N}(\underline{0}, \bar{\mathbf{V}})$ *where* $\bar{\mathbf{V}}$ *is the solution of*

$$\mathbf{A}\bar{\mathbf{V}} + \bar{\mathbf{V}}\mathbf{A} = a\mathbf{C}. \qquad (14)$$

Proof By the mean value theorem and the finite matrix $\mathbf{U}_1'(\underline{\theta})$, one can show that for $\underline{\theta}$ near $\underline{\theta}^*$

(1) $k_1\|\underline{\theta} - \underline{\theta}^*\|^2 \leq (\underline{\theta} - \underline{\theta}^*)^T\underline{U}_1(\theta) \leq k_3\|\underline{\theta} - \underline{\theta}^*\|^2$

(2) $\|\underline{U}_1(\underline{\theta})\| \leq k_2\|\underline{\theta} - \underline{\theta}^*\| + k_4$ where $k_j, j = 1, \ldots, 4$ are constants.

(3) $\mathsf{E}\left[\|\delta M_n\|^2\right] < \infty$ under assumption (a).

(4) By Taylor expansion around $\underline{\theta}^*$, it can be shown that $\underline{U}_1(\underline{\theta}) = \mathbf{A}(\underline{\theta} - \underline{\theta}^*) + o(\|\underline{\theta} - \underline{\theta}^*\|)$.

From (1),(2),(3),(4) and **Theorems 5.8, 5.9** of [5], we conclude the asymptotic normality of $\underline{\theta}^n$. □

Remark: The results of **Theorem 2** hold for the recursive SAGE algorithm if $\mathbf{U}_1'(\underline{\theta})$ in (b) is replaced by $\mathbf{U}_2'(\underline{\theta})$ and $\mathbf{D}_{\mathrm{EM}}$ is replaced by $\mathbf{D}_{\mathrm{SAGE}}$ and $\mathbf{A} = \mathcal{I}_{\mathrm{SAGE}}(\underline{\theta}^*)^{-1}\mathcal{I}(\underline{\theta}^*)$.

As noted in [4], the asymptotic variances $\mathbf{V}$ or $\bar{\mathbf{V}}$ are a measure of the rate of convergence. The optimal covariance is achieved when $\mathcal{I}_{\mathrm{EM}}(\underline{\theta}^n)$ in (1) is replaced by $\mathcal{I}(\underline{\theta}^n)$ [2]. However, the inverse of the augmented information matrices $\mathcal{I}_{\mathrm{EM}}(\underline{\theta}^n)$ and $\mathcal{I}_{\mathrm{SAGE}}(\underline{\theta}^n)$ are in general much easier to compute than $\mathcal{I}(\underline{\theta}^n)$. By proper choice of the augmentation scheme, the recursive procedures (1) and (8) can be implemented easily.

5. APPLICATION TO DOA ESTIMATION

In previous sections we were only concerned with the theoretical aspects of the recursive EM and SAGE. To show their potential in solving practical problems, we will derive recursive procedures for DOA estimation under a deterministic signal model and independent Gaussian noise. In this case, the array output $\underline{X}(n), n = 1, 2, \ldots$ are independent but not identically distributed. This difficulty can be overcome by considering the concentrated log–likelihood function. The purpose of the recursive procedures is to find the maxima of the concentrated log–likelihood function. The information matrices are used as scaling factors to improve the convergence rate.

Consider an array of N sensors receiving signals generated by M far field narrowband sources. For signals arriving from $\underline{\theta} = (\theta_1, \ldots, \theta_M)$ the array output $\underline{X}(n)$ at the time n can be described as

$$\underline{X}(n) = \mathbf{H}(\underline{\theta})\, \underline{s}(n) + \underline{U}(n), \tag{15}$$

where $\mathbf{H}(\underline{\theta}) = [\underline{d}(\theta_1), \ldots, \underline{d}(\theta_M)] \in \mathbb{C}^{N\times M}$ contains M steering vectors $\underline{d}(\theta_m) \in \mathbb{C}^{N\times 1} (m = 1, \ldots, M)$, $\underline{s}(n) = [s_1(n), \ldots, s_M(n)]^T \in \mathbb{C}^{M\times 1}$, $\underline{U}(n) \in \mathbb{C}^{N\times 1}$ denote signal waveforms, noise vector, respectively. $\underline{s}(n)$ is considered as deterministic, unknown. $\underline{U}(n)$ is independent $\mathcal{N}(\underline{0}, \nu\mathbf{I})$ distributed with ν being known. The problem is to estimate $\underline{\theta}$ from the observations $\underline{X}(1), \underline{X}(2), \ldots$. Let $\mathbf{P}(\underline{\theta}) = \mathbf{H}(\underline{\theta})(\mathbf{H}(\underline{\theta})^H\mathbf{H}(\underline{\theta}))^{-1}\mathbf{H}(\underline{\theta})^H$. Omitting the constant terms, the concentrated log–likelihood of $\underline{X}(n)$ is given by

$$L(\underline{\theta}) = \frac{1}{\nu}\underline{X}(n)^H\mathbf{P}(\underline{\theta})\underline{X}(n). \tag{16}$$

Taking first derivative of $L(\underline{\theta})$, the m–th element of the gradient vector $\underline{g}(\theta)$ can be written approximately as

$$g_m(\underline{\theta}) = -\frac{2}{\nu}\Re\left[\left(\underline{d}'(\theta_m)\hat{s}_m(n)\right)^H(\underline{X}(n) - \mathbf{H}(\underline{\theta})\hat{\underline{s}}(n))\right] \tag{17}$$

where $\hat{\underline{s}}(n) = (\mathbf{H}(\underline{\theta})^H\mathbf{H}(\underline{\theta}))^{-1}\mathbf{H}(\underline{\theta})^H\underline{X}(n)$, $\hat{s}_m(n)$ is the m–th element of $\hat{\underline{s}}(n)$ and $\underline{d}'(\theta_m) = \frac{\partial}{\partial\theta_m}\underline{d}(\theta_m)$.

The EM algorithm has the augmented data

$$\underline{Y}(n) = [\underline{Y}_1(n)^T \ldots \underline{Y}_m(n)^T \ldots \underline{Y}_M(n)^T]^T \tag{18}$$

where $\underline{Y}_m(n) = \underline{d}(\theta_m)s_m(n) + \underline{U}_m(n)$. $\underline{U}_m(n)$ is $\mathcal{N}(\underline{0}, \frac{\nu}{M}\mathbf{I})$ distributed. Let $\underline{d}''(\theta_m) = \frac{\partial^2}{\partial\theta_m^2}\underline{d}(\theta_m)$. The information matrix $\mathcal{I}_{\mathrm{EM}}(\underline{\theta})$ is a diagonal matrix with the m–th diagonal element

$$[\mathcal{I}_{\mathrm{EM}}(\underline{\theta})]_{mm} = \frac{2}{\nu}\Re\Big[(-\underline{d}''(\theta_m)\hat{s}_m(n))^H(\underline{X}(n) - \mathbf{H}(\underline{\theta})\hat{\underline{s}}(n)) + M\|\underline{d}'(\theta_m)\hat{s}_m(n)\|^2\Big] \tag{19}$$

The augmented data used by SAGE is given by

$$\underline{Z}_m(n) = \underline{d}(\theta_m)s_m(n) + \underline{U}(n). \tag{20}$$

The corresponding $\mathcal{I}_{\mathrm{SAGE}}(\underline{\theta})$ is a diagonal matrix with the m–th diagonal element

$$[\mathcal{I}_{\mathrm{SAGE}}(\underline{\theta})]_{mm} = \frac{2}{\nu}\Re\Big[(-\underline{d}''(\theta_m)\hat{s}_m(n))^H(\underline{X}(n) - \mathbf{H}(\underline{\theta})\hat{\underline{s}}(n)) + \|\underline{d}'(\theta_m)\hat{s}_m(n)\|^2\Big]. \tag{21}$$

The recursive EM is then obtained by using (17) and (19) in the recursion (1). The recursive SAGE can be implemented by using (17), (21) and (8).

6. SIMULATIONS

The recursive procedures developed in the previous section are applied to simulated data. The narrowband signals are received by an uniformly linear array with 15 sensors. The signals are generated by 3 sources with equal power located at $\underline{\theta}_{\mathrm{true}} = [24^\circ\ 28^\circ\ 43^\circ]$. Note that the first two sources have only half a beamwidth separation. The initial estimate is given by $\underline{\theta}^0 = [19^\circ\ 32^\circ\ 48^\circ]$. The signal to noise ratio (SNR) is kept at 0, 10 dB. Each experiment runs through 50 Monte Carlo trials. The step size $a_n = an^{-\alpha}$ is chosen to be $3n^{-0.6}$. Each recursion is performed from $n = 1$ to 100.

The mean values of estimates from 50 Monte Carlo trials are presented in fig. 1 and fig. 2. The Mean Squared Errors (MSE) are displayed in fig. 3 and fig. 4. With increasing time, both algorithms lead to decreasing MSE. They also converge faster at the higher SNR. Given the same step size a_n, recursive SAGE has a better convergence rate than recursive EM. The diagonal augmented information matrices $\mathcal{I}_{\mathrm{EM}}(\underline{\theta})$, $\mathcal{I}^{[c]}_{\mathrm{SAGE}}(\underline{\theta})$ for this problem make the recursions (1), (8) very easy to compute. It is possible to use these recursions for real time processing.

7. CONCLUSION

The problem of recursive parameter estimation using augmented data was studied in this work. We proposed recursive EM and SAGE algorithms to facilitate sequential processing of arriving

data. It was proved that under mild conditions sequences of estimates generated by the proposed algorithms are strongly consistent and asymptotically normal distributed. In addition, we applied them to DOA estimation and obtained good simulation results. With this example we demonstrated that recursive EM and SAGE algorithms can be useful numerical tools for practical applications.

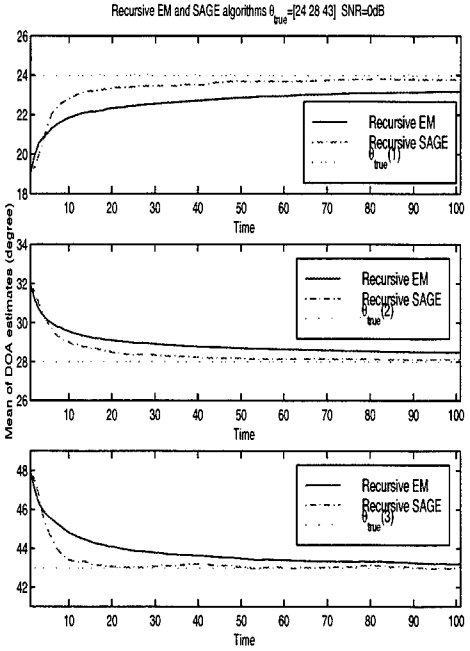


Fig.1. Recursive EM and SAGE with application to DOA Estimation. Mean of DOA estimates. $\underline{\theta}_{\mathbf{true}} = [24^\circ\ 28^\circ\ 43^\circ]$, SNR=0dB.

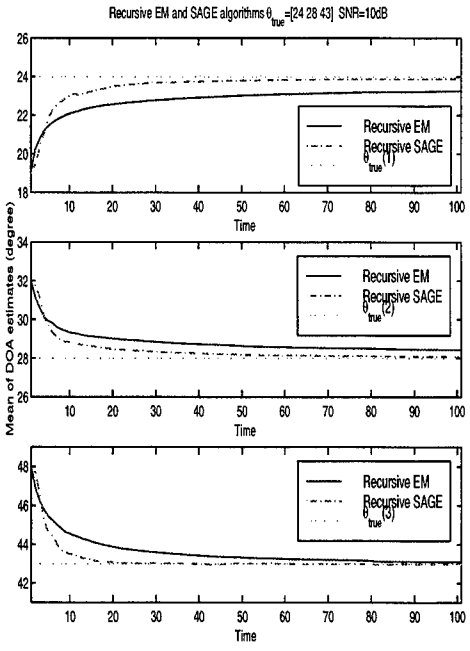


Fig. 2. Recursive EM and SAGE with application to DOA Estimation. Mean of DOA estimates. $\underline{\theta}_{\mathbf{true}} = [24^\circ\ 28^\circ\ 43^\circ]$, SNR=10dB.

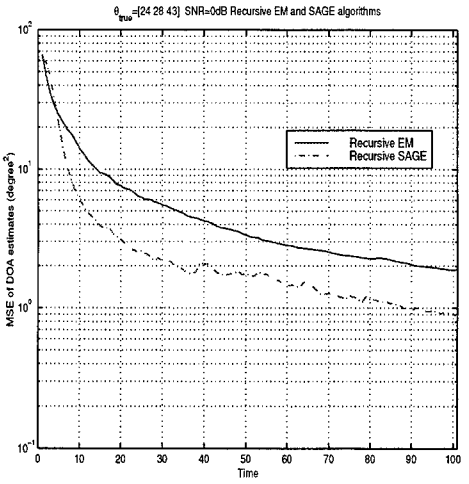


Fig. 3. MSE at each recursion. SNR=0dB.

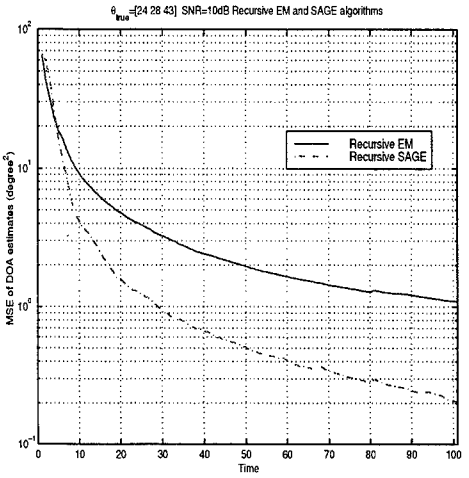


Fig. 4. MSE at each recursion. SNR=0dB.

8. REFERENCES

[1] A. P. Dempster, N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

[2] Václav Fabian. On asymptotically efficient recursive estimation. *The Annals of Statistics*, 6(4):854–866, 1978.

[3] Jeffrey A. Fessler and Alfred O. Hero. Space–alternating generalized expectation–maximization algorithm. *IEEE Trans. Signal Processing*, 42(10):2664–2677, October 1994.

[4] Harold J. Kushner and G. George Yin. *Stochastic Approximation Algorithms and Applications*. Applications of Mathematics. Springer, 1997.

[5] Georg Ch. Pflug. *Optimization of Stochastic Models*. Kluwer Academic Publishers, Boston, Dordrecht, London, 1996.

[6] D. M. Titterington. Recursive parameter estimation using incomplete data. *J. R. Statist. Soc. B*, 46(2):257–267, 1984.

[7] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, first edition, 1985.

CONVERGENCE PERFORMANCE OF SUBBAND ARRAYS FOR SPATIO-TEMPORAL EQUALIZATION

Yimin Zhang[†], *Kehu Yang* [‡], *and Moeness G. Amin*[†]

[†] Department of Electrical and Computer Engineering,
Villanova University, Villanova, PA 19085
E-mail: zhang,moeness@ece.villanova.edu

[‡] ATR Adaptive Communications Research Laboratories,
Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
E-mail: yang@acr.atr.co.jp

ABSTRACT

Subband arrays have been proposed as useful means to realize joint spatio-temporal domain equalization in digital mobile communications. They are used to mitigate channel impairment problems caused by the inter-symbol interference (ISI) and co-channel interference (CCI). In this paper, we propose the normalized subband array and locally orthogonalized subband array techniques for channel equalizations. The least square mean (LMS) algorithm is used for adaptation. The convergence performance of the proposed techniques is analyzed and compared with that of conventional space-time adaptive processing (STAP) techniques. It is shown that subband decompositions provide great flexibility in implementing spatio-temporal equalizations. Both analytical and numerical simulation results demonstrate that the proposed subband array techniques substantially improve the convergence performance without significant additional computations.

1. INTRODUCTION

For high-speed digital wireless networks, the communication channels are often frequency-selective, and the inter-symbol interference (ISI) becomes highly pronounced. Another pressing problem in mobile communication is the co-channel interference (CCI), which is generated due to the frequency reuse in cellular systems. Space-time adaptive processing (STAP) systems prove useful in suppressing both the ISI and CCI, leading to increased capacity and range [1].

Subband adaptive arrays have been proposed as alternative to STAP for a variety of purposes. In [2] – [4], the authors proposed to use subband arrays to realize joint spatial and temporal domain equalization. In [5], the steady state mean square error (MSE) performance was analyzed for different feedback schemes of subband arrays.

The work of Y. Zhang and M. G. Amin is supported by the Office of Naval Research under Grant N00014-98-1-0176 and the Air Force Research Laboratory under Grant no. F30602-00-1-0515.

In this paper, we propose the normalized least mean square (LMS) subband array and the locally orthogonalized LMS subband array techniques, and analyze their convergence properties in comparison with that of conventional STAP techniques. It is shown that the subband decomposition offers flexible implementation of the spatio-temporal equalization. The analysis and numerical simulations demonstrate that the proposed subband array techniques substantially improve the convergence performance without significant additional computations.

2. SIGNAL MODEL

We consider a base station using an antenna array of N sensors with P users, where $P < N$. The signal of interest is denoted by $s_1(m), m \in (-\infty, \infty)$, whereas the signals from other users are $s_p(m)$, $p = 2, ..., P$. Assume the symbol period T is common for the P users. Then, the received data vector at the array, $\underline{x}(t) = [x_1(t), x_2(t), \cdots, x_N(t)]^T$, can be expressed as

$$\underline{x}(t) = \sum_{p=1}^{P} \sum_{m=-\infty}^{\infty} s_p(m)\underline{h}_p(t - mT) + \underline{b}(t) \qquad (1)$$

where

$s_p(m)$: information symbol of the pth user,

$\underline{h}_p(t)$: channel response vector (including pulse shaping filter) of the pth user,

$\underline{b}(t)$: additive noise vector.

We make the following assumptions.

A1) The user signals $s_p(m), p = 1, 2, ..., P$, are wide-sense cyclo-stationary and i. i. d. (independent and identically distributed) with $E[s_p(m)s_p^*(m)] = 1$.

A2) All channels $\underline{h}_p(t), p = 1, 2, ..., P$, are linear, quasi-static, and of a finite duration within $[0, D_pT]$. That is, $\underline{h}_p(t) = 0, p = 1, 2, ..., P$, for $t > D_pT$ or $t < 0$.

A3) The noise vector $\underline{b}(t)$ is zero-mean, temporally and spatially white with

$$E[\underline{b}(t)\underline{b}^T(t + \tau)] = \mathbf{0}, \text{ for any } \tau,$$

and

$$E[\underline{b}(t)\underline{b}^H(t+\tau)] = \sigma\delta(\tau)\mathbf{I}_N,$$

where the superscripts ^{T} and ^{H} denote transpose and conjugate transpose, respectively, $\delta(\cdot)$ is the delta function, and $\mathbf{I}_N$ is the $N \times N$ identity matrix.

The data vector $\underline{x}(t)$ is sampled at $T_s = T/J$, where $J \geq 1$ is an integer representing the oversampling rate. Define

$$\tilde{x}(n) = [\underline{x}^T(nT) \quad \underline{x}^T(nT - T_s) \quad \cdots \quad \underline{x}^T(nT - (J-1)T_s)]^T$$

as the $JN \times 1$ input signal by considering the J oversampling branches as virtual channels. Accordingly,

$$\tilde{x}(n) = \sum_{p=1}^{P} \sum_{m=-\infty}^{\infty} s_p(m)\tilde{h}_p(n-m) + \tilde{b}(n) \tag{2}$$

where

$$\tilde{h}_p(n) = [\underline{h}_p^T(nT) \quad \underline{h}_p^T(nT - T_s) \quad \cdots \quad \underline{h}_p^T(nT - (J-1)T_s)]^T$$

and

$$\tilde{b}(n) = [\underline{b}^T(nT) \quad \underline{b}^T(nT - T_s) \quad \cdots \quad \underline{b}^T(nT - (n-1)T_s)]^T.$$

By storing M symbols (i.e., JM snapshots) of the data vectors, we obtain the following expression

$$\mathbf{x}(n) = \sum_{p=1}^{P} \mathbf{H}_p\mathbf{s}_p(n) + \mathbf{b}(n), \tag{3}$$

where

$$\mathbf{x}(n) = [\tilde{x}^T(n) \quad \tilde{x}^T(n-1) \quad \cdots \quad \tilde{x}^T(n-M+1)]^T$$

$$\mathbf{H}_p = \begin{bmatrix} \tilde{h}_p(0) & \cdots & \tilde{h}_p(D_p) & 0 & \cdots & \cdots & 0 \\ 0 & \tilde{h}_p(0) & \cdots & \tilde{h}_p(D_p) & 0 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & \cdots & 0 & \tilde{h}_p(0) & \cdots & \tilde{h}_p(D_p) \end{bmatrix}$$

$$\mathbf{s}_p(n) = [s_p(n) \quad s_p(n-1) \quad \cdots \quad s_p(n-M-D_p+1)]^T$$

and

$$\mathbf{b}(n) = [\tilde{b}^T(n) \quad \tilde{b}^T(n-1) \quad \cdots \quad \tilde{b}^T(n-M+1)]^T.$$

3. SPACE-TIME ADAPTIVE PROCESSING

Denote $\mathbf{w}(n)$ as the weight vector of a STAP system corresponding to the received data vector $\mathbf{x}(n)$. Then, the output of the STAP system becomes

$$y(n) = \mathbf{w}^H(n)\mathbf{x}(n). \tag{4}$$

By assuming that the reference signal is the ideal replica of the desired signal, the error signal is

$$e(n) = s_1(n-v) - y(n) = s_1(n-v) - \mathbf{w}^H(n)\mathbf{x}(n), \tag{5}$$

where v denotes some proper time delay [6]. By applying the LMS algorithm to above system, the weight vector is updated by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu_0\mathbf{x}(n)e^*(n) \tag{6}$$

where $\mu_0 > 0$ is the step size. To ensure convergence, the step size μ_0 should satisfy [7]

$$\mu_0 < 2/\lambda_{\max} \tag{7}$$

where $\lambda_{\max}$ is the maximum eigenvalue of the following covariance matrix

$$\mathbf{R} = E[\mathbf{x}\mathbf{x}^H] = \sum_{p=1}^{P} \mathbf{H}_p\mathbf{H}_p^H + \sigma\mathbf{I}_{JMN} \tag{8}$$

and $E[\cdot]$ denotes statistical expectation operator. As it is often a complicated problem to estimate the maximum eigenvalue, the step size is, in practice, determined by the total power of the received signal vector, as

$$\mu_0(n) = \mu/P(n), \tag{9}$$

where $\mu < 2$ is a constant, and $P(n) = E[\mathbf{x}^H(n)\mathbf{x}(n)]$ is the power, which can be estimated recursively by the following equation

$$P(n+1) = \alpha P(n) + (1-\alpha)\mathbf{x}^H(n)\mathbf{x}(n) \tag{10}$$

for $0 < \alpha < 1$.

The time constants of the convergence are given by [8]

$$\tau_i = \frac{1}{4\mu\lambda_i}, \quad 1 \leq i \leq JMN \tag{11}$$

where λ_i's are the eigenvalues of the covariance matrix $\mathbf{R}$. Comparing (11) with (7), it is well-known that the eigenvalue spread $\lambda_{\max}/\lambda_{\min}$ should be small to warrant fast convergence. However, the eigenvalue spread is usually large due to the channel dispersions and the high signal correlation across the virtual channels [6]. Several methods, including self-orthogonalizing methods and subspace-based methods, have been proposed to reduce the eigenvalue spread and to improve the convergence rate [7]. Nevertheless, these methods are complicated in the sense that they require either matrix inversion or eigen-decomposition of a large covariance matrix $\mathbf{R}$. In the next section, we consider simpler approaches based on subband decomposition.

4. SUBBAND ARRAYS

4.1. Subband Decomposition

Subband decomposition is realized by using a set of analysis and synthesis filters. Discrete Fourier transform (DFT) and modified-QMF filter banks are examples of perfect reconstructed and near-perfect reconstruction filter banks, respectively [4]. Decimation can be applied post the analysis filters to reduce the processing data rate, provided that the decimation rate does not exceed the number of subbands. Decimation often reduces the steady state system performance due to aliasing. Furthrt, perfect reconstruction property can be easily destroyed if adaptive techniques

are employed after decimation due to changes in the aliasing characteristics. In this paper, no decimation is applied for subband signal components. Also, synthesis filters are not considered for simplicity.

Let the subband decomposition devide the data sequence at the output of ith virtual channel, $\tilde{x}_i(n)$, into K subband sequences, $x_i^{(1)}(n), \cdots, x_i^{(K)}(n)$, where the superscript (k) denotes the data component at the kth subband. We define

$$\mathbf{x}_T(n) = \left[\left(\underline{\mathbf{x}}_T^{(1)}(n) \right)^T, \cdots, \left(\underline{\mathbf{x}}_T^{(K)}(n) \right)^T \right]^T$$

as the data vector for the subband arrays with

$$\underline{\mathbf{x}}_T^{(k)}(n) = \left[x_1^{(k)}(n), x_2^{(k)}(n) \cdots, x_N^{(k)}(n) \right]^T.$$

The filter banks can be designed to reduce the cross-correlation between signal components in adjacent subbands. The correlation reduction permits the processing the subband signals individually. In the following, two subband array techniques are proposed.

4.2. Normalized LMS Subband Arrays

Processing the signal vector for a subband array, $\mathbf{x}_T(n)$, by the weight vector

$$\mathbf{w}_T(n) = \left[\left(\mathbf{w}_T^{(1)}(n) \right)^T \left(\mathbf{w}_T^{(2)}(n) \right)^T \cdots \left(\mathbf{w}_T^{(K)}(n) \right)^T \right]^T,$$

the output of the subband array becomes

$$y_T(n) = \mathbf{w}_T^H(n)\mathbf{x}_T(n) = \mathbf{w}_T^H(n)\mathbf{T}\mathbf{x}(n) \qquad (12)$$

and the error signal is given by

$$e_T(n) = s_1(n) - y_T(n) = s_1(n) - \mathbf{w}_T^H \mathbf{x}_T(n). \qquad (13)$$

The adaptive subband arrays can implement the normalized LMS algorithm, where the weight vector is updated by

$$\mathbf{w}_T(n+1) = \mathbf{w}_T(n) + \mu \mathbf{P}^{-1} \mathbf{x}_T(n) e_T^*(n), \qquad (14)$$

where $\mathbf{P} = diag[P^{(k)}, k = 1, 2, \cdots, K]$ is a diagonal matrix with the signal power at the kth subband $P^{(k)} = E\left[(\mathbf{x}_T^{(k)})^H \mathbf{x}_T^{(k)} \right]$ as its kth diagonal element. $P^{(k)}$ is often estimated recursively by

$$P^{(k)}(n+1) = \alpha P^{(k)}(n) + (1-\alpha)(\mathbf{x}_T^{(k)}(n))^H \mathbf{x}_T^{(k)}(n). \qquad (15)$$

Unlike the STAP, where only one step size is defined, in the subband array, the equivalent step size

$$\mu^{(k)}(n) = \mu / P^{(k)}(n) \qquad (16)$$

is used for the kth subband. Proper normalization of the step size can, indeed, greatly reduce the eigenvalue spread of the covariance matrix, specifically in the case when the signal arrivals have different power at different frequencies. The change in the signal power spectrum may be simply caused by the frequency selectivity of the channels. The eigenvalue spread due to unflat signal spectrum can be well compensated by adjusting the step sizes, provided that the signal correlation between different subbands is small.

However, subband array processing may still suffer from high signal correlations across the output of each virtual channel, which limit the convergence improvement. Below, we propose locally orthogonalized LMS subband arrays for further convergence improvement.

4.3. Locally Orthogonalized LMS Subband Arrays

The locally orthogonalized LMS subband arrays perform eigen-decomposition separately at each subband and determine the step sizes based on their respective eigenvalues. It is important to note that, unlike the subspace-based LMS subband array techniques, where the eigen-decomposition of the covariance matrix $\mathbf{R}$ of size $JMN \times JMN$ could be computationally prohibitive, in the proposed method, the matrix at each subband is of dimension $JN \times JN$, which is considerably smaller and more amenable to fast computations.

Let $\mathbf{R}_T^{(k)} = E[\mathbf{x}_T^{(k)}(n)(\mathbf{x}_T^{(k)}(n))^H]$ denote the signal covariance matrix of $\mathbf{x}_T^{(k)}(n)$ defined at the kth subband. Similar to the power estimation, $\mathbf{R}_T^{(k)}$ can also be estimated recursively by

$$\mathbf{R}_T^{(k)}(n+1) = \beta \mathbf{R}_T^{(k)}(n) + (1-\beta)\mathbf{x}_T^{(k)}(n)(\mathbf{x}_T^{(k)}(n))^H, \qquad (17)$$

where $0 < \beta < 1$. Let $\mathbf{\Lambda}^{(k)} = diag\left[\lambda_1^{(k)}, \cdots, \lambda_N^{(k)} \right]$ be the diagonal eigenvalue matrix of $\mathbf{R}_T^{(k)}(n)$ and $\mathbf{F}^{(k)}$ the matrix with the corresponding eigenvectors as its columns. The vector $\mathbf{x}_T^{(k)}(n)$ can be decorrelated by using the following orthogonal projection,

$$\mathbf{x}_o^{(k)}(n) = \left(\mathbf{F}^{(k)} \right)^H \mathbf{x}_T^{(k)}(n). \qquad (18)$$

The weight vector at the kth subband, denoted as $\mathbf{w}_o^{(k)}$, can be updated similar to the normalized LMS subband array as

$$\mathbf{w}_o^{(k)}(n+1) = \mathbf{w}_o^{(k)}(n) + \mu \left(\mathbf{\Lambda}^{(k)} \right)^{-1} \mathbf{x}_o^{(k)}(n) e_o^*(n), \qquad (19)$$

where $e_o(n)$ is the error signal at the system output.

5. SIMULATION RESULTS

Computer simulations are performed to demonstrate the effectiveness of the proposed subband array techniques in terms of convergence performance improvement over the conventional STAP systems. A three-element array with uniform circular arrangement is employed. The interelement spacing is $\sqrt{3}$ wavelength, and the oversampling factor is $J = 2$. One desired user and two cochannel interferers are considered. All of the user signals are modulated by QPSK with FIR raised-cosine pulse shaping filtering, where the rolloff factor is set to 1.0. Six rays are randomly generated for each user signal. The different parameters are listed in Tables 1-3, where θ, ϕ, τ, and ξ are, respectively, the elevation angle of arrival (AOA), azimuth AOA, time delay, and propagation loss. The input SNR of the direct ray is 10 dB for each user signal.

In this simulation example, 24 taps are used for the STAP system, and the DFT filter bank is used to generate 24 subband bins for the subband arrays. The output residual error power is shown in Fig. 1. It is evident from this figure that the power normalization across the subbands alone does not greatly improve the convergence performance, whereas the improvement of the convergence performance by the locally orthogonalized subband array is evident.

6. CONCLUSION

In this paper, two simple subband array techniques were proposed to perform spatio-temporal equalization with improved convergence performance compared with conventional STAP techniques. The normalized LMS subband arrays and the locally orthogonalized LMS subband arrays respectively decorrelate the impinging signals in the temporal domain and the spatio-temporal domains, resulting in improved convergence performance with proper adjustment of the step sizes. The importance of decorrelating the signal arrivals across the virtual channels was emphasized.

7. REFERENCES

[1] A. J. Paulraj and C. B. Papadias, "Space-time processing for wireless communications," *IEEE Signal Processing Magazine*, vol. 14, no. 6, pp. 49–83, Nov. 1997.

[2] Y. Zhang, K. Yang, and M. G. Amin, "Adaptive subband arrays for multipath fading mitigation," in *Proc. IEEE Antennas and Propagation Society Int. Symp.*, Atlanta, GA, pp. 380–383, June 1998.

[3] Y. Zhang, K. Yang, and Y. Karasawa, "Subband CMA adaptive arrays in multipath fading environment," *IEICE Trans. Commun.*, vol. J82-B, no. 1, pp. 97–108, Jan. 1999.

[4] Y. Zhang, K. Yang, and M. G. Amin, "Adaptive array processing for multipath fading mitigation via exploitation of filter banks," *IEEE Trans. Antennas Propagat.*, vol. 49, no. 4, pp. 505–516, April 2001.

[5] Y. Zhang, K. Yang, and Y. Karasawa, "Performance analysis of subband arrays," in *Proc. Int. Symp. on Antennas and Propagation*, Fukuoka, Japan, pp. 903–906, Aug. 2000.

[6] K. Yang, Y. Zhang, and Y. Mizuguchi, "A signal subspace-based subband approach to space-time adaptive processing for mobile communications," *IEEE Trans. Signal Processing*, vol. 49, no. 2, pp. 401–413, Feb. 2001.

[7] S. Haykin, *Adaptive Filter Theory*, 3rd Ed. New Jersey: Prentice Hall, 1996.

[8] B. Widrow, J. McCool, M. G. Larimore, and C. R. Johnston Jr., "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 64, pp. 1151–1162, Aug. 1976.

Table 1: Parameters of the desired signal

No.	θ (deg)	ϕ (deg)	τ (sym)	ξ
1	12.3	24.6	0	1.0
2	8.3	19.1	0.62	$-0.18 - j0.77$
3	26.7	6.2	1.72	$-0.51 - j0.59$
4	24.0	56.8	5.06	$0.68 + j0.30$
5	9.3	13.6	7.78	$-0.13 - j0.60$
6	26.6	37.3	7.90	$0.33 - j0.31$

Table 2: Parameters of interferer 1

No.	θ (deg)	ϕ (deg)	τ (sym)	ξ
1	8.6	33.6	0	1.0
2	19.0	53.4	2.31	$0.85 + j0.18$
3	12.4	66.6	2.65	$2.23 - j0.68$
4	29.0	49.5	5.30	$0.62 - j0.32$
5	22.1	53.5	5.62	$-0.27 + j0.50$
6	26.2	73.7	7.55	$0.05 - j0.56$

Table 3: Parameters of interferer 2

No.	θ (deg)	ϕ (deg)	τ (sym)	ξ
1	6.6	120.6	0	1.0
2	12.2	149.8	4.04	$0.14 + j0.90$
3	6.3	135.4	6.04	$0.05 + j0.67$
4	14.7	131.7	6.56	$-0.49 + j0.42$
5	17.9	133.5	9.65	$0.28 - j0.38$
6	7.0	134.1	11.02	$0.10 + j0.42$

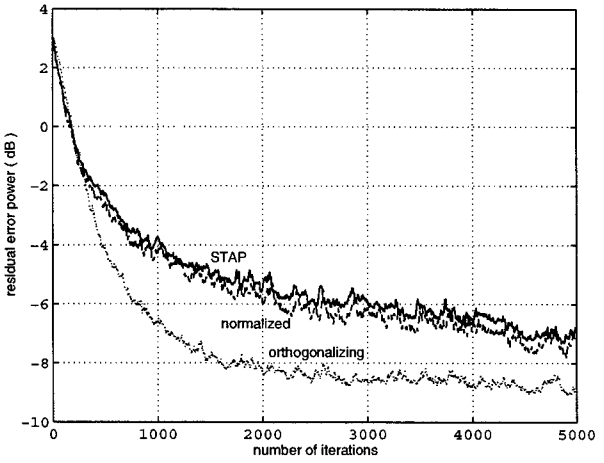


Fig. 1 Comparison of the residual error power of STAP and subband arrays.

SIMULATION AND PERFORMANCE BOUNDS FOR REAL-TIME PREDICTION OF THE MOBILE MULTIPATH CHANNEL

Paul D. Teal and Rodney G. Vaughan

Communications Team, Industrial Research Limited
PO Box 31-310, Lower Hutt, New Zealand
Ph: +64 4 569 0116, Fx: 0754, Email: p.teal@irl.cri.nz

ABSTRACT

The fading encountered in multipath mobile communication channels is often the primary cause of degradation in communication system performance. There are many situations in mobile communications in which it would be advantageous for a communications system to have real time information on how a signal will fade in advance of the fade actually occurring. This paper looks at the ways in which this real time prediction of the mobile channel can be achieved. Physical models of mobile channels which allow prediction are discussed. Algorithms based on these models, and their performance, are presented. Performance bounds for model based prediction based on the Cramer Rao bound are also derived. The algorithms are also applied to measured channel data.

1. INTRODUCTION

IN multipath propagation environments, the signal received by an antenna has a complex Gaussian distribution, providing there are many multipath components. The fading envelope, which has a Rayleigh or Rician distribution, is a dominant feature in mobile radio communications.

There are many situations in mobile communications in which it would be advantageous for a communications system to have information on how a signal will fade in advance of the fade actually occurring. If the timing of a fade is known sufficiently in advance, there will be sufficient time for corrective action to be negotiated between the transmitter and receiver [1,2]. This corrective action may for example take the form of change of time-slot (for a TDMA system), change of frequency (for a FDMA system), change of power level, or change of coding scheme. Several approaches to continuous power adjustment have been suggested, but these suffer when the channel information available to the transmitter is obsolete. The information will be obsolete for a rapidly varying channel. Channel prediction can overcome this problem. A similar argument applies to improvement of transmit antenna diversity systems and the Multiple-Input Multiple-Output (MIMO) systems which require channel information at the transmitter.

The statistical modelling of the channels is often taken to imply that the channel variation is random, i.e., cannot be predicted more than the correlation distance (or time) of the changing channel. However, by introducing a physical model for the multipath, it becomes possible, in principle, to apply signal processing to predict the channel over distances much longer than the correlation distance. The idea is to model the channel, and use the samples along a known spatial trajectory to estimate the channel model parameters. The channel model can then be used to extrapolate beyond the region of the measurements.

In this paper, a model for the fading channel of a radio communications signal is presented, and bounds for prediction based on the model are established.

2. NARROWBAND FADING CHANNEL MODEL

The model primarily used in this paper is that of far-field scatterers of a narrowband signal surrounding the receiver. This model is derived and used for example in [3,4]. The measurement segment becomes in effect a synthetic array, and the relative delay to each of the elements of this array is used as the basis for localising the scatterers. It is assumed that samples $\{r_m\}$ of the the complex channel gain can be simply derived from the receiver (i.e., the data sequence is known at the receiver, or a decision feedback mechanism is used). These samples may be described using:

$$r_m = \sum_{n=1}^{N} \zeta_n e^{j\frac{2\pi}{\lambda}\Delta_x \sin\theta_n} + \eta_m \qquad (1)$$

where

- r_m is the mth complex channel sample,
- N is the number of discrete scatterers present,
- θ_n is the bearing of the scatterer location
- Δ_x is the distance travelled by the receiver between channel samples,
- λ is the signal carrier wavelength,
- η_m is a sampled complex Gaussian noise process of zero mean, assumed to be white (or whitened) with variance σ_η^2 so that $E\{\eta_{m_1}\eta^*_{m_2}\} = \sigma_\eta^2\delta_{m_1,m_2}$,
- ζ_n is the complex attenuation of a far field scattering point at angle θ. This term includes factors such as space loss and polarisation mismatch.

Equation (1) can be seen to be equivalent to the model of complex sinusoids in noise. The *Doppler frequency* of scatterer n is defined as $\varpi_n = \frac{2\pi}{\lambda}\Delta_x \sin\theta_n$. The model can be expressed in matrix forms as

$$\mathbf{r} = \mathbf{A}(\boldsymbol{\varpi})\boldsymbol{\zeta} + \boldsymbol{\eta} \qquad (2)$$

where

$$\mathbf{A}(\boldsymbol{\varpi}) = [\mathbf{a}(\varpi_1), \mathbf{a}(\varpi_2), \ldots \mathbf{a}(\varpi_N)]$$
$$\mathbf{a}(\varpi) = \left[e^{j\varpi 0}, e^{j\varpi 1}, \ldots, e^{j\varpi(M-1)}\right]^T,$$

$$\boldsymbol{\varpi} = [\varpi_1, \ldots, \varpi_N]^T \qquad \mathbf{r} = [r_1, \ldots, r_M]^T$$
$$\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_N]^T \qquad \boldsymbol{\eta} = [\eta_1, \ldots, \eta_M]^T$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

and M is the number of complex samples used for the prediction. The channel samples $\{r_m\}$, called the *measurement segment,* are known. The problem is to use only this information to estimate as accurately as possible the values in the *prediction segment* $\{r_m | M+1 \leq m \leq M_2\}$. This is achieved by first estimating the parameters of the model N, $\boldsymbol{\zeta}$, and $\boldsymbol{\varpi}$, and using the estimated values to extrapolate the complex sinusoids.

The minimum description length (MDL) criterion for estimation of the "model order" $\widehat{N}$ is known to be both unbiased and consistent. The version of MDL presented in [5] is used in this paper.

The maximum likelihood estimates of the remaining parameters can be shown to be $\widehat{\boldsymbol{\zeta}} = \mathbf{A}(\widehat{\boldsymbol{\varpi}})^{+}\mathbf{r}$ where

$$\widehat{\boldsymbol{\varpi}} = \arg\max_{\boldsymbol{\varpi}} \quad \mathbf{r}^H \mathbf{A}(\boldsymbol{\varpi})\mathbf{A}(\boldsymbol{\varpi})^{+}\mathbf{r} \tag{3}$$

and the superscript $^{+}$ represents the Moore-Penrose generalised inverse. The maximum likelihood solution requires an N dimensional search which is generally impractical. Subspace methods of estimating $\boldsymbol{\varpi}$ have been shown (e.g., see references in [6]) to achieve results close the Cramer Rao bound provided the signal-to-noise ratio (SNR) is adequate, and these have been used in this paper.

3. CRAMER RAO BOUND

3.1. Bound Formulation

From equation (2), it can be seen that $\mathbf{r}$ is a complex Gaussian random vector with mean

$$\boldsymbol{\mu} = \sum_{n=1}^{N} \mathbf{r}_n = \sum_{n=1}^{N} \zeta_n e^{j\varpi_n \mathbf{m}} \tag{4}$$

and variance $\mathbf{C} = \sigma_\eta^2 \mathbf{I}$. Calculation of the bound is facilitated by expressing the attenuation as the real parameters amplitude and phase; thus $\zeta_n = \varsigma_n e^{j\psi_n}$, so the parameter vector $\boldsymbol{\xi}$ is $(\varsigma_1, \psi_1, \varpi_1, \ldots, \varsigma_N, \psi_N, \varpi_N)^T$. Using Bangs' formula [7], the elements of the Fisher information matrix are given by

$$[\mathbf{J}(\boldsymbol{\xi})]_{ij} \quad = \quad \mathrm{tr}\left(\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \xi_i}\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \xi_j}\right) + 2\,\mathrm{Re}\left(\frac{\partial \boldsymbol{\mu}^H}{\partial \xi_i}\mathbf{C}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \xi_j}\right)$$

Differentiating with respect to each of the parameters, the Fisher information matrix can be shown to consist of 3×3 blocks of the form:

$$\frac{2}{\sigma_\eta^2}\,\mathrm{Re}\left\{ e^{j(\psi_{n_1}-\psi_{n_2})} \begin{pmatrix} 1 & j\varsigma_{n_2} & j\varsigma_{n_2} \\ -j\varsigma_{n_1}^* & \varsigma_{n_1}^*\varsigma_{n_2} & \varsigma_{n_1}^*\varsigma_{n_2} \\ -j\varsigma_{n_1}^* & \varsigma_{n_1}^*\varsigma_{n_2} & \varsigma_{n_1}^*\varsigma_{n_2} \end{pmatrix} \odot \sum_m \left[e^{jm(\varpi_{n_2}-\varpi_{n_1})} \begin{pmatrix} 1 & 1 & m \\ 1 & 1 & m \\ m & m & m^2 \end{pmatrix} \right] \right\} \tag{5}$$

where n_1 and n_2 are the row and column indices for the 3×3 blocks, and $\odot$ represents element by element multiplication.

To obtain an estimate of typical prediction performance, this Fisher Information Matrix is calculated using parameters taken from the following distributions: the *complex* amplitudes ζ_n are independent identically distributed (iid) zero-mean complex Gaussian variables with variance of 1, and θ_n are iid uniform over $(\frac{-\pi}{2}, \frac{\pi}{2}]$ with $0 < \frac{2\pi}{\lambda}\Delta_x < \pi$.

The error between the actual and predicted complex envelope is

$$|\epsilon[m]|^2 = \left| \sum_{n=1}^{N} \varsigma_n e^{j\psi_n} e^{j\varpi_n m} - \sum_{n=1}^{N} \hat{\varsigma}_n e^{j\hat{\psi}_n} e^{j\hat{\varpi}_n m} \right|^2 . \tag{6}$$

Taking a first order approximation for $\epsilon[m]$ and defining ϵ_ς, ϵ_ψ and ϵ_ϖ as the difference between the actual and estimated values of each of the symbols,

$$\epsilon[m] \quad \approx \quad \sum_{n=1}^{N} \frac{\partial r[m]}{\partial \varsigma_n}\epsilon_{\varsigma_n} + \frac{\partial r[m]}{\partial \psi_n}\epsilon_{\psi_n} + \frac{\partial r[m]}{\partial \varpi_n}\epsilon_{\varpi_n}. \tag{7}$$

Taking the expectation

$$E\left\{|\epsilon[m]|^2\right\} \approx \sum_{n_1=1}^{N}\sum_{n_2=1}^{N} \varsigma_{n_1}\varsigma_{n_2} e^{j(\psi_{n_1}-\psi_{n_2})} e^{jm(\varpi_{n_1}-\varpi_{n_2})} h_{n_1 n_2},$$

where $h_{n_1 n_2}$ is the sum of the elements of the matrix

$$\begin{bmatrix} \frac{1}{\varsigma_{n_1}\varsigma_{n_2}} & \frac{-j}{\varsigma_{n_1}} & \frac{-jm}{\varsigma_{n_1}} \\ \frac{j}{\varsigma_{n_2}} & 1 & m \\ \frac{jm}{\varsigma_{n_2}} & m & m^2 \end{bmatrix} \odot \left(\mathbf{J}^{-1}\right)_{n_1 n_2} \tag{8}$$

and $\left[\mathbf{J}^{-1}\right]_{n_1 n_2}$ is the 3×3 block of the Fisher inverse with n_1 and n_2 being the row and column indices for the 3×3 blocks. This assumes of course that the bound is nearly achieved so that $E\{(\xi - \widehat{\xi})(\xi - \widehat{\xi})^H\} - \mathbf{J}^{-1}$ is not merely a positive semi-definite matrix [8, p82] as required by the Cramer-Rao inequality, but actually close to the zero matrix.

It is immediately apparent that for a large prediction range (large m), the most critical parameter is the frequency ϖ, since the variance of this estimate is multiplied by m^2 to calculate the overall prediction error.

3.2. Invertibility of the Fisher Information Matrix

A simulation of the scenario described above lead to the discovery that a large proportion of the Fisher information matrices were very poorly conditioned. It is conjectured in [9] that the Fisher information matrix is singular only if two or more of the tone frequencies are equal, modulo 2π, assuming M is large enough. Conditions affecting the likelihood of the Fisher information matrix being close to singular are the number of scatterers present N, the number of elements in the virtual array M, and the length of the virtual array $(M-1)\Delta_x$.

The number of scatterers for which the parameters may be reliably estimated for a given measurement segment length was investigated. The number of scatterers in a given scenario was reduced by eliminating scatterers for which ϖ_{n_1} was close to that of another scatterer ϖ_{n_2}, until the LINPACK reciprocal condition estimate [10] was larger than 10^{-15}. The results are shown in figure 1. If the actual number of scatterers N' is larger than the number N presented in figure 1, then the parameters of some of the scatterers cannot be estimated, even though two scatterers which are very close in Doppler frequency may be effectively modelled as one scatterer.

The power of the scatterers not parametrised has two effects. First it decreases the overall effective SNR, since some of the scatterers effectively become noise, and secondly it causes an increase in the error between the predicted and actual signal.

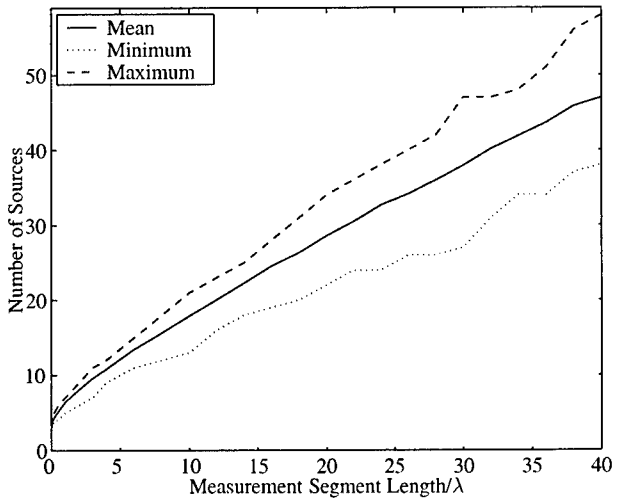


Figure 1: The number of scatterers which can be modelled as a function of the measurement segment length, using invertibility of the Fisher Information Matrix as criterion. The shaded region is that for which the number of real parameters to be estimated is less than or equal to the number of real measurements available.

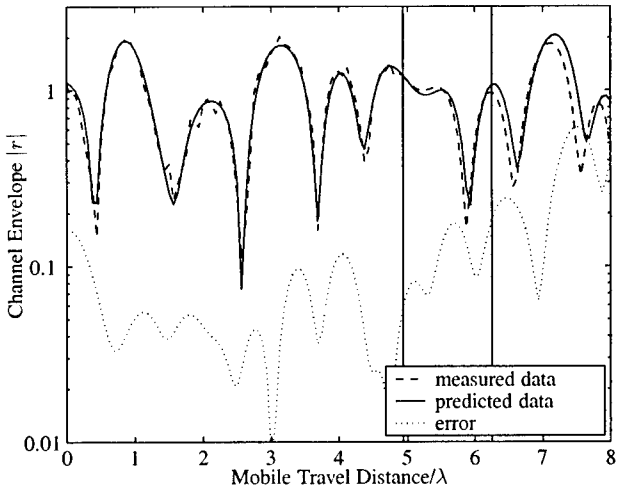


Figure 2: Example of Prediction Behaviour. Only the measured data to the left of the first vertical line is used for prediction. The region between the two vertical lines is where the predicted envelope is within 20% of the actual envelope. The RMS level of the *measured* envelope is 1, so prediction is said to "fail" when the error first exceeds 0.2

If $\epsilon_1[m]$ is the error that would be expected in the case where $N = N'$ and for unity SNR, and P is the power of the "ignored" scatterers where the power of all scatterers is unity, the expected error now becomes

$$E\{|\epsilon[m]|^2\} = (\sigma_\eta^2 + P)E\left(|\epsilon_1[m]|^2\right) + P. \quad (9)$$

3.3. Performance Measures

In the simulations of section 4 the performance criterion is the distance (in wavelengths) for which the predicted and actual signal envelopes differ by less than 20% of the root mean square (RMS) value of the envelope in the measurement segment. As can be seen in the example of prediction behaviour presented in figure 2, this may be a pessimistic criterion in many occasions, since the error even in the measurement segment may exceed 20%, especially when the SNR is small. The performance measure used in [1] (distance for which the predicted and actual signal envelopes are within 5% of the maximum amplitude value in the measurement segment) was used for very high SNR, and is not practical for the SNR values considered here.

The performance measure used in evaluating the CRLB was slightly different. Equations 8 and 9 yield an expected square error $E\{|\epsilon|^2\}$. Prediction in this case was said to be valid to the point where this error exceeded 0.04. Where there are more than about 10 scatterers present, those that are modelled as noise result in the mean square error being always larger than this number, even thought prediction will not necessarily always fail. For $N \leq 10$ and for long measurement distances, however there is reasonable agreement between the bound and the simulations, as shown by comparison of figure 3 and the simulations of figure 4.

4. SIMULATION

Simulations were used to provide an indication of the performance of the subspace algorithms when used to enable prediction using the far-field discrete scatterers model. The SNR was 20 dB with expected signal power of unity, with $M = 40$, and $\widehat{N}$ chosen by the MDL criterion. The number of independent scenarios used to find the mean performance presented in the graphs was 3000.

In figure 4 (unmarked lines) the prediction performance is presented as a function of the length (in wavelengths) of the measurement segment. A selection of subspace methods was used (MUSIC, ESPRIT, PCLP and Minimum Norm), and whichever of these yielded the lowest Mean Square Error (MSE) of the predicted channel in the measurement segment was selected for each scenario.

Prediction beyond about 0.2λ is not achieved until the length of the measurement segment exceeds 2λ - even when the actual scatterer locations are known perfectly.

For longer measurement segments significant prediction performance was obtained, the prediction range rising rapidly (between 1.2 and 1.7th power) with measurement segment length. Where the number of scatterers is very large however, the prediction performance actually *decreases* when the measurement segment length increases above about 1λ. This is actually an artefact of the spacing between samples increasing. The spacing is the same in the prediction segment as in the measurement segment and so the prediction length is observed as zero if it is less than the inter-sample distance.

A practical limitation to long range prediction appears to be the number of significant scatterers being large, which means that a long measurement segment is required. In practice, long measurement segments become impractical because it is unlikely that any physical scenario will remain static over trajectories of several wavelengths, except perhaps at very high carrier frequencies.

5. MEASUREMENTS

The prediction algorithms were applied to measured data as well as to synthesised data. The measurements were taken at several frequencies, mostly 1.92 and 5.9 GHz. There were three different locations used - inside a laboratory, outdoors, and inside a large workshop.

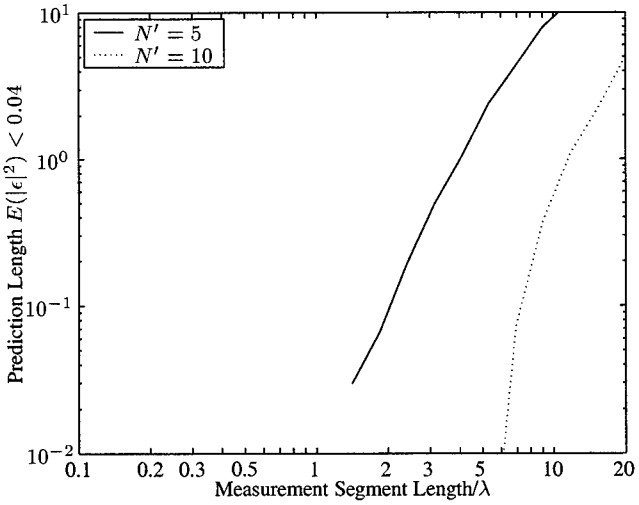


Figure 3: Prediction range based on CRLB

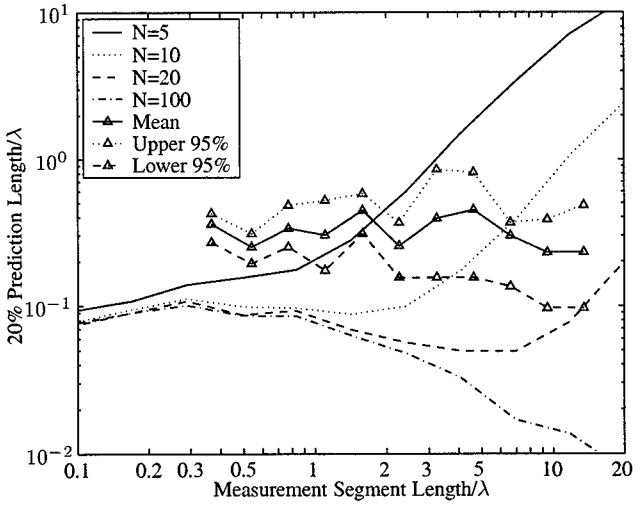


Figure 4: Prediction Performance versus Measurement Segment Length for both simulated data (points unmarked) and measured data (points marked with triangles) using subspace parameter estimation.

The subspace algorithms were applied to a series of data points measured in the laboratory (an irregularly shaped room of dimensions approximately 19 × 10 × 3 metres) at 1.92 GHz. This example is typical of measurements made in this and the other two locations. The channel was measured at 999 points spaced evenly over a 3 metre distance. Averaging of 50 scenarios was obtained by starting the measurement segment at different points in the data set, and by using different but relatively close frequencies. The likelihood of each frequency estimate $\widehat{\varpi}$ was increased by using a gradient method before estimating the amplitudes $\widehat{\zeta}$.

The results are shown in the triangle-marked lines of figure 4, with upper and lower 95% confidence limits. It is obvious that prediction does not improve significantly with increasing measurement segment length. There are two likely explanations of this. The first is that the scenario is not sufficiently static. The scatterers may be too close to the receiver for the far field model to be valid, or the scatterers may themselves move. The second is that the model is valid, but the number of scatterers is large. As the simulations show, if the scatterers are many, prediction is very limited. It would be expected that if there are many scatterers, the field becomes equivalent to a diffuse field, and prediction is truly confined to the correlation distance. If the scatterers are uniformly distributed in angle, this correlation distance is approximately $\frac{\lambda}{5}$ [e.g.,3].

6. CONCLUSIONS

A channel model based on narrow-band, far-field discrete scatterers has been presented. This model in principle allows long term prediction of channel behaviour based only on *a priori* samples of the channel.

A form of the CRLB for prediction error has been derived and applied, and the performance has also been evaluated on simulated channels.

It is widely assumed that the number of significant scatterers is small [11]. This assumption has been found to be critical to the viability of long range prediction. Several statistical measures of the channel are the same for a small or a large number of scatterers. However, the former can be parametrised to obtain long-range prediction, whereas the latter cannot. Evidence to date suggests that the number of significant scatterers in many situations is large.

7. REFERENCES

[1] J. B. Andersen, J. Jensen, S. H. Jensen, and F. Frederiksen, "Prediction of future fading based on past measurements," in *VTC-99/FALL, Delft, The Netherlands*, Sep 1999.

[2] A. Duel-Hallen, S. Hu, and H. Hallen, "Long-range prediction of fading signals," *IEEE Signal Proc. Mag.*, pp. 62–75, May 2000.

[3] R. H. Clarke, "A statistical theory of mobile-radio reception," *The Bell System Technical Journal*, pp. 957–1000, July-August 1968.

[4] P. D. Teal, R. Raich, and R. G. Vaughan, "Prediction of fading in the mobile multipath environment," *IEEE Trans. Vehicular Technology*, 2000. submitted.

[5] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. on Signal Processing*, vol. ASSP-33, pp. 387–392, April 1985.

[6] O. J. Micka and A. J. Weiss, "Estimating frequencies of exponentials in noise using joint diagonalization," *IEEE Trans. on Signal Processing*, vol. SP-47, pp. 341–348, Feb 1999.

[7] W. J. Bangs, *Array Processing with Generalized Beamformers*. PhD thesis, Yale University, New Haven, CT, 1971.

[8] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communciations and Control*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[9] D. Rife and R. Boorstyn, "Multiple tone parameter estimation from discrete-time observations," *Bell System Technical Journal*, vol. 55, pp. 1389–1410, Nov 1976.

[10] J. Wilkinson and C. Reinsch, *Handbook for Automatic Computation*, vol. 2. New York: Springer-Verlag, 1971.

[11] J.-K. Hwang and J. H. Winters, "Sinusoidal modeling and prediction of fast fading processes," in *Proceedings GLOBECOM*, vol. 2, (Sydney, Australia), pp. 892–897, Nov 1998.

ROBUSTNESS OF THE FINITE-LENGTH MMSE-DFE WITH RESPECT TO CHANNEL AND SECOND-ORDER STATISTICS ESTIMATION ERRORS

Athanasios P. Liavas

Department of Computer Science, University of Ioannina, 45110, Ioannina, Greece.
E-mail: liavas@cs.uoi.gr.

ABSTRACT

The filters constituting the minimum mean square error decision-feedback equalizer (MMSE-DFE), as well as related performance measures, can be computed by assuming perfect knowledge of the channel impulse response and the input and noise second-order statistics (SOS). In practice, we estimate the unknown channel and SOS, and inevitable estimation errors arise. We model estimation errors as small perturbations, i.e., of order ϵ, with ϵ a sufficiently small positive number, and we study the behavior of the MMSE-DFE under mismatch by performing a first-order perturbation analysis. We prove that the excess MSE induced by $O(\epsilon)$ estimation errors is $O(\epsilon^2)$, uncovering important robustness properties associated with the MMSE-DFE.

1. INTRODUCTION

The finite-length minimum mean square error decision-feedback equalizer (MMSE-DFE) has proved to be an efficient structure toward ISI mitigation in packet-based communication systems. The MMSE-DFE is determined by the feedforward and the feedback filter, which can be computed by assuming perfect knowledge of the channel impulse response and the input and additive channel noise second-order statistics (SOS) [1].

In practice, the channel impulse response and the input and noise SOS are unknown and we estimate them either by training or blindly. Thus, inevitable estimation errors arise. The robustness of the MMSE-DFE with respect to mismatch was first considered in [2], where the authors developed closed-form expressions for the "perturbed" MMSE-DFE filters and the corresponding performance measures. However, for the evaluation of these expressions they have to resort to computer simulations. Consequently, we feel that the analysis of [2] does not provide much analytical insight into the behavior of the MMSE-DFE under mismatch.

This work was supported by the EPETII Program of the Greek Secretariat for Research and Technology.

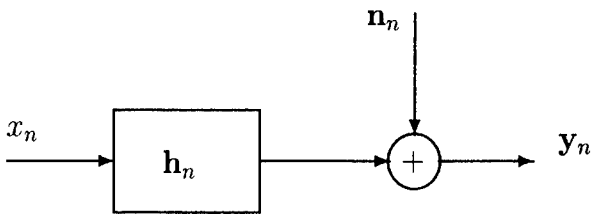


Figure 1: Channel model.

We model the channel impulse response and SOS estimation errors as perturbations of order ϵ, with ϵ being a small positive number, and we study the behavior of the MMSE-DFE under mismatch by using a first-order perturbation analysis. We show that the excess mean square error induced by $O(\epsilon)$ errors is $O(\epsilon^2)$. Simulations show that the range of ϵ for which our first-order analysis remains valid depends on the SNR.

2. FINITE-LENGTH MMSE-DFE

2.1. Channel Model

We consider the baseband discrete-time fractionally sampled noisy communication channel modeled by the ν-th order 1-input/p-output linear time-invariant system depicted in Fig. 1. Its input-output relation is

$$\mathbf{y}_n = \sum_{i=0}^{\nu} \mathbf{h}_i x_{n-i} + \mathbf{n}_n,$$

where x_n denotes the input sequence and the $p \times 1$ vectors $\mathbf{y}_n$, $\mathbf{n}_n$ and $\mathbf{h}_i$ denote, respectively, the terms of the output, noise and channel finite impulse response sequences. We define the impulse response parameter vector $\mathcal{H}_\nu \triangleq \left[\mathbf{h}_0^H \ \cdots \ \mathbf{h}_\nu^H \right]^H$, where superscript ^{H} denotes Hermitian transpose. The data vector

$$\mathbf{y}_{n:n-N_f+1} \triangleq \left[\mathbf{y}_n^H \ \cdots \ \mathbf{y}_{n-N_f+1}^H \right]^H$$

can be expressed as

$$\mathbf{y}_{n:n-N_f+1} = \mathbf{H}\,\mathbf{x}_{n:n-N_f-\nu+1} + \mathbf{n}_{n:n-N_f+1}$$

0-7803-7011-2/01/$10.00 ©2001 IEEE

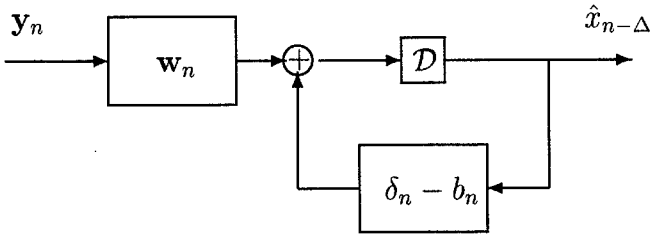


Figure 2: Finite-length DFE.

where the $pN_f \times (\nu + N_f)$ matrix $\mathbf{H}$ is defined by

$$\mathbf{H} \triangleq \begin{bmatrix} \mathbf{h}_0 & \cdots & \cdots & \mathbf{h}_\nu & & \\ & \ddots & & & \ddots & \\ & & \mathbf{h}_0 & \cdots & \cdots & \mathbf{h}_\nu \end{bmatrix}$$

and the definitions of $\mathbf{x}_{n:n-N_f-\nu+1}$ and $\mathbf{n}_{n:n-N_f+1}$ are obvious.

2.2. Finite-length MMSE-DFE

Our aim is to recover (a delayed version of) the input sequence x_n by passing the noisy output data $\mathbf{y}_n$ through the finite-length DFE depicted in Fig. 2. The DFE is determined by the following parameter vectors:

1. $\mathbf{w} \triangleq \left[\mathbf{w}_0^H \ \cdots \ \mathbf{w}_{N_f-1}^H \right]^H$, which denotes the p-input/1-output length-N_f feedforward filter;
2. $\mathbf{b} \triangleq \left[1 \ b_1^* \ \cdots \ b_{N_b}^* \right]^H$, which determines the single-input/single-output length-N_b strictly causal feedback filter. The settings of the feedback filter are $\{-b_1^*, \ldots, -b_{N_b}^*\}$, where superscript * denotes complex-conjugate.

Assuming that *the past decisions are correct* and considering delay Δ, the error between the desired output and the input to the decision device $\mathcal{D}$ is given by [1]

$$\begin{aligned} e_n &\triangleq x_{n-\Delta} - \left(\sum_{i=0}^{N_f-1} \mathbf{w}_i^H \mathbf{y}_{n-i} - \sum_{i=1}^{N_b} b_i^* x_{n-\Delta-i} \right) \\ &= \mathbf{b}^H \mathbf{x}_{n-\Delta:n-\Delta-N_b} - \mathbf{w}^H \mathbf{y}_{n:n-N_f+1} \\ &= \widetilde{\mathbf{b}}^H \mathbf{x}_{n:n-N_f-\nu+1} - \mathbf{w}^H \mathbf{y}_{n:n-N_f+1} \end{aligned}$$

where we have defined $\widetilde{\mathbf{b}} \triangleq \left[\mathbf{0}_{1\times\Delta} \ \mathbf{b}^H \ \mathbf{0}_{1\times s} \right]^H$, with $\mathbf{0}_{i\times j}$ denoting the $i\times j$ zero matrix and $s \triangleq N_f+\nu-\Delta-N_b-1$. We simplify notation by omitting the subscripts from $\mathbf{x}_{n:n-N_f-\nu+1}$, $\mathbf{y}_{n:n-N_f+1}$ and $\mathbf{n}_{n:n-N_f+1}$.

The MMSE-DFE settings are computed by minimizing the mean square error (MSE) $\mathcal{E}\left[|e_n|^2\right]$

$$\begin{aligned} \mathbf{MSE} &\triangleq \mathcal{E}\left[\left(\widetilde{\mathbf{b}}^H \mathbf{x} - \mathbf{w}^H \mathbf{y} \right) \left(\mathbf{x}^H \widetilde{\mathbf{b}} - \mathbf{y}^H \mathbf{w} \right) \right] \\ &= \widetilde{\mathbf{b}}^H \mathbf{R}_{xx} \widetilde{\mathbf{b}} - \widetilde{\mathbf{b}}^H \mathbf{R}_{xy} \mathbf{w} - \mathbf{w}^H \mathbf{R}_{yx} \widetilde{\mathbf{b}} + \mathbf{w}^H \mathbf{R}_{yy} \mathbf{w} \quad (1) \end{aligned}$$

where

$$\mathbf{R}_{xx} \triangleq \mathcal{E}[\mathbf{x}\mathbf{x}^H], \ \mathbf{R}_{xy} \triangleq \mathcal{E}[\mathbf{x}\mathbf{y}^H] = \mathbf{R}_{xx}\mathbf{H}^H = \mathbf{R}_{yx}^H$$

$$\mathbf{R}_{yy} \triangleq \mathcal{E}[\mathbf{y}\mathbf{y}^H] = \mathbf{H}\mathbf{R}_{xx}\mathbf{H}^H + \mathbf{R}_{nn}, \ \mathbf{R}_{nn} \triangleq \mathcal{E}[\mathbf{n}\mathbf{n}^H].$$

At the optimal settings, $\mathcal{E}\left[e_n \mathbf{y}^H\right] = \mathbf{0}_{1\times pN_f}$, yielding

$$\mathbf{R}_{yx}\widetilde{\mathbf{b}} = \mathbf{R}_{yy}\mathbf{w} \Longrightarrow \mathbf{w} = \mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\widetilde{\mathbf{b}}.$$

Substituting the above expression for $\mathbf{w}$ into (1), we obtain $\mathbf{MSE} = \widetilde{\mathbf{b}}^H \mathbf{R}\, \widetilde{\mathbf{b}}$, where

$$\mathbf{R} \triangleq \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}. \quad (2)$$

If we define

$$\mathbf{R}_\Delta \triangleq \left[\mathbf{0}_{(N_b+1)\times\Delta} \ \mathbf{I}_{N_b+1} \ \mathbf{0}_{(N_b+1)\times s} \right] \mathbf{R} \begin{bmatrix} \mathbf{0}_{\Delta\times(N_b+1)} \\ \mathbf{I}_{N_b+1} \\ \mathbf{0}_{s\times(N_b+1)} \end{bmatrix}$$

where $\mathbf{I}_i$ denotes the $i \times i$ identity matrix, then the MSE is expressed as $\mathbf{MSE} = \mathbf{b}^H \mathbf{R}_\Delta \mathbf{b}$ and it can be shown that it is minimized for [1]

$$\mathbf{b}_o = \frac{\mathbf{R}_\Delta^{-1}\,\mathbf{e}_0}{\mathbf{e}_0^H\,\mathbf{R}_\Delta^{-1}\,\mathbf{e}_0}, \quad \mathbf{w}_o = \mathbf{R}_{yy}^{-1}\,\mathbf{R}_{yx}\,\widetilde{\mathbf{b}}_o \quad (3)$$

where $\mathbf{e}_0$ is the vector with 1 at the first position and zeros elsewhere. The minimum MSE is given by

$$\mathbf{MMSE} \triangleq \widetilde{\mathbf{b}}_o^H \mathbf{R}\, \widetilde{\mathbf{b}}_o = \mathbf{b}_o^H \mathbf{R}_\Delta \mathbf{b}_o = \frac{1}{\mathbf{e}_0^H\,\mathbf{R}_\Delta^{-1}\,\mathbf{e}_0}. \quad (4)$$

3. MMSE-DFE: PERFORMANCE ANALYSIS UNDER MISMATCH

3.1. The framework

Let us assume that an estimation procedure has furnished the estimates $\{\widehat{\mathbf{h}}_i\}_{i=0}^{\hat{\nu}}$, leading to the impulse response vector estimate $\widehat{\mathcal{H}}_{\hat{\nu}} \triangleq \left[\widehat{\mathbf{h}}_0^H \ \cdots \ \widehat{\mathbf{h}}_{\hat{\nu}}^H \right]^H$. We consider the case $\hat{\nu} \leq \nu$. In order to compare the unequal length vectors $\mathcal{H}_\nu$ and $\widehat{\mathcal{H}}_{\hat{\nu}}$, we augment $\widehat{\mathcal{H}}_{\hat{\nu}}$ with leading and trailing zeros, obtaining

$$\widehat{\mathcal{H}}_\nu^{m_1} \triangleq \left[\mathbf{0}_{1\times pm_1} \ \widehat{\mathcal{H}}_{\hat{\nu}}^H \ \mathbf{0}_{1\times p(\nu-\hat{\nu}-m_1)} \right]^H$$

whose length equals the length of $\mathcal{H}_\nu$. Then, we define $m_1^* \triangleq \arg\min_{m_1} \|\mathcal{H}_\nu - \widehat{\mathcal{H}}_\nu^{m_1}\|_2$, where $\|\cdot\|_2$ denotes, depending on the argument, the matrix or vector 2-norm. That is, pm_1^* is the number of leading zeros we must insert in front of $\widehat{\mathcal{H}}_{\hat{\nu}}$, so that the augmented impulse response vector estimate becomes closest to $\mathcal{H}_\nu$. In the sequel, we shall work with the augmented

impulse response vector estimate $\widehat{\mathcal{H}}_\nu \triangleq \widehat{\mathcal{H}}_\nu^{m_1^*}$. We note that working with $\widehat{\mathcal{H}}_\nu$ instead of $\widehat{\mathcal{H}}_{\hat{\nu}}$ amounts simply to insertion of an extra delay of m_1^* time units.

We consider our channel estimate as being good if $\mathcal{H}_\nu$ and $\widehat{\mathcal{H}}_\nu$ are close to each other. In terms of the associated filtering matrices, we express this condition as:

$$\Delta\mathbf{H} \triangleq \widehat{\mathbf{H}} - \mathbf{H}, \quad \|\Delta\mathbf{H}\|_2 \leq \epsilon$$

where ϵ is a small positive number. The estimation errors in the input and noise SOS can be expressed in an analogous manner, that is

$$\Delta\mathbf{R}_{xx} \triangleq \widehat{\mathbf{R}}_{xx} - \mathbf{R}_{xx}, \quad \|\Delta\mathbf{R}_{xx}\|_2 \leq \epsilon$$

$$\Delta\mathbf{R}_{nn} \triangleq \widehat{\mathbf{R}}_{nn} - \mathbf{R}_{nn}, \quad \|\Delta\mathbf{R}_{nn}\|_2 \leq \epsilon.$$

Different quantities may be known with different accuracies, i.e., the ϵ's in the above expressions may be different. In this case, ϵ is the biggest of these values.

3.2. MMSE-DFE: Perturbation analysis

Under mismatch, efforts toward computation of $\mathbf{R}_{yy}$, $\mathbf{R}_{yx}$ and $\mathbf{R}_{xy}$ lead to:

$$\widehat{\mathbf{R}}_{yy} \triangleq \widehat{\mathbf{H}}\widehat{\mathbf{R}}_{xx}\widehat{\mathbf{H}}^H + \widehat{\mathbf{R}}_{nn}, \; \widehat{\mathbf{R}}_{xy} \triangleq \widehat{\mathbf{R}}_{xx}\widehat{\mathbf{H}}^H = \widehat{\mathbf{R}}_{yx}^H.$$

Then, efforts toward computing $\mathbf{R}$ and $\mathbf{R}_\Delta$ give:

$$\widehat{\mathbf{R}} \triangleq \widehat{\mathbf{R}}_{xx} - \widehat{\mathbf{R}}_{xy}\widehat{\mathbf{R}}_{yy}^{-1}\widehat{\mathbf{R}}_{yx}$$

$$\widehat{\mathbf{R}}_\Delta \triangleq \left[\mathbf{0}_{(N_b+1)\times\Delta} \; \mathbf{I}_{N_b+1} \; \mathbf{0}_{(N_b+1)\times s}\right] \widehat{\mathbf{R}} \begin{bmatrix} \mathbf{0}_{\Delta\times(N_b+1)} \\ \mathbf{I}_{N_b+1} \\ \mathbf{0}_{s\times(N_b+1)} \end{bmatrix}.$$

The resulting "optimal" filters are given by

$$\widehat{\mathbf{b}}_o = \frac{\widehat{\mathbf{R}}_\Delta^{-1}\,\mathbf{e}_0}{\mathbf{e}_0^H\,\widehat{\mathbf{R}}_\Delta^{-1}\,\mathbf{e}_0}, \quad \widehat{\mathbf{w}}_o = \widehat{\mathbf{R}}_{yy}^{-1}\,\widehat{\mathbf{R}}_{yx}\,\widetilde{\widehat{\mathbf{b}}}_o \qquad (5)$$

where $\widetilde{\widehat{\mathbf{b}}}_o$ is the appropriately zero-padded version of $\widehat{\mathbf{b}}_o$ (recall the definition of $\widetilde{\mathbf{b}}$ in terms of $\mathbf{b}$).

Assuming correct past decisions, the corresponding MSE can be expressed as

$$\begin{aligned}\widehat{\mathbf{MMSE}} &\triangleq \mathcal{E}\left[\left(\widetilde{\widehat{\mathbf{b}}}_o^H\mathbf{x} - \widehat{\mathbf{w}}_o^H\mathbf{y}\right)\left(\mathbf{x}^H\widetilde{\widehat{\mathbf{b}}}_o - \mathbf{y}^H\widehat{\mathbf{w}}_o\right)\right] \\ &= \widetilde{\widehat{\mathbf{b}}}_o^H\,\mathbf{R}_{xx}\widetilde{\widehat{\mathbf{b}}}_o - \widetilde{\widehat{\mathbf{b}}}_o^H\,\mathbf{R}_{xy}\widehat{\mathbf{w}}_o - \widehat{\mathbf{w}}_o^H\mathbf{R}_{yx}\,\widetilde{\widehat{\mathbf{b}}}_o \\ &\qquad + \widehat{\mathbf{w}}_o^H\mathbf{R}_{yy}\widehat{\mathbf{w}}_o.\end{aligned}$$

Substituting into the above equation the expression for $\widehat{\mathbf{w}}_o$ given in (5), we obtain

$$\widehat{\mathbf{MMSE}} = \widetilde{\widehat{\mathbf{b}}}_o^H\,\mathcal{R}\,\widetilde{\widehat{\mathbf{b}}}_o \qquad (6)$$

where

$$\begin{aligned}\mathcal{R} \triangleq\; &\mathbf{R}_{xx} - \mathbf{R}_{xy}\widehat{\mathbf{R}}_{yy}^{-1}\widehat{\mathbf{R}}_{yx} - \widehat{\mathbf{R}}_{xy}\widehat{\mathbf{R}}_{yy}^{-1}\mathbf{R}_{yx} \\ &+\widehat{\mathbf{R}}_{xy}\widehat{\mathbf{R}}_{yy}^{-1}\mathbf{R}_{yy}\widehat{\mathbf{R}}_{yy}^{-1}\widehat{\mathbf{R}}_{yx}. \qquad (7)\end{aligned}$$

Our aim is to assess the excess MSE, $\widehat{\mathbf{MMSE}}-\mathbf{MMSE}$, introduced by the channel and SOS estimation errors. To that end, we first relate $\mathcal{R}$ and $\mathbf{R}$.

Theorem 1: *Matrices $\mathcal{R}$ and $\mathbf{R}$ satisfy*

$$\mathcal{R} = \mathbf{R} + O(\epsilon^2). \qquad (8)$$

The proof, which is can be constructed by using first-order perturbation expansions, can be found in [3].

We can now relate $\widehat{\mathbf{MMSE}}$ and $\mathbf{MMSE}$.

Theorem 2: *Quantities $\widehat{\mathbf{MMSE}}$ and $\mathbf{MMSE}$ satisfy*

$$\widehat{\mathbf{MMSE}} = \mathbf{MMSE} + O(\epsilon^2).$$

Proof: We define $\Delta\mathbf{b}_o \triangleq \widehat{\mathbf{b}}_o - \mathbf{b}_o$. Using (6), (8) and (4) and ignoring higher-order error terms, we obtain

$$\begin{aligned}\widehat{\mathbf{MMSE}} &= \widetilde{\widehat{\mathbf{b}}}_o^H\,\mathcal{R}\,\widetilde{\widehat{\mathbf{b}}}_o = \widetilde{\widehat{\mathbf{b}}}_o^H\,\mathbf{R}\,\widetilde{\widehat{\mathbf{b}}}_o = \widehat{\mathbf{b}}_o^H\mathbf{R}_\Delta\widehat{\mathbf{b}}_o \\ &= \left(\mathbf{b}_o^H + \Delta\mathbf{b}_o^H\right)\mathbf{R}_\Delta\left(\mathbf{b}_o + \Delta\mathbf{b}_o\right) \\ &= \mathbf{MMSE} + \left(\Delta\mathbf{b}_o^H\mathbf{R}_\Delta\mathbf{b}_o + \mathbf{b}_o^H\mathbf{R}_\Delta\Delta\mathbf{b}_o\right).\end{aligned}$$

From the definition (3) of $\mathbf{b}_o$, we obtain that the vector $\mathbf{R}_\Delta\mathbf{b}_o$ is a multiple of $\mathbf{e}_0$. By construction, the first element of $\Delta\mathbf{b}_o$ is identically zero, since the first element of both $\mathbf{b}_o$ and $\widehat{\mathbf{b}}_o$ is 1. Thus, the terms inside the parenthesis vanish, to prove theorem 2. □

Theorem 2 says that the MMSE-DFE is very robust with respect to small channel and SOS mismatch.

4. SIMULATIONS

In our simulations, we use the communication channel whose impulse response is plotted in Fig. 3. It models a multipath scenario, and is derived by oversampling, by a factor of 2, the continuous-time channel impulse response $h(t) = p(t) - 0.4\,p(t-0.7)$, where $p(t)$ is the truncated raised-cosine pulse, with roll-off factor $\beta = 0.22$. The truncation interval is $[-3T,\ 3T]$, where T denotes the symbol period, and the sampling instants are the integer multiples of $T/2$. The input is a BPSK signal, taking, with equal probability, the values ± 1, yielding $\mathbf{R}_{xx} = \mathbf{I}$. At the multi-channel output, we add temporally and spatially white noise with variance σ_n^2. Hence, $\mathbf{R}_{nn} = \sigma_n^2\mathbf{I}$. We define the SNR as:

$$\mathrm{SNR} \triangleq 10\log_{10}\frac{\mathcal{E}[\|\mathbf{w}_n\|_2^2]}{\mathcal{E}[\|\mathbf{n}_n\|_2^2]}$$

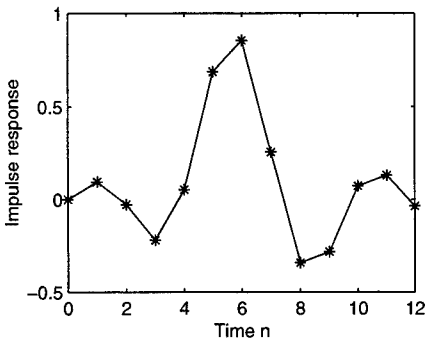


Figure 3: Fractionally sampled channel impulse response.

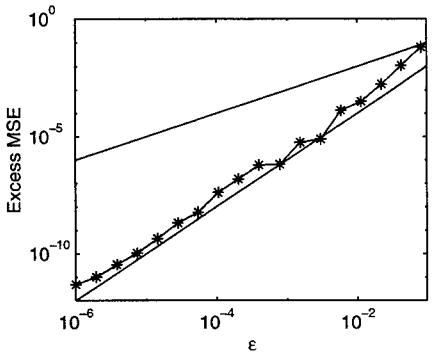


Figure 4: Excess MSE ('*-') versus ϵ, for SNR=12 dB; upper and lower lines: functions ϵ and ϵ^2, respectively.

where $\mathbf{w}_n$ is the noiseless channel output at time n. Using the knowledge of the channel impulse response and the input and noise SOS, we compute the optimal filters $\mathbf{b}_o$ and $\mathbf{w}_o$, as well as the **MMSE**, for $N_f = 8$, $N_b = 4$ and all possible delays. In order to relate the range of ϵ, for which our first-order analysis remains valid, to the size of the unperturbed quantities, we perturb $\{\mathbf{h}_i\}_{i=0}^{6}$, $\mathbf{R}_{xx}$ and $\mathbf{R}_{nn}$ using random perturbations, such that:

$$\max\left(\|\Delta\mathbf{H}\|_2, \|\Delta\mathbf{R}_{xx}\|_2, \|\Delta\mathbf{R}_{nn}\|_2\right) = \epsilon.$$

The sizes of the unperturbed quantities are: $\|\mathbf{H}\|_2 = 1.4580$, $\|\mathbf{R}_{xx}\|_2 = 1$ and $\|\mathbf{R}_{nn}\|_2 = \sigma_n^2$. Using inaccurate data, we compute $\widehat{\mathbf{b}}_o$, $\widehat{\mathbf{w}}_o$ and $\widehat{\mathbf{MMSE}}$. In Fig. 4, we plot the excess MSE for (a typical) delay $\Delta = 3$ and SNR=12 dB ($\sigma_n^2 \approx 0.049$), versus ϵ. In the same figure, we plot the functions ϵ and ϵ^2 (upper and lower line, respectively). We observe that for $\epsilon \in [0, \epsilon^*)$, with $\epsilon^* \approx .0546$, the excess MSE shows a quadratic dependence on ϵ. That is, in this range, our first-order analysis is valid and the excess MSE is remarkably small.

In Fig. 5, we depict the corresponding plots for SNR=25 dB ($\sigma_n^2 \approx 0.0025$). It is clear that, now, our analysis is valid for a smaller range of ϵ, i.e., $\epsilon \in [0, \epsilon^*)$, with $\epsilon^* \approx 0.0026$. We observe that, in both cases, our

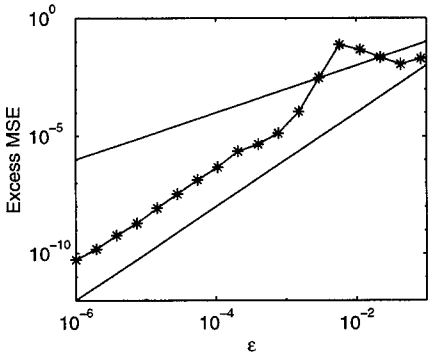


Figure 5: Excess MSE ('*-') versus ϵ, for SNR=25 dB; upper and lower lines: functions ϵ and ϵ^2, respectively.

analysis remains valid for $\epsilon \in [0, \epsilon^*)$, with $\epsilon^* \approx \sigma_n^2$. Thus, for fixed $\mathbf{H}$ and $\mathbf{R}_{xx}$, the range of ϵ for which our analysis remains valid decreases for increasing the SNR.

In order to isolate the effects of the estimation errors in each one of the quantites of interest, i.e., $\{\mathbf{h}_i\}_{i=0}^{6}$, $\mathbf{R}_{xx}$ and $\mathbf{R}_{nn}$, we performed experiments where we perturbed only one quantity at a time [3]. We observed that for $\epsilon \in [0, \epsilon^*)$, with $\epsilon^* \approx 0.3$, the MMSE-DFE was insensitive to estimation errors in $\mathbf{H}$ and $\mathbf{R}_{xx}$, because the excess MSE induced by size-ϵ perturbations on these quantities was very close to, and in many cases smaller than, ϵ^2. We observed that this happened irrespective of the SNR. On the other hand, the MMSE-DFE was more sensitive to errors occuring in $\mathbf{R}_{nn}$, especially at high SNR. This was to be expected since for fixed $\|\mathbf{H}\|_2$ and $\|\mathbf{R}_{xx}\|_2$, $\|\mathbf{R}_{nn}\|_2$ becomes smaller for increasing the SNR. Hence, the size of the perturbations $\|\Delta\mathbf{R}_{nn}\|_2$ tolerated by a first-order analysis decreases as well.

References

[1] N. Al-Dhahir and J. M. Cioffi, "MMSE Decision-Feedback equalizers: Finite-length results," *IEEE Trans. Information Theory*, vol. 41, no. 4, pp. 961–975, July 1995.

[2] N. Al-Dhahir and J. M. Cioffi, "Mismatched finite-complexity MMSE desicion feedback equalizers," *IEEE Trans. Signal Processing*, vol. 45, no. 4, pp. 935–944, April 1997.

[3] A. P. Liavas, "On the robustness of the finite-length MMSE-DFE with respect to channel and second order statistics estimation errors," submitted to the IEEE Trans. Signal Processing, February 2001.

Performance Evaluation of Blind Channel Estimation using a Frequency Domain Base-band Communication Model

Saman S. Abeysekera & Patrick K. S. Ong

School of Electrical and Electronic Engineering,
Nanyang Technological University, Nanyang Avenue, SINGAPORE 639798.
E-mail: esabeysekera@ntu.edu.sg

Abstract: *A frequency domain approach for evaluating estimation bounds (Cramer-Rao bounds) of a base-band communication channel model parameters is presented. It is assumed that the estimation process uses 2nd order statistics of an over-sampled signal with the exploitation of signal cyclo-stationary properties. The described frequency domain approach provides useful insight into the channel estimation problem and is independent of the adopted parameter estimation technique. Furthermore, the proposed frequency domain technique does not depend on the noise probability density function and could be easily extended for evaluating the estimation performance in the presence of colored noise.*

1. Introduction

The inter-symbol interference resulting due to the presence of multi-path in digital communication channels needs to be equalized for successful data transfer. Classically, equalization is achieved using a training sequence but this has the drawback of reducing the data rate. The increasing interest in digital mobile communication and digital broadcasting require high data rates and thus blind equalization techniques are preferred over the use of training sequences. Blind channel equalization could be achieved via adaptive algorithms, however, they result in very slow convergence rates. In overcoming these difficulties, fast blind channel estimation and equalization has been proposed which are obtained using the second-order statistics of the process [1][2].

Over sampling the signal above the Nyquist rate and using cyclo-stationary properties of the process, various second-order statistics based channel estimation algorithms have been described in literature [4]. Attempts have also been made to obtain the bounds of such estimation methods. Most of these bounds are obtained using time domain techniques that are tedious and also depend on the method used for estimation. This paper proposes to evaluate the estimation bounds using a frequency domain approach and describes the procedure in detail. A simulation example is shown to justify the assumptions used in deriving the bounds.

2. Channel Model

Consider a baseband communication system with the following channel model.

$$x(t)=\sum_{\alpha=-\infty}^{\infty}h(\alpha)u(t-\alpha)+n(t)\ , \qquad (1)$$

where t takes discrete values and it is assumed that the sampling rate is normalized to 1; $h(\cdot)$ is the channel impulse response; and $n(t)$ is an additive noise process. In equation (1), the information symbols, s_k, that are transmitted at T intervals are given by

$$u(t)=\sum_{k=-\infty}^{\infty}s_k\delta(t-kT)\ . \qquad (2)$$

Equation (1) can be equivalently expressed in the frequency domain as

$$X(\omega)=H(\omega)U(\omega)+N(\omega)\ . \qquad (3)$$

In the following discussion we will demonstrate how the channel frequency response $H(\omega)$ can be estimated using $X(\omega)$.

Assume that $X(\omega)$ is estimated from the received signal $x(t)$ via the use of DFT of length L (selected as an integer multiple of T). That is $X(\omega)$ is estimated at L distinct frequency points in the frequency interval of $(0-2\pi)$ given as

$$X(k)=X(\omega)\big|_{\omega=2\pi k/L}=H(k)U(k)+N(k) \qquad 0\le k\le L-1 \quad (4)$$

We will now examine some correlation properties of $X(k)$. In order to do this consider the correlation properties of $U(k)$.

3. Correlation Properties of $U(k)$

$U(k)$ is given by

$$U(k)=\sum_{n=0}^{L-1}u(n)e^{-jnk\,2\pi/L}\ . \qquad (5)$$

Consider the product $\left[U(k_1)U^*(k_2)\right]$, and evaluate its statistical mean as,

$$E\left[U(k_1)U^*(k_2)\right]=\sum_{n=0}^{L-1}\sum_{m=0}^{L-1}E\left[u(n)u^*(m)\right]e^{-j(nk_1-mk_2)2\pi/L} \qquad (6)$$

By the substitution for $u(n)$ from equation (2) and using cyclo-stationary properties we get

$$E\left[u(n)u^*(m)\right]=E\left[s_n s_m^*\right]\sum_{k=-\infty}^{\infty}\delta(t-kT)=\delta(n-m)\sum_{k=-\infty}^{\infty}\delta(t-kT) \quad (7)$$

Substituting this in equation (6) we get

$$E\left[U(k_1)U^*(k_2)\right]=\sum_{n=0}^{L-1}\sum_{k=-\infty}^{\infty}\delta(n-kT)e^{-jn(k_1-k_2)2\pi/L}\ . \qquad (8)$$

Equation (8) can be simplified to obtain

$$E\left[U(k_1)U^*(k_2)\right]=\sum_{n=0}^{1/T-1}e^{-jn(k_1-k_2)2\pi T/L}\ . \qquad (9)$$

Note that the right hand side of equation (9) approaches to zero, asymptotically, if (k_1-k_2) is not an integer multiple of $1/T$. Therefore, in the limit $L\to\infty$,

$$E\left[U(k_1)U^*(k_2)\right]=\begin{cases}\frac{L}{T} & (k_1-k_2)=\frac{\varepsilon L}{T}\\ 0 & other-wise\end{cases} \quad (10)$$

where ε denotes an integer. Now, choosing $k_2=k_1+\varepsilon_1 L/T$ and $k_4=k_3+\varepsilon_2 L/T$, the following expression for covariance of $\left[U(k_1)U^*(k_2)\right]$ can also be obtained (see [5] for details).

$$Covariance\left[U(k_1)U^*(k_2),U(k_3)U^*(k_4)\right]$$

$$=\frac{1}{T^2}\left[\delta(k_1-k_3-\varepsilon' L/T)+\delta(k_1+k_3-\varepsilon'' L/T)\right]. \quad (11)$$

Again ε' and ε'' in equation (11) denote integers.

Assuming $n(t)$ of equation 1 as an independent white Gaussian noise sequence of variance σ^2, the following relations can be obtained for the noise spectrum $N(k)$ [5].

$$E\left[N(k_1)N^*(k_2)\right]=\sigma^2\delta(k_1-k_2) \quad (12)$$

$$E\left[N(k_1)N^*(k_2)N^*(k_3)N(k_4)\right]= \quad (13)$$
$$\sigma^4\left[\begin{matrix}\delta(k_1-k_2)\delta(k_3-k_4)+\\ \delta(k_1-k_3)\delta(k_2-k_4)+\delta(k_1)\delta(k_2)\delta(k_3)\delta(k_4)\end{matrix}\right]$$

4. Channel Estimation using the Data Covariance Matrix

Now consider the input data covariance matrix $R_x(t,\tau)$ which consists of elements $r_x(t,\tau)$ that are defined as,

$$r_x(t,\tau)=E\left[x(t)x^*(t-\tau)\right]. \quad (14)$$

(Note that above data covariance matrix is used for channel identification using the subspace method of channel estimation [2].)

Performing a 2-d Discrete Fourier Transform (DFT) on $R_x(t,\tau)$ we obtain

$$\Gamma_x^k(\nu)\Leftrightarrow\underset{\tau\to\nu \quad t\to k}{2d-DFT}\left\{R_x(t,\tau)\right\}. \quad (15)$$

It is noted here that we can estimate the performance of channel estimation, e.g. Cramer-Rao bounds (CRBs) of estimates, by using either $r_x(t,\tau)$ or $\Gamma_x^k(\nu)$. Both methods would result in identical measures as the transformation in equation (15) is one to one. It is further noted here that in all reported literature the CRBs are obtained in the time domain using $R_x(t,\tau)$. In the following we propose to estimate the CRBs of channel estimation in the frequency domain using $\Gamma_x^k(\nu)$.

To do this we first note that $\Gamma_x^k(\nu)$ can be expressed as [1]

$$\Gamma_x^k(\nu)=X(\nu+k2\pi/T)X^*(\nu), \quad 0\le k\le T-1 \quad (16)$$

Suppose $\Gamma_x^k(\nu)$ is evaluated using L data samples at $2\pi/L$ frequency points, i.e. at $\nu=m2\pi/L$ for $0\le m\le L$, we get,

$$\Gamma_x^k(m)=X(m2\pi/L+k2\pi/T)X^*(m2\pi/L), \quad (17)$$

Now using the results of previous section, i.e. equations (3) and (10) - (13), the following can be obtained.

$$E\left[\Gamma_x^k(m)\right]=H((m+k\,L/T)2\pi/L)H^*(m2\pi/L)+T\sigma^2\delta(k) \quad (18)$$

and

$$Covariance\left[\Gamma_x^{k_1}(m_1),\Gamma_x^{k_2\,*}(m_2)\right]=$$
$$H((m_1+k_1\,L/T)2\pi/L)H^*(m_1\,2\pi/L)\times$$
$$H((m_2+k_2\,L/T)2\pi/L)H^*(m_2\,2\pi/L)\times$$
$$\frac{\left[\delta(m_1-m_2-\varepsilon' L/T)+\delta(m_1+m_2-\varepsilon'' L/T)\right]}{M}, \quad (19)$$
$$+\frac{T^2\sigma^4}{M}\delta(k_1-k_2)\delta(m_1-m_2)\times$$
$$[1+\delta(k_1)\delta(m_1)+\delta(k_1)\delta(m_1-L)]$$

where σ^2 is the power of the additive white Gaussian noise process. Note that in above equations (18) and (19) it is assumed that M data segments have been used in the estimation of $R_x(t,\tau)$.

5. Probability Distribution of $\Gamma_x^k(m)$.

We have presented the expectation and covariance properties of $\Gamma_x^k(m)$ in equation (18) and (19), respectively. In this section we will investigate the probability distribution of $\Gamma_x^k(m)$. Typically, the number of data segments M used in the estimation of $R_x(t,\tau)$ is large. Hence we can assume that the probability distribution of $\Gamma_x^k(m)$ obey a complex Gaussian distribution. The following simulation results justify this claim. The simulation results of Figure 1 are obtained using the channel model discussed in reference [2]. A binary independent symbol sequence (±1) (BPSK) was used in the simulations. (The number of virtual channels T = 4; the width of DFT temporal window $L=10\times T$; the degree of ISI = 4; the number of data symbols used = 500; SNR= 30dB.) The probability distribution of $\Gamma_x^k(m)$ (at $k=1,m=11$) was obtained using 10,000 iterations. It can be seen from Figure 1 that both real and imaginary parts of $\Gamma_x^k(m)$ obey a Gaussian distribution.

6. Cramer-Rao Bounds for Channel Estimation

For the derivation of Cramer-Rao bound (CRB) for the channel estimation, consider $\Gamma_x^k(m)$ as the complex observation vector,

$$y=\Gamma_x^k(m) \qquad 0\le m\le L-1, \quad 0\le k\le T-1, \quad (20)$$

and the parameters need to be estimated as $\theta_m = [\mathrm{Re}\{H(m2\pi/L)\}, \mathrm{Im}\{H(m2\pi/L)\}]$, $0 \le m \le L-1$.

As $\Gamma_x^k(m)$ obeys a complex Gaussian distribution we can now obtain the Fisher Information matrix for the estimation as [3],

$$[I(\theta)]_{ij} = trace\left\{C_{yy}^{-1}(\theta)\frac{\partial C_{yy}(\theta)}{\partial \theta_i}C_{yy}^{-1}(\theta)\frac{\partial C_{yy}(\theta)}{\partial \theta_j}\right\} + 2\,\mathrm{Re}\left\{\frac{\partial m_y^H(\theta)}{\partial \theta_i}C_{yy}^{-1}(\theta)\frac{\partial m_y(\theta)}{\partial \theta_j}\right\}, \quad (21)$$

where $m_y(\theta)$ and $C_{yy}(\theta)$ are the expected value and covariance of $\Gamma_x^k(m)$ as described in equation (18) and (19), respectively. $[I(\theta)]^{-1}$ then provides the required CRB for channel estimation.

For example, consider the previously noted channel model of reference [2]. For the quaternary-QAM signal format with additive white Gaussian noise (SNR=25 dB), using equation (21) the channel parameter estimation bounds could be obtained. The final estimation bound can be calculated as

$$H_b = \frac{1}{L}\sum_{m=0}^{L-1}|b_m|^2, \quad (22)$$

where b_m is the CR bound of parameter θ_m, which was obtained from the diagonal elements of $[I(\theta)]^{-1}$. The definition in equation (22) is appropriate because the parameters associated with bounds b_m, i.e. θ_m is related to the variance of the channel impulse response as $\sum_{m=0}^{L-1}|\theta_m|^2$.

7. Discussion

A frequency domain approach for evaluating estimation bounds (Cramer-Rao bounds) of a base-band communication channel model parameters is presented in the paper. The frequency domain approach of evaluating the CRB is the novel part of the reported work. Usually CRB is estimated in the time domain using correlation matrix $R_x(t,\tau)$. The CRB evaluated from either $R_x(t,\tau)$ or $\Gamma_x^k(m)$ should produce identical results, as the latter is the 2-d Fourier transform of the former. Moreover, the described frequency domain approach provides useful insight into the channel estimation problem and also is independent of the adopted parameter estimation technique. The proposed frequency domain technique does not depend on the noise probability density function. Furthermore, the CRB derivation technique could be easily extended for evaluating the estimation performance in the presence of colored noise, by modifying equations (12) and (13).

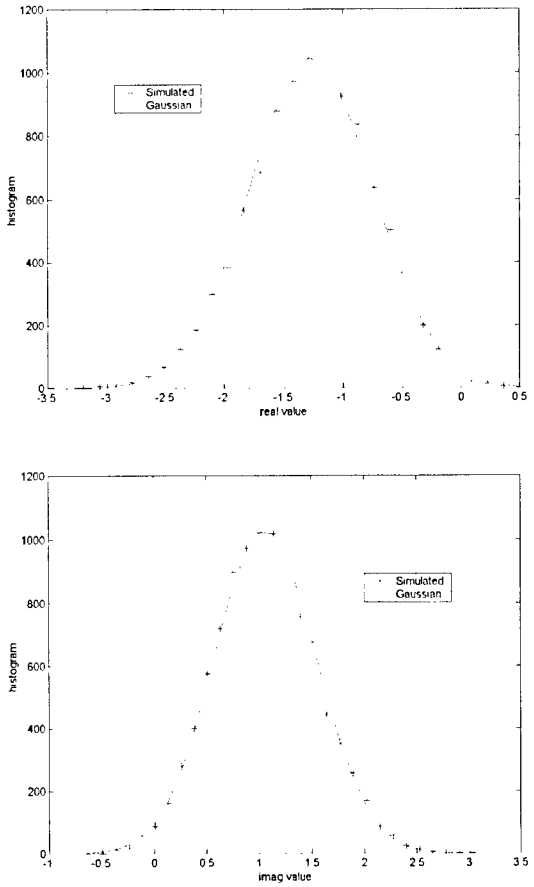


Figure 1: A Comparison of the probability distribution of $\Gamma_x^k(m)$ $(k=1, m=11)$ obtained via simulations with a Gaussian distribution. Simulations for both real and imaginary parts of $\Gamma_x^k(m)$ are shown.

8. REFERENCES

[1] L. Tong, G. Xu and T. Kailath , "A new approach to blind identification and equalization of multipath channels", Proceedings of the 25th Asilomar Conference, Pacific Grove, California, U.S.A., pp. 856-860, 1991.

[2] E. Moulines, P. Duhamel, J-F Cordoso and S. Mayrague, "Subspace Method for the Blind Identification of Mutichannel FIR Filters", IEEE Transactions on Signal Processing, SP-43, pp. 516-525, 1995.

[3] S. M. Kay, Fundamentals of Statistical Signal Processing and Estimation Theory, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[4] L. Tong and S. Perreau, "Multichannel Blind Identification: From Subspace to Maximum Likelihood Methods", Proceedings of the IEEE, vol. 86, pp. 1951-1968, 1998.

[5] B. Porat, *Digital Processing of Random Signals: Theory and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

JOINT CHANNEL ESTIMATION AND DECODING OF SPACE-TIME TRELLIS CODES

Jianqiu Zhang and Petar M. Djurić

Department of Electrical and Computer Engineering
State University of New York at Stony Brook
Stony Brook, NY, 11794-2350
e-mail: {jizzhang, djuric }@ece.sunysb.edu

ABSTRACT

In this paper, we discuss the application of sequential importance sampling (SIS) to joint channel estimation and decoding of space-time trellis codes (STTCs). First, we present the dynamic state space model (DSSM) of the system, and then we briefly review the theory of SIS. A special case of SIS, a combination of SIS and Kalman filtering, is shown through simulations to be a viable approach to the problem which addresses time-varying flat-fading environments. Our solution is admissible if phase ambiguity can be avoided. We show that the phase ambiguity can be reduced by carefully designing the STTC modulation constellations.

1. INTRODUCTION

Space-time coding (STC) introduced by Tarokh et al. [1] exploits spatial and temporal diversity and thus provides a framework for increased data rates in wireless communications. Among families of space time codes, STTCs have more advantages than ST block codes as pointed out in [1]. It is generally assumed that STC will be used in fading environments. Therefore, while decoding, it is necessary to estimate the channel state information (CSI), i.e., the fading coefficients of the channel. Most of the time in the literature, it is assumed that the CSI is available through sending pilot signals periodically from the transmit to the receive side. Here we consider the problem of joint estimation of the CSI and decoding of STTC when pilot signals are not available. Because the problem is highly nonlinear, it is hard to apply for its resolution conventional methods such as the extended Kalman filter. Recently in [2] and [4], SIS has been used to address this type of communication problems. In [2], a scheme was developed for the joint detection and decoding of convolutional codes via combination of SIS and Kalman filtering. In this paper, we generalize their scheme and apply it to STTC decoding when the channel is assumed to be flat-fading, time-varying, and is modeled by auto-regressive moving-average (ARMA) processes. We propose a quadruple 8-PSK constellation to a delay diversity STTC that greatly reduces the effect of phase ambiguity on the symbol error rate (SER).

This work was supported by the National Science Foundation under Awards CCR-9903120 and CCR-0082607.

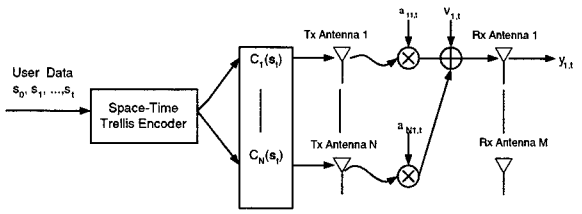


Figure 1: Space-Time Trellis Code System

2. SYSTEM DESCRIPTION

Suppose a communication system employs N transmit antennas and M receive antennas as in Figure 1. A sequence of user data symbols, $s_0, ..., s_t$, where $s_t \in \mathcal{A}$ and $\mathcal{A}$ is the set of all possible user data symbols, is put through a space-time trellis encoder. The new user state vector of the space-time trellis encoder at time t is determined according to the state transition equation,

$$\mathbf{s}_t = Z(\mathbf{s}_{t-1}, s_t) \tag{1}$$

where $\mathbf{s}_{t-1}$ is the previous user state and s_t is the new user symbol. Based on the current user state, the encoder generates a code vector that consists of N symbols, $\mathbf{c}(\mathbf{s}_t)^T = [c_1(\mathbf{s}_t) \ ... \ c_N(\mathbf{s}_t)]$, to be transmitted by antennas where $c_i(\cdot)$ denotes the modulation function for the ith antenna.

Let $\alpha_{nm,t}$ be the fading coefficient from the nth transmit antenna to the mth receive antenna at time t. The fading coeffcient can be modeled as an ARMA process that matches the power spectral density of the channel. An ARMA (r_1, r_2) process can be represented as

$$\begin{aligned}\alpha_{nm,t} + \phi_1 \alpha_{nm,t-1} \cdots \phi_{r_1} \alpha_{nm,t-r_1} \\ = \rho_0 u_{nm,t} + \cdots + \rho_{r_2} u_{nm,t-r_2}\end{aligned} \tag{2}$$

where $u_{nm,t}$ is an i.i.d. random complex Gaussian process that drives the ARMA process, and $\{\phi_i\}$ and $\{\rho_i\}$ are known AR and MA coefficients. We assume that all channel coefficients have the same power spectral density and therefore their AR and MA coefficients are identical.

The next section shows that a DSSM representation of the STTC system is convenient for the derivation of the SIS algorithm. For convenience, we assume $r_1 = r_2 = r$; otherwise zeros can be padded to the coefficients to make the orders equal. Also, we introduce the channel state vector $\mathbf{h}_{nm,t}^T = [h_{nm,t} \cdots h_{nm,t-r}]$, which is of dimension $(r+1)$. Then a state transition equation can be constructed according to

$$\mathbf{h}_{nm,t} = \mathbf{F}_0 \mathbf{h}_{nm,t-1} + \mathbf{g} u_{nm,t} \tag{3}$$

where

$$\mathbf{F}_0 = \begin{bmatrix} -\phi_1 & -\phi_2 & \cdots & -\phi_r & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{g} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The fading coefficient is then represented as

$$\alpha_{nm,t} = \mathbf{o}^T \mathbf{h}_{nm,t}$$

where $\mathbf{o}^T = [\rho_0 \ \rho_1 \ \cdots \ \rho_r]$.

Now we arrange all the fading coefficients at time t into a single $NM \times 1$ vector, $\boldsymbol{\alpha}_t = [\alpha_{11,t} \cdots \alpha_{N1,t} \cdots \alpha_{1M,t} \cdots \alpha_{NM,t}]^T$, and define the $NM(r+1) \times 1$ channel state vector as $\mathbf{h}_t = [\mathbf{h}_{11,t} \cdots \mathbf{h}_{N1,t} \cdots \mathbf{h}_{1M,t} \cdots \mathbf{h}_{NM,t}]^T$. Then we can express all the fading coefficients in a compact form, i.e.,

$$\mathbf{h}_t = \mathbf{F}\mathbf{h}_{t-1} + \mathbf{G}\mathbf{u}_t \tag{4}$$

$$\boldsymbol{\alpha}_t = \mathbf{O}\mathbf{h}_t \tag{5}$$

where $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{O}$ are obtained from $\mathbf{F}_0$, $\mathbf{g}$, and $\mathbf{o}$, respectively. For example,

$$\mathbf{O}^T = \begin{bmatrix} \mathbf{o} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{o} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{o} \end{bmatrix}$$

where $\mathbf{0}$ is an $(r+1) \times 1$ all zero vector, and the matrix $\mathbf{O}$ is of dimension $NM(r+1) \times NM(r+1)$.

With ideal time and frequency information and flat fading, the received signals are simply the product of fading coefficients and users signals embedded in noise. At time t we can write,

$$\begin{aligned} \mathbf{y}_t &= \mathbf{C}(\mathbf{s}_t)\boldsymbol{\alpha}_t + \mathbf{v}_t \\ &= \mathbf{C}(\mathbf{s}_t)\mathbf{O}\mathbf{h}_t + \mathbf{v}_t \end{aligned} \tag{6}$$

where $\mathbf{y}_t = [y_{1,t} \cdots y_{M,t}]^T$ is the received signal vector at all M receive antennas, and $\mathbf{v}_t = [v_{1,t} \cdots v_{M,t}]^T$ is the observation noise vector. The code matrix is an $M \times NM$ matrix constructed from the code vector $\mathbf{c}(\mathbf{s}_t)$, or

$$\mathbf{C}(\mathbf{s}_t)^T = \begin{bmatrix} \mathbf{c}(\mathbf{s}_t) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{c}(\mathbf{s}_t) & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{c}(\mathbf{s}_t) \end{bmatrix}$$

where $\mathbf{0}$ is an $N \times 1$ all zero vector. The state transition equations (1) and (4) and the observation equation (6) together form the DSSM representation of the STTC system. We can see that the DSSM describes both the channel state vector $\mathbf{h}_t$ and the user state vector $\mathbf{s}_t$ as Markov processes hidden in the observation $\mathbf{y}_t$.

3. COMBINED SIS AND KALMAN FILTERING

In this section, we review the theory of combined SIS and Kalman filtering, which we propose to apply to the STTC system. Define $\mathbf{s}_{0:t} = \{\mathbf{s}_0, ..., \mathbf{s}_t\}$ as the set of all user state vectors up to time t, and let $\mathbf{y}_{0:t}$ be defined similarly. In the Bayesian framework, all the information about the user data symbols is contained in the posterior density, $p(\mathbf{s}_{0:t}|\mathbf{y}_{0:t})$. The evaluation of the expected value of a function of the user states $\xi(\mathbf{s}_{0:t})$, using the posterior density, involves high dimensional integration and is almost impossible to carry out analytically. However, if we have samples from the posterior density, $\mathbf{s}_{0:t}^{(j)}$, where $j = 1, 2, ..., J$ is the sample index, we can approximate the expectation using Monte Carlo integration

$$E[\xi(\mathbf{s}_{0:t}|\mathbf{y}_{0:t})] \cong \frac{1}{W_t} \sum_{j=1}^{J} \xi(\mathbf{s}_{0:t}^{(j)}) w_t^{(j)} \tag{7}$$

where $w_t^{(j)}$ is the weight associated with the jth sample and $W_t = \sum_{j=1}^{J} w_t^{(j)}$ is the sum of the weights. Most of the time, however, taking samples from $p(\mathbf{s}_{0:t}|\mathbf{y}_{0:t})$ is a difficult task itself, and we have to resort to importance sampling, i.e., drawing samples from an importance function $\pi(\mathbf{s}_{0:t}|\mathbf{y}_{0:t})$ that may render the task easier. Then, samples drawn from the importance function are weighted according to

$$w_t^{(j)} = \frac{p(\mathbf{s}_{0:t}^{(j)}|\mathbf{y}_{0:t})}{\pi(\mathbf{s}_{0:t}^{(j)}|\mathbf{y}_{0:t})}. \tag{8}$$

However, even the direct importance sampling from the distribution $\pi(\mathbf{s}_{0:t}|\mathbf{y}_{0:t})$ is difficult. Fortunately, the posterior density function can be factored, that is,

$$p(\mathbf{s}_{0:t}|\mathbf{y}_{0:t}) \propto p(\mathbf{s}_{0:t-1}|\mathbf{y}_{0:t-1}) \times p(\mathbf{y}_t|\mathbf{s}_{t-1}) \tag{9}$$

and this provides us with a possibility to evaluate the posterior recursively. Indeed, if we select an importance function in the form of

$$\pi(\mathbf{s}_{0:t}|\mathbf{y}_{0:t}) = \pi(\mathbf{s}_{0:t-1}|\mathbf{y}_{0:t-1})\pi(\mathbf{s}_t) \tag{10}$$

as new observations become available, we can evaluate the importance weights recursively.

The SIS algorithm can be implemented according to the following scheme:

For $t = 0, 1, 2, ...$

1. Sample $\mathbf{s}_t^{(j)} \sim \pi(\mathbf{s}_t)$ and set $\mathbf{s}_{0:t}^{(j)} = (\mathbf{s}_{0:t-1}^{(j)}, \mathbf{s}_t^{(j)})$, where $j = 1, ..., J$.

2. Evaluate the importance weights according to (8). It can be shown that when the importance function is in the form of (10), the weights can be obtained recursively from

$$w_t^{(j)} = w_{t-1}^{(j)} \frac{p(\mathbf{y}_t|\mathbf{s}_t^{(j)}) p(\mathbf{s}_t^{(j)}|\mathbf{s}_{t-1}^{(j)})}{\pi(\mathbf{s}_t^{(j)})}. \quad (11)$$

The selection of the importance function affects the efficiency of the SIS algorithm in terms of the number of samples needed to approximate the distributions of interest. The optimum importance function is

$$\pi(\mathbf{s}_t) = p(\mathbf{s}_t|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t})$$

and a proof of it is provided in [3]. In our case when the CSI is unknown, the optimum importance function can be further expended according to

$$\begin{aligned} p(\mathbf{s}_t|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t}) &= \int p(\mathbf{s}_t, \mathbf{h}_t|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t}) d\mathbf{h}_t \\ \propto & \int p(\mathbf{y}_t|\mathbf{s}_t, \mathbf{h}_t) p(\mathbf{h}_t|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t-1}) d\mathbf{h}_t \\ & \times p(\mathbf{s}_t|\mathbf{s}_{0:t-1}^{(j)}). \end{aligned} \quad (12)$$

Inspecting (4) and (6), we can see that the DSSM, given the user state vectors, is linear and Gaussian in the channel state vectors. Hence, the Gaussian probability density function $p(\mathbf{h}_t|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t-1})$ can be obtained by computing the mean and the covariance matrix of $\mathbf{h}_t$ using the prediction step of the Kalman filter. Because the likelihood function $p(\mathbf{y}_t|\mathbf{s}_t, \mathbf{h}_t)$ is Gaussian as well, the integration in (12) can be carried out analytically. As a result, we propose to use a combined SIS and Kalman filtering algorithm, which is conceptually the same as the one from [2]. The algorithm is composed of the following steps:

1. For $j = 1, ..., J$, use the prediction step of the Kalman filter to obtain $p(\mathbf{h}_t|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t-1})$ and evaluate the proposal density using (12).

2. Draw samples and compute their weights as in the case of common SIS algorithms.

3. For $j = 1, ..., J$, use the update step of the Kalman filter to obtain the density function $p(\mathbf{h}_t|\mathbf{s}_{0:t}^{(j)}, \mathbf{y}_{0:t})$, which is needed for the next round of iteration.

An important characteristic of transmitted signals in some communication systems, such as the one addressed here, is that future observations $\mathbf{y}_{t+1:t+p}$ often hold information about the current user state vector $\mathbf{s}_t$. As a result, another posterior density of interest is $p(\mathbf{s}_{0:t}|\mathbf{y}_{0:t+p})$. One can obtain it by first finding $p(\mathbf{s}_{0:t+p}|\mathbf{y}_{0:t+p})$ and then marginalizing with respect to $\mathbf{s}_{t+1:t+p}$. The density $p(\mathbf{s}_{0:t+p}|\mathbf{y}_{0:t+p})$ can be approximated using the combined SIS and Kalman filtering algorithm described above, which is equivalent to the delayed weight method in [2]. Another way of solving this problem is to use the delayed importance function $p(\mathbf{s}_t|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t+p})$, which takes into account all the relevant observations. The drawback of this importance function lies in its computational complexity as we can see from the expression for the density $p(\mathbf{s}_t|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t+p})$, where

$$\begin{aligned} & p(\mathbf{s}_t|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t+p}) \\ \propto & \sum_{\mathbf{s}_{t+1,t+p} \in \mathcal{A}^p} \prod_{\tau=0}^{p} \int p(\mathbf{y}_{t+\tau}|\mathbf{s}_{t+\tau}, \mathbf{h}_{t+\tau}) \\ & \times p(\mathbf{h}_{t+\tau}|\mathbf{s}_{t:t+\tau}, \mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t+\tau-1}) d\mathbf{h}_{t+\tau} \\ & \times p(\mathbf{s}_{t:t+p}|\mathbf{s}_{0:t-1}^{(j)}). \end{aligned} \quad (13)$$

To evaluate this expression, one has to perform predictive Kalman filtering for all possible future sequences of user state vectors $\mathbf{s}_{t:t+p}$. The complexity of the algorithm is proportional to the size of the set $\mathcal{A}^p$. The delayed sample importance function should be weighted with respect to the posterior density $p(\mathbf{s}_{0:t}|\mathbf{y}_{0:t+p})$ which turns out to be

$$w_t^{(j)} \propto w_{t-1}^{(j)} \frac{\sum_{\mathbf{s}_t} p(\mathbf{s}_t^{(j)}|\mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t+p})}{p(\mathbf{y}_{t:t+p-1}|\mathbf{s}_{0:t-1}^{(j)})}. \quad (14)$$

The denominator can be evaluated similarly as the importance function. When dealing with delayed estimation, intuitively the number of delayed samples and delayed weight should be selected so that all relevant observations are taken into account according to the constraint length (memory) of the STTC.

4. PHASE AMBIGUITY

The fading coefficients and the modulated user data are all unknowns that have to be estimated and there may be multiple pairs of estimates with identical posterior probabilities. If $(\mathbf{C}(\mathbf{s}_t), \mathbf{h}_t)$ is an estimate of (6), a phase shifted version $(\mathbf{C}(\mathbf{s}_t)\mathbf{\Theta}, \mathbf{\Theta}^{-1}\mathbf{h}_t)$, where $\mathbf{\Theta}$ is a phase shifting matrix, will be an equally acceptable estimate. For example, consider the 8-PSK delay diversity STTC with two transmit antennas as described in [1]. The codes for the two transmit antennas are

$$c_1(\mathbf{s}_t) = e^{-j\frac{3\pi}{4}s_t} \quad c_2(\mathbf{s}_t) = e^{j\frac{\pi}{4}s_{t-1}}$$

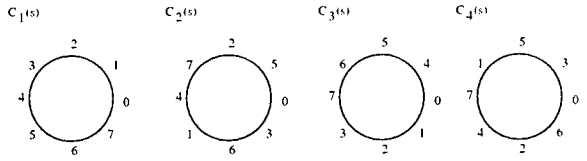


Figure 2: Constellation of Space Time Trellis Code

where $\mathbf{s}_t = [s_t, s_{t-1}]^T$. The phase shifting matrix in this case is $\mathbf{\Theta} = \text{diag}\{e^{j\theta_1}, e^{j\theta_2}\}$. However, as we use a recursive algorithm that evolves with time, a phase shifted version of the channel state estimation with phase shifting matrix $\mathbf{\Theta} = \text{diag}\{e^{j\theta_1}, e^{j\theta_2}\}$ is plausible over a time period $t : t+q$, where $q \in \mathbb{N}$ are natural numbers, if only the following condition is met

$$c_1^{-1}(c_1(s)e^{j\theta_1}) = c_2^{-1}(c_2(s)e^{j\theta_2}) \quad \forall s \in \mathbf{s}_{t:t+q} \qquad (15)$$

i.e., given $\mathbf{\Theta}$, a legitimate code vector sequence $\mathbf{C}(\mathbf{s}_t)\mathbf{\Theta}$, $\cdots$, $\mathbf{C}(\mathbf{s}_{t+q})\mathbf{\Theta}$ can be found. The above condition can be satisfied given several pairs of (θ_1, θ_2) in the original modulation constellation $\forall q$. As a result, phase ambiguity will occur and simulation had shown that it can cause a break down of the algorithm. One may consider allowing use of a bigger variety of constellations to avoid the occurrence of (15) $\forall q$, while guaranteeing certain amount of coding gain. If the condition cannot be avoided completely, one should try to prevent it for larger values of q. It is a challenging task to design constellations that best explore the spatial and time diversities with the added requirement of reduced phase ambiguity. We propose four ad-hoc designed 8-PSK constellations where each time-slot is divided into two sub-slots with two transmit antennas. It is considered that the fading coefficients of the channel will not change between the two sub-time slots. The constellations of c_1 to c_4 are shown in Figure 2. Phase ambiguity is greatly reduced by using this new constellation.

5. SIMULATION

We simulated a two transmit antenna and one receive antenna STTC system. The ARMA process describing the CSI was chosen the same as in [2]. We used the same resampling process as ascribed in [2] and resampled for every 5 steps. The number of delayed weight and delayed sample was one.

The result of the joint detection and decoding algorithm is shown in Figure 3 and it was compared with the genie-aided case when an additional stream of known user data was sent through the same channel for the direct estimation of the channel. Channel estimates in terms of mean and covariance were obtained using Kalman filtering. Then, they were employed in the combined SIS and Kalman filtering algorithm replacing channel estimates obtained from the density function $p(\mathbf{h}_t | \mathbf{s}_{0:t-1}^{(j)}, \mathbf{y}_{0:t-1})$. The genie-aided case serves

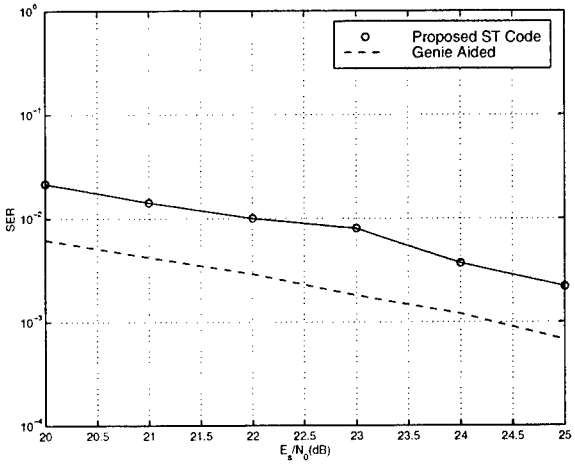


Figure 3: Simulation Result

as a lower bound and the proposed algorithm is 3 dB away from the bound. The genie-aided case assumes that the additional information about the channel are obtained without any cost. However, in practice it is almost never the case. If we take into account the additional amount of energy used for channel estimation in the genie-aided case, the 3 dB difference in the simulation result can be explained. For every simulated point, at least 100 symbol errors were accumulated.

6. REFERENCES

[1] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time Codes for High Data Rate Wireless Communications: Performance Criterion and Code Construction," *IEEE Transactions on Information Theory*, pp. 744-765, Mar. 1998.

[2] R. Chen, X. Wang, and J. S. Liu, "Adaptive Joint Detection and Decoding in Flat-Fading Channels via Mixture Kalman Filtering," *IEEE Transactions on Information Theory*, pp. 2079-2094, Sept. 2000.

[3] A. Doucet, S. Godsill, C. Andrieu, "On Sequential Monte Carlo Sampling Methods for Bayesian Filtering," *Statistics and Computing*, vol. 10, No. 3, pp. 197-208, 2000.

[4] J. H. Kotecha and P. M. Djurić, "Sequential Monte Carlo Sampling Detector for Rayleigh Fast-fading Channels," *International Conference on Acoustics Speech and Signal Processing*, 2000.

A MODIFIED CONSTANT MODULUS ALGORITHM FOR ADAPTIVE CHANNEL EQUALIZATION FOR QAM SIGNALS

Moeness Amin[†], *Lin He*[†], *Charles Reed, Jr*[††], *Robert Malkemes*[††]

[†] Department of Electrical and Computer Engineering
Villanova University, Villanova PA 19085, USA

[††] Integrated Circuit System Laboratory
Sarnoff Corporation, Princeton NJ 08543-5300, USA

ABSTRACT

A modified constant modulus algorithm (MCMA) for adaptive equalization of wireless indoor channel for QAM signals is presented. The algorithm minimizes an error cost function that includes both amplitude and phase of the equalizer output. In addition to the amplitude-dependent term that is provided by the conventional constant modulus algorithm (CMA), the cost function includes a signal constellation matched error (CME) term. This term speeds up convergence and allows the equalizer to switch to Decision Directed (DD), or any soft-decision mode, faster than the CMA applied alone. The constellation-matched error term is constructed using polynomials with desirable properties. The MCMA is applied to a decision feedback equalizer and shown to provide improved performance over dual mode techniques.

1. INTRODUCTION

The Constant Modulus Algorithm (CMA) [1, 2, 3, 4, 5] is a blind technique that achieves channel equalization without the need of a training sequence. The use of CMA in the initial phase of adaptive equalization and then switching (dual mode) to DD or another constellation-matched algorithm, at appropriate values of mean-square error (MSE), is typically performed to improve both the global and local convergence properties [3, 5, 6, 7]. In [6], the constellation-matched error (CME) term is the magnitude-square of finite order polynomial of the equalizer output with zeros at the signal message points, whereas in [7], this term is the complement of the sum of Gaussian functions centered at the message points.

Unlike dual-mode techniques, the proposed modified CMA (MCMA) is similar, in concept, to the "Stop-and-Go" algorithm in that the equalizer integrates a constellation-matched error term in a continuos manner during both initialization and convergence. However, in the MCMA, the cost function is constructed from two separate well defined error terms, one is identical to the CMA case, and the other corresponds to a DD mode, or any other constellation matched error function.

The proposed MCMA belongs to stochastic gradient descent schemes. It performs well for QAM and under dynamic channels, as it converges to acceptable levels of MSE faster than the CMA if applied alone. At these levels, adaptation may proceed with only the signal constellation matched error term of the cost function and without the CMA part. The latter may cause residual errors at local convergence regions, specifically for high order QAM.

The paper provides the framework for defining and selecting an appropriate CME term. This term should satisfy three main desirable properties, namely, uniformity, symmetry, and zero/maximum penalties at the zero/maximum deviations from the QAM symbols. These properties serve to shape the cost function in a manner that is not biased towards any specific alphabet, and to properly alert the adaptive equalizer when high error values are produced. It is noted that neither CME functions defined in [6, 7] strictly satisfies these properties over the entire extent of constellation region. On the other hand, an even-power cosinusoidal CME function satisfies the above conditions, yields a simple gradient, and bounds on the associated adaptation step is inversely proportional to its power.

Section 2 of this paper presents the proposed modified constant modulus algorithm. In Section 3, the local convergence properties are analyzed. The simulation performance of the MCMA is presented in Section 4.

2. MODIFIED CONSTANT MODULUS ALGORITHM

The general form of cost function for the modified constant modulus algorithm is given by

$$J(\boldsymbol{w}) = E\{(|z_k|^2 - A)^2 + \beta(g(z_{kr}) + g(z_{ki}))\} \quad (1)$$

$$A = \frac{E\{|I_k|^4\}}{E\{|I_k|^2\}} \quad (2)$$

where z_k is the received baseband complex signal, z_{kr} and z_{ki} are the real and imaginary parts of z_k respectively, and I_k is the transmitted symbol. The first term in (1) is the amplitude error function, which is the cost function of conventional CMA. $g(x)$ is a constellation matched error function and β is a weighting factor that trades off amplitude and constellation-matched errors. We define $g(x)$ as a polynomial function. This polynomial must strive to satisfy three important key properties in the range $-2Ld \le x \le 2Ld$ for QAM signals,

$$\begin{aligned} s_x &= (2m_x - 1)d, \quad m_x = -L+1, \cdots, L \\ s_y &= (2m_y - 1)d, \quad m_y = -L+1, \cdots, L \end{aligned} \quad (3)$$

where (s_x, s_y) is the symbol point, L is an integer number and $2d$ is the minimum distance between symbols.

0-7803-7011-2/01/$10.00 ©2001 IEEE

Property 1:

The polynomial should be uniform in that it does not favor or penalize information symbols over others, which justifies its periodic behavior,

$$g(x) = g(x + 2ld) \tag{4}$$

where l is an integer.

Property 2:

The polynomial should be symmetric around each alphabet, i.e.,

$$g(s_x + x) = g(s_x - x) \forall_t : 0 \le x \le d \tag{5}$$

Property 3:

The maximum value, which is normalized to one, is reached at the center point in between two consecutive alphabets. The minimum values are zeros and only occur at the constellation points. Accordingly, the cost function places the highest penalty at the maximum deviation and no penalty for zero errors. That is,

$$g(s_x \pm d) = 1 \text{ and } g(s_x) = 0 \tag{6}$$

The above three polynomial properties, although desirable, may consume a large number of degrees of freedom which translate into high polynomial order and complex error gradient expression. The latter increases algorithm computations per iteration update and may render the algorithm unattractive for real-time implementations.

The gradient recursion for the equalizer weight vector, $\boldsymbol{w}$, can be formulated as [8]

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \mu \nabla J(\boldsymbol{w})\big|_{\boldsymbol{w} = \boldsymbol{w}_k} \tag{7}$$

where μ is a step size. The derivative of $J(\boldsymbol{w})$ can be carried out term by term from (1). The first term is directly from the CMA. Its derivative is

$$\frac{d}{d\boldsymbol{w}} E\{(|z_k|^2 - A)^2\} = 4E\{(|z_k|^2 - A) z_k^* \boldsymbol{x}_k\} \tag{8}$$

The z_{kr} and z_{ki} are expressed explicitly in terms of $\boldsymbol{w}$,

$$\begin{aligned} z_{kr} &= \frac{z_k + z_k^*}{2} = \frac{\boldsymbol{w}^H \boldsymbol{x}_k + \boldsymbol{x}_k^H \boldsymbol{w}}{2} \\ z_{ki} &= \frac{z_k - z_k^*}{2j} = \frac{\boldsymbol{w}^H \boldsymbol{x}_k - \boldsymbol{x}_k^H \boldsymbol{w}}{2j} \end{aligned} \tag{9}$$

then it is easy to show that

$$\frac{d}{d\boldsymbol{w}} z_{kr} = \boldsymbol{x}_k, \qquad \frac{d}{d\boldsymbol{w}} z_{ki} = -j\boldsymbol{x}_k \tag{10}$$

Based on (10), it can be shown that

$$\frac{d}{d\boldsymbol{w}} E\{g(z_{kr}) + g(z_{ki})\} = E\{\eta_k \boldsymbol{x}_k\} \tag{11}$$

where

$$\eta_k = \frac{d}{dx} g(x)\Big|_{x = z_{kr}} - j \frac{d}{dx} g(x)\Big|_{x = z_{ki}} \tag{12}$$

According to (8), (10) and (12), the derivative of $J(\boldsymbol{w})$ is obtained as

$$\nabla J(\boldsymbol{w}) = 4E\left\{\left(|z_k|^2 - A) z_k^* + \frac{\beta}{4}\eta_k\right)\boldsymbol{x}_k\right\} \tag{13}$$

Therefore, the modified CMA updating equation is

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - 4\mu\varphi_k \boldsymbol{x}_k \tag{14}$$

and

$$\varphi_k = (|z_k|^2 - A) z_k^* + \frac{\beta}{4}\eta_k \tag{15}$$

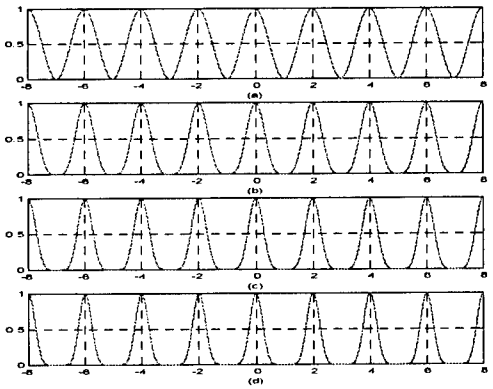


Figure 1. $g_c(x)$ (a) n=1, (b) n=2, (c) n=3, (d) n=4.

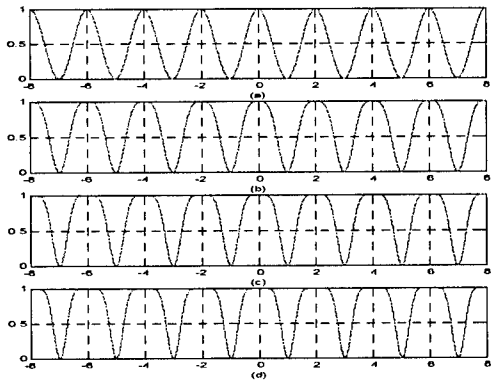


Figure 2. $g_s(x)$ (a) n=1, (b) n=2, (c) n=3, (d) n=4.

In [10], finite order polynomials are designed to form the desired CME term. It can readily be known that even power cosinusoidal functions, representing infinite-order polynomials, shown in figure 1, satisfy the above three properties [10, 11],

$$g_c(x) = \cos^{2n}\left(\frac{x}{2d}\pi\right) \tag{16}$$

where n is an integer. It is also found that

$$g_s(x) = 1 - \sin^{2n}\left(\frac{x}{2d}\pi\right) \tag{17}$$

satisfy the same three properties, as shown in figure 2. The functions in (16, 17) are polynomials of infinite order, but have a simple closed form representations and their values can be obtained using look-up tables. Comparing Fig.1(b) and Fig.2(b), Fig.1(c) and Fig.2(c), and Fig.1(d) and Fig.2(d), it is clear that

for $n>1$, $g_c(x)$ is flat at the symbol points, whereas $g_s(x)$ is sharp. This behavior has a desirable effect on the performance of the MCMA, as shown in the simulation examples.

3. LOCAL CONVERGENCE PROPERTIES

It is shown that after MCMA converges to acceptable levels of MSE, adaptation may proceed with only the signal constellation matched error term of the cost function and without the CMA part. The latter may cause residual errors at local convergence regions, specifically for high order QAM. In this section, we analyze the local convergence properties of the MCMA using (17). The analysis closely follows that given in [7]. We denote $\boldsymbol{s}$ and $\boldsymbol{x}$ as the input and output channel sequences. If the channel matrix is $\boldsymbol{C}$, then $\boldsymbol{x} = \boldsymbol{Cs} + \boldsymbol{v}$, where $\boldsymbol{v}$ is additive white noise. Denoting $\boldsymbol{w}$ as the equalizer weight vector, the output of the equalizer is $\boldsymbol{w}^H\boldsymbol{x}$. The input sequence is of independent identically distributed (iid) symbols. We denote $\boldsymbol{w}_0$ the ideal equalizer weighting vector such that $\boldsymbol{w}_0^H\boldsymbol{Cs}$ equals to one of the constellation points. In general, the equalizer vector is $\boldsymbol{w} = \boldsymbol{w}_0 + \Delta\boldsymbol{w}$, where $\Delta\boldsymbol{w}$ represents the perturbation. Assuming a small deviation $\Delta\boldsymbol{w}$ of the equalizer weights with respect to the ideal vector $\boldsymbol{w}_0$ and choosing $g_s(x)$ as the constellation matched function, we obtain

$$J_f(\boldsymbol{w}) = E\{[1-\sin^{2n}(\frac{z_r}{2d}\pi)]+[1-\sin^{2n}(\frac{z_i}{2d}\pi)]\} \quad (18)$$

Using Taylor series expansion around the message point,

$$J_f(\boldsymbol{w}) \approx E\{n(\frac{\pi}{2d})^2[(z_r-s_x)^2+(z_i-s_y)^2]\} \quad (19)$$

where we have ignored the terms of high orders. The above equation can be expressed as

$$J_f(\boldsymbol{w}) \approx n(\frac{\pi}{2d})^2 E\{\left|\Delta\boldsymbol{w}^H\boldsymbol{Cs}+\boldsymbol{w}^H\boldsymbol{v}\right|^2\}$$

$$\approx n(\frac{\pi}{2d})^2\{\sigma_s^2\Delta\boldsymbol{w}^H\boldsymbol{CC}^H\Delta\boldsymbol{w}+\sigma_n^2\boldsymbol{w}^H\boldsymbol{w}\} \quad (20)$$

where σ_s^2, σ_n^2 indicate the average power of the transmitted sequence and the noise variance, respectively. The respective gradient is

$$\nabla J_f \approx 2n(\frac{\pi}{2d})^2\sigma_s^2\{\boldsymbol{CC}^H\Delta\boldsymbol{w}+\frac{\sigma_n^2}{\sigma_s^2}\boldsymbol{w}_0^H\} \quad (21)$$

In noise-free environment,

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \mu_f 2n(\frac{\pi}{2d})^2\sigma_s^2\boldsymbol{CC}^H\Delta\boldsymbol{w}_k \quad (22)$$

Subtracting $\boldsymbol{w}_0$ from both sides of (22),

$$\Delta\boldsymbol{w}_{k+1} = (\boldsymbol{I}-\mu_f 2n(\frac{\pi}{2d})^2\sigma_s^2\boldsymbol{CC}^H)\Delta\boldsymbol{w}_k \quad (23)$$

where $\boldsymbol{I}$ is the identity matrix. This recursive rule converges if the parameter μ_f is chosen to satisfy the following inequality

$$0<\mu_f<\frac{1}{n(\frac{\pi}{2d})^2\sigma_s^2\lambda_{\max}} \quad (24)$$

where λ_{max} is the largest eigenvalue of $\boldsymbol{CC}^H$. From the above inequality, we observe that small values of n allow the selection of high values of μ_f, for the same channel, i.e., λ_{max}. At the same time, small values of n correspond to flat nulls of the cost function, which affects the MSE performance.

4. SIMULATIONS

In the simulations, the transmitted signal is 64 QAM. The channel is frequency selective and comprised of two multipaths with complex coefficients [1.0000 0.1294-j0.4830]. The signal-to-noise ratio (SNR) is 30dB. Decision feedback equalizers were used which have 16 taps for the feed-forward filter and 16 taps for the feedback filter. Figure 3 shows the performance comparison of the conventional CMA and the MCMA, implementing equations (16, 17). It is evident from the MSE curves that the MCMA improves the equalizer performance by offering faster convergence and smaller misadjustment. Upon convergence, the Symbol Error Rate (SER) for CMA becomes 1.3×10^{-2}, whereas for the MCMA, the SER is less than 10^{-4}. The main reason for such improvement is that the MCMA considers both the modulus and constellation properties of the transmitted signals.

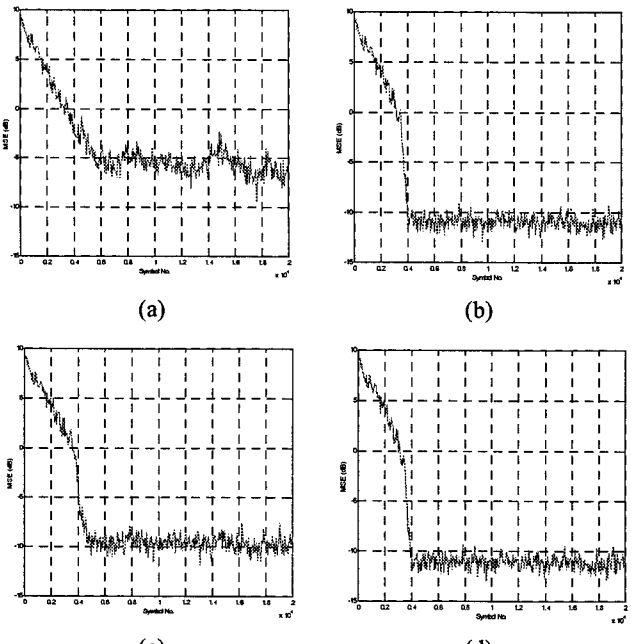


Figure 3. Performance comparison for CMA and MCMA. (a) CMA; (b) MCMA, $g_c(x)=g_s(x)$, n=1; (c) MCMA, $g_c(x)$, n=3; (d) MCMA, $g_s(x)$, n=3.

To determine the effect of the cosinusoidal power term on MCMA performance, we compare figs 3 (b,c,d). The mean square error (MSE) is 0.0812 for n=1, where $g_c(x)=g_s(x)$. For n=3, the MSE is 0.1102 for $g_c(x)$, and 0.0783 for $g_s(x)$. It is clear that Fig.4(c) has the best performance while Fig.4(b) has

the worst performance. Therefore, for the MCMA, the constellation- matched function with a sharp behavior at the symbol points enhances the performance at high SNRs.

The above simulation has demonstrated that the MCMA performs well for QAM, as it converges to acceptable levels of MSE much faster than the CMA, if applied alone. At these levels, adaptation may proceed with only the signal constellation matched error term of the cost function and without the CMA part. Figure 4 compares the performance of CMA and MCMA switching to DD after convergence (3000 samples). It is shown that although the final MSE is same for the two cases, but using MCMA before switching need less convergence time, which because the MCMA converges faster and have less misadjustment than CMA.

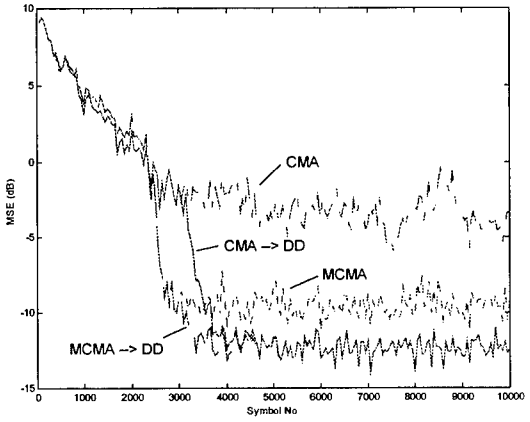


Figure 4. Performance comparison for dual-mode algorithms with CMA and MCMA.

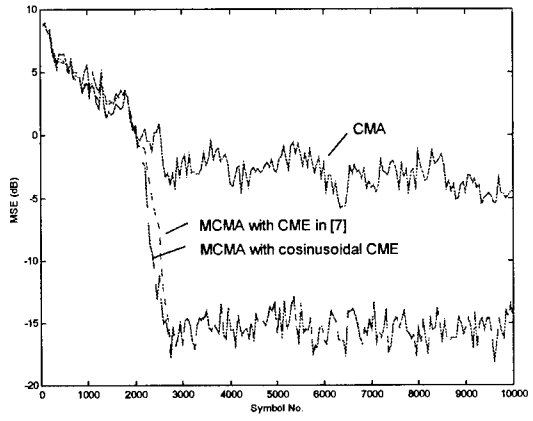


Figure 5. Performance comparison for CMA and MCMA with different constellation-matched functions (SNR = ∞).

In [7], the constellation-matched term is the complement of a sum of Gaussian functions centered at the symbol points, which also approximately satisfies the three desired properties (4-6). This CME can provide almost the same performance as the cosinusoidal CME, if the corresponding parameters are choosing properly. Figure 5 compares the performance for MCMA for both CMEs. It is evident that the two CMEs have equal performance. However, it should be noted that using cosinusoidal functions can greatly reduce computations, because in [7], the CME involves the sum of Gaussian functions centered at every constellation point, leading to computations that are proportional to the number of symbol points, while the cosinusoidal functions have no such dependence.

5. CONCLUSION

In the paper, we have presented an equalizer that minimizes a cost function made up of blind and constellation-dependent terms. The adaptive implementation is referred to as modified constant modulus algorithm. It is shown that the MCMA leads to faster convergence and smaller misadjustment than the CMA. This property permits one to switch to constellation-dependent mode much faster than the case when CMA is applied alone during the initialization process. The paper also presented a framework to select the constellation-matched error term suitable for cost function minimizations.

6. REFERENCES

[1] J.R. Treichler and B.G. Agree, "A new approach to multipath correction of constant modulus signals," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 31, No.2, pp.459-471, April 1983.

[2] Y.S. Choi, H.Wang and D.I. Song, "Adaptive blind equalization coupled with carrier recovery for HDTV modem," IEEE Trans. Consumer Electronics, Vol.39, No.3, pp.386-391, August 1993.

[3] J. Shynk, etc., "A comparative performance study of several blind equalization algorithms," SPIE, San Diego, CA, July 1991.

[4] M. Ghosh, "Blind decision feedback equalization for terrestrial television receivers," Proceedings of the IEEE, Vol.86, No.10, October 1999.

[5] C. Johnson, Jr., P. Schniter, T. Endres, J. Behn, D. Brown and R. Casas, "Blind equalization using the CM criterion: A review," Proceeding of the IEEE, Vol.86, No.10, October 1998.

[6] T.H. Li, "A Blind Equalization for Nonstationary Discrete-valued Signals," IEEE Transactions on Signal Processing, Vol.45, pp.247-254, January 1997.

[7] S. Barbarossa and A. Scaglione, "Blind Equalization Using Cost Functions Matched to the Signal Constellation," Proc. 31st Asilomar Conf. Sig. Sys. Comp., Pacific Grove, CA, Vol.1, pp.550-554, November 1997.

[8] T. Kailath, Adaptive Filter Theory, Prentice-Hall Information and System Science Series, Englewood, New Jersey 07632.

[9] Z. Xu and P. Liu, "Demodulation of Amplitude Modulation Signals in the Presence of Multipath," SSAP-2000, pp.33-37, August 2000.

[10] L. He, M. Amin and Charles Reed, Jr., "Adaptive equalization techniques for indoor dynamic wireless communication channels," SPIE, Orlando, FL, April 2001.

[11] L. He, M. Amin and Y. Zhang, "A Two-Antenna Adaptive Equalizer for Indoor Wireless Channel," Technique Report, Sarnoff Corporation, September 2000.

A PERFORMANCE COMPARISON OF FULLBAND AND DIFFERENT SUBBAND ADAPTIVE EQUALISERS

Hafizal Mohamad[1], *Stephan Weiss*[1], *Markus Rupp*[2], *and Lajos Hanzo*[1]

[1] Dept. Electronics & Computer Science, University of Southampton, UK
[2] Wireless Research Lab / Bell-Labs, Lucent Technologies, Holmdel, NJ, USA
{hm99r,sw1,lh}@ecs.soton.ac.uk, rupp@lucent.com

ABSTRACT

We present two different fractionally spaced (FS) equalisers based on subband methods, with the aim of reducing the computational complexity and increasing the convergence rate of a standard fullband FS equaliser. This is achieved by operating in decimated subbands at a considerably lower update rate and by exploiting the prewhitening effect that a filter bank has on the considerable spectral dynamics of a signal received through a severely distorting channel. The two presented subband structures differ in their level of realising the feedforward and feedback part of the equaliser in the subband domain, with distinct impacts on the updating. Simulation results pinpoint the faster convergence at lower cost for the proposed subband equalisers.

1. INTRODUCTION

Linear channel distortions caused by multipath propagation and limited bandwidth lead to inter-symbol interference (ISI) at the receiver, which in many cases results in a high bit error rate in the detection. Therefore, many different adaptive equalisation structures have been proposed in the past in order to compensate for these channel distortions in the receiver. Most popular amongst the subset of linear or minimum-mean-square error (MMSE) equalisers are currently fractionally spaced (FS) architectures [1], whereby the equalisation filter operates at a rate higher than the symbol rate.

A standard fractionally spaced equaliser is shown in Fig. 1. The structure operates the feedforward (FF) part of the equaliser at an oversampled rate, here twice the symbol rate. In the flow graph in Fig. 1, the FF part is implemented as a polyphase structure [2] the two polyphase components running $a_0[n]$ and $a_1[n]$ of the adaptive FF filter at the lower symbol rate. The two filters $a_0[n]$ and $a_1[n]$ are excited by the two polyphase components of the oversampled channel output $x[m]$. The feedback (FB) part of the equaliser is symbol spaced. This is due to the equation error formulation or the decision feedback mode of the equaliser. In the FB part, the adaptive filter $b[n]$ can be excited by either a training signal (switch position 1) — a copy of the transmitted symbol sequence $u[n]$ delayed by Δ periods — or in decision feedback mode (switch position 2). All FF and FB parts $a_0[n]$, $a_1[n]$ and $b[n]$ are adaptive and updated by a suitable algorithm at the symbol rate based on an appropriate criterion of the equalisation error $e[n]$.

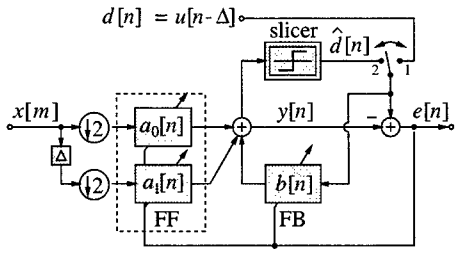


Figure 1: Fractionally spaced equaliser with a polyphase representation of the FF part.

A fractionally spaced equaliser may suffer from considerable computational complexity due to the requirement for long filters if the channel exhibits severe distortions [3], and from slow convergence due to strong spectral dynamics at the input to the equaliser [4]. These characteristics have previously triggered the application of subband techniques to FS equalisers [5], based on the computational reduction, prewhitening, and parallelisation properties of the subband approach [6, 7, 8]. In this contribution, we evaluate two different subband architectures for FS equalisers. This includes a novel scheme for including the equaliser's feedback section into the subband domain, and the incorporation of decision directed subband equaliser structures to track channel alterations after initial equaliser training.

This paper is organised as follows. In Sec. 2, we briefly describe the channel characteristics and motivate subband decompositions. Then, we introduce the proposed subband adaptive equaliser structures and discuss the complexity issue in Sec. 3. In Sec. 4, we present some simulation results to demonstrate the performance of the subband approach.

2. CHANNEL CHARACTERISTICS AND SUBBAND DECOMPOSITIONS

For the popularly applied least mean square (LMS) type algorithm in equalisation, the convergence speed is inversely proportional to the eigenvalue spread of its input signal [9]. In turn the eigenvalue spread of a signal can be approximated by the ratio between the maximum and minimum value of its power spectral density (PSD). As an example for the spectral dynamics that can be encountered, we consider a severely dispersive channel given in Fig. 2. The

0-7803-7011-2/01/$10.00 ©2001 IEEE

selected channel with a delay spread of approximately 100 symbol periods exhibits additional spectral zeros that reduce the equaliser convergence performance, and also encompasses the transmit and receive filters, that impose a low-pass characteristic on the PSD.

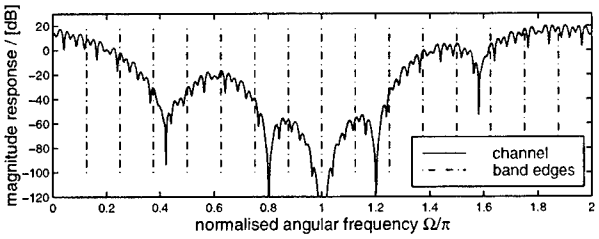


Figure 2: Channel spectral dynamics characteristic with transmit- and receive filter.

A general decomposition into K frequency bands decimated by N (so-called "subbands") is shown in Fig. 3. The filters in both analysis and synthesis bank are bandpass filters, which, together with the decimation process yield a prewhitening of the subband signals compared to the input. Further, computational savings arise due to an N times lower update rate and lower filter orders compared to fullband implementations. For adaptive filtering applications, adaptive filterings can be operated in each band independently, which lends itself to a parallel implementation. As a drawback, subband structures however introduce aliasing that limits the algorithm performance. Therefore, oversampled filter banks (OSFB) with and oversampling ratio $K/N > 1$ are preferred here [5, 6]. An example of $K = 16$ subband channel is indicated by the band edges in Fig. 2, where the eigenvalue spread within each band is reduced. Therefore, the faster convergence of the algorithm is expected with subband decompositions.

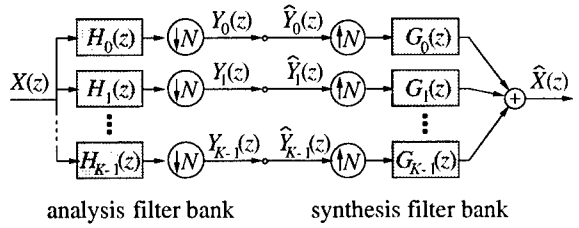


Figure 3: K-channel filter bank decimated by N with analysis filters $H_k(z)$ and synthesis filters $G_k(z)$.

An additional benefit of the subband implementation is that an impulse response in the decimated domain can be modelled with less coefficients than required in the fullband case due to the increased sampling period, achieving similar modelling capabilities. In general, this decreases the necessary filter length by a factor of N, whereby a moderate overhead of prototype filter coefficients has to be taken into account as in the subband domain potentially fractional delays have to be modelled [7]. The length of subband filter coefficients is given by

$$L_{\text{Subband}} = \frac{L_{\text{Fullband}} + L_p}{N} \tag{1}$$

where L_p denotes for the length of the prototype filter.

3. SUBBAND ADAPTIVE EQUALISER STRUCTURES

In this section, we introduce two different subband adaptive equaliser structures and discuss the complexity issues of the equalisers. For the subband implementation, we utilise OSFBs as described in reference [10].

3.1. Structure I

For subband equaliser structure I, the FF part of the fullband equaliser in Fig. 1, is projected into subbands. The resulting architecture is shown in Fig. 4, whereby **H** and **G** denote analysis and synthesis filter bank blocks including decimation and expansion as given in Fig. 3. The system blocks $\mathbf{A}_0$ and $\mathbf{A}_1$ are diagonal polynomial matrices representing independent filters within each of the K subbands. As the FB part has to be performed at symbol rate, the error is evaluated based on the FF outputs reconstructed by **G**, and is projected back into the subband domain to update the filters in $\mathbf{A}_0$ and $\mathbf{A}_1$.

A drawback of the update procedure for the FF part is, that the error signal contains a transfer path. This transfer path can be approximated by a delay identical to L_p/N. This delay has been reported to result in degraded convergence speed [11]. To overcome this problem, a modification of the structure I architecture will be introduced by integrating the FB part into subbands.

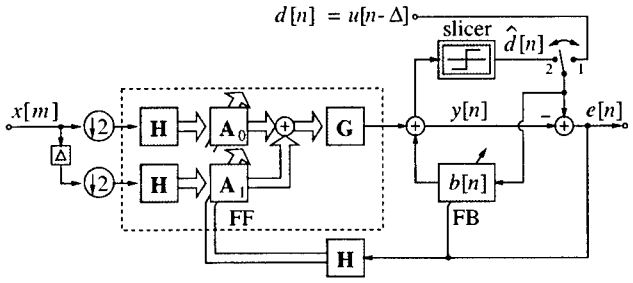


Figure 4: Adaptive equaliser structure I with the FF part in subband.

3.2. Structure II

A subband equaliser structure II is shown in Fig. 5, which has the aim to overcome slow convergence due to the error transfer path in structure I. The error signal is now formed in the subband domain and can be used to delaylessly update both the FF and FB parts. Similarly to structure I, **B** is of diagonal polynomial form holding the adaptive FB filters running independently within each subband.

In structure II architecture, all adaptive filters are updated by the immediately formed subband errors *at the*

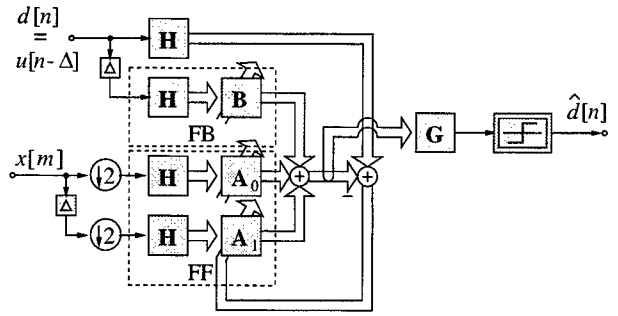


Figure 5: Adaptive equaliser structure II with both the FF and FB parts in subbands.

same time. This is expected to provide improved convergence characteristics over structure I. However, as the error is calculated in the subband domain, this structure can only be used in training mode. The decision directed learning mode — switch position 2 in the fullband structure in Fig. 1 and the subband structure I in Fig. 4 — requires a non-linearity that cannot be transferred into the subband domain. Therefore, if decision directed mode was to be performed, structure I would have to be selected. By appropriate subband projections, the FB filter $b[n]$ in Fig. 4 can be reconstructed from **B** in Fig. 5.

3.3. Computational Complexity

The complexity of a fullband equaliser implementation in terms of multiply-accumulates (MACs) when using an NLMS algorithm for updating is approximately given by

$$C_{\text{fullband}} = 4 \cdot 2(L_{\text{FF}} + L_{\text{FB}}) = 8(L_{\text{FF}} + L_{\text{FB}}) \qquad (2)$$

where the factor of 4 accounts for the required complex valued arithmetic. The feedforward and feedback filter lengths are represented by L_{FF} and L_{FB}, respectively.

For our subband equaliser implementations, the complexity of the filter banks has to be considered. In a fast implementation, one analysis or synthesis filter bank operation cost

$$C_{\text{filterbank}} = \frac{1}{N} \cdot (2L_p + 4K\log_2 K) \qquad (3)$$

MACs per fullband sampling period [10].

Thus the complexity of subband structure I with the FF part in subband and 4 filter bank operations is

$$C_{\text{subband,I}} = \frac{K}{N} 4 \cdot 2(L_{\text{FF}}) + 4 \cdot 2(L_{\text{FB}}) + 4C_{\text{filterbank}}. \qquad (4)$$

For subband structure II, we require

$$C_{\text{subband,II}} = \frac{K}{N} 4 \cdot 2(L_{\text{FF}} + L_{\text{FB}}) + 5C_{\text{filterbank}} \qquad (5)$$

due to operating both FF and FB parts in subbands in the structure and executing 5 filter banks.

4. SIMULATIONS AND RESULTS

The channel characteristic in Fig. 2 has been used to test the fullband and subband equalisers introduced in Sec. 3. Quadrature amplitude modulation (QAM) signals are used in our simulation. A normalised least mean square (NLMS) algorithm is employed for adaptation of the fullband and subband structure II adaptive filters, while a delay-NLMS is used in subband structure I. The normalised step size of $\tilde{\mu} = 0.4$ is set for all equaliser structures. The delay Δ for the different systems is set such that the FF part targets almost only the pre-cursor, while the FB part of the equaliser eliminates the post-cursor. For the subband structures, the OSFBs split the fullband signal into $K = 16$ channels decimated by $N = 14$, with $L_p = 448$.

The filter length of the subband equalisers is selected according on (1). The number of coefficients of the different structures — L_{FF} refers to the filter in the FF part, and L_{FB} to the FB part of the equaliser — is listed in Tab. 1.

Equaliser structure	L_{FF}	L_{FB}
Fullband	500	100
Structure I	70	100
Structure II	70	40

Table 1: Number of coefficients in the FF and FB parts of the different simulated equaliser structures.

The performance of the three — fullband, and subband structure I and II — equaliser systems is assessed in terms of achieved mean squared error (MSE) and bit error rate (BER), whereby both the learning characteristic as well as the steady state are of interest.

4.1. Convergence Behaviour

The MSE learning characteristic of the three systems is presented in Fig. 6. The curves are averaged over an ensemble of 25 runs with a random 64-QAM input signal $u[n]$ in the absence of channel noise. In terms of convergence rate, the subband structures exhibit a convergence speed that is approximately twice as fast as the fullband equaliser. Whereby subband structure II attains a faster initial MSE convergence performance over structure I. It is indicative that both subband structure I and II attain a considerably better steady-state error performance than the fullband system.

4.2. Bit Error Rates

We further examine the performance of the fullband equaliser and subband structure II in terms of BER for various levels of QAM over the previous channel, which now is disturbed by noise at variable SNR. The noise is independent of the transmitted signal. An additive white Gaussian noise is coloured by the receive filter. The BER performance results for 4-, 16-, 64-, and 256-QAM over variable SNR are shown in Fig. 7. The displayed BER values are taken for the steady-state case after adapting the equalisers for $5 \cdot 10^5$ symbol periods. In general, the fullband equaliser is superior particularly for lower modulation levels at low SNR.

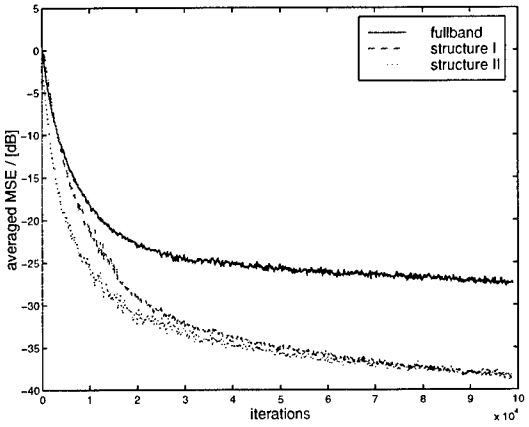


Figure 6: MSE performance for fullband and subband (structure I and II) equalisers for a noise free channel.

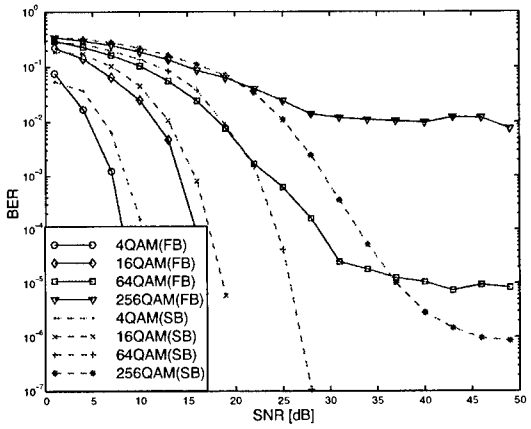


Figure 7: BER performance of fullband (FB) and subband (SB) structure II over variable channel SNR for various modulation levels.

A clear advantage for the steady-state performance of the subband structure can be noted for higher QAM levels (64-QAM and 256-QAM) at higher SNR above 25 dB.

4.3. Computational Cost Comparison

The filter lengths of the proposed subband structures — selected according to (1) — are given in Tab. 1. These filter lengths have been set to achieve similar modelling capabilities of the fullband and different subband structures. The computational complexity of the equaliser structures — calculated according to (2), (4), and (5) — are displayed in the second column of Tab. 2. The third column in Tab. 2 represents the computational cost comparison for the subband equalisers implementations compared to the fullband realisation. Subband structure I and II only require 39% and 29%, respectively, of the fullband equaliser's computational complexity.

Equaliser structure	MACs	% of Fullband
Fullband	4800	100%
Structure I	1882	39%
Structure II	1416	29%

Table 2: Computational cost comparison for different equaliser structures.

5. CONCLUSIONS

This paper has introduced structures for subband adaptive equalisation and presented some simulation results. An important indication from these results is that for severely distorting channels subband equalisers can attain a faster convergence rate and better steady-state error than their fullband counterpart with a gain in BER for high SNR when operating in higher level QAM modes. The subband equalisers were implemented at a reduced computational cost compared to the fullband system.

6. REFERENCES

[1] J. R. Treichler, I. Fijalkow, and C. R. Johnson, "Fractionally Spaced Equalizers: How Long Should They Really Be?," *IEEE SP Mag*, 13(3):65–81, May 1996.

[2] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, 1993.

[3] S. Weiss, S. R. Dooley, R. W. Stewart, and A. K. Nandi, "Adaptive Equalization in Oversampled Subbands," *IEE Elec. Let.*, 34(15):1452–1453, July 1998.

[4] M. Rupp, "On the Learning Behaviour of Decision Feedback Equalizers," in *Asilomar Conf. SSC*, Monterey, CA, Oct. 1999.

[5] S. Weiss, M. Rupp, and L. Hanzo, "A Fractionally Spaced DFE with Subband Decorrelation," in *Asilomar Conf. SSC*, Monterey, CA, Nov. 2000.

[6] W. Kellermann, "Analysis and Design of Multirate Systems for Cancellation of Acoustical Echoes," in *Proc. ICASSP*, 5:2570–2573, New York, 1988.

[7] A. Gilloire and M. Vetterli, "Adaptive Filtering in Subbands with Critical Sampling: Analysis, Experiments and Applications to Acoustic Echo Cancellation," *IEEE Trans SP*, 40(8):1862–1875, Aug. 1992.

[8] J. J. Shynk, "Frequency-Domain and Multirate Adaptive Filtering," *IEEE SP Mag.*, 9(1):14–37, Jan. 1992.

[9] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, Englewood Cliffs, New York, 1985.

[10] S. Weiss and R. W. Stewart, "Fast Implementation of Oversampled Modulated Filter Banks," *IEE Elec. Let.*, 36(17):1502–1503, Aug. 2000.

[11] M. Rupp and R. Frenzel, "Analysis of LMS and NLMS Algorithms with Delayed Coefficient Update under the Presence of Spherically Invariant Processes," *IEEE Trans SP*, 42(3):668–672, Mar. 1994.

SIMULATION OF WIDEBAND MOBILE RADIO CHANNELS USING SUBSAMPLED ARMA MODELS AND MULTISTAGE INTERPOLATION*

Dieter Schafhuber, Gerald Matz, and Franz Hlawatsch

Institute of Communications and Radio-Frequency Engineering, Vienna University of Technology
Gusshausstrasse 25/389, A-1040 Vienna, Austria
Tel.: +43 1 58801 38973, Fax: +43 1 58801 38999, E-mail: dschafhu@aurora.nt.tuwien.ac.at
web: http://www.nt.tuwien.ac.at/dspgroup/time.html

ABSTRACT

We present a technique for simulating time-varying mobile radio channels. This technique is specifically suited to the small relative Doppler bandwidths of wideband channels encountered in CDMA and OFDM communications. A "subsampled" ARMA innovations filter and multistage interpolation are used to achieve an accurate and computationally efficient approximation of specified or measured Doppler spectra (scattering functions). We discuss the calculation of the ARMA coefficients and the optimal design of the multistage interpolator. Simulation results demonstrate the excellent performance of the proposed channel simulator.

1. INTRODUCTION

Computer simulation of mobile radio channels is of great importance for the development and evaluation of mobile communications systems. A discrete-time channel model that is convenient for channel simulation is the time-varying tapped delay line (FIR filter) with input-output relation [1, 2]

$$y[n] = \sum_{m=0}^{M-1} h_m[n]\, x[n-m]\,.$$

Here, $x[n]$ is the channel input signal, $y[n]$ is the channel output signal, $h_m[n]$ is the channel's time-varying impulse response (with m the delay index and n the time index), and $M-1$ is the maximum delay. For *wide-sense stationary uncorrelated scattering* (WSSUS) channels, each tap weight sequence $h_m[n]$ is a stationary random process with autocorrelation function $r_m[l] = \mathrm{E}\{h_m[n+l]\, h_m^*[n]\}$, and different tap weight processes $h_m[n]$, $h_{m'}[n]$ are uncorrelated [1, 2]. The power spectra of the $h_m[n]$,

$$S_m(\nu) = \sum_{l=-\infty}^{\infty} r_m[l]\, e^{-j2\pi\nu l}, \qquad m = 0, 1, \cdots, M-1\,,$$

are termed the channel's *Doppler spectra* or *scattering function* [1, 2]. Here, ν is the Doppler frequency normalized by the sampling frequency. For wideband CDMA and OFDM systems, the sampling frequency is significantly higher than the channel's maximum Doppler shifts. This results in extremely small Doppler bandwidths, i.e., the Doppler spectra $S_m(\nu)$ have extremely narrowband lowpass characteristics.

From the discussion above, it follows that the simulation of a WSSUS channel amounts to generating realizations of M uncorrelated, stationary tap processes $h_m[n]$ ($m = 0, 1, \cdots, M-1$) whose second-order statistics should conform to the specified $r_m[l]$ or, equivalently, $S_m(\nu)$. In this paper, we consider a tap process generator based on an *autoregressive moving-average* (ARMA) innovations filter [3, 4] that is driven by stationary white Gaussian noise. To avoid the high ARMA model order that would normally be needed for achieving the small relative Doppler bandwidths encountered in wideband CDMA and OFDM systems, we propose a "subsampled" ARMA innovations filter that is designed using a subsampled autocorrelation

$$r'_m[n] \triangleq r_m[nL]\,. \tag{1}$$

To compensate for the subsampling, the ARMA filter is followed by a *multistage interpolator* [5] for which we propose an MSE-optimal design. A block diagram of the resulting tap process generator is shown in Fig. 1. The generic structure of a channel simulator consisting of innovations filters and interpolators was previously considered in [6].

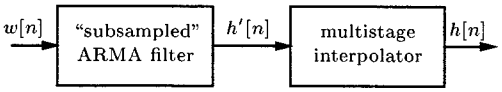


Figure 1: *Structure of the proposed channel simulator: Generating a realization of a tap weight process* $h_m[n]$.

The proposed channel simulator has numerous advantages over other channel simulation techniques [1, 7–10]: the ARMA modeling approach allows accurate approximation of arbitrary Doppler spectra; arbitrarily long tap sequences can be generated online; new realizations are generated in each simulation run; the simulated channel is guaranteed to be Rayleigh fading; and finally, the multistage interpolator allows for efficient implementation, simplified interpolator filter design, and easy adjustment of the Doppler bandwidth without modification of the ARMA filter.

The rest of the paper is organized as follows. Section 2 discusses the subsampled ARMA innovations filter and presents methods for calculating the filter coefficients. Section 3 considers the multistage interpolator and its optimal design. Finally, simulation results are provided in Section 4.

2. ARMA INNOVATIONS FILTER

The subsampled ARMA innovations filter is an IIR filter described by the difference equation [3]

$$\sum_{k=0}^{P} a[k]\, h'[n-k] = \sum_{k=0}^{Q} b[k]\, w[n-k]\,, \quad \text{with } a[0] \triangleq 1\,, \tag{2}$$

*This work was supported by FWF grant P11904-TEC.

0-7803-7011-2/01/$10.00 ©2001 IEEE

where $a[k]$ and $b[k]$ are the autoregressive (AR) and moving-average (MA) coefficients, respectively, P and Q are the AR and MA model orders, respectively, $h'[n]$ is the filter output (the subsampled tap process; note that we suppress the subscript m in $h'_m[n]$ etc.), and $w[n]$ is the filter input that is chosen as white Gaussian noise. Multiplying (2) by $h'^*[n-l]$ and taking expectations yields [3]

$$\sum_{k=0}^{P} a[k]\, r'[l-k] = \sum_{k=0}^{Q} b[k]\, c^*[k-l]\,, \quad l \in \mathbb{Z}\,, \tag{3}$$

where $r'[n] = r[nL]$ is the subsampled autocorrelation and $c[n]$ is the impulse response of the ARMA filter. This is a nonlinear equation in the ARMA coefficients $a[k]$ and $b[k]$ since $c[n]$ depends on $a[k]$ and $b[k]$.

In the frequency domain, (3) becomes

$$S'(\nu) = \frac{B(\nu)}{A(\nu)}\, C^*(\nu) = \frac{|B(\nu)|^2}{|A(\nu)|^2}\,, \tag{4}$$

where $S'(\nu)$, $A(\nu)$, $B(\nu)$, and $C(\nu) = B(\nu)/A(\nu)$ are the Fourier transforms of $r'[n]$, $a[n]$, $b[n]$, and $c[n]$, respectively.

For calculation of the AR and MA coefficients, the *specified* (subsampled) autocorrelation and Doppler spectrum are substituted for $r'[n]$ in (3) and for $S'(\nu)$ in (4), respectively. Typically, the ARMA model will only provide an approximation to the specified $r'[n]$ and $S'(\nu)$, and thus (3) and (4) will be satisfied only approximately.

2.1. Calculation of the AR Coefficients

Usually, the AR coefficients $a[n]$ are estimated from higher-lag (and, thus, smaller) values of $r'[n]$ that are not influenced by the MA model part [3]. However, here we include the central (largest) values of $r'[n]$ since we observed this to yield better accuracy of the overall ARMA approximation. This approach means that we first fit an AR filter and then fit an MA filter to the resulting residual autocorrelation. Formally setting $Q = 0$ and noting that the ARMA filter impulse response $c[n]$ is causal, (3) for $l = 1, 2, \cdots, N$ together with $a[0] = 1$ yields the Yule-Walker equation [3]

$$\mathbf{R}\,\mathbf{a} = -\mathbf{r}\,, \tag{5}$$

where

$$\mathbf{R} = \begin{bmatrix} r'[0] & r'[-1] & \dots & r'[-P+1] \\ r'[1] & r'[0] & \dots & r'[-P+2] \\ \vdots & \vdots & & \vdots \\ r'[N-1] & r'[N-2] & \dots & r'[N-P] \end{bmatrix},$$

$\mathbf{a} = [a[1] \cdots a[P]]^T$, and $\mathbf{r} = [r'[1] \cdots r'[N]]^T$, with N the number of samples of $r'[n]$ that are used for the calculation. Equation (5) is a system of N linear equations in the P unknowns $a[n]$. For $N \geq P$, the least-squares solution of (5) (stabilized by diagonal loading [11, Chap. 7.4]) is given by

$$\mathbf{a} = -(\tilde{\mathbf{R}}^H \tilde{\mathbf{R}})^{-1} \tilde{\mathbf{R}}^H \mathbf{r}\,, \quad \text{with } \tilde{\mathbf{R}} = \mathbf{R} + \gamma\, \mathbf{I}\,.$$

Here, γ is a suitable loading parameter ensuring that all poles of the AR filter are inside the unit circle. For good results, N must be chosen much larger than P.

Criteria for selecting the AR model order P are discussed in [3]. However, it is also noted in [3] that these criteria seem to work well only for a true AR process. For a Doppler bandwidth of 10^{-4} and subsampling factor $L = 1000$, we obtained good results with $P = 10 \cdots 50$ and $N = 3P \cdots 10P$.

2.2. Calculation of the MA Coefficients

Once the AR coefficients $a[n]$ have been determined, we can proceed to calculate the MA coefficients $b[n]$. With *Durbin's method* [3,4], the MA modeling problem is transformed into two AR modeling problems of which one has significantly higher order than the MA model order Q. However, since in our case $Q = 100 \cdots 1000$ to ensure good approximation accuracy, Durbin's method would be extremely expensive. Therefore, here we propose a modified version of the extended Prony method described in [3]. Our method is also related to the Blackman-Tukey spectral estimator [4].

Equation (4) can be rewritten as

$$|B(\nu)|^2 = |A(\nu)|^2\, S'(\nu)\,. \tag{6}$$

Basically, $b[n]$ can be obtained by causal factorization of $|B(\nu)|^2$. In the time domain, (6) reads

$$\beta[n] = \alpha[n] * r'[n]\,, \tag{7}$$

with $\alpha[n] = a[n] * a^*[-n]$ and $\beta[n] = b[n] * b^*[-n]$. But from $\beta[n] = b[n] * b^*[-n]$, it follows that $\beta[n]$ should have finite support $[-Q, Q]$ and a nonnegative real Fourier transform. Therefore, we "correct" (7) by applying a Bartlett window to the right-hand side,

$$\beta[n] = \begin{cases} (\alpha[n] * r'[n])\,(1 - \frac{|n|}{Q+1})\,, & |n| \leq Q \\ 0\,, & |n| > Q\,. \end{cases} \tag{8}$$

This enforces both finite support $[-Q, Q]$ and a nonnegative real Fourier transform (since the Fourier transforms of both $\alpha[n] * r'[n]$ and $1 - \frac{|n|}{Q+1}$ are nonnegative real). Finally, $b[n]$ is obtained by causal factorization of $\beta[n]$ [12, App. A]. To this end, the cepstrum of $\beta[n]$ in (8) is calculated [13]. Transforming the cepstrum's causal part back into the original domain yields the MA coefficients $b[n]$, $n = 0, 1, \cdots, Q$. The overall technique was observed to produce similar results as Durbin's method (see Section 4) at significantly reduced computational complexity.

Criteria for selecting the MA model order Q are discussed in [3]. We found that for good approximation accuracy, the MA filter should have the same length as the central (dominant) part of the subsampled autocorrelation $r'[n]$. For a Doppler bandwidth of 10^{-4} and subsampling factor $L = 1000$, we obtained good results with $Q = 100 \cdots 1000$.

3. MULTISTAGE INTERPOLATOR

To compensate for the subsampling of $r[n]$ in (1), the output $h'[n]$ of the subsampled ARMA filter is interpolated by the subsampling factor L. If L is chosen as a composite number, i.e., $L = \prod_{k=0}^{K-1} L_k$, a particularly efficient *multistage* interpolator [5] can be used. Here, interpolation by L is performed by K successive interpolator stages with interpolation factors L_k. Each interpolator stage is represented using a polyphase decomposition. The input-output relations of the individual interpolator stages are [5]

$$h^{(k+1)}[n] = \sum_{i=0}^{L_k - 1} u_i^{(k)} \left[\left\lfloor \frac{n}{L_k} \right\rfloor - i \right], \quad k = 0, 1, \cdots, K-1$$

(where $\lfloor \xi \rfloor$ denotes the largest integer $\leq \xi$), with

$$u_i^{(k)}[n] = \sum_{l=-V_k}^{V_k - 1} p_i^{(k)}[l]\, h^{(k)}[n-l]\,, \quad i = 0, 1, \cdots, L_k - 1\,. \tag{9}$$

Here, $h^{(k)}[n]$ and $h^{(k+1)}[n]$ are the input and output, respectively, of the kth interpolator stage (in particular, $h^{(0)}[n] = h'[n]$ and $h^{(K)}[n] = h[n]$, cf. Fig. 1), and $p_i^{(k)}[n]$ and $u_i^{(k)}[n]$ are the impulse response (of length $2V_k$) and output sequence, respectively, of the ith polyphase filter of the kth interpolator stage.

We propose a mean square error (MSE) optimal design of the multistage interpolator that is analogous to the "deterministic MSE design" described in [5, Sec. 4.3.6]. The kth interpolator stage is designed such that the MSE

$$\mathrm{MSE}^{(k)} \triangleq \mathrm{E}\{|h^{(k+1)}[n] - \tilde{h}^{(k+1)}[n]|^2\}$$

is minimized. Here, $\tilde{h}^{(k+1)}[n]$ is the output of an ideal lowpass interpolator with appropriate cutoff frequency.

The output signals of the polyphase branches within the kth interpolator stage are nonoverlapping in time and thus orthogonal. Hence, $\mathrm{MSE}^{(k)} = \sum_{i=0}^{L_k-1} \mathrm{MSE}_i^{(k)}$ with

$$\mathrm{MSE}_i^{(k)} = \mathrm{E}\{|u_i^{(k)}[n] - \tilde{u}_i^{(k)}[n]|^2\},$$

where $\tilde{u}_i^{(k)}[n] = \sum_{l=-\infty}^{\infty} \tilde{p}_i^{(k)}[l]\, h^{(k)}[n-l]$ (cf. (9)) is the output signal of an ideal interpolator polyphase filter with transfer function $\tilde{P}_i^{(k)}(\nu) = e^{j2\pi i\nu/L_k}$ [5]. This means that the individual polyphase filters $p_i^{(k)}[n]$ can be designed independently by separately minimizing the MSE components $\mathrm{MSE}_i^{(k)}$. It is easily verified that $\mathrm{MSE}_i^{(k)}$ can be expressed in the frequency domain as

$$\mathrm{MSE}_i^{(k)} = \int_{-1/2}^{1/2} \left|P_i^{(k)}(\nu) - \tilde{P}_i^{(k)}(\nu)\right|^2 S^{(k)}(\nu)\, d\nu. \quad (10)$$

Here, $S^{(k)}(\nu) = \sum_{n=-\infty}^{\infty} r^{(k)}[n]\, e^{-j2\pi\nu n}$, with $r^{(k)}[n] = r[nI_k]$, $I_k = \prod_{i=k}^{K-1} L_i$, is the Doppler spectrum of the input process of the kth interpolator stage. Inserting $P_i^{(k)}(\nu) = \sum_{n=-V_k}^{V_k-1} p_i^{(k)}[n]\, e^{-j2\pi\nu n}$ and $\tilde{P}_i^{(k)}(\nu) = e^{j2\pi i\nu/L_k}$ into (10) and setting the derivatives of the resulting expression with respect to $p_i^{(k)*}[n]$ $(n = -V_k, \cdots, V_k - 1)$ equal to zero [3], we obtain the following equations for the MSE-optimal interpolator coefficients $p_i^{(k)}[n]$,

$$\sum_{l=-V_k}^{V_k-1} p_i^{(k)}[l]\, r^{(k)}[n-l] = r^{(k+1)}[nL_k+i], \quad n = -V_k, \cdots, V_k-1.$$

This can be written as

$$\mathbf{R}^{(k)}\, \mathbf{p}_i^{(k)} = \mathbf{r}_i^{(k+1)}, \quad (11)$$

with the vectors $\mathbf{p}_i^{(k)} = [p_i^{(k)}[-V_k] \cdots p_i^{(k)}[V_k-1]]^T$, $\mathbf{r}_i^{(k+1)} = [r^{(k+1)}[-V_kL_k+i] \cdots r^{(k+1)}[(V_k-1)L_k+i]]^T$ and the Hermitian Toeplitz matrix $\mathbf{R}^{(k)}$ with first row $[r^{(k)}[0] \cdots r^{(k)}[-2V_k+1]]$. Since $\mathbf{R}^{(k)}$ does not depend on i, (11) is best solved by explicitly inverting $\mathbf{R}^{(k)}$ using the Levinson algorithm [3, 4], which has to be done only once per interpolator stage k.

The MSE-optimal multistage interpolator explicitly depends on the exact shape of the specified Doppler spectrum $S(\nu)$. If this dependence is undesired, one may employ a suboptimal default design using a rectangular $S(\nu)$ that is constant within the Doppler bandwidth and zero outside.

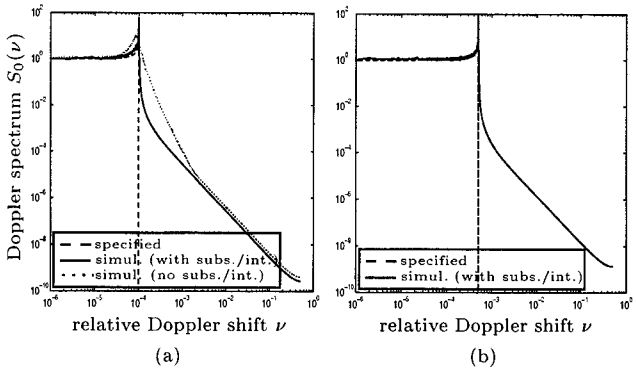


Figure 2: *Simulation of a single-tap channel: (a) Specified (Jakes) Doppler spectrum, simulated Doppler spectrum, and simulated Doppler spectrum obtained without subsampling/interpolation; (b) simulated Doppler spectrum obtained at the output of the second interpolator stage.*

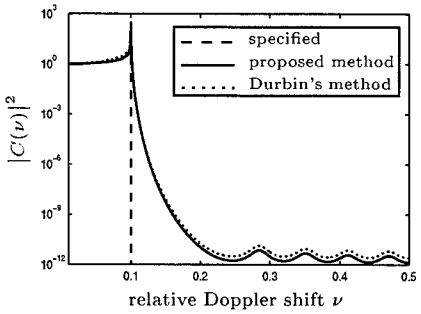


Figure 3: *Squared magnitude of the frequency responses of the subsampled ARMA filter obtained with the proposed MA design method and with Durbin's method.*

4. SIMULATION RESULTS

4.1. Simulation 1: Jakes Doppler Spectrum

Fig. 2 and Fig. 3 consider the simulation of a single-tap channel ($M = 1$) with a Jakes Doppler spectrum [1]

$$S_0(\nu) = \begin{cases} \frac{\nu_{\max}}{\sqrt{\nu_{\max}^2 - \nu^2}}, & |\nu| \le \nu_{\max}, \\ 0, & \text{else.} \end{cases}$$

The relative Doppler bandwidth was $\nu_{\max} = 10^{-4}$ (corresponding, e.g., to an absolute Doppler bandwidth of 100 Hz when a sampling frequency of 1 MHz is used). The subsampling/interpolation factor was chosen as $L = 1000$. An ARMA filter of order $P = 20$, $Q = 100$ was designed using parameters $N = 200$ and $\gamma = 6 \cdot 10^{-10}$. The multistage interpolator comprised 3 stages with $L_0 = 5$, $L_1 = 40$, $L_2 = 5$ and polyphase filter lengths $V_0 = 5$, $V_1 = 2$, $V_2 = 10$. It was designed using a rectangular default Doppler spectrum.

Fig. 2(a) shows the specified (Jakes) Doppler spectrum $S_0(\nu)$ of the tap process $h_0[n]$ as well as an estimate of the Doppler spectrum of the simulated channel (estimated from 200 realizations of $h_0[n]$ with length 10^6 each.) It can be seen that the approximation is very accurate. Fig. 2(a) also shows the estimated Doppler spectrum obtained with an ARMA channel simulator that used the same ARMA model order ($P = 20$, $Q = 100$) but *no* subsampling/interpolation. It is seen that for a given ARMA model order, the sub-

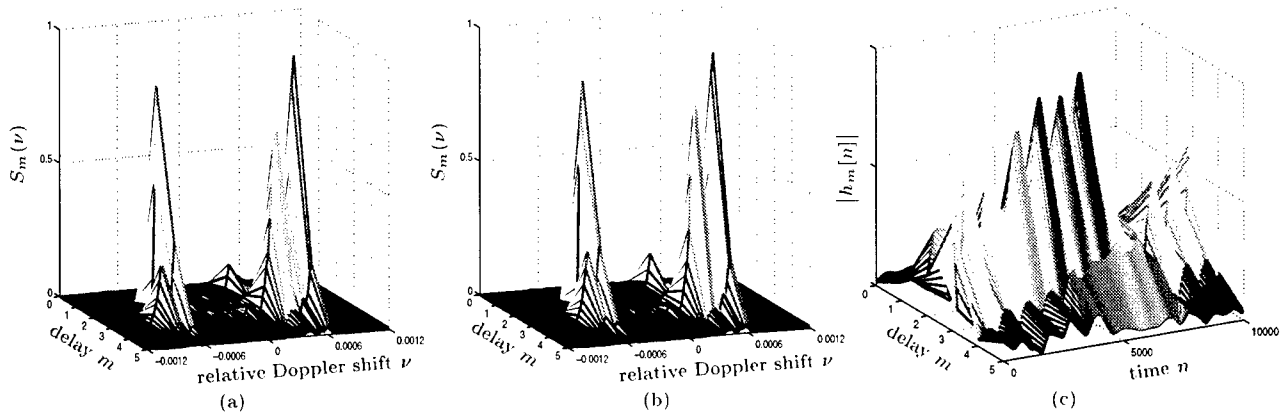


Figure 4: *Simulation of a realistic channel: (a) Specified scattering function (estimated from measured channel data); (b) simulated scattering function (estimated from realizations of the simulated impulse response); (c) magnitude of a realization of the simulated impulse response.*

sampling/interpolation technique yields a substantial performance improvement. (In fact, without subsampling/interpolation, an MA order of $Q = 10^4 \cdots 10^5$ would be needed to obtain a comparable performance.)

Fig. 2(b) shows an estimate of the Doppler spectrum of the simulated tap process at the output of the second interpolator stage (i.e., before the final interpolator stage). Since the interpolation factor of the final interpolator stage is $L_2 = 5$, the tap process after the second interpolator stage has Doppler bandwidth $L_2\nu_{\max} = 5 \cdot 10^{-4}$. This shows that it is possible to simultaneously generate channels with equal Doppler profile but different Doppler bandwidths.

Fig. 3 compares the frequency responses of the subsampled ARMA filter obtained with our MA design method (see Subsection 2.2) and with Durbin's method [3, 4]. Identical AR coefficients were used. Note that these frequency responses do not include the interpolator; the Doppler bandwidth of 0.1 in Fig. 3 corresponds to a Doppler bandwidth of 10^{-4} after interpolation by $L = 1000$. It can be seen that our efficient method achieves slightly better stop-band attenuation than Durbin's method.

4.2. Simulation 2: Realistic Channel

Fig. 4(a) shows a specified scattering function that was estimated [14] from channel data measured in a suburban area.[1] To each one of the $M = 6$ specified Doppler spectra $S_m(\nu)$, $m = 0, 1, \cdots, 5$, we designed a corresponding subsampled ARMA filter of order $P = 50$ and $Q = 1000$ using parameters $N = 500$ and $\gamma = 10^{-5}$. The subsampling/interpolator parameters were as in Subsection 4.1. Fig. 4(b) shows an estimate of the scattering function $S_m(\nu)$ derived from 200 realizations of the simulated impulse response (tap processes) $h_m[n]$, $m = 0, 1, \cdots, 5$, each of length $5 \cdot 10^5$. It is seen that the channel simulator achieves a good approximation of the specified scattering function. Finally, a segment of a simulated impulse response $h_m[n]$ is shown in Fig. 4(c).

5. CONCLUSIONS

We have presented a technique for simulating time-varying mobile radio channels that is specifically suited to the small relative Doppler bandwidths encountered in wideband CDMA and OFDM communications. The combination of a "subsampled" ARMA innovations filter with a multistage interpolator was shown to yield substantial advantages regarding accuracy, efficiency, and flexibility.

[1] Courtesy of T-Nova Deutsche Telekom Innovationsgesellschaft mbH, Technologiezentrum Darmstadt, Germany.

REFERENCES

[1] W. C. Jakes, *Microwave Mobile Communications*. New York: Wiley, 1974.

[2] P. A. Bello, "Characterization of randomly time-variant linear channels," *IEEE Trans. Comm. Syst.*, vol. 11, pp. 360–393, 1963.

[3] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs (NJ): Prentice Hall, 1992.

[4] S. M. Kay, *Modern Spectral Estimation*. Englewood Cliffs (NJ): Prentice Hall, 1988.

[5] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs (NJ): Prentice Hall, 1983.

[6] J. K. Cavers, *Mobile Channel Characteristics*. Boston (MA): Kluwer, 2000.

[7] M. F. Pop and N. C. Beaulieu, "Limitations of sum-of-sinusoids fading channel simulators," *IEEE Trans. Comm.*, vol. 49, pp. 699–708, April 2001.

[8] P. Höher, "A statistical discrete-time model for the WSSUS multipath channel," *IEEE Trans. Veh. Technol.*, vol. 41, no. 4, pp. 461–468, 1992.

[9] G. Wetzker, U. Kaage, and F. Jondral, "A simulation method for Doppler spectra," in *5th IEEE Int. Symposium on Spread Spectrum Techniques and Applications*, pp. 517–521, 1998.

[10] G. B. Giannakis and C. Tepedelenlioğlu, "Basis expansion models and diversity techniques for blind identification and equalization of time-varying channels," *Proc. IEEE*, vol. 86, pp. 1969–1986, Oct. 1998.

[11] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing*. Englewood Cliffs (NJ): Prentice Hall, 1993.

[12] J. Salz, "Optimum mean-square decision feedback equalization," *Bell Syst. Tech. J.*, vol. 52, pp. 1341–1371, Oct. 1973.

[13] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1975.

[14] H. Artés, G. Matz, and F. Hlawatsch, "An unbiased scattering function estimator for fast time-varying channels," in *Proc. 2nd IEEE Workshop on Signal Processing Advances in Wireless Communications*, (Annapolis, MD), pp. 411–414, May 1999.

A MIXED MAP/MLSE RECEIVER FOR CONVOLUTIONAL CODED SIGNALS TRANSMITTED OVER A FADING CHANNEL

Langford B. White

Dept. of Electrical and Electronic Engineering
Adelaide University
Adelaide, SA 5005, Australia.
email : Lang.White@adelaide.edu.au

Robert J. Elliott

Department of Mathematical Sciences
University of Alberta
Edmonton, Alberta, Canada, T6G 2G1
email: relliott@ualberta.ca

ABSTRACT

This paper addresses the problem of estimating a rapidly fading convolutionally coded signal such as might be found in a wireless telephony or data network. We model both the channel gain and the convolutionally coded signal as Markov processes, and thus the noisy received signal as a hidden Markov process (HMP). Two now-classical methods for estimating finite-state hidden Markov processes are the Viterbi algorithm and the *a posteriori* probability (APP) filter. A hybrid recursive estimation procedure is derived whereby one hidden process (the encoder state in our application) is estimated using a Viterbi-type (ie sequence based) cost and the other (the fading process) using an APP based cost such as maximum *a posteriori* probability. Using simulations, performance of the optimal scheme is compared with a number of suboptimal techniques - decision directed Kalman and HMP predictors, and Kalman filter and HMP filter per-survivour processing (PSP) techniques. Superior performance of the optimal scheme is demonstrated with little extra computational requirement compared to the PSP techniques.

1. INTRODUCTION

In wireless telephony and data networks, propagation characteristics of the radio channel give rise to often rapid fluctuations in the received signal power [1]. For multilevel signalling constellations such as Pulse Amplitude Modulation (PAM) and Quadrature Amplitude Modulation (QAM), it is necessary for the receiver to have a good estimate of the instantaneous channel power gain in order to properly demodulate the signal. For many practical channels, the channel power gain may vary so quickly, that gain estimation methods based on a static model of the channel gain (eg adaptive methods, maximum likelihood) may not track sufficiently quickly to permit demodulation of the signal. Thus dynamic models for the channel gain should be applied in such cases. Dynamic models will give rise to estimation structures which are designed to track more quickly, and thus should improve performance. In this paper, we specify a finite state Markov chain to model the amplitude gain process.

Most wireless telecommunications signals employ forward error correction (FEC) at the physical layer to give protection against symbol errors introduced by noise on the channel. The most common type of FEC is convolutional coding [2]. Convolutional coding works by adding redundancy (linear dependence) into the transmitted symbol stream by multiple input - multiple output linear FIR filtering (modulo 2). The maximum delay in the filter is called the constraint length of the encoder. An encoder which produces n output bits for each m input bits is called a rate m/n encoder. Commonly used rates are 1/2, 3/4, 5/6 and 7/8, however for some applications (eg deep space communications) rates as low as 1/128 might be used. In this paper, we consider only rate $1/n, n \geq 2$ encoding. A convolutionally encoded signal may be represented as a hidden Markov model (HMM) with state consisting of all the input bits stored in the encoder memory, and observation consisting of the output symbol stream. The transition structure of the state is highly constrained. For example for a rate 1/2 encoder of constraint length M has 2^M states (corresponding to all possible combinations of the M stored input bits in the encoder), but there are only 2 possible transitions from each state, corresponding to the 2 possibilities for the next input bit. In such highly constrained problems, it is recognised that *Maximum Likelihood Sequence Estimation* (MLSE) should be used, leading to the well-known Viterbi algorithm (VA) [4], where it is demonstrated that MLSE yields (asymptotically) the optimal error performance.

In this paper we model our received signal as the product of the channel gain process and the convolutionally encoded process observed in additive white Gaussian noise. Thus we have an HMP dependent on two underlying Markov chains, one being the state of the convolutional encoder, and the other being the state of the channel gain process. We derive a optimal mixed estimation algorithm, whereby we seek MLSE for the encoder state, and maximum *a posteriori* probability (MAP) estimates for the channel gain process. Such an algorithm clearly involves joint estimation of both underlying Markov process states, albeit with different criteria used to determine each component. The MLSE for the encoder then allows us to extract the original input bit sequence.

As a comparison, we use two classes of suboptimal approaches. The simplest class is a decoupled structure consisting of an estimator for the channel gain process, combined with a standard MLSE algorithm applied to estimate the encoder state. This structure mimics in some sense the usual *automatic gain control* (AGC) commonly used in receivers. Decision feedback of delayed symbols is used to parameterise the channel gain estimator. The other suboptimal methods used are based on *Per-survivour Processing* (PSP) [8]. Here a bank of amplitude estimators are used ; each associated with a surviving candidate optimal path from the MLSE. There is no requirement for feedback of delayed (or otherwise) symbols with these PSP methods. Within each class, we inves-

0-7803-7011-2/01/$10.00 ©2001 IEEE

tigate the performance of 2 types of amplitude estimators. The first class is based on an AR(1) model for the amplitude process, and results in a Kalman filter based amplitude estimator. The same AR(1) model is used to derive the Kalman filter based PSP method similar to [9] (which also addresses the frequency selective fading case). In each case, the second order statistics of the Markov chain amplitude process are used to parameterise the Kalman filter(s). The other type of estimator uses the finite state Markov chain model itself to derive the corresponding HMP filter(s) for the amplitude process in both decision feedback and PSP modes of operation. Performance of the optimal and the 4 suboptimal techniques is compared with the aid of simulated 4 level Pulse Amplitude Modulated (PAM) signals.

2. SIGNAL AND CHANNEL MODEL

We will consider convolutionally coded signals with constraint length M. Denote by X_k the length M binary vector being the convolutional encoder state at sample time k. This process follows a 'shift-register' type behaviour so that for $k \geq 0$,

$$X_{k+1} = S\, X_k + e_1\, b_{k+1}. \tag{1}$$

Here S is the $M \times M$ shift matrix with $S_{ij} = 1$ if $i = j + 1$, and zero otherwise, and e_1 is the unit vector in $\mathbb{R}^M$ with unity in the first position. The sequence $\{b_k\}$ denotes the input binary message stream which is independent and takes the values 0 and 1 with equal probability. Consequently, the state space of X has $2^M \equiv N^{(2)}$ binary vectors. This state space can be identified with the set $\{e_1^{(2)}, \ldots, e_{N^{(2)}}^{(2)}\}$ of unit vectors in $\mathbb{R}^{N^{(2)}}$. We shall write $X^{(2)}$ for the version of X defined in the canonical space $\{e_1^{(2)}, \ldots, e_{N^{(2)}}^{(2)}\}$. Each basis vector $e_i^{(2)}$ corresponds to one binary vector in $\{0,1\}^M$. Each binary vector corresponds to a decimal integer, so we shall choose the (decimal) under i so that e_i is associated with the corresponding binary vector. Any vector X_k has only two possible successor states. The transition matrix $A^{(2)}$ for $X^{(2)}$, therefore, is sparse with elements

$$a_{ij}^{(2)} = \begin{cases} 0.5 & i = 2j \bmod N^{(2)} \\ 0.5 & i = 2j+1 \bmod N^{(2)} \\ 0 & \text{else.} \end{cases} \tag{2}$$

The encoder operates at rate $1/P$, $P \geq 2$, [1] with generator matrix $G : \{0,1\}^M \to \{0,1\}^P$. Suppose $\mathcal{M} : \{0,1\}^P \to \{q_1, \ldots, q_{2^P}\}$ where q_i is real, denotes the modulation operation. Its task is to map the 2^P possible values of the encoder output onto 2^P real symbol values which may be transmitted. With minor modifications we can also handle complex modulation types such as Quadrature Amplitude Modulation (QAM). The transmitted signal is then

$$x_k = \mathcal{M}(GX_k \bmod 2). \tag{3}$$

The transmitted signal is propagated through a flat fading channel, which acts on the channel as a multiplicative gain [1]. The fading process is here modelled as a finite state Markov chain taking values in the set $\{a_1, a_2, \ldots, a_{N^{(1)}}\}$ where $0 = a_1 < a_2 < \cdots < a_{N^{(1)}}$. [2] We provide some justification for the choice of such a model in [7]. An additional reason for such a choice is the applicability of an estimation theory based on the Expectation-Maximisation algorithm [3] which we address in forthcoming work [6]. Write $a = (a_1, \ldots, a_{N^{(1)}})^T \in \mathbb{R}^{N^{(1)}}$ and suppose the chain determining the fading dynamics is $X^{(1)} = \{X_k^{(1)}\}$ where $X_k^{(1)}$ takes values in the (canonical) set of unit vectors

$$\{e_1^{(1)}, \ldots, e_{N^{(1)}}^{(1)}\} \in \mathbb{R}^{N^{(1)}}. \tag{4}$$

Then the real value (gain) associated with the fading channel at time t is $\langle X_k^{(1)}, a\rangle$. We suppose the transition matrix $A^{(1)} = (a_{ji}^{(1)})$ of $X^{(1)}$ has entries

$$a_{ji}^{(1)} = \begin{cases} p, & i = 1,\ j = 2 \\ q, & i = 2,\ j = 1 \\ \lambda, & i = 2, \ldots, N^{(1)} - 1,\ j = i + 1 \\ \mu, & j = 2, \ldots, N^{(1)} - 1;\ i = j + 1 \end{cases} \tag{5}$$

with diagonal elements $a_{ii}^{(1)}$ chosen so that each column of $A^{(1)}$ sums to unity. The received signal is given by

$$y_k = \langle X_k^{(1)}, a\rangle\, \mathcal{M}(GX_k \bmod 2) + \sigma\, n_k \tag{6}$$

where the $\{n_k\}$ is a sequence of independent normal $\mathcal{N}(0,1)$ random variables, and σ^2 is the noise power. When X_k is in the state corresponding to the vector $e_i^{(2)}$, write $d = (d_1, \ldots, d_{N^{(2)}})^T$, with

$$d_i = \mathcal{M}(GX_k \mod 2) \tag{7}$$

so that $\mathcal{M}(GX_k \mod 2) = d_i = \langle X_k^{(2)}, d\rangle$. Our observation process can be thus written in terms of the canonical state variables as

$$y_k = \langle X_k^{(1)}, a\rangle\, \langle X_k^{(2)}, d\rangle + \sigma\, n_k. \tag{8}$$

We assume all parameters $a, d, \sigma^2, p, q, \lambda$ and μ are known. Adaptive estimation is addressed in [6].

2.1. Optimal Demodulation

Given the observations $\mathcal{Y}_k = \{y_0, y_1, \ldots, y_k\}$ we wish to obtain recursive estimates for $X_k^{(1)}$ and $X_k^{(2)}$, perhaps with some delay $\Delta \geq 0$. If one was interested in minimum variance or maximum *a posteriori* probability (MAP) estimation of both the underlying Markov chain states, one would proceed to determine a recursive update for the joint *a posteriori* probabilities

$$q_k(i,j) = \Pr\left\{X_{k-\Delta}^{(1)} = e_i^{(1)}, X_{k-\Delta}^{(2)} = e_j^{(2)} | \mathcal{Y}_k\right\}, \tag{9}$$

and then compute the associated conditional expectations or MAP estimates.. In the usual Viterbi algorithm (dynamic programming), computation of (9) is replaced by a sequential maximisation over all possible sample paths of $X_t^{(1)}$ and $X_t^{(2)}$ for $t = 0, \ldots, k - \Delta$. The new mixed estimation procedure proposed in this paper consists of using the *a posteriori* probability estimates for the $X^{(1)}$ process, coupled with a Viterbi maximum likelihood sequence estimation criterion for $X^{(2)}$. Formally, this means considering quantities of the form

[1] More general rates can also be dealt with using a multiple input version of (1).

[2] The zero amplitude state is included to permit detection of the presence of the signal or otherwise, if desired.

$$\tilde{q}_k(i,j) = \max_{X_0^{(2)},\ldots,X_{k-1}^{(2)}} \Pr\left\{X_0^{(2)},\ldots,X_{k-1}^{(2)}, X_k^{(1)} = e_i^{(1)}, X_k^{(2)} = e_j^{(2)} | \mathcal{Y}_k\right\}. \quad (10)$$

Candidate optimal sequences for $X_k^{(2)}$ are obtained in the usual MLSE manner, except that the each time in the backtracking phase, MAP estimates are obtained for $X_k^{(1)}$ by maximisation (over i) of (10). We have the following recursion [7] :

$$\tilde{q}_{k+1}(i,j) = \max_{1\le\ell\le N^{(2)}} a_{j\ell}^{(2)} \sum_{n=1}^{N^{(1)}} a_{in}^{(1)}\, \tilde{q}_k(n,\ell)\, \frac{\phi\left(\frac{y_{k+1}-a_i d_j}{\sigma}\right)}{\sigma\phi(y_{k+1})}. \quad (11)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$. Initialisation at $k = 0$ is given by

$$\tilde{q}_0(i,j) = \frac{\phi\left(\frac{y_0-a_i d_j}{\sigma}\right)}{\sigma\,\phi(y_0)}\, \pi^{(1)}(i)\,\pi^{(2)}(j), \quad (12)$$

where $\pi^{(1)}$ and $\pi^{(2)}$ are the initial probability distributions for $X^{(1)}$ and $X^{(2)}$ respectively. At each time point, we keep track of the maximising index in (11), ie let

$$\Psi_k(i,j) = \operatorname*{argmax}_{1\le\ell\le N^{(2)}} a_{j\ell}^{(2)} \sum_{n=1}^{N^{(1)}} a_{in}^{(1)}\, \tilde{q}_k(n,\ell). \quad (13)$$

We also keep track of the maximising value of $X^{(1)}$ for each value of $X^{(2)}$,

$$\eta_k(j) = \operatorname*{argmax}_{1\le i\le N^{(1)}} \tilde{q}_k(i,j). \quad (14)$$

Estimates $\hat{X}_{k-\Delta|k}^{(1)}$ and $\hat{X}_{k-\Delta|k}^{(2)}$ (of $X_{k-\Delta}^{(1)}$ and $X_{k-\Delta}^{(2)}$ respectively) are produced by backtracking by a fixed number Δ samples at each time $k > \Delta$:

$$\begin{aligned} (i*, j*) &:= \operatorname{argmax}_{i,j}\, \tilde{q}_k(i,j) \\ \text{For } s = k-1,\ldots,k-\Delta & \\ j* &:= \Psi_{s+1}(i*, j*), \qquad i* := \eta_s(j*) \\ \text{Then :} & \\ \hat{X}_{k-\Delta}^{(2)} &= e_{j*}^{(2)}, \qquad \hat{X}_{k-\Delta}^{(1)} = e_{i*}^{(1)}. \end{aligned} \quad (15)$$

The backtracking delay is necessary to enable proper construction of the maximum likelihood sequence. This delay is chosen sufficiently large that all candidate optimal sequences backtracking from time k have merged at time $k-\Delta$. Thus in order to apply the algorithm, the quantities $\tilde{q}_k(i,j)$ are initialised at time $k = 0$ according to (12), and updated for each time $k > 0$ via (11). At each time we also retain maximising indices via (13) and (14). Backtracking also takes place at each time $k > \Delta$ according to (15) to extract the desired estimates.

2.2. Reduced Complexity Filters

The reader is referred to [7] for details of the various suboptimal filters used here.

3. SIMULATIONS

In this section we present results of simulation experiments used to compare 6 demodulators applied to the fading convolutionally coded signal described above. The performance of the optimal scheme, the Kalman and HMP PSP techniques, the Kalman and HMP predictor based methods, and usual MLSE with the amplitude process known to the receiver were compared. The Kalman and HMM predictor methods used decision delay $\Delta = 1$, which we argue later is the best value to choose, at least in the Kalman case. In our experiments, we did not observe any statistically significant difference between the performance of the Kalman filter based methods and the corresponding HMP based methods, ie the Kalman predictor method performed similarly to the HMP predictor method, and similarly for the PSP techniques.

The resulting Bit Error Rates (BER) are shown in figure 1. Figure 2 repeats for a more rapidly varying amplitude case. It is seen that in both cases, the predictor based methods perform the worst, with the PSP methods yielding performance in between that of the predictor methods and the optimal method. The optimal technique performs quite close to the case where the receiver knows the fading process exactly. The performance gain in using the optimal filter appears to increase for higher SNRs.

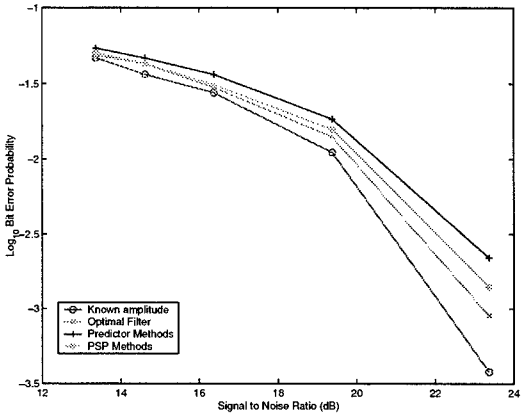


Fig. 1. Bit Error Rate Performance for Filter ($\mu = \lambda = 0.05$)

We also examined the error behaviour of the Kalman predictor method as a function of the parameter Δ. Recall that $\Delta \ge 1$ denotes the time lag (in samples) until we make a decision about the encoder state. This value is used to predict the amplitude process (gain) value forward from the Kalman filter to the Viterbi decoder. Figure 3 shows rather interesting behaviour in that the smallest possible $\Delta = 1$ resulted in the best overall BER performance. Here $\mu = \lambda = 0.1$, and the SNR was 29 dB. Clearly, larger smoothing lags, which one would normally expect to result in better state estimates (for the encoder process) [5] are not resulting in better performance of the overall scheme. We may conclude that the behaviour evident in figure 3 is due to the poor prediction performance of the Kalman method. This is to be expected since it is not generally possible to accurately predict a discrete state HMM. We conclude that some sort of joint estimation procedure (either explicit as in our optimal approach, or implicit as in PSP) is really necessary to obtain reasonable performance with

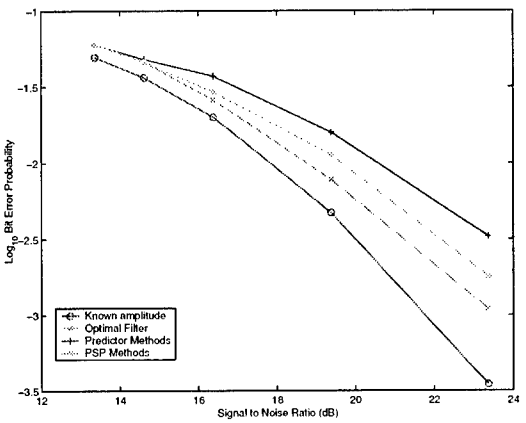


Fig. 2. Bit Error Rate Performance for Filter ($\mu = \lambda = 0.2$)

the model we have assumed for the fading channel amplitude process. Computational complexity for the decision directed methods is low since only one amplitude tracking filter is required. For PSP and optimal processing, the complexity is greater by approximately the number of convolutional encoder states (ie 2^M) since that number of tracking filters are required.

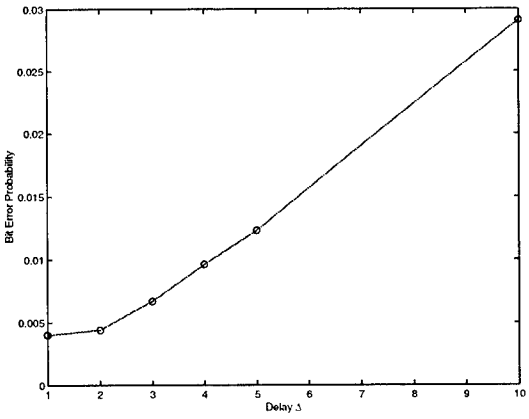


Fig. 3. Effect of parameter Δ on the Kalman predictor method

4. CONCLUSION

In this paper we have derived the optimal filter for a hidden Markov process consisting of the product of two statistically independent underlying Markov chains observed in additive white Gaussian noise, which may have state dependent moments. We apply a mixed estimation criterion in order to formulate the filter. We seek the *Maximum Likelihood Sequence* corresponding to one of the underlying chains, and *a posteriori* probabilities (APPs) for the other underlying chain. This mixed criterion is motivated by a particular application, namely the demodulation of a rapidly fading convolutionally coded communications signal. The signal is decoded using maximum likelihood sequence estimation (MLSE). Estimation of the fading process is performed according to the maximum *a posteriori* probability criterion, requiring computation of APPs. The performance of the optimal filter for this example is compared to a more conventional approach consisting of decoupled estimators for each underlying chain. These estimators are standard MLSE implemented via the Viterbi algorithm for the convolutionally coded part, and a decision-directed predictor for the gain process. The case where the gain process is known to the receiver is used as a benchmark. We also compare performance with a per-survivour processing (PSP) technique which has computational complexity less than the optimal method, but greater than the simple prediction technique. In both the prediction and PSP methods, we examined both Kalman and hidden Markov process based approaches, and found no significant difference in performance between them in each case. The PSP approach has been addressed in [9], which also considers frequency selective fading. Simulations show that the predictor methods performs worst but the optimal filter illustrates minimal performance degradation as compared to the known amplitude case. The PSP technique offers performance between that of the simple prediction method, and the optimal method. In this paper, we have not addressed the issue of estimating the fading process model parameters. This problem is being addressed in current work [6]. We have also not addressed frequency selective fading here, but indicate that the same idea as presented here could be applied to such cases, albeit with a substantial increase in computational requirements.

5. REFERENCES

[1] W. C. Jakes, *Microwave mobile communications*, New York, USA : Wiley, 1974.

[2] B. Sklar, *Digital Communications, Fundamentals and Applications*. Englewood Cliffs, NJ : Prentice-Hall, 1988.

[3] R.J. Elliott, L. Aggoun and J.B. Moore, *Hidden Markov Models: Estimation and Control*, New York, NY : Springer-Verlag, 1995.

[4] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", *IEEE Trans. Information Theory*, v. IT-13, 1967, pp. 260-269.

[5] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Englewood Cliffs, USA : Prentice-Hall, 1979.

[6] L. B. White and R. J. Elliott, "Adaptive Parameter Estimation for Rapidly Fading Convolutional Coded Signals", in preparation.

[7] L. B. White and R. J. Elliott, "A Mixed MAP/MLSE receiver for convolutional coded signals transmitted over a fading channel", *IEEE Trans. Signal Processing*, in review.

[8] R. Raheli, A. Polydoros and C-K. Tzou, "Per-survivor processing : A general approach to MLSE in uncertain environments", *IEEE Trans. on Communications*, v. 42, no. 2/3/4, 1995, pp. 354-364.

[9] M. E. Rollins and S. J. Simmons, "Simplified per-survivor Kalman processing in fast frequency-selective fading channels", *IEEE Trans. Communications*, v. 45, no. 5, May 1997, pp. 544-553.

[10] L. B. White, "A comparison of optimal and Kalman filtering for a hidden Markov process", *IEEE Signal Processing Letters*, v. 5, no. 5, pp. 124-126, May 1998.

ORTHOGONAL EXTENSIONS OF AR PROCESSES WITHOUT ARTIFICIAL DISCONTINUITIES FOR SIZE-LIMITED FILTER BANKS

M.E. Domínguez Jiménez

E.T.S.I. Industriales
Universidad Politécnica de Madrid
28006 MADRID (Spain)
E-mail: edominguez@etsii.upm.es

N. González Prelcic

E.T.S.E.Telecomunicación
Universidade de Vigo
36200 VIGO (Spain)
E-mail: nuria@tsc.uvigo.es

ABSTRACT

This work is concerned with extension techniques of finite signals for subband processing using tree-structured filter banks. In many applications it is desirable that the selected extension defines an orthogonal transform. Although it is clear that periodization solves this problem, some complications arise when using this technique: spurious high frequencies or artificial discontinuities appear in the transform vector. Considering AR processes as input signals, the solution of this problem is an algorithm for the generation of alternative orthogonal signal extensions which do not introduce artificial discontinuities in the subband signals. Experimental results that illustrate the effectiveness of the proposed design method are discussed briefly.

1. INTRODUCTION AND NOTATION

The use of tree-structured paraunitary filter banks for processing finite length signals needs of specific techniques for handling the boundaries in order to ensure perfect reconstruction and orthogonality [1, 2, 5, 6, 7]. In the previous literature two approaches have been proposed to overcome this problem: signal extension methods [6] and boundary filters [5]. Among the traditional signal extensions the only one that preserves the orthogonality of the transformation is the periodization, but it is known to introduce artificial discontinuities in the transform domain, a very annoying effect in many applications. On the other hand, although the works on boundary filters provide alternative solutions that also lead to orthogonal transformations, the problem of artificial discontinuities is not solved either.

In our recent work [3, 4] we have proposed an algorithm for the design of orthogonal extensions different to the classical periodization, but the new solutions does not necessarily provide transformations which do not present these spurious high frequencies. Thus, this paper is a continuation of our previous work; considering AR processes as input signals, it provides an efficient scheme for the design of orthogonal extensions which do not introduce artificial discontinuities in the transform domain.

Before starting, we summarize the notation and recall a few results from our previous papers which are necessary to follow the development of the new algorithm. We will consider only real valued signals and filters. Boldface lowercase letters will denote vectors and boldface uppercase ones will denote matrices. We use $\mathbf{H}_{m\times n}$ to represent an m rows n columns matrix; the Nth-order null and identity matrices are respectively denoted by $\mathbf{O}_N$ and $\mathbf{I}_N$.

We will consider the paraunitary filter bank given by the low pass filter $\mathbf{h} = [h(0), h(1), \cdots, h(L-1)]$ and the associated high pass filter $\mathbf{g} = [h(L-1), -h(L-2), \cdots, -h(0)]$, assuming that $L = 2K + 2$, with K even. From these filters we can construct the matrix $\mathbf{H}_{m\times(m+2K)}$. As shown in the previous literature [5], $\mathbf{H}_{m\times(m+2K)}$ can be written as a block Toeplitz form:

$$\begin{bmatrix} \mathbf{A}_K & \dots & \mathbf{A}_0 & \mathbf{0}_2 & \dots & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{A}_K & \dots & \mathbf{A}_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \dots & \ddots & \mathbf{0}_2 \\ \mathbf{0}_2 & \dots & \mathbf{0}_2 & \mathbf{A}_K & \dots & \mathbf{A}_0 \end{bmatrix}$$

where, for all $j = 0, \dots, K$,

$$\mathbf{A}_j = \begin{bmatrix} h(2j+1) & h(2j) \\ -h(L-2j-2) & h(L-2j-1) \end{bmatrix}.$$

In the other way, $\mathbf{H}_{K\times 3K}$ can be split into three block-Toeplitz submatrices of order K: $\mathbf{H}_{K\times 3K} = [\mathbf{D}\ \mathbf{E}\ \mathbf{F}]$. $\mathbf{D}$ and $\mathbf{F}$ are, respectively, upper and lower block triangular matrices [3]. Moreover, we can write $\mathbf{D} = \mathbf{Q_1}\mathbf{K_D}\mathbf{P_1}$, $\mathbf{F} = \mathbf{Q_0}\mathbf{K_F}\mathbf{P_0}$ and

$$\mathbf{E} = \mathbf{Q_1}\mathbf{K_D}\mathbf{C}\mathbf{P_0} - \mathbf{Q_0}\mathbf{K_F}\mathbf{C}^T\mathbf{P_1},$$

where $[\mathbf{Q_0}\ \mathbf{Q_1}]$ and $[\mathbf{P_0}^T\mathbf{P_1}^T]$ are orthogonal, and $\mathbf{K_F}$, $\mathbf{K_D}$ and $\mathbf{C}$ are square matrices of order $K/2$.

Let us consider a finite signal $\mathbf{x}$ of even length $N \geq 2K$: $\mathbf{x} = [\mathbf{x}(0), \mathbf{x}(1), \cdots, \mathbf{x}(N-1)]^T = [\mathbf{x}_a^T\ \mathbf{x}_c^T\ \mathbf{x}_b^T]^T$, where

0-7803-7011-2/01/$10.00 ©2001 IEEE

$\mathbf{x}_a$ and $\mathbf{x}_b$ contain, respectively, the first and last K components of $\mathbf{x}$, and $\mathbf{x}_c$ the remaining central ones. We define an extension of $\mathbf{x}$ as the vector $\mathbf{x}_e = [\mathbf{x}_l^T, \mathbf{x}^T, \mathbf{x}_r^T]^T$. We will study linear extensions of the type $\mathbf{x}_l = \mathbf{C}^{l,a}\mathbf{x}_a + \mathbf{C}^{l,b}\mathbf{x}_b$ and $\mathbf{x}_r = \mathbf{C}^{r,a}\mathbf{x}_a + \mathbf{C}^{r,b}\mathbf{x}_b$, where $\mathbf{C}^{l,a}$, $\mathbf{C}^{l,b}$ and $\mathbf{C}^{r,a}$, $\mathbf{C}^{r,b}$ are, respectively, the left and right extension matrices.

Throughout this paper we want to study the transformation of the extended vector, i.e., $\mathbf{y}_e = \mathbf{H}_{N\times(N+2K)}\mathbf{x}_e$. This amounts to processing the signal $\mathbf{x}_e$ by means of the analysis filter bank given by $\mathbf{h}$ and $\mathbf{g}$, only retaining the N central output samples. The whole transformation of the original signal $\mathbf{x}$ can be expressed as $\mathbf{y}_e = \mathbf{G}\mathbf{x}$. It has been proven [3] that the transformation is orthogonal if and only if

$$\mathbf{G} = \begin{bmatrix} \mathbf{DC}^{l,a}+\mathbf{E} & \mathbf{F}\,\mathbf{0}_{K\times(N-3K)} & \mathbf{DC}^{l,b} \\ & \mathbf{H}_{(N-2K)\times N} & \\ \mathbf{FC}^{r,a} & \mathbf{0}_{K\times(N-3K)}\,\mathbf{D} & \mathbf{E}+\mathbf{FC}^{r,b} \end{bmatrix}$$

and there exists a unitary matrix $\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \\ \mathbf{V}_3 & \mathbf{V}_4 \end{bmatrix}$ of order K such that

$$\begin{aligned} \mathbf{DC}^{l,a} &= \mathbf{Q}_1(\mathbf{V}_1 - \mathbf{K_D C})\mathbf{P_0} \\ \mathbf{DC}^{l,b} &= \mathbf{Q}_1\mathbf{V}_2\mathbf{P}_1 \\ \mathbf{FC}^{r,a} &= \mathbf{Q_0}\mathbf{V}_3\mathbf{P_0} \\ \mathbf{FC}^{r,b} &= \mathbf{Q_0}(\mathbf{V}_4 + \mathbf{K_F C}^T)\mathbf{P}_1. \end{aligned} \tag{1}$$

2. DESIGN OF ADAPTIVE ORTHOGONAL TRANSFORMS WITHOUT ARTIFICIAL DISCONTINUITIES

Our aim is to construct an orthogonal extension which does not introduce artificial discontinuities in the transform domain. We consider that the input signal corresponds to an AR process, so that we can assume that when extending the original signal by using linear prediction techniques, the subband vector does not present spurious high frequencies. Unfortunately, the extension obtained by means of linear prediction is not orthogonal. Therefore, our design problem can be formulated as the search for the orthogonal extension that leads to a transform vector as similar as possible to the transform resulting from an extension by linear prediction.

2.1. Design of the transformation matrix

Let $\mathbf{x} = [\mathbf{x}_a^T, \mathbf{x}_{a'}^T, \mathbf{x}_{c'}^T, \mathbf{x}_{b'}^T, \mathbf{x}_b^T]^T$ be a signal of length N, being $\mathbf{x}_a, \mathbf{x}_{a'}, \mathbf{x}_{b'}, \mathbf{x}_b$ of length $K = L/2 - 1$. Let us consider the signal $\mathbf{x}_{pr}$ which comes from extending $\mathbf{x}$ at each border by means of a Kth order linear predictor. In this way, if $\mathbf{x}$ is an AR process of order K, so is $\mathbf{x}_{pr}$. In other words, if $\mathbf{r}$ is the vector containing the K autoregression coefficients, we construct $\mathbf{x}_{pr}$ by appending K samples $\mathbf{C_r x}_b$ at the right border of $\mathbf{x}$. $\mathbf{C_r}$ is the K-th order power of the companion matrix associated to $\mathbf{r}$. Analogously, we can obtain the vector $\mathbf{l}$ and extend the left border with the vector $\mathbf{C_l x}_a$[1]. In this way,

$$\mathbf{x}_{pr} = [(\mathbf{C_l x}_a)^T, \mathbf{x}_a^T, \mathbf{x}_{a'}^T, \mathbf{x}_{c'}^T, \mathbf{x}_{b'}^T, \mathbf{x}_b^T, (\mathbf{C_r x}_b)^T]^T.$$

On the other hand, let us consider any orthogonal extension $\mathbf{x}_e$ from $\mathbf{x}$,

$$\mathbf{x}_e = [\mathbf{x}_l^T, \mathbf{x}_a^T, \mathbf{x}_{a'}^T, \mathbf{x}_{c'}^T, \mathbf{x}_{b'}^T, \mathbf{x}_b^T, \mathbf{x}_r^T]^T.$$

Now, we impose the respective transform vectors, $\mathbf{Hx}_{pr}$ and $\mathbf{Hx}_e = \mathbf{Gx} = \mathbf{y}_e$, to be as close as possible; this problem can be formulated as the minimization of the euclidean norm

$$\|\mathbf{Hx}_{pr} - \mathbf{y}_e\|.$$

Both vectors are equal except from the first and last K samples; thus, the first K coefficients of the error vector are

$$(\mathbf{DC}^{l,a} - \mathbf{DC_l})\mathbf{x}_a + \mathbf{DC}^{l,b}\mathbf{x}_b$$

whereas the last K coefficients can be written as

$$(\mathbf{FC}^{r,b} - \mathbf{FC_r})\mathbf{x}_b + \mathbf{FC}^{r,a}\mathbf{x}_a$$

Now we must find the extension matrices $\mathbf{C}^{l,a}$, $\mathbf{C}^{l,b}$, $\mathbf{C}^{r,a}$, $\mathbf{C}^{r,b}$ which minimize the norm $\|\mathbf{Hx}_{pr} - \mathbf{y}_e\|^2$. By using the identities (1) and the fact that $\mathbf{Q_0}$ and $\mathbf{Q_1}$ have orthonormal columns, we get the following expression for the norm to be minimized:

$$\begin{aligned} &\|((\mathbf{V}_1 - \mathbf{K_D C})\mathbf{P_0} - \mathbf{K_D P_1 C_l})\mathbf{x}_a + \mathbf{V}_2\mathbf{P}_1\mathbf{x}_b\|^2 + \\ &+ \left\|((\mathbf{V}_4 + \mathbf{K_F C}^T)\mathbf{P}_1 - \mathbf{K_F P_0 C_r})\mathbf{x}_b + \mathbf{V}_3\mathbf{P_0}\mathbf{x}_a\right\|^2, \end{aligned}$$

which can also be written in a matrix-vector form:

$$\left\| \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \\ \mathbf{V}_3 & \mathbf{V}_4 \end{bmatrix} \begin{bmatrix} \mathbf{P_0}\mathbf{x}_a \\ \mathbf{P}_1\mathbf{x}_b \end{bmatrix} - \begin{bmatrix} \mathbf{K_D}(\mathbf{P_1 C_l} + \mathbf{C P_0})\mathbf{x}_a \\ \mathbf{K_F}(\mathbf{P_0 C_r} - \mathbf{C}^T\mathbf{P}_1)\mathbf{x}_b \end{bmatrix} \right\|^2.$$

Taking into account that the first matrix, $\mathbf{V}$, is unitary, we can finally formulate the minimization problem as:

$$\min_{\mathbf{V}\ \text{unitary}} \|\mathbf{Va} - \mathbf{b}\|$$

being

$$\mathbf{a} = \begin{bmatrix} \mathbf{P_0}\mathbf{x}_a \\ \mathbf{P}_1\mathbf{x}_b \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{K_D}(\mathbf{P_1 C_l} + \mathbf{C P_0})\mathbf{x}_a \\ \mathbf{K_F}(\mathbf{P_0 C_r} - \mathbf{C}^T\mathbf{P}_1)\mathbf{x}_b \end{bmatrix}.$$

The Cauchy-Schwartz inequality and the fact that $\mathbf{V}$ is unitary, assure that

$$\|\mathbf{Va} - \mathbf{b}\|^2 \geq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\|\|\mathbf{b}\| = (\|\mathbf{a}\| - \|\mathbf{b}\|)^2.$$

[1] Note that there exists a clear relationship between $\mathbf{l}$ and $\mathbf{r}$.

The minimum error $|\|\mathbf{a}\| - \|\mathbf{b}\||$ is reached if and only if $\mathbf{Va}$ and $\mathbf{b}$ are proportional, so it suffices to take

$$\mathbf{V} \text{ unitary such that } \mathbf{Va} = \frac{\|\mathbf{a}\|}{\|\mathbf{b}\|}\mathbf{b}. \qquad (2)$$

In order to build a Kth order unitary matrix $\mathbf{V}$, we take first the normalized vectors $\mathbf{a_1} = \mathbf{a}/\|\mathbf{a}\|$ and $\mathbf{b_1} = \mathbf{b}/\|\mathbf{b}\|$; secondly, by using a Gram-Schmidt procedure, we build any unitary basis of $\mathbb{R}^K$ whose first vector is $\mathbf{a_1}$, and similarly, any unitary basis beginning with $\mathbf{b_1}$. The matrix which transforms the first basis into the second one is $\mathbf{V}$. If we want $\mathbf{V}$ to be a product of rotations, as in Givens parameterization, the number of degrees of freedom for constructing $\mathbf{V}$ is $\frac{(K-1)(K-2)}{2}$ (although $\mathbf{V}$ has size K, the condition $\mathbf{Va_1} = \mathbf{b_1}$ lead us to the problem of building a unitary matrix of size $K-1$, so we need, $\frac{1}{2}(K-1)(K-2)$ parameters).

Remark: Although there exist infinite unitary matrices $\mathbf{V}$ which satisfy (2) and, hence, for which the minimum error is reached, we must remark that all of them lead to a unique transform vector $\mathbf{y}_e = \mathbf{Gx}$. Thus, the proposed method provides a unique subband transform vector $\mathbf{y}_e$, related to the original signal $\mathbf{x}$. Nevertheless, the lack of unicity of $\mathbf{V}$ implies that the expression of the new adaptive orthogonal matrix $\mathbf{G}$ is not unique. Its first and last K rows may vary. Moreover, if we regard that these rows contain the border filters associated to the orthogonal transform, we conclude that this method lead to the construction of an infinite number of orthogonal boundary filters which do not introduce artificial discontinuities. In other words, the proposed solution can be considered as the first design method for this kind of adaptive orthogonal boundary filters.

2.2. Error estimation

The expression for the minimum error in the frequency domain is $e_f = |\|\mathbf{a}\| - \|\mathbf{b}\||$, which depends only on the signal and the prototype filter. The method proposed in the previous section provides the orthogonal extension that minimizes this error. On one hand, we know that

$$\begin{aligned}\|\mathbf{b}\|^2 &= \|\mathbf{K_D}(\mathbf{P_1C_l} + \mathbf{CP_0})\mathbf{x}_a\|^2 \\ &+ \left\|\mathbf{K_F}(\mathbf{P_0C_r} - \mathbf{C}^T\mathbf{P_1})\mathbf{x}_b\right\|^2,\end{aligned}$$

$$\|\mathbf{a}\|^2 = \|\mathbf{P_0}\mathbf{x}_a\|^2 + \|\mathbf{P_1}\mathbf{x}_b\|^2;$$

and defining $m = \|\mathbf{b}\|^2 - \|\mathbf{a}\|^2$, it can be shown, by using the relations between $\mathbf{K_D}, \mathbf{K_F}$ and $\mathbf{C}$ [3], that

$$-\left(\|\mathbf{P_0}\mathbf{x}_a\|^2 + \|\mathbf{P_1}\mathbf{x}_b\|^2\right) \le m \le \|\mathbf{C_l}\mathbf{x}_a\|^2 + \|\mathbf{C_r}\mathbf{x}_b\|^2.$$

On the other hand,

$$e_f = \left|\|\mathbf{a}\| - \sqrt{\|\mathbf{a}\|^2 + m}\right| \le \sqrt{|m|};$$

therefore, we can write the following bound for the error:

$$e_f \le \max\left(\left\|\begin{bmatrix}\mathbf{P_0}\mathbf{x}_a \\ \mathbf{P_1}\mathbf{x}_b\end{bmatrix}\right\|, \left\|\begin{bmatrix}\mathbf{C_l}\mathbf{x}_a \\ \mathbf{C_r}\mathbf{x}_b\end{bmatrix}\right\|\right).$$

We can observe that this bound depends only on the behavior of the original signal near its edges: if the absolute values of the original samples are small at the borders, so will the error bound.

2.3. Generation of the extended signal

We are interested in the design of the associated extended vector in the time domain $\mathbf{x}_e = [\mathbf{x}_l^T, \mathbf{x}^T, \mathbf{x}_r^T]^T$. Let $\alpha = \|\mathbf{a}\|/\|\mathbf{b}\|$, and let $\mathbf{V}$ be any unitary matrix such that $\mathbf{Va} = \alpha\mathbf{b}$. The submatrices of $\mathbf{V}$ satisfy

$$\begin{cases}\mathbf{V_1P_0}\mathbf{x}_a + \mathbf{V_2P_1}\mathbf{x}_b = \alpha\mathbf{K_D}(\mathbf{P_1C_l} + \mathbf{CP_0})\mathbf{x}_a, \\ \mathbf{V_3P_0}\mathbf{x}_a + \mathbf{V_4P_1}\mathbf{x}_b = \alpha\mathbf{K_F}(\mathbf{P_0C_r} - \mathbf{C}^T\mathbf{P_1})\mathbf{x}_b.\end{cases}$$

If we left multiply these identities by $\mathbf{Q_1}$ and $\mathbf{Q_0}$, respectively, and apply (1), we get

$$\begin{cases}\mathbf{D}\mathbf{x}_l = \alpha\mathbf{DC_l}\mathbf{x}_a + (\alpha-1)\,\mathbf{Q_1K_DCP_0}\mathbf{x}_a, \\ \mathbf{F}\mathbf{x}_r = \alpha\mathbf{FC_r}\mathbf{x}_b - (\alpha-1)\,\mathbf{Q_0K_F}\mathbf{C}^T\mathbf{P_1}\mathbf{x}_b.\end{cases}$$

From these identities we derive that $\mathbf{Dx}_l, \mathbf{Fx}_r$ (and, therefore, the whole transform vector $\mathbf{y}_e$), no longer depend on the expression of the matrix $\mathbf{V}$, whenever $\mathbf{V}$ verifies (2). But, in the time domain, we obtain that $\mathbf{x}_l, \mathbf{x}_r$ are not completely determined: there exist arbitrary vectors $\mathbf{m}_1, \mathbf{m}_2$ such that

$$\begin{cases}\mathbf{x}_l = \alpha\mathbf{C_l}\mathbf{x}_a + (\alpha-1)\,\mathbf{P}_1^T\mathbf{CP_0}\mathbf{x}_a + \mathbf{P_0}^T\mathbf{m}_1 \\ \mathbf{x}_r = \alpha\mathbf{C_r}\mathbf{x}_b - (\alpha-1)\,\mathbf{P}_0^T\mathbf{C}^T\mathbf{P_1}\mathbf{x}_b + \mathbf{P_1}^T\mathbf{m}_2\end{cases}$$

so the extended vector $\mathbf{x}_e$ cannot be defined in a unique way. However, there is only one choice that best approximates $\mathbf{x}_{pr}$; in effect,

$$\|\mathbf{x}_e - \mathbf{x}_{pr}\|^2 = \|\mathbf{x}_l - \mathbf{C_l}\mathbf{x}_a\|^2 + \|\mathbf{x}_r - \mathbf{C_r}\mathbf{x}_b\|^2$$

and it can be easily shown that both norms are minimized by taking $\mathbf{m}_1 = (1-\alpha)\mathbf{P_0C_l}\mathbf{x}_a$ and $\mathbf{m}_2 = (1-\alpha)\mathbf{P_1C_r}\mathbf{x}_b$. For this choice, we obtain

$$\begin{cases}\mathbf{x}_l = \mathbf{C_l}\mathbf{x}_a + (\alpha-1)\,\mathbf{P_1}^T(\mathbf{CP_0} + \mathbf{P_1C_l})\mathbf{x}_a \\ \mathbf{x}_r = \mathbf{C_r}\mathbf{x}_b - (\alpha-1)\,\mathbf{P_0}^T(\mathbf{C}^T\mathbf{P_1} - \mathbf{P_0C_r})\mathbf{x}_b\end{cases}$$

Again, the error in the time domain $e_t = \|\mathbf{x}_e - \mathbf{x}_{pr}\|$ is proportional to $\alpha - 1 = \frac{\|\mathbf{a}\| - \|\mathbf{b}\|}{\|\mathbf{b}\|}$, and this quantity only depends on the signal and the prototype filters, not on the orthogonal extension.

3. EXPERIMENTAL RESULTS

We have applied the proposed method to a great variety of finite signals, considering Daubechies filters of different lengths as prototype filters. Some results that illustrate the performance of our method are shown in the figures presented here. The first test signal is the cubic spline presented in Figure 1(a), which corresponds to an AR-4 process. The output of the two channel cell obtained using the orthogonal extension proposed in this paper and length 10 Daubechies filters is displayed in Figure 1(b), while in Figure 1(c) we can observe the transform vector when using a periodic extension. It can be clearly observed that no artificial discontinuities appear when using our extension algorithm. A more realistic signal, that corresponds to an audio frame, is shown in Figure 2(a). The transform vectors using our orthogonal extension algorithm and periodization can be observed in Figures 2(b) and 2(c) respectively. In this case the transformations have been performed using length 26 Daubechies filters. Again, the performance of our method overcomes the periodic extension technique.

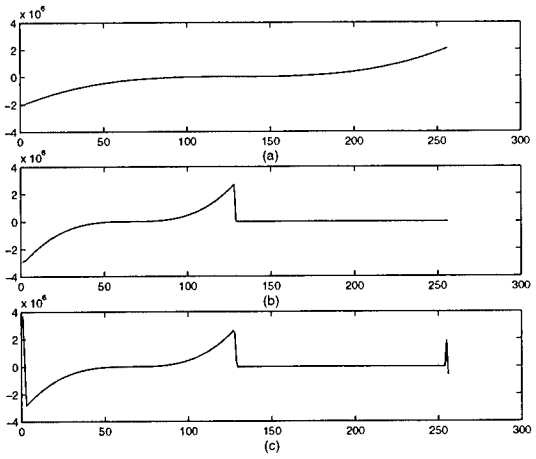


Fig. 1. (a) Cubic signal; (b) transform vector using our orthogonal extension method; (c) transform vector using periodic extension.

4. CONCLUSIONS

In this paper we have developed a technique for processing finite length signals with paraunitary filter banks without introducing artificial discontinuities in the subband signals. The two issues that have been considered are the design of the optimal transformation matrix and the generation of the corresponding extended signal. The design procedure is formulated as an optimization problem that can be analytically solved, so that the theory can be clearly developed. The absence of artificial discontinuities in the transform domain is clear from our tests, providing a great improvement in relation to existing orthogonal signal extension methods and boundary filters.

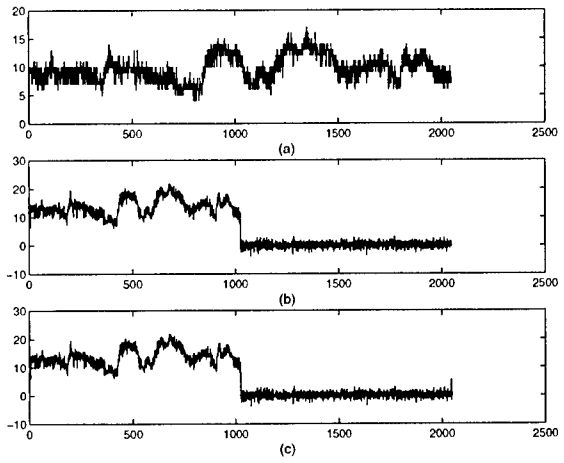


Fig. 2. (a) Original audio frame; (b) transform vector using the proposed orthogonal extension method; (c) transform vector using periodization.

5. REFERENCES

[1] R. H. Bamberger, S. L. Eddins, V. Nuri, "Generalized symmetric extensions for size-limited multirate filter banks", IEEE Transactions on Image Processing, Vol.3, pp. 82-87, Jan. 1994.

[2] A. Cohen, I. Daubechies, P. Vial, "Wavelets on the interval and fast wavelet transforms", Journal of Applied and Computational Harmonic Analysis, Vol. 1, pp 54-81, Dec. 1993.

[3] M. E. Domínguez, N. González Prelcic, "Design of non-expansionist and orthogonal extension methods for tree-structured filter banks", Proceedings of ICASSP, Vol I, pp. 532-535, 2000.

[4] M. E. Domínguez, N. González Prelcic, "New orthogonal extension methods for tree-structured filter banks", Proceedings of EUSIPCO, Vol. II, pp. 1073-1076, 2000.

[5] C. Herley, M. Vetterli, "Orthogonal time-varying filter banks and wavelet packets", IEEE Transactions on Signal Processing, Vol.42, n. 10, pp. 2650-2663, Oct. 1994.

[6] G. Karlsson and M. Vetterli, "Extension of finite length signals for sub-band Coding", Signal Processing, Vol. 17, pp. 161-168, 1989.

[7] R. L. de Queiroz, K. R. Rao, " On reconstruction methods for processing finite-length signals with paraunitary filter banks", IEEE Transactions on Signal Processing, Vol.43, pp. 2407-2410, Oct. 1995.

M-BAND PERFECT-RECONSTRUCTION LINEAR-PHASE FILTER BANKS

X. M. Xie, S. C. Chan and T. I. Yuk
Department of Electrical and Electronic Engineering
University of Hong Kong, Pokfulam Road, Hong Kong.
{xmxie, scchan, tiyuk}@eee.hku.hk

ABSTRACT

This paper studies the design of *M*-channel perfect-reconstruction (PR) linear-phase (LP) filter banks (FBs) with $M = 2^K$ using a tree-structured FB. It is based on a observation of Fliege [1], the length of the analysis filters is decreased by a factor of two when the depth of the tree is increased by one, while its transition bandwidth is increased by the same factor. A lattice-based 2-channel LP FB is chosen because the frequency responses of the lowpass and highpass analysis (synthesis) filters can be designed to be closely symmetric to the other around $\pi/2$. By properly selecting the filter length, transition bandwidth, and stopband attenuation of the 2-channel PR LP FBs at each level of the tree structure, it is possible to design uniform PR LP FB with excellent frequency characteristic and much lower system delay.

I. INTRODUCTION

Perfect-reconstruction (PR) linear-phase (LP) filter banks (FBs) are used in a wide range of applications, such as data compression, communications, and image and speech coding. Fig. 1 shows the block diagram of a critically decimated *M*-band uniform FB. The input signal is first decomposed by *M* analysis filters $H_i(z)$. The outputs are then decimated by a factor of *M* to form *M* subband signals. In the synthesis bank, the subband signals are upsampled by a factor of *M* before passing through the synthesis filters $F_i(z)$ to reconstruct the processed signal. The theory and design of PR FB have been widely studied in the literature [2]. One efficient structure is the cosine-modulated (CMFB), where the analysis filters (synthesis filters) are obtained by cosine modulation of a prototype filter. Due to the cosine modulation, the implementation and design complexities of CMFB are very low compared with a general PR FBs. Unfortunately, the classical CMFB proposed in [3] does not have LP analysis and synthesis filters, which is desirable in some applications. More recently, a new class CMFB using a different cosine and sine modulations are proposed [4]. Although the analysis and synthesis filters are LP, its frequency support is considerably different from that of uniform FB and there is considerable overlap between the passband of the low frequency analysis filters. Another popular class of LP *M*-channel uniform FBs is the linear-phase paraunitary filter bank (LPPUFB) [5], where the LP FB is parameterized as a cascade of delays and unitary matrices, which can further be parameterized as a series of planar rotations. The design of LPPUFB can be very involved because of the large number of design parameters and the highly nonlinear dependency of the frequency response on the rotation parameters [5-8]. This usually limits the stopband attenuation of the FB. Another commonly used method to construct PR FB is to cascade sets of PR FBs with smaller number of channels in a tree structure [9]. For example, an 8-channel PR FB can be obtained by cascading sets of 2-channel PR FBs in a tree structure with 3 levels as shown in Fig. 2. The output from the previous level is further decomposed using the analysis filters in that level into two more channels. In general, all the 2-channel PR FBs can differ from each other and they can be either linear- or nonlinear-phase. In wavelet transform and most tree-structured FBs considered in the literature, the same set of PR FB is used throughout the tree structure. Two significant drawbacks of this structure, from the viewpoint of designing a uniform FB, are the high system delay and the asymmetric transition band of the analysis filters. The latter usually results in a higher filter order to satisfy a given stopband attenuation and transition bandwidth, which further increases the total system delay. This is illustrated in Fig. 3 using a 2-channel PR FB with filter length $N = 128$. It can be seen that transition bandwidth are unequal and the system delay rapidly increases to 889 samples. In [1], Fliege has shown that the system delay of a tree-structured FB can be drastically reduced by having non-identical analysis filters in each level of the tree. More precisely, the length of the filters should decrease by a factor of two when going from one level to the other, while their transition bandwidth should increase by the same factor. In this paper, we further study this novel idea in the design of *M*-channel LP uniform FB using the lattice-based two-channel LP FB proposed in [10]. The main reason in choosing the latter is that the frequency responses of the lowpass and highpass analysis (synthesis) filters can be designed to be closely symmetric to the other around $\pi/2$. By properly selecting the filter length, transition bandwidth, and stopband attenuation of the 2-channel PR LP FBs at each level of the tree structure, it is possible to design uniform LP FB with excellent frequency characteristic and much lower system delay. For example, a uniform 8-channel PR LP FB with the same worse-case transition bandwidth requirement and stopband attenuation as the previous one (N=128) can be achieved with the new structure with a much lower implementation complexity and system delay of 377 samples (Fig. 4). The savings also increase linearly with the depth of the tree structure. The resulting FB, therefore, serves as useful alternative to the LPPUFB for designing LP FB with N a powers of two number and more generally M a composite number. Though the design of the component LPPUFBs in the latter case will become more complicated than the 2-channel LP FB in the former, it is still much simpler than designing directly an *M*-channel LPPUFB. The rest of the paper is organized as follows: Section II is devoted to the proposed tree-structured PR LP FB. The design procedure and some design examples are given in Section III. Conclusions are drawn in Section IV.

II. TREE-STRUCTURED PR LP FBS

First of all, let's consider an 8-channel tree-structured uniform FB constructed by cascading 2-channel PR FBs as shown in Fig. 2. $H_0^{(k)}(z)$ and $H_1^{(k)}(z)$ are respectively the lowpass and highpass analysis filters of the 2-channel PR FB at the k-th level of the tree structure, where $1 \le k \le K$, and K is the total number of levels in the tree. In the synthesis bank, the subband signals are recombined successively, two at a time, by a set of synthesis filters $F_i^{(k)}(z)$. From the noble identity, we know that $H(z)$ followed by a decimator with a ratio of two is equivalent to $H(z^2)$ preceding the decimator. Therefore, the tree-structured FB can be redrawn as an 8-channel uniform FB shown in Fig. 1 by moving the analysis filters to the right hand side of the tree structure, leading to M analysis filters $H_m(z)$, m=1,...,M. For convenience, let's treat the index "i" in $H_i^{(k)}(z)$ as the k-th

0-7803-7011-2/01/$10.00 ©2001 IEEE

digits in a weighted binary representation and denote it by b_k. The equivalent transfer function obtained by passing the signal though the branch $H_{b_1}^{(1)}(z)$, $H_{b_2}^{(2)}(z)$,..., $H_{b_K}^{(K)}(z)$ can then be labeled as $H_m(z)$,where $m = b_1 + 2b_2 + \cdots + 2^{K-1}b_K = (b_1,.....b_K)_2$. The resulting analysis filters $H_m(z)$ and synthesis filters $F_m(z)$ $(0 \le m \le M-1)$ can then be written as

$$H_m(z) = H_{b_1}^{(1)}(z^{2^0})H_{b_2}^{(2)}(z^{2^1})\ldots H_{b_k}^{(k)}(z^{2^{k-1}})\ldots H_{b_K}^{(K)}(z^{2^{K-1}}) \quad (1)$$

$$F_m(z) = F_{b_1}^{(1)}(z^{2^0})F_{b_2}^{(2)}(z^{2^1})\ldots F_{b_k}^{(k)}(z^{2^{k-1}})\ldots F_{b_K}^{(K)}(z^{2^{K-1}}) \quad (2)$$

It is clear that the whole system is PR if the 2-channel FBs in each level are PR. Further, if $H_i^{(k)}(z)$ and $F_i^{(k)}(z)$ are LP, then so are the filters $H_m(z)$ and $F_m(z)$.

As mentioned earlier, if the same set of FB is employed at all levels in the tree-structured FB, then the frequency responses of the overall analysis filters are not identical due to the up-sampling of z in moving the analysis filters to left of the decimators, Fig. 3. As a result, higher implementation complexity is required to achieve a given transition bandwidth and stopband attenuation. This also significantly increases the system delay of the FB. For example, if the lattice-based PR LP FB in [10] is used as the 2-channel FB, the system delay D of the 3-level tree-structured FB is given by

$$D = N^{(1)} + 2(N^{(2)} + 2(N^{(3)} - 1) - 1) - 1. \quad (3)$$

where $N^{(k)}$ is the length of FB in the k -th level. If we set $N^{(1)} = N^{(2)} = N^{(3)} = 128$, then the resulting tree-structured FB, as shown in Fig. 3, will have a system delay of 889 samples. It can be seen that 2-channel PR FBs at different levels of the tree structure contribute differently to the total system delay. Each component is linearly proportional to the length of the FB used and its scalar constant grows exponentially with the depth or level of the FB in the FB tree. To reduce the total system delay, it is therefore advantageous to reduce the length of the FB when the level increases.

From the design of 1-D FIR filter using the Kaiser window method, we know that the length of the filter N, stopband attenuation A, and transition bandwidth $\Delta\omega$ are related by the following formula

$$N = \frac{A-8}{2.285 \cdot \Delta\omega}. \quad (4)$$

For a given stopband attenuation and passband ripple, the filter length is inversely proportional to the transition bandwidth. From (4), it can be seen that the worse case stopband attenuation of $H_m(z)$ is equal to the worse case stopband attenuation of its factors $H_{b_1}^{(1)}(z^{2^0})$,...,$H_{b_k}^{(K)}(z^{2^{K-1}})$ and its transition bandwidth will depend on those of $H_{b_1}^{(1)}(z^{2^0})$,...., and $H_{b_K}^{(K)}(z^{2^{K-1}})$. Due to the upsampling by a factor of 2^{k-1}, the transition bandwidth of $H_{b_k}^{(i)}(z^{2^{k-1}})$ will be 2^{k-1} times narrower than that of $H_{b_1}^{(i)}(z)$. Thus, to achieve a uniform transition bandwidth, the length of 2-channel PR FB should be reduced by a factor of 2 when we go from one level to the other. In other words,

$$N^{(1)} = 2 \times N^{(2)} = 4 \times N^{(3)} \cdots = 2^{K-1} \times N^{(K)} \quad (5)$$

$$\Delta^{(1)} = \Delta^{(2)}/2 = \Delta^{(3)}/4 = \cdots = \Delta^{(K)}/2^{K-1}. \quad (6)$$

The system delay D (for K=3) is now reduced to $D = 3N^{(1)} - 7$, which grows only linearly with the depth of the tree. The arithmetic complexity, in terms of the numbers of multiplications and additions per unit time (MPUs and APUs), also reduced from 99 MPUs and 291 APUs to 59 MPUs and 171 APUs. This novel technique has been mentioned in [1] but unfortunately the selection of the 2-channel PR FB and detail design examples are missing. In this work, we shall show that this approach can be used to obtain M-channel LP PR uniform FB ($M = 2^K$, m a positive integer) with very good frequency characteristic, using the lattice-based 2-channel LP PR FB proposed in [10]. Although the efficient structure by Phoog et al [11] can also be used in a similar manner, which is very attractive because of their low design and implementation complexities [12], the frequency responses of its associated lowpass and highpass filters are not quite symmetric to each other. Therefore, the frequency characteristic at the transition band edges will start to degrade when we they are cascaded to form a tree structure with large number of channels. If one is comfortable with nonlinear-phase FIR filters (i.e. only passband linear-phase), then the CQF [9] and the general low-delay 2-channel FB [13] can also be employed. The latter will further reduce the total system delay of the FB. Interested readers are referred to [14] for more details regarding their design and factorization. We now consider the design of the 2-channel lattice-based LP FB and some design examples.

III. DESIGN PROCEDURE AND EXAMLES

A. DESIGN PROCEDURE

For a 2-channel critically decimated FB, the PR condition is given by

$$H_0^{(k)}(-z)H_1^{(k)}(z) - H_0^{(k)}(z)H_1^{k}(-z) = \beta \cdot z^{-d}, \quad (7)$$

where d is a positive integer and β is a nonzero constant. The synthesis filters are given by $F_0^{(k)}(z) = H_1^{(k)}(-z)$ and $F_1^{(k)}(z) = -H_0^{(k)}(-z)$. For our LP FB, $H_0^{(k)}(z)$ and $H_1^{(k)}(z)$ are chosen respectively to be symmetric and antisymmetric having the same filter length, which is an even number. Instead of optimizing the lattice coefficients, which involves highly nonlinear objective function, the coefficients of filters $H_0^{(k)}(z)$ and $H_1^{(k)}(z)$ are obtained by solving the following constrained optimization

$$\min_{h} \phi^{(k)} = \sigma\int_0^{\pi-\omega_s^{(k)}} (1-|H_0^{(k)}(e^{j\omega})|)^2 d\omega + (1-\sigma)\int_{\omega_s^{(k)}}^{\pi} |H_0^{(k)}(e^{j\omega})|^2 d\omega$$
$$+ (1-\sigma)\int_0^{\pi-\omega_s^{(k)}} |H_1^{(k)}(e^{j\omega})|^2 d\omega + (1-\sigma)\int_{\omega_s^{(k)}}^{\pi} (1-|H_1^{(k)}(e^{j\omega})|)^2 d\omega$$

subject to (7), (8)

where $\omega_s^{(k)}$ $(\pi/2 < \omega_s^{(k)} < \pi)$ and $\pi - \omega_s^{(k)}$ are the stopband cut off frequencies of $H_0^{(k)}(z)$ and $H_1^{(k)}(z)$, respectively, σ is a weighting constant from 0 to 1, and h is the vector containing the free variables in the impulse response. The transition bandwidth is $\Delta^{(k)} = 2\omega_s^{(k)} - \pi$. It is assumed that the FBs at stage k are all identical. The constrained optimization is solved using the DCONF subroutine in the IMSL library. On average, it takes 150 iterations for convergence and the violation of PR constraints is of the order of 10^{-15}.

Given the number of channel $M = 2^K$, transition bandwidth $\Delta^{(1)}$ and stopband attenuation A. The 2-channel PR LP FB at the first level is first designed by the above method to satisfy the given specification. Suppose that a filter length of $N^{(1)}$ is required. The (K-1) 2-channel PR LP FBs at the other levels can be designed with parameters given in (5) and (6).

B. DESIGN EXAMPLES

We now present two examples: i) a two-level tree-structured FB with $K = 2$ and M=4, and ii) a three-level tree-structured FB with $K = 3$ and M=8. For simplicity, $N^{(1)}$, $N^{(2)}$,

and $N^{(3)}$ are chosen as 128, 64 and 32, respectively. The frequency responses of the three two-channel LP FBs are shown in Fig.3 (a), (b) and (c). It can be seen that they have approximately the same stopband attenuation but successively wider transition band. The frequency responses of the 4-band and 8-band LP analysis FBs obtained by cascading these 2-channel FBs in a tree structure are shown in Fig. 6 and Fig. 4, respectively. It can be seen that frequency characteristic of this LP 8-channel PR FB is very good and a stopband attenuation over 50 dB can be readily obtained. The design parameters of $H_0^{(k)}(z)$ are summarized in Table 1. As a final remark, we shall contrast the relative merits of this tree-structured FB and the LPPUFB. As mentioned earlier, the LPPUFB usually involves considerable number of parameters, especially when the number of channel and filter length is large. The objective function is also a highly nonlinear function of the planar rotation parameters. All these somewhat limits the stopband attenuation of the FB that can be designed. The proposed tree-structured FB is relatively easy to design because the 2-channel LP FBs can be designed separately. This limits the number of parameters in each sub problem to a reasonable value. Moreover, the design of 2-channel LP PR FB is much easier than designing a LPPUFB and a number of efficient design methods are already available. On the other hand, the major disadvantage of the tree-structured FB is the restriction that the number of channel M is a powers-of-two number. Though it is also possible to form tree-structured FB by cascading FBs with 2, 3 and larger number of channels using a similar approach, there is still a fundamental limitation on the number of channel that can be designed.

IV. CONCLUSION

A method for designing M-channel LP PR FB with $M = 2^K$ using a tree-structured FB is presented. It is based on a previous observation of Fliege, where the length of the analysis filters is decreased by a factor of two when the depth of the tree is increased by one, while its transition bandwidth is increased by the same factor. A lattice-based 2-channel LP FB is chosen because the frequency responses of the lowpass and highpass analysis (synthesis) filters can be designed to be closely symmetric to the other around $\pi/2$. By properly selecting the filter length, transition bandwidth, and stopband attenuation of the 2-channel PR LP FBs at each stage of the tree structure, it is possible to design uniform PR LP FB with excellent frequency characteristic and much lower system delay.

REFERENCES

[1] N. J. Fliege, Multirate digital signal processing. John Wiley & Sons Ltd, 1995.

[2] P. P. Vaidyanathan, Multirate systems and filter banks. Englewood Cliffs, NJ: Prentice Hall, c1992.

[3] R. D. Koilpillai and P. P. Vaidyanathan, "Cosine-modulated FIR filter banks satisfying perfect reconstruction," *IEEE Trans. SP.*, Vol. 40, pp. 770-783, Apr. 1992.

[4] Y. P. Lin and P. P. Vaidyanathan, "Linear phase cosine modulated maximally decimated filter banks with perfect reconstruction," *IEEE Trans. SP.*, Vol. 42, no.11, pp. 2525-2539, Nov. 1995.

[5] A. K. Soman, P. P. Vaidyanathan, and T. Q. Nquyen, "Linear-phase paraunitary filter banks: Theory, factorizations and applications," *IEEE Trans. SP.*, Vol. 41, pp. 3480-3496, Dec. 1993.

[6] T. D. Tran and T. Q. Nguyen, "On M-channel linear-phase FIR filter banks and application in image compression," *IEEE Trans. SP.*, Vol. 45, pp. 2175-2187, Sept. 1997.

[7] R. L. de Queiroz, T. Q. Nguyen and K. R. Rao, "The GenLOT: Generalized linear-phase lapped orthogonal transform," *IEEE Trans. on SP.*, Vol. 40, pp. 497 –507, Mar. 1996.

[8] T. D. Tran, R. L. de Queiroz and T. Q. Nguyen, "Linear-phase perfect reconstruction filter bank: lattice structure, design, and application in image coding," *IEEE Trans. on SP.*, Vol. 48, pp. 133 –147, Jan. 2000.

[9] M. J. T. Smith and T. P. Barnwell III, "Exact reconstruction techniques for tree-structured sub-band coders," *IEEE Trans. ASSP.*, pp. 434-441, June 1986.

[10] T. Q. Nguyen and P. P. Vaidyanathan, "Two-channel perfect-reconstruction FIR QMF structures which yield linear-phase analysis and synthesis filters," *IEEE Trans. SP.*, Vol. 37, pp. 676-690, May 1989.

[11] S. M. Phoong, C. W. Kim, P. P. Vaidyanathan and R. Ansari, "A new class of two-channel biorthogonal filter banks and wavelet bases," *IEEE Trans. SP.*, Vol. 43, pp. 649-665, Mar. 1995.

[12] J. S. Mao, S. C. Chan and K. L. Ho, "Design of two-channel PR FIR filter banks with low system delay," *Proc. IEEE ISCAS'2000*, Vol. 1, pp. 627 –630.

[13] K. Nayebi, T. P. Barnwell III and M. J. T. Smith, "Low delay FIR filter banks: design and evaluation," *IEEE Trans. SP.*, Vol. 42, pp. 24-31, Jan. 1994.

[14] K. S. Pun, S. C. Chan, X. M. Xie, K. L. Ho and T. I. Yuk, "On the efficient realization and design of multiplier-less two-channel perfect reconstruction FIR filter banks," in student conference, *IEEE ICASSP'2001.*

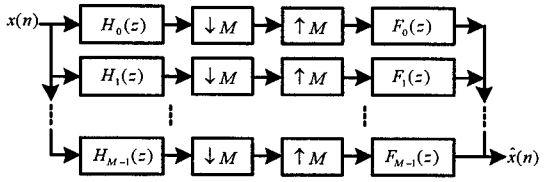


Fig. 1. The block diagram of a critically decimated uniform M-band FB.

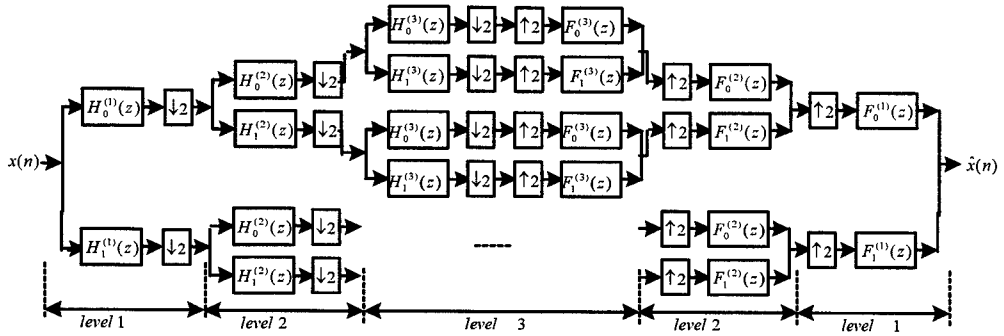


Fig. 2. 3-level maximally decimated tree-structured FB.

$H_0^{(k)}(z)$ for 4-band uniform FB		$H_0^{(1)}(z),\ k=1$	$H_0^{(2)}(z),\ k=2$
$H_0^{(k)}(z)$ for 8-band uniform FB	$H_0^{(1)}(z),\ k=1$	$H_0^{(2)}(z),\ k=2$	$H_0^{(3)}(z),\ k=3$
Length of filter $N^{(k)}$	128	64	32
Stopband cut off frequency $\omega_s^{(k)}$	$0.275\times 2\pi$	$0.3\times 2\pi$	$0.35\times 2\pi$
Transition bandwidth $\Delta^{(k)}$	$0.025\times 2\pi$	$0.05\times 2\pi$	$0.1\times 2\pi$
Parameters to be optimized	128	64	32
Implementation complexity	33 MPUs 97 APUs	17 MPUs 49 APUs	9 MPUs 25 APUs

Table 1. Parameters of the lowpass filters in tree-structured FB.

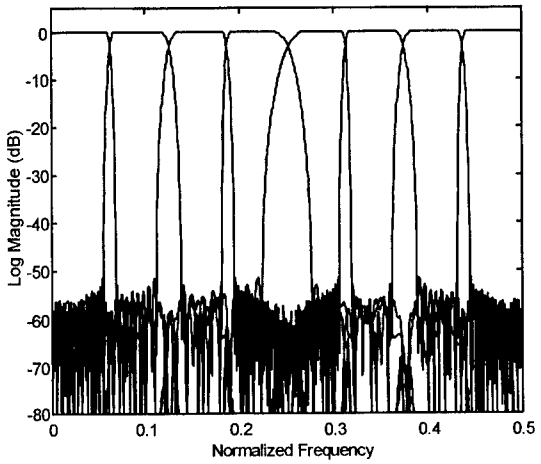


Fig. 3. The frequency responses of 3-level tree-structured FB by conventional method.

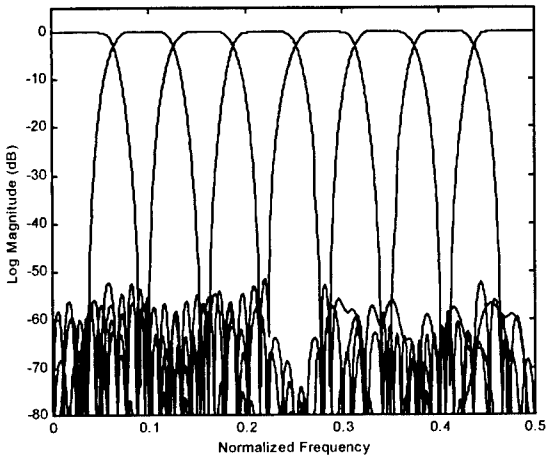
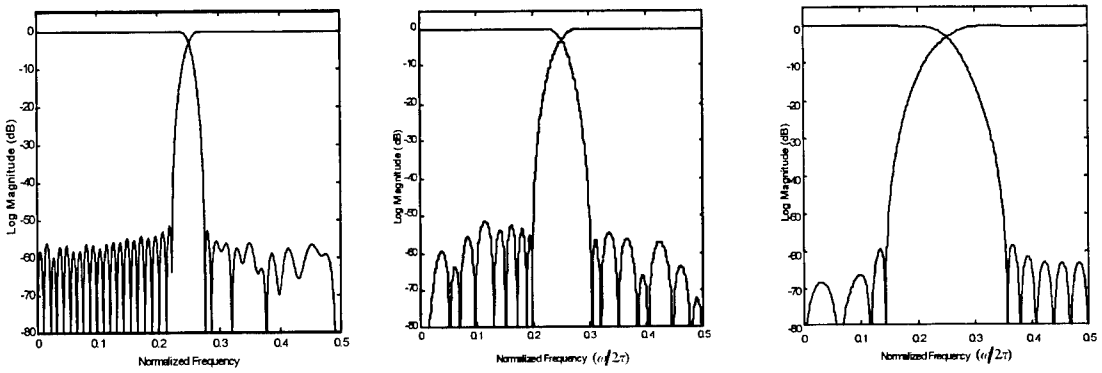


Fig. 4. The frequency responses of 3-level tree-structured FB by the proposed method.



(a) (b) (c)

Fig. 5. Frequency responses of { $H_0^{(k)}(z)$ and $H_1^{(k)}(z)$ } in (a) level 1, (b) level 2, and (c) level 3.

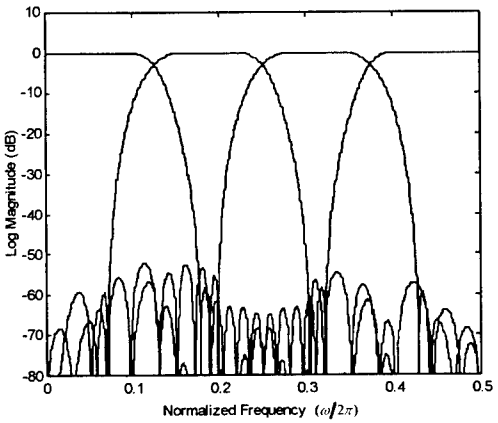


Fig. 6. The frequency responses of 4-band FB.

IMPROVEMENT OF FACTORIZATION FOR TWO-CHANNEL PERFECT RECONSTRUCTION FIR FILTER BANKS

Shi Guangming and Jiao Licheng
Key Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China

ABSTRACT

Recently, there is an increasing interest in designing structurally perfect reconstruction (PR) filter banks because the system can be implemented by using sum of powers-of-two (SOPOT) coefficients. The structurally PR filter banks can be designed by factorization based on lifting scheme. But there exist some problems that will be addressed in this paper. Improvement of the factorization to solve the problems is proposed. The procedures of proof for the improvement are given. Final, the given examples show that the proposed method is effective.

I. INTRODUCTION

Perfect reconstruction (PR) multirate filter banks (FB), which is used for division of a signal into frequency bands and the reconstruction of the signal from the individual bands, have important applications in signal analysis, signal coding and the design of wavelet bases [1~4]. Many researchers [5~11] proposed a number of constrained optimization techniques for designing linear-phase and low-delay PR two-channel filter banks. But there is a problem that the filter banks so obtained are in general pseudo PR. To overcome the problem, structurally PR filter banks are desired. An efficient lattice type PR-QMF bank that "structurally" ensures the PR property was reported by Vaidynathan and Hoang [6]. But it is difficult to design with a general-purposed nonlinear optimization technique because of some reasons such as nonconvergence, step-size selection and local suboptimality. Another type PR filter banks based on algebraic formulation are proposed by Pinchon [12]. The filter bank can be factorized with a cascade of N blocks. However, unfortunately, only linear phase filter banks were discussed. Actually, a general factorization of PR filter banks can be used with a lifting scheme. The lifting scheme first proposed by Donoho [13] is also available for designing a structurally PR filter banks. An important advantage, however, is that it can also be used in biorthogonal wavelet. Early progress in lifting scheme has been focused on the design of discrete wavelet transform or two band subband filtering. Such lifting scheme utilizes the Euclidean algorithm for polynomial in factorization. Vetterli [14] employed the Euclidean algorithm and the close connection between Diophantine equations and the PR conditions to parameterize all solutions of highpass filters with a given lowpass filter. Daubechies [15] applied the lifting scheme to design the discrete wavelet filter bank. This factorization, which is based on the lifting scheme, is also used for the general two-channel PR filter banks if the determinant of the polyphase matrix is equal to constant multiplies of signal delays. It can be used to convert a numerically optimized nearly PR filter bank into a structurally PR system. But there possibly exist non-causal polynomial and two analysis filters have the similar frequency response after factorization.

In paper, an improvement of the factorization to solve the problems is proposed. The procedures of deriving are also given. The paper is organized as follows: the formerly factorization method for two-channel filter bank is described in Section 2. The improvement algorithm of the factorization and some constrained problems are addressed in Section 3. Design procedures and several examples, including linear phase and low delay filter banks, are given in Section 4. Finally, the conclusions are drawn in Section 5.

II. Factorization of Two-Channel FIR Filter Banks Using Lifting Scheme

Consider the basic structure of a two-channel FIR filter bank with analysis filters $\{H_0(z),\ H_1(z)\}$ and synthesis filter $\{F_0(z),\ F_1(z)\}$ in Fig.1.

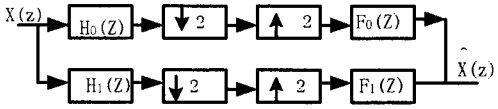


Figure 1. The two-channel Filter bank

The relationship between the input $x(z)$ and the output $\hat{x}(z)$ is given by

$$\hat{x}(z) = T(z)x(z) + A(z)x(-z) \qquad (1)$$

where $T(z) = \frac{1}{2}[H_0(z)F_0(z) + H_1(z)F_1(z)]$ and

$A(z) = \frac{1}{2}[H_0(-z)F_0(z) + H_1(-z)F_1(z)]$.

Setting $F_0(z) = -H_1(-z)$ and $F_1(z) = H_0(-z)$, the aliasing term $A(z)$ is equal to zero. The condition to achieve perfect reconstruction with FIR synthesis filters after a FIR analysis section can be expressed as:

$$H_{00}(z)H_{11}(z) - H_{01}(z)H_{10}(z) = \beta \cdot z^{-d}, \qquad (2)$$

which is called Bezout identities [5], where $H_0(z) = H_{00}(z^2) + z^{-1}H_{01}(z^2)$, $H_1(z) = H_{10}(z^2) + z^{-1}H_{11}(z^2)$, β is some constant, and system delay parameter d is some integer. In general, the design problem of the two-channel PR filter bank is formulated as a constrained non-linear optimization problem but not being robust to coefficient quantization. One problem with the constrained optimization approach is that the filter bank is not completely PR. A method to solve this problem is factorization using lift scheme shown as follows.

Let the analysis FIR filters be $H_0(z) = \sum_{n=0}^{L_0-1} h_0(n)z^{-n}$ and $H_1(z) = \sum_{n=0}^{L_1-1} h_1(n)z^{-n}$. Without loss of generality, suppose that $|H_1(z)| \geq |H_0(z)|$ ($|H(z)|$, the degree of a Laurent polynomial, is defined as $|H(z)| = P_1 - P_0$ if $H(z) = \sum_{n=P_0}^{P_1} h(n)z^{-n}$). $H_{10}(z)$ and $H_{11}(z)$ can be expressed as

$$H_{10}(z) = \tilde{H}_{10}(z) - Q(z)H_{00}(z),$$
$$H_{11}(z) = \tilde{H}_{11}(z) - Q(z)H_{01}(z), \qquad (3)$$

where $\tilde{H}_{10}(z)$ and $\tilde{H}_{11}(z)$ are satisfied Eq.(2) and $Q(z)$ is some real polynomials. Therefore, all the highpass filters, $H_1(z)$, that are "complementary" to the lowpass filter $H_0(z)$, are given by $H_1(z) = \tilde{H}_1(z) - Q(z^2)H_0(z)$. Using

0-7803-7011-2/01/$10.00 ©2001 IEEE

the Euclidean algorithm on $H_{00}(z)$ and $H_{01}(z)$, one gets following matrix product

$$\begin{bmatrix} H_{00}(z) \\ H_{01}(z) \end{bmatrix} = \prod_{i=1}^{n} \begin{bmatrix} q_i(z) & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} K \\ 0 \end{bmatrix}, \quad \text{where} \quad q_i(z) \quad \text{are}$$

polynomials, K is a non-zero constant, and n is some integer. The particular solution $\tilde{H}_{10}(z)$ and $\tilde{H}_{11}(z)$ can be obtained by constructing the following polyphase matrix $\tilde{\mathbf{P}}(z)$ with determinant 1.

$$\tilde{\mathbf{P}}(z) = \begin{bmatrix} H_{00}(z) & \tilde{H}_{10}(z) \\ H_{01}(z) & \tilde{H}_{11}(z) \end{bmatrix}$$

$$= \prod_{i=1}^{n} \begin{bmatrix} q_i(z) & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} K & 0 \\ 0 & (-1)^n z^{-d} \beta / K \end{bmatrix}, \tag{4}$$

so we have

$$\begin{bmatrix} H_{00}(z) & H_{10}(z) \\ H_{01}(z) & H_{11}(z) \end{bmatrix} = \begin{bmatrix} H_{00}(z) & \tilde{H}_{10}(z) \\ H_{01}(z) & \tilde{H}_{11}(z) \end{bmatrix} \begin{bmatrix} 1 & -Q(z) \\ 0 & 1 \end{bmatrix}$$

$$= \prod_{i=1}^{n} \begin{bmatrix} q_i(z) & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} K & 0 \\ 0 & (-1)^n z^{-d} \beta / K \end{bmatrix} \begin{bmatrix} 1 & -Q(z) \\ 0 & 1 \end{bmatrix}.$$

Therefore, the two-channel filter bank $H_0(z)$ and $H_1(z)$ can be factorized in to some $q_i(z)$ and constants. The advantage of this factorization is that is robust from coefficient quantification. But since the factorization is implemented by simple long division algorithm for Laurent polynomial, there exist two problems: first problem is that $\tilde{H}_1(z)$ may be a low pass filter by Eq.(4) and $H_1(z)$ from Eq.(3) is not always high pass filter when $H_0(z)$ and $H_1(z)$ have same length; second problem is the $q_i(z)$ may be non-causal.

III. IMPROVEMENT OF THE FACTORIZATION ALGORITHM

Euclidean algorithm is based on the long division for Laurent polynomial [14~17]. Let us analysis the long division. Consider two causal Laurent polynomials $a_0(z)$ and $b_0(z)$ with $|a_0(z)| \geq |b_0(z)|$, assume that their first coefficients of z^0 are non-zeros, then there always exists a Laurent polynomial $q_1(z)$ (the quotient), and a Laurent polynomial $r_0(z)$ (the remainder) with $|r_0(z)| < |b_0(z)|$, so that

$$a_0(z) = q_1(z) b_0(z) + r_0(z). \tag{5}$$

The quotient $q_1(z)$ and the remainder $r_0(z)$ can be calculated by long division. If $q_1(z) b_0(z)$ has to match the beginning or end terms of $a_0(z)$, the division is called fore-long division or back-long division, respectively.

If the fore-long division is taken, then

$$\begin{bmatrix} a_0(z) \\ b_0(z) \end{bmatrix} = \begin{bmatrix} q_1(z) & z^{-k} \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} b_0(z) \\ r_0(z) \end{bmatrix}, \tag{6}$$

or If the back-long division is used, then

$$\begin{bmatrix} a_0(z) \\ b_0(z) \end{bmatrix} = \begin{bmatrix} q_1(z) & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} b_0(z) \\ r_0(z) \end{bmatrix}, \tag{7}$$

where $k = |q_1(z)| + 1$, and the coefficient of z^0 in $r_0(z)$ is nonzero. Hence, the division is non-unique. Based on the above study, a fully PR two-channel filter banks $H_0(z)$ and $H_1(z)$ can be factorized as follows:

$$\begin{bmatrix} H_{00}(z) \\ H_{01}(z) \end{bmatrix} = \prod_{i=1}^{j} \begin{bmatrix} q_i(z) & z^{-k_i} \\ 1 & 0 \end{bmatrix} \prod_{i=j+1}^{n} \begin{bmatrix} q_i(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} K \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} H_{10}(z) \\ H_{11}(z) \end{bmatrix} = \prod_{i=1}^{j} \begin{bmatrix} q_i(z) & z^{-k_i} \\ 1 & 0 \end{bmatrix} \prod_{i=j+1}^{n} \begin{bmatrix} q_i(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -Q(z) \\ (-1)^{n-1} \beta / K \end{bmatrix} \tag{8}$$

Since $H_0(z)$ and $H_1(z)$ are constrained by Eq.(2), the numbers of taking back-long division j for factorization are also constrained.

Proposition: Suppose that there are j terms of $q(z)$ which are obtained by fore-long division and $n - j$ terms of $q(z)$ by back-long division after factorization, then number j and $|q_i(z)|$ $\{i = 1, \cdots, j\}$ are satisfied the following constrained condition.

$$\sum_{i=1}^{j} (|q_i(z)| + 1) = d. \tag{9}$$

Proof: We define $D_{-1} = H_{00}$, $D_0 = H_{01}$, $A_{-1} = H_{10}$ and $A_1 = H_{11}$. In this notation (2) becomes

$$D_{-1}(z) A_0(z) - A_{-1}(z) D_0(z) = \beta z^{-d}. \tag{10}$$

Now use fore-long division starting with the pair $D_{-1}(z), D_0(z)$. The first step gives

$$D_{-1}(z) = q_1(z) D_0(z) + z^{-k_1} D_1(z).$$

Also do one division of the pair $A_{-1}(z), A_0(z)$, denoting the remainder $A_1(z)$, $A_{-1}(z) = p_1(z) A_0(z) + z^{-l_1} A_1(z)$.

The $A_1(z)$ and $D_1(z)$ are nonzero polynomial in z^0. If we choose $|p_1(z)| = |q_1(z)|$, then $l_1 = k_1 = |q_1(z)| + 1$.

Together these equations give

$$\begin{aligned} \beta z^{-d} &= D_{-1}(z) A_0(z) - A_{-1}(z) D_0(z) \\ &= (q_1(z) - p_1(z)) A_0(z) D_0(z) \\ &\quad + z^{-k_1} (D_1(z) A_0(z) - A_1(z) D_0(z)). \end{aligned}$$

Since the coefficient of $A_0(z) D_0(z)$ in z^0 is nonzero, we must have $p_1(z) = q_1(z)$, and hence

$$D_0(z) A_1(z) - A_0(z) D_1(z) = -\beta z^{-(d-k_1)}$$

Since this is of the same form, but of lower degree, than the equation that we started with (10), we can compare the second step of this factorization with the fore-long division of $A_0(z), A_1(z)$ and this gives $p_2(z) = q_2(z)$ when $d - k_1 - k_2 > 0$. The result is that we get a succession of Bezout identities

$$D_{j-1}(z) A_j(z) - A_{j-1}(z) D_j(z) = (-1)^j \beta z^{-(d - \sum_{i=1}^{j} k_i)},$$

which are of decreasing degree. We find in turn that $p_1(z) = q_1(z), \cdots p_j(z) = q_j(z)$. When $d - \sum_{i=1}^{j} k_i = 0$, $D_{j-1}(z) A_j(z) - A_{j-1}(z) D_j(z) = (-1)^j \beta$. It is clear that if the fore-long division is continually used at this time, then $p_{j+1}(z) \neq q_{j+1}(z)$. The back-long division must be used instead. We have

$D_{j-1}(z) = q_{j+1}(z) D_j(z) + D_{j+1}(z)$, and

$A_{j-1}(z) = p_{j+1}(z) A_j(z) + A_{j+1}(z)$, where

$|D_j(z)| > |D_{j+1}(z)|, |A_j(z)| > |A_{j+1}(z)|$. The $A_{j+1}(z)$ and $D_{j+1}(z)$ are also nonzero polynomial in z^0.

At same way, let $|p_{j+1}(z)| = |q_{j+1}(z)|$,

$$(-1)^j \beta = D_{j-1}(z)A_j(z) - A_{j-1}(z)D_j(z)$$
$$= (q_{j+1}(z) - p_{j+1}(z))A_j(z)D_j(z)$$
$$+ (D_{j+1}(z)A_j(z) - A_{j+1}(z)D_j(z)).$$

Since $|A_j(z)D_j(z)| \geq |D_{j+1}(z)A_j(z)|$ and $|A_j(z)D_j(z)| \geq |D_{j+1}(z)A_j(z)|$, we must have $p_{j+1}(z) = q_{j+1}(z)$.

The rest can be deduced by analogy till $D_{n-1}(z)$ is monomial (a constant K), namely $|D_{n-1}(z)| = 0$. We have $D_{n-2}(z) = q_n(z)D_{n-1}(z)$ and $D_n(z) = 0$. Substituting them in to $D_{n-2}(z)A_{n-1}(z) - A_{n-2}(z)D_{n-1}(z) = (-1)^{n-1}\beta$, We get $A_{n-2} = q_n(z)A_{n-1}(z) + (-1)^{n-1}\beta / D_{n-1}(z)$. Now, let $Q(z) = -A_{n-1}(z)$, $K = D_{n-1}(z)$, the whole factorization can be formulated matrix form as.

$$\begin{bmatrix} H_{00}(z) & H_{10}(z) \\ H_{01}(z) & H_{11}(z) \end{bmatrix} = \underbrace{\begin{bmatrix} q_1(z) & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} q_j(z) & z^{-k_j} \\ 1 & 0 \end{bmatrix} \cdots}_{n} \begin{bmatrix} K & -Q(z) \\ 0 & (-1)^{n-1}\beta/K \end{bmatrix} \quad (11)$$

Actually, we can also use back-long division in the first step of factorization. The order of fore- or back-division to be taken can be arbitrarily arranged during the procedures of factorization.

IV. Procedure of Design and Examples

The procedure of the factorization:

(1) First step is to design a nearly PR Filter banks { $H_0(z)$ and $H_1'(z)$ }. Note that d should be equal to and less than $(1/4)\cdot(N_0 + N_1) - 1$, respectively, for linear-phase and low-delay filer banks, where N_0 and N_0 are the length of $H_0(z)$ and $H_1'(z)$. The design problem can be formulated as a constrained non-linear optimization problem, which can be solved by the NCONF/DCONF subroutine in the IMSL library. The pass band ripples and the stop band attenuation of the low pass and high pass filters can be minimized, and under PR condition constrained an objective function Φ_{near} to be minimized is as follows:

$$\min_{\mathbf{h}} \Phi_{near} = \sigma \int_0^{\omega_{p0}} (1 - |H_0(e^{j\omega})|^2)^2 d\omega + (1-\sigma) \int_{\omega_{s0}}^{\pi} |H_0(e^{j\omega})|^2 d\omega$$
$$+ \sigma \int_0^{\omega_{s1}} |H_1'(e^{j\omega})|^2 d\omega + (1-\sigma) \int_{\omega_{p1}}^{\pi} (1 - |H_1'(e^{j\omega})|^2)^2 d\omega$$

subjected to the PR condition in (2). (12)

Here, $\boldsymbol{h}$ is the vector containing the impulse responses of $H_0(z)$ and $H_1'(z)$; σ is a weighting constant from 0 to 1 which is used to control the relatively important of the error in the stop band and pass band; ω_{p_0} and ω_{p_1} are the pass band cut-off frequencies of $H_0(z)$ and $H_1'(z)$; ω_{s_0} and ω_{s_1} are the stop band cut-off frequencies of $H_0(z)$ and $H_1'(z)$

(2) Factorizing $H_0(z)$ and $H_1'(z)$ via the improvement factorization algorithm in Section 3, we have.

$$\begin{bmatrix} H_{00}(z) \\ H_{01}(z) \end{bmatrix} = \prod_{i=1}^{j} \begin{bmatrix} q_i(z) & z^{-k_i} \\ 1 & 0 \end{bmatrix} \prod_{i=j+1}^{n} \begin{bmatrix} q_i(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} K \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} H'_{10}(z) \\ H'_{11}(z) \end{bmatrix} = \prod_{i=1}^{j} \begin{bmatrix} p_i(z) & z^{-k_i} \\ 1 & 0 \end{bmatrix} \prod_{i=j+1}^{n} \begin{bmatrix} p_i(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -Q(z) \\ (-1)^{n-1}\beta/K \end{bmatrix} \quad (13)$$

(3) For nearly PR, $q_i(z)$ and $p_i(z)$ so obtained by step (2) are not exactly same. Substituting $p_i(z)$ with $q_i(z)$ into Eq.(13), a new filter $H_1(z)$ can be obtained by Eq.(11). $H_0(z)$ and $H_1(z)$ will constructed a structurally PR filter bank.

Using above factorization, we can find that $q_i(z)$ is always causal so long as $H_0(z)$ and $H_1(z)$ are causal. And $H_1(z)$ can be guaranteed to be highpass filter because its frequency response is resemble to that of $H_1'(z)$.

We now present several design examples. The first one is a low-delay PR FB with length, $N_0 = N_1 = 24$, where N_i is the length of the filter $H_0(z)$. The factorization matrix is as follows.

$$\begin{bmatrix} H_{10}(z) & H_{00}(z) \\ H_{11}(z) & H_{01}(z) \end{bmatrix} = \prod_{i=1}^{4} \begin{bmatrix} q_i(z) & z^{-k_i} \\ 1 & 0 \end{bmatrix} \prod_{i=5}^{13} \begin{bmatrix} q_i(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} K & -Q(z) \\ 0 & (-1)^{n-1}\beta/K \end{bmatrix}$$

The system delay is 15 ($d = 7$) samples. Fig. 2 plots the frequency response of its analysis banks. Frequency responses of the optimized filters before and after factorization are shown in dashed and solid lines, respectively. They are fairly close to each other. The coefficients of the filter bank are shown in Table 1. The second one is a linear phase filter bank with length $N_0 = N_1 = 32$. It is factorized as follows matrix.

$$\begin{bmatrix} H_{10}(z) & H_{00}(z) \\ H_{11}(z) & H_{01}(z) \end{bmatrix} = \prod_{i=1}^{8} \begin{bmatrix} q_i(z) & z^{-k_i} \\ 1 & 0 \end{bmatrix} \prod_{i=9}^{17} \begin{bmatrix} q_i(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} K & -Q(z) \\ 0 & (-1)^{n-1}\beta/K \end{bmatrix}$$

Their frequency responses of the optimized filters before and after factorization are shown in dashed and solid lines, respectively in Fig 3 and Their coefficients is listed in Table 2.

V. CONCLUSION

In this paper, an improvement of factorization technique approach to design the two-channel filter bank is presented. It can avoid some problems in the formerly factorization. The design results suggest that a structurally PR filter bank with good frequency characteristics can be obtained.

REFERENCES

[1] Horng B R, Samueli H. The Design of Low-Complexity Linear-Phase FIR Filter Banks Using Powers-of-Two Coefficients with an Application to Subband Image Coding. IEEE Trans. on Circuits and System for Video Technology, 1991, 318

[2] Gilloire A, Vetterli M. Adaptive Filtering in Subbands with Critical Sampling: Analysis, experiments, and application to Acoustic Echo Cancellation. IEEE Trans. on SP., 1992, 1862

[3] S. M. Phoong, C. W. Kim, P.P. Vaidyanathan, R. Ansari, 'A New Class of Two-Channel biorthogonal Filter Banks and Wavelet Bases', IEEE Trans. SP., Vol. 43, No. 3, Mar. 1995.

[4] Azimi-Sadjadi M R, Charleston S. A New Time Delay Estimation in Subbands for Resolving Multiple Specular Reflexions. IEEE Trans. on Signal Processing 1998, 3398

[5] Vaidyanathan P. P. Multirate systems and filter banks. Englewood Cliffs, NJ: Prentice Hall, 1993

[6] Vaidyanathan P.P. Hoang P.Q. Lattice Structures for Optimal Design and Robust Implementation of Two-Channel Perfect Reconstruction QMF Banks. IEEE Trans. On ASSP, 1988, 81-94

[7] Coh C.K., Lim Y.C. An efficient algorithm to design weighted minimax perfect reconstruction quadrature mirror filter bank. IEEE Trans. On Signal Proce., 1999, 3303-3314

[8] Hiroshi Ochi, Morihiko Ohta, et al, Linear Programming Design of Two-channel Perfect-Reconstruction Biorthogonal Filter Banks—Linear phase and Low delay. IEEE International Symposium on Circuits and System, 1997, 969-972

[9] Parviz Saghizadch, Willson N. Using Unconstrained Optimization in the Design of Two-channel Perfect-Reconstruction Linear-Phase FIR Filter Banks. Proc. Of IEEE Conf. on SP 1995, 1053-1055

[10] Yang S.J. Lee J.H. Chieu B.C. Minimax design of two-channel low-delay perfect-reconstruction FIR filter banks. IEE Proceeding Vis. Image Signal Process, Vol. 146, No.1 February, 1999.

[11] Chao Her-Chang Two-channel filter banks satisfying low-delay and perfect reconstruction design. Signal Processing 80 (2000) 465-479

[12] Pinchon D.,Siohan P. Analysis, Design, and Implementation of Two-channel Linear-Phase Filters: A New Approach. IEEE Trans. On Signal Processing, 1998, pp. 1814-1826

[13] D.L.Donoho, Interpolating wavelet transforms, Preprint, Department of Statistics, Stanford University, 1992

[14] Martin Vetterli , Cormac Herley. Wavelets and Filter Banks: Theory and Design. IEEE Trans. on SP, 1992, 2207-2232.

[15] Daubechies I.C., Sweldents W. Factoring Wavelet transform into lifting steps. J.Fourier Analy Application ,1998, 247-269,

[16] Khansari M.R.K., Dubois E. Padé table, continued fraction expansion, and perfect reconstruction filter banks. IEEE Trans. On SP, 1996, 1955-1963.

[17] Kok C.W., Nguycn T. Q. Chinese Remainder theorem: Recent Trend and New Results in Filter Banks Design, Proc. Of EUSIPCO-96, Trieste, Italy, 1996, 755-758

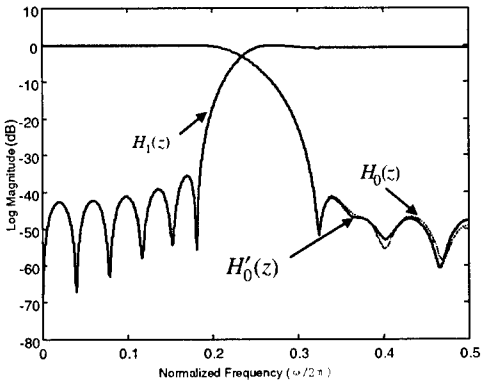


Fig. 2 Frequency responses for the low delay filter bank with length 24: before(solid line) and after factorization (dashed line)

Table 1: The coefficients of low delay filter bank with length 24 after factorization

$q_i(z)$	z^0	z^{-1}	k_i
$i=1$	2.3870e-1		1
$i=2$	-2.4796e+0	-9.1321e+0	2
$i=3$	1.0264e-2	-1.3706e-1	2
$i=4$	6.0653e-1	4.1425e+0	2
$i=5$	5.5442e+0	9.4486e-1	
$i=6$	9.7053e-1	4.6411e-1	
$i=7$	-5.7880e-1	1.4550e-1	
$i=8$	-4.6366e+0	-1.8024e+0	
$i=9$	1.6395e-1	-4.8520e-2	
$i=10$	-1.8739e+1	1.1209e+1	
$i=11$	-2.5538e-2	-6.8751e-3	
$i=12$		-6.0549e+0	
$i=13$	3.7595e-2		
K = 5.9299e-1		β = -2.5735e-1	
$Q(z)$ =1.2204e+1			

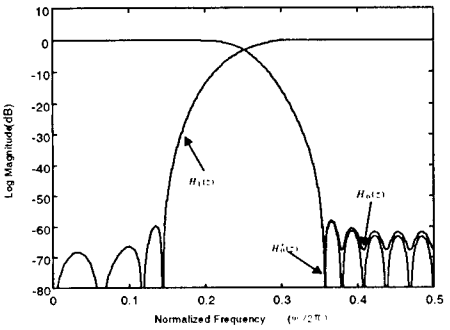


Fig. 3 Frequency responses for linear phase filter bank with length 32 before(solid line) and after factorization (dashed line)

Table 2: The coefficients of linear phase filter bank with length 32 after factorization

$q_i(z)$	z^0	z^{-1}	k_i
$i=1$	-1.0196e+1		1
$i=2$	2.1131e-2	5.7234e-2	2
$i=3$	9.7279e+0	1.6716e+1	2
$i=4$	3.2152e-2	1.1469e-1	2
$i=5$	-3.5707e+0	9.5284e+1	2
$i=6$	-5.3194e-4	-9.4726e-3	2
$i=7$	-1.5939e+3	1.0795e+3	2
$i=8$	-5.6660e-5	-2.2936e-4	2
$i=9$	3.2639e+3	-1.6344e+3	
$i=10$	-8.9678e-5	-6.3634e-4	
$i=11$	-3.5044e+3	-9.5698e+2	
$i=12$	3.0716e-3	-2.9728e-4	
$i=13$	2.6832e+2	1.9539e+1	
$i=14$	-1.6007e-3	1.2674e-4	
$i=15$	6.0213e+5	7.2528e+4	
$i=16$		2.7713e-7	
$i=17$	-6.4185e+5		
K = -3.7091e-3		β = 4.9075e-001	
$Q(z)$ =1.9095e-3			

SUBBAND ADAPTIVE GENERALIZED SIDELOBE CANCELLER FOR BROADBAND BEAMFORMING

Wei Liu, Stephan Weiss, and Lajos Hanzo

Communications Research Group
Department of Electronics & Computer Science
University of Southampton, SO17 1BJ, U.K.
{w.liu, s.weiss, l.hanzo}@ecs.soton.ac.uk

ABSTRACT

In this paper, we propose a novel subband adaptive broadband beamforming architecture based on the generalised sidelobe canceller (GSC), in which we decompose each of the tapped delay-line signals feeding the adaptive part of the GSC and the reference signal into subbands and perform adaptive minimisation of the mean squared error in each subband independently. Besides its lower computational complexity, this new subband adaptive GSC outperforms its fullband counterpart in terms of convergence speed because of its prewhitening effect. Simulations based on different kinds of blocking matrices with different orders of derivative constraints are presented to support these findings.

1. INTRODUCTION

Adaptive beamforming has found many applications in various areas ranging from sonar and radar to wireless communications. It is based on a technique where, by adjusting the weights of a sensor array with attached filters, a prescribed spatial and spectral selectivity is achieved. Fig. 1 shows a beamformer with M sensors receiving a signal of interest from the direction of arrival (DOA) angle ϑ.

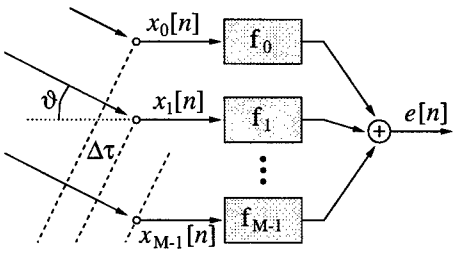


Fig. 1: A signal impinging from an angle ϑ onto a beamformer with M sensors.

To perform beamforming with high interference rejection and resolution, arrays with a large number of sensors and filter coefficients have to be employed. To facilitate real-time implementation, various methods are employed to reduce the computational complexity, such as the partially adaptive beamforming [1], wavelet-based beamforming [2] and subband beamforming [3]. In the latter, the received sensor signals are first split into decimated subbands, then an independent beamformer is applied to each subband. The advantage arises from the processing in decimated subbands, although at the expense of having to project constraints into the subband domain as well.

We here focus on a linearly constrained minimum variance (LCMV) beamformer, which can be efficiently implemented as a generalized sidelobe canceller (GSC) [4, 5]. Different from [3], instead of performing beamforming in subbands by decomposing the input sensor signals, we employ subband adaptive filtering techniques for the adaptive process of the GSC structure only. Specifically, noting that there are in total $M - S$ input tapped delay-lines for the adaptive part of the GSC, we decompose each of the tap-delay line signals and the reference signal $d[n]$ into K subbands by a K-channel filter banks as shown in Fig. 3 and perform adaptive minimisation in each subband. Simulation results with different blocking matrices and different order of derivative constraints show that this new method outperforms the fullband counterpart in addition to its very low computational complexity.

The rest of this paper is organised as follows: Section 2 is a brief review of GSC-based broadband beamforming based on a generalized sidelobe canceller with derivative constraints. In Section 3, we introduce the proposed subband-based GSC structure. Simulation and results will be given in Section 4 and conclusions are drawn in Section 5.

0-7803-7011-2/01/$10.00 ©2001 IEEE

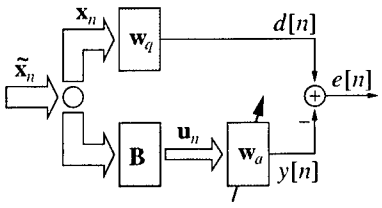


Fig. 2: Structure of a generalized sidelobe canceller.

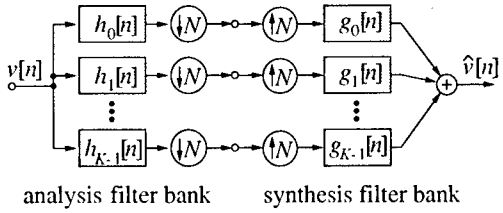


Fig. 3: K channel filter banks with decimation N.

2. GENERALIZED SIDELOBE CANCELLER

An LCMV beamformer performs the minimization of the variance or power of the output signal with respect to some given spatial and spectral constraints. For a beamformer with M sensors and J filter taps following each sensor as shown in Fig. 1, the output $e[n]$ can be expressed as:

$$e[n] = \mathbf{w}^{\mathrm{H}} \cdot \mathbf{x}_n \tag{1}$$

where coefficients and input sample values are defined as

$$\mathbf{w} = \left[\mathbf{w}_0^{\mathrm{T}} \ \mathbf{w}_1^{\mathrm{T}} \ \ldots \ \mathbf{w}_{J-1}^{\mathrm{T}}\right]^{\mathrm{H}} \tag{2}$$

$$\mathbf{w}_l = \left[w_0[l] \ w_1[l] \ \ldots \ w_{M-1}[l]\right]^{\mathrm{T}} \tag{3}$$

$$\mathbf{x}_n = \left[\tilde{\mathbf{x}}_n^{\mathrm{T}} \ \tilde{\mathbf{x}}_{n-1}^{\mathrm{T}} \ \ldots \ \tilde{\mathbf{x}}_{n-J+1}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{4}$$

$$\tilde{\mathbf{x}}_n = \left[x_0[n] \ x_1[n] \ \ldots \ x_{M-1}[n]\right]^{\mathrm{T}} \ . \tag{5}$$

The data vector $\tilde{\mathbf{x}}_n$ is a time slice as given in Fig. 1. A coefficient $w_m[l]$ is defined to sit at the tap position l of the mth filter f_m. The LCMV problem can now be formulated as [6]

$$\min_{\mathbf{w}} \mathbf{w}^{\mathrm{H}}\mathbf{R}_{xx}\mathbf{w} \qquad \text{subject to} \qquad \mathbf{C}^{\mathrm{H}}\mathbf{w} = \mathbf{f} \ , \tag{6}$$

where $\mathbf{R}_{xx}$ is the covariance matrix of observed array data in $\mathbf{x}_n$, $\mathbf{C} \in \mathbb{C}^{MJ \times SJ}$ is a constraint matrix and $\mathbf{f} \in \mathbb{C}^{SJ}$ is the constraining vector. The constraint matrix here imposes derivative constraints of order $S-1$ [7],

$$\mathbf{C} = \left[\hat{\mathbf{C}}_0 \ldots \hat{\mathbf{C}}_{S-1}\right] \qquad \text{with} \quad \hat{\mathbf{C}}_i = \begin{bmatrix} \mathbf{c}_i & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{c}_i \end{bmatrix} \tag{7}$$

with $\mathbf{c}_i = \left[(-m_0)^i \ (1-m_0)^i \ \cdots \ (M-1-m_0)^i\right]^{\mathrm{T}}$ and a phase origin point m_0.

The constrained optimisation of the LCMV problem in (6) can be conveniently solved using a GSC. The GSC performs a projection of the data onto an unconstrained subspace by means of a blocking matrix $\mathbf{B}$ and a quiescent vector $\mathbf{w}_{\mathrm{q}}$. Thereafter, standard unconstrained optimisation algorithms such as least mean square (LMS) or recursive least squares (RLS) algorithms can be invoked [8]. Fig. 2 shows the principle of a GSC, where the desired signal $d[n]$ is obtained via $\mathbf{w}_{\mathrm{q}}$,

$$d[n] = \mathbf{w}_{\mathrm{q}}^{\mathrm{H}} \cdot \mathbf{x}_n \qquad \text{with} \quad \mathbf{w}_{\mathrm{q}} = \mathbf{C}(\mathbf{C}^{\mathrm{H}}\mathbf{C})^{-1}\mathbf{f} \ . \tag{8}$$

The input signal $\mathbf{u}_n = \left[u_0[n] \ u_1[n] \ \ldots \ u_{M-S-1}[n]\right]^{\mathrm{T}}$ to the following multichannel adaptive filter (MCAF) is obtained by $\mathbf{u}_n = \mathbf{B}^{\mathrm{H}}\tilde{\mathbf{x}}_n$, whereby the $M \times (M-S)$ blocking matrix $\mathbf{B}$ must satisfy

$$\tilde{\mathbf{C}}^{\mathrm{H}}\mathbf{B} = \mathbf{0} \qquad \text{where} \quad \tilde{\mathbf{C}} = [\mathbf{c}_0 \ \cdots \mathbf{c}_{S-1}] \ . \tag{9}$$

In the next section, we will focus on the multiple-input optimisation process and introduce our subband adaptive GSC structure by employing the subband adaptive filtering techniques.

3. SUBBAND ADAPTIVE GENERALIZED SIDELOBE CANCELLER

Subband decompositions for adaptive filtering applications are commonly based on oversampled modulated filter banks (OSFB) as shown in Fig. 3 , where the input signal is divided into K frequency bands by analysis filters and then decimated by a factor N. Due to oversampling, i.e. $N < K$, a low alias level in the subband signals can be achieved. This is important since aliasing will limit the performance of an subband adaptive filtering (SAF) system [9]. Due to its lower update rate and fewer coefficients to represent an impulse response of a given length, the subband implementation only necessitates K/N^2 (K/N^3) of the operations required for a fullband adaptive algorithm with a complexity of $\mathcal{O}(L_{\mathrm{a}})$ ($\mathcal{O}(L_{\mathrm{a}}^2)$), where L_{a} is the total number of coefficients in the fullband realisation [3].

When applying SAF techniques to the MCAF in the GSC structure in Fig. 2, the subband setup as shown in Fig. 4 arises. There, the blocks labelled A perform an OSFB analysis operations, splitting the signal into K frequency bands each running at an N times lower sampling rate compared to the fullband input to the

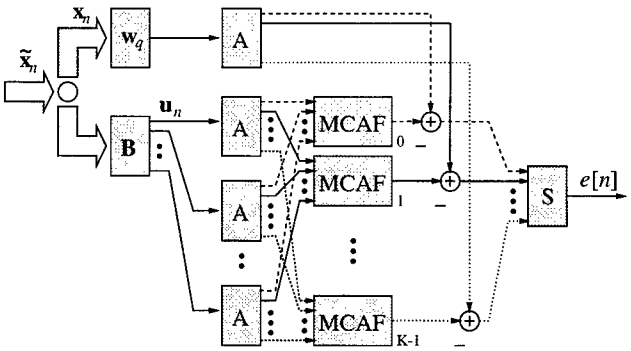


Fig. 4: Subband adaptive GSC; an independent MCAF is applied to each subband.

block. Within each subband, an independent MCAF is operated, and a synthesis filter bank, labelled S, recombines the different subsystem outputs to a fullband beamformer output $e[n]$.

In addition to the lower computational complexity of this subband adaptive GSC, it promises faster convergence speed for LMS-type adaptive algorithms because of the pre-whitening effect of the input signal. Next, we will give some simulation results to demonstrate the performance of our subband adaptive GSC.

4. SIMULATIONS AND RESULTS

In our simulation, we use a beamformer with $M = 15$ sensors and $J = 60$ coefficients for each attached filter. Each of the input signals $u_i[n]$ $(i = 0, 2, \cdots, M-S-1)$ and the reference signal $d[n]$ are divided into $K = 8$ subbands by an oversampled GDFT filter bank [10] with decimation factor $N = 6$ as characterised in Fig. 5. This subband adaptive GSC is constrained to received a signal of interest from broadside, which is white Gaussian with unit variance. The beamformer should adaptively suppress a broadband interference signal covering the frequency interval $\Omega = [0.25\pi; 0.75\pi]$ from $\vartheta = 30°$ and with a signal-to-interference ratio (SIR) of -24 dB. The sensor signals are corrupted by additive Gaussian noise at an SNR of 20 dB.

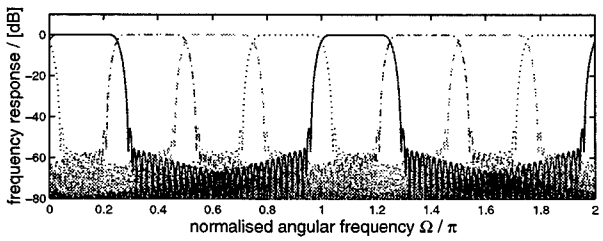


Fig. 5: Magnitude response of $K = 8$ channel filter bank decimated by $N = 6$.

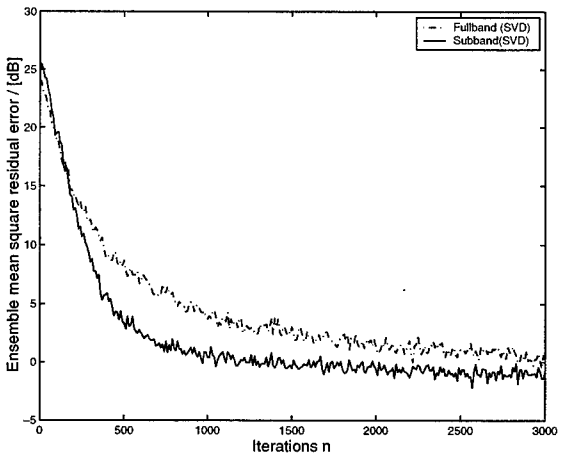


Fig. 6: Learning curves for simulation I ($S = 2$).

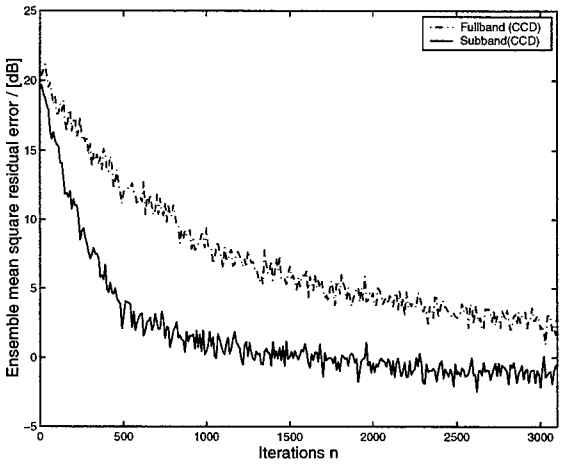


Fig. 7: Learning curves for simulation II ($S = 2$).

In order to compare the performance of our subband method with its fullband counterpart, we give four examples based on two commonly used approaches for building the blocking matrix, each with two different orders of constraints. The first approach is based on the cascaded columns of difference (CCD) method [11], the second on a singular value decomposition (SVD) [5]. The four examples are: (I) SVD method with first order derivative constraints ($S = 2$), (II) CCD method with $S = 2$, (III) SVD method with zero order derivative constraints ($S = 1$), (IV) CCD method with $S = 1$.

The step size in the NLMS adaptation for the first two examples is set to $\tilde{\mu} = 0.30$, and to $\tilde{\mu} = 0.20$ for examples (III) and (IV). Simulation results for these four cases are shown in Fig. 6 to Fig. 9, respectively. As a performance criterion, these figures display the ensemble mean square value of the residual error, which is defined as the difference between the beamformer output $e[n]$ and the appropriately delayed desired signal received from broadside.

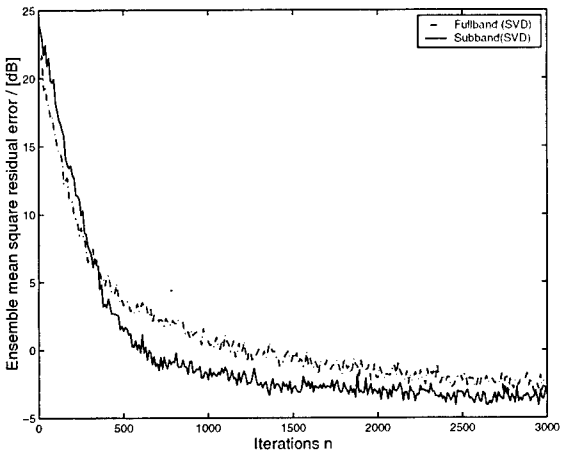


Fig. 8: Learning curves for simulation III ($S = 1$).

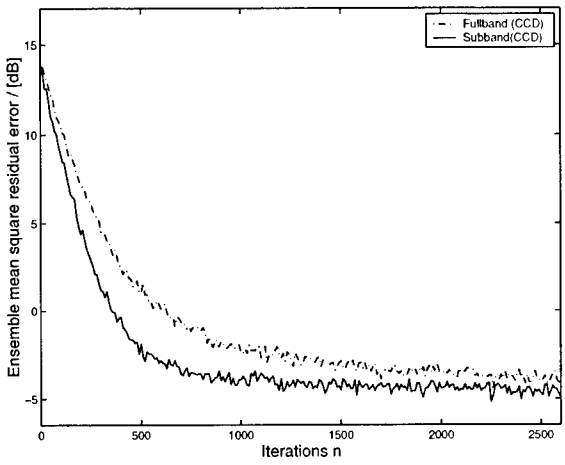


Fig. 9: Learning curves for simulation IV ($S = 1$).

From these results we can see that the subband adaptive method always has a faster convergence speed because of its pre-whitening effect. Comparing Fig. 6 with Fig. 7 and Fig. 8 with Fig. 9, we see the fullband performance changes according to different building of the blocking matrix, whereas the subband method has a relatively uniform performance independent of settings. With the added benefit of its low computational complexity due to processing in decimated subbands, the presented subband method outperforms the traditional fullband implementation.

5. CONCLUSIONS

A novel subband adaptive Generalized Sidelobe Canceller for broadband beamforming has been proposed. By employing subband adaptive filtering techniques, the computational complexity is greatly reduced. Moreover, the new method can also achieve a faster convergence speed because of its pre-whitening effect. Superiority of this new method to fullband implementation has been demonstrated by four examples based on different approaches for the blocking matrix and different orders of derivative constraints.

6. REFERENCES

[1] D. J. Chapman, "Partial Adaptivity for Large Arrays," *IEEE Trans AP*, 24(9):685–696, Sept. 1976.

[2] Y. Y. Wang, W. H. Fang, and J. T. Chen, "Improved Wavelet-Based Beamformers with Dynamic Subband Selection," in *Proc. IEEE AP-S Int. Symp.*, 1999.

[3] S. Weiss, R. W. Stewart, M. Schabert, I. K. Proudler, and M. W. Hoffman, "An Efficient Scheme for Broadband Adaptive Beamforming," in *Asilomar Conf. SSC*, I:496–500, Monterey, CA, Nov. 1999.

[4] L. J. Griffith and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Trans AP*, 30(1):27–34, Jan. 1982.

[5] K. M. Buckley and L. J. Griffith, "An Adaptive Generalized Sidelobe Canceller with Derivative Constraints," *IEEE Trans AP*, 34(3):311–319, Mar. 1986.

[6] O. L. Frost, III, "An Algorithm for Linearly Constrained Adaptive Array Processing," *Proc. IEEE*, 60(8):926–935, Aug. 1972.

[7] K.C. Huarng and C.C. Yeh, "Performance Analysis of Derivative Constraint Adaptive Arrays with Pointing Errors," *IEEE Trans AP*, 40(8):975–981, Aug. 1992.

[8] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Englewood Cliffs, 2nd edition, 1991.

[9] S. Weiss, R. W. Stewart, A. Stenger, and R. Rabenstein, "Performance Limitations of Subband Adaptive Filters," in *Proc. EUSIPCO*, III:1245–1248, Rodos, Greece, Sep. 1998,

[10] S. Weiss and R. W. Stewart, *On Adaptive Filtering in Oversampled Subbands*, Shaker Verlag, Aachen, Germany, 1998.

[11] N. K. Jablon, "Steady State Analysis of the Generalized Sidelobe Canceler by Adaptive Noise Canceling Techniques," *IEEE Trans AP*, 34(3):330–337, Mar. 1986.

AN EFFICIENT DESIGN OF FRACTIONAL-DELAY DIGITAL FIR FILTERS USING THE FARROW STRUCTURE

Carson K.S. Pun, Y.C. Wu, S.C. Chan, and K.L. Ho
Department of Electrical and Electronic Engineering,
The University of Hong Kong
Email: kspun@eee.hku.hk, ycwu@eee.hku.hk, scchan@eee.hku.hk, klho@eee.hku.hk

ABSTRACT

Fractional-delay digital filter (FD-DF), implemented using the Farrow structure, is very attractive in providing online tuning delay of digital signals. This paper proposes a new method for the design of such Farrow-based FD-DF using sum-of-powers-of-two (SOPOT) coefficients. Using the SOPOT coefficient representation, coefficient multiplication can be implemented with limited number of shifts and additions. Design examples show that the proposed method can greatly reduce the design time and complexity of the Farrow structure while providing comparable phase and amplitude responses.

I. INTRODUCTION

Fractional-delay digital filters (FD-DF) are very useful in delaying signals, which is required in many applications such as software radio, digital modems, arbitrary sampling rate conversion, time-delay estimation, etc. The Farrow structure [1] is particularly attractive because it can provide variable signal delay, making high-speed online tuning feasible. The basic principle of the Farrow structure is to approximate the impulse response of an ideal fractional-delay digital filter with delay μ by polynomial interpolation from the impulse responses of a limited set of fractional-delay digital filters with delays equally spaced within a range usually chosen to be μ = [-0.5,0.5]. To implement the Farrow structure, the signal will pass through these sub-filters and multiply with the appropriate powers of d to produce the output, Figure 1. The number of sub-filters required is equal to the order of the polynomial approximation used plus one. For precise control of the signal delay, the length of these sub-filters and the order of polynomial approximation will be considerable, requiring a large number of multipliers to implement this structure. As a result, higher power dissipation and larger area for VLSI implementation is expected.

In this paper, a novel algorithm for designing the Farrow structure with sum-of-powers-of-two (SOPOT) coefficients is proposed. SOPOT representation of filter coefficients is an attractive method for VLSI or hardware implementation of digital filters because multiplication of SOPOT coefficients can be implemented efficiently using hard-wired shifters and adders only (i.e. multiplier-less). More precisely, all the coefficients of the sub-filters in the Farrow structure are represented in SOPOT form and are implemented as additions and hardwired shifts. To further reduce the number of adders required in this structure, a transposed form of the sub-filters in the Farrow structure is employed which allows us to implement all the SOPOT multiplications with a single multiplier-block (MB) [3]. The application of MB to the efficient implementation of interpolated filters and filter banks were reported in [3]. Unfortunately, the design of such multiplier-less Farrow structure was missing. Using MB, the redundancy in the multiplication of the input with all the constant multipliers can be fully explored through the reuse of the immediate results generated. In principle, it is possible to remove all the redundancy found in the constant multipliers leading to a significant reduction in the number of adders required to implement the Farrow structure. The proposed design algorithm consists of two different steps: A FD-DF filter with real-valued coefficients is first designed. A flexible and efficient "random search" algorithm is then employed to search for the SOPOT coefficients while minimizing some criteria such as the number of SOPOT terms used subjected to a given frequency specification. This random search algorithm is similar to the mutation of genetic algorithm (GA) and the random walk in stimulated annealing. The main difference here is that we have limited its search space to a small neighborhood of the real-valued solution. This greatly shortens the search time to a few minutes. Our experience also indicates that excellent SOPOT solutions can be obtained in a reasonably time even when the filters involved are IIR. This is very difficult to achieve by GA even with design time several orders of magnitude longer [9]. The latter is mainly due to high sensitivities of the poles. Another advantage of this algorithm is that it can also be used to minimize directly the hardware cost such as adder cells of the filters, taking into account round-off and overflow noise [9]. There are many methods for designing FD-DF with real-valued coefficients [1][4][8]. In this work, the prototype fractional delay filters for the Farrow structure are designed using complex Chebyshev approximation, which is readily available in MATLAB. They are then interpolated to obtain the sub-filter coefficients for generating the MB. Design examples show that more than half of the adders in the SOPOT coefficients can be reduced with slight or negligible degradation in frequency responses, representing significant saving in hardware resources and power dissipation.

This paper is organized as follows: the efficient Farrow structure with SOPOT coefficients and multiplier block is introduced in Section II. Its design algorithm will be described in Section III followed by several examples in Section IV. Finally, conclusions are drawn in Section V.

II. THE EFFICIENT FARROW STRUCTURE

As mentioned earlier, one problem with the implementation of variable fractional-delay digital filters is the dependence of the impulse responses of the FD-DF on the delay parameter μ. More precisely, the output of the FD-DF, $y[(m+\mu)T]$, is given by

$$y[(m+D+\mu)T] = \sum_{i=0}^{N-1} x[(m-i)T] \cdot h_\mu(i), \qquad (1)$$

where $x[mT]$ is the input signal sampled at a period T, $h_\mu(i)$ is the FD-DF with delay $D+\mu$ and D is an integer constant, and N is the length of $h_\mu(i)$. To avoid the implementation of a large number of filters with different delays, Farrow [1] proposed to approximate each impulse response $h_\mu(i)$ with the following P^{th} order polynomial in delay value μ such that the delay control is independent of the filter coefficients.

$$h_\mu(i) = \sum_{n=0}^{P} b_n(i)\mu^n . \qquad (2)$$

Substituting (2) into (1) gives

$$y[(m+D+\mu)T] = \sum_{n=0}^{P} \left[\sum_{i=0}^{N-1} x[(m-i)T] \cdot b_n(i) \right] \mu^n . \qquad (3)$$

Figure 1 shows the Farrow structure for implementing equation (3), where the input signal is passed through a number of sub-filters $b_n(i)$, $n = 0,\ldots,P$, and is multiplied by the appropriate powers of μ to produce the output. Though the Farrow structure is very useful in providing a continuous value of signal delay, it still requires large number of multiplications for implementation,

0-7803-7011-2/01/$10.00 ©2001 IEEE

especially when P and N are large to provide very precise control of the frequency characteristics of the FD-DF. One method to avoid the expensive multipliers is to convert the filter coefficients in the following SOPOT representation

$$\hat{b}_n(i) = \sum_{k=1}^{L} b_{n,k}(i) \cdot 2^{a_k} , \quad (4)$$

with $b_{n,k}(i) \in \{-1,1\}$ and $a_k \in \{-l,.....,-1,0,1,.....,l\}$, where l is a positive integer and its value determines the range of the coefficients. L is the number of terms used in the coefficient approximation and is usually limited to a small number. The coefficient multiplications can therefore be implemented as limited shifts and additions, resulting in a significant reduction in implementation complexity. Very often, there is also significant redundancy in these SOPOT coefficients, which appears as common sub-expressions among different SOPOT coefficients. Due to the z-operator, it is somewhat difficult to remove these sub-expressions without increasing much shift registers. Fortunately, thanks to the transposed form of the sub-filters, the Farrow structure can be rewritten as in Figure 2. In this new structure, the input is multiplied to a number of constant coefficients. Hence, the common sub-expressions within the SOPOT coefficients can be eliminated [5][6] using a single multiplier-block, which further reduces the complexity of the Farrow structure.

A number of methods were proposed for designing the Farrow structure-based FD-DF [1]. Given these real-valued coefficients of the Farrow structure, it remains to determine the SOPOT coefficients $\hat{b}_n(i)$ that satisfy a given specification with the minimum number of adders in the multiplier-block. Commonly used distortion measures are the least squares and the minimax criterion. Without loss of generality, we shall employ the minimax criteria in this paper. Let $H(e^{j\omega},\mu)$ and $\hat{H}(e^{j\omega},\mu)$ be the frequency responses of the real-valued Farrow structure and its SOPOT counterpart, then the design problem can be stated as follows:

Given a set of initial Farrow coefficients $b_n(i)$, the maximum number of terms L in each coefficient and the dynamic range l of the coefficients, determine the SOPOT coefficients $\hat{b}_n(i)$ such that the maximum value of phase response error δ_p is minimized subject to a given peak amplitude error $\delta_a < \varepsilon$, where

$$\delta_p = \max_{0<\omega<\pi,|\mu|<0.5} \left[\left| \frac{\arg\{H(e^{j\omega},\mu)\} - \arg\{\hat{H}(e^{j\omega},\mu)\}}{\omega} \right| \right],$$

$$\text{subject to } \delta_a = \max_{0<\omega<\pi,|\mu|<0.5} \left| H(e^{j\omega},\mu) - \hat{H}(e^{j\omega},\mu) \right| < \varepsilon . \quad (5)$$

III. Design Procedure

The design of Farrow structure is first designing the prototype filters with specific fractional delay, then through interpolating these prototype filters to acquire the subfilters of the Farrow structure. The Farrow structure prototype filters are designed using Complex Chebyshev Approximation which is readily available in Matlab as cremez. For example, if the interpolation order is 3, then there are 4 subfilters. That is we are required to design a batch of prototype filters all with equal length (say 10 or more) with frequency response of $H_i(e^{j\omega}) = e^{-j\Delta_i\omega}, i = 1,...10$ and $\Delta_i = -0.5 + (i-1)/9$. For the same impulse coefficients, these will then be interpolate using a third order polynomial using least-square. Repeat each prototype filters coefficients for this interpolation procedure and these final polynomial coefficients are the initial full-precision Farrow structure filter coefficients.

The optimization procedure consists of two stages. First, the SOPOT coefficients of the initial Farrow structure such that the performance measure in (5) is minimized using a random search technique. Then, the minimum number of adders needed in the multiplier block is determined. The generation of the multiplier-block from the SOPOT coefficients follows the algorithms proposed in [3]. Let $\boldsymbol{b}_I$ be the vector containing the initial values $b_n(i)$'s of the Farrow structure. The principle of the random search algorithm is to generate random candidate SOPOT coefficients in the neighborhood of $\boldsymbol{b}_I$ so as to search for the optimal discrete solution. More precisely, a new coefficient vector $\boldsymbol{b}_{NEW}$ is generated by adding to it a random vector to the original coefficient vector $\boldsymbol{b}_I$ as follows

$$\boldsymbol{b}_{NEW} = \lfloor \boldsymbol{b}_I + \alpha \cdot \boldsymbol{b}_R \rfloor_{SOPOT} . \quad (6)$$

where α is a scale factor which control the size of neighborhood to be searched, b_R is a vector with its elements being random numbers in the range $[-1,1]$, and $\lfloor \cdot \rfloor_{SOPOT}$ is the rounding operation which convert its argument to the nearest SOPOT coefficients with maximum number of terms in each coefficient being L and dynamic range l. The performance measures δ_p and δ_a of the new coefficients are then calculated. The set that yields the minimum phase error with a given peak amplitude error ε is the optimum solution under the constraints of L and l. As this is a random search algorithm, the longer the searching time, the higher the chance of founding the optimal solution.

There are several advantages of this algorithm. First of all, with the computational power of nowadays personal computer (PC) the time for obtaining high quality solutions is manageable. In fact, for the problem considered here, the overall design time only takes less than 5 minutes to complete on a typical Pentium-400 PC using Matlab 5.3., including both the design of SOPOT coefficients and the multiplier-block design. Secondly, it is applicable to problem with general objective function probably with very complicated inequality constraints. Moreover, a set of possible solutions representing different tradeoffs between computational complexity and performance will be generated during the search. Therefore, it helps one to achieve an appropriate tradeoff for a given application. It is also possible to combine the two stages together to improve the performance but the computational time will be greatly increased.

IV. Design Examples

Example 1

As a simple example, the famous cubic Lagrange interpolator [7], with coefficients shown in Table 1, is implemented using the proposed algorithm. The passband bandwidth under optimization is from 0 to 0.4π. The original peak ripple error and phase delay error are, respectively, 0.048769 and 0.008179. By multiplying all the coefficients by 6, they can be converted to simple integers. Using the concept of multiplier-block, the additions in implementing the multipliers 3 and 6 can be shared to reduce the total number of adders. The final Farrow structure implemented using the multiplier-block is shown in Figure 3. (*The ">>n" sign in Figure 3, means a hard-wired shift towards the LSB for n-bit. As for the "<<n" sign, it means a hard-wired shift towards the MSB for n-bit.*) It requires only 3 adders including the scaling (1/6) in SOPOT coefficients at the output.

Example 2

As another example, let's consider the coefficients provided by Farrow in [1]. The SOPOT coefficients obtained by the random search algorithm are shown in Table 2. The bandwidth under consideration for this filter is from 0 to 0.6π. The original peak ripple error and phase delay error are 0.006271 and 0.0032,

respectively; whereas for the SOPOT Farrow structure, the peak ripple error and phase delay error are 0.005371 and 0.0046, respectively. After common sub-expressions elimination, the multiplier-block requires only 13 adders compared favorably with 32 real multiplications in the original Farrow structure. These results show that the number of adders can be drastically reduce by using multiplier-block. The resultant Farrow structure filter has a much lower complexity than the real-valued Farrow structure but providing nearly the same phase delay and amplitude response.

Example 3

Our last example will be on a Farrow structure with higher polynomial order. The prototype filters are designed using complex Chebyshev approximation and interpolated by a 5th order polynomial. The bandwidth under consideration is from 0 to 0.75π. The SOPOT coefficients are shown in Table 3. After common-expression elimination, the sub-filters require only 18 adders to achieve a peak ripple error of 0.026376 and maximum phase delay error of 0.0059. The frequency responses of the proposed structure and its real-values counterpart are shown in Figure 4. It can be seen that they are very close to each other. The performance and arithmetic complexity of the various implementations are summarized in Table 4.

	$b_3(\cdot)$	$b_2(\cdot)$	$b_1(\cdot)$	$b_0(\cdot)$
0	1/6	0	-1/6	0
1	-½	½	1	0
2	½	-1	-½	1
3	-1/6	½	-1/3	0

Table 1. Coefficients of the Lagrange-based FD-DF.

	$b_3(\cdot)$	$b_2(\cdot)$	$b_1(\cdot)$	$b_0(\cdot)$
0	$-2^{-6}+2^{-8}$	$2^{-4}-2^{-7}$	0	$-2^{-6}+2^{-10}$
1	$2^{-4}+2^{-6}-2^{-8}$	$-2^{-2}+2^{-5}+2^{-8}$	-2^{-6}	$2^{-4}-2^{-7}$
2	$-2^{-1}+2^{-3}-2^{-5}$	$2^{-1}+2^{-3}+2^{-7}$	$2^{-3}-2^{-5}+2^{-8}$	$-2^{-3}-2^{-5}-2^{-9}$
3	$2^{-0}-2^{-3}+2^{-5}$	$-2^{-1}+2^{-5}+2^{-8}$	$-2^{-0}-2^{-2}+2^{-5}$	$2^{-1}+2^{-3}-2^{-7}$
4	$-2^{-0}+2^{-3}-2^{-5}$	$-2^{-1}+2^{-5}+2^{-8}$	$2^{-0}+2^{-2}-2^{-5}$	$2^{-1}+2^{-3}-2^{-7}$
5	$2^{-1}-2^{-3}+2^{-5}$	$2^{-1}+2^{-3}+2^{-7}$	$-2^{-3}+2^{-5}-2^{-8}$	$-2^{-3}-2^{-5}-2^{-9}$
6	$-2^{-4}-2^{-6}+2^{-8}$	$-2^{-2}+2^{-5}+2^{-8}$	2^{-6}	$2^{-4}-2^{-7}$
7	$2^{-6}-2^{-8}$	$2^{-4}-2^{-7}$	0	$-2^{-6}+2^{-10}$

Table 2. SOPOT coefficients for the proposed Farrow structure.

	$b_5(\cdot)$	$b_4(\cdot)$	$b_3(\cdot)$	$b_2(\cdot)$	$b_1(\cdot)$	$b_0(\cdot)$
1	$2^{-5}-2^{-7}$	$-2^{-3}+2^{-6}$ -2^{-10}	$-2^{-5}-2^{-7}$	$2^{-3}+2^{-5}$ $+2^{-7}$	2^{-7}	$-2^{-5}-2^{-9}$
2	$-2^{-4}-2^{-6}$	$2^{-2}-2^{-5}$ -2^{-7}	$2^{-3}+2^{-6}$ $+2^{-10}$	$-2^{-1}+2^{-3}$ $+2^{-6}$	$-2^{-5}+2^{-9}$	$2^{-4}+2^{-6}$ -2^{-10}
3	$2^{-2}-2^{-4}$ -2^{-8}	$-2^{-2}-2^{-5}$ -2^{-7}	$-2^{-1}-2^{-6}$ -2^{-8}	$2^{-0}-2^{-2}$ $+2^{-5}$	$2^{-3}-2^{-7}$	$-2^{-2}+2^{-4}$ $+2^{-7}$
4	$-2^{-2}-2^{-5}$ -2^{-7}	$2^{-3}+2^{-5}$ -2^{-8}	$2^{-0}+2^{-4}$ $+2^{-8}$	$-2^{-1}-2^{-5}-$ 2^{-7}	$-2^{-0}-2^{-2}$	$2^{-1}+2^{-3}$ -2^{-9}
5	$2^{-2}+2^{-5}$ $+2^{-7}$	$2^{-3}+2^{-5}$ -2^{-8}	$-2^{-0}-2^{-4}$ -2^{-8}	$-2^{-1}-2^{-5}-$ 2^{-7}	$2^{-0}+2^{-2}$	$2^{-1}+2^{-3}$ -2^{-9}
6	$-2^{-2}+2^{-4}$ $+2^{-8}$	$-2^{-2}-2^{-5}$ -2^{-7}	$2^{-1}+2^{-6}$ $+2^{-8}$	$2^{-0}-2^{-2}$ $+2^{-5}$	$-2^{-3}+2^{-7}$	$-2^{-2}+2^{-4}$ $+2^{-7}$
7	$2^{-4}+2^{-6}$	$2^{-2}-2^{-5}$ -2^{-7}	$-2^{-3}-2^{-6}$ -2^{-10}	$-2^{-1}+2^{-3}$ $+2^{-6}$	$2^{-5}-2^{-9}$	$2^{-4}+2^{-6}$ -2^{-10}
8	$-2^{-5}+2^{-7}$	$-2^{-3}+2^{-6}$ -2^{-10}	$2^{-5}+2^{-7}$	$2^{-3}+2^{-5}$ $+2^{-7}$	-2^{-7}	$-2^{-5}-2^{-9}$

Table 3. SOPOT coefficients for the proposed Farrow structure.

	Ex 1.	Ex 2.	Ex 3.
Real-valued (Multipliers)	4	32	48
SOPOT (Adders)	N/A	48	78
Multiplier-block (Adders)	3	13	18
% of adders reduction	25%	72.92%	76.92%
Design Time used on Pentium-400 (Minutes)	N/A	4	6

Table 4. Comparison between various implementation schemes.

V. Conclusion

A new method for the design Farrow-based FD-DF using sum-of-powers-of-two (SOPOT) coefficients is proposed. This method has the advantage of fast design time with good frequency response of the Farrow structure and able to reduce the no. of terms of SOPOT coefficient in order to reduce hardware complexity. Design examples show the robustness of this method for designing Farrow structure filters with different specifications.

References

[1] C. W. Farrow, "A continuously variable digital delay element", in *Proc. of ISCAS 1988*, pp.2641-2645.

[2] A. G. Dempster and N. P. Murphy, "Efficient interpolators and filter banks using multiplier blocks," *IEEE Trans. Signal Processing*, vol. 48. pp. 257-261, Jan. 2000

[3] A.G. Dempster and M.D. Macleod, "Use of minimum-adder multiplier blocks in FIR digital filters," *IEEE Trans. CAS. II*, vol. 42, no.9, pp. 569-577, Sep. 1995.

[4] T. I. Laakso, V. Valimaki, M. Karjalainen and U. K. Laine, "Splitting the unit delay, tools for fractional delay filter design," *IEEE Signal Processing Magazine*, pp.30-60, Jan. 1996.

[5] M. Potkonjak, M. B. Srivastava and A. P. Chandrakasan, "Multiple constant multiplications: efficient and versatile framework and algorithms for exploring common subexpression elimination," *IEEE Trans. Computer-aided Design*, vol. 15, pp.151-165, Feb. 1996.

[6] R. Pasko, P. Schaumont, V. Derudder and D. Durackova, "Optimization method for broadband modem FIR filter design using common subexpression elimination," in *Proc. of Symposium on System Synthesis 1997*, pp.100-106.

[7] L. Erup, F. Gardner, and R. A. Harris, "Interpolation in digital modems — part II: implementation and performance," *IEEE Trans. Commun.*, vol. 41, pp. 908-1008, Jun. 1993.

[8] J. Vesma and T. Saramaki, "Design and properties of polynomial-based fractional delay filters," in *Proc. ISCAS 2000*, pp.104 –107.

[9] Carson K.S. Pun, S C. Chan and K. L. Ho., "Efficient design of a class of multiplier-less perfect reconstruction two-channel filter banks and wavelets with prescribed output accuracy," to appear in *11th IEEE Workshop on Statistical Signal Processing 2001.*

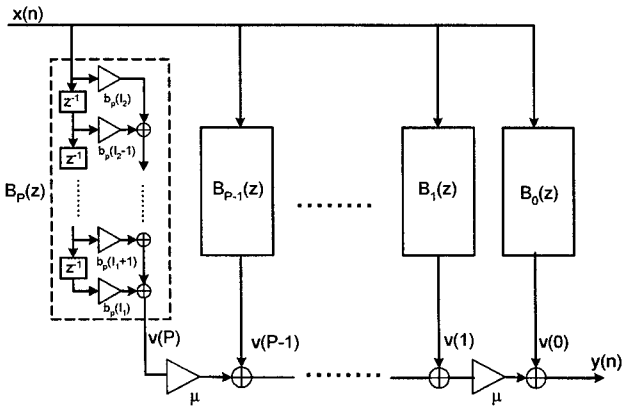


Figure 1. Original Farrow structure.

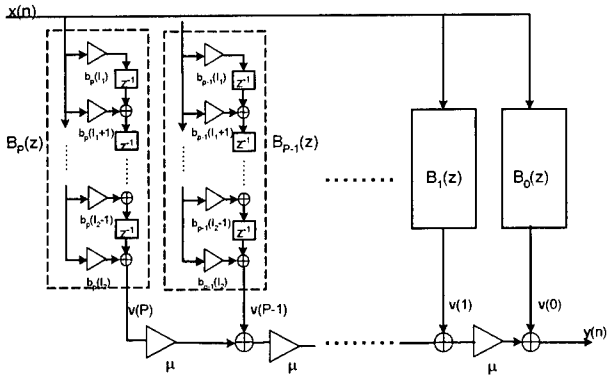


Figure 2. Proposed implementation of the Farrow structure.

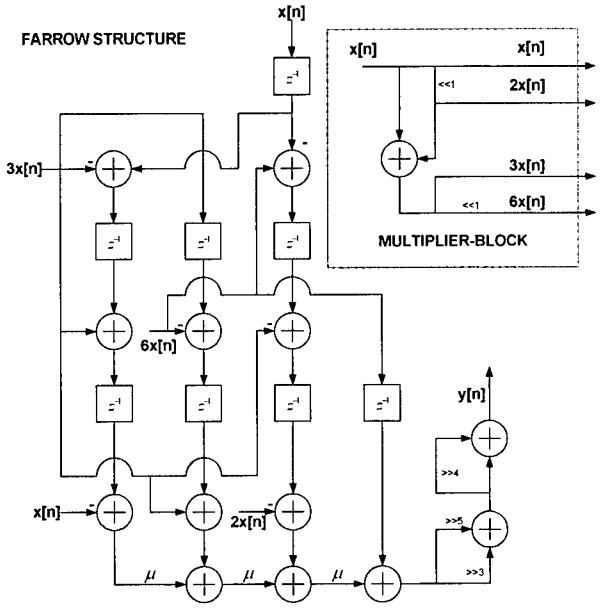
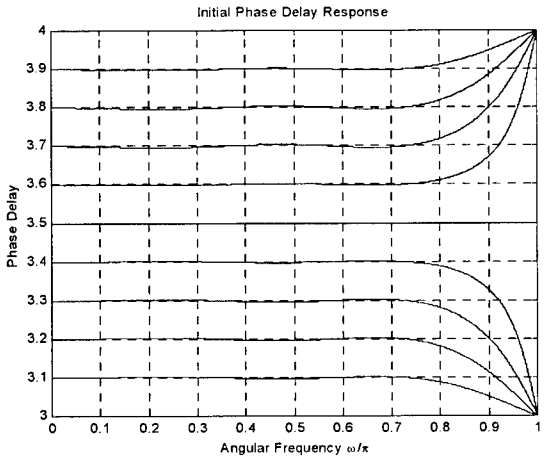
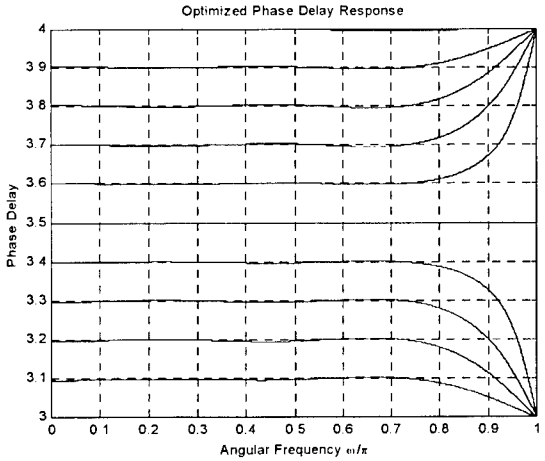
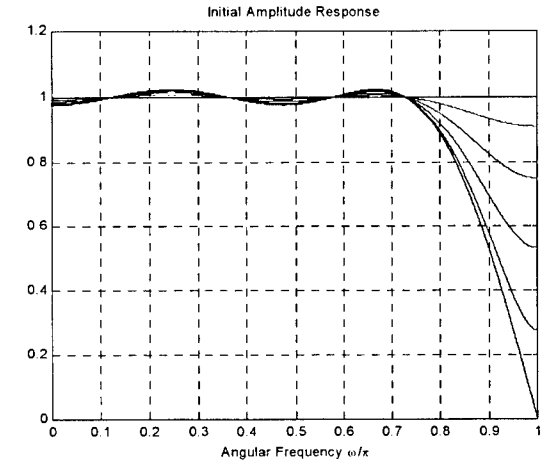


Figure 3. Farrow structure of Lagrange interpolator in Example 1.







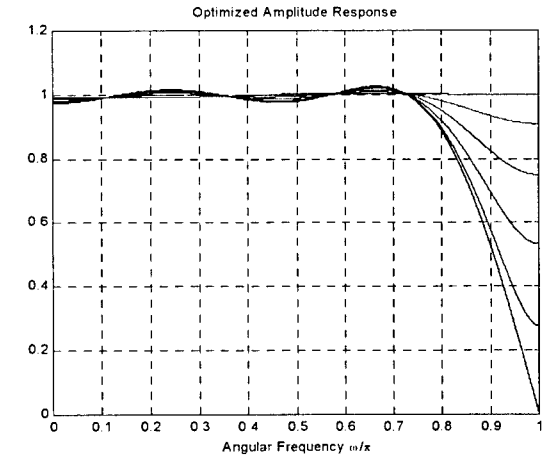


Figure 4. Frequency and phase delay responses in Example 3.

EFFICIENT DESIGN OF A CLASS OF MULTIPLIER-LESS PERFECT RECONSTRUCTION TWO-CHANNEL FILTER BANKS AND WAVELETS WITH PRESCRIBED OUTPUT ACCURACY

Carson, K.S. Pun, S.C. Chan and K.L. Ho

Department of Electrical and Electronic Engineering,
The University of Hong Kong

Email: kspun@eee.hku.hk, scchan@eee.hku.hk, klho@eee.hku.hk

ABSTRACT

This paper proposes a novel algorithm for the design and hardware reduction of a class of multiplier-less two-channel PR filter banks (FBs) using sum-of-powers-of-two (SOPOT) coefficient. It minimizes a more realistic hardware cost, such as adder cells, subject to a prescribe output accuracy taking into account of the rounding and overflow effects, instead of using just the SOPOT terms as in conventional method. Furthermore, by implementing the filters in the FBs using multiplier-block (MB), significant overall saving in hardware resources can be achieved. An effective random search algorithm is also proposed to solve the design problem, which is also applicable to PR IIR FBs with highly nonlinear objective functions.

I. INTRODUCTION

Perfect reconstruction (PR) multirate filter banks (FB) have important applications in signal analysis, signal coding and the design of wavelet bases. A number of techniques for designing linear-phase and low-delay PR two-channel filter banks are now available [1][2][3]. Recently, there is an increasing interest in designing PR filter banks with very low implementation complexity. One of the applications is to provide efficient hardware implementation of the 9/7 wavelet filter for the JPEG2000 standard. FBs using sum of powers-of-two (SOPOT) coefficients are particularly attractive for VLSI or hardware implementation because multiplication of SOPOT coefficients can be implemented efficiently using hard-wired shifters and adders only (i.e. multiplier-less). The design of such SOPOT PR FBs using the 2-channel lossless lattice structure and genetic algorithm was studied in [6]. Another family of multiplier-less PR two-channel FIR/IIR FB and wavelets, using SOPOT coefficients and the structure in [1], was studied recently by the authors in [4][7]. They are attractive because of their low hardware and design complexities. Furthermore, the PR condition is structurally imposed and is robust to coefficient quantization.

It is well known that there are two sources of error in implementing a digital filter: coefficient round-off error and signal round-off error [10]. Coefficient round-off error happens when the real-valued coefficients of the filter, obtained say by the Park-McClellan algorithm, are rounded to their fixed-point representations to simply the hardware implementation. The frequency response of the filter is therefore changed, and might not satisfy the specification any more. On the other hand, signal round-off error occurs when overflow occurs due to insufficient internal wordlength and improper scaling; and when rounding is performed for long intermediate data after multiplications with the filter coefficients. Signal round-off error is usually more difficult to handle in hardware implementation because complicated hardware for detecting overflows, etc., would significantly slow down the throughput of the system. The SOPOT FBs mentioned above are free from coefficient round-off noise because the FBs are optimized using the SOPOT coefficients as variables. Unfortunately, most of these methods only focused on minimizing the number of SOPOT terms to meet a given frequency specification, and pay little attention to signal round-off error. In order to satisfy a given output accuracy, one usually employs a fixed and long wordlength for all intermediate data, which means increased hardware complexity. Therefore, the design problem should be to minimize the hardware complexity of the system while satisfying the given frequency specification and the output accuracy. The hardware complexity could be the number of adder cells and registers used in the FBs, which is related to the exact wordlength used for each intermediate data. The output accuracy of a digital filter is usually specified statistically by its output noise power due to the rounding operations performed, using a given noise model. For fine quantization, round-off noise is usually modeled as white and is uncorrelated with the signal and other noise sources. To satisfy a given output accuracy (say 16-bit), one has to determine the appropriate scaling and wordlength of each intermediate data to avoid signal overflow and to achieve a noise power less than the given specification (say -96dB for 16-bit accuracy).

The purpose of this paper is to provide a solution to the above problem, with particular emphasis on the SOPOT FBs that we have proposed in [4][7]. This class of PR FBs is chosen because the required stopband attenuation and system delay can easily be achieved using simple design formula for order estimation and the efficient Park-McClellan design algorithm. Using the real-valued coefficients so obtained as initial guess, the SOPOT coefficients and the internal wordlength of all intermediate data are jointly optimized using a novel random search algorithm to minimize some measure of the hardware complexity, while satisfying the given specification. In this work, both the number of adders and their adder cells are minimized because they constituted over 70% of the total hardware cost as compared with other components such as latches. The random search algorithm is similar to the mutation of genetic algorithm (GA) and the random walk in stimulated annealing. The main difference here is that we have limited its search space to a small neighborhood of the real-valued solution obtained in [4][7] using the Park-McClellan algorithm. This greatly shortens the search time to a few minutes. Moreover, for IIR FBs, excellent SOPOT solutions can be obtained in a reasonably time, which cannot be achieved by GA even with design time several orders of magnitude longer. The latter is mainly due to high sensitivities of the poles. The number of adders required to implement the SOPOT multiplications is further reduced by using the technique of "multiplier-block" (MB) [9]. By using MB, redundancy in the SOPOT coefficients is removed. Design examples demonstrated that our design method is very efficient and capable of reducing dramatically the hardware complexity of the FBs, while meeting the given specifications. More difficult 2-channel SOPOT IIR PR FBs can also be designed using the proposed method. Our paper is organized as follows: in section II, the SOPOT FBs considered and the MB technique will be described. The round-off-noise and overflow problems will be addressed in Section III. Section IV is devoted to the 'Random search' design algorithm. This is followed by several design examples in Section V. Finally conclusions are drawn in section VI.

II. 2-CHANNEL PR SOPOT FB

Fig. 1 shows the structure of the PR FB proposed in [5]. The functions $\alpha(z)$ and $\beta(z)$ can be linear-phase FIR, nonlinear-phase FIR, or IIR functions, without affecting the PR conditions. It can be shown that the lowpass and highpass analysis filters are given by $H_0(z)=(z^{-2N}+z^{-1}\beta(z^2))/2$ and $H_1(z)=-\alpha(z^2)H_0(z)+z^{-2M-1}$, respectively. It is also possible to realize wavelet bases from this FBs by imposing certain regularity condition on $H_0(z)$ and $H_1(z)$. Details regarding their design can be found in [4]. In the multiplier-less FB [4][7], each coefficient in $\alpha(z)$ and $\beta(z)$ is represented as the following sum of powers-of-two coefficients (SOPOT) or canonical signed digits (CSD), $b=\sum_{k=0}^{L-1} a_k \cdot 2^{-b_k}$, where a_k is either 1 or –1, and $b_k \in \{-l_L,....1,0,...l_U\}$. The larger the numbers l_L, l_U, and L, the closer the SOPOT approximation will be to the original real number. In practice, the number of non-zero terms is usually kept to a small number while satisfying a given specification so that the multiplication can be implemented as a limited number of shift and add (subtract) operations, giving rise to multiplier-less realization. Multiplier-less filter banks and wavelet bases with linear-phase and low system delay can be obtained from this structure by searching for the SOPOT coefficients using the genetic algorithm [6][7]. As mentioned earlier, the number of adders needed to implement $\alpha(z)$ and $\beta(z)$ can further be reduced

0-7803-7011-2/01/$10.00 ©2001 IEEE

by rewriting them in transposed form. It can be seen that instead of multiplying the delayed input samples with the filter coefficients as in the direct form, the input sample is now multiplied with all the coefficients. This can be efficiently implemented using a multiplier block (MB) [9]. Let's consider a simple example with two filter coefficients: 3 and 21. The SOPOT representations of these two numbers are: $3=2^1+1$ and $21=2^4+2^2+1$. This requires 3 adders and 3 shifts. If implemented in a MB, the multiplication of the input with the coefficient 3 will also be generated by decomposing 3 as 2^1+1, requiring one addition. The multiplication with 21, however, can be simplified by re-using the intermediate result generated by the first filter coefficient '3' as $21=3\cdot 7=3\cdot(2^4-1)$. Actually, the intermediate result, after multiplication by 3, is multiplied by 7, which requires one less adder than generating 21 directly. In principle, it is possible to remove all the redundancy found in the constant multipliers leading to a realization with the *minimum number of adders*. This can drastically reduce the number of adders required for realizing such FBs when there is a large number of filter coefficients to be implemented in the transposed form FIR structure (around 50% in our example).

III. Round-Off Noise and Overflow Analyses

1. Analysis of Round-off Noise

As mentioned earlier, round-off noise occurs when rounding is performed during arithmetic computation. In fixed-point arithmetic, round-off operation is usually performed after multiplication to limit the wordlength of the intermediate data in order to save hardware resources. Round-off error is thus generated. Due to the difficulty in analyzing exactly the rounding error, they are usually treated as white random process, uncorrelated with the signal and other noise sources. For rounding operation, quantization noise will have zero mean and a variance $\sigma^2=\Delta^2/12$, where Δ is the quantization step-size, which is determined by the number of fractional bits that is retained after multiplication.

Consider the transposed form FIR filter in figure 2. The blocks **D** and **Q{.}** represent respectively a register and the round-off operator. Any signal in this filter, for example the input signal $x[n]$, has a fixed-point representation of the form $<n|m>$, which means that the total wordlength is $n+m$ bits where n represents the integer bits (including the sign bit) and m the fractional bits. For notation convenience, any signal will be represented as $x[n]:<n|m>$, meaning that it has n integer bits and m fractional bits. Now, consider the input sample $x[n]:<1|7>$, which is a 8-bit number gated into the digital filter at every clock cycle. It will be multiplied by $h[0]:<1|9>$ and $h[1]:<1|7>$. If no rounding is performed, the fixed-point formats of the products $x[n]h[0]$ and $x[n]h[1]$ will be $<1|16>$ and $<1|14>$, respectively. Suppose that the products are rounded by the operator **Q{.}** to the format: $<1|14>$. Since the wordlength of $x[n]h[1]$ before and after rounding is equal, so there is no round-off noise ($e_1[n]=0$). While for the signal $x[n]h[0]$, the wordlength is shortened from $<1|16>$ to $<1|14>$, hence, a round-off noise, $e_0[n]\neq 0$, with a power of $P=(2^{-13})^2/12$ is generated. In general, if R, the number of bits in the fractional part of the fixed-point representation, is rounded to B ($B<R$), then the round-off noise power is given by: $P_e=2^{-2(B-1)}/12$ [10]. If there are M such rounding noise sources in the transposed form, the total noise power at the output is given simply by their sum: $P_{total}=\sum_{k=1}^{M}P_{e_k}=\sum_{k=1}^{M}2^{-2(B_k-1)}/12$. For a general digital filter, the k*th* noise source might pass through a transfer function with z-transform $H_k(z)=\sum_{n=0}^{L_k}h(n)z^{-n}$, then the total output noise power is $P_{total}=\sum_{k=1}^{M}P_{e_k}|H(0)|^2$, assuming that they are uncorrelated. The output accuracy, in terms of the number of fractional bits, is therefore given by $(1/6)\cdot 10\cdot\log_{10}(P_{total})$. In general, to have 16-bit output accuracy, the output noise-power must be below –96 dB level. From these results, we can see that, the larger the number of noise sources, the lower will be the accuracy of the computation. The noise power can however be reduced by increasing the wordlength for the fractional bits, at the expense of increased hardware complexity.

2. Preventing Overflow

Another important source of error is signal overflow [10], which occurs when the allocated wordlength in the integer part is insufficient to represent correctly the fixed-point representation of the output after addition (such as the adders in Fig. 2). In order to avoid overflow, we must allocate more bits to the integer part of the register (say **D** in Fig.2). We are given the option to retain or decrease the number of bits in the fractional part, depending on the required accuracy. To determine whether overflow will occur for a given adder, we can compute certain measures of the transfer function from the input to this particular adder. Here, we prefer a more conservative measure using the absolute sum of the impulse response, i.e. L1 scaling. For example, let x_{max} be the maximum input to a FIR transposed form digital filter $H(z)=\sum_{k=0}^{L}h(k)z^{-k}$ as shown Fig. 2. Then the maximum (or worse case) value at the output of the lth adders of the FIR filter is $d_l=\left(\sum_{k=1}^{l}|h(k)|\right)x_{max}, l=0,\ldots L$. From these values, it is possible to determine the required integer wordlength at each position to avoid any overflow. The number of fractional bits will be optimized to satisfy the given output accuracy. It should be noted that there are other scaling method such as L2 scaling which can also be used. However, there is still a small probability that overflow will occur. In digital signal processor, special hardware is usually used to detect the present of overflow and the result will be clipped to the maximum/minimum values of the representation (saturation arithmetic).

IV. The Design Algorithm

Our design method consists of two parts. First, the parameters of the filters $\alpha(z)$ and $\beta(z)$ such as their coefficients and their order (parameters N and M) are determined from the frequency specification (system delay, stopband attenuation, cutoff frequencies) using the method in [4]. Then, the SOPOT coefficients are determined using a random search algorithm to generate the MB (see 1 below). The hardware complexity of the FBs are then minimized while maintaining the output accuracy using the noise models mentioned earlier (see 2 below).

1. *Search for the SOPOT filter coefficients.*

The optimization procedure consists of two stages. First, a random search algorithm, to be discussed in the sequel, is used to search for the SOPOT coefficients of $\alpha(z)$ and $\beta(z)$ such that a given performance measure is minimized. Then, the minimum number of adders needed in the multiplier block is determined. The generation of the multiplier-block from the SOPOT coefficients follows the algorithms proposed in [9]. Let $\mathbf{x}_I$ be the vector containing the real-valued coefficients of $\alpha(z)$ and $\beta(z)$ obtained by the method in [4]. The principle of the random search algorithm is to generate random candidate SOPOT coefficients in the neighborhood of $\mathbf{x}_I$ so as to search for the optimal discrete solution. More precisely, a new coefficient vector $\mathbf{x}_{NEW}$ is generated by adding to it a random vector to the original coefficient vector $\mathbf{x}_I$ to form $\mathbf{x}_{NEW}=\left[\mathbf{x}_I+\alpha\cdot\mathbf{x}_R\right]_{SOPOT}$, where α is a scale factor which controls the size of the neighborhood to be searched, $\mathbf{x}_R$ is a vector with its elements being random numbers in the range $[-1,1]$, and $[\cdot]_{SOPOT}$ is the rounding operator which converts its argument to the nearest SOPOT coefficients with maximum number of terms in each coefficient being L and dynamic range l_U and l_L. The following objective function, which is the minimax error between the desired frequency response $H_d(e^{j\omega})$ and the

frequency response $H(e^{j\omega},\hat{\mathbf{x}})$ calculated using the candidate $\hat{\mathbf{x}}$ in the frequency band of interest $\omega \in S$, is minimized:

$$score = \max_{\omega \in S} \left(\left| H(e^{j\omega},\hat{\mathbf{x}}) - H_d(e^{j\omega}) \right| \right). \quad (1)$$

The process is repeated with different vector $\hat{\mathbf{x}}$ so that the SOPOT space in the neighborhood of $\hat{\mathbf{x}}$ is sampled randomly. Since the sampled solutions are close to the real-valued optimal solution, their frequency responses will also be close to the ideal one, but with different hardware complexity. The set that yields the minimum score with a given number of terms is recorded. As this is a random search algorithm, the longer the searching time, the higher the chance of finding the optimal solution.

2. Minimization of the filter banks hardware structures with prescribed output accuracy

After the MB is generated, the maximum wordlength of all the products, x[n]h[i], i=0,...L, in Fig. 2, is calculated. If we do not perform any rounding using the operator **Q**{.}, and sufficient wordlength is allocated to all adders, then there is no rounding error. Of course, this will require excessive hardware cost, especially when the output accuracy is low. Our goal is to determine the format of the rounded signals, **Q**{x[n]h[i]}, i=0,...,L, to satisfy the output accuracy. Suppose that the formats are stored in a vector $\boldsymbol{\delta}$. Given the rounded output format of the MB, $\boldsymbol{\delta}$, one can determine, using the method described in Section III.2, the formats of the registers, **D**'s, and the structure of the adders, in order to avoid any overflow. The fractional part for those scaled output, to prevent overflow, can either retain its wordlengh or reduce it by one as mentioned in Section III.2. This option is stored in a vector $\boldsymbol{\delta}_f$, to be optimized together with $\boldsymbol{\delta}$. The noise power at the filter output of the filter is readily computed accordingly to the analysis described in Section III.1. Note the output noise power from $\alpha(z)$ and $\beta(z)$ will be evaluated and their contributions at the lowpass (and highpass) analysis filters will be properly summed, using their respective power transfer functions mentioned in Section III.1. Our design algorithm seeks to lower the wordlength of each intermediate data and hence the complexity format as specified in $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_f$ to minimize the hardware cost. Using $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_f$, the hardware cost, C, given by the adder cells in the MB and the subsequent adders in Fig. 2 can be evaluated. In summary, the design problem is

$$\min_{(\boldsymbol{\delta},\boldsymbol{\delta}_f)} C(\boldsymbol{\delta},\boldsymbol{\delta}_f) \text{ subject to } P_{total} \le P_{spec}, \quad (2)$$

where P_{total} is the output noise power at the lowpass and highpass filters and P_{spec} is the specified output accuracy. Using a random search algorithm similar to that mentioned in Section IV.1, the vector $(\boldsymbol{\delta},\boldsymbol{\delta}_f)$ is searched in the neighborhood of their full precision values $(\boldsymbol{\delta},\boldsymbol{\delta}_f)_\infty$ (that is no rounding) for feasible solutions that satisfying the given output accuracy. The one with the minimum hardware cost $C(\boldsymbol{\delta},\boldsymbol{\delta}_f)$ is declared as the solution of this problem. There are several advantages of this algorithm. First of all, with the computational power of nowadays personal computer (PC) the time for obtaining high quality solutions is manageable, especially when an initial real-valued solution is available by some means. In fact, for the problem considered here, the overall design time is less than 10 minutes using a Pentium-400 PC with Matlab 5.3, including both the design of SOPOT coefficients, generation of the MB and the internal wordlength allocation. Secondly, it is applicable to problems with general objective functions probably with very complicated inequality constraints, as illustrated in this work. It is also possible to combine the search with the MB generation processes together for better performance but the computational time will be greatly increased. We now present a few design examples.

V. Design Examples

5.1. Two-channel PR FBs with $\beta(z)$ and $\alpha(z)$ FIR filters

To demonstrate the effectiveness of our algorithm for solving the complicated design problem, a two-channel FB with the following frequency specification is designed: passband and stopband cutoff frequencies ω_p= 0.4π, and ω_s=0.6π, respectively; stopband attenuation is 39 dB, system delay = 23. From the design procedure in [4], the parameters N and M are determined to be 3 and 8, respectively. The wordlength of the input is 8-bit and is normalized to be less than 1, i.e. in <1|7> format. The required output accuracy is at least 16-bit for fractional part without overflow. The frequency response of the final SOPOT FB is shown in figure 3, and the details of its optimized structure are summarized in table 1. The reduction of the number of adders obtained by using MBs to implement $\beta(z)$ and $\alpha(z)$ is around 50%. It can also be observed that the number of adder cells is significantly reduced by 27% (compared with a fixed wordlength of 24 bits using MBs to satisfying the same output accuracy) using the proposed random search algorithm to minimize the necessary internal wordlength, while satisfying the prescribed output accuracy of 16-bit. The overall design takes about 10 minutes on a typical Pentium-533 computer.

5.2. Two-channel PR FB with $\beta(z)$ IIR and $\alpha(z)$ FIR

To demonstrate the effectiveness of our random search algorithm in designing SOPOT PR IIR FB, $\beta(z)$ is chosen as an IIR filter while $\alpha(z)$ as an FIR filter. In order to guarantee the stability of the IIR filter, the denominator of $\beta(z)$ is factorized as a lattice structure and the magnitude of the lattice coefficients are forced to be less than 1. They are then used as optimization variable in the random search algorithm. The design specifications are: passband cutoff frequency ω_p =0.4π, stopband cutoff frequency ω_s =0.6π. N and M are determined to be 4 and *11*, respectively. The real-valued filter coefficients are obtained by the method in [11]. The SOPOT coefficients of the FBs obtained by the proposed algorithm are shown in table 2, and the frequency response is shown in figure 4. The frequency characteristic is very good despite the high nonlinearity of the objective function for the IIR FBs. From our experience, similar results cannot be achieved by GA even with design time several orders of magnitude longer. The latter is mainly due to high sensitivities of the poles. Since only the SOPOT coefficient optimization is performed, the computation time is much shorter, only 6 minutes in this case. The hardware structure is omitted here due to page length limitation.

VI. Conclusion

A novel algorithm for the design and hardware reduction of a class of multiplier-less two-channel PR FBs using SOPOT is presented. It minimizes a more realistic hardware cost, such as adder cells, subject to a prescribe output accuracy taking into account rounding and overflow effects. Further, by implementing the filters in the FBs using multiplier-block (MB), significant overall saving in hardware resources can be achieved. An effective random search algorithm is also proposed to solve the design problem, which is also applicable to PR IIR FB with highly nonlinear objective functions.

References

[1] W. Liu, S. C. Chan, and K. L. Ho, "Low-delay perfect reconstruction two-channel FIR/IIR filter banks and wavelet bases with SOPOT coefficients," in *Proc. IEEE ICASSP'2000, Istanbul, Turkey*, May 2000.

[2] G. Schuller and M. J. T. Smith, "A new algorithm for efficient low delay filter bank design," in *Proc. IEEE ICASSP'95*, vol. 2, pp. 1472 –1475, May. 1995.

[3] R. Gandhi and S. K. Mitra, "Design of two-channel low delay perfect reconstruction filter banks," in *Proc. 32nd International Asilomar Conference on Signals, Systems & Computers*, vol. 2, pp. 1655 –1659, Nov. 1998.

[4] J. S. Mao, S. C. Chan, W. Liu, and K. L. Ho, "Design and multiplier-less implementation of a class of two-channel PR FIR filterbanks and wavelets with low system delay," *IEEE Trans. SP.*, vol. 48, pp. 3379 – 3394, Dec. 2000.

[5] S. M. Phoong, C. W. Kim, P.P. Vaidyanathan and R. Ansari, "A new class of two-channel biothogonal filter banks and wavelet bases," *IEEE Trans. SP.*, vol. 43, pp. 649-664, Mar. 1995.

[6] S. Sriranganathan, D. R. Bull, and D. W. Redmill, "The design of low complexity two-channel lattice-structure perfect-reconstruction filter banks using genetic algorithms," in *Proc. IEEE ISCAS'97*, vol. 4, pp. 2393 –2396, Jun. 1997.

[7] W. Liu, S. C. Chan, and K. L. Ho, "Multiplier-less Low-delay FIR and IIR Wavelet Filter Banks with SOPOT Coefficients," in *Proc. IEEE ICASSP'2000, Istanbul, Turkey*, May 2000.

[8] I. Daubechies and W. Sweldens, "Factoring wavelet transform into lifting steps," *J. Fourier Anal. Appl.*, vol. 4, no. 3, pp. 247-269, 1998.

[9] A. G. Dempster and M. D. Macleod, "Use of minimum-adder multiplier blocks in FIR digital filters," *IEEE Trans. CAS. II*, vol. 42, pp. 569-577, Sept. 1995.

[10] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing*. 2nd Edition, NJ, Prentice Hall.

[11] S. C. Chan, J. S. Mao and K. L. Ho, "A new design method for two-channel perfect reconstruction IIR filter banks," *IEEE Signal Processing Letters*, vol. 7, pp. 221-223, Aug. 2000

	$H_0(z)$ –39.084dB, avg. SOPOT term = 2.38 β(z) (nonlinear FIR filter)			$H_1(z)$ –39.48dB, avg. SOPOT term = 2.13 α(z) (nonlinear FIR filter)		
n		PWL	Reg. (d_{n+1})		PWL	Reg. (d_{n+1})
0	$2^{-5}+2^{-7}$	<1|14>	<1|14>	-2^{-7}	<2|18>	<2|18>
1	$-2^{-3}+2^{-6}$ -2^{-8}	<1|15>	<1|15>	$2^{-5}-2^{-7}$	<2|18>	<2|18>
2	$2^{-1}+2^{-5}$ $+2^{-8}$	<1|15>	<1|15>	$-2^{-4}+2^{-6}$ -2^{-8}	<2|17>	<2|18>
3	$2^{-0}-2^{-2}$ -2^{-6}	<1|13>	<2|15>	$2^{-4}+2^{-6}$ $+2^{-8}$	<2|17>	<2|18>
4	$-2^{-2}-2^{-4}$ $+2^{-7}$	<1|14>	<2|15>	$-2^{-2}+2^{-4}$ $+2^{-7}$	<2|17>	<2|18>
5	$2^{-2}-2^{-5}$ -2^{-7}	<1|14>	<2|15>	$2^{-1}+2^{-3}$ -2^{-6}	<2|18>	<3|18>
6	$-2^{-3}-2^{-5}$	<1|12>	<3|15>	$2^{-1}+2^{-3}$ $+2^{-5}$	<2|18>	<4|18>
7	$2^{-3}-2^{-7}$	<1|14>	<3|15>	$-2^{-3}+2^{-5}$	<2|18>	<4|18>
8	$-2^{-3}+2^{-5}$ $+2^{-8}$	<1|15>	<3|15>	$2^{-3}-2^{-9}$	<2|17>	<4|18>
9	2^{-4}	<1|11>	<3|15>	$-2^{-4}-2^{-6}$	<2|17>	<4|18>
10	$-2^{-5}-2^{-6}$ $+2^{-8}$	<1|15>	<3|15>	$2^{-4}-2^{-6}$ $+2^{-9}$	<2|17>	<4|18>
11	$2^{-5}-2^{-7}$	<1|14>	<3|15>	$-2^{-5}+2^{-8}$	<2|17>	<4|18>
12	-2^{-6}	<1|13>	<3|15>	2^{-6}	<2|19>	<4|19>
13				-2^{-7}	<2|17>	<4|19>
14				2^{-8}	<2|17>	<4|19>

Design Results	
Overflow Possibility	Zero
Input Format	<1|7>
Output Accuracy (fractional side)	-96.6dB (accuracy > 16-bit)
Output Wordlength	23-bit
Number of Adders in the MB	
β(z)	α(z)
10	9
Esimated number of adder cells (with fixed wordlength of 24-bit using MBs)	1104
Estimated number of adder cells (with optimized wordlength using MBs)	825 (saved 27%)

Table 1. Filter banks results of example 5.1. PWL : product word length. Reg. : register.

	$H_0(z)$ -51.809dB, avg. term = 3.36	$H_1(z)$ -51.25dB, avg. term = 3.25	
n	β(z) (numerator direct form)	n	α(z) (Linear Phase FIR Filter)
0	$-2^{-6}+2^{-9}+2^{-11}$	0,15	$-2^{-8}+2^{-10}$
1	$2^{-6}-2^{-10}$	1,14	$2^{-7}+2^{-12}$
2	$-2^{-4}+2^{-8}+2^{-10}$	2,13	$-2^{-6}-2^{-9}+2^{-11}-2^{-15}$
3	$2^{-1}-2^{-3}-2^{-6}+2^{-8}-2^{-10}$	3,12	$2^{-5}+2^{-9}$
4	$2^{-1}-2^{-1}+2^{-3}-2^{-6}+2^{-9}$	4,11	$-2^{-4}+2^{-8}+2^{-10}-2^{-13}$
5	$2^{-1}-2^{-2}-2^{-3}+2^{-5}+2^{-10}$	5,10	$2^{-3}-2^{-5}+2^{-7}+2^{-12}$
6	$2^{-1}+2^{-3}-2^{-5}-2^{-8}$	6,9	$-2^{-2}+2^{-4}-2^{-7}-2^{-11}$
n	β(z) (denominator with lattice coefficients)	7,8	$2^{-1}+2^{-3}+2^{-7}-2^{-9}$
0	2^{-0}		
1	$2^{-0}-2^{-4}+2^{-8}+2^{-9}-2^{-11}$		
2	$2^{-0}-2^{-2}+2^{-5}+2^{-10}$		
3	$2^{-2}-2^{-8}-2^{-11}$		
4	$-2^{-5}-2^{-7}$		
5	$2^{-7}+2^{-8}-2^{-10}$		
6	$-2^{-9}-2^{-11}$		

Table 2. Filter banks results of example 5.2.

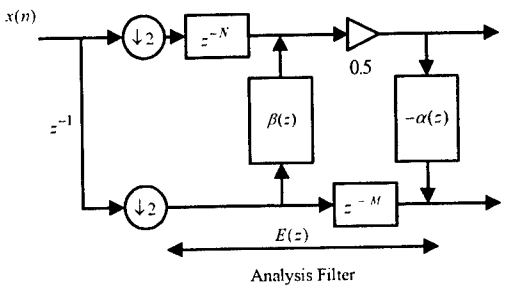


Fig.1. The biothogonal filter banks (analysis filter).

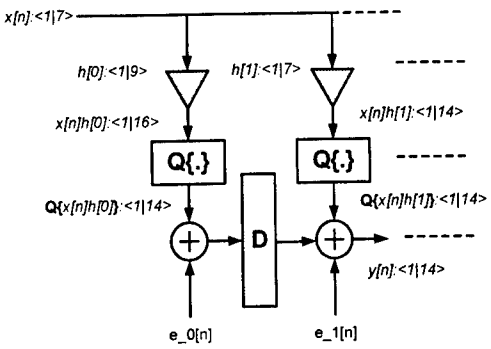


Fig. 2. Typical digital FIR filters with round-off noise model.

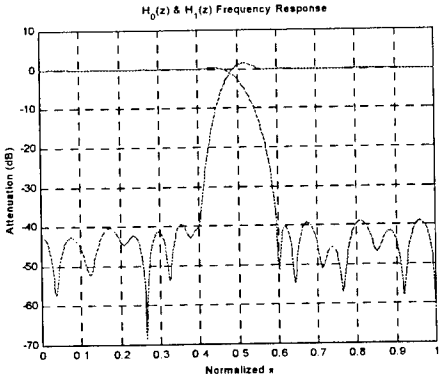


Fig. 3. Frequency responses of the two-channel FB in example 5.1.

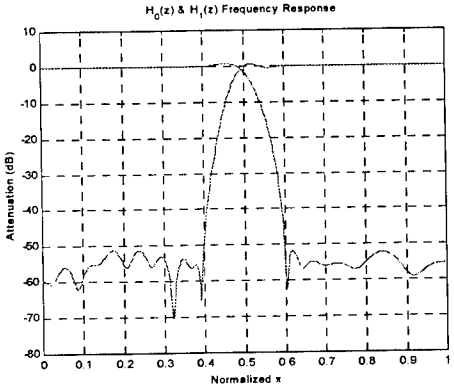


Fig. 4. Frequency responses of the two-channel FB in example 5.2.

OPTIMAL BIORTHOGONAL FILTER BANKS WITH MINIMIZATION OF QUANTIZATION NOISE AMPLIFICATION

Arunkumar M

Sasken Communication Technologies Limited, Bangalore, India

Anamitra Makur

ECE Dept., Indian Institute of Science, Bangalore, India

ABSTRACT

In this work we find out the optimal biorthogonal filter bank, in the ideal case, employing the proposed minimization of quantization noise amplification. In the ideal case it turns out to be a paraunitary filter bank which completely decorrelates the input signal, followed by scalar *DPCM* on each of the subbands. Its coding gain is shown to be equal to that of the ideal scalar *DPCM* coder acting on the original signal, irrespective of the number of channels. But the previously known optimal biorthogonal filter bank attains this coding gain only when the number of channels tends to infinity. The coding gain advantage of the new optimal filter bank structure, in the FIR case, is verified. The minimization of quantization noise amplification is also used to maximize the coding gain of a given biorthogonal filter bank. The coding gain improvements are verified for low bit-rates by measuring the reconstruction error introduced in coding an AR(1) source.

1. INTRODUCTION

The coding gain of a biorthogonal subband coder using the *additive uncorrelated white noise* model for the quantizers is given by the equation (1) for the optimum bit allocation case, where σ_x^2 denotes the input signal variance, $\sigma_{x_i}^2$'s denote the subband variances, $F_i(e^{j\omega})$ denotes the frequency response of the ith synthesis filter and M is the number of channels. Figure (1) shows the block diagram of a subband coder using the polyphase representation of the filter bank [9].

$$CG = \frac{\sigma_x^2}{\left(\prod_{k=0}^{M-1} \sigma_{x_k}^2 \int_{-\pi}^{\pi} |F_k(e^{j\omega})|^2 \frac{d\omega}{2\pi}\right)^{\frac{1}{M}}} \qquad (1)$$

The denominator of the coding gain expression contains terms equal to the energy of the synthesis filters, which are same as the squared norm of the synthesis filters. By norm of an FIR filter, what is meant is the magnitude of the vector whose components are the filter coefficients. These terms represent the quantization noise amplification taking place in the filter bank. When the white quantization noise vector passes through the synthesis polyphase matrix $\mathbf{R}(z)$, it becomes coloured and the variance is amplified. Then one would think that if the noise psd is appropriately modified, the amplification can be minimized and the coding gain can be maximized. Modifying the psd of the quantization noise vector can be done by passing the quantization noise through a colouring filter which is a multiple input multiple output filter. Let us denote it by $\mathbf{A}(z)$. This quantization noise filtering is carried out by adding an appropriate linear combination of the previously known quantization noise components to the present signal component being quantized.

In [8] the problem of optimal paraunitary filter banks is solved and a construction method is given for any arbitrary psd. The optimal biorthogonal filter bank has been shown to be the optimal paraunitary filter bank followed by a scalar filter in each channel which is the ideal half whitening filter for that channel [6]. The half-whitening arises because of the quantization noise amplification occurring in a biorthogonal filter bank.

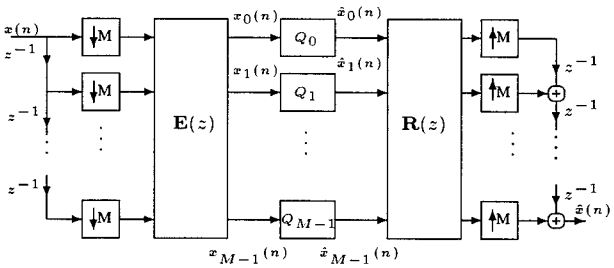


Fig. 1. Subband coder using a PR filter bank.

2. OPTIMAL BIORTHOGONAL FILTER BANK WITH MINIMIZATION OF QUANTIZATION NOISE AMPLIFICATION

Figure (2) shows how the quantization noise modification is done in a filter bank employing *gain plus additive noise* model for quantizers. D is a diagonal matrix having diagonal elements equal to $1/\alpha_i$'s. Note that the *additive uncorrelated white noise* quantizer model is a special case arising when α_i's are 1. Now we consider what would be the optimal biorthogonal filter banks if we do the modified quantization. The *additive uncorrelated white noise* model is assumed for the quantizers. Let the analysis polyphase matrix be $\mathbf{\Lambda}(z)\mathbf{E}(z)$, where $\mathbf{E}(z)$ is the polyphase matrix corresponding to the paraunitary filter bank which completely decorrelates the subband signals. Since the signals are completely decorrelated, it is assumed that there may not be any loss of generality in restricting $\mathbf{\Lambda}(z)$ to be diagonal. And we prove it in the subsequent section by showing that this structure indeed achieves the maximum coding gain possible, that is the gain of the ideal *DPCM* coder [4]. Now the problem is to find out the optimum colouring filter $\mathbf{A}(z)$. To avoid delay-free loops, the restriction on $\mathbf{A}(z)$ is that its zeroth order coefficient matrix must be lower triangular with 1's along the diagonal. Considering the fact that the paraunitary part of the synthesis filter does not introduce any noise am-

0-7803-7011-2/01/$10.00 ©2001 IEEE

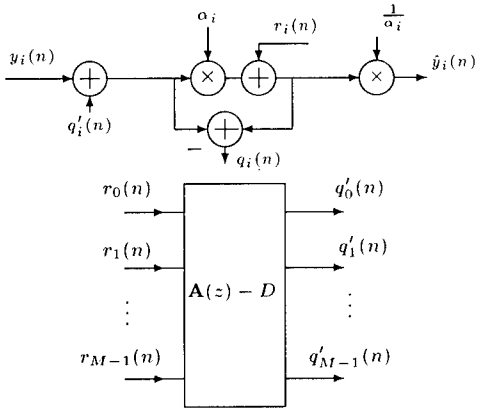


Fig. 2. Performing quantization noise filtering in a PR filter bank employing *gain plus additive noise model* of the quantizer.

plification, it is seen that we have to compensate only for the amplification by the diagonal matrix $\mathbf{\Lambda}^{-1}(z)$. It is easily seen that, to minimize the norms of the filters represented by the columns of $\mathbf{\Lambda}^{-1}(z)\mathbf{A}(z)$, $\mathbf{A}(z)$ also needs to be diagonal. Let $A_i(z)$ be the ith diagonal element of $\mathbf{A}(z)$. $A_i(z)$'s are scalar polynomials with 1 as the constant coefficient. The quantization noise variance introduced at the output by the channel i is given by,

$$\sigma_{q_i}^2 = c\,2^{-2b_i} \int_{-\pi}^{\pi} S_{x_i x_i}(e^{j\omega}) \left|\lambda_i(e^{j\omega})\right|^2 \frac{d\omega}{2\pi} \int_{-\pi}^{\pi} \left|A_i(e^{j\omega})\lambda_i^{-1}(e^{j\omega})\right|^2 \frac{d\omega}{2\pi} \quad (2)$$

where $x_i(n)$ is the ith subband signal at the output of $\mathbf{E}(z)$, c is the constant of proportionality of the quantizers and $\lambda_i(z)$ is the ith diagonal entry of $\mathbf{\Lambda}(z)$. Using Cauchy-Schwarz inequality, $\sigma_{q_i}^2$ is lower bounded by B where

$$B = c\,2^{-2b_i} \left(\int_{-\pi}^{\pi} \left|A_i(e^{j\omega})\right| S_{x_i x_i}^{\frac{1}{2}}(e^{j\omega}) \frac{d\omega}{2\pi} \right)^2 \quad (3)$$

Due to uncorrelated assumption, the total output noise variance is the average of $\sigma_{q_i}^2$ from each channel. Therefore, to minimize the total output noise, each $\sigma_{q_i}^2$ is to be minimized. Applying Cauchy-Schwarz inequality to B, we get

$$\left(\int_{-\pi}^{\pi} \left|A_i(e^{j\omega})\right| S_{x_i x_i}^{\frac{1}{2}}(e^{j\omega}) \frac{d\omega}{2\pi} \right)^2 \leq \int_{-\pi}^{\pi} S_{x_i x_i}(e^{j\omega}) \left|A_i(e^{j\omega})\right|^2 \frac{d\omega}{2\pi} \quad (4)$$

Thus, to minimize B, it is enough to minimize the right hand side of equation (4). $A_i(z)$ is restricted to have 1 as the constant coefficient. Such an $A_i(z)$ which minimizes the right hand side of equation (4) is known from the theory of linear prediction [4]. So for any order, $A_i(z)$ is the optimal linear predictor of the ith subband signal. In the infinite order case $\frac{1}{A_i(z)}$ becomes the spectral factor of the psd of the ith subband and

$$\left|A_i(e^{j\omega})\right| = \frac{G_i}{S_{x_i x_i}^{\frac{1}{2}}(e^{j\omega})} \quad (5)$$

where G_i^2 is the variance of the output of the ideal predictor when the input is $x_i(n)$. In this case it can be seen that, the inequality (4) is satisfied with equality. The condition for $\sigma_{q_i}^2$ achieving its lower bound B is,

$$\left|\lambda_i(e^{j\omega})\right|^2 = k_i \frac{\left|A_i(e^{j\omega})\right|}{S_{x_i x_i}^{\frac{1}{2}}(e^{j\omega})} \quad (6)$$

which simplifies using equation (5) to

$$\left|\lambda_i(e^{j\omega})\right| = \frac{(k_i G_i)^{\frac{1}{2}}}{S_{x_i x_i}^{\frac{1}{2}}(e^{j\omega})} \quad (7)$$

Thus, neglecting the scaling, the optimum $\lambda_i(z)$ as well as the optimum $A_i(z)$ is the whitening filter of the ith subband signal. A scaling on the $\lambda_i(z)$ will not affect the quantization noise as its effect is undone by the $\lambda_i^{-1}(z)$. The coding gain expression depends only on the magnitude of $\lambda_i(z)$'s and $A_i(z)$'s and their phase is irrelevant. So by choosing $\lambda_i(z)$ as equal to the optimum $A_i(z)$, it can be ensured that stable inverse filter exists, if $A_i(z)$ is designed to be optimal minimum phase linear predictor [4].

Using equations (5) and (7), it can be shown that the output noise variance, with optimum bit allocation, is

$$\sigma_{\mathbf{q}}^2 = c 2^{-2\bar{b}} \left(\prod_{i=0}^{M-1} \sigma_{y_i}^2 \right)^{\frac{1}{M}} \quad (8)$$

where $\bar{b}$ is the average bit-rate and $y_i(n)$ is the ith subband signal at the output of $\lambda_i(z)$. Therefore, the noise amplification has been completely eliminated. While in presence of noise amplification, the optimal biorthogonal filter bank only achieves half-whitening, the proposed optimal biorthogonal filter bank without noise amplification achieves full whitening. Therefore, we name it the full-whitening (FW) filter bank.

The FW structure can be used for finite order as well. Though no finite order paraunitary filter bank can completely decorrelate the subbands, given a paraunitary filter bank, its coding gain can be increased using this structure. Since $\lambda_i(z) = A_i(z)$, the filtering by $\lambda_i(z)$ and the modification of the quantization noise can be more easily done, using the scalar *DPCM* structure on each channel. It can be shown that coding gain improvement is ensured.

In the case of ideal optimum linear predictor it has been shown that

$$\sigma_{y_i}^2 = \gamma_{x_i}^2 \sigma_{x_i}^2 \quad (9)$$

where $\gamma_{x_i}^2$ is the spectral flatness measure for $S_{x_i x_i}(e^{j\omega})$ [4]. It has also been shown that

$$\gamma_{x_i}^2 = \frac{1}{\sigma_{x_i}^2} exp\left(\int_{-\pi}^{\pi} log_e \left[S_{x_i x_i}(e^{j\omega})\right] \frac{d\omega}{2\pi} \right) \quad (10)$$

In the following theorem we show that the M channel optimal FW filter bank, for any M, achieves the highest possible coding gain for any filter bank.

Theorem 1 *The optimal FW filter bank introduced above attains the gain of the ideal DPCM coder.*

Proof: The coding gain of the ideal *DPCM* coder is given by

$$CG_{DPCM} = \frac{\sigma_x^2}{exp\left(\int_{-\pi}^{\pi} log_e \left[S_{xx}(e^{j\omega})\right] \frac{d\omega}{2\pi} \right)} \quad (11)$$

Using equations (10) and (8) we get,

$$CG_{FW} = \frac{\sigma_x^2}{\left(\prod_{i=0}^{M-1} exp\left(\int_{-\pi}^{\pi} log_e\left[S_{x_i x_i}(e^{j\omega})\right]\frac{d\omega}{2\pi}\right)\right)^{\frac{1}{M}}} \tag{12}$$

So we have to show that both the denominators are equal. If D_{FW} is the denominator in equation (12), then

$$D_{FW} = exp\left(\frac{1}{M}\sum_{0\le i\le M-1}\int_{-\pi}^{\pi} log_e\left[\,S_{x_i x_i}(e^{j\omega})\right]\frac{d\omega}{2\pi}\right) \tag{13}$$

Since $\mathbf{E}(z)$ completely decorrelates the subbands, the corresponding subband filter responses are non-overlapping, because the psd of $x(n)$ is assumed to be non-zero everywhere. Being paraunitary, the filters have flat top. So it can be seen that the ith subband filter, say $H_i(z)$ selects a segment, not necessarily contiguous, of $MS_{xx}(e^{j\omega})$ of total width $\frac{2\pi}{M}$, where $S_{xx}(e^{j\omega})$ is the psd of the original scalar signal $x(n)$. The decimation operation causes its $M-1$ images, each shifted by an amount $\frac{2\pi}{M}$ to be added to it, then a stretching by a factor of M and a scaling by $\frac{1}{M}$. The Nyquist-M property of the filter ensures that no two of the images overlap. So it can be seen that each $S_{x_i x_i}(e^{j\omega})$ is obtained by taking the sections of the original psd selected by $H_i(z)$'s, stretching by a factor of M and reordering. Thus the area under any function of $S_{x_i x_i}(e^{j\omega})$ is the same as M times the area of the same function of $S_{xx}(e^{j\omega})$ over the support region of $H_i(z)$. The factor M comes because of the stretching. So,

$$D_{FW} = exp\left(\int_{-\pi}^{\pi} log_e\left[S_{xx}(e^{j\omega})\right]\frac{d\omega}{2\pi}\right) \tag{14}$$

This is because, the support of $H_i(z)$'s are disjoint and together they fill the entire $(-\pi,\pi)$. Thus, $CG_{DPCM} = CG_{FW}$ and the theorem is proved. []

3. OPTIMUM COLOURING FILTER FOR A GIVEN BIORTHOGONAL FILTER BANK

For a finite order biorthogonal filter bank the colouring filter is not diagonal. In this part the optimum implementable colouring filter such that the quantization noise amplification is minimized is found out, for a given filter bank.

The coding gain given by the equation (1) assumes the *additive uncorrelated white noise* model for the quantizers. The derivation of the optimum $\mathbf{A}(z)$ is done using the more general *gain plus additive noise* model for the quantizers, which is more suitable at low bit-rates [4]. To avoid delay-free loops, $\mathbf{A}(z)$ should be causal and its zeroth order coefficient matrix A_0 should be upper triangular or its permutations. In the absence of $\mathbf{A}(z)$, the output noise vector corrupting the blocked reconstructed signal vector will be $\mathbf{r}(n)$ multiplied by $\frac{1}{\alpha_i}'s$ and filtered through $\mathbf{R}(z)$, the synthesis polyphase matrix. With $\mathbf{A}(z)$, it will be $\mathbf{r}(n)$ filtered through $\mathbf{R}(z)\mathbf{A}(z)$, where the diagonal elements of the constant coefficient of $\mathbf{A}(z)$ are $(\frac{1}{\alpha_0}, \frac{1}{\alpha_1}, \ldots, \frac{1}{\alpha_{M-1}})$. It can be shown that [1] the optimum $\mathbf{A}(z)$ having the required properties, which minimizes the reconstruction error variance is obtained from the projectors which project each synthesis filter onto the space spanned by the filters above it and all the synthesis filters delayed by multiples of M samples. The resulting $\mathbf{A}(z)$ is FIR and the coding gain CG_1 obtained, relative to the case where no colouring filter is used is given by equation (15),

$$\frac{CG1}{CG} = \left(\frac{\prod_{i=0}^{M-1}\frac{1}{\alpha_i{}^2}\mathbf{f}_i{}^T\mathbf{f}_i}{\prod_{i=0}^{M-1}\mathbf{f}_i'{}^T\mathbf{f}_i'}\right)^{\frac{1}{M}} \tag{15}$$

where $\mathbf{f}_i$ is the vector representing the ith synthesis filter impulse response, and $\mathbf{f}_i'$ is the vector representing the residual after projection of $\frac{1}{\alpha_i{}^2}\mathbf{f}_i$. This coding gain improvement can be shown [1] to be independent of the order of performing the quantization among the subband signals.

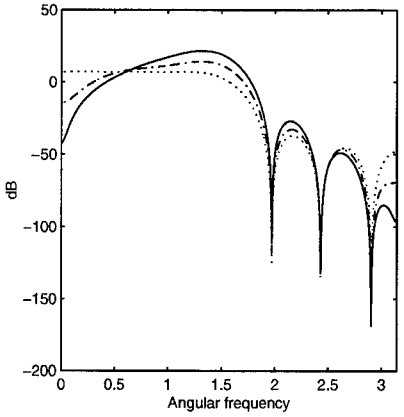


Fig. 3. Response of the low pass analysis filters used in the simulation. Bold curve is for the FW filter bank, the dash-dotted curve is for the half whitening filter bank and the dotted curve is for the paraunitary filter bank.

4. SIMULATION RESULTS

In the following table the coding gain of different schemes are compared for 2 channel case for an AR(1) source with correlation coefficient $\rho = 0.95$. While column 2 gives the idead coding gain, the coding gain for finite order case is given in column 3. The deatils are: paraunitary filter bank of order 11, half-whitening filter bank using the same paraunitary filters cascaded with half-whitening filters of order 3, and full-whitening filter bank same as half-whitening case with coloring filters $A_0(z)$ and $A_1(z)$ of order 3.

scheme	coding gain	
	ideal	FIR
DPCM	10.11 dB	
paraunitary [8]	5.96 dB	5.56 dB
half-whitening [6]	8.16 dB	7.75 dB
full-whitening	10.11 dB	9.78 dB

The figure (3) shows the low pass analysis filter responses of the half-whitening, FW, and the paraunitary filter bank used for the simulation. The dip of the low pass analysis filter bank near $\omega = 0$ clearly shows the whitening action of the filter bank. The full whitening solution has almost twice as much dip in dB as the half whitening solution.

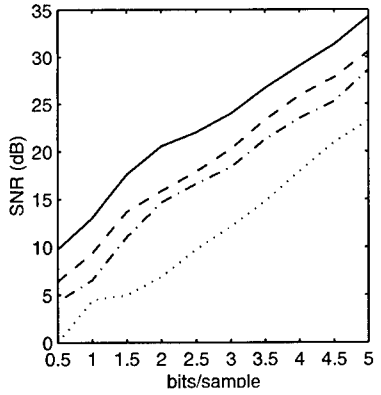


Fig. 4. SNR versus bit-rate curves for actual quantization of an AR-1 source having ρ=0.95. Bold curve is for the FW filter bank, dashed curve is for the half whitening filter bank, dash-dotted curve is for the paraunitary filter bank and dotted curve is for the scalar *PCM*.

The actual coding gain improvement may vary from the theoretical ones, especially at low bit rates because the quantizer model becomes less accurate and we used the high-resolution quantization assumption that the variance of signal plus the quantization noise filtered through the colouring filter is not much different from the variance of the original signal. Simulations of actual quantization carried out proves that good gain is obtained even at very low bit rates. The variation of SNR with bit-rate is given in figure (4). Coding gain improvement for a given filter bank is illustrated by simulations carried out on the same AR source. The table given below gives the theoretical coding gain improvement obtained for several standard two-channel biorthogonal filter banks, along with the measure of non-orthogonality introduced in Lightstone *et al* [3].

Filter Bank	Relative Gain (dB)	Measure of non-orthogonality
Egger-Li 4-12 [5]	0.5195	1.2874
Legall 3-5 [2]	0.1633	0.3887
Moulin 1-3 [7]	0.8805	2.9671
Moulin 5-11 [7]	1.2689	2.5587

The actual relative gain values agree with the theoretical gains even at low bit rates. We also consider the variation of the improvement in coding gain with the order of the colouring filter $\mathbf{A}(z)$. Though the optimum $\mathbf{A}(z)$ is FIR, a lower order $\mathbf{A}(z)$ may be preferred because of complexity considerations. It is seen that largest increase in the relative gain comes when the order is increased from 0 to 1. Moreover, good improvement is obtained even with first order $\mathbf{A}(z)$.

5. CONCLUSION

In this work we proposed the optimal biorthogonal filter bank with minimization of quantization noise amplification. In the ideal case this led to a filter bank which attained the maximum possible coding gain, namely, the coding gain of the ideal *DPCM* coder. On the contrary, the previously known optimal biorthogonal filter bank achieved this bound only when the number of channels tends to infinity. Given a biorthogonal bank, the coding gain was optimized by minimizing the quantization noise amplification by modifying the quantization noise. By doing simulation of the actual quantization on an AR(1) source using finite order filters, we showed that very good coding gain advantage is obtained for the optimal biorthogonal filter bank structure in the finite order case even at very low bit-rates.

6. REFERENCES

[1] Arunkumar M. , "Minimization of Quantization Noise Amplification in Biorthogonal Subband Coders," ME Thesis, Dept. of Electrical Engineering, Indian Institute of Science, Bangalore, [January 2001].

[2] D. LeGall and A. Tabatabai, "Subband coding of images using symmetric short kernel filters and arithmetic coding techniques," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1988, pp. 761–764.

[3] M. Lightstone, E. Majani, and S. K. Mitra, "Low bit-rate design considerations for wavelet-based coding," Technical Report CIPR 95-02, Department of ECE, University of California, Santa Barbara, CA, 1995.

[4] N. S. Jayant and P. Noll, "Digital Coding of Waveforms: Principles and Applications to Speech and Video," Englewood Cliffs, NJ: Prentice-Hall Signal Processing Series, 1984.

[5] O. Egger and W. Li, "Subband coding of images using asymmetrical filter banks," *IEEE Transactions on Image Processing*, Vol. 4, April 1995, pp. 478–485.

[6] Pierre Moulin, Mihai Anitescu and Kannan Ramchandran, "Theory of Rate-Distortion-Optimal, Constrained Filterbanks-Application to IIR and FIR Biorthogonal Designs," *IEEE Transactions on Signal Processing*, Vol. 48, No. 4, April 2000, pp. 1120-1131.

[7] P. Moulin, "A Multiscale Relaxation Algorithm for SNR maximization in nonorthogonal subband coding," *IEEE Transactions on Image Processing*, Vol. 4, September. 1995, pp. 1269–1281.

[8] P. P. Vaidyanathan, "Theory of Optimal Orthonormal Subband Coders," *IEEE Transactions on Signal Processing*, Vol. 46, No. 6, June 1999, pp. 1528–1544.

[9] P. P. Vaidyanathan, "Multirate Systems and Filter Banks," Prentice-Hall Inc. : Englewood Cliffs, 1993.

DOUBLY ORTHOGONAL WAVELET PACKETS FOR DS-CDMA COMMUNICATION

Hongbing Zhang and H. Howard Fan

GIRD Systems, Inc.
Cincinnati, OH 45220
and
Department of ECECS, University of Cincinnati
Cincinnati, OH 45221-0030, USA
Email: h.fan@uc.edu

Alan Lindsey

Air Force Research Lab/IFGC
Rome, NY 13441-4505, USA

ABSTRACT

Wavelet packets have been found a promising candidate for user signature waveforms in code division multiple access communication systems. The waveforms are usually chosen from an orthonormal basis so that there will be no interference between different users. However, timing errors may cause these signature waveforms to lose orthogonality to each other. In this paper, we describe a signal set which utilizes double orthogonality based on wavelet packets and binary Walsh codes. This double orthogonality produces auto- and cross-correlations that are much better than the conventional wavelet packet sets and are comparable to the pseudo random binary codes. The double orthogonality may also enable easy implementation in low complexity receiver design.

Key Words Multirate Processing, Wavelets, CDMA

1. INTRODUCTION

Wavelet Packets have properties that make them a good candidate for spreading codes in a Code Division Multiple Access (CDMA) system. By arbitrary prunning of a binary wavelet packet construction tree, an orthonormal and complete wavelet packet basis set can be constructed effectively. This provides perfect spreading codes that have zero cross-correlations, thereby eliminating multiple access interference in the absence of synchronization error. Wavelet packet based methods also have the advantage of naturally enabling multirate communication, and much work was devoted to user signature waveform designs using wavelet packets for improving cross-correlation properties over pseudo random codes [1]. Learned *et al* [2] not only used wavelet packets as user signature waveforms but also designed an optimal joint detector which achieved a lower complexity compared with conventional CDMA optimal receiver designs. Lindsey [3] found that by carefully choosing the wavelet packet basis, a wider selection of time-frequency tilings of wavelet packets could achieve a much better match of the transmission signal with the channel. Based on this observation, a method called wavelet packet modulation (WPM) was proposed and proved to have significant improvement of communication performance over Quadrature Amplitude Modulation (QAM) [9]. All of these works assume perfect timing of the signature waveforms.

The approaches in the literature using wavelet packets as spreading user waveforms differ from the conventional CDMA systems in the following ways. First, they are more like FDMA or TDMA systems. Each user mainly occupies relatively a small portion of the available bandwidth, or transmit mainly in a small portion of the symbol duration. Thus, compared with the conventional CDMA system, this kind of multiple access may suffer from narrow band or impulsive interference if there is no information about the interference available at hand. In addition, the narrow band waveforms tend to behave like periodic functions, i.e., their autocorrelations have more than one peak. This makes the synchronization task difficult in the receiver. Since the waveform set is generated from the nodes of the lowest level (beginning with a length 1 signal) and some higher levels of a binary wavelet packet tree, some of the waveforms are simply shifted versions of one another. Thus, this approach requires good, if not perfect, synchronization.

However, timing error cannot be ignored in some cases, such as in reverse link communication in a cellular system. Wong *et al* [4] investigated the timing error effect and derived an algorithm to optimize the wavelet packet design in a wavelet packet division multiplexing system. It has been shown that a lower error probability than commonly used wavelet packets can be achieved using the optimum design. Hetling *et al* [1] [5] investigated the possible interference from another user waveform for asynchronous communication channel. Sesay *et al* [6] also investigated the multiuser interference from the perspective of auto- and cross-correlation functions and error probability in a waveform division multiple access system. Other researchers also discussed the interference in such a spread spectrum system [7]-[8]. Most of these works propose alternative wavelet packet filter designs to reduce the multiuser interference.

In this paper, we describe and investigate a doubly orthogonal wavelet packet set, which utilizes wavelet packets and binary Walsh codes. This double orthogonality produces much better auto- and cross-correlations than the conventional wavelet packet sets and may also enable low complexity receiver design. Computer simulation results confirm the effectiveness of this new waveform design.

2. DOUBLY ORTHOGONAL WAVELET PACKET SET

We propose a set of user waveforms as a candidate of spreading codes for a CDMA system. The code waveforms should have good autocorrelation and cross-correlation properties, and the autocorrelation functions should have only one narrow peak. This ensures the initial acquisition and the following tracking of synchronization. The cross-correlations between any pair of waveforms in the set should be small enough so that the mul-

This work is supported in part by AFRL/IFKD under Contract F30602-00-C-0086 through GIRD Systems, Inc.

0-7803-7011-2/01/$10.00 ©2001 IEEE

tiple access interference due to other users can be maintained at minimum. The proposed doubly orthogonal wavelet packet waveforms have the desired correlation properties.

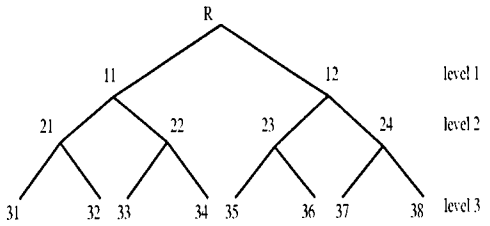


Figure 1: Binary wavelet packet tree structure

Wavelet packet waveforms are generated by up-sampling and filtering impulses from certain nodes of the binary wavelet packet tree. Figure 1 shows the binary wavelet packet tree structure. To generate wavelet packet waveforms, we begin from a certain node and go up to the root of the tree by up-sampling and filtering an impulse signal. The level and position of the node determine how many times of the up-sampling and filtering process are taken and types of the filter, usually a low-pass or high-pass quadrature mirror filter. From eight level-3 nodes of the tree, we can generate eight wavelet packet waveforms. The length of the generated waveforms is determined by the length of the input impulses. The shortest waveform which can be generated from a level-3 node is length 8, if the input impulse has a length of 1. However, the filtering process will make the generated waveforms to fold several times. Higher filter level results in more folding. As a consequence, some of the waveforms from different nodes become just shifted versions of each other, and may not be used in an asynchronous system.

Now consider the proposed doubly orthogonal wavelet packet waveforms. For simplicity, we consider a 64 user CDMA system. Instead of generating all 64 waveforms from 64 level-6 nodes, we divide the users into 8 groups, named A through H, each containing eight users. We also chop each symbol into 8 chips in the time domain. Each group of 8 users associates with one length 8 Walsh code as the chip code for each symbol interval. Due to the orthogonality of Walsh code, these 8 groups of users have waveforms orthogonal to each other. The 8 users in one group are assigned orthogonal waveforms based on wavelet packets. If we generate the wavelet packet waveforms from all the eight level-3 nodes of a wavelet packet tree, we can generate 8 orthogonal wavelet packet waveforms. We name these wavelet packet waveforms from 1 to 8. The proposed doubly orthogonal waveforms are thus generated by mapping the 8 wavelet packet waveforms to each of the 8 chips of the Walsh code. Eight different ordering possibilities of the mapping enable us to fit 8 users in one Walsh code. Thus totally we can generate 64 different user waveforms. The mapping is simply done by multiplying the wavelet packet waveforms with the Walsh code chip value, i.e., 1 or -1.

Figure 2 is an example of the mapping matrix which defines the 8 different orders of mapping wavelet packet waveforms to Walsh code chips of all 1's, i.e., user group A. The numbers shown in the 8 × 8 matrix is the wavelet packet waveform numbers (1 to 8). Each column in the mapping matrix corresponds to one Walsh code chip, or one time slot. Each row in the matrix corresponds to one possible order of mapping 8 wavelet packet waveforms to 8 chips. This defines a unique user waveform. Eight rows define eight user waveforms in one user group. For example, the third row specifies that the waveform for user 3 in the group is generated by concatenating wavelet packet waveforms 3, 7, 4, 5, 8, 2, 6, and 1. This particular order makes the waveform unique. Other users have different mapping orders so that the generated waveforms differ from that of user 3. Note that in each time slot the 8 users have been mapped with 8 different wavelet packet waveforms. This ensures that the 8 user waveforms in the same user group are orthogonal to each other. Since all of the rows define mappings of all the 8 wavelet packet waveforms to the Walsh chips, all user waveforms will occupy the entire frequency bandwidth as well as all the time slots.

chipslots	1	2	3	4	5	6	7	8
*user*1	1	2	3	4	5	6	7	8
*user*2	2	1	6	8	4	3	5	7
*user*3	3	7	4	5	8	2	6	1
*user*4	4	8	7	2	6	1	3	5
*user*5	5	4	2	7	3	8	1	6
*user*6	6	5	8	1	2	7	4	3
*user*7	7	6	5	3	1	4	8	2
*user*8	8	3	1	6	7	5	2	4

Figure 2: Chip wavelet packet numbers in Group A of 8 users. The Walsh code is all 1's.

Since different users occupy distinct wavelet packet waveforms in any of the time slots, it is desirable to represent the signal using permutation notations. Using the above example, the eight rows or eight columns in the mapping matrix are different permutations of $X_8 = \{1,2,3,4,5,6,7,8\}$. Note that absence of repetition of the 8 wavelet packet waveforms in each column is important to ensure orthogonality, whereas such absence in each row is not essential, although desirable. The constructed signal set of group A users using the above example is

$$s_{A,k}(n) = \sum_{i=1}^{8} P_{3,\sigma_i(k)}(n - 8(i-1)) \quad k = 1, ..., 8 \qquad (1)$$

where $\sigma_i(k)(i = 1,2,...,8)$ are eight permutations of X_8 for the k*th* user, and $P_{3,l}(l = 1,2,...,8)$ are eight wavelet packet waveforms each with length 8.

For user Group B, the Walsh code is, e.g., 1,1,1,1,-1,-1,-1,-1. Then the mapping matrix would be the same as Figure 2, except that in the last four columns all wavelet packet waveforms need to be multiplied by -1. Other user groups follow Figure 2 and the corresponding Walsh codes in a similar way. Thus the constructed signal set of the k*th* user in the j*th* group is

$$s_{j,k}(n) = \sum_{i=1}^{8} W_j(i) P_{3,\sigma_i(k)}(n - 8(i-1))$$
$$j = A, B, ..., H, k = 1, 2, ..., 8 \qquad (2)$$

where $W_j(i), (i = 1,2,...,8)$ is the j*th* length 8 Walsh code. Since the Walsh codes form an orthogonal basis, user waveforms with same wavelet packet mapping orders but in different groups are also orthogonal to each other. For example, user 1 in group A and B have same wavelet packet mapping order, but because the wavelet packet waveforms are multiplied by two orthogonal Walsh codes, these two waveforms are orthogonal to each other. It is easy to see that user waveforms in different groups and with different mapping orders are also orthogonal to each other. Thus, the 64 user waveforms form an orthogonal set.

This algorithm can be generalized to achieve a tradeoff between the autocorrelation and cross-correlation properties of the

waveforms. If the desired length of the waveforms is $N = 2^{j+k}$, we can divide the users into $M = 2^j$ groups, and generate length of $L = 2^k$ wavelet packet waveforms. An orthogonal waveform set can be formed by combining the wavelet packet waveforms according to the above algorithm. The number of waveforms in the set is $N = M \times L$. A tradeoff between the autocorrelation and cross-correlation properties can be achieved with different combinations of M and L. In general, a smaller M and a larger L results in better cross-correlation but poorer autocorrelation, and vice versa.

3. CORRELATION PROPERTIES OF THE WAVEFORMS

Now, we investigate the autocorrelation and cross-correlation properties of the waveforms proposed in the last section. As an example, we choose the Daubechies 4 wavelet as the mother wavelet from which a wavelet packet tree is constructed. The reason is that the order of the filter is lower than other wavelets because Daubechies wavelets have minimum size of support. The correlation functions we are to investigate are discrete periodic auto- and cross-correlation functions defined as

$$R_i(k) = \frac{1}{N} \sum_{n=0}^{N-1} s_i(n) s_i(n+k) \tag{3}$$

and

$$C_{ij}(k) = \frac{1}{N} \sum_{n=0}^{N-1} s_i(n) s_j(n+k) \tag{4}$$

where N is the waveform length. We have also investigated the averaged cross-correlation functions defined as

$$\bar{C}_i(k) = \frac{1}{N-1} \sum_{j=1, j \neq i}^{N} C_{ij}(k) \tag{5}$$

Figure 3 gives an example of the autocorrelation function of a length 64 doubly orthogonal wavelet packet waveform. We can see that this autocorrelation has a single narrow peak. This is similar to the autocorrelation function of the length 63 Gold code given in Figure 4. Figure 5 gives an example of the autocorrelation function of a length 64 wavelet packet waveform, which is much worse.

Figure 6 gives an example of the cross-correlation function between a pair of doubly orthogonal wavelet packet waveforms. Note that the cross-correlation is zero when the relative shift of the two waveforms is zero. Compared with the cross-correlation function of Gold codes given in Figure 7, we can find that the cross-correlation of the proposed waveforms is in the same level with that of Gold codes, but not as regularly distributed. For the conventional wavelet packets, the cross-correlation is much better than Gold codes on the average [1]. However, in the wavelet packet set many waveforms are the shifted versions of one another, which gives poor cross-correlation. Figure 8 gives such an example of the cross-correlation function between a pair of length 64 wavelet packet waveforms. Since these two waveforms are shifted versions of each other, the cross-correlation not only has large values but also has value 1 for some relative shifts. This is not the case for the doubly orthogonal wavelet packet waveforms.

Figure 9 gives an example of the averaged cross-correlation function of one doubly orthogonal wavelet packet waveform. Compared with the averaged cross-correlation function of Gold code given in Figure 10, we can find the proposed waveforms are at a similar level. Figure 11 gives the averaged cross-correlation function of a length 64 wavelet packet waveform. We can see that the cross-correlation of the doubly orthogonal wavelet packet waveform is higher than that of a conventional wavelet packet waveform. However, the proposed waveforms do not have any large cross-correlation values as the conventional wavelet packet waveforms in Figure 8.

4. REFERENCES

[1] Kenneth Hetling, Gary Saulnier and Pankaj Das, "Spreading codes for wireless spread spectrum communications", *Proceedings of ICC'96*, vol. 1, pp 68-72.

[2] Rachel E. Learned, Alan S. Willsky, and Don M. Boroson, "Low complexity optimal joint detection for oversaturated multiple access communications", *IEEE Trans. on Signal Processing*, vol. 45, no. 1, pp113-123, January 1997.

[3] Alan R. Lindsey, "Wavelet packet modulation for orthogonally multiplexed communications", *IEEE Trans. on Signal Processing*, vol. 45, no. 5, pp1336-1339, May 1997.

[4] Kon Max Wong, Jiangfeng Wu, Tim N. Davidson, and Qu Jin, "Wavelet packet division multiplexing and wavelet packet design under timing error effects", *IEEE trans. on Signal Processing*, vol. 45, no. 12, pp 2877-2890, Dec. 1997.

[5] Kenneth Hetling et al, "A PR-QMF (wavelet) based spread spectrum communications system", *Proceedings of MILCOM'94*, vol. 3, pp760-764.

[6] A. B. Sesay et al, "Waveform division multiple-access", *IEE Proceddings-I*, vol. 140, no. 3, pp176-184, June 1993.

[7] Fred Daneshgaran and Marina Mondin, "Wavelets and scaling functions as envelope waveforms for modulation", *Proceedings of IEEE-SPIE*, pp 504-507, 1994.

[8] D. Mihai Ionescu and Mark A. Wickert, "On the performance of a CDMA system with user signatures based on packet wavelets in multipath channels", *Proceedings of VTC'97*, vol. 1, pp 392-396.

[9] William Wayne Jones, "A unified approach to orthogonally multiplexed communication using wavelet bases and digital filter banks", Ph.D Dissertation, Ohio University, August 1994.

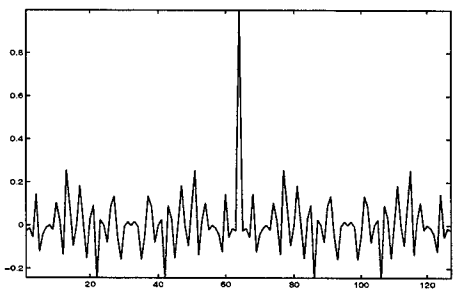


Figure 3: Autocorrelation of a DOWP waveform

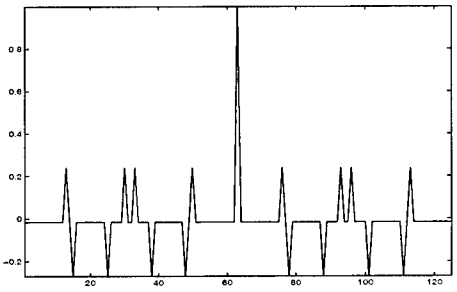


Figure 4: Autocorrelation of a length 63 Gold code

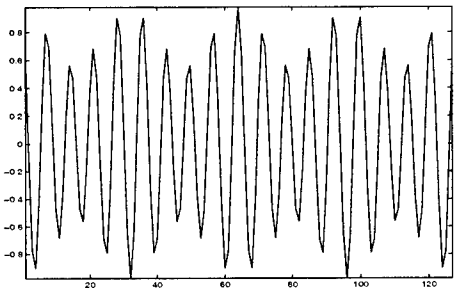


Figure 5: Autocorrelation of a WP waveform

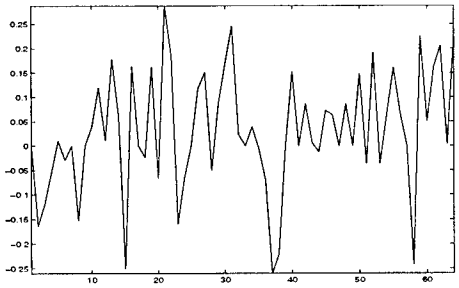


Figure 6: Cross-correlation between a pair of DOWP waveforms

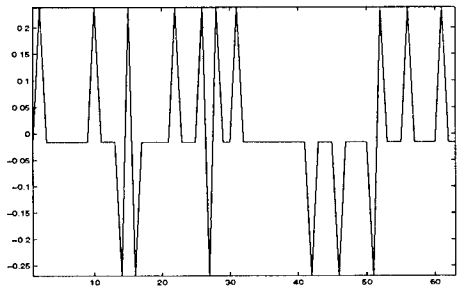


Figure 7: Cross-correlation between a pair of Gold codes

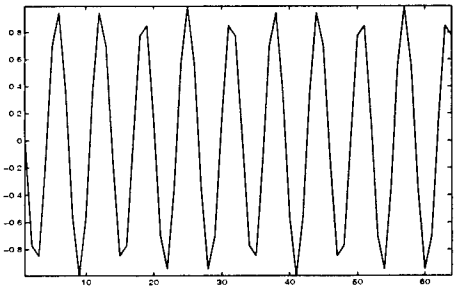


Figure 8: Cross-correlation between a pair of WP waveforms

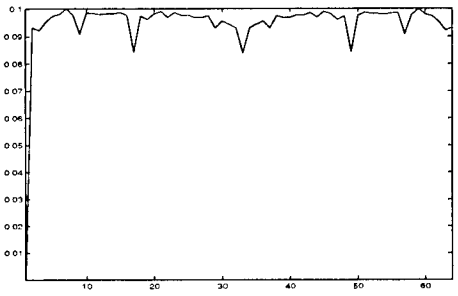


Figure 9: Averaged cross-correlation of a DOWP waveform

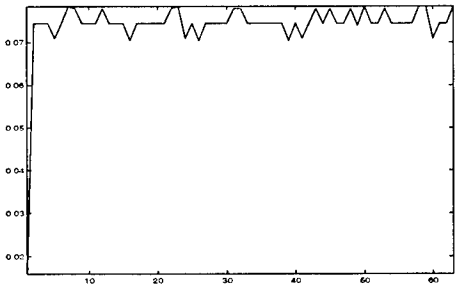


Figure 10: Averaged cross-correlation of a Gold code

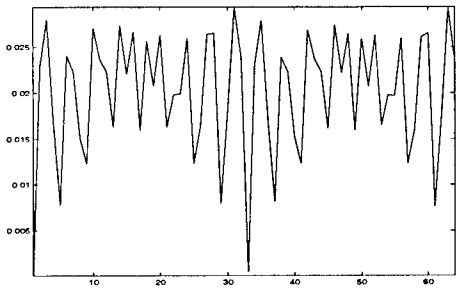


Figure 11: Averaged cross-correlation of a WP waveform

AUTHOR'S INDEX

Abeysekera, Saman S. 417
Abeysekera, Saman S. 556
Abrahamsson, R. 150
Abramovich, Yuri I. 229
Ahmed, Osama A. 317
Ahn, Byungha 142
Ali, Hassan 401
Alieva, Tatiana 321
Alieva, Tatiana 325
Alvarez, A. 500
Amblard, Pierre-Olivier 66
Amin, Moeness G. 313
Amin, Moeness G. 361
Amin, Moeness G. 544
Amin, Moeness G. 563
An, Senjian 265
Andersson, Tomas 409
Andrieu, C. 38
Andrieu, Christophe 14
Andrieu, Christophe 309
Arunkumar, M 603
Athley, Fredrik 516
Attallah, S. 206

Barkat, B. 206
Barnett, Adrian G. 437
Bastiaans, Martin J. 321
Bastiaans, Martin J. 325
Belhouari, Sofiane Brahim 433
Belouchrani, Adel 333
Belouchrani, Adel 444
Belouchrani, Adel 448
Bergman, N. 34
Bezerianos, A. 257
Bezerianos, A. 261
Bi, Guoan 210
Bilgutay, Nihat 82
Blázquez, Raúl 138
Bohlin, Patrik 440
Böhme, Johann F. 536
Böhme, Johann F. 540
Boonstra, A.J. 365
Borgnat, Pierre 66
Bradaric, Ivan 277
Bradley, M. 150
Brcich, Ramon F. 26
Brown, Christopher L. 166
Brown, Christopher L. 58
Bugallo, Mónica F. 198
Burgos, Mateo 138

Cantoni, Antonio 357
Castedo, Luis 114
Castedo, Luis 198
Chakraborty, Mrityunjoy 373
Chambers, Jonathon A. 94
Champagne, Benoit 504
Chan, S. C. 583
Chan, S. C. 595
Chan, S. C. 599
Chang, Xianwen 293
Chavanne, R. 512
Chen, Chii-Horng 118
Chen, Chii-Horng 293
Cheriet, Mohamed 333
Chi, Chong-Yung 118
Chi, Chong-Yung 293
Chia, Nicholas K. K. 30
Choi, Daebum 142
Choi, Seungjin 444
Chong, Lucy L. 385
Chung, Pei Jung 540
Cichocki, A. 237
Cichocki, Andrzej 273
Cichocki, Andrzej 444
Cirillo, Luke A. 166
Collins, Leslie 162
Collins, Michael J. 241
Costa, Antonio H. 337
Courville, Marc de 389

Dasgupta, Soura 269
Davy, Manuel 305
Davy, Manuel 309
Debbah, Merouane 389

Deng, Tian-Bo 353
Descombes, X. 22
Dhibi, Youssef 70
Diamantaras, Konstantinos I. 277
Djuric, Petar M. 42
Djuric, Petar M. 429
Djuric, Petar M. 559
Doucet, A. 14
Doucet, Arnaud 305
Doucet, Arnaud 309
Doucet, Arnaud 341
Drouiche, K. 500
Duong, Sinh 154

Elliott, Robert J. 575
Enescu, Mihai 289

Fabrizio, G. A. 134
Faizakov, Avi 213
Fan, H. Howard 607
Faulkner, M. 106
Fitzgerald, W. J. 38
Flandrin, Patrick 66
Fong, William 18
Forssell, U. 34
Friedmann, Jonathan 221

Gao, Ping 162
Garcia, Francisco M. 178
Ge, Lijia 504
Georgiev, Pando 273
Gershman, Alex B. 217
Gershman, Alex B. 536
Geva, A. B. 329
Gharieb, R. R. 237
Ghirmai, Tadesse 42
Giannakis, Georgios B. 13
Giannakis, Georgios B. 381
Giannakis, Georgios B. 393
Godsill, Simon J. 245
Godsill, Simon J. 496
Godsill, Simon 18
Goldberg, Jason 221
Gomez, P. 500
Grajal, Jesús 138
Grajal, Jesús 174
Green, Matthew 345
Gretton, Arthur 305
Gretton, Arthur 341
Guangiming, Shi 587
Gunnarsson, F. 34
Gustaffson, F. 34

Haan, Nicholas M. 245
Habersat, J. 150
Hachem, Walid 389
Hall, Peter 1
Hamza, A. Ben 130
Hamza, A. Ben 90
Händel, Peter 377
Händel, Peter 409
Händel, Peter 425
Hanzo, Lajos 567
Hanzo, Lajos 591
He, Lin 563
He, Yun 130
Herbrich, Ralf 341
Hernández, Miguel Ángel Lagunas 528
Herrmann, Frank 456
Hill, Simon I. 488
Hlawatsch, Franz 571
Ho, K. L. 595
Ho, K. L. 599
Hopgood, James R. 492
Hua, Y. 265
Hua, Yingbo 401
Huang, Dawei 421
Huang, L. 106
Huang, Yufei 42

Iserte, Antonio Pascual 528
Iskander, D. R. 241
Iskander, D. Robert 182
Ito, Mabo R. 154
Izzetoglu, Meltem 82

Jakobsen, Kaj B. 146
Jansson, J. 34

Jiménez, M. E. Dominguez 579
Jung, K. 468

Kadambe, S. 126
Kadtke, James B. 186
Kaiser, Thomas 70
Kannan, B. 86
Karlsen, Brian 146
Karlsson, R. 34
Kasilingam, Dayalan 337
Katkovnik, Vladimir 532
Kawamura, Toshiya 464
Kempen, L. van 158
Khan, Mohammad Asmat Ullah 472
Kim, J. 468
Kim, K. I. 468
Kim, Yonghoon 532
Kingsbury, N.G. 480
Ko, Hanseok 142
Koivunen, V. 281
Koivunen, Visa 289
Kotecha, Jayesh H. 429
Krim, Hamid 130
Krim, Hamid 90
Krolik, Jeffrey L. 297

Lamba, T. S. 373
Lane-Glover, A. T. 102
Larsen, Jan 146
Larsson, E. G. 150
Larsson, Erik G. 393
Larzabal, Pascal 301
Leduc, Jean-Pierre 484
Lee, Dominic S. 30
Lee, Seungsin 62
Lee, Ta-Sung 110
Leitao, Jose M. N. 50
Leitão, José M.N. 202
Lennartsson, Ron K. 186
Leshem, Amir 190
Leyman, A. Rahim 210
Leyman, A. Rahim 233
Li, J. 150
Li, J. 476
Li, Jian 393
Liavas, Athanasios P. 552
Licheng, Jiao 587
Lieshout, M. N. M. van 22
Lin, Gau-Joe 110
Lin, Zhiping 194
Lindsey, Alan R. 361
Lindsey, Alan 607
Liu, Guoqing 393
López, Gustavo 138
López-Risueño, Gustavo 174
Lopez-Valcarce, Roberto 269
Loubaton, Philippe 389
Lourtie, Isabel M. G. 178
Lui, Wei 591
Lundberg, Magnus 170
Lundin, Henrik 377
Luo, Zhi-Quan 217

Maksymonko, G. 150
Makur, Anamitran 603
Malkemes, Robert 563
Manton, Jonathan H. 225
Manton, Jonathan H. 401
Manton, Jonathan H. 405
Mao, Jian 504
Martinez, R. 500
Matz, Gerald 571
McFee, John E. 154
Médynski, D. 512
Mengersen, Kerrie L. 46
Meraim, K. Abed 512
Meraim, Karim Abed 285
Meraim, Karim Abed 448
Messer, Hagit 221
Messer, Hagit 78
Míguez, Joaquín 114
Míguez, Joaquín 198
Milstein, Laurence B. 385
Mohamad, Hafizal 567
Morelande, Mark R. 182
Morelande, Mark R. 241
Moura, Jose M. F. 2
Mourad, N. 237
Mu, Weifeng 313
Muquet, Bertrand 381

Nagarajan, Krishnamurthy 74
Nandi, Asoke K. 456
Neira, Ana I. Pérez 528
Nieto, V. 500
Nordebo, Sven 357
Nordholm, Sven 357
Nordlund, P-J 34
Nur, Darfiana 46

O'Droma, Mairtin 504
Oja, H. 281
Onaral, Banu 82
Ong, Patrick K .S. 556
Ong, S. H. 476
Ozdemir, Ozgur 98

Palaniappan, R. 249
Paul, J. S. 257
Paul, J. S. 261
Pelin, Per 440
Pentek, Aron 186
Perreau, Sylvie 122
Pervin, Suraiya 373
Pesavento, Marius 217
Pesavento, Marius 536
Petropulu, Athina P. 277
Petropulu, Athina P. 54
Pettitt, Tony 421
Policker, S. 329
Prelcic, N. Gonzalez 579
Pun, Carson K. S. 595
Pun, Carson K. S. 599
Punskaya, E. 38

Raich, Raviv 221
Ranheim, Anders 440
Rao, Raghuveer 62
Raveendran, P. 249
Rayner, Peter J. W. 305
Rayner, Peter J. W. 341
Rayner, Peter J. W. 488
Rayner, Peter J. W. 492
Reed Jr., Charles 563
Reeves, T .H. 480
Rodellar, V. 500
Rupp, Markus 567
Russell, Kevin L. 154

Safavi, Anahid 285
Sahli, H. 158
Salah, Hicham Bousbia 448
Sando, Simon 421
Sanei, S. 233
Sanei, S. 476
Sano, Akira 524
Sanz, José M. 138
Saruwatari, Hiroshi 464
Sattar, F. 452
Saulnier, Gary J. 397
Schafhuber, Dieter 571
Schölkopf, Bernhard 341
See, Chong Meng Samson 520
Seyedi, Alireza 397
Shereshevski, Yoav 78
Sherman, D. 257
Shikano, Kiyohiro 464
Silva, Francisco A. S. 50
Sirbu, Marius 289
Siyal, M. Y. 452
Skoglund, Mikael 377
Skoglund, Mikael 409
Sørensen, Helge B. D. 146
Sousa, Fernando M.G. 202
Spencer, Nicholas K. 229
Stankovic, Ljubiša 321
Stankovic, Ljubiša 325
Stoica, R. 22
Svantesson, Thomas 508
Svensson, Lennart 170

Tabrikian, Joseph 213
Tan, Chan-Choo 110
Tantum, Stacy 162
Teal, Paul D. 548
Teunissen, P. J. G 4
Thakor, N. V. 3
Thakor, N. V. 257
Thakor, N. V. 261
Thakor, Nitish V. 253

Tong, S. 257
Tong, S. 261
Torlak, Murat 98
Tournier, Marc Chenu 301
Turley, M. D. 134

Vardarajan, Vijay 297
Vaughan, Rodney G. 548
Vázquez, Gregori 413
Veen, Alle-Jan van der 190
Veen, Alle-Jan van der 365
Veen, Alle-Jan van der 460
Vesin, Jean-Marc 433
Villares, Javier 413
Visuri, S. 281
Völcker, Björn 425

Wang, Junfeng 337
Wang, Wenyi 369
Wee, L. C. 452
Weiss, Stephan 567
Weiss, Stephan 591
Wennström, Mattias 508
White, Langford B. 349
White, Langford B. 575
Wolfe, Patrick J. 488
Wolfe, Patrick J. 496
Wolff, Rodney C. 437
Wolff, Rodney C. 46
Wong, Kon Max 536
Wu, Y. C. 595

Xie, X. M. 583
Xin, Jingmin 524

Yan, Xiang 337
Yang, Chunhua 210
Yang, Kehu 544
Yang, Xueshi 54
Yazici, Birsen 82
Yen, L. C. 452
Yeredor, Arie 78
Yuk, T. I. 583
Yuvapoositanon, Peerapol 94

Zang, Zhuquan 357
Zarzoso, Vicente 456
Zerubia, J. 22
Zhang, Hongbing 607
Zhang, Hongxuan 253
Zhang, Jianqiu 559
Zhang, Yimin 313
Zhang, Yimin 544
Zhao, Liang 361
Zheng, F. C. 106
Zheng, Yuanjin 194
Zhou, G. Tong 74
Zhou, Shengli 381
Zhu, Y. 261
Zhu, Yi-sheng 253
Zoubir, A. M. 102
Zoubir, Abdelhak M. 166
Zoubir, Abdelhak M. 26