UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1995

SUMMER RESEARCH EXTENSION PROGRAM FINAL REPORTS

VOLUME 4A

WRIGHT LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES
5800 Uplander Way
Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Major David Hart

Program Manager, RDL
Scott Licoscos

Program Administrator, RDL
Gwendolyn Smith

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Bolling Air Force Base
Washington, D.C.

May 1996

20010319 035

AQM01-06-1068

REPORT DOCUMENTATION PAGE

Form Approved

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existin the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for r Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduct

AFRL-SR-BL-TR-00-

0699

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE
	May, 1996	

4. TITLE AND SUBTITLE
1995 Summer Research Program (SRP), Summer Research Extension Program (SREP), Final Report, Volume 4A, Wright Laboratory

5. FUNDING NUMBERS
F49620-93-C-0063

6. AUTHOR(S)
Gary Moore

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Research & Development Laboratories (RDL)
5800 Uplander Way
Culver City, CA 90230-6608

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research (AFOSR)
801 N. Randolph St.
Arlington, VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The United States Air Force Summer Research Program (SRP) is designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by the faculty members, graduate students, and high school students being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory (AFRL) Technical Directorates and Air Force Air Logistics Centers (ALC). AFOSR also offers its research associates (faculty only) an opportunity, under the Summer Research Extension Program (SREP), to continue their AFOSR-sponsored research at their home institutions through the award of research grants. This volume consists of the SREP program background, management information, statistics, a listing of the participants, and the technical report for each participant of the SREP working at the AF Wright Laboratory.

14. SUBJECT TERMS
Air Force Research, Air Force, Engineering, Laboratories, Reports, Summer, Universities, Faculty, Graduate Student, High School Student

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract	**PR** - Project
G - Grant	**TA** - Task
PE - Program Element	**WU** - Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with....; Trans. of....; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD -	See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE -	See authorities.
NASA -	See Handbook NHB 2200.2.
NTIS -	Leave blank.

Block 12b. Distribution Code.

DOD -	Leave blank.
DOE -	Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
	Leave blank.
NASA -	Leave blank.
NTIS -	

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self- explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

PREFACE

This volume is part of a five-volume set that summarizes the research of participants in the 1995 AFOSR Summer Research Extension Program (SREP). The current volume, Volume 1 of 5, presents the final reports of SREP participants at Armstrong Laboratory, Phillips Laboratory, Rome Laboratory, Wright Laboratory, Arnold Engineering Development Center, Frank J. Seiler Research Laboratory, and Wilford Hall Medical Center.

Reports presented in this volume are arranged alphabetically by author and are numbered consecutively -- e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3, with each series of reports preceded by a management summary. Reports in the five-volume set are organized as follows:

VOLUME	TITLE
1A	Armstrong Laboratory (part one)
1B	Armstrong Laboratory (part two)
2	Phillips Laboratory
3	Rome Laboratory
4A	Wright Laboratory (part one)
4B	Wright Laboratory (part two)
5	Arnold Engineering Development Center Frank J. Seiler Research Laboratory Wilford Hall Medical Center

1995 SREP FINAL REPORTS

Armstrong Laboratory

VOLUME 1

Report #	Report Title Author's University	Report Author
1	Determination of the Redox Capacity of Soil Sediment and Prediction of Pollutant University of Georgia, Athens, GA	Dr. James Anderson Analytical Chemistry AL/EQ
2	Finite Element Modeling of the Human Neck and Its Validation for the ATB Villanova University, Villanova, PA	Dr. Hashem Ashrafiuon Mechanical Engineering AL/CF
3	An Examination of the Validity of the Experimental Air Force ASVAB Composites Tulane University, New Orleans, LA	Dr. Michael Burke Psychology AL/HR
4	Fuel Identification by Neural Networks Analysis of the Response of Vapor Sensitive Sensors Arrays Edinboro University of Pennsylvania, Edinboro, PA	Dr. Paul Edwards Chemistry AL/EQ
5	A Comparison of Multistep vs Singlestep Arrhenius Integral Models for Describing Laser Induced Thermal Damage Florida International University, Miami, FL	Dr. Bernard Gerstman Physics AL/OE
6	Effects of Mental Workload and Electronic Support on Negotiation Performance University of Dayton, Dayton, OH	Dr. Kenneth Graetz Psychology AL/HR
7	Regression to the Mean in Half Life Studies University of Main, Orono, ME	Dr. Pushpa Gupta Mathematics & Statistics AL/AO
8	Application of the MT3D Solute Transport Model to the Made-2 Site: Calibration Florida State University, Tallahassee, FL	Dr. Manfred Koch Geophysics AL/EQ
9	Computer Calculations of Gas-Phase Reaction Rate Constants Florida State University, Tallahassee, FL	Dr. Mark Novotny SupercompComp. Res. I AL/EQ
10	Surface Fitting Three Dimensional Human Scan Data Ohio University, Athens, OH	Dr. Joseph Nurre Mechanical Engineering AL/CF
11	The Effects of Hyperbaric Oxygenation on Metabolism of Drugs and Other Xenobioti University of So. Carolina, Columbia, So. Carolina	Dr. Edward Piepmeier Pharmaceutics AL/AO
12	Maintaining Skills After Training: The Role of Opportunity to Perform Trained Tasks on Training Effectiveness Rice University, Houston, TX	Dr. Miguel Quinones Psychology AL/HR

1995 SREP FINAL REPORTS

Armstrong Laboratory

VOLUME 1 (cont.)

Report #	Report Title Author's University	Report Author
13	Nonlinear Transcutaneous Electrical Stimulation of the Vestibular System University of Illinois Urbana-Champaign, Urbana,IL	Dr. Gary Riccio Psychology AL/CF
14	Documentation of Separating and Separated Boundary Layer Flow, For Application Texas A&M University, College Station, TX	Dr. Wayne Shebilske Psychology AL/HR
15	Tactile Feedback for Simulation of Object Shape and Textural Information in Haptic Displays Ohio State University, Columbus, OH	Dr. Janet Weisenberger Speech & Hearing AL/CF
16	Melatonin Induced Prophylactic Sleep as a Countermeasure for Sleep Deprivation Oregon Health Sciences University, Portland, OR	Mr. Rod Hughes Psychology AL/CF

1995 SREP FINAL REPORTS

Phillips Laboratory

VOLUME 2A

Report #	Report Title Author's University	Report Author
1	Investigation of the Mixed-Mode Fracture Behavior of Solid Propellants University of Houston, Houston, TX	Dr. K. Ravi-Chandar Aeronautics PL/RK
2	Performance Study of ATM-Satellite Network SUNY-Buffalo, Buffalo, NY	Dr. Nasser Ashgriz Mechanical Engineering PL/RK
3	Characterization of CMOS Circuits Using a Highly Calibrated Low-Energy X-Ray Source Embry-Riddle Aeronautical Univ., Prescott, AZ	Dr. Raymond Bellem Computer Science PL/VT
4	Neutron Diagnostics for Pulsed Plasmas of Compact Toroid-Marauder Type Stevens Institute of Tech, Hoboken, NJ	Dr. Jan Brzosko Nuclear Physics PL/WS
5	Parallel Computation of Zernike Aberration Coefficients for Optical Aberration Correction University of Houston-Victoria, Victoria, TX	Dr. Meledath Damodaran Math & Computer Science PL/LI
6	Quality Factor Evaluation of Complex Cavities University of Denver, Denver, CO	Dr. Ronald DeLyser Electrical Engineering PL/WS
7	Unidirectional Ring Lasers and Laser Gyros with Multiple Quantum Well Gain University of New Mexico, Albuquerque, NM	Dr. Jean-Claude Diels Physics PL/LI
8	A Tool for the Formation of Variable Parameter Inverse Synthetic Aperture Radar University of Nevada, Reno, NV	Dr. James Henson Electrical Engineering PL/WS
9	Radar Ambiguity Functionals Univ. of Massachusetts at Lowell, Lowell, MA	Dr. Gerald Kaiser Physics PL/GP
10	The Synthesis and Chemistry of Peroxonitrites Peroxonitrous Acid Univ. of Massachusetts at Lowell, Lowell, MA	Dr. Albert Kowalak Chemistry PL/GP
11	Temperature and Pressure Dependence of the Band Gaps and Band Offsets University of Houston, Houston, TX	Dr. Kevin Malloy Electrical Engineering PL/VT
12	Theoretical Studies of the Performance of Novel Fiber-Coupled Imaging Interferom University of New Mexico, Albuquerque, NM	Dr. Sudhakar Prasad Physics PL/LI

1995 SREP FINAL REPORTS

Phillips Laboratory

VOLUME 2B

Report #	Author's University	Report Author
13	Static and Dynamic Graph Embedding for Parallel Programming Texas AandM Univ.-Kingsville, Kingsville, TX	Dr. Mark Purtill Mathematics PL/WS
14	Ultrafast Process and Modulation in Iodine Lasers University of New Mexico, Albuquerque, NM	Dr. W. Rudolph Physics PL/LI
15	Impedance Matching and Reflection Minimization for Transient EM Pulses Through University of New Mexico, Albuquerque, NM	Dr. Alexander Stone Mathematics and Statics PL/WS
16	Low Power Retromodular Based Optical Transceiver for Satellite Communications Utah State University, Logan, UT	Dr. Charles Swenson Electrical Engineering PL/VT
17	Improved Methods of Tilt Measurement for Extended Images in the Presence of Atmospheric Disturbances Using Optical Flow Michigan Technological Univ., Houghton, MI	Mr. John Lipp Electrical Engineering PL/LI
18	Thermoluminescence of Simple Species in Molecular Hydrogen Matrices Cal State Univ.-Northridge, Northridge, CA	Ms. Janet Petroski Chemistry PL/RK
19	Design, Fabrication, Intelligent Cure, Testing, and Flight Qualification University of Cincinnati, Cincinnati, OH	Mr. Richard Salasovich Mechanical Engineering PL/VT

1995 SREP FINAL REPORTS

Rome Laboratory

VOLUME 3

Report #	Author's University	Report Author
1	Performance Study of an ATM/Satellite Network Florida Atlantic University, Boca Raton, FL	Dr. Valentine Aalo Electrical Engineering RL/C3
2	Interference Excision in Spread Spectrum Communication Systems Using Time-Frequency Distributions Villanova University, Villanova, PA	Dr. Moeness Amin Electrical Engineering RL/C3
3	Designing Software by Reformulation Using KIDS Oklahoma State University, Stillwater, OK	Dr. David Benjamin Computer Science RL/C3
4	Detection Performance of Over Resolved Targets with Non-Uniform and Non-Gaussian Howard University, Washington, DC	Dr. Ajit Choudhury Engineering RL/OC
5	Computer-Aided-Design Program for Solderless Coupling Between Microstrip and Stripline Structures Southern Illinois University, Carbondale, IL	Dr. Frances Harackiewicz Electrical Engineering RL/ER
6	Spanish Dialect Identification Project Colorado State University, Fort Collins, CO	Dr. Beth Losiewicz Psycholinguistics RL/IR
7	Automatic Image Registration Using Digital Terrain Elevation Data University of Maine, Orono, ME	Dr. Mohamed Musavi Engineering RL/IR
8	Infrared Images of Electromagnetic Fields University of Colorado, Colorado Springs, CO	Dr. John Norgard Engineering RL/ER
9	Femtosecond Pump-Probe Spectroscopy System SUNY Institute of Technology, Utica, NY	Dr. Dean Richardson Photonics RL/OC
10	Synthesis and Properties B-Diketonate-Modified Heterobimetallic Alkoxides Tufts University, Medford, MA	Dr. Daniel Ryder, Jr. Chemical Engineering RL/ER
11	Optoelectronic Study of Seniconductor Surfaces and Interfaces Rensselaer Polytechnic Institute, Troy, NY	Dr. Xi-Cheng Zhang Physics RL/ER

1995 SREP FINAL REPORTS

Wright Laboratory

VOLUME 4A

Report #	Author's University	Report Author
1	An Investigation of the Heating and Temperature Distribution in Electrically Excited Foils Auburn University, Auburn, AL	Dr. Michael Baginski Electrical Engineering WL/MN
2	Micromechanics of Creep in Metals and Ceramics at High Temperature Wayne State University, Detroit, MI	Dr. Victor Berdichevsky Aerospace Engineering WL/FI
3	Development of a Fluorescence-Based Chemical Sensor for Simultaneous Oxygen Quantitation and Temp. Measurement Columbus College, Columbus, GA	Dr. Steven Buckner Chemistry WL/PO
4	Development of High-Performance Active Dynamometer Sys. for Machines and Drive Clarkson University, Potsdam, NY	Dr. James Carroll Electrical Engineering WL/PO
5	SOLVING $z(t)=1n[A\cos(w_1t)+B\cos(w_2)+C]$ Transylvania University, Lexington, KY	Dr. David Choate Mathematics WL/AA
6	Synthesis, Processing and Characterization of Nonlinear Optical Polymer Thin Films University of cincinnati, Cincinnati, OH	Dr. Stephen Clarson Matls Science& Engineering WL/ML
7	An Investigation of Planning and Scheduling Algorithms for Sensor Management Embry-Riddle Aeronautical University, Prescott, AZ	Dr. Milton Cone Comp. Science & Engineering WL/AA
8	A Study to Determine Wave Gun Firing Cycles for High Performance Model Launches Louisiana State University, Baton Rouge, LA	Dr. Robert Courter Mechanical Engineering WL/MN
9	Characterization of Electro-Optic Polymers University of Dayton, Dayton, OH	Dr. Vincent Dominic Electro Optics Program WL/ML
10	A Methodology for Affordability in the Design Process Clemson Univeristy, Clemson, SC	Dr. Georges Fadel Mechanical Engineering WL/MT
11	Data Reduction and Analysis for Laser Doppler Velocimetry North Carolina State University, Raleigh, NC	Dr. Richard Gould Mechanical Engineering WL/PO

Report #	Author's University	Report Author
12	Hyperspectral Target Identification Using Bomen Spectrometer Data University of Dayton, Dayton, OH	Dr. Russell Hardie Electrical Engineering WL/AA
13	Robust Fault Detection and Classification Auburn University, Auburn, AL	Dr. Alan Hodel Electrical Engineering WL/MN
14	Multidimensional Algorithm Development and Analysis Mississippi State University, Mississippi State University, MS	Dr. Jonathan Janus Aerospace Engineering WL/MN
15	Characterization of Interfaces in Metal-Matrix Composites Michigan State University, East Lansing, MI	Dr. Iwona Jasiuk Materials Science WL/ML
16	TSI Mitigation: A Mountaintop Database Study Lafayette College, Easton, PA	Dr. Ismail Jouny Electrical Engineering WL/AA
17	Comparative Study and Performance Analysis of High Resolution SAR Imaging Techniques University of Florida, Gainesville, FL	Dr. Jian Li Electrical Engineering WL/AA

1995 SREP FINAL REPORTS

Wright Laboratory

VOLUME 4B

Report #	Author's University	Report Author
18	Prediction of Missile Trajectory University of Missouri-Columbia, Columbia, MO	Dr. Chun-Shin Lin Electrical Engineering WL/FI
19	Three Dimensional Deformation Comparison Between Bias and Radial Aircraft Tires Cleveland State University, Cleveland, OH	Dr. Paul Lin Mechanical Engineering WL/FI
20	Investigation of A1GaAs/GaAs Heterojunctin Bipolar Transistor Reliability Based University of Central Florida, Orlando, FL	Dr. Juin Liou Electrical Engineering WL/EL
21	Thermophysical Invariants From LWIR Imagery for ATR University of Virginia, Charlottesville, VA	Dr. Nagaraj Nandhakumar Electrical Engineering WL/AA
22	Effect of Electromagnetic Environment on Array Signal Processing University of Dayton, Dayton, OH	Dr. Krishna Pasala Electrical Engineering WL/AA
23	Functional Decomposotion of Binary, Multiple-Valued, and Fuzzy Logic Portland State University, Portland, OR	Dr. Marek Perkowski Electrical Engineering WL/AA
24	Superresolution of Passive Millimeter-Wave Imaging Auburn University, Auburn, AL	Dr. Stanley Reeves Electrical Engineering WL/MN
25	Development of a Penetrator Optimizer University of Alabama, Tuscaloosa, AL	Dr. William Rule Engineering Science WL/MN
26	Heat Transfer for Turbine Blade Film Cooling with Free Stream Turbulence-Measurements and Predictions University of Dayton, Dayton, OH	Dr. John Schauer Mech. & Aerosp. Engineering WL/FI
27	Neural Network Identification and Control in Metal Forging University of Florida, Gainesville, FL	Dr. Carla Schwartz Electrical Engineering WL/FI
28	Documentation of Separating and Separated Boundary Layer Flow, for Application University of Minnesota, Minneapolis, MN	Dr. Terrence Simon Mechanical Engineering WL/PO
29	Transmission Electron Microscopy of Semiconductor Heterojunctions Carnegie Melon University, Pittsburgh, PA	Dr. Marek Skowronski Matls Science & Engineering WL/EL

1995 SREP FINAL REPORTS

Wright Laboratory

VOLUME 4B (cont.)

Report #	Author's University	Report Author
30	Parser in SWI-PROLOG Wright State University, Dayton, OH	Dr. K. Thirunarayan Computer Science WL/EL
31	Development of Qualitative Process Control Discovery Systems for Polymer Composite and Biological Materials University of California, Los Angeles, CA	Dr. Robert Trelease Anatomy & Cell Biology WL/ML
32	Improved Algorithm Development of Massively Parallel Epic Hydrocode in Cray T3D Massively Parallel Computer Florida Atlantic University, Boca Raton, FL	Dr. Chi-Tay Tsai Engineering Mechanics WL/MN
33	The Characterization of the Mechanical Properties of Materials in a Biaxial Stress Environment University of Kentucky, Lexington, KY	Dr. John Lewis Materials Science Engineering WL/MN

1995 SREP FINAL REPORTS

VOLUME 5

Report #	Author's University	Report Author
	Arnold Engineering Development Center	
1	Plant-Wide Preventive Maintenance and Monitoring Vanderbilt University	Mr. Theodore Bapty Electrical Engineering AEDC
	Frank J. Seiler Research Laboratory	
1	Block Copolymers at Inorganic Solid Surfaces Colorado School of Mines, Golden, CO	Dr. John Dorgan Chemical Engineering FJSRL
2	Non-Linear Optical Properties of Polyacetylenes and Related Barry University, Miami, FL	Dr. M. A. Jungbauer Chemistry FJSRL
3	Studies of Second Harmonic Generation in Glass Waveguides Allegheny College, Meadville, PA	Dr. David Statman Physics FJSRL
	Wilford Hall Medical Center	
1	Biochemical & Cell Physiological Aspects of Hyperthermia University of Miami, Coral Gables, FL	Dr. W. Drost-Hansen Chemistry WHMC

1995 SUMMER RESEARCH EXTENSION PROGRAM (SREP) MANAGEMENT REPORT

1.0 BACKGROUND

Under the provisions of Air Force Office of Scientific Research (AFOSR) contract F49620-90-C-0076, September 1990, Research & Development Laboratories (RDL), an 8(a) contractor in Culver City, CA, manages AFOSR's Summer Research Program. This report is issued in partial fulfillment of that contract (CLIN 0003AC).

The Summer Research Extension Program (SREP) is one of four programs AFOSR manages under the Summer Research Program. The Summer Faculty Research Program (SFRP) and the Graduate Student Research Program (GSRP) place college-level research associates in Air Force research laboratories around the United States for 8 to 12 weeks of research with Air Force scientists. The High School Apprenticeship Program (HSAP) is the fourth element of the Summer Research Program, allowing promising mathematics and science students to spend two months of their summer vacations working at Air Force laboratories within commuting distance from their homes.

SFRP associates and exceptional GSRP associates are encouraged, at the end of their summer tours, to write proposals to extend their summer research during the following calendar year at their home institutions. AFOSR provides funds adequate to pay for SREP subcontracts. In addition, AFOSR has traditionally provided further funding, when available, to pay for additional SREP proposals, including those submitted by associates from Historically Black Colleges and Universities (HBCUs) and Minority Institutions (MIs). Finally, laboratories may transfer internal funds to AFOSR to fund additional SREPs. Ultimately the laboratories inform RDL of their SREP choices, RDL gets AFOSR approval, and RDL forwards a subcontract to the institution where the SREP associate is employed. The subcontract (see Appendix 1 for a sample) cites the SREP associate as the principal investigator and requires submission of a report at the end of the subcontract period.

Institutions are encouraged to share costs of the SREP research, and many do so. The most common cost-sharing arrangement is reduction in the overhead, fringes, or administrative charges institutions would normally add on to the principal investigator's or research associate's labor. Some institutions also provide other support (e.g., computer run time, administrative assistance, facilities and equipment or research assistants) at reduced or no cost.

When RDL receives the signed subcontract, we fund the effort initially by providing 90% of the subcontract amount to the institution (normally $18,000 for a $20,000 SREP). When we receive the end-of-research report, we evaluate it administratively and send a copy to the laboratory for a technical evaluation. When the laboratory notifies us the SREP report is acceptable, we release the remaining funds to the institution.

2.0 THE 1995 SREP PROGRAM

SELECTION DATA: A total of 719 faculty members (SFRP Associates) and 286 graduate students (GSRP associates) applied to participate in the 1994 Summer Research Program. From these applicants 185 SFRPs and 121 GSRPs were selected. The education level of those selected was as follows:

1994 SRP Associates, by Degree			
SFRP		GSRP	
PHD	MS	MS	BS
179	6	52	69

Of the participants in the 1994 Summer Research Program 90 percent of SFRPs and 25 percent of GSRPs submitted proposals for the SREP. Ninety proposals from SFRPs and ten from GSRPs were selected for funding, which equates to a selection rate of 54% of the SFRP proposals and of 34% for GSRP proposals.

1995 SREP: Proposals Submitted vs. Proposals Selected			
	Summer 1994 Participants	Submitted SREP Proposals	SREPs Funded
SFRP	185	167	90
GSRP	121	29	10
TOTAL	306	196	100

The funding was provided as follows:

Contractual slots funded by AFOSR	75
Laboratory funded	14
Additional funding from AFOSR	11
Total	100

Six HBCU/MI associates from the 1994 summer program submitted SREP proposals; six were selected (none were lab-funded; all were funded by additional AFOSR funds).

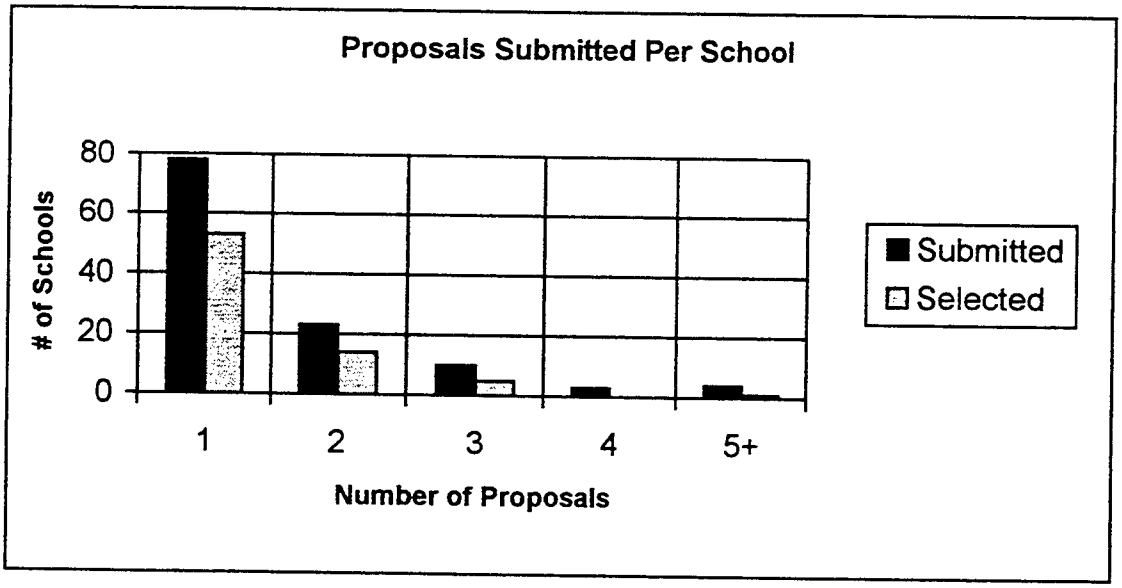
Proposals Submitted and Selected, by Laboratory		
	Applied	Selected
Armstrong Laboratory	41	19
Arnold Engineering Development Center	12	4
Frank J. Seiler Research Laboratory	6	3
Phillips Laboratory	33	19
Rome Laboratory	31	13
Wilford Hall Medical Center	2	1
Wright Laboratory	62	37
TOTAL		

Note: Phillips Laboratory funded 3 SREPs; Wright Laboratory funded 11; and AFOSR funded 11 beyond its contractual 75.

The 306 1994 Summer Research Program participants represented 135 institutions.

Institutions Represented on the 1994 SRP and 1995 SREP		
Number of schools represented in the Summer 92 Program	Number of schools represented in submitted proposals	Number of schools represented in Funded Proposals
135	118	73

Forty schools had more than one participant submitting proposals.



The selection rate for the 78 schools submitting 1 proposal (68%) was better than those submitting 2 proposals (61%), 3 proposals (50%), 4 proposals (0%) or 5+ proposals (25%). The 4 schools that submitted 5+ proposals accounted for 30 (15%) of the 196 proposals submitted.

Of the 196 proposals submitted, 159 offered institution cost sharing. Of the funded proposals which offered cost sharing, the minimum cost share was $1000.00, the maximum was $68,000.00 with an average cost share of $12,016.00.

Proposals and Institution Cost Sharing		
	Proposals Submitted	Proposals Funded
With cost sharing	159	82
Without cost sharing	37	18
Total	196	100

The SREP participants were residents of 41 different states. Number of states represented at each laboratory were:

States Represented, by Proposals Submitted/Selected per Laboratory		
	Proposals Submitted	Proposals Funded
Armstrong Laboratory	21	13
Arnold Engineering Development Center	5	2
Frank J. Seiler Research Laboratory	5	3
Phillips Laboratory	16	14
Rome Laboratory	14	7
Wilford Hall Medical Center	2	1
Wright Laboratory	24	20

Eleven of the 1995 SREP Principal Investigators also participated in the 1994 SREP.

ADMINISTRATIVE EVALUATION: The administrative quality of the SREP associates' final reports was satisfactory. Most complied with the formatting and other instructions provided to them by RDL. Ninety seven final reports and two interim reports have been received and are included in this report. The subcontracts were funded by $1,991,623.00 of Air Force money. Institution cost sharing totaled $985,353.00.

TECHNICAL EVALUATION: The form used for the technical evaluation is provided as Appendix 2. ninety-two evaluation reports were received. Participants by laboratory versus evaluations submitted is shown below:

	Participants	Evaluations	Percent
Armstrong Laboratory	23[1]	20	95.2
Arnold Engineering Development Center	4	4	100
Frank J. Seiler Research Laboratory	3	3	100
Phillips Laboratory	19[2]	18	100
Rome Laboratory	13	13	100
Wilford Hall Medical Center	1	1	100
Wright Laboratory	37	34	91.9
Total			

Notes:

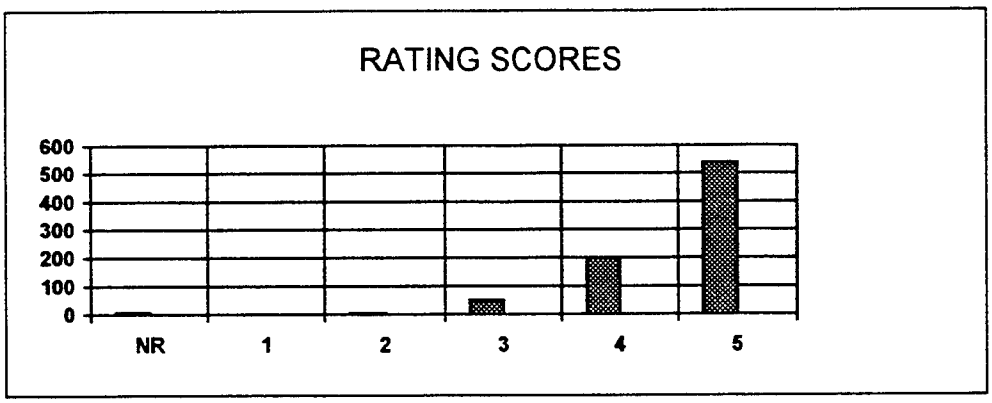
1: Research on two of the final reports was incomplete as of press time so there aren't any technical evaluations on them to process, yet. Percent complete is based upon 20/21=95.2%

2: One technical evaluation was not completed because one of the final reports was incomplete as of press time. Percent complete is based upon 18/18=100%

3: See notes 1 and 2 above. Percent complete is based upon 93/97=95.9%

The number of evaluations submitted for the 1995 SREP (95.9%) shows a marked improvement over the 1994 SREP submittals (65%).

PROGRAM EVALUATION: Each laboratory focal point evaluated ten areas (see Appendix 2) with a rating from one (lowest) to five (highest). The distribution of ratings was as follows:



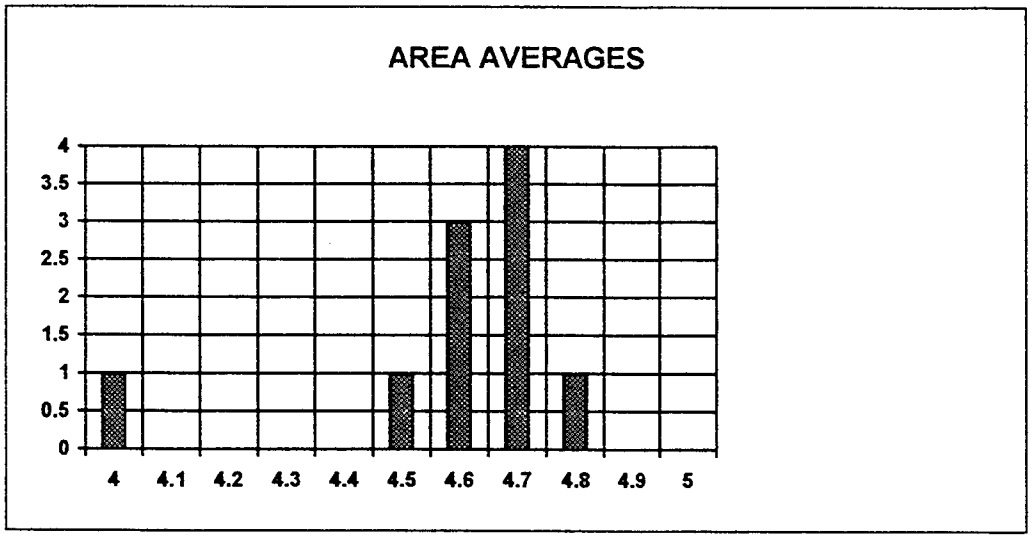
Rating	Not Rated	1	2	3	4	5
# Responses	7	1	7	62 (6%)	226 (25%)	617 (67%)

The 8 low ratings (one 1 and seven 2's) were for question 5 (one 2) "The USAF should continue to pursue the research in this SREP report" and question 10 (one 1 and six 2's) "The

one-year period for complete SREP research is about right", in addition over 30% of the threes (20 of 62) were for question ten. The average rating by question was:

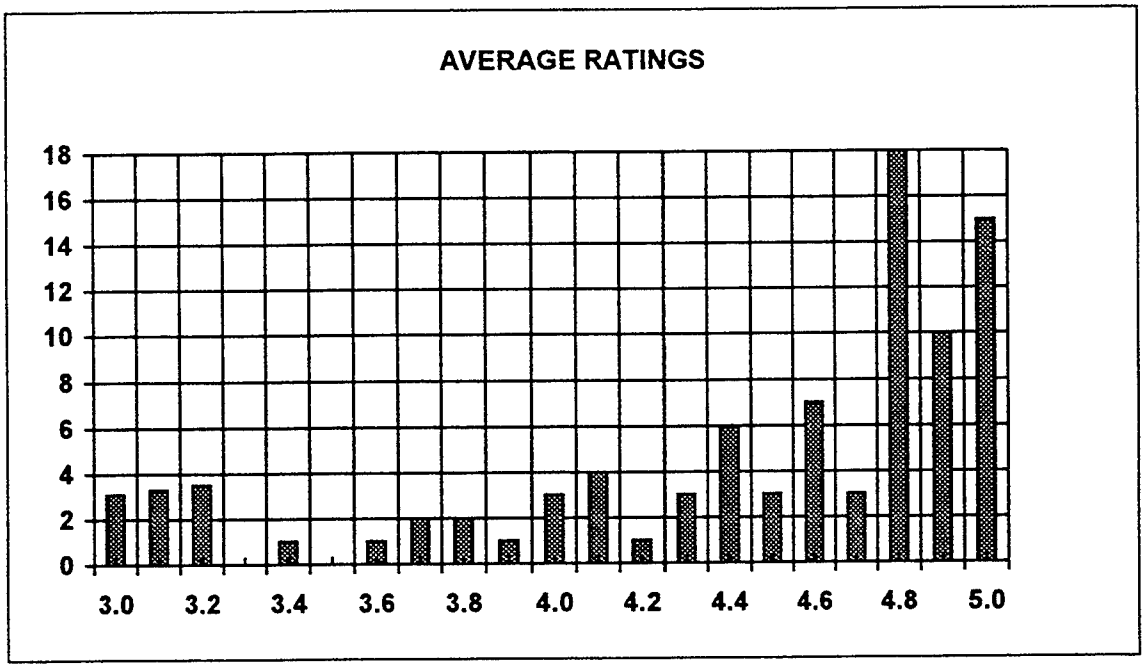
Question	1	2	3	4	5	6	7	8	9	10
Average	4.6	4.6	4.7	4.7	4.6	4.7	4.8	4.5	4.6	4.0

The distribution of the averages was:



Area 10 "the one-year period for complete SREP research is about right" had the lowest average rating (4.1). The overall average across all factors was 4.6 with a small sample standard deviation of 0.2. The average rating for area 10 (4.1) is approximately three sigma lower than the overall average (4.6) indicating that a significant number of the evaluators feel that a period of other than one year should be available for complete SREP research.

The average ratings ranged from 3.4 to 5.0. The overall average for those reports that were evaluated was 4.6. Since the distribution of the ratings is not a normal distribution the average of 4.6 is misleading. In fact over half of the reports received an average rating of 4.8 or higher. The distribution of the average report ratings is as shown:



It is clear from the high ratings that the laboratories place a high value on AFOSR's Summer Research Extension Programs.

3.0 SUBCONTRACTS SUMMARY

Table 1 provides a summary of the SREP subcontracts. The individual reports are published in volumes as shown:

Laboratory	Volume
Armstrong Laboratory	1A, 1B
Arnold Engineering Development Center	5
Frank J. Seiler Research Laboratory	5
Phillips Laboratory	2
Rome Laboratory	3
Wilford Hall Medical Center	5
Wright Laboratory	4A, 4B

1995 SREP SUB-CONTRACT DATA

Report Author Author's University	Author's Degree	Sponsoring Lab	Performance Period	Contract Amount	Univ. Cost Share
Anderson , James **Analytical Chemistry** **University of Georgia, Athens, GA**	**PhD** 95-0807	**AL/EQ**	**01/01/95 12/31/95** Determination of the Redox Capacity of Soil Sediment and Prediction of Pollutant	**$25000.00**	**$1826.00**
Ashrafiuon , Hashem **Mechanical Engineering** **Villanova University, Villanova, PA**	**PhD** 95-0800	**AL/CF**	**01/01/95 12/31/95** Finite Element Modeling of the Human Neck and Its Validation for the ATB Model	**$25000.00**	**$19528.00**
Burke , Michael **Tulane University** **Tulane University, New Orleans, LA**	**PhD** 95-0811	**AL/HR**	**01/01/95 09/30/95** An Examination of the Validity of the New Air Force ASVAB Composites	**$25000.00**	**$1818.00**
Edwards , Paul **Chemistry** **Edinboro Univ of Pennsylvania, Edinboro, PA**	**PhD** 95-0808	**AL/EQ**	**01/01/95 12/31/95** Fuel Identification by Neural Networks Analysis of the Response of Vapor Sensiti	**$25000.00**	**$5000.00**
Gerstman , Bernard **Physics** **Florida International Universi, Miami, FL**	**PhD** 95-0815	**AL/OE**	**01/01/95 12/31/95** A Comparison of Multistep vs Singlestep Arrhenius Integral Models for Describing	**$24289.00**	**$2874.00**
Graetz , Kenneth **Department of Psychology** **University of Dayton, Dayton, OH**	**PhD** 95-0812	**AL/HR**	**01/01/95 12/31/95** Effects of Mental Workload and Electronic Support on Negotiation Performance	**$25000.00**	**$0.00**
Gupta , Pushpa **Mathematics** **University of Maine, Orono, ME**	**PhD** 95-0802	**AL/AO**	**01/01/95 12/31/95** Regression to the Mean in Half Life Studies	**$25000.00**	**$2859.00**
Koch , Manfred **Geophysics** **Florida State University, Tallahassee, FL**	**PhD** 95-0809	**AL/EQ**	**12/01/94 04/30/95** Application of the MT3D Solute Transport Model to the Made-2 Site: Calibration	**$25000.00**	**$0.00**
Novotny , Mark **Supercomputer Comp Res. I** **Florida State University, Tallahassee, FL**	**PhD** 95-0810	**AL/EQ**	**01/01/95 12/31/95** Computer Calculations of Gas-Phase Reaction Rate Constants	**$25000.00**	**$0.00**
Nurre , Joseph **Mechanical Engineering** **Ohio University, Athens, OH**	**PhD** 95-0804	**AL/CF**	**01/01/95 12/31/95** Surface Fitting Three Dimensional Human Head Scan Data	**$25000.00**	**$20550.00**
Piepmeier , Edward . **Pharmaceutics** **University of South Carolina, Columbia, SC**	**PhD** 95-0801	**AL/AO**	**01/01/95 12/31/95** The Effects of Hyperbaric Oxygenation on Metabolism of Drugs and Other Xenobioti	**$25000.00**	**$11740.00**
Quinones , Miguel **Psychology** **Rice University, Houston, TX**	**PhD** 95-0813	**AL/HR**	**01/01/95 12/31/95** Maintaining Skills After Training: The Role of Opportunity to Perform Trained T	**$25000.00**	**$4000.00**
Riccio , Gary **Psychology** **Univ of IL Urbana-Champaign, Urbana, IL**	**PhD** 95-0806	**AL/CF**	**01/01/95 05/31/95** Nonlinear Transcutaneous Electrical Stimulation of the Vestibular System	**$22931.00**	**$0.00**
Shebilske , Wayne **Dept of Psychology** **Texas A&M University, College Station, TX**	**PhD** 95-0814	**AL/HR**	**01/01/95 12/31/95** Cognitive Factors in Distr Training Effects During Acquisition of Complex Skills	**$25000.00**	**$5614.00**

1995 SREP SUB-CONTRACT DATA

Report Author Author's University	Author's Degree	Sponsoring Lab	Performance Period	Contract Amount	Univ. Cost Share
Weisenberger , Janet **Dept of Speech & Hearing** **Ohio State University, Columbus, OH**	**PhD** 95-0805	**AL/CF** Tactile Feedback for Simulation of Object Shape and Textural Information in Hapt	**01/01/95 12/31/95**	**$25000.00**	**$12234.00**
Hughes , Rod **Psychology** **Oregon Health Sciences University, Portland, OR**	**MA** 95-0803	**AL/CF** Melatonin Induced Prophylactic Sleep as a Countermeasure for Sleep Deprivation	**01/01/95 12/31/95**	**$25000.00**	**$0.00**
Bapty , Theodore **Electrical Engineering** **Vanderbilt University, Nashville, TN**	**MS** 95-0848	**AEDC/E** Plant-Wide Preventive Maintenance & Monitoring	**01/01/95 12/31/95**	**$24979.00**	**$0.00**
Dorgan , John **Chemical Engineering** **Colorado School of Mines, Golden, CO**	**PhD** 95-0834	**FJSRL/F** Block Copolymers at Inorganic Solid Surfaces	**01/01/95 12/31/95**	**$25000.00**	**$0.00**
Jungbauer , Mary Ann **Chemistry** **Barry University, Miami, FL**	**PhD** 95-0836	**FJSRL/F** Non-Linear Optical Properties of Polyacetylenes and Related Substituted Compound	**01/01/95 12/31/95**	**$25000.00**	**$24714.00**
Statman , David **Physics** **Allegheny College, Meadville, PA**	**PhD** 95-0835	**FJSRL/F** Studies of Second Harmonic Generation in Glass Waveguides	**01/01/95 12/31/95**	**$25000.00**	**$6500.00**
, Krishnaswamy **Aeronautics** **University of Houston, Houston, TX**	**PhD** 95-0818	**PL/RK** Mixed-Mode Fracture of Solid Propellants	**01/01/95 12/31/95**	**$24993.00**	**$8969.00**
Ashgriz , Nasser **Mechanical Engineering** **SUNY-Buffalo, Buffalo, NY**	**PhD** 95-0816	**PL/RK** Effects of the Jet Characteristics on the Atomization and Mixing in A Pair of Im	**01/01/95 12/31/95**	**$25000.00**	**$22329.00**
Bellem , Raymond **Computer Science** **Embry-Riddle Aeronautical Univ, Prescott, AZ**	**PhD** 95-0817	**PL/VT** Experimental Studies of the Effects of Ionizing Radiation on Commerically Proces	**12/01/94 11/30/95**	**$20000.00**	**$8293.00**
Brzosko , Jan **Nuclear Physics** **Stevens Institute of Tech, Hoboken, NJ**	**PhD** 95-0828	**PL/WS** Neutron Diagnostics for Pulsed Plasmas of Compact Toroid - Marauder Type	**11/01/94 02/01/95**	**$24943.00**	**$0.00**
Damodaran , Meledath **Math & Computer Science** **University of Houston-Victoria, Victoria, TX**	**PhD** 95-0831	**PL/LI** Parallel Computation of Zernike Aberration Coefficients for Optical Aber Correct	**01/01/95 12/31/94**	**$24989.00**	**$9850.00**
DeLyser , Ronald **Electrical Engineering** **University of Denver, Denver, CO**	**PhD** 95-0877	**PL/WS** Quality Factor Evaluation of Complex Cavities	**01/01/95 12/31/95**	**$25000.00**	**$46066.00**
Diels , Jean-Claude **Physics** **University of New Mexico, Albuquerque, NM**	**PhD** 95-0819	**PL/LI** Unidirectional Ring Lasers and Laseer Gyros with Multiple Quantum Well Gain Medi	**01/01/95 12/31/95**	**$25000.00**	**$0.00**
Henson , James **Electrical Engineering** **University of Nevada, Reno, NV**	**PhD** 95-0820	**PL/WS** Automatic Feature Extraction and Assessment of Wideband Range-Doppler Imagery of	**01/01/95 12/31/95**	**$25000.00**	**$0.00**
Kaiser , Gerald **Physics** **University of Mass/Lowell, Lowell, MA**	**PhD** 95-0821	**PL/GP** Multiresolution Analysis with Physical Wavelets	**01/01/95 12/31/95**	**$25000.00**	**$5041.00**

1995 SREP SUB-CONTRACT DATA

Report Author Author's University	Author's Degree	Sponsoring Lab	Performance Period	Contract Amount	Univ. Cost Share
Kowalak , Albert **Chemistry** **University of Massachusetts/Lo, Lowell, MA**	**PhD** 95-0822	**PL/GP**	**01/01/95 12/31/95** The Synthesis and Chemistry of Peroxonitrites and Peroxonitrous Acid	**$24996.00**	**$4038.00**
Malloy , Kevin **Electrical Engineering** **University of New Mexico, Albuquerque, NM**	**PhD** 95-0829	**PL/VT**	**01/01/95 12/31/95** Temperature & Pressure Dependence of the Band Gaps & Band Offsets	**$24999.00**	**$0.00**
Prasad , Sudhakar **Physics** **University of New Mexico, Albuquerque, NM**	**PhD** 95-0823	**PL/LI**	**01/01/95 12/31/95** Theoretical Studies of the Performance of Novel Fiber-Coupled Imaging Interferom	**$25000.00**	**$11047.00**
Purtill , Mark **Mathematics** **Texas A&M Univ-Kingsville, Kingsville, TX**	**PhD** 95-0824	**PL/WS**	**01/01/95 12/31/95** Static and Dynamic Graph Embedding for Parallel Programming	**$25000.00**	**$100.00**
Rudolph , Wolfgang **Physics** **University of New Mexico, Albuquerque, NM**	**PhD** 95-0833	**PL/LI**	**01/01/95 12/31/95** Ultrafast Process and Modulation in Iodine Lasers	**$24982.00**	**$6000.00**
Stone , Alexander **Mathematics & Statistics** **University of New Mexico, Alburquerque, NM**	**PhD** 95-0827	**PL/WS**	**01/01/95 12/31/95** Impedance Matching And Reflection Minimization For Transient EM Pulses Through D	**$24969.00**	**$0.00**
Swenson , Charles **Dept of Electrical Engnr** **Utah State University, Logan, UT**	**PhD** 95-0826	**PL/VT**	**01/01/95 12/31/95** Low Power Retromodulator based Optical Transceiver for Satellite Communications	**$25000.00**	**$25000.00**
Lipp , John **Electrical Engineering** **Michigan Technological Univ, Houghton, MI**	**MS** 95-0832	**PL/LI**	**01/01/95 12/31/95** Improved Methods of Tilt Measurement for Extended Images in the Presence of Atmo	**$24340.00**	**$15200.00**
Petroski , Janet **Chemistry** **Cal State Univ/Northridge, Northridge, CA**	**BA** 95-0830	**PL/RK**	**10/01/94 12/31/94** Thermoluminescence of Simple Species in Molecular Hydrogen Matrices	**$4279.00**	**$0.00**
Salasovich , Richard **Mechanical Engineering** **University of Cincinnati, Cincinnati, OH**	**MS** 95-0825	**PL/VT**	**01/01/95 12/31/95** Design, Fabrication, Intelligent Cure, Testing, and Flight Qualification of an A	**$25000.00**	**$4094.00**
Aalo , Valentine **Dept of Electrical Engr** **Florida Atlantic University, Boca Raton, FL**	**PhD** 95-0837	**RL/C3**	**01/01/95 12/31/95** Performance Study of an ATM/Satellite Network	**$25000.00**	**$13120.00**
Amin , Moeness **Electrical Engineering** **Villanova University, Villanova, PA**	**PhD** 95-0838	**RL/C3**	**01/01/95 12/31/95** Interference Excision in Spread Spectrum Communication Systems Using Time-Freque	**$25000.00**	**$34000.00**
Benjamin , David **Computer Science** **Oklahoma State University, Stillwater, OK**	**PhD** 95-0839	**RL/C3**	**01/01/95 12/31/95** Designing Software by Decomposition using KIDS	**$24970.00**	**$0.00**
Choudhury , Ajit **Engineering** **Howard University, Washington, DC**	**PhD** 95-0840	**RL/OC**	**11/30/94 10/31/95** Detection Performance of Over Resolved Targets with Non-Uniform and Non-GAussian	**$25000.00**	**$0.00**
Harackiewicz , Frances **Electrical Engineering** **So. Illinois Univ-Carbondale, Carbondale, IL**	**PhD** 95-0841	**RL/ER**	**01/01/95 12/31/95** Computer-Aided-Design Program for Solderless Coupling Between Microstrip and Str	**$23750.00**	**$29372.00**

1995 SREP SUB-CONTRACT DATA

Report Author Author's University	Author's Degree	Sponsoring Lab	Performance Period	Contract Amount	Univ. Cost Share
Losiewicz , Beth **Psycholinguistics** **Colorado State University, Fort Collins, CO**	**PhD** 95-0842	**RL/IR**	**01/01/95 12/31/95** Spanish Dialect Identification Project	**$25000.00**	**$4850.00**
Musavi , Mohamad **University of Maine, Orono, ME**	**PhD** 95-0843	**RL/IR**	**01/01/95 12/31/95** Automatic Image Registration Using Digital Terrain Elevation Data	**$25000.00**	**$12473.00**
Norgard , John **Elec & Comp Engineering** **Univ of Colorado-Colorado Sprg, Colorado**	**PhD** 95-0844	**RL/ER**	**01/01/95 12/31/95** Infrared Images of Electromagnetic Fields	**$25000.00**	**$2500.00**
Richardson , Dean **Photonics** **SUNY Institute of Technology, Utica, NY**	**PhD** 95-0845	**RL/OC**	**01/01/95 12/31/95** Femtosecond Pump-Probe Spectroscopy System	**$25000.00**	**$15000.00**
Ryder, Jr. , Daniel **Chemical Engineering** **Tufts University, Medford, MA**	**PhD** 95-0846	**RL/ER**	**01/01/95 12/31/95** Synthesis and Properties B-Diketonate-Modified Heterobimetallic Alkoxides	**$25000.00**	**$0.00**
Zhang , Xi-Cheng **Physics** **Rensselaer Polytechnic Institu, Troy, NY**	**PhD** 95-0847	**RL/ER**	**01/01/95 12/31/95** Optoelectronic Study of Seniconductor Surfaces and Interfaces	**$25000.00**	**$0.00**
Drost-Hansen , Walter **Chemistry** **University of Miami, Coral Gables, FL**	**PhD** 95-0875	**WHMC/**	**01/01/95 12/31/95** Biochemical & Cell Physiological Aspects of Hyperthermia	**$25000.00**	**$8525.00**
Baginski , Michael **Electrical Engineering** **Auburn University, Auburn, AL**	**PhD** 95-0869	**WL/MN**	**01/01/95 12/31/95** An Investigation of the Heating and Temperature Distribution in Electrically Exc	**$24995.00**	**$10098.00**
Berdichevsky , Victor **Aerospace Engineering** **Wayne State University, Detroit, MI**	**PhD** 95-0849	**WL/FI**	**01/01/95 12/31/95** Micromechanics of Creep in Metals and Ceramics at High Temperature	**$25000.00**	**$0.00**
Buckner , Steven **Chemistry** **Colullmbus College, Columbus, GA**	**PhD** 95-0850	**WL/PO**	**01/01/95 12/31/95** Development of a Fluorescenece-Based Chemical Sensor for Simultaneous Oxygen Qua	**$24900.00**	**$8500.00**
Carroll , James **Electrical Engineering** **Clarkson University, Potsdam, NY**	**PhD** 95-0881	**WL/PO**	**01/01/95 12/31/95** Development of HIgh-Performance Active Dynamometer System for Machines and Drive	**$24944.00**	**$38964.00**
Choate , David **Mathematics** **Transylvania University, Lexington, KY**	**PhD** 95-0851	**WL/AA**	**01/01/95 12/31/95** SOLVING z(t)=ln{/A[cos(wlt)]+B[sin(w2t)]+C}	**$24993.00**	**$8637.00**
Clarson , Stephen **Materials Sci & Eng** **University of Cincinnati, Cincinnati, OH**	**PhD** 95-0852	**WL/ML**	**12/01/94 11/30/95** Synthesis, Processing and Characterization of Nonlinear Optical Polymer Thin Fil	**$25000.00**	**$15000.00**
Cone , Milton **Comp Science & Elec Eng** **Embry-Riddel Aeronautical Univ, Prescott, AZ**	**PhD** 95-0853	**WL/AA**	**01/01/95 12/31/95** An Investigation of Planning and Scheduling Algorithms for Sensor Management	**$25000.00**	**$11247.00**
Courter , Robert **Mechanical Engineering** **Louisiana State University, Baton Rouge, LA**	**PhD** 95-0854	**WL/MN**	**01/01/95 12/31/95** A Study to Determine Wave Gun Firing Cycles for High Performance Model Launches	**$25000.00**	**$3729.00**

1995 SREP SUB-CONTRACT DATA

Report Author Author's University	Author's Degree	Sponsoring Lab	Performance Period	Contract Amount	Univ. Cost Share
Dominic , Vincent **Electro Optics Program** **University of Dayton, Dayton, OH**	**PhD** 95-0868	**WL/ML**	**01/01/95 12/31/95** Characterization of Electro-Optic Polymers	**$25000.00**	**$12029.00**
Fadel , Georges **Dept of Mechanical Engnr** **Clemson University, Clemson, SC**	**PhD** 95-0855	**WL/MT**	**01/01/95 12/31/95** A Methodology for Affordability in the Design Process	**$25000.00**	**$8645.00**
Gould , Richard **Mechanical Engineering** **North Carolina State Univ, Raleigh, NC**	**PhD** 95-0856	**WL/PO**	**01/01/95 12/31/95** Data Reduction and Analysis for laser Doppler Velocimetry	**$24998.00**	**$9783.00**
Hardie , Russell **Electrical Engineering** **Univsity of Dayton, Dayton, OH**	**PhD** 95-0882	**WL/AA**	**01/01/95 12/31/95** Hyperspectral Target Identification Using Bomen Spectometer Data	**$24999.00**	**$7415.00**
Hodel , Alan **Electrical Engineering** **Auburn University, Auburn, AL**	**PhD** 95-0870	**WL/MN**	**01/01/95 12/31/95** Robust Falut Tolerant Control: Fault Detection and Classification	**$24990.00**	**$9291.00**
Janus , Jonathan **Aerospace Engineering** **Mississippi State University, Mississippi State,**	**PhD** 95-0871	**WL/MN**	**01/01/95 12/31/95** Multidimensional Algorithm Development & Analysis	**$25000.00**	**$7143.00**
Jasiuk , Iwona **Dept of Materials Science** **Michigan State University, East Lansing, MI**	**PhD** 95-0857	**WL/ML**	**01/01/95 12/31/95** Characterization of Interfaces in Metal-Matrix Composites	**$25000.00**	**$0.00**
Jouny , Ismail **Electrical Engineering** **Lafayette College, Easton, PA**	**PhD** 95-0880	**WL/AA**	**01/01/95 12/31/95** TSI Mitigation: A Mountaintop Database Study	**$24300.00**	**$5200.00**
Li , Jian **Electrical Engineering** **University of Florida, Gainesville, FL**	**PhD** 95-0859	**WL/AA**	**10/10/95 12/31/95** Comparative Study and Performance Analysis of High Resolution SAR Imaging Techni	**$25000.00**	**$4000.00**
Lin , Chun-Shin **Electrical Engineering** **University of Missouri-Columbi, Columbia, MO**	**PhD** 95-0883	**WL/FI**	**01/01/95 12/31/95** Prediction of Missile Trajectory	**$25000.00**	**$2057.00**
Lin , Paul **Mechanical Engineering** **Cleveland State University, Cleveland, OH**	**PhD** 95-0860	**WL/FI**	**01/01/95 12/31/95** Three Dimensional Deformation Comparison Between Bias and Radial Aircraft Tires	**$25000.00**	**$6886.00**
Liou , Juin **Electrical Engineering** **University of Central Florida, Orlando, FL**	**PhD** 95-0876	**WL/EL**	**01/01/95 12/31/95** Investigation of AlGaAs/GaAs Heterojunction Bipolar Transister Reliability Based	**$25000.00**	**$11040.00**
Nandhakumar , Nagaraj **Electrical Engineering** **University of Virginia, Charlottesville, VA**	**PhD** 95-0861	**WL/AA**	**01/01/95 12/31/95** Thermophysical Invariants fro, LWIR Imagery for ATR	**$24979.00**	**$4500.00**
Pasala , Krishna **Dept of Electrical Engr** **University of Dayton, Dayton, OH**	**PhD** 95-0879	**WL/AA**	**01/01/95 12/31/95** Effect of Electromagmetic Enviornment on Array Signal Processing	**$25000.00**	**$1078.00**
Perkowski , Marek **Dept of Electrical Engnr** **Portland State University, Portland, OR**	**PhD** 95-0878	**WL/AA**	**01/01/95 09/15/95** Functional Decomposition of Binary,Multiple-Valued, & Fuzzy Logic	**$24947.00**	**$18319.00**

1995 SREP SUB-CONTRACT DATA

Report Author Author's University	Author's Degree	Sponsoring Lab	Performance Period	Contract Amount	Univ. Cost Share
Reeves , Stanley **Dept of Electrical Engnr** **Auburn University, Auburn, AL**	**PhD** 95-0862	**WL/MN**	**01/01/95 12/31/95** Superresolution of Passive Millimeter-Wave Imaging	**$25000.00**	**$0.00**
Rule , William **Engineering Mechanics** **University of Alabama, Tuscaloosa, AL**	**PhD** 95-0872	**WL/MN**	**01/01/95 12/31/95** Development of a Penetrator Optimizer	**$24968.00**	**$14576.00**
Schauer , John **Mech & Aerosp Eng** **University of Dayton, Dayton, OH**	**PhD** 95-0873	**WL/PO**	**11/01/94 11/30/95** Heat Transfer for Turbine Blade Film Cooling with Free Stream Turbulence - Measu	**$25000.00**	**$7428.00**
Schwartz , Carla **Electrical Engineering** **University of Florida, Gainesville, FL**	**PhD** 95-0863	**WL/FI**	**01/01/95 12/31/95** Neural Network Identification and Control in Metal Forging	**$25000.00**	**$0.00**
Simon , Terrence **Dept of Mechanical Engineering** **University of Minnesota, Minneapolis, MN**	**PhD** 95-0864	**WL/PO**	**01/01/95 12/31/95** Documentation of Separating and Separated Boundary Layer Flow, for Application	**$24966.00**	**$3996.00**
Skowronski , Marek **Solid State Physics** **Carnegie Melon University, Pittsburgh, PA**	**PhD** 95-0865	**WL/EL**	**01/01/95 12/31/95** Transmission Electron Microscopy of Semiconductor Heterojunctions	**$25000.00**	**$6829.00**
Thirunarayan , Krishnaprasad **Computer Science** **Wright State University, Dayton, OH**	**PhD** 95-0866	**WL/EL**	**01/01/95 12/31/95** VHDL-93 Parser in SWI-PROLOG: A Basis for Design Query System	**$25000.00**	**$2816.00**
Trelease , Robert **Dept of Anatomy & Cell Bi** **University of California, Los Angeles, CA**	**PhD** 95-0867	**WL/ML**	**12/01/94 12/01/95** Development of Qualitative Process Control Discovery Systems for Polymar Composi	**$25000.00**	**$0.00**
Tsai , Chi-Tay **Engineering Mechanics** **Florida Atlantic University, Boca Raton, FL**	**PhD** 95-0874	**WL/MN**	**01/01/95 12/31/95** Improved Algorithm Development of Massively Parallel Epic Hydrocode in Cray T3D	**$24980.00**	**$0.00**
Lewis , John **Materials Science Engrng** **University of Kentucky, Lexington, KY**	**MS** 95-0858	**WL/MN**	**01/01/95 12/31/95** The Characterization of the Mechanical Properties of Materials in a Biaxial Stre	**$25000.00**	**$13833.00**

APPENDIX 1:

SAMPLE SREP SUBCONTRACT

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
1995 SUMMER RESEARCH EXTENSION PROGRAM
SUBCONTRACT 95-0837

BETWEEN

Research & Development Laboratories
5800 Uplander Way
Culver City, CA 90230-6608

AND

Florida Atlantic University
Department of Electrical Engineering
Boca Raton, FL 33431

REFERENCE: Summer Research Extension Program Proposal 95-0837
Start Date: 01-01-95 End Date: 12-31-95
Proposal Amount: $25,000.00

(1) **PRINCIPAL INVESTIGATOR:** Dr. Valentine A. Aalo
Department of Electrical Engineering
Florida Atlantic University
Boca Raton, FL 33431

(2) **UNITED STATES AFOSR CONTRACT NUMBER:** F49620-93-C-0063

(3) **CATALOG OF FEDERAL DOMESTIC ASSISTANCE NUMBER (CFDA):12.800**
PROJECT TITLE: AIR FORCE DEFENSE RESEARCH SOURCES PROGRAM

(4) **ATTACHMENT 1 REPORT OF INVENTIONS AND SUBCONTRAACT**
2 CONTRACT CLAUSES
3 FINAL REPORT INSTRUCTIONS

*****SIGN SREP SUBCONTRACT AND RETURN TO RDL*****

1. BACKGROUND: Research & Development Laboratories (RDL) is under contract (F49620-93-C-0063) to the United States Air Force to administer the Summer Research Program (SRP), sponsored by the Air Force Office of Scientific Research (AFOSR), Bolling Air Force Base, D.C. Under the SRP, a selected number of college faculty members and graduate students spend part of the summer conducting research in Air Force laboratories. After completion of the summer tour participants may submit, through their home institutions, proposals for follow-on research. The follow-on research is known as the Summer Research Extension Program (SREP). Approximately 61 SREP proposals annually will be selected by the Air Force for funding of up to $25,000; shared funding by the academic institution is encouraged. SREP efforts selected for funding are administered by RDL through subcontracts with the institutions. This subcontract represents an agreement between RDL and the institution herein designated in Section 5 below.

2. RDL PAYMENTS: RDL will provide the following payments to SREP institutions:
   - 80 percent of the negotiated SREP dollar amount at the start of the SREP research period.
   - The remainder of the funds within 30 days after receipt at RDL of the acceptable written final report for the SREP research.

3. INSTITUTION'S RESPONSIBILITIES: As a subcontractor to RDL, the institution designated on the title page will:

a. Assure that the research performed and the resources utilized adhere to those defined in the SREP proposal.

b. Provide the level and amounts of institutional support specified in the SREP proposal..

c. Notify RDL as soon as possible, but not later than 30 days, of any changes in 3a or 3b above, or any change to the assignment or amount of participation of the Principal Investigator designated on the title page.

d. Assure that the research is completed and the final report is delivered to RDL not later than twelve months from the effective date of this subcontract, but no later than December 31, 1998. The effective date of the subcontract is one week after the date that the institution's contracting representative signs this subcontract, but no later than January 15, 1998.

e. Assure that the final report is submitted in accordance with Attachment 3.

f. Agree that any release of information relating to this subcontract (news releases, articles, manuscripts, brochures, advertisements, still and motion pictures, speeches, trade associations meetings, symposia, etc.) will include a statement that the project or effort depicted was or is sponsored by: Air Force Office of Scientific Research, Bolling AFB, D.C.

g. Notify RDL of inventions or patents claimed as the result of this research as specified in Attachment 1.

h. RDL is required by the prime contract to flow down patent rights and technical data requirements to this subcontract. Attachment 2 to this subcontract

contains a list of contract clauses incorporated by reference in the prime contract.

4. All notices to RDL shall be addressed to:

RDL AFOSR Program Office
5800 Uplander Way
Culver City, CA 90230-6609

5. By their signatures below, the parties agree to provisions of this subcontract.

Abe Sopher
RDL Contracts Manager

Signature of Institution Contracting Official

Typed/Printed Name

Date

Title

Institution

Date/Phone

ATTACHMENT 2
CONTRACT CLAUSES

This contract incorporates by reference the following clauses of the Federal Acquisition Regulations (FAR), with the same force and effect as if they were given in full text. Upon request, the Contracting Officer or RDL will make their full text available (FAR 52.252-2).

FAR CLAUSES	TITLE AND DATE
52.202-1	DEFINITIONS
52.203-3	GRATUITIES
52.203-5	COVENANT AGAINST CONTINGENT FEES
52.203-6	RESTRICTIONS ON SUBCONTRACTOR SALES TO THE GOVERNMENT
52.203-7	ANTI-KICKBACK PROCEDURES
52.203-8	CANCELLATION, RECISSION, AND RECOVERY OF FUNDS FOR ILLEGAL OR IMPROPER ACTIVITY
52.203-10	PRICE OR FEE ADJUSTMENT FOR ILLEGAL OR IMPROPER ACTIVITY
52.203-12	LIMITATION ON PAYMENTS TO INFLUENCE CERTAIN FEDERAL TRANSACTIONS
52.204-2	SECURITY REQUIREMENTS
52.209-6	PROTECTING THE GOVERNMENT'S INTEREST WHEN SUBCONTRACTING WITH CONTRACTORS DEBARRED, SUSPENDED, OR PROPOSED FOR DEBARMENT
52.212-8	DEFENSE PRIORITY AND ALLOCATION REQUIREMENTS
52.215-2	AUDIT AND RECORDS - NEGOTIATION
52.215-10	PRICE REDUCTION FOR DEFECTIVE COST OR PRICING DATA

52.215-12	SUBCONTRACTOR COST OR PRICING DATA
52.215-14	INTEGRITY OF UNIT PRICES
52.215-8	ORDER OF PRECEDENCE
52.215.18	REVERSION OR ADJUSTMENT OF PLANS FOR POSTRETIREMENT BENEFITS OTHER THAN PENSIONS
52.222-3	CONVICT LABOR
52.222-26	EQUAL OPPORTUNITY
52.222-35	AFFIRMATIVE ACTION FOR SPECIAL DISABLED AND VIETNAM ERA VETERANS
52.222-36	AFFIRMATIVE ACTION FOR HANDICAPPED WORKERS
52.222-37	EMPLOYMENT REPORTS ON SPECIAL DISABLED VETERAN AND VETERANS OF THE VIETNAM ERA
52.223-2	CLEAN AIR AND WATER
52.223-6	DRUG-FREE WORKPLACE
52.224-1	PRIVACY ACT NOTIFICATION
52.224-2	PRIVACY ACT
52.225-13	RESTRICTIONS ON CONTRACTING WITH SANCTIONED PERSONS
52.227-1	ALT. I - AUTHORIZATION AND CONSENT
52.227-2	NOTICE AND ASSISTANCE REGARDING PATIENT AND COPYRIGHT INFRINGEMENT

52.227-10	FILING OF PATENT APPLICATIONS - CLASSIFIED SUBJECT MATTER
52.227-11	PATENT RIGHTS - RETENTION BY THE CONTRACTOR (SHORT FORM)
52.228-7	INSURANCE - LIABILITY TO THIRD PERSONS
52.230-5	COST ACCOUNTING STANDARDS - EDUCATIONAL INSTRUCTIONS
52.232-23	ALT. I - ASSIGNMENT OF CLAIMS
52.233-1	DISPUTES
52.233-3	ALT. I - PROTEST AFTER AWARD
52.237-3	CONTINUITY OF SERVICES
52.246-25	LIMITATION OF LIABILITY - SERVICES
52.247-63	PREFERENCE FOR U.S. - FLAG AIR CARRIERS
52.249-5	TERMINATION FOR CONVENIENCE OF THE GOVERNMENT (EDUCATIONAL AND OTHER NONPROFIT INSTITUTIONS)
52.249-14	EXCUSABLE DELAYS
52.251-1	GOVERNMENT SUPPLY SOURCES

DOD FAR CLAUSES	DESCRIPTION
252.203-7001	SPECIAL PROHIBITION ON EMPLOYMENT
252.215-7000	PRICING ADJUSTMENTS
252.233-7004	DRUG FREE WORKPLACE (APPLIES TO SUBCONTRACTS WHERE THERE IS ACCESS TO CLASSIFIED INFORMATION)
252.225-7001	BUY AMERICAN ACT AND BALANCE OF PAYMENTS PROGRAM
252.225-7002	QUALIFYING COUNTRY SOURCES AS SUBCONTRACTS
252.227-7013	RIGHTS IN TECHNICAL DATA - NONCOMMERCIAL ITEMS
252.227-7030	TECHNICAL DATA - WITHOLDING PAYMENT
252.227-7037	VALIDATION OF RESTRICTIVE MARKINGS ON TECHNICAL DATA
252.231-7000	SUPPLEMENTAL COST PRINCIPLES
252.232-7006	REDUCTIONS OR SUSPENSION OF CONTRACT PAYMENTS UPON FINDING OF FRAUD

APPENDIX 2:

SAMPLE TECHNICAL EVALUATION FORM

SUMMER RESEARCH EXTENSION PROGRAMTECHNICAL EVALUATION

SREP NO: 95-0811
SREP PRINCIPAL INVESTIGATOR: Dr. Michael Burke

Circle the rating level number, 1 (low) through 5 (high), you feel best evaluate each statement and return thecompleted form by mail to:

RDL
Attn: 1995 SREP Tech Evals
5800 Uplander Way
Culver City, CA 90230-6608
(310) 216-5940 or (800) 677-1363

1.	This SREP report has a high level of technical merit.	1 2 3 4 5
2.	The SREP program is important to accomplishing the lab's mission.	1 2 3 4 5
3.	This SREP report accomplished what the associate's proposal promised.	1 2 3 4 5
4.	This SREP report addresses area(s) important to the USAF.	1 2 3 4 5
5.	The USAF should continue to pursue the research in this SREP report.	1 2 3 4 5
6.	The USAF should maintain research relationships with this SREP associate.	1 2 3 4 5
7.	The money spent on this SREP effort was well worth it.	1 2 3 4 5
8.	This SREP report is well organized and well written.	1 2 3 4 5
9.	I'll be eager to be a focal point for summer and SREP associates in the future.	1 2 3 4 5
10.	The one-year period for complete SREP research is about right.	1 2 3 4 5

---

11. If you could change any one thing about the SREP program, what would you change.

12.What would you definitely NOT change about the SREP program?

USE THE BACK FOR ANY ADDITIONAL COMMENTS.

Laboratory: Armstrong Laboratory
Lab Focal Point: Linda Sawin Office Symbol: AL/HRMI
Phone: (210) 536-3876

Michael Baginski report unavailable at time of publication.

MICROMECHANICS OF DIFFUSIONAL CREEP

Dr. Victor L. Berdichevsky
Professor
Department of Mechanical Engineering

Wayne State University
Detroit, MI 48202

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

Wayne State University

December 1995

Micromechanics of Diffusional Creep

V. Berdichevsky
Mechanical Engineering Department
Wayne State University, Detroit, MI 48202

Abstract

In polycrystal materials at high temperatures and low stresses, creep occurs mostly due to diffusion vacancies through the grain bodies and over the grain boundaries. A continuum theory of vacancy motion is considered to analyze diffusional creep on microscopical level. A linear version of such theory was formulated by Nabarro, Herring, Coble and Lifshitz. This theory is revised here from the perspectives of continuum mechanics and presented in a thermodynamically consistent nonlinear form. A certain difficulty which has been overcome in this endeavor is the absence of Lagrangian coordinates in diffusional creep, the major building block of any theory in continuum mechanics.

A linearized version of the theory is studied for the case of bulk diffusion. The derivation of macro-constitutive equations is considered using the homogenization technique. It is shown that macroequations are nonlocal in time and nonlocality is essential in primary creep. For secondary creep polycrystals behaves as viscoelastic body. For secondary creep, a variational principle is found which determines microfields and macromoduli in stress-strain rate constitutive equations. Two dimensional honeycomb microstructure and single crystal deformation are studied numerically by a finite element method.

Micromechanics of Diffusional Creep

V. Berdichevsky

1 Introduction

Predictions of mechanical behavior of solids can be roughly classified as short-term and long-term predictions. In short-term prediction, the behavior can be elastic or plastic, depending on the level of stress. For sufficiently low stresses, solids behave elastically. However, over long time periods, even for very low stresses, solids develop irreversible deformations. This phenomenon is called creep.

Three points are worthy stressing in discussion of creep. First, everything creeps. Actually, solids creep even at zero external load, due to the fact that practically no polycrystalline body is in thermodynamic equilibrium. Second, creep is an energy driven phenomenon. Materials creep in order to decrease energy (or other thermodynamical potential, depending on the external conditions). Energy of a polycrystal, for example, can be decreased by moving grain boundaries. This occurs in reality, but very slowly, by means of thermodynamic fluctuations. The rate of change is magnified significantly by elevating the temperature and/or applying an external load. Third, mechanisms of creep are stress-temperature dependent.

Two major creep mechanisms are movement of dislocations and diffusion of vacancies. A typical deformation mechanism map is shown in the stress-temperature plane in Fig. 1. Above the curve (high stresses) the dominating mechanism is dislocation motion. Below the curve (low stresses) deformations occur due to diffusion of vacancies. It is believed that at low temperatures, vacancies move mostly over the grain boundaries (Coble creep), while for high temperatures, motion vacancies through the lattice dominate (Herring-Nabarro creep or bulk diffusional creep). Diffusional creep is the leading phenomenon in many technological processes at high temperatures. Superplasticity, sintering, and void formation, occur mostly due to diffusional creep. In this paper we focus on a thermodynamically consistent theory of diffusional creep. The foundations of this theory were laid down by Nabarro [1], Herring [2], Coble [3], and Lifshitz [4]. Extensive reviews of various aspects of creep theory can be found in [5] – [24].

Mechanism of plastic deformation caused by bulk diffusional creep can be viewed as follows. Let a monocrystal be loaded by an external force (Fig. 2). Consider the right edge of the monocrystal. A surface external force might be thought as a set of forces applied to each atom of the very right column of atoms (Fig. 2a). Due to thermal fluctuations some of the atoms of this column can jump to a new equilibrium position (Fig. 2b). Then the next atoms may jump into the vacant places, and we see that vacancies entered into the crystal body. Then vacancies can migrate inside the body and leave the body at the free surface (Fig. 2d).

Motion of vacancies is accompanied by the corresponding motion of material in the opposite direction. The moved material is shaded in Fig. 2e. Since motion of vacancies is dispersed over material one observe an effective elongation of the specimen (Fig. 2f).

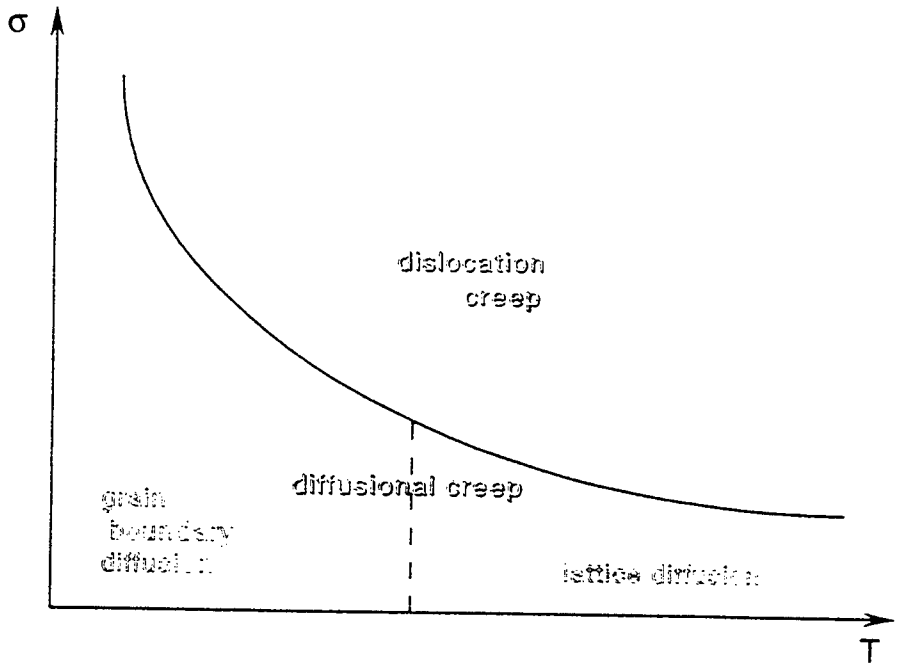


Figure 1: Deformation mechanism map.

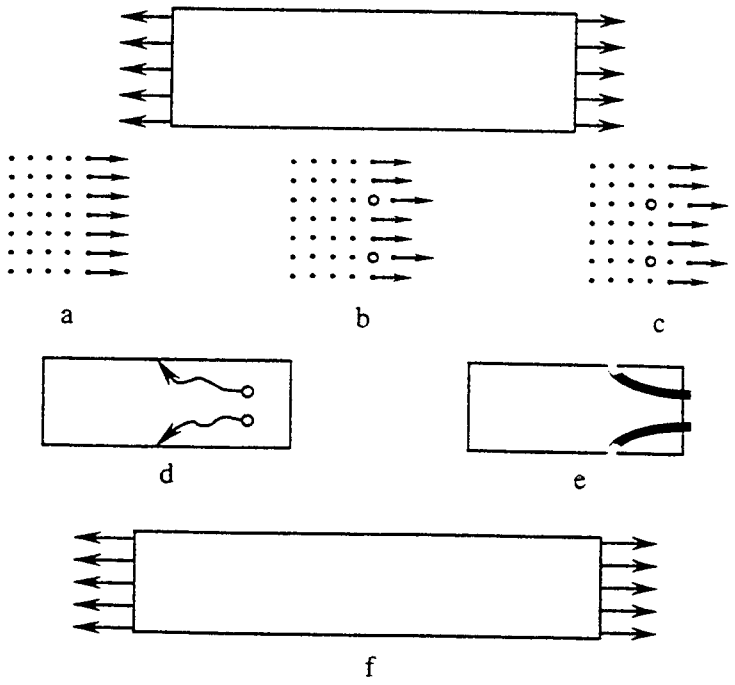


Figure 2: Mechanism of plastic deformation caused by bulk diffusional creep.

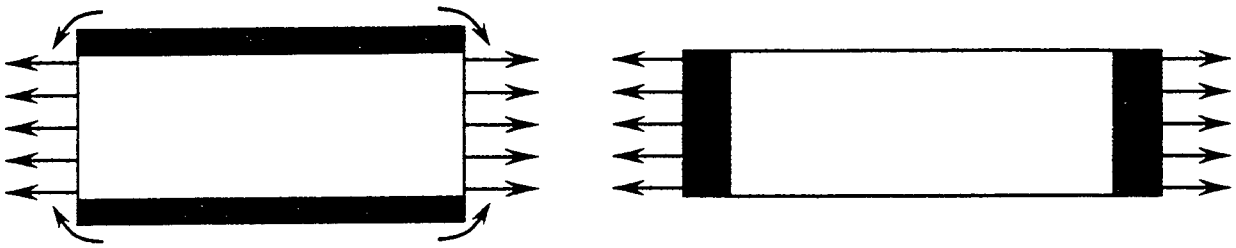


Figure 3: Boundary diffusion.

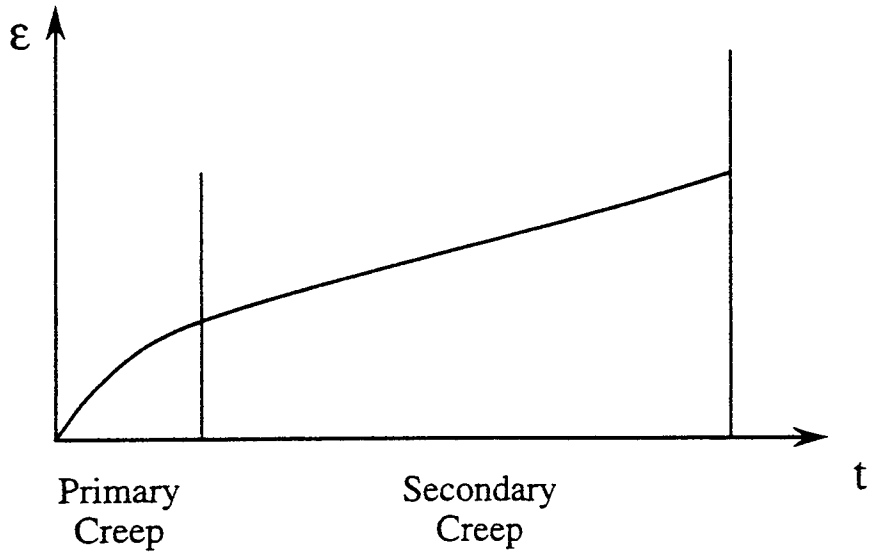


Figure 4: Typical creep strain-time dependence.

In the case of boundary diffusion material flows over the boundaries from unloaded to loaded pieces of the boundary, and that yields some macroscopical plastic deformation. This process is shown schematically in Fig. 3; the moving material is shadowed.

A typical strain-time dependence for constant stresses is shown in Fig. 4. It is clearly observed two different regimes of the plastic flow. At the first moment strains grow fast, then the strain rate decays until it approaches some limit value. These two regimes are referred to as primary and secondary creep.

The aim of this paper is to construct the microequations of diffusional creep in the framework of continuum mechanics and develop a homogenization procedure to derive macroequations of creep.

There are a number of reasons to pursue these goals. First of all, a phenomenological approach to derivation of macroequations for creep provides too many options. Realization of our program may help to choose the right one.

Second, the problem seems challenging from the perspectives of continuum mechanics. Looking at the sketch of boundary diffusional creep shown in Fig. 5, one may observe that the basic notion of continuum

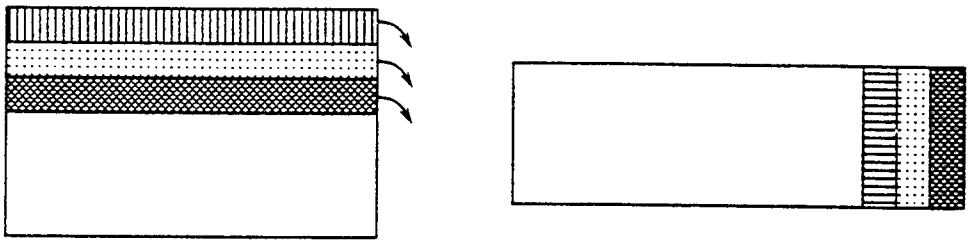


Figure 5: Mixing by boundary diffusion.

mechanic, Lagrangian coordinates, cannot be used in this case. Really, material points which were on the grain boundary moves into the grain body: this is in clear contradiction to the main postulate of continuum mechanics [25, 26] on the existence of a diffeomorphism between the deformed and undeformed states, and, as a consequence, to the existence of Lagrangian coordinates. If a continuum deformation were a diffeomorphism, the material points, which are on the boundary, stay on the boundary forever. Lagrangian coordinates are used in continuum mechanics, for example, in the definition of velocity: one has to say "velocity of what" is introduced. We suggest a way to overcome this difficulty.

Third, a theory of diffusional creep must be a building block for the theory of dislocational climb which is, at the moment, in a premature stage.

The contents of the paper is as follows. The contents of the paper is as follows. Section 2 describes the main feature of the model for bulk difffusional creep, which is the existence of plastic displacement field. This is an unusual situation in plasticity. The general kinematic relations for the bulk diffusion and surface diffusion is given in Section 3. In the central Section 4 the closed system of equations of diffusional creep is developed from thermodynamical considerations. The linear version of the general theory is presented in Section 5. In the rest of the paper linear theory of bulk diffusional creep is studied aiming to derive macroscopic laws for grain structure starting from micromodel, which is referred to as homogenization problem. In the Section 6 the formulation of homogenization problem is given for a particular case of periodic grain structure. The theorem of uniqueness of the solution is proven, which is an evidence of correctness of the basic equations. In Section 7 the general type of macroscopic constitutive relations is established. Secondary creep is considered in Section 8. It is proven ,that under constant loads the transient solution tends to a steady state solution,and the closed system of equations is found which allows one to find the macrocharacteristics of the secondary creep without "tracing" the transient solution. Numerical example of solving this system is presented in Section 9. Dimension analysis of the equations and numerical modeling

of the transient process are discussed in Section 10.

2 Micromechanics of Bulk Diffusional Creep: A Logical Skeleton of the Theory.

Logic structure of the theory is especially simple in the case of bulk diffusional creep, and before going to a detailed discussion, we outline it briefly.

In creep theory, new required functions appear: plastic strains $\varepsilon_{ij}^{(p)}$. The key point of the bulk diffusional creep is that plastic strains are compatible: there exist plastic displacement $w_i^{(p)}$ such that (in linear case)

$$\varepsilon_{ij}^{(p)} = \frac{1}{2}\left(\frac{\partial w_i^{(p)}}{\partial x_j} + \frac{\partial w_j^{(p)}}{\partial x_i}\right) \tag{2.1}$$

Here and in the following small Latin indices run values 1, 2, 3 and correspond to projections on the Cartesian axis of the observer frame; x_i are the observer coordinates.

The consistency of plastic deformation is a pure kinematical property. This is an assumption which aims to model the process of deformation shown schematically in Fig. 2e.

In contrast to a general creep theory where six additional equations are to be given for six unknown functions $\varepsilon_{ij}^{(p)}$, in bulk diffusional creep one has to give only three additional equations for $w_i^{(p)}$.

It is clear that the plastic rate $\dot{w}_i^{(p)}$ should be related to vacancy motion. Some kinematical and thermodynamical consideration shows that the corresponding relation (in its simplest version) is

$$\dot{w}_i^{(p)} = D\frac{\partial c}{\partial x_i} \tag{2.2}$$

where c is vacancy concentration, dot denotes time derivative and D is the diffusion coefficient.

Equation (2.2) reduces the number of closing equations to one: an equation for vacancy concentration c. This last equation is the diffusion equation for c

$$\frac{\partial c}{\partial t} = D\Delta c \tag{2.3}$$

Equations (2.1) - (2.3) should be complemented by usual equations of elasticity and provided with the boundary conditions.

Now we proceed to a detailed consideration.

3 Continuum Kinematics.

We are going to model in terms of continuum mechanics the following physical phenomenon. If an external load is applied to an atomic lattice containing a cloud of vacancies, it appears a direction of preferable migration of vacancies. Motion of vacancies causes the motion of atoms in opposite direction. The motion

of atoms is perceived by an observer as an irreversible plastic deformation of the material. Our first step is to establish a kinematical relation which relates motion of vacancies and motion of the material.

We model motion of vacancies and material by two continua with velocities u_i and v_i correspondingly. We assume that vacancies are not created inside the material and can come only from the boundary. Then, as we shall argue,

$$v_i^{(e)} = (1-c)\, v_i + c u_i \tag{3.1}$$

where $v_i^{(e)}$ is an "elastic" velocity. If the elastic velocity $V_i^{(e)}$ is zero, the relation (3.1) expresses velocity of material (atoms) v_i in terms of velocity of vacancies u_i and vacancy concentration c.

Usually, vacancy concentration is negligible in comparison to the unity. Nevertheless, we keep factor $(1-c)$ until the final calculations in order to underline the physical origination of various terms.

Equation (3.1) is a postulate which is motivated by the following reasons.

Consider a piece of crystal lattice, a "representative volume of material," and think of v_i as the average velocity of all the atoms of this piece

$$v_i = \frac{1}{N_a} \sum_{\alpha} v_i^{\alpha} \tag{3.2}$$

where N_a is the number of atoms, v_i^{α} is the velocity of the α -th atom, and summation is taken over all of atoms of the piece. Similarly, velocity of vacancies is the average value of velocities of all vacancies:

$$u_i = \frac{1}{N_v} \sum_{\alpha} u_i^{\alpha} \tag{3.3}$$

Here N_v is the number of vacancies and u_i^{α} is the velocity of the α-th vacancy. Volume average velocity $\tilde{v}_i$ is, by definition,

$$\tilde{v}_i = \frac{1}{N} \left(\sum_{\alpha} v_i^{\alpha} + \sum_{\alpha} u_i^{\alpha} \right) \tag{3.4}$$

where N is the total number of lattice sites

$$N = N_a + N_v \tag{3.5}$$

It follows from (3.2) - (3.5) that

$$\tilde{v}_i = (1-c)\, v_i + c u_i \tag{3.6}$$

where the volume fraction of vacancies c is, by definition,

$$c = \frac{N_v}{N} \tag{3.7}$$

Relation (3.6) holds for mixture of any two substances. Now we must express in some way the fact that we are dealing with diffusion of vacancies. We may assume that in the process of position exchange between an atom and a vacancy the velocities of the atom and the vacancy are equal in magnitude and opposite in sign. Therefore, in accordance with (3.4), $\tilde{v}_i = 0$. Then (3.6) links the velocities of atoms and vacancies. It is clear that atoms and vacancies might have a common additional velocity. Then $\tilde{v}_i$ is not zero and equal to this additional velocity. The additional velocity is not related to the process of vacancy diffusion and,

hence, the irreversible deformation. We identify this velocity with "elastic" velocity and denote it by $v_i^{(e)}$. Then (3.6) takes the form (3.1).

Note that the term "elastic" velocity is not quite exact. If one defines elasticity as part of deformation which disappears after unloading then velocity $v_i^{(e)}$ might have a contribution form a plastic rigid motion, a motion of the monocrystal after unloading as a rigid body. However, we take some liberty in terminology to simplify the notations and use the term elastic velocity for the sum of the "real" elastic velocity and plastic velocity of rigid motion.

The flux of vacancies relative to material J_i is given by

$$J_i = c\left(u_i - v_i^{(e)}\right) \tag{3.8}$$

In accordance with (3.1) and (3.8) material velocity v_i can be expressed in terms of elastic velocity and vacancy flux as

$$v_i = v_i^{(e)} - \frac{1}{1-c} J_i \tag{3.9}$$

This is a key kinematical relation.

Since vacancies can be generated only on the boundary, vacancy concentration obeys the conservation law

$$\frac{\partial c}{\partial t} + \frac{\partial c u_i}{\partial x_i} = 0 \tag{3.10}$$

Equations (3.1), (3.8) - (3.10) form the basic kinematical relations of bulk diffusional creep. Now we are going to incorporate in this picture the surface diffusion.

Denote by V_0 the initial state of monocrystal with zero stresses. Let V be the deformed state of the monocrystal. Region V depends on time. We refer both states to some Cartesian coordinates x^i . Besides, we introduce in the region V_0 some coordinates curvilinear in general, ξ^a, which, in a "usual" situation, play the role of Lagrangian coordinates. Indices a, b, c run values 1, 2, 3 and correspond to projections on the axis ξ^a. There is one-to-one correspondence between observer's coordinates x_i and coordinates ξ^a

$$x^i = x \leq circ^i\left(\xi^a\right) \tag{3.11}$$

Without loss of generality mapping (3.11) may be identical, however, it is convenient to leave it without specifications because coordinates x^i and ξ^a obey to different groups of transformations [27]. This is why we use another group of Latin indices, a, b, c, in the notation for Lagrangian coordinates.

At each moment of time t, there is mapping of region V_o to region V

$$x^i = x^i\left(\xi^a, t\right) \tag{3.12}$$

If this mapping is a diffeomorphism then ξ^a are Lagrangian coordinates. In this case, if a point ξ^a lies on the boundary ∂V_o of the region V_o, its image is on the boundary ∂V of the region V for all instants t. Velocity is defined as velocity of the particleξ^a: $v^i = \partial x^i\left(\xi^a, t\right)/\partial t$. This is a classical kinematical scheme of continuum mechanics (see, for example, [25, 26, 27]). As one sees from Fig. 5, this is not the case for boundary diffusion creep, and we have to change the kinematical scheme. We introduce as a "primary" kinematical object

the region V which is changed in time. In this region two velocity fields, material velocity v^i and vacancy velocity u^i are defined. If mapping (3.12) were a diffeomorphism and $v^i = {\partial x^i(t,\xi^\alpha)}/{\partial t}$, then the normal velocity of the boundary surface ∂V is equal to $v^i n_i$. In the case of boundary diffusion these velocities are different. We denote the difference by u:

$$v_{boundary} = v^i n_i + u \tag{3.13}$$

Velocity u is caused by the material flow over the boundary. It appears as independent kinematical characteristics. However, "more fundamental" characteristics might be introduced as primary characteristics of boundary diffusion: boundary mass flux J^α. Boundary mass flux is defined in the following way. Mass of material is conserved in the boundary flow, therefore, a law of conservation of mass should take place. Denote by J^α the vector of mass flow on the surface. Greek indices run values 1, 2 and correspond to projections on the boundary surface. If γ is a curve on the boundary and ν_α is the unit normal vector to γ at a point P, then the scalar $J^\alpha \nu_\alpha \Delta s$ means the mass flow through the arc of γ of the length Δs at the point P. Let ρ be the mass density of material. Then the law of conservation of mass has the form

$$\rho u = \nabla_\alpha J^\alpha \tag{3.14}$$

where ∇_α is the covariant derivative on the surface ∂V.

Mass density obeys also the law of conservation of mass inside the region V

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho v^i}{\partial x^i} = 0 \tag{3.15}$$

Equations (3.14), (3.15) and (3.13) provide the conservation of mass in volume V

$$\frac{\partial}{\partial t}\int\limits_{V(t)} \rho d^3x = \int\limits_V \frac{\partial \rho}{\partial t} d^3x + \int\limits_{\partial V} \rho v_{boundary} d^2x = -\int\limits_V \frac{\partial \rho v^i}{\partial x^i} d^3x + \int\limits_{\partial V} \rho v_{boundary} d^2x$$

$$= \int\limits_{\partial V} \rho \left(v_{boundary} - v^i n_i\right) d^2x = \int\limits_{\partial V} \rho u d^2x = \int\limits_{\partial V} \nabla_\alpha J^\alpha d^2x = 0$$

It is natural to consider J^α as primary characteristics of boundary diffusion, then velocity u is determined by equation (3.14).

Now we come to the point where we have to introduce displacements. It is natural to define a field of elastic displacements $w_i^{(e)}(t,x)$ which has the domain $V(t)$. Vector $w_i^{(e)}(t,x)$ means the displacement of a monocrystal from the thought unloaded state to the actual state $V(t)$. If there are no plastic strains, the displacement $w_i^{(e)}(t,x)$ relates to velocity by the formula

$$\frac{\partial w_i^{(e)}}{\partial t} + v_k^{(e)} \frac{\partial w_i^{(e)}}{\partial x_k} = v_i^{(e)} \tag{3.16}$$

Equation (3.16) can be rewritten as

$$\left(\delta_i^k - \frac{\partial w_i^{(e)}}{\partial x_k}\right) v_k^{(e)} = \frac{\partial w_i^{(e)}}{\partial t} \tag{3.17}$$

The latter relation can be considered as a system of linear equations with respect to velocity $v_k^{(e)}$, if the displacement field is known. We keep formulas (3.16), (3.17) as the definition of the vector of elastic displacements if velocity $v_k^{(e)}$ is considered as a primary quantity. Remember that, by our convention, plastic (and, hence, elastic) deformations are consistent for diffusional creep and vector of elastic displacements exists.

4 Thermodynamics of Diffusional Creep.

We derive the basic equations of diffusional creep following the usual thermodynamical approach: we assume an expression for free energy of the material and construct the equations in a way to warrant the negativeness of the time derivatives of free energy.

Free energy F of a polycrystal has, by our assumption, an energy density per unit volume F:

$$\mathcal{F} = \int\limits_{V(t)} F d^3x \tag{4.1}$$

We accept that energy density F is a function of gradient of elastic displacement $w_{i,j}^{(e)}$ ($\equiv {\partial w_i^{(e)}}/{\partial x^i}$), vacancy concentration c and temperature T:

$$F = F\left(w_{i,j}^{(e)},\, c,\, T\right) \tag{4.2}$$

Temperature T is maintained constant.

Note that the assumption (4.2) taken together with the definition of elastic displacements (3.16) extracts a special class of models. For example, if elastic displacement is defined, instead of (3.16), by the formula

$$\frac{\partial w_i^{(e)}}{\partial t} + v_k \frac{\partial w_i^{(e)}}{\partial x_k} = v_i^{(e)} \tag{4.3}$$

which may have some motivations, we would arrive at the class of models, which differs from the considered one in nonlinear case.

Let us find time derivative of free energy. We assume first that region $V(t)$ is occupied by a monocrystal and all fields are smooth inside V. We have

$$\frac{d\mathcal{F}}{dt} = \int\limits_V \left(\frac{\partial F}{\partial w_{i,j}^{(e)}} \frac{\partial}{\partial x^j} w_{i,t}^{(e)} + \frac{\partial F}{\partial c} \frac{\partial c}{\partial t} \right) d^3x + \int\limits_{\partial V} \left(v^i n_i + u\right) d^2x \tag{4.4}$$

After substituting in (4.4) the expression for ${\partial c}/{\partial t}$ from (3.10) and integration by parts we obtain

$$\frac{d\mathcal{F}}{dt} = \int\limits_V \left[-\left(\frac{\partial}{\partial x_k} \frac{\partial F}{\partial w_{i,k}^{(e)}} \right) \left(\delta_{im} - w_{i,k}^{(e)}\right) v^{(e)m} + \left(c v^{(e)i} + J^i \right) \frac{\partial}{\partial x^i} \frac{\partial F}{\partial c} \right] d^3x +$$
$$\int\limits_{\partial V} \left[\frac{\partial F}{\partial w_{i,j}^{(e)}} n_j \left(\delta_{ik} - w_{i,k}^{(e)}\right) v^{(e)k} - \frac{\partial F}{\partial c} \left(c v^{(e)i} + J^i \right) n_i + F\left(v^i n_i + u\right) \right] d^2x \tag{4.5}$$

Here we expressed also $\partial w_i^{(e)}/\partial t$ in terms of elastic velocity from (3.17).

For further transformations we need an identity [27]

$$\left(\frac{\partial}{\partial x_k}\frac{\partial F}{\partial w_{i,k}^{(e)}}\right)\left(\delta_{im}-w_{i,m}^{(e)}\right)=\frac{\partial}{\partial x_k}\left(\frac{\partial F}{\partial w_{i,m}^{(e)}}\left(\delta_{im}-w_{i,m}^{(e)}\right)+F\delta_m^k\right)-\frac{\partial F}{\partial c}\frac{\partial c}{\partial x^m} \tag{4.6}$$

This identify can be checked by direct inspection. Using (4.6) and (3.9) we can rewrite (4.5) in the form

$$\frac{d\mathcal{F}}{dt}=\int\limits_V\left(-\frac{\partial\sigma^{km}}{\partial x^k}v_m^{(e)}+J^i\frac{\partial}{\partial x^i}\frac{\partial F}{\partial c}\right)d^3x+\int\limits_{\partial V}\left(\sigma^{ij}n_jv_i^{(e)}-\left(\frac{\partial F}{\partial c}+\frac{F}{1-c}\right)J^in_i+Fu\right)d^2x \tag{4.7}$$

Here we introduced a notation

$$\sigma_i^j=\frac{\partial F}{\partial w_{m,j}^{(e)}}\left(\delta_{mi}-w_{m,i}^{(e)}\right)+\left(F-c\frac{\partial F}{\partial c}\right)\delta_i^j \tag{4.8}$$

It is seen from this expression that σ_i^j have the sense of components of stress tensor.

Assume that J^i do not depend on $v_m^{(e)}$. Since σ_i^j do not depend on $v_m^{(e)}$ as well, and $v_m^{(e)}$ can be chosen arbitrary, (4.7) can comply with negativeness of dF/dt if and only if the equilibrium equations hold

$$\frac{\partial\sigma_i^j}{\partial x^j}=0 \tag{4.9}$$

The simplest expression for the vacancy flux which does not contradict to the negativeness of dF/dt is

$$J^i=-D^{ij}\frac{\partial}{\partial x^j}\frac{\partial F}{\partial c} \tag{4.10}$$

where D^{ij} is a positive tensor.

Consider now the boundary terms. Let V be a polycrystal. Denote by Σ the grain boundary surface. Then the surface terms in dF/dt take the form

$$\int\limits_\Sigma\left(\left[\sigma^{ij}v_i^{(e)}-\frac{\partial F}{\partial c}+\frac{F}{1-c}J^j\right]n_j+[Fu]\right)d^2x \tag{4.11}$$

where for any quantity A the symbol $[A]$ means the difference of A at two sides of the surface Σ.

Let us present the surface force $\sigma^{ij}n_j$ as a sum of normal force $\sigma_{nn}n^i$ $\left(\sigma_{nn}\equiv\sigma^{ij}n_in_j\right)$ and tangent traction. Similarly, $v_i^{(e)}$ is the sum of the normal velocity $v_n^{(e)}n_i$ $\left(v_n^{(e)}\equiv v_i^{(e)}n^i\right)$ and the tangent one. Then

$$\sigma^{ij}v_i^{(e)}n_j=\sigma^{\alpha j}v_\alpha^{(e)}n_j+\sigma_{nn}v_n^{(e)}$$

(Greek indices α, β, γ run values 1, 2 and correspond to projection on the tangent plane to Σ. Using also (3.9) we rewrite (4.11) in the form

$$\int\limits_\Sigma\left(\left[\sigma^{\alpha j}v_\alpha^{(e)}\right]n_j+[\sigma_{nn}(v_n+u)]+\left[\left(\frac{\sigma_{nn}-F}{1-c}-\frac{\partial F}{\partial c}\right)J^i\right]n_j-[(\sigma_{nn}-F)u]\,d^2x\right) \tag{4.12}$$

It is natural to require continuity of the total normal velocity of the adjacent grains

$$[v_n + u] = 0 \tag{4.13}$$

Since σ_{nn} (as well as other "generalized forces" in (4.12)) does not depend on velocity, it is necessary that σ_{nn} be continuous:

$$[\sigma_{nn}] = 0 \tag{4.14}$$

Normal vacancy flux $J^i n_i$ can be arbitrary, and vacancies on two sides of the grain boundary seem being produced with an independent rate. Therefore, it is natural to accept that the corresponding coefficient at $J^i n_i$ in (4.12) are zeros: at both sides of the boundary surface

$$\frac{\sigma_{nn} - F}{1 - c} - \frac{\partial F}{\partial c} = 0 \tag{4.15}$$

In accordance with (3.14), the last term in (4.12) can be written as

$$\int_{\sum} [(\sigma_{nn} - F)\, u]\, d^2x = \int_{\sum} \left[\frac{\sigma_{nn} - F}{\rho} \nabla_\alpha J^\alpha\right] d^2x = -\int_{\sum} \left[J^\alpha \nabla_\alpha \frac{\sigma_{nn} - F}{\rho}\right] d^2x \tag{4.16}$$

Here we integrated by part and dropped the term on the polycrystal boundary. Finally,

$$\frac{dF}{dt} = \int_V J^i \frac{\partial}{\partial x^i} \frac{\partial F}{\partial c} d^3x + \int_{\partial V} \left(\left[\sigma^{\alpha j} v_\alpha^{(e)}\right] n_j - \left[J^\alpha \nabla_\alpha \frac{\sigma_{nn} - F}{\rho}\right]\right) d^2x \tag{4.17}$$

There are different models which obey the negativeness of (4.17). The most plausible version is based on the assumptions that $\sigma^{\alpha j} n_j$ are continuous and surface fluxes of material J^α are independent on both sides of Σ. Then, neglecting reciprocal effects, one can put

$$\sigma^{\alpha j} n_j = -\mu^{\alpha\beta} \left[v_\beta^{(e)}\right] \quad \text{on } \sum \tag{4.18}$$

$$J^\alpha = d^{\alpha\beta} \nabla_\beta \left(\frac{\sigma_{nn} - F}{\rho}\right) \quad \text{on each side of } \sum \tag{4.19}$$

Note that, in contrast to σ_{nn}, energy density is not continuous on Σ, therefore material fluxes J^α are different on both sides of Σ. However, this a nonlinear effect.

The equations derived in this section close the system of equations of diffusional creep.

5 Linearized theory.

In the linear case the system of equations is simplified essentially. First of all, in this case one can neglect the changes of region V in the process of deformation. Second, kinematical relations take a simple form

$$v_i^{(e)} = \frac{\partial w_i^{(e)}}{\partial t} = v_i + c u_i \tag{5.1}$$

$$v_i = v_i^{(e)} - J_i \tag{5.2}$$

$$\frac{\partial c}{\partial t} + \frac{\partial J_i}{\partial x_i} = 0 \tag{5.3}$$

$$\rho u = \nabla_\alpha J^\alpha \tag{5.4}$$

Third, energy density is a quadratic function of elastic strains

$$\varepsilon_{ij}^{(e)} = 1/2 \left(\frac{\partial w_i^{(e)}}{\partial x^j} + \frac{\partial w_j^{(e)}}{\partial x^i} \right), \tag{5.5}$$

and deviation $s = c - c_o$ of vacancy concentration from its equilibrium value c_o (for brevity from now on the function s will be referred to as vacancy concentration)

$$F = \frac{1}{2} A^{ijke} \varepsilon_{ij}^{(e)} \varepsilon_{ke}^{(e)} + \frac{1}{2} A s^2 + \text{function of } T \tag{5.6}$$

Here A^{ijke} are Young moduli while A is an additional material constants. From some statistical reasoning [5]

$$A = \frac{\rho_o T}{m c_o} \tag{5.7}$$

where m is the mass of one atom, ρ_o is the mass density of an ideal lattice. In (5.6) we neglect an interaction term $A^{ij} \varepsilon_{ij}^{(e)} (c - c_o)$.

In accordance with (4.8), stress tensor σ^{ij} in linear theory has the form

$$\sigma^{ij} = A^{ijke} \varepsilon_{ke}^{(e)} \tag{5.8}$$

It obeys the equilibrium equations

$$\frac{\partial \sigma^{ij}}{\partial x^j} = 0 \tag{5.9}$$

Vacancy flux J^i is given by (4.10)

$$J^i = -A D^{ij} \frac{\partial s}{\partial x^j} \tag{5.10}$$

Therefore, equation (5.3) transforms to usual diffusion equation

$$\frac{\partial s}{\partial t} = \frac{\partial}{\partial x^i} \left(A D^{ij} \frac{\partial s}{\partial x^j} \right) \tag{5.11}$$

We assume that diffusion constants obey the positive definiteness condition

$$D^{ij} \xi_i \xi_j \geq D \xi_i \xi_i \quad \text{for } \forall\, \xi_i \xi_i > 0. \tag{5.12}$$

On the grain boundary we have from (4.13), (4.14), (4.15), (4.18) and (4.19)

$$[v_n] = 0 \tag{5.13}$$

$$[\sigma_{nn}] = 0 \tag{5.14}$$

$$\sigma_{nn} = As \text{ at each side of grain boundary} \tag{5.15}$$

$$\sigma^{\alpha j} n_j = -\mu^{\alpha\beta} \left[\frac{\partial w_\beta^{(e)}}{\partial t} \right] \tag{5.16}$$

$$J^\alpha = \frac{d^{\alpha\beta}}{\rho} \nabla_\beta \sigma_{nn} \text{ at each side of grain boundary} \tag{5.17}$$

It follows from (5.4) and (5.17) the law of growth of grain boundaries due to boundary diffusion

$$\rho u = \nabla_\beta \frac{d^{\alpha\beta}}{\rho} \nabla_\alpha \sigma_{nn} \tag{5.18}$$

Equations (5.1) - (5.18) form a closed system of equations of diffusional creep.

6 Homogenization Problem.

From now on we shall consider a special case of the linearized theory, formulated in Section 5, when there is no boundary diffusion and hence the only irreversible deformation is due to the bulk vacancy diffusion. Formally this means that coefficients $d^{\alpha\beta}$ in (4.19) are supposed to be zero, which eliminates equations (5.17), (5.4) and (5.18) from the system (5.1) - (5.18).

Further we assume that constants $\mu^{\alpha\beta}$ in boundary conditions (5.16) are zero, which neglects the tangent stresses at the grain boundary:

$$\sigma^{\alpha j} n_j = 0 \text{ at each side of grain boundary} \tag{6.1}$$

This is equivalent to an additional assumption that the process of shear stress relaxation at the grain boundaries is much faster than the bulk diffusion process and completes immediately after the load is applied, so that the adjacent grains can slide without resistance along their common boundary.

In the absence of the boundary diffusion the deformation of the region V is described by the displacement field $w_i(x^i, t)$, defined in V and related to the velocity v_i by the formula

$$v_i(\boldsymbol{x}, t) = \dot{w}_i(\boldsymbol{x}, t) \quad \boldsymbol{x} \in V \tag{6.2}$$

We introduce also the plastic displacements, which are determined by the flux J_i by means of relation

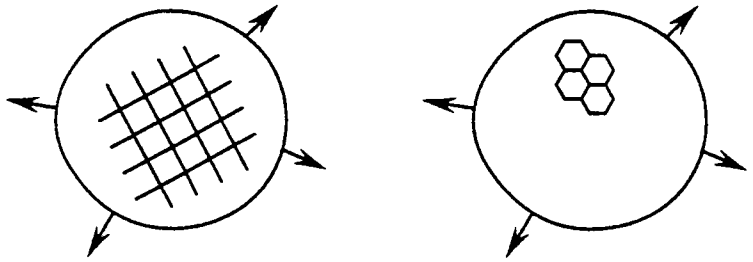
$$\dot{w}_i^{(p)}(\boldsymbol{x}, t) = -J_i(\boldsymbol{x}, t) \quad \boldsymbol{x} \in V \tag{6.3}$$

Than the displacements w_i are the sum of the elastic and plastic displacements

$$w_i = w_i^{(e)} + w_i^{(p)}, \tag{6.4}$$

The similar is also true for the strains:

$$\varepsilon_{ij} = 1/2 \left(\frac{\partial w_i}{\partial x^j} + \frac{\partial w_j}{\partial x^i} \right), \quad \varepsilon_{ij}^{(p)} = 1/2 \left(\frac{\partial w_i^{(p)}}{\partial x^j} + \frac{\partial w_j^{(p)}}{\partial x^i} \right)$$
$$\varepsilon_{ij} = \varepsilon_{ij}^{(e)} + \varepsilon_{ij}^{(p)} \tag{6.5}$$



a. Rectangular Microstructure. b. Honycomb Microstructure.

Figure 6: Microstructures.

Instead of (5.13), the continuity condition of normal displacement will be employed

$$[w_n] = 0 \tag{6.6}$$

Condition (5.13) follows from (6.6) but not vise versa. The difference is that (6.6) excludes the possibility that the normal displacements are discontinuous at the moment $t = 0$ when the load is applied.

It is also necessary to complement the equations above with initial conditions for vacancy concentration and plastic displacements:

$$s(\boldsymbol{x}, t) = 0, \quad \boldsymbol{x} \in V, \ t = 0, \tag{6.7}$$

$$\boldsymbol{w}^{(p)}(\boldsymbol{x}, t) = 0, \quad \boldsymbol{x} \in V, \ t = 0. \tag{6.8}$$

The closed system of equations in the considered case of the absence of the boundary diffusion and zero shear boundary stresses consists of the equations: (5.5), (5.8) - (5.11), (5.14), (5.15), (6.1), (6.3) - (6.8).

Consider a polycrystal body containing a huge number of grains. We are going to derive a theory for prediction of mechanical behavior of the body. The experience gained in averaging of random structures shows that the most results for bodies with random and periodic structures are qualitatively similar. (See, for example, [27]). Therefore we consider a body with a periodic microstructure (Fig. 6) loaded with some constant or variable traction. The problem is to find microfields of elastic and plastic deformations and macroscopical constitutive equations.

For simplicity and consistency with the performed numerical modeling, only the 2-D plane strain case of regular hexagonal periodical microstructure (Fig. 6, b)). will be considered. The reason is that with boundary condition (6.1) not all microstructures can withstand the instantaneous application of the external traction. For example, rectangular microstructure (Fig. 6, a)) can not be loaded by shear stresses, applied parallel to the grain boundary. In other words, any macrodeformation of the structure should be the result of the application of macrostresses. Here we decided to pick up one structure which possesses the necessary properties rather than to formulate general restrictions on the grain geometry, which though could be done in 2-D as well as in 3-D case. An accurate formulation of that property will be done at the end of this Section after the formulation of the homogenization problem.

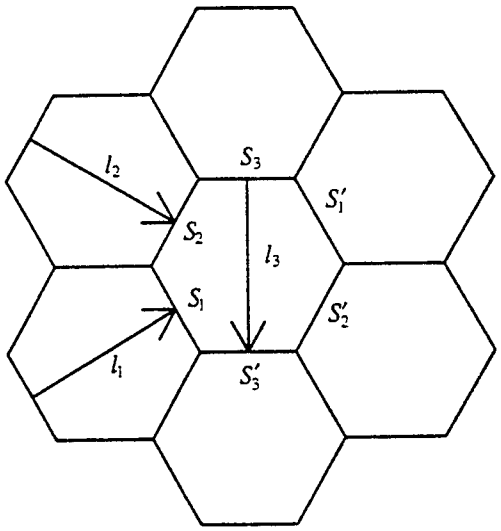


Figure 7: Hexagonal structure. The translation vectors mapping the corresponding parts of the cell boundaries, shown by arrows.

We consider the asymptotical statement of homogenization problem when the period of microstructure L tends to zero, and averaged equations are the corresponding limit equations (see, for example, [27]).

Before presenting the results, some description of the periodic structure is to be done.

We assume that the grains coincide with the cells of the periodic structure. Let ω^+ be an arbitrary cell, and ϵ be half of the distance between the opposite hexagon edges, which will be taken for the characteristic size of the grain. The boundary $\partial\omega^+$ of the cell ω^+ is comprised of three pairs of lines S_1, S_1', S_2, S_2', S_3, S_3' such that for every line S_α, there exists a translation $l_\alpha \in G$, mapping S_α onto S_α'. This notation is explained on Fig. 7.

The periodical regular hexagonal grain structure M is obtained by translation of that cell by all elements of translation symmetry group, generated by vectors l^1 and l^2:

$$G = \left\{ l^{mk} \,\middle|\, l^{mk} = ml^1 + kl^2, \quad m, k = 0, \pm 1, \pm 2, \ldots \right\}. \tag{6.9}$$

For $l \in G$ we denote by $\omega(l)$ the image of the cell ω^+ under the translation l. Different cells $\omega(l)$ may have in common the boundary points only, and the union of the cells covers the whole plane. Obviously, the translation $-l_\alpha$ maps S_α' onto S_α. Thus the periodic structure induces the certain mapping of the cell boundary $\partial\omega^+ \Longleftrightarrow \partial\omega^+$, which will be used for the formulation of the boundary conditions. For every point $\boldsymbol{x} \in \partial\omega^+$ we denote by $\boldsymbol{l}(\boldsymbol{x})$ the corresponding translation vector. The points $\boldsymbol{x}$ and $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{l}(\boldsymbol{x})$ will be referred to as the corresponding points. Note that $\boldsymbol{l}(\boldsymbol{x})$ is constant within each line S_α, S_α'.

The unit normal $\boldsymbol{n}$ to the cell boundary is assumed to be directed outward the cell, therefore at the corresponding points $\boldsymbol{x}$ and $\boldsymbol{x}'$ we have

$$\boldsymbol{n}(\boldsymbol{x}) + \boldsymbol{n}(\boldsymbol{x}') = 0, \quad \boldsymbol{l}(\boldsymbol{x}) = -2\epsilon\boldsymbol{n}(\boldsymbol{x}). \tag{6.10}$$

Let $f(\boldsymbol{x})$ be an arbitrary function, which is continuous within each grain, but may be discontinuous at the grain boundaries. Function $f(\boldsymbol{x})$ is called periodic if

$$f(\boldsymbol{x}+\boldsymbol{l}) = f(\boldsymbol{x}) \quad \text{for any } \boldsymbol{x} \in \dot{\omega}^{+} \text{ and for any } \boldsymbol{l} \in G. \tag{6.11}$$

Here $\dot{\omega}^{+}$ is the interior of a cell ω^{+}.

If function $f(\boldsymbol{x})$ is known within any cell, it can be extended to the whole space by the formula (6.11). From now on the term "periodic function" will be used in the sense of the above definition, unless otherwise is explicitly indicated.

Denote by ω^{-} the cell, such that $S_i = \omega^{+} \cap \omega^{-}$. It follows from (6.11) and the definition of the corresponding points that

$$[f] \equiv f^{+} - f^{-} \equiv f(\boldsymbol{x}^{+}.t) - f(\boldsymbol{x}^{-},t) = f(\boldsymbol{x},t) - f(\boldsymbol{x}',t) \quad \text{for } \boldsymbol{x} \in S_i. \tag{6.12}$$

Thus, for periodic functions the discontinuity conditions can be expressed in terms of function values within one cell, which allows to formulate the cell problem. Instead of using of formal procedure of homogenization (See, for example, [27]) we prefer here "intuitive" approach, which seems to be easier in our particular case.

Averaged equations by its physical sense relate a macroscopically homogeneous deformation of a "large" (compared with grain size ϵ) specimen to averaged stresses. Instead of "large" specimen the whole plane is considered. One may assume that macrostrains $\bar{\varepsilon}_{ij}(t)$ are given as functions of time and macrostresses $\bar{\sigma}^{ij}(t)$ should be found, or vise versa. For definiteness, let us consider the case when macrostresses are given.

If there were no grain boundaries, the homogeneous plain deformation history would be generated by the displacement field

$$\bar{w}_i(\boldsymbol{x},t) = \bar{\varepsilon}_{ij}(t)\, x^j \tag{6.13}$$

The grain structure results in additional periodical displacements $W_i(\boldsymbol{x},t)$, so that total displacements are given by the sum

$$w_i(\boldsymbol{x},t) = \bar{\varepsilon}_{ij}(t)\, x^j + W_i(\boldsymbol{x},t) \tag{6.14}$$

Since the first term in (6.14) is obviously continuous over space coordinates, it follows from (6.6) and (6.12) that the field $W_i(x,t)$ satisfies the condition

$$W_n(\boldsymbol{x},t) + W_n(\boldsymbol{x}',t) = 0 \text{ for } \boldsymbol{x} \in \partial\omega^{+} \Rightarrow \dot{W}_n(\boldsymbol{x},t) + \dot{W}_n(\boldsymbol{x}',t) = 0. \tag{6.15}$$

Vacancy concentration s is a periodic function. With (6.12) taken into account, equations (5.14) and (5.15) link the normal stress values and vacancy concentration at the corresponding points of the boundary:

$$\sigma_{nn}(\boldsymbol{x}.t) = \sigma_{nn}(\boldsymbol{x}',t) \quad \text{for } x \in \partial\omega^{+}, \tag{6.16}$$

$$s(\boldsymbol{x},t) = s(\boldsymbol{x}',t) \quad \text{for } \boldsymbol{x} \in \partial\omega^{+}. \tag{6.17}$$

The macrostresses, or averaged stresses, are defined by formula

$$\bar{\sigma}^{ij}(t) = \frac{1}{|\omega^+|}\int\limits_{\omega^+}\sigma^{ij}(\boldsymbol{x},t)\,d^2x \tag{6.18}$$

The full set of equations is as follows: (5.5), (5.8) - (5.11), (5.14), (5.15), (6.1) (6.3) - (6.5), (6.14) - (6.18). For further reference that system is denoted as system P. Initial conditions for the system P are (6.7) and (6.8). It is implied that all equations included in system P and initial conditions should be satisfied in the cell ω^+.

Now we are going to show, that the chosen microstructure can not be subjected to instantaneous macrodeformation, if stresses are zero. With zero stresses and zero vacancy concentration s, the elastic strain coincides with the total strain and is equal to zero, hence the displacement field w_i within cell ω^+ is rigid body motion:

$$w_i = \bar{\varepsilon}_{ij}x^j + W_i = \lambda e_{ij}x^j + a_i, \boldsymbol{x} \in \omega^+,$$

$$\lambda,\ a_i = const,\quad e_{11} = e_{22} = 0,\ e_{12} = -e_{21} = 1. \tag{6.19}$$

Relation (6.19) allows to express the displacement W_i in terms of macrostrains and rigid body motion:

$$W_i = -\bar{\varepsilon}_{ij}x^j + \lambda e_{ij}x^j + a_i. \tag{6.20}$$

Plugging of (6.20) into continuity condition (6.15) yields

$$0 = W_n(\boldsymbol{x}) + W_n(\boldsymbol{x}') =$$

$$= \left(-\bar{\varepsilon}_{ij}x^j + \lambda e_{ij}x^j + a_i\right)n^i(\boldsymbol{x}) + \left(-\bar{\varepsilon}_{ij}x'^j + \lambda e_{ij}x'^j + a_i\right)n^i(\boldsymbol{x}') =$$

$$= \left(-\bar{\varepsilon}_{ij}x^j + \lambda e_{ij}x^j + a_i\right)n^i(\boldsymbol{x}) - \left(-\bar{\varepsilon}_{ij}x'^j + \lambda e_{ij}x'^j + a_i\right)n^i(\boldsymbol{x}) =$$

$$= \left(-\bar{\varepsilon}_{ij} + \lambda e_{ij}\right)\left(x^j - x'^j\right)n^i(\boldsymbol{x}) = -\left(-\bar{\varepsilon}_{ij} + \lambda e_{ij}\right)l^j(\boldsymbol{x})\,n^i(\boldsymbol{x}) =$$

$$= \bar{\varepsilon}_{ij}l^j(\boldsymbol{x})\,n^i(\boldsymbol{x}) - e_{ij}l^j(\boldsymbol{x})\,n^i(\boldsymbol{x}) =$$

$$= \bar{\varepsilon}_{ij}l^j(\boldsymbol{x})\,n^i(\boldsymbol{x}) - \lambda\boldsymbol{n}(\boldsymbol{x})\otimes\boldsymbol{l}(\boldsymbol{x}) = \bar{\varepsilon}_{ij}l^j(\boldsymbol{x})\,n^i(\boldsymbol{x}),\quad \boldsymbol{x} \in \partial\omega^+. \tag{6.21}$$

The vector product $\boldsymbol{n}\otimes\boldsymbol{l}$ in (6.21) vanishes because these vectors are collinear at each boundary point (See (6.10)). Since normal is constant within each edge of the hexagon, (6.21) provides three homogeneous linear equations with respect to three macrostrain components $\bar{\varepsilon}_{ij}$. The direct checking shows that its determinant is not zero, which implies that all macrostrains have to be zero.

Let R be the set of periodical displacement field V_i, defined at the cell ω^+ by formula for rigid body motion

$$V_i = \lambda e_{ij}x^j + a_i,\quad \lambda,\ a_i = const, \tag{6.22}$$

and extended to the whole plane by periodicity condition (6.11). Under the displacement V_i each cell shifts by the constant vector a_i and rotates around its center by the angle λ . It follows from(6.19) - (6.21) that any such a field satisfies the continuity condition (6.15) and does not produce macrodeformation. Fig. 9 illustrates the movement of the cells. The holes that one can see at the corners of the hexagons, is the second order effect and is ignored by the small deflection theory used here.

Theorem 1.

Consider the solution of the system P with the initial conditions (6.7) and (6.8). Macrostresses $\bar{\sigma}^{ij}(t)$ are given functions of time. The total and elastic displacements of this solution are defined with the accuracy of the arbitrary displacement field from set R. All the other components of the solution , such as vacancy concentration, plastic displacements and strains, elastic strains, macrostrains and stresses are uniquely defined.

Proof.

Introduce the notations

$$I(t) = \frac{1}{2}\int\limits_{\omega^*} As^2 d^2x + \frac{1}{2}\left[\varepsilon^{(e)}, \varepsilon^{(e)}\right],$$

$$\left[\varepsilon^{(e)}, \varepsilon^{(e)}\right] \equiv \int\limits_{\omega^+} A^{ijkt}\varepsilon_{ij}^{(e)}\varepsilon_{kt}^{(e)} d^2x,$$

$$[\nabla s, \nabla s] \equiv \int\limits_{\omega^*} AD^{ij}\frac{\partial s}{\partial x^i}\frac{\partial s}{\partial x^j} d^2x \tag{6.23}$$

Since the system P is linear, it is sufficient to prove, that if macrostresses are zero, than the system P with initial conditions (6.7) and (6.8) has only zero solution for all components with the exception of displacements, which belong to the set R. At the initial moment $t = 0$ the plastic displacement is zero due to (6.8) , the displacements coincide with the elastic displacements and since the macrostresses are zero, the macrostrains are also zero at the moment $t = 0$. Hence

$$I(0) = 0. \tag{6.24}$$

Using inequality (C.6) for $t^* = 0$ from Appendix C, we obtain that functional $I(t)$ is zero for $t \geq 0$. Hence

$$s(t) = 0, \quad \varepsilon_{ij}^{(e)} = 0 \quad t \geq 0 \quad \Rightarrow \quad \sigma_{ij} = 0, \varepsilon_{ij}^{(p)} = 0. \tag{6.25}$$

Hence the displacement of the cell is rigid body motion, given by the formula (6.19). It was proven above that in order to satisfy the continuity condition (6.15), the macrostrains has to be zero. The uniqueness theorem is proven.

Remark 1. Let us consider the loading case when the non-zero macrostresses are applied only at some time interval $[0, t^*]$, and were removed afterwards. Then from inequality (C.6) we conclude that vacancy concentration s and elastic strains $\varepsilon_{ij}^{(e)}$ exponentially tend to zero, hence the stresses also tend to zero. In other words, after unloading the residual stresses are relaxing to zero exponentially with respect to time.

We conclude this Section with the presentation of averaged stresses in terms of values of normal microstresses at the grain boundary (See Appendix B):

$$\bar{\sigma}^{ij} = \frac{\epsilon}{|\omega^+|} \int\limits_{\partial\omega^+} \sigma_{nn}\, n^i\, n^j dx. \tag{6.26}$$

Relation (6.26) is valid for arbitrary stress field satisfying equilibrium equations (5.9) and boundary conditions (6.1) , (6.16). With (5.15) taken into account, the averaged stresses can be expressed in terms of the values of vacancy concentration at the grain boundary:

$$\bar{\sigma}^{ij} = \frac{A\epsilon}{|\omega^+|} \int\limits_{\partial\omega^+} s\, n^i\, n^j dx. \tag{6.27}$$

It can be checked (See Appendix B) that for arbitrary constant C, the following identity holds

$$C\delta^{ij} = \frac{\epsilon}{|\omega^+|} \int\limits_{\partial\omega^+} C\, n^i\, n^j dx. \tag{6.28}$$

It follows from (6.28) and (6.27), that if vacancy concentration is constant over the grain boundary than the corresponding macrostress tensor is spherical and the plane is under hydrostatic compression or tension.

7 Boltzman Superposition Principle and Macroequations.

As it has been already stated above, the macromodel should provide the relations between macrostresses and macrostrains $\bar{\sigma}^{ij}(t)$ and $\bar{\varepsilon}_{ij}(t)$. It seems almost obvious, that any parabolic type linear system such as P satisfies Boltzman superposition principle and hence the stress-strain relation would involve an integral operator.

Let us first assume that at $t = 0$ the unit tension along axis x^1 is instantaneously applied to the polycrystal and remains unchanged for $t > 0$. Than the only non-zero stress component is $\bar{\sigma}^{11}(t) = 1$. Denote by $\aleph(t)$ the solution of system P with initial conditions (6.7) and (6.8), corresponding to load case under consideration:

$$\begin{aligned} \aleph(t) \;=\; & \{\bar{\varepsilon}_{ij}(t),\, \sigma^{ij}(\cdot,t),\, \varepsilon_{ij}(\cdot,t),\, \varepsilon_{ij}^{(e)}(\cdot,t),\, \varepsilon_{ij}^{(p)}(\cdot,t), \\ & s(\cdot,t),\, W_i(\cdot,t),\, w_i(\cdot,t),\, w_i^{(e)}(\cdot,t)\, w_i^{(p)}(\cdot,t)\}. \end{aligned} \tag{7.1}$$

Solution $\aleph(t)$ is defined only for $t \geq 0$. Let us formally define it for $t < 0$:

$$\aleph(t) \;=\; 0 \quad \text{for } t < 0. \tag{7.2}$$

If the same tension $\bar{\sigma}^{11}(t) = 1$ is applied at some time $t_1 > 0$, than the solution is obviously equal to $\aleph(t - t_1)$ for $t \geq 0$. Let us stress that $\aleph(t - t_1) = 0$ for $t < t_1$ because of the definition (7.2).

The next step is to consider the load history, when at a discrete moments $t_i = i\Delta$, $i = 1, 2, \ldots, k$ a tension increments $d\bar{\sigma}^{11}(t_1)$, $d\bar{\sigma}^{11}(t_2)$, $\ldots, d\bar{\sigma}^{11}(t_k)$ are applied. Than at any particular time t, $t_m < t < t_{m+1}$, the total tension $\bar{\sigma}^{11}(t_i)$ is given by the formula

$$\bar{\sigma}^{11}(t) \;=\; \sum_{i=1}^{m} d\bar{\sigma}^{11}(t_i), \tag{7.3}$$

and the solution is given by the sum

$$\sum_{i=1}^{m} d\bar{\sigma}^{11}(t_i)\aleph(t-t_i). \tag{7.4}$$

Extension of the formula (7.4) to a continuous loading process provides the following formula for the solution:

$$\int_0^t \dot{\bar{\sigma}}^{11}(\xi)\,\aleph(t-\xi)\,d\xi. \tag{7.5}$$

Let us denote by $R_{ij,kt}(t)$ the macrostrain $\bar{\varepsilon}_{ij}(t)$ corresponding to the application of the macrostress $\bar{\sigma}^{kt}(t) = 1,\ t > 0$. The values $R_{ij,kt}(t)$ at $t = 0$ are components of the tensor of elastic compliances of polycrystal. Because of that it is convenient to decompose $R_{ij,kt}(t)$ into the sum

$$R_{ij,kt}(t) = R_{ij,kt}(0) + K_{ij,kt}(t), \quad K_{ij,kt}(0) = 0 \tag{7.6}$$

By its mechanical sense the function $K_{ij,kt}(t)$ is the $\bar{\varepsilon}_{ij}$ creep strain component caused by constant load $\bar{\sigma}^{kt}(t) = 1$, while the other macrostress components are equal to zero. Than for an arbitrary loading process it holds

$$\begin{aligned}\bar{\varepsilon}_{ij}(t) &= \int_0^t R_{ij,kt}(t-\xi)\,\dot{\bar{\sigma}}^{kt}(\xi)\,d\xi = \\ &= R_{ij,kt}(0)\,\bar{\sigma}^{kt}(t) + \int_0^t \frac{\partial K_{ij,kt}(t-\xi)}{\partial t}\bar{\sigma}^{kt}(\xi)\,d\xi\end{aligned} \tag{7.7}$$

Equations of the type (7.7) are widely used for creep modeling of polymers and concrete.

So we arrive to the conclusion: in order to find macrostrains, caused by arbitrary loading process, it is necessary and sufficient to know instantaneous elastic moduli tensor $R_{ij,kt}(0)$ and creep tensor $K_{ij,kt}(t)$, which components are creep strains caused by the corresponding constant macrostresses. Thus, in numerical modeling or experiment one may consider only loading cases when constant load is instantaneously applied to the body and remains unchanged. This is nothing else but classical experiment to find the creep property of material.

Inversion of (7.7) renders

$$\begin{aligned}\bar{\sigma}^{ij}(t) &= \int_0^t Q^{ij,kt}(t-\xi)\,\dot{\bar{\varepsilon}}_{kt}(\xi)\,d\xi = \\ &= Q^{ij,kt}(0)\,\bar{\varepsilon}_{kt}(t) + \int_0^t \frac{\partial Z^{ij,kt}(t-\xi)}{\partial t}\bar{\varepsilon}_{kt}(\xi)\,d\xi,\end{aligned} \tag{7.8}$$

$$Q^{ij,kt}(t) = Q^{ij,kt}(0) + Z^{ij,kt}(t), \quad Z^{ij,kt}(0) = 0 \tag{7.9}$$

Here $Q^{ij,kt}(t)$ is macrostrees component $\sigma_0^{ij}(t)$ caused by instantaneous application of macrostrain $\bar{\varepsilon}_{kt}(t) = 1$, while all the other macrostrain components are equal to zero. Tensor $Q^{ij,kt}(0)$ is the elastic moduli tensor of polycrystal.

It is worth mentioning that creep curves $K_{ij.kl}(t)$ for small values of t have an asymptotics

$$K_{ij.kl}(t) \sim t^{1/2}, \quad \frac{\partial K_{ij.kl}(t)}{\partial t} \sim t^{-1/2} \tag{7.10}$$

and hence creep rate tends to infinity as $t^{-1/2}$ when t tends to zero:

$$\dot{\bar{\varepsilon}}_{ij}(t) \sim t^{-1/2}, \quad t \to 0 \tag{7.11}$$

An important feature of the constitutive equations (7.7), (7.8) is that these relations are not local: there is a memory of the history of the process. This means that local theories of primary creep are not adequate at least in the case of bulk diffusional creep.

8 Secondary creep

Generally speaking, the macroscopic constitutive equations are given by the integral operators (7.7) or (7.8). However, for "slow" loading processes and developed creep it is possible to use as an approximation the creep law

$$\dot{\bar{\varepsilon}}_{ij} = E_{ijkl}\,\bar{\sigma}'^{kl}, \qquad \bar{\sigma}'^{kl} = \bar{\sigma}^{kl} - \delta^{kl}\,\bar{\sigma}^{ss}/2, \tag{8.1}$$

or

$$\bar{\sigma}'^{ij} = e^{ijkl}\,\dot{\bar{\varepsilon}}_{kl}. \tag{8.2}$$

Also incompressibility condition is imposed

$$\dot{\bar{\varepsilon}}_{kk} = 0, \tag{8.3}$$

which reflects physically obvious fact that there is no volume change due to bulk vacancy diffusion. Tensor e^{ijkl} is the inverse tensor to E_{ijkl}.

The macrocharacteristics of the secondary creep E_{ijkl} are the limits of the creep rates $\dot{K}_{ijkl}(t)$ when $t \to \infty$. The fact that under the applied constant macrostresses the creep rates tend to some constants when $t \to \infty$ will be formulated and justified below and constitutes the basis of the approximation (8.1) - (8.3).

We start from formal description of how to compute the constants involved into secondary creep law (8.2). It turns out that they may be found from the following variational principle.

Let $\dot{\bar{\varepsilon}}_{ij}$ be an arbitrary constant macroscopic creep rates, satisfying the incompressibility condition (8.3). Denote by $J(s)$ the following functional of function $s(\boldsymbol{x})$

$$J(s) = \frac{1}{2}[\nabla s, \nabla s] - \epsilon \int\limits_{\partial\omega^+} \dot{\bar{\varepsilon}}_{ij} n^i n^j s\, d\boldsymbol{x}. \tag{8.4}$$

Here the notation (6.23) is used. Consider the minimization problem

$$J(s) \to \min_s . \tag{8.5}$$

Minimum is sought on the set of all functions s obeying the constraints (6.17). It follows from (6.28) and (8.3) that linear with respect to s term in (8.4) is zero for $s = const.$ hence the solution s^* of the problem is determined up to an arbitrary constant. We fix this constant by the condition

$$\int\limits_{\partial\omega^+} s^*(\boldsymbol{x})\, n^k\, n^k\, d\boldsymbol{x} = 0. \tag{8.6}$$

The necessary and sufficient condition of minimum is the following identity, which should hold for every function satisfying the condition (6.17)

$$[\nabla s^*, \nabla s] = \epsilon \int\limits_{\partial\omega^+} \dot{\bar{\varepsilon}}_{ij} n^i n^j s\, d\boldsymbol{x} \quad \text{for } \forall\ s:\ s(\boldsymbol{x}) = s(\boldsymbol{x}'),\ \ \boldsymbol{x} \in \partial\omega^+. \tag{8.7}$$

Differential form of the problem (8.5) is derived from (8.7) :

$$\frac{\partial}{\partial x^i} A D^{ij} \frac{\partial s}{\partial x^j} = 0 \qquad \boldsymbol{x} \in \omega^+, \tag{8.8}$$

$$[J_n](\boldsymbol{x}) = -2\epsilon \dot{\bar{\varepsilon}}_{ij} n^i(\boldsymbol{x})\, n^j(\boldsymbol{x}),\ \ \boldsymbol{x} \in \partial\omega^+.$$

$$[J_n](\boldsymbol{x}) \equiv J_n(\boldsymbol{x}) + J_n(\boldsymbol{x}') \qquad \boldsymbol{x} \in \partial\omega^+,$$

$$J_n(\boldsymbol{x}) \equiv -A D^{ij} \frac{\partial s(\boldsymbol{x})}{\partial x^j} n_i(\boldsymbol{x}) \quad \boldsymbol{x} \in \partial\omega^+. \tag{8.9}$$

After the solution s^* of the variational problem (8.4), (6.17), (8.5), (8.6) is found, the deviator of macrostresses is defined by the formula (6.27) which takes the form

$$\bar{\sigma}'^{ij} = \frac{A\epsilon}{|\omega^+|} \int\limits_{\partial\omega^+} s^*\, n^i\, n^j d\boldsymbol{x}. \tag{8.10}$$

Macrostresses $\bar{\sigma}'^{ij}$ are deviatoric due to condition (8.6) since

$$\bar{\sigma}'^{kk} = \frac{A\epsilon}{|\omega^+|} \int\limits_{\partial\omega^+} s^*\, n^k\, n^k d\boldsymbol{x} = 0. \tag{8.11}$$

The solution s^* depends linearly on the parameters $\dot{\bar{\varepsilon}}_{ij}$. Hence by plugging this solution in (8.10) one obtains macrostresses in terms of creep velocities $\dot{\bar{\varepsilon}}_{ij}$, i.e. the relation (8.2). With more details,consider two solutions, corresponding to two linear independent loading cases:

$$\begin{aligned} &s^{12} = s^{21} \ \text{ corresponds to } \dot{\bar{\varepsilon}}_{12} = \dot{\bar{\varepsilon}}_{21} = 1/2,\ \dot{\bar{\varepsilon}}_{11} = 0,\ \dot{\bar{\varepsilon}}_{22} = 0, \\ &s^{11} = -s^{22} \text{ corresponds to } \dot{\bar{\varepsilon}}_{12} = \dot{\bar{\varepsilon}}_{21} = 0,\ \dot{\bar{\varepsilon}}_{11} = -\dot{\bar{\varepsilon}}_{22} = 1/2. \end{aligned} \tag{8.12}$$

Than the solution s^* is the linear combination

$$s^* = \dot{\bar{\varepsilon}}_{ij} s^{ij} \tag{8.13}$$

Substitution of (8.13) into (8.10) provides the formulas for the macrocharacteristics e^{ijkl}:

$$e^{ijkl} = \frac{A\epsilon}{|\omega^+|} \int\limits_{\partial\omega^+} \left(s^{kl}\right) n^i n^j dx. \tag{8.14}$$

It is obvious that only two constants among e^{ijkl} are independent.

So far it was shown how to find deviator of macrostresses if macroscopic constant incompressible creep rates are given. Let us prove that the secondary creep law is reversible. Multiplying (8.10) by $\dot{\bar{\varepsilon}}_{ij}$ we obtain after summation over repeated indices and using (8.7):

$$\bar{\sigma}'^{ij}\dot{\bar{\varepsilon}}_{ij} = \frac{A\epsilon}{|\omega^+|} \int\limits_{\partial\omega^+} s^* \dot{\bar{\varepsilon}}_{ij} n^i n^j dx = \frac{A}{|\omega^+|} \left[\nabla s^*, \nabla s^*\right]. \tag{8.15}$$

The left side of the relation (8.15) is zero if and only if all creep rates $\dot{\bar{\varepsilon}}_{ij} = 0$, which means that the matrix of the quadric form $\dot{\bar{\varepsilon}}_{ij} = e^{ijkt}\dot{\bar{\varepsilon}}_{ij}\dot{\bar{\varepsilon}}_{kt}$ is positively definite, hence the law (8.2) may be inverted.

Now we can describe how to find creep rates and vacancy concentration for secondary creep. Let macrostresses $\bar{\sigma}^{0ij}$ are given constants. First the deviator of tensor $\bar{\sigma}^{0ij}$ should be calculated

$$\bar{\sigma}'^{0ij} = \bar{\sigma}^{0ij} - \delta^{ij}p, \quad p = \bar{\sigma}^{0kk}/2. \tag{8.16}$$

Then the creep rates $\dot{\bar{\varepsilon}}_{0ij}$ are found satisfying the creep law (8.1) - (8.3). The vacancy concentration s^0 is the sum of the constant p and the solution of variational problem (8.5), corresponding to creep rates $\dot{\bar{\varepsilon}}^0_{ij}$:

$$s^0(\boldsymbol{x}) = p + \dot{\bar{\varepsilon}}^0_{ij} s^{ij}(\boldsymbol{x}). \tag{8.17}$$

The last step to define the microcharacteristics of the secondary creep is to determine the elastic strains and stresses within the cell ω^+. The normal stresses at the cell boundary are determined from ({ref5.12), since the vacancy concentration s^0 is found:

$$\sigma_{nn}(\boldsymbol{x}) = As^0(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\omega^+. \tag{8.18}$$

Formulas (6.1) and (8.18) define surface tractions at the grain boundary.

Thus the elastic displacements, elastic strains and stresses inside of the cell may be found from the solution of the elasticity problem (5.5), (5.8), (5.9), (6.1) and (8.18), if the principal vector and moment produced by surface tractions are zero, which they are as it is shown in Appendix B. Denote this solution as $w_i^{0(e)}$, $\varepsilon_{ij}^{0(e)}, \sigma^{0ij}$.

At this point all the characteristics of secondary creep are determined.

Theorem 3.

Under constant applied macrostresses $\bar{\sigma}^{0ij}$ the solution of the system P with initial conditions (6.7), (6.8) reveals the following asymptotic behavior:

$$s(\boldsymbol{x}, t) \rightarrow s^0(\boldsymbol{x}),$$

$$\dot{\bar{\varepsilon}}_{ij}(t) \rightarrow \dot{\bar{\varepsilon}}^0_{ij},$$

$$\varepsilon^{(e)}_{ij}(\boldsymbol{x}.t) \rightarrow \varepsilon^{0(e)}_{ij}(\boldsymbol{x}),$$

$$\sigma^{ij}(\boldsymbol{x},t) \rightarrow \sigma^{0ij}(\boldsymbol{x}). \tag{8.19}$$

Proof.

It is shown in Lemma 2 , Appendix C, that the difference between two arbitrary solutions of the system P, corresponding to the same loading process $\bar{\sigma}^{ij}(t)$, tends to zero in the following sense:

$$s^1(\boldsymbol{x},t) - s^2(\boldsymbol{x},t) \rightarrow 0.$$

$$\dot{\bar{\varepsilon}}^1_{ij}(t) - \dot{\bar{\varepsilon}}^2_{ij}(t) \rightarrow 0,$$

$$\varepsilon^{1(e)}_{ij}(\boldsymbol{x},t) - \varepsilon^{2(e)}_{ij}(\boldsymbol{x},t) \rightarrow 0,$$

$$\sigma^{1ij}(\boldsymbol{x},t) - \sigma^{2ij}(\boldsymbol{x},t) \rightarrow 0. \tag{8.20}$$

Let us stress that solutions need not to satisfy initial conditions (6.7), (6.8) and need not to have the same initial conditions. This means that if some particular solution of the system P is found, than any other solution tends to it, regardless of the initial conditions. Thus, to find the asymptotics of the solution of the problem it is sufficient to find some particular solution of the system P. We shall use upper case index "0" for all quantities related to this solution. This implies that introduced above functions with the same index are part of this particular solution.

Let us first define macrostrains as a constant strain rate process:

$$\bar{\varepsilon}^0_{ij}(t) = \dot{\bar{\varepsilon}}^0_{ij} t \tag{8.21}$$

Second, define the plastic displacement. Since the plastic displacement velocity is expressed in terms of the vacancy concentration from (6.3), (5.10), the only freedom left is to define the plastic displacements at $t = 0$. We pose

$$w^{0(p)}_i(\boldsymbol{x},0) = -w^{0(e)}_i(\boldsymbol{x}) \quad \boldsymbol{x} \in \omega^+. \tag{8.22}$$

Than

$$w^{0(p)}_i(\boldsymbol{x},t) = -w^{0(e)}_i(\boldsymbol{x}) - tJ^0_i(\boldsymbol{x}), \quad J^0_i \equiv -AD^{ij}\frac{\partial s^0}{\partial x^j}, \quad \boldsymbol{x} \in \omega^+ \quad \boldsymbol{x} \in \omega^+. \tag{8.23}$$

Third, since the elastic and plastic displacements are defined over the cell, the additional displacement in the presentation (6.14) ought to be as follows:

$$W^0_i(\boldsymbol{x},t) = -t\dot{\bar{\varepsilon}}^0_{ij}x^j - tJ^0_i(\boldsymbol{x}) \quad \boldsymbol{x} \in \omega^+. \tag{8.24}$$

To conclude the construction of the particular solution, it is necessary to check the continuity condition (6.15). It is obviously holds at $t = 0$, and hence it is enough to check second condition in (6.15) for $t > 0$. It follows from (8.9) ,(8.23) that

$$\left[\dot{W}_n^0\right] = -\dot{\bar{\varepsilon}}_{ij}^0 n^i n^j - \left[j_n^0\right] = 0. \tag{8.25}$$

Theorem is proven.

Remark 1. Let us normalize the diffusivity tensor:

$$D^{ij} = D\tilde{D}^{ij}, \tag{8.26}$$

where D is some characteristic value of tensor D^{ij}, and introduce dimensionless coordinates

$$y^i = \frac{x^i}{\epsilon} \tag{8.27}$$

which maps the cell ω^+ onto unit cell Ω. The functional (8.4) is transformed to

$$J(s) = \frac{1}{2}\int\limits_{\Omega} \tilde{D}^{ij}\frac{\partial s \partial s}{\partial y^i \partial y^j}d^2y - \frac{\epsilon^2}{AD}\int\limits_{\partial\Omega} \dot{\bar{\varepsilon}}_{ij} n^i n^j s d^y. \tag{8.28}$$

Than secondary creep macrocharacteristics can be represented as follows:

$$e^{ijkl} = \tilde{e}^{ijkl}\frac{\epsilon^2}{D}, \quad E_{ijkl} = \tilde{E}_{ijkl}\frac{D}{\epsilon^2}, \tag{8.29}$$

where dimensionless constants $\tilde{e}^{ijkl}$ and $\tilde{E}_{ijkl}$ depend on the constants $\tilde{D}^{ij}$ and the unit cell shape only. An important consequence is that secondary creep rates do not depend on the elastic properties and even on the value of the constant A. Elastic properties influence only stress microfields.

9 Numerical Results for Secondary Creep

For definiteness, it was assumed that grains are isotropic, and hence only four physical constants are needed: Young modulus E, Poison ratio ν, the constant A in (5.15), diffusivity constant D in (8.26) (with $\tilde{D}^{ij} = \delta^{ij}$), and the grain size ϵ.

Secondary Creep Rates. In creep, the periodic hexagonal structure behaves isotropically. Thus the creep law (8.1) contains just one macrocharacteristics - the viscosity μ:

$$\bar{\sigma}'^{ij} = \mu\dot{\bar{\epsilon}}^{ij}. \tag{9.1}$$

The dimension analysis of the cell problem shows that μ depends on the grain size ϵ and the diffusivity coefficient D only

$$\mu = a\frac{\epsilon^2}{D} \tag{9.2}$$

where a is some constant.Numerical simulations give the following value of the constant a for hexagonal structure

$$a = 0.059 \tag{9.3}$$

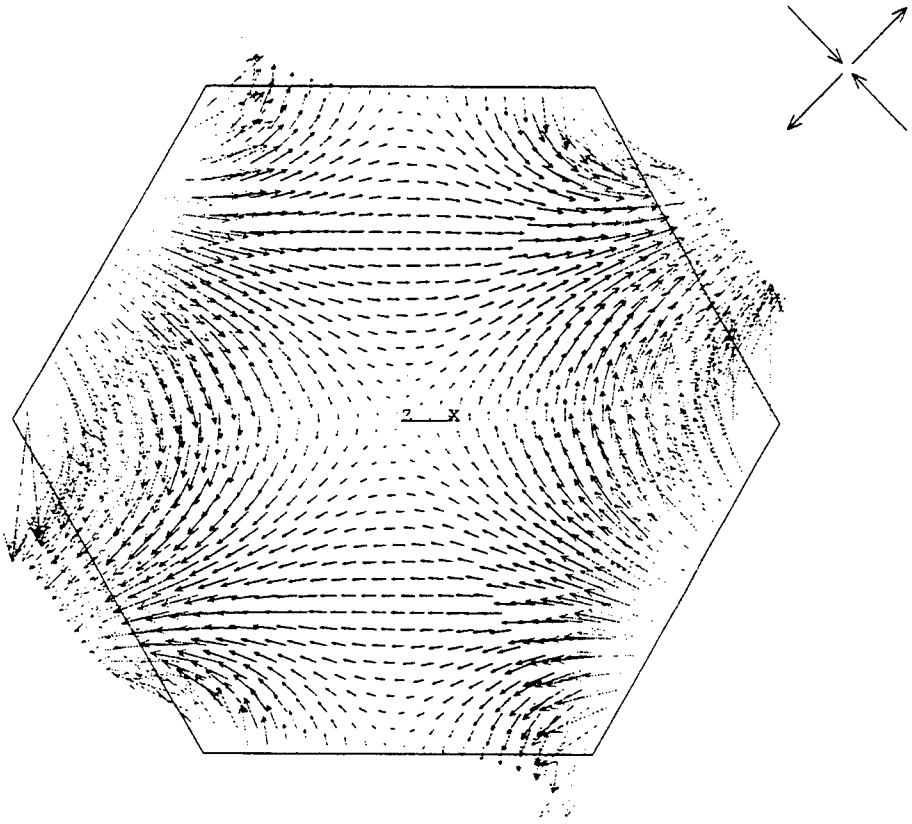


Figure 8: Creep velocity distribution during the secondary creep

Formulas (9.2), (9.3) inspire an assumption that the similar relation between macro- and microcharacteristics takes place for the random structure as well, where ϵ is the averaged grain size and D is the characteristic diffusion coefficient of monocrystals, while the coefficient a is of the order of unity.

Microdeformation. Distribution of creep velocity over the cell in the regime of secondary creep is shown in Fig. 8. The orientation of shear stress applied is given at the right top of Fig. 8. It is seen that there are three pairs of opposite cell sides with different properties. Material departs from one pair of sides and arrives at the other pair of sides. The remaining two sides consist of two pieces: material leaves one piece and arrives at the other one.

10 Dimensions Analysis and Transition Time to Secondary Creep.

Let E be some characteristic value of tensor A^{ijkt}. Similar to (8.26), normalize tensor tensor A^{ijkt} using the value E:

$$\tilde{A}^{ijkt} \equiv \frac{A^{ijkt}}{E}. \tag{10.1}$$

Let us assume that dimensionless parameters $\tilde{A}^{ijkt}$ and $\tilde{D}^{ijkt}$ remain unchanged in our analysis. Than a solution of the system P depends on four constants: E D, A (See (5.15)), and ϵ - characteristic grain size.

Our intent is to transform the system P to dimensionless form. In addition to dimensionless space coordinates y^i (See (8.26) introduce intrinsic time τ and normalized displacements and flux:

$$\tau = t\frac{AD}{\epsilon^2}; \quad \dot{f} \equiv \frac{\partial f}{\partial \tau},$$

$$\tilde{\boldsymbol{w}} = \tfrac{1}{\epsilon}\boldsymbol{w}, \tilde{\boldsymbol{W}} = \tfrac{1}{\epsilon}\boldsymbol{W} \;\; \tilde{\boldsymbol{w}}^{(e)} = \tfrac{1}{\epsilon}\boldsymbol{w}^{(e)}, \;\; \tilde{\boldsymbol{w}}^{(p)} = \tfrac{1}{\epsilon}\boldsymbol{w}^{(p)}, \qquad \tilde{J}_i = \frac{1}{\epsilon}\frac{\epsilon^2}{AD}J_i,$$

$$\tilde{\sigma}^{ij} = \tfrac{1}{A}\sigma^{ij} \tag{10.2}$$

Vacancy concentration and strains need not to be normalized. Than system P is reduced to the system $\tilde{P}$:

$$\varepsilon_{ij}^{(e)} = 1/2\left(\frac{\partial \tilde{w}_i^{(e)}}{\partial y^j} + \frac{\partial \tilde{w}_j^{(e)}}{\partial y^i}\right), \tag{10.3}$$

$$\tilde{\sigma}^{ij} = e\tilde{A}^{ijke}\varepsilon_{ke}^{(e)} \tag{10.4}$$

$$\frac{\partial \tilde{\sigma}^{ij}}{\partial y^j} = 0 \tag{10.5}$$

$$\tilde{J}^i = -\tilde{D}^{ij}\frac{\partial s}{\partial y^j} \tag{10.6}$$

$$\frac{\partial s}{\partial \tau} = \frac{\partial}{\partial y^i}\left(\tilde{D}^{ij}\frac{\partial s}{\partial y^j}\right) \tag{10.7}$$

$$\tilde{\sigma}_{nn} = s, \;\; \boldsymbol{y} \in \partial\Omega \tag{10.8}$$

$$\tilde{\sigma}^{\alpha j}n_j = 0 \quad \boldsymbol{y} \in \partial\Omega \tag{10.9}$$

$$\dot{\tilde{w}}_i^{(p)}(\boldsymbol{y},\tau) = -\tilde{J}_i(\boldsymbol{y},\tau) \quad \boldsymbol{y} \in \Omega \tag{10.10}$$

$$\tilde{w}_i = \tilde{w}_i^{(e)} + \tilde{w}_i^{(p)}. \tag{10.11}$$

$$\varepsilon_{ij} = 1/2\left(\frac{\partial \tilde{w}_i}{\partial y^j} + \frac{\partial \tilde{w}_j}{\partial y^i}\right), \quad \varepsilon_{ij}^{(p)} = 1/2\left(\frac{\partial \tilde{w}_i^{(p)}}{\partial y^j} + \frac{\partial \tilde{w}_j^{(p)}}{\partial y^i}\right)$$

$$\varepsilon_{ij} = \varepsilon_{ij}^{(e)} + \varepsilon_{ij}^{(p)} \tag{10.12}$$

$$\tilde{w}_i(\boldsymbol{y},\tau) = \bar{\varepsilon}_{ij}(\tau)\, y^j + \tilde{W}_i(\boldsymbol{y},\tau) \tag{10.13}$$

$$\tilde{W}_n(\boldsymbol{y},\tau) + \tilde{W}_n(\boldsymbol{y}',\tau) = 0 \text{ for } \boldsymbol{y} \in \partial\Omega \Rightarrow \dot{\tilde{W}}_n(\boldsymbol{y},\tau) + \dot{\tilde{W}}_n(\boldsymbol{y}',\tau) = 0. \tag{10.14}$$

$$\tilde{\sigma}_{nn}(\boldsymbol{y},\tau) = \tilde{\sigma}_{nn}(\boldsymbol{y}',\tau) \quad \text{for } \boldsymbol{y} \in \partial\Omega, \tag{10.15}$$

$$s(\boldsymbol{y},\tau) = s(\boldsymbol{y}',\tau) \quad \text{for } \boldsymbol{y} \in \partial\Omega. \tag{10.16}$$

$$\bar{\tilde{\sigma}}^{ij}(\tau) = \frac{1}{|\Omega|}\int_{\Omega} \tilde{\sigma}^{ij}(\boldsymbol{y},\tau)\, d^2y \tag{10.17}$$

Initial conditions :

$$s(\boldsymbol{y},\tau) = 0, \quad \tilde{\boldsymbol{w}}^{(p)}(\boldsymbol{y},\tau) = 0, \quad \boldsymbol{y} \in \Omega, \quad \tau = 0. \tag{10.18}$$

We see that the only dimensionless parameter, $e = E/A$, remains in the equations. To get a feeling what may be the actual value of parameter e, let us consider copper at 1000 K temperature. It is known that equilibrium value of vacancy concentration varies in broad range is $C_0 \sim 10^{-8} - 10^{-4}$. Than it follows from (5.7) that $e \sim 0.01 - 100$.

Let us study numerically how the solution depends on parameter e. For simplicity computations were done for the problem of compression of a single crystal by an absolutely rigid frictionless stamps (see [Lifshitch]). Region Ω is a square, the characteristic size is the distance from its center to the edges (See Fig. 10). Vertical crystal edges are free. For simplicity let us assume that crystal is isotropic. Under this assumption the compression of the crystal will not result in stamp rotation, and from symmetry considerations we may assume that the displacement of the cell center is zero. Let $\bar{\varepsilon}(\tau)$ be the vertical displacement of the upper stamp , which is the unknown function and which is analogous to the macrostrain in system P. The normal average stress $\bar{\tilde{\sigma}}$ at the contact between the stamps and the crystal surface serves as an analog to the macrostresses.The stamp is loaded by the constant force, such that the average stress is equal to -1.

$$\bar{\tilde{\sigma}} \equiv \frac{1}{2}\int_{-1}^{1} \tilde{\sigma}_n(y^1,1,\tau)\, dy^1 = \frac{1}{2}\int_{-1}^{1} \tilde{\sigma}_n(y^1,-1,\tau)\, dy^1 = -1, \tag{10.19}$$

The system of equations of the problem of compression of a single crystal is comprised from (10.3) - (10.12), (10.19), initial conditions (10.18) plus boundary conditions:

$$\tilde{\sigma}_{nn}(\pm 1, y^2, \tau) = 0, \quad -1 \le y^2 \le 1. \tag{10.20}$$

$$\tilde{w}_n(y^1, \pm 1, \tau) = \pm\bar{\varepsilon}(\tau), \quad -1 \le y^1 \le 1. \tag{10.21}$$

e	0.1		1		10	
τ	2.5	0.8	0.65	0.3	0.5	0.25
$\dot{\bar{\varepsilon}}$	1.74	2	1.72	2	1.73	2
Asymptotic Value of $\dot{\bar{\varepsilon}} = 1.7$						

Table 1: Stabilization of the creep rate for various values of parameter e.

The Theorems 1 - 3 can be proven for this problem as well.

The steady state solution for secondary creep can be obtained in closed form, and the value of steady state creep rate is $\dot{\bar{\varepsilon}} = 1.7$. Hence the analog of the formulas (9.1) - (9.3) in this case is

$$\bar{\sigma} = 0.588\dot{\bar{\varepsilon}}\frac{\epsilon^2}{D}. \tag{10.22}$$

One may notice that the numerical coefficients in (9.1) - (9.3) and in (10.22) are of the same order of magnitude.

Let us discuss numerical results for transient solution. Parameter e values were chosen to be 0.1, 1 .10, which is in the middle of expected range.

1. As it was expected, for small values of dimensionless time τ creep rates fit very well the asymptotic $\sim \tau^{-1/2}$ (See Fig. 11 -13).

2. With τ increased, steady state creep rate of 1.7 is achieved (see Table 1). Practically steady state is reached at $\tau \sim 1$ (see Table 1).

3. Parameter e somewhat affects the transition time necessary to reach the steady state creep rate. The smaller the e the larger the transition time. However the modeling results does not allow to conclude what kind of dependency is it . As one can see from the Table 1, the transition time for $e = 0.1$ is much larger, than for $e = 1$, but there is no noticeable difference between cases with $e = 1$ and 10.

4. At the first moment of load application, the only non-zero stress component is $\bar{\sigma}^{22}$, and it is equal to -1 over Ω. With creep developed, stresses tends to the limit, with does not depend on parameter e, as it should be because of the Theorem 3. Fig 14 shows stress distribution at the stamp-crystal contact for $e = 10$, $\tau = 0.5$. Stars mark asymptotical steady state stress distribution. Transition time to steady state stresses is of the same magnitude, as the transition time needed for creep rate to become constant.

11 Conclusions.

Three interesting outcomes of this study seem worthy noting.

First, the constitutive macroequations of diffusional creep turn out to be nonlocal. It is not seen how to eliminate the nonlocality by introducing additional internal variables. Probably, the elimination of the nonlocality on the macroscale is impossible in principle. Since this seems to be the case, a search for the adequate local constitutive equations for creep is hardly to be successful.

Second, there is an intrinsic material time $\tau = tDA/\varepsilon^2$. Strain- time dependence (for constant stresses) is universal for intrinsic time in the sense that it does not depend on the material and on the temperature (temperature dependence penetrates through the material constants D and A).

Third, as the variational principle shows, the creep rates do not depend on the elastic properties in secondary creep: only diffusion constants, the grain size and the grain geometry are important. Formula (7.2) is an example of such a dependency.

Acknowledgments.

The author thanks R. Bagley, P. Hazzledine for the useful discussions and B. Shoykhet and V. Sutyrin for collaboration. The support of this research by Structural Division of Wright-Paterson Laboratory and AFOSR grant F49620-94-1-0127 and Summer Extension Program Contract 95-0849 is greatly appreciated.

Appendix

A The Effect of Grain Boundary Stress Relaxation on Apparent Elastic Modulus

The assumption (6.1) that shear stresses at the grain boundaries can be neglected in creep problem is believed to be correct by many authors. It would be interesting to find an experimental evidence that such an effect is real. It may not be an easy task, because the numerical modeling revealed surprisingly low influence of grain shear stress relaxation on apparent elastic modulus. With more details, the averaged elastic properties were computed for periodic structure described in Section 6. Boundary conditions (6.1), (6.15) and (6.16) were applied and averaged elastic moduli were calculated from the solution of periodical elasticity problem. For definiteness it was assumed that grains are isotropic and hence only two elastic constants need to be calculated. More so, it is obvious that if hydrostatic pressure tensor is applied to the plane, the structure does not "feel" the cuts made, and hence the bulk modulus of the polycrystal is the same as the bulk modulus of the grain itself:

$$\frac{E^*}{2(1-\nu^*)} = \frac{E}{2(1-\nu)} \tag{A.1}$$

where E, ν and E^*, ν^* are Young's modulus and Poisson's ratio of the grain and the polycrystal correspondingly. Because of that the ratios E^*/E and G^*/G depend only on the Poisson's ratio.

Results are listed in Table 2, and as one can conclude the Young and shear moduli drop no more than 20% as a result of shear stress relaxation.

ν	0.3	0.34	0.45
E^*/E	0.830	0.829	0.828
G^*/G	0.806	0.811	0.823
ν^*	0.339	0.370	0.459

Table 2: Apparent elastic moduli of polycrystal with fully relaxed shear grain boundary stresses.

B Properties of periodical stress fields, satisfying the equilibrium conditions.

Let us prove formulas (6.26) and (6.28), which holds for arbitrary stress field satisfying equilibrium equations (5.9) and boundary conditions (6.16), (6.1).

Multiplying (5.9) by x^k and integrating over cell ω^+ we get

$$0 = \int\limits_{\omega^+} \frac{\partial \sigma^{ij}}{\partial x^j} x^k d^2x = -\int\limits_{\omega^+} \sigma^{ij}\delta_j^k d^2x + \int\limits_{\partial\omega^+} \sigma^{ij} n_j x^k dx \Rightarrow$$

$$\int\limits_{\omega^+} \sigma^{ik} d^2x = \int\limits_{\partial\omega^+} \sigma^{ij} n_j x^k dx = \int\limits_{\partial\omega^+} \sigma_{nn} n^i x^k dx =$$

$$= \sum_{r=1}^{3} \left(\int\limits_{S_r} \sigma_{nn}(\boldsymbol{x},t)\, n^i(\boldsymbol{x},t)\, x^k dx + \int\limits_{S_r'} \sigma_{nn}(\boldsymbol{x}',t)\, n^i(\boldsymbol{x}',t)\, x'^k dx' \right)$$

$$= \sum_{r=1}^{3} \left(\int\limits_{S_r} \sigma_{nn}(\boldsymbol{x},t)\, n^i(\boldsymbol{x})\, x^k dx - \int\limits_{S_r} \sigma_{nn}(\boldsymbol{x},t)\, n^i(\boldsymbol{x})\, (x^k + l^k(\boldsymbol{x}))\, dx \right) =$$

$$= -\sum_{r=1}^{3} \int\limits_{S_r} \sigma_{nn}(\boldsymbol{x},t)\, n^i(\boldsymbol{x})\, l^k(\boldsymbol{x})\, dx = -\tfrac{1}{2} \int\limits_{\partial\omega^+} \sigma_{nn} n^i l^k dx =$$

$$= \tfrac{\epsilon}{2} \int\limits_{\partial\omega^+} \sigma_{nn} n^i n^k dx \qquad \text{(B.1)}$$

The presentation (6.26) follows from (B.1) and definition (6.18). In order to prove (6.28) let us notice, that spherical tensor $\sigma^{ij} = C\delta^{ij}$ may be substituted in (6.26), which is reduced in this case to (6.28).

Let s be an arbitrary function, satisfying periodicity condition (6.17). Let us define surface tractions on the cell boundary $\partial\omega^+$ by formulas (6.1) and (5.15). Than the principal vector $\boldsymbol{F}$ and principal moment M applied to the grain due to these tractions, are zeros.

$$\boldsymbol{F} = \int\limits_{\partial\omega^+} \sigma_{nn}(\boldsymbol{x})\, \boldsymbol{n}(\boldsymbol{x})\, dx = 0. \qquad \text{(B.2)}$$

$$M = \int\limits_{\partial\omega^+} \sigma_{nn}(\boldsymbol{x})\, \boldsymbol{n} \otimes \boldsymbol{x} dx = \sum_{r=1}^{3} \left(\int\limits_{S_r} \sigma_{nn}(\boldsymbol{x}) \left(\boldsymbol{n}(\boldsymbol{x}) \otimes \boldsymbol{x} + \boldsymbol{n}(\boldsymbol{x}') \otimes \boldsymbol{x}'\right) dx \right) =$$

$$= \sum_{r=1}^{3} \left(\int_{S_r} \sigma_{nn}(\boldsymbol{x}) \boldsymbol{n}(\boldsymbol{x}) \otimes (\boldsymbol{x} - \boldsymbol{x}') \, dx \right) = - \sum_{r=1}^{3} \left(\int_{S_r} \sigma_{nn}(\boldsymbol{x}) \boldsymbol{n}(\boldsymbol{x}) \otimes \boldsymbol{l}(\boldsymbol{x}) \, dx \right) = 0. \tag{B.3}$$

The last term in (B.3) is zero because of (6.10).

C Asymptotic behavior of the solution at large time.

Lemma 1.

For every solution of the system P the following identity holds:

$$\frac{dI}{dt} + A[\nabla s, \nabla s] = |\omega^+| \, \bar{\sigma}^{ij} \dot{\bar{\epsilon}}_{ij} \tag{C.1}$$

Proof. It follows from (5.11), (5.15) that

$$Q \equiv \frac{dI}{dt} + A[\nabla s, \nabla s] = A[\nabla s, \nabla s] + \int_{\omega^-} As \frac{\partial s}{\partial t} d^2x + \int_{\omega^+} A^{ijkt} \epsilon_{ij}^{(e)} \dot{\epsilon}_{kt}^{(e)} d^2x =$$

$$= A[\nabla s, \nabla s] + \int_{\omega^-} A \frac{\partial}{\partial x^i} \left(D^{ij} \frac{\partial s}{\partial x^j} \right) d^2x + \int_{\omega^+} \sigma^{kt} \dot{\epsilon}_{kt}^{(e)} d^2x =$$

$$= A[\nabla s, \nabla s] - A[\nabla s, \nabla s] + \int_{\partial\omega^+} \sigma_{nn} A D^{ij} \frac{\partial s}{\partial x^j} n_i dx + \int_{\omega^-} \sigma^{kt} \dot{\epsilon}_{kt}^{(e)} d^2x =$$

$$= \int_{\partial\omega^-} \sigma_{nn} \dot{w}_n^{(p)} dx + \int_{\omega^+} \sigma^{kt} \dot{\epsilon}_{kt}^{(e)} d^2x \tag{C.2}$$

With (6.5) and (6.14) the elastic strains are expressed in terms of averaged strains, plastic strains and strains $\epsilon_{ij}(\boldsymbol{W})$ generated by field $\boldsymbol{W}$:

$$\epsilon_{ij}^{(e)} = \bar{\epsilon}_{ij} + \epsilon_{ij}(\boldsymbol{W}) - \epsilon_{ij}^{(p)}, \quad \epsilon_{ij}(\boldsymbol{w}) \equiv \frac{1}{2} \left(\frac{\partial w_i}{\partial x^j} + \frac{\partial w_j}{\partial x^i} \right),$$

$$\dot{\epsilon}_{ij}^{(e)} = \dot{\bar{\epsilon}}_{ij} + \epsilon_{ij}(\dot{\boldsymbol{W}}) - \dot{\epsilon}_{ij}^{(p)}. \tag{C.3}$$

The substitution of (C.3) into (C.2) yields

$$Q = \int_{\partial\omega^+} \sigma_{nn} \dot{w}_n^{(p)} dx + \int_{\omega^-} \sigma^{ij} \left(\dot{\bar{\epsilon}}_{ij} + \epsilon_{ij}(\dot{\boldsymbol{W}}) - \dot{\epsilon}_{ij}^{(p)} \right) d^2x =$$

$$= \int_{\partial\omega^-} \sigma_{nn} \dot{w}_n^{(p)} dx + \int_{\omega^-} \sigma^{ij} \dot{\bar{\epsilon}}_{ij} d^2x + \int_{\partial\omega^+} \sigma_{nn} \left(\dot{W}_n - \dot{w}_n^{(p)} \right) dx =$$

$$= |\omega^+| \, \bar{\sigma}^{ij} \dot{\bar{\epsilon}}_{ij} + \frac{1}{2} \int_{\partial\omega^+} \sigma_{nn} \left[\dot{W}_n \right] dx = |\omega^+| \, \bar{\sigma}^{ij} \dot{\bar{\epsilon}}_{ij}. \tag{C.4}$$

Boundary conditions (6.16) and (6.15) were used in the derivation (C.4). The Lemma 1 is proven.

Lemma 2.

let us assume that for $t \geq t^0$ macrostresses are equal to zero:

$$\bar{\sigma}^{ij}(t) = 0, \quad t \geq t^*, \tag{C.5}$$

Than for an arbitrary solution of the system P the following estimations hold:

$$I(t) \leq e^{-\beta(t-t^*)} I(t^*), \quad t \geq t^*. \ \beta = const > 0, \tag{C.6}$$

and the following components of the solution tend to zero:

$$s \to 0, \ \varepsilon_{ij}^{(e)} \to 0, \ \sigma^{ij} \to 0, \ \dot{\varepsilon}_{ij} \to 0, \ t \to 0. \tag{C.7}$$

Proof.

For $t \geq t^*$ the identity (C.1) is reduced to the following:

$$\frac{dI(t)}{dt} + A[\nabla s, \nabla s] = 0 \tag{C.8}$$

It follows from (C.5) and (6.27) that

$$\int\limits_{\partial\omega^+} s n^k n^k dx = \int\limits_{\partial\omega^+} s dx = 0, \ t \geq t^*. \tag{C.9}$$

Than the following inequalities hold [28]

$$\int\limits_{\omega^+} A s^2 d^2x \leq C_2 [\nabla s, \nabla s], \tag{C.10}$$

$$\int\limits_{\partial\omega^+} A^2 s^2 dx \leq C_3 [\nabla s, \nabla s], \tag{C.11}$$

Adding if necessary to the solution some field $V \in R$ we may modify the elastic solution so that at each moment t averaged over the cell ω^+ elastic displacements and rotation are zero:

$$\int\limits_{\omega^+} w_i^{(e)} d^2x = 0, \tag{C.12}$$

$$\int\limits_{\omega^+} \left(\frac{\partial w_1^{(e)}}{\partial x^2} - \frac{\partial w_2^{(e)}}{\partial x^1} \right) d^2x = 0 \tag{C.13}$$

Than it takes place [29]

$$\int\limits_{\partial\omega^+} \left(w_n^{(e)} \right)^2 dx \leq C_4 \left[\epsilon^{(e)}, \epsilon^{(e)} \right], \ C_4 = const. \tag{C.14}$$

Let us prove that

$$\left[\epsilon^{(e)}, \epsilon^{(e)} \right] \leq C_5 [\nabla s, \nabla s]. \tag{C.15}$$

With (5.9), (6.1) and (5.15) taken into account, the left side of (C.15) is reduced to

$$\left[\epsilon^{(e)}, \epsilon^{(e)} \right] = \int\limits_{\omega^+} \sigma^{ij} \epsilon_{ij}^{(e)} d^2x = \int\limits_{\partial\omega^+} \sigma^{ij} n_j w_i^{(e)} dx = \int\limits_{\partial\omega^+} \sigma_{nn} n_j w_n^{(e)} dx =$$

$$\int\limits_{\partial\omega^-} A s n_j w_n^{(e)} dx \le \sqrt{\int\limits_{\partial\omega^-} A^2 s^2 dx} \sqrt{\int\limits_{\partial\omega^-} \left(w_n^{(e)}\right)^2} \le C_6 \sqrt{\left[\epsilon^{(e)}, \epsilon^{(e)}\right]} \sqrt{[\nabla s, \nabla s]}. \tag{C.16}$$

The inequalities in (C.16) follow from (C.14), (C.11). Estimation (C.15) follows from (C.16).

Combining (C.15) and (C.10) we obtain

$$A\left[\nabla s, \nabla s\right] \ge \beta I(t), \quad \beta = const \tag{C.17}$$

It follows from (C.8) and (C.17) that

$$0 = \frac{dI(t)}{dt} + A\left[\nabla s, \nabla s\right] \ge \frac{dI(t)}{dt} + \beta I(t), \tag{C.18}$$

which results in the basic relation

$$I(t) \le e^{-\beta(t-t^*)} I(t^*) \Rightarrow I(t) \to 0, \quad t \to \infty. \tag{C.19}$$

The first three statements in (C.7) follow immediately from (C.18). Since elastic strains and vacancy concentration tend to zero, the same is true for elastic and plastic strain rates, if the solution of the system P is sufficiently smooth. Hence

$$\Delta_{ij} \equiv \dot{\varepsilon}_{ij} + \varepsilon_{ij}\left(\dot{W}\right) = \dot{\varepsilon}_{ij}^{(e)} + \dot{\varepsilon}_{ij}^{(p)} \to, \quad t \to \infty. \tag{C.20}$$

Using Levi-Chivita formulas, we obtain the following presentation:

$$\dot{W}_i = -\dot{\varepsilon}_{ij} x^j + \lambda e_{ij} x^j + a_i + T(\Delta_{ij}), \quad \lambda, a_i = const, \quad T(\Delta_{ij}) \to 0. \tag{C.21}$$

The second and third terms in (C.21) represent the rigid body displacement, the last term stands for Levi-Chivita integrals. Substitution of (C.21) into continuity condition (6.15) is done similar to evaluation (6.21) and yields

$$\dot{\varepsilon}_{ij} l^j(\boldsymbol{x}) n^i(\boldsymbol{x}) + T_n(\Delta_{ij}), \quad T_n(\Delta_{ij}) \to 0, \quad t \to \infty. \tag{C.22}$$

Since normal is constant over S_r, $r = 1.2.3$, the relation (C.22) provides three different conditions, which may be considered as a system of linear equations with respect to three components $\dot{\varepsilon}_{ij}$.The determinant of this system is not zero, and than it follows from (C.22) that $\dot{\varepsilon}_{ij} \to 0$, $t \to \infty$. Lemma 2 is proven.

References

[1] Nabarro, R. R. N., "Deformation of Crystals by the Motion of Single Ions," In *Report of a Conference on Strength of Solids*. The Physical Soe., 1948, pp. 75 – 90.

[2] Herring, C., "Diffusional Viscosity of a Polycrystalline Solid," *J. Appl. Phys.*, Vol. 21, 1950, pp. 437 – 45.

[3] Coble, R. L., "A Model for Boundary-Diffusion Controlled Creep in Polycrystalline Materials," *J. Appl. Phys.*, Vol. 34, 1963, pp. 1679 – 82.

[4] Lifshitz, I. M., "On the Theory of Diffusion-Viscous Flow of Polycrystalline Bodies," *Soviet Physics JETP*. Vol. 17, 1963, pp. 909 – 20.

[5] Poirier, J. P., *Creep of Crystals*, Cambridge University Press, 1985.

[6] Rabotnov, Yu., *Creep Problems in Structural Members*, John Wiley & Sons, 1969.

[7] Flynn, C. P., *Point Defects and Diffusion*, Clarendon Press, 1972.

[8] Agulo-Lopez, F., Catlow, C. R. A., and Townsend, P. D., *Point Defects in Materials*, Academic Press, 1988.

[9] Shesterikov, S. and Lokotshendo, A., "Creep and Long term strength of Materials," *Itogi Nauki*, Vol. 13, 1980.

[10] Crawford, J. H. and Slifkin, L. M., *Points Defects in Solids*, Plenum Press, 1975.

[11] Gruber, B., editor, *Theory of Crystal Defects. Proc. Summer School in Hrarany, 1964*. Academic Press, 1966.

[12] Ashby, M. F. S. and Veral, R. A., "Diffusion accommodated flow and super plasticity," *Acta Metall.*, Vol. 21, 1973, pp. 149 – 63.

[13] Murch, G.E. and Nowick, A.S., editors, *Diffusion in Crystalline Solids*. Academic Press, 1984.

[14] Christian, J. W., *The Theory of Transformations in Metals and Alloys*, Pergamon Press, 1975.

[15] Varotsos, P. A. and Alexopoulos, K. D., *Thermodynamics of Point Defects and Their Relation with Bulk Properties* , North-Holland, 1986.

[16] Friedel, J., *Dislocations*, Pergamon Press, 1964.

[17] Dottrell, A. H., *Dislocations and Plastic Flow inCrystals*, Oxford, 1958.

[18] Girifalco, L. A., *Statistical Physics of Materials*, J. Wiley & Sons, 1973.

[19] Kittel, C., *Introduction to Solid State Physics*, J.Wiley & Sons, 1976.

[20] Evans, H. E., *Mechanisms of Creep Fracture*, Elsevier Applied Sciences, 1984.

[21] Martin, J. W. and Doherty, R. D., *Stability of Microstructure in Metallic Systems*, Cambridge Univ. Press, 1976.

[22] Gomer, R. and Smith, C. S., editors, *Structure and Properties of Solid Surfaces.*, Univ. of Chicago Press, 1953.

[23] Rabotnov, Yu., *Mechanics of Deformable Solids*, 1965

[24] Gittus, J., *Creep, Viscoelasticity and Creep Fracture in Solids*, J. Wiley & Sons, 1975.

[25] Sedov, L., *Continuum Mechanics*, 1970

[26] Trusdell, C., *Continuum Mechanics*, 1960

[27] Berdichevsky, V., *Variational Prinicples of Continuum Mechanics*, Nauka, Moscow, 1983.

[28] Sobolev, S. L., *Applications of Functional Analysis in Mathematical Physics R.I., 1963*. volume 7. American Math. Society, Providence, R. I., 1963, Translation from Russian.

[29] Friedrichs, K. O., "On the boundary value problem of theory of elasticity and Korn's inequality," *Ann. Math.*, Vol. 48, No. 2, 1947.

DEVELOPMENT OF A FLUORESCENCE-BASED CHEMICAL SENSOR FOR SIMULTANEOUS OXYGEN QUANTITATION AND TEMPERATURE MESUREMENT IN FUELS

Steven W. Buckner
Assistant Professor
Department of Chemistry and Geology

Columbus College
Columbus, GA 31907

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

Columbus College

December 1995

DEVELOPMENT OF A FLUORESCENCE-BASED CHEMICAL SENSOR FOR SIMULTANEOUS OXYGEN QUANTITATION AND TEMPERATURE MEASUREMENT IN FUELS

Steven W. Buckner
Assistant Professor
Department of Chemistry and Geology
Columbus College

Abstract

A new method was developed for the measurement of temperature in liquids. The technique is based on the fluorescence lifetime of intermolecular excimers of pyrene-based probes (1,3-bis-(1'-pyrenyl)propane (BPYP) and 1,10-bis-(1'-pyrenyl)decane (BPYD)) as a function of temperature. Work during the contract period included construction of a time-resolved fluorimeter, measurement of fluorescence spectra of probe molecules in pure solvents and aviation fuel, generation of calibration curves for the temperature response of the probes, and studies of the probe in fuel cells at high temperature. Fluorescence spectra of the probes in fuel show the profound effect of background fluorescence on the relative intensities in the monomer and excimer regions that precludes steady-state measurements of real fuels. Calibration curves for the lifetimes as a function of temperature show the lifetime to be a linearly decreasing function over the range of 110°C to 220°C .

DEVELOPMENT OF A FLUORESCENCE-BASED CHEMICAL SENSOR FOR SIMULTANEOUS OXYGEN QUANTITATION AND TEMPERATURE MEASUREMENT IN FUELS

Steven W. Buckner

Introduction

The effects of temperature on physical, chemical and biological processes makes its measurement an important aspect of most experiments and simulations. The simplest thermometers such as thermocouples are also the least expensive but are not useful for many applications. These include applications where the sample is physically inaccessible to a temperature probe, very small samples volumes, and cases where the probe would perturb the experiment, such as a flowing liquid. To alleviate these problems a variety of optical thermometric probes have been developed.

Linear and non-linear Raman scattering methods have been developed which are based on lineshape analysis from which temperatures are extracted based on a Boltzman population analysis[1,2]. These techniques are typically quite expensive and are affected by background fluorescence from the sample. Another group of methods has been developed based on steady-state fluorescence measurements of probes which have temperature dependent spectra. Melton and coworkers[3], and Stuffelbeam[4]

have recently shown excimer fluorescence to be useful for temperature measurement. The fluorescence spectrum of pyrene shows a pronounced change in the excimer region (around 500 nm) as a function of temperature. The photochemical sequence for emission by the excimer is shown in reaction (6). After photoexcitation of the

$$A + h\nu \rightleftharpoons A^* + A \underset{k_2}{\overset{k_1}{\rightleftharpoons}} (A)_2^* \dashrightarrow 2A + h\nu \qquad (6)$$

monomer, collision of a ground state pyrene molecule with the excited fluorophore forms the excimer. The excimer can fluoresce (the excimer fluorescence is red shifted from that of the monomer) or back dissociate into the ground and excited state monomers. The equilibrium constant for the formation of the excimer ($k_{eq} = k_1/k_2$) is a decreasing function of temperature, as expected for an exoergic dimerization reaction. The diffusion rate in solution is an increasing function of temperature which increases the rate at which dimerization can take place, but the free energy dependence of the excimer formation is greater. Thus, with increasing temperature the integrated intensity of the excimer fluorescence decreases relative to the monomer fluorescence, as shown in Figure 1. Temperature is obtained from the ratio of the fluorescence intensity in the excimer region relative to that in the monomer region (I_e/I_m.) But there are serious drawbacks to this approach.

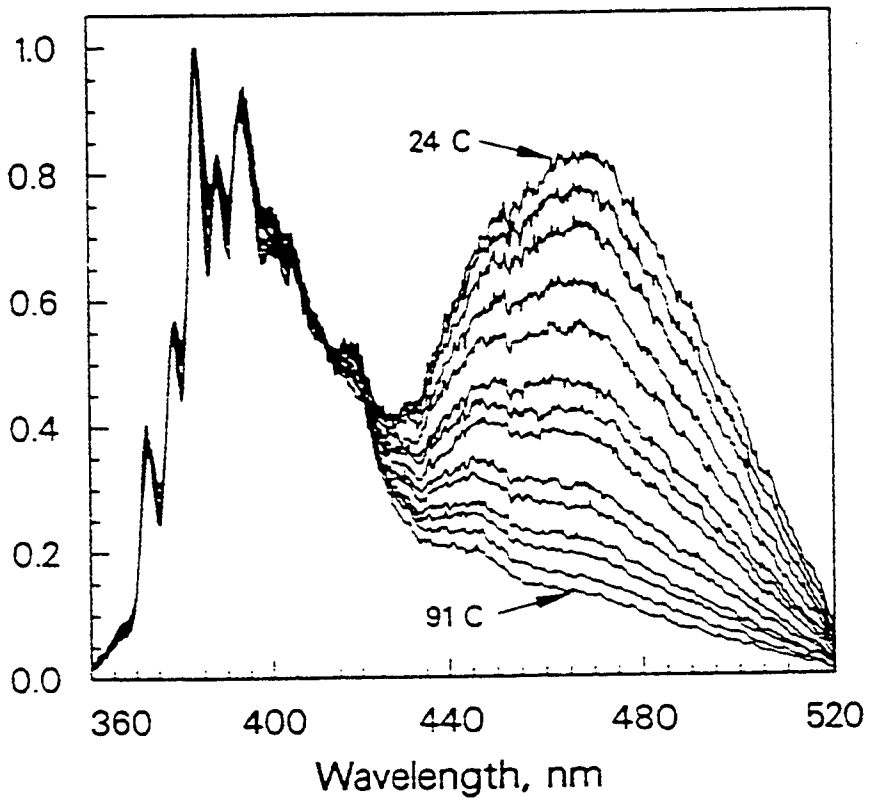


Figure 1. Temperature Dependence of the Pyrene Fluorescence Spectrum Showing the Excimer Region (taken from ref. 4).

An entire fluorescence spectrum must be obtained for each temperature. The spectrum is obtained in a time-independent fashion, so the fluorescence of the fuel and the probe will overlap. These techniques were demonstrated with pure solvents with no background fluorescence. Simple spectral subtraction will not be possible if thermal stressing of the fuel results in chemical reaction of the fluorophores in the fuel. Consumption or production of any of the fluorophores during stressing will result in time-dependent changes of the fuel background fluorescence spectrum. This will directly affect I_e/I_m which will yield an erroneous temperature.

Another problem arises with the use of intermolecular excimers of pyrene. The signal is a strong function of the concentration of pyrene (k_{eq} = [pyrene]^2/[pyrene dimer]). The temperature is computed as a ratio of the intensity of the excimer to monomer fluorescence, which depends on total pyrene concentration. Thus, the pyrene concentration must be carefully controlled. If pyrene is consumed or formed in any reaction during the observation time the measurement is invalid. Finally, the pyrene concentration must be maintained in the 0.1% to 10% (w/w) range. This high level of pyrene doping in the fuel is a concern and may affect other chemistry in the fuel. Melton has used the intramolecular excimer coupled with the above I_e/I_m ratio method to lower the concentration necessary to form the excimer. The intramolecular excimer approach has not found success however, as there are concentration effects on the ratio of I_e/I_m which have to date only allowed its use in

high purity hexadecane. Others solvents do not have significantly large regions where I_e/I_m is concentration independent and in commercially obtained hexadecane there are sufficient impurities that upon thermal stressing near 200°C spectral impurities appear making the fluorescence spectra nonreproducible.

Another fluorescence-based approach has been developed by Ben-Amotz and coworkers[5]. The fluorescence spectrum of N,N'-bis(2,5-di-tert-butylphenyl)-3,4,9,10-perylenedicarboximide (BDBP) shows shifts in the fluorescence maximum and in the intensity of the blue tail of the spectrum as a function of temperature. BDBP produces useful signal at ppb levels. However, this method is also performed in a time-independent manner and suffers from the problem of fluorescence background interference. Studies were only performed in high purity solvents and oils. Any background fluorescence from the sample would affect the intensity measurements. Background changes during thermal stressing would also shift the intensity ratios.

During the contract period, a new method for temperature measurement has been developed which avoids the problems encountered in the techniques discussed above. The method is based on the intramolecular excimer fluorescence from 1,3-bis-(1'-pyrenyl)propane (BPYP) and 1,10-bis-(1'-pyrenyl)decane (BPYD). The excimer fluorescence is strongly red shifted from that of the fuel[6], which allows for partial separation of the fuel and probe signals. The excimer also has a longer fluorescence lifetime than the fuel. Time-resolved measurement of the fluorescence intensity

allows for further separation of the fuel and probe.

Experimental Approach

All fluorescence measurements were made on a home-built fluorimeter described below[7]. A block diagram of the instrument is shown in Figure 2. Excitation of samples was accomplished using a N_2 laser (Laser Sciences VSL-337) with an output power of 120μJ/pulse, a pulsewidth of 3 nsec, and a maximum pulse rate of 20 Hz. A small portion (4%) of the pump beam is split off with a glass flat into a photodiode (Texas Instruments TIED-56) which is fast-wired to yield a 200 psec risetime[8]. The output from the photodiode initiates data acquisition by a Tektronix TDS 350 digital storage oscilloscope with a 1 GHz sampling rate and a 200 MHz analog bandwidth. After the glass flat the pump beam is focussed into a sample cell. For high temperature studies, the cell consist of a 250 ml round bottom pyrene reaction flask with a distillation column, a thermocouple for reference temperature measurement, a teflon stirrer to ensure uniform temperature throughout the cell, and a nitrogen gas inlet for degassing the liquid. All samples were degassed prior to study to prevent O_2 quenching of the excited state. Future studies will investigate the combined effects of temperature and oxygen concentration. The fluorescence emission from the sample is collected by a two lens system and coupled into an Optometrics DMC1-02 monochromator set to

pass 520 nm light. The fluorescent photons are then directed onto an RCA 931A photomultiplier tube which has been fast-wired to yield a rise-time of ~1.3 nsec[9]. 256 fluorescence decay curves were averaged for each point on the calibration curves shown in this work.

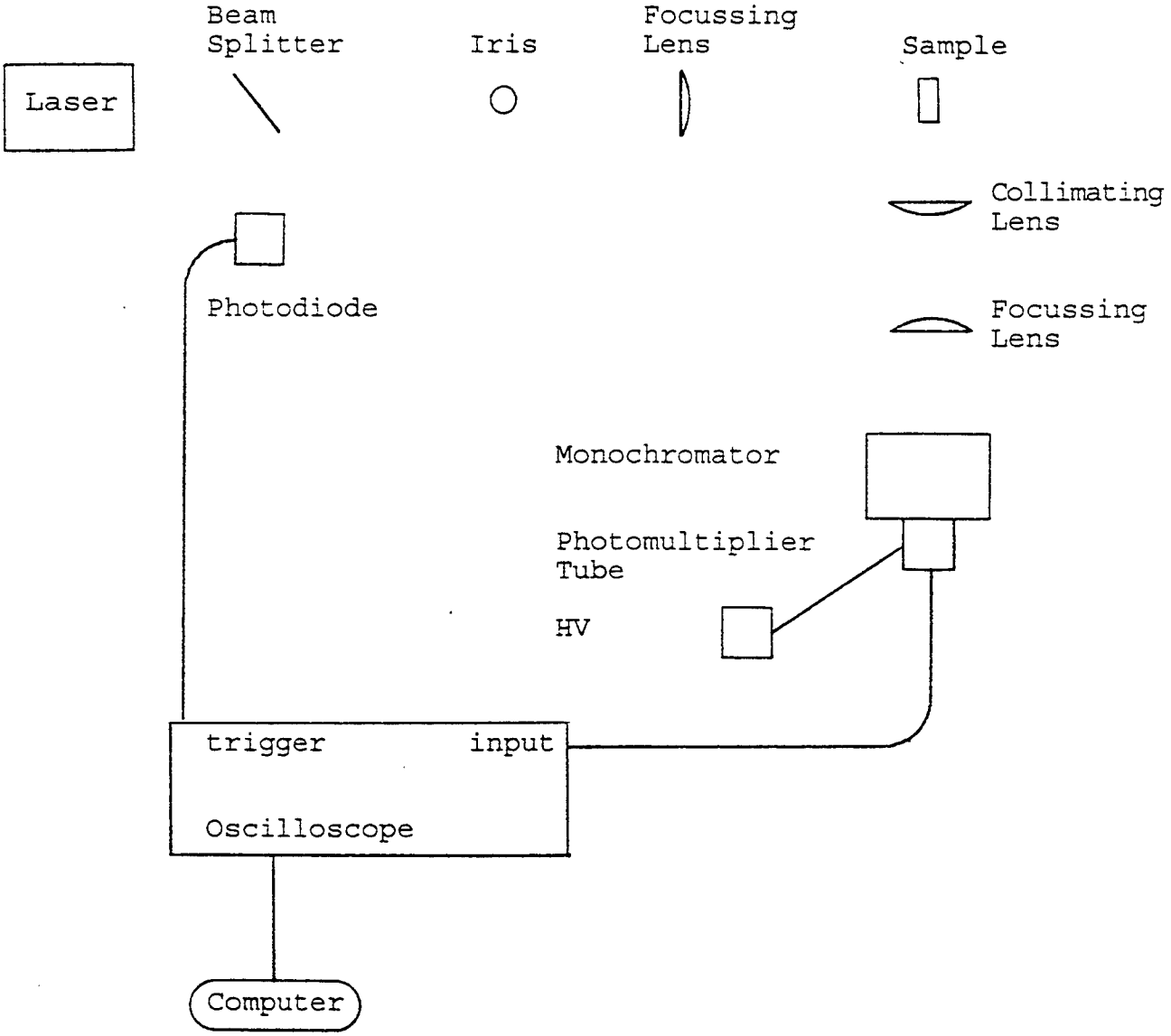


Figure 2. Diagram of the Time-Resolved Fluorimeter.

The fuel used in this work was POSF-2827, a Jet A aviation fuel. Dodecane and cyclohexane were obtained from Fisher Scientific and used without further purification. The pure hydrocarbons did not produce any detectable fluorescence. 1,3-bis-(1'-pyrenyl)propane (BPYP) and 1,10-bis-(1'-pyrenyl)decane (BPYD) were purchased from Molecular Probes and used without further purification. Dodecane was doped with 400 ppb BPYP for the lifetime studies. Cyclohexane and the fuel were doped with 5 to 10 ppm BPYP and BPYD for the fluorescence spectral studies.

Results and Discussion

The fluorescence emission spectra of BPYP and BPYD dissolved in cyclohexane are shown in Figures 3 and 4. The excitation wavelength is 337 nm. The high energy band with overlaying vibrational peaks corresponds to emission from the pyrene monomers. The peaks around 500 nm are for emission from the intramolecular excimers. The time-resolved emission in the excimer band of BPYP (dissolved in dodecane) is shown in Figure 5. The decay shows multiexponential behavior arise from competition between emission from the monomer and excimer and interconversion between the two forms. A kinetic scheme which accounts for this behavior, initially proposed by Zachariasse and coworkers, is shown below[10].

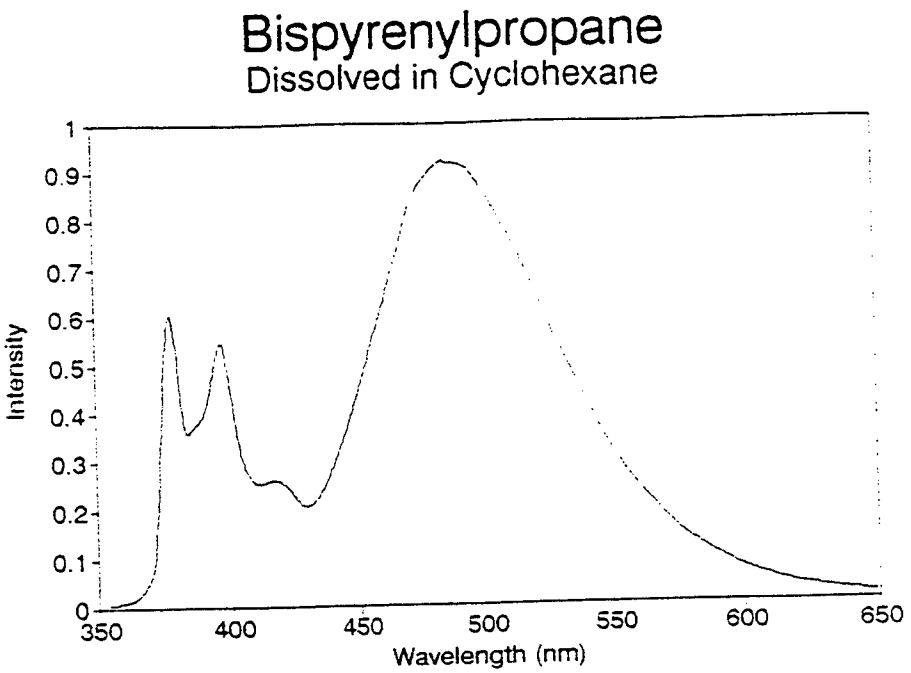


Figure 3. Fluorescence Spectrum of BPYP Dissolved in Cyclohexane.

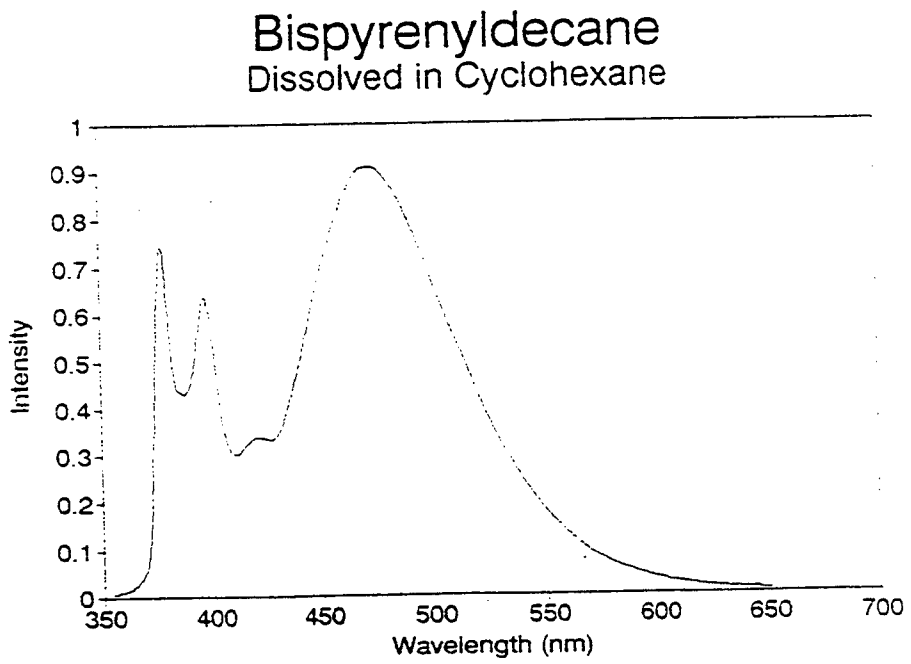


Figure 4. Fluorescence Spectrum of BPYD Dissolved in Cyclohexane

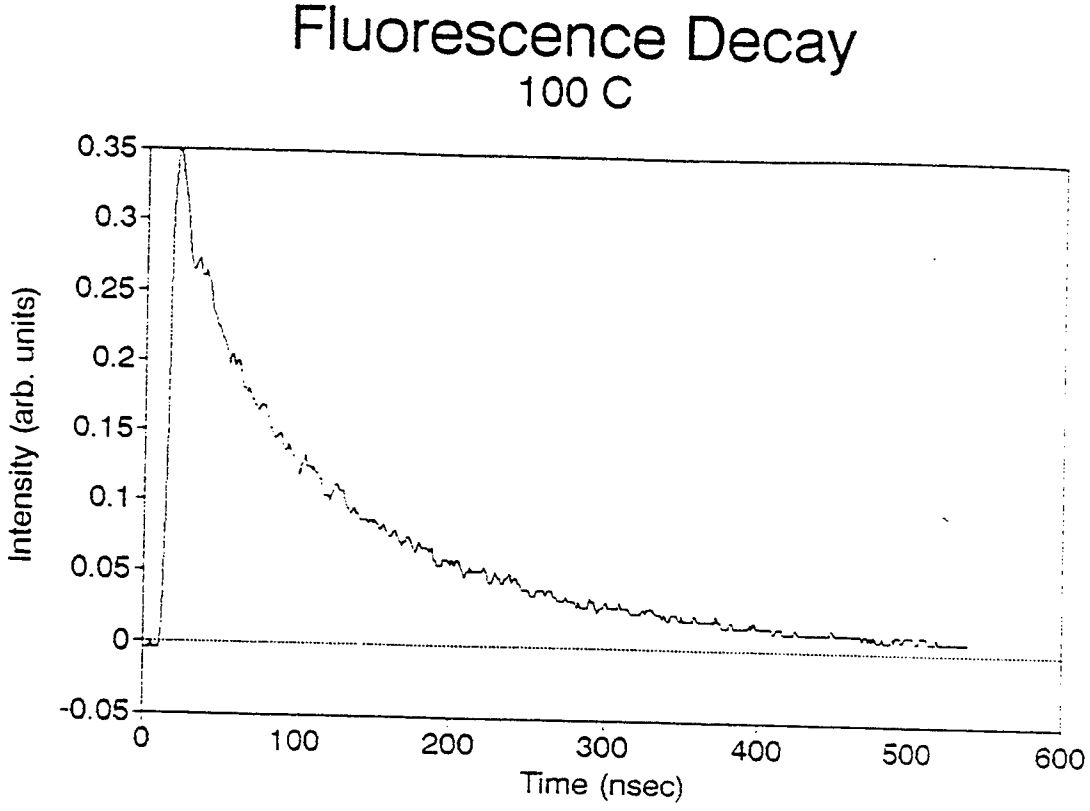
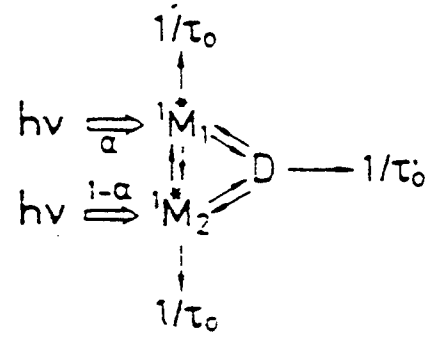
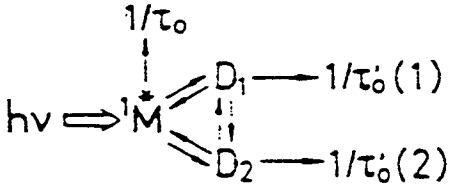


Figure 5. Time Resolved Emission from BPYP Dissolved in Dodecane (337 nm excitation wavelength, 520 nm emission wavelength)





All three decay constants are a function of viscosity and temperature. For the current experiments we are only concerned with the utility of the lifetime in temperature measurements. As a first step we will fit the intermediate portion of the fluorescence decay (between 40 and 200 nanoseconds, corresponding to k_2).

Figures 6 and 7 show the effect of temperature on the excimer lifetime fit between 40 and 200 nanoseconds. The lifetime is an increasing function of temperature up to 110°C where it reaches a maximum. Above 120°C the lifetime is a linearly decreasing function of temperature. The slope of the high temperature region is -0.5 nsec/°C. Similar behavior is observed for the ratio of the excimer fluorescence intensity to the monomer fluorescence intensity, as reported by Melton and coworkers[3(c)]. The increase in lifetime at low temperatures is caused by an increase in the rate of formation of the excimer from the excited monomer. At sufficiently high temperatures the monomer and excimer equilibrate. Further increases in temperature shift the equilibrium from the excimer back towards the dimer[10].

Figure 7 functions as a temperature calibration curve. There are a few points to consider regarding the utility of this curve. First, the intensity based excimer fluorescence approaches show similar functional forms for their calibration curves. The excimer relative intensity increases to a maximum near 100°C and then decreases linearly. Second, each temperature does not produce

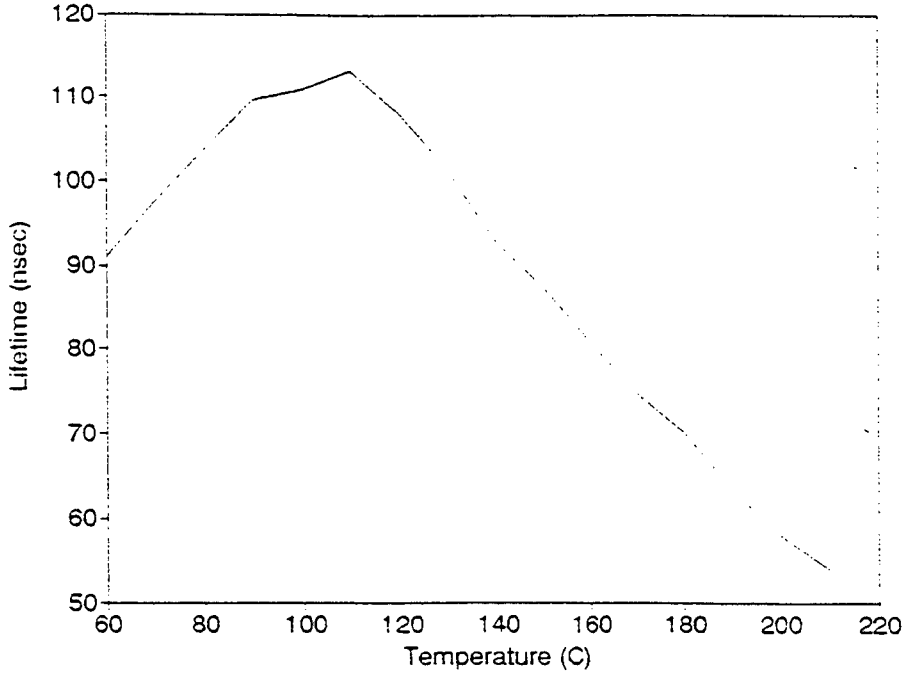


Figure 6. Fluorescence Decays for BPYP in Dodecane as a Function of Temperature.

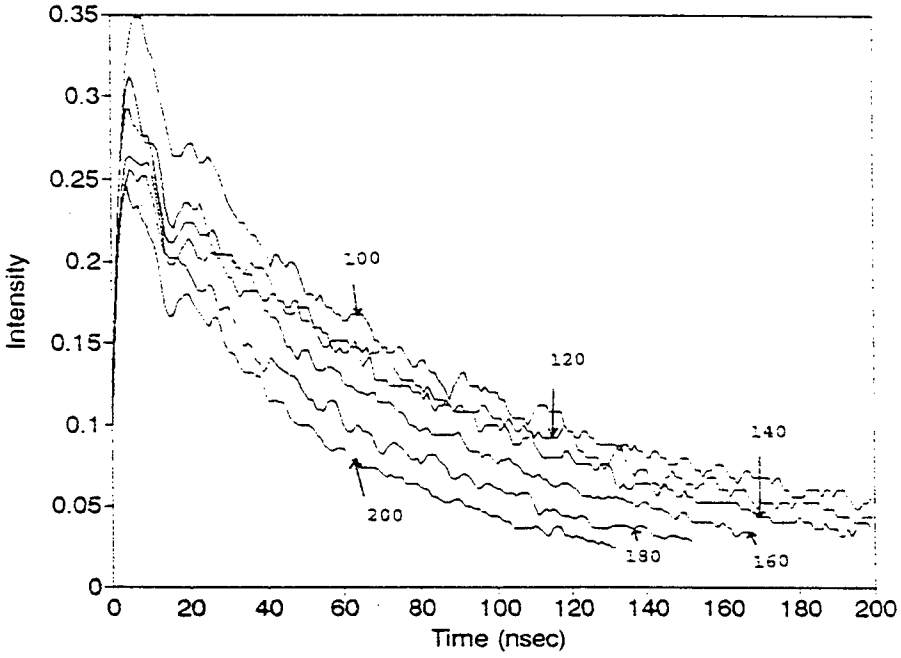


Figure 7. Lifetime of BPYP in Dodecane as a Function of Temperature

a characteristic lifetime. Thus, a lifetime of 105 nanoseconds can be produced from a sample with two different temperatures. This difficulty is only a problem at intermediate temperatures near 100°C, however. At higher temperatures there is a simple linear one-to-one relationship between temperature and lifetime. Third, there is a limitation to the upper temperature limit which can be measured on the current apparatus. The present fluorimeter has an instrument response function of 6 nanoseconds (FWHM). Assuming the linear lifetime-temperature relation continues out to high temperatures (an assumption which must fail beyond 320°C as the lifetime can never go negative), the excimer will have a lifetime on the same order as or shorter than the instrument response above 300°C.

Though the intensity and lifetime based approaches have similar calibration curves there are two significant advantages to the lifetime approach. First, the intensity based approach is dependent on the concentration of probe fluorophore. Thus, if the probe is consumed in thermal or photochemical reactions, the intensity will change at a constant sample temperature. In contrast, the lifetime is independent of probe fluorophore concentration in the range used in these experiments. Second, background fluorescence can dramatically affect the intensity-based approach. For example, Melton and coworkers report variations in the fluorescence spectrum before and after heating commercial hexadecane due to thermal reactions occurring at higher

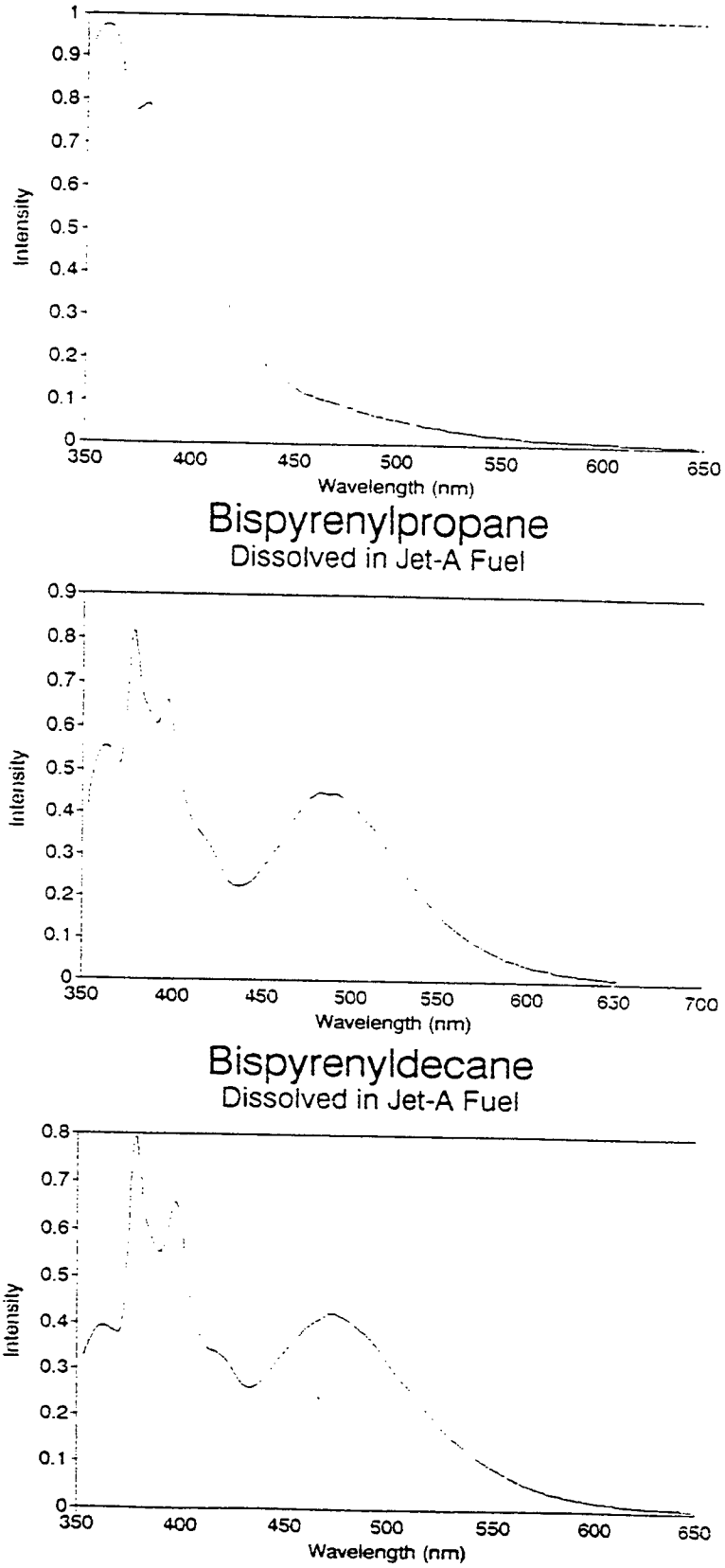


Figure 8. Fluorescence Spectra of (a) Pure Jet-A Fuel (b) BPYP Dissolved in Fuel (c) BPYD Dissolved in Fuel.

temperatures[3(b)]. Only after distillation of the solvent prior to study were reproducible results obtained. For the lifetime method this is no problem. Reproducible results were observed before and after heating with commercially used solvents requiring no distillation. The affect of background fluorescence can also be seen in the steady-state fluorescence spectrum of BPYP and BPYD in Jet-A fuel. Spectra for neat fuel and fuel with both probes (at a concentration of 5 ppm) are shown in Figure 8. These can be contrasted with the spectra of the fluorophores in cyclohexane shown in Figures 3 and 4. The emission of the fuel strongly overlaps with the monomer emission from the probes. Further, the high optical density of the fuel appears to decrease the total quantum yield for the probes dissolved in fuel. Finally, there are shifts in the fluorescence maxima from cyclohexane to fuel. These shifts might also occur upon thermal stressing of the fuel. The intensity-based approaches would clearly suffer in the presence of real fuels. It is conceivable that intensity based calibration curves could be prepared with the probe in fuel. However, upon thermal stressing and irradiation with ultraviolet light, the fluorescence spectrum of the fuel will change and prevent comparison to a standard calibration curve. In any event, there is no evidence that the intensity-based approach has any utility in anything except highly pure solvents. We have used the excimer lifetime-based approach in fuels during thermal stressing. Though the data have not been fit as of this writing, the data preliminarily appear to show a correlation between the excimer

lifetime and the fuel temperature. The very weak emission intensity of the fuel in the excimer region provides no significant interference.

Future Directions

The work presented here suggests the excimer lifetime method will be useful for optical temperature measurements. Future work will involve a comparison of all the decay constants for the excimer as a function of time and further studies in real fuel systems. A final goal is the simultaneous quantitation of oxygen and temperature using the three decay constants in the excimer emission.

References

1) The application of CARS to gas phase systems is discussed in: Attal-Tretout, B.; Bouchardy, P.; Magre, P.; Pealat, M.; Taran, J.P.; Appl. Phys. B, **1990**, 51, 17.

2) The application of linear Raman to temperature measurements is demonstrated in: Kip, B.J.; Meier, R.J.; Appl. Spectrosc., **1990**, 44, 707.

3) (a) Murray, A.M.; Melton, L.A.; Applied Optics, **1985**, 24, 2783. (b) Gossage, H.E.; Melton, L.A.; Applied Optics, **1987**, 26, 2256. (c) Hanlon, T.R.; Melton, L.A.; J. Heat. Trans., **1992**, 114, 450.

4) Stuffelbeam, J.H.; Appl. Spec., **1989**, 43, 274.

5) Schrum, K.F.; Williams, A.M.; Haerther, S.A.; Ben-Amotz, D.; Anal. Chem., **1994**, 66, 2788.

6) (a) Zachariasse, K.A.; Kuhnle, W.; Weller, A.; Chem. Phys. Lett., **1978**, 59, 375. (b) Smith, T.A.; Shipp, D.A.; Scholes, G.D.; Ghiggino, K.P.; J. Photochem. Photobiol. A: Chem., **1994**, 80, 151. (c) Zachariasse, K.A.; Duveneck, G.; Kuhnle, W.; Reynders, P.; Striker, G.; Chem. Phys. Lett.; **1987**, 133, 390.

7) Lakowicz, J.R.; "Principles of Fluorescence Spectroscopy",

Plenum:New York, 1983.

8) Haris, J.M.; Barnes, Jr., W.T.; Gustafson, T.L.; Bushaw, T.H.; Lytle, F.E.; Rev. Sci. Instrum., **1980**, 51, 988.

9) Harris, J.M.; Lytle, F.E.; McCain, T.C.; Anal. Chem., **1976**, 48, 2095.

10) Zachariasse, K.A.; Duveneck, G.; Busse, R.; J. Am. Chem. Soc., **1984**, 106, 1045.

CU-RACE:
Clarkson University Real-Time Acquisition and Control Environment

James J. Carroll
Assistant Professor
Department of Electrical and Computer Engineering

Clarkson University
Potsdam, NY 13699-5720

Final Report for:
Summer Faculty Research Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Wright Laboratory

December 1995

CU-RACE:
Clarkson University Real-Time Acquisition and Control Environment

James J. Carroll
Assistant Professor
Department of Electrical and Computer Engineering
Clarkson University

Abstract

The Clarkson University Real-Time Acquisition and Control Environment, CU-RACE, was designed for the simulation, real-time implementation, and analysis of real-world data acquisition and control systems. The software allows users to model systems in a variety of languages, implement systems on a variety of digital signal processing (DSP) hardware, perform on-line system parameter modifications, and record/display system data in real-time. These capabilities allow users to rapidly design and sophisticated data acquisition and control systems.

CU-RACE:
Clarkson University Real-Time Acquisition and Control Environment

James J. Carroll

1. Overview of the CU-RACE Environment

1.1 What is CU-RACE?

CU-RACE is a Microsoft Windows 3.x/95/NT based environment capable of simulating and implementing state-of-the-art user-defined data acquisition and control algorithms for real-world systems. Using the CU-RACE environment, an engineer can begin implementing signal processing or control algorithms without an extensive DSP programming background.

Within CU-RACE, all systems developed consist of module definition files and a binary project file. The module definition files are text based system models and control (signal processing) algorithms that describe the system operation. Each file describes either a continuous time-domain module, a discrete time-domain module, or the connection between module inputs and outputs. The number of possible continuous or discrete modules is limited only to the available memory in the PC and DSP system. The module definition files are written in a particular language which corresponds to an installed language library. The present version of CU-RACE installs a library utilizing a language developed for the simulation of nonlinear systems, called Simnon.

1.2 Interacting with CU-RACE

1.2.1 The Main Window

The main CU-RACE window, as seen in Figure 1.0, utilizes the Microsoft Windows multiple document interface (MDI) design. The environment window contains a menu, a toolbar for accessing certain menu items, a status bar for indicating system setting and status, and a client area for displaying module definition text windows and real-time graph windows.

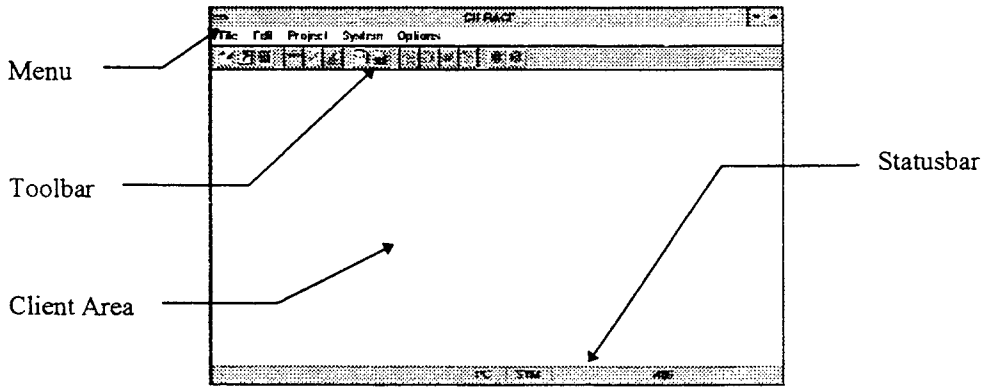


Figure 1.0: CU-RACE Environment

1.2.2 The Toolbar

To ease the building and running of user-defined system models, the toolbar provides a short-cut to commonly accessed menu items. These buttons include:

New module file

Open module file

Save module file

Cut selected text

Copy selected text

Paste clipboard at cursor

Open project file

Close current project

Data routing dialog

Modules to implement dialog

Set DSP I/O dialog

Change parameter on-line

Run current system

Stop current system

1.2.3 The Status Bar

The status bar, as seen in Figure 1.1, displays current system environment settings. It also provides brief descriptions of menu items when traversing menu lists and clicking toolbar buttons. When implemented systems are running, the status bar's "lost DSP buffer counter" displays the number of data buffers sent from the DSP that could not be accepted at that particular time, i.e., sometimes the environment is busy, due to overhead in logging and updating real-time in graphical displays, and cannot accept new (i.e., incoming) data. Note, having the operating system multitask too many programs within Windows can also cause CU-RACE to drop DSP data buffers. The line number display updates as users edit module definitions. Since system lines are indicated when parser errors occur, this display eases the module debug process. The system host, PC or DSP, indicates the location of the system when ran. The system mode, simulation or implementation, is also shown. The last item displayed, i.e., the system model, indicates the type of system host. With a DSP host for the system, the equipment manufacturer is displayed. The processor type is displayed for a PC host.

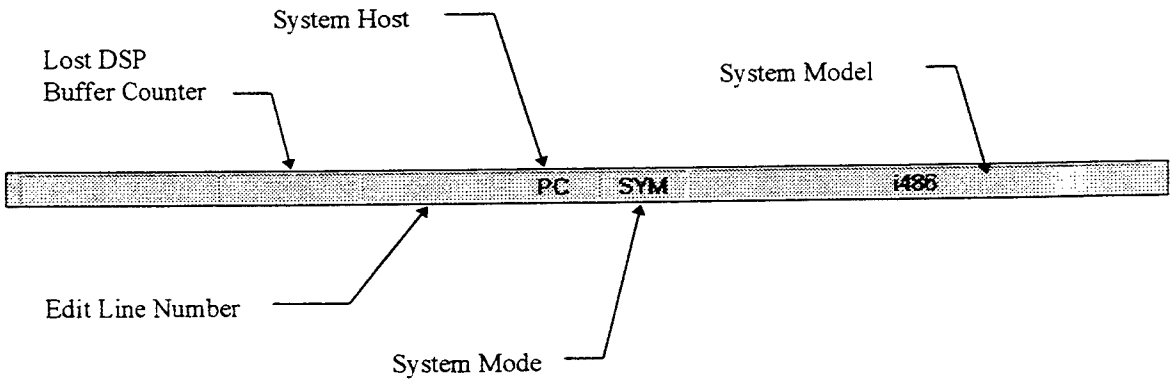


Figure 1.1: Environment Status Bar

1.3 Using CU-RACE

The CU-RACE flow chart in Figure 1.2 shows the steps necessary to build projects, simulate and implement systems on DSP hardware and log data. The steps outlined in the chart give the user manual section number containing the detailed description for that step.

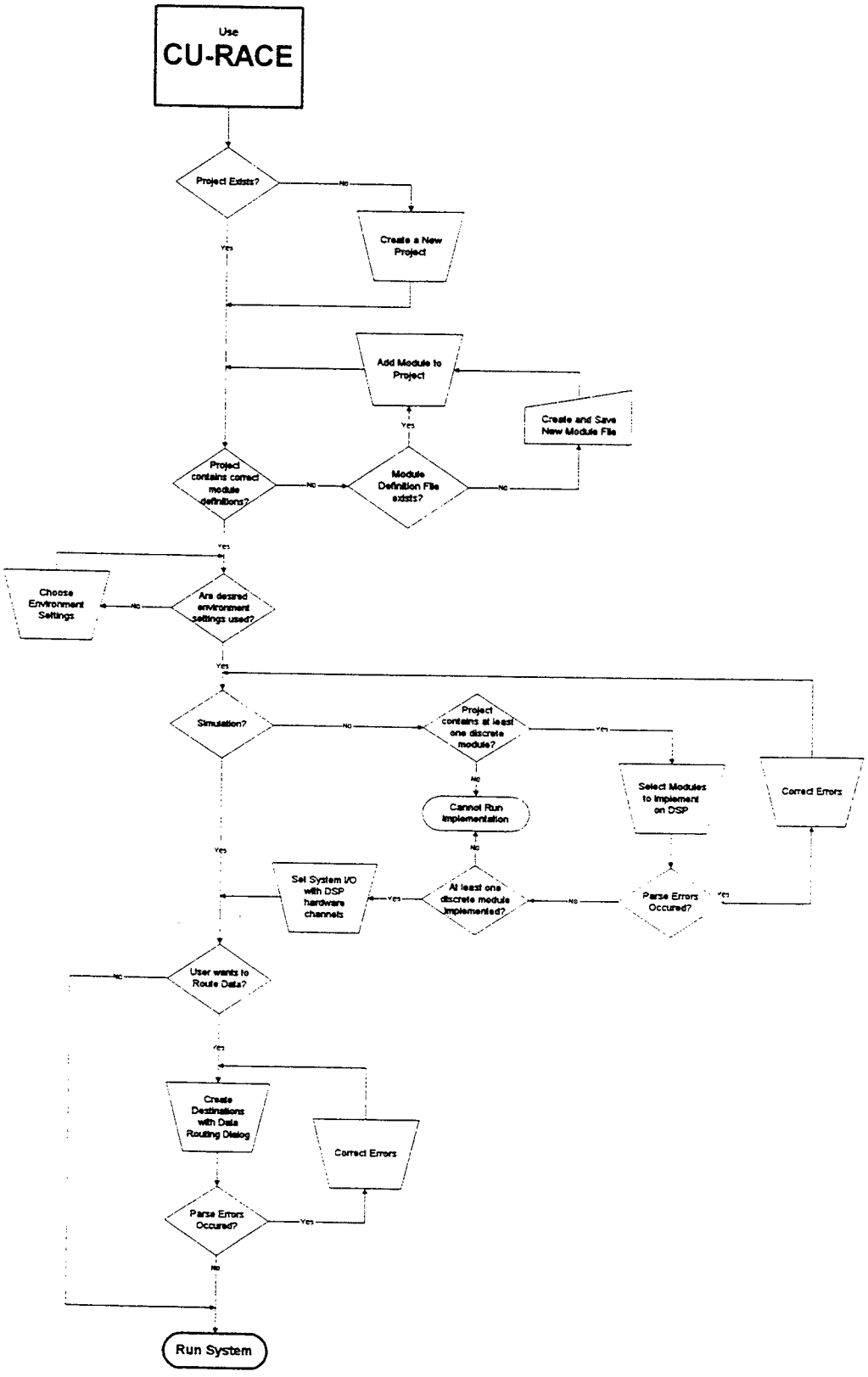


Figure 1.2: Flow Chart for using CU-RACE.

2. Working with Projects

2.1 Project Definition

A project contains all information about the system under test (SUT). Visually, the project displays a window similar to Figure 2.0, which displays the list of module definition files that describe the SUT. The project also contains the current CU-RACE system settings, such as environment settings, data logging settings, implemented discrete modules, and the DSP hardware and I/O channels associated with the SUT.

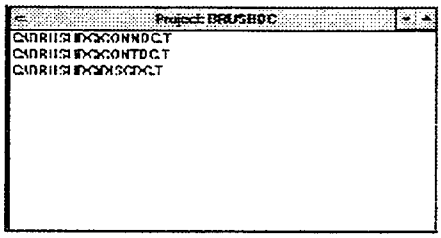


Figure 2.0: Project Window

2.2 Creating a Project

To create a project, select "Project | New" from the CU-RACE menu. The user is then prompted to select a name and location for the project. The default extension for project files is ".ILI". The location/directory of the new project will be used for storing the project file and any new data logging files created while working with this project.

To complete the project setup, the text-based module definition files describing the system components are added to the project. Select "Project | Add" for adding these files to the project. The modules added to the project can be stored in any directory. Selecting "Project | Remove" will remove the highlighted module definition from the project window of the current project. Note, this will not delete the file, only remove it from the project. Once added to the project, a module definition can be displayed in an edit window by double-clicking on the module in the project window.

Only one module describing the connections between continuous and discrete modules is permitted. During simulation studies, there must be at least one continuous or discrete module defined. Multiple module systems require a connection module which describes the connections between module inputs and outputs. For real-time DSP implementations, there is a minimum requirement of one discrete, one continuous, and one connection module in a project.

2.3 Opening and Closing Projects

Once created, closed projects can be opened by selecting either "Project | Open" from the CU-RACE menu or by clicking the button on the toolbar. Environment selections, data routing destinations, implemented modules, and DSP I/O channels are automatically set to the state at which the project was last run. The CU-RACE environment will close and save any currently open project in the event another project is opened or a new project is created.

When closing a project, CU-RACE will save all environment settings related to the project at that time. To close a project, select "Project | Close" from the menu or click the button on the toolbar. The project file is constructed and maintained by CU-RACE. Users are not advised to edit or alter project files directly, i.e., using an editor outside of CU-RACE, as uncorrectable file corruption can occur.

3. Environment Options

3.1 Environment Settings

How user-defined systems are built and run are defined in the "Environment Options" dialog shown in Figure 3.0. To open this dialog, select "Options | Environment" from the menu. The areas where adjustments are made to the environment are system mode, simulation and integration routine settings, active DSP equipment, current language parser, and the system logging rate.

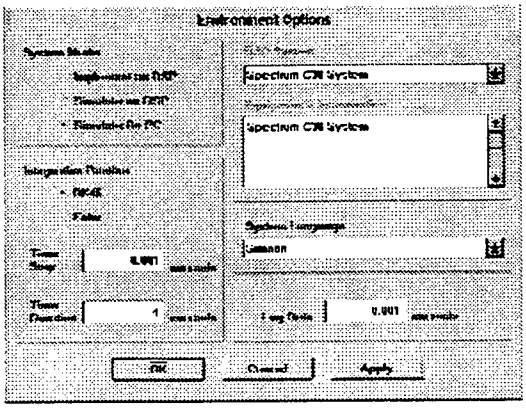


Figure 3.0: CU-RACE Environment Settings Dialog

3.2 System Modes

Use the system mode selection to select if the system is to be simulated on the PC or DSP equipment, or to have discrete components implemented on DSP equipment with hardware I/O capabilities. Select "Simulate on PC" in order to validate system models and/or algorithms. System simulation assists in debugging algorithms and confirms the user's understanding of physical systems before real-time DSP implementation is attempted.

3.3 Integration Routines

When running simulations, user's must set the continuous module integration routine, time step of simulation and duration of simulation. Euler and RK45 are the two possible methods for numerical integration within the CU-RACE environment, as described briefly below [1].

3.3.1 Euler

The Euler routine uses a simple single step method where system states y_n are updated according to the equation:

$$y_{n+1} = y_n + h_n f(y_n, t_n) \tag{3.0}$$

3.3.2 RK45

RK45, an integration routine with varying time step, is a more accurate technique than Euler but with a much higher cost in the number of calculations per simulation. Within each call to RK45, system states are updated by first calculating temporary system values, k, by

$$k_i = h_n f(y_n + \sum_{j=1}^{6} \beta_{ij} k_j, t_n + \alpha_i h_n) \qquad i = 1,...,6. \tag{3.1}$$

where α_i, β_{ij}, γ_i, and γ_i^* are constants defined as

α_i	β_{ij}					γ_i	γ_i^*
0						16/135	25/216
1/4	1/4					0	0
3/8	3/32	9/32				6656/1282 5	1408/2565
12/13	1932/2197	- 7200/2197	7296/2197			28561/564 30	2197/4104
1	439/216	-8	3680/513	-845/4104		-9/50	-1/5
1/2	-8/27	2	- 3544/2565	1859/4104	-11/40	2/55	0

The new state values are then found by calcutating

$$y_{n+1} = y_n + \sum_{i=1}^{6} \gamma_i k_i \tag{3.2}$$

where y_{n+1}, y_n, γ_i, k_i represent the new state value, current state value, constant, and previously calculated k term. A second state value, y^*, is calculated by

$$y_{n+1}^* = y_n + \sum_{i=1}^{6} \gamma_i^* k_i \tag{3.3}$$

in order to produce and error value, E, by

$$E = \sum_{i=1}^{6} (\gamma_i - \gamma_i^*) k_i . \tag{3.4}$$

Based on the difference of this error and the tolerance setting for RK45, the current time step is varied and another pass of the algorithm is started. If the number of passes through the algorithm exceeds 2000, the simulation will report an error. A system causing this reaction is too complex for an RK45 simulation. The system may be simulated by increasing the tolerance and re-running. But this may result in a less accurate solution of the system.

3.4 Time Step

The time step setting is used to define the incremental change of system time for each pass through the selected integration routine. It is also the accuracy for which discrete modules can be simulated.

3.5 Time Duration

A system will run for the time duration set in the environment. It is possible to stop a running simulation by selecting "System | Stop" or clicking the button on the toolbar.

3.6 DSP Selection

The DSP hardware used for real-time implementation is selected from the list of installed DSP equipment found on the top right corner of this dialog box. The window below this setting displays a brief description of the selected DSP system's hardware capabilities. If there is no DSP equipment installed in the PC, a "none" DSP selection and "Simulate on PC" mode is set is for the current system mode.

3.7 Language Libraries

The system language settings allows the user to select the library CU-RACE will use for reading module definitions. The selected library should correspond to the language used in creating the module definitions. A default library for the Simnon language is installed with the initial version of CU-RACE.

3.8 Log Period

The log period is used to control the time at which system data is recorded for files and/or real-time graph windows. This setting is used for both simulation and implementation of systems. When running simulations, it is necessary to choose a log period which is a common multiple of the time step.

4. Data Routing

In order to save or display system data values during simulation or DSP implementation it is necessary to define a "destination" to which the data is routed. To create or modify a destination, open the Data Routing dialog, shown in Figure 4.0, by selecting "System | Data Routing" from the menu or by clicking the button located on the toolbar.

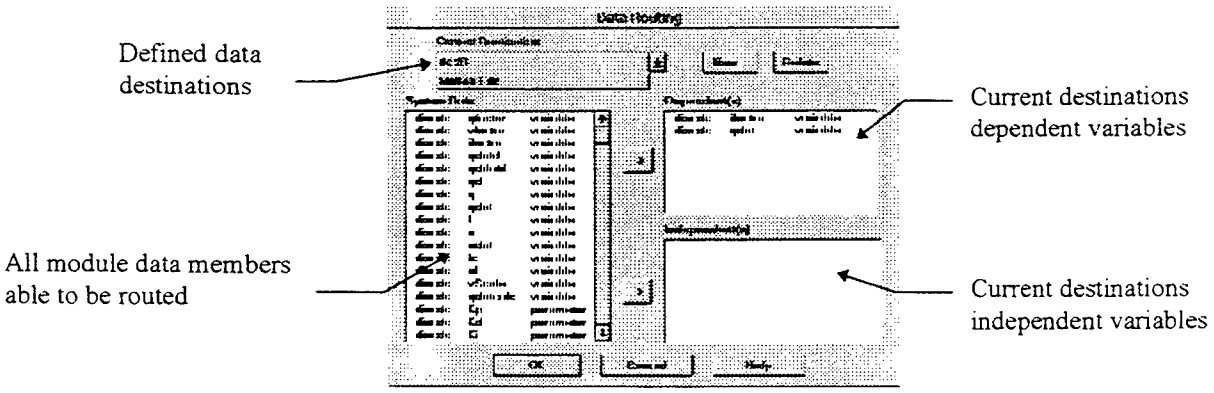


Figure 4.0: Data Routing Dialog

4.1 Creating a Destination

Routing data requires only a few simple steps. First, create a new destination by clicking New... on the dialog box. In response, the "New Destination" dialog of Figure 4.1 appears.

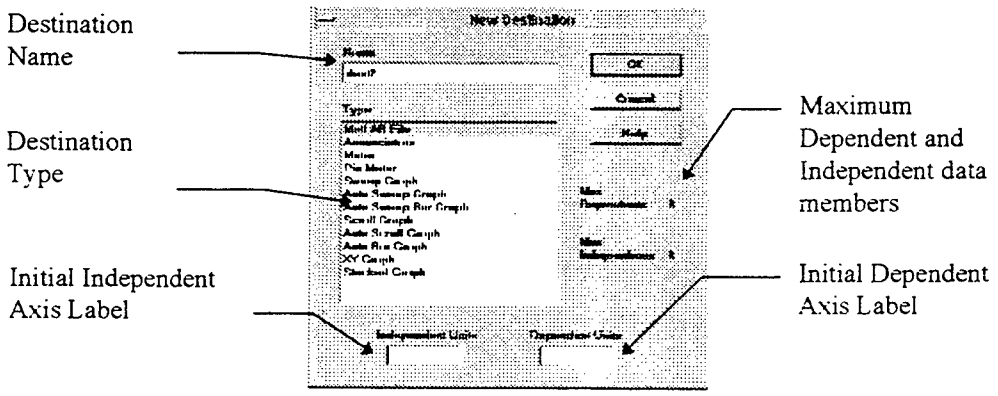


Figure 4.1: New Destination Dialog

The new destination name, type, and labels are set with this dialog. A default name, which can be modified, is assigned in the name control. It is not necessary to define axis labels for real-time graph destinations, the default is no name. There are 13 destination types to choose from. Figure 4.2 lists an example of each type.

On selecting the type, the maximum number of dependent and independent variables for that type is displayed to the right of the list. Only the Matlab and XY Plot allow user assigned independent data. All other types display their data with respect to the system time. For graph destinations, axis labels can be assigned to the graph by typing a label name in the boxes at the bottom of this dialog box.

After a new destination is defined, its name will appear in the current destination list at the top of the "Data Routing" dialog.

Destination Type	Max. Dependents	Maximum Independents	Display
Matlab Binary File	8	8	N/A
ANNUNCIATOR	8	0	
METER	8	0	
PIE METER	8	0	
SWEEP GRAPH	8	0	
AUTO SWEEP GRAPH	8	0	
AUTO SWEEP BAR GRAPH	8	0	
SCROLL GRAPH	8	0	
AUTO SCROLL GRAPH	8	0	
AUTO BAR GRAPH	8	0	
LOGIC GRAPH	8	0	
XY GRAPH	8	8	
STACKED GRAPH	8	0	

Figure 4.2: Destination Types

4.2 Deleting Destinations

If an error is made when creating the new destinations name, type or labels, the destination can be removed by pressing the Delete button.

4.3 Assigning Module Data to Destinations

The final step in creating a destination is to assign module data to the dependents and/or independents list. To do so, select the desired data or group of data from the "System Data" list and press the -> key next to the dependent or independent list. The dependents and independents list will not

allow data members to be assigned if the maximum number of items has been reached for that destination type. If an assignment is made in error, double-clicking the module data within the dependents or independents list will remove that selection.

5. Implementing Modules

5.1 What does it mean?

In CU-RACE, implementing a module means running that module on DSP hardware. This implies giving CU-RACE the following information: (1) the discrete module definitions to be executed on the DSP, and (2) which hardware I/O channels to be used by the modules for data I/O. The DSP code generation and hardware communication is performed automatically by CU-RACE and its installed DSP libraries.

5.2 Setting Modules as Implemented

To set a defined discrete module as "implemented", press the button found on the toolbar. The "MODULES TO IMPLEMENT" dialog, shown in Figure 5.0, allows the user to select which defined discrete modules are to be implemented on the DSP hardware.

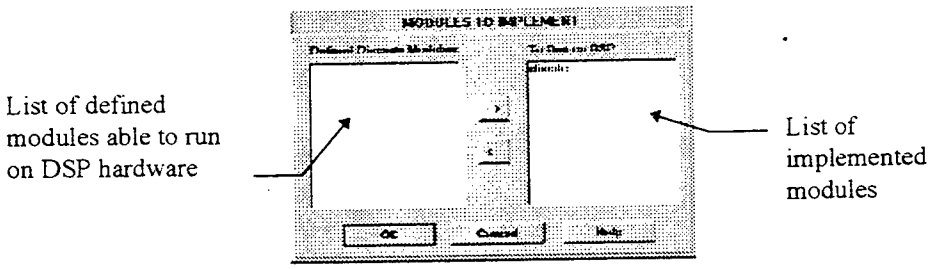


Figure 5.0: Modules-to-Implement Dialog

In order to have a discrete module implemented, select that module (or group of modules) in the "Defined Discrete Modules" list then click the button. This moves the selected module(s) to the "To Run on DSP" list. Use the button to move implemented module(s) to a non-implemented state.

6. DSP I/O Assignments

6.1 Interacting with DSP Hardware

Modules being implemented on DSP equipment (i.e., implemented modules) must have their input and output variables assigned to other implemented modules input/output variables via a connecting module or to physical hardware I/O channels.

When implemented modules have connections to other implemented modules, the connect becomes internal to the DSP hardware and does not require a hardware I/O setting. For implemented modules having connections with continuous modules or other non-implemented discrete modules, the user must assign connections between that module variables and DSP I/O channels.

Module input data will have units according to the type DSP channel. For example, a typical encoder used for measuring rotational angle will route data with units of counts. If a discrete module definition requires data in units of radians from such an encoder channel, that channel's associated input data must first be converted to radians based on the number of counts per revolution of that encoder. It is the user's responsibility to perform all necessary unit conversions on defined module I/O data.

6.2 Assigning Module Data to DSP Channels

To assign implemented module variables to physical hardware channels, select "System | Set DSP I/O" or press the button on the toolbar. The SET I/O dialog shown in Figure 6.0 appears.

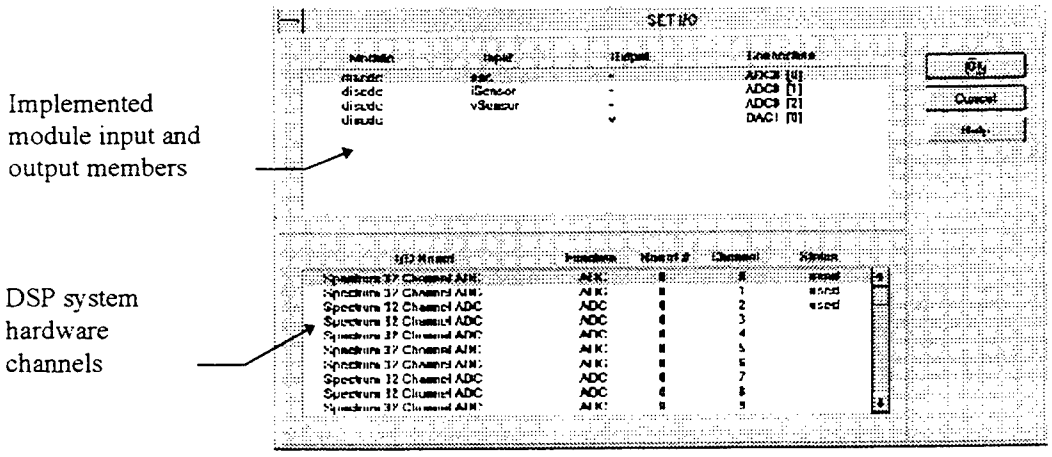


Figure 6.0: SET I/O Dialog Box

To create a connection, select a channel from the DSP hardware list and then double-click the data item that you want to associate with the selected channel. Once the connection is created, the data item displays the associated function, board number, and channel in its "Connection" column. Likewise, the newly associated DSP channel displays "used" in its "status" column. Input data members only make assignments with DSP channels of function ADC, ENC, and DIN type. Output data members can only make assignments with DSP channels of function ADC and DOUT type.

If you make a connection error, double-clicking on the associated DSP channel removes the connection. Also, previously assigned DSP channels, if selected, are reassigned if you double-click on other module data.

7. On-Line Parameter Changes

7.1 Why Change Parameters?

The ability to alter values of system data items is often desirable when simulating or implementing a system. This can aid the user by adjusting the system model or control algorithm quickly. To do so, click the button on the toolbar to open the "Modules Parameters" dialog, shown in Figure 7.0.

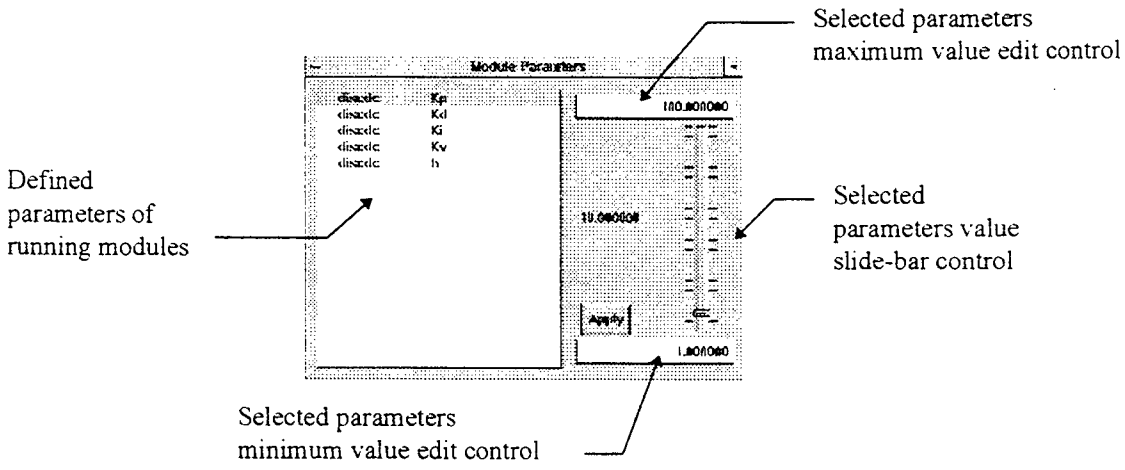


Figure 7.0: Module Parameters Dialog

7.2 Changing Parameter Values

To change the current value of a parameter, select that parameter from the list and then increase or decrease the current value by moving the thumb control, ▭, on the slide bar. The new value is displayed just left of the slide-bar. Once the slide-bar control is in focus, you can use the arrow up, arrow down, page up, and page down keys to cause incremental value changes.

You can change the maximum and minimum value limits of a parameter by editing the edit controls just above and below the slide-bar. You must click the "Apply" button in order to put the new limits into effect. If you apply a new maximum limit that is lower than the current value, the new maximum is set to the current value; likewise, a new minimum value greater than the current value is set to the current value.

Appendix A: CU-RACE Installation

In order to begin working with the CU-RACE program on Windows 3.x/95/NT operating systems, follow these few simple steps:

1. Create a new directory, such as C:\CURACE, off of the root directory of the hard drive.
2. Copy curace.exe, cuerror.dll, cusys.dll, and dspsys.dll into the C:\CURACE directory.
3. Copy qcbasef.dll and qcrtf.dll into the C:\CURACE directory.
4. Since the Simnon language parser is included with this version, copy its file, cusimnon.dll, into the C:\CURACE directory.
5. Create an ASCII text file, named cu-race.ini, containing the following text:

```
[DSP SYSTEMS]
LANG1=C:\CURACE\CUSIMNON.DLL
CURLANG=1
```

This file initializes CU-RACE with the Simnon language library. At least one language library must be installed and setup before you can use CU-RACE. Following the first running of CU-RACE, additional environment settings will be automatically added to this initialization file. After completing these steps, you can design and simulate systems within CU-RACE.

To add additional language libraries, add to the cu-race.ini file:

LANG# = *language library path and filename*
where # is the integer number of the language library

The numbers for all listed libraries should begin at one and increment by one. For example, after you have the Simnon library installed, adding an ACSL language parser would be as simple as copying the new library into the C:\CURACE directory and adding to the cu-race.ini file:

```
LANG2=C:\CURACE\CUACSL.DLL
```

To implement systems on DSP equipment, follow the appropriate installation instructions included with the CU-RACE DSP Library. Similar to adding new language libraries, the DSP libraries will add to the cu-race.ini file:

SYSTEM# = C:\....*library name*
CURSYS = #

PHASE ANGLE ZERO

David B. Choate
Associate Professor
Department of Mathematics

Transylvania University
Lexington, KY 40508

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

Transylvania University

December 1995

PHASE ANGLE ZERO

David B. Choate
Associate Professor
Department of Mathematics

Abstract

A solution is given for the classical equation

$$z(t) = \ln[A\cos(w_1 t) + B\cos(w_2) + C] \ .$$

PHASE ANGLE ZERO

David Choate

I. The Solution For Phase Angle Zero .

a. The frequency w_2.

To solve the classical problem of determining the frequencies w_1 and w_2 given that

$$z(t) = \ln[A\cos(w_1 t) + B\sin(w_2) + C] \qquad (1.1)$$

we begin by setting t = 1. Then

$$e^{z(1)} = A\cos(w_1) + B\sin(w_2) + C \qquad (1.2)$$

If t = -1, then

$$e^{z(-1)} = A\cos(w_1) - B\sin(w_2) + C \qquad (1.3)$$

Subtracting (1.3) from (1.2) gives

$$e^{z(1)} - e^{z(-1)} = 2B\sin(w_2) \ . \qquad (1.4)$$

If we let t = 2 in (1.1), then

$$e^{z(2)} = A\cos(2w_1) + B\sin(2w_2) + C \qquad (1.5)$$

If we let t = -2 in (1.1), then

$$e^{z(-2)} = A\cos(2w_1) - B\sin(2w_2) + C \qquad (1.6)$$

Subtracting (1.6) from (1.5) yields

$$e^{z(2)} + e^{z(-2)} = 2B\sin(2w_2). \qquad (1.7)$$

Recall the identity

$$\sin(2w_2) = 2\sin(w_2)\cos(w_2) \ . \qquad (1.8)$$

Substituting (1.8) into (1.7) gives

$$e^{z(2)} - e^{z(-2)} = 2\cos(w_2)[2B\sin(w_2)] \qquad (1.9)$$

Substituting (1.4) into (1.9) gives

$$e^{z(2)} - e^{z(-2)} = [2\cos(w_2)[e^{z(1)} - e^{z(-1)}] \qquad (1.10)$$

Assuming that w_1 and w_2 are distinct and that $w_2 \neq n\pi$, n an integer, we may divide to obtain

$$w_2 = \cos^{-1}\{(e^{z(2)} - e^{z(-2)})/[2(e^{z(1)} - e^{z(-1)})] \qquad (1.11)$$

b. The frequency w_1.

Adding (1.2) and (1.3) gives

$$e^{z(1)} + e^{z(-1)} = 2A\cos(w_1) + 2C \qquad (1.12)$$

Letting t = 0 in (1.1) yields

$$C = e^{z(0)} - A \qquad (1.13)$$

Substituting (1.13) into (1.12) gives

$$e^{z(1)} + e^{z(-1)} = 2A\cos(w_1) + 2[e^{z(0)} - A] \qquad (1.14)$$

Adding (1.5) and (1.6) and then substituting in (1.13) gives

$$e^{z(2)} + e^{z(-2)} - 2e^{z(0)} = 2A[\cos(2w_1) - 1] \qquad (1.15)$$

Recall the identity

$$\cos(2w_1) = 2\cos^2 w_1 - 1 \tag{1.16}$$

Substituting (1.16) into (1.15) gives

$$e^{z(2)} + e^{z(-2)} - 2e^{z(0)} = 4A[\cos^2 w_1 - 1] \tag{1.17}$$

Solving (1.14) for A gives

$$A = [e^{z(2)} + e^{z(-2)} - 2e^{z(0)}]/[2(\cos w_1 - 1)] \tag{1.18}$$

provided that $w_1 \neq n\pi$, n an integer.

Substituting (1.18) into (1.17) gives

$$e^{z(2)} + e^{z(-2)} - 2e^{z(0)} = 2[e^{z(1)} + e^{z(-1)} - 2e^{z(0)}](\cos w_1 + 1) \tag{1.19}$$

And so

$$w_1 = \cos^{-1}\{[e^{z(2)} + e^{z(-1)} - 2e^{z(0)}]/2[e^{z(1)} + e^{z(-1)} - 2e^{z(0)}]\} \qquad (1.20)$$

provided w_1 and w_2 are distinct and that $w_1 \neq n\pi$.

c. The constants A, B and C .

Substituting (1.20) into (1.14) gives

$$A = \frac{[e^{z(1)} + e^{z(-1)} - 2e^{z(0)}]}{[e^{z(2)} + e^{z(-2)} - 4e^{z(1)} - 4e^{z(-1)} + 6e^{z(0)}]} \qquad (1.21)$$

Substituting (1.21) into (1.13) yields

$$C = e^{z(0)} - \frac{e^{z(1)} + e^{z(-1)} - 2e^{z(0)2}}{e^{z(2)} + e^{z(-2)} - 4e^{z(1)} - 4e^{z(-1)} + 6e^{z(0)}} \qquad (1.22)$$

Solving (1.2) for B gives

$B = [e^{z(1)} - A\cos w_1 - C]/\sin w_2$ where

A, C, $\cos w_1$ are found through equations

(1.21), (1.22) and (1.11) and where

$$\sin w_2 = \sqrt{\{1 - [e^{z(2)} - e^{z(-2)}]^2/[4[e^{z(1)} - e^{z(-1)}]^2\}}$$

can be derived from (1.11).

II. The General Solution

i. Negative time values were used to avoid ghastly calculations in the derivation. But the values -2, -1, 0, +1, +2 may be replaced by any five consecutive; for example, 1, 2, 3, 4, 5 .

ii. After deriving similar expression for fixed nonzero phase angels, it may be possible to guess the formula given an arbitrary phase angle.

SYNTHESIS OF A NOVEL SECOND ORDER NONLINEAR OPTICAL MATERIAL

Stephen J. Clarson
Associate Professor
Department of Mechanical Engineering

University of Cincinnati
Cincinnati, OH 45221

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

University of Cincinnati

December 1995

SYNTHESIS OF A NOVEL SECOND ORDER NONLINEAR OPTICAL MATERIAL

Stephen J. Clarson & Lawrence L. Brott
Department of Materials Science and Engineering
University of Cincinnati

ABSTRACT

Synthesis of second order nonlinear optical (NLO) polymers represents an exciting field with the resulting chromophore containing materials being used in such devices as frequency doublers or electro-optical computers. In this research, a novel NLO chromophore was developed by incorporating two fluorene molecules in its backbone with thiophene and pyridine end groups that act as electron donating and withdrawing groups respectively. Long alkyl chains were attached to the C-9 carbons on the fluorene backbone to aid in the chromophore's solubility in the host polymer.

SYNTHESIS OF A NOVEL SECOND ORDER NONLINEAR OPTICAL MATERIAL

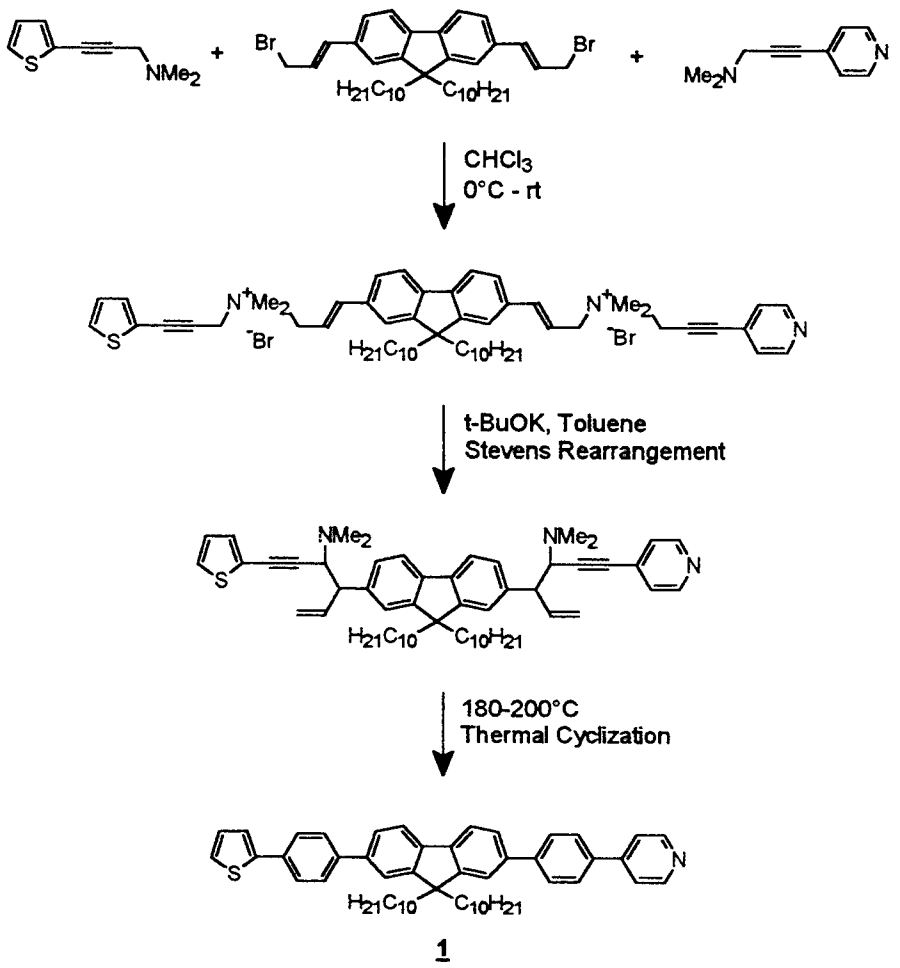
Stephen J. Clarson & Lawrence L. Brott
Department of Materials Science and Engineering
University of Cincinnati

1.0 INTRODUCTION

1.1 PREVIOUS RESEARCH

During the summer of 1994 at WL/MLBP in Dayton, OH (RDL/AFOSR Summer Program 1994), one of our goals was to synthesize a second order nonlinear optical chromophore **1** based on the fluorene molecule (see figure 1) [1,2]. This chromophore was designed to be thermally stable by having thiophene and pyridine rings as electron donor and acceptors, respectively, on opposite sides of the molecule. In addition, these rings were separated by four phenyl rings to increase the nonlinearity. Finally, two long alkyl chains were added to aid in solubility in common organic solvents.

The synthesis of this molecule was accomplished by reacting a thiophene propargyl amine and a pyridine propargyl amine with a bis(allyl bromide)fluorene to form a salt (see figure 1) [3]. The bonds were then rearranged with t-BuOK in a procedure known as a Stevens rearrangement. The chromophore was finally completed with a thermal cyclization with an overall yield of about 10%. The disappointing yield was due to the several combinations the amines could react with the fluorene in addition to the several possibilities of bond rearrangement.



1

Figure 1 Synthesis of second order nonlinear optical chromophore **1**.

1.2 CARBON-CARBON BOND FORMATION

After reviewing the results, it was determined that a new approach was needed. Instead of creating phenyl rings through thermal cyclization, it was decided to simply link rings together with carbon-carbon coupling. These couplings can be accomplished by either a Grignard reaction or by one of two types of palladium catalyzed tin reactions (see table 1).

GRIGNARDS:									
TECHNIQUE 1:	R—MgBr	+	Br—R'	⟶	R—R'				
TRIBUTYL TIN:									
TECHNIQUE 2:	R—Br	+	$(Bu)_3SnSn(Bu)_3$	$\xrightarrow{Pd}$	R—Sn(Bu)$_3$	+	Br—R'	$\xrightarrow{Pd}$	R—R'
TECHNIQUE 3:	R—Li	+	$(Bu)_3SnCl$	⟶	R—Sn(Bu)$_3$	+	Br—R'	$\xrightarrow{Pd}$	R—R'

Table 1 Carbon-carbon coupling techniques.

When performing the first technique, an arylbromide is stirred with magnesium shavings to produce the Grignard, which is then reacted with a different arylbromide to form a carbon-carbon bond [4,5,6]. However, when performing a tributyltin-coupling, two different methods can be used. When starting with an arylbromide, hexabutylditin must be stirred with a palladium catalyst to create an aryl-tributyltin. This product is further reacted with another arylbromide with additional palladium to create the desired carbon-carbon bond [7,8]. However, a simpler method is to start with a lithiated ring and simply stir it with tributyltinchloride to produce an aryl-tributyltin compound [9]. This third technique is more desirable than the second since palladium is used only once and there is no risk of forming a diaryl byproduct during the first step of the coupling reactions.

Even though each of these procedures yield the desired products in a relatively few steps, there are some complications that need to be addressed. For example, Grignard's require that all starting materials and glassware be extremely dry; any moisture in the solvent will easily quench the reaction. In addition, great amounts of heat can be generated and one must be careful not to let the reaction get out of control. Although the tin reactions are more tolerant of water, they are toxic and should be handled carefully. Unlike the Grignard's which should be used as soon as they are made, tin compounds can be stored relatively easily, and some can even be purified by column chromatography.

1.3 GENERAL SCHEME

It was decided that instead of trying to synthesize compound 1 again, it would be easier to make a similar chromophore 5 with even greater solubility (see figure 2). The synthesis of this molecule could be broken down into three steps: 1) coupling a thiophene ring with a fluorene molecule, 2) coupling a pyridine ring to a fluorene molecule, and 3) joining the resulting products together to produce the desired chromophore 5. With this in mind, the initial task was to complete the thiophene-fluorene molecule.

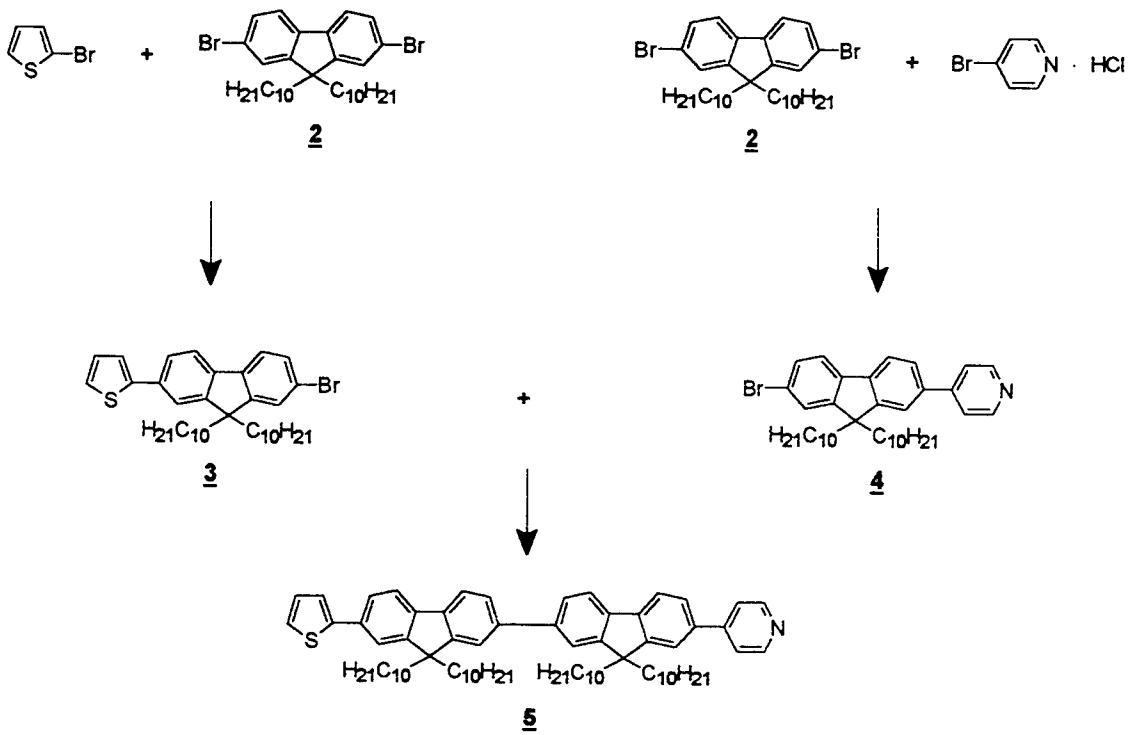


Figure 2 General plan for the synthesis of the chromophore 5.

2.0 THIOPHENE / FLUORENE COUPLING

2.1 EARLY EXPERIMENTS

Research began by making a Grignard from 2-bromothiophene and trying to react that with dibromofluorene 2 (see figure 3).

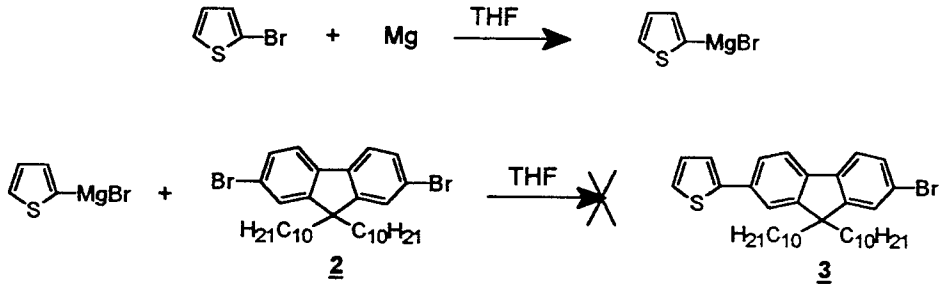
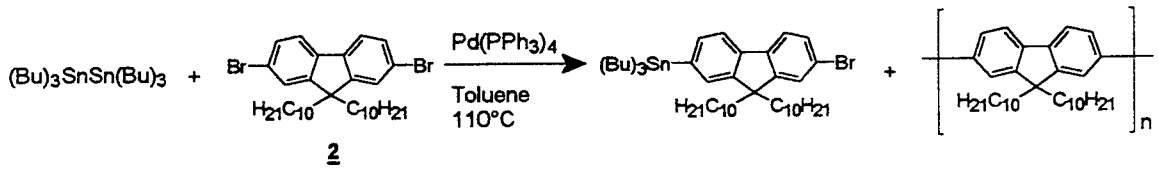


Figure 3 First attempt at coupling thiophene with fluorene.

While making the Grignard was not difficult, it was unreactive towards the fluorene. A similar experiment was planned by making a Grignard using fluorene 2 and reacting it with the bromothiophene. Despite several attempts at making this Grignard, results were frustrated, even after several well-known "tricks" were used [4]. This approach was abandoned in favor of adding

tributyltin to one side of the dibromofluorene **2** using the second coupling technique discussed in section 1.2 (refer to table 1). This approach worked too well in that once the product was formed, the head would go on to react with the tail of another fluorene to produce a polyfluorene (see figure 4).

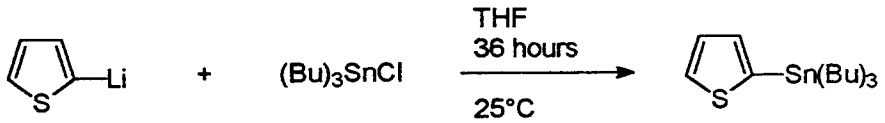


2

Figure 4 Attempt at adding tributyltin to fluorene.

2.2 THIOPHENE-TIN SYNTHESIS

A fourth approach was to react tributyltinchloride with 2-lithiothiophene using the third coupling technique (see figure 5). By stirring the two starting compounds in THF (dried over sodium and freshly distilled) for 36 hours at room temperature under nitrogen, the desired product was made with yields of 47%.



6

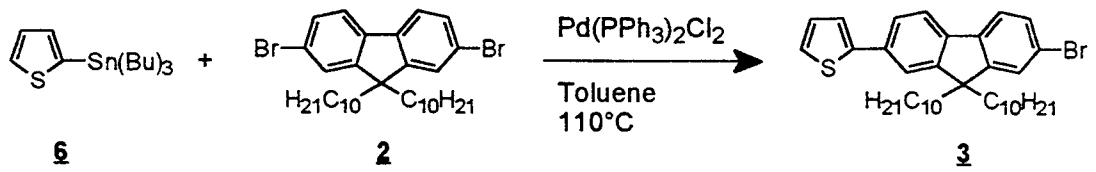
Figure 5 Adding tributyltin to thiophene.

In order to remove any unreacted tin compound, the THF was rotavaped off, and then hexanes added. An aqueous potassium fluoride solution was added to the flask and stirred vigorously for an hour. The layers were separated with the organic layer being dried with $MgSO_4$, filtered, and the product was then purified by column chromatography.

2.3 THIOPHENE-FLUORENE SYNTHESIS

The coupling of the thiophene-tin **6** and the dibromofluorene **2** was accomplished by refluxing the starting materials in toluene (dried, freshly distilled, and degassed) with 0.05 mole equivalents of $Pd(PPh_3)_2Cl_2$ for eight hours (see figure 6). In general, the reaction was complete when the solution turned black. The workup included vigorously stirring the solution with a saturated aqueous solution of potassium fluoride, filtering, separating the two layers, drying the

toluene over $MgSO_4$, and finally filtering. After removing the toluene, the product 3 was purified by column chromatography.

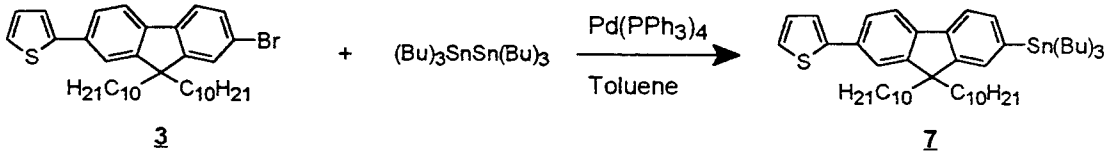


6 2 3

Figure 6 Coupling of thiophene and fluorene.

2.4 SYNTHESIS OF THIOPHENE-FLUORENE-TIN

Tin was added to compound 3 by using the second coupling technique (refer to table 1), as described in figure 7.



3 7

Figure 7 Synthesis of the thiophene-fluorene-tin.

By stirring compound 3 with hexabutylditin and 0.05 mole equivalents of $Pd(PPh_3)_4$, the desired product 7 was made. Purification was done by alumina column chromatography.

3.0 PYRIDINE / FLUORENE COUPLING

3.1 EARLY EXPERIMENTS

With the success of the thiophene-tin couplings, efforts were quickly focused on making the 4-(tributyltin)pyridine. Interestingly, many of the published procedures begin by reacting 4-bromopyridine with tributyltinsodium, even though the 4-bromopyridine is unstable and only commercially available as a hydrochloride salt [10,11]. One paper suggested stirring the pyridine salt in an aqueous NaOH solution and diethyl ether at 0°C. After separating the layers, the ether layer was to be stirred with $MgSO_4$ at 0°C for two hours before being filtered and used. Instead of using tributyltinsodium, it was decided that an attempt would be made at lithiating the neutralized 4-bromopyridine with n-BuLi and then reacting the resulting product with tributyltinchloride as described in the third coupling technique in section 1.2 (refer to table 1).

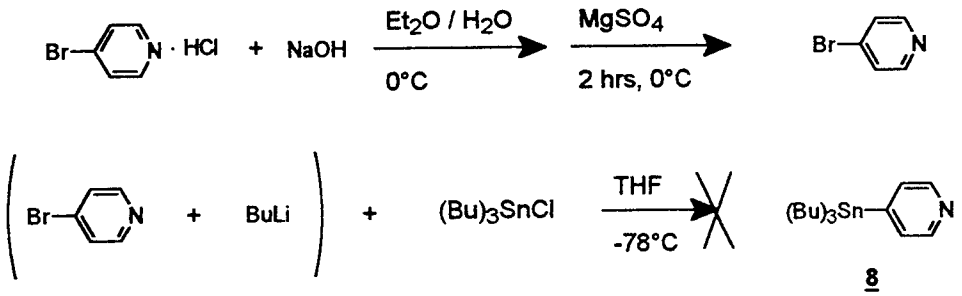


Figure 8 Attempt at adding tributyltin to 4-bromopyridine.

Unfortunately, enough water remained in the ether layer after neutralizing the pyridine salt that the butyllithium was quenched.

3.2 PYRIDINE-TIN SYNTHESIS

As a result, it was decided to neutralize and lithiate the pyridine salt with two equivalents of butyllithium as described in figure 9.

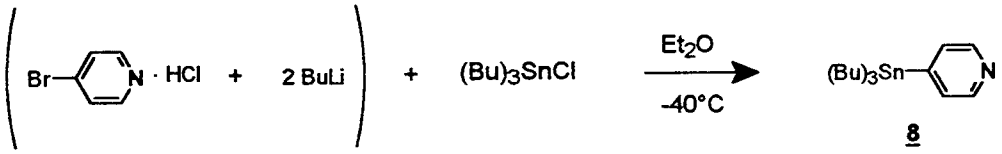
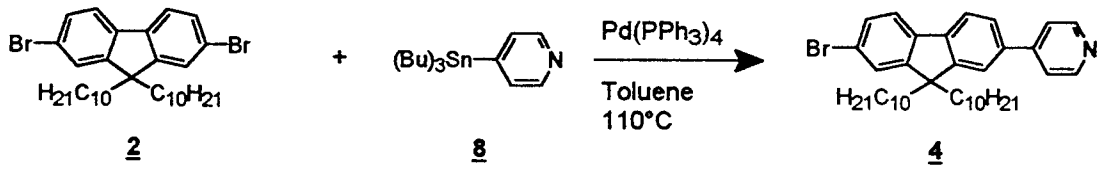


Figure 9 Adding tributyltin to pyridine.

A slurry of the pyridine salt was vigorously stirred in dry diethylether which was cooled to -40°C. Butyllithium was slowly dripped into the solution so as to not raise the internal temperature. The mixture was allowed to stir an additional twenty minutes before the tributyltinchloride was added. The product **8** was finally purified by alumina column chromatography with a yield of 82%.

3.3 PYRIDINE-FLUORENE SYNTHESIS

The 4-(tributyltin)pyridine **8** was coupled to fluorene **2** in the same way as the 2-(tributyltin)thiophene **6** was to fluorene **2** (refer to figure 5) except that $Pd(PPh_3)_4$ was used as the catalyst. The yield of 58% could probably have been improved had excess fluorene been used to discourage the formation of a dipyridyl-fluorene.

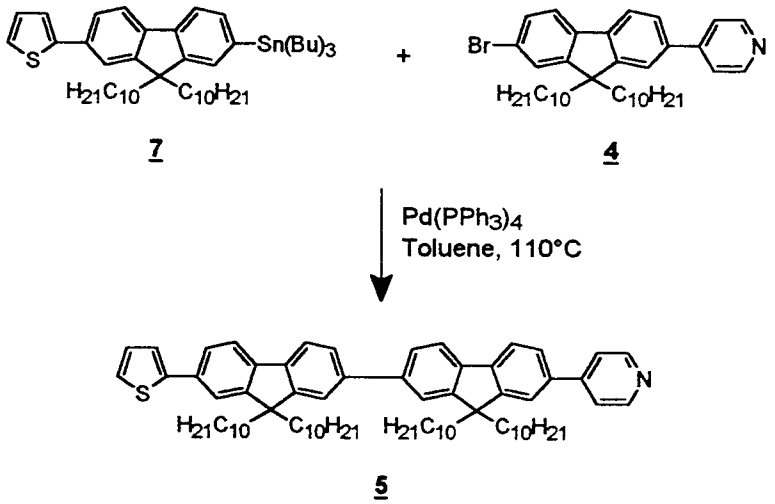


2 8 4

Figure 10 Coupling of pyridine and fluorene.

4.0 CHROMOPHORE SYNTHESIS

The final step in the chromophore 5 synthesis was to couple compounds 7 and 4 together by refluxing the two in toluene (freshly distilled and degassed) and 0.05 mole equivalents of $Pd(PPh_3)_4$ under N_2 for eight hours. Yield was 24% since the chromophore was purified twice by column chromatography before being characterized.



7 4

5

Figure 11 Final step in the chromophore synthesis.

5.0 CONCLUDING REMARKS

The synthetic methods needed to yield a novel second order NLO material 5 containing two fluorene groups has been successfully developed. Under a University of Cincinnati / WL/WPAFB educational partnership agreement, the synthesis and purification of this compound has been carried out. This alternate approach to making chromophore appears to afford better yields with about the same number of steps. In addition, the techniques developed here are general and can be applied whenever there is a need to add a thiophene or pyridine ring to a brominated site.

6.0 ACKNOWLEDGMENTS

It is a pleasure to thank the individuals who have made our research with WL/MLBP both a fruitful and pleasurable experience. In particular, thanks go to Mr. Bruce Reinhardt, Dr. Bob Evers, Dr. Ted Helminiak, Ms. Lisa Denny, Ms. Marilyn Unroe, Dr. Jay Bhatt, Ms. Ann G. Dillard, and Dr. Ram Kannan for all their help and kind hospitality.

BIBLIOGRAPHY

1. Prasad, P.N. and Williams, D.., *Introduction to Nonlinear Optical Effects in Molecules and Polymers*, Wiley-Interscience, New York, 1991.
2. Clarson, S.J. and Brott, L.L., *Synthesis of Novel Second and Third Order Nonlinear Optical Materials*, RDL Summer Faculty Research Program 94-0138, Final Report 1994.
3. Reinhardt, B.A. and Unroe, M.R., *Synthesis*, **1987**, 981.
4. Lai, Y., *Synthesis*, **1981**, 585.
5. Rieke, R.D., *Accounts of Chemical Research*, **1977**, 10, 301-306.
6. Rieke, R.D. and Bales, S.E., *Journal of the American Chemical Society*, **1974**, 96(6), 1775-1781.
7. Eaborn, C, Azarian, D., Dua, S. and Walton, D., *Journal of Organometallic Chemistry*, **1976**, 117, C55-C57.
8. Kosugi, M., Shimizu, K., Ohtani, A. and Migita, T., *Chemistry Letters*, **1981**, 829-830.
9. Bailey, T.R., *Tetrahedron Letters*, **1986**, 27(37), 4407-4410.
10. Yamamoto, Y. and Yanagi, A., *Chemical and Pharmaceutical Bulletin*, **1982**, 30(5), 1731-1737.
11. Yamamoto, Y. and Yanagi, A., *Heterocycles*, **1981**, 16(7), 1161-1164.

A GENETIC ALGORITHM SCHEDULER FOR THE SENSOR MANAGER

Milton L. Cone
Associate Professor
Department of Computer Science/Electrical Engineering

Embry-Riddle Aeronautical University
3200 Willow Creek Road
Prescott, AZ 86301-3720

Final Report for:
Summer Faculty Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Wright Laboratory

December 1995

1. INTRODUCTION

Combat aircraft carry many different sensors to perceive the outside world. Radars and missile warning receivers are two examples. Historically these sensors have worked independently of one another. When there weren't very many sensors, the crew members could operate them, deciding which sensors to employ, what mode to operate them in, when to use them, and what the data they collected meant. Over the years, sensors have proliferated. Each generation performs new functions with new modes and moves more rapidly between different functions and modes. Crew members can no longer be expected to manage all of the onboard sensors. Enter the need for a sensor manager.

The sensor manager's job is to "select the right sensor to perform the right service on the right object at the right time" [Musick & Malhotra, 1994]. It has five primary functions. These are:

- Generate potential options (taskings) for the sensors given sensor status and availability
- Prioritize the options given sensor performance and situation needs
- Develop a sensor schedule that best uses the sensor assets to accomplish the current mission goals
- Communicate desired actions to the individual sensor managers

The sensor manager works in a very dynamic, multiprocessor environment. This complicates the scheduling problem. There are several approaches to scheduling. [Stankovic, et. al., 1995] and [El-Rewini, etal., 1995] review several results. There are the classical scheduling theory results, such as earliest deadline first or rate monotonic scheduling, that apply to single processor systems. For multiprocessor systems, approaches to static scheduling are understood but difficult to implement. Work on dynamic scheduling in multiprocessor systems is just beginning. Many of these approaches have been applied to the job-shop scheduling problem (JSSP).

The JSSP has many things in common with the sensor scheduling problem.

- Each has multiple jobs that must be scheduled on a limited number of shared resources (machines or sensors). Generally these resources cannot do multiple jobs simultaneously.
- A job may consist of several tasks that must be completed in order. The jobs compete with each other for time on the resources. Different tasks take different amounts of time on the same resource.
- Often, the goal is to schedule all of the tasks in a minimum amount of time while maintaining any precedence relations among the tasks.

In the general JSSP there are *j* jobs each consisting of several tasks to be accomplished on *m* machines. The plan, which is known before scheduling begins, assigns each task to a machine with a known execution time and establishes the order in which a job moves from machine to machine. There are several benchmark JSSP problems on which researchers try their algorithms. Typical is the 6x6 benchmark problem, the statement of which is shown below, [Muth & Thompson, 1963]. In this benchmark, job 1 takes 1 unit of time on machine 3, then moves to machine 1 where it takes 3 units of time. A job must complete in the order shown, e.g. job 1 goes from machine 3 to 1 to 2 to 4 to 6 and finally to 5.

	(m,t)	(m,t)	(m,t)	(m,t)	(m,t)	(m,t)
Job 1	3,1	1,3	2,6	4,7	6,3	5,6
Job 2	2,8	3,5	5,10	6,10	1,10	4,4
Job 3	3,5	4,4	6,8	1,9	2,1	5,7
Job 4	2,5	1,5	3,5	4,3	5,8	6,9
Job 5	3,9	2,3	5,5	6,4	1,3	4,1
Job 6	2,3	4,3	6,9	1,10	5,4	3,1

Table 1. The 6x6 benchmark problem.

The minimum time to complete all tasks in the JSSP 6x6 problem is 55. [Fang, et al., 1993] showed how a genetic algorithm could achieve this result. The approach used here is a variation of their algorithm.

2. OVERVIEW

The purpose of this study is to develop metrics to quantify the performance of the genetic algorithm on the 6 x 6 JSSP. These metrics will be used to study the effect of changing the mutation and crossover rates and of starting the scheduling process with an initial seed schedule. The postulate is that there should be a combination of mutation and crossover rates that yields a faster convergence to the optimal solution or to an average value of makespan (makespan is the time when the last task completes if the first task starts at time 0) closer to the optimal value of 55 and that seeding should improve the results. The metrics will also be used to see if varying crossover and mutation rates while the algorithm executes helps speed convergence. Finally, code will be developed that handles scheduling both hard and soft deadline tasks in the genetic algorithm framework.

3. THE REPRESENTATION

A chromosome is a string containing j x m numbers ranging from 1 to j. For the 6 x 6 problem, a chromosome is made up of 36 numbers between 1 and 6. Each number appears 6 times, once for each machine. A typical chromosome might be

(4,5,2,6,4,6,6,6,3,4,2,3,4,1,2,5,3,2,5,3,3,6,6,1,1,4,1,2,5,5,4,1,2,3,5,1).

A schedule is built by taking the first gene, 4, and scheduling the machine assigned in the task sequence for job 4. In this case, it is machine 2. Next, the first task for job 5 is scheduled, then job 2 etc. When job 4 appears for the second time, the machine assigned to the second task for job 4, which is machine 1, is scheduled. Crossover occurs between adjacent chromosomes. Any number of genes may be crossed. The crossover operator used here is similar to the PMX operator in [Goldberg & Lingle, 1985] as modified by [Cone, 1995]. The basic idea of the PMX crossover operator is to swap a sequence of genes between two chromosomes then map the two new chromosomes into valid schedules. For example, if the second and third genes (5 and 2) were exchanged with the next chromosome in the typical chromosome above and the two new genes were 1 and 3, then 1 and 3 would take the place of 5 and 2 and 5 would replace the first occurrence of 1 in the string and 2 would replace the first occurrence of 3. The new typical chromosome would be

(4,1,3,6,4,6,6,6,3,4,2,2,4,5,2,5,3,2,5,3,3,6,6,1,1,4,1,2,5,5,4,1,2,3,5,1).

The mutation operator simply exchanges randomly picked genes within any one chromosome. The strength of this approach is that crossover and mutation always lead to valid schedules.

The approach to schedule building, although similar to [Fang, et al., 1993], seems simpler since the schedule builder here does not have to maintain a circular list of uncompleted jobs. The schedule builder steps through the list of genes that makes up a chromosome, taking the next available task for the job encountered in the chromosome. For example, if the first five genes in a chromosome are (4,1,3,6,4...) then the first task for jobs 4, 1, 3, and 6 would be scheduled then the second task for job 4 etc. When the scheduler places a task on the schedule, it first finds the time when the previous task in that job completes as the current task can not start before then. Next it starts from that time and works forward on the schedule for the machine for which this task is assigned, looking for an open slot of time long enough to fit the task into. If there are no slots big enough, the task is scheduled after the last task on this machine finishes.

4. THE GENETIC ALGORITHM

A genetic algorithm scheduler starts by generating an initial population of chromosomes on which the genetic operators of crossover and mutation can work. Each chromosome is evaluated for its fitness. The chromosomes with better evaluations are more likely to survive into the next generation although the process is stochastic so that a weak chromosome's survival probability is still finite. This study started with an initial population of 50 schedules. This population was either randomly generated or seeded with one schedule with a makespan of 58 combined with 49 randomly chosen schedules. The purpose of seeding a known schedule is to see if a good schedule speeds the convergence to the optimal solution. The genetic algorithm stops after a set number of

schedules is generated. After some experimenting, it was found that under most conditions 6000 schedules are enough to allow the genetic algorithm to find a schedule with a makespan of 55. If the optimal makespan was not found in 6000 trials extending the number of trials did not significantly improve the chances of finding an optimal solution.

The genetic algorithm used for this study is called GENESIS. It was developed by [Grefenstette, et al., 1991] and is one of the most popular of the genetic algorithm implementations.

5. THE METRICS

Four measures were used to judge the performance of the genetic algorithm scheduler. The first is the trial in which the genetic algorithm first found the optimal solution of 55 for the schedule. Twenty-five experiments were run, each starting with a different random number. The average trial was then calculated from the twenty-five experiments. The smaller this number, the faster, on average, the algorithm finds a good solution. While most experiments found the best solution not all did. This leads to the second measure.

The second measure is the average of the best schedules for each of the 25 experiments. If every schedule found the optimal result then the average schedule evaluation is 55. If the optimal solution was not found in 6000 trials then the trial was set at 6000 and the average trial calculated. Since that process skews the averages, the average trial in conjunction with the average schedule evaluation is a better judge of whether any improvement has occurred in the algorithm. The closer the average schedule evaluation is to 55, the more meaningful the comparison between average trial numbers.

The third metric is the average evaluation of the 50 schedules determined in the final generation. Each generation consists of 50 schedules. In the 6000 trials there are 120 generations. The third metric is the average evaluation of those final 50 schedules averaged over the 25 experiments that comprise a study. When the best evaluation found for all 25 experiments is 55, then the generation average for the generation in which the last 55 was found is also given. The third metric shows whether there is a correlation between how fast the optimal solution was found and the average performance of a generation. The presumption is that the lower the generatrion average, the sooner the optimal solution should be found (a lower first metric).

The last metric is the bias. Bias is a measure of the convergence of the schedules in a generation to one schedule. The higher the bias the more all schedules look the same. If the bias is one then all schedules in a generation are identical. Bias might be used to adjust the crossover and mutation rates while the genetic search is ongoing. Early in the search the bias should be high preventing the algorithm from converging on one suboptimal solution. Later, the rates can be adjusted to focus the search in one area. Bias

is calculated by finding the percentage of 1's or 0's in each bit position for a chromosome then averaging over a chromosome. In each bit position the percentage used to compute the average is the higher of the 1's and 0's percentage. Thus the bias can not fall below 0.500. The number reported in the tables is the average bias for the last generation, averaged over the 25 experiments. As in the third metric, when the second metric is 55 then the bias is also reported for the generation in which the second metric reached 55.

6. RESULTS

Parametric Study of Crossover and Mutation Rates

A series of studies were conducted to examine the effects of different crossover and mutation rates on the four metrics. A study consists of 25 identical experiments except each is started with a different initial random number. The results are recorded in Table 1.

The top number in each cell is the average trial number in which the optimal solution for the 6 x 6 JSSP was found. If the optimal value of 55 was not found in 6000 trials then the trial number was taken as 6000 and the average calculated. This tends to skew the average trial to a lower value than expected since 6000 is used instead of the larger correct value.

The second number in each cell is the average of the best evaluations found over the 25 experiments. If each experiment found the optimal result then the average is 55. The closer the average is to 55, the more useful the average trial number is for comparisons with other cells. For example, the cell with 55.04 found the optimal 55 in all but one of the 25 experiments. In that one experiment the best result was 56. Therefore the average trial number has some significance when compared to those studies with average schedule evaluations of 55. On the other hand, the average of the best evaluations in cell (0,1) is 56.8, a comparatively high number. The average trial number, 4934, is also very high, even though most of the experiments found a near optimal solution.

The third number is the population average at the end of the experiment for the last 50 schedules. In cell (0,3) the population average is 67.46. This number is computed by averaging the evaluations in the last generation for an experiment, then averaging those numbers over the 25 experiments that make up a study. The number, 67.45, in parenthesis is the same number computed at the generation where the last optimal result, 55, was found.

The last number is the bias. A high bias as in cell (0,0) at 0.955 means that most of the schedules had the same genes in the same order. A low bias such as 0.667 in cell (0,5) means that there was still a lot of variety in the 50 schedules that make up a generation.

The smallest bias is 0.500 and the largest bias is 1.0. A bias of 1.0 means that all of the schedules are the same.

mutation crossover	0.0	0.001	0.01	0.1	0.9	1.0
0.0	5762 59.00 59.02 0.955	4934 56.80 57.58 0.965	2274 55.28 59.60 0.818	1223 55.00 67.46 (67.45) 0.686 (0.682)	925 55.00 69.24 (68.48) 0.668 (0.666)	1666 55.00 68.55 (68.87) 0.667 (0.666)
0.2	3374 55.72 56.60 0.959	2587 55.56 58.96 0.895	2111 55.04 63.79 0.767	1053 55.00 67.55 (67.94) 0.681 (0.679)	2031 55.04 68.98 0.668	1113 55.00 68.63 (68.90) 0.668 (0.667)
0.4	2553 55.24 62.29 0.850	1961 55.16 63.11 0.822	1470 55.00 65.45 (65.44) 0.755 (0.748)	892 55.00 67.91 (67.98) 0.679 (0.679)	1595 55.00 68.69 (68.87) 0.668 (0.666)	1692 55.04 68.72 0.667
0.6	2152 55.04 63.10 0.831	1315 55.00 63.76 (64.38) 0.816 (0.798)	1550 55.00 65.98 (66.15) 0.744 (0.745)	1292 55.00 68.52 (68.41) 0.680 (0.679)	1346 55.00 69.08 (69.09) 0.667 (0.666)	2133 55.04 69.04 0.668
0.8	1262 55.08 63.30 0.835	1278 55.00 63.94 (65.18) 0.817 (0.784)	1158 55.00 66.45 (66.40) 0.748 (0.740)	1439 55.00 68.56 (68.14) 0.678 (0.678)	1364 55.00 69.34 (69.34) 0.666 (0.666)	1699 55.00 69.20 (69.18) 0.667 (0.667)
1.0	1368 55.12 63.41 0.833	1001 55.00 64.60 (64.53) 0.811 (0.805)	1215 55.00 66.51 (66.54) 0.744 (0.745)	1715 55.00 68.50 (68.57) 0.676 (0.676)	1573 55.04 69.05 0.666	1645 55.08 69.27 0.667

Table 1. Average trial number, average schedule evaluation for best schedule found, average population and bias. Parenthesis indicates values when last 55 found. Twenty-five experiments each study. JSSP 6 x 6. Trial set to 6000 if optimal result of 55 not found.

The next table, Table 2, is similar to Table 1, except that the initial population includes a schedule with an evaluation of 58. The purpose of this investigation is see if a seed schedule will improve the performance of the genetic algorithm by finding the optimal schedule sooner (metric one) or will lower the average of the best schedules found (metric two).

mutation crossover	0.0	0.001	0.01	0.1	0.9	1.0
0.0	6000 57.96 57.99 0.963	4778 56.44 57.02 0.972	1997 55.00 59.24 (59.16) 0.813 (0.834)	612 55.00 67.61 (67.66) 0.686 (0.687)	965 55.00 68.78 (68.65) 0.667 (0.667)	1185 55.00 68.99 (68.75)0.66 8 (0.666)
0.2	4107 55.64 56.58 0.959	3145 55.56 58.17 0.926	2112 55.16 64.19 0.760	1021 55.00 67.80 (68.09) 0.683 (0.681)	2005 55.08 68.71 0.666	1655 55.12 68.81 0.667
0.4	2179 55.16 61.23 0.864	2687 55.40 62.66 0.830	1815 55.08 65.49 0.746	1018 55.00 68.04 (68.11) 0.680 (0.680)	1224 55.00 69.34 (69.34) 0.667 (0.667)	1779 55.00 69.12 (69.10) 0.667 (0.667)
0.6	1985 55.04 62.84 0.841	1527 55.00 63.60 (63.81) 0.820 (0.813)	1670 55.00 66.06 (66.21) 0.746 (0.746)	919 55.00 68.19 (68.24) 0.679 (0.679)	1083 55.00 69.29 (68.88) 0.667 (0.667)	1418 55.00 69.10 (69.22) 0.668 (0.666)
0.8	858 55.00 62.98 (66.00) 0.835 (0.768)	1423 55.00 64.61 (64.61) 0.812 (0.812)	831 55.00 66.14 (66.70) 0.744 (0.742)	1917 55.04 68.57 0.678	1992 55.08 69.22 0.666	1456 55.00 68.88 (69.52) 0.666 (0.668)
1.0	1866 55.12 63.54 0.831	944 55.00 63.64 (65.53) 0.819 (0.775)	942 55.00 66.52 (66.96) 0.748 (0.735)	1045 55.00 69.20 (68.67) 0.678 (0.678)	1419 55.00 68.78 (69.25) 0.666 (0.666)	1794 55.00 69.59 (69.59) 0.667 (0.667)

Table 2. Average trial number, average schedule evaluation for best schedule found, average population, and bias. Twenty-five experiments each study. JSSP 6 x 6. Trial set to 6000 if optimal result of 55 not found. Initial seed schedule evaluation 58.

The last table, Table 3, contains the results for the same conditions as Table 1, except that a different random seed starts the process. The results from Table 1 and Table 3 should be approximately the same if the random process has stabilized.

mutation crossover	0.0	0.001	0.01	0.1	0.9	1.0
0.0	5761 58.60 58.3 0.969	4533 56.60 57.32 0.968	1949 55.08 59.06 0.820	593 55.00 67.27 (67.37) 0.685 (0.683)	1701 55.00 68.80 (68.86) 0.668 (0.667)	1226 55.04 68.79 0.667
0.2	3996 56.20 57.74 0.950	3155 55.52 58.41 0.927	1842 55.08 63.74 0.763	816 55.00 67.88 (67.66) 0.682 (0.682)	848 55.00 69.21 (68.82) 0.667 (0.668)	1455 55.00 68.84 (68.83) 0.667 (0.666)
0.4	2531 55.36 61.23 0.864	1735 55.08 63.03 0.820	1365 55.00 65.34 (65.40) 0.748 (0.748)	1106 55.00 67.82 (68.21) 0.677 (0.679)	1740 55.08 69.06 0.666	1053 55.00 69.05 (69.15) 0.666 (0.666)
0.6	1378 55.04 62.94 0.842	1737 55.04 63.66 0.815	1472 55.00 65.78 (66.29) 0.745 (0.743)	866 55.00 68.02 (68.59) 0.679 (0.678)	1896 55.08 68.93 0.668	1806 55.04 68.96 0.667
0.8	940 55.00 63.46 (65.35) 0.831 (0.779)	1132 55.00 63.88 (64.81) 0.820 (0.796)	1075 55.00 66.54 (66.61) 0.742 (0.743)	1161 55.00 68.22 (68.56) 0.680 (0.677)	1344 55.00 69.31 (68.81) 0.667 (0.667)	1842 55.04 69.22 0.667
1.0	1260 55.04 64.62 0.807	1528 55.08 64.46 0.814	1036 55.00 66.42 (67.13) 0.747 (0.740)	1302 55.00 68.55 (68.99) 0.678 (0.677)	1556 55.04 69.16 0.668	1382 55.00 69.10 (69.44) 0.667 (0.667)

Table 3. Average trial number, average schedule evaluation for best schedule found, average population, and bias. Same as Table 1 except different initial seed.

Hard Deadlines

Next, code was developed to examine how the genetic algorithm handles soft and hard deadlines. A hard deadline is taken to mean one that must start at a fixed time. For example, the third task of Job 2 which takes 10 units of time on machine 5 is a hard deadline if it had to begin at time 20. A soft deadline can start at any time but still must follow the task precedence assigned to it. In this example the first two tasks of Job 2 would precede the hard deadline.

The scheduling algorithm consists of moving the hard deadline task and all tasks that must precede it to the front of the schedule where they are assigned first. The original schedule that exists in the population is not changed. A procedure senses when the hard deadline task needs to be assigned The start time for that task is set to the right value. In case the hard deadline cannot be met, a flag is set that penalizes the makespan evaluation by an arbitrary factor. The factor 1.1 is used here. If a schedule evaluates to 55 but cannot meet its hard deadline then the best evaluation it can have is 60.5. This penalty is large enough to keep the schedule from being a major factor in subsequent generations but still keeps a finite possibility that it can survive. The advantage of this approach is that schedules are continuously evaluated so that if none of the schedules meets the hard deadline, the schedule that comes closest to meeting it can be used.

Another approach to scheduling hard deadlines is to leave the hard deadline task in its current position. When it comes time to assign the hard deadline, the task is put into the correct time slot if it exists. The problem with this approach is that it admits many more illegal schedules. The more tasks assigned before the hard deadline is handled, the more likely that the time slot will be unavailable.

Still another approach is to only move the hard deadline task to the front and assign it first. The problem with this approach is that there is a good chance the preceding tasks will not be completed by the time the hard deadline must start. Moving all of the tasks in a job preceding the hard deadline to the front maximizes the chance of ending up with a legal schedule.

Variable Crossover and Mutation Rates

The final results come from varying the crossover and mutation rates during the execution of the genetic algorithm. High crossover and mutation rates mean that there is a considerable disruption in the schedules. This is good early in the sampling of the schedule space when schedules from diverse areas prevent premature convergence to a non-optimal solution. With succeeding generations, this turbulence needs to be reduced so that good solutions are refined.

The first attempt at changing crossover and mutation rates during execution of the genetic search consisted of deterministically changing both. A high crossover rate at the

beginning and a high mutation rate at the end should speed the convergence. The mutation rate started at 0.001 and grew towards 1.0 and the crossover started at 1.0 and decayed towards 0. A decay or growth factor of 0.95 was used. This causes the crossover rate to drop to 95% of its previous value with each generation. The result is:

mean trial number = 956
mean make-span = 55.04
bias = 0.667
final generation average = 69.09

The bias stayed at about 0.669 and finished at 0.667 and the current generation average stayed about 69 finishing at 69.09.

Next the mutation and crossover rates were reversed with every other condition unchanged. The mutation rate started at 1.0 and decayed towards 0.0 and the crossover rate started at 0.001 and grew towards 1.0. In this case:

mean trial number = 1252
mean make-span = 55.08
bias = 0.876
final generation average = 56.66

The next variation consisted of repeating the previous run where the mutation rate decays and the crossover rate grows except that the growth factor was set to 0.8577 and the mutation and crossover rates were reset every 30 generations. The 0.8577 growth factor allowed the mutation and crossover rates to cover the same range in 30 generations as the previous values did in the full run of 120 generations. Resetting the values every 30 generations introduces diversity into the population to increase the chances of finding the optimal schedule. The result is:

mean trial number = 1070
mean make-span = 55.00
bias = 0.719 (0.668)
final generation average = 64.14 (68.84)

The values is parenthesis are the values when all experiments reached an optimal schedule.

Generally, better results happen when the bias is low, somewhere around 0.667. The next scheme consists of toggling the mutation rate between two values depending on the value of the bias. If the bias gets to 0.7 a higher mutation rate is used. This introduces more turbulence into the population driving the bias down. With the crossover rate constant at

0.6, the mutation rate was toggled between 0.01 and 0.1. The result is:

mean trial number = 1318
mean make-span = 55.00
bias = 0.695 (0.697)
final generation average = 67.68 (67.21)

The last strategy consisted of introducing a moving window for crossover. [Fang, et al., 1993] observed that at the start, most of the improvement in the evaluation happens when genes are swapped at the beginning of the chromosome. This helps the chromosome find good places to search for the optimal solution. Later on when the beginning genes of the chromosome are in place, it is more important to concentrate the crossover at the end of the chromosome so that good schedules can be fine-tuned. A window was set up that covered 18 genes. Crossover was restricted to be within those 18 genes. The window started at the beginning of a chromosome and systematically moved across the chromosome until, at the last generation, crossover can only occur within the last 18 genes. Thus for generation 60, half way through, the 18 middle genes could be used for crossover. Not all 18 genes have to be used. A random generator establishes how may genes are actually used and where in the window crossover occurs.

Table 4. shows the results of the moving window strategy. Runs for crossover rate of 0.0 are not shown since they are the same as Table 1.

mutation crossover	0.0	0.001	0.01	0.1	0.9	1.0
0.0						
0.2	4386 56.20	3142 55.60	2026 55.12	904 55.00	1309 55.00	1101 55.00
0.4	3501 55.84	1725 55.16	1539 55.00	896 55.00	1345 55.00	1248 55.00
0.6	2385 55.44	2801 55.16	1575 55.00	1175 55.00	1462 55.08	1655 55.04
0.8	2621 55.48	2166 55.28	1076 55.00	1024 55.00	2060 55.04	1537 55.04
1.0	2385 55.28	1940 55.24	1162 55.00	1630 55.12	1551 55.04	1381 55.00

Table 4. Mean trial number and mean make-span for moving window.

7. Discussion

Even when the mutation and crossover rates are poorly suited for this problem, genetic

search does a good job at finding a solution. There are

$$\frac{36!}{(6!)^6} = 2.67x10^{24}$$

possible, valid schedules for this problem. This number is found by calculating the

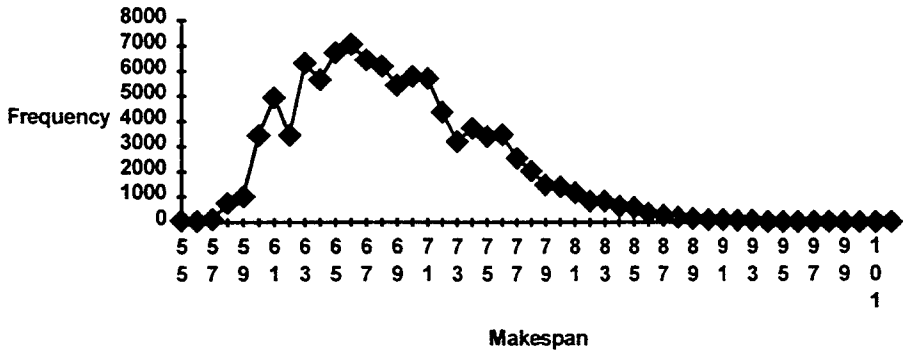


Figure 1. Frequency distribution for 100,000 randomly chosen schedules.

number of possible combinations as if each schedule is unique and then dividing by the number of schedules that are duplicates.

The number of solutions that produce the optimal schedule is unknown but can be estimated by randomly sampling the solution space. Figure 1 shows the distribution of solutions for 100,000 randomly chosen schedules with replacement. In the 100,000 schedules, there were 71 schedules that evaluated to 55. This means that on average a random draw will find the optimal result once in every 1408 tries. The mode of the distribution is 66, the mean is 69.1 and the median is 68.

Assuming that 25 experiments are enough to characterize the performance of the genetic algorithm, then Table 5 compares Tables 1 and 3 showing which combinations of crossover and mutation found the optimal solution. The 1 indicates that Table 1 converged to 55, the 3 means Table 3 and the B means both. It is clear that mutation is critical to finding the optimal solution. Without it, none of the experiments were completely successful at finding the optimal solution. Mutation rates of 0.01 and 0.1 were the best at finding the optimal solution, seemingly independent of the crossover rate. A crossover rate of 0.8 showed the most consistent performance. Four of the six studies were always successful at getting to the optimal solution.

There is a definite amount of mixing that must go on for the best performance of the genetic algorithm. As the rate of mixing is increased by going to higher crossover rates,

mutation crossover	0.0	0.001	0.01	0.1	0.9	1.0
0.0				B	B	1
0.2				B	3	B
0.4			B	B	1	3
0.6		1	B	B	1	
0.8	3	B	B	B	B	1
1.0		1	B	B		3

Table 5. Studies from Tables 1 and 3 that found the optimal solution in every experiment.

the mutation rate for the most robust performance decreases. It also appears that too much mixing can have a detrimental effect on the ability of the genetic algorithm to find the optimal solution as only one study in the 1.0 column and two studies in the 0.9 column consistently reached the optimal result. The detrimental effect is more pronounced for the lower mixing rates since most of the studies in the two right hand columns only missed getting to 55 in one or two experiments and those experiments found a 56 result.

No B's in the first column of Table 5 indicate that crossover by itself is not enough to generate the best performance. On the other hand, X's in the first row show mutation alone can produce the optimal result. Until recently crossover was considered the more important operator. This result supports the contention that the mutation operator is critical to the proper functioning of the genetic algorithm.

How does the genetic algorithm compare to randomly picking schedules? Generally where the genetic algorithm found the optimal solution it outperformed random selection. In the 14 cases in Table 5 where the genetic algorithm reached the optimal solution in every experiment, four from Table 1 and three from Table 3 exceeded the random draw. Of those seven, four performed significantly worse than a random draw. Picking a mutation rate of 0.1 and a crossover rate of 0.4 as a combination of values that gives good performance under a variety of conditions and sets in the middle of combinations that also perform well, genetic search found the optimal solution in 892 and 1106 trials compared to the random 1408 trials. This is a 21% to 37% improvement.

Genetic algorithms are notorious for failing to find the final steps to the optimal solution. In fact hill climbing may be combined with genetic search, where hill climbing is used to traverse those final steps. If the requirement is relaxed to finding a schedule that is either 55 or 56, then the genetic algorithm performance might be expected to improve significantly over a random draw. A random draw of 100,000 schedules found either 55 or 56, 103 times. This is on average once every 971 schedules. The mutation/crossover operator combination of 0.4/0.1 found schedules of 55 or 56 on average of every 588 trials for the conditions of Table 1. This is a 39.1 % improvement. While better than the

improvement for the optimal case, it is still not the improvement one might hope for considering the overhead cost associated with the genetic algorithm calculation. The reason for the lack of significant improvement is attributed to the huge number of different solutions that produce the optimal result. There are approximately $2x10^{21}$ different, optimal solutions to this problem. The genetic algorithm needs a harder problem in order show more improvement. The purpose of the simpler problem is to allow insight into the workings of genetic search. This should help in analysing the more difficult problems.

Next consider Tables 1 and 2. Does the seed of a schedule that evaluates to 58 help the genetic algorithm find the optimal solution and find it faster? Of the 36 combinations of crossover and mutation rates studied 17 showed improvement, 16 had performance decreases, and 3 stayed about the same. Six combinations were helped to find the optimal solution for all 25 experiments that did not find the optimal solution in Table 1. Four combinations that evaluated to 55 in Table 1 actually performed worse with the seed. Although, generally, only one, two or three experiments found 56 as the best solution instead of 55 with the rest of the 25 experiments converging on 55. With the mixed performance, it is hard to conclude that an initial seed schedule improves the genetic algorithm performance. On average, though, a seed schedule does not seem to hurt performance either. This phenomena has been observed by other researchers experimenting with genetic algorithms. More information does not always help. The genetic algorithm seems to perform better than other algorithms when there is less knowledge about the environment.

Variable Crossover and Mutation Rates

The approach to speed convergence rates by increasing the mutation rate while decreasing the crossover rate during the execution of the genetic algorithm worked better than decreasing the mutation rate while increasing the crossover rate. The mean trial number was 956 versus 1252. The similar value from Table 1. for mutation rate of 0.001 and crossover rate of 1.0 is 1001. If the one value of 56 is deleted from the experiment then the mean trial value is 745 and if the final value of 56 is accepted as a 55 then mean trial value drops to 735. One of the reasons that the first approach worked better than the second is that the selection process tends to eliminate schedules that are different. This effectively reduces the gene pool on which the crossover operator can work. Crossover is essentially happening on the same schedules. A high mutation rate at the end helps introduce new chromosomes into the population assuring that some diversity survives. Crossover is most effective in the beginning of the search when there is a large diversity of schedules for cross breeding. The bias value reflects the diversity. In the first approach the bias stayed at about its initial value of 0.667. In the second approach the bias increased to 0.876. This high value means that most of the schedules look very much alike.

The next strategy to improve the convergence rate also tries to introduce diversity into the population of schedules. A normal run consists of 120 generations of 50 trials each for a total of 6000 trials. If the optimal value has not been found in 30 generations, the mutation and crossover values are reset to their initial values. A high initial value for mutation rate is used so that a maximum number of new schedules are introduced into the population. The 30 generation value is arbitrary. The mean trial number fell for this approach from the 1252 above to 1070. This approach also had the advantage that all of the experiments found the optimal solution. Notice that the bias was intermediate at 0.719 but was low when the last optimal value was found. Most of the time a low bias around 0.667, gives better results. A larger diversity in the population prevents premature convergence of the population to a local, but non-optimal peak.

The next approach tries to hold the bias down by switching to a higher mutation rate when the bias starts to build up. This is the first approach here to incorporate feedback into a strategy. The switching kept the bias down. It never got above 0.7 indicating substantial diversity in the population. But the technique failed to significantly improve on the performance of a random draw of schedules, 1318 versus 1408.

The last approach uses a moving window to localize the effect of crossover. Of the 30 combinations run, a comparison to the values in Table 1 show that the mean trial number is greater in Table 4 for 14 entries and less for 16. Generally the moving window did better for the higher mutation rates. To find the optimal result the genetic operator has to perform well in the end game. This is especially true of the moving window which concentrates on fine tuning the schedules at the end. When the mutation rates are small. selection tends to eliminate all but a few of the best schedules. This makes the crossover operator ineffective since schedules are being crossed that are about the same. When the mutation rates are high, different schedules are continually being introduced into the population. This helps the genetic algorithm find the optimal solution. The best performing combination of mutation and crossover rates in Table 1, cell (2,3), was also the best in Table 4. In fact the mean trial number was almost identical, 892 compared to 896. It is also important to note that the genetic algorithm uses the same sequence of random numbers in Table 4 as Table 1. The changes made to the genetic algorithm to introduce the moving window did not use the random number generator. Therefore any changes in performance are as a result of the moving window.

8. Conclusions

The distribution of schedules for the 6x6 JSSP has been given. The optimal solution, 55, comprises approximately 0.071% of the 2.67×10^{24} possible schedules. That means there are 1.90×10^{21} different correct solutions. With that many correct solutions, any technique, even random search, works well.

The lessons for the sensor manager are that the performance of genetic search can be tuned by changing the crossover and mutation rates. Of the two, the mutation rate is the

more important. Tuning depends on the details of the problem. If the solution space is heavily populated with optimal solutions, as the 6x6 JSSP is, then any technique works well. The simpler the technique, the better, since it generally will be more robust and execute faster. A random search is simple and executes fast. If the solution space is not heavily populated with optimal solutions, then a more powerful technique, like genetic search, needs to be used and the overhead of the genetic algorithm calculation can be justified. The less knowledge there is about the problem the more likely a technique like genetic algorithms will be more effective. When there is detailed knowledge about the problem then a more powerful technique that takes advantage of this knowledge will outperform techniques like random and genetic search. The best choice of approaches depends on the details of the problem

The strategies to improve the performance of genetic search all showed some improvement in the mean trial number. None seemed to be the ideal solution that makes a major improvement in the genetic algorithm performance. All seemed to help. When the genetic algorithm scheduler is applied to the sensor manager then these strategies can be tested to find out which works best.

9. Future Work

There are several problems that have yet to be addressed. This section will present those problems and suggest ways that they may be resolved.

The first is timing. The genetic algorithm will have to operate in a real-time environment. Thus the algorithm needs to run as fast as possible. One of the advantages of the random search is that it is very fast. Several schedules can be tried while the genetic algorithm is handling its overhead. As shown in this paper the genetic algorithm finds the optimal solution in fewer trials than the random search, but may take longer to find that solution due to overhead. One way to speed up the genetic algorithm is to shorten the evaluation process. In the example presented here, the genetic algorithm had to build a schedule for each of the trials. If there is a way to rate the schedules short of a complete evaluation then the process can be accelerated. One possible way is to observe that a dot product of two schedules in some way represents how close one schedule is to another. For example, consider a schedule made up of six numbers (1, 2, 3, 4, 5, 6). Another schedule made up of the six numbers in the same order, (1, 2, 3, 4, 5, 6) when multiplied together like a dot product with the first element in each schedule multiplied together and added to the second elements multiplied together, etc., yields the number 89. This is the maximum of any permutation of the order of these six numbers. If the sequence (1, 2, 3, 4, 5, 6) is the optimal sequence then any other permutation yields a smaller number. Thus the dot product is a way to judge how close two schedules are to each other.

An evaluation function could be constructed that takes the population for a generation and randomly chooses four schedules for evaluation. The remaining schedules would

receive the same evaluation as the one of the four schedules to which they come closest. At the end the final population would be evaluated for the best solution. A solution of this type, if it works, should accelerate the genetic search execution speed significantly. The principle at work here is that perfect knowledge is not necessary in a random process to get good results.

The second problem is uncertain execution times. A schedule is built assuming perfect knowledge of the execution times of a task. At some level of abstraction in the scheduling process these times cease to be accurately known. A schedule based on best knowledge may also work with the changed execution times. There are generally some gaps in every schedule that can cushion the increased execution times. If the execution times decrease then the current schedule works fine although it may no longer be the optimal schedule. A possible solution to this problem is to use multiple models. Each model represents a different set of execution times. The schedule to be picked is the one from the models that is closest to the best solution of all the individual problems. A measure like the dot product might be used to judge which solutions are closest.

The third problem is to integrate the pilot into the operation of the sensor manager. Some recent work [Tarn, et. al. - 1995] have shown that event scheduling makes it easier to integrate human operators into the operation of a remote robotics system. In this case the scheduler is scheduling events instead of working a timeline. In fact that is how the genetic algorithm scheduler works for the 6x6 JSSP problem. The genetic algorithm is manipulating events, i.e. the order in which the tasks are going to be accomplished, only at the evaluation does the timeline get developed. If this event driven approach can be developed in terms of the sensor manager, then it may be possible to seamlessly integrate the pilot into the operation. For example, in the JSSP problem suppose sometime during the execution of the schedule the operator decides that he wants the fifth task of job 6 which currently takes 4 time units on machine 5 to run on machine 3 where it takes 2 units of time. All he has to do is change the event not the schedule. The interaction is with events not with the timeline.

References

M. L. Cone (1995). Genetic algorithms and the sensor manager scheduler. *RDL Final Report, 1995*, To Be Published.

H. El-Rewini, H. H. Ali, & T. Lewis (1995). Task scheduling in multiprocessor systems. *Computer*, Dec.

H-L Fang, P. Ross, & D. Corne (1993). A promising genetic algorithm approach to job-shop scheduling, rescheduling, and open-shop scheduling problems. In S. Forrest (ed) *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 375-382. San Mateo, CA: Morgan Kaufman.

D. E. Goldberg & R. Lingle Jr. (1985). Alleles, loci, and the traveling salesman problem. In J. Grefenstette (ed) *Proceedings of the First International Conference on Genetic Algorithms and their Applications*, pages 154-159. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

J. J. Grefenstette, L. Davis, & D. Cerys (1991). *GENESIS and OOGA: Two Genetic Algorithm Systems.* TSP, Melrose MA.

S. Musick & R. Malhotra (1994). Chasing the elusive sensor manager. *Proceedings of the NAECON, May 1994.* Dayton, Ohio.

J. F. Muth & G. L. Thompson (1963). *Industrial Scheduling.* Prentice Hall, Englewood Cliffs, NJ.

J. A. Stankovic, M. Spuri, M. Di Natale, & G. C. Buttazzo (1995). Implications of classical scheduling results for real-time systems. *Computer*, Jun.

T. J. Tarn, B. K. Ghosh, & N. Xi (1995). Sensor referenced control and planning: theory and applications. Short course at 34th IEEE Confenernce on Decision and Control, New Orleans, LA.

INTERIOR BALLISTICS OF THE WAVE GUN

Robert W. Courter
Associate Professor
Department of Mechanical Engineering

and

Jason Hugenroth
Graduate Student
Department of Mechanical Engineering

University of Cincinnati
Cincinnati, OH 45221

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

University of Cincinnati

December 1995

INTERIOR BALLISTICS OF THE WAVE GUN

Robert W. Courter
Associate Professor
Department of Mechanical Engineering
Louisiana State University

and

Jason J. Hugenroth
Graduate Student
Department of Mechanical Engineering

Abstract

A specialized light gas gun firing cycle called Wave Gun, developed by Thomas Dahm of Astron Research and Engineering, is investigated as an improved model launcher for free flight aeroballistic ranges. Specifically, it is desired to decrease model loading while increasing or maintaining muzzle velocity. In partial fulfillment of this goal a limited experimental program has been performed for the purpose of validating the interior ballistics code to be used in the study. Data to be used for code validation have been obtained from four shots of the Astron Wave Gun. The measured parameters include pressure histories in the propellant chamber, nozzle entrance, nozzle exit and three axial launch tube locations. In addition, the time and speed of first piston compression and muzzle velocity are recorded. The simulation has been validated to the point that it provides useful results for investigating the potential of the Wave Gun firing cycle. A preliminary parametric study of the firing cycle has been initiated. The gun geometry for the study is based on Eglin Air Force Base's current light gas gun model launcher. The results indicate that the Wave Gun firing cycle can be applied to improve launcher performance. For the Eglin Gun this would require fabrication of a new pump tube and barrel section.

1. Introduction

1.1 Background

1.1.1 Aeroballistics Range Facilities

Aeroballistics range facilities offer scientists and engineers a unique tool for research in hypervelocity aerodynamics. In essence, the aeroballistic range consists of a launcher for propelling aerodynamic models, and an instrumented range. The instrumented range consists of a series of noninvasive data collection stations at various distances along the flight path of the model. These allow determination of model speed, position and orientation. From this, fundamental aerodynamic characteristics of the model can be obtained. It is obvious that the launcher plays a critical role in the overall capability of an aeroballistics range facility. The launcher must be able to launch aerodynamic models to sufficiently high velocities while maintaining the integrity of the launch package. This translates to maximizing velocity and minimizing launch loads, while not exceeding the material constraints of the gun system.

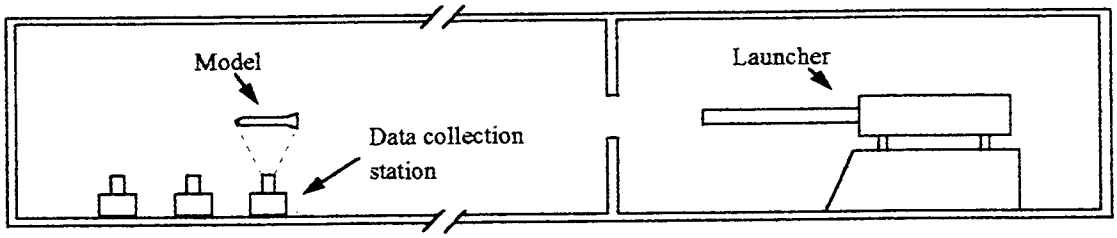


Figure 1: Aeroballistic Range Facility

1.1.2 Conventional Guns

In its simplest form a gun consists of a propellant chamber, a barrel or launch tube and a projectile (See Figure 2). The pressure in the propellant chamber is raised above the ambient by some means, usually the combustion of a solid propellant. This high pressure accelerates the projectile down the barrel where it leaves at a velocity termed the muzzle velocity. The fundamental equations governing the

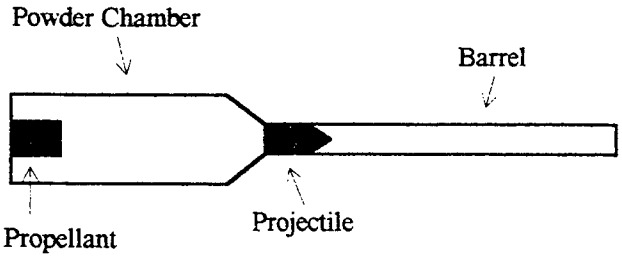


Figure 2: Conventional Gun System

velocity of a projectile being propelled down a gun barrel can be found by applying Newton's 2nd law to the projectile.

$$M\frac{du_p}{dt} = Mu_p\frac{du_p}{dx_p} = p_p A \tag{1}$$

where M is the mass, u_p is the instantaneous velocity, p_p is the instantaneous pressure at the base of the projectile and A is the cross sectional area of the barrel (barrel friction and air pressure downstream of the projectile are neglected here). Integrating Eq. (1) over the length of the barrel (L), gives the muzzle velocity

$$u_m = \sqrt{2\bar{p}\,AL/M} \tag{2}$$

where $\bar{p}$ is the average base pressure defined as

$$\bar{p} \equiv \frac{1}{L}\int_0^L p_p dx_p \tag{3}$$

From Eq. (2) it can be seen that to increase muzzle velocity the value of the terms under the square root sign must be increased. The values of the cross sectional area, barrel length and projectile mass are largely dictated by the design requirements for the gun. This leaves increasing the average base pressure

as the only way to increase muzzle velocity for a given gun configuration. The maximum value for p_p being determined by material constraints, it is desired to minimize the ratio of maximum base pressure to average base pressure. This ratio is termed the piezometric efficiency and is defined as

$$\eta \equiv \frac{p_{max}}{\bar{p}} \tag{4}$$

It is easy to see that the maximum muzzle velocity will be achieved when the average base pressure is constant and equal to the maximum allowable gun pressure. This corresponds to a piezometric efficiency of one. Seigel[1] shows that for a preburned propellant gun[i] with an ideal gas propellant and no chamberage[ii] the pressure drop behind the projectile is approximately

$$\frac{p}{p_o} = e^{-\gamma u/a_o} \tag{5}$$

where p_o is the initial propellant pressure, a_o is the initial propellant sound speed, u is the gas speed and γ is the ratio of specific heats. This equation applies to all parts of the expanding gas, including that at the projectile base, and is valid as long as the projectile leaves the barrel before a reflected rarefaction wave reaches the projectile[iii]. Despite the idealized constraints, Eq. (5) yields an important result which can be carried over to non-idealized guns. This is that the nondimensional pressure of the expanding gas in a gun is a function of the nondimensional velocity u/a_o and the ratio of specific heats. In order to maximize gun performance it is necessary to minimize the ratio of γ/a_o. Large changes in γ are not possible; for an ideal gas this range is from about 1.67 to 1.0[Ref. 20]. Therefore, a propellant gas with a

[i] Preburned propellant gun refers to a gun in which the maximum pressure has been reached in the propellant chamber before the projectile is allowing to move.
[ii] A gun in which the diameter of the barrel is smaller than the diameter of the propellant chamber is said to have chamberage.
[iii] A continuous series of rarefaction waves emanate from the base of the projectile into the propellant gas at the speed of sound when the projectile motion begins.

high initial sound speed is required to improve gun performance. For an ideal gas the sound speed is a function of the square root of the temperature divided by the molecular weight. This translates into the need for a low molecular weight gas at high temperature. The desire for low molecular weight can be immediately appreciated in light of the fact that the potential energy stored in the high pressure gas is expended accelerating both the gas and the projectile[2]. Accordingly, in a light gas less of the total energy is required to accelerate the gas with the obvious benefit that more energy is available to accelerate the projectile. Unfortunately, the molecular weight of the combustion products of available chemically reacting propellants is rather high. (Approximately 28 for nitrocellulose-based propellants[2].) Attempting to increase the temperature of the propelling gas introduces further complications, such as accelerated erosion of gun components due to higher temperatures. These formidable technological barriers limited the muzzle velocity of laboratory guns to a modest 10,000 fps until the late 1940's and the development of the two stage light-gas gun.

1.1.3 The Two Stage Light-Gas Gun

Some limitations of conventional guns are circumvented by inserting a second stage between the propellant chamber and the barrel. The second stage, or pump tube, is filled with a light gas, (usually hydrogen or helium) to some specified pressure. The light gas is sealed upstream by a movable piston and downstream by a frangible diaphragm.

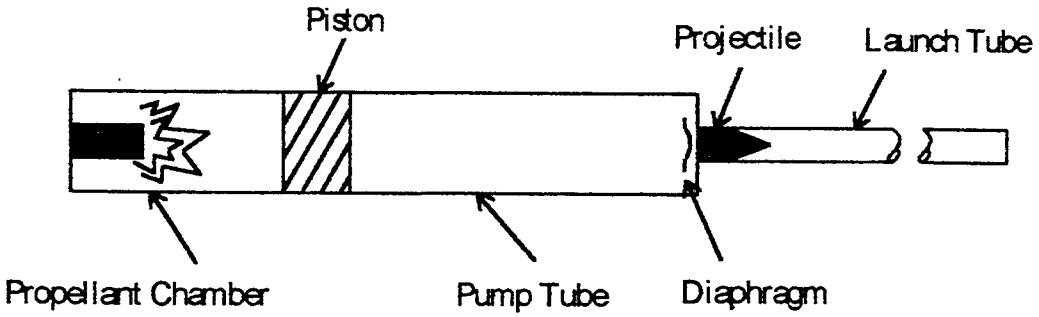


Figure 3: Two Stage Light Gas Gun

The operation of the light gas gun consists of raising the pressure in the propellant chamber by some means. This is typically accomplished by combustion of chemical propellants. The high pressure accelerates the piston which compresses the light gas. The diaphragm is used to prevent projectile motion

until a sufficiently high pressure is reached in the pump tube. Premature projectile acceleration would result in insufficient compression of the light gas and an attendant loss in gun performance. Similarly, some method is usually employed to retard the piston's initial motion until a specified pressure is reached in the propellant chamber. Once the pressure in the pump tube reaches the specified value, the diaphragm ruptures. This allows the light gas to expand, accelerating the projectile to higher velocities than can be achieved in conventional guns. The first light gas gun was developed at the New Mexico School of Mines in the late 1940's[3]. This gun could launch 4.5 gram projectiles to velocities approaching 12000 fps. More recently gun designers have been able to increase maximum muzzle velocities to over 37,000 fps[4].

Current light gas gun launchers typically operate on what has come to be known as the "isentropic compression" cycle. This cycle employs a heavy piston and a high volume pump tube with a low pressure gas charge, usually on the order of 200 to 600 psi. In this mode of operation the heavy piston moves slowly, nearly isentropically compressing the light gas. The large volume pump tube is used to maximize the compression ratio, hence increasing the temperature rise in the light gas with an attendant increase in the sound speed. Near the end of the piston's travel, the high pressure in the pump tube ruptures the diaphragm allowing the light gas to expand and accelerate the model.

1.1.4 Wave Gun

In 1981, in an effort to weaponize the light gas gun, Thomas Dahm[4] of Astron Research and Engineering invented a unique firing cycle. This cycle employed a light piston and a low volume pump tube with a high gas charge pressure. The small length to diameter ratio of the pump tube combined with a long propellant burn and the light piston sets up an oscillation of the piston near the downstream end of the pump tube. This piston oscillation results in multiple high pressure pulses which reach the projectile base before it leaves the launch tube.

In this mode of operation a conventional propellant charge is again used to accelerate the piston which is restrained from initial motion. When the pressure in the propellant chamber reaches the piston start pressure a rapid acceleration of the piston occurs. This rapid acceleration of the light piston increases the volume of the powder chamber at a rate such that the evolution of powder gases is not sufficient to maintain the propellant chamber pressure. Consequently, a peak pressure is reached in the

propellant chamber shortly after piston motion begins. This is followed by a high pressure in the pump tube. Unlike the isentropic compression cycle, at the end of the first Wave Gun compression the pressure in the pump tube is not sufficient to rupture the diaphragm. Instead, the piston rebounds due to the high pressure in the pump tube and the low pressure in the propellant chamber. The continued burning of the propellant again accelerates the piston downstream, this time rupturing the diaphragm. At this point a second rebound occurs followed by a third compression stroke which reaches the projectile before it leaves the barrel[i]. This firing cycle type with its light piston and small pump tube volume made it an attractive candidate for weaponization. A consequence of this firing cycle is that high velocities can be attained with lower peak accelerations than occur in a conventional LGG. This feature makes the Wave Gun an attractive candidate for use in aeroballistic ranges where there is substantial interest in decreased loading of aerodynamic models.

1.2 Motivation for Research

Current aeroballistic range facilities performance limits are dictated not by a launcher's maximum attainable muzzle velocity, but rather by the maximum loading that the sometimes fragile models can withstand while being accelerated down the launch tube (barrel). Considerable effort has been expended by the aeroballistics community toward the development of a launcher firing cycle that will produce high muzzle velocities with moderate model loading. Emerging technologies such as the electromagnetic rail gun and two-stage hybrid launcher[6] have had limited success but are not yet sufficiently mature for aeroballistic research applications. The standard isentropic compression cycle produces a peak acceleration on the model which decays as the model moves downstream. This acceleration cannot exceed the model loading limits and is therefore what limits launcher performance. Limited by the same maximum model loading, the Wave Gun has the potential of producing multiple acceleration peaks (see Figure 4), with correspondingly higher muzzle velocities. This performance capability was noted by Dahm and Randall[4] in their effort to weaponize the LGG. This potential, as it applies to high performance model launchers, has not been explored fully.

[i] This cycle is termed a 2-3 cycle since the projectile is accelerated on the second and third Wave Gun compressions. Other cycles such as a 1-2 or a 3-4 cycle are possible.

Due to the complexity of the various physical phenomena occurring in the Wave Gun (or any light gas gun), it is impractical to use solely analytical methods to study it. An entirely experimental research program would prove costly and inefficient. It is therefore desirable to employ a numerical interior ballistics simulation to investigate the Wave Gun's potential. However, the difficulty of predicting accurately all of the variables needed to model the Wave Gun necessitates the execution of a limited experimental program for the purpose of validating the numerical simulation.

1.3 Objectives

It is desired to perform a limited experimental program on the Wave Gun to obtain data for the purpose of validating and modifying the interior ballistics simulation. The simulation can then be used for exploring the Wave Gun concept. Specifically, it is desired to apply this concept to the LGG model launcher used at Eglin Air Force Base's Aeroballistics Range Facility. Initial studies on the Eglin gun will be used to assess the feasibility and usefulness of applying the Wave Gun firing cycle to the Eglin model launcher.

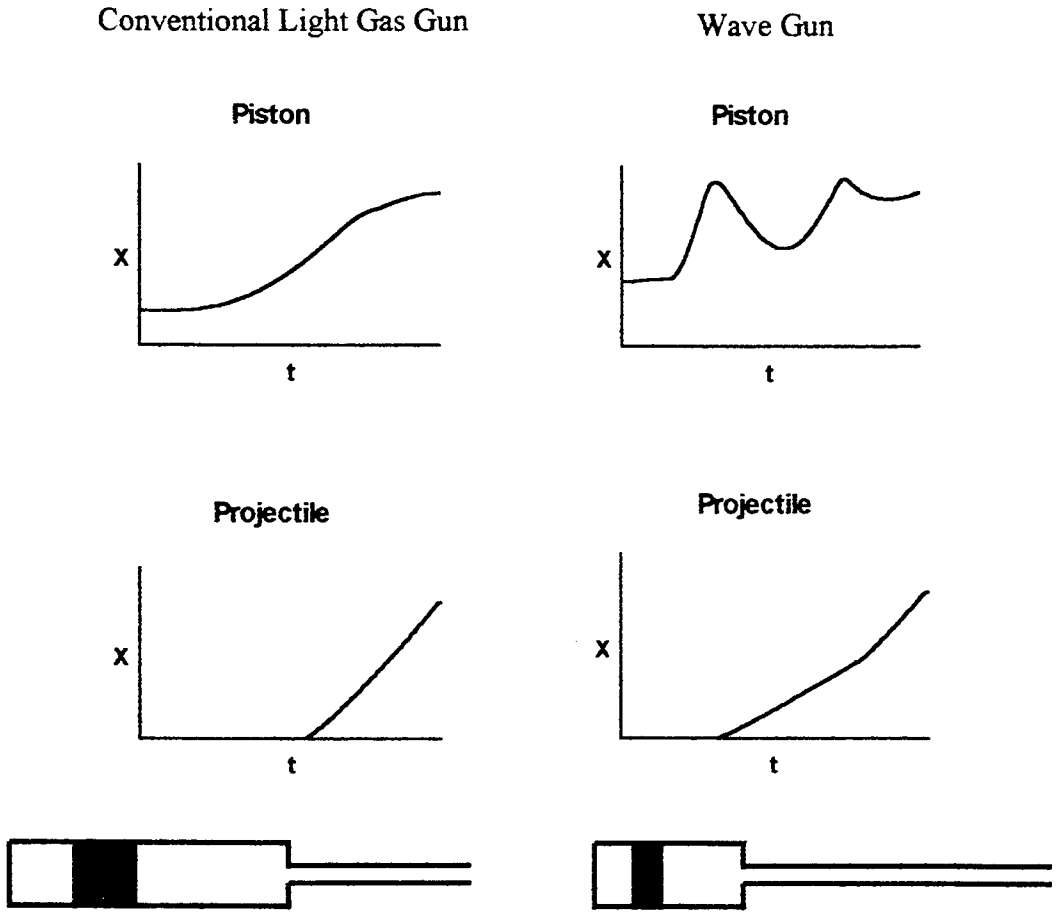


Figure 4: Firing cycle comparison

2. Firing Cycle Simulation

Arnold Engineering Development Center's (AEDC) current light gas gun code has been obtained for the purpose of investigating Wave Gun's capabilities[i]. The code was originally developed by Piacesi, Gates and Seigel[7] and has been extensively modified by DeWitt[8]. The heart of the simulation is a quasi-one-dimensional, hydrodynamic algorithm based on the "q" method of Von Neumann and Richtmyer[9]. The "q" method is a Langrangian scheme which allows for the automatic treatment of shocks through the addition of an artificial dissipative term to the governing equations for adiabatic flow. The governing equations with the addition of the q-parameter are shown below:

Energy equation

$$\frac{\partial E}{\partial t} = -(p+q)\frac{\partial V}{\partial t} \tag{6}$$

Equation of state

$$p = p(E,V) \tag{7}$$

Equation of motion

$$\frac{\partial u}{\partial t} = -\frac{\partial (p+q)}{\partial M} \cdot A(x) \tag{8}$$

where the mass M is

$$M = \int_0^x \rho(x) \cdot A(x)\, dx \tag{9}$$

where q is defined as

$$q = \frac{C_o^2}{V}\left(\frac{\partial u}{\partial j}\right)^2 \quad \text{, if} \quad \frac{\partial u}{\partial j} < 0 \qquad\qquad q = 0 \quad \text{, if} \quad \frac{\partial u}{\partial j} \geq 0 \tag{10}$$

The differential equations are discretized and applied to the gun system by dividing the space into regions

[i] The AEDC code can be obtained from Dr. Robert W. Courter, Department of Mechanical Engineering, Louisiana State University.

for propellant, piston and light gas, each having its own equation of state and initial conditions. Each region is further subdivided into zones to achieve a desired computational resolution. Mass points are concentrated at the interface of zones such that half the mass of a zone is represented at its left and right interfaces. Shown below is a simple schematic of the one-dimensional finite difference grid.

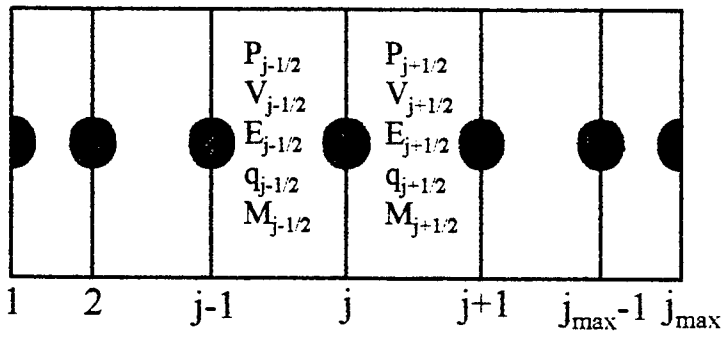
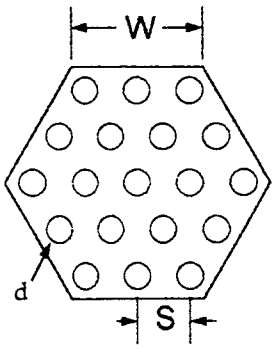


Figure 5: "q" method point mass model

The initial conditions of velocity, pressure, density, specific volume and internal energy are specified for each zone. The finite difference equations, which appear in Appendix A, are numerically integrated to obtain the new values of the variables.

The code includes de Saint Robert's power law equation for propellant burning, virial-type real gas model for the light gas (helium for the present case, see Appendix B), dissipative influences such as gas friction and heat transfer as well as piston sliding friction and deformation models. Modeling of the actual M30/19 propellant grain (Figure 6) is accomplished by tabulating the surface area of the grain at progressive stages of burn. This is necessary since the evolution of powder gases is a function of the rate of burn into the grain surface and the burn area of the propellant grain. The former is determined by de Saint Robert's equation, and the latter is calculated on the basis of uniform surface burning. In addition the present author has tailored the code to meet the requirements of the present program. These modifications include the handling of projectile sliding friction and the addition of a piston start rupture, pressure,which are not modeled in the original code. Furthermore, the projectile release or diaphragm which is modeled as an instantaneous event in the original code, has been changed.



where: W = 0.195313 in
S = 0.078125 in
d = 0.02 in
length = 0.5 in (not shown)

Figure 6: M30/19 propellant grain

Both the piston and projectile releases are now modeled as having a linearly decaying retarding force which acts over the distance required for the parts to move free of restraint. The projectile sliding friction is modeled as a retarding force which is a function of the base pressure times the contact area and a friction coefficient (see Appendix C).

Input for the simulation is read from the data set LINPUT.DAT. The main input parameters can be summarized as follows:

1) Code initialization:

Number of regions

Number of zones associated with each region

Equation of state for each region (solids, liquids, gas)

Printout cycle

2) Gun geometry:

Propellant chamber volume

Pump tube length and bore

Piston length

Area reduction geometry

Launch tube bore and length

3) Propellant properties:

Density and grain volume

Burn rate coefficient and exponent

Combustion gas properties

4) Shot conditions:

Propellant load

Piston start pressure

Light gas charge pressure

Model start pressure

Piston material and density

Launch package weight

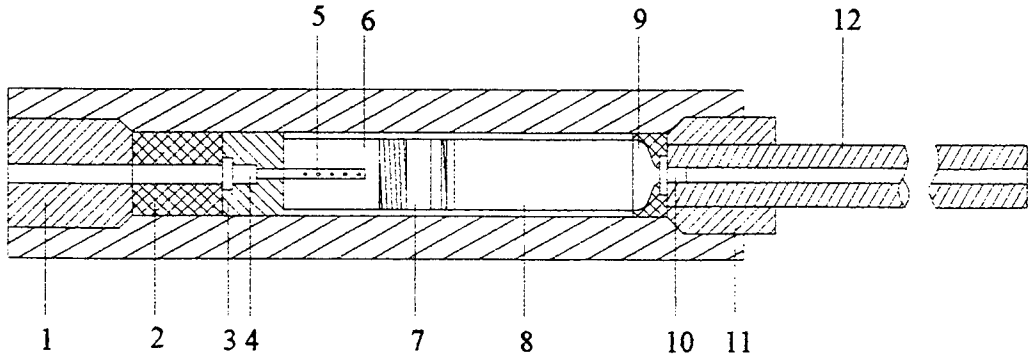
The simulation provides complete temporal and spatial output of velocity, pressure, internal energy and reduced specific volume. The projectile velocity and acceleration are also recorded.

3. The Experimental Program

3.1 Experimental Facility Description

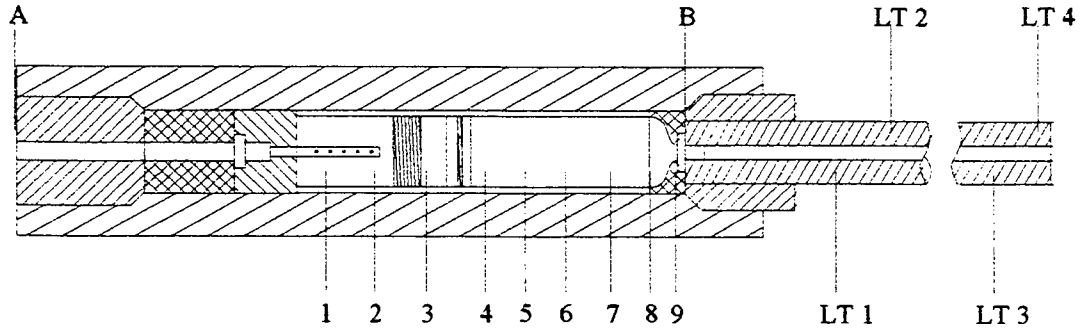
The present experimental facility utilizes the Astron 30 mm Wave Gun originally designed to investigate light gas gun weaponization. It was recommissioned in July, 1994, at Eglin Air Force Base in support of the present research program. In order to achieve the flexibility needed in the original research program, the Wave Gun uses a massive steel pressure vessel to house the gun's internal parts. This design allows the internal geometry of the launcher to be modified by inserting sleeves of various size into the pressure vessel. A schematic is shown in Figure 7. Starting upstream, the internal parts consist of the outer breech plug, spacer, inner breech plug and igniter assembly, propellant chamber, piston, pump tube, nozzle, projectile, barrel nut and barrel. Although the facility originally had three sets of internal parts, only one configuration is available at present.

The outer pressure vessel contains nine instrumentation ports, (hereafter referred to as gun ports 1-9) with an additional four instrumentation ports located in the eight foot launch tube (Figure 8). Gun ports (GP) 2,8,9 and launch tube ports (LT) 1,2,3 contained quartz type piezoelectric pressure transducers which measured pressure histories in the propellant chamber, nozzle entrance, nozzle exit, and three axial launch tube locations, respectively. The transducer signals were sent to charge amplifiers before being recorded by a High Techniques (HT) 600 digital oscilloscope. A novel approach to obtaining piston velocities including oscillations was implemented by Astron Research and Engineering for this gun. It consisted of placing a metal band around the piston which was to shunt a circuit between the pump tube wall and several insulated probes which barely protruded into the pumptube from different axial locations. The data were then multiplexed and sent to a digital waveform recorder. Unfortunately the resulting data were obscured with many anomalies, making the data unreadable. The cause was thought to be motion of the internal sleeves during the shot. Although this method seems workable with some refinement, a simpler, albeit less informative, approach was adopted. Instead, GP 5 and GP 6 contained breakwires which consisted of a wooden dowel with a thin loop of insulated wire glued axially along the dowel and a



No.	Part	Length (cm)	Diameter (cm)	No.	Part	Length (cm)	Diameter (cm)
1	Breech plug (outer)	-	-	7	Piston	10.29	11.43
2	Spacer	16.04	13.29	8	Pump tube	41.81	11.43*
3	Breech plug (inner)	12.07	13.29	9	Nozzle	15.24	11.43*
4	Igniter	-	-	10	Model	5.72	3.00
5	Spit tube	-	-	11	Barrel nut	-	-
6	Propellant chamber	28.19	11.43*	12	Launch tube	243.84	3.00
					* Inside diameters		

Figure 7: Wave Gun test apparatus



No.	Location (cm)*	Use	No.	Location (cm)**	Use
1	45.72	Not active	LT 1	45.72	Pressure transducer
2	60.96	Pressure transducer	LT 2	76.20	Pressure transducer
3	71.12	Not active	LT 3	137.16	Pressure transducer
4	81.28	Not active	LT 4	198.12	Not active
5	91.44	Break wire			
6	101.60	Break wire			
7	111.76	Not active			
8	121.92	Pressure transducer	* Measured from A		
9	128.11	Pressure transducer	** Measured from B		

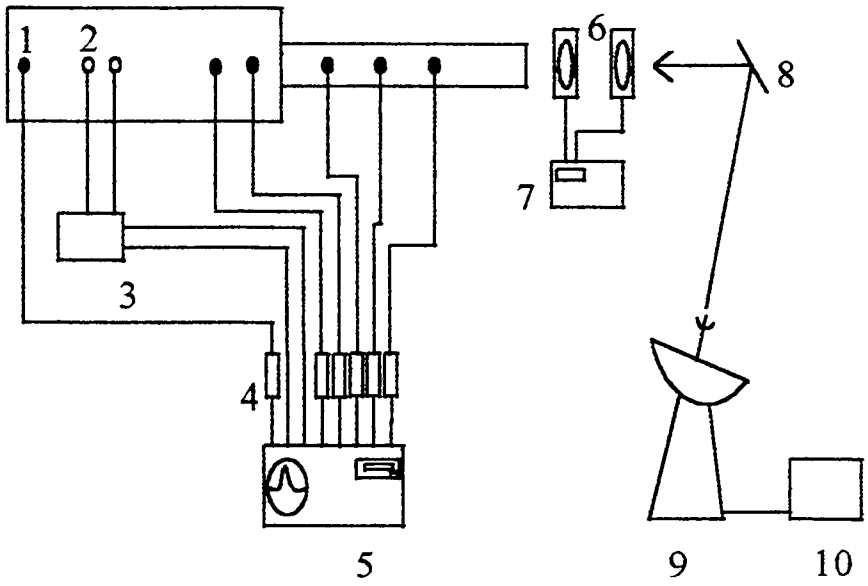
Figure 8: Wave Gun instrumentation ports

3/8 in. bolt to which the dowel was epoxied. The breakwire assembly was screwed into the pumptube to measure the time of the first piston compression. The dowel acted as a stiffener for the wire so that it would shear cleanly upon piston passage. The breakwire signals were also recorded by the HT600 and all signals were on the same time base. The remaining instrumentation ports were inactive. Muzzle velocity was recorded by either two infrared sky screens or Doppler radar. The sky screens were placed a known distance apart and were used to trigger counters as the projectile passed overhead. Using this time and distance an average velocity was obtained. The radar unit, which was used when available, also has the ability to measure inbore velocity and acceleration. Neither the sky screens nor the radar were on the same time base as the other instrumentation. Figure 9 shows the instrumentation setup.

Loading and firing of the gun consists of inserting the nozzle and pump tube sleeves into the outer tube. Then a projectile is inserted into the upstream end of the launch tube and the barrel nut is screwed onto the pressure vessel. The model itself is an aluminum cylinder with an integral flange which serves as the diaphragm (Figure 10). The flange thickness can be any size up to a half inch, with a steel washer used to insure an adequate gas seal with the nozzle if the flange is less than a half inch thick. Next, the inner breech and igniter assembly is threaded onto the upstream end of the propellant chamber. The propellant charge is bagged and inserted around the spit tube. The polypropylux piston (Figure 11) is threaded on one end so that it can be screwed onto the downstream section of the propellant chamber. This makes it possible to vary the piston start pressure by varying the number of threads engaged. The propellant chamber assembly is then inserted. The pump tube, nozzle and propellant chamber must all be carefully aligned both axially and radially so that the instrumentation ports in the sleeves align with the gun ports. Final assembly consists of inserting the spacer and outer breech plug and simultaneously torqueing the breech plug and barrel nut without disturbing the sleeves' alignment. The helium can then be added through a fill port located in the pump tube. The gun is then fired using a 20 mm electric primer.

3.2 Experimental Results

Since being commissioned at Eglin AFB in July of 1994, nine Wave Gun shots have been fired. The first two shots occurred during the 1994 Summer Research Program and are discussed in Reference



No.	*Instrument*	*Model*
1	Piezoelectric pressure transducers	Kistler 60704
2	Breakwires	-
3	Breakwire power supply	-
4	Charge amplifiers (1-4)	Kistler 504E4
	Charge amplifiers (5-6)	PCB 463A
5	Digital oscilloscope	Hi Techniques HT-600
6	Sky Screens	-
7	Sky Screen counter	-
8	Sacrificial mirror	-
9	Radar head	Opus Electronics
10	Radar analyzer	Terma DR-5000

Figure 9: Experimental setup

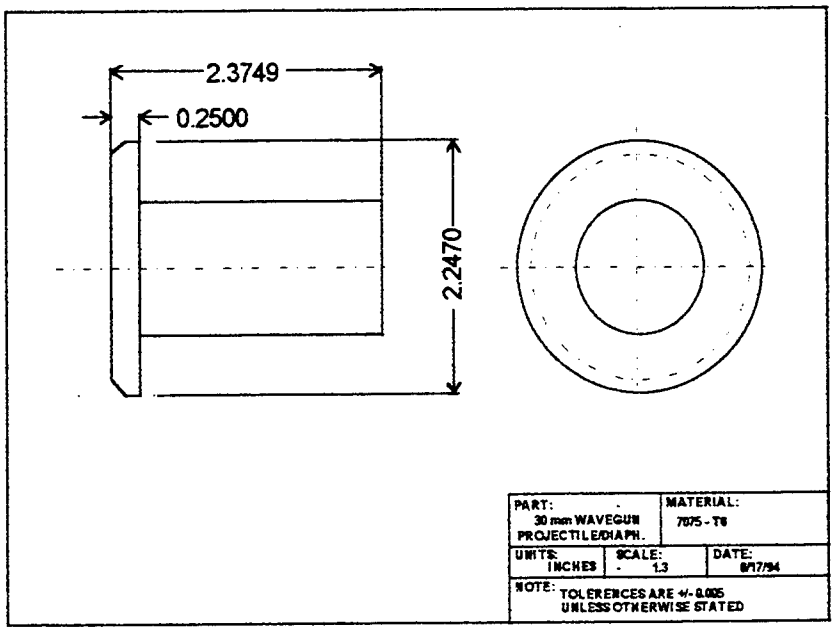


Figure 10: Wave Gun projectile

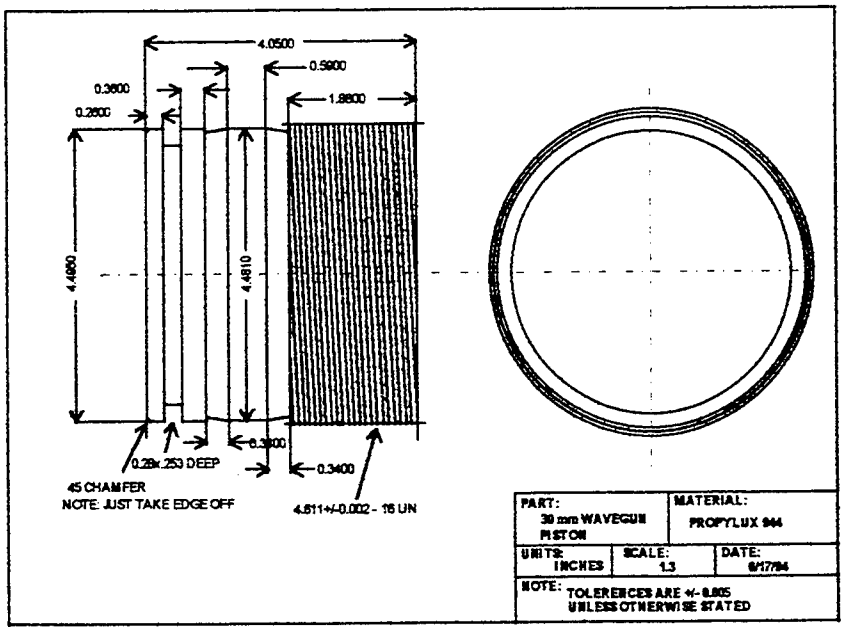


Figure 11: Wave Gun piston

10. Of the remaining seven shots three yielded little or no data. These three shots are considered useless for the purpose of code validation and are omitted from the following discussion. The four "good" shots (those yielding useful data) consisted of two pairs of nearly identical shots. All shots used M30/19 MP as the main propellant with a 16-17 gm FFFG black powder primer charge in the spit tube. The instrumentation was triggered from LT 1. This was done to alleviate triggering problems that might occur in the gun ports if the alignment of the internal parts was off. The HT600 has the ability to record events that occur before and after the trigger. This is crucial since the trigger occurs after a sizable portion of the ballistic event has occurred. The HT600 was set to sample at 500 kHz for 32 ms, which gave over three times the sweep time needed, with the actual ballistic event taking less than 10 ms. The parameters for the four shots used in the present experimental program are shown in Table 1 with the corresponding results in Table 2.

Table 1: Shot parameters

Shot number	*Shot date*	*Propellant weight (gm)*	*Model weight (gm)*	*Distance piston screwed in (in)*	*Flange/ Diaphragm thickness (in)*	*Helium charge pressure (psi)*
1	4/14/95	1280	111.6	1.88	1/4	2300
2	6/21/95	1280	111.6	1.88	1/4	2700
3	6/26/95	800	111.6	1.88	1/8	2200
4	7/06/95	800	111.6	1.88	1/8	2200

Table 2: Shot results

Shot number	*Propellant chamber pressure at piston start (psi)*	*Model start pressure (psi)*	*Cycle type*	*Muzzle velocity (fps)*
1	12000	40000	3-4	5626
2	12000	44000	3-4	6113
3	7500	25000	2-3	4143
4	8000	30000	1-2	3733

The shot parameters for shots 1 and 2 differ only in the helium charge pressure for reasons discussed later in this section. For Shot 1 full sweeps were recorded at all pressure port locations. However, this shot was performed before the HT600 oscilloscope was equipped to handle all eight channels. The result was that LT 2 and LT 3 were not on the same time base as the rest of the traces and therefore proved useless in code validation. In addition the breakwires did not perform properly and no

data were recovered from them. It should be noted that the breakwire assembly discussed in the previous section was actually a modification in response to the results of Shot 1 (the former assembly lacked the dowel stiffener). The muzzle velocity (see Table 2) was recorded using the infrared sky screens placed ten feet apart, with the first one approximately 10 feet from the muzzle. This setup is necessary to prevent muzzle blast from triggering the sky screens. The polypropylux piston was completely extruded and shot down range. For Shot 2 all eight channels were recorded successfully on the HT600 oscilloscope. The muzzle velocity was successfully recorded using the radar, and the piston was found to have wedged in the nozzle. Shot 3 was a success with the exception that GP 5 and GP 9 did not return data. It was discovered after the shot that rain water had leaked onto the electrical contact of the GP 9 transducer due to improper sealing of the transducer port between shots. The breakwire bolt in GP 5 fractured and subsequently blew out before piston passage occurred. The muzzle velocity was determined from radar. On Shot 4 LT 3 failed to return data for reasons never determined. It was decided for this shot however, to attempt to obtain inbore velocities using the radar. Although this capability was known to exist from the beginning of the experimental program, difficulties with the rest of the instrumentation belated its use. This setup consisted quite simply of using an ordinary mirror to give the radar head a line of sight down the launch tube. It was also necessary to apply a correction factor to the data based upon the barrel bore and radar frequency to account for the barrel acting as a wave guide. The results were pleasing with a continuous track starting from approximately 250 fps to the muzzle velocity being recorded. The inability to monitor the initial model movement is a limitation of the radar head that was used and was expected. The data recovered also indicated two clear acceleration peaks of 34000 and 70000 G's respectively, confirming the Wave Gun cycle concept. As Tables 1 and 2 indicate, the last two shots, intended to be identical, in fact executed different cycle types. This may be the result of a severe helium leak occurring through one or both of the breakwire ports. It was noted that after charging the pump tube to approximately 2600 psi a severe leak developed which was said by the range technician to have "stabilized" at 2200 psi. There is, needless to say, some ambiguity in how "stable" the leak became and what the actual pressure was at shot time, noting that several minutes elapse between final loading and shooting. However, the piston speed data (Table 4) does not support this hypothesis. A low charge

pressure would in turn produce a high piston speed, but the experimental value is nearly half the predicted value. For both Shot 3 and Shot 4 the piston was found wedged in the nozzle.

As with the initiation of virtually any experimental program, technical and logistical difficulties were encountered which impeded progress. One such difficulty was simply scheduling a time when the necessary equipment and personnel would be available to perform a shot. Also some wear of the gun components is evident, the most significant being at the meeting face of the pump tube and nozzle where erosion is occurring. A multitude of hairline cracks on the outside surface of the nozzle are, also visible. This wear has not affected gun performance nor has it compromised safety. In addition the wear seems to be increasing at a decreasing rate. However, the most dominant problem that has plagued the experimental program to date has been the inability to obtain high pressure helium. The small pump tube volume requires the helium charge pressure to be exceedingly high, 2500-4000 psi, compared to about 200 psi for standard light gas guns. The only suitable helium source found was from the Navy Dive School in Panama City, Florida, where two standard scuba tanks (80 ft^3 standard air at 3000 psi) were obtained for the tests. Staging from standard helium bottles this allowed for a maximum pressure of 2700 psi in the pump tube. The helium deficiency was also related to other difficulties. Leaks frequently occurred in the gun when the helium pressure was much greater than 2000 psi. When this occurred a judgment call had to be made on how severe the leak was and whether to dump the helium and secure the leak or continue the test. Dumping the helium and securing the leak meant that the shot would have to be made at a lower pressure or completely aborted until the scuba bottles could be refilled.

4. Numerical Studies

4.1 Wave Gun Simulations and Code Validation

Most of the initial inputs to the simulation code were based upon first principles or documented empirical values. The two most notable exceptions to this were the piston and model start pressures, which were derived from the experimental pressure histories. In regard to these values a large degree of variability was noted for the piston start pressures. It became apparent that the modeling of the piston and projectile releases as instantaneous events was not adequate. Based on these observations the piston and projectile breakout models were modified as outlined in section 2. The initial values for the piston and model starts were then based on the theoretical pressures required to shear the threads and flange, respectively. These values proved to be good initial estimates as can be seen in Table 3. The theoretical values were calculated using the distortion energy theory to determine the ultimate shear strength of the respective materials. The final values decided on for use in the simulation were the ones which gave the best match with the experimental data.

Table 3: Shot start pressure comparison

	Theoretical pressure (psi)	*Pressure used in simulation (psi)*
Piston screwed in 1.88 inches	4038	4350-5800
1/4 inch model flange thickness	2901	2900
1/8 inch model flange thickness	1451	1200

Note that it was not possible to determine the actual piston and model start pressures based solely on the experimental pressure histories. This is because the pressure in the powder chamber continues to rise as the piston slowly breaks free of restraint, and similarly the pump tube pressure continues to rise as the model shears its flange. What was done for the purpose of code validation was to match, as well as possible, the first peak in the powder chamber pressure history for the simulation to that of the experimental pressure history. This provides a close match at initial piston motion. Table 4 provides a summary of the simulation results and the respective experimental values. For the high performance shots (Shots 1 and 2, Figures 12-13) the first compression and rebound by the piston are predicted exceptionally well. After the second compression, however, the powder chamber pressure begins to be underpredicted. This underprediction results in an exaggerated piston rebound, which is indicated by underpressures in the pump tube occurring during piston rebound. Dahm and Randall[4] experienced this as well and attributed it to erosive burning. Erosive burning is the phenomenon whereby the burning rate of a solid propellant increases due to high gas velocity near the surface of the propellant grain. The conditions for

Table 4: Experimental and numerical results

Shot number		*Powder chamber pressure at piston start (psi)*	*Pump tube pressure at model start (psi)*	*Average piston speed between GP 5 & GP 6 (fps)*	*Maximum pressure (psi)*	*Muzzle velocity (fps)*
1	Numerical	11362	67370	1089	108750	6030
	Experimental	12000	40000	------	76000	5626
2	Numerical	11848	58715	1043	94631	6031
	Experimental	12000	44000	1089	82000	6113
3	Numerical	8402	24889	750	37192	4580
	Experimental	7500	25000	------	25000	4143
4	Numerical	7808	24584	1048	135180	3709
	Experimental	8000	30000	617	30000	3684

erosive burning are very favorable for this particular firing cycle. One can visualize the situation that arises when the propellent grains,which are being convected downstream with the flow, suddenly experience a flow reversal due to the piston rebound. The heavy, fast moving propellant grains will not be able to change directions as quickly as the lighter propellant gas. This would result in high relative velocities near the grain surface. However, although the conditions for erosive burning are favorable in this type of cycle, it is usually considered insignificant in propellants with large grain structures, such as the present case. At any rate, propellant gases appear to be evolving more quickly than predicted. Attempts to adjust the documented burn rate coefficient and exponent accordingly have not had the desired effect. Despite the underprediction of the propellant gas evolution and the excessive piston rebound, the timing of the pressure waves in the gunports is predicted very well.

The launch tube data are used in part to determine the arrival, or more correctly, the passage of the model, which is indicated by a sudden sharp pressure rise at the respective pressure port. It is also essential for identifying the cycle type for the experimental case. For Shot 1 data were only recovered for LT 1. The pressure curves here indicate the model arrival to be about a half a millisecond too fast. This is further reflected by a muzzle velocity that is predicted a little more that 7% high. Some of this variance, however, can be attributed to the use of sky screens to obtain muzzle velocity. Sky screens, which rely on the reflection of an infrared beam off the projectile, can be triggered by muzzle blast and must be placed some distance downstream of the muzzle to work properly. For Shot 2 the model arrival times are very well predicted for all launch tube locations. This agrees well with the prediction of muzzle

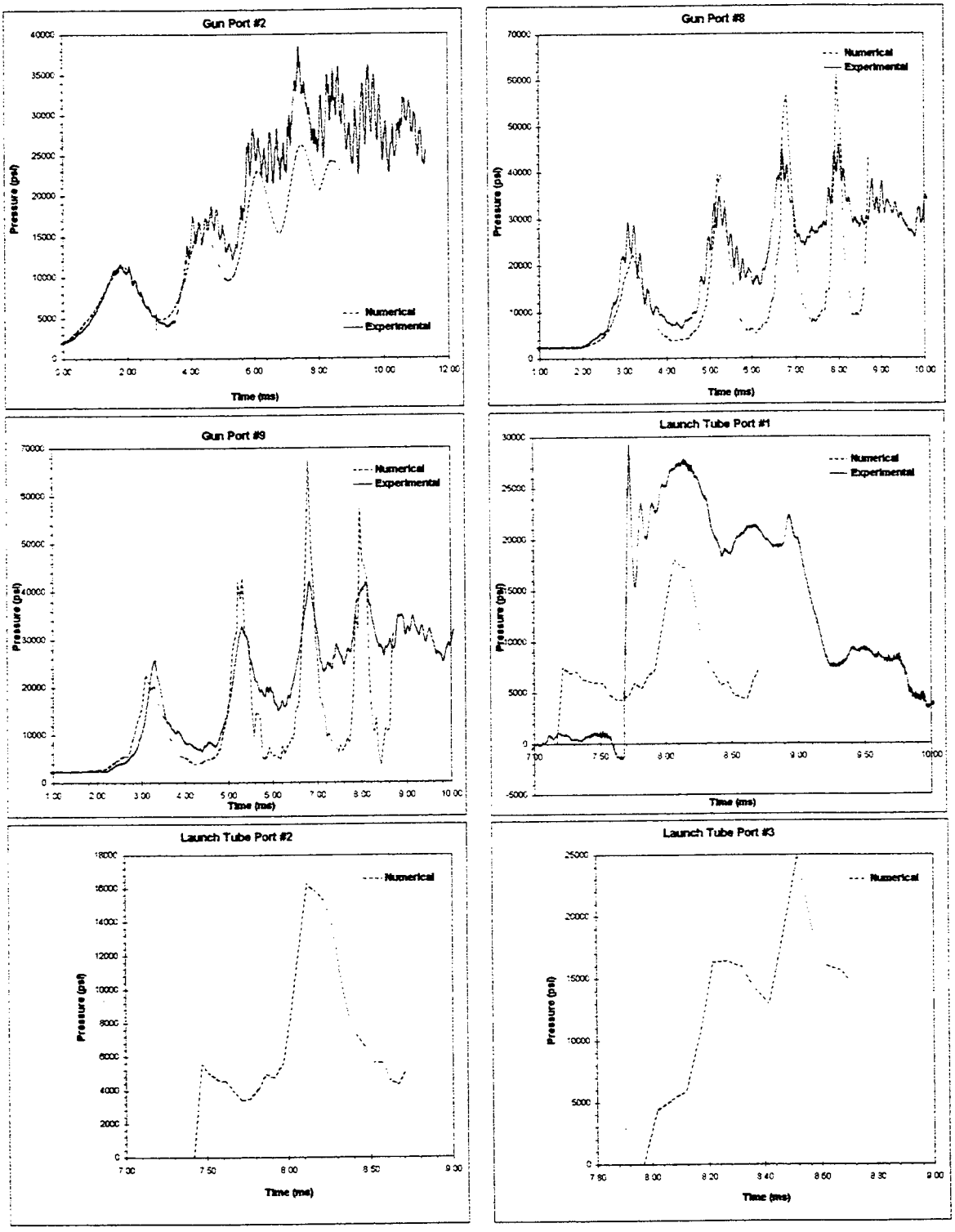


Figure 12: Shot 1

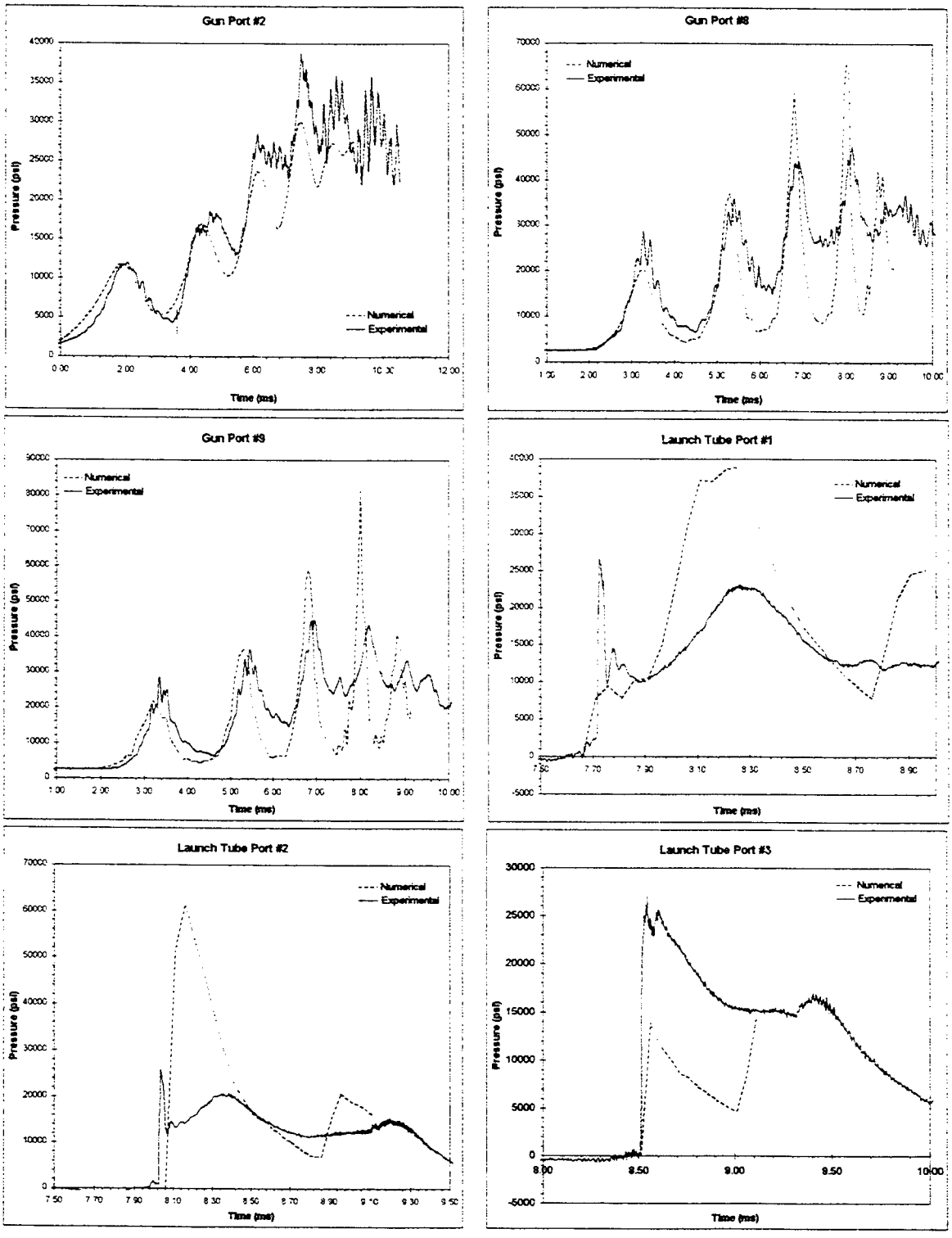


Figure 13: Shot 2

velocity within 2% of actual. Qualitatively and quantitatively the similarity in the launch tube data ends with the model arrival times. It is thought that the underprediction of the propellant burning coupled with the length of the ballistic event, (four piston compressions) may be the culprit. This is, however, highly speculative and at present there is no clear explanation why the launch tube data are so poorly simulated for these shots. For Shot 2 the piston speed is predicted about 4% slow. For Shot 1 the experimental piston speed data were lost. The predicted value was slightly higher than for Shot 2, owing to the lower helium charge pressure, and seems very reasonable.

The pressure histories for the entire cycle seem for the most part to be better predicted for the third and fourth shots (Figures 14-16). The deficiency in propellant gas evolution noted for the first two shots is also not as evident. This may be due in part to the fact that these shots execute fewer cycles preventing inadequacies of the numerical model from compounding. The pressure histories at the launch tube locations seem better predicted for these lower performance shots in that they are qualitatively more similar. The magnitudes for the same are over predicted with a corresponding overprediction of the muzzle velocities. The peak accelerations on the model, which were measured for Shot 4 (Figure 16), are overpredicted by significantly more than 100% by the simulation. This is especially troublesome in light of the fact that the present research is specifically interested in model accelerations. The discrepancy between the accuracy of the velocities and accelerations initially may seem counter intuitive. This disagreement is made clearer when one considers two things: the underpredicted projectile acceleration during piston rebound, and the smoothing effect of integrating acceleration to get velocity.

The simulation in its present state has shown the ability to predict cycle timing, cycle type and velocities fairly well. The pressure histories are also fairly well predicted for significant portions of the ballistic event and are comparable to results obtained in other numerical studies[4,5]. Considering the complexity of the ballistic event and the relative simplicity of the simulation, the results are acceptable and, more importantly, useful. Even though further improvements could probably be made to the simulation it would be desirable to have data from several more experimental shots to do so. The experimental shots are costly and time consuming and are not deemed appropriate at present.

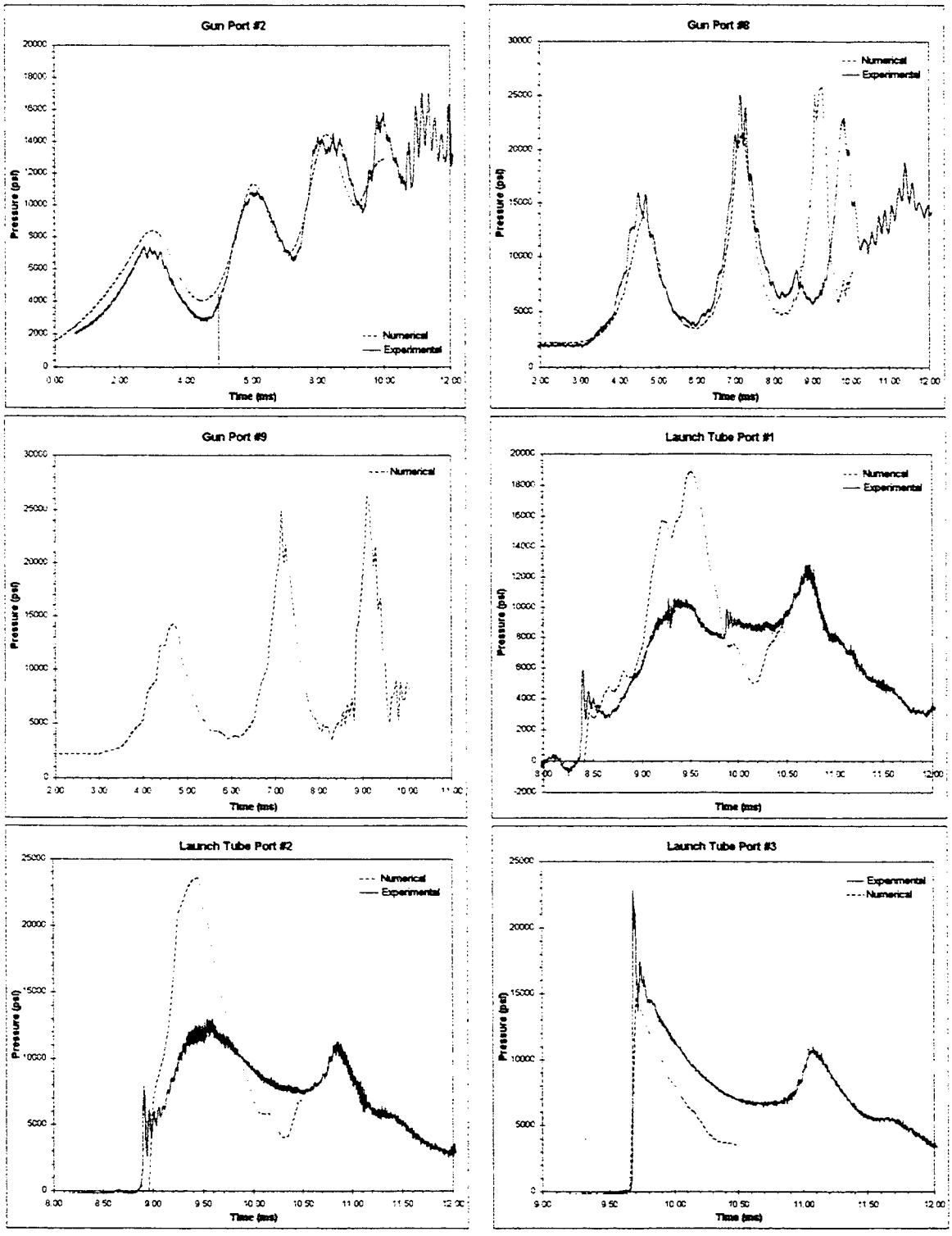


Figure 14: Shot 3

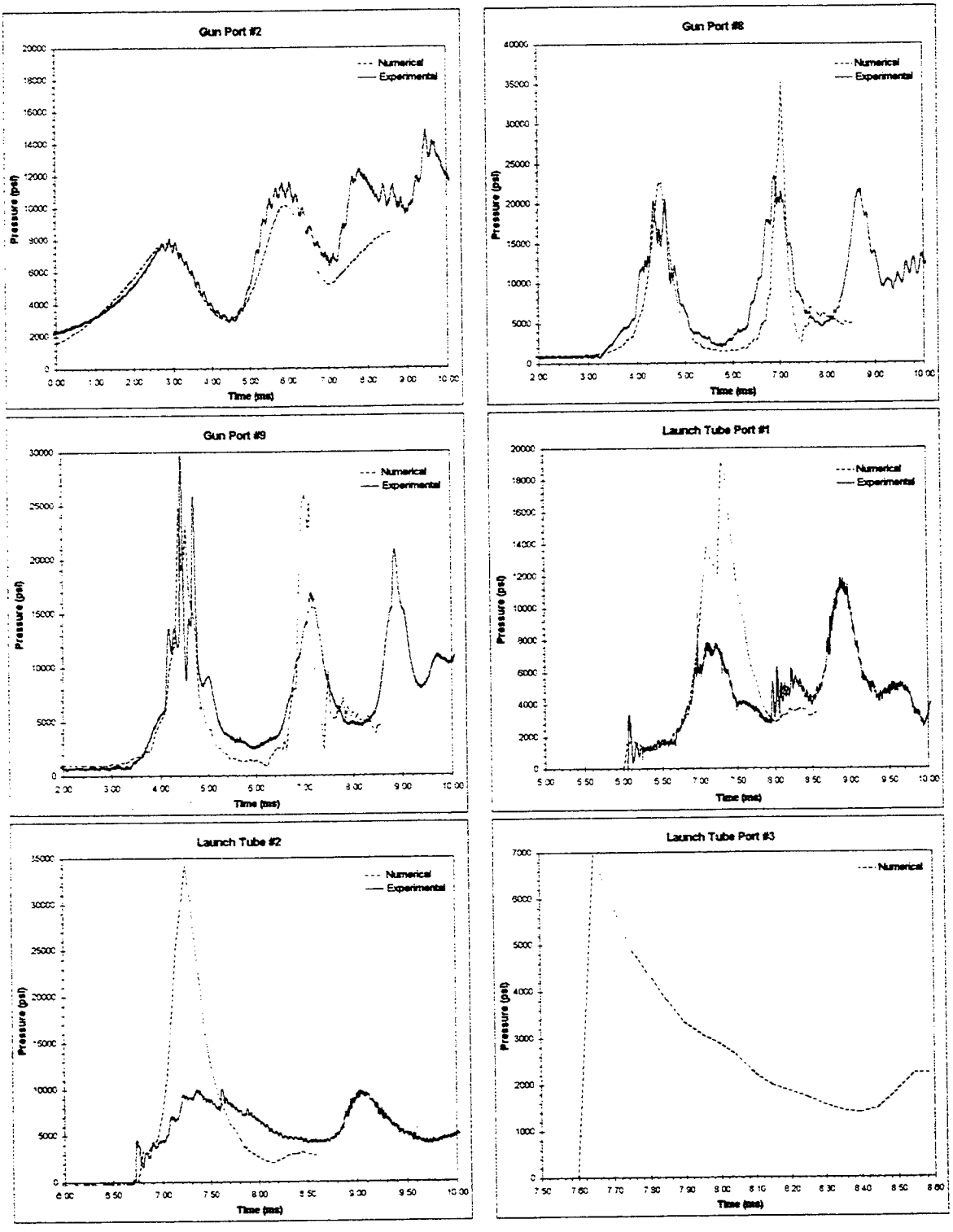
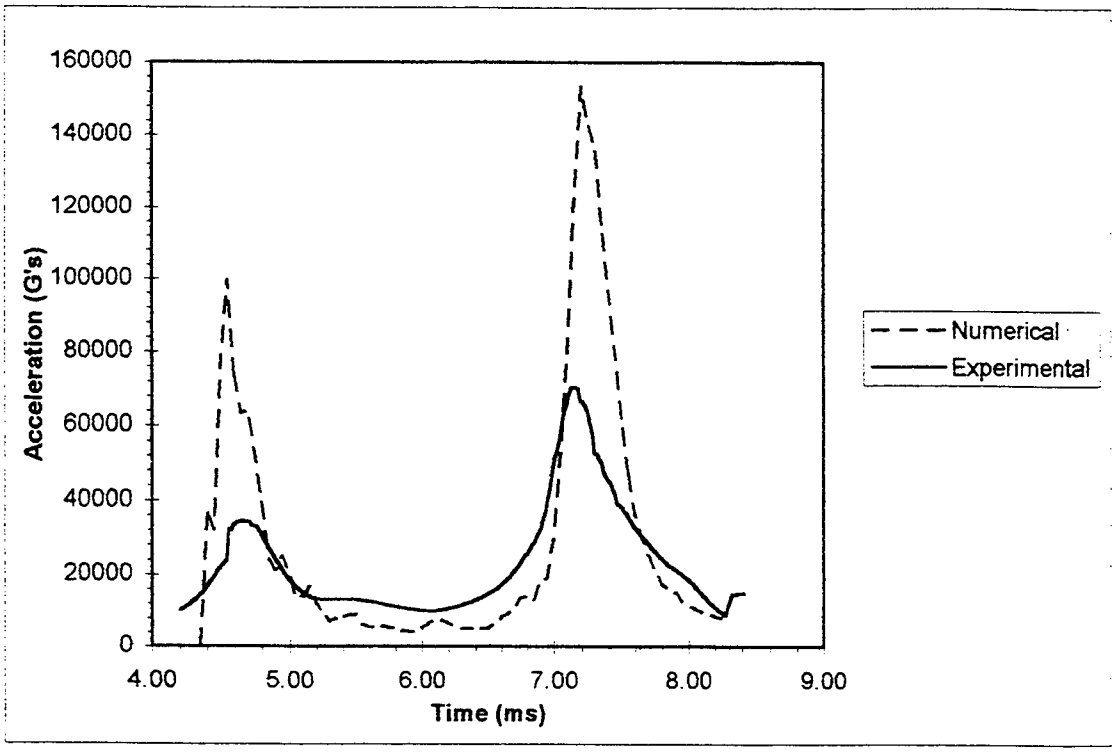


Figure 15: Shot 4



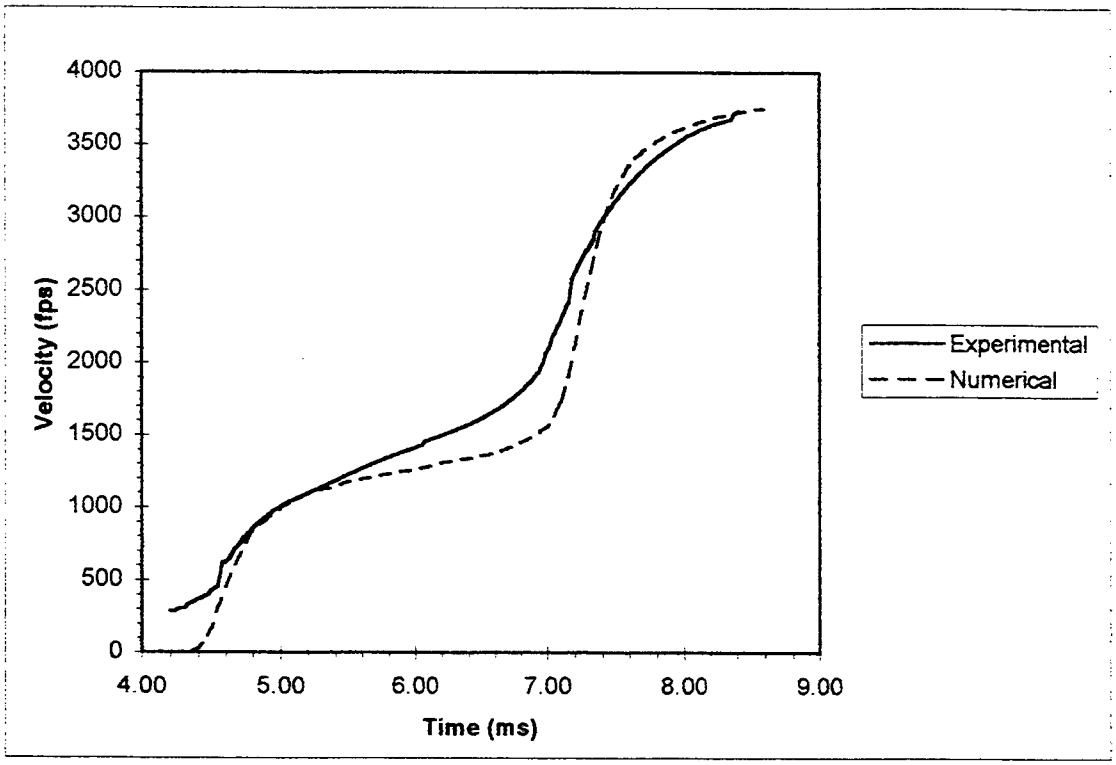


Figure 16: Shot 4 radar data

It is felt that the simulation's performance is adequate for an initial study of the Wave Gun firing cycle's potential to reduce model loading.

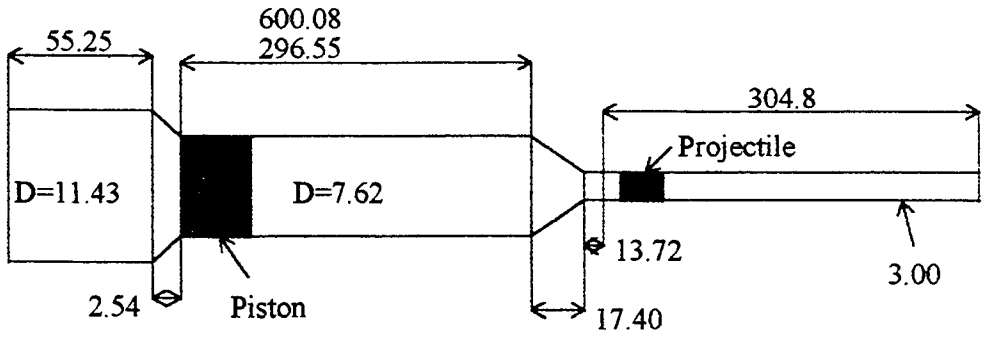
4.2 Eglin Light Gas Gun Simulations

A limited parametric study was performed on the LGG (Figure 17) used at Eglin Air Force Base's Aeroballistic Range Facility. It was desired to explore the usefulness and feasibility of adapting the Wave Gun firing cycle to the Eglin facility. First, it was necessary to simulate the isentropic compression cycle on which the Eglin Gun now operates. Unfortunately, the Eglin Gun is not instrumented, and therefore, no pressure history data for this gun are available. Consequently, the simulated results could only be compared to the experimental muzzle velocities. Shot parameters and muzzle velocities were obtained from a shot log of actual free flight aeroballistic tests at the Eglin facility. After some tailoring of the propellant burn rate exponent, satisfactory simulations were obtained. Seven simulations were run with muzzle velocities ranging from 3560 feet per second to 5928 feet per second. These are for comparison to the Wave Gun type firing cycle to be adapted to the Eglin Gun. These results are shown in Table 5.

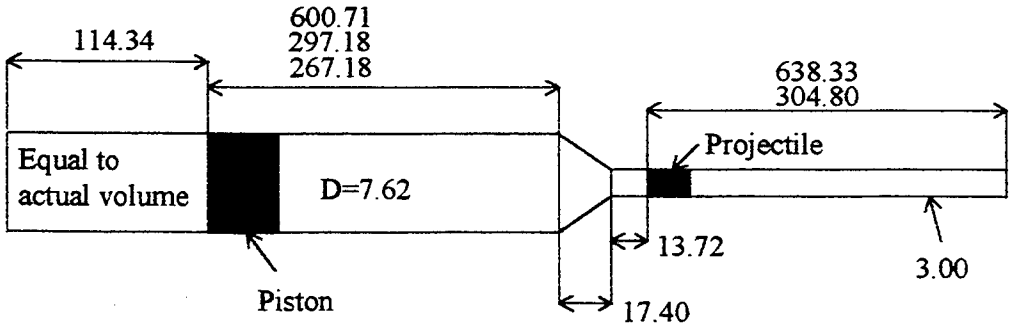
Table 5: Isentropic compression cycle simulation results

Propellant charge (lbm)	*Helium Charge pressure (psi)*	*Shot start pressure (ksi)*	*Maximum gun pressure (psi)*	*Muzzle velocity (fps)*	*Maximum projectile acceleration (G's)*
0.92592	514	4	13378	3560	24260
1.08024	514	4	17308	3963	31922
1.23456	514	4	21574	4371	40726
1.38888	514	4	26484	4770	49373
1.54320	514	4	32523	5164	57600
1.69752	514	4	38008	5551	68636
1.85184	514	4	45024	5928	81809

The Wave Gun Cycle (WGC) was adapted to the Eglin Gun by decreasing the piston weight and increasing the helium charge pressure. For the isentropic case the plastic piston is made heavy by inserting a lead weight in its core. The reduction in piston weight is effected by removing this lead and by reducing its length to the smallest value possible which will still prevent it from tumbling in the pump tube. A length equal to its diameter was decided to be sufficient to prevent tumbling. The resulting



ACTUAL GUN GEOMETRY



MODELED GUN GEOMETRY

Figure 17: Eglin Gun geometry

piston weight was 0.7 lbm. The role of pump tube length on gun performance for the WGC was the first parameter considered. There are at present two pump tubes available for the Eglin gun, a long one with a length of 601 cm and a short one with a length of 297 cm, both with a diameter of 7.62 cm. Simulations were run with helium charge pressures ranging from 1200 to 2700 psi for both the long and short pump tubes. The results indicated better gun performance for the short pump tube (Figures 18,19). Based on these results a locally[i] optimal pump tube length was sought. This was accomplished by considering pump tube lengths ranging from 207 cm to 297 cm, with an optimal of 267 cm found. The performance of the 267 cm pump tube surpassed that of the two longer pump tubes (Figures 20,21) and was therefore chosen for comparing the two cycle types. Table 6 and Table 7 shows the results of the pump tube length comparison.

A consequence of the small volume pump tube used for the WGC is that the launch tube length can be increased while total gun length is conserved. This is a convenient result since the total gun length is limited by the size of the gun room in which the Eglin launcher is housed. This benefit was exploited in the parametric study where the total gun length is held constant at 1051 cm. Using the 267 cm pump tube various simulations were run to duplicate the range of velocities predicted for the isentropic case, while attempting to minimize acceleration. Adjustments to the propellant charge, helium charge pressure and model shot start pressure were used to achieve the desired velocity range. Table 8 below shows the results for the specified shot parameters.

Table 6: Wave Gun Cycle results

Propellant charge (lbm)	*Helium Charge pressure (psi)*	*Shot start pressure (ksi)*	*Maximum gun pressure (psi)*	*Muzzle velocity (fps)*	*Maximum projectile acceleration (G's)*	*Cycle type*
1.54320	3600	8	11565	4041	17010	1-2
2.00616	3600	12	14182	4379	21276	2-3
2.46912	3600	12	17932	4720	30536	1-2
2.46912	2700	20	23095	5123	44800	2-3
2.77776	2700	20	27160	5381	53303	2-3
3.08460	2400	20	33995	5591	62214	2-3
3.08460	2700	20	32596	5619	63972	2-3
3.70368	3600	24	41306	6077	81035	2-3

[i] Previous work has indicated that a global parameter optimization is extremely difficult due to numerous local minima associated with the complex interactions of the various shot parameters (see Reference 12).

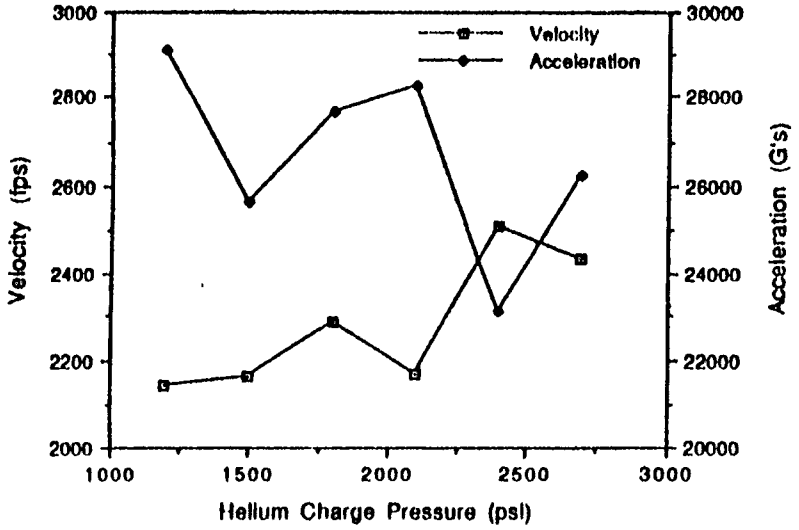


Figure 18: 601 cm Pump Tube.
8 ksi Shot Start Pressure.

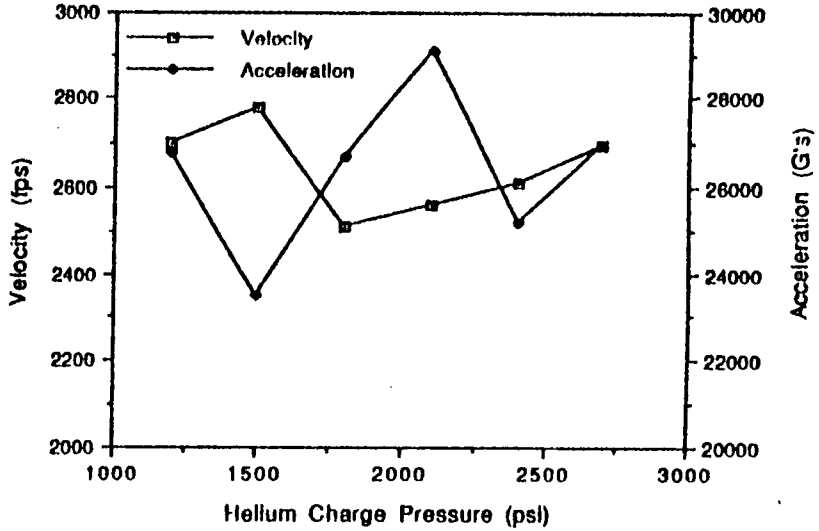


Figure 19: 297 cm Pump Tube.
8 ksi Shot Start Pressure.

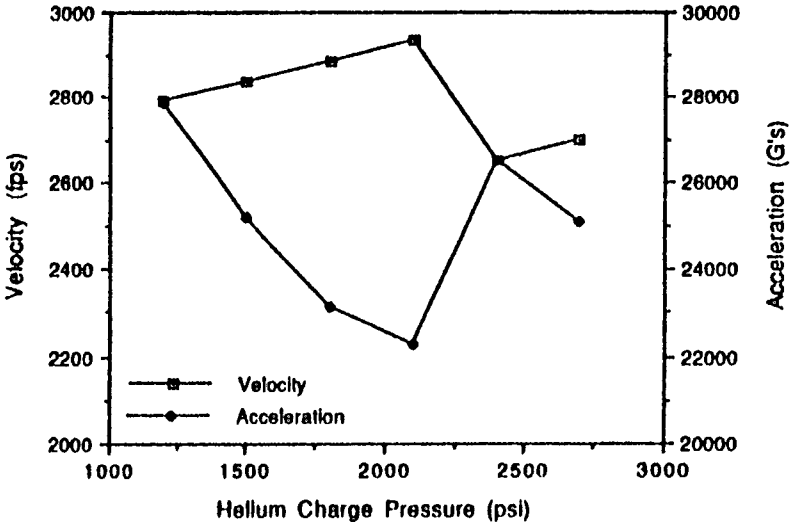


Figure 20: 267 cm Pump Tube.
8 ksi Shot Start Pressure.

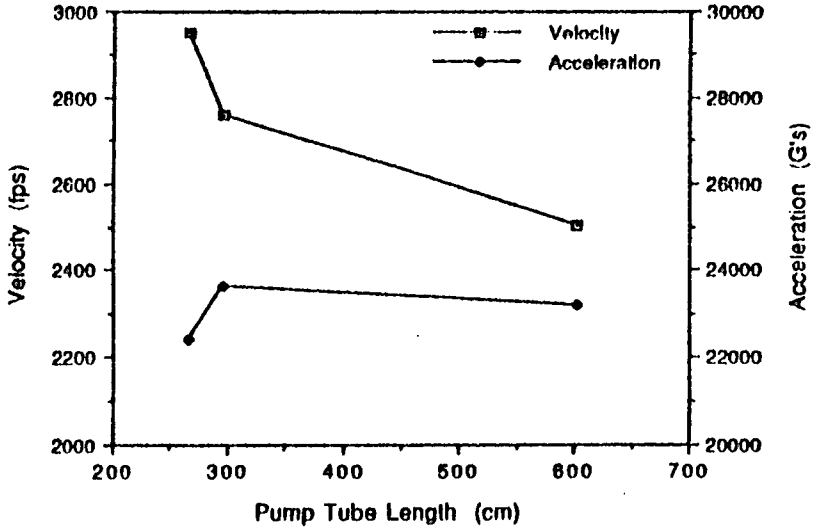


Figure 21: Pump Tube Length Comparison

Table 7: Pump tube length comparison - 8 ksi shot start pressure

Shot Parameters: Propellant load - 1.54320, Shot start pressure - 8 ksi, Piston start pressure - 1 ksi			
Pump tube length (cm)	*Helium charge pressure (psi)*	*Muzzle velocity (fps)*	*Maximum acceleration (G's)*
601	1200	2135	29141
	1500	2160	25444
	1800	2278	27655
	2100	2171	28339
	2400	2519	23198
	2700	2449	26235
297	1200	2698	26780
	1500	2776	23630
	1800	2515	26650
	2100	2563	29131
	2400	2608	25154
	2700	2692	26924
267	1200	2792	27844
	1500	2845	25156
	1800	2894	23127
	2100	2956	22419
	2400	2685	26453
	2700	2709	25148

Table 8: Pump tube length comparison - 12 ksi shot start pressure

Shot Parameters: Propellant load - 1.54320, Shot start pressure - 12 ksi, Piston start pressure - 1 ksi			
Pump tube length (cm)	*Helium charge pressure (psi)*	*Muzzle velocity (fps)*	*Maximum acceleration (G's)*
297	1200	3020	33694
	1500	2993	31802
	1800	2697	26837
	2100	2899	27873
	2400	2689	27424
	2700	2751	26863
267	1200	3189	38397
	1500	3049	30246
	1800	2796	27952
	2100	2957	27738
	2400	3103	28694
	2700	2843	26059

In comparing these results with those of the isentropic compression cycle (Table 5) it is seen that for comparable velocities the maximum acceleration is significantly lower for the WGC. It is worth mentioning again that these results are for the case where total gun length is conserved. For the 267 cm pump tube this translates to a launch tube length about twice that used for the isentropic case. Lest the reader think that the increased launch tube length produced obvious results, it is worth noting that if the Eglin Gun is operated in the isentropic compression mode using both the long launch tube and 601 cm pump tube the results are comparable to those of the WGC. In other words to achieve the same performance level as the WGC a gun operating on the isentropic compression cycle would have to be over ten feet longer than the gun currently used in Eglin's Aeroballistic Range Facility.

5. Conclusions and Recommendations

5.1 Conclusions

The interior ballistics simulation has been validated to the point that it is producing meaningful results. Furthermore the results are comparable to those in other numerical studies such as those found in References 4 and 5. It is felt that further improvements may be made by adjusting the propellant burn modeling to make up for the deficiency in propellant gas evolution. Currently the burn rate into the grain surface, which is determined using de Saint Robert's power law equation, is assumed uniform on the entire grain surface. This may be a poor assumption when the grain becomes fragmented and should be investigated further. In addition a phenomenon which is known to be occurring in the Astron Wave Gun is the flow and subsequent rapid cooling of gases in the annular regions between the sleeves and the outer pressure vessel. This effect, which is a consequence of the high ratio of surface area to volume in the annulus, was verified by Astron Research and Engineering using closed bomb tests. This work is detailed in Reference 11 where it is suggested that this effect could be modeled as a virtual leak.

Some promising results have been found for the application of the WGC to reduce model loading. These results indicate that it would be necessary to fabricate additional parts for the WGC to be applied effectively to the Eglin LGG model launcher. Obviously it would be preferable to identify a set of shot parameters which would improve performance and require no gun modifications. Based on the present results this scenario seems unlikely. However, the possibility cannot be ruled out completely without a more thorough parametric investigation. It is unfortunate that a definitive answer regarding the use of the Eglin Launcher in the WGC mode has not been forthcoming from the present study. However, though it now appears that the WGC does not inherently produce lower model loading, it is also evident that some gains can be made with appropriate launcher geometries. The present investigators intend to continue the parametric study of this cycle to extend its envelope of applicability for the type of testing that is performed in the Eglin aeroballistic facility.

5.2 Recommendations

The results of the simulations indicate that significant improvement in launcher performance should be attainable utilizing the WGC. It would be advantageous at this point to initiate a limited experimental study using Eglin's current light gas gun. This would be useful in locating and correcting inadequacies in the simulation which are certain to occur when a new gun system is modeled. It is not feasible to modify the Eglin gun to accommodate pressure-sensing instrumentation. However, the authors are aware of two pieces of equipment that the Eglin facility currently has at its disposal which would be suitable for inbore measurements requiring no modifications to the gun. These are the doppler radar which was used in the present experimental program and a VISAR laser interferometer. Data from such tests would provide support for an extensive parametric study involving pump tube volume, helium charge pressure, propellant type and loading, and piston and model start pressures. In this study emphasis should be placed on finding a suitable configuration which would require the least number of modifications to the present gun system.

6. Nomenclature

A	Barrel or launch tube cross-sectional area
$A(x)$	Cross-sectional area of gun at x
a_o	Initial sound speed in driver gas
C_o	Constant used to specify thickness of shock region
E	Specific internal energy
j	Mass point number
L	Barrel or launch tube length
M	Mass
n	cycle number
P	pressure
p	pressure in driver gas
p_{max}	Maximum base pressure
p_o	Initial pressure in driver gas
p_p	Projectile base pressure
$\overline{p}$	Average base pressure
q	Artificial viscosity
t	Time
u	Velocity of driver gas
u_m	Muzzle velocity
u_p	Projectile velocity
V	specific volume
x	Spatial coordinate
x_p	Distance projectile has traveled down barrel
η	Piezometric efficiency
$\rho,\rho(x)$	Density
γ	Ratio of specific heats

7. References

1. Seigel, A. E., *Theory of High-Muzzle-Velocity Guns*, Progress in Astronautics and Aeronautics, Volume 68, pp. 143-144, October, 1978.

2. Charters, A.C., Denardo, P.b., and Rossow, V.J., *Development of a Piston-Compressor Type Light-Gas Gun for the Launching of Free-Flight Models as High-Velocity*, NACA Technical Note No. 4143, November, 1957.

3. Crozier, W.D., and Hume, W., *High-Velocity Light Gas Gun*, Journal of Applied Physics, Volume 28, No. 8, pp. 892-984, 1957.

4. Dahm, T.J., and Randall, D.S., *The Wave Gun Concept for a Hyperveloctiy Rapid-Fire Weapon*, Astron Research and Engineering Mountain-View, California, January, 1984.

5. Groth, C.P.T. and Gottlieb, J.J., *Numerical Study of a Two-Stage Light-Gas Hypervelocity Projectile Launchers*, UTIAS Report No. 327, CN ISSN 0082-5255, October, 1988.

6. Dahm, T.J. and Watson, J.D., *Analysis and Design of a Two-Stage Hybrid Launcher*, Final Report, contract No. DNA 001-76-C-0407, July, 1977.

7. Piacesi, R., Gates, D.F. and Seigel, A.E., *Computer Analysis of Two-Stage Hypervelocity Model Launchers*, NOLTR 62-87, Naval Ordinance Laboratory, White Oak Maryland, August, 1963.

8. Cable, A.J. and DeWitt, J.R., *Optimizing and Scaling of Hypervelocity Launchers and Comparison with Measured Data*, , Arnold Engineering Development Center, April, 1967.

9. Von Neumann, J. and Richtmyer, R.D., *A Method for the Numerical Calculation of Hydrodynamic Shocks*, Journal of Applied Physics, Vol. 21, March, 1950.

10. Courter R.W. and Hugenroth J.J., *A Research Plan for Evaluating Wave Gun as a Low-Loading Model Launcher For High Speed Aeroballistics Tests*, Final Report RDL/AFOSR, Summer Research Program, Eglin A.F.B., Florida, August, 1994.

11. Randall, D.S., *Wave Gun Development During the Past Year*, A presentation to the Aeroballistic Range Association, Quebec, P. Q., Canada, September 9-12, 1986.

12. Lorin, Jean-Yves S., *Optimization of the Firing-Cycle of a Light-Piston Light-Gas Gun*, A Thesis, Louisiana State University, December 1992.

CHARACTERIZATION OF ELECTRO-OPTIC POLYMERS

Vincent G. Dominic
Assistant Professor
Center for Electro-Optics

University of Dayton
300 College Park Ave.
Dayton, Ohio 45469-0245

Final Report for:
1995 Summer Research Extension Program
Wright Laboratories
Wright-Patterson Air Force Base

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Wright Laboratory

December 1995

CHARACTERIZATION OF ELECTRO-OPTIC POLYMERS

Vincent G. Dominic
Assistant Professor
Center for Electro-Optics
University of Dayton

Abstract

We present a simple experimental procedure that uses a slowly-rotating Fabry-Perot étalon to measure simultaneously the electro-refraction and electro-absorption in a poled polymer. Both effects generally contribute to the electro-optic signal from such material systems and can be distinguished by rotating the sample and observing asymmetric peaks in the signal. The experimental results show the expected increase in both electro-refraction and electro-absorption as the probe wavelength approaches the absorption band of the chromophore. Furthermore, the dispersion of the complex electro-optic coefficient displays a periodic variation that we attribute to multiple-étalon interference. The stratified nature of the thin-film structure causes the multiple-reflection interference. This artifact will pollute most of the standard electro-optic characterization techniques for poled-polymer films.

We also report on initial studies of in-plane poling experiments for making an electro-optic probe that will sense ultrafast electrical signals on circuit boards. Our system is different than previous probes because the electro-optic sensing material will be a poled polymer attached to the end of a fiber. Such a probe can be placed to investigate any spot on a circuit board and allows multiple probes simply by using multiple fibers. We have found that in-plane poling of polymers is nontrivial because of breakdown problems at the polymer/air interface.

CHARACTERIZATION OF ELECTRO-OPTIC POLYMERS

Vincent G. Dominic

Introduction

This effort concentrated on two different aspects of electro-optic (*EO*) thin-film poled polymers. The majority of the effort reported here concerns a novel twist on the Fabry-Perot characterization technique for poled polymer films. We found that by utilizing a reasonably "thick" (1 mm) étalon containing both the active polymer layer as well as the glass substrate we could determine both the electric-field-induced change in the refractive index (*electro-refraction* $\rightarrow$ *ER*) as well as the change in the absorption coefficient (*electro-absorption* $\rightarrow$ *EA*) of the polymer. We need only rotate the sample $\pm 2.5°$ from normal to view five étalon resonances. The *ER* signal component switches sign on either side of a resonance so that if only *ER* is present one expects equal positive and negative excursions on the *EO* signal. The vertical asymmetry that we often observe is therefore indicative of either electro-absorption or a measurement artifact - multiple étalon interference. Our technique is wonderfully simple and helps distinguish the generally unwanted absorptive behavior near a resonance from the desired refractive response. Part I describes this work.

We also pursued a device application aimed at utilizing the favorable linear and nonlinear properties of thin-film polymer systems. Our device goal was a moveable electro-optic probe capable of ultrafast electrical detection. To accomplish this goal studied in-plane poling of thin layers of polymer films. In-plane poling means that both electrodes are coplanar so that the polar axis of the poled polymer lies in the plane of the layer. This proved more difficult than expected because of electrical breakdown in the substrate and in the air. We recently overcame this problem and are progressing towards our goal of the *EO* probe. Part II describes this work

Part I: Complete Electro-Optic Characterization of Thin-Film Poled Polymers

While investigating Mach-Zehnder and Michelson[1,2] methods of measuring the electro-optic properties of thin-film poled polymers we noticed disturbingly erratic behavior of electro-optic coefficients (r_{ij}). In particular we noticed a non-physical variation of the *EO* coefficient as we translated the probe beam location within the poled region. The test structure that we used in these measurements is shown in Fig. 1. The poling electrodes are the indium tin oxide (*ITO*) layer and the gold electrode/mirror.

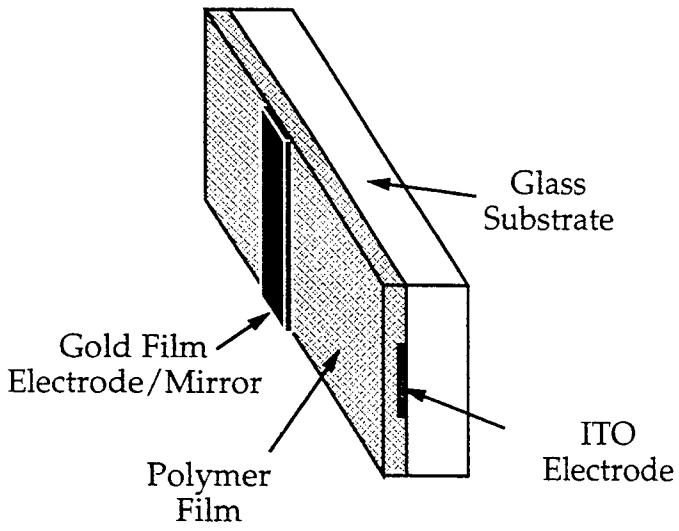


Figure 1 Schematic view of the poled-polymer thin-film test structure. The poled region of the polymer lies between the intersection of the *ITO* and gold electrodes. Typical thickness dimensions are (substrate ~1 mm, *ITO* ~60 nm , the polymer ~2 μm, and the gold ~35 - 50 nm).

The polymer that lies near the center of the region between the *ITO* electrode and the gold electrode is assumed to be uniformly poled. However, in a Michelson interferometric geometry we observed the spatial dependence shown in Fig. 2 for the electro-optic signal at normal incidence. Since the polymer was spin coated onto the glass slide, the thickness of the layer is certainly nonuniform. The data of Fig. 2 indicates that a measurement artifact associated with interference is polluting the Michelson measurement of the *EO* signal. During investigation of this artifact we included a motorized rotation stage to accurately position the sample at normal incidence. As the sample rotated we noticed Fabry-Perot resonances in the transmitted light signal and asymmetric resonance behavior in the *EO* signal. This observation led to a detailed investigation of electro-refraction, electro-absorption, and multiple-étalon interference.

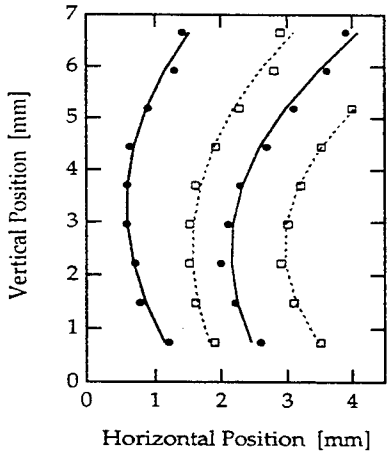


Figure 2 Electro-optic modulation signal vs. position across the poled region of the Dow polymer sample labeled TP86: minimum signals (●), maximum signals(□). The test structure was translated in the

transverse plane as the positions of the maximum and minimum signals were recorded.

Many techniques for characterizing electro-optic polymers have been developed. The measurement methods include: Michelson and Mach-Zehnder interferometric techniques,[1,2] attenuated total reflection techniques,[3,4] ellipsometric/polarimetry techniques,[5-10] and Fabry-Perot étalon modulation schemes.[11-16] We developed a method that is a variation of previous Fabry-Perot characterization methods. The primary difference is that we use a much thicker étalon (millimeters instead of microns) whose thickness includes the glass substrate of the sample. The primary advantage of the thicker étalon is that it displays multiple resonance peaks for small rotation angles. We typically observe transmissivity curves with five resonance peaks within a 5° rotation (±2.5° from normal). The multiple resonances allow accurate determination of the étalon parameters. Another significant advantage is that by rotating through several resonance peaks we can discern whether multiple effects contribute to the modulation signal. Typically, this is evidenced by a vertically offset modulation signal. Additional advantages of our method are a simple experimental setup, results interpretation is straightforward, and the sample structure is consistent with that used by many other measurement techniques. For example, the sample is readily switched into a Mach-Zehnder or ellipsometric/reflection setup to verify results.

Generally, electro-optic devices require a relatively large, phase-only modulation. This accounts for the emphasis that is placed on the field induced changes in the refractive index, Δn, denoted herein as electro-refraction (*ER*). Relatively few papers discuss the concurrent field-induced change in absorption, $\Delta\alpha$, denoted electro-absorption (*EA*).[7-9,17-22] The composite electro-optic (*EO*) effect results from both phase and amplitude modulations of the probe light beam. With a complex electro-optic coefficient the real part governs the phase modulations and the imaginary part describes the amplitude modulations. Decomposition into phase and amplitude effects becomes particularly important when the probe wavelength approaches the chromophore absorption band.[9,23] Unfortunately, the resonant enhancement of electro-refraction is accompanied by increased electro-absorption. If neglected, the interaction of the competing effects may lead to significant over- or under-estimation of the electro-optic coefficient. We also found that unwanted surface reflections give signal asymmetries that can be incorrectly interpreted as electro-absorption. These multiple surface reflections are characterized with a simple model showing how this artifact imposes the appearance of electro-absorption onto a purely electro-refractive signal. By measuring the wavelength dispersion of the complex electro-optic coefficient, the spurious multiple-reflection artifact (which oscillates with wavelength) can be distinguished from the underlying electro-refraction and electro-absorption.

Sample description and experimental setup

Figure 3 shows the sample geometry and the experimental arrangement that we used for the electro-optic characterization experiments. The substrate is a ≈1 mm thick glass slide coated with the transparent conductor indium tin oxide (*ITO*). The *ITO* is etched so that a narrow stripe extends the length of the slide. This stripe acts as one electrical contact. The polymer is then spin-coated on top of the *ITO* with a thickness of approximately 2 μm. Once the polymer dries, a thin stripe of gold is evaporated on top, perpendicular to the *ITO* stripe. This gold strip serves as the second electrical contact for the sample. This structure sandwiches a ≈25 mm^2 rectangular region of the polymer between the *ITO* and gold contacts. This region of the sample is poled by first heating the polymer to its glass transition temperature and then applying a strong dc electric field (≈100 V/μm). The torque on the dipolar chromophore molecules causes a macroscopic polar alignment within the polymer. After decreasing the temperature, the cooled polymer semi-permanently maintains the non-centrosymmetric alignment of the chromophores within the electro-optically active poled region.

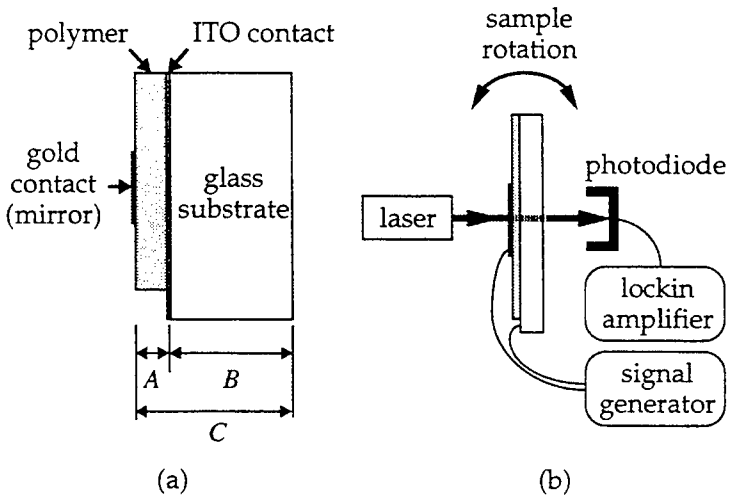


Figure 3 Schematic view of the generic poled-polymer sample geometry (not to scale) and the experimental arrangement. (a) Sample description: The sample consists of the following layers: gold electrode/mirror, polymer layer, *ITO* contact, and the glass substrate. The substrate is typically ≈1 mm thick, the *ITO* ≈60 nm thick, the polymer ≈2 μm thick, and the gold ≈35–50 nm thick. (b) A laser beam transmitted through the poled region of the Fabry-Perot structure is collected by a photodiode. The sample is slowly rotated through ±2.5° as the lockin measures the angular dependence of both the average photodiode signal and the modulation signal (light signal variation imposed by the signal generator). A computer (not shown) controls the sample rotation and acquires the data.

After poling, the sample is attached to a motorized rotation stage and slowly rotated ±2.5° while the transmitted light beam is monitored, as shown schematically in Fig. 3b. The full angular dependence of both the transmittance and the field-induced transmittance modulation is measured. Both Δn and $\Delta\alpha$ of the poled polymer are measured in this simple way. References 7-9 and 21-22 describe techniques that also measure the real and imaginary components of the complex electro-optic coefficient, but not with the simplicity of the present method. Recently, Ziari *et al.* suggested an elegant method based on Young's double-slit experiment that also measures both the phase and

amplitude modulation effects in poled polymers.[17] Their technique requires coplanar-poled samples.

We direct a laser beam through the poled region of the sample and collect the transmitted light with a photodiode. The gold and *ITO* serve as contacts for applying the modulation voltage. The gold and glass/air interface provide mirrors for the étalon. A sinusoidal voltage (±16 V, 5 kHz) applied across the ≈2 μm thick polymer layer induces phase and amplitude changes in the sample. A Stanford Research SR530 lockin amplifier measures both the average and the time-varying components of the transmitted light signal while a controlled actuator slowly rotates the sample through ±2.5° (0° = normal incidence). We are careful to insure that the rotational axis is centered on the incidence spot to prevent translation of the beam across the poled region as the sample rotates. This is important because the spin-coating process results in a polymer layer with varying thickness (see Fig. 2). Even sub-wavelength thickness variations can strongly affect the Fabry-Perot's throughput. The input beam is vertically polarized so that the $\tilde{r}_{13} = r_{13} + i\, s_{13}$ component of the complex electro-optic coefficient is measured (using the notation of ref. 7). The incidence angle is small enough so that the contribution of the $\tilde{r}_{33}$ component can be ignored even if we utilize *p*–polarized light.

Single-étalon, *ER/EA*-interference model

Let us, for the moment, ignore some of the reflecting surfaces in our sample and model the sandwich structure with only two reflections: the gold film and the air/substrate interface. This corresponds to étalon *C* in Fig. 3a. The étalon thickness includes the electro-optically active polymer layer, the *ITO*, and the glass substrate. The reflectivities and transmisivities for the three reflecting surfaces (1 = gold, 2 = polymer/*ITO*/glass interface, 3 = air/glass interface) are labeled r_j, t_j (j = 1,2,3), and we assume here that $r_2 = 0$. The refractive indices of the glass and polymer are denoted as n_g and n_p, respectively. The *s*–polarized probe beam selects the ordinary refractive index of the polymer $n_p = n_o$. The layer thicknesses are L_g and L_p and the glass substrate is assumed to have no absorption at any of the wavelengths of interest. The *ITO* is very thin compared to either the polymer or the glass layer and is therefore ignored in this simple model. The intensity transmission of this simplified Fabry-Perot étalon is:

$$I_{trans} = \frac{I_{inc}\, e^{-\alpha_p L_p} \left| t_1\, t_3 \right|^2}{\text{denom}} \tag{1a}$$

where:

$$\text{denom} = \left(1 + r_1\, r_3\, e^{-\alpha_p L_p}\right)^2 - 4\, e^{-\alpha_p L_p}\, r_1\, r_3 \sin^2(\delta), \tag{1b}$$

α_p is the absorption in the polymer and:

$$\delta = \frac{2\pi}{\lambda}\left\{ L_g \sqrt{n_g^2 - \sin^2\Theta} + L_p \sqrt{n_p^2 - \sin^2\Theta} \right\} \tag{2}$$

with the wavelength λ and the external angle of incidence Θ

The applied modulation voltage induces a change in both n_p and α_p. In the case were the absorption is reasonably small such that $\alpha_p \lambda << 1$, Clays and Schildkraut[7] showed that the field-induced perturbations Δn_p and $\Delta \alpha_p$ may be written in terms of the real and imaginary parts of the complex electro-optic coefficient $\tilde{r}_{13}$:

$$\Delta n_p = -\frac{1}{2} n_p^3 \, r_{13} \, E_{applied} \tag{3a}$$

and:

$$\Delta \alpha_p = -\frac{2\pi}{\lambda} n_p^3 \, s_{13} \, E_{applied} \tag{3b}$$

where $E_{applied}$ is the applied modulation field:

$$E_{applied} = \frac{V_{applied}}{L_p} \cos(2\pi f t) \tag{4}$$

and f is the modulation frequency.

The differential change in the transmitted intensity caused by the field-induced modulation is:

$$\Delta I_{trans} = \frac{I_{trans}}{\text{denom}} \left\{ \begin{array}{l} \left(-\Delta\alpha_p \, L_p\right)\left(1 - r_1 \, r_3 \, e^{-\alpha_p L_p}\right)\left(1 + r_1 \, r_3 \, e^{-\alpha_p L_p}\right) \\ + \; 4 \, r_1 \, r_3 \, e^{-\alpha_p L_p} \; (\sin\delta) \; \dfrac{2\pi}{\lambda} \dfrac{\Delta n_p \, n_p \, L_p}{\sqrt{n_p^2 - \sin^2\Theta}} \end{array} \right\}. \tag{5}$$

The first term in braces is the electro-absorptive effect while the second is the electro-refractive. Equations 1 and 5 provide a theoretical model to fit the measured data. Notice that after Eqn. 1 is utilized to fit the transmittance signal, all the parameters except Δn_p and $\Delta \alpha_p$ are determined. Equations 3 & 5 show that both Δn_p and $\Delta \alpha_p$ contribute linearly to the *1f* lockin signal.

The linear dependence of the electro-optic perturbations is verified by varying the amplitude of the applied modulation voltage. The electro-strictive, electro-mechanical, and Kerr effects are ignored because they will appear as *2f* signals on the lockin $\left(\propto \left| E_{applied} \right|^2 \right)$. By contrast, a piezoelectric effect, ΔL_p, will contribute to the *1f* signal. Inspection of Eqn. 1 shows that the functional form of the piezoelectric effect is precisely the same as Eqn. 5 but with the Δ shifted to L_p in both terms within braces. The large numerical factor $2\pi/\lambda$ in the second term indicates that ΔL_p will give essentially the same angular dependence as Δn_p. To determine whether there is a piezoelectric contribution, a second sample made with the same polymer but with a more reflective gold layer is placed in a Michelson interferometer with the gold contact as one of the retroreflecting mirrors. With the sample in this orientation the light does not pass through the polymer and therefore electro-refraction and electro-

absorption cannot contribute to the signal. Any piezoelectric effect in the polymer will move the gold mirror and thus contribute to the lockin signal. In this experimental configuration we observed no evidence of piezoelectricity in this polymer system.

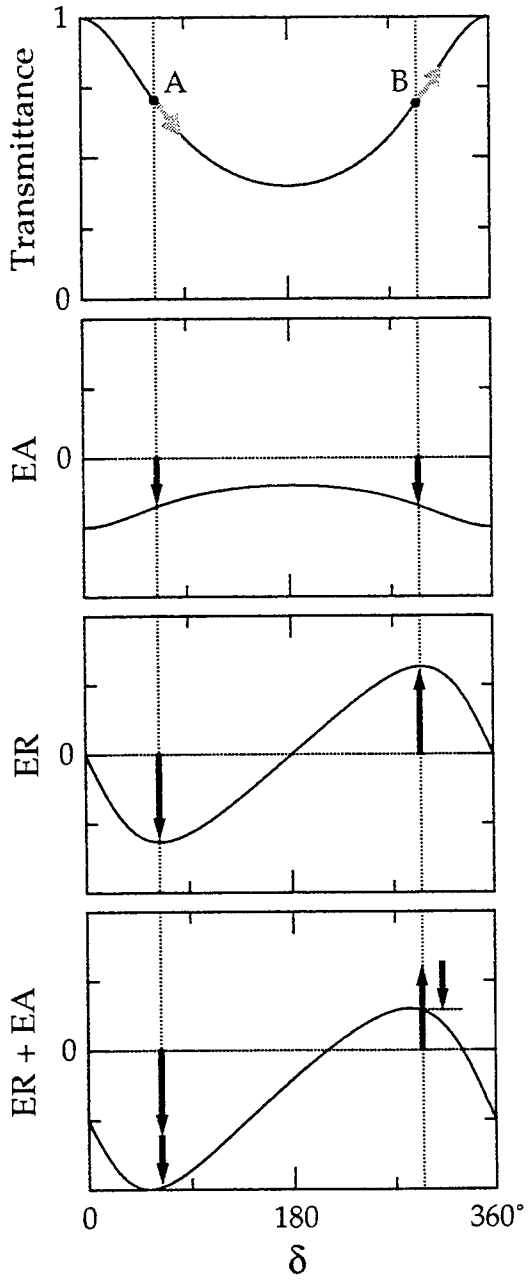


Figure 4 This figure shows how the electro-refractive (*ER*) signal and the electro-absorptive signal (*EA*) interfere to give an asymmetric response as the Fabry-Perot structure is rotated. Part (a) shows the low finesse ($\mathcal{F}$ = 1.5) transmittance variation with δ and indicates two points where the *ER* signal magnitude is maximized. The arrows indicate how a positive Δn affects the transmittance. Part (b) shows the expected *ER* response (positive Δn) while part (c) shows the *EA* response (positive $\Delta\alpha$). The total signal, *ER*+*EA*, in part (d) displays a strong vertical asymmetry.

The angular dependence of the electro-absorptive term is determined by the variation of I_{trans} whereas the electro-refractive term varies according to $I_{trans} \sin\delta$. This gives a strong asymmetry to the magnitude of the lockin signal peaks as the incidence angle changes such that δ shifts by $\approx\pi$. If Eqn. 1b is re-written as:

$$\text{denom} = 1 + \left(r_1 r_3 e^{-\alpha_p L_p}\right)^2 + 2 e^{-\alpha_p L_p} r_1 r_3 \cos\delta \qquad (6)$$

then because of the $\cos\delta$ term, the transmitted intensity is at the mid-visibility point of the Fabry-Perot resonance when δ is approximately $\pi/2$ or $3\pi/2$. At both these points, labeled *A* and *B* in Fig. 4 below, the electro-absorptive (*EA*) contribution to the lockin signal has the same magnitude and sign whereas the electro-refractive (*ER*) signal is of equal magnitude but opposite sign. This behavior is illustrated in Fig. 4. At point *A* there is constructive interference (the *ER* & *EA* signal components add) while at point *B* there is destructive interference (the *ER* & *EA* signal components subtract). Thus, when both electro-refraction and electro-absorption contribute, the electro-optic signal magnitude is unequal on either side of a Fabry-Perot resonance. Such asymmetries have been observed by others.[21,22]

Experimental results

Most of the data below is gathered using a golden yellow sample labeled TP86 supplied by the Dow Chemical Company. Figure 5 displays the angular variation of the average photodiode signal along with the magnitude of the *If* lockin signal as the poled-polymer Fabry-Perot sample is rotated. The data are represented by dots and the theoretical curve fits using Eqns. 1 and 5 are displayed as solid lines. The average signal shows the expected low-finesse Airy function behavior. According to the simple model discussed above, there is no electro-absorptive contribution to the modulation signal because the positive and negative peaks of the lockin signal are of equal magnitude. The signal variation with incidence angle is shaped as one would expect for a purely electro-refractive electro-optic effect. The advantage of our thick étalon (mm) versus the thin (μm) étalons studied previously[11-16] is the appearance of many resonances while tuning over small angles ($\Theta \leftrightarrow \pm 2.5°$). This gives ample data to precisely determine the Δn variation. The data sets are analyzed by first using Eqn. 1 to fit the average signal angular dependence. This is a three parameter fit where the incident optical power, the cavity finesse, and the thickness of the polymer are allowed to vary. The polymer thickness is restricted to vary by less than $\pm\lambda$ and the finesse does not change by more than 10% for any of our probe wavelengths. Since the air/glass interface has low reflectivity the finesse of the cavity is only ≈1.37. After fitting the average photodiode signal, all parameters are held fixed except Δn and $\Delta\alpha$ which are then fit to the variations in the lockin signal.

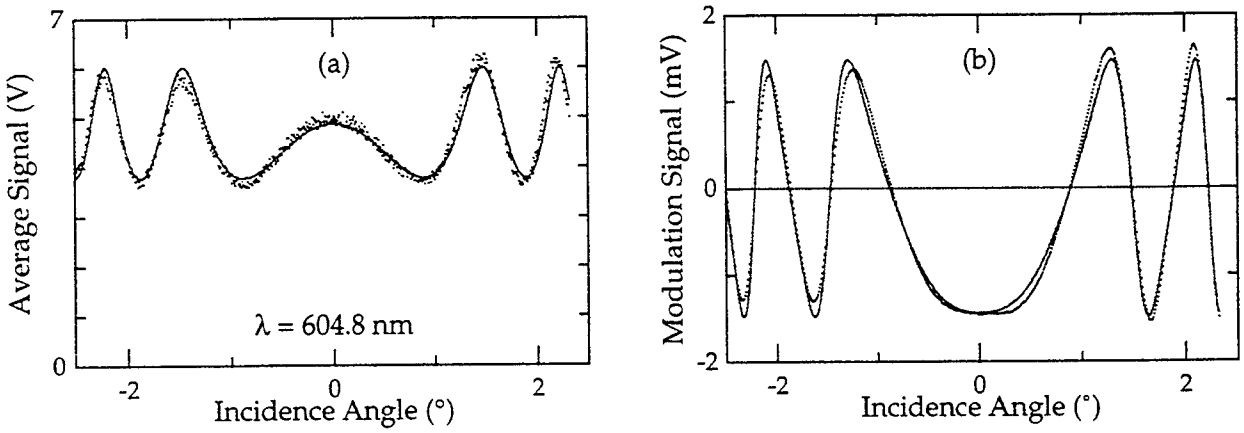


Figure 5 (a) Average photodiode signal and theoretical fit (dots=data, line=fit) as the étalon is rotated with λ = 604.8 nm. (b) Lockin signal dependence on the incidence angle with the raw data (dots) and the theoretical fit (solid line). The average signal shows the expected low-finesse étalon behavior and the lockin signal displays symmetric peaks that reach their maximum where the slope of the average signal is highest. The modulation signal has equal magnitude on either side of a Airy-function transmittance resonance.

By tuning the probe laser to λ = 594.1 nm, the appearance of the lockin signal is dramatically altered. According to the simple model, the vertical asymmetry of the modulation signal apparent in Fig. 6a is the signature of the electro-absorptive effect interfering with the electro-refractive effect. If the laser is tuned even closer to the absorption edge of the chromophore then electro-absorption completely dominates the measurement.[15] This is illustrated in Fig. 6b below where a probe laser wavelength of λ = 543.5 nm is used on the golden yellow sample whose absorption edge lies at approximately λ = 520 nm. Notice that the lockin signal never crosses zero (compare to Fig. 5b, 6a).

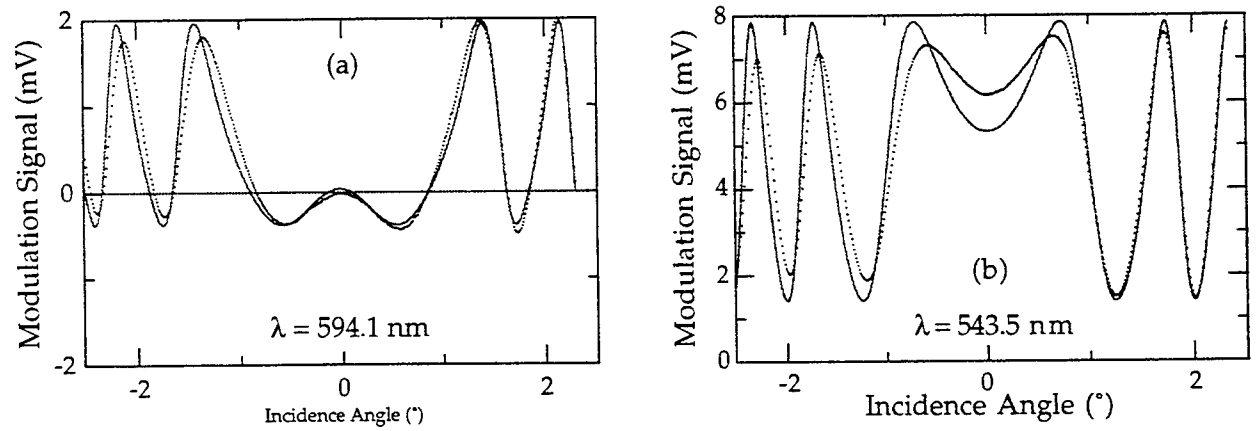


Figure 6 (a) Lockin signal variation as the étalon is rotated in a λ = 594.1 nm probe beam. The peaks of the lockin signal are quite asymmetric: the positive excursion is about four times larger than the negative excursion. (b) Lockin signal dependence on the incidence angle at λ = 543.5 nm. In this case the signal never crosses zero and electro-absorption completely dominates the measurement. In both (a) and (b) the dots are raw data and the solid curve is the theoretical fit using the single étalon, *ER/EA*-interference model. The average signals incident on the photodiode displayed low-finesse étalon behavior much like Fig. 5a.

A tunable HeNe laser was then used to roughly determine the dispersion of the complex electro-optic coefficient. At a probe wavelength of $\lambda = 543.5$ nm we measured $\tilde{r}_{13} \approx (5 + i\,1.25)$ pm/V. Both the electro-refractive (*ER*) and electro-absorptive (*EA*) contributions to the signal increased dramatically as the probe wavelength approached the edge of the chromophore absorption band. This behavior is generally expected (see, for example, the two-level models developed in refs. 9 and 23). Curiously, the measured *EA* coefficient switched sign at a probe wavelength of $\lambda = 612$ nm. This is difficult to explain since the absorption itself monotonically decreases for wavelengths longer than $\lambda = 520$ nm (see Fig. 7). In fact, further investigations revealed that the *EA* coefficient magnitude varied and switched sign as the probe beam was moved to different spots within the poled region. This erratic behavior may be caused by the nonuniform thickness of the film created during the spin-coating process, as illustrated in Fig. 2 above. Since the electro-optic coefficient represents a fundamental property of the material, it should be independent of the polymer layer thickness, as was nicely demonstrated in ref. 14. Furthermore, the actual electro-absorption for $\lambda > 590$ nm is very small. Therefore, we don't expect significant electro-absorption at these wavelengths. However, the modulation signal data at both $\lambda =$ 612 and 632.8 nm displayed a strong vertical asymmetry. We show below that the unwanted reflectivity from the polymer/*ITO*/glass interface leads to vertically offset lockin signals that can be incorrectly interpreted as the *EA* effect. Further evidence for this unwanted reflectivity is shown in the measured absorbance spectra of the TP86 polymer system using a Perkin-Elmer UV/VIS spectrophotometer. Figure 7 below shows clear Fabry-Perot étalon resonances throughout the spectrum. The periodicity of the resonances predicts well the polymer layer thickness. Thus, although the $\lambda =$ 543.5 nm data in Fig. 6b is clearly indicative of electro-absorption, the longer wavelength data is polluted by multiple reflections that give the appearance of electro-absorption.

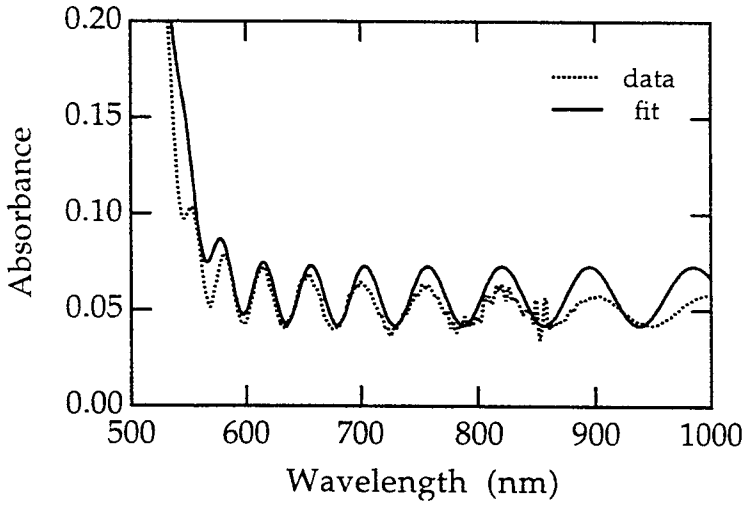


Figure 7 Fabry-Perot resonance oscillations in the absorption spectra of the TP86 polymer sample. The solid line is a fit based on an étalon filled with material of n=1.7 and ≈2.9 μm. thick. This corresponds to the distance between the gold mirror and the glass/*ITO*/polymer interface.

To better understand the multiple étalon problem, the electro-optic dispersion of the TP86 sample was re-measured with a dye laser. Pumping a rhodamine-6G dye with an argon laser allows tuning of the probe wavelength over a range from λ = 565–627.5 nm. Figure 8 shows the wavelength dependence of the real and imaginary parts of the complex electro-optic coefficient $\tilde{r}_{13}$ determined at 2.5 nm intervals. Notice that the imaginary part (*EA*-component) shows a clear periodic variation with a period of ≈30 nm. This variation is unphysical based on the monotonic decrease of the absorption over this wavelength range. The oscillation is caused by the additional étalons (*A* & *B* in Fig. 3a) created by the unwanted polymer/*ITO*/glass reflection. Also note that the sinusoidal variation of both the *ER* and *EA* components are superimposed on a curve that decreases at longer wavelengths. The true absorption spectra of this sample (underneath the étalon resonances shown in Fig. 7) monotonically decreases for wavelengths $\lambda > 520$ nm. The oscillation period in the imaginary part of the electro-optic coefficient corresponds to an étalon with the thickness and refractive index of the TP86 polymer (*A* in Fig. 3a):

$$\Delta\lambda = \lambda^2/(2\, n_p\, L_p) = 36 \text{ nm} \tag{7}$$

where $L_p \approx 2.9$ μm and $n_p \approx 1.7$. The polymer thickness was determined by fitting an Airy function to the weak, periodic oscillation observed in the sample absorption spectra. The periodic variation in the absorbance is also caused by the thin *A* étalon. Figure 9 shows a series of measurements in the dispersion determination; each graph displays both the average photodiode signal variation as well as the electro-optic modulation signal for various dye laser probe wavelengths. Notice that the functional fits are uniformly good across the entire data set.

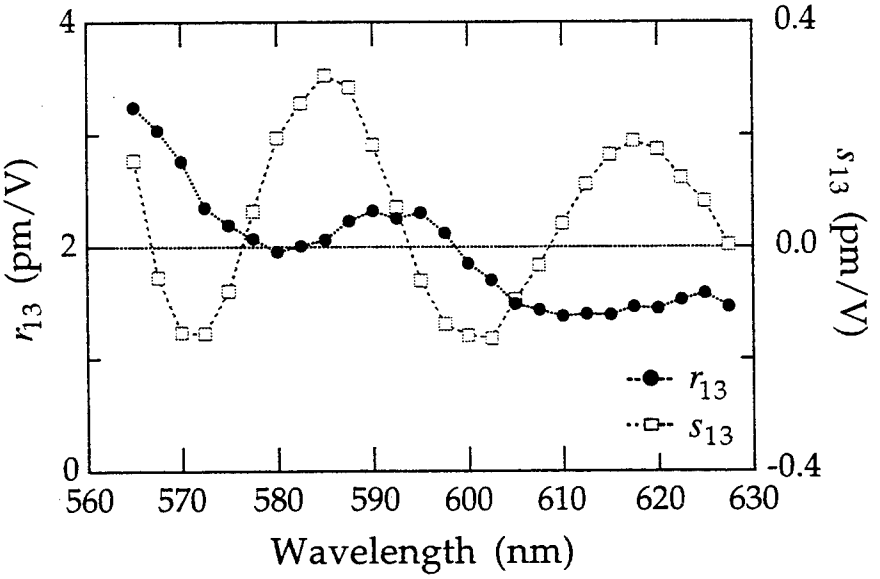


Figure 8 Dispersion of the real r_{13} and imaginary s_{13} components of the complex electro-optic coefficient for TP86. The oscillation in both components is caused by a multiple étalon artifact. The increase of the electro-refractive component r_{13} at shorter wavelengths suggests the true nature of the dispersion.

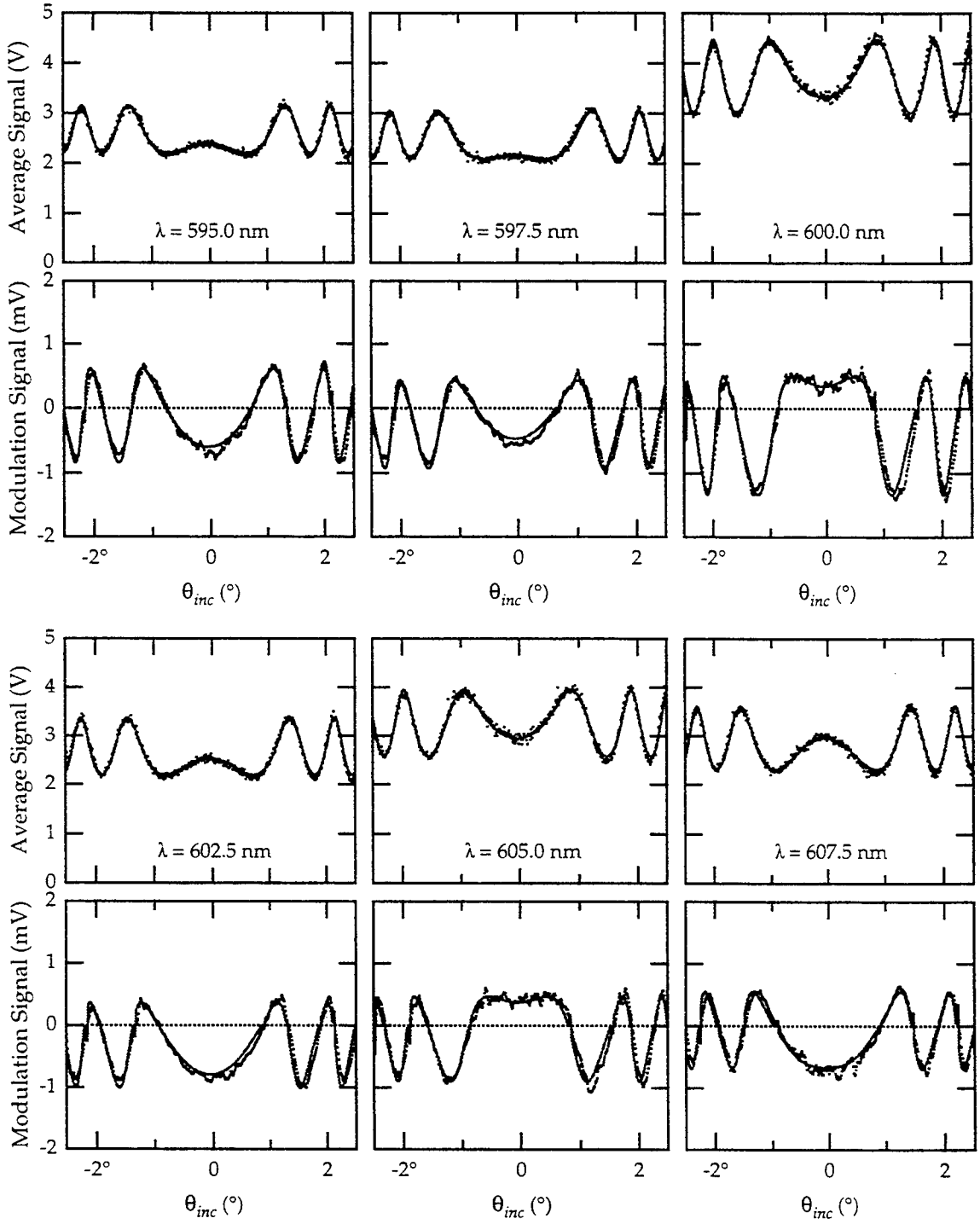


Figure 9 Angular variation of the average photodiode and electro-optic modulation signals from the Fabry-Perot electro-optic measurement technique (λ = 595.0 nm to 607.5 nm in 2.5 nm increments). The experimental data are represented by dots and theoretical curve fits by lines. The theoretical fits use the single étalon, *ER/EA*-interference model. Note that the complete data set extended from λ = 565.0 nm to 627.5 nm in 2.5 nm increments.

Multiple étalon, no-electro-absorption model

The previously discussed model ignored the unwanted reflectivity of the polymer/*ITO*/substrate interface. This reflectivity gives rise to the unphysical oscillatory behavior of the *EA* coefficient s_{13} in Fig. 8. The discussion above focused on Étalon *C*. The thickness of étalon *A* is approximately equal to the polymer thickness and the transmissivity of this étalon will remain essentially constant over the narrow angular range probed in our experiments. Therefore, the electro-refractive signal from this étalon by itself remains constant as the sample rotates ±2.5°. By contrast, the electro-refractive signal from étalon *C* contains several resonant peaks as the incidence angle varies. The electro-refractive signal (*C*) switches sign on either side of a resonance and therefore the signal from étalons *A* and *C* will add constructively/destructively on either side of a resonance. Notice that étalon *B* does not contribute to the modulated signal since the spacer layer (glass) is inactive. The interference of the two electro-refractive signal components (étalons *A* and *C*) gives asymmetric peaks to the lockin signal because of the essentially constant offset contributed by étalon *A*.

Thus, the multiple-étalon interference can give a signal that appears to be electro-absorption but in fact is not. This point has been carefully explored by the authors of refs. 8 & 9 in the context of the ellipsometry/reflection geometry frequently used to measure the electro-optic properties of poled polymers. This effect is probably also the source of the asymmetric electro-optic signals observed in the Mach-Zehnder experiment of Norwood *et al.*[2] Vertically asymmetric peaks in the modulation signal, on either side of an étalon resonance, are also apparent in the data of C. H. Wang *et al.*[15] These authors utilize a Fabry-Perot at normal incidence and scan the wavelength (rather than the incidence angle) to study the étalon resonance behavior.

The importance of this multiple reflection effect must be emphasized: most of the common electro-optic characterization techniques (Fabry-Perot, ellipsometry/reflection, Mach-Zehnder) for polymer thin films are susceptible to pollution when multiple reflections are present. The popular method of Teng & Man[5] is certainly affected (ref. 7-9) and one should be careful in interpreting the coefficients determined without accounting for this spurious effect. With respect to the dispersion measurements shown in Fig. 8, near the absorption band (λ = 543.5 nm, Fib. 6b) the signal is clearly dominated by electro-absorption. The unwanted surface reflection seems to add a periodic oscillation to s_{13} with a magnitude of ≈0.2 pm/V. Thus, the *EA* coefficient is considered unpolluted when $s_{13} >> 0.2$ pm/V. We conclude that the data shown in Fig. 8 contains both the artificial oscillation from multiple-étalon interference and a real electro-absorptive effect that increases at shorter wavelengths.

We extend the single étalon results to model the coupled Fabry-Perot cavity interference using the

method described in ref. 24. First the transmission and reflection coefficients for the last two layers of the sample are determined as if they were the only layers present. The reflectivity of the j^{th} layer (bounded by reflecting surfaces r_j and r_{j+1}) is:

$$r_{j,j+1} = \frac{r_j + r_{j+1} e^{i\delta_j}}{1 + r_j r_{j+1} e^{i\delta_j}} \tag{8}$$

where:

$$\delta_j = \frac{4\pi}{\lambda} n_j L_j \cos\phi_j \tag{9}$$

and n_j, L_j, & ϕ_j are the refractive index, thickness, and propagation angle in the j^{th} layer. The transmissivity of this layer is:

$$t_{j,j+1} = \frac{t_j t_{j+1} e^{i\delta_j/2}}{1 + r_j r_{j+1} e^{i\delta_j}}. \tag{10}$$

Once the field reflectivity and transmissivity are found for the j^{th} layer, this layer is replaced by an effective surface having the properties dictated by Eqns. 8–10. This process is iterated until all layers are included.[24]

In this model, the thickness of the *ITO* layer is ignored but its reflectivity is included to give three reflecting surfaces (see Fig. 3a). The polymer is the only electro-optically active layer. Using Eqns. 8-10 and the approximation that $r_2 << r_3 < r_1$, we can show that:

$$\Delta I_{trans} \approx \frac{I_{trans}}{\text{denom}} \left\{ \frac{4 r_1 r_2 (\sin\gamma)}{\text{denom}} + 4 r_1 r_3 (\sin\delta) \right\} \frac{2\pi}{\lambda} \frac{\Delta n_p n_p L_p}{\sqrt{n_p^2 - \sin^2\Theta}}. \tag{11}$$

where δ is the same as Eqn. 2, denom is given by Eqn. 1b, and:

$$\gamma = \frac{2\pi}{\lambda} L_p \sqrt{n_p^2 - \sin^2\Theta}. \tag{12}$$

Note that the second term in braces gives the same electro-refractive signal as Eqn. 5. Étalon *A* produces the first term in braces. This term gives a vertical offset to the signal since it does not switch sign as quickly with Θ as the sinδ term. Thus, the interference of the first term with the second generates the vertical asymmetry of the electro-optic signal. The wavelength periodicity of the sinγ term, given by Eqn. 7, approximately matches the oscillation period of the data in Fig. 8.

Equation 11 is analogous to Eqn. 5 because it shows two terms that interfere to produce vertically asymmetric lockin signals. However, the incidence angle (Θ) dependence is slightly different between the two models. We examine the modulation signal data for λ = 612 nm to compare the two angular

dependencies. When the TP86 sample is probed with this wavelength, we expect very little contribution from electro-absorption since this wavelength is far from the chromophore absorption edge. However, asymmetric peaks in the lockin signal are observed as shown in Fig. 10. This electro-optic signal dependence is well fit with either the single-étalon, *ER/EA*-interference model (Fig. 10a) or the multiple-étalon, no-electro-absorption model (Fig. 10b). The average transmittance signal functional fits using either model are also quite satisfactory.

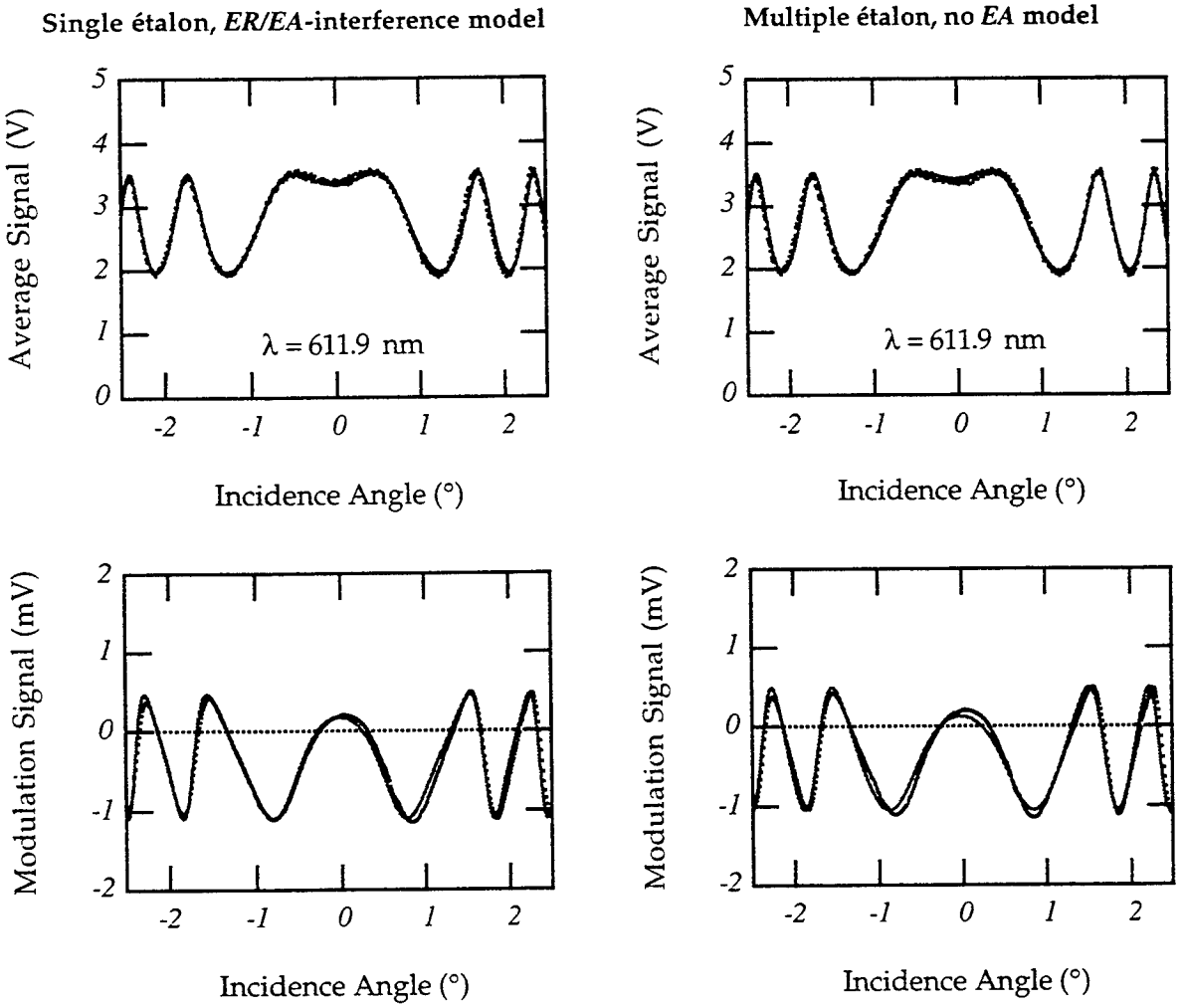


Figure 10 (left) At a probe wavelength $\lambda = 612$ nm the lockin signal (bottom) is unexpectedly asymmetric since the absorption is essentially zero for $\lambda > 550$ nm. The transmittance (average photodiode signal, top) and the asymmetric modulation signal (bottom) are well fit by the single-étalon, *ER/EA*-interference equation (Eqn. 5). (right) The same data are also well fit by the multiple-étalon, no *EA* model.

We further investigated the similarity between the two models by numerically simulating (at a particular wavelength) the angular dependence of the transmittance signal and the electro-optic modulation signal from the multiple-étalon model. This simulation uses the complete multiple-étalon model, not the approximation given by Eqn. 11. The simulated data is then fit with Eqns. 1 and 5 from

the single-étalon, *ER/EA*-interference model. The functional fits to the model data for both the average transmittance and the modulation signal are quite good. Our simulation generates data with the multiple-étalon/no-electro-absorption model for a series of probe wavelengths and fits each of these with the single-étalon, *ER/EA*-interference model. The purely real electro-optic coefficient $r_{13}(\lambda)$ utilized in the multiple-étalon model depends on the wavelength of the simulated probe beam. We used the dispersion model from ref. [9] to fit the data from r_{13} shown in Fig. 8 above. We also increased the simulated thickness of the polymer layer to 3.4 μm to match the periodicity in the measured data. Figure 11 summarizes the results of the simulated dispersion. The fitted electro-absorptive coefficient s_{13} oscillates with a period of ≈30 nm. This shows that the apparent oscillation of the *EA* coefficient in Fig. 8 above is indeed caused by multiple étalon effects and not by actual variations of the coefficient.

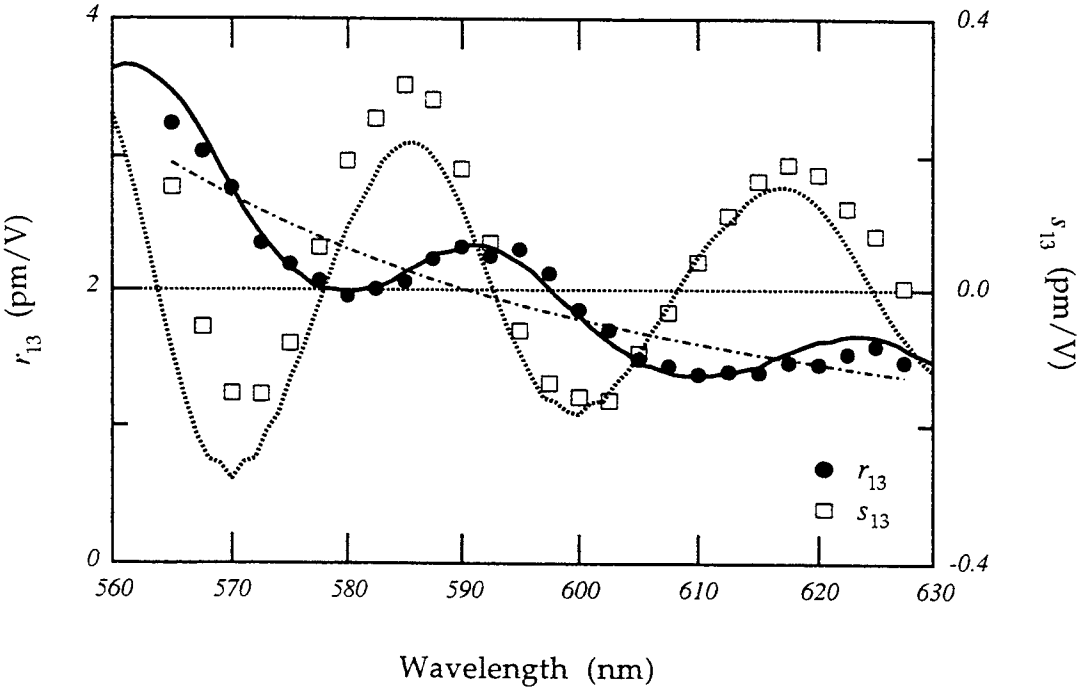


Figure 11 Result of the simulation described in the text. Simulation data is generated at a single wavelength with the multiple-étalon, no-electro-absorption model and then fit with the single-étalon, *ER/EA*-interference model. All parameters in the generation of the data are fixed except the wavelength, which is varied to determine the simulated dispersion of r_{13} and s_{13}. The field reflectivities are 0.7, 0.08, & 0.2 for the gold layer, the polymer/*ITO*/glass interface, and the air/glass interface, respectively. The electro-optic coefficient assumed for the simulation depends on the probe wavelength. We used the dispersion formula given by Chollet *et al.* (ref. [9]) to determine the dispersion of the real part of the electro-optic coefficient r_{13}. The functional fit is shown by the dash-dot line. Notice that a reflectance as low as 0.6% from the *ITO* layer gives the appearance of electro-absorption as large as 0.2 pm/V.

Because both models yield extremely nice predictions of the signals' angular dependence, combining all effects into a single comprehensive model will be quite difficult because of competition to fit the asymmetry. A better approach, currently under investigation, is to increase the finesse of the primary cavity and decrease the importance of the unwanted surface reflection. To reduce the influence

of the unwanted polymer/*ITO*/glass surface reflection, the air/glass interface is coated with gold thus increasing its reflectivity. We coated our TP86 sample in this manner but then damaged it during measurements.

The authors of ref. [9] utilized a complicated method based on the ellipsometric technique of Teng and Man to measure the dispersion of the complex electro-optic coefficient in a Disperse Red 1/PMMA polymer system. Their method requires detailed knowledge of the dispersion of the linear optical coefficients (refractive index, absorption) over the wavelength range of the measurement. Their data provides a good benchmark for verifying our simpler method. We constructed a DR1/PMMA guest-host polymer sample. The sample is quite similar to the test structure shown in Fig. 1, except that two gold mirrors are coated onto the sample. This improves the finesse and should suppress the artificial appearance of *EA*-like effects in our data. Unfortunately, in the wavelength range of our dye laser the absorption of DR1 is relatively large so that the true electro-absorption is substantial. Because of this, we cannot determine whether the increased finesse removed the multiple-étalon artifact or whether it was simply swamped by the much larger *EA* effect. Figure 12 shows the measured dispersion of the real and imaginary parts of the electro-optic coefficient of the DR1/PMMA system. Figure 13 shows a series of measurements in this DR1/PMMA sample at different probe wavelengths. The figure displays both the variation of the transmittance as well as the lockin signal. Notice that the finesse is higher than in Fig. 9 and that the single-étalon, *ER/EA*-interference model fits are again quite satisfactory.

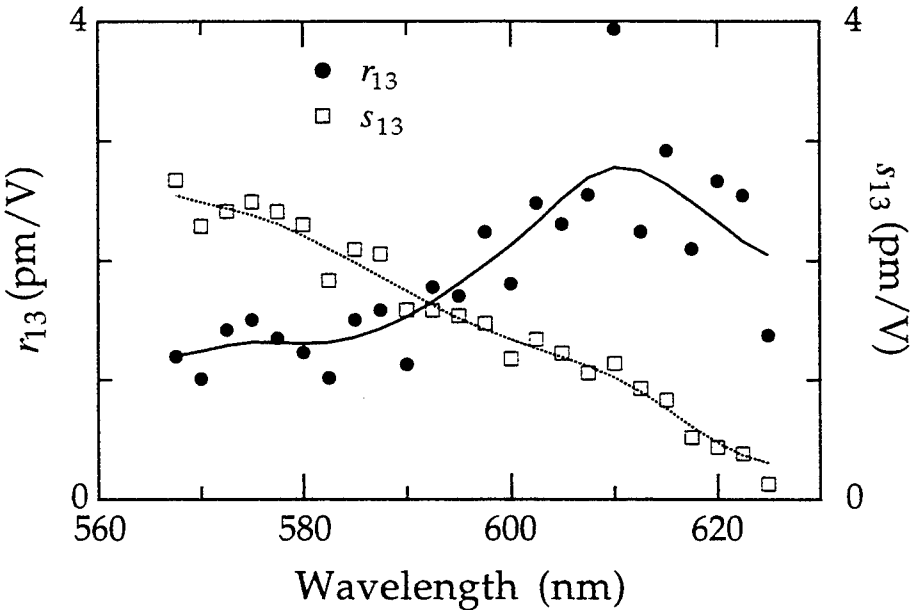


Figure 12 Dispersion of the real r_{13} and imaginary s_{13} components of the complex electro-optic coefficient for a guest/host DR1/PMMA sample. The lines are guides to the eye. Notice that the *EA* coefficient s_{13} is much larger than in TP86 (Fig. 8) and that there is no apparent oscillation in the values.

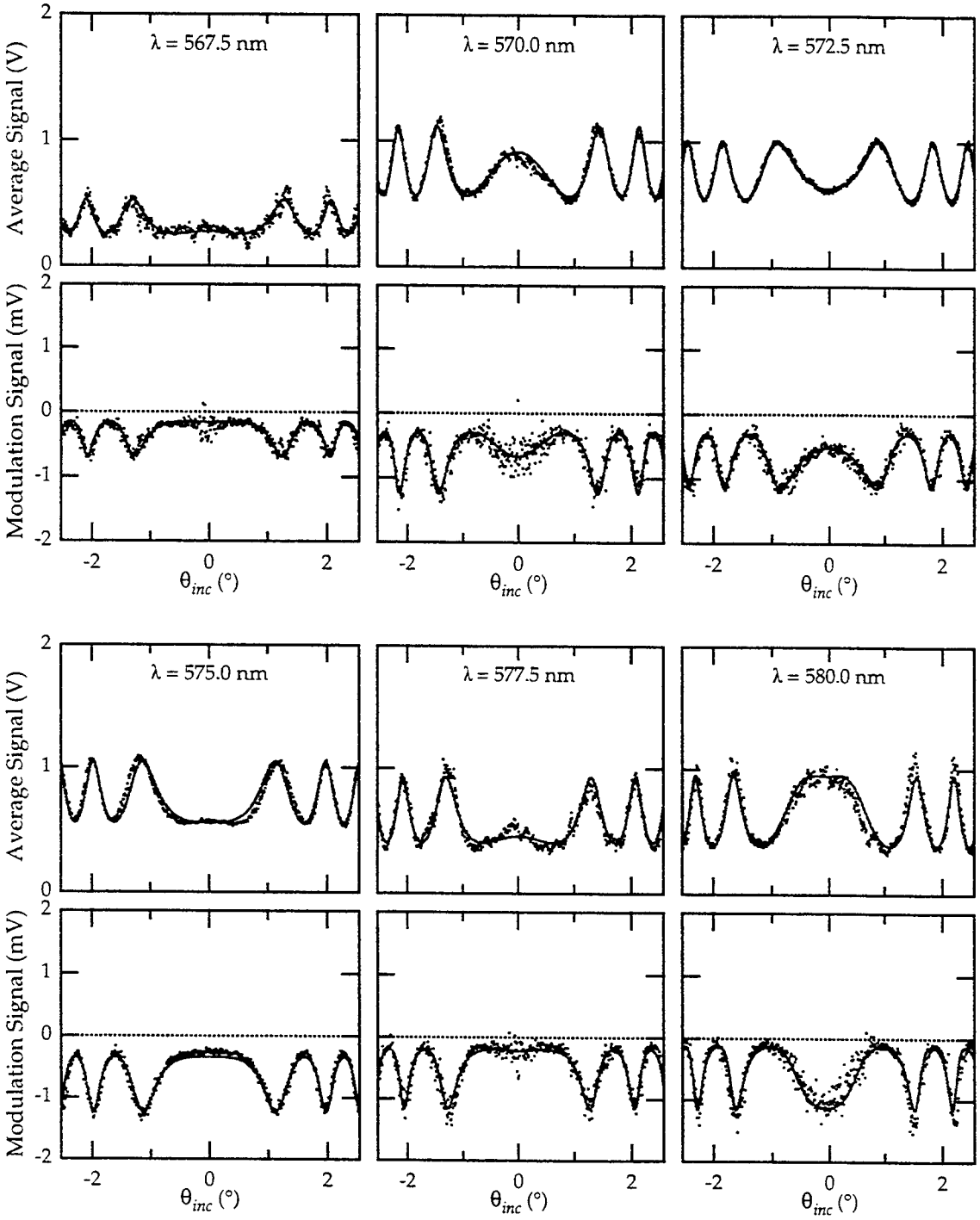


Figure 13 Angular variation of the average photodiode and electro-optic modulation signals from the high-finesse DR1/PMMA Fabry-Perot electro-optic measurement technique (λ = 567.5 nm to 580.0 nm in 2.5 nm increments). The experimental data are represented by dots and theoretical curve fits by lines. The theoretical fits use the single étalon, *ER/EA*-interference model. Note that the complete data set extended from λ = 567.5 nm to 625 nm in 2.5 nm increments. The lockin (modulation) signal does not switch sign on either side of an étalon resonance. This is indicative of strong electro-absorption, as also seen in Fig. 6b for the TP86 sample.

The fact that the lockin signal does not switch sign on either side of an étalon resonance indicates that electro-absorption is the dominant field-induced effect over the wavelength range of the measurement. Such a strong *EA* effect is expected when the probe wavelength approaches the absorption edge of the chromophore. We compared our data to that of Chollet *et al.* (ref. [9]) to explore the validity of our technique. There is quite a large discrepancy between their measurements and Fig. 12 above. The polymer systems are quite similar except that their DR1 chromophore is attached as a sidechain to the PMMA polymer backbone whereas our DR1 chromophore is simply a guest in the host PMMA. The absorption spectra for the sidechain-attached and guest/host systems are almost identical, however. We are exploring the disagreement between the data of ref. [9] and ours by measuring the dispersion of the complex *EO* coefficient over a wider wavelength range.

For wavelengths where *EA* is not expected to be large we propose that high finesses étalons will eliminate the appearance of artificial asymmetries in the data. A simple alternative to creating a high finesse étalon is to ignore oscillations in the imaginary part of the electro-absorption coefficient on the order of 0.2 pm/V since this is the magnitude of the spurious oscillation in the electro-absorption coefficient for the surface reflectivities of our samples. We have also seen that when s_{13} is much greater than 0.2 pm/V one can safely conclude that the asymmetry in the lockin signal arises from electro-absorption and is not an artifact of the unwanted reflections. One final possibility is to alter the sample structure to closely match the refractive indices of the various layers. In any case, in-plane poling[17] might be used to eliminate the need for the *ITO* layer altogether.

Conclusions

We use Fabry-Perot étalons with electro-optically active spacer layers to modulate the transmitted signal via both electro-refraction Δn and electro-absorption $\Delta\alpha$. As we rotate the étalon with a controlled stage the dependence of the modulated light signal determines the complex electro-optic coefficient $\tilde{r}_{13}$. The beauty of this method is its simplicity and the fact that strong electro-absorption ($s_{13} >> 0.2$ pm/V) is immediately recognized by asymmetric modulation signals with unequal peaks on either side of a Fabry-Perot resonance. Unfortunately, multiple (>2) surface reflections cause spurious effects in our experiments that masquerade as electro-absorption. This multiple étalon effect is rather difficult to avoid since reflectances as low as 0.6% produce variations in the measured parameters. The dispersion of the complex electro-optic coefficient shows both the real increase in electro-refraction and electro-absorption as the wavelength approaches the chromophore absorption edge as well as an artificial variation caused by internal reflections. The unwanted étalon problem can appear not only in Fabry-Perot experiments but also in ellipsometric and interferometric (Mach-Zehnder, Michelson) experiments. We regard with caution results that don't properly account for this

spurious effect. We believe that the multiple-étalon artifact may be overcome by coating the air/glass interface to increase its reflectivity. In this manner the finesse of the thick étalon cavity is increased which subsequently reduces the interference from the thin étalon.

References

1. K. D. Singer, M. G. Kuzyk, W. R. Holland, J. E. Sohn, S. J. Lalama, R. B. Comizzoli, H. E. Katz, and M. L. Schilling, "Electro-optic phase modulation and optical second-harmonic generation in corona-poled polymer films," Appl. Phys. Lett. **53**, 1800-1802 (1988).
2. R. A. Norwood, M. G. Kuzyk, and R. A. Keosian, "Electro-optic tensor ratio determination of side-chain copolymers with electro-optic interferometry," J. Appl. Phys. **75** (4), 1869-1874 (1994).
3. V. Dentan, Y. Levy, M. Dumont, P. Robin, and E. Chastaing, "Electro-optical properties of ferroelectric polymers studied by attenuated total reflectance," Opt. Commun. **69**, 379-383 (1989).
4. S. Herminghaus, B. A. Smith, and J. D. Swalen, "Electro-optic coefficient in electric-field-poled polymer waveguides," J. Opt. Soc. Am. B **8**, 2311 (1991).
5. C. C. Teng and H. I. Man, "Simple Reflection Technique for Measuring the Electro-Optic Coefficient of Poled Polymers," Appl. Phys. Lett. **56** (18), 1734-1736 (1990).
6. J. Schildkraut, "Determination of the electro-optic coefficient of a poled polymer film," Appl. Opt. **29** (19), 2839-2841 (1990).
7. K. Clays and J. S. Schildkraut, "Dispersion of the complex electro-optic coeficient and electrochromic effects in poled polymer films," J. Opt. Soc. Am. B **9** (12), 2274-2282 (1992).
8. Y. Levy, M. Dumont, E. Chastaing, P. Robin, P. A. Chollet, G. Gadret, and F. Kajzar, "Reflection method for electro-optical coefficient determination in stratified thin film structures," Mol. Cryst. Liq. Cryst. Sci. Technol. - Sec. B: Nonlinear Optics **4** (4), 1-19 (1993).
9. P.-A. Chollet, G. Gadret, F. Kajzar, and P. Raimond, "Electro-optic coefficient determination in stratified organized molecular thin films: application to poled polymers," Thin Solid Films **242**, 132-138 (1994) and P.-A. Chollet, G. Gadret, F. Kajzar, and P. Raimond, "Determination of the dispersion of the linear and quadratic electro-optic coefficient in thin films by electric field-induced modulation ellipsometry," SPIE vol. **2143**, pp. 54-67 (1994).
10. Y. Shuto and M. Amano, "Reflection measurement technique of electro-optic coefficients in lithium niobate crystals and poled polymer films," J. Appl. Phys. **77** (9), 4632-4638 (1995).
11. V. I. Sokolov, D. B. Kushev, and V. K. Subasniev, "Proposed interference method for determining the signs of electro-optical coefficients," Sov. Phys. Crystallogr. **18** (2), 200-201 (1973).
12. H. Uchiki and T. Kobayashi, "New determination of electro-optic constants and relevant nonlinear susceptibilities and its application to doped polymer," J. Appl. Phys. **64** (5), 2625-2629 (1988).
13. C. A. Eldering, A. Knoesen, and S. T. Kowel, "Use of Fabry-Perot devices for the characterization of polymeric electro-optic films," J. Appl. Phys. **69**, 3676-3686 (1991).
14. R. Meyrueix, J. P. Lecomte, and G. Tapolsky, "A Fabry Perot Interferometric Technique for the electro-optical characterization of nonlinear optical polymers," Nonlin. Opt. **1**, 201-211 (1991).
15. C. H. Wang, B. S. Wherrett, J. P. Cresswell, M. C. Petty, T. Ryan, S. Allen, I. Ferguson, M. G. Hutchings, and D. P. Devonald, "Observation of electro-optic and electroabsorption modulation in a Langmuir-Blodgett film Fabry-Perot étalon," Opt. Lett. **20** (14), 1533-1535 (1995).
16 J. P. Cresswell, M. C. Petty, C. H. Wang, B. S. Wherrett, Z. Ali-Adib, P. Hodge, T. G. Ryan, S. Allen, "An electro-optic Fabry-Perot through-plane-modulator based on a Langmuir-Blodgett film," Opt. Comm. **115**, 271-275 (1995).
17. M. Ziari, S. Kalluri, S. Garner, W. H. Steier, Z. Liang, L. R. Dalton, and Y. Shi, "Novel electro-optic measurement technique for coplanar electrode poled polymers," in *Nonlinear Optical Properties of Organic Materials VIII, ed. G. R. Möhlmann*, SPIE **2527** (1995).
18. J. L. Stevenson, S. Ayers, and M. M. Faktor, "The linear electrochromic effect in meta-nitroaniline," J. Phys. Chem. Solids **34**, 235-239 (1973).
19. R. H. Page, M. C. Jurich, B. Reck, A. Sen, R. J. Twieg, J. D. Swalen, G. C. Bjorklund, and C. G.

Willson, "Electrochromic and optical waveguide studies of corona-poled electro-optic polymer films," J. Opt. Soc. Am. B **7** (7), 1239-1250 (1990).
20. A. Horvath, H. Bassler, and G. Weiser, "Electroabsorption in conjugated polymers," Phys. Stat. Sol. B **173**, 755-764 (1992).
21. F. Qiu, K. Misawa, X. Cheng, A. Ueki, and T. Kobayashi, "Determination of complex tensor components of electro-optic constants of dye-doped polymer films with a Mach-Zehnder interferometer," Appl. Phys. Lett. **65** (13), 1605-1607 (1994).
22. H. Ono, K. Misawa, K. Minoshima, A. Ueki, and T. Kobayashi, "Complex electro-optic constants of dye-doped polymer films determined with a Mach-Zehnder interferometer," J. Appl. Phys. **77** (10), 4935-4940 (1995).
23. K. D. Singer, M. G. Kuzyk, and J. E. Sohn, "Second-order nonlinear-optical processes in orientationally ordered materials: relationship between molecular and macroscopic properties," J. Opt. Soc. Am. B **4** (6), 968-976 (1987).
24. O. S. Heavens, *Optical Properties of Thin Solid Films,* (Dover Publications, New York, 1965).

Part II: The Electro-Optic Probe

The rapid advance of high-speed electronic processing is driven by society's voracious appetite for information. The operational frequency of electronic circuits (*e.g.* the clock rate of computer central processor chips) continues to grow rapidly. The traditional method for inspecting such electronic circuits utilizes metallic probes that physically contact the circuit board and consequently affect its operation. We began development of an all-optical technique to probe the behavior of ultrafast electronic circuits using the emerging technology of electro-optic polymers. The proposed method is non-contacting and does not adversely affect the electronic circuit operation. We will integrate our electro-optic probe with fiber optic delivery so that our system can be utilized in rapid testing during production line assembly. This electro-optic circuit probe will greatly simplify testing and analysis of future generation high-speed electronic circuits.

Electro-optic probes constructed of dielectric and electro-optic materials can be noninvasive since they don't contact or capacitively load a circuit. Recently there has been a great deal of work on using electro-optic crystals as a method of monitoring electrical signals on boards.[1-6] This work has recently turned to poled-polymer films for improvement of the electro-optic probing effect.[7-10] The advantages of poled-polymer films is their large, fast electro-optic response as well as their low permitivity (small low-frequency dielectric constant). The placement of a low permitivity, high resistivity polymer layer near a fast integrated circuit is relatively noninvasive and nonloading. The fastest response time measured to date with a poled-polymer electro-optic (*EO*) probe is 460 GHz.[9]

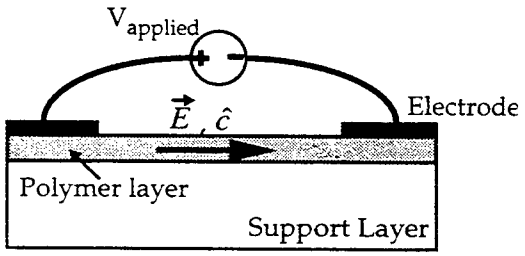


Fig. 14 Transverse poling of a polymer layer. The support layer gives mechanical support to the polymer during poling and during removal of the electrodes after poling. We will attach this structure to the tip of a polarization preserving fiber to create the *EO* Probe.

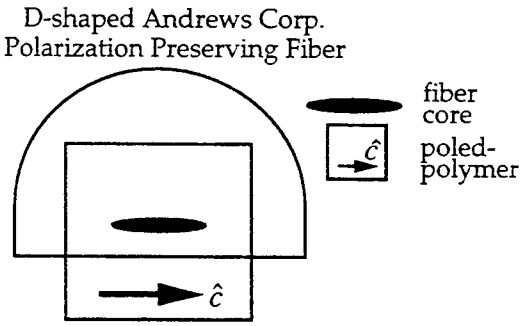


Fig. 15 During attachment of the poled polymer layer onto the tip of the polarization-preserving fiber the polar axis of the electro-optic polymer must align to one of the polarization axes of the fiber, as shown above. This is an end-view of the fiber tip.

All the polymer *EO* probes to date have been affixed to the circuit board under investigation. This is ideal for laboratory experiments but inappropriate for real life test conditions in which one would

like to fabricate a limited number of probes that can be reused for many circuit boards. We suggested a method to achieve just such an *EO* probe with poled polymers. We fabricated a thin poled-polymer layer that can be attached to the end of a polarization maintaining fiber. The tip of the fiber, when placed in the vicinity of a changing electrical signal will serve as an *EO* probe. Since both the polymer layer and the fiber are low permitivity, the circuit behavior will not be adversely affected.

Manufacturing the fiber *EO* probe requires in-plane poling of the polymer layer.[11-17] In-plane poling was unexpectedly challenging and we devoted significant effort towards achieving large *EO* nonlinearities in the in-plane geometry. The primary difficulty is catastrophic, sample-destroying electrical breakdown during poling attempts. The polymer was poled using two electrodes separated as shown in Fig. 14. The polar axis of the resulting electro-optic effect will lie in the plane of the thin polymer layer.[9] This is crucial since without this poling arrangement we must resort to oblique angles of incidence which precludes a fiber system. After poling, we attach the film to the tip of a well-cleaved optical fiber. During attachment the polar axis of the poled polymer must be accurately aligned to one of the axes of the polarization maintaining fiber. A "D" shaped fiber with the polarization axes aligned parallel and perpendicular to the flat of the "D" is well suited for this. We inspect the shape of the fiber cladding and align the *EO* probe accordingly. This is shown schematically in Fig. 15.

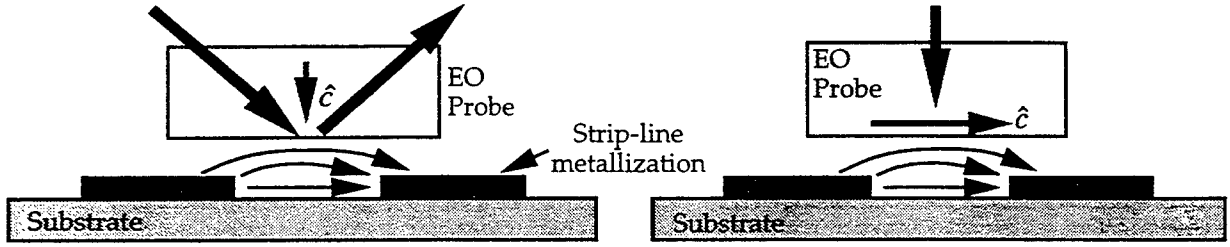


Fig. 16 The left diagram shows the typical method of aligning the poled-polymer polar axis with respect to the field orientation. In this case the signal is only proportional to the weak component of the electric field along the polar axis direction. With transverse poling (shown on the right), the probe senses the much larger transverse electric field component.

Now consider probing a coplanar waveguide circuit. As the electrical signal propagates down the waveguide the electric fields point primarily from one stripe of circuit metallization to the other and are thus strongest in the plane of the circuit. Such fields must, of course, fringe upwards in the vertical direction and most systems measure this smaller field component with the smaller electro-optic coefficient r_{13}. Because of our proposed poling arrangement, we can measure the strong transverse field component with the larger electro-optic coefficient r_{33} (see Fig. 16). To measure the transient electrical signal one launches light from a very quiet semiconductor laser into a polarization-preserving fiber with the input polarization fixed at 45° to the fiber axes. The light travels through the fiber, out through the poled-polymer layer, and strikes one of the circuit metallizations. Part of the reflected light is collected by the fiber (after again passing through the polymer) and travels back through the

fiber to the detection system. The polarization state of the emergent light is altered because the light polarized along the polymer polar axis suffers a different phase shift in the polymer (because of the electrical signal and the Pockels effect) than the perpendicularly polarized light. The detection arm uses a Babinet-Soleil compensator to make the polarization circular in the absence of any applied electrical signal. After the compensator, the reflected beam passes through a Wollaston prism to separate the beam into two orthogonally polarized beams that impinge on two balanced, high-speed PIN photodiodes. The electrical signal in the circuit imposes a change in the returned polarization such that the two detectors see unbalanced powers. This differential detection scheme (overviewed in Fig. 17) cancels common-mode noise.

For an initial test of our proposed *EO* Probe apparatus we will apply a sinusoidal variation to the circuit and utilize a lockin amplifier (in differential mode) to detect the signal. A more general apparatus will require a low noise differential amplifier rather than a lockin, but this will follow proof-of-principle experiments. One difficulty with the system as proposed will be the problem of insuring that the fiber is pointed at a circuit metallization. We monitor the average power of the returned signal to determine whether the fiber is pointing at a nice reflector, such as a metallization. Of course, the thickness of the polymer and support layer is crucial in determining how much light the fiber will collect after reflection.

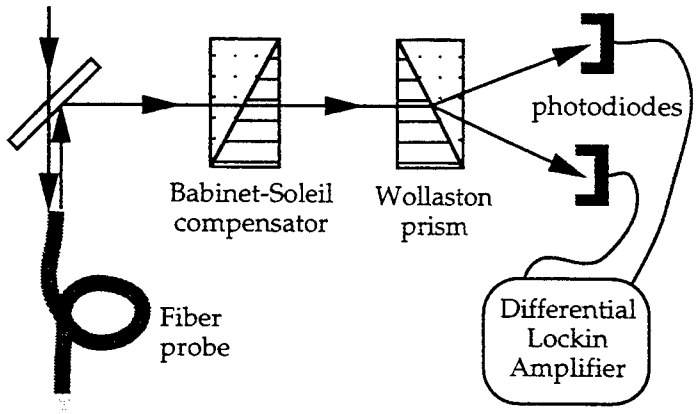


Fig. 17 Electro-optic probe experimental layout. Light is launched with equal efficiency into both polarization modes of a polarization-preserving optical fiber. After propagating through the poled-polymer (immersed in the field region) the light is reflected and re-traverses the fiber. The returned light passes through a compensator and then a Wollaston prism. Any alteration of the returned polarization state results in unequal intensities incident on the two balanced detectors. This is measured differentially by the lockin amplifier.

Progress and Problems

One of the major hurdles that we faced in trying to develop a fiber optic electric field probe using an *EO* polymer layer as the field-sensitive medium is the necessity to pole the sample transversely to its thin dimension. This gives the optimum orientation of the polar axis for detecting electric fields whose strongest component lies in the plane of the circuit board. The in-plane poling proved to be difficult because of catastrophic electric breakdown in either the substrate supporting the polymer layer or along the interface between the polymer and air. We switched from soda-lime glass microscope slides to quartz slides so that the substrate would have lower conductivity and higher dielectric strength. This solved the problem of breakdown in the substrate. Next we struggled with developing polymer overcladdings (see Fig. 18) so that the electric field at the air/polymer interface is reduced to non-catastrophic levels. As a rule of thumb, the electric field from an electrode gap will fringe out sideways from the gap a distance of $\approx W_{gap}$ (the width of the gap). Thus we must cover our ≈3 μm active polymer layer with 20-30 μm of highly resistive overcladding polymer.

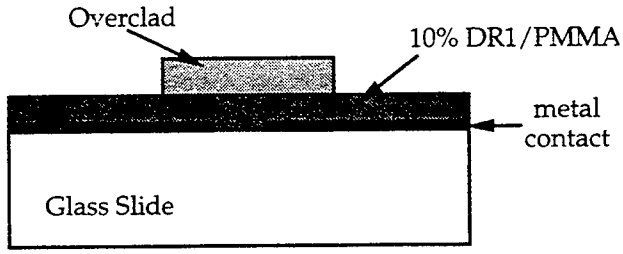


Figure 18 Schematic drawing of the in-plane poling geometry for the electro-optic polymer used in the *EO* probe. The gap in the metal contact is $W_{gap} \approx 20$ μm. This structure is inverted from that of Fig. 14 because the electrodes are underneath the polymer layer. The final device structure will include a layer between the glass and metal so that the structure can be peeled up and reattached to the fiber.

Figure 18 shows a typical sample for in-plane poling. First we evaporate gold to a thickness of several hundred angstroms onto a quartz slide. Next we use photolithography and etching to define several electrode stripes/gaps onto our 1" dia. samples. The photolithographic mask pattern contained five strips about 2 cm long. Each strip has a gap and the gaps are approximately 5, 10, 15, 20, and 25 μm wide. These values are generally obtained using 50 seconds of UV exposure in our mask aligner and 4:1 deionized water and 350 developer for 75 seconds. This geometry makes contacting the electrodes simple. A thick (10-15 μm) layer of 5% PMMA was used as an undercoat layer between the glass coverslip and the polymer to allow us to peal up the polymer after it has been poled so that we can affix it to a fiber tip. Over this substrate/electrode definition we spin coat 10% (wt.) Disperse Red 1/ PMMA guest/host polymer. Initially we used chloroform as the PMMA solvent but we later found that cyclopentanone gives much smoother, more uniform samples while also allowing higher DR1 concentrations. After drying, the samples were poled by placing them on a heating element and

applying voltage through a probe station contacting the electrodes. Without any overcoat polymer we always observed breakdown in the air over the electrode gap that ablated the polymer. The high voltage supply (max. 3000 V) reached poling fields of ≈100 V/μm, even for the largest gaps.

Thus we began a search for an appropriate overclad material. Overcoating the gaps and electrodes with photoresist failed due to the dark color, making it difficult to later transmit a laser beam through the sample. Many of the overcoats that we tried failed because the solvent in the new layer tended to redisolve the active PMMA layer. Five-minute epoxy and overclads of PMMA have this problem. A few samples with both an undercoat and overcoat of 5% PMMA were successfully poled at 100V/μm and 100°C. However the problem of the chloroform in the overcoat redissolving the DR1/PMMA continued causing the electrodes to lift up and move. A thin layer of photoresist was used between the electrodes and the overcoat to prevent changes in the gap position. This was achieved by spin coating the photoresist on the sample after it had been processed, covering the areas near the gaps with overcoat, and then using developer to remove the photoresist to contact the electrodes.

The above sample configuration did not work consistently. The overcoat area frequently developed many small bubbles when heated causing the sample to either short at low voltage or alter the gap size so the sample would not pole. Placing the sample, already overcoated, in the vacuum oven over night to remove excess solvent from the PMMA overcoat did not prevent the bubbling. The problem appeared to be in the thin layer of photoresist used between the electrodes and overcoat in the gap area. When photoresist was placed on a glass slide and covered with PMMA overcoat bubbles appeared when heated. PMMA overcoat alone did not bubble when heated.

The simplest solution is to find a new polymer that is a good dielectric and dissolves in a solvent that does not redissolve PMMA. After consulting with Srinath Kalluri (USC), we found that polystyrene could be dissolved in heated cyclohexane, a solvent that did not dissolve PMMA. To help dissolve the polystyrene a small amount (10:1) of butyl acetate, which does dissolve PMMA, was added to the cyclohexane. The overcoat layer dissolved the DR1/PMMA leaving the polymer an orange color rather than the usual red. Using a 20:1 combination for the solvent causes only slight discoloration. Low percentages of polystyrene (<5%) provided smoother overcoat layers. Higher percentages for the overcoat bubbled when heated. The lower percentages did not bubble but multiple layers were required to prevent break down when voltage was applied. We have recently (Jan. 1996) succeeded in poling several samples prepared in this way. The electro-optic coefficient measurement is discussed below and the *EO* coefficient correlates well with samples poled by parallel-plate techniques and measured with the method of Teng and Man.[18,19]

The samples that we poled for use in the *EO* probe must be characterized in terms of the electro-optical response before attaching them to the fiber tip. To measure the electro-optic activity of our in-plane poled polymer samples we used a variation of the common Teng and Man ellipsometric technique.[18,19] Figure 19 below shows the transmission geometry setup required for this measurement. The output of a laser diode is polarized and attenuated and then focused with a 20X objective onto the poled gap of the sample. The throughput light is collimated with a 10X objective. This light then passes through a Babinet-Soleil compensator (which is set so that in the absence of any voltage applied to the sample, the compensator produces a circularly polarized output state) and then an analyzer. We insure that the beam is centered in the gap by observing the back reflected light with the CCD camera and TV monitor. An HP8116 signal generator applies ±16V sinusoidal voltage across the ≈25 μm gap. The lockin amplifier detects the *EO*-induced change in the photodiode signal.

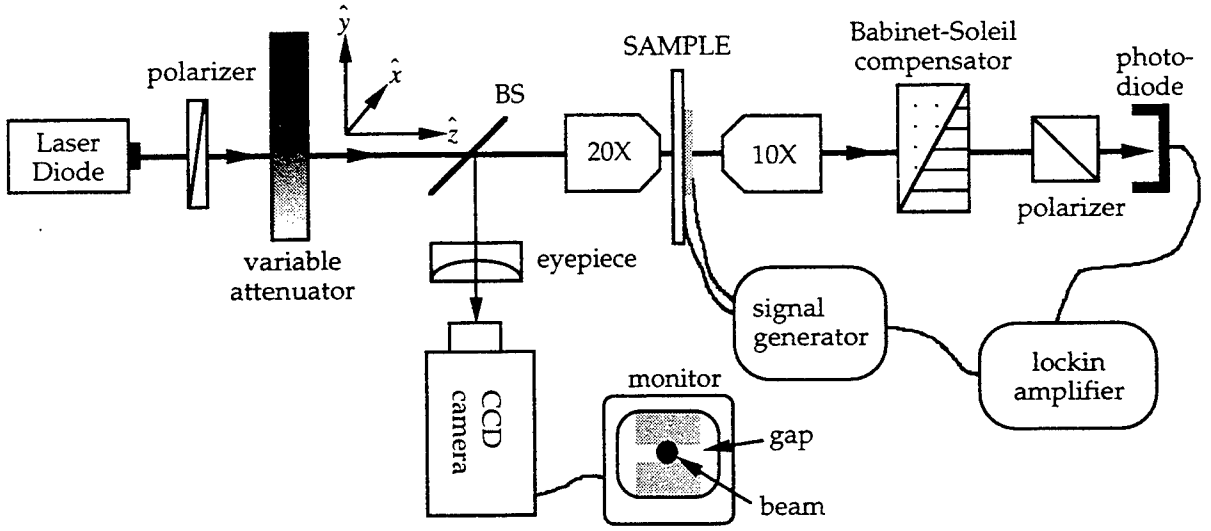


Figure 19 *EO* measurement system for the in-plane poled samples. The system is the transmission analog to the Teng and Man reflection measurement. The input polarization lies at 45° from the *x*-*y* axes so that it splits equally into ordinary and extraordinary components in the polymer (whose polar axis is assumed to lie along $\hat{y}$. A CCD camera helps insure that the optical beam passes through the tiny gap in the poling electrodes.

Typically the DR1/PMMA polymer layer is 3 μm thick. Since the gap width is much greater than the thickness of the active polymer layer we can assume that both the poling and modulating electric fields are uniform throughout the active polymer layer. This means that we can ignore fringe-field effects and assume that the gap is uniformly poled with the polar axis aligned with $\hat{y}$ (shown in the figure). The probe beam polarization state (before entering the poled polymer sample) is at 45° with respect to the *xy* axes so that the ordinary and extraordinary polarizations within the sample are equally excited. The applied voltage will alter the ordinary and extraordinary refractive indices according to:

$$\Delta n_o = -\frac{1}{2} n_o^3 r_{13} \frac{V_{applied}}{W_{gap}} \tag{13a}$$

$$\Delta n_e = -\frac{1}{2} n_o^3 r_{33} \frac{V_{applied}}{W_{gap}} \tag{13b}$$

where W_{gap} is the width of the gap. Because $3r_{13} = r_{33}$, the change in the ordinary refractive index is smaller than the change in the extraordinary index. The polarization state emerging from the polymer layer is altered by this field-induced birefringence.

If we assume that before applying a modulation field to the polymer that the ordinary and extraordinary refractive indices are equal $n_o = n_e = n$, then the change in the intensity after the analyzer ΔI_{after} is given simply as:

$$\frac{\Delta I_{after}}{I_{after}} = -n^3 \left(r_{33} - r_{13}\right) \frac{2\pi}{\lambda} T_{poly} \frac{V_{applied}}{W_{gap}} \tag{14}$$

where T_{poly} is the thickness of the active polymer layer and the compensator is set for maximum linearity and sensitivity which implies that half the incident light makes it to the photodiode when the applied voltage is zero: $I_{after} = I_{incident}/2$. Generally, one uses the assumption that $r_{33} = 3r_{13}$ and we then find that:

$$r_{33} = -\frac{3\lambda}{4\pi n^3 V_{applied}} \frac{\Delta I_{after}}{I_{after}} \frac{W_{gap}}{T_{poly}}. \tag{15}$$

Notice that we need accurate measurements of both the gap width and the polymer thickness. The electrode gap width is measured with a microscope and the polymer thickness is determined with a DekTak profilometer. We have successfully characterized our in-plane films with this method. A recent 10% DR1/PMMA sample poled at about 100 V/μm at T = 68°C for 10 minutes gave an EO coefficient of ≈6 pm/V. This corresponds well with the value we obtain utilizing a sample with the same DR1 doping percentage poled through the thickness and measured in the standard Tend & Man ellipsometric technique.

Conclusions

Our recent success in overcoming the in-plane poling problem has re-invigorated our efforts to create a fiber-based EO probe for high-speed electronic circuit measurements. We are continuing this effort even after the AFOSR grant expires. We will, of course, acknowledge the grant when we succeed in publishing this work. We also expect to utilize our in-plane poled samples to further investigate the

ER and *EA* effects described in Part I since samples poled in this manner allow access to both $\tilde{r}_{13}$ and $\tilde{r}_{33}$. In-plane poling will also eliminate the need for the high refractive index *ITO* layer that caused the spurious artifacts in the *ER/EA* measurements as well.

References

1. J. A. Valdmanis and G. Mourou, "Subpicosecond electrooptic sampling: Principles and applications," IEEE J. Quantum Electr. QE-22, 69-78 (1986).
2. B. H. Kolner and D. M. Bloom, "Electrooptic sampling in GaAs integrated circuits," IEEE J. Quantum Electr. **QE-22**, 79-93 (1986).
3. K. J. Weingarten, M. J. W. Rodwell, and D. M. Bloom, "Picosecond optical sampling of GaAs integrated circuits," IEEE J. Quantum Electr. **QE-24**, 198-220 (1988).
4. U. D. Keil and D. R. Dykaar, "Electro-optic sampling and carrier dynamics at zero propagation distance," Appl. Phys. Lett. **61** (13), 1504-1506 (1992).
5. D. R. Dykaar, R. F. Kopf, U. D. Keil, E. J. Laskowski, and G. J. Zydzik, "Electro-optic sampling using an aluminum gallium arsenide probe," Appl. Phys. Lett. **62** (15) 1733-1735 (1993).
6. G. Baur and G. Sölkner, "Laser diode based electro-optic measurement system with high voltage resolution," Rev. Sci. Instrum. **64** (4), 1081-1084 (1993).
7. T. Nagatsuma, T. Shibata, E. Sano, and A. Iwata, "Subpicosecond sampling using a noncontact electro-optic probe," J. Appl. Phys. **66** (9), 4001-4009 (1989).
8. J. I. Thackara, D. M. Bloom, and B. A. Auld, "Electro-optic sampling of poled organic media," Appl. Phys. Lett. **59** (10), 1159-1161 (1991).
9. P. M. Ferm, C. W. Knapp, C. Wu, J. T. Yardley, B.-B. Hu, X.-C. Zhang, and D. H. Auston, "Femtosecond response of electro-optic poled polymers," Appl. Phys. Lett. **59** (21), 2651-2653 (1991).
10. T. Nagatsuma, M. Yaita, and M. Shinagawa, "External electro-optic sampling using poled polymers," Jpn. J. Appl. Phys. **31** L1373-L1375 (1992).
11. J. W. Wu, J. F. Valley, S. Ermer, E. S. Binkley, J. T. Kenney, G. F. Lipscomb, and R. Lytel, Appl. Phys. Lett. **58**, 225 (1991).
12. A. Nahata, J. Shan, J. T. Yardley, and C. Wu, J. Opt. Soc. Am. B **10**, 1553 (1993).
13. A. Nahata, C. Wu, C. Knapp, V. Lu, J. Shan, and J. T. Yardley, Appl. Phys. Lett. **64**, 3371 (1994).
14. S. Yitzchaik, G. Berkovic, and V. Krongaus, J. Appl. Phys. **70**, 3949 (1991).
15. M. Stähelin, C. A. Walsh, D. M. Burland, R. D. Miller, R. J. Twieg, and W. Volksen, J. Appl. Phys. **73**, 8471 (1993).
16. A. Otomo, G. I. Stegeman, W. H. G. Horsthuis, and G. R. Möhlmann, "Strong field, in-plane poling for nonlinear optical devices in highly nonlinear side change polymers," Appl. Phys. Lett. **65** (19), 2389-2391 (1994).
17. M. Ziari, S. Kalluri, S. Garner, W. H. Steier, Z. Liang, L. R. Dalton, and Y. Shi, "Novel electro-optic measurement technique for coplanar electrode poled polymers," in *Nonlinear Optical Properties of Organic Materials VIII, ed. G. R. Möhlmann*, SPIE **2527** (1995).
18. C. C. Teng and H. I. Man, "Simple Reflection Technique for Measuring the Electro-Optic Coefficient of Poled Polymers," Appl. Phys. Lett. **56** (18), 1734-1736 (1990).
19. J. Schildkraut, "Determination of the electro-optic coefficient of a poled polymer film," Appl. Opt. **29** (19), 2839-2841 (1990).

APPENDIX

(Preprint of Jour. Appl. Phys. submission)

Electro-refraction and electro-absorption in poled polymer Fabry-Perot étalons

Ned F. O'Brien and Vince Dominic
Center for Electro-Optics
University of Dayton
Dayton, Ohio 45469-0245

and

Stephen Caracci
Wright Laboratories - Materials Directorate
WL/MLPO Bldg. 651
3005 P St. STE 6
WPAFB, Ohio 45433-7707

Abstract

We present a simple experimental procedure that uses a slowly-rotating étalon to measure simultaneously the electro-refraction and electro-absorption in a poled polymer. Both effects generally contribute to the measured signal from such material systems and can be distinguished by rotating the sample and observing asymmetric peaks in the signal. The experimental results show the expected increase in both electro-refraction and electro-absorption as the probe wavelength approaches the absorption band of the chromophore. Furthermore, the dispersion of the complex electro-optic coefficient displays a periodic variation that we attribute to multiple-étalon interference. The stratified nature of the thin-film structure causes the multiple-reflection interference. This artifact will pollute most of the standard electro-optic characterization techniques for poled-polymer films.

PACS #: 42.70Jk, 78.66, 78.20Jq, 42.70Nq

Introduction

The low dielectric constant of polymer films makes them quite useful for high-speed opto-electronic applications such as electro-optic switching or modulation. Moreover, poled-polymer fabrication is compatible with integrated circuit board technology and polymer devices are potentially less expensive than lithium niobate electro-optic components. The development of better electro-optic polymers and some initial device implementations have spurred tremendous interest in recent years. Research has focused on improving the lifetime stability of the induced nonlinearity - especially at elevated temperatures - and developing chromophore / polymer systems with sizable nonlinearities. In either case, improvement requires a reliable means of characterizing the polymer's electro-optical properties for feedback to the molecular engineering process.

Many techniques reported in recent years measure the nonlinear optical properties of organic polymer thin film devices. Such techniques include: Michelson and Mach-Zehnder interferometric techniques,[1,2] attenuated total reflection techniques,[3,4] ellipsometric/polarimetry techniques,[5-10] and Fabry-Perot étalon modulation schemes.[11-16] Our method is a variation of previous Fabry-Perot measurement techniques. The primary difference is that we use a much thicker étalon (millimeters instead of microns) whose thickness includes the glass substrate of the sample. The primary advantage of the thicker étalon is that it displays multiple resonance peaks for small rotation angles. We typically observe transmissivity

curves with 5 resonance peaks within a 5° rotation. The multiple resonances allow accurate determination of the étalon parameters. Another significant advantage is that by rotating through several resonance peaks we can discern whether multiple effects contribute to the modulation signal. Typically, this is evidenced by a vertically offset modulation signal. Additional advantages of our method is that the experimental setup is simple, interpretation of the results is straightforward, and the sample structure is consistent with that used by many other measurement techniques. For example, the sample is readily switched into a Mach-Zehnder or ellipsometric/reflection setup to verify results.

Most proposed electro-optic devices require a relatively large, phase-only modulation. This accounts for the emphasis that is placed on the field induced changes in the refractive index, Δn, denoted herein as electro-refraction (ER). Relatively few papers discuss the concurrent field-induced change in absorption, $\Delta\alpha$, denoted electro-absorption (EA).[7-9,17-22] The composite electro-optic (EO) effect results from both phase and amplitude modulations of the probe light beam. With a complex electro-optic coefficient, the real part governs the phase modulations and the imaginary part describes the amplitude modulations. Decomposition into phase and amplitude effects becomes particularly important when the probe wavelength approaches the chromophore absorption band.[9,23] Unfortunately, the resonant enhancement of electro-refraction is accompanied by increased electro-absorption. If neglected, the interaction of the competing effects may lead to significant over- or under-estimation of the electro-optic coefficient. We also found that unwanted surface reflections give signal asymmetries that can be incorrectly interpreted as electro-absorption. These multiple surface reflections are characterized with a simple model showing how this artifact imposes the appearance of electro-absorption onto a purely electro-refractive signal. By measuring the wavelength dispersion of the complex electro-optic coefficient, the spurious multiple-reflection artifact (which oscillates with wavelength) can be distinguished from the underlying electro-refraction and electro-absorption.

Sample description and experimental setup

Figure 1 shows the sample geometry and the experimental arrangement that we used for the electro-optic characterization experiments. The substrate is a ≈1 mm thick glass slide coated with the transparent conductor indium tin oxide (ITO). The ITO is etched so that a narrow stripe extends the length of the slide. This stripe acts as one electrical contact. The polymer is then spin-coated on top of the ITO with a thickness of approximately 2 μm. Once the polymer dries, a thin stripe of gold is evaporated on top, perpendicular to the ITO stripe. This gold strip serves as the second electrical contact for the sample. This structure sandwiches a ≈25 mm^2 rectangular region of the polymer between the ITO and gold contacts. This region of the sample is poled by first heating the polymer to its glass transition temperature and then applying a strong dc electric field (≈100 V/μm). The torque on the dipolar chromophore molecules causes a macroscopic polar alignment within the polymer. After decreasing the temperature, the cooled polymer semi-permanently maintains the non-centrosymmetric alignment of the chromophores within the electro-optically active or poled region.

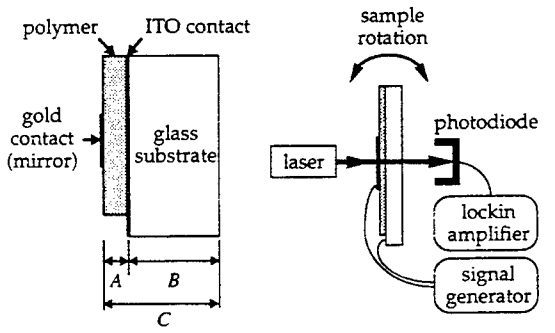


Figure 1) Schematic view of the experimental arrangement and the generic poled-polymer sample geometry (not to scale). (a) Sample description: The sample consists of the following layers: gold electrode/mirror, polymer layer, ITO contact, and the glass substrate. The substrate is typically ≈1 mm thick, the ITO ≈60 nm thick, the polymer ≈2 μm thick, and the gold ≈35–50 nm thick. (b) A laser beam transmitted through the poled region of the Fabry-Perot structure is collected by a photodiode. The sample is slowly rotated through ±2.5° as the lockin measures the angular dependence of both the average photodiode signal and the modulation signal (light signal variation imposed by the signal generator). A computer (not shown) controls the sample rotation and acquires the data.

After poling, the sample is attached to a motorized rotation stage and slowly rotated ±2.5° while the transmitted light beam is monitored, as shown schematically in Fig. 1b. The full angular dependence of both the transmittance and the field-induced modulation of the transmittance is measured. Both Δn and $\Delta\alpha$ of the poled polymer are measured in this simple way. References 7-9 and 21-22 describe techniques that also measure the real and imaginary components of the complex electro-optic coefficient, but not with the simplicity of the present method. Recently, Ziari *et al.* suggested an elegant method based on Young's double-slit experiment that also measures both the phase and amplitude modulation effects in poled polymers.[17] Their technique was designed for coplanar-poled polymer samples.

We direct a laser beam through the poled region of

the sample and collect the transmitted light with a photodiode. The gold and ITO serve as contacts for applying the modulation voltage. The gold and glass/air interface provide mirrors for the étalon. A sinusoidal voltage (±16 V, 5 kHz) applied across the ≈2 μm thick polymer layer induces phase and amplitude changes in the sample. A Stanford Research SR530 lockin amplifier measures both the average and the time-varying components of the transmitted light signal while a controlled actuator slowly rotates the sample through ±2.5° (0° = normal incidence). We are careful to insure that the rotational axis is centered on the incidence spot to prevent translation of the beam across the poled region as the sample rotates. This is important because the spin-coating process results in a polymer layer with varying thickness. Even sub-wavelength thickness variations can strongly affect the Fabry-Perot's throughput. The input beam is vertically polarized so that the $\tilde{r}_{13} = r_{13} + i\, s_{13}$ component of the complex electro-optic coefficient is measured (using the notation of ref. 7). The incidence angle is small enough so that the contribution of the $\tilde{r}_{33}$ component can be ignored even if we utilize *p*–polarized light.

Single-étalon, *ER/EA*-interference model

Let us, for the moment, ignore some of the reflecting surfaces in our sample and model the sandwich structure with only two reflections: the gold film and the air/substrate interface. This corresponds to étalon *C* in Fig. 1a. The étalon thickness includes the electro-optically active polymer layer, the ITO, and the glass substrate. The reflectivities and transmisivities for the three reflecting surfaces (1 = gold, 2 = polymer/ITO/glass interface, 3 = air/glass interface) are labeled r_j, t_j (j = 1,2,3), and we assume here that $r_2 = 0$. The refractive indices of the glass and polymer are denoted as n_g and n_p, respectively. The *s*–polarized probe beam selects the ordinary refractive index of the polymer $n_p = n_o$. The layer thicknesses are L_g and L_p and the glass substrate is assumed to have no absorption at any of the wavelengths of interest. The ITO is very thin compared to either the polymer or the glass layer and is therefore ignored in this simple model. The intensity transmission of this simplified Fabry-Perot étalon is:

$$I_{trans} = \frac{I_{inc}\, e^{-\alpha_p L_p} \left| t_1\, t_3 \right|^2}{\text{denom}} \tag{1a}$$

where:

$$\text{denom} = \left(1 + r_1\, r_3\, e^{-\alpha_p L_p}\right)^2 - 4\, e^{-\alpha_p L_p}\, r_1\, r_3 \sin^2(\delta) \tag{1b}$$

where α_p is the absorption in the polymer and:

$$\delta = \frac{2\pi}{\lambda}\left\{ L_g \sqrt{n_g^2 - \sin^2\Theta} + L_p \sqrt{n_p^2 - \sin^2\Theta} \right\} \tag{2}$$

with the wavelength λ and the external angle of incidence Θ

The applied modulation voltage induces a change in both n_p and α_p. In the case were the absorption is reasonably small such that $\alpha_p\, \lambda << 1$, Clays and Schildkraut[7] showed that the field-induced perturbations Δn_p and $\Delta\alpha_p$ may be written in terms of the real and imaginary parts of the complex electro-optic coefficient $\tilde{r}_{13}$:

$$\Delta n_p = -\frac{1}{2} n_p^3\, r_{13}\, E_{applied} \tag{3a}$$

and:

$$\Delta\alpha_p = -\frac{2\pi}{\lambda} n_p^3\, s_{13}\, E_{applied} \tag{3b}$$

where $E_{applied}$ is the applied modulation field:

$$E_{applied} = \frac{V_{applied}}{L_p} \cos(2\pi f t) \tag{4}$$

and f is the modulation frequency.

The differential change in the transmitted intensity caused by the field-induced modulation is:

$$\Delta I_{trans} = \frac{I_{trans}}{\text{denom}} * \left\{ -\Delta\alpha_p\, L_p \left(1 - r_1\, r_3\, e^{-\alpha_p L_p}\right)\left(1 + r_1\, r_3\, e^{-\alpha_p L_p}\right) + \frac{8\pi}{\lambda} r_1\, r_3\, e^{-\alpha_p L_p} \sin\delta \frac{\Delta n_p\, n_p\, L_p}{\sqrt{n_p^2 - \sin^2\Theta}} \right\}. \tag{5}$$

The first term in braces is the electro-absorptive effect while the second is the electro-refractive. Equations 1 and 5 provide a theoretical model to fit the measured data. Notice that after Eqn. 1 is utilized to fit the transmittance signal, all the parameters except Δn_p and $\Delta\alpha_p$ are determined. Equations 3 & 5 show that both Δn_p and $\Delta\alpha_p$ contribute linearly to the *1f* lockin signal.

The linear dependence of the electro-optic perturbations is verified by varying the amplitude of the applied modulation voltage. The electro-strictive, electro-mechanical, and Kerr effects are ignored because they will appear as *2f* signals on the lockin $\left(\propto \left|E_{applied}\right|^2\right)$. By contrast, a piezoelectric effect, ΔL_p, will contribute to the *1f* signal. Inspection of Eqn. 1 shows that the functional form of the piezoelectric effect is precisely the same as Eqn. 5 but with the Δ shifted to L_p in both terms within braces. The large numerical factor $2\pi/\lambda$ in the second term indicates that ΔL_p will have essentially the same

angular dependence as Δn_p. To determine whether there is a piezoelectric contribution, a second sample made with the same polymer but with a more reflective gold layer is placed in a Michelson interferometer with the gold contact as one of the retroreflecting mirrors. With the sample in this orientation the light does not pass through the polymer and therefore electro-refraction and electro-absorption cannot contribute to the signal. Any piezoelectric effect in the polymer will move the gold mirror and thus contribute to the lockin signal. In this experimental configuration we observed no evidence of piezoelectricity in this polymer system.

The angular dependence of the electro-absorptive term is determined by the variation of I_{trans} whereas the electro-refractive term varies according to $I_{trans} \sin\delta$. This gives a strong asymmetry to the magnitude of the lockin signal peaks as the incidence angle changes such that δ shifts by $\approx\pi$. If Eqn. 1b is re-written as:

$$\text{denom} = 1 + \left(r_1 r_3 e^{-\alpha_p L_p}\right)^2 + 2 e^{-\alpha_p L_p} r_1 r_3 \cos\delta \qquad (6)$$

then because of the $\cos\delta$ term, the transmitted intensity is at the mid-visibility point of the Fabry-Perot resonance when δ is approximately $\pi/2$ or $3\pi/2$. At both these points, labeled *A* and *B* in Fig. 2 below, the electro-absorptive (*EA*) contribution to the lockin signal is the same (magnitude and sign) whereas the electro-refractive (*ER*) signal is of equal magnitude but opposite sign. This behavior is illustrated in Fig. 2. At point *A* there is constructive interference (the *ER* & *EA* signal components add) while at point *B* there is destructive interference (the *ER* & *EA* signal components subtract). Thus, when both electro-refraction and electro-absorption contribute, the electro-optic signal magnitude is unequal on either side of a Fabry-Perot resonance. Such asymmetries have been observed by others.[21,22]

Experimental results

Most of the data below is gathered using a golden yellow sample labeled TP86 supplied by the Dow Chemical Company. Figure 3 displays the angular variation of the average photodiode signal along with the magnitude of the *1f* lockin signal as the poled-polymer Fabry-Perot sample is rotated. The data are represented by dots and the theoretical curve fits using Eqns. 1 and 5 are displayed as solid lines. The average signal shows the expected low-finesse Airy function behavior. According to the simple model discussed above, there is no electro-absorptive contribution to the modulation signal because the positive and negative peaks of the lockin signal are of equal magnitude. The signal variation with incidence angle is shaped as one would expect for a purely electro-refractive electro-optic effect. The advantage of our thick étalon (mm) versus the thin (μm) étalons studied previously[11-16] is the appearance of many resonances while tuning over small angles ($\Theta \leftrightarrow \pm 2.5°$). This gives ample data to precisely determine the Δn variation. The data sets are analyzed by first using Eqn. 1 to fit the average signal angular dependence. This is a three parameter fit where the incident optical power, the cavity finesse, and the thickness of the polymer are allowed to vary. The polymer thickness is restricted to vary by less than $\pm\lambda$ and the finesse does not change by more than 10% for any of our probe wavelengths. Since the air/glass interface has low reflectivity the finesse of the cavity is only ≈1.37. After fitting the average photodiode signal, all parameters are held fixed except Δn and $\Delta\alpha$ which are then fit to the variations in the lockin signal.

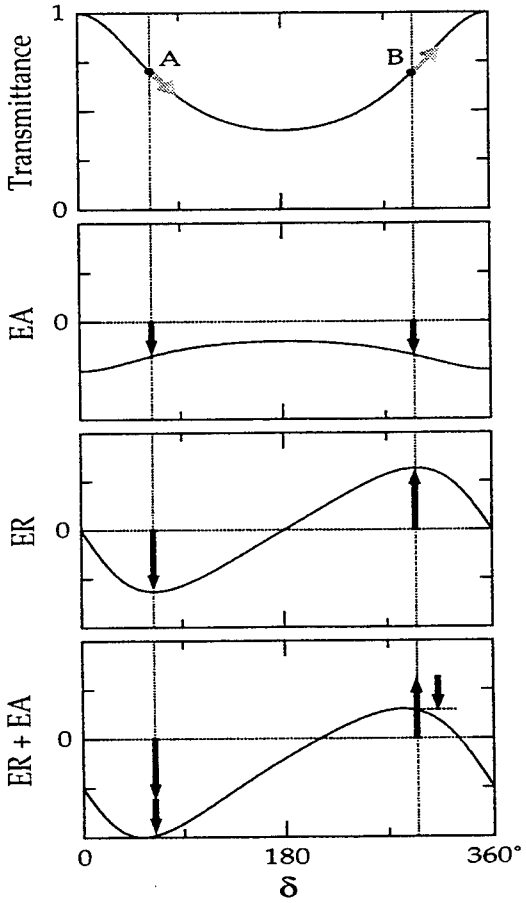


Figure 2) This figure shows how the electro-refractive (*ER*) signal and the electro-absorptive signal (*EA*) interfere to give an asymmetric response as the Fabry-Perot structure is rotated. Part (a) shows the low finesse ($\mathcal{F} = 1.5$) transmittance variation with δ and indicates two points where the *ER* signal magnitude is maximized. The arrows indicate how a positive Δn affects the transmittance. Part (b) shows the expected *ER* response (positive Δn) while part (c) shows the *EA* response (positive $\Delta\alpha$). The total signal, *ER+EA*, in part (d) displays a strong vertical

asymmetry.

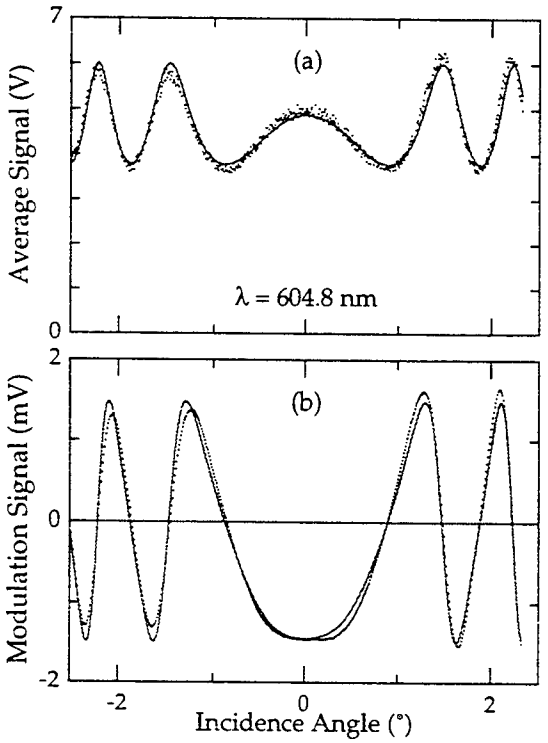


Figure 3) (a) Average photodiode signal and theoretical fit (dots=data, line=fit) as the étalon is rotated with λ = 604.8 nm. (b) Lockin signal dependence on the incidence angle with the raw data (dots) and the theoretical fit (solid line). The average signal shows the expected low-finesse étalon behavior and the lockin signal displays symmetric peaks that reach their maximum where the slope of the average signal is highest. The modulation signal has equal magnitude on either side of a Airy-function transmittance resonance.

By tuning the probe laser to λ = 594.1 nm, the appearance of the lockin signal is dramatically altered. According to the simple model, the vertical asymmetry of the modulation signal apparent in Fig. 4a is the signature of the electro-absorptive effect interfering with the electro-refractive effect. If the laser is tuned even closer to the absorption edge of the chromophore then electro-absorption completely dominates the measurement.[15] This is illustrated in Fig. 4b below where a probe laser wavelength of λ = 543.5 nm is used on the golden yellow sample whose absorption edge lies at approximately λ = 520 nm. Notice that the lockin signal never crosses zero (compare to Fig. 3b, 4a).

A tunable HeNe laser was then used to roughly determine the dispersion of the complex electro-optic coefficient. At a probe wavelength of λ = 543.5 nm we measured $\tilde{r}_{13} \approx (5 + i\,1.25)$ pm/V. Both the electro-refractive (*ER*) and electro-absorptive (*EA*) contributions to the signal increased dramatically as the probe wavelength approached the edge of the chromophore absorption band. This behavior is generally expected (see, for example, the two-level models developed in refs. 9 and 23). Curiously, the measured *EA* coefficient switched sign at a probe wavelength of λ = 612 nm. This is difficult to explain since the absorption itself monotonically decreases for wavelengths longer than λ = 520 nm. In fact, further investigations revealed that the *EA* coefficient magnitude varied and switched sign as the probe beam was moved to different spots within the poled region. This erratic behavior may be caused by the nonuniform thickness of the film created during the spin-coating process. Since the electro-optic coefficient represents a fundamental property of the material, it should be independent of the polymer layer thickness, as was nicely demonstrated in ref. 14. Furthermore, the actual electro-absorption for $\lambda > 590$ nm is very small. Therefore, we don't expect significant electro-absorption at these wavelengths. However, the modulation signal data at both λ = 612 and 632.8 nm displayed a strong vertical asymmetry. We show below that the unwanted reflectivity from the polymer/ITO/glass interface leads to vertically offset lockin signals that can be incorrectly interpreted as the *EA* effect. Thus, although the λ = 543.5 nm data is clearly indicative of electro-absorption, the longer wavelength data is polluted by multiple reflections that give the appearance of electro-absorption.

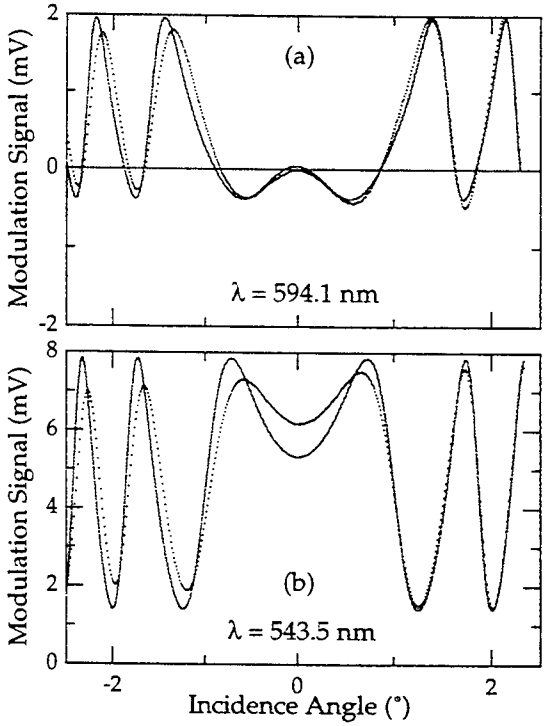


Figure 4) (a) Lockin signal variation as the étalon is rotated in a λ = 594.1 nm probe beam. The peaks of the lockin signal are quite asymmetric: the positive excursion is about four times larger than the negative excursion. (b) Lockin signal dependence on the incidence angle at λ = 543.5 nm. In this case the signal never crosses zero and electro-absorption completely dominates the measurement. In both (a) and (b) the dots are raw data and the solid curve is the theoretical fit using the single étalon, *ER/EA*-interference model. The average

signals incident on the photodiode displayed low-finesse étalon behavior much like Fig. 3a.

To better understand the multiple étalon problem, the electro-optic dispersion of the TP86 sample was re-measured with a dye laser. Pumping a rhodamine-6G dye with an argon laser allows tuning of the probe wavelength over a range from λ = 565–627.5 nm. Figure 5a shows the wavelength dependence of the real and imaginary parts of the complex electro-optic coefficient $\tilde{r}_{13}$ determined at 2.5 nm intervals. Notice that the imaginary part (*EA*-component) shows a clear periodic variation with a period of ≈30 nm. This variation is unphysical based on the monotonic decrease of the absorption over this wavelength range. The oscillation is caused by the additional étalons (*A* & *B* in Fig. 1a) created by the unwanted polymer/ITO/glass reflection. Also note that the sinusoidal variation of both the *ER* and *EA* components are superimposed on a curve that decreases at longer wavelengths. The absorption spectra (measured with a Perkin-Elmer spectrophotometer) of this sample displays a monotonically decreasing absorbance for wavelengths $\lambda > 520$ nm. The oscillation period in the imaginary part of the electro-optic coefficient corresponds to an étalon with the thickness and refractive index of the TP86 polymer (*A* in Fig. 1a):

$$\Delta\lambda = \lambda^2/(2 n_p L_p) = 36\ \text{nm} \qquad (7)$$

where $L_p \approx 2.9$ μm and $n_p \approx 1.7$. The polymer thickness was determined by fitting an Airy function to the weak, periodic oscillation observed in the sample absorption spectra. The periodic variation in the absorbance is also caused by the thin *A* étalon.

Multiple étalon, no-electro-absorption model

The previously discussed model ignored the unwanted reflectivity of the polymer/ITO/substrate interface. This reflectivity gives rise to the unphysical oscillatory behavior of the *EA* coefficient s_{13} in Fig. 5a. Étalon *C* was the étalon discussed above. The thickness of étalon *A* is approximately equal to the polymer thickness and the transmissivity of this étalon will remain essentially constant over the narrow angular range probed in our experiments. Therefore, the electro-refractive signal from this étalon by itself remains constant as the sample rotates ±2.5°. By contrast, the electro-refractive signal from étalon *C* contains several resonant peaks as the incidence angle varies. The electro-refractive signal (*C*) switches sign on either side of a resonance and therefore the signal from étalons *A* and *C* will add constructively/ destructively on either side of a resonance. Notice that étalon *B* does not contribute to the modulated signal since the spacer layer (glass) is inactive. The interference of the two electro-refractive signal components (étalons *A* and *C*) gives asymmetric peaks to the lockin signal because of the essentially constant offset contributed by étalon *A*.

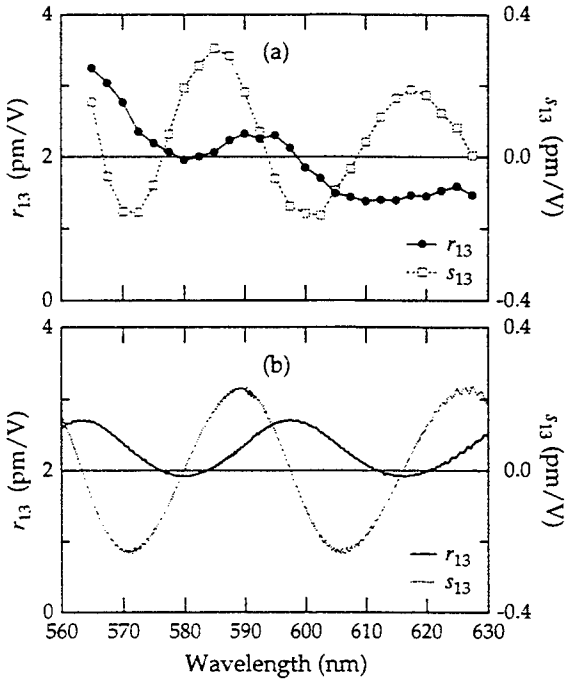


Figure 5) (a) Dispersion of the electro-refractive r_{13} and electro-absorptive s_{13} effects of TP86 measured with a tunable dye laser. Notice that both the electro-absorption and the electro-refraction oscillate periodically. Both components of the total electro-optic coefficient are superimposed on curves that increase at shorter wavelengths. The oscillation behavior is an artifact caused by unwanted surface reflections. The true behavior of the *ER* and *EA* coefficients is contained in the underlying monotonic increase at shorter wavelengths. (b) Result of the simulation described in the text. Simulation data is generated at a single wavelength with the multiple-étalon, no-electro-absorption model and then fit with the single-étalon, *ER/EA*-interference model. All parameters in the generation of the data are fixed except the wavelength, which is varied to determine the simulated dispersion of r_{13} and s_{13}. The field reflectivities are 0.7, 0.08, & 0.2 for the gold layer, the polymer/ITO/glass interface, and the air/glass interface, respectively. The electro-optic coefficient assumed for the simulation is $\tilde{r}_{13}$ = 2 pm/V. Notice that a reflectance as low as 0.6% from the ITO layer gives the appearance of electro-absorption as large as 0.2 pm/V.

Thus, the multiple-étalon interference can give a signal that appears to be electro-absorption but in fact is not. This point has been carefully explored by the authors of refs. 8 & 9 in the context of the ellipsometry/reflection geometry frequently used to measure the electro-optic properties of poled polymers. This effect is probably also the source of the asymmetric electro-optic signals observed in the Mach-Zehnder experiment of Norwood *et al.*[2] Vertically asymmetric peaks in the modulation signal, on either side of an étalon resonance, are also apparent in the data of C. H. Wang *et al.*[15] These authors utilize a Fabry-Perot at normal incidence and scan the wavelength (rather than the incidence angle) to study the étalon resonance behavior.

The importance of this multiple reflection effect must be emphasized: most of the common electro-optic characterization techniques (Fabry-Perot, ellipsometry/ reflection, Mach-Zehnder) for polymer thin films are susceptible to pollution when multiple reflections are present. The popular method of Teng & Man[5] is certainly affected (ref. 7-9) and one should be careful in interpreting the coefficients determined without accounting for this spurious effect. With respect to the dispersion measurements shown in Fig. 5a, near the absorption band (λ = 543.5 nm, Fib. 4b) the signal is clearly dominated by electro-absorption. The unwanted surface reflection seems to add a periodic oscillation to s_{13} with a magnitude of ≈0.2 pm/V. Thus, the EA coefficient is considered unpolluted when s_{13} >> 0.2 pm/V. We conclude that the data shown in Fig. 5a contains both the artificial oscillation from multiple-étalon interference and a real electro-absorptive effect that increases at shorter wavelengths.

We extend the single étalon results to model the coupled Fabry-Perot cavity interference using the method described in ref. 24. First the transmission and reflection coefficients for the last two layers of the sample are determined as if they were the only layers present. The reflectivity of the j^{th} layer (bounded by reflecting surfaces r_j and r_{j+1}) is:

$$r_{j,j+1} = \frac{r_j + r_{j+1}\, e^{i\delta_j}}{1 + r_j\, r_{j+1}\, e^{i\delta_j}} \qquad (8)$$

where:

$$\delta_j = \frac{4\pi}{\lambda} n_j\, L_j \cos\phi_j \qquad (9)$$

and n_j, L_j, & ϕ_j are the refractive index, thickness, and propagation angle in the j^{th} layer. The transmissivity of this layer is:

$$t_{j,j+1} = \frac{t_j\, t_{j+1}\, e^{i\delta_j/2}}{1 + r_j\, r_{j+1}\, e^{i\delta_j}}. \qquad (10)$$

Once the field reflectivity and transmissivity are found for the j^{th} layer, this layer is replaced by an effective surface having the properties dictated by Eqns. 8–10. This process is iterated until all layers are included.[24]

In this model, the thickness of the ITO layer is ignored but its reflectivity is included to give three reflecting surfaces (see Fig. 1a). The polymer is the only electro-optically active layer. Using Eqns. 8-10 and the approximation that $r_2 << r_3 < r_1$, we can show that:

$$\Delta I_{trans} \approx \frac{I_{trans}}{\text{denom}} * \left\{ \frac{r_1 r_2 \sin\gamma}{\text{denom}} + r_1 r_3 \sin\delta \right\} \frac{8\pi}{\lambda} \frac{\Delta n_p\, n_p\, L_p}{\sqrt{n_p^2 - \sin^2\Theta}} \qquad (11)$$

where δ is the same as Eqn. 2, denom is given by Eqn. 1b, and:

$$\gamma = \frac{2\pi}{\lambda} L_p \sqrt{n_p^2 - \sin^2\Theta}. \qquad (12)$$

Note that the second term in braces gives the same electro-refractive signal as Eqn. 5. Étalon A produces the first term in braces. This term gives a vertical offset to the signal since it does not switch sign as quickly with Θ as the $\sin\delta$ term. Thus, the interference of the first term with the second generates the vertical asymmetry of the electro-optic signal. The wavelength periodicity of the $\sin\gamma$ term, given by Eqn. 7, approximately matches the oscillation period of the data in Fig. 5a.

Equation 11 is analogous to Eqn. 5 because it shows the two terms that interfere to produce vertically asymmetric lockin signals. However, the incidence angle (Θ) dependence is slightly different between the two models. We examine the modulation signal data for λ = 612 nm to compare the two angular dependencies. When the TP86 sample is probed with this wavelength, we expect very little contribution from electro-absorption since this wavelength is far from the chromophore absorption edge. However, asymmetric peaks in the lockin signal are observed as shown in Fig. 6. This electro-optic signal dependence is well fit with either the single-étalon, *ER/EA*-interference model (Fig. 6a) or the multiple-étalon, no-electro-absorption model (Fig. 6b). The average transmittance signal is not shown, but the functional fits using either model are also quite satisfactory.

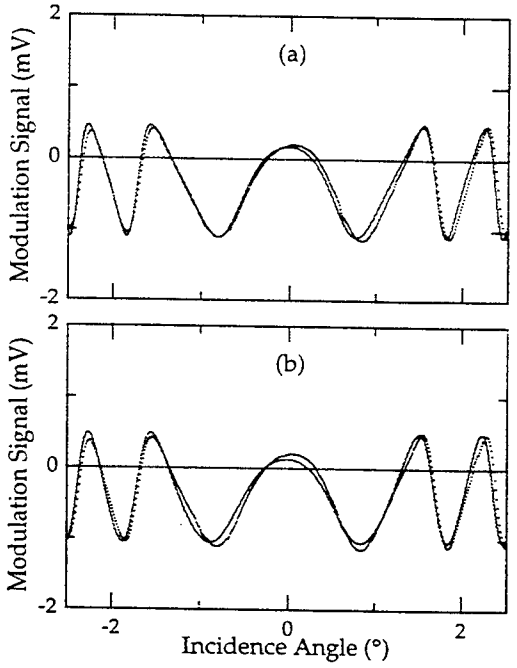


Figure 6) (a) At a probe wavelength λ = 612 nm the lockin signal is unexpectedly asymmetric since the absorption is essentially zero for λ > 550 nm. The asymmetric signal is well fit by the single-étalon, *ER/EA*-

interference equation (Eqn. 5). (b) The same data as in (a) is also well fit by the multiple-étalon, no-electro-absorption model.

We further investigated the similarity between the two models by numerically simulating (at a particular wavelength) the angular dependence of the transmittance signal and the modulation (electro-optic) signal from the multiple-étalon model. This simulation uses the complete multiple-étalon model, not the approximation given by Eqn. 11. The simulated data is then fit with Eqns. 1 and 5 from the single-étalon, *ER/EA*-interference model. The functional fits to the model data for both the average transmittance and the modulation signal are quite good. Our simulation generates new data with the multiple-étalon/no-electro-absorption model for a series of probe wavelengths and fits each of these with the single-étalon, *ER/EA*-interference model. A summary of the results of the simulation is shown in Fig. 5b. The fitted electro-absorptive coefficient s_{13} oscillates with a period of ≈36 nm. This shows that the apparent oscillation of the *EA* coefficient in Fig. 5a above is indeed caused by multiple étalon effects and not by actual variations of the coefficient.

Because both models yield extremely nice predictions of the signals' angular dependence, combining all effects into a single comprehensive model will be quite difficult because of competition to fit the asymmetry. A better approach, currently under investigation, is to increase the finesse of the primary cavity and decrease the importance of the unwanted surface reflection. To reduce the influence of the unwanted polymer/ITO/glass surface reflection, the air/glass interface will be coated with gold thus increasing its reflectivity. A simple alternative is to ignore oscillations in the imaginary part of the electro-absorption coefficient on the order of 0.2 pm/V since this is the magnitude of the spurious oscillation in the electro-absorption coefficient for the surface reflectivities of our samples. Also, when s_{13} is much greater than 0.2 pm/V one can safely conclude that the asymmetry in the lockin signal arises from electro-absorption and is not an artifact of the unwanted reflections. One final possibility is to alter the sample structure to closely match the refractive indices of the various layers. In this case, in-plane poling[17] might be used to eliminate the need for the ITO layer altogether.

Conclusions

We use Fabry-Perot étalons with electro-optically active spacer layers to modulate the transmitted signal via both electro-refraction Δn and electro-absorption $\Delta\alpha$. As we rotate the étalon with a controlled stage the dependence of the modulated light signal determines the complex electro-optic coefficient $\tilde{r}_{13}$. The beauty of this method is its simplicity and the fact that strong electro-absorption (s_{13} >> 0.2 pm/V) is immediately recognized by asymmetric modulation signals with unequal peaks on either side of a Fabry-Perot resonance. Unfortunately, multiple (>2) surface reflections cause spurious effects in our experiments that masquerade as electro-absorption. This multiple étalon effect is rather difficult to avoid since reflectances as low as 0.6% produce variations in the measured parameters. The dispersion of the complex electro-optic coefficient shows both the real increase in electro-refraction and electro-absorption as the wavelength approaches the chromophore absorption edge as well as an artificial variation caused by internal reflections. The unwanted étalon problem can appear not only in Fabry-Perot experiments but also in ellipsometry/reflection and Mach-Zehnder experiments. We regard with caution results that don't properly account for this spurious effect. We are currently working to overcome the multiple-étalon artifact by coating the air/glass interface to increase its reflectivity. In this manner we can increase the finesse of the thick étalon cavity and decrease the interference from the thin étalon.

Acknowledgments

This research was supported by an AFOSR Summer Research Extension Program (grant #95-0868). Bob Gulotty (The Dow Chemical Company) kindly supplied the TP86 sample. We also thank Ken Hopkins, Pat Hemenger, and John Detrio for supporting this research.

References

1. K. D. Singer, M. G. Kuzyk, W. R. Holland, J. E. Sohn, S. J. Lalama, R. B. Comizzoli, H. E. Katz, and M. L. Schilling, "Electro-optic phase modulation and optical second-harmonic generation in corona-poled polymer films," Appl. Phys. Lett. **53**, 1800-1802 (1988).
2. R. A. Norwood, M. G. Kuzyk, and R. A. Keosian, "Electro-optic tensor ratio determination of side-chain copolymers with electro-optic interferometry," J. Appl. Phys. **75** (4), 1869-1874 (1994).
3. V. Dentan, Y. Levy, M. Dumont, P. Robin, and E. Chastaing, "Electro-optical properties of ferroelectric polymers studied by attenuated total reflectance," Opt. Commun. **69**, 379-383 (1989).
4. S. Herminghaus, B. A. Smith, and J. D. Swalen, "Electro-optic coefficient in electric-field-poled polymer waveguides," J. Opt. Soc. Am. B **8**, 2311 (1991).
5. C. C. Teng and H. I. Man, "Simple Reflection Technique for Measuring the Electro-Optic Coefficient of Poled Polymers," Appl. Phys. Lett. **56** (18), 1734-1736 (1990).
6. J. Schildkraut, "Determination of the electro-

optic coefficient of a poled polymer film," Appl. Opt. **29** (19), 2839-2841 (1990).
7. K. Clays and J. S. Schildkraut, "Dispersion of the complex electro-optic coeficient and electrochromic effects in poled polymer films," J. Opt. Soc. Am. B **9** (12), 2274-2282 (1992).
8. Y. Levy, M. Dumont, E. Chastaing, P. Robin, P. A. Chollet, G. Gadret, and F. Kajzar, "Reflection method for electro-optical coefficient determination in stratified thin film structures," Mol. Cryst. Liq. Cryst. Sci. Technol. - Sec. B: Nonlinear Optics **4** (4), 1-19 (1993).
9. P.-A. Chollet, G. Gadret, F. Kajzar, and P. Raimond, "Electro-optic coefficient determination in stratified organized molecular thin films: application to poled polymers," Thin Solid Films **242**, 132-138 (1994).
10. Y. Shuto and M. Amano, "Reflection measurement technique of electro-optic coefficients in lithium niobate crystals and poled polymer films," J. Appl. Phys. **77** (9), 4632-4638 (1995).
11. V. I. Sokolov, D. B. Kushev, and V. K. Subasniev, "Proposed interference method for determining the signs of electro-optical coefficients," Sov. Phys. Crystallogr. **18** (2), 200-201 (1973).
12. H. Uchiki and T. Kobayashi, "New determination of electro-optic constants and relevant nonlinear susceptibilities and its application to doped polymer," J. Appl. Phys. **64** (5), 2625-2629 (1988).
13. C. A. Eldering, A. Knoesen, and S. T. Kowel, "Use of Fabry-Perot devices for the characterization of polymeric electro-optic films," J. Appl. Phys. **69**, 3676-3686 (1991).
14. R. Meyrueix, J. P. Lecomte, and G. Tapolsky, "A Fabry Perot Interferometric Technique for the electro-optical characterization of nonlinear optical polymers," Nonlin. Opt. **1**, 201-211 (1991).
15. C. H. Wang, B. S. Wherrett, J. P. Cresswell, M. C. Petty, T. Ryan, S. Allen, I. Ferguson, M. G. Hutchings, and D. P. Devonald, "Observation of electro-optic and electroabsorption modulation in a Langmuir-Blodgett film Fabry-Perot étalon," Opt. Lett. **20** (14), 1533-1535 (1995).
16 J. P. Cresswell, M. C. Petty, C. H. Wang, B. S. Wherrett, Z. Ali-Adib, P. Hodge, T. G. Ryan, S. Allen, "An electro-optic Fabry-Perot through-plane-modulator based on a Langmuir-Blodgett film," Opt. Comm. **115**, 271-275 (1995).
17. M. Ziari, S. Kalluri, S. Garner, W. H. Steier, Z. Liang, L. R. Dalton, and Y. Shi, "Novel electro-optic measurement technique for coplanar electrode poled polymers," in *Nonlinear Optical Properties of Organic Materials VIII, ed. G. R. Möhlmann*, SPIE **2527** (1995).
18. J. L. Stevenson, S. Ayers, and M. M. Faktor, "The linear electrochromic effect in meta-nitroaniline," J. Phys. Chem. Solids **34**, 235-239 (1973).
19. R. H. Page, M. C. Jurich, B. Reck, A. Sen, R. J. Twieg, J. D. Swalen, G. C. Bjorklund, and C. G. Willson, "Electrochromic and optical waveguide studies of corona-poled electro-optic polymer films," J. Opt. Soc. Am. B **7** (7), 1239-1250 (1990).
20. A. Horvath, H. Bassler, and G. Weiser, "Electroabsorption in conjugated polymers," Phys. Stat. Sol. B **173**, 755-764 (1992).
21. F. Qiu, K. Misawa, X. Cheng, A. Ueki, and T. Kobayashi, "Determination of complex tensor components of electro-optic constants of dye-doped polymer films with a Mach-Zehnder interferometer," Appl. Phys. Lett. **65** (13), 1605-1607 (1994).
22. H. Ono, K. Misawa, K. Minoshima, A. Ueki, and T. Kobayashi, "Complex electro-optic constants of dye-doped polymer films determined with a Mach-Zehnder interferometer," J. Appl. Phys. **77** (10), 4935-4940 (1995).
23. K. D. Singer, M. G. Kuzyk, and J. E. Sohn, "Second-order nonlinear-optical processes in orientationally ordered materials: relationship between molecular and macroscopic properties," J. Opt. Soc. Am. B **4** (6), 968-976 (1987).
24. O. S. Heavens, *Optical Properties of Thin Solid Films,* (Dover Publications, New York, 1965).

A METHODOLOGY FOR AFFORDABILITY IN THE DESIGN PROCESS

Georges M. Fadel
Associate Professor

Charles Kirschman and Pierre Grignon
Graduate Students

Mechanical Engineering Department
Clemson University
202 Fluor Daniel Engineering Innovation Building
Clemson, SC 29634-0921

Final Report for:
Summer Faculty Research Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

December, 1995

A METHODOLOGY FOR AFFORDABILITY IN THE DESIGN PROCESS

Georges M. Fadel
Associate Professor
Mechanical Engineering Department
Clemson University

ABSTRACT

With the decrease in availability of funds and the tighter economy, the emphasis on the issue of affordability in design is becoming of critical importance, especially for weapons systems. Unfortunately, the cost issue in early conceptual design has been given little consideration. Yet, it is this consideration of affordability at the conceptual design stage that provides the highest leverage, since design modifications become more and more costly as the design evolves and the design freedom is reduced. This fact does not deter from the need to carry the issue of cost versus performance – which may be considered as a synonym of affordability – throughout the design process. Methodologies to identify technologies that drive the cost (cost drivers) are needed, as well as methodologies that identify the performance drivers. This research builds upon the work performed by the P.I. during his summer faculty research sabbatical at the Wright Labs in the summer of 1994. It expands the idea of design metrics and applies them to different stages of the design process based on functional decomposition. The main targets of the work aim at identifying metrics that can provide some measure of the flexibility of a design, as well as its relative value.

A taxonomy for functional decomposition of mechanical components is described. This decomposition is used as a basis on which to apply metrics. Three metrics are proposed: ***Please***, which includes meeting a performance level or exceeding it, ergonomic issues, maintainability and possibly recyclability issues; ***Protect*** which encompasses all issues related to safety of the user, the environment, etc.; and ***Icost***, a synonym of affordability which in our application is the inverse of the cost of a product. The application of these metrics at conceptual stage – essentially derived from knowledge and experience – and later in the design, according to more formalized methods based on the utility theory, are described. A software system that encapsulates the ideas is detailed, and additional work is recommended to evaluate and assess the method proposed.

INTRODUCTION

The financial resources available in our country for defense purposes have been reduced significantly in recent years. Yet, the demands on our armed forces keep increasing, and weapons systems need to be kept at the forefront of technology. For this reason, the design of new systems which originally was driven by performance while neglecting cost, is being re–examined. The emphasis on the issue of affordability is becoming of critical importance. Unfortunately, the cost issue in early conceptual design has been given little consideration. Yet, it is this consideration of affordability at the conceptual design stage that provides the highest leverage, since design modifications become more and more costly as the design evolves and the design freedom is reduced. This fact does not deter from the need to carry the issue of cost versus performance – which is directly related to affordability – throughout the design process. Methodologies to identify technologies that drive the cost (cost drivers) are needed, as well as methodologies that identify the performance drivers. This research builds upon the work performed by the P.I. during his summer faculty research sabbatical at the Wright Labs in the summer of 1994. It expands the idea of design metrics and applies them to different stages of the design process based on functional decomposition. The main objectives of the work were to develop a set of metrics that can be used at the conceptual level, when the amount of information about the design is very small, and then revisit the metrics when more information becomes available.

PREVIOUS WORK

The Air Force has defined affordability in its Science and Technology Affordability White Paper [S&T, 93] as follows: "*A Technology is considered affordable if it meets the customer's requirements, is within the customer's budget, and has the best value among available alternatives.*" The keywords in this definition are technology, customer requirements, budget, value and alternatives. In order to access the affordability of a technology, we must therefore consider its life–cycle costs, its ability to meet the customer's requirements, and be able to compare it to alternative technologies. To determine life–cycle costs, we need tools, rules and metrics in order to evaluate the technology during the initial development of some system where the leverage is the highest, or later, during preliminary design when form and function are defined. The technology for cost reduction will result from the ability to translate experience and heuristic information into models that illustrate the sensitivity of cost, performance and environmental impact to a design parameter.

Raymer [Raymer, 92] has devoted one chapter in his book to the cost analysis issue in aircraft design. Life cycle costs, including research development test and evaluation costs, operation and maintenance costs, and various cost estimating methods are mentioned. In particular, Raymer expands on DAPCA IV or Development and Procurement Cost of Aircraft (Rand Corporation). He lists several equations in constant 1986 Dollars that relate the various cost components to weight and other characteristics of the airplane. They are the only cost estimators identified in the literature survey that have been derived for use at the *conceptual* design stage (It is assumed that aircraft manufacturers have their own methods, but the author did not have access to any such information). It is interesting to note that these equations relate cost in a near linear way to weight, velocity, thrust and other performance measures. However, intuitively, if we graphically relate

performance to cost, we should obtain curves such as displayed in Figure 1. One such performance-cost curve should show when a new technology is needed. The figure describes how a technology becomes more and more costly when more performance is needed. At a certain point, there is a need to move to another technology that might be more or less affordable, but that allows an increase in performance. Where is the point at which a new technology has to be introduced? Can we generate these curves from the information at hand and estimate at what slope a new technology needs to be introduced? Are the weight, thrust, and number of aircraft the correct performance measures to use at the conceptual stage? Should instead increased capabilities be the performance measures, and the expected performance gains versus cost for development and maturation be the affordability metric? These are questions that need to be resolved.

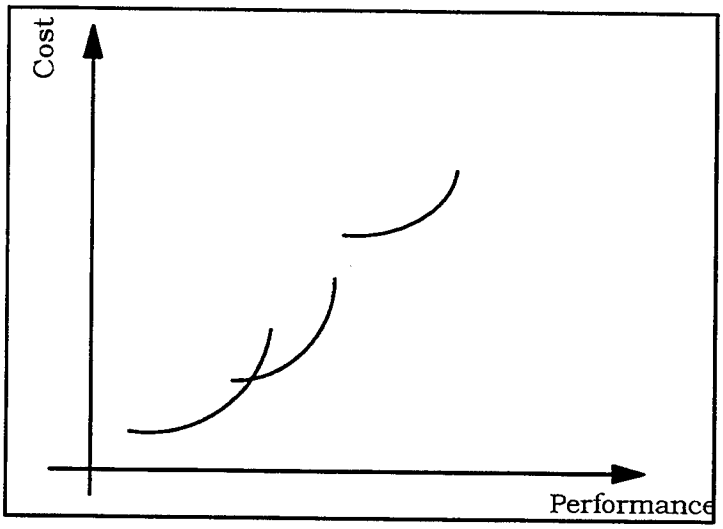


Figure 1. Performance – Cost graph

The issue of *costing* later in the *design* stage has been implemented in many methods. The Boothroyd-Dewhurst design for assembly method (DFA) [Boothr, 90] considers assembly issues and qualitatively links cost to the number of parts. The DFA also includes some measures of simplicity of design that directly affect cost. Another less known methodology dealing with assembly issues is the Hitachi Assemblability Evaluation Method or AEM. This proprietary method is based on the selection of assembly element symbols and their combination which results in a numerical rating [Boothr, 88]. Cognition [SDRC, 94] implemented an expert system based component costing approach which is provided either with Cognition's ACIS based performance modeler Mechanical Advantage II, or with SDRC's I-DEAS solid modeling package. In both cases, the software allows for customization and has an extensive database of process costs that result in a quantitative figure at the design stage (This database needs further development since only a few manufacturing processes are presently commercially available.) Costing databases need to be continuously updated as labor rates and material costs change, and as new manufacturing processes become available. Note that this method depends on a knowledge of the form of the design, and therefore cannot be used at the conceptual stage.

Ed Dean, from NASA Langley, has extensively studied the issue of cost in design, especially related to the NASA mission. He states [Dean, 89] that a measurable relationship exists between

system attributes and the cost of the system. He then identifies the cost drivers to be the attributes of a function and the requirements or constraints of the design, and corrects the weighted sum of the metrics by using an exponential factor to normalize for temporal effects, and a power factor to normalize for economic quantities. In other subsequent papers [Dean, 91][Dean, 92][Dean, 95] and [Dean, 96], he explores elements of designing for cost, function cost analysis, parametric cost analysis and value engineering.

Other costing related research at the process level is abundant. Many textbooks exist that deal with the cost of manufacturing parts [Trucks, 74], [Gunthe, 71], [VDI, 72], and papers can be found that treat most processes. For instance, Poli, Escudero and Fernandez [Poli, 88] describe the relationship between part complexity and molding tool costs; Knight and Poli [Knight, 85] present a systematic approach to the producibility and cost of forging design; Gutowski et al [Gutows, 94] propose a theoretical cost model for advanced composite fabrication. All these models can be very helpful to obtain a first estimate of a process cost. They usually presume that you know the final form of your design, and often do not include tooling costs, availability of tools, etc.

Thus, at the conceptual level, very rough estimates of cost could be derived using previous experience that relate some performance levels such as weight and thrust to cost. These estimates rarely result in even orders of magnitude correct costs when novel technologies are involved; instead they are more appropriate for mature and well understood technologies. Later in the design process, at the component level, if the form or shape is known, cost information can be derived but only if data has been gathered and made available through a costing database. Some of the issues that need work are standardization, uniform costing procedures, and access to suppliers' cost models. What is lacking is a better cost model for putting components together such as assemblies. Boothroyd and Dewhurst [Boothr, 90] developed a set of rules for assemblies intended to reduce costs. These rules include the minimization of part count, the avoidance of threaded fasteners, the avoidance of flexible elements, the design for automation, the standardization of dimensions. When a machine such as a printer is designed, many of these rules significantly affect the overall design. However, in the design of a weapon system such as an aircraft, many of these rules are difficult if not impossible to implement at certain levels because of the complexity of the systems.

How to handle the decision making process to minimize cost in complex designs is a methodology that has to be formalized at the different granularity levels (Assembly to subassemblies to components or parts). Cost has to be tied to producibility, support, recyclability in order to present the designer with a complete image of the impact of the design. Even the cost modelers available (Cognition's for instance) will not easily adapt themselves to what-if scenarios. Engineers have to perform multiple runs of the software considering different manufacturing alternatives and comparing them. Furthermore, the software is not at a level where it can assist the engineer in the decision making process. Note that the present feature-based modelers can be customized to help derive cost at the embodiment design stage, the stage where the form of the design is chosen. However, if different processes are considered for manufacturing, then different features will be characterized. For instance, in a molding process, a rib is a feature witch is inserted to add support and has some thickness and height characteristics, whereas in a machining operation, the material around the rib must be removed, and the rib may not even be mentioned as a feature.

Another major concern when considering cost is the identification of relationships between features and parameters. In a car for instance, one part can be affected by the geometry or behavior (temperature or vibration) of roughly 1200 to 1800 parts [Sferro, 94]. Therefore, the modification of a part or the machining or tolerancing of a part, could affect a number of others and alter a cost equation. These relationships are investigated and established from the customer's point of view in Quality Function Deployment (QFD) houses of quality. These QFD matrices allow the designer to identify how some function or parameter affects the perception of quality for a customer. These perceptions are uniquely qualitative and do not consider cost or other metrics. It might be advantageous to consider houses of quality that incorporate cost, maintainability or other design parameters that provide the designer with feedback on relationships other than quality. Still, it is the task of the engineer to identify the sensitivities of cost or maintainability to a change in an engineering characteristic, and there is no formalized method to derive all these relationships and ensure that no significant one is overlooked.

As mentioned earlier, affordability entails the evaluation of a technology and its comparison to its competitors. For such a comparison, some measure of the producibility of a technology is needed. Producibility in engineering design is affected by three main factors: *design, process* and *materials.* In *design*, the issues of manufacturability, simplicity, maintainability, recyclability, and ease of assembly are some critical measures that indicate the level of producibility of a product. These measures apply to either the design of components (manufacturability, recyclability and simplicity), or to assemblies (ease of assembly, maintainability). When considering the *process*, tolerances, variability, labor, equipment or tooling requirements are significant issues and metrics; and when considering the *materials*, their selection, properties and links to the design issues are independent variables that affect both sets of metrics listed earlier for the *design* and the *process.*

The issue of *producibility* is investigated by Motorola in its "Six Sigma Approach" [Harry, 92], and an implementation is described in Texas Instruments' design for producibility of active array radar modules [TI, 93]. The approach ties the *process* variability to *producibility* in terms of expected yield. It is based on statistical Process Control (SPC) or on estimates of expected defects of processes. It seems to be presently applied mainly to processes related to circuit board manufacturing and should be investigated for other manufacturing processes. The Six Sigma approach allows quantification of the degree to which a *process* is under control, and the goal of the method is to reduce the number of defects in a product at production time.

The Taguchi method [Taguch, 93] stresses quality and relates the design parameters and *process* characteristics to the quality of a component. In this method, statistical measures are used again to estimate some deviation from a desired value and a loss function is used to quantify the quality of a part by translating the loss of quality to some equivalent added cost to the customer.

Both methods address the producibility at the component manufacturing level. The Boothroyd–Dewhurst method introduced earlier deals with assembly producibility questions and relates that producibility to cost.

Thus, costing techniques are either available, or can be developed for materials and processes. Producibility of designs can be assessed if the form and material are known, but the issue of affordability at the conceptual design stage is still not resolved.

Considering metrics more generally, a myriad of research targets the selection of the "best" design. Most generally, Otto [Otto, 93] describes measurement theory and its applications to design. He concludes that measurement theory is quite general and can be applied to design. Pugh [Pugh, 91] describes a comparison technique where one alternative is chosen as the datum and other alternatives are qualitatively compared to it. This technique is applied at the conceptual stage to choose which concept to pursue. Ullman [Ullman, 92] restates this work, and adds that it is important to use both criteria and alternatives that are at the same level of abstraction. He recommends using the customer's wants and needs developed in the House of Quality with the abstract concepts present in conceptual design.

Several researchers have focused on one aspect of a design and shown techniques to improve a design with respect to that one performance criterion. These "Design for X" strategies are important to include the downstream issues early in the design process. Boothroyd and Dewhurst [Booth, 88] mentioned earlier describe techniques to measure designs based on the cost of assembly. Similarly, Ishii and Eubanks [Ishii, 93] describe a Design for Life Cycle philosophy as well as a system to implement this philosophy to mechanical systems. Whitney [Whitne, 88] describes a case study illustrating the cost savings of "Design for Manufacturing." In all of these cases, the designs are measured with respect to only one aspect of the downstream processes.

Otto and Antonsson [Otto, 91] describe a design strategy where the designer balances between conservative and aggressive design. In the former, a weak aspect of the design is improved, possibly to the detriment of a stronger aspect. In aggressive design, improvement in an already strong area is sufficient to compensate for weaknesses elsewhere. They also discuss the use of preferences in a design to explicitly state the trade–off strategy. This permits some elasticity in the design parameters and allows the designer to incorporate subjective data.

Thurston [Thurst, 91] has pursued several aspects of design measurement. The basis of her work is the multi–attribute utility theory which, in this context, allows the designer to emphasize preferences in certain attributes. These preferences are used to construct a design evaluation equation. This equation is then used to determine the net change in a design as the levels of the individual attributes change. Utility theory is used to help the designer choose between finding a single alternative with certainty or choosing a lottery with certain chances of better or worse outcomes. This determines the willingness of the designer to make trade–offs between attributes.

Diteman and Stauffer [Ditema, 93] have also explored multi–attribute utility theory in design concept selection. A set of dimensions are preference ranked in pairs to choose the relative ordering of the dimensions. For example, the designer might rate "performance" over "aesthetics", but rank "safety" over "performance". In this way, an ordered list of dimensions is produced, The utility theory is used within each dimension to rate the criteria with the weights totalling 100%. Because the preferences can vary, the same criteria can be used for many different problems.

Another aspect of Diteman and Stauffer's [Ditema, 93] research is that they have compared an Absolute Rating Method (ARM) to a Relative Comparison Method (RCM) through experimentation. The difference between the two methods is how many alternatives are considered at once. In the ARM, a single alternative is ranked based on its ability to meet a particular criteria; in the RCM, all alternatives are ranked from best to worst. The authors conclude that the ARM is more accurate because it is more consistent across designers, but less sensitive to small variations than the

RCM. They also postulate that the RCM is not a good technique when there are more than three or four alternatives.

A different form of evaluation is presented in the Quality Function Deployment (QFD) work (Hauser and Clausing [Hauser, 88], Sullivan [Sulliv, 86]). The advantage of QFD is that it attempts to apply the eventual customer metric to the design in the early stages. In this work, a "House of Quality" is developed which relates what the customer desires (Customer Attributes) to the Design Requirements that the engineers use. The ratings are typically presented as "Strong Effect", "Moderate Effect", "Weak Effect", and "No Effect". This technique shows areas of weakness in a design as compared to competitors products, as well as which engineering factors to change to gain the most improvement in customer satisfaction.

Locasio and Thurston [Locasi, 94] have shown one way to use the House of Quality to generate quantitative values that can be used in design. Their technique is based on the optimization of equations found by applying the multi-attribute utility theory to the House of Quality. Suh [Suh, 90] describes how to use the House of Quality to determine the Functional Requirements for a product. Bascaran and Telez [Bascar, 94] describe a technique for extending the link between QFD and Axiomatic Design strategies. The customer attributes are divided into Functional Requirements which should have matching Design Parameters and Market Requirements. In addition, Quality Characteristics are added in what they call the "Voice of the Engineer" section. All of these techniques have used and built on the concept that the customer's needs and wants be directly used in the design process. However, QFD requires a lot of market research which is difficult to do at the conceptual stages of design. QFD is better suited to redesign than original design [Suh, 90].

[Fadel, 94] presented a method to deal with affordability at the conceptual level. The method, which is based on the identification of cost drivers in the house of quality, is intended to help the designer make decisions very early in the design process. Although it does not guarantee "affordable" designs, it helps the thought process of the designer and allows him or her to compare alternatives before embodiment design is initiated. As mentioned in the report, such a method needs to be used in tandem with affordability tools later in the design process. In particular, tools such as the DFA of Boothroyd Dewhurst, Cognition's cost estimator, and producibility tools such as Six Sigma and Taguchi have to be used extensively to further control quality and cost.

One suggestion of the report is to consider developing a functional decomposition of the design, and to associate to it some metric which would help in assessing cost and performance. This research develops this idea, and, through the development of a type of taxonomy for functional design, establishes a set of metrics that can be used to assess the design.

In the next section, the proposed methodology to deal with affordability as one of the metrics at the design stage is detailed and applied to sample mechanical problems.

PROPOSED METHODOLOGY

Products are defined by their function. For this reason, a designer typically starts with an objective which is a functional description of a product. This primary functional objective is then broken down into several descriptions. Sub–systems (assemblies of physical parts or components) are selected to perform the specific sub–functions. This process results in a hierarchy of functions that the designer can use to develop the final product.

Each of these sub–functions can become a root function for a given sub–system. The design process can then be applied to this sub–system in a manner that is mostly independent of the overall design, and each of the current level sub–system main functions can be partitioned in a hierarchical manner into smaller subfunctions at lower levels. However, this process must stop at some point. A logical stopping point in the decomposition is when the designer reaches an *elemental mechanical function* [Kirsch, 96]. An elemental mechanical function may be defined to be a function that has a specific form associated with it, as opposed to one that must be broken up into smaller forms to accomplish the function. For example, an electric motor is a standard mechanical component commonly used in designs. A designer should not try to partition the function of a motor, "Create Rotary Motion", into its sub–functions since the finished design will only call for the motor, not the components of the motor.

Additionally, a very low level function may only be accomplished by one form, while a higher level function gives the designer more freedom to manipulate the design and generate new ideas. For instance, consider the case of the simple gear train shown in Figure 2. At the highest level, the function of the gear train is to *transform rotary motion.* The sub–functions that can be derived are *transmit torque* for each gear, *support* for each shaft, and *connect* each of the shafts to maintain relative distance. To make any improvement, the designer would need to consider the base functions and try to improve each individual gear, a difficult task.

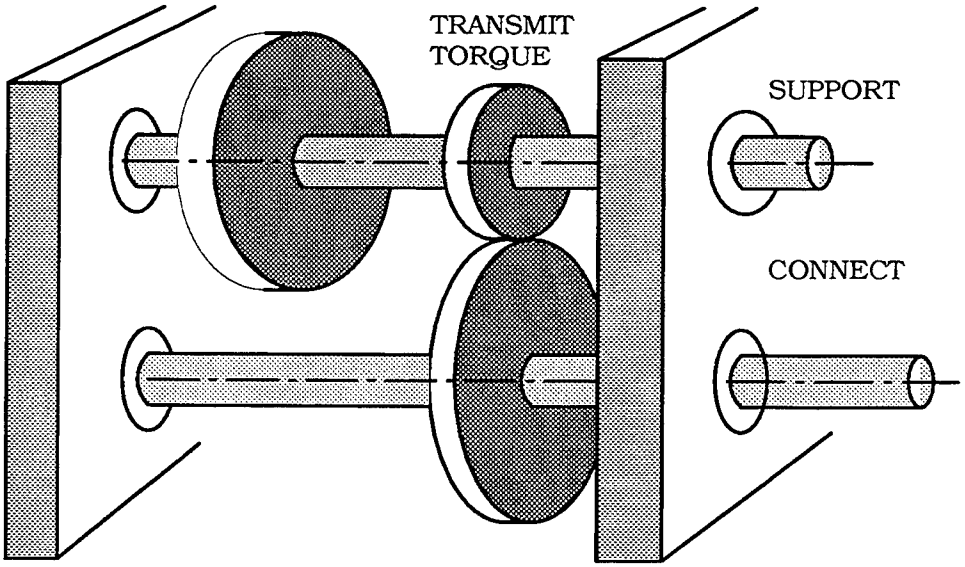


Figure 2. Simple Gear Train

Now consider the case where the entire gear train is one function, *transform rotary motion*. In this case, improvements become more obvious. For instance, the gear train could be replaced by a planetary gear system. Other possibilities are a harmonic gear or pulleys and belts. Each of these components transforms rotary motion, decreasing or increasing velocity while increasing or decreasing torque. However, these alternatives would not be obvious at the lowest functional description level where only gears could be substituted, not the whole assembly.

The elemental functional description is best suited to the selection of the components used in the design. The form used for support and enclosure is dependant upon the components, and is usually chosen after the component selection is complete. This is not to say that the enclosure will not impose constraints on the design of the components; rather that the final form of the enclosure must follow from the collection of components selected.

Function–Based Taxonomy

After analyzing work by Collins and coauthors on helicopter design [Collin, 76], and considering consumer products such as an automobile, a drill, a lawn mower, four basic types of functions were derived. These are related to the concepts of ***Motion***, ***Power/matter***, ***Control***, and ***Enclosure***.

From these four basic types of functions, a taxonomy is proposed. Most (if not all) parts of a mechanical design can be classified into these four categories. These seem to describe the "basic" function of a subsystem. The categories are shown in Figure 3 below. The triangle symbolizes the

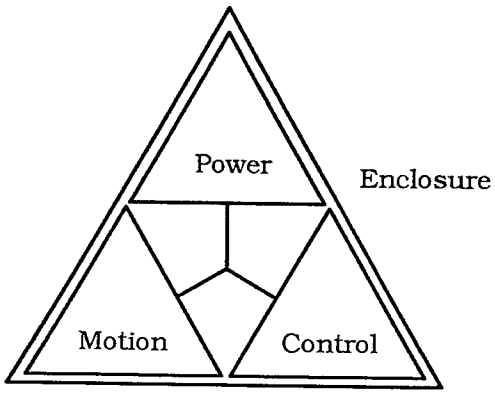


Figure 3. Four Categories of Mechanical Function.

equal importance of each of the functions, and their relations to each other. It describes how the enclosure function is the geometric aspect of the design that ties the different components together to accomplish the main design function.

This classification system provides a good tight description of the basic functional subsystem to which a part belongs. However, it is not sufficient to describe the overall function of a part or the design intent. Once again, Collins work was studied to extract different connotations for each functional group.

The classification system is then extended to cover more descriptive mechanical functions, as shown in Table I below. For instance, power is commonly available in electrical, mechanical, or

chemical forms. It can be created, stored, supplied, transmitted, and dissipated. These concepts form the basis of the power function set. Similar lists are developed for motion, control, and enclosure.

Table I. Basic Function Groups.

Control	• Power, Motion, Information • Continuous, Discreet • Modification, Indication • User–supplied, Internal Feedback
Power / Matter	• Store, intake, Expel, Modify, Transmit, Dissipate • Electrical, Mechanical, Other
Enclose	• Cover, View, Protect • Removable, Permanent • Support, Attach, Connect, Guide, Limit
Motion	• Rotary, Linear, Oscillatory, Other • Create, Convert, Modify, Dissipate, Transmit • Flexible, Rigid

After establishing the verb–adjective combination, the information content is still insufficient; for example "transmit electrical power" needs a direction, such as "to heat". Therefore, several "directions" are proposed. Now, all of the ideas above plus the directions can be converted into sentence form as shown in Table II.

These sentences lead to about 150 combinations of elemental mechanical functions. It should be noted that the phrases in brackets above do not apply to all of the other options; generally they apply to one or two of the possibilities.

This classification system provides a simple and direct way of choosing a function for a design. It should be noted that objective functions are different from these elemental functions since they are higher level. Objective functions encompass several elemental functions. An example of this is the function of a drill, which is designed to "Drill Hole". This function is not on the above list; it is a high–level function that is the sum of the many lower functions of the individual sub–components of the drill: battery for power storage, motor for creation of rotary motion, control mechanism, and enclosure.

This decomposition technique is beneficial at the initial stages of design and as a teaching tool. The hierarchy shown in Figure 4 takes into account the four types of functions, since almost all mechanical designs incorporate them. Matter storage and transport is not included in the initial diagram since it is not as widely applicable; however it can be easily added. Control is applied separately to each function rather than being treated as a separate and independent function. The power–control function combination could be accomplished by a switch for instance, while the motion–control function combination may be obtained using a brake.

Table II Elemental Mechanical Functions

Create Convert Modify Transmit Dissipate	Rotary Linear Oscillatory Other	**Motion**	[to	Rotary Linear Oscillatory Other	Motion]	.
Store Intake Expel Modify Transmit Dissipate	[Electrical Mechanical Other]	**Power/ Matter**	[to	Control Heat Move]		.
Continuous Discrete	User Feedback	**Control** of	Power Motion Information	resulting in	Modification Indication	.
Support Attach Connect Guide Limit Cover Protect View	Part X which is	Moving Stationary	[to Part Y which is	Moving Stationary]	in a Removable Permanent	Manner

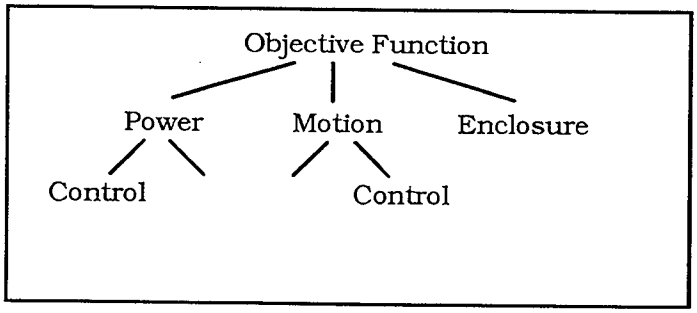


Figure 4. Functional Hierarchy for Mechanical Design

In summary, a decomposition method for Function-based design is proposed. The "taxonomy" is derived from four major functions that describe mechanical designs. These are the motion, power, control and enclosure functions. This technique that allows the designer to hierarchically decompose designs starting from the main objective of the product which is the top level function should be tested on aeronautical designs. Using the method, the designer is given the freedom to

select appropriate functions and their form counterpart. The form tree would parallel the function tree, conforming to the independence axiom proposed by Suh [Suh, 90]. Should a single function accomplish more than one function, its multiple occurrence in the form tree would indicate to the designer which functions are affected by the particular form and would prompt the designer to reflect on possible changes that may affect more than the targeted modification. The ability to stop at the elemental function level provides added robustness and a broader selection of components to accomplish a specific function. Once a design is decomposed and forms selected to accomplish the functions, the next questions are how do we evaluate the designs, and what are the metrics that we could use.

METRICS

The representation of the "Value" of a design in the design space is a difficult question. The primary question is "How do we measure the value of a design?" This question may be very difficult to answer in an absolute sense. However, in a relative sense, designs can be compared. When presented with two competing products, a designer typically knows how one design is better than the other. One may be much stronger in performance, another may be very inexpensive to build. It is extremely difficult to declare that a design is the "best"; however, it is often possible to point out that one is better in some respects than another.

To further complicate things, consider the statement from Fowler (1990), that "the value of anything is variable, depending on environment and attitude". This implies that if two designs are compared at two different points in time, the relative values can change. Consider comparing an American and a Japanese automobile just before and just after the start of the oil embargo in the 1970's. When gas was cheap, the American car was more valuable with greater luxury and more power. But after the dramatic rise in gasoline prices, the foreign vehicles were sought because of the low operating cost.

The choice of metrics for the value axis is difficult. The value of a product is subject to change over time, and is different depending on who is evaluating it. When considering the various voices involved in the design, the following set evolves: Performance is the *Voice of the Customer*. This is how well the product performs the functions required of it. Ergonomics are also customer driven, since they interface with the product. Safety is the *Voice of Responsibility*. Engineers have ethical as well as legal responsibilities to protect the public. Another aspect of the design is producing it; typically this is called Design for Manufacturing (DFM). DFM is used here to mean all of the DFX theories that pertain to manufacturing of the product. In this sense, the 'X' can mean Assembly, Forging, Casting, etc., it is the *Voice of Manufacturing*. Finally, Icost is defined as the inverse of cost. Icost can be thought of as the *Voice of Management*, who always want the product to cost less to produce. Naturally, all the voices are also overlapping, the customer for instance, is concerned with product safety and maybe recyclability. Cost affects everyone. So, how does one measure value?

When comparing the effects of a change in one metric, a table such as Table III below can be developed. This table shows the effects of increasing one of the five metrics on the others. The notable values are the 0's, which occur when the change in one attribute does not affect the other one. This means that the orthogonal metrics can be viewed as perpendicular, as shown in Figure

5. Other metrics are placed at some angle with respect to each other. These angles depend on the design, and the figure illustrates a change in the design, with the center of the circle representing the current design. The arrows pointing outwards signify increase in a particular metric. Thus, for a product on the market, value is increased by moving away from the origin in the figure below (Figure 5). The comparison of two products from different manufacturers cannot however be accomplished with this representation scheme.

Table III. Effects of Increasing One Metric on Other Metrics

	Performance	DFM	Ergonomics	Safety	Icost
Performance		↓	0	↓	↓
DFM	↓		↓	0	↑
Ergonomics	0	↓		↑	↓
Safety	↓	0	↑		↓

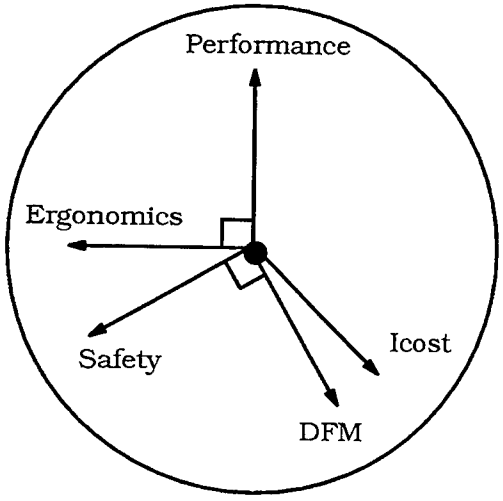


Figure 5. Five metrics.

The perpendicular metrics could be merged into a single metric, or all the metrics can be separately considered. In the case where they are merged, performance and ergonomics form the Pleasure metric. Similarly, DFM and safety can be merged to form the Protection metric, although this is not a very good name. These merged metrics will form the base of the following work without loss of generality.

Thus, in order to help select the more affordable, safer and more pleasing component from those available to the designer, three metrics are proposed: they are ***Please***, ***Protect***, and ***ICost***. The first two are based on work done by Ouellette [Ouelle, 92], and the third was added as a balance because of its necessary role in the quest for affordable designs. These measures are sufficiently general as to be applicable at the earliest stages of design. In fact, these metrics also help to bring the voice of the customer into the decision process at a very early stage, before all formal design parameters are chosen. The metrics are also applicable later in the design process, and

can be used to evaluate a finished product. In later stages of design, measures such as maintainability can be considered part of both please and protect, to different degrees. Application of these metrics later in design is the topic of continuing research .These metrics can be described as follows:

The first metric, ***Pleasure***, refers to the way that a product increases a customer's satisfaction. Admittedly, this is a fairly nebulous goal, as different people have different tastes. For instance, consider the vast array of available automobiles. Some people are happy with a small sporty car, while others prefer a larger sedan. A company's goal is to sell products, and to do so the product must please the customer. Therefore, the burden falls on the designer to determine what the customer wants, and to appeal to a particular chosen customer base. Techniques such as interviews and focus groups [Ullman, 92] can help in this regard, as they do in QFD.

Pleasure has many different aspects as it pertains to design. The primary requirement is that the product must meet performance requirements. If it does not, then it is not a successful product. However, many products exceed the minimum performance levels and provide added features which increase the customer's satisfaction with the product. Fowler [Fowler, 90] states that "supporting functions which are not essential to the performance of the main function are essential to increase product acceptance by satisfying the wants of the user." He also states that supporting functions are typically intangible, but account for 75% of the cost of the product. These supporting functions invariably form the basis for the users decision to buy a product according to Fowler. The four functions that Fowler uses to describe this are *assure dependability, assure convenience, enhance product,* and *please senses.*

Another important element of the pleasure metric is "human factors" such as the user interface and ergonomics. Furthermore, such things as maintainability, reliability, and recyclability, to name a few, may have a great impact on the customer's pleasure with a product. Trying to compare a product or even a component of that product based on all of these factors is extremely difficult at an early stage of design. It is much easier to think about the customer's pleasure directly when comparing alternatives.

The second metric, ***Protection***, is a measure of the degree of safety present in the design. Note that while a certain level of safety may be mandated by law, additional measures may contribute to customer satisfaction. For instance, most current cars come with air bags for the driver. However, some manufacturers also add air bags for the passenger, special crash–resistant frames, and space frames in the doors. The former is a protection presently required by law, while the latter is a collection of protection features that make a product more desirable. Again, the amount of value that these features add is dependant on the particular customer. The designer is responsible to ensure that the product does not open the company up to liability; it is also his/her ethical responsibility to ensure the product is safe.

The term ***icost*** was chosen because, as a metric, an increase is easier to graphically display. Icost is an inverse function of cost, which will increase as cost decreases. The cost value is a not the consumer price, since that is typically out of the engineer's hands. Instead, this cost value is the price of the manufactured product. This is the cost of materials plus processing for a manufactured part, or the component cost for a part that will be purchased from another manufacturer. In the case of the latter, the cost for a reasonably large number of components should be considered

based on the application and predicted sales. These costs will usually have a direct impact on the consumer price, but they are more measurable by the engineer.

It can be said that these three metrics are balanced because typically a design cannot be improved in one of these areas without hurting at least one of the others. This is because the measures are not orthogonal, as shown in Figure 6. The angles between the three axes are not constant but will change depending on the particular design. Because the metrics are not orthogonal, some of the goals of two of the metrics may be coincident, such as serviceability which will increase both please and protect.

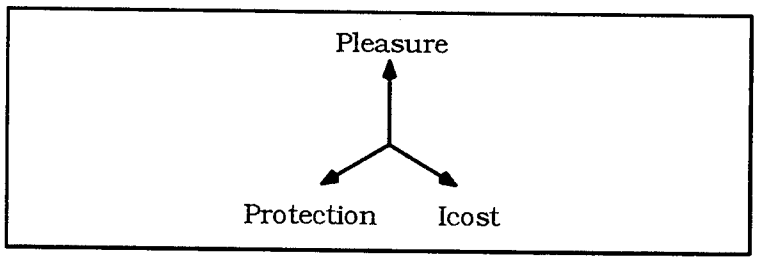


Figure 6. Pleasure, Protection, and Icost as Balanced Metrics.

As a simple example of these metrics, consider purchasing a vehicle. The primary function of the vehicle is to transport, and the consumer has already provided the constraints that it must carry multiple passengers and luggage over long distances in any weather. These constraints have narrowed the field away from skateboards, ice skates, bicycles, etc., and into the wide array of cars and similar vehicles. If the buyer is strictly concerned with pleasure, they will most likely buy an exotic sports car that travels fast and looks very nice or an extremely luxurious car. If their only concern is safety, they may buy a gunless tank or similar vehicle which can survive almost anything. However, if cost is the primary concern, then the consumer will buy a Yugo or similar vehicle (probably used) that will be very economical. Thus in summary,

- The first metric, ***Please***, refers to the way that a product pleases the customer. Admittedly, this is a fairly nebulous goal, as different people judge products according to different standards. However, a relative measure can be incorporated in the design process, and issues such as noise reduction or other ergonomic factors could be correlated to an increase in the *please* metric.
- The second metric, ***Protect***, also contributes to customer satisfaction as well as to the safety issues in a product. It also includes some of the design for Manufacturability issues.
- The third metric is ***ICost*** or inverse of cost. This is not the consumer *cost*, rather the *cost* of materials and manufacturing, and could be extended to cover the life–cycle *cost* of a part that has to be produced or purchased.

One thing that does hold true in design is that it takes effort to change the design. This is primarily effort on the part of the design team, including engineers and manufacturing personnel. If a current design is being used either as a prototype or in production, it requires no effort to use this design. However, in order to add more value to the design, effort must be expended. This leads

to the view shown in Figure 7 below, where the functions graphed represent any increasing function.

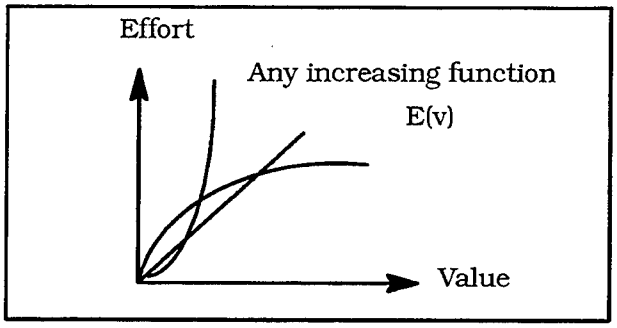


Figure 7. Effort vs. Value.

In the real world, there is a diminishing return for the effort that is put in. With some effort, the design can be improved; but as more effort is expended, it becomes increasingly difficult to improve one area. This means that the equation E(v) should be increasing slowly a first, but become steeper later. A simple equation which fits these criteria is a parabola. In this case, the equation is:

$$E(v) = kv^2 \qquad (1)$$

where k is the constant that determines the steepness of the parabola.

So, we arrive at a system where we have three metrics: Pleasure, Protection, and Inverse Cost (Icost). These metrics provide a basis for comparing designs in a relative manner. But applying these in the two–dimensional scheme above is difficult – the design team will want to consider all three metrics simultaneously. This leads to the design shell, as depicted in Figure 8. This shell is defined by the three metrics as parabolas on the surface.

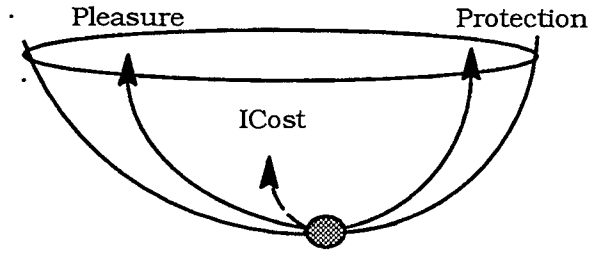


Figure 8. Design Shell.

This shell is not round, since the parabolas will not have the same constant. In this scheme, the bottom of the bowl is the current design. Effort must be expended to increase the value. How much effort depends on which value the design team wishes to increase.

There is a tendency for an increase in one of the metrics to cause a decrease in at least one of the others. But, as a particular metric is increased further and further, it requires more and more effort on the part of the design team to accomplish the goal. When the effort required for an improvement is too great, the best way to improve customer satisfaction is by changing the technolo-

gy. This problem is addressed as it relates to affordability in [Fadel, 94] and is mentioned in the introduction and described in Figure 1.

At no time is the exact size of the bowl known, only the slopes in different directions. As a technology matures, the slope of the bowl increases, making improvements more difficult to design. Even though these improvements still increase the value of the product, the design effort adds too much cost for it to be economically feasible for the company.

Figure 9a equates the diameter of the value bowl with the flexibility of the technology. A large, shallow bowl has lots of room for improvement in all areas, which is evidence of a very flexible technology. A steeper bowl, however, has very little room for improvement before a large expenditure of effort is required to improve the product. This technology dies much more quickly as it is surpassed. Often this steeper bowl is indicative of a mature technology, one that is well understood but has little remaining room for improvement.

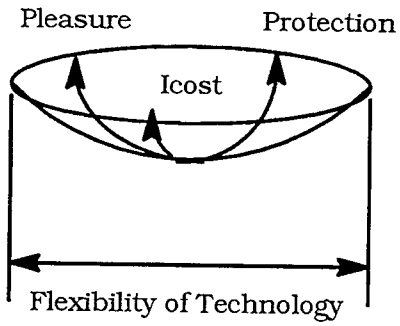


Figure 9a. Standard Design Shell.

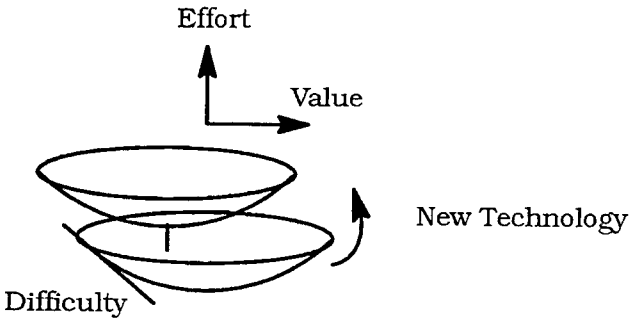


Figure 9b. Changing Design Shells.

Also, Figure 9b illustrates the fact that the jump from one technology to another does not necessarily mean that all three metrics will improve. There is overlap in the bowls, and the value of the product may remain the same or even decrease with a technological jump. However, the new technology should provide more room for improvement in all directions, allowing the designer to improve the product beyond the value in the previous technology.

Also shown in Figure 9b is the difficulty. This is defined as the slope of the value bowl. The difficulty is how much effort is required to improve this particular combination of one or more metrics. Bowls are described as continuous surfaces. Therefore, although bowls are not round, the difficulty should vary smoothly over the surface. Difficulty can be thought of as the derivative of the curvature of the side, or

$$\text{Difficulty} = \frac{\text{Effort Required}}{\text{Increase in Metric}} \qquad (2)$$

A leap in manufacturing technology may also change the bowls. Consider a company moving from producing 100 units per year to 10,000 units per year. In this case, the affordability will increase as the unit cost decreases, without changing the design of the product. This shows that it is important to consider the technology of manufacturing as part of the design.

Combining the Metrics for the Relative Rating of Alternatives

Having defined three possible metrics to assess the design, the next step is to identify ways to combine them in order to have some basis to select one design over another. The representation of the value bowl is not ideally suited to graphically represent the "relative" value of one design versus another. In fact, the local coordinate system of the value bowl needs to be separated from the global coordinate system of relative value since locally, an increase in value means moving away from the current design whereas two different designs are located at some relative location with respect to each other on a different scale. Therefore, we have chosen the utility theory which allows to combine various metrics in a non–linear fashion [Thurst, 93] as the tool needed to convert from the three metrics to the one that can be used to compare designs.

Another advantage of this approach is that the metrics can evolve as more information is available about the design. In fact, the utility theory can be used for each of the individual metrics to arrive at some number. Should definite data be available, cost data can be directly obtained by an additive process instead, whereas the other two metrics would still benefit from the technique that allows the designer to make tradeoffs, here between metrics.

IMPLEMENTATION

The following steps summarize the proposed implementation plan for the function based methodology that will result in a better appreciation of a product's cost and performance.

- The first step in the process is to create a design specification that includes as many constraints and criteria as are necessary. These become the basis for the decisions on which type of component to choose.
- The second step is to develop the functional hierarchy to some level. The hierarchy should probably extend one or two levels, and typically should have been extended past the choice of physical principles.
- Now the basic functions are available and can be used in the evaluation process.
- The next step is to start with one of the functions. Using an automated listing or the designer's own knowledge and experience, he or she should develop a list of physical embodiments for the particular function. Each of the components should remain generic if possible, such as "electric motor", rather than a specific model or type. This enables the designer to compare at a higher level rather than just selecting one of two motors from competing companies.
- Once a list is generated, the designer should eliminate any possibilities that are completely unfeasible. The designer must be careful not to let personal prejudices affect judgment and prematurely remove a viable candidate.

- The next step is to create the matrix comparing the different criteria and behavior to the chosen metrics, *Please*, *Protect*, and *ICost*. This table should be populated by the engineer in a manner similar to a House of Quality. Each of the factors should be assigned a weight as to how much it affects the *Please*, *Protect*, and *ICost* as seen from the eyes of the customer. (Typically, the weights should be – Greatly Affects, – Moderately Affects, – Mildly Affects, and – Does Not Affect.) The numeric values associated with these are often 9, 3, 1, and 0 respectively.

- Now numbers are generated for each of the design alternatives, and for each of the metrics. These can be combined using the utility theory, and a relative ranking can be obtained.

In order to permit the designer to experiment with the proposed taxonomy and metrics, a sample implementation of a design system was developed. The system was written using Tcl and the Tk Toolkit [Ouster, 94]. Tcl is a powerful scripting language that provides a set of generic programming procedures and is user extensible. Furthermore, Tcl can be embedded in C code for additional functionality. Tk is a Graphical User Interface (GUI) tool that is written using Tcl and provides the necessary commands for building interfaces for the X Window System.

Extensions to these languages make many tasks simple. The primary extension used for this work was the Tree Widget [Bright, 1994]. This widget was used to handle the graphical description of the hierarchies.

The system has been designed to work with hierarchic decomposition of functions. It allows the designer to add functions in two ways. First, the designer can add a higher–level function, such as "Drill Hole" . This type of function is added via the keyboard, as shown in Figure 10.

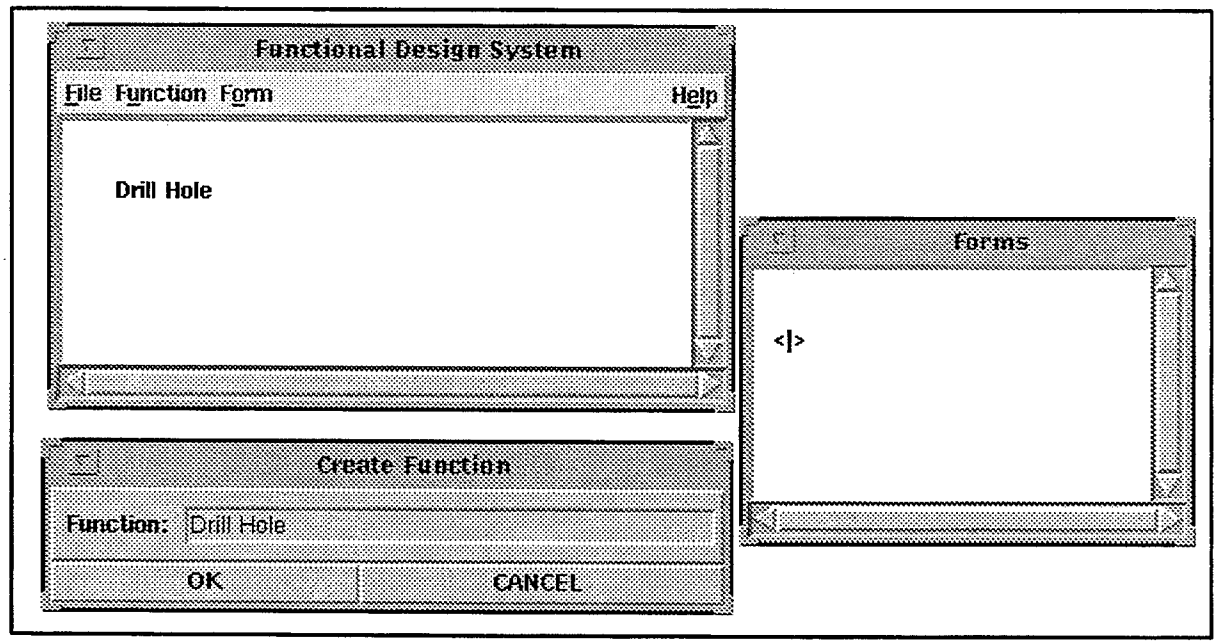


Figure 10. Arbitrary Function Addition.

Since there is nothing in the tree, it is added in the Function tree as the root node. A place holder is added in the Form tree, which may later be replaced by a specific form. The form tree is the embodiment of the design which can be

Taxonomy functions can be added to the hierarchy through the menu choices, as shown in Figure 11. In this figure, the functional insertion box for the Power functions is illustrated. Click-

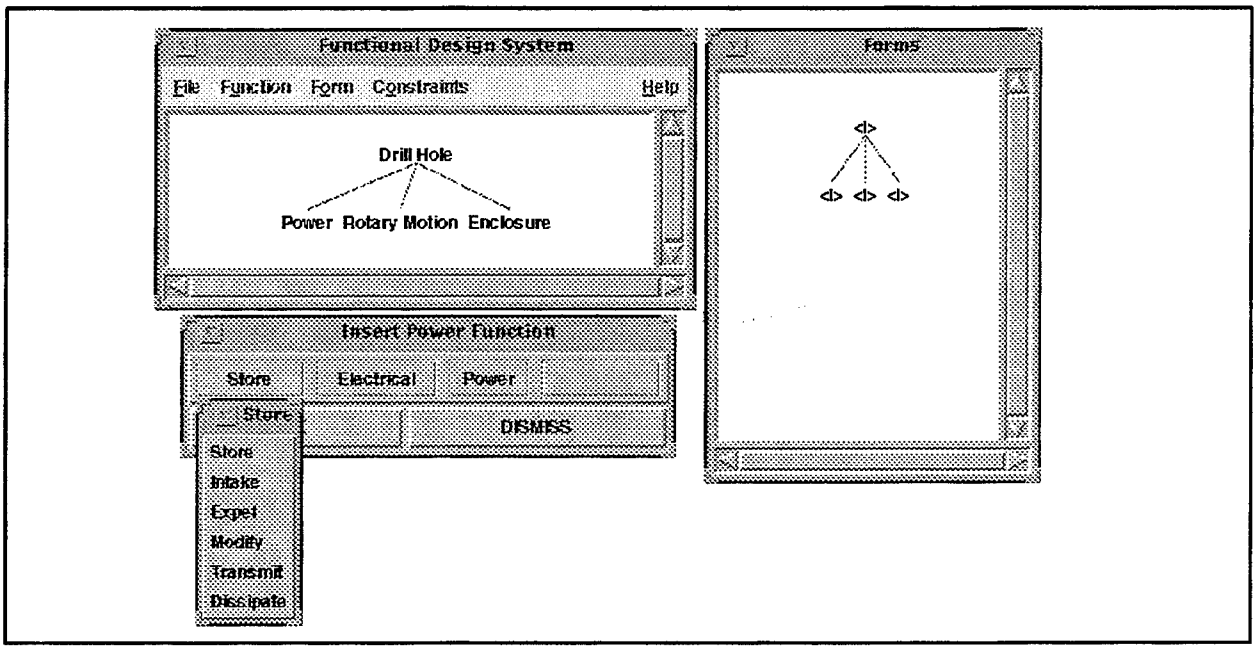


Figure 11. Taxonomy Function Addition.

ing on one of the sentence choices, such as "Store", produces a menu of all choices available for that position. For example, under "Store" are the choices "Intake Expel Modify Transmit Dissipate". Choosing one of the other options will cause that name to appear in the sentence in place of the current option. Again, a place holder is inserted into the Form tree to be replaced by a corresponding form at a later time.

Once the level of functions is found, a set of generic forms can be applied, as shown in Figure 12. For a selected function, a list of possible forms is presented. The designer can choose between the forms, or add a new form to the list based on experience. This permits a set of generic forms to be added to the design in a simple manner. Should the same generic function accomplish more than one function, it can be entered several times, warning the designer that Suh's independence axiom has been violated [Suh, 90].

Using the Metrics in Design

Now that the design is decomposed, how can one select the embodiment that best satisfies the customer and results in the most affordable product? The design system is just a sample of how a taxonomy can be applied. The metrics described earlier have a great potential for evaluating the customer's desire for a product, and they can be applied at any level of design on top of this taxonomy. In this report, the discussion is limited to the application of these metrics to generic form selection at the conceptual design stage.

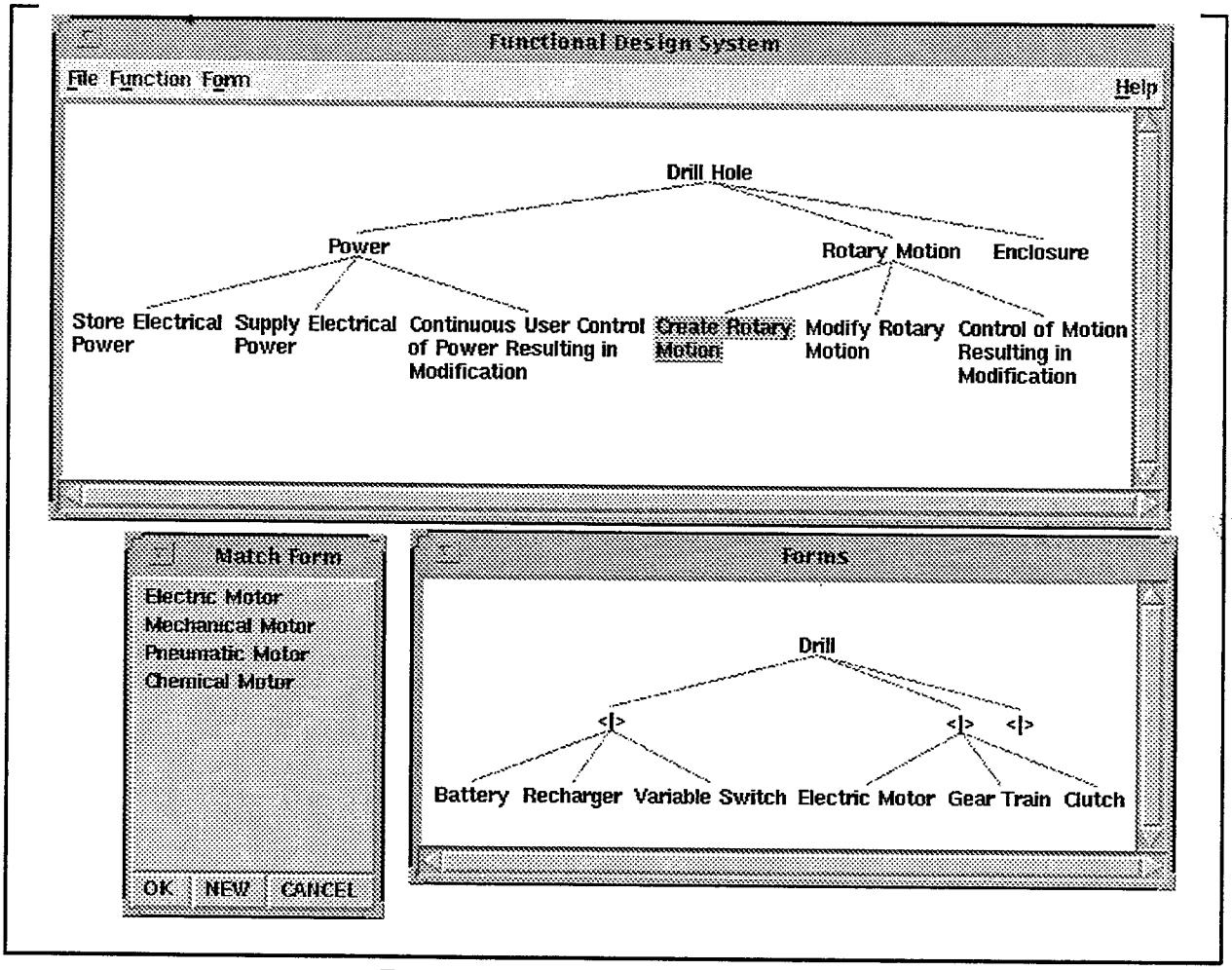


Figure 12. Form Selection.

One of the first steps in a design is to create a design specification that includes all necessary constraints and criteria. These become the basis for the decisions on which type of component to choose. Typically this is done through a technique like QFD where the desires of the customer are brought in and converted into measurable engineering requirements.

Also, behaviors should be identified at this stage. Behaviors of a component or a system can be thought of as the "side-effects" of an item. For instance, an electric motor creates rotary motion. However, a behavior of the motor is that it generates heat, so it could be used as a room heater, albeit an inefficient one. The standard behaviors identified during this research are weight, size, heat, noise, vibration and radiation. Some or all of these may be moved into constraints or criteria by the designer, such as setting a limit on the amount of vibration that can be produced.

Having the constraints, criteria, and behaviors, the designer develops a functional specification and typically decomposes the problem into simpler subproblems. These subproblems can be further decomposed, until a functionally simple level is reached.

If the designer does not have a preset choice, the different forms need to be compared to determine the best choice. This decision should be based on both the ability of the forms to satisfy the constraints, criteria, and behaviors, and the effects that the constraints, criteria, and behaviors have on the metrics of Pleasure, Protection, and Icost. Each of these effects can be considered

separately and combined to create a three dimensional evaluation box as shown in Figure 13 below.

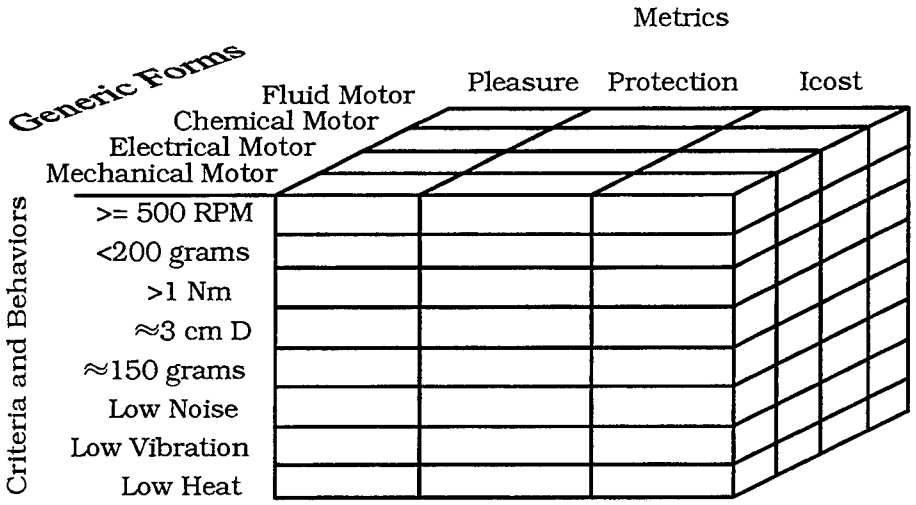


Figure 13. Three–Dimensional Evaluation Box.

A desirability equation permits the designer to compare generic forms in a non–linear manner. It is similar to a non–linear weighted sums approach in that it provides the designer with a manner of expressing preference that is not restricted to a proportional approach.

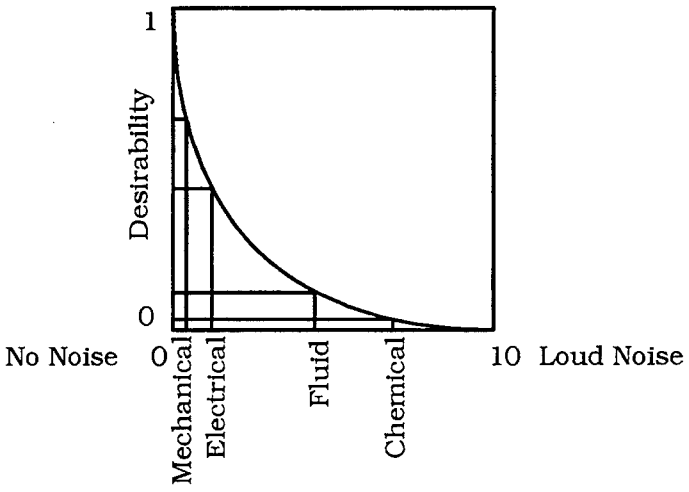


Figure 14 Example Desirability Values for Motors.

A desirability equation provides a means for the designer to measure the value of some physical quantity on a absolute scale but express his/her preference on a non–linear relative scale. For example, the noise values for four different motors are shown in Figure 14. Here the motors are given an absolute value for noise on a scale of 0 to 10 where 0 is no noise and 10 is loud noise. Although the motors are ranked absolutely based on their noise, the designer has chosen to put a premium on quietness and has chosen an inversely proportional curve for the desirability equa-

tion. This arrangement give the quieter motors a much higher score in the table. The resulting values are normalized to a scale of 0 to 1 so that all comparisons are equal.

There are many different types of desirability equations. A set of equations that fulfill most engineering needs is shown in Figure 15. The simplest equation is binary which is used for constraints that must be satisfied at a specific value. A step function is used when there is a value below which the solution is unacceptable, which is often the case in constraints. Proportional equations are linear, and are equivalent to having no equation at all. They are used to normalize the values of the rankings.

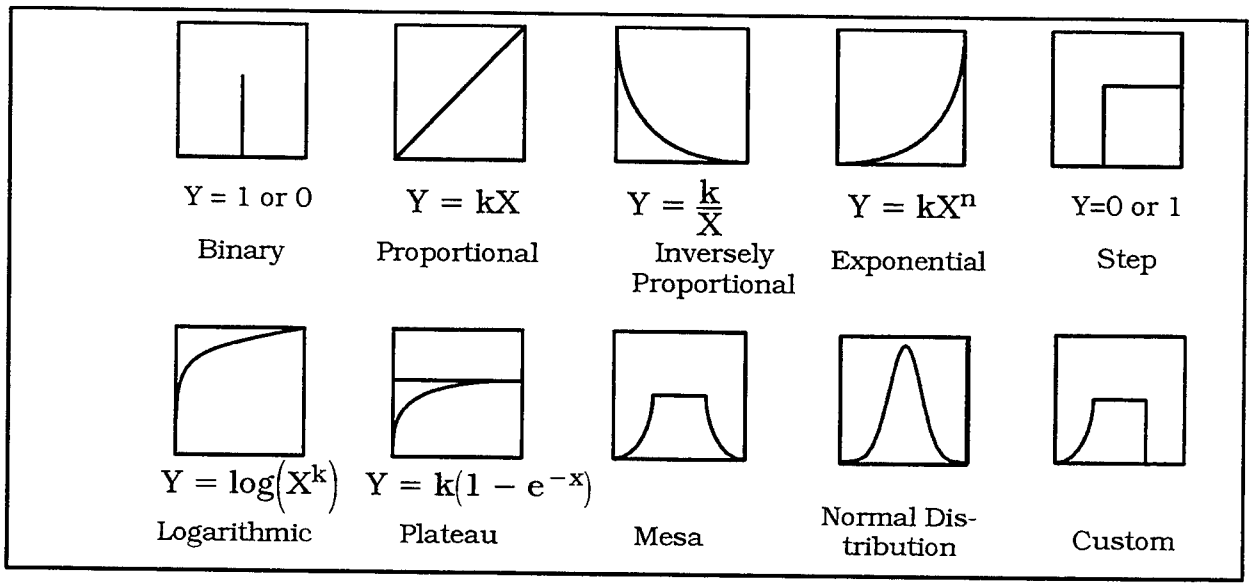


Figure 15. Typical Desirability Equations.

The exponential and inversely proportional equations provide a way to emphasize one end of the scale much more than the other. They asses high penalties for values far from one end. The designer can choose the slope of the curve which controls the magnitude of the penalty assessed. The logarithmic function provides a rapid rise as the performance first increases, but then levels off quickly. The plateau equation is similar to the logarithmic, except that it is much flatter at higher values. Both of these equations are useful for instances when some of an attribute is good, but a lot of it is not better. For example, making a printer somewhat quiet is important, but the value of making it silent is low.

The Mesa function [Fadel, 93] is unusual in that it shows equal preference within a range of values, but assesses steep penalties for values outside that range. It is important because often the designer does not know the exact value desired, but does know a region where it will fall. The normal distribution provides a preference for a particular value, and penalties assessed via the laws of probability. Any of these desirability equations can be combined over various ranges to form custom equations.

These desirability equations have been also implemented in Tc/Tkl, and Figures 16 and 17 are snapshots of the computer screen available to the user.

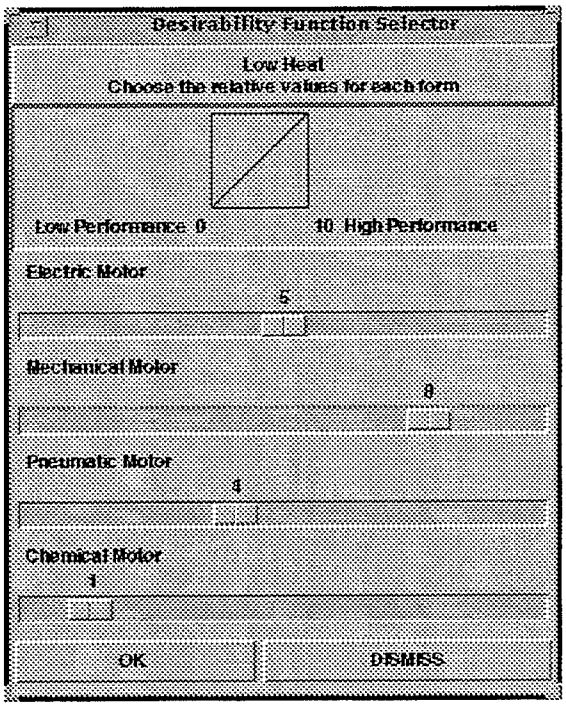


Figure 16. Desirability Function Selector

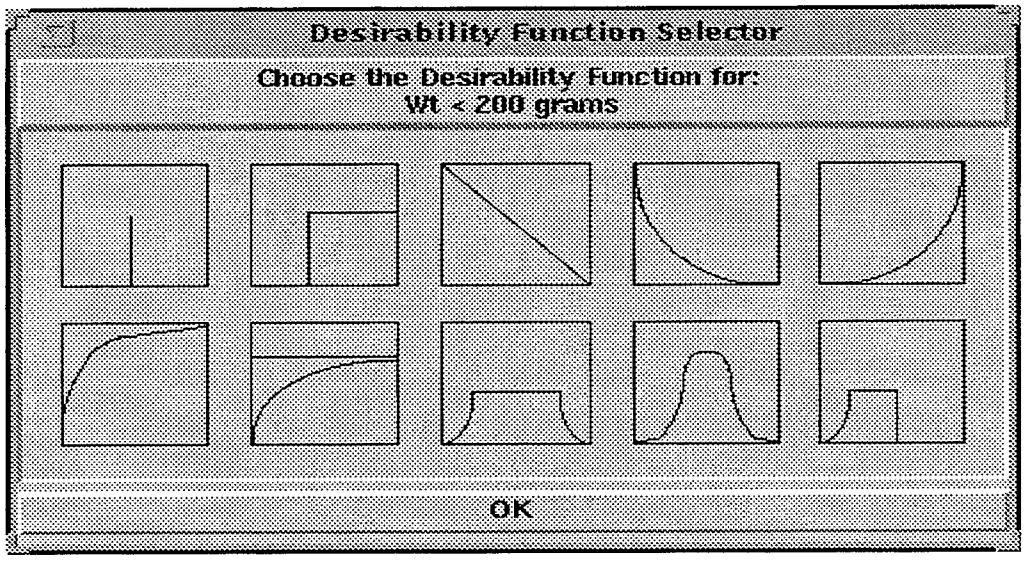


Figure 17. Desirability Function Selector

This Evaluation box is also implemented in the Tk–Tcl system and is illustrated in Figures 18, 19, and 20.

CONCLUSIONS

A taxonomy for Function–based design is proposed. The taxonomy is derived from four major functions that describe mechanical designs, and is used as the foundation for metrics that include

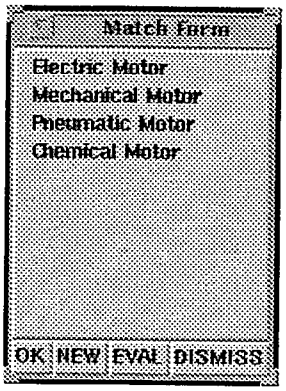


Figure 18. Matching The Forms for Evaluation

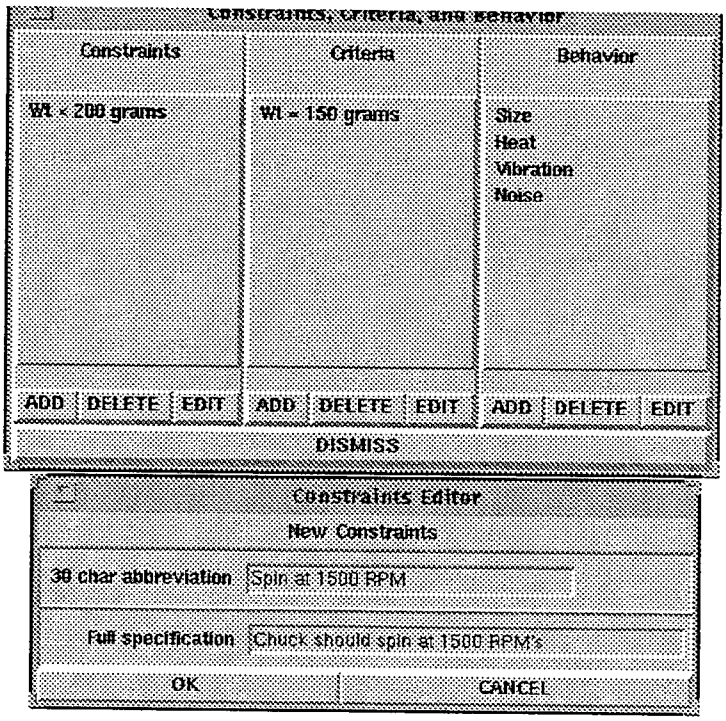


Figure 19 Establishing Constraints, Criteria and Behavic

cost and safety. These functions are the motion, power, control and enclosure functions. A taxonomy is proposed that allows the designer to hierarchically decompose designs starting from the main objective of the product which is the top level function. Using the taxonomy, the designer is given the freedom to select appropriate functions and their form counterpart. The ability to stop at the elemental feature level provides added robustness and a broader selection of components to accomplish a specific function. The proposed taxonomy allows the designer to include his or her intent in the functional description, and the fact that this initial stage of conceptual design stops at the elemental level frees the designer from considering details that are probably irrelevant at this stage of the design. The application of cost, pleasing and safety metrics using a QFD type methodology ensures that the voice of the customer is taken into account during the design. It also provides with a method to compare designs and have some idea of the cost drivers in the design.

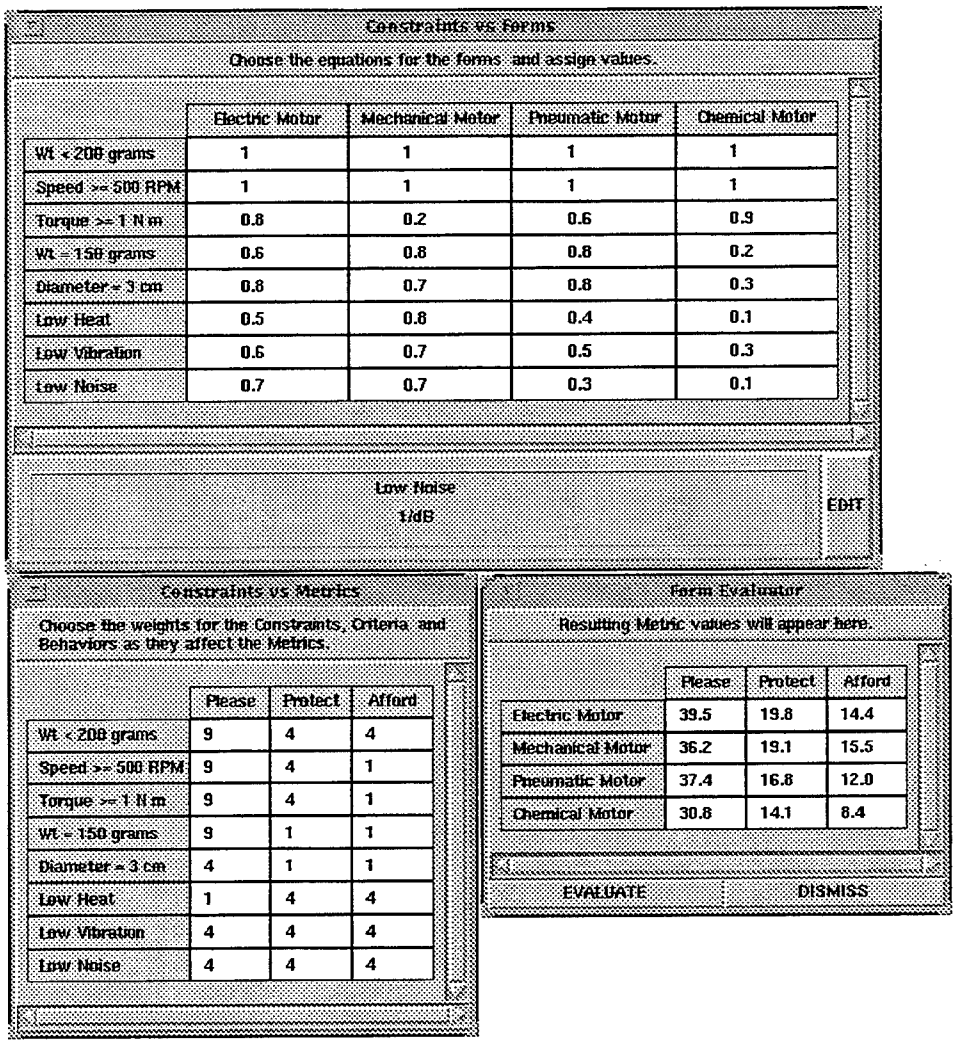


Figure 20. The 3D Evaluation Box after Processing

To this end, a different approach to "measuring" design is provided. The approach is aimed at the selection of individual components, at the generic level. It is here where many of the important decisions that affect the final form are made. This approach utilizes the constraints, criteria, and behaviors identified by the engineers to develop relative preference values.

Another advantage of the proposed system is the small amount of information handled by the designer at any point. Rather than simply choosing an electric motor over all other types, the designer can compare the alternatives based on the various constraints, criteria, and behaviors. This limits the amount of information that the designer must keep current in his or her brain for each ranking.

Furthermore, the value for an entire design is a composite of the parts, and the technique discussed provides a way to optimize from the bottom up. When each of the components is chosen with the customer in mind, the overall product should be better. However, techniques for combining these values to achieve an overall value for the metrics are still being explored. It is hoped that an intelligent combination of the individual metrics will provide insight as to how one design compares with another. Techniques such as utility theory may be applicable to this problem.

New metrics to compare components and designs are also proposed. Current metrics are typically aimed at the designer, as in the DFX strategies. However, the metrics Please, Protect, and Icost are much more closely related to the customer than many of the currently used metrics. These offer a way to estimate customer satisfaction of a product, and to bring that voice down to the level of generic form selection. These three metrics also provide a very balanced description of the customer's desires.

Finally, relating the work to the issue of affordability, we have attempted to establish a methodology for design that includes the aspect of affordability. This methodology is applied at the early level of conceptual design when the highest payoff is possible. The formalized technique facilitates the comparison of forms that accomplish particular functions. It allows designers to measure the flexibility and potential of their designs, It allows them to target increase in value, whether related to cost, pleasing the customer, or providing more safety, and it can be extended for use in design embodiment when more information is available and more accurate numbers can be obtained. The methodology still has many holes and needs to be tested, this report is the first attempt at bringing several ideas together to deal with affordability differently from what is currently attempted.

REFERENCES

[Basca, 94] E. Bascaran and C. Tellez, **"The Use of the Independence Design Axiom as an Enhancement to QFD"**, Proceedings of the 1994 Design Theory and Methodology –DTM '94 Conference, ASME DE–Vol 68, 1994, pp 63–69.

[Booth, 88] Boothroyd, G. and Dewhurst, P.," **Product Design for Manufacture and Assembly"**, Manufacturing Engineer, April, 1988, p.p. 42–46.

[Booth, 90] Boothroyd, G. and Dewhurst, P., **Product Design for Assembly Handbook**, Boothroyd Dewhurst Inc, Wakefield, RI 1990.

[Bright, 94] Brighton, A.,**"Tree Widget"**, Manual page for software, available from ftp.aud.alcatel.com/pub/tcl/extensions, tree3.6.2.tar.gz, 1994.

[Collin, 76] J. A. Collins, B. T. Hagan, and H. M. Bratt, **"The Failure-Experience Matrix – A Useful Design Tool"**, *Transactions of the ASME, Series B, Journal of Engineering in Industry*, Vol 98, Aug 1976, pp 1074–1079.

[Chow, 78] Chow, W.C. **Cost Reduction in Product Design**, Van Nostrand Reinhold, 1978

[Dean, 89] Dean, Edwin B., **"Parametric Cost Analysis: A Design Function"**, Transactions of the American Association of Cost Engineers, 33rd Annual Meeting, San Diego, CA June, 1989.

[Dean, 91] Dean, Edwin B. and Resit Unal, **"Designing for Cost"**, Presented at the American Association of Cost Engineers, 1991.

[Dean, 92] Dean, Edwin B. and Resit Unal, **"Elements of Designing for Cost"**, Presented at the AIAA Aerospace Design Conference Irvine, CA 1992. AIAA–92–1057.

[Dean, 95] Dean, Edwin B., **"Parametric Cost Deployment"**, Proceedings of the 7th symposium on Quality Function Deployment, Novi, MI 1995.

[Dean, 96] Dean, Edwin B., Web resource on **design for Competitive Advantage** at the NASA Langley Research center, http://mijuno.larc.nasa.gov/Default.html

[Ditema, 93] M. L. Diteman and L. A. Stauffer, **"Testing of Relative Comparison Methods for Evaluating Design Concept Alternatives"**, Proceedings of the 1993 Design Theory and Methodology –DTM '93 Conference, ASME DE-Vol 53 1993, pp 149–155.

[Fadel, 93] G. M. Fadel and S. Cimtalay, **"Automatic Evaluation of Move-Limits in Structural Optimization"**, *Structural Optimization*, Vol. 4, 1993

[Fadel, 94] **"A Methodology for Affordability in the Design Process"**, Final Report for Summer Faculty Research Program, Wright Laboratory, Air Force Office of Scientific Research, August, 1994.

[Fowler, 90] T. C. Fowler, **Value Analysis in Design**, Van Nostrand Reinhold, 1990.

[Gutows, 94] Gutowski, T., Hoult, D., Dillon, G. et al, **"Development of a Cost Model for Advanced Composite Fabrication"** Submitted to Composite Manufacturing, May 1994.

[Gunthe, 71] Gunther, W., **Die Grundlagen der Wertanalyse**, Z. VDI 113, 238–241, 1971.

[Ishii, 93] K. Ishii and C. F. Eubanks, **"Life-cycle Evaluation of Mechanical Systems"**, Proceedings of the 1993 NSF Design and Manufacturing Systems Conference, Charlotte, NC, Jan. 6–8, 1993, pp. 575–579.

[Harry, 92] Harry, M. and Lawson, J.R., **"Six Sigma Producibility Analysis and Process Characterization"** Motorola University Press, 1992.

[Hauser, 88] J. R. Hauser and D. Clausing, **"The House of Quality"**, *Harvard Business Review*, May–June 1988, pp 63–73.

[Kirsch, 96] C. F. Kirschman, G. M. Fadel and C. C. Jara–Almonte, **"A Function-Based Taxonomy for Mechanical Design"**, submitted to *DTM '96 Conference*, September, 1996.

[Knight, 85] Knight, W.A. and Poli, C., **"A Systematic Approach to Forging Design"**, Machine Design, 57, January 24, 1985

[Locasi, 94] A. Locasio and D. L. Thurston, **"Quantifying the House of Quality for Optimal Product Design"**, Proceedings of the 1994 Design Theory and Methodology –DTM '94 Conference, ASME DE-Vol 68, 1994, pp 43–54.

[Michae, 89] Michaels, J.V., Wood, W.P. **Design to Cost**, Wiley Interscience, 1989.

[Otto, 91] K. N. Otto and E. K. Antonsson, **"Trade-Off Strategies in Engineering Design"**, *Research in Engineering Design*, Vol 3, 1991, pp 87–103.

[Otto, 93] K. N. Otto, **"Measurement Foundations for Design Engineering Methods"**, Proceedings of the 1993 Design Theory and Methodology –DTM '93 Conference, ASME DE-Vol 53 1993, pp 157–165.

[Ouellet, 92] M. P. Ouellette, **"Form Verification for the Conceptual Design of Complex Mechanical Systems"**, Masters of Science Thesis, Georgia Institute of Technology, April 1992.

[Ouster, 94] Ousterhout, J.K., **Tcl and the Tk Toolkit**, Addison–Wesley, 1994.

[Poli, 88] Poli, C, Escudero, J, Poli, C. and Fernandez, R. "**How Part Design Affects Injection Molding Tool Costs**", Machine Design, 60, November 24, 1988

[Pugh, 91] S. Pugh, **Total Design: Integrated Methods for Successful Product Engineering**, Addison Wesley, 1991.

[Raymer, 92] Raymer, Daniel, **Aircraft Design A Conceptual Approach**, AIAA 1992

[S&T, 93] **S&T Affordability White Paper**, Air Force, Wright Laboratories, Mantech, Concurrent Engineering, pg.2 July 1993.

[Sferro, 94] Sferro, Peter, Personal communication, FORD Alpha Manufacturing Center, June 1994.

[Staple, 94] Staples, J. W., "**Optimizing the Allocation of Resources during Preliminary Automobile Design**," M.S. Thesis, Georgia Tech, 1994.

[Suh, 90] Nam Suh, **The Principles of Design** , Oxford University Press, New York, 1990.

[Sulliv, 86] L. P. Sullivan, "**Quality Function Deployment**", *Quality Progress*, June 1986, pp 39–50.

[Taguch, 93] Taguchi, G., **Introduction to Quality Engineering – Designing Quality into Products and Processes**, New York, UNIPUB/Quality Resources, 1986.

[Thurst, 91] D. L. Thurston, "**A Formal Method for Subjective Design Evaluation with Multiple Attributes**", *Research in Engineering Design*, Vol 3, 1991, pp 105–122.

[Thurst, 93] Thurston, D.L. and Essington, S.K., "**A Tool for Optimal Manufacturing Design Decisions**", Manufacturing Review, Vol. 6., No. 1., March 1993.

[Trucks, 74] Trucks, H.E. **Designing for Economical Production**, SME 1974.

[Ullman, 92] D. G. Ullman, **The Mechanical Design Process**, McGraw–Hill, 1992.

[VDI, 72] VDI–Taschenbuch 135, **Wertanalyse – Idee, Methode, System 4.** Dusseldorf, VDI Verlag, 1972.

[Whitne, 88] D. E. Whitney, "**Manufacturing by Design**", *Harvard Business Review*, July–August, 1988.

[Whitne, 94] D. E. Whitney, "**Some Differences Between VLSI and Manufacturing Design**", Proceedings of the NSF Workshop on New Paradigms for Manufacturing, Arlington, VA, May, 1994, pp 61–64.

DATA REDUCTION AND ANALYSIS FOR LASER DOPPLER VELOCIMETRY

Dr. Richard D. Gould
Associate Professor
Mechanical and Aerospace Engineering Dept.

Mechanical and Aerospace Engineering Dept.
North Carolina State University
Raleigh, NC 27695

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

North Carolina State University

December 1995

DATA REDUCTION AND ANALYSIS FOR LASER DOPPLER VELOCIMETRY

Richard D. Gould
Associate Professor
Mechanical and Aerospace Engineering Dept.
North Carolina State University

ABSTRACT

The laser Doppler Velocimeter (LDV) data analysis software used by the experimental research branch at Wright Laboratory (WL/POPT) was found to be deficient in many ways. A new LDV data analysis software program was developed to resolve these deficiencies. The program named TSISTAT was written in the FORTRAN language and was written to be portable so that it can run on UNIX workstations, mini-computers or Intel processor based personal computers. The program reads TSI IFA750 acquired raw laser Doppler velocimetry files and calculates turbulence statistics up to the third moment for mixed turbulence quantities and up to the fourth moment for homogeneous turbulence quantities for up to three velocity components. A batch processing capability has been included so that up to 100 files can be processed during each execution. A menu screen allows the user to select various run time options. In addition, it constructs and prints automatically scaled histograms of velocity PDFs with a labeled summary table of all pertinent information. These histograms and summaries can be formatted to print as an ASCII file, a postscript file or a Hewlett Packard PCL 5 file. Turbulence statistics are written to three formatted data files for later use by plotting routines or analysis programs. These statistics can be normalized with a constant reference velocity or can be normalized with different reference velocity for each data point processed during the batch job by reading a user created file which contains the reference velocities for each data point. Prior to calculating the turbulence statistics, data points lying outside of ±3 standard deviations are discarded. This default value of 3 can be overridden by the user at run time. The program allows for three velocity bias corrections, allows for non-orthogonal laser beam angle correction, and calculates the statistical uncertainties for all quantities using the jackknife method. Lastly, the user can override the fringe spacing, frequency shift and laser wavelength in the raw data files by reading a user created file which contains the correct values for these parameters. The source code was delivered to the scientists at WL/POPT.

DATA REDUCTION AND ANALYSIS FOR LASER DOPPLER VELOCIMETRY

Dr. Richard D. Gould

1. INTRODUCTION

The purpose of this proposed research was to develop software which can be used to analyze the raw laser Doppler velocimeter (LDV) data obtained using a Thermal Systems Incorporated (TSI) IFA750 signal processor. The software is necessary to provide additional important turbulence statistics and other capabilities not found in the standard TSI FIND data acquisition and analysis software package which is shipped with the IFA 750 signal processor.

The new software, named TSISTAT, has all the capabilities of the TSI FIND software including the ability to analyze up to three velocity components, with time between data and residence time data if these options were selected at data acquisition time, and will calculate all the standard turbulence statistics (i.e. mean velocities, standard deviation, Reynolds stresses, turbulent triple products). In addition, up to 100 files can be analyzed at one time using the batch processing capability of this new program. A menu screen allows the user to select various run time options. Auto-scaled histograms and formatted and fully labeled summary tables are created for each data point and are formatted for one, two or three-component velocity measurements with this program. These histograms and summaries can be formatted to print as an ASCII file, a postscript file or a Hewlett Packard PCL 5 file. High and low pass data filtering based on user selected number of standard deviations is also included. The number of discarded points and revised turbulence statistics are calculated with this option. Also, the correct velocity components relative to an orthogonal laboratory coordinate system can be obtained even if the laser beams were rotated (on purpose or inadvertently) relative to this laboratory coordinate system during data acquisition. More importantly, the correct velocity can be obtained even if the laser beam pairs were not aligned to be perfectly orthogonal, as is the case with most LDV systems. Three velocity bias correction methods, the McLaughlin-Tiederman correction, the residence time correction and the time between data correction, were included as options in this new software. Lastly, the capability to override the default fringe spacing, frequency shift and laser wavelength is also possible with this software.

This software was developed to be portable so it can be run on the UNIX based workstations and mini-computers at Wright Laboratory in the Advanced Propulsion and Power

Directorate (WL/POPT) and also on Intel processor based personal computers. A modular approach was used so that subsequent modification can be accomplished easily. The fully commented FORTRAN source code consisting of one main program, 32 subroutines and 4 function programs, totaling 5039 lines of code, was provided to WL/POPT so that future additions and modifications can be performed.

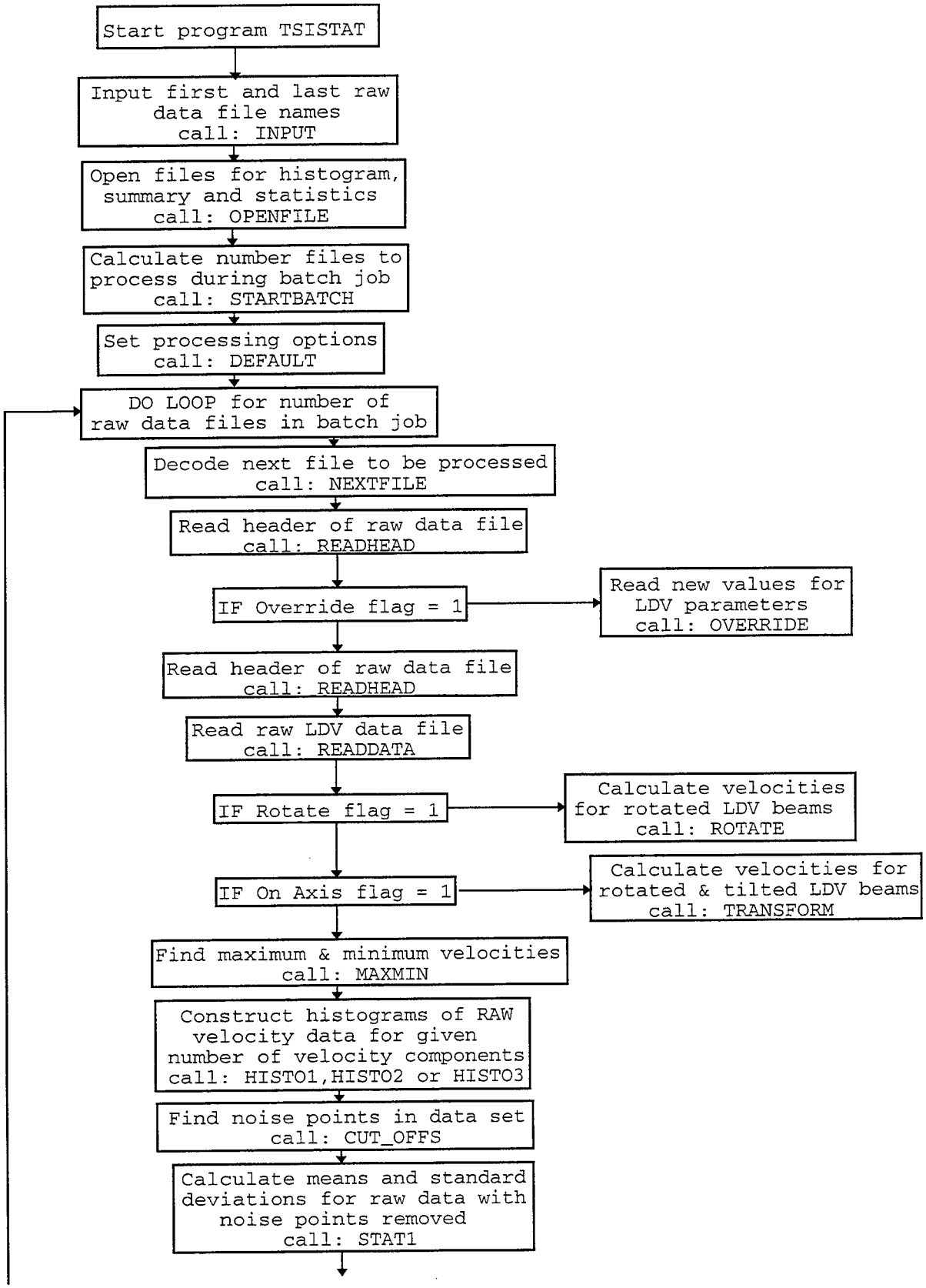
2. PROGRAM STRUCTURE

A modular program structure with one main calling program, named TSISTAT, was selected for this project. The FORTRAN source code listing of the main program, showing this structure, is given in the Appendix. A flow chart of the main program is given below in Figure 1. A named common block structure was used to "pass" variables between subroutines instead of passing the variables directly between the calling program and the various subroutines using an argument. This method was selected to save memory allocation space for the "passed" variables and also allows all the subprograms which include the declared named common block access to the variables and control over their values. If variables were passed in an argument list in a subroutine call they would need to be dimensioned in both the calling program and the subroutine. This requires twice the memory space for variable storage than is required using common blocks. Ten descriptive, named common blocks were defined, instead of one large common block holding all the variables, so that only the required named common blocks for each subroutine need to be declared. It is believed that this structure makes the program more readable. A list of all the named common blocks used is given below.

```
common /ldvdat/ u(5120),v(5120),w(5120),ttu(5120),ttv(5120)
                ,ttw(5120),curtime(5120),dpoints
common /limits/ u_min,u_max,v_min,v_max,w_min,w_max,u_cut_l
                ,v_cut_l,w_cut_l,u_cut_h,v_cut_h,w_cut_h,sigma
                ,tmax,srate,isamp,inoise,i3sig
common /ldvset/ df1,df2,df3,fs1,fs2,fs3,wlen1,wlen2,wlen3,del,eps
common /processor/ nowpdp,c,nokdp,numctr,ctype1,ctype2,ctype3
                  ,samptim,coinwind,mode,tbd,ttime,axis
common /defaults/ uref,angle_u,angle_v,angle_w,iover,irot,hunit
                  ,inorm,iprint,ips,ipcl,imt,irt,itbd,t1,t2,t3,t4
                  ,t5,t6
common /stats/ ubar,vbar,wbar,stdev_u,stdev_v,stdev_w,ufrac,vfrac
               ,wfrac,ti_u,ti_v,ti_w,uv,uw,vw,uuu,vvv,www,uuv,uuw
               ,uvv,uww,vww,vvw,uuuu,vvvv,wwww,u95,v95,w95
common /sums/ sum2_u,sum3_u,sum4_u,sum2_v,sum3_v,sum4_v,sum2_w
              ,sum3_w,sum4_w,sum_uv,sum_uw,sum_vw,sum_uuv,sum_uuw
              ,sum_vvw,sum_uvv,sum_uww,sum_vww
common /uncert/ ustd95j,vstd95j,wstd95j,uu95j,vv95j,ww95j,uv95j
                ,uw95j,vw95j,uuu95j,vvv95j,www95j,uuv95j,uuw95j
                ,uvv95j,uww95j,vww95j,vvw95j
common /position/ x,y,z

common /files/ nmlen,sfname,fnamein,extname
```

TSISTAT PROGRAM FLOW CHART



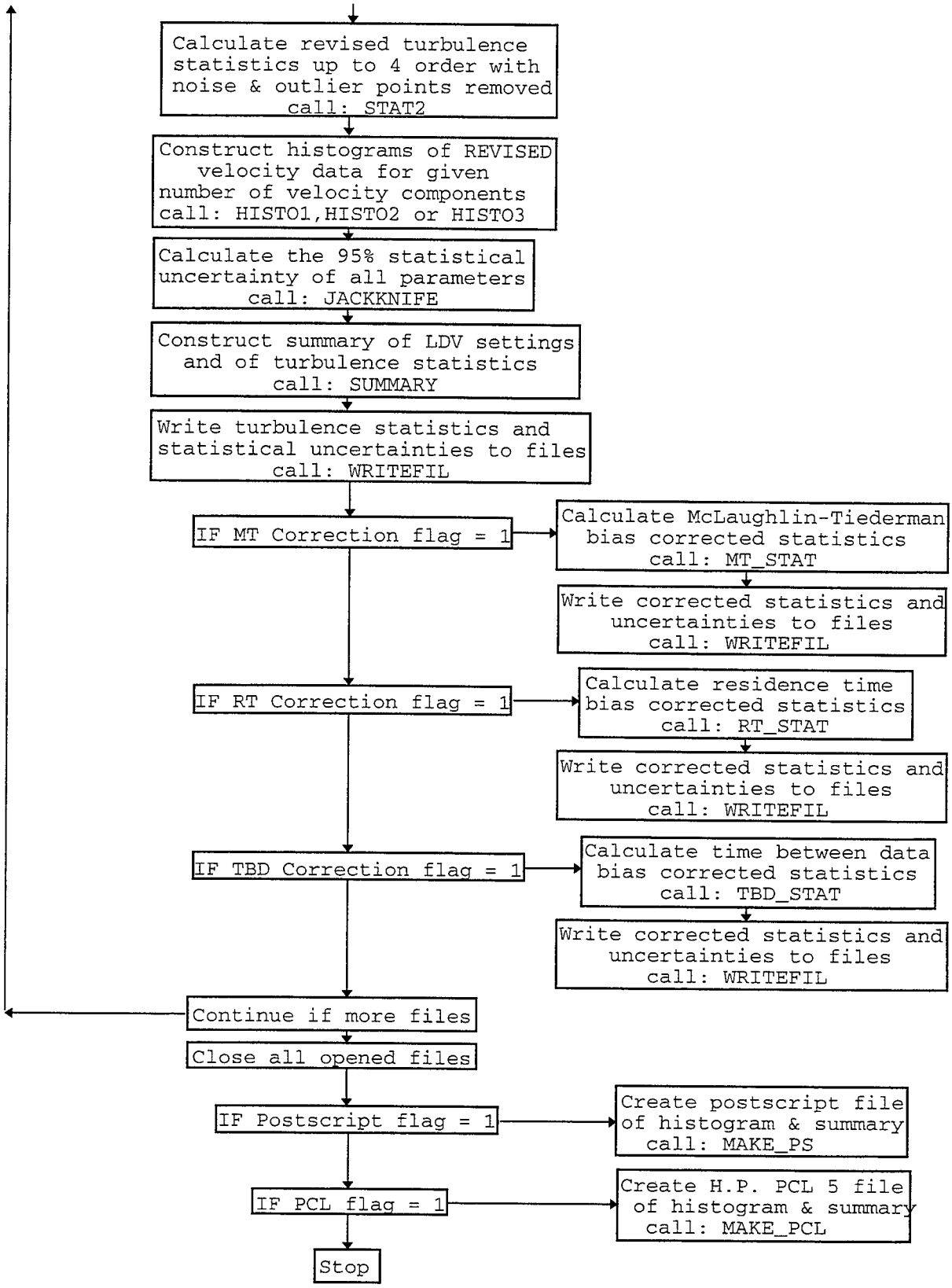


Figure 1. Flow chart of program TSISTAT.

Numerous files are opened, assigned file names and closed during program execution. A summary table of file allocation is given below. Note that the name "family" in the file names below is replaced during program execution with the actual family name that was assigned to the raw data files being analyzed.

Table 1. File allocation.

Unit number	File name	Opening subroutine	Closing subroutine	Read or write	File type
1	family.R**	Words_2 - Words_7	Words_2 - Words_7	Read	direct
2	family.OVR	Overwrite	Overwrite	Read	sequential
7	family.PRT	Default	Program Tsistat	Write	sequential
9	family.R**	Nextfile	Readheader	Read	sequential
10	family.STM	Openfile	Program Tsistat	Write	sequential
11	family.STR	Openfile	Program Tsistat	Write	sequential
12	family.STT	Openfile	Program Tsistat	Write	sequential
13	family.MTM	Default	Program Tsistat	Write	sequential
14	family.MTR	Default	Program Tsistat	Write	sequential
15	family.MTT	Default	Program Tsistat	Write	sequential
16	family.RTM	Default	Program Tsistat	Write	sequential
17	family.RTR	Default	Program Tsistat	Write	sequential
18	family.RTT	Default	Program Tsistat	Write	sequential
19	family.TBM	Default	Program Tsistat	Write	sequential
20	family.TBR	Default	Program Tsistat	Write	sequential
21	family.TBT	Default	Program Tsistat	Write	sequential
22	family.REF	Default	Program Tsistat	Read	sequential
20	family.PRT	Make_ps Make_pcl	Make_ps Make_pcl	Read	sequential
21	family.EPS family.PCL	Make_ps Make_pcl	Make_ps Make_pcl	Write	sequential

3. SUBROUTINE DESCRIPTIONS

A brief description of all the subroutines and function programs contained in this data analysis program are given below.

Subroutine INPUT

Subroutine prints header to screen and queries user for the name of the starting and ending TSI IFA 750 generated raw data files. An example of this screen is given in Figure 2.

```
  ************************************************************************
  *    Program reads TSI IFA750 acquired raw laser Doppler velocimetry   *
  *    files and calculates turbulence statistics, prints histograms of  *
  *    velocity PDFs, allows for 3 velocity bias corrections, allows for *
  *    non-orthogonal beam correction, writes statistics to data files   *
  *    and calculates the statistical uncertainties for all quantities   *
  *    using the jackknife method. Batch file processing is also possible.*
  *                                                                      *
  *                                                                      *
  *    Developed by: Richard D. Gould                                    *
  *                  Jan. 16, 1996                                       *
  *                  version 1.0, All rights reserved by author          *
  ************************************************************************


Enter first file name to be processed(JUNK.R00)
test.r00
Enter extension of last file to be processed(R30)
r13


        14 files will be processed during batch job!
             starting with :     test.r00
             ending with :       test.r13

        Histograms and summary statistics will be stored in file :     test.PRT
        Statistics will be stored in files -     test.STM
                                                 test.STR
                                                 test.STT

        Should we continue batch job using these files?(y/[n])
y
```

Figure 2. Screen produced be subroutines INPUT AND STARTBATCH.

Subroutine OPENFILE

Subroutine opens a file for writing histograms and formatted summary tables of turbulence statistics. Subroutines HISTO1, HISTO2 OR HISTO3 create the histograms, while subroutine SUMMARY creates the summary tables. This subroutine also opens three files where the revised ensemble averaged turbulence statistics are written. Subroutine WRITEFIL writes the formatted statistics to these opened files. The revised turbulence statistics and histogram summaries for all raw data files processed during each batch job are written sequentially to these opened files. The same "family" name as that of the raw data files being processed during this batch job is used as

the "family" name for these new histogram and statistics files. A summary of the files opened in this subroutine is given in Table 2.

Table 2. File description for subroutine OPENFILE.

Unit number	File name	Data in file
7	family.PRT	ASCII file of histograms and summary table for each data point processed during batch job.
10	family.STM	x,y,z position, reference velocity and ensemble averaged mean velocity statistics
11	family.STR	x,y,z position, reference velocity and ensemble averaged Reynolds stress statistics
12	family.STT	x,y,z position, reference velocity and ensemble averaged turbulent triple product statistics

Subroutine STARTBATCH

Subroutine decodes the starting and ending file names and determines how many files are to be processed during current execution of the program. The histograms and statistics of all the files processed during this batch job will be included in the output files. Start and end file numbers and number of files are passed back to calling program. Figure 2 shows the screen produced by this subroutine for an example where TEST.R00 is the first raw data file and TEST.R13 is the last raw data file to process during this batch job.

Subroutine DEFAULT

Subroutine defines default settings used to control file processing and gives menu driven interactive control to change these settings prior to batch processing. The menu screen produced by this subroutine is shown in Figure 3 and has six options. The first option concerns the selection of a reference velocity. A reference velocity can be input as a constant value for all files processed during the current batch job or can be input by reading a user provided input file called 'family'.REF which may have a different reference velocity for each raw data file being processed. If a reference velocity data file is selected it must contain a reference velocity for each data point processed during the current batch job. One line from the reference velocity file is read each time subroutine WRITEFIL is called just prior to writing the normalized turbulence statistics. The read statement and format of the reference velocity file is:

```
      read(22,10) uref
10  format(f10.4)
```

It is important to note that additional reference velocities have to be included in this file if any of the velocity bias correction statistics are selected. This is because they utilize subroutine WRITEFIL also. This can be seen in the source code listing shown in the Appendix. The default sets Uref=1.0, that is no normalization.

```
    The current settings are as follows:

         Histogram and statistics summary is connected to unit: 7

         1.) Normalization: No normalization used,     Uref =   1.000

         2.) Histogram file format: ASCII     : test.PRT

         3.) Cut-off number of standard deviations: 3.0

         4.) LDV probe volume correction angles(u,v,w):  (  .000,  .000,  .000)

         5.) Velocity bias correction: No bias correction will be made

         6.) LDV fringe spacing,freq. shift,wavelength: will NOT be overridden

    Do you wish to change any of these settings(y/n)?
n
```

Figure 3. Menu screen produced by subroutine DEFAULT.

Option 2 specifies the format of the histogram and summary pages. Users can select ASCII, postscript or HP PCL 5 format. Note that an ASCII file is always produced and is written to unit 7 to a file named "family".PRT. This file is read by subroutines MAKE_PS or MAKE_PCL if postscript or PCL 5 formats, respectively, are desired. Option 3 allows high and low pass limits to be selected for the revised data by prescribing the maximum number of standard deviations the velocity PDFs can span. The default value is 3 standard deviations. Velocity components for a laboratory orthogonal coordinate system can be calculated even when the LDV lasers beam were rotated relative to this coordinate system during data acquisition (*i.e.* +45 and +135 deg. beam orientation) by inputting the beam rotation angles using option 4 in the menu. Correction for non-orthogonal beam orientation can also be made using this option by inputting the actual beam angles relative to the x-axis in the laboratory coordinated system (*i.e.* +0.93, +89.9, 0.0) input here. The default is set for no beam rotation and for perfect beam orthogonality. Option 5 instructs the program to calculate turbulence statistics using either one of three or all three velocity bias correction schemes. The three velocity bias corrections are: the McLauglin-Tiederman bias correction, the residence time correction and the time between data correction. If selected corrected turbulence statistics are written to data files having the same

family name as the raw data files processed during the present batch job, but with extension of MT*, RT* and TB* for McLauglin-Tiederman, Residence time and Time between data corrected statistics. The * is either an M, R of T for mean velocities, Reynolds stresses or triple products. Table 1 summarizes the file allocation if this option is selected. Note the point about the additional entries required in the reference velocity file if bias corrected turbulence statistics are desired. The default is set for no velocity bias corrected turbulence statistics. Lastly, option 6 allows the LDV fringe spacing(s), frequency shift(s) and laser wavelength(s) to be overwritten and used in the subsequent calculations if desired. The override file, if needed, is discussed in the subroutine OVERRIDE section below. The default is set for no override. A summary of the option flags used in subroutine DEFAULT is given in Table 3.

Table 3. Control flag definitions in subroutine DEFAULT.

Function	Flag variable	Flag setting definitions
Normalization flag(Option 1)	inorm	1 = no normalization, Uref = 1.0 2 = normalize with constant 3 = read file for reference velocity
Histogram and summary page format flag (Option 2)	iprint	1 = ASCII format (family.PRT) 2 = Postscript format (family.EPS) 3 = PCL 5 format (family.PCL)
LDV beam rotation flag (Option 4)	irot	0 = no beam rotation 1 = beam rotation(input angles from keyboard)
McLaughlin-Tiederman velocity bias correction flag	imt	0 = do not calculate M-T corrected statistics 1 = calculate M-T corrected statistics
Residence time velocity bias correction flag	irt	0 = do not calculate M-T corrected statistics 1 = calculate M-T corrected statistics
Time between data velocity bias correction flag	itbd	0 = do not calculate M-T corrected statistics 1 = calculate M-T corrected statistics
LDV parameter override flag (Option 6)	iover	0 = do not override LDV parameters 1 = read file to override LDV parameters

Subroutine NEXTFILE

Subroutine encodes the integer file number in the batch process DO LOOP into the character string in the extension name of the next raw data file to be processed during batch job. Once the file name is determined it is opened as unit 9. The current file number and raw data file name in the batch job is written to screen so that the user can monitor progress.

Subroutine READHEADER

Subroutine reads the header of each TSI IFA750 raw data file according to the documentation given in Appendix A of the TSI FIND software manual (Version 4). This subroutine was adapted from the source code provided by TSI with the FIND software distribution. The header contains 152 lines of 20 length character strings and one line of consisting of 14 length character string. See FIND manual for more details. These 20 length characters strings are decoded into floating point numbers using function ATOF20 and into integer numbers using function ATOI20. The important variables read from the header are listed below in Table 4.

Table 4. LDV and signal processor settings.

Variable name	Description
mode	1 = random, 0 = coincidence
numctr	number of signal processors
tbd	0 = off, 1 = on, 2 = even time
onaxis	0 = off, 1 = on
nokdp2	numbers kilo data points
nokdp	number of data points
ttime	transit time: 0 = off, 1 = on
ctype1-3	processor type: 1=1990,2=1980,3=ifa550,4=ifa750
fd1, fd2, fd3	fringe spacing (nanometers)
fs1, fs2, fs3	frequency shift (MHz)
wlen1, wlen2, wlen3	wavelength (micrometers)
x, y, z	position
coinwind	coincidence window(microseconds)
samptim	even time(microseconds)
del, eps	rotation about z and tilt about x angles

Subroutine OVERRIDE

Subroutine reads a file having the family name of the raw data files currently being processed during the batch job and the extension of .OVR if the override flag is set to one by subroutine DEFAULT. This file contains user defined values for the fringe spacing, frequency shift and laser wavelength for all three velocity components. This file is read after the header file of each raw data file is read and thus these new values override the ones defined in the header file. These values are overridden for each file processed in the present batch job. This option should be selected only if the information stored in the header file is incorrect. The read and format statements used in subroutine OVERRIDE are listed below. The program expects the

fringe spacing in units of microns, the frequency shift in units of Mhz and laser wavelength in units of nanometers. This file should always contain three lines even if only one or two components are being analyzed.

```
      read(2,100) df1,df2,df3
      read(2,100) fs1,fs2,fs3
      read(2,100) wlen1,wlen2,wlen3
100   format(3f10.4)
```

An example a *.OVR file is listed below. The single line above the three data lines is only used to show column numbers and should not be included in the "family".OVR data file.

```
12345678901234567890123456789

   1.8000    1.7000    1.6000
  39.0000   38.0000   37.0000
 510.0000  480.0000  460.0000
```

Subroutine READDATA

Subroutine determines the number of words transferred(i.e. nowpdp) with each LDV realization based on number of signal processors, mode of operation and, whether the time between data word or transit time words are transferred. Once the number of words per realization is found this subroutine calls the proper subroutine to read this data sequence. Table 5 lists the names of the specific subroutines and the word pattern for the given number of words per data point. This subroutine was adapted from the source code provided by TSI with the FIND software distribution.

Table 5. Subroutine names to read raw LDV data.

Number of words per data point	Subroutine name	Word pattern (see subroutines CONVERTA, and CONVERTB for word formats)
2	WORDS_2	aword, bword
3	WORDS_3	aword, bword, tbd word
4	WORDS_4	aword, bword, aword, bword
5	WORDS_5	aword, bword, aword, bword, tbd word
6	WORDS_6	aword, bword, aword, bword, aword, bword
7	WORDS_7	aword, bword, aword, bword, aword, bword, tbd word

Subroutine ROTATE

Subroutine calculates laboratory coordinate x, y and z-axis velocities from LDV measurements made when the LDV beams were rotated relative to this coordinate system. For example, in many cases it is possible to make measurements nearer to a wall or solid surface if

the beams in a two-component LDV system are rotated +45.0 and +135.0 degrees with respect to the x-axis in the laboratory coordinate system. This option is selected in subroutine DEFAULT. If the IROT flag is set to one this subroutine is called and all velocities are transformed to the laboratory coordinate system. This subroutine calculates the correct orthogonal velocity components even if the laser beams are rotated (or misaligned) such that they are non-orthogonal. The beam orientation angles are input by the user from menu option 4 in subroutine DEFAULT. All angles are referenced to the laboratory coordinate x-axis. That is, the x-axis is 0 degrees and the y-axis is 90 degrees. . In its current form the correction for the z-axis is a simple cosine projection of the z-component. Figure 4 shows two sets of LDV beams that are nonperfectly orthogonal where, θ_b, θ_g are the deviation angles of the blue and the green beams from being orthogonal to the laboratory coordinate system, x-y.

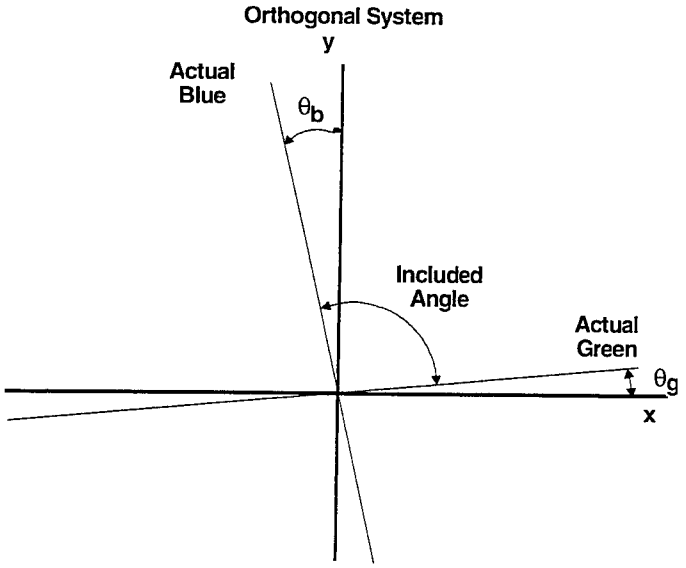


Figure 4. Departure from orthogonal system.

Note that subroutine DEFAULT requires that you input θ_g for the u direction angle and $90^\circ + \theta_b$ for the v direction angle in this example. The following equations were used to perform the coordinate transformation for an LDV system where, the x & y-coordinates may be rotated relative to the laboratory coordinate system and may also be non-orthogonal,

$$U = \frac{U_b \sin\theta_g - U_g \sin\theta_b}{\sin(\theta_g - \theta_b)} \qquad V = \frac{U_b \cos\theta_g - U_g \cos\theta_b}{\sin(\theta_b - \theta_g)} \qquad W = U_v \cos\theta_v$$

where; U_g, U_b, and U_v are the velocities measured by the rotated system, U, V, and W are the transformed velocities relative to the orthogonal laboratory coordinate system, g, b, and v denote the rotated beams, θ_g = the rotation angle of beam g relative to the x-direction axis, θ_b= the rotation angle relative to the x-direction axis, and θ_v = the rotation angle relative to the z-direction axis.

Subroutine TRANSFORM

Subroutine calculates the velocities when the beams are rotated relative to the laboratory coordinate system x-axis by the angle δ, and the system is tilted by the angle ε. These two angles are read from the TSI raw data file header. If they are set during data acquisition the ONAXIS flag is set to 1. This subroutine is called by the main program only if ONAXIS = 1. Note that the rotation algorithm used below assumes that the two beam sets are perfectly orthogonal. Figure 5 shows a typical rotated beam system where δ = the rotation angle of the LDV system in the xy plane or about the z axis where positive rotation is counterclockwise (looking from the left). δ = 0 defines an LDV system whose x-axis is parallel to the laboratory x-axis.

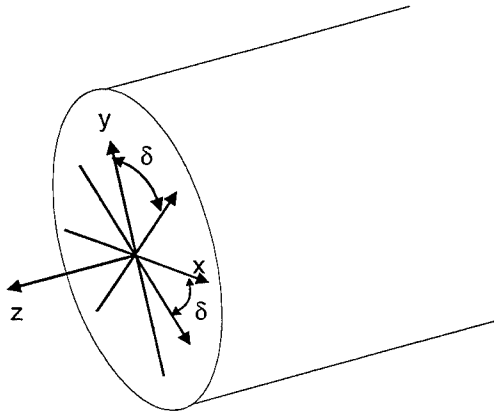


Figure 5. Rotation about the z-axis, or in the xy-plane.

The set of equations used to perform the coordinate transformation for a rotated system about the z-axis (rotating the system in the xy plane) and/or a tilted system about the x-axis (i.e tilting out of the xy-plane) are given below. Figure 6 shows the angles δ and ε used in these equations.

$$U = U_1 \cos\delta - U_2 \sin\delta \qquad V = U_1 \cos\varepsilon \sin\delta + U_2 \cos\varepsilon \cos\delta - U_3 \sin\varepsilon$$

$$W = U_1 \sin\varepsilon \sin\delta + U_2 \sin\varepsilon \cos\delta + U_3 \cos\varepsilon$$

where δ is described above and ε = the tilt of the system out of the xz-plane or about the x-axis where positive ε is a tilt above the x-z-plane and negative ε is below the x-z-plane. $\varepsilon = 0$ defines an LDV system whose axis is in the xz-plane. U, V, and W are the of the transformed velocities relative to the laboratory coordinate system.

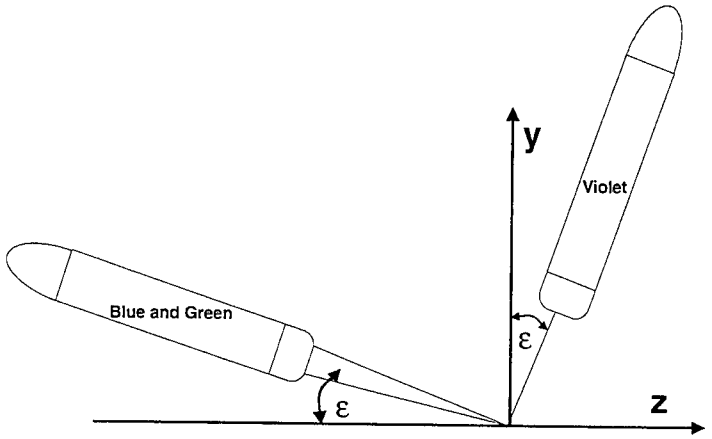


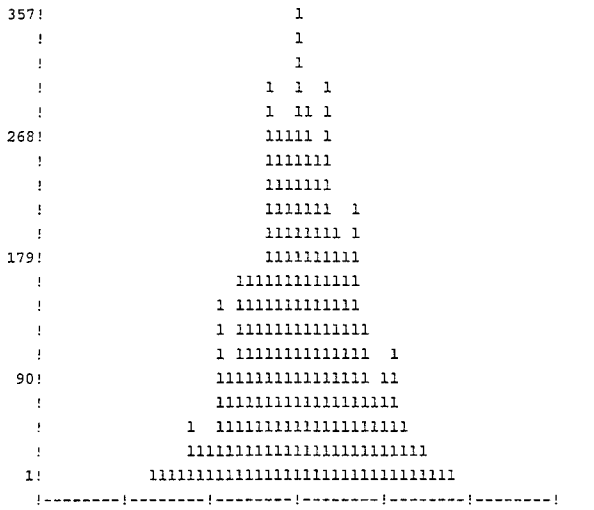
Figure 6. Tilt of the system about the x-axis.

Subroutine MAXMIN

Subroutine finds the maximum and minimum velocities for up to three velocity components. It is called after all transformations have been performed so that the actual maximum and minimum orthogonal velocity components are found. These values are used by the HISTOgram subroutines to print the raw data PDFs and also by the CUT_OFF subroutine.

Subroutines HISTO1, HISTO2 and HISTO3

Subroutines constructs a histograms 20 units high with 56 bins for one and two component velocity measurements and with 36 bins for three component measurements. The actual maximum and minimum velocities in the distribution are printed on each end of the graph. HISTO1, HISTO2 or HISTO3 are called for one-component, two-component or three-component measurements respectively. The total width of the one component formatted histogram is 80 columns while the total width of the two and three-component formatted histograms is 132 columns. Examples of the histogram format and summary printout are given for a one-component data file, a two-component data file, and a three-component data file in Figures 7 through 9, respectively.



file name - test.R00 4096 samples 1 channel

x coor.(mm)	.000	total time(s)	.92	sample rate(hz)	4434.3	mode: coincident	
y coor.(mm)	.000	noise pts	0	filtered samples	4069	tbd: tbd on	
z coor.(mm)	.000	3.0 sigma pts	27	rotation(deg)	.000	tilt(deg)	.000

u mean(m/s)	30.414	uuu (m3/s3)	.0136
+-	.023	+-	.0446
u std dev(m/s)	.7323	uu (m2/s2)	.5363
+-	.0158	+-	.0232
u local ti	.0241		

Figure 7. Histogram and summary page example for one-component measurements.

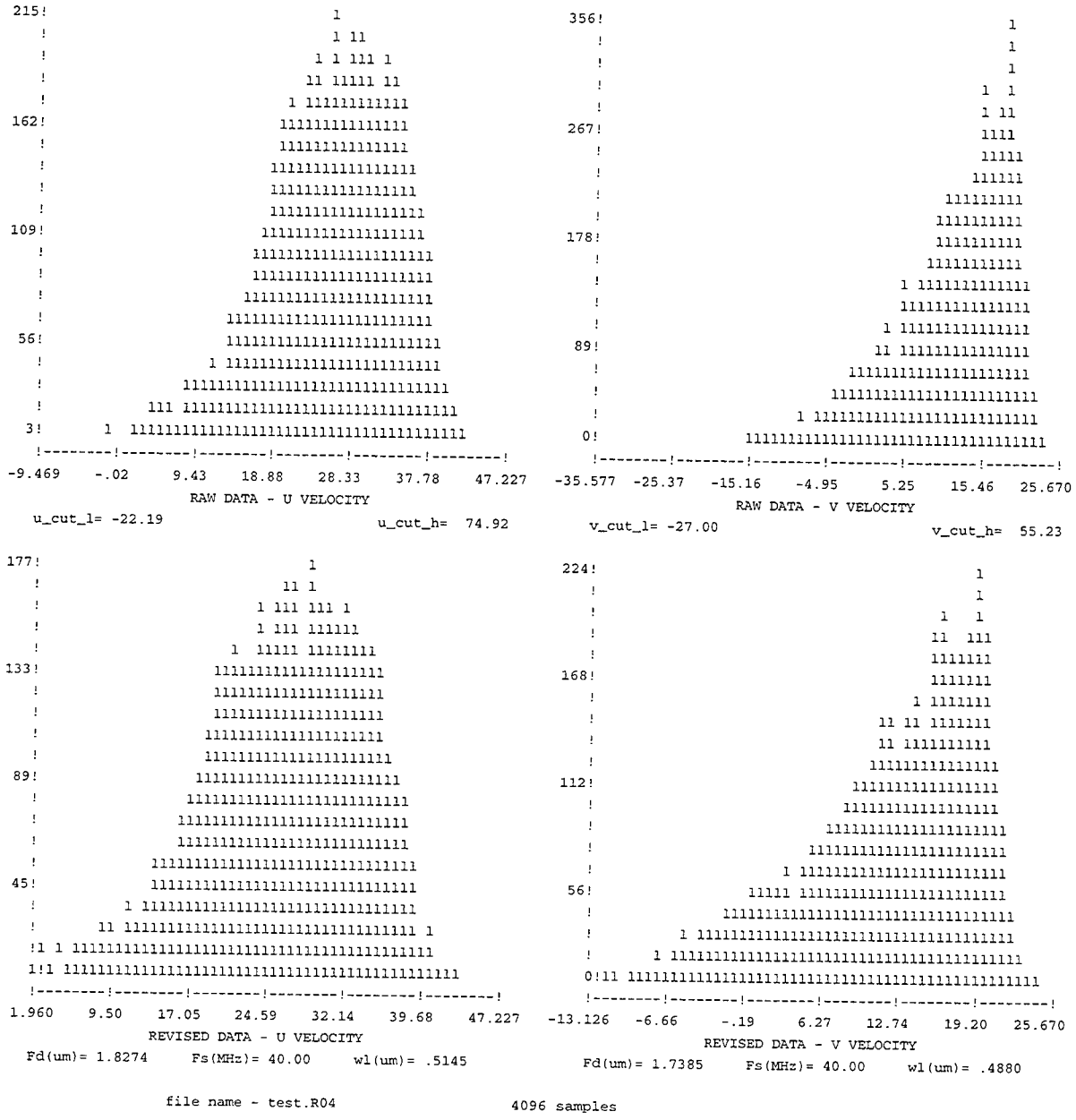


Figure 8. Histogram and summary page example for two-component measurements.

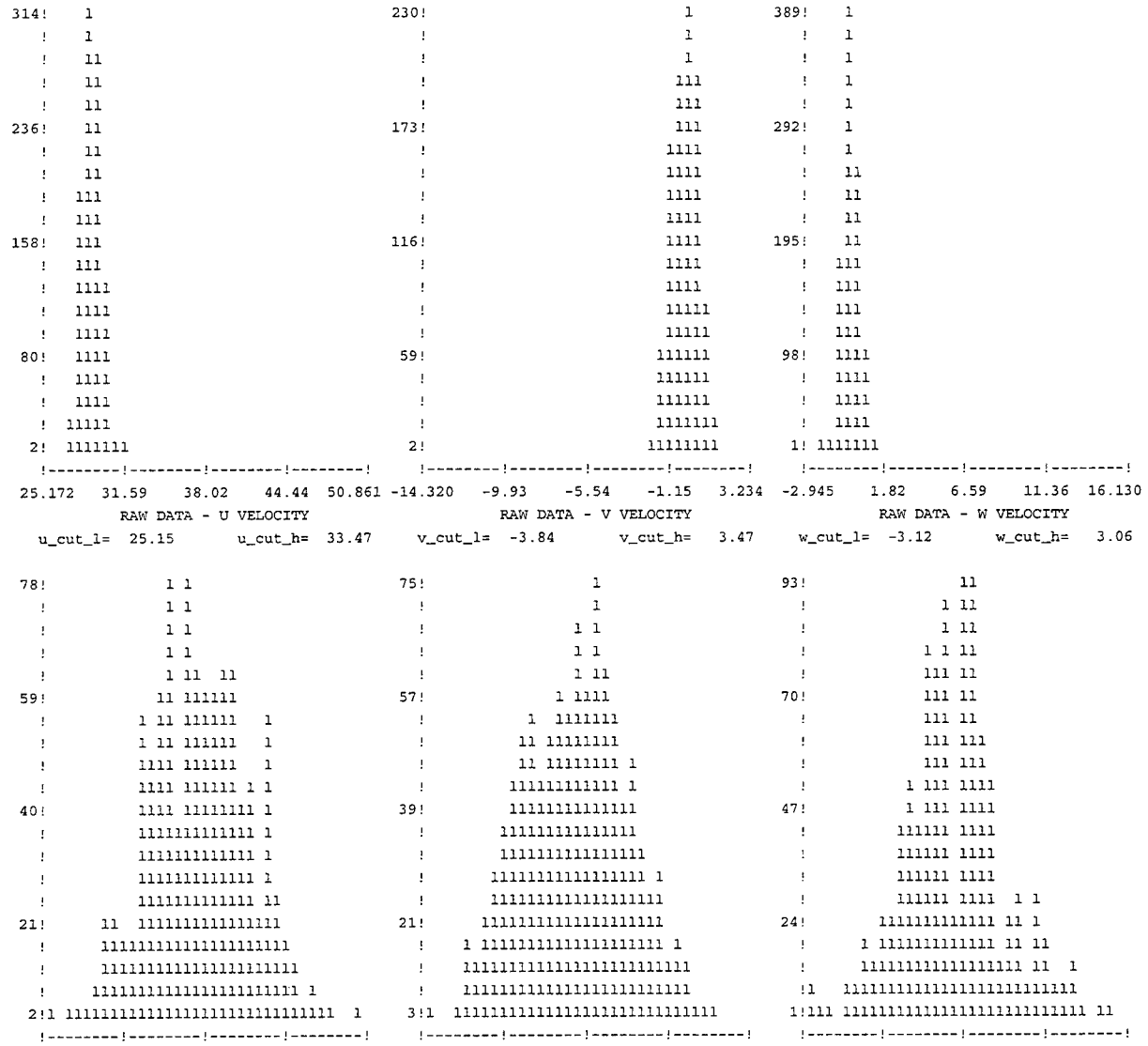


Figure 9. Histogram and summary page example for three-component measurements.

Subroutine CUT_OFF

Subroutine numerically integrates the area under the raw data PDFs to find an estimate of the mean and standard deviation of the raw data. Only histogram bins with more than 5% of the total number of samples are used in this integration so that spurious data do not corrupt the estimates. Upper and lower cutoff limits were set by adding and subtracting, respectively, 2.5 times the half-width of the data which met this 5 % threshold. Applying this technique to a Gaussian distribution is equivalent to setting cutoff limits which correspond to ± 4.1 standard deviations and thus its application removes only spurious data. This method is a variation of the method suggested by Meyers (1988). For a properly operating LDV very few data points should be discarded by this subroutine. This subroutine finds high and low cutoff values for each velocity component. These are used by subroutine STAT1 to discard noise outliers.

Subroutine STAT1

Subroutine calculates the preliminary mean and standard deviations for up to three velocity components. First, any noise points which lie outside of the high and low limits estimated by subroutine CUT_OFF are discarded. Note that if a noise point from one component is discarded its sister components for multiple component measurements are also discarded. The number of discarded points is also stored in variable, INOISE. Lastly, new high and low limits are defined by multiplying SIGMA (defined using menu option 3 in subroutine DEFAULT) by the standard deviation of each velocity component after the noise points have been discarded.

Subroutine STAT2

Subroutine calculates the revised ensemble averaged turbulence statistics. Mixed moments up to third order and homogenous moments up to fourth order are calculated. Points lying outside of the high and low limits defined in subroutine STAT1 have been removed prior to these calculations. 95% confidence bounds for the mean velocities are calculated once the variance of the revised data is known. New maximum and minimum velocities are found which are then used by the HISTOgram subroutine. Finally, the velocity and current time arrays are reordered to include only revised data. If time information was stored in the raw data file it is used to calculated the sample time and average data rate. Note that all the summed moments are stored as common variables in COMMON /SUMS/ for use by the jackknife subroutine.

Subroutine JACKKNIFE

Subroutine calculates the 95% statistical uncertainty of all turbulence statistics using a resampling method called the Jackknife method. It uses the summed values of all the revised statistics calculated in subroutine STAT2 as a starting point. One data point is removed from the data set and the summed moment of each statistic using this new data set is found. This is repeated until the summed moments of all the statistics of interest are calculated when each of the isamp data points is removed one at a time from the data set. The uncertainty is found by finding the variance of all these summed moments of the new data sets with one data point removed relative to the summed moments of the entire sample set. The summed moments of the original data set were stored to save computational effort. A brief description of the jacknife method is given below. For more details see Benedict and Gould (1996).

Given a data set $\mathbf{x} = (x_1, x_2, \ldots x_N)$ and some statistical estimator, $\hat{\theta}$, determined from this original data set $\mathbf{x}$, the jackknife makes use of the N data subsets that leave out one measurement at a time from the original data set giving $i = 1,2,\ldots N$ new jackknife data sets

$$\mathbf{x}_{jack,i} = (x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots x_N)$$

called *jackknife samples*. These new data sets are used to form N *jackknife replications,* $\hat{\theta}_{jack,i}$, of the statistical estimator of interest, $\hat{\theta}$. The jackknife estimate of variance for $\hat{\theta}$ is defined as

$$\mathrm{var}\left(\hat{\theta}\right)_{jack} = \frac{N-1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_{jack,i} - \overline{\hat{\theta}_{jack}}\right)^2$$

where

$$\overline{\hat{\theta}_{jack}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_{jack,i}$$

and an approximate 95% confidence interval for the estimator, $\hat{\theta}$, is then given by $\hat{\theta} \pm 1.96\left[\mathrm{var}\left(\hat{\theta}\right)_{jack}\right]^{1/2}$.

The jackknife in its general form requires N^2 calculations per variance estimate which would require roughly the computing time of a power spectrum *without* benefit of the fast Fourier transform. At the expense of generality, however, the jackknife may be structured so as to reduce the number of calculations from N^2 to N. This is a very significant reduction in computing time (when N is measured in thousands) and, in many cases, is well worth the added programming chore. This reduction in computation is accomplished by expanding the individual jackknife centralized moment statistics to create a series of summations composed entirely of

noncentralized variables. A simple example is afforded by the jackknife estimate of the variance for $\overline{u^2}$ which would be written as

$$\hat{\theta}_{jack,i} = \overline{u^2_{jack,i}} = \frac{1}{N-1}\sum_{\substack{j=1 \\ j\neq i}}^{N}\left(U_j - \overline{U_{jack,i}}\right)^2$$

where the summation is over $N-1$ samples since each jackknife replication leaves out one sample from the data set. The equation above may then be recast as

$$\overline{u^2_{jack,i}} = \frac{1}{N-1}\left[\sum_{\substack{j=1 \\ j\neq i}}^{N} U_j^2 - 2\overline{U_{jack,i}}\sum_{\substack{j=1 \\ j\neq i}}^{N} U_j + (N-1)\left(\overline{U_{jack,i}}\right)^2\right]$$

The individual terms are then summed only once over $j = 1, 2, \ldots, N$ for each overall variance estimate and merely decremented by the appropriate power of U_j for each jackknife replication. Note that the mean for each jackknife data set, $\overline{U_{jack,i}}$, is itself a noncentralized variable so that it too requires only one sum variable in order to determine its N values. This procedure may be carried out for any statistic based on central moments. The cost, of course, being that a single jackknife algorithm no longer applies to every statistic. New sum variables must be programmed for each new statistic of interest.

Subroutine SUMMARY

Subroutine prints a formatted summary table including the file name, number of data points in the file, x, y, z location, LDV settings, signal processor settings, and all the revised turbulence statistics and 95% uncertainties to the histogram and summary file: "family".PRT. In its current use, subroutine SUMMARY is called after the revised histograms are printed so that the histograms (both raw data and revised data) and summary table for each data point occupy only one page consisting of 66 lines. The maximum width of the summary table is 132 columns. Examples of the summary printout are given for a one-component data file, a two-component data file, and a three-component data file in Figures 7 through 9, respectively.

Subroutine WRITEFIL

Subroutine writes normalized turbulence statistics to files connected to unit numbers passed to it. The statistics are written to three data files so that the length of each one is limited. Mean statistics are written to unit n1, second order moments are written to unit n2, and third order moments are written to unit n3. Table 1 summarizes the file allocation and naming scheme. The statistics are all normalized with the reference velocity defined in subroutine DEFAULT prior to

writing in this subroutine. If non-normalize values are needed be sure that the reference velocity is set to 1.0 in subroutine DEFAULT. The write and format statements used for writing these statistics for one-component, two-component or three-component measurements are listed below.

```
c
c     Write 1 component statistics to three statistics files
c

      write(n1,100) x,y,z,uref,ubar/uref,u95/uref,ufrac
  100 format(3f8.3,4(1x,1pe11.4))
      write(n2,110) x,y,z,uref,stdev_u/uref,ustd95j/uref
     &,stdev_u*stdev_u/ur2,uu95j/ur2,ti_u
  110 format(3f8.3,6(1x,1pe11.4))
      write(n3,120) x,y,z,uref,uuu/ur3,uuu95j/ur3
  120 format(3f8.3,3(1x,1pe11.4))
c
c     Write 2 component statistics to three statistics files
c
      write(n1,200) x,y,z,uref,ubar/uref,u95/uref,ufrac,vbar/uref
     &,v95/uref,vfrac
  200 format(3f8.3,7(1x,1pe11.4))
      write(n2,210) x,y,z,uref,stdev_u/uref,ustd95j/uref
     &,stdev_u*stdev_u/ur2,uu95j/ur2,stdev_v/uref,vstd95j/uref
     &,stdev_v*stdev_v/ur2,vv95j/ur2,uv/ur2,uv95j/ur2
  210 format(3f8.3,11(1x,1pe11.4))
      write(n3,220) x,y,z,uref,uuu/ur3,uuu95j/ur3,uuv/ur3
     &,uuv95j/ur3,uvv/ur3,uvv95j/ur3,vvv/ur3,vvv95j/ur3
  220 format(3f8.3,9(1x,1pe11.4))

c     write 3 component statistics to three statistics files
c
      write(n1,300) x,y,z,uref,ubar/uref,u95/uref,ufrac,vbar/uref
     &,v95/uref,vfrac,wbar/uref,w95/uref,wfrac
  300 format(3f8.3,10(1x,1pe11.4))
      write(n2,310) x,y,z,uref,stdev_u*stdev_u/ur2,uu95j/ur2
     &,stdev_v*stdev_v/ur2,vv95j/ur2,stdev_w*stdev_w/ur2,ww95j/ur2
     &,uv/ur2,uv95j/ur2,uw/ur2,uw95j/ur2,vw/ur2,vw95j/ur2
  310 format(3f8.3,13(1x,1pe11.4))
      write(n3,320) x,y,z,uref,uuu/ur3,uuu95j/ur3,uuv/ur3
     &,uuv95j/ur3,uvv/ur3,uvv95j/ur3,vvv/ur3,vvv95j/ur3
     &,uuw/ur3,uuw95j/ur3,uww/ur3,uww95j/ur3
     &,vvw/ur3,vvw95j/ur3,vww/ur3,vww95j/ur3,www/ur3,www95j/ur3
  320 format(3f8.3,19(1x,1pe11.4))
```

Subroutine MT_STAT

Subroutine calculates McLaughlin-Tiederman(1973) velocity bias corrected turbulence statistics. Note that the sum variables used by subroutine STAT2 are reused here and are no longer valid for standard ensemble averaged statistics calculations. This subroutine called only if IMT is set to 1 in subroutine DEFAULT.

Subroutine RT_STAT

Subroutine calculates Hoesel-Rodi(1977) residence time velocity bias corrected turbulence statistics. Note that the sum variables used by subroutine STAT2 are reused here and are no longer valid for standard ensemble averaged statistics calculations. This subroutine is called only if IRT is set to 1 in subroutine DEFAULT.

Subroutine TBD_STAT

Subroutine calculates Barnett and Bentley(1974) time between data velocity bias corrected turbulence statistics. Note that the sum variables used by subroutine STAT2 are reused here and are no longer valid for standard ensemble averaged statistics calculations. This subroutine is called only if ITBD is set to 1 in subroutine DEFAULT.

Subroutine MAKE_PS

Reads ASCII file of histogram and statistics summary ("family".PRT) produced by subroutines HISTO1, HISTO2, HISTO3 and SUMMARY and creates a postscript language file having the same family name, but with an .EPS extension.

Subroutine MAKE_PCL

Reads ASCII file of histogram and statistics summary ("family".PRT) produced by subroutines HISTO1, HISTO2, HISTO3 and SUMMARY and creates an HP PCL 5 formatted file having the same family name, but with an .PCL extension.

Subroutines WORDS_2, WORDS_3, WORDS_4, WORDS_5, WORDS_6, WORDS_7

Subroutines read 2,3,4,5,6 or 7 LDV data words for each velocity realization stored in the TSI IFA750 raw data file according to the documentation given in the Appendix of the TSI FIND software manual (Version 4). This subroutine was adapted from the source code provided by TSI with the FIND software distribution. Many modifications where made including writing velocities and time to arrays as opposed to reading the raw data file each time a new statistic is calculated. In addition, calculation and storage of the velocity as opposed to the Doppler frequency, storage of the current time as opposed to the time between data and calculation and storage of the residence time were added. Only the first 5120 points in the raw data file are read in this implementation of the program so that the arrays are not overwritten and also so that the final executable program can be run under DOS using less than 480 kilobytes. Table 5 lists the

raw data word pattern for each of these subroutines. These subroutines are called by subroutine READDATA and call Subroutine CONVERTA and functions CONVERTB and CONVERTR.

Function ATOF20

Function converts a 20 length ASCII character string to a single precision (real*4) floating point number. This program was adapted from the one provided by TSI in the FIND Version 4.0 software distribution. It has been modified to work with the Microsoft Version 5.1 FORTRAN compiler and also has been modified to handle exponential notation. This function is called by subroutine READHEADER.

Function ATOI20

Function converts a 20 length ASCII character string to an integer (int*2) number. This program was adapted from the one provided by TSI in the FIND version 4.0 software distribution. This function is called by subroutine READHEADER.

Subroutine CONVERTA

Subroutine breaks down the "AWORD" transferred by the TSI IFA 750 to obtain the number of cycles in the Doppler burst (bits 0-7) and the processor address (bits 8-9). This subroutine was adapted from the source code provided by TSI with the FIND version 4.0 software distribution. It is called by subroutines WORDS_2,WORDS_3,...WORDS_7.

Function CONVERTB

Function converts the "BWORD" transferred by the TSI IFA 750 which is composed of a 12 bit mantissa and a 4 bit exponent into the time a particle takes to traverse NCYLCLEs in the Doppler burst. Ncycles is obtained from the AWORD using subroutine CONVERTA. This subroutine was adapted from the source code provided by TSI with the FIND version 4.0 software distribution. It is called by subroutines WORDS_2,WORDS_3,...WORDS_7.

Function CONVERTR

Function converts the residence time word transferred by the TSI IFA 750 which is composed of a 12 bit mantissa and a 4 bit exponent. Documentation for the structure of this word can be found in the Appendix of the TSI FIND version 4.0 software manual. The scale factor give the residence time in nanoseconds.

4. REFERENCES

Barnett, D.O. and Bentley, H.T., (1974) "Statistical Bias of Individual Realization Laser Velocimeters," Engineering Extension Series (Purdue University), **144**, *Proc. of the Second International Workshop on Laser Velocimetry*, pp. 428-444.

Benedict, L. H. and Gould R. D., (1996) "Uncertainty Estimates for Any Turbulence Statistic,"L. H. accepted in the *Proceedings of the Eighth International Symposium on Applications of Laser Techniques to Fluid Mechanics*, Lisbon, Portugal, July 8-11, 1996.

Hoesel, W. and Rodi, W., (1977) "New Biasing Elimination Method for Laser Doppler Velocimeter Counter Processing," *Rev. Sci. Instrum.*, **48**, pp. 910-919.

McLaughlin, D.K. and Tiederman, W.G., (1973) "Bias Correction for Individual Realization of Laser Anemometer Measurements in Turbulent Flows," *Physics of Fluids*, **16**, pp. 2082-2088.

Meyers, J. F., (1988) "Laser Velocimeter Data Acquisition and Real Time Processing using a Microcomputer," Proceedings of the 4th International Symposium on Applications of Laser Anemometry to Fluid Mechanics, Lisbon, Portugal, p. 7.20.

APPENDIX

SOURCE CODE LISTING OF MAIN PROGRAM TSISTAT

```
      program tsistat                                                    tsistat001
c                                                                        tsistat002
c*********************************************************************** tsistat003
c     Program reads TSI IFA750 acquired raw laser Doppler velocimetry    tsistat004
c     files and calculates turbulence statistics, prints hisograms of   tsistat005
c     velocity PDFs, allows for 3 velocity bias corrections, allows for  tsistat006
c     non-orthogonal beam correction, writes statistics to data files    tsistat007
c     and calculates the statistical uncertainties for all quantities    tsistat008
c     using the jackknife method. Batch file processing is also possible tsistat009
c                                                                        tsistat010
c     Developed by: Richard D. Gould                                     tsistat011
c                   Feb. 13, 1996                                        tsistat012
c                   version 1.0, All rights reserved by author           tsistat013
c*********************************************************************** tsistat014
c                                                                        tsistat015
      real*8 u,v,w,ttu,ttv,ttw,u_min,u_max,v_min,v_max,w_min,w_max       tsistat016
     &      ,u_cut_l,v_cut_l,w_cut_l,u_cut_h,v_cut_h,w_cut_h             tsistat017
      real*4 df1,df2,df3,fs1,fs2,fs3,wlen1,wlen2,wlen3,del,eps,x,y,z     tsistat018
     &      ,sigma,tmax,srate,coinwind,samptim,uref,angle_u,angle_v      tsistat019
     &      ,angle_w                                                     tsistat020
      integer*4 curtime,nokdp,nowpdp,c,dpoints,isamp,inoise,i3sig        tsistat021
      integer*2 numctr,ctype1,ctype2,ctype3,nmlen,nstart,nend,numfil     tsistat022
     &         ,jfile,ifile,iover,irot,hunit,inorm,iprint,ips,ipcl,imt  tsistat023
     &         ,irt,itbd,u10,u11,u12,u13,u14,u15,u16,u17,u18,u19,u20,u21 tsistat024
      character*3 extname                                                tsistat025
      character*12 sfname,fnamein                                        tsistat026
      character*1 tbd,mode,ttime,axis,in                                 tsistat027
      character*40 t1,t2,t3,t4,t5,t6                                     tsistat028
c                                                                        tsistat029
      common /ldvdat/ u(5120),v(5120),w(5120),ttu(5120),ttv(5120)        tsistat030
     &               ,ttw(5120),curtime(5120),dpoints                    tsistat031
      common /limits/ u_min,u_max,v_min,v_max,w_min,w_max,u_cut_l        tsistat032
     &               ,v_cut_l,w_cut_l,u_cut_h,v_cut_h,w_cut_h,sigma      tsistat033
     &               ,tmax,srate,isamp,inoise,i3sig                      tsistat034
      common /ldvset/ df1,df2,df3,fs1,fs2,fs3,wlen1,wlen2,wlen3,del,eps  tsistat035
      common /processor/ nowpdp,c,nokdp,numctr,ctype1,ctype2,ctype3      tsistat036
     &                  ,samptim,coinwind,mode,tbd,ttime,axis            tsistat037
      common /defaults/ uref,angle_u,angle_v,angle_w,iover,irot,hunit    tsistat038
     &                 ,inorm,iprint,ips,ipcl,imt,irt,itbd,t1,t2,t3,t4  tsistat039
     &                 ,t5,t6                                           tsistat040
      common /position/ x,y,z                                            tsistat041
      common /files/ nmlen,sfname,fnamein,extname                        tsistat042
      data u10,u11,u12,u13,u14,u15,u16,u17,u18,u19,u20,u21                tsistat043
     & /10,11,12,13,14,15,16,17,18,19,20,21/                             tsistat044
c                                                                        tsistat045
*********************************************************************** tsistat046
c     begin main program                                                 tsistat047
c*********************************************************************** tsistat048
c                                                                        tsistat049
      call input                                                         tsistat050
      call openfile                                                      tsistat051
      call startbatch(nstart,nend,numfil,in)                             tsistat052
      if(in .eq. 'y ') go to 1                                           tsistat053
      if(in .eq. 'Y ') go to 1                                           tsistat054
      go to 900                                                          tsistat055
    1 ifile=0                                                            tsistat056
      call default                                                       tsistat057
c                                                                        tsistat058
c*********************************************************************** tsistat059
c     do loop 1000 controls batch processing                             tsistat060
c*********************************************************************** tsistat061
c                                                                        tsistat062
```

```
      do 1000 jfile=nstart,nend                                          tsistat063
      call nextfile(jfile,ifile)                                         tsistat064
      call readheader                                                    tsistat065
      if(iover .eq. 1) call override                                     tsistat066
      call readdata                                                      tsistat067
      if(irot .eq. 1 .and. numctr .gt. 1) call rotate                    tsistat068
      if(axis .eq. 1) call transform                                     tsistat069
      call maxmin                                                        tsistat070
c                                                                        tsistat071
c     print histograms of raw data                                       tsistat072
c                                                                        tsistat073
      if(numctr .lt. 3) write(hunit,'(a)') ' '                           tsistat074
      if(numctr .eq. 1 .and. mode .eq. '0') call histo1(t1,hunit)        tsistat075
      if(numctr .eq. 2 .and. mode .eq. '0') call histo2(t1,t2,hunit)     tsistat076
      if(numctr .eq. 3 .and. mode .eq. '0') call histo3(t1,t2,t3,hunit)  tsistat077
      if(mode .eq. '1') call histo1(t1,hunit)                            tsistat078
c                                                                        tsistat079
      call cut_offs(hunit)                                               tsistat080
      call stat1                                                         tsistat081
      call stat2                                                         tsistat082
c                                                                        tsistat083
c     print histograms of revised data                                   tsistat084
c                                                                        tsistat085
      if(numctr .eq. 1 .and. mode .eq. '0') call histo1(t4,hunit)        tsistat086
      if(numctr .eq. 2 .and. mode .eq. '0') call histo2(t4,t5,hunit)     tsistat087
      if(numctr .eq. 3 .and. mode .eq. '0') call histo3(t4,t5,t6,hunit)  tsistat088
      if(mode .eq. '1') call histo1(t1,hunit)                            tsistat089
c                                                                        tsistat090
      call jackknife                                                     tsistat091
      call summary(hunit)                                                tsistat092
      call writefil(u10,u11,u12)                                         tsistat093
      if(imt .eq. 1) call mt_stat                                        tsistat094
      if(imt. eq. 1) call writefil(u13,u14,u15)                          tsistat095
      if(irt .eq. 1 .and. ttime .eq. '1') call rt_stat                   tsistat096
      if(irt. eq. 1 .and. ttime .eq. '1') call writefil(u16,u17,u18)     tsistat097
      if(itbd .eq. 1 .and. tbd .eq. '1') call tbd_stat                   tsistat098
      if(itbd. eq. 1 .and. tbd .eq. '1') call writefil(u19,u20,u21)      tsistat099
c                                                                        tsistat100
 1000 continue                                                           tsistat101
  900 close(7)                                                           tsistat102
      close(10)                                                          tsistat103
      close(11)                                                          tsistat104
      close(12)                                                          tsistat105
      if(imt .eq. 1) then                                                tsistat106
      close(13)                                                          tsistat107
      close(14)                                                          tsistat108
      close(15)                                                          tsistat109
      endif                                                              tsistat110
      if(irt .eq. 1) then                                                tsistat111
      close(16)                                                          tsistat112
      close(17)                                                          tsistat113
      close(18)                                                          tsistat114
      endif                                                              tsistat115
      if(itbd .eq. 1) then                                               tsistat116
      close(19)                                                          tsistat117
      close(20)                                                          tsistat118
      close(21)                                                          tsistat119
      endif                                                              tsistat120
      if(inorm .eq. 3) close(22)                                         tsistat121
c                                                                        tsistat122
      if(ips .eq. 1) call make_ps                                        tsistat123
      if(ipcl .eq. 1) call make_pcl                                      tsistat124
c                                                                        tsistat125
      stop                                                               tsistat126
      end                                                                tsistat127
```

Band Selection and Performance Analysis for Multispectral Target Detectors Using Truthed Bomem Spectrometer Data

Russell C. Hardie
Assistant Professor
Department of Electrical Engineering

and

Ajay Kanodia
Graduate Research Assistant
Department of Electrical Engineering

University of Dayton
Dayaton, OH 45469

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

University of Dayton

December 1995

BAND SELECTION AND PERFORMANCE ANALYSIS FOR MULTISPECTRAL TARGET DETECTORS USING TRUTHED BOMEM SPECTROMETER DATA

Ajay Kanodia
Graduate Research Assistant

Russell C. Hardie
Assistant Professor
Department of Electrical Engineering
University of Dayton

Abstract

This report investigates the problem of spectral band selection for multispectral target detection techniques. Based on truthed high-resolution spectrometer data, a detailed empirical study is performed. Both detector performance and optimal band selection is analized for a wide range of targets and background pairs under a variety of conditions. In particular, we investigate multispectral detector performance as a function of spectral band selection and other factors such as noise, spectral bandwidth, time of day, and even season. Bomem spectrometer data is used exclusively in this study. The Bomem spectrometer collects data in the range of 2.86 μm to 14.32 μm distributed over 728 spectral bands.

The multispectral detectors used are based on Bayes classifiers. Therefore, the selection of optimal spectral bands is investigated using the Mahalanobis distance and Bhattacharyya distance criteria. We believe that these are appropriate quantitative measures of target/background class separability for Bayes classifiers. Given the truthed Bomem data, these metrics provide a systematic method for performing spectral band selection and providing quantitative measures of detector performance.

BAND SELECTION AND PERFORMANCE ANALYSIS FOR MULTISPECTRAL TARGET DETECTORS USING TRUTHED BOMEM SPECTROMETER DATA

Ajay Kanodia

Russell C. Hardie

1 Introduction

Passive multispectral imaging techniques can be very useful for target detection and surveillance. Such techniques rely on differential reflectances and emissivities between a target and background over a discrete set of wavelengths. Multispectral techniques can be used for large area target searches in a variety of backgrounds. These techniques will likely be called upon where single band systems perform poorly or where a high level of automation is required. New sensor technology is continuously advancing and high spectral resolution imaging systems are being developed. Such data should allow for discrimination and detection of a much larger number of target classes than with low spectral resolution data. Thus, algorithm development for multispectral target detection and identification must keep pace with advancing sensor technology. A good overview of basic multispectral imaging techniques can be found in [12].

One effective type of imaging system used in aircraft today for target detection is a forward looking infrared (FLIR) system. These systems utilize broad-band infrared imagery. Figure 1 shows a classification of the electro-magnetic spectrum for reference. FLIR systems employ a scanning or staring sensor array which produces a single frame intensity image. This imagery generally provides good target discrimination in both day and night. However, during broad-band thermal crossover, when the background and target have similar temperatures, the target may not be easily detectable. In addition, single band systems rely heavily on the spatial recognition capabilities of the human user. Thus, these systems generally have a low level of automation and require high spatial resolution. In many cases, multispectral information can dramatically improve the detectability of certain targets. Increased spectral resolution can compensate for lower spatial resolution in some cases. Exploitation of spatial and spectral properties of a target should provide the best performance.

Several basic problems exist in the exploitation of spectral information for target detection. One is spectral band selection. With the potential of having many spectral bands for target detection,

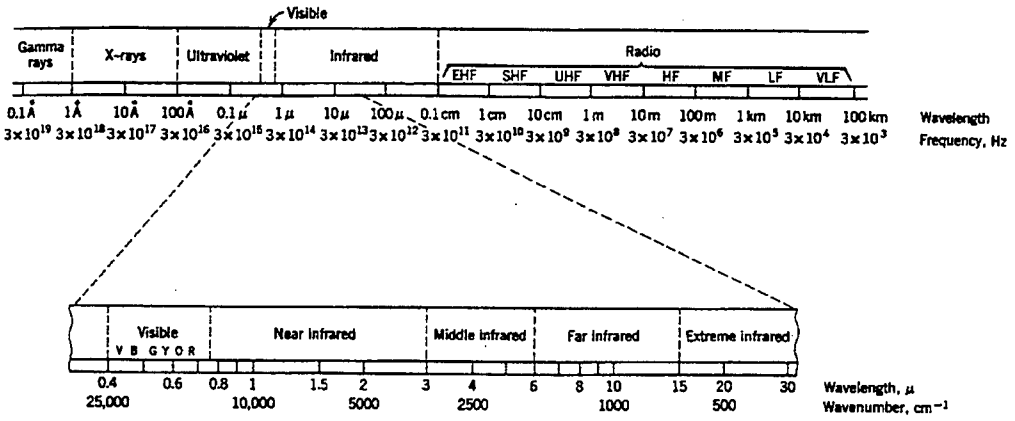


Figure 1: Electromagnetic spectrum classification.

it is important to find a manageable number of spectral bands which provide the greatest utility in discriminating targets and backgrounds of interest. The choice of spectral bands may depend critically of many factors. These include the target and background of interest, the level of sensor noise, spectral bandwidth, time of day, season and other factors. This report presents a detailed empirical study investigating many of these factors. Some previous research in multispectral band selection using statistical distance measures has been studied in [15]. Band selection specifically for classification of soil organic matter content using a Karhunen-Loeve based method is presented in [8]. General feature selection methods are discussed in [2]. Other types of multispectral feature extraction is presented in [1, 9, 10, 11]. Still other significant challenges exist for multispectral target detection. These include: atmospheric distortion and variation; solar illumination differences; spectral changes due to viewing angle; within scene sensor calibration; and scene-to-scene calibration. All of these issues must be addressed in order to fully exploit multispectral data.

Nonwithstanding these problems, automated multispectral target detection promises to be highly successful in a number of applications. One successful target detection approach uses an adaptive constant-false-alarm-rate (CFAR) multispectral detector. The adaptive CFAR detector is proposed and analyzed in [13, 14, 16]. This method, based on a Bayes classifier, shows promising results. Other detectors are described in [3].

The main focus of this report is the problem of spectral band selection for multispectral target detection techniques. Based on truthed high-resolution spectrometer data, a detailed empirical study is performed. Both detector performance and optimal band selection is analized for a wide

range of targets and background pairs under a variety of conditions. These conditions include: varying noise levels; different spectral bandwidths; different times of day; and different seasons. We explore and quantify many of these effects in an attempt to understand their relative importance and impact on band selection and detector performance. Bomem spectrometer data is used exclusively in this study. The Bomem spectrometer collects data in the range of 2.86 μm to 14.32 μm distributed over 728 spectral bands.

The type of classifier used will drive the band selection process. The selected bands and the feature space must contain the most relevant information for a given classifier. The multispectral detectors studied here are based on a Gaussian Bayes classifier and have been described in [3]. Therefore, the selection of optimal spectral bands is investigated using the Mahalanobis distance and Bhattacharyya distance criteria. We believe that these are appropriate quantitative measures of target/background class separability for Bayes classifiers. Given the truthed Bomem data, these metrics provide a systematic method for performing spectral band selection and providing quantitative measures of detector performance.

The remainder of this report is organized as follows. Section 2 describes the multispectral classifiers used in our analysis. The spectral band selection process using statistical distance metrics is described in Section 3. The results of our study are provided in Section 4. Specifically, the effect of noise, bandwith, time of day, season, and type of target and background is quantitatively evaluated. Finally, some conclusions are provided in Section 5.

2 Multispectral Target Detectors

In this section, the multispectral target detectors are described. More information can be found in [3]. Note that an excellent treatment of statistical pattern recognition is provided in [2].

Before discussing detectors, some notation must be introduced. Consider a passive multispectral sensor which collects an N-band multispectral image denoted $\mathbf{x}(n_1, n_2)$ where

$$\mathbf{x}(n_1, n_2) = [x_1(n_1, n_2), x_2(n_1, n_2), \ldots, x_N(n_1, n_2)]. \tag{1}$$

The indices n_1 and n_2 are spatial indices. In our analysis, the spatial samples do not lie on a regular grid and the indices n_1 and n_2 are not used explicitly. Thus, x_i for $i = 1, 2, \ldots, N$ is the radiance value for a single pixel from spectral band i (with some predetermined wavelength).

2.1 Linear and Quadratic Classifiers

The following analysis deals with a two class problem. It will be assumed that there is one target class and one background class. Let hypothesis H_b be that the observation vector x (a single multispectral observation vector) belongs to the background class and let hypothesis H_t be that x belongs to the target class. Each multispectral observation will be classified in this way creating a class map.

The Bayes classifier, which provides the minimum probability of classification error, is defined by the following likelihood ratio:

$$l(\mathrm{x}) = \frac{p(\mathrm{x}|H_t)}{p(\mathrm{x}|H_b)} \overset{H_t}{>} \frac{P_b}{P_t}. \tag{2}$$

The function $p(\mathrm{x}|H_b)$ and $p(\mathrm{x}|H_t)$ are the conditional probability density functions for the observed spectral vector x. The variables P_b and P_t are the *a priori* probabilities for H_b and H_t respectively. Equation (2) states that if the likelihood ratio $l(\mathrm{x})$ is greater than $\frac{P_b}{P_t}$, than one should declare that x belongs to the target class. Otherwise, x belongs to the background class. It is often useful to define the the minus-log-likelihood ratio from (2) yielding

$$-\ln\{l(\mathrm{x})\} = -\ln\{p(\mathrm{x}|H_b)\} + \ln\{p(\mathrm{x}|H_t)\} \overset{H_t}{<} \ln\frac{P_t}{P_b}. \tag{3}$$

Now consider the case where the target and background have Gaussian pdfs. Let the target mean in N spectral space be μ_t and the $N \times N$ covariance be Σ_t. The background mean and covariance will be denoted μ_b and Σ_b respectively. In this case, the decision rule in (3) becomes

$$h(\mathrm{x}) = \frac{1}{2}(\mathrm{x}-\mu_b)^T\Sigma_b^{-1}(\mathrm{x}-\mu_b) - \frac{1}{2}(\mathrm{x}-\mu_t)^T\Sigma_t^{-1}(\mathrm{x}-\mu_t) \overset{H_t}{>} \ln\frac{P_t}{P_b} - \frac{1}{2}\ln\frac{|\Sigma_b|}{|\Sigma_t|}. \tag{4}$$

A Gaussian spectral distribution model will be assumed here.

The decision boundary defined by (4) is, in general, a quadratic function in x. In the case where the covariance of the target and background are equal, that is $\Sigma_t = \Sigma_b = \Sigma$, then the decision boundary reduces to a linear function in x. The decision rule for this case becomes

$$h(\mathrm{x}) = \frac{1}{2}(\mu_t - \mu_b)^T\Sigma^{-1}\mathrm{x} \overset{H_t}{>} \ln\frac{P_t}{P_b} - \frac{1}{2}(\mu_b^T\Sigma^{-1}\mu_b - \mu_t^T\Sigma^{-1}\mu_t). \tag{5}$$

If $\mu_b = \mathbf{0}$, then the first term in (5) can be interpreted as a target spectral matched filter. The output of this matched filter is thresholded to perform classification. The spectral matched filter approach was used in [13, 14, 16]. However, this is suboptimal when the covariance matrices of the target and background are not equal. This result follows since the linear classifier does not exploit covariance differences.

2.1.1 The Bhattacharyya Distance

Computing the exact probability of error for a Bayes classifier may be very difficult in general. However, one relatively simple upper bound on the Bayes classification error is the *Bhattacharyya* bound given by

$$P_{error} \leq \sqrt{P_t P_b} e^{-B}, \tag{6}$$

where the parameter B is the *Bhattacharyya* distance (B-distance) given by

$$B = \frac{1}{8}(\mu_t - \mu_b)^T \left(\frac{\Sigma_t + \Sigma_b}{2}\right)^{-1} (\mu_t - \mu_b) + \frac{1}{2}\ln \frac{\left|\frac{\Sigma_t + \Sigma_b}{2}\right|}{\sqrt{|\Sigma_t||\Sigma_b|}}. \tag{7}$$

The B-distance itself is an excellent measure of class separability for the two class problem. The larger the B-distance, the more separable the two classes are and the better the performance one can expect with an optimal classifier. Since the basic advantage of multispectral data over single band data is increased class separability between a given target and background, it is important to have a good quantitative measure of this. The author proposes the adoption of this metric for evaluating the separability of various classes of materials and objects in spectral space.

The B-distance in (7) has two terms which will be separately denoted as:

$$B_1 = \frac{1}{8}(\mu_t - \mu_b)^T \left(\frac{\Sigma_t + \Sigma_b}{2}\right)^{-1} (\mu_t - \mu_b) \tag{8}$$

$$B_2 = \frac{1}{2}\ln \frac{\left|\frac{\Sigma_t + \Sigma_b}{2}\right|}{\sqrt{|\Sigma_t||\Sigma_b|}}. \tag{9}$$

The first term, B_1, disappears when $\mu_t = \mu_b$ and the second term, B_2, is zero when $\Sigma_t = \Sigma_b$. Therefore, B_1 reflects class separability due to the spectral mean difference between the target and the background. The term B_2 provides a measure of class separability due to covariance difference. Thus, the B-distance can be used to compare the performance of a linear classifier, which only exploits mean differences, to the optimal quadratic classifier, which exploits mean and covariance differences.

2.2 Single Hypothesis Detector

So far we have considered the case of a background and target class. To implement such a classifier some information about both classes must be known or estimated. In some cases very little may be known about the target *a priori*. In such cases, a single hypothesis detector (SHD) may be useful. In the SHD we only test one hypothesis: the multispectral observation belongs to the background class (H_b). If x is not declared a background observation, it may represent a target observation.

This method is not as efficient as the linear or quadratic classifier, particularly for higher spectral dimensions. However, for many reasons, this may be a highly practical alternative [13, 14].

The decision statistic for the SHD is the Mahalanobis distance (M-distance). The M-distance is defined as

$$M(\mathrm{x}) = (\mathrm{x} - \mu_b)^T \Sigma_b^{-1} (\mathrm{x} - \mu_b). \tag{10}$$

After the M-distance is calculated for an observation vector, a threshold is applied to perform classification. Thus, if the observation vector lies beyond a set M-distance away from the background mean, it is declared to be not background (and assumed to be target). The threshold is determined by the desired probability of the false alarm. In particular, the threshold for the desired probability of false alarm can be calculated using the pdf of the background class. For the Gaussian case, the density function of the M-distance in two-dimensional spectral space is given by exponential distribution function [2]

$$p_M(y) = e^{-\frac{y}{2}} u(y). \tag{11}$$

The probability that a multispectral observation from the background class has an M-distance less than a threshold T is given by

$$\int_0^T e^{-\frac{y}{2}} dy = 1 - P_{fa}. \tag{12}$$

Integrating (12), we get

$$1 - e^{-\frac{T}{2}} = 1 - P_{fa}. \tag{13}$$

Finally solving for T yields

$$T = -2 \log P_{fa}. \tag{14}$$

Applying this threshold to the M-distance, forms an elliptical decision boundary in the spectral observation space. All spectral observation vectors inside this boundary are declared to be background observations.

In addition to serving as a discriminant function for the SHD, the M-distance also provide a quantitative measure of class separability appropriate for the SHD. To use the M-distance in this fashion, the observation, x, in (10) is replaced by the target mean. A related metric for class separability is signal to clutter ratio (SCR) [13, 14, 4, 16]. SCR is simply the square root of the M-distance and can be written as

$$SCR = [(\mu_b - \mu_t)^T \Sigma_b{}^{-1} (\mu_b - \mu_t)]^{1/2}. \tag{15}$$

When $N = 2$, the output SCR, now termed as dual SCR can be written as

$$SCR = SCR_1 \left[(1-\rho^2)^{-1}(1-2\rho R + R^2)\right]^{1/2}, \tag{16}$$

where SCR_1 is signal to clutter ratio measured in the best of two single bands, R is called the target and background color ratio and is defined as $R = \frac{SCR_2}{SCR_1}$, with $|SCR_1| \geq |SCR_2|$.

To compute the B-distance, M-distance, or SCR, the class means and covariances are required. In most practical cases, these will have to be estimated from truthed training data. In this case the sample mean and covariance can be used. If P spectral vectors from each class are available, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P\}$, the sample mean estimate is given by

$$\hat{\mu} = \frac{1}{P}\sum_{k=1}^{P} \mathbf{x}_k. \tag{17}$$

The unbiased sample covariance estimate is given by

$$\hat{\Sigma} = \frac{1}{P-1}\sum_{k=1}^{P} (\mathbf{x}_k - \hat{\mu})^T(\mathbf{x}_k - \hat{\mu}). \tag{18}$$

A "rule of thumb" is to have $\mathcal{O}(N^2)$ samples, where N is the number of spectral bands, with which to make the mean and covariance estimates [2]. All of the class means and covariances used in this study are estimated is this way.

To illustrate the operation of the quadratic, linear and single hypothesis detector, a typical Bomem data set is analized in Fig. 2. Figure 2a shows a quadratic decision boundary for a two spectral band space. Here the target is a camouflaged truck and the background is soil. Multiple spatial samples for each are shown and the optimal boundary is calculated using the sample mean and covariance estimates from these data. A linear classifier decision boundary is illustrated in Fig. 2b for the same data. Finally, the elliptical decision boundary for the SHD is illustrated in Fig. 2c. In general the performance of the quadratic classifier exceeds that of a linear classifier which exceeds that of the SHD.

3 Data Acquisition and Spectral Band Selection

Our study is based on truthed data collected with the Bomem MB-100 Fourier Transform Spectrometer, which is shown in Fig. 3. A brief description of the measurement setup for data collection is included in this section. The band selection process using the M-distance and B-distance is also discussed.

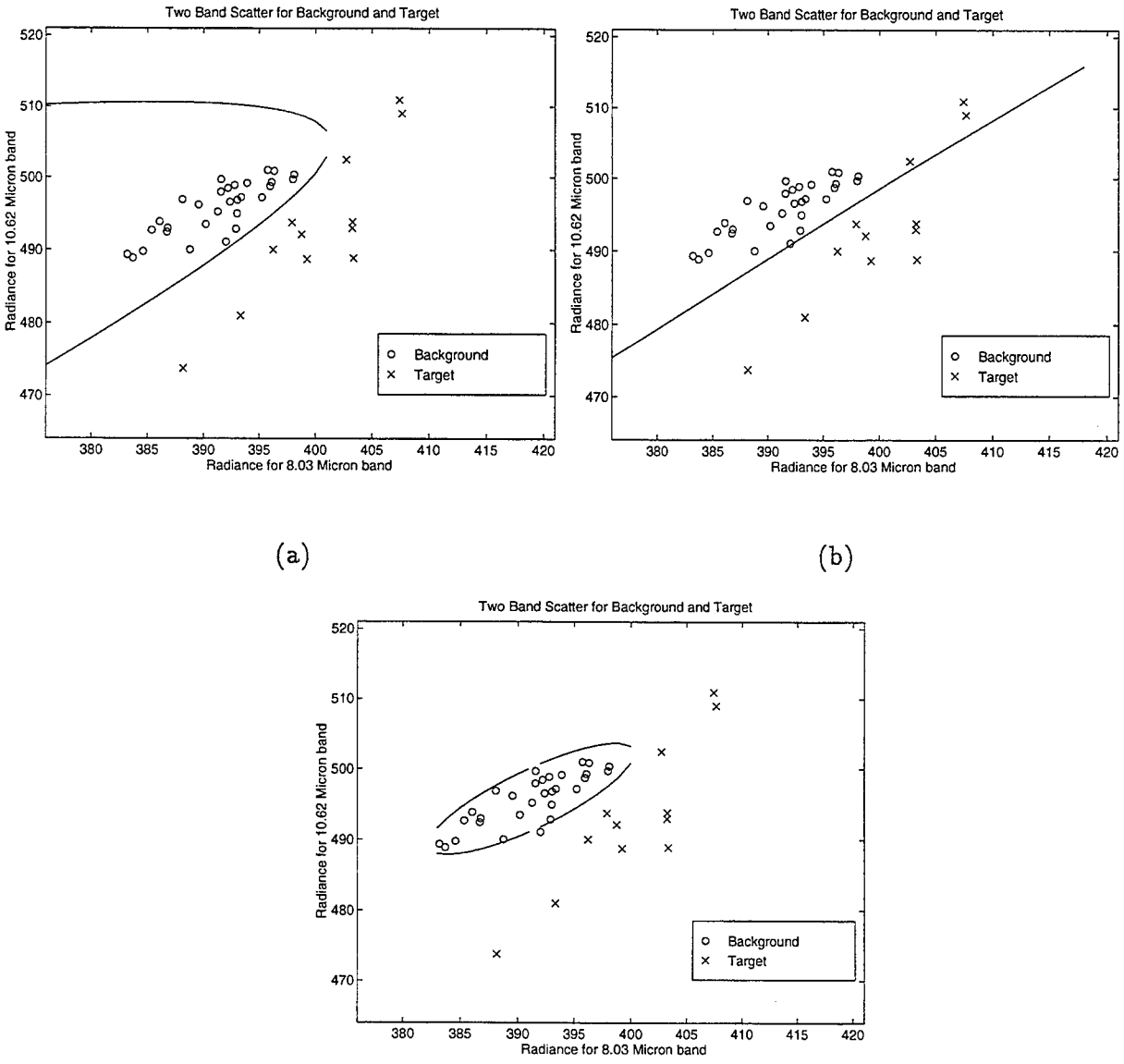


Figure 2: *Decision boundaries for the background and target pair of soil and camouflaged truck respectively. (a) Decision boundary for a quadratic classifier, (b) decision boundary for a linear classifier, (c) decision boundary for a single hypothesis detector.*

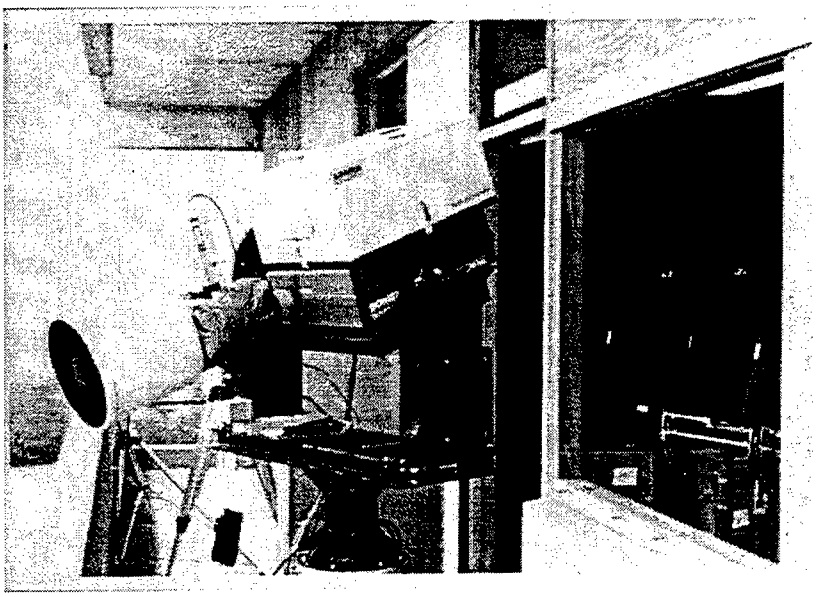


Figure 3: *Bomem MB-100 Fourier Transform Spectrometer at the WPAFB Avionics Laboratory tower.*

3.1 Bomem MT-100 Fourier Transform Spectrometer

The Bomem MB-100 FTS with telescope and 2-axis computer controlled pedestal operating from the WPAFB Avionics Laboratory tower is shown in Fig. 3. The computer not only works as a control unit but also acts as a data acquisition system. The system is automated and it takes approximately 5-10 minutes to collect 25 samples of the object under consideration. The system also consists of a bore sighted visible wavelength video camera with a tape recorder and two Electro Optical Industries model blackbody sources. These sources are used to calibrate the spectroscope before and after the data collection. The Bomem spectrometer is not an imaging spectrometer. It performs single spatial point measurements. However, multiple spatial measurements of each object are generally collected. The spectral resolution of the Bomem is very high which allows for a thorough study of band selection and multispectral target detector performance. A more detailed description of the device can be found in [4, 6].

The Bomem spectrometer collects data from a wavelength range of 2.86 μm to 14.32 μm, which corresponds to wave numbers 3502.09839 and 698.10553, respectively. There are 728 bands linearly distributed in the wave number. Since wave number are the reciprocal of the wavelength, the 728 bands are distributed non-linearly between the wavelengths as shown in Fig. 4(top). Band number 1 correspond to the wavelength 14.32 μm and Band number 728 correspond to the wavelength 2.86

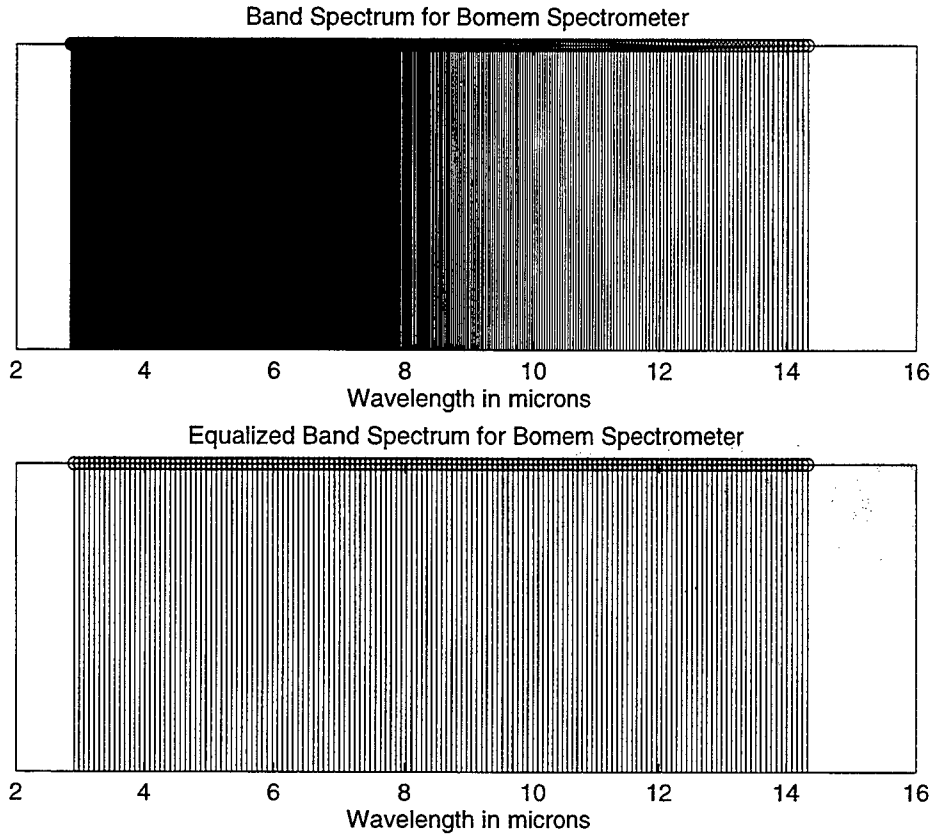


Figure 4: *(top) Spectral band distribution of the Bomem spectrometer before equalization. (bottom) Spectral band distribution of the Bomem spectrometer after equalization. One line represents a spectral band center wavelength.*

μm. To eliminate very narrow bands and reduce the total number of bands, the data are run through an equalization process. Here the data are averaged to make the bands uniformly distributed in wavelengths. After equalization the number of bands is reduced to 146. The equalized band center wavelengths are shown in the Fig. 4(bottom). An example to illustrate how the equalization effects actual observed spectra is provided in Fig. 5. The figure shows the mean radiance spectrum for spectral observations of scrub, before and after equalization. After equalization some of the narrow band features are smoothed.

3.2 Band Selection

In the band selection process, we wish to find a small subset of spectral bands which provide the best target/background discrimination. The optimal subset of spectral bands will depend on many factors including the type of classification algorithm used. For the Bayes classifiers, it is appropriate for the selection of the optimal bands to be based on the B-distance criterion. Thus, for a specific

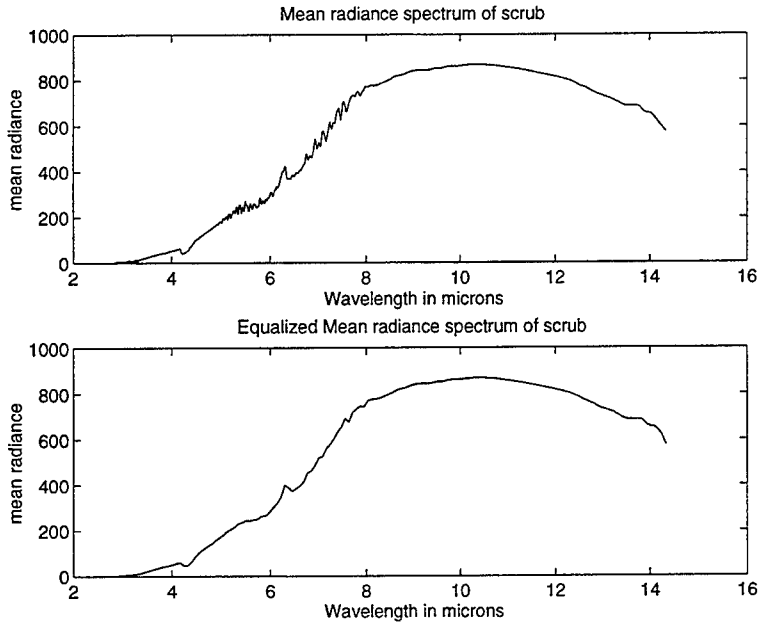


Figure 5: *(top) Mean radiance spectrum of scrub before equalization. (bottom) Mean radiance spectrum of scrub after equalization.*

truthed data set, we define the optimal subset of bands to be those which maximize the B-distance criterion. For the SHD, the M-distance criterion is appropriate. The hope is that these bands will perform well on other statistically similar untruthed data.

Here we focus on a two band analysis since it lends itself to visualization and it is relatively simple. Multiple band selection using a forward sequential technique is described in [3]. For the two band case, the B-distance can be calculated for all possible two band combinations for a given target and background using the equations

$$B(\lambda_1, \lambda_2) = B_1(\lambda_1, \lambda_2) + B_2(\lambda_1, \lambda_2), \tag{19}$$

where,

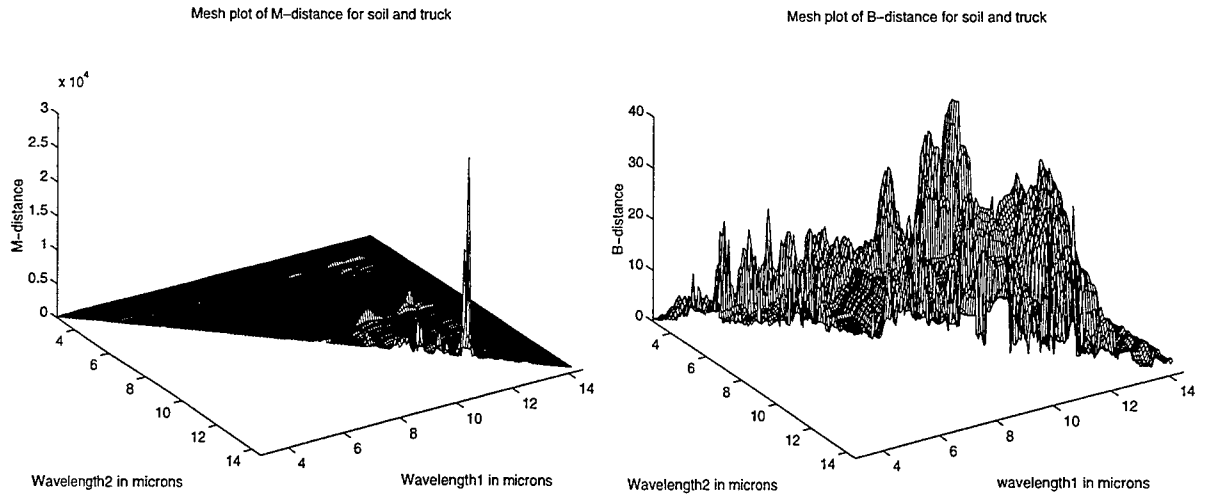
$$B_1(\lambda_1, \lambda_2) = \frac{1}{8}(\mu_t(\lambda_1, \lambda_2) - \mu_b(\lambda_1, \lambda_2))^T \left(\frac{\Sigma_t(\lambda_1, \lambda_2) + \Sigma_b(\lambda_1, \lambda_2)}{2}\right)^{-1} (\mu_t(\lambda_1, \lambda_2) - \mu_b(\lambda_1, \lambda_2)) \tag{20}$$

$$B_2(\lambda_1, \lambda_2) = \frac{1}{2}\ln \frac{\left|\frac{\Sigma_t(\lambda_1, \lambda_2) + \Sigma_b(\lambda_1, \lambda_2)}{2}\right|}{\sqrt{|\Sigma_t(\lambda_1, \lambda_2)||\Sigma_b(\lambda_1, \lambda_2)|}}. \tag{21}$$

Similarly, the M-distance can be computed as

$$M(\lambda_1, \lambda_2) = [\mu_t(\lambda_1, \lambda_2) - \mu_b(\lambda_1, \lambda_2)]^T [\Sigma_b(\lambda_1, \lambda_2)]^{-1} [\mu_t(\lambda_1, \lambda_2) - \mu_b(\lambda_1, \lambda_2)]. \tag{22}$$

Here λ_1 and λ_2 represent the center wavelengths of the two spectral bands used. Thus, center wavelengths are selected which maximize (19) or (22). These bands should provide the best class



(a) (b)

Figure 6: *Mesh plots showing the (a) M-distance and (b) B-distance as a function of spectral band center wavelength for a camouflaged truck target and soil background.*

separability for the appropriate classifier. These band center wavelengths will be referred to as the optimal band pair for a particular target and the background data set.

To illustrate the selection of a best band pair using the B-distance and M-distance criteria, a typical example is shown in Fig. 6 for the background and target pair of soil and camouflaged truck. The horizontal axes in these plots represent the center wavelengths of the equalized spectral bands used. The height in Fig. 6a represents the M-distance for every possible band pair, while the height in Fig. 6b represents the B-distance. The band pair corresponding to the peak would represent the optimal band pair for those target and background observations.

The best band pair tends to vary for different target and background pairs and on the specific observation conditions. In particular, we find that the best band pair depends the target, background, sensor noise level, spectral bandwidth, time of the day, season and other factors. We explore and quantify many of these effects in an attempt to understand their relative importance and impact on band selection and detector performance in Section 4.

4 Detector Robustness and Band Selection

In this section we study the robustness of the multispectral detectors. In particular we use the statistical distance metrics to perform optimal band selection and evaluate detector performance.

As stated earlier, several factor will influence band selection and detector performance. We consider the following: sensor noise level; spectral bandwidth; time of day; season; and target type. The data used for the analysis are from collection experiments at

1. White Sands Missile Range (WSMR), New Mexico. The collection occurred between 6 Jan. 1993 and 12 Jan. 1993 [5].

2. Wright-Patterson Air Force Base (WPAFB), Ohio. The data acquisition occurred between 24 Sept. 1993 and 29 Nov. 1993 [7].

The WSMR data contain measurements of different tanks, trucks, painted panels, desert scrub, soil, grass and mixed backgrounds. The data collection experiment at WPAFB spanned several months providing data suitable for seasonal analysis. During the span of the WPAFB collection, the leaves on trees changed color from green to yellow and then fell from the trees. Thus, a significant change in vegetation backgrounds is observed. The WPAFB data contain observations of an M35 truck, painted panels, grass, trees, and soil. Our analysis begins with sensor noise effects using WSMR data.

4.1 Noise Analysis

When any multispectral data are collected, some noise is invariably introduced. This noise can greatly effect the performance of detectors and can influence the best band selection. To understand the effect of noise on detector performance and band selection, a empirical study has been performed and the results are presented in this section.

The background and target for this analysis are scrub and camouflaged truck respectively. The data used are from the WSMR collection and were acquired on 10 Jan. 1993 at 7:26 AM (file:asxzza) [5]. To simulate the effect of additional additive noise, the estimated covariance of the background data Σ_b is altered by adding a small noise term to it. The modified covariance can be written as

$$\Sigma_b' = \Sigma_b + \sigma^2 I, \tag{23}$$

where σ^2 is the noise variance to be added and I is the identity matrix. Since our analysis deals with two bands, the background covariance is a 2×2 matrix and so is the identity matrix. This simulates the effect that noise will have on observations with a sensor with higher noise levels then the Bomem.

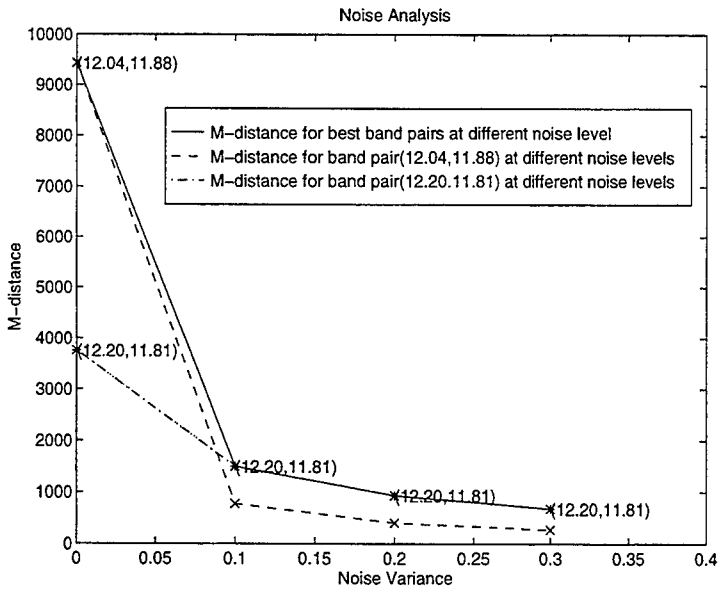


Figure 7: *M-distance vs. noise variance for several band pairs including the best band pair at each noise level. Here the target and background are camouflaged truck and scrub, respectively (file wsmr/asxzza).*

Using the altered background covariance, the M-distance is calculated for each band pair the optimal band pair is found. This is done for noise variances from $\sigma^2 = 0.1$ to $\sigma_2^2 = 0.3$. Figure 7 shows a plot of M-distance as a function of noise for various band pairs. Note that by adding a small noise (e.g., $\sigma_2^2 = 0.1$), the M-distance drops significantly. The best band pair also shifts slightly, but remains in the far infrared region. Note that lower M-distance means a poorer target/background class separability and decreased detector performance. Figure 8 shows the effect of noise on M-distance for all band pairs. Clearly, class separability drops for all band pairs and the peak changes slightly in Figs. 8a and 8b.

The trend observed in our analysis is that the optimal band pair will change depending on the level of sensor noise. To understand this, note that a strong background correlation generally increases target/background class separability as does a large target/background mean difference. In very low noise levels, strong spectral correlation can be preserved during data acquisition and exploited. With higher levels of additive white noise, some correlation is lost, reducing the M-distance. However, the impact of noise on the target/background mean difference is much smaller. Thus, in the case of noise, band selection tends to favor bands which have a larger target/background mean difference.

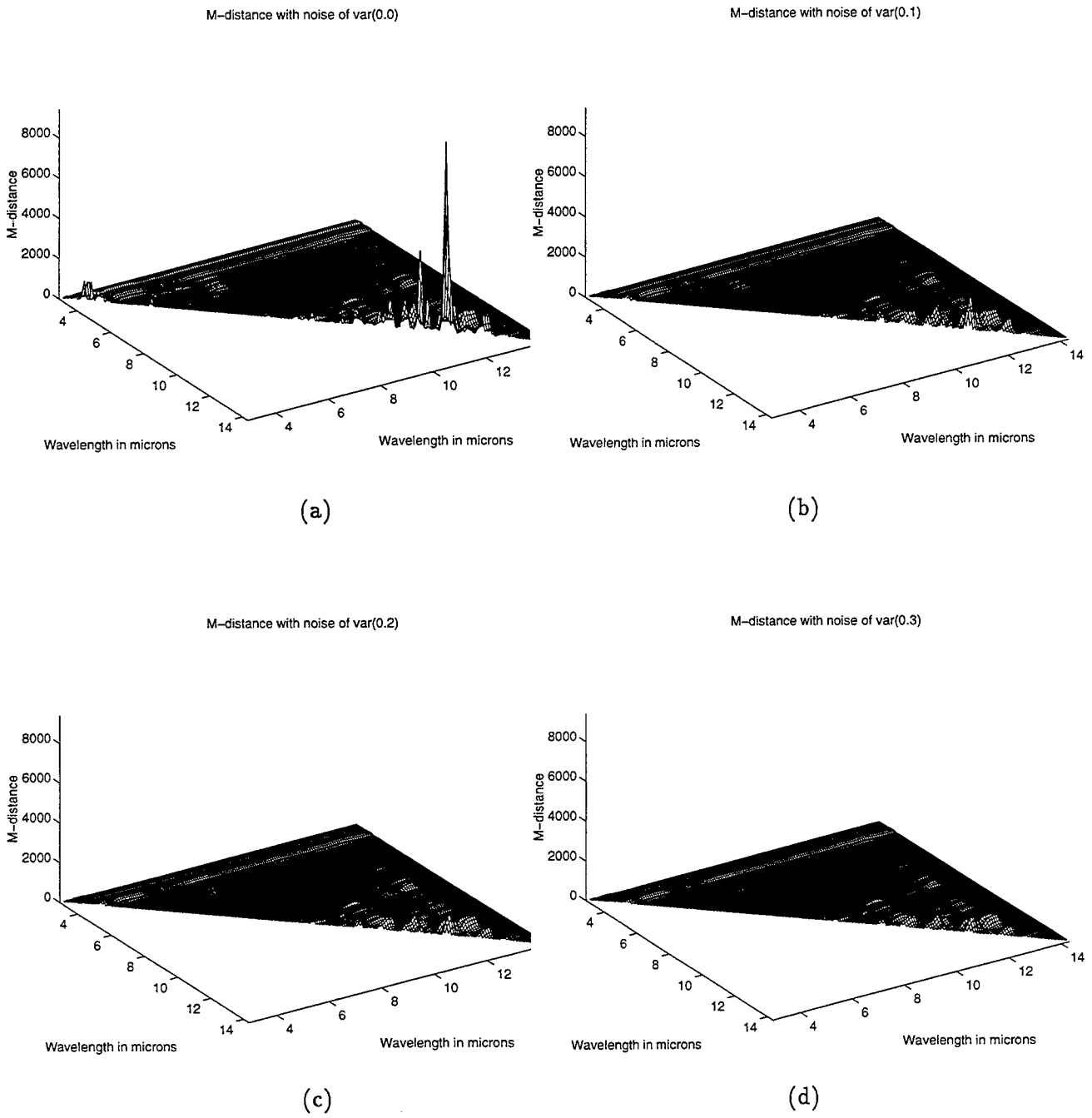
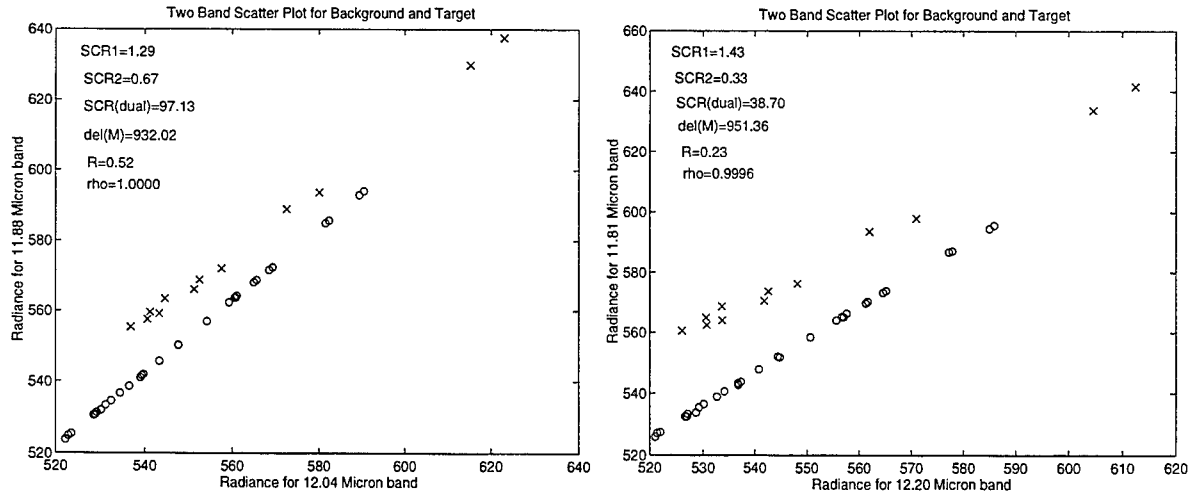


Figure 8: *M-distances for the target and background pair of camouflaged truck and scrub with (a) no additional noise (b) noise of variance 0.1 (c) noise of variance 0.2 (d) noise of variance 0.3.*



(a) (b)

Figure 9: *Scatter plot of best band pair selected for the scrub and truck data set (a) when no noise is present (b) assuming noise of variance 0.1 is present.*

This can be demonstrated by examining the scatter plots for the best band pairs with different levels of noise. Such scatter plots are shown in Fig. 9. Figure 9a shows the scatter plot for the observations from the optimal band pair selected with no additional noise. Figure 9b shows the scatter plot for the observations from the optimal band pair assuming additive noise variance $\sigma^2 = .1$. To aid in interpreting these plots, several statistical parameters are provided along with each. To measure the target/background mean difference we define

$$\Delta\mu = ||\mu_b - \mu_t||^2 = (\mu_b - \mu_t)^T(\mu_b - \mu_t). \tag{24}$$

This parameter, (del(M)), is seen to increase from Figs. 9a to 9b. The term ρ is the background correlation coefficient and it decreases from Figs. 9a to 9b. Thus, by "trading" background correlation for increased target/background mean difference, the detector will be more robust in the presence of noise. Note that the mesh plots in Fig. 8 show a dramatic decrease in the peak M-distance with additive noise. However, some of the surrounding band pairs do not decrease significantly. These other band pairs do not provide the best performance when no noise is present, but they are robust in noise. These band pairs rely less on background correlation which is lost in the presence of sensor noise. The single band and dual band SCRs are also shown in Fig. 9 along with the target/background color R.

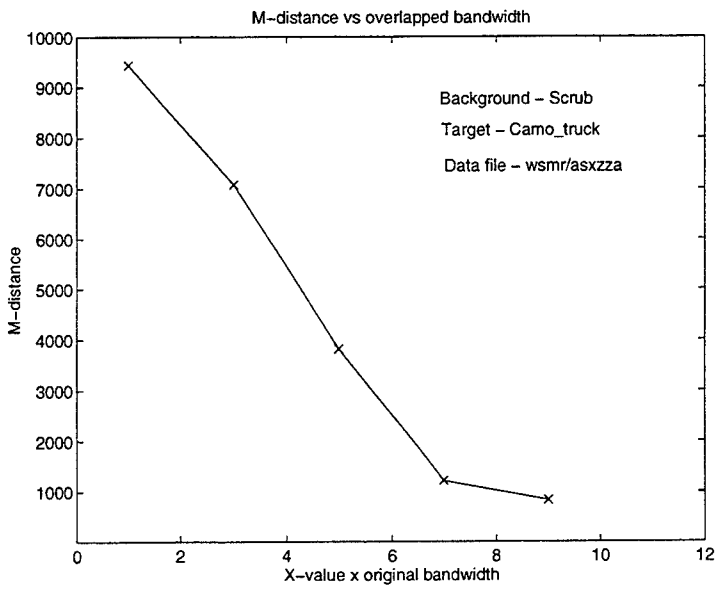


Figure 10: *M-distance vs. effective bandwidth in overlapping analysis.*

4.2 Bandwidth Analysis

This section addresses the issues of spectral bandwidth. In particular, we quantify class separability as a function of bandwidth for a specific target and background pair. Two types of analysis are done here. Overlapping bandwidth analysis is where the bandwidth remains the same for all the bands but are allowed to "overlap" in bandwidth range. In the non-overlapping bandwidth analysis, the spectral bands may not overlap. Thus, the non-overlapping analysis is a subset of the overlapping analysis.

We begin with the overlapping analysis which is done with the background and target set of scrub and camouflaged truck. These data are from the WSMR data collection acquired on 10 Oct. 1993 and 7:54 AM (file: asxzza) [5]. The effective bandwidth is increased by averaging the data within the appropriate bandwidth range. At each bandwidth step, the M-distance for all two-band combinations is calculated and the maximum M-distance is found. This maximum M-distance versus the effective bandwidth is plotted in Fig. 10. Note that as the bandwidth increases, the M-distance decreases. Thus, the narrow band features provide the best discrimination for this target background pair. Similar results have been seen for other targets and backgrounds. The results of the non-overlapping analysis applied to the same target and background pair are shown in Fig. 11.

An interesting result, observed in Fig. 12, is that with the increase of bandwidth, the peak in

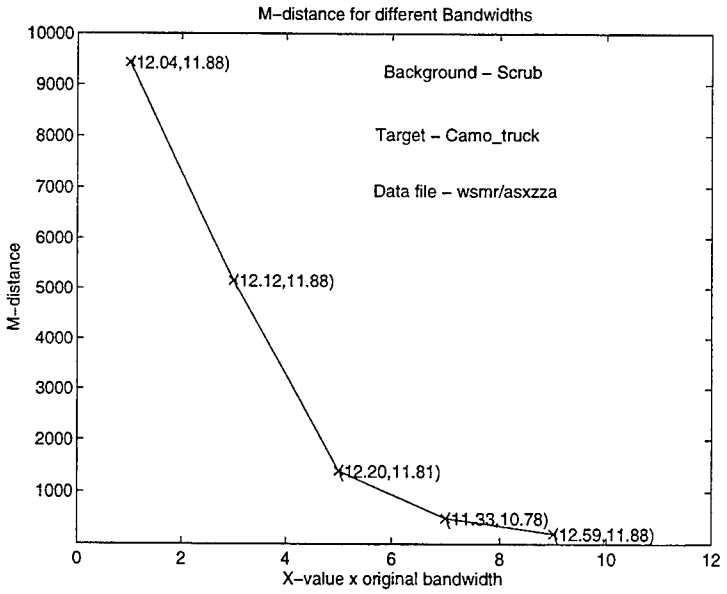


Figure 11: *M-distance vs. effective bandwidth in non-overlapping analysis.*

the M-distance plot remains at approximately the same center wavelength. Figure 12a shows the M-distance plot for the original bandwidth. The M-distance plots for bandwidths of 3, 5 and 7 times the original are shown in Figs. 12b, 12c and 12d, respectively.

4.3 Time-of-Day Analysis

The spectral response patterns depend on the solar illumination and the temperature of the object. These factors vary with the time of day, and consequently, band selection may change. The sensitivity of the multispectral detectors to time of day is investigated here. The time analysis is performed using the WPAFB data sets described in Table 1. These data have been selected because they contain the same target and background at different times of the same day. The exact times are provided in Table 1. For our analysis, we selected grass as the background and an M35 open truck as the target.

Figure 13 shows the entire mean spectrum of the target and background at several different times. By observing these mean plots, it can be seen that the mean difference is maximum at 9:20 AM. Consequently, the M-distance for the optimal band pair is also relatively large during this time. Furthermore, it can be seen that at early hours of the day (around 6:30 AM), both the background and the target are cool and by 9:30 AM both the target and the background have heated up significantly. This expressed an an overall increase in the infrared spectrum. However, note that the target heats up more than the background. After noon, both bodies cool down and

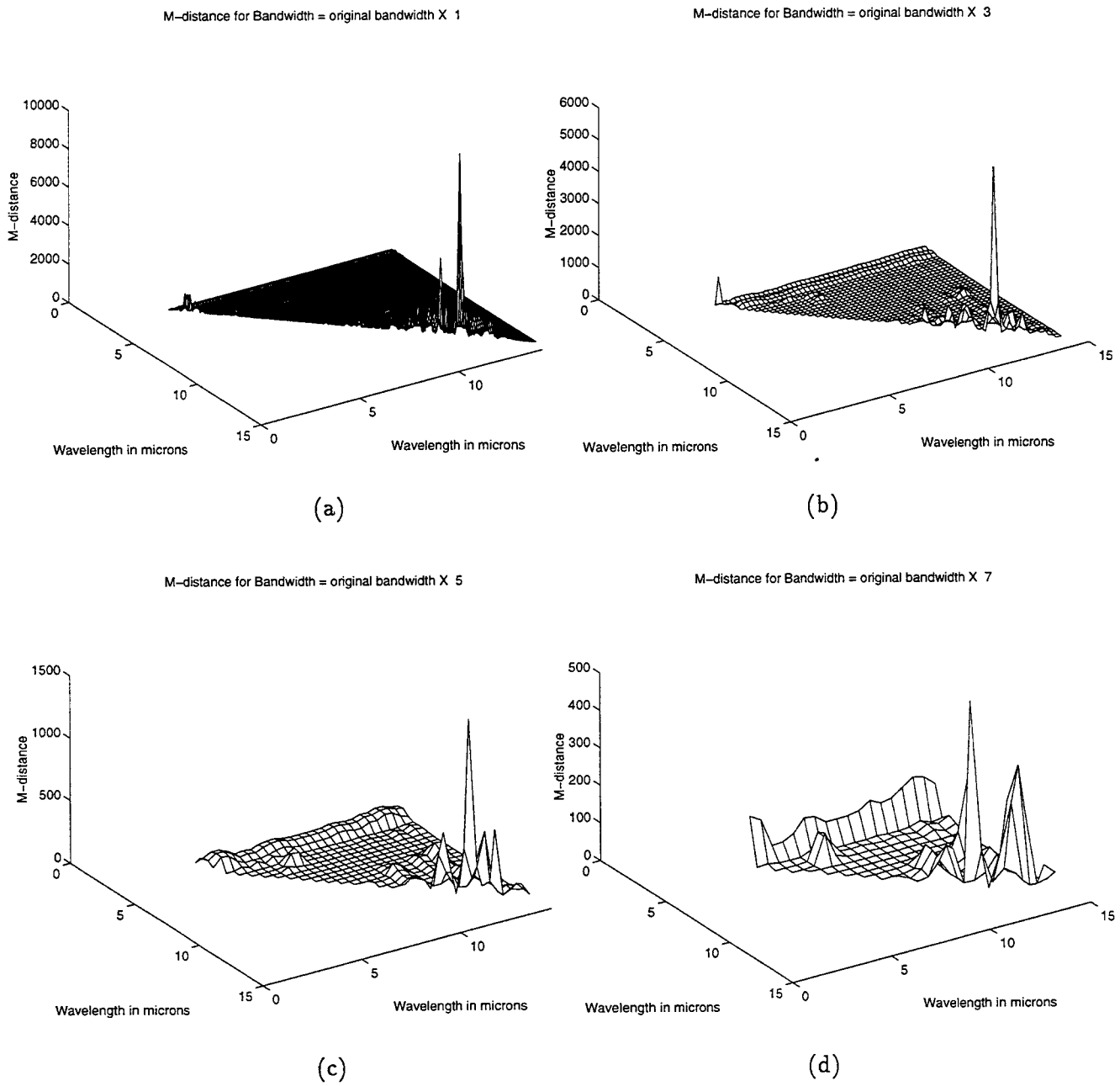


Figure 12: *M-distance for the (a) original data (b) data with 3 times original bandwidth (c) data with 5 times original bandwidth (d) data with 7 times original bandwidth.*

Table 1: *WPAFB data files used for time of day analysis.*

Time Analysis (Test for Robustness and performance)				
Filename	DATE/TIME	TYPE	TARGET	Background
E10071	10-7-93,6:45	Contrast	M35 truck	grass
E10073	10-7-93,9:20	Contrast	M35 truck	grass
E10077	10-7-93,11:55	Contrast	M35 truck	grass
E10078	10-7-93,20:40	Contrast	M35 truck	grass

the target cools down much faster than the background.

This change in the mean difference between target and background over time results in a change in the optimal band pair. To illustrate this, Fig. 14 shows the M-distance mesh plots for the target and background at the different times. Note that the peak occurs at different wavelength pairs at each time of day. Figure 15 shows a plot of M-distance using the optimal band pair at each time for the same target and background. The best target/background discrimination is obtained at 12 noon.

4.4 Seasonal Analysis

For seasonal analysis, data collected in the month of October and November are used. These data have been selected to investigate the effect of fall vegetation change on band selection. During this time interval the spectral response patterns of the vegetation change significantly. The background and target selected are grass and carc panel respectively. The acquisition time and date of these data are provide in Table 2. Note that the times the data are collected are between 10 AM to 11 AM.

Figure 16 shows a plot of the maximum M-distance and corresponding best band pair computed for the grass and carc panel on different dates. As can be seen, the best classification result is obtained in the mid October. It can also be observed that the mean spectrum difference between the target and the background is larger in October than September. This is illustrated in Fig. 17 which shows the entire mean spectrum for the carc panel and grass on different dates. Finally, the M-distance plots for the carc panel and grass on the different dates are shown in Fig. 18. Here it

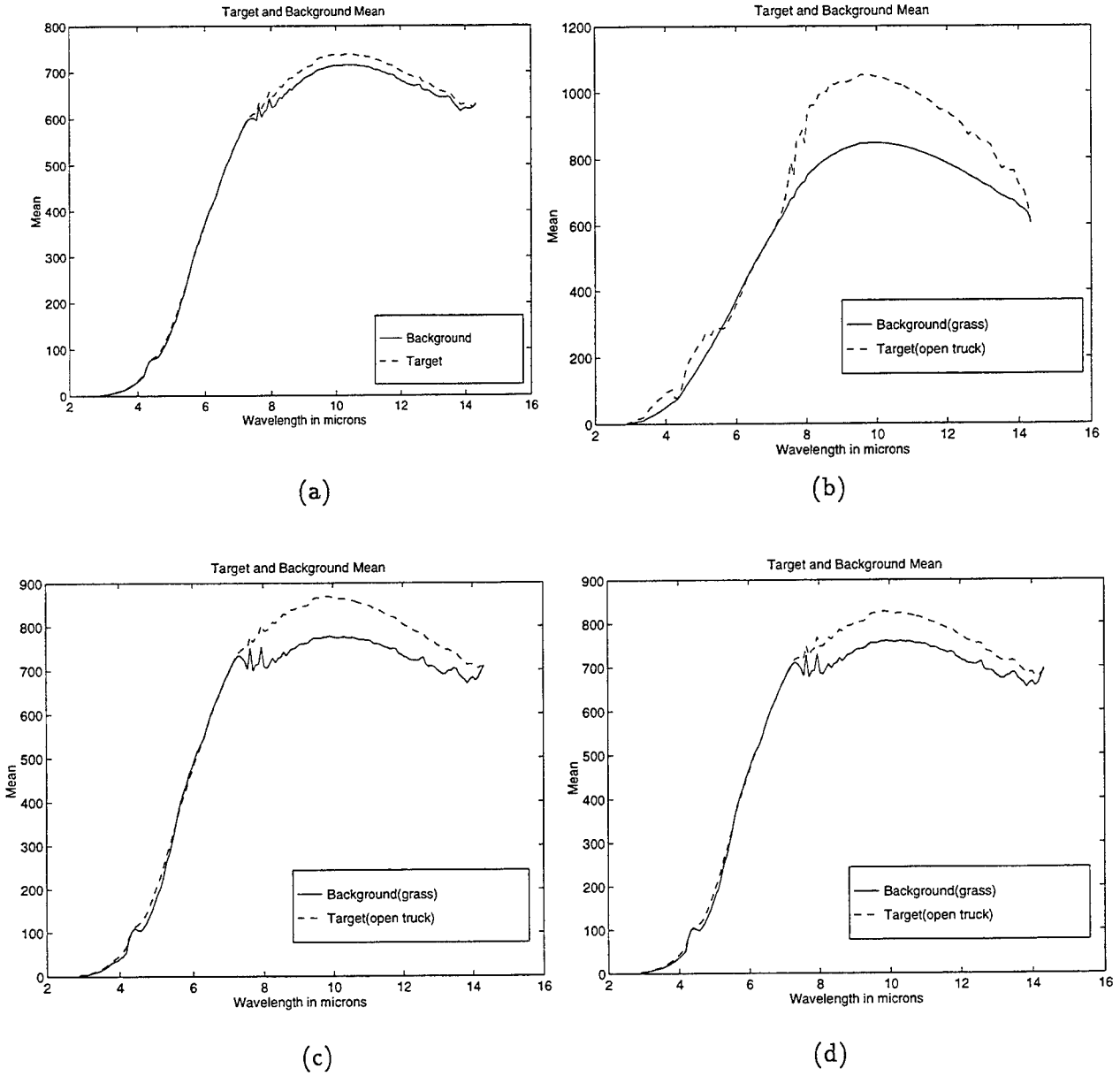


Figure 13: *Mean spectrum plot for the open truck and grass at (a) 6:45 AM (b) 9:20 AM (c) 11:55 AM (d) 8:40 PM.*

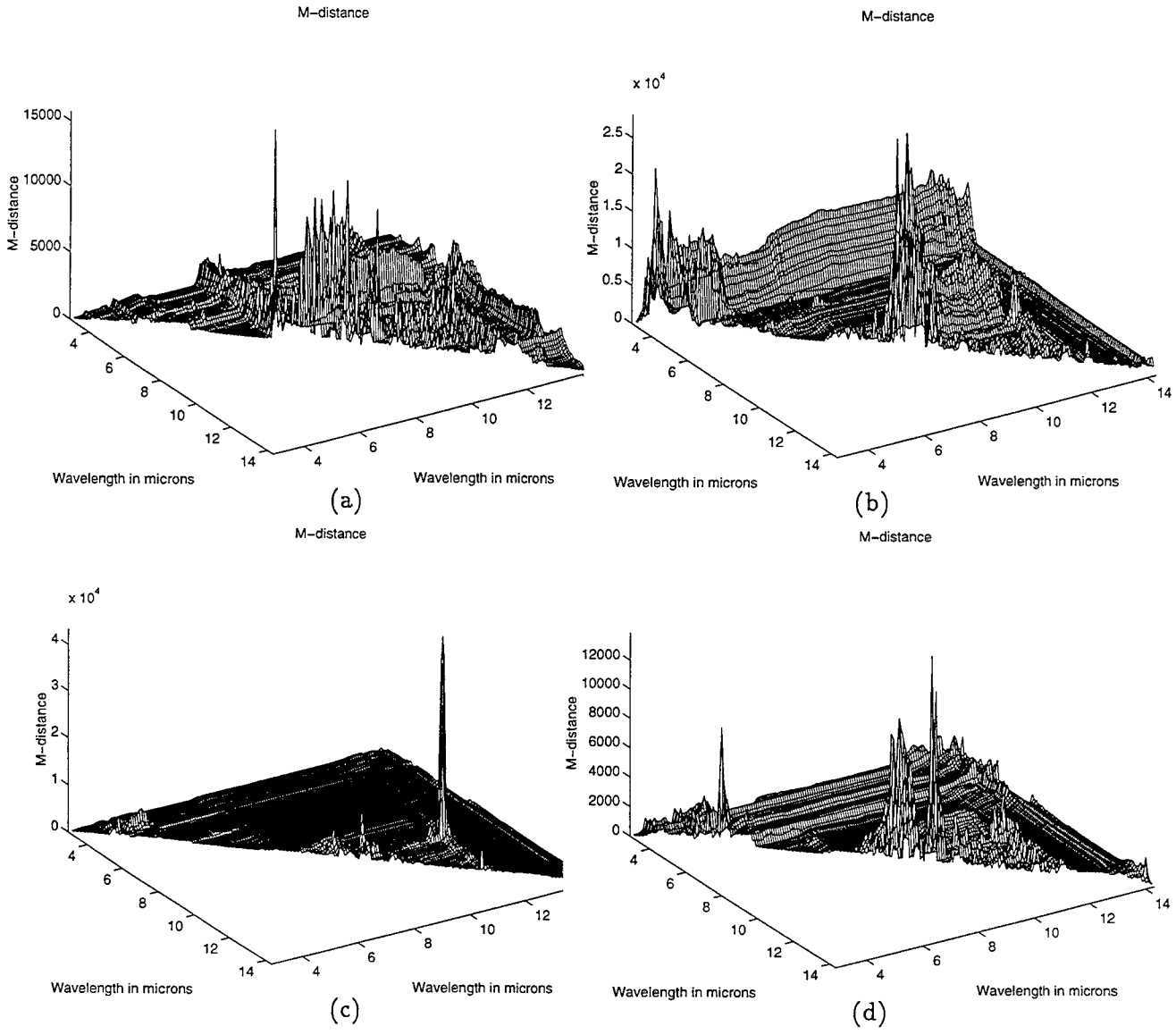


Figure 14: *M-distance plot for open truck and grass at (a) 6:45 AM (b) 9:20 AM (c) 11:55 AM (d) 8:40 PM.*

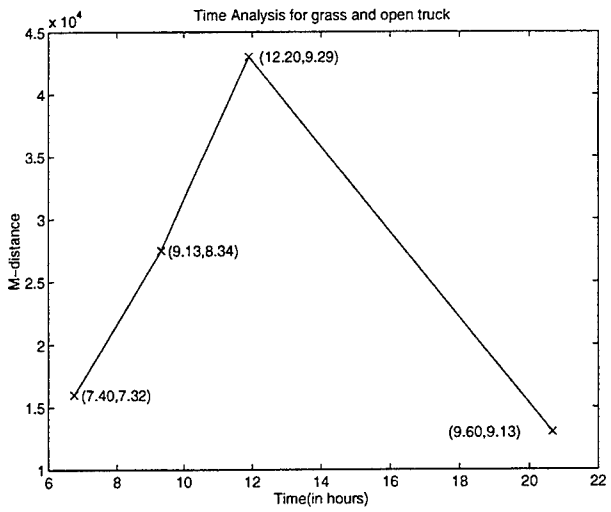


Figure 15: *Maximum M-distance and corresponding band pair for grass and the open truck calculated at different times of the day.*

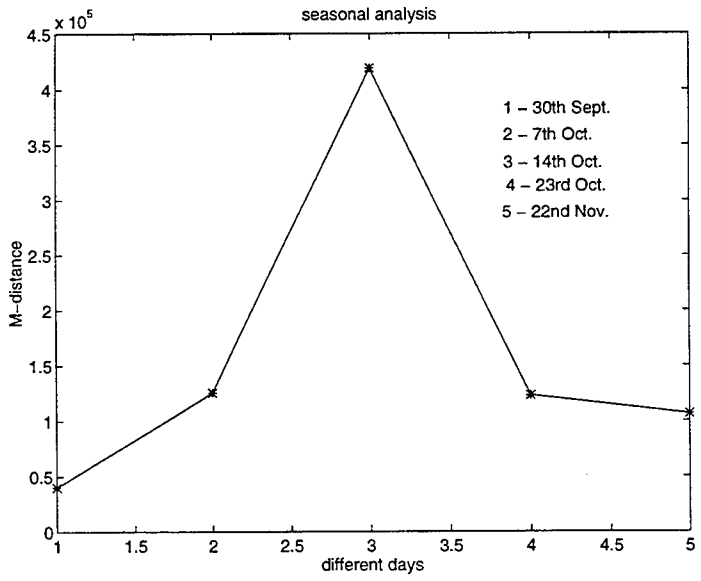


Figure 16: *Maximum M-distance for grass and carc panel on different dates.*

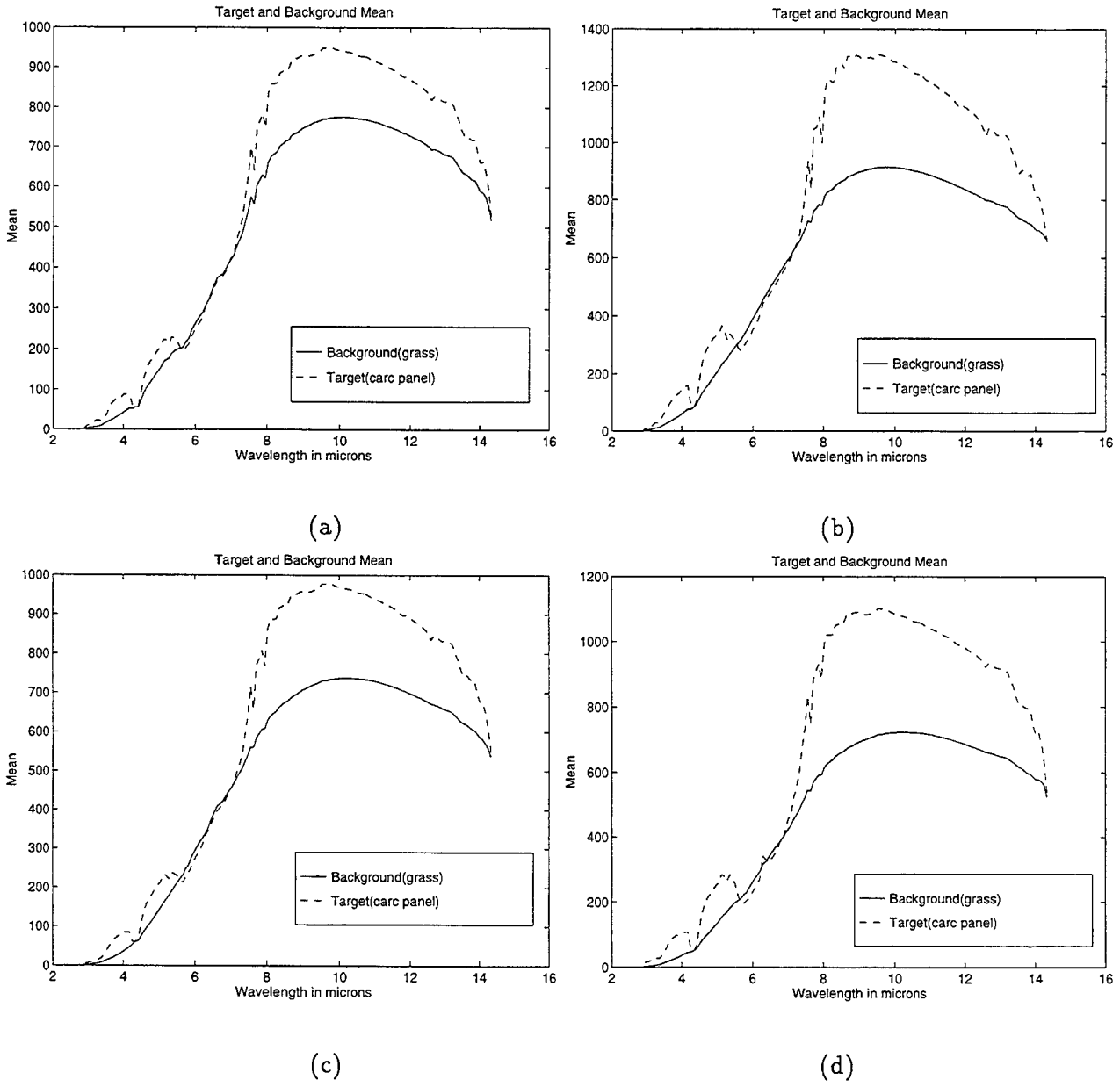


Figure 17: *Mean spectrum plot for carc panel and grass data collected on (a) 30 Sept.(b) 7 Oct. (c) 14 Oct. (d) 23 Oct.*

Table 2: *WPAFB data files used for seasonal analysis.*

Seasonal Analysis (Test for Robustness and performance)				
Filename	DATE/TIME	TYPE	TARGET	Background to use
E09305	9-30-93,11:05	Correlation	Carc Panel	grass
E10074	10-7-93,10:20	Contrast	Carc Panel	grass
E10144	10-14-93,10:00	Contrast	Carc Panel	grass
E10224	10-23-93,10:02	Contrast	Carc Panel	grass
E11225	11-22-93,9:55	contrast	Carc Panel	grass

can be seen that the optimal band pair changes considerably on different dates.

4.5 Different Targets in Same Background

In this section, we investigate detector performance and band selection with different targets. A common scrub background is paired with each target in this analysis. Table 3 contains a list of all the targets and the background used here. These are from a WSMR collection experiment.

The best band pair for each target with the scrub background is computed using the M-distance criteria. The optimal band pair for one target is then applied to the other targets to test the robustness of this band pair. The results of this analysis with all targets listed in Table 3 are shown in Fig. 19. The height of each bar represents the M-distance for the target specified on one axis and the band pair specified on the other axis. The band pair column labeled "overall best band pair" contains the best band pair for each target with the scrub background. Thus, this column contains the largest possible M-distance for each target. The rest of the band pair columns show how well a band pair selected for one target works on a different target. For example, the band pair column labeled "T1-(9.76, 3.778)" refers to the band pair $9.76\mu m, 3.778\mu m$ which is the optimal band pair for target "T1," the specular panel. The rest of the band pair columns are labeled similarly.

Note that the band pair which provides optimal discrimination between the specular target and scrub does not perform well on any other target. Also note that the targets T7-T14 are easily discriminated from the scrub using band pairs in the $10\mu m - 12\mu m$ infrared region. The

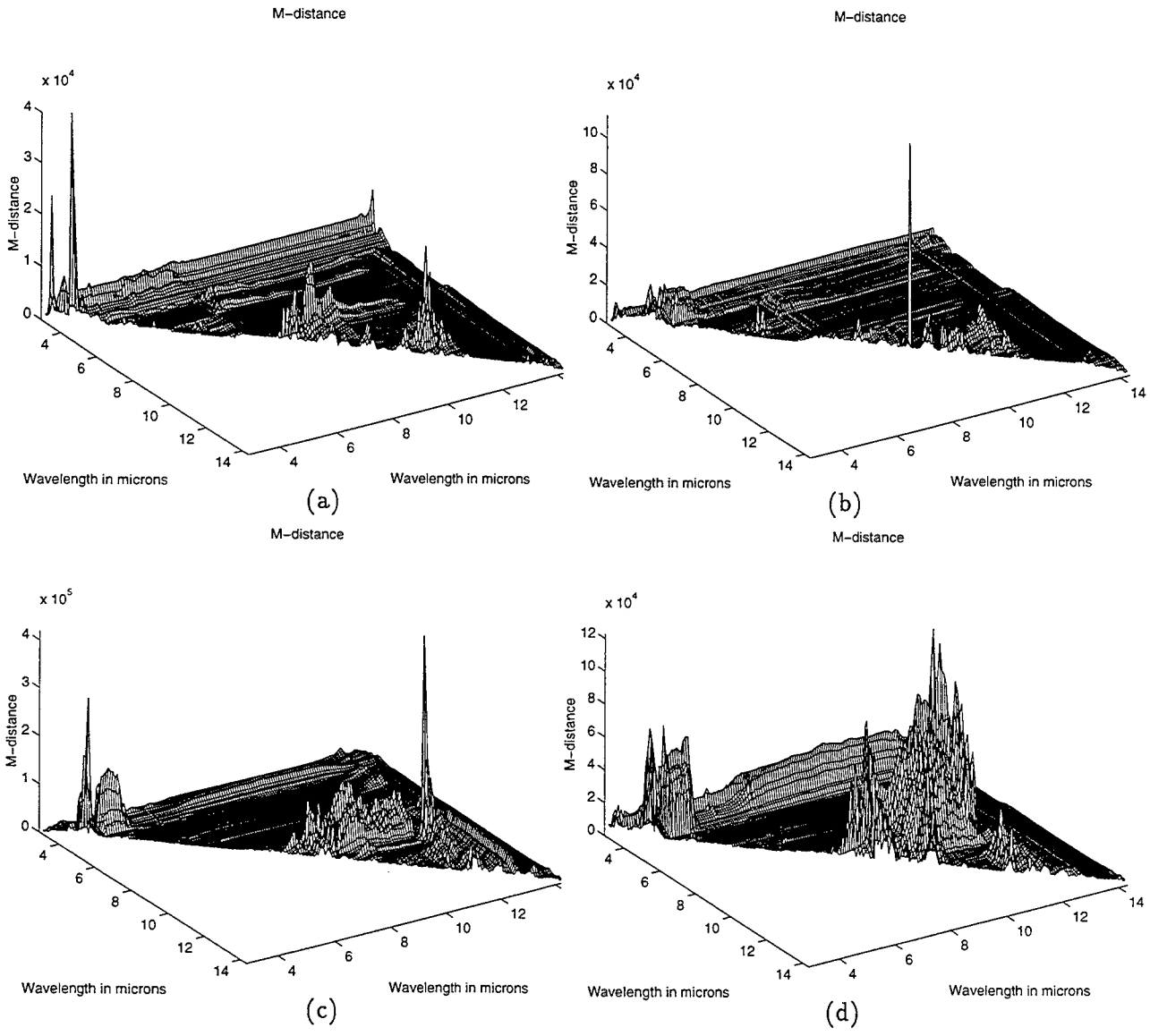


Figure 18: *Mesh plot of the M-distance calculated for the carc panel and grass data collected on (a) 30 Sept. (b) 7 Oct. (c) 14 Oct. (d) 23 Oct.*

Table 3: *WSMR data file used for the study of different targets with a common background of scrub.*

Data file for the analysis of different targets in same background	
File	Contents
atfzwa	Camo_Tank1 0 11 Tank2 12 23 Camo_Truck 24 35 Hertz_Truck_Cab 36 38 Hertz_Truck_Back 39 50 BB_Trailer 51 62 White_Truck 63 74 Rental_Car 75 82 CARC_Tan 83 84 Low_E_Tan 85 86 Diffuse 87 88 Specular 89 90 CARC_Green 91 92 Cylinder 93 94 Sky 95 99 Blackbody 100 102 Scrub 103 132

Table 4: *WSMR and WPAFB data files used in multiple target/background analysis.*

Data files used to find a overall sub-optimal band pair				
File	Date	Time	Background	Target
atfzwa	1-10-93	18:36	scrub	tank2
e10077	10-7-93	11:55	trees	carc
arxzza	1-10-93	6:42	soil	camo_truck
asxzza	1-10-93	7:26	scrub	camo_truck
e10078	10-7-93	20:40	grass	carc
atfzwc	1-10-93	18:36	scrub	BB-trailor
e10071	10-7-93	6:45	trees	crac
e10073	10-7-93	9:20	trees	net
e10144	10-14-93	10:00	trees	open_truck
e09305	9-30-93	11:05	grass	carc
e10224	10-23-93	10:02	trees	emissive panel
e10074	10-23-93	10:02	trees	net
arxzzb	1-10-93	20:40	soil	camo_tank1
a10285	10-28-93	10:12	trees	diffuse panel
e11294	11-29-93	10:09	trees	obscured truck
asxzzd	1-10-93	18:36	scrub	camo_tank1

camouflaged objects are less separable from the background. For example the optimal band pair for the CARC panel and scrub is $9.13\mu m, 8.894\mu m$ and the M-distance for this is lower than for T7-T14.

4.6 Multiple Target-Background Pairs

Up until this point, we have investigated band selection and detector performance for one target/background pair at a time. However, in many applications band selection will be performed with many possible targets and backgrounds in mind. We explore two approaches to this problem. The first involves the M-distance criteria and the second involves the B-distance criteria. The M-distance analysis is based on how many target/background pairs a particular band pair can provide below a set probability of false alarm for. The B-distance analysis selects spectral bands which minimize the probability of error given many possible target/background scenarios. The target set of 16 target/background pairs used in this analysis is listed in Table 4. A wide variety of targets and backgrounds is considered here. Furthermore, the collection times and seasons span a wide range.

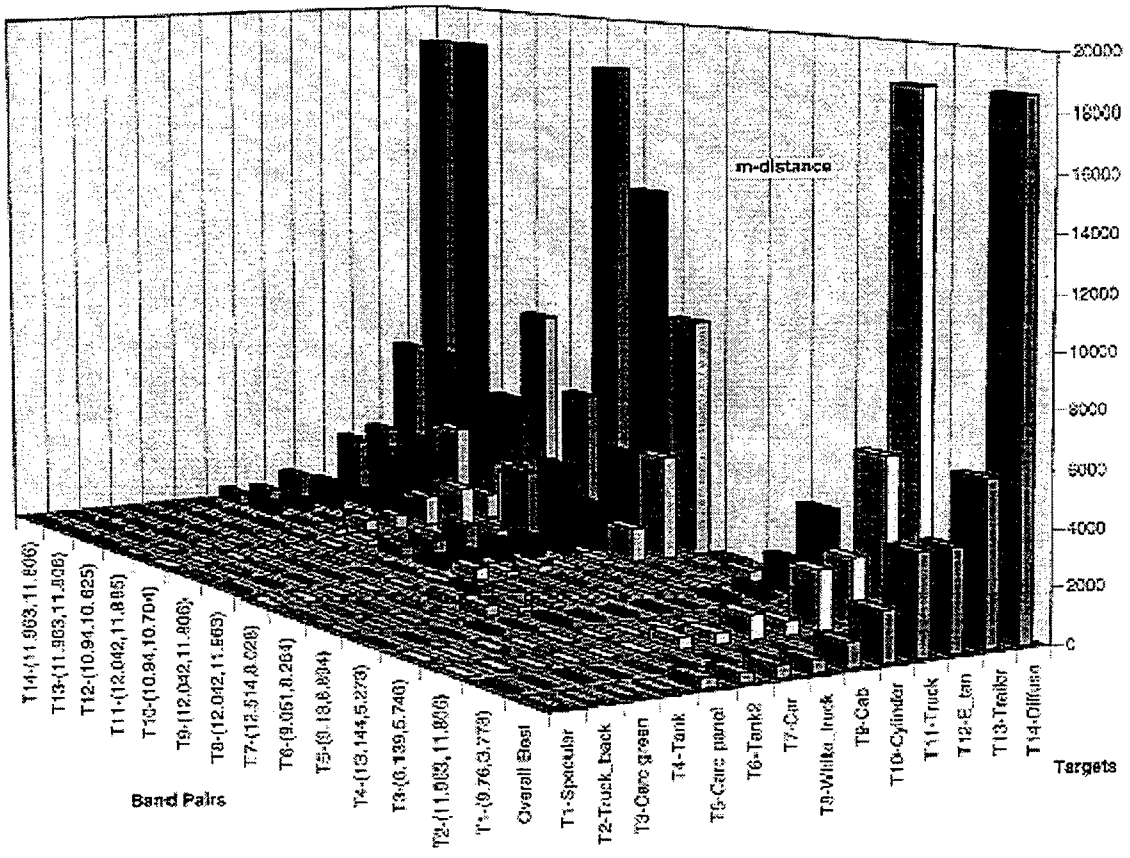
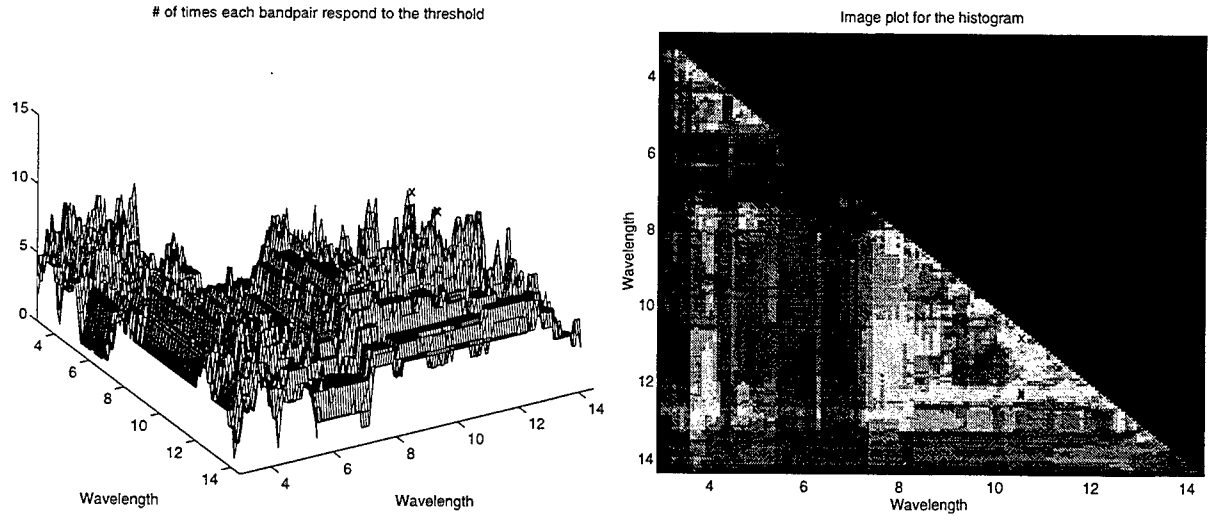


Figure 19: *M-distance for different targets with a common background of scrub. The "Targets" axis shows each target analized from T1 (specular panel) - T14 (diffuse panel). The "Band pairs" axis shows the specific band pairs used. Each band pair is the optimal band pair for one of the targets. The target for which each band pair is optimal is listed along with the band pair wavelengths on the "Band pairs" axis.*



(a) (b)

Figure 20: *(a) Mesh plot showing the number of target/background pairs for which each band pair yields below a specified probability of false alarm. (b) Image showing the same information where bright pixels represents the highest number of target/background pairs and consequently the best band pairs.*

4.6.1 M-Distance Analysis

The first step inthe M-distance analysis is to compute the M-distance for every band pair for each target/background pair. A desired probability of false alarm P_{fa} is specified. The number of target/background pairs for which a particular band pair yields a P_{fa} below the threshold is counted. This result may vary from $0-16$ in our analysis. The band pair yielding the most would be the globally optimal band pair for the data studied under the M-distance criteria. The results of this analysis are shown in Fig. 20. Three band pairs provide the best results: $(10.468\mu m, 10.783\mu m)$, $(10.468\mu m, 12.199\mu m)$, and $(10.468\mu m, 12.278\mu m)$. These peak values are indicated with an 'x' in Fig. 20a. Fig. 20b shows the same information as Fig. 20a in grayscale image form where bright pixels represents the highest number of target/background pairs, and consequently, the best band pairs.

4.6.2 B-Distance Analysis

Another approach that can be used to find a band pair which performs well with several target/background pairs uses the B-distance. Since the B-distance relates to probability of error, we seek the band pair that minimizes the overall probability of error. To begin, let us define different targets as t_i where $i = 1, 2, \ldots, N_t$. Similarly let b_j represent the different backgrounds, where $j = 1, 2, \ldots, N_b$. Now let us define $Pr(t_i, b_j)$ as the probability that the present scenario of interest involves detecting target t_i with background b_j. Also let Pe_{ij} be the probability of error and C_{ij} the cost of error for that scenario. Using cost terms allows one to specify that error with one target/background pair is more or less important than with another. Next, define the *a priori* probabilities as follows:

$$P_{ij}{}^{t} = Pr(t_i|t_i, b_j), \tag{25}$$

$$P_{ij}{}^{b} = Pr(b_j|t_i, b_j). \tag{26}$$

Now the total probability of error for a specific band pair can be written as

$$Pe(\lambda_1, \lambda_2) = \sum_{i=1}^{N_t} \sum_{j=1}^{N_b} Pe_{ij}(\lambda_1, \lambda_2) Pr(t_i, b_j) \tag{27}$$

and total cost can be written as

$$C(\lambda_1, \lambda_2) = \sum_{i=1}^{N_t} \sum_{j=1}^{N_b} Pe_{ij}(\lambda_1, \lambda_2) Pr(t_i, b_j) C_{ij}. \tag{28}$$

Using Bhattacharyya bound for the total error we get

$$Pe_{ij}(\lambda_1, \lambda_2) = \sqrt{P_{ij}{}^{t} P_{ij}{}^{b}}\, e^{-B_{ij}(\lambda_1, \lambda_2)}, \tag{29}$$

where

$$B_{ij} = B_{ij_1} + B_{ij_2}, \tag{30}$$

and

$$B_{ij_1} = \frac{1}{8} \left[(\mu_{ti} - \mu_{bj})^T \left(\frac{\Sigma_{ti} + \Sigma_{bj}}{2} \right)^{-1} (\mu_{ti} - \mu_{bj}) \right], \tag{31}$$

$$B_{ij_2} = \frac{1}{2} \left[ln \frac{\frac{|\Sigma_{ti} + \Sigma_{bj}|}{2}}{\sqrt{|\Sigma_{ti}|\,|\Sigma_{bj}|}} \right]. \tag{32}$$

Now the total cost is given by

$$C(\lambda_1, \lambda_2) = \sum_{i=1}^{N_t} \sum_{j=1}^{N_b} \sqrt{P_{ij}^{t} P_{ij}^{b}}\, e^{-B_{ij}(\lambda_1, \lambda_2)} Pr(t_i, b_j) C_{ij}. \tag{33}$$

Now consider the special case where $C_{ij} = C_0$ for all i and j and $Pr(t_i, b_j) = \frac{1}{N}$, where N is the total number of combinations of targets and backgrounds ($N = N_t N_b$). Also let $P^t_{ij} = P^t$ and $P^b_{ij} = P^b$ for all i and j. This implies that all the targets are equally likely to be present and all the backgrounds are equally likely to be present. In this case,

$$C(\lambda_1, \lambda_2) = \sqrt{P^t P^b} \frac{1}{N} C_0 \sum_{i=1}^{N_t} \sum_{j=1}^{N_b} e^{-B_{ij}(\lambda_1, \lambda_2)}. \tag{34}$$

For minimum cost, the sums of exponential terms $e^{-B_{ij}(\lambda_1, \lambda_2)}$ should be minimized. For example, if $N_t = N_b = 2$, the following must be minimized:

$$\sum_{i=1}^{N_t} \sum_{j=1}^{N_b} e^{-B_{ij}(\lambda_1, \lambda_2)} = e^{-B_{11}(\lambda_1, \lambda_2)} + e^{-B_{12}(\lambda_1, \lambda_2)} + e^{-B_{21}(\lambda_1, \lambda_2)} + e^{-B_{22}(\lambda_1, \lambda_2)}. \tag{35}$$

That is, the wavelengths λ_1 and λ_2 which minimize the cost must be determined.

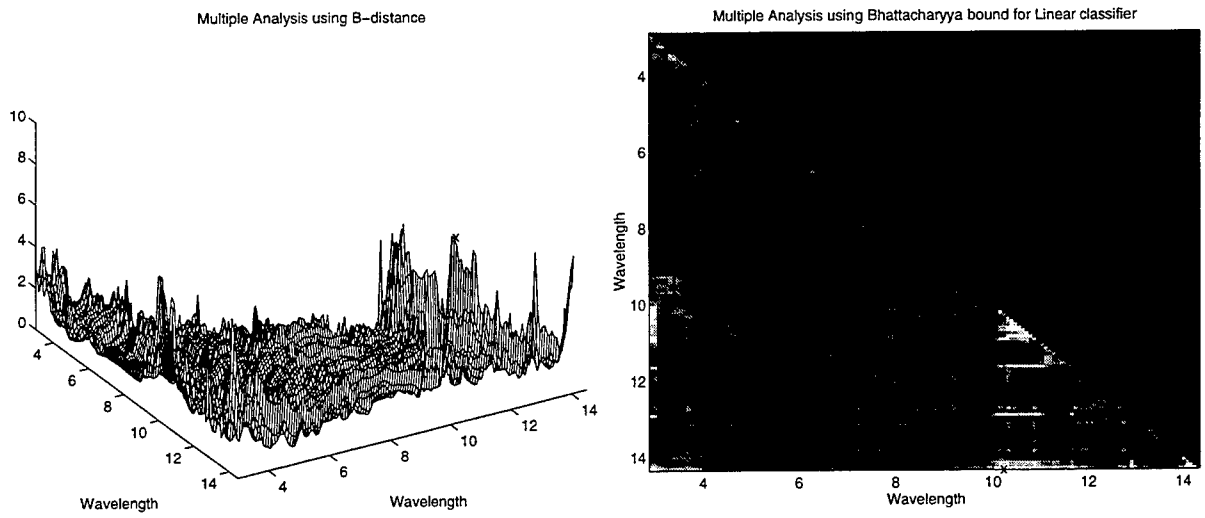
In our multiple target/background pair analysis, we consider only the B_1-distance. This relates to the probability of error for a linear classifier. The cost function results for the target/background pairs in Table 4 are shown in Fig. 21. However, to facilitate comparison with the M-distance results, the cost function is reversed so that a large value is actually a low cost. Figure 21a shows the mesh plot of the scaled inverse cost function. The peak of this function indicates the band pair which gives the minimum cost. This peak occurs at the band pair $(10.15\mu m, 14.32\mu m)$ and is designated with an 'x.' Figure 21b shows an image plot for the inverse cost function where the lowest cost band pairs points are shown as bright pixels. Secondary band pair peaks are seen to occur in the $10\mu m - 11\mu m$ range.

5 Conclusions

A detailed empirical study of factors effecting band selection and multispectral detector performance is presented here. Band selection and detector performance are analized using the statistical distance measures B-distance and M-distance. These measures quantify the level class separability between a target and background. Given truthed Bomem spectrometer data, band pairs which maximize the B-distance of M-distance can be selected by an exhaustive search.

In this study, several observations have been made which we believe will generalize to other data sets. These observations are summarized below.

- As noise levels increase, any strong correlation in the data is most severely effected. Thus, when significant sensor noise is expected, bands which offer large target/background mean



(a) (b)

Figure 21: *(a) Mesh plot of the scaled inverse cost function for each band pair and the 'x' indicates the peak at the band pair* $(10.15\mu m, 14.32\mu m)$. *(b) Image plot for the inverse cost function where the lowest cost bands are shown as bright pixels.*

difference will tend to be more robust than those with the strongest spectral correlation. If little noise is present, strong correlation in certain bands is observed in many natural backgrounds. If this correlation is preserved in data acquisition, large class separability is possible.

- In our analysis of specific target/background pairs, narrow spectral bands provided the largest class separability. That is, class separability was observed to decrease with increased bandwidth. This is due to the fact that narrow band spectral features tend to be more pronounced than broad band features. However, narrow spectral bands must be more carefully tuned to a particular target and background.

- In general, the vehicle targets against soil and vegetation backgrounds were most easily discriminated near noon, although the maximum target/background mean spectrum difference occurred earlier in the day. The lowest discrimination of the times tested was found to be at 6:45 AM (near thermal crossover).

- At WPAFB, the vegetation backgrounds were most easily discriminated from the vehicles

and panels during mid October.

- For the variety of targets tested against a scrub background, band pairs in the far infrared ($10\mu m - 12\mu m$), generally provided the best discrimination.
- Based on the set of target and backgrounds listed in Table 3, the overall optimal band pair using B-distance analysis is ($10.15\mu m, 14.32\mu m$). Other secondary band pair peaks have been found in the $10\mu m - 12\mu m$ range.

Many other factors also effect band selection and detector performance. The methodology used here and in [3] can provide a means of quantitatively evaluating such factors. All of the analysis presented here has been performed using a custom written MATLAB environment tailored to Bomem spectrometer data.

References

[1] C. T. Chen and D. Landgrebe, "A Spectral Feature Design System for the HIRIS/MODIS Era," *IEEE Transaction on Geoscience and Remote Sensing*, Vol. 27, No. 6, November 1989.

[2] K. Fukunaga, *Statistical Pattern Recognition*, Sec. Ed., Academic Press, 1990.

[3] R. C. Hardie, "Adaptive Quadratic Classifier for Multispectral Target Detection," *Summer Faculty Research Program, Wright Laboratory*, Dayton, August 1994.

[4] J. E. Thomas, "Multispctral Detection of Ground Targets in Highly Correlated Background," *Masters Thesis*, AFIT, March 1995.

[5] C. R. Schwartz and M.T. Eismann, "White Sands Missile Range Measurement Catalog," *Infrared Multispectral Sensor Program : Feild Measurements, Analysis And Modelling* Vol. 6, December 1994.

[6] J. N. Cederquist, C. R. Schwartz and M. T. Eismann, "Redstone Arsenal Measurement Catalog," *Infrared Multispectral Sensor Program : Feild Measurements, Analysis And Modelling* Vol. 2, October 1993.

[7] K. K. Ellis, "Wright-Patterson Air Force Base Seasonal Measurements Catalog," *Thermal Infrared Multispectral Program* Vol. 1, May 1994.

[8] T. L. Henderson, A. Szilagyi, M. F. Baumgardner, C. Chen and D. A. Landgrebe, "Spectral Band Selection for Classification of Soil Organic Matter Content," *Soil Science Society of America Journal* Vol. 53, No. 6, November 1989.

[9] X. Jia and J. Richards, "Efficient Maximum Likeliood Classification for Imaging Spectrometer Data Sets," *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 32, No. 2, March 1994.

[10] B. Kim and D. Landgrebe, "Hierarchical Classifier Design in High-Dimensional, Numerous Class Cases," *IEEE Transaction on Geoscience and Remote Sensing*, Vol. 29, No. 4, July 1991.

[11] C. Lee and D. Landgrebe, "Feature Extraction Based on Decision Boundaries," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 4, April 1993.

[12] J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*, Springer-Verlag, 1986.

[13] I. S. Reed and X. Yu, "Adaptive Multiple-Band CFAR Detection of and Optical Pattern with Unknown Spectral Distribution," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 38, No. 10, Oct. 1990.

[14] A. D. Stocker, I. S. Reed and X. Yu, "Multi-Dimensional Signal Processing for Electro-Optical Target Detection," Proc. SPIE Int. Soc. Opt. Eng., Vol. 1305, Apr., 1990.

[15] P. H. Swain and R. C. King, "Two Effective Feature Selection Criteria for Multispectral Remote Sensing," *Proceedings of the First Int. Joint Conf. Patt. Recogn.*, November 1973, pp. 536-540.

[16] X. Yu, I. S. Reed and A. D. Stocker, "Comparative Performance Analysis of Adaptive Multispectral Detectors," *IEEE Trans. on Signal Processing*, Vol. 41, No. 8, Aug. 1993.

ROBUST FAULT TOLERANT CONTROL: FAULT DETECTION AND ESTIMATION

A. S. Hodel
Associate Professor
Department of Electrical Engineering
Auburn University

200 Broun Hall
Auburn AL 36849

Final report for:

Summer Research Extension Program
AFOSR Contract F49620-93-C-0063
Sponsored by:

Air Force Office of Scientific Research
Bolling Air Force Base, DC
and
Armstrong Laboratory

December 1995

ROBUST FAULT TOLERANT CONTROL: FAULT DETECTION AND ESTIMATION

A. S. Hodel
Associate Professor
Department of Electrical Engineering

Abstract

Real-world applications of computational intelligence (CI) can enhance the fault detection and identification capabilities of a missile guidance and control system. A simulation of a bank-to-turn cruciform missile demonstrates that actuator failure causes the missile performance to degrade. When an actuator for one of the fins becomes sticky, the missile may miss the target. Both the zero order moment term, Z, and the fin rate term, F, show changes during actuator failure. The types of changes observed are similar to variations present in olfactory neural tissue which are shown to be effectively analyzed using neural networks as a filter for clustering and smoothing. Failure of one fin actuator can be detected by using such filters, depicting filter output as "fuzzy numbers." For each term or "sensor" monitored, the fuzzy number output measures the difference between ideal, steady state or recent windows and current windows. Sensor integration is accomplished using a comparing neural network on scaled fuzzy numbers. The Z term fuzzy numbers are used to quickly detect a fault, and the F term is combined with the result for confirmation. The F term is further analyzed to isolate the failed fin. The Z term is tested for high variation between steady state and observed windows. The F term is tested for a high range of rates for the four fins. After the Z term has exhibited high variation for about 0.05 sec, the F term of the failed fin typically drops below the fin rates of the other three fins, creating a large range of rates, and by the end of 2.0 sec, indicating which fin has failed. Of course this pattern is impacted by variations in the missile, the target and the guidance laws employed. These variations make use of a crisp decision boundary less satisfactory than use of the neural-fuzzy risk analysis suggested. Thus, fuzzy numbers are generated and analyzed to 1) detect a fault and 2) determine which fin (if any) failed. Simulations address the following questions: 1) What degree of actuator failure is required in order for *detection* to occur, 2) What degree of actuator failure is required for fault *detection* and *isolation* to occur, 3) Do failure detection and/or isolation require data from both zero order moment and fin rate terms? A suite of target trajectories are simulated, and properties and limitations of the approach reported. In some cases, detection of the failed actuator occurs within 0.1 sec., and isolation occurs within 0.2 sec. If a failed fin is detected early, the guidance law parameters can be modified. If the sticky fin can be identified early, the missile can be rolled to move the sticky fin away from the rudder position, and the fin mapping logic (T Matrix) can be modified. Suggestions for further research are offered.

ROBUST FAULT TOLERANT CONTROL: FAULT DETECTION AND ESTIMATION

A. S. Hodel

1. Introduction

This report addresses the application of robust multivariable control techniques toward the development of an autonomous, robust, fault-tolerant control system for a guided missile. More precisely, the focus of this research was to establish practical techniques for the purpose of *detecting* and, where possible, *identifying* sensor and actuator faults in a missile guidance system. The development of a fault-tolerant control system requires a complete *systems engineering* approach in order to adequately characterize the parameters and uncertainties in the underlying problem as well as to describe a satisfactory control system. A preliminary conception of our proposed work is shown in Figure 1. The solid lines indicate typical missile control loop blocks as currently implemented. The thick dotted lines indicate the scope of the work completed under this research proposal. The thin dotted lines indicate work that remains to be accomplished (gain scheduling based on health monitoring information). A health monitoring system should permit robust operation in the presence of faulty sensors/actuators. The function of such a system would comprise four tasks:

1. continuous monitoring of system performance with respect to design objectives
2. rapid and reliable *detection* of system faults when they occur, without "false alarms."
3. classification/identification of the system faults when detected, and
4. controller reconfiguration (estimator, autopilot, guidance law) in response to fault classification.

This report largely addresses items 1-3 above for the case of a single "sticky" guidance fin.

The investigation summarized in this report primarily addressed the use of *computational intelligence* techniques (e.g., fuzzy logic, artificial neural networks, etc.) for the rapid detection and classification of faults in missile operation. As such, in consultation with our Eglin AFB point of contact it was determined early in the contract that a literature review would be required, culminating in a broad tutorial on available techniques (see Section 2). Following this, preliminary experiments were performed on a missile simulator, modified for the purpose of this investigation so that a selected fin would fail 0.3 seconds into missile flight. Results of these simulations are summarized in Section 3. Conclusions of our study are presented in Section 4.

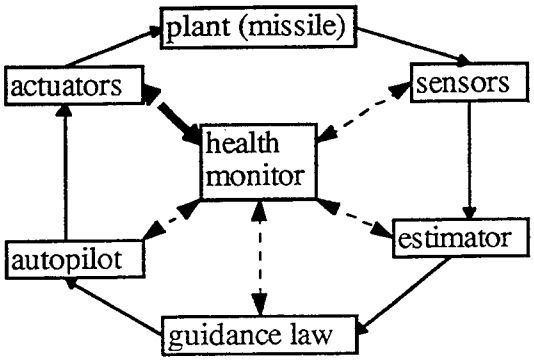


Figure 1: Block diagram conception of fault detection/identification system

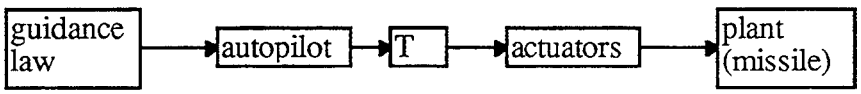
2. Literature survey

The problem of health monitoring, particularly failure detection/classification, is one that lends itself to treatment in a computational intelligence (CI) paradigm. That is, the fundamental problem involved is one of *pattern recognition* and appropriate *classification* in response to given input data. Because of the enormous number of techniques and tools available for application toward a general health-monitoring paradigm, it was decided to develop a general survey of available CI tools and strategies; the fruit of this work is a chapter [12] in the forthcoming *CRC Handbook of Industrial Electronics*. This work delineates several general classes of CI tools: neural networks (NN), fuzzy logic (FZ), evolutionary systems (ES), and virtual reality (VR), each of which provides useful functions in a control design and analysis environment. Related to these tools are several classes of algorithms for feature extraction, or *classification*:: vector quantization (VQ), adaptive resonance theory (ART), self-organizing maps (SOM), learning vector quantization (LVQ), etc. Each of these tools is summarized in [12]. We include a preprint of this chapter in this report for completeness (Appendix A). It should be emphasized that this paper provides a valuable contribution to the literature: this paper not only discusses the tools available, but also the applications to which they are relevant and the steps of the engineering design process in which they should be applied. As such, this work should prove useful to researchers and engineers who seek to effectively employ CI techniques in their respective tasks and projects.

3. Experimental results

The problem addressed in this work is one frequently occurring in aerospace engineering. Available data is noisy and scarce, but proper analysis is critical. Traditional statistical analysis in such situations frequently incorporates nonparametric statistics [1]. Measures of data range and distribution are sought. Many subjective decisions are used to focus the analysis on important data elements having significant impact on the results. Approaches to examining data cluster characteristics are suggested in the following section. These techniques are illustrated in discussion of an analysis of olfactory neurons of a rat [2] and of health monitoring in the fin actuators of a missile guidance system. Subjectivity in analysis can be a powerful tool, if property identified and incorporated when appropriate. Extensions to this work involve the use of a "memorizing" program to automatically monitor domain expert use of computer tools. Decisions and timing can be recorded, analyzed and projected into the design of more interactive analysis systems that incorporate multiple senses. Future applications for these approaches are discussed in a later section.

The basis of the work performed under this contract is as follows. Consider the missile guidance subsystem shown below:



where the matrix T is the *mapping fin* logic used to map the elevator, rudder, and aileron commands $[\delta_e \quad \delta_r \quad \delta_a]^T$ of the autopilot to the command signals $d_c = [d_1 \quad d_2 \quad d_3 \quad d_4]^T$ sent to the four

guidance fin actuators. Since there is an extra degree of freedom in d_c, the matrix T may be selected among a wide range of choices. For example, one may select T based on a zero-moment fin configuration for a hypothetical cruciform missile as

$$\begin{bmatrix}\delta_e\\ \delta_r\\ \delta_a\\ 0\end{bmatrix}=\begin{bmatrix}1 & 1 & -1 & -1\\ 1 & -1 & -1 & 1\\ -1 & -1 & -1 & -1\\ 1 & -1 & 1 & -1\end{bmatrix}\begin{bmatrix}d_1\\ d_2\\ d_3\\ d_4\end{bmatrix} \text{ to obtain } \begin{bmatrix}d_1\\ d_2\\ d_3\\ d_4\end{bmatrix}=T\begin{bmatrix}\delta_e\\ \delta_r\\ \delta_a\\ 0\end{bmatrix} \text{ where } T=\begin{bmatrix}1 & 1 & -1 & 1\\ 1 & -1 & -1 & -1\\ -1 & -1 & -1 & 1\\ -1 & 1 & -1 & -1\end{bmatrix}.$$

Should a fin become "sticky," modeled in our simulations as having a high damping ratio in its actuator dynamics, the zero-moment term of the above equations is no longer satisfied. Our simulations indicate that a sufficiently sticky actuator will cause a missile to fail in its mission. However, based on the use of fuzzy risk analysis, we are able to detect rapidly a single fin failure of this nature, which may enable the use of a gain-scheduled controller for degraded missile performance.

The remainder of this section is organized as follows. First, in Section 3.1 we outline approaches to the problem of detection and isolation of a single failed missile fin. Following this, in Section 3.2, we discuss the results of simulations based on a modified version of the missile simulation developed in [14]. (Much of this section is drawn from the paper [13].)

3.1 Approaches to the problem

Mechanisms suggested for approaching this problem are three-fold: (1) neural network clustering, (2) fuzzy risk analysis for evaluation, and (3) the use of programs that "memorize" an expert's analytic habits. For controls applications, "cloning" an expert (training an artificial neural network to mimic expert data) can initialize a process. Tracking adherence to non-stationary set-points can monitor system state variable shifts. Fuzzy risk analysis can encourage tuning the system to approach sensible performance measures [3,4].

Noisy, scarce data in critical scenarios produces problems often ignored or handled in a very subjective manner. Treating certain aspects of such a controls problem with a combination of neural and fuzzy system tools can increase the system reliability and document the subjectivity for future reuse or modification.

3.2 Results

Examples of analysis health monitoring of a missile guidance system are presented to illustrate these suggestions. (Additional applications of this analysis are presented in [2], where data are analyzed that were drawn from tests of olfactory neurons in the rat. These experiments led to a search for a technique for blending conflicting data from experimental iterations.)

Since the results of the work with Josephson [2] are strongly related to the health monitoring problem results presented in this report, we briefly summarize [2]. Testing for clustering of the data was performed using a simple version of a self-organizing map (SOM) [6]. These clusters were analyzed by using the cluster number as input to a simple multi-layer perceptron (MLP), and the measured data as the output. For each data sample, a training vector was included where the input was

the data class and the output was the measured data. A two-hidden layer architecture was chosen. Enough hidden layer nodes were included to insure rapid and accurate training for the case of no conflicting data. Identical results from each experiment (data class) would thus produce a set of training vectors where each input had a unique output. Such a training set would train quickly to very small rms and maximum errors. The discrepancy between this ideal state and the rms and max errors for conflicting data is a measure of the *degree of conflict* or *consistency* within and between the clusters.

Since the *training performance* is the product of this experiment, the trained network weights are discarded after propagating the training input once and (quickly) recomputed for the next set of data. Consideration is given to Cover's theorems [7,8,9] and the impact of potential memorization of data patterns is neutralized. Training a network to perform on a generalized set of input data is not an objective of this experiment. Instead, the *rms error* of a network which should train quickly but does not is valued.

Fixing the training time at a short time (long enough for a perfect dataset to train) makes practical the implemention of this technique in real-time. The discrepancy between the ideal and observed rms error gives a measure of the size of the variation within clusters. Propagating the cluster number gives a centroid for the cluster. This centroid is an "average" that treats outliers well. The centroid reflects the distribution of "most" of the cluster points, and the maximum error indicates the size of an outlier error. Visualizing the size and closeness of these error balls gives a firm measure for comparing clusters. In the olfactory data, the most distinctive clusters had small error balls and did not overlap much. In the less obvious clusters (according to nonparametric statistics) the error balls were large, and clusters overlapped. Data visualization helped confirm expert observations of distinct neuroanatomical rat olfactory tracts.

Further tests of the technique described showed promising results on multi-criteria decision data based on a rough set experiment. Likewise, the Iris data gave good results. In both cases, the analysis pointed out patterns in the data overlooked by other commentaries, but obviously present. Linking the centroids and error balls from the ANN trained on conflicting data with fuzzy systems and evolutionary systems seems very natural. Using a quickly trained ANN to evaluate the characteristics of a cluster at each time step in a system can indicate slippage from expected relationships or past history. This type of pattern recognition can be invaluable in control systems such as those governing actuators on missiles.

For the detection of an anomaly within 10 msec, missile fin positions and rates are monitored. Using data windows of 10 measurements in 10 msec, comparisons of the "centroids" and ranges of these windows gives an idea of the similarity of these two "clusters" of data. Similar windows are ignored, but differences in windows indicate a problem. If an actuator becomes sticky, response to acceleration commands slows. This immediately sets the zero-moment term of the control algorithm off balance See Figure 2. Fin rates are impacted by smaller than expected accelerations See Figure 3. Other fins compensate by making wide swings and oscillating. Sometimes this action is enough to allow success of the mission, but often not. Since maneuvers also cause swings in the zero moment

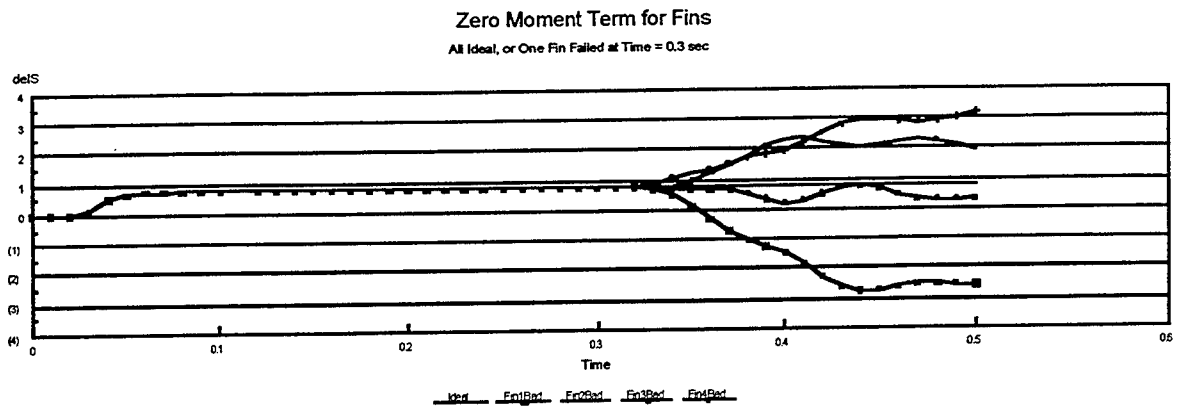


Figure 2. Zero Moment Term for Fins.

term and variations in fin rate among fins, care must be taken in assuming that cluster differences indicate failures. In this particular example, however, a missile can roll to move a questionable fin away from the sensitive rudder-control role, and continue to operate. Thus, a fuzzy number measure of cluster difference seems appropriate as a decision aide to the automatic control system.

Fuzzy risk analysis [4] compares two fuzzy numbers (fuzzy clusters) by subtracting their central points to obtain the risk central point, and taking the worst case and best case differences as risk end points [10]. Examples of these calculations and triangular fuzzy number diagrams illustrate the possibilities and cautions. See Figures 4-5. Numerous traditional methods for estimating a center and spread for a set of numbers exist. In the missile example presented, the data is fairly well-behaved. In cases where conflicting data are available, outliers should be noted, but not allowed to over-bias the estimation of the center. Combining the olfactory and missile example techniques seems a promising approach to health monitoring of a diabetic's dosage regulation system.

4. Conclusions

The handling of data with extreme variance and discontinuities shows promise for use in estimating the status of a diabetic patient in dosage calibration schedules [11]. Intelligent interfaces between patients and computers can be used to monitor the success of treatment and to suggest dosage adjustments or call a physician. Automatically flagging disturbances in expected or recent clusters should be a useful safeguard for many patients since trends are usually obscured by widely fluctuating data.

In summary, there are a number of potential applications in medicine and in industry for this simple filter which gives a centroid, rms error ball and maximum error for a data cluster. The centroid and the error ball size can be treated as a fuzzy number and input to other elements of a computational intelligence system. Data visualization can help in development, but the prime end use of this technique may be as a filter in an intelligent virtual reality system.

Acknowledgments

Much of the labor in this study was performed by Mrs. Mary Lou Padgett, Senior Research Associate in the Department of Electrical Engineering at Auburn University, to whom the principal investigator

expresses fond thanks and gratitude. Thanks are also due to Dr. Walter Karplus, UCLA Department of Computer Science, who served as a consultant on this project. Additional contributions by E. Josephson, J. Evers and A. M. A. Albisser are also much appreciated.

REFERENCES

[1] Conover, W. J. *Practical nonparametric statistics*, 2nd Ed. Wiley & Sons (1980).

[2] E. M. Josephson, M. L. Padgett and D. Buxton. "ANNs: A new tool for neuroanatomical tract-tracing," *1994 Proc. Workshop on Neural Networks, Fuzzy Systems, Evolutionary Systems and Virtual Reality: Academic/Industrial/NASA/Defense*, Washington DC, Dec. 1994 (in press).

[3] Werbos, Paul A. 1996. "Backpropagation to Control," Neural Networks Section, *CRC Handbook on Industrial Electronics*, CRC Press and IEEE Press (in press).

[4] Cooper, A. 1996. "Fuzzy Risk Analysis," Neural Networks Section, *CRC Handbook on Industrial Electronics*, CRC Press and IEEE Press (in press).

[5] M. L. Padgett, "Clustering, simulation and neural networks in real-world applications, Neural Networks, *SPIE Aerospace Sensing*, Orlando, FL April, 1995

[6] Kohonen, T. "The Self-Organizing Map" *Proceedings of the IEEE*, Sept. (1990).in Lau, C (ed.), *Neural Networks Theoretical Foundations and Analysis*, IEEE Press, 1992.

[7] S. K. Rogers and M. Kabrisky. *An introduction to biological and artificial neural networks for pattern recognition*, Vol. TT4, SPIE, 1991).

[8] C. E. Martin, S. K. Rogers, D. W. Ruck, "Nonparametric Bayes error estimation for HRR target identification" *Applications of Neural Networks V*, SPIE Vol. 2243, pp. 2-10 (1994).

[9] Klein, Lawrence A. *Sensor and data fusion concepts and applications*. Vol. TT14. SPIE (1993).

[10] Padgett, M. L. and W. D. Padgett, "Simulation and Computational Intelligence in Real-World Applications," *Simulation*, Vol. 65, No. 1, July, 1995, pp. 5-10.

[11] Albisser, A. M. A. *Diabetes Simulator, V. 1.65* Better Control Medical Computers, Inc., (1994).

[12] M. L. Padgett, P. J. Werbos, and T. Kohonen, "Strategies and Tactics for the Application of Neural Networks to Industrial Electronics." *CRC Handbook of Industrial Electronics*, J. D. Irwin, Editor, To Appear.

[13] M. L. Padgett, "A practical filter for conflicting or subjective data," Proc. Australia and New Zealand Intelligent Information Systems Conference, Perth, Western Australia, Nov. 27, 1995.

[14] M. E. Wallis and J. J. Feeley, "Bank-to-turn missile/target simulation on a desktop computer," *Proceedings of the SCS Western Multiconference on Modeling and Simulation on Microcomputers*, 1989, pp. 79-84.

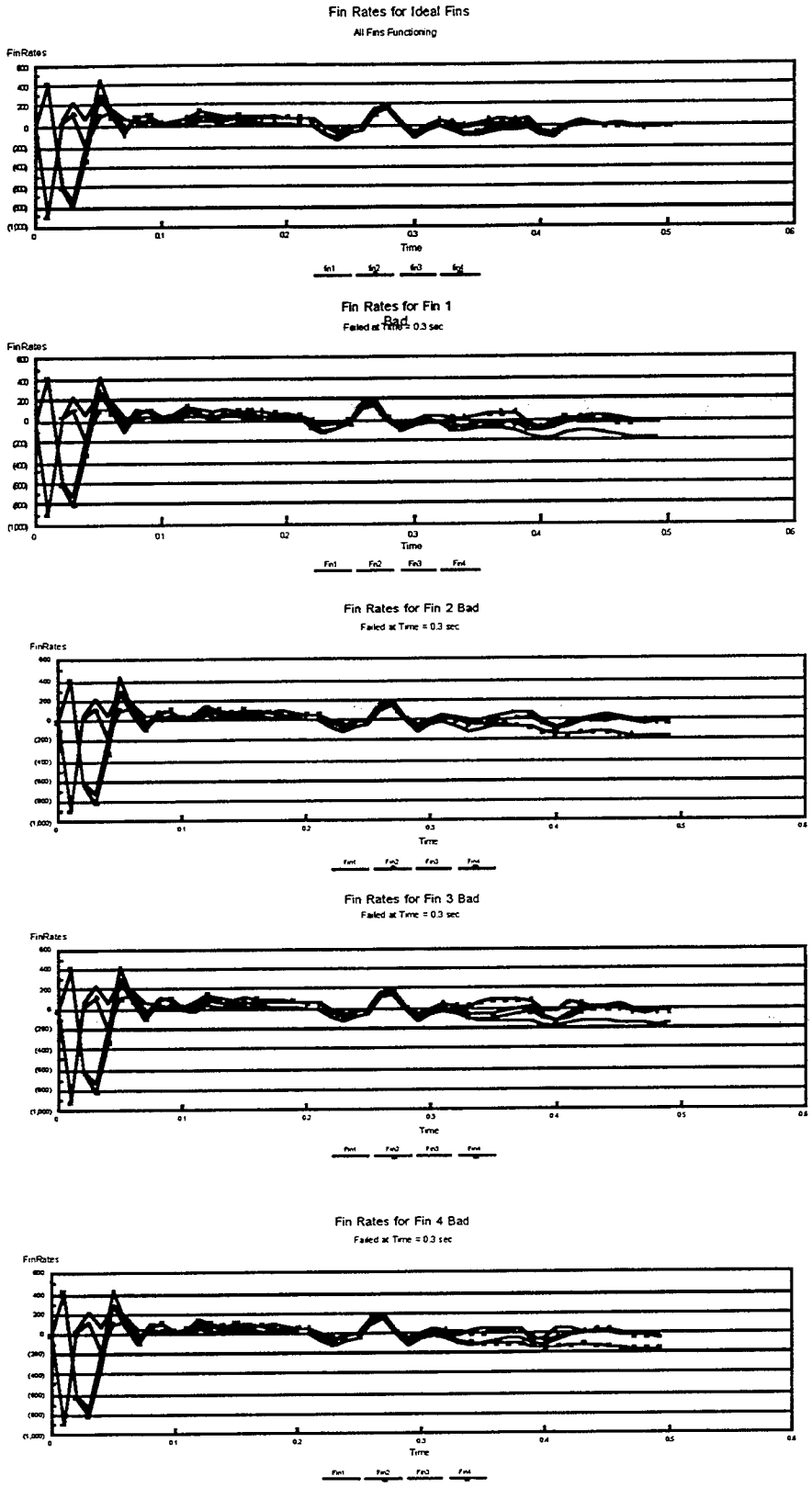


Figure 3. Fin Rates for Ideal Fins and for single failures in fin 1, 2, 3 or 4.

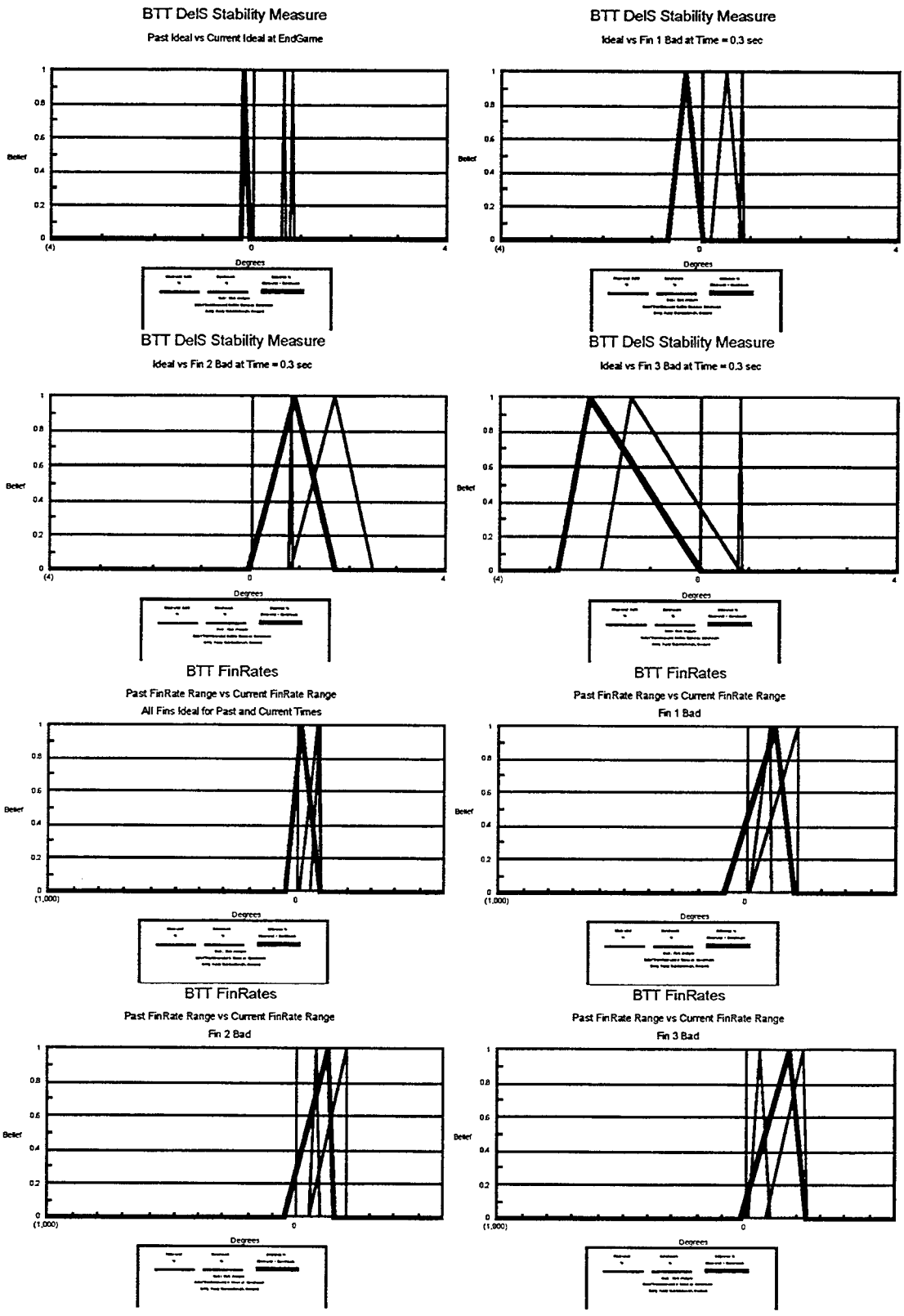


Figure 4. Fuzzy Stability Measure for Zero Moment Term, DelS and for Fin Rates.

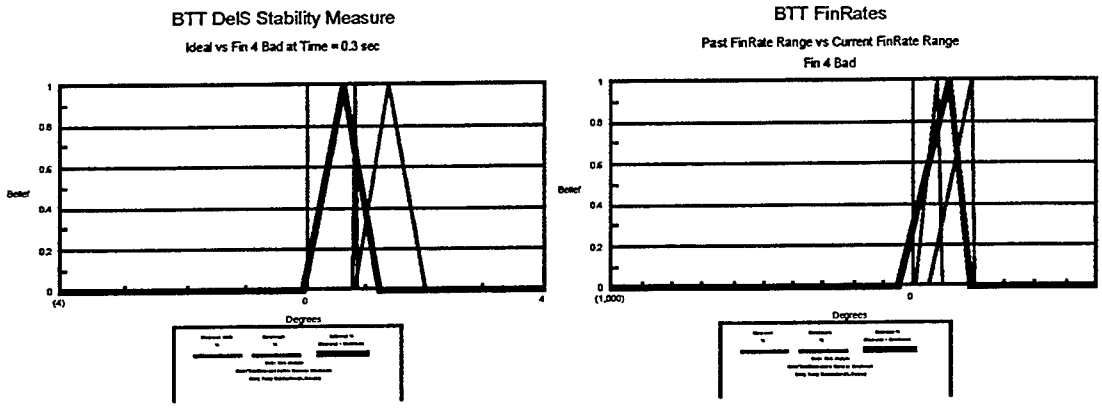


Figure 5. a. Fuzzy Stability Measure; b. Fuzzy Fin Rates

MULTIDIMENSIONAL ALGORITHM DEVELOPMENT AND ANALYSIS

J. Mark Janus
Assistant Professor
Department of Aerospace Engineering

Mississippi State University
P.O. Drawer A
Mississippi State University, MS 39762

Final Report for:
Summer Research Extension Program

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

Mississippi State University

December 1995

MULTIDIMENSIONAL ALGORITHM DEVELOPMENT AND ANALYSIS

J. Mark Janus
Assistant Professor
Department of Aerospace Engineering
Mississippi State University

Abstract

An integral part of advanced warhead design is blast enhancement against targets. Enhancement of wave interaction with solid surfaces involves the modeling of complex physics and can be achieved only through accurate simulation, followed by design modification. A detailed investigation of interaction phenomena leads to a better understanding of the underlying physics involved and subsequently to design modifications. To achieve a detailed solution with sufficient accuracy for a real–world complex configuration, one must use a large number of computational nodes (or cells). The majority of computational tools available to the researcher today simulate the flow physics by utilizing theory which was developed in one–dimension, yet apply it (in a split fashion) in three–dimensions. This lack of incorporating multidimensional physics into the model introduces errors and thus reduces accuracy. The 1980's saw a vast improvement in the "high–resolution" capabilities of flow models. This was due to an increase in the physics (in addition to a decrease in the non–physical numerical errors) introduced into the numerical algorithms. Beginning in the mid–eighties, a segment of the computational community embarked upon the development of the next generation of computational field solvers. These would account for the multidimensional nature of the media as well as retain some of the properties of their "high–resolution" predecessors. Most of this effort naturally lent itself to unstructured data and thus went into algorithms based on unstructured datasets. Unfortunately, much of this latest technology has not found its way into algorithms using structured datasets. The base software is a standard two–dimensional finite–volume high–resolution approximate Riemann solver incorporating Roe averaging. Modifications were made to this software in order to utilize a decoupling technique and multidimensional advection algorithm(s) presented in literature. This report focuses on an implementation of this multidimensional decoupling procedure utilizing fluctuation–splitting theory as well as presenting the implementation of another "genuinely multidimensional upwind solver". Both implementations use cell–centered quadrilateral–based data rather than unstructured cell–vertex triangle–based data.

MULTIDIMENSIONAL ALGORITHM DEVELOPMENT AND ANALYSIS

J. Mark Janus

Introduction

In the gulf war, the Air Force amply demonstrated that precise placement of munitions was effective in inflicting battle damage as well as reducing collateral damage. Todays post–Gulf war modern military will be to asked to defend the nation with fewer, yet more precise, weapons. In this light, the ability of a single warhead to do its job the first time, every time, has been elevated to a higher level as never before. This increased reliance on mission success is a result of the likely reduction in the shear number of warheads due to defense cutbacks. In addition, it is in the best interest of the Air Force to reduce the number of sorties required to accomplish a task, thus reducing exposure to enemy fire. The evolution of warhead delivery has seen it go from hand–dropped shells back in WWI to the laser–guided weaponry seen today. It was made evident in the Gulf war that delivery systems have been improved to the point where individual warheads can be "flown" right down a ventilation shaft if need be. Though this was the case, some of the systems used where essentially conventional "dumb" bombs with sophisticated guidance systems mounted on them. Specialized warhead development leading to advanced warhead design is one avenue which will lead to a capability to do equivalent damage with fewer weapons.

The next generation of warhead will likely be designed utilizing the great power of computational methods to shed new light on the underlying complex physics involved. Enhancement of wave interaction with solid surfaces involves the modeling of complex physics and can be achieved only through accurate simulation and design modification. An accurate computational simulation of this phenomena involves the detailed modeling of a blast wave impinging on a solid surface. To accomplish this task for a real–world complex configuration a researcher will be required to use a large number of computational nodes (or cells) to maintain sufficient accuracy. The computational tools (e.g. flow solvers) available to the researcher today simulate the physics by utilizing theory which was developed in one–dimension, and apply it (in a split fashion) in three–dimensions. This lack of incorporating multidimensional physics into the model introduces errors and thus reduces accuracy (again requiring an increased mesh density). The 1980's saw a vast improvement in the "high–resolution" capabilities of flow models. This was due to an increase in the physics (in addition

to a decrease in the non–physical numerical errors) introduced into the numerical algorithms. Beginning in the mid–eighties, a segment of the computational community embarked upon the development of the next generation of computational field solvers. These would account for the multidimensional nature of the media as well as retain some of the properties of their "high–resolution" predecessors. Most of this effort naturally lent itself to unstructured data and thus went into algorithms based on unstructured datasets. Unfortunately, much of this latest technology has not found its way into algorithms using structured datasets (such as the EAGLE and BEGGAR flow solvers).

Around 1986, Professor Charles Hirsch, et.al. presented a rather significant technique to optimally decouple the multidimensional Euler equations (inviscid equations of fluid motion) [1]. Although seemingly a rather impressive contribution to the computational community, the implementation of this mathematical technique has been quite cumbersome for those researchers so inclined to utilize it. In as such, the broad acceptance of this approach has not yet been borne out. Since that time, investigations into the true multidimensional modeling of the flow physics has yielded some intriguing (yet often complicated) new philosophical approaches to solving the flow in more than one spatial dimension.

Prior to Professor Hirsch's effort, some other researchers [2], [3] had set out on the quest to quantify and rectify the errors produced when one–dimensional operators are applied in a split fashion to facilitate the solution of a multidimensional field problem. From the late eighties to the present much interest has been generated toward the inclusion of multidimensional physics in flow modeling. This is evidenced by the numerous articles available in recent literature each describing the authors own interpretation of how Mother Nature operates and the basics on how to implement this theory in algorithmic (or software) form. Some approaches lean toward the simplicity of utilizing flowfield information to determine the orientation of the one–dimensional Riemann problem [3], [4], [5], [6], [7], while others utilize a fully multidimensional wave decomposition in determining an interface flux function [8], [9], [10]. Skeptics of these algorithmic pioneers would say that a truly successful (versatile) implementation of a multidimensional approach has been questionable (primarily lacking efficiency, accuracy not warranted, etc.). The bottom line has been that you could get comparable quality solutions by increasing the mesh density and using a more conventional algorithm. Bear in mind, that for spatial resolution, a doubling in mesh density (in each of three dimensions) translates to nearly an order of magnitude increase in computational mesh and subsequently runtime. Therefore improvements in the spatial resolution properties of an algorithm yield a significant payoff in computational resources required to accomplish a task.

The effort here has been to develop and analyze a flow model incorporating multidimensional physics with only limited modifications to existing conventional flow software. For this effort the flow domain is restricted to two–dimensions. The base software is a standard two–dimensional finite–volume high–resolution approximate Riemann solver incorporating Roe averaging. Modifications were made to this software in order to utilize Hirsch's decoupling technique [1] and the multidimensional advection algorithm(s) presented in [12]. Two multidimensional decoupling implementations were investigated; the first utilizes the fluctuation–splitting theory outlined in [12] yet does so on cell–centered quadrilateral–based data, the second utilizes the "auxiliary variable" form of the Euler equations to form a genuinely multidimensional upwind solver [13].

Methodology #1

The two–dimensional Euler equations in conservative form are written:

$$\boldsymbol{Q}_t + \boldsymbol{F}_x + \boldsymbol{G}_y = \boldsymbol{0} \quad . \tag{1}$$

where $\boldsymbol{Q}$ is the vector of conserved variables and $\boldsymbol{F}$ and $\boldsymbol{G}$ are the flux vectors:

$$\boldsymbol{Q} = \begin{bmatrix} \varrho \\ \varrho u \\ \varrho v \\ \varrho E \end{bmatrix} \quad , \quad \boldsymbol{F}_x = \begin{bmatrix} \varrho u \\ \varrho u^2 + p \\ \varrho uv \\ \varrho uH \end{bmatrix} \quad , \quad \boldsymbol{G}_y = \begin{bmatrix} \varrho v \\ \varrho uv \\ \varrho v^2 + p \\ \varrho vH \end{bmatrix} \quad .$$

In [1], Hirsch proceeds to develop a system of diagonal, decoupled transport equations which are "fully equivalent to the original system of Euler equations in conservative form", This system can be written as:

$$\boldsymbol{W}^*_t + D^x \boldsymbol{W}^*_x + D^y \boldsymbol{W}^*_y = \boldsymbol{S} \quad . \tag{2}$$

where $\boldsymbol{W}^*$ is a vector of advected quantities (entropy, a component of velocity, and two acoustic–like variables), D^x and D^y are diagonal matrices of advection speeds, and $\boldsymbol{S}$ is a source term:

$$\partial \boldsymbol{W}^* = \begin{bmatrix} \partial\varrho - \frac{1}{c^2}\partial p \\ \boldsymbol{l}^{(1)} \cdot \partial \boldsymbol{V} \\ \boldsymbol{k}^{(2)} \cdot \partial \boldsymbol{V} + \frac{\partial p}{\varrho c} \\ -\boldsymbol{k}^{(2)} \cdot \partial \boldsymbol{V} + \frac{\partial p}{\varrho c} \end{bmatrix} \quad , \quad \boldsymbol{S} = \begin{bmatrix} 0 \\ -\frac{c}{2}\,(\boldsymbol{l}^{(1)} \cdot \nabla)\,(W^3 + W^4) \\ -c\,(\boldsymbol{l}^{(1)} \cdot \nabla)\,W^2 \\ -c\,(\boldsymbol{l}^{(1)} \cdot \nabla)\,W^2 \end{bmatrix} \quad .$$

Hirsch showed that a particular choice of the vectors $\boldsymbol{k}^{(1)}$ and $\boldsymbol{k}^{(2)}$ will minimize the source term and hence will optimally decouple the system (note: the vectors $\boldsymbol{l}^{(1)}$ and $\boldsymbol{l}^{(2)}$ are perpendicular to $\boldsymbol{k}^{(1)}$ and $\boldsymbol{k}^{(2)}$, respectively and also, in general, the system holds for any selection of $\boldsymbol{k}^{(1)}$ and $\boldsymbol{k}^{(2)}$). Hirsch

showed that in order to minimize the source terms, $\boldsymbol{k}^{(1)}$ needs to be aligned locally to the pressure gradient and that $\boldsymbol{k}^{(2)}$ is related to the strain–rate tensor. The choice of these vectors was part of the investigation here and will be discussed in more detail shortly. The solution evolves in time according to Eq. (2) and is written as a set of scalar equations of the form:

$$W^{*k}_t + \lambda^k \cdot \nabla W^{*k} = S^k \quad . \tag{3}$$

The relationship (transformation matrix) between the vector of advected quantities and the conservative variables is given in [1] as:

$$P = \begin{vmatrix} 1 & 0 & \frac{\varrho}{2c} & \frac{\varrho}{2c} \\ u & \frac{\varrho k_y^{(2)}}{K} & \frac{\varrho}{2cK}(uK + ck_x^{(1)}) & \frac{\varrho}{2cK}(uK - ck_x^{(1)}) \\ v & \frac{-\varrho k_x^{(2)}}{K} & \frac{\varrho}{2cK}(vK + ck_y^{(1)}) & \frac{\varrho}{2cK}(vK - ck_y^{(1)}) \\ \frac{\boldsymbol{V}\cdot\boldsymbol{V}}{2} & \frac{\varrho}{K}(\boldsymbol{l}^{(2)} \cdot \boldsymbol{V}) & \frac{\varrho}{2cK}(HK + c\boldsymbol{V}\cdot\boldsymbol{k}^{(1)}) & \frac{\varrho}{2cK}(HK - c\boldsymbol{V}\cdot\boldsymbol{k}^{(1)}) \end{vmatrix} ,$$

with

$$K = \boldsymbol{k}^{(1)} \cdot \boldsymbol{k}^{(2)} \quad , \tag{4}$$

$$\partial \boldsymbol{Q} = P \; \partial \boldsymbol{W}^* \quad . \tag{5}$$

Note, in this matrix the dot product of the two vectors $\boldsymbol{k}^{(1)}$ and $\boldsymbol{k}^{(2)}$ appears in the denominator, thus this dot product cannot be allowed to go to zero (i.e. the choice of vectors cannot be such that the vectors are perpendicular to one another). If by Hirsch's optimum selection criteria this should occur, then the second vector is chosen as $\boldsymbol{k}^{(2)} = \boldsymbol{k}^{(1)}$. One area of interest here is in the choice of the second wave vector $\boldsymbol{k}^{(2)}$. Some ambiguity as to the "proper" choice to minimize the source term is found in literature. Furthermore none of the approaches to minimize the source term associated with the second wave vector appear to actually zero the term. Several approaches were investigated in this effort. To further add to this issue, the advection "direction" of the source term is also open for interpretation. In the work by Bermudez and Vazquez [14], various approaches were studied involving upwind methodologies for hyperbolic conservation laws with source terms. Some very interesting results were found regarding the treatment of the source term and the resulting numerical properties of the solution (eg. erroneous wave propagation speeds associated with a particular handling of the source term). In the attractive approach described in [14], the source term is split

into two parts with the directional splitting coming from the systems eigenstructure. It is believed that the handling of the source terms in this effort is consistent with that directional splitting concept.

Fundamental to Hirsch's development of Eq. (2) and the optimal decoupling procedure is the diagonalization of a linear sum of the flux Jacobian matrices. The flux Jacobian matrices originate from casting the Euler equations in quasilinear form beginning with the conservative form, Eq. (1). Maintaining discrete conservation while utilizing equations which are not in conservative form requires instituting a particular linearization procedure [15]. As stated in [12], for a triangle–based mesh, this results in taking the arithmetic average of the values of Roe's parameter vector **Z** [16] given as:

$$\boldsymbol{Z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} \sqrt{\varrho} \\ \sqrt{\varrho}\,u \\ \sqrt{\varrho}\,v \\ \sqrt{\varrho}\,H \end{bmatrix} \quad , \tag{6}$$

at the triangle vertices, thus

$$\overline{\boldsymbol{Z}} = \frac{\boldsymbol{Z}_1 + \boldsymbol{Z}_2 + \boldsymbol{Z}_3}{3} \quad . \tag{7}$$

Therefore except for the gradient terms, all data in the decoupled equations (e.g. wave speed and the transformation matrix P) is constructed from this specially averaged data, **Z**. This is analogous to the use of "Roe averaged" variables in one dimension [16]. The procedure implemented here on quadrilateral grids for using this theory, which requires the use of triangular "elements" in two–dimensions, will be explained next.

Consider the quadrilateral grid shown in Fig. 1, with cell centers indicated. Triangular elements with data stored at the vertices (as is necessary for utilizing the theory presented in [12]) can be formed by simply connecting the cell centers with a mesh and then taking the diagonal (either one) of the newly formed sub–grid, see Fig. 2. The coordinates of the primary mesh cell centers can be obtained geometrically by finding the intersection of the diagonals of each primary cell. Each primary cell center is the common vertex of six triangular elements from the sub–grid, refer to Fig. 2. The procedure outlined in [12] for the advection of scalar variables according to Eq. (3) can now be utilized. The scheme referred to as LDA (Low Diffusion A) was used in this study.

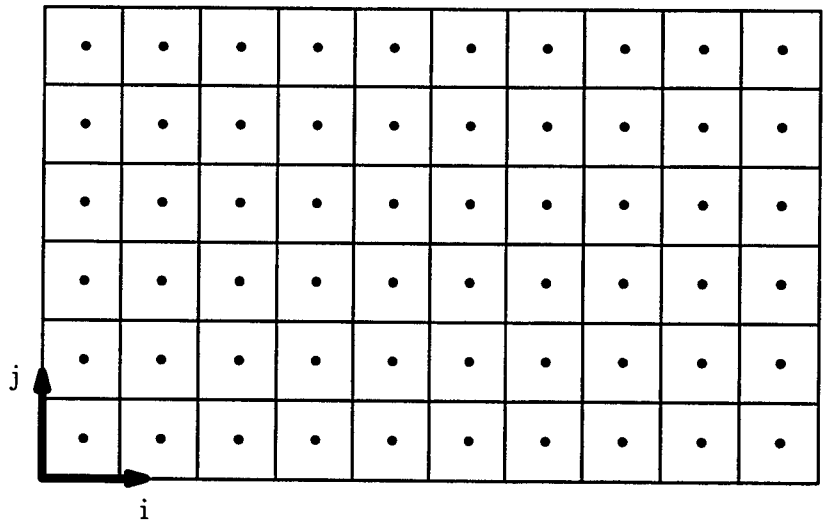


Figure 1. Primary Quadrilateral Grid

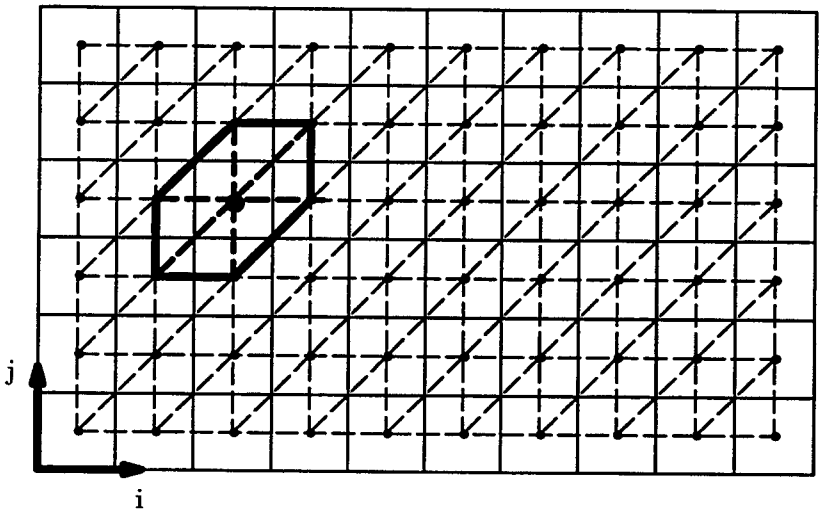


Figure 2. Sub–grid Development

The value of the dependent variables for each cell center was updated according to the following:

$$\boldsymbol{Q}_{i,j}^{n+1} = \boldsymbol{Q}_{i,j}^{n} - \frac{\Delta t}{A_{i,j}} \sum_{T} \sum_{k} \beta_{T,i,j}^{k} A_T \left(\nabla W^{*k} \cdot \bar{\lambda}^{k} \right) \overline{\boldsymbol{R}}^{k} \quad , \tag{8}$$

where the summation index T carries over all triangles having (i,j) as a common vertex, while $\beta_{T,i,j}^{k}$ represents the fraction of the residual of the k^{th} wave in sub–grid element T sent to (i,j). $A_{i,j}$ is the area of the primary cell, whereas A_T is the area of the sub–grid triangular element T, and the bar ($\bar{}$) quantities indicate evaluation at the special "average" state of the sub–grid element T. $\overline{\boldsymbol{R}}^{k}$ are the columns of the matrix P. Presently, the boundaries are maintained using the primary grid,

phantom cells, and characteristic variable boundary conditions [17]. This appears to work yet may be the source of some solution difficulties encountered during this effort.

The modifications to the base software included a routine to locate the coordinates of each cell center, a routine to determine the flowfield gradients for each triangular element of the sub–grid and subsequently the vectors $\boldsymbol{k}^{(1)}$ and $\boldsymbol{k}^{(2)}$ based on this data, and a replacement routine for that which previously computed the residual based on a flux balance of each cell (finite volume).

Methodology #2

A slightly different (and possibly easier to follow) approach toward multidimensional upwinding was investigated during the course of this project. Consider the following quasilinear non–conservative form of the Euler equations in the auxiliary variables (s, u, v, p)

$$\begin{aligned} & s_t + us_x + vs_y = 0 \\ & \varrho v_t + \varrho u v_x + \varrho v v_y + p_y = 0 \\ & \varrho u_t + \varrho u u_x + \varrho v u_y + p_x = 0 \\ & p_t + up_x + vp_y + \varrho c^2 (u_x + v_y) = 0 \end{aligned} \tag{9}$$

where $ds = d\varrho - \dfrac{dp}{c^2}$.

The fluctuations of the system can be written as

$$\boldsymbol{r} = \boldsymbol{r}^x + \boldsymbol{r}^y \tag{10}$$

where

$$\boldsymbol{r}^x = -A_T \overline{A} \cdot \left(\hat{s}_x,\ \hat{\varrho u}_x,\ \hat{\varrho v}_x,\ \hat{p}_x \right)^T \tag{11}$$

$$\boldsymbol{r}^y = -A_T \overline{B} \cdot \left(\hat{s}_y,\ \hat{\varrho u}_y,\ \hat{\varrho v}_y,\ \hat{p}_y \right)^T \tag{12}$$

with

$$\overline{A} = \begin{vmatrix} \overline{u} & 0 & 0 & 0 \\ 0 & \overline{u} & 0 & 1 \\ 0 & 0 & \overline{u} & 0 \\ 0 & \overline{c}^2 & 0 & \overline{u} \end{vmatrix}, \quad \overline{B} = \begin{vmatrix} \overline{v} & 0 & 0 & 0 \\ 0 & \overline{v} & 0 & 0 \\ 0 & 0 & \overline{v} & 1 \\ 0 & 0 & \overline{c}^2 & \overline{v} \end{vmatrix},$$

and

$$
\begin{aligned}
\overline{u} &= \frac{\overline{z}_2}{\overline{z}_1} \\
\overline{v} &= \frac{\overline{z}_3}{\overline{z}_1} \\
\overline{H} &= \frac{\overline{z}_4}{\overline{z}_1} \\
\overline{c}^2 &= (\gamma - 1)\left[\overline{H} - \frac{1}{2}\left(\overline{u}^2 + \overline{v}^2\right)\right] \\
\hat{\varrho_x} &= 2\ \overline{z}_1(z_1)_x \\
\hat{\varrho u_x} &= \overline{z}_1(z_2)_x - \overline{z}_2(z_1)_x \\
\hat{\varrho v_x} &= \overline{z}_1(z_3)_x - \overline{z}_3(z_1)_x \\
\hat{p_x} &= \frac{\gamma - 1}{\gamma}\left[(\overline{z}_4(z_1)_x + \overline{z}_1(z_4)_x) - (\overline{z}_2(z_2)_x + \overline{z}_3(z_3)_x)\right]
\end{aligned}
\tag{13}
$$

where $z_i,\ i = 1, 2, 3, 4$ were defined in the last section.

The corresponding terms involving derivatives in the y direction can be written in an analogous manner. Introducing the matrix

$$
C_a = \begin{vmatrix}
1 & 0 & 0 & \frac{1}{\overline{c}^2} \\
\overline{u} & 1 & 0 & \frac{\overline{u}}{\overline{c}^2} \\
\overline{v} & 0 & 1 & \frac{\overline{v}}{\overline{c}^2} \\
\frac{\overline{u}^2 + \overline{v}^2}{2} & \overline{u} & \overline{v} & \frac{1}{\gamma - 1} + \frac{\overline{u}^2 + \overline{v}^2}{2\overline{c}^2}
\end{vmatrix},
$$

Referring to Fig. 3, the fluctuation is then distributed according to the following formulae:

$$
\begin{aligned}
A_{i,j}Q_{i,j}^{n+1} &= A_{i,j}Q_{i,j}^{n} + \frac{\Delta t}{2}C_a\left(\boldsymbol{r}^x + \tilde{\boldsymbol{r}}^x\right) \\
A_{i+1,j}Q_{i+1,j}^{n+1} &= A_{i+1,j}Q_{i+1,j}^{n} + \frac{\Delta t}{2}C_a\left[\left(\boldsymbol{r}^x - \tilde{\boldsymbol{r}}^x\right) + \left(\boldsymbol{r}^y + \tilde{\boldsymbol{r}}^y\right)\right] \\
A_{i+1,j+1}Q_{i+1,j+1}^{n+1} &= A_{i+1,j+1}Q_{i+1,j+1}^{n} + \frac{\Delta t}{2}C_a\left(\boldsymbol{r}^y - \tilde{\boldsymbol{r}}^y\right)
\end{aligned}
\tag{14}
$$

where

$$
\begin{aligned}
\tilde{\boldsymbol{r}}^x &= sign(\overline{A})\boldsymbol{r}^x \\
\tilde{\boldsymbol{r}}^y &= sign(\overline{B})\boldsymbol{r}^y
\end{aligned}
\tag{15}
$$

yields a scheme which is essentially a standard dimensionally split method with a slight variation in the Roe linearization procedure.

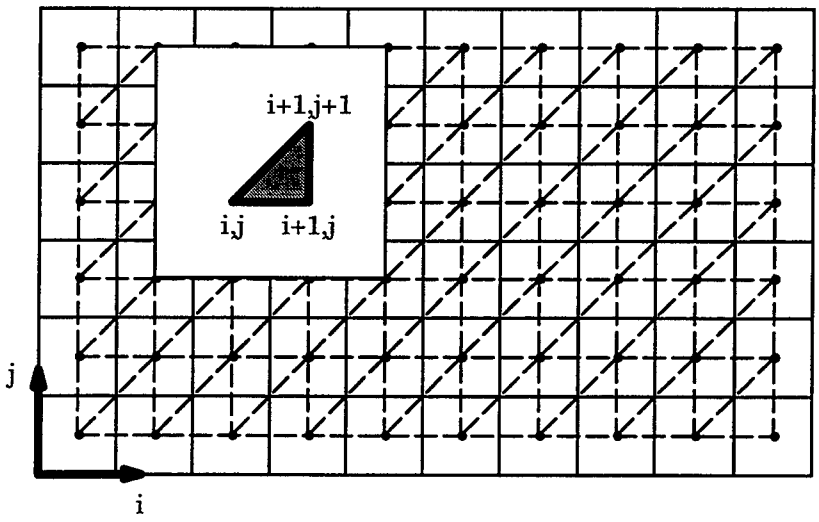


Figure 3. Sub–grid Triangulation I

A genuinely two–dimensional, linearity preserving, second–order accurate scheme can be constructed by the introduction of the following vectors in place of r^x, r^y, respectively in Eqs. (14),

$$r_i^{x^*} = r_i^x + \Psi(q_i) r_i^y \tag{16}$$

$$r_i^{y^*} = r_i^y + \frac{\Psi(q_i)}{q_i} r_i^x$$

for $i = 1, 2, 3, 4$, with

$$q_i = -\frac{r_i^x}{r_i^y} \tag{17}$$

where Ψ is a (non–compressive) limiter.

Results

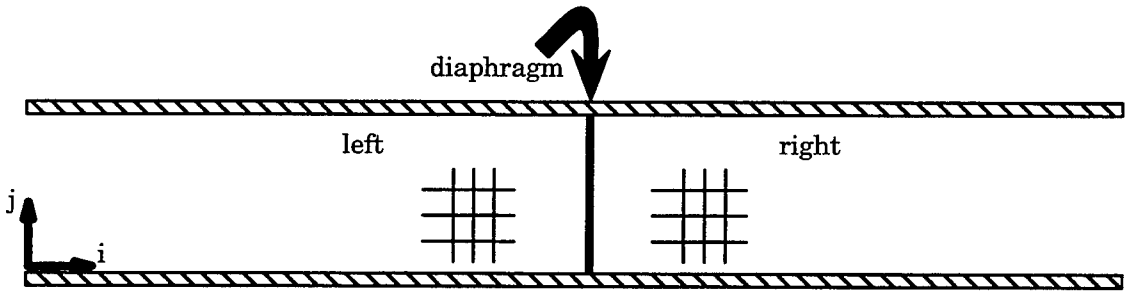


Figure 4. Shock Tube Configuration

In order to test the software developed in this effort, two test cases were employed. The first was that of a simple shock tube modeled with a two–dimensional domain. Although the initial condition and solution are strictly one–dimensional (involving planar wave fronts traveling axially), this is a good test to determine adequacy of propagation of waves not aligned with the grid (by creating

grids which slant) and is sufficiently simple to detect anomalous software behavior. The conditions across the shock tube diaphragm (i.e. the initial conditions) were a left to right pressure ratio of 10 to 1, a left to right density ratio of 8 to 1, and still air (no flow), see Fig. 4. Also, the upper and lower boundary conditions are handled two separate ways. The first trials utilized a periodic boundary condition to remove any influence of a "wall" boundary condition, and the second trials utilized a phantom cell type boundary to enforce the conditions consistent with a solid wall.

For a vertical grid (i.e. the cells of the quadrilateral grid are squares), the solutions coming from the fluctuation–split code and the multidimensional upwind code were identical to that produced from a one–dimensional finite–volume approximate Riemann solver using Roe averaging (see Fig. 5). The image shown in the figure was developed using a technique referred to as numerical

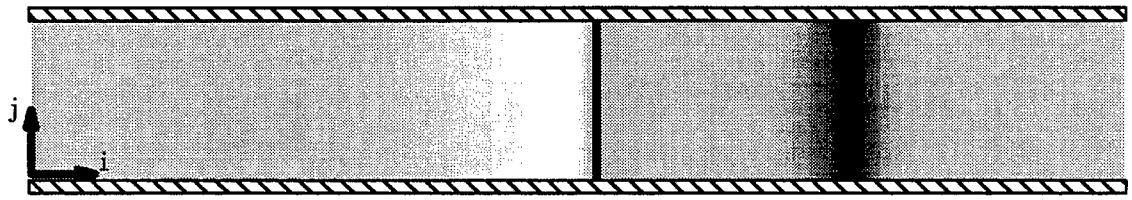


Figure 5. Standard Roe Solver Using Vertical Grid

Schlieren imagery. This approach essentially depicts density gradients (dotted with the velocity field) as light and dark regions to visualize the solution for this test case. When the grid is slanted (using periodic boundary conditions top and bottom), the solution deteriorates (note the broadened shock front) for the first–order standard Roe solver as shown in Fig. 6. The fluctuation–split solver and the multidimensional upwind solver both exhibit narrower shock fronts than the standard solver, refer to Figs. 7 and 8, respectively.The slanted grid was also tested with true impermeable surface conditions for the top and bottom boundaries. Some degree of solution degradation is observed in the images of all three approaches, although the fluctuation–split method appears to yield the most satisfactory results.

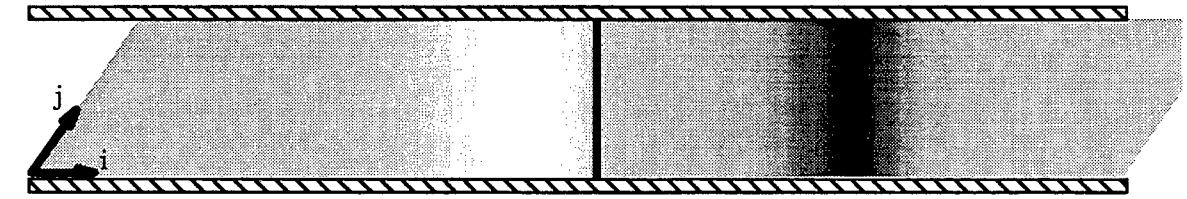


Figure 6. Standard Roe Solver Using Slanted Grid and Periodic BC

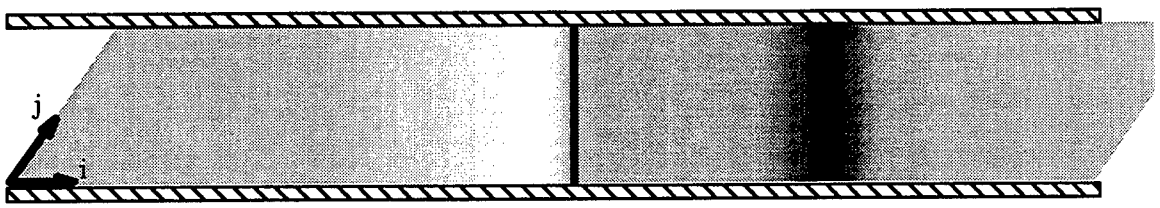


Figure 7. Fluctuation–split Solver Using Slanted Grid and Periodic BC

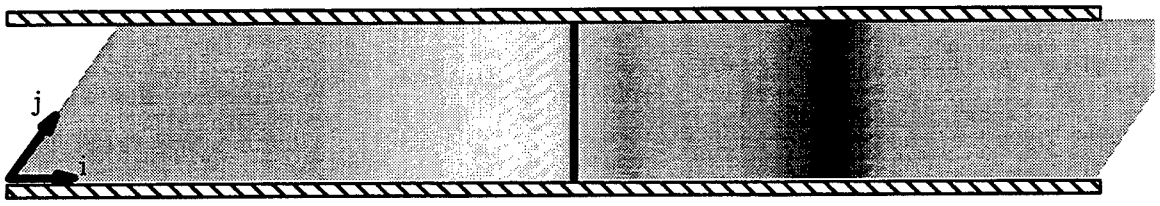


Figure 8. Multidimensional Upwind Solver Using Slanted Grid and Periodic BC

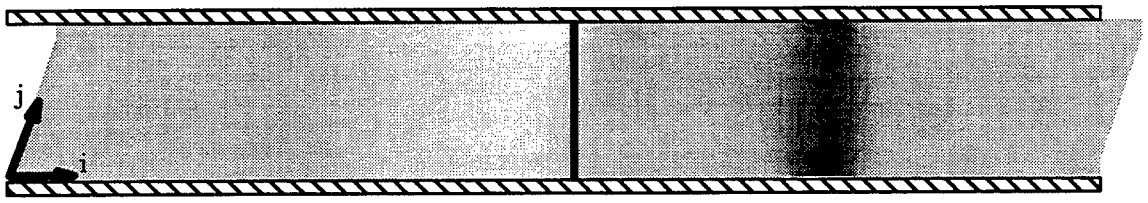


Figure 9. Standard Roe Solver Using Slanted Grid and Wall BC

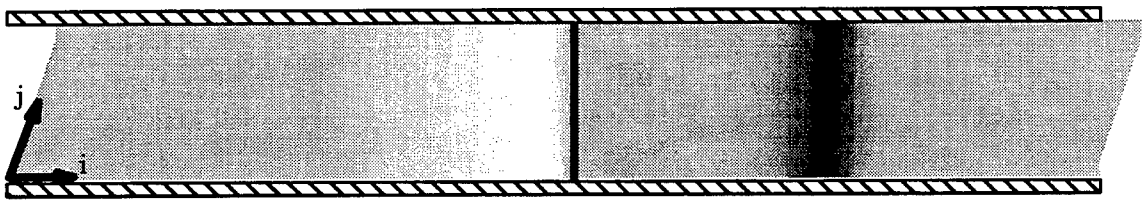


Figure 10. Fluctuation–split Solver Using Slanted Grid and Wall BC

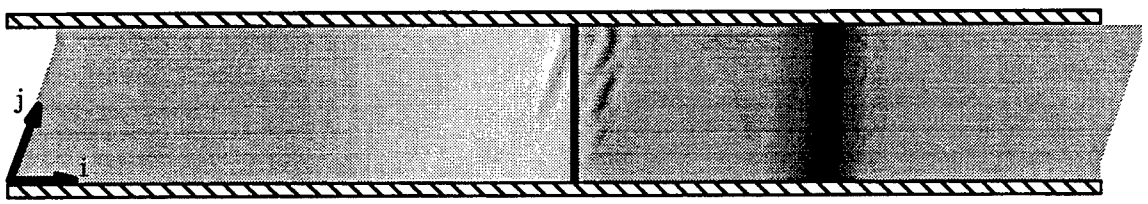


Figure 11. Multidimensional Upwind Solver Using Slanted Grid and Wall BC

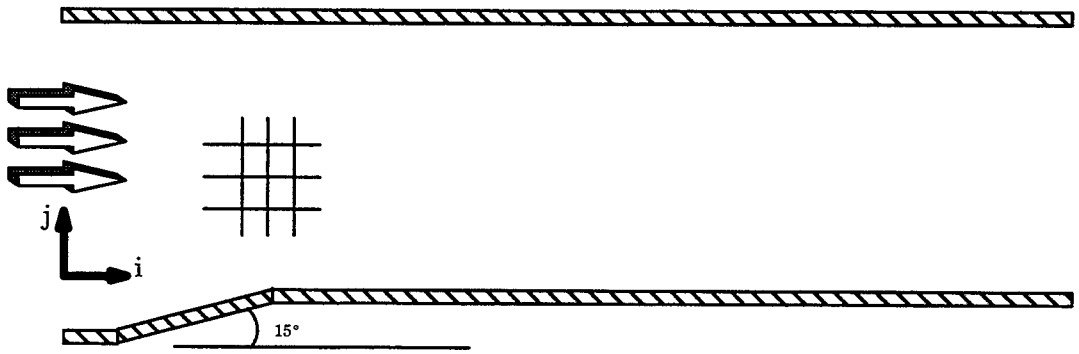


Figure 12. 15° Confined Ramp Configuration

The second test case was that of a 15° confined ramp with an inlet Mach number of 1.9, see Fig. 12. The 15° turn angle results in an oblique shock cutting across the channel at approximately 48° to horizontal. This oblique shock then reflects off the upper wall (again turning the flow), but this time the turn angle is too great to be accomplished with an oblique shock and thus a Mach stem forms with a normal shock at the upper wall. This is followed downstream with further reflections and interaction with an expansion fan emanating from the ramps downstream corner. Due to the complexity of this two–dimensional flowfield an exact solution is difficult to obtain (one could possibly use method of characteristics). Rather, as a comparison 'standard', a solution from a conventional high–resolution algorithm with a fine grid and third–order accuracy is shown in Fig. 13. The

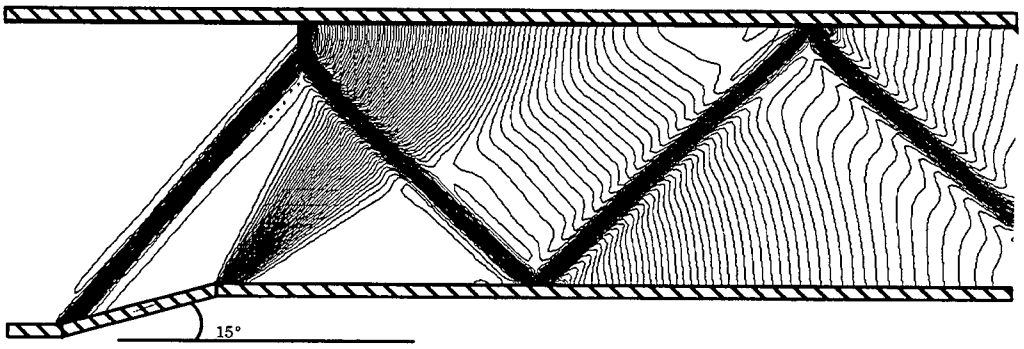


Figure 13. Pressure Contours of Third–Order Fine Grid Solution (standard Roe solver)

figure shows pressure contours making the oblique shocks and expansion fan obvious. Consider also the numerical Schlieren image shown in Fig. 14. This solution was passed through the routine which determines the wave propagation vectors $\boldsymbol{k}^{(1)}$ and $\boldsymbol{k}^{(2)}$ to qualitatively assess the performance of that routine, see Figs. 15 and 16.

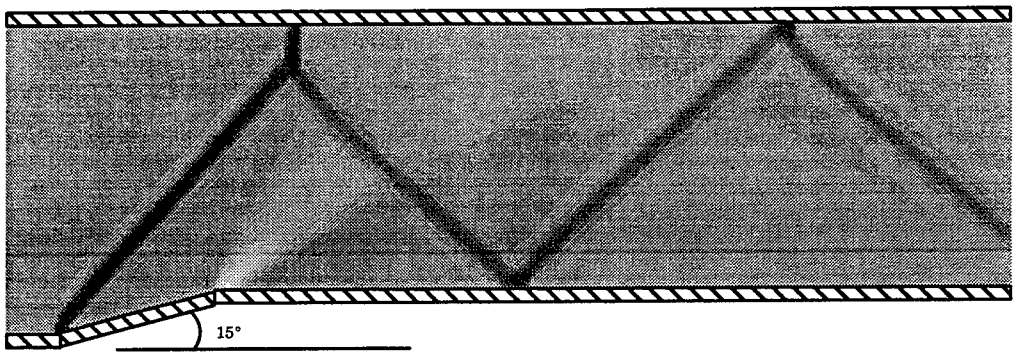


Figure 14. Numerical Schlieren of Third–Order Fine Grid Solution (standard Roe solver)

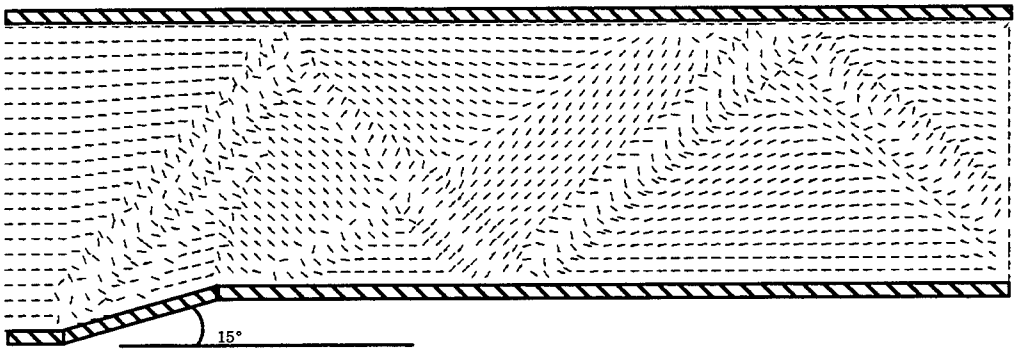


Figure 15. First Wave Vector $\boldsymbol{k}^{(1)}$ (aligned with pressure gradient)

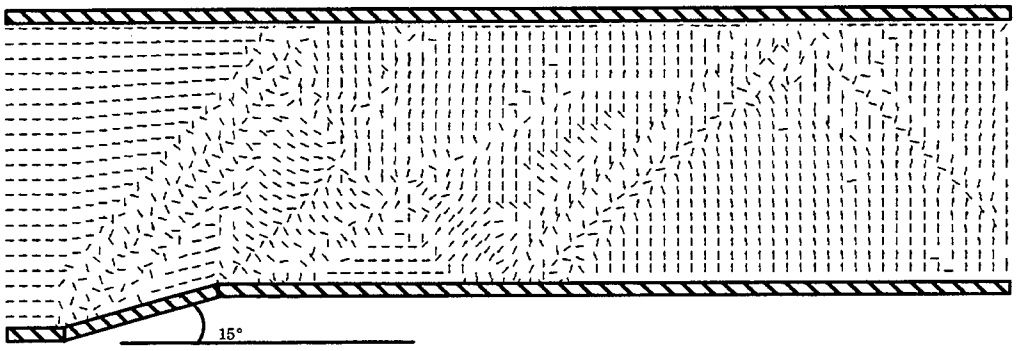


Figure 16. Second Wave Vector $\boldsymbol{k}^{(2)}$ (chosen here to truly zero the source term)

Achieving a converged solution using the software modified with the fluctuation–splitting theory has thus far proven to be difficult to say the least. One problem area investigated was the selection of the wave vectors. As the solution evolves in time, the wave vectors are constantly adjusting to it and as such cause the solution to further change. Other researchers have noted that this feedback hinders convergence. Unfortunately, infrequently updating the wave vectors has not corrected the

problem and sometimes results in solution divergence. Consider the solution shown in Figs. 18 and 17. It should be noted that the solution produced using the fluctuation split code has only one–

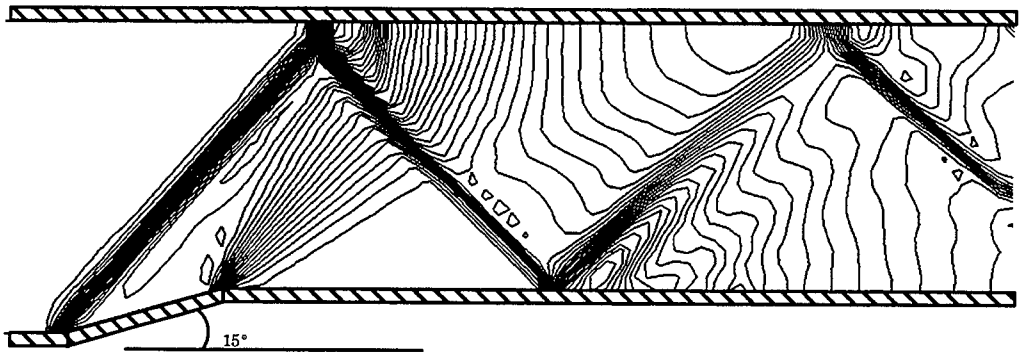


Figure 17. Pressure Contours Using Fluctuation–split Solver

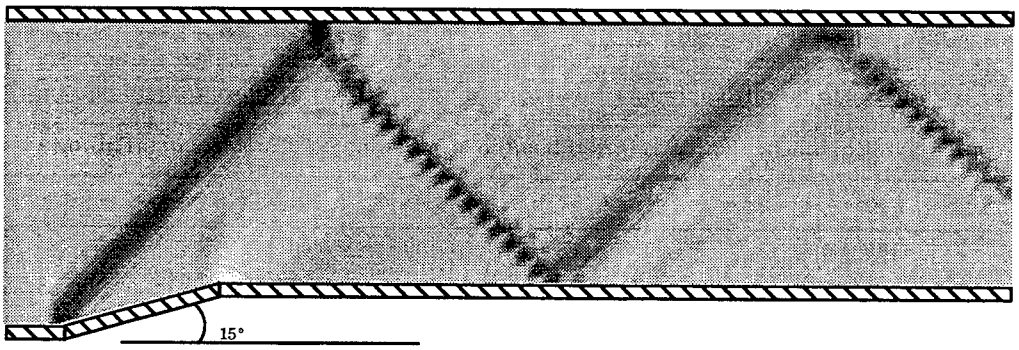


Figure 18. Numerical Schlieren Using Fluctuation–split Solver

quarter the mesh cells of that shown in Fig. 13. For true comparison purposes, consider the solution shown in Fig. 19. This solution was produced using a conventional algorithm, first–order accuracy

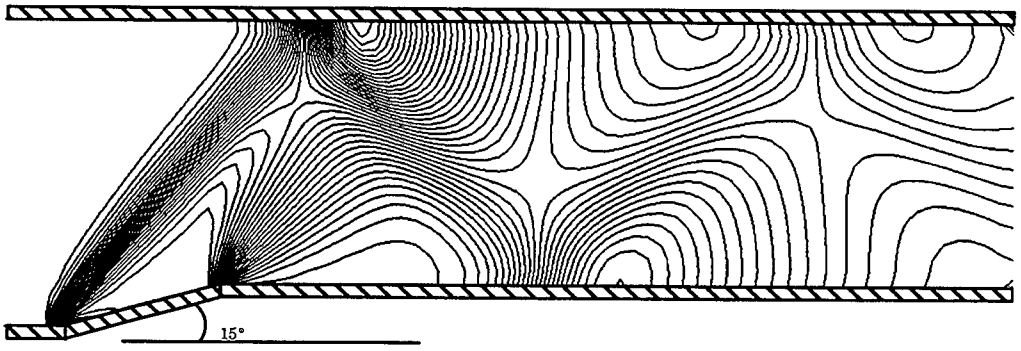


Figure 19. Pressure Contours of First–Order Coarse Grid Solution (standard Roe Solver)

and the same primary grid as that used for the fluctuation split code. Note the degradation in reso-

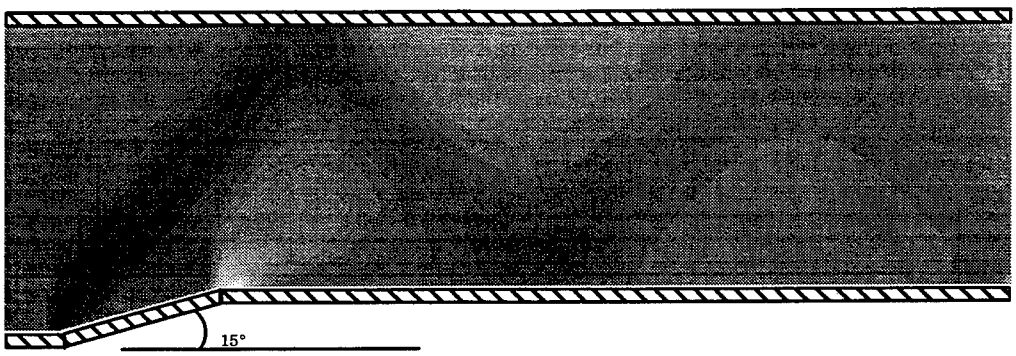


Figure 20. Numerical Schlieren of First–Order Coarse Grid Solution (standard Roe Solver)

lution for the oblique shocks is so great they begin to be indistinguishable on down the channel. Although this is the case, the angles (both incident and reflected) seem to agree with the fine grid solution quite well.

Conclusions

The continuation of this project has led to determining some of the problem(s) with the implementation which causes convergence difficulties. The solution feedback through the wave vectors ($\boldsymbol{k}^{(1)}$ and $\boldsymbol{k}^{(2)}$) is a definite source of this problem. As the solution evolves in time the wave vectors are continually adjusting to the solution, yet the solutions numerical properties depend largely on the choice of wave vector directions. This feedback has been mentioned in literature as a source of convergence degradation. In addition, under certain circumstances no wave vector directions can be found which completely decouple the equations and thus the source term is left nonzero. An effort toward the proper selection of wave vector direction to ensure a minimal source term and an effort toward the proper numerical treatment (propagation) of the source term were investigated. Results to date indicate the procedure to distribute the source term used here is consistent with the findings of other parties studying hyperbolic equations with source terms.

The present boundary conditions in use are those involving standard phantom cell boundary maintenance. This is the preferred mode of operation and appears to be compatible with the altered algorithm. Tests showed that the mirror cell concept is a viable approach when using the multidimensional algorithm, yet due to the detailed use of cell center information the precise construction of the phantom cell is crucial to proper boundary condition implementation. This is in contrast to the standard Roe solver in which the details of the phantom cell can remain ambiguous and still yield satisfactory results.

Recent multidimensional experience has indicated that conventional algorithms in general do not live up to their traditional high–resolution capabilities. The effort here has been to develop and analyze a flow model incorporating multidimensional physics with only limited modifications to existing conventional flow software. For this effort the flow domain has been restricted to two–dimensions. The base software was a conventional finite–volume high–resolution approximate Riemann solver incorporating Roe averaging. Modifications were made to this software in order to utilize Hirsch's decoupling technique and multidimensional advection algorithm(s) presented in literature. This report focuses on an implementation of the multidimensional decoupling procedure which utilizes the fluctuation splitting theory outlined in literature yet does so on cell–centered quadrilateral–based data rather than on unstructured cell–vertex triangle–based data. As evidenced by the results shown here, the code is still in the research phase. In addition, another multidimensional procedure was investigated. This second methodology appears to be less complicated conceptually yet the results obtained to–date are not quite as sharp as that obtained from the fluctuation–split code. Several goals of this project have been met, (eg. the modifications to the base software included just three new routines with one replacing an original routine). Thus far the solutions obtained show the implementation has promise, and how they stack up to solutions produced from more conventional software (compared side by side) has been shown.

References

[1] Ch. Hirsch, C. Lacor and H. Deconinck. Convection algorithms based on a diagonalization procedure for the multidimensional euler equations. AIAA Paper No. 87–1163, 1987.

[2] S. Davis. A rotationally biased upwind difference scheme for the Euler equations. *Journal of Computational Physics*, 56:65–92, 1984.

[3] P.L. Roe. Discrete models for the numerical analysis of time–dependent multidimensional gas dynamics. *Journal of Computational Physics*, 63:458–476, 1986.

[4] D. W. Levy. Use of a rotated Riemann solver for the two–dimensional Euler equations. Ph.D. thesis, University of Michigan, 1990.

[5] A. Dadone and B. Grossman. A rotated upwind scheme for the Euler equations. AIAA Paper No. 91–0635, 1991.

[6] Y. Tamara and K. Fuji. A multidimensional upwind scheme for the Euler equations on structured grids. In M. M. Hafez, editor, *4th ISCFD Conference*, U. C. Davis, 1991.

[7] D. A. Kontinos and D. S. McRae. An explicit, rotated upwind algorithm for solution of the Euler/Navier Stokes equations. AIAA Paper No. 91–1531–CP, 1991.

[8] I. Parpia. A planar oblique wave model for the Euler equations. AIAA Paper No. 91–1545, 1991.

[9] C. Rumsey, B. van Leer and P. L. Roe. A multidimensional flux function with applications to the Euler and Navier Stokes equations. *Journal of Computational Physics*, 105:306–323, 1993.

[10] K. G. Powell, T.J. Barth and I. H. Parpia. A solution scheme for the Euler equations based on a multidimensional wave model. AIAA Paper No. 93–0065, 1993.

[11] J. Mark Janus and Animesh Chatterjee. On the use of a wake integral method for computational drag analysis. AIAA Paper No. 95–0535, to be presented 1995.

[12] H. Paillere, H. Deconinck, R. Struijs, P. L. Roe, L. M. Mesaros and J. D. Muller. Computations of inviscid compressible flows using fluctuation–splitting on triangular meshes. AIAA Paper No. 93–3301–CP, 1993.

[13] Sidilkover, D. and Roe,P.L. Unification of some advection schemes in two dimensions. NASA CR–195044, February, 1995.

[14] Bermudez, Alfredo and Vazquez, M^ Elena. Upwind methods for hyperbolic conservation laws with source terms. *Computers Fluids*, Vol. 23, No. 8, pp. 1049–1071, 1994.

[15] P. L. Roe, R. Struijs and H. Deconinck. A conservative linearization of the multidimensional Euler equations. *to appear, Journal of Computational Physics*, 1993.

[16] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43:357–372, 1981.

[17] J. M. Janus. The development of a three dimensional split flux vector euler solver with dynamic grid applications. M.S. thesis, Mississippi State University, 1984.

The Slice Compression Test: A Finite Element Analysis

Iwona Jasiuk
Professor
Department of Materials Science and Mechanics

and

Khalid Alzebdeh
Ahmed Al-Ostaz
Dang Xing

Michigan State University
East Lansing, MI 48824-1226

Final Report for:
Summer Research Extension Program
Wright Patterson Air Force Base

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base
Washington, D.C.

and

Wright Patterson AFB

December 1995

THE SLICE COMPRESSION TEST: A FINITE ELEMENT ANALYSIS

Iwona Jasiuk
Professor
Department of Materials Science and Mechanics
Michigan State University

Abstract

We conducted a finite element analysis of the slice compression test (Shafry, Brandon and Terasaki, 1989) applied to metal-matrix composites. This test, designed to characterize interfaces in composite materials, involves pressing a composite material specimen on a soft plate so that fibers are forced to protrude and intrude into the soft plate leaving a permanent deformation. We focused on the Ti-6Al-4V matrix with SCS-6 silicon carbide fibers composite system. First, we investigated numerically the effect of several parameters (various model geometries, boundary conditions and others) involved in a finite element model of the test on local stress fields. This was achieved by using the ANSYS5.1 software package. Then, we conducted a fracture analysis at an interface between the fiber and the matrix using the ABAQUS software and employed contact elements at the fiber/matrix interface. We incorporated a stress-based criterion for crack initiation and frictional slipping. The analysis was performed in three stages: i) thermal analysis to account for thermal residual stresses developed during a manufacturing process of a bulk composite from which the slice composite specimen was cut out, ii) mechanical loading in a form of a pressure applied gradually on a rigid plate placed on the top end of the specimen (full load was 600MPa), and iii) mechanical unloading from the full load state to a zero state of loading. In the outputs of this study we focused our attention on the fiber protrusion length, the permanent indent depth in the base-plate material, and the debonding length along the fiber-matrix interface. The characterization of interfacial properties relied on parametric studies. In our study, we assumed several different interfacial properties and performed the analysis to obtain the output parameters. We compared these outputs to the experimental ones obtained at the WPAFB in the Materials Division under the direction of Dr. D. Miracle.

SUMMARY

We conducted a finite element analysis of the slice compression test (Shafry, Brandon and Terasaki, 1989) applied to metal-matrix composites. This test, designed to characterize interfaces in composite materials, involves pressing a composite material specimen on a soft plate so that fibers are forced to protrude and intrude into the soft plate leaving a permanent deformation. We focused on the Ti-6Al-4V matrix with SCS-6 silicon carbide fibers composite system. First, we investigated numerically the effect of several parameters (various model geometries, boundary conditions and others) involved in a finite element model of the test on local stress fields. This was achieved by using the ANSYS5.1 software package. Then, we conducted a fracture analysis at an interface between the fiber and the matrix using the ABAQUS software and employed contact elements at the fiber/matrix interface. We incorporated a stress-based criterion for crack initiation and frictional slipping. The analysis was performed in three stages: i) thermal analysis to account for thermal residual stresses developed during a manufacturing process of a bulk composite from which the slice composite specimen was cut out, ii) mechanical loading in a form of a pressure applied gradually on a rigid plate placed on the top end of the specimen (full load was 600MPa), and iii) mechanical unloading from the full load state to a zero state of loading. In the outputs of this study we focused our attention on the fiber protrusion length, the permanent indent depth in the base-plate material, and the debonding length along the fiber-matrix interface. The characterization of interfacial properties relied on parametric studies. In our study, we assumed several different interfacial properties and performed the analysis to obtain the output parameters. We compared these outputs to the experimental ones obtained at the WPAFB in the Materials Division under the direction of Dr. D. Miracle.

1. INTRODUCTION

In composite materials a fiber-matrix *interface* plays a crucial role and influences both local stresses and effective properties of composites (see e.g. Drzal and Madhukar, 1993). For example, the stiffness and strength depend on the load transfer across the interface, the toughness is affected by the fiber pull-out or crack deflection mechanisms, and the ductility is influenced by the relaxation of high stresses near the interface (Clyne and Withers, 1993). Thus, it is not surprising that much of the recent research in mechanics of composite materials has focused on interfaces, but due to the complexity of this subject many issues remain unresolved. A fundamental problem in this area is that of characterization of interfaces. There are several tests which are being used to measure the properties of interfaces in composite materials. These include a *fragmentation test*, a *pull-out test*, a *droplet test*, a *push-out test*, a *push-in* (micro-indentation) *test*, a *transverse* test, a *slice compression test*, and other tests (Drzal and Herrera-Franco, 1991).

The Slice Compression Test (SCT) was introduced by Shafry, Brandon and Terasaki (1989) to characterize interfaces in ceramic-matrix composites. It involves a polished slice of a unidirectional composite material cut perpendicular to reinforcing fibers and compressed between two blocks (Fig. 1). The top block has a high modulus and high strength while the bottom block has a low modulus and low yield stress. Under a critical load the crack initiates at the fiber-matrix interface, and the fibers debond and then protrude making permanent imprints on the plastically deforming bottom block. Upon the release of loading the fibers retract back into the composite a small amount which depends on the coefficient of friction at the fiber-matrix inter-

face. The depths of the imprints, the protruded lengths, and the load at which the initial indents occurred are the main outputs of this test which can be easily measured. However, the challenge is how to interpret these results and thus there is a need for the mechanics solution.

The SCT has been studied theoretically and experimentally, in the context of ceramic-matrix composites, by Shafry, Brandon and Terasaki (1989), Kagawa and Honda (1991), Hsueh (1993, 1994, 1995), Lu and Mia (1994), and Hsueh, Brandon and Shafry (1996). However, theoretical analyses of this test were approximate and involved simplifying assumptions. Experiments posed numerous challenges, which included non-uniform fiber distributions, variations in fiber orientations, rotation of specimens during experiments, and possible matrix crushing during the test.

Recently, the SCT was used to characterize interfaces in metal-matrix composites. This was an experimental study, conducted at the Wright Patterson Air Force Base Laboratory (WPAFB) (Waterbury et al., 1995, 1996). It focused on two composite systems which had very different interface properties: the Ti-6Al-4V matrix and either Textron SCS-6 silicon carbide fibers or Amercom, AC3, carbon coated fibers. It involved carefully prepared model composite systems with very low volume fractions of (non-interacting) fibers, which were aligned and accurately oriented perpendicular to the polished surfaces.

In this report, we present the results of a finite element analysis of the slice compression test applied to metal-matrix composites. We used the numerical approach because of a complex geometry and a nonlinear nature of this boundary value problem. This non-linearity was due to both plasticity and contacts (friction) between surfaces. Our numerical study was conducted in parallel to the experimental effort at the WPAFB and we considered the same composite system involving Ti-6Al-4V matrix and Textron SCS-6 silicon carbide fibers. The objective of our study was to understand the influence of several parameters in the test, which included the geometry, the boundary conditions, the coefficients of friction between the composite specimen and plates, and the parameters characterizing the fiber-matrix interface (normal and shear strengths and the coefficient of friction) on the stress fields, the crack lengths and the depths of imprints. The ultimate goal was to predict the critical shear strength τ^f and the coefficient of friction μ at the matrix-inclusion interface in the above mentioned metal-matrix composites from the permanent imprint depths measured experimentally.

2. THE FINITE ELEMENT MODEL

For simplicity and to make the problem computationally tractable we modeled the geometry of an experimental set-up as an axisymmetric problem in a form of a composite cylinder, as shown in Fig. 2a, consisting of a fiber (of radius equal to unity, which corresponded to 75μm in the experiment) and the surrounding matrix in a form of the concentric cylinder representing a composite sample, and the two cylinders on top and bottom representing stiff and soft blocks, respectively. This complex composite cylinder consisted of:

1) the matrix, Ti-6Al-4V, a linear elastic-strain hardening material,
2) the fiber, SCS-6 silicon carbide, a linear elastic material,
3) the soft bottom plate, brass, a linear elastic-strain hardening material, and
4) the stiff top plate, silicon nitride, a linear elastic material (same as the fiber).

The material properties of fiber, matrix and brass are given in Tables 5-7 in the Appendix.

We conducted the numerical analysis of the slice compression test using the commercially available finite element software packages ANSYS 5.1 (1992) and ABAQUS 5.5 (1995). Initially, we used ANSYS and explored the effects of a model geometry, plastic properties of the bottom plate and of a pre-cracked fiber/matrix interface on the local stress fields. Most of these studies were based on a perfect bonding condition at all interfaces involved in the model. In ANSYS, we used six-node triangular elements with a very fine mesh at the regions of interest (see Fig. 3a). On the pre-existing crack at matrix/fiber interface, we used the surface-to-surface contact element CONTAC48 which accounted implicitly for the frictional resistance. The same contact element type was generated between the testing plates and the specimen. Because the fracture analysis in ANSYS had limited capabilities, at a later stage of this project we switched to ABAQUS to complete the fracture analysis. We chose ABAQUS program for its capabilities to handle non-linear and contact problems as well as its powerful fracture analysis features. In this analysis we used axisymmetric quadrilateral elements to model all four phases (see Fig. 3b). Due to computer time and space, we chose a rather crude mesh. We introduced contact elements (pairs) at fiber/matrix interface which allowed debonding and slip with Coulomb's friction. At both composite/bottom plate and composite/top plate interfaces we specified contact pairs which allowed slip with friction.

3. MODELING CHALLENGES

The SCT is a very complex test from both experimental and numerical perspectives. In the numerical analysis we made several simplifications in order to make the problem computationally tractable. For example we needed to decide upon the proper model which would make the problem less complicated, but still fairly representative to the actual experimental setup. In this regard, we conducted several parametric studies prior to the fracture analysis which included: (a) various model geometries (b) several types of boundary conditions, (c) parametric study of a tangential plastic modulus of base-plate, and (d) several lengths of precracked interface. We used ANSYS 5.1 software package under the following conditions: (i) only mechanical loading was applied, i.e, thermal residual stresses were excluded, (ii) perfect bonding condition at fiber/matrix interface was assumed except at a precrack's surface (if it existed), and (iii) several interface conditions were assumed between the specimen and testing plates.

3.1 Model Geometry

The geometry of the complex composite cylinder model greatly influenced the results which motivated us to decide upon proper dimensions of components (fiber, matrix, and two plates) in the test. In particular, radial dimensions influenced both interfacial normal and shear stresses since they controlled a confinement on the fiber/matrix interface. The wider was the sample the more closely we could represent the specimen in its experimental set-up but at a price of a high computer time and space. The crack length was smaller than that for narrower geometries (if the outer surface was traction-free). We considered here a number of cases:

a) A model with dimensions: L = 15, W =10, L_T = 1, L_B = 5, R =1, boundary conditions as shown in Fig. 2a, and mechanical loading of 600MPa was termed as "basic model/problem". The stress contours obtained using a nonlinear static analysis are shown in Fig. 4-a ,b ,c & d. We show that this geometry overestimates the actual interfacial stresses (see Fig. 5a & b) and fiber

protrusion length (Fig. 5c) which suggests the use of models with wider widths. Normal stress along composite/brass interface is shown in Fig. 5d which indicates that nonuniform distribution is attained.

b) A thicker cylinder model containing two fibers (Fig. 6) was considered to explore the effect of such geometry on the interfacial stresses and on the fiber protrusion length. Same static pressure of 600 MPa was applied. Results showed that fiber/matrix interfacial shear and normal stresses decreased. Normal composite/base-plate interfacial stresses and protrusion length also decreased dramatically. Although we gained a better representation, the drawback of this new model is its high computational time and space as well as in loosing a detailed information at the regions of interest (lower portion of fiber/matrix interface and a region right beneath the fiber where intrusion occurs).

c) Another approach based on one of the effective medium theories (a generalized self-consistent model) (Christensen and Lo, 1979) was investigated. In this approach a third concentric cylinder having effective elastic properties was added to the basic composite cylinder model, (see Fig. 7a). Effective elastic Young's modulus and Poisson's ratio of the additional concentric cylinder were calculated as E = 116490 MPa, $\nu = 0.299$ which showed a minute deviation from those of the matrix since fiber volume fraction was very small (1%). As an example, we show the results for a model having a concentric cylinder of radial width equal to 20. This approach was used in (Ananth and Chandra, 1995) and is advantageous over the approach in (b) in the sense that it eliminates the problem of stress concentration at some regions of less interest. In connection to this study, we introduced more confinement to the bottom plate by extending its width beyond the composite specimen (see Fig. 7b). This led to a stress concentration at the right lower specimen corner.

d) Since properties of the effective concentric cylinder were so close to those of matrix, then just varying the width of matrix from 10 (as in the basic model) to 20, 30, 40, or 60 while keeping a unit fiber radius was another option to be considered. After each run, we recorded: fiber/matrix interfacial normal and shear stresses, fiber and matrix normal stresses at base-plate/composite interface, and a total fiber protrusion length, as given in Table 1.

Table 1: Interfacial stresses and fiber protrusion length

W (Width)	σ_{xx} (MPa)	σ_{xy} (MPa)	σ_{yy} (f) center/ average	σ_{yy} (m) average	Protrusion/ R
10	683	424	-728/ -750	-586	.078
20	575	329	-656/ -670	-534	.0030
30	482	329	-679/ -690	-550	.0023
40	411	326	-677/ -690	-560	.0019

Table 1: Interfacial stresses and fiber protrusion length

W (Width)	σ_{xx} (MPa)	σ_{xy} (MPa)	σ_{yy} (f) center/ average	σ_{yy} (m) average	Protrusion/ R
60	366	318	-668/ -685	-556	.0020

Investigating results in the above table, we observe that:
1) The case W = 10 (W is a radial width) yields higher stresses and protrusion length compared to the rest of cases. Therefore, the traction free boundary condition applied in the basic model highly overestimates the actual stresses which indicates that the ratio W/R is (is a fiber radius) an important parameter in the analysis.

2) σ_{xx} and σ_{xy} decrease when W increases due to the additional confinement as expected.
3) When $W \geq 20$ σ_{xy} stabilizes i.e. converges to an almost constant value which justifies the choice of radial width W = 20.

4) The fiber protrusion length decreases as W increases. One should observe that as W changes from W = 10 to W = 20 protrusion length decreases about 3 times. The difference tends to be smaller when we go for higher widths.

Based on these observations, we concluded that a model in which W = 20, or W = 30 was a sufficiently representative in the numerical analysis. Nevertheless, higher widths led to a smaller discrepancy in outputs especially in σ_{xx} but at the cost of a higher computer time and space. Additional justification for the use of 20 width is that the fracture mode II is the dominant one in this problem (depends on τ_{xy}).

3.2 Boundary Conditions on the Outer Surface

Boundary conditions at the composite's outer surface influenced the output significantly. While in the basic model we had a traction-free boundary condition, another possible approach was to use a displacement boundary condition in such a way that it accounted for the effect coming from the unmodelled portion of specimen. However, this was a non-trivial task, since it required a solution of a corresponding boundary value problem with larger dimensions which would serve as an input to the basic model. Due to the presence of the easily deforming base material this displacement was non-uniform. Also, we needed to repeat this process for each load step. In particular, we implemented this process as follows. Based on the wider composite cylinder model, we solved for the displacement fields (u_x and u_y) at the section which corresponded to the outer cylinder's surface in the basic model. Then, considering the narrower geometry (basic model), we imposed obtained displacements on its outer surface and performed the new analysis. This procedure was performed under the mechanical loading step (p = 600 MPa) assuming perfect boundary conditions at all interfaces. As an example, displacements (u_x and u_y) as shown in Fig. 8a & b, were obtained from the model of width equal to 30 (see section 3.1).

Fiber/matrix interfacial stresses in this model are shown in Fig. 9a and b, while the fiber protrusion length is shown in Fig. 9c. When we compared these results to those obtained from the basic model when the displacement fields were imposed (in an avareged sense) on its outer surface (see Figs. 10 a, b, c & d), we observed that the two sets were comparable. Although this procedure helped in reducing the size of model, still it required a considerable effort. For this reason we did not consider this procedure in remaining analyses.

3.3 Plasticity of the Base-Plate

For simplicity, when analytical solutions of the slice compression test were developed, some researchers (Hsueh, 1993, Lu and Mia, 1994) assumed that the soft base-plate produces a uniform traction condition on the specimen's lower edge. Generally, our finite element results did not support this observation. In the contrary, average fiber normal stress was found to be higher in the fiber than that in the matrix. To investigate the effect of plastic properties of the base-plate on the interfacial stresses as well as on the fiber protrusion length, we performed analyses in which we varied the tangential modulus of the base-plate according to the schematic graph shown in Fig. 11. The summary of output results are provided in Table 2.

Table 2: Average normal stress σ_{yy} (in MPa) at composite/ base-plate interface and fiber protrusion length

E^t (MPa)	σ_{yy} (matrix) (MPa)	σ_{yy} (fiber) (MPa)	Protrusion/ R
4000	-599.	-764.	0.062
2000	-595.	-729.	0.10
1000	-598.	-701.	0.14
100	-597.	-600.	0.30

From Table 2, we observe that the protrusion length increases when the tangential modulus of the base increases. This may be due to the fact that the higher material flow permits higher fiber protrusion into the base-plate. It is only at the elastic-perfectly plastic limit (approximated by $E^t = 100$), that a uniform traction in an averaging sense on the composite's lower edge is attained. The case of $E^t = 0$ (i.e. exact elastic-plastic case) showed convergence problems. We conclude here that the choice of plastic properties of the base-plate does significantly affect test results.

3.4 Precracked Fiber/Matrix Interface

This part of our study was conducted using the ANSYS software under the mechanical loading step only. The objective was to introduce a pre-existing crack along the fiber/matrix interface assuming that it was developed in the thermal loading step analysis (see Fig. 12a, 13a). Inclusion of the thermal residual stresses into the model was a challenge and could not achieved by ANSYS as a separate load step. We used the contact elements (CONCT48) at the crack surfaces

along fibre/matrix interface, the composite/base-plate interface, and the composite/rigid plate interface. Two crack lengths were introduced in two distinct cases: D or 2D (D is the fiber diameter) along the matrix/fiber interface near the soft plate. The contact elements made the problem computationally nontrivial. The contact surface parameters needed to be chosen with care in order to preclude two problems: divergent solution and fiber over-penetration into the base-plate.

Results of normal and shear interfacial stresses as well as the fiber protrusion in the case of the crack length D are shown in Figs 12 b, c & d, while for the crack length 2D are shown in Figs. 13b, c & d. We observe how the crack length affects the fiber protrusion length. As the crack length increases the fiber protrudes more into the base-plate as one may expect.

3.5 Thermal Loading Analysis

The challenge here was how to model residual thermal stresses developed during a manufacturing process of a bulk composite from which the composite slice was cut out and tested. Also, in our finite element model of SCT, the thermal analysis was applicable to the composite cylinder specimen only, i.e. the two testing plates needed to be excluded. Thermal residual stresses developed in the composite during the manufacturing process when the composite was cooled down from a reference temperature of $T_{ref} \cong 625^{o}C$ to the room temperature ~25°C. These stresses were mainly due to a mismatch in coefficients of thermal expansion (CTE) of matrix and fiber. However, since we considered just a slice of a composite, then these residual stresses could be affected by a cutting process. To account for this cutting effect, we implemented the thermal analysis in two steps as was done by Anath and Chandra (1995) also in the context of titanium-matrix composites. These two steps may be summarized as follows (see Fig. 14):

Step 1

To model the actual thermal residual stresses in the bulk composite, we applied at both ends of our cylinder model a unidirectional constant displacement on the upper edge (u_z = constant) a long with a temperature change of $\Delta T = -600^{\circ}C$. The constant displacement boundary condition was imposed so that it produced a generalized plane strain state (i.e. a very long cylinder with its both ends free to deform). In the generalized plane strain case, ε_z is constant and for thermal loading it produces a zero resultant force at both ends of the cylinder. For this problem, we determined this constraint by an analytical elasticity solution which involved solving the following equations simultaneously in order to compute the unknown constants C_1, C_3 and C_4 (e.g., Timoshenko and Goodier, 1980).

$$\frac{-\alpha^f E^f}{(1-\nu^f)}\frac{\Delta T}{2} + \frac{E^f}{(1+\nu^f)(1-2\nu^f)}C_1 - \frac{E^m}{(1+\nu^m)(1-2\nu^m)}C_3 - \frac{E^m}{(1+\nu^m)a^2}C_4 = 0 \qquad (1)$$

$$-\frac{\alpha^m \Delta T}{2(1-\nu^m)}\left(1-\frac{a^2}{b^2}\right) + \frac{C_3}{(1+\nu^f)(1-2\nu^f)} - C_3 a - \frac{C_4}{a} = 0 \qquad (2)$$

$$\frac{-\alpha^f E^f}{(1-\nu^f)}\frac{\Delta T}{2}+\frac{E^f}{(1+\nu^f)(1-2\nu^f)}\frac{C_1}{}-\frac{C_4}{(1+\nu^m)b^2}=0 \qquad (3)$$

In above equations, α is the thermal coefficient of expansion, E is the Young's modulus of elasticity, ν is the Poisson's ratio. The superscripts *f, m* stand for fiber and matrix, respectively and *a, b* are the fiber and matrix radii, respectively.

Upon solution of above equations, the constants were found as C_1= -0.00152, C_3 = -0.00209, C_4= -0.00143. Then, we substituted them back into the σ_{rr} and $\sigma_{\theta\theta}$ field expressions of fiber and matrix (not included here). For the plane strain case, the unidirectional (σ_z) stress is related to the in-plane stresses according to the formula

$$\sigma_z^i = \nu^i(\sigma_r^i+\sigma_\theta^i)-E^i\alpha^i\Delta T+E^i\varepsilon_z \qquad (4)$$

where superscript *i* denotes fiber (*f*) or matrix (*m*).

In the generalized plane strain at the constrained edge of the cylinder equilibrium must be satisfied

$$\int_{A_f}\sigma_z^f dA_f+\int_{A_m}\sigma_z^m dA_m=0 \qquad (5)$$

From Eqn. (5) we found that $\varepsilon_z = -0.00556$, $u_z = \varepsilon_z L$ (see Fig. 15a)

Interfacial stresses σ_{xx} and τ_{xy} obtained are shown in Figs. 15b & c. We observe from Fig. 15c, that this problem yields zero shear stresses everywhere in the model, as expected from the elasticity solution. On the other hand, normal stress along the top edge of specimen consists of two constant blocks in the fiber and matrix, respectively, for which equation (5) holds (see Fig. 15b).

Step2

In this step, we removed the constant displacement-constraint condition on the specimen's edge and re-equilibrated the solution of step (1). Thus, we represented the composite slice of length L cut out from the bulk composite (of step 1) with modified thermal stresses accordingly. As a result, normal and shear stresses along the fiber/matrix interface developed as shown in Figs. 14a and b).

Examining the results of step 2 of the above thermal analysis we observe that stresses are

identical to those of a thermal analysis under $\Delta T = -600^{o}C$ load and boundary conditions as shown in Fig. 2b. Contour plots of σ_{xx} and τ_{xy} in the latter problem are shown in Figs. 16a & b, while normal and shear stresses along interface are shown in Fig. 16c & d. This is may be due to the fact that under this thermal load of $\Delta T = -600^{o}C$ stresses are still within their elastic limit, i.e. no plasticity occurs, except for singular points (A, B) at the traction-free surfaces (end points at the fiber-matrix interface). Therefore, we conclude that there is no need to account for the effect of cutting at $\Delta T = -600C^{\circ}$. Nevertheless, for higher temperatures one may need to account for this cutting effect. For example, Anath and Chandra (1995) showed that there was a need for such analysis at $\Delta T = -900°C$.

4. FRACTURE ANALYSIS

Direct (automatic) fracture analysis was implemented using ABAQUS software and contact elements were generated at the interfaces. At the fiber-matrix interface we assumed that the crack initiation and propagation would occur under a combined action of normal tensile and shear stresses in a form of fracture modes I and II, respectively. In the analysis we used a quadratic stress based failure criterion (Tsai, 1965; Hashin, 1980) which accounts for both modes

$$\left(\frac{\sigma_r}{\sigma^f}\right)^2 + \left(\frac{\tau}{\tau^f}\right)^2 \geq 1 \tag{6}$$

where σ^f is the tensile strength of the interface resisting crack opening, and τ^f is the failure shear stress, while σ_r is normal tensile stress, and τ is shear stress across interface. After debonding, interfacial frictional stress develops according to Coulomb's law, thus, τ^f in (6) is modified according to

$$\tau^f = \tau^{cr} + \mu p \qquad \text{where} \quad p = -\sigma_r \ \text{if} \ \sigma_r < 0, \quad p = 0 \ \text{if} \ \sigma_r > 0 \tag{7}$$

where τ^{cr} is a shear strength of the interface which we are after, p is a compressive contact pressure at the interface, and μ is a coefficient of friction.

In the analysis we assumed given strengths of the interface σ^f, τ^f and the coefficient of friction μ (which enters after the crack initiation) as well as the locations of the possible initial crack tips, which we specified to be at the bottom and top points (lines) at the fiber-matrix interface and denoted by A and B in Fig. 2b. Then, we loaded the composite in small increments up to a maximum load. At each increment the program solved for the interfacial stresses and checked if the stresses at the specified crack tip locations satisfied the fracture criterion (6). If they did, the crack initiated, i.e. debonding (slip) took place. At this stage time steps were specified to be very small (we took each step as 0.001 in the 0 to 1 time scale). This check was repeated at each next such small load step, as long as the failure criterion (6) was met, thus representing the crack propagation upon the increasing load. The crack propagation stopped if the stresses at the crack tip did not satisfy (6) anymore. The load steps were gradually increased until the stresses satis-

fied again the fracture criterion (6) and the crack resumed propagating.

We simulated the whole test in three load steps performed in a sequence in such a way that the output of a previous step was incorporated as the input in a following step. These steps included:

(1) the thermal loading which involved cooling down from a reference temperature of $625^{o}C$, at which the composite was assumed to be stress-free, to a room temperature of $25^{o}C$, which gave $\Delta T = -600^{o}C$. Note that the actual processing temperature in composites is higher, over $800^{o}C$, but for the small model composite samples which were cooled very slowly the lower ΔT is more representative (Waterbury, 1994).

(2) the mechanical loading involving an elasto-plastic static analysis in which a uniform pressure was applied on the top stiff plate and increased gradually by small increments from 0 to 600 MPa.

(3) the mechanical unloading in which the load was decreased gradually by small increments from the maximum load reached to a state of zero loading.

In the next sections we discuss the three stages of loading and some selected results.

4.1 Thermal Fracture Analysis (Step 1)

In the thermal stress analysis in addition to solving for stress fields we were also interested in predicting the crack length at the fiber-matrix interface caused by the thermal loading. In order to understand the influence of interfacial parameters σ^{f}, τ^{f} and μ on the crack length in the composite due to thermal loading we carried out a parametric investigation. Results are summarized in Table 3.

Table 3: Crack length as a function of interfacial properties

μ	σ^{f} (MPa)	τ^{f} (MPa)	$\frac{l}{R}$
1.0	90000	100	2.0
0.75	1	100	2.2
0.75	300	100	2.2
0.75	400	150	1.2
0.75	200	150	1.2
0.75	1	200	0.2

Table 3: Crack length as a function of interfacial properties

μ	σ^f (MPa)	τ^f (MPa)	$\frac{l}{R}$
0.50	1	100	2.8
0.50	200	100	2.8
0.25	90000	100	4.0
0.25	200	50	> 4.0

where l is the crack length and R is the fiber radius.

Upon examining the results in Table 3, we observe the following:

i) The interfacial normal strength σ^f does not contribute to the crack propagation caused by thermal loading, which indicates that only fracture mode II occurs. This is due to the fact that the radial stress is in compression except for a very small zone at the fiber-matrix interface (at the free surface) as observed from our finite element results and analytical elastic solutions (e.g. Kurtz and Pagano, 1991).

ii) The increase in the coefficient of friction μ causes the crack length to decrease. For example, as μ is increased from 0.25 to 1.0 the crack length decreases by a factor of 2. Since the larger frictional resistance builds up on the crack surfaces this reduces the tendency of further cracking.

iii) The increase in the interfacial shear strength τ^f results in shorter crack lengths. It is interesting to note that an increase of τ^f by a factor of 2 leads to a decrease in the crack length by a factor of 11.

Thus, we concluded that the shear strength τ^f is the key parameter which controls the crack initiation and its final length, the coefficient of friction μ also plays an important role, but the normal strength σ^f does not influence the results.

4.2 Mechanical Loading/Unloading Analysis (Steps 2 & 3)

After the calculation of thermal residual stresses we applied the mechanical pressure of 600MPa on the top stiff plate so that it generated a uniform strain on specimen's upper edge. As indicated earlier, we increased the load gradually in small increments to capture the crack growth. Then, we unloaded the specimen gradually by very small increments to a zero loading state, and took a record of the fiber protrusion length, crack lengths along the matrix-fiber interface and the depth of permanent imprints in the brass-plate. In order to characterize the interfacial properties, we varied the inputs for the interface characteristics (τ^f, σ^f, and μ) and compared the output quantities to its experimental counterparts. We estimated the interfacial properties as those inputs

which their output quantities matched the experimental ones with a certain accuracy. This process was run using the following geometric model:

* Dimensions:

$L = 15$, $W = 10$, $L_T = 1$, $L_B = 5$.

(where L, W are the axial and radial dimensions of the composite cylinder, respectively, and L_T, L_B are the thicknesses of the top and bottom plates, respectively)

* Loads and boundary conditions:

Thermal load of $\Delta T = -600^{o}C$ and pressure of 600 MPa and a traction-free boundary condition at cylinder's outer edge.

After the full mechanical loading was completed, the maximum fiber protrusion (equal to the indentation depth) was reached while the permanent indent depth was recorded after the unloading step (step 3). We observed that the fiber retracted back by a small amount due to the action of frictional forces on the crack surface. In addition, we reported the crack length at top and bottom regions of matrix/fiber interfaces after steps 1 & 3 although we believe that the occurrence of top crack is just an artifact which is influenced by the boundary condition specified on the top-plate/ composite interface. We found that a perfectly bonded interface would eliminate such a crack. A typical crack growth history is depicted in Fig. 17 in which the load step from 0 to 1 corresponds to a thermal load of ΔT from 0 to $-600^{o}C$, from 1 to 2 to a pressure which increases from 0 to 600 MPa, and from 2 to 3 to the release of pressure to zero. The bottom curve, denoted by 1, corresponds to the upper end, while the top curve corresponds to the lower end. It is interesting to note that at some load intervals crack length is constant although the load is increasing. This is due to changes in interfacial stresses, which vary along the fiber length, and are affected by loading and frictional resistance.

Table 4: Protrusion and crack lengths as a function of interfacial properties

μ	σ^f (MPa)	τ^f (MPa)	Protrusion/ R	$\frac{l}{R}$ (step 1) bottom/top	$\frac{l}{R}$ (Step 3) bottom/top	$\frac{l}{R}$ (Step 3) total
0.75	**200**	150	0.0432	1.2/1.0	7.121/1.6	8.721
		200	0.0301	0.1/0.2	4.152/3.58	7.732
		300	0.0185	0/0	2.165/0	2.165
		400	**0.0160**	**0/0**	**1.769/0**	**1.769**
		450	0.0152	0/0	1.671/0	1.671
		600	0.0150	0/0	1.576/0	1.576
	400	**400**	**0.015**	**0/0**	**1.769/0**	**1.769**

Table 4: Protrusion and crack lengths as a function of interfacial properties

μ	σ^f (MPa)	τ^f (MPa)	Protrusion/ R	$\frac{l}{R}$ (step 1) bottom/top	$\frac{l}{R}$ (Step 3) bottom/top	$\frac{l}{R}$ (Step 3) total
0.50	200	200	0.0285	0.1/0.2	4.745/2.80	7.545
		400	0.0169	0/0	1.769/0	1.769
		600	0.0150	0/0	1.576/0	1.576
0.25	**200**	**200**	**0.0262**	**0.1/0.2**	**5.141/2.4**	**7.541**
		400	0.0172	0/0	1.769/0	1.769
		450	0.0154	0/0	1.671/0	1.671
		600	0.0150	0/0	1.572/0	1.572
	400	**200**	**.0301**	**0.1/0.2**	**5.141/2.4**	**7.541**

A summary for results of this parametric study is outlined in Table 4. Several observations are made here:

(1) As one may expect, the fiber protrusion (and crack) lengths decrease as we increase τ^f. The crack initiation and propagation are mainly controlled by the shear interfacial strength and thus mode II failure is dominant in the test.

(2) When we increase the friction coefficient μ, the fiber protrusion and crack lengths decrease. However, the difference between results of $\mu = 0.25$ and $\mu = 0.75$ is not large which suggests that also this test is less sensitive to the coefficient of friction parameter.

(3) Outputs are clearly insensitive to the normal strength σ^f of interface which indicates that this test cannot estimate the interfacial normal strength.

(4) When we compare these results to those obtained experimentally at the WPAFB (Waterbury et. al. 1995, 1996) we find similar trends but the experimental results lie a little lower. For example, at the specimen thickness of 1.125 mm (equivalent to L = 15 in our model), the experimental fiber protrusion length $l = 0.0104$ mm. Our numerical results higher by about a factor of 2. This is due to the narrower width of specimen which is considered in our calculations. Thus, a wider concentric composite cylinder should be used.

(5) We considered another geometry in which specimen had extended dimensions L = 30, W = 20, $L_T = 2$, $L_B = 10$. The final crack length was 11.4R and both maximum and permanent indent

depths were 0.058R and 0.054R, respectively. By a comparison with the corresponding results in Table 4, we observed a big effect of the model geometry on the output (a parametric geometry study suggested to use W = 20, but due to the highly needed computer memory it was discarded). For example, an increase in length L leads to the larger crack and protrusion lengths (and intrusion depths) (Waterbury et al., 1995, 1996), while an increase in radial width W decreases protrusion length as it adds more confinement at the interface. Thus, these two geometric parameters have a reverse effect.

Other quantities of interest are stresses along the fiber-matrix and brass/composite interfaces. Stresses σ_{xx} (normal) and σ_{xy} (shear) are causes of debonding along fiber/matrix interface. Von Mises stresses along fiber/matrix interface are used to measure plasticity at the interface while σ_{yy}, σ_{xy} along the brass/composite interface affect the fiber protrusion length. Along the brass/composite interface there is a non-uniform normal stress σ_{yy} with a maximum occurring at the fiber center. If the distribution of this stress is averaged over the matrix and fiber regions, respectively, the two averages are not equal. This normal stress plays a major role in introducing additional shear stresses at the fiber-matrix interface (as discussed in section 3.5) which contribute to further debonding across the fiber-matrix interface and fiber intrusion into the base plate.

Finally, this analysis was challenging due to the complex nature and many parameters involved in the test. However, a considerable progress have been achieved. Nevertheless, to match the numerical simulation results to those obtained experimentally (thus a better characterization of interface will be possible) a wider composite model should be used. Similar fracture parametric study based on such new geometry should be done. The success of this analysis depends on the available computer capabilities. This work is under current investigation.

5. CONCLUSIONS

In this project we analyzed the slice compression test using a finite element method. We used ANSYS5.1 to implement parametric studies to investigate the influence of various parameters in the model on the local stress fields and ABAQUS at later stage to carry out the fracture analysis in three load steps: (i) thermal loading of $\Delta T = -600^{o}C$ to incorporate thermal residual stresses in the bulk composite cylinder, (ii) mechanical loading of 600 MPa gradually applied on top rigid plate, and (iii) mechanical gradual unloading to a zero load state. We employed contact elements at the fiber-matrix interface, which incorporated stress-based criterion for crack initiation and allowed frictional slip at contact surfaces after debonding occurred. Both interfaces between the specimen and composite cylinder were completely debonded. This problem was highly challenging due to the non-linearity in material properties and model geometry involved in the analysis. Furthermore, the existence of the two testing plates between which the composite specimen was compressed as well as the nature of load steps make the analysis further complicated from a numerical perspective. Our numerical results showed a reasonable success toward the global goal which was the characterization of interfacial properties in metal matrix composites. For a chosen thickness of composite cylinder (W = 10), a slightly higher shear strength was obtained. If the thickness were increased, the experimental fiber protrusion and interfacial crack lengths would most likely be reproduced, and thus, a numerical estimation would be possible. This will be the subject of our calculations in near future. Our results show that the slice compression test is capa-

ble of characterizing the shear strength and the coefficient of friction of fiber/matrix interface.

ACKNOWLEDGEMENTS

We would like to thank Dr. M. D. Waterbury and Prof. A. Gawecki for helpful discussions on the slice compression test. Additionally, I. J. would like to thank Drs. D. Miracle and M. D.Waterbury for suggesting this problem and for numerous discussions on the slice compression test during her 1994 summer visit at the Materials Division at the WPAFB in Dayton. This work was supported by the AFOSR Summer Research Program follow-up research grant.

REFERENCES

ABAQUS. Version 5.5. 1995. Hibbit, Karlsson & Sorensen, Inc., USA.

Anath, C. R. and N. Chandra. 1995. "Numerical Modeling of Fiber Push-out Test in Metallic and Intermetallic Matrix Composites - Mechanics of the Failure Process," *Journal of Composite Materials,* 29 (11):1488-1514.

Clyne, T. W. and P. J. Withers. 1993. *An Introduction to Metal Matrix Composites,* Cambridge University Press, Cambridge.

Christensen, R.M. and Lo, K.H., 1979 "Solutions for Effective Shear properties in Three Phase Sphere and Cylinder Models," Journal of the Mechanics and Physics of Solids, 27:315-330.

Drzal, L. T. and P. J. Herrera-Franco. 1991. "Composite Fiber-Matrix Bond Tests," *Engineered Materials Handbook, Adhesives and Sealants*, ASM International, 3:391-405.

Drzal, L. T. and M. Madhukar. 1993. "Fibre-matrix Adhesion and its Relationship to Composite Mechanical Properties," *Journal of Materials Science,* 28:569-610.

Hashin Z., 1980. "Failure Criteria for Unidirectional Fiber Composites," *Journal of Applied Mechanics*, 47:329-334.

Hsueh, C. H. 1993. "Analyses of Slice Compression Tests for Aligned Ceramic Matrix Composites," *Acta Metallurgica et Materialia,* 41:3585-3593.

Hsueh, C. H. 1994. "Slice Compression Tests Versus Fiber Push-in Tests," *Journal of Composite Materials,* 28:638-655.

Hsueh, C. H. 1995. "Analyses of Slice Compression Tests for Aligned Ceramic Matrix Composites - II. Type II Boundary Condition," *Acta Metallurgica et Materialia,* 43(4):1407-1413.

Hsueh, C. H., D.G. Brandon and N. Shafry. 1996. "Experimental and Theoretical Aspects of Slice Compression Tests," *Materials Science and Engineering,* preprint.

Kagawa, Y. and K. Honda. 1991. "A Protrusion Method for Measuring Fiber/Matrix Sliding Fric-

tional Stresses in Ceramic Matrix Composites," *Ceram. Eng. Sci. Proc.* 12:1127-1138.

Kurtz, R.D. and N.J. Pagano. 1991. "Analysis of the Deformation of a Symmetrically-loaded Fiber Embedded in a Matrix Material," *Composites Engineering*, 1:13-27.

Lu, G.Y. and Y.W. Mai. 1994. "A Theoretical Model for the Evaluation of Interfacial Properties of Fibre-Reinforced Ceramics with the Slice Compression Test," *Composites Science and Technology* 51:565-574.

Shafry, N., D. G. Brandon and M. Terasaki. 1989. "Interfacial Friction and Debond Strength of Aligned Ceramic Matrix Composites," in *Euro-Ceramics, Engineering Ceramics,* ed. G. de With, R.A. Terpstra, and R. Metselaar, 3:453-457, Applied Science Publishers, UK.

Timoshenko, S.P. and Goodier, J.N. 1987. *Theory of Elasticity*, 3rd ed, McGraw Hill, New York.

Tsai, S. W., 1965. "Strength Characteristics of Composite Materials," *NASA CR*-224.

Waterbury, M. D. 1994. private communication.

Waterbury, M., D. Tilly, W. Kralik and D. Miracle. 1995. "Evaluation of TMC Interface Properties by the Slice Compression Test," *Proceedings of ICCM-10*, Whistler, B.C. Canada, August 1995, VI:719-726.

Waterbury, M., D. Tilly, W. Kralik and D. Miracle, 1996. "Slice Compression Testing of Titanium Matrix Composites," *Acta Materialia*, submitted.

FIGURES

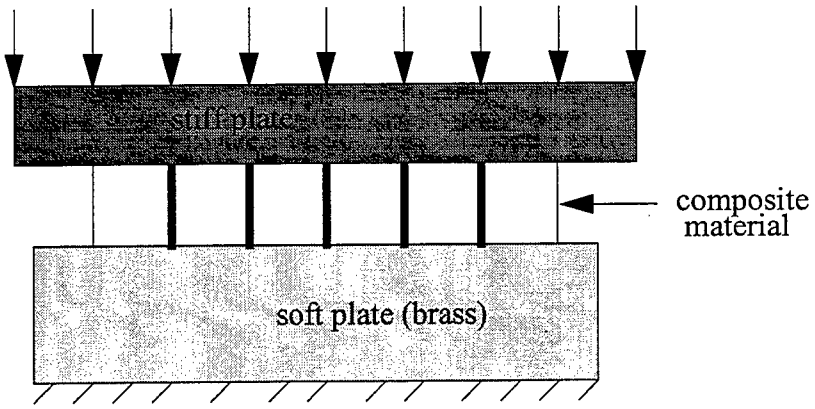


Fig. 1 A sketch of the experimental set-up for the slice compression test.

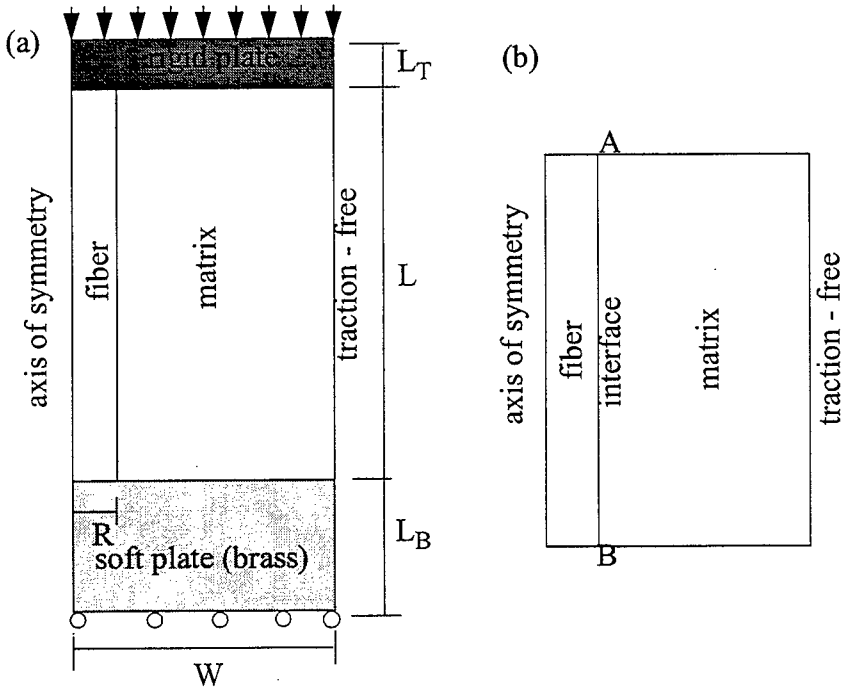
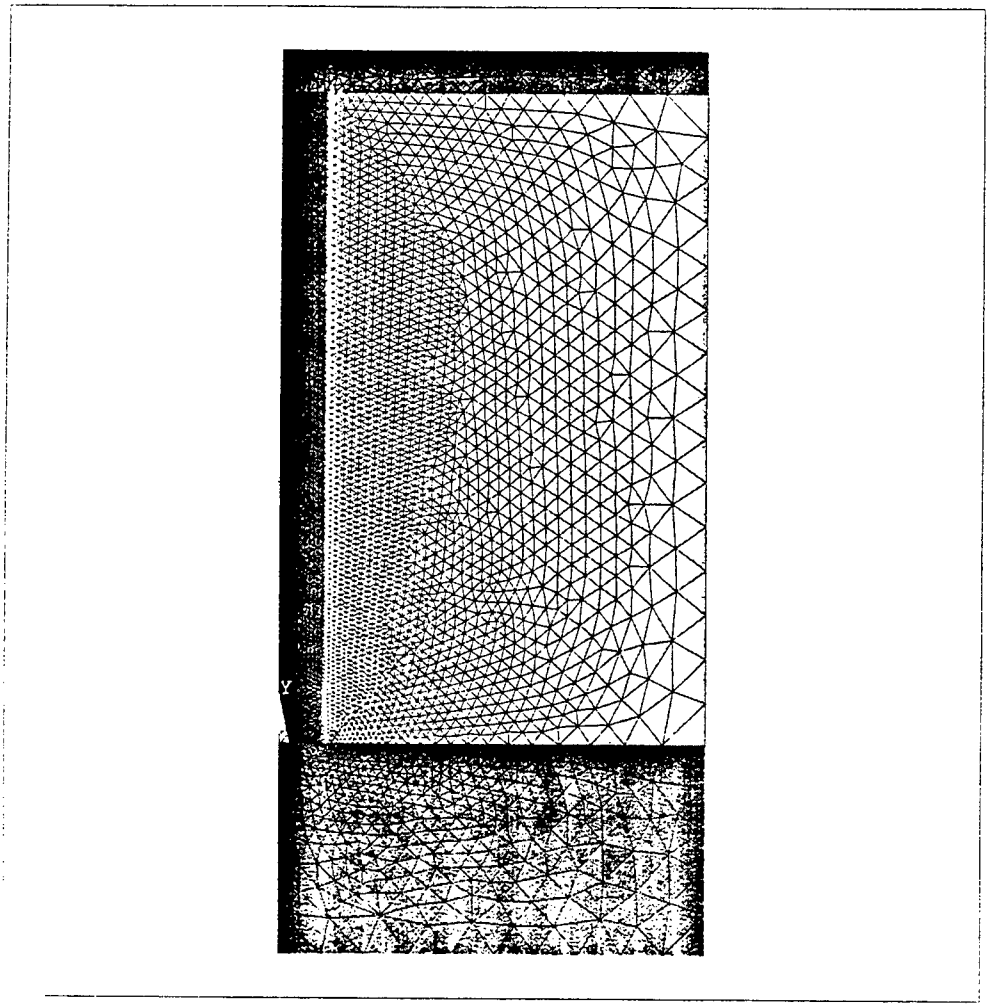
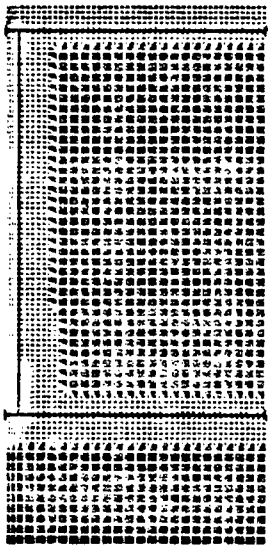


Fig. 2 A finite element model of the slice compression test for (a) thermal/mechanical, and (b) thermal analyses.



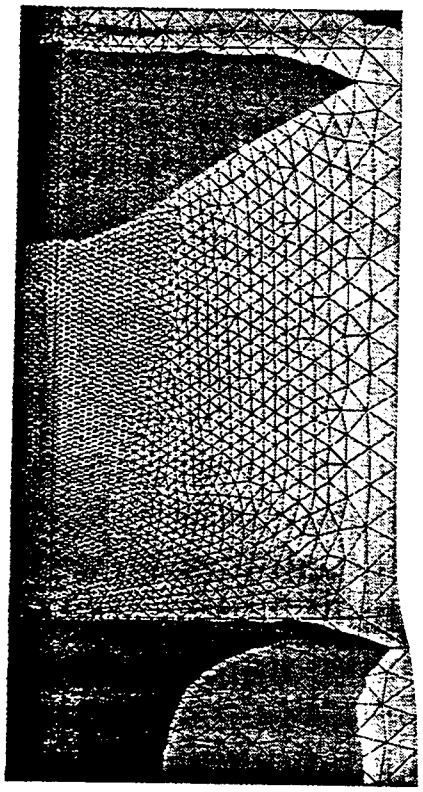
(a)



(b)

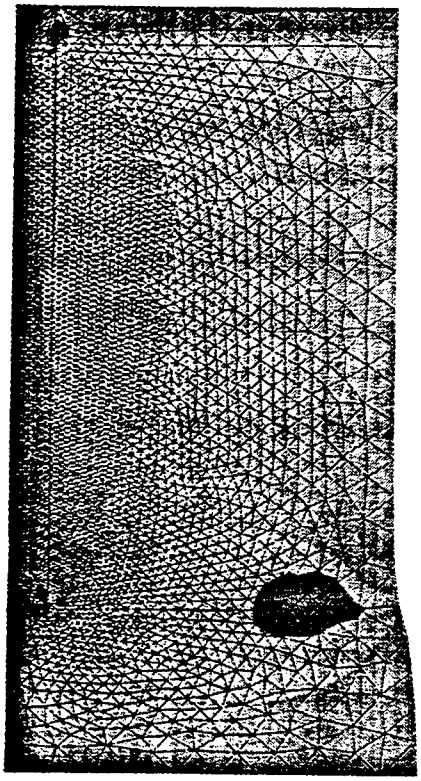
Fig. 3 The finite element mesh used in (a) ANSYS, and (b) ABAQUS programs.

(a)



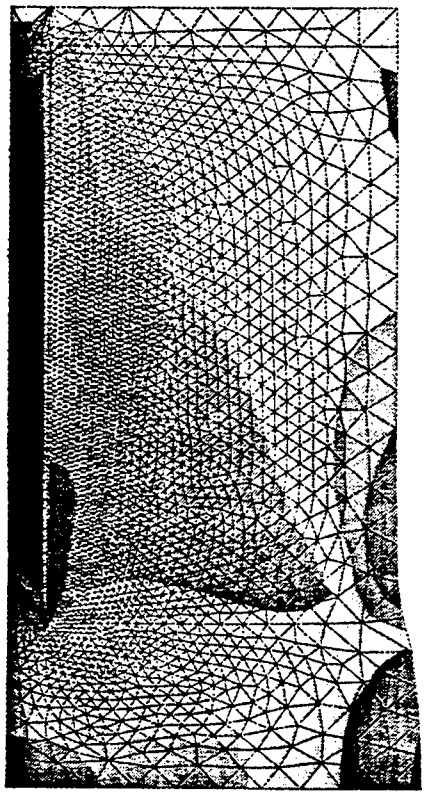
ANSYS 5.1
OCT 19 1995
20:48:13
PLOT NO. 13
NODAL SOLUTION
STEP=1
SUB =6
TIME=1.2
SX (AVG)
RSYS=0
DMX =0.27099
SMN =-0.418E+09
SMX =0.131E+10
-0.418E+09
-0.227E+09
-0.350E+08
0.157E+09
0.348E+09
0.540E+09
0.731E+09
0.923E+09
0.111E+10
0.131E+10

(b)



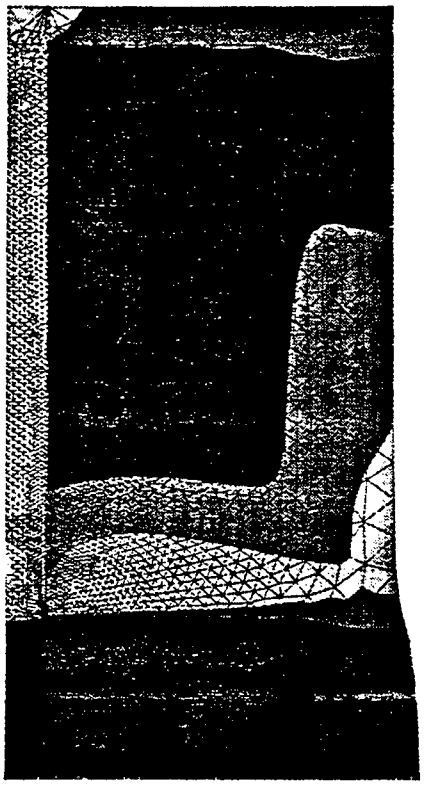
ANSYS 5.1
OCT 19 1995
20:48:48
PLOT NO. 15
NODAL SOLUTION
STEP=1
SUB =6
TIME=1.2
SXY (AVG)
RSYS=0
DMX =0.27099
SMN =-0.367E+09
SMX =0.448E+09
-0.367E+09
-0.276E+09
-0.186E+09
-0.951E+08
-0.457E+07
0.860E+08
0.177E+09
0.267E+09
0.358E+09
0.448E+09

(c)



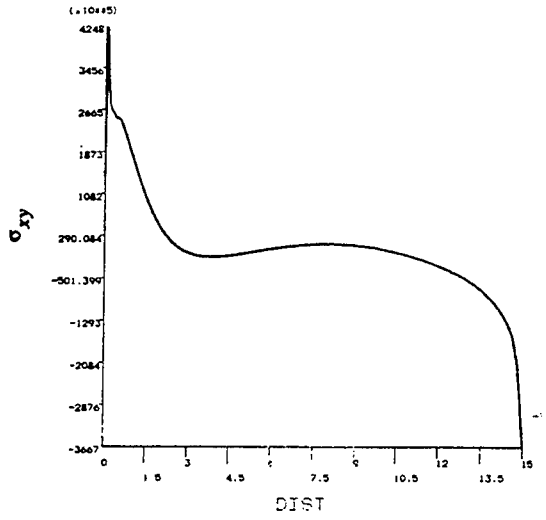
ANSYS 5.1
OCT 19 1995
20:48:31
PLOT NO. 14
NODAL SOLUTION
STEP=1
SUB =6
TIME=1.2
SY (AVG)
RSYS=0
DMX =0.27099
SMN =-0.161E+10
SMX =-0.223E+09
-0.161E+10
-0.146E+10
-0.130E+10
-0.115E+10
-0.994E+09
-0.840E+09
-0.686E+09
-0.532E+09
-0.378E+09
-0.223E+09

(d)

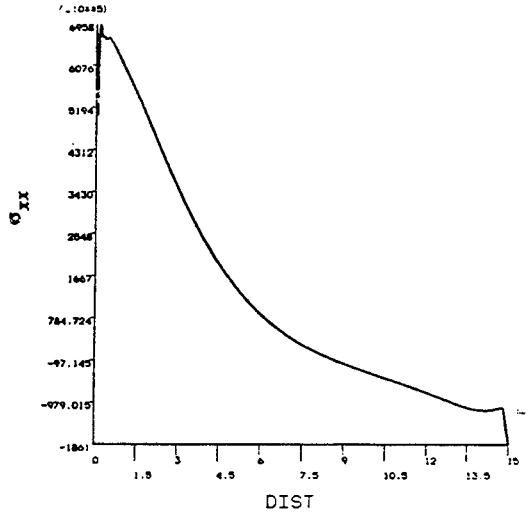


ANSYS 5.1
OCT 19 1995
20:49:09
PLOT NO. 16
NODAL SOLUTION
STEP=1
SUB =6
TIME=1.2
SEQV (AVG)
DMX =0.27099
SMN 0.384E+09
SMX 0.237E+10
0.384E+09
0.605E+09
0.825E+09
0.105E+10
0.127E+10
0.149E+10
0.171E+10
0.193E+10
0.215E+10
0.237E+10

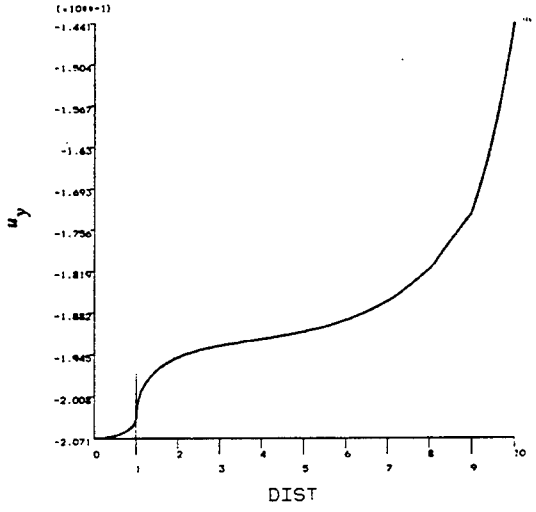
Fig. 4 Contour plots for (a) σ_{xx}, (b) σ_{xy}, (c) σ_{yy}, and (d) σ_{eff} (von Mises) for the basic model.



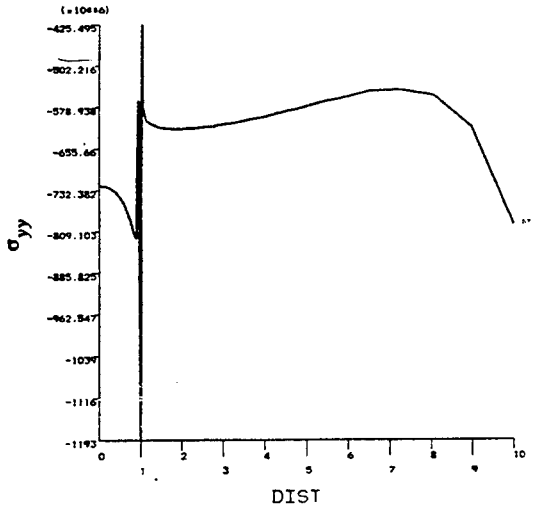
(a)



(b)



(c)



(d)

Fig. 5 (a) Normal stress, (b) shear stress long fiber/matrix interface; (c) normal deflection and (d) normal stress along the composite/brass interface in the basic model.

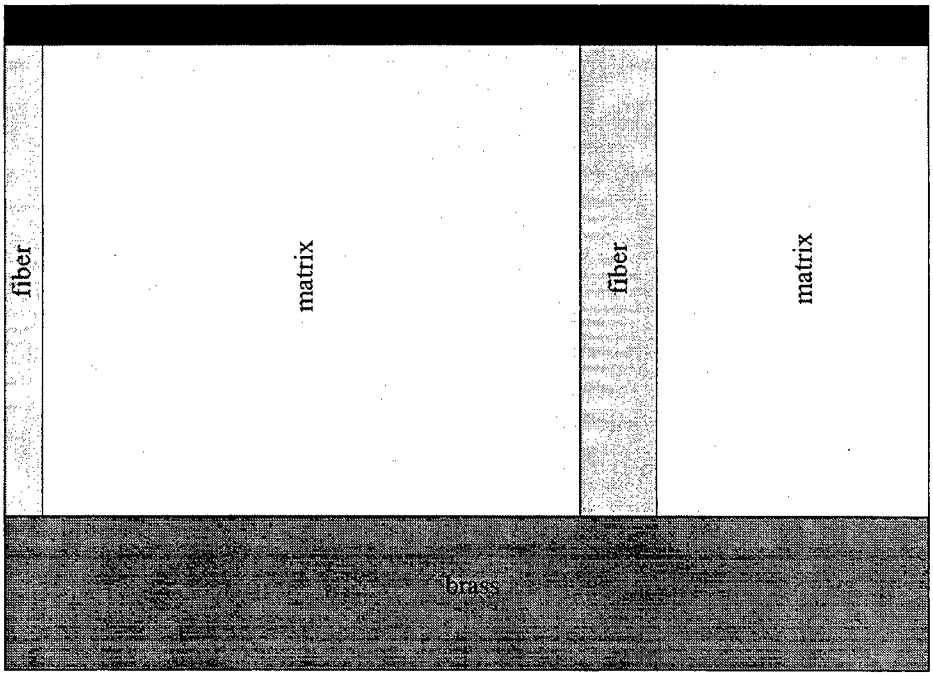
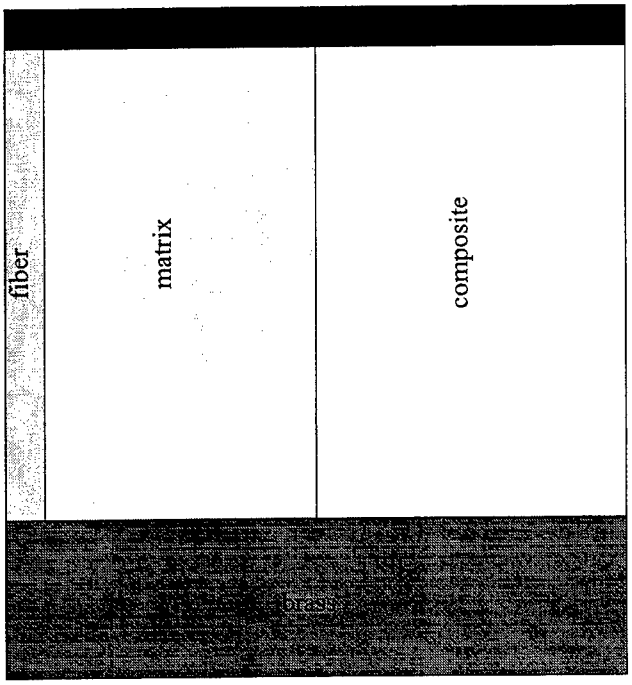


Fig. 6 A geometric model accounting for fiber interaction.

(a)



(b)

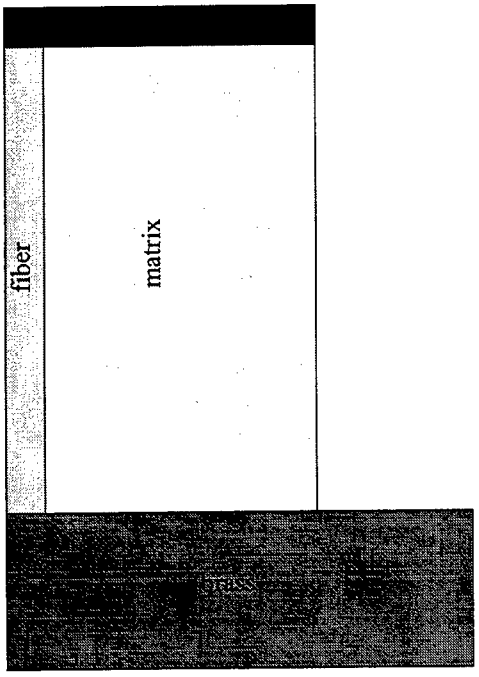
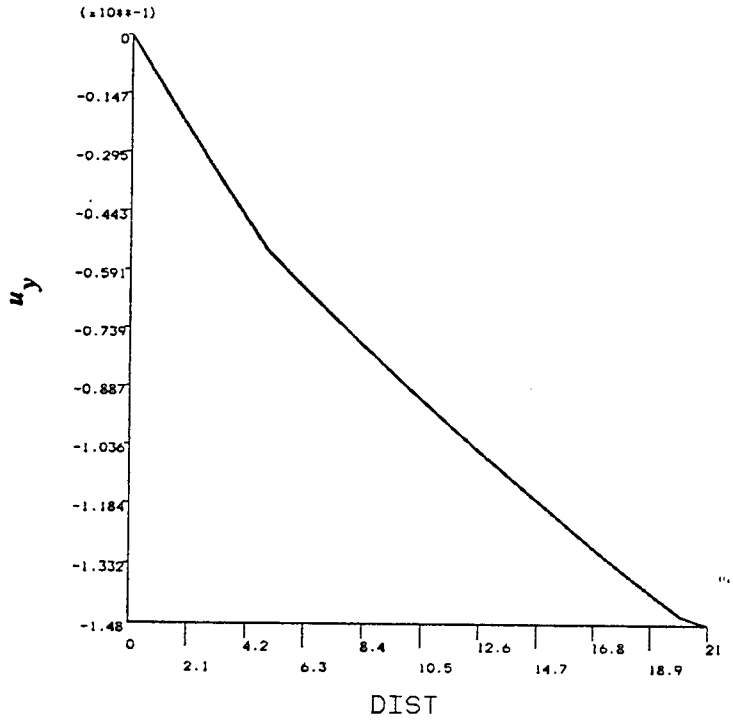


Fig. 7 (a) The generalized self-consistent composite model, and (b) the basic model with an extended brass-plate beyond the specimen.

(a)



(b)

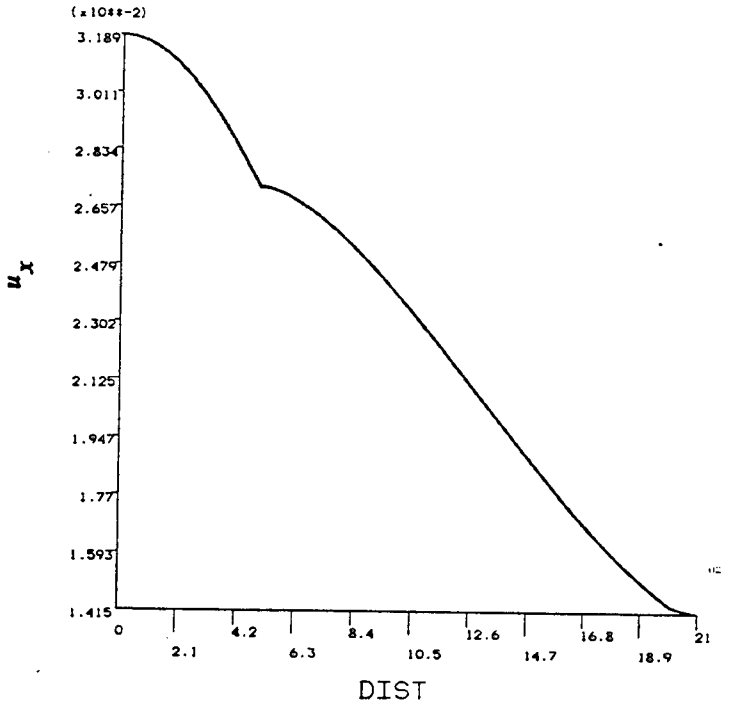
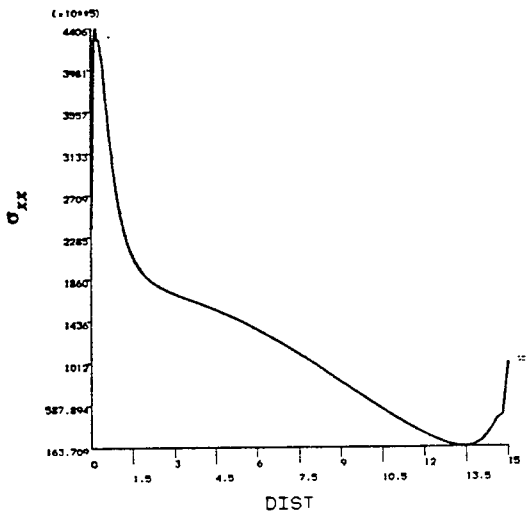
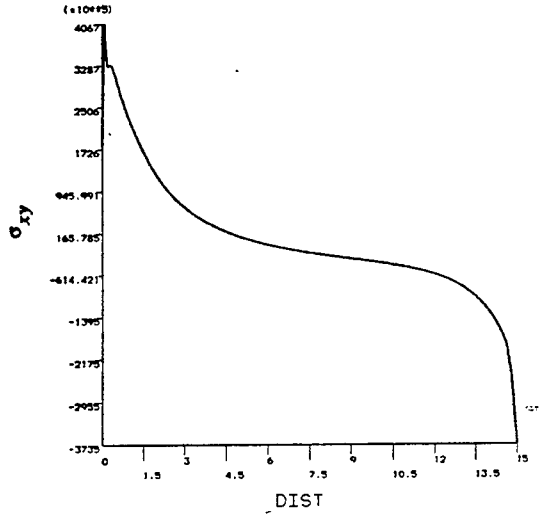


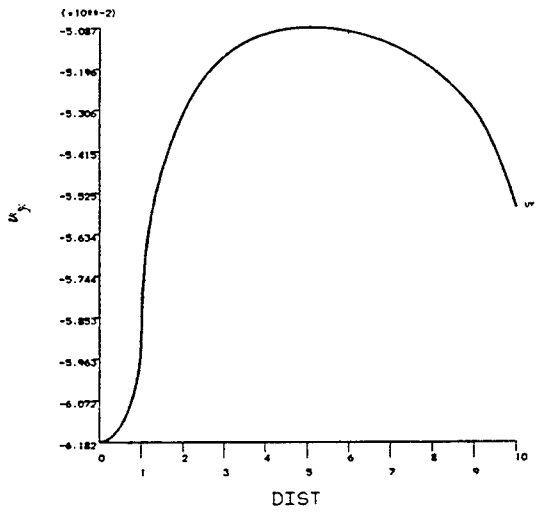
Fig. 8 (a) Normal displacement, and (b) transverse displacement along a section in an extended model (radially) corresponding to the free edge in the basic model.



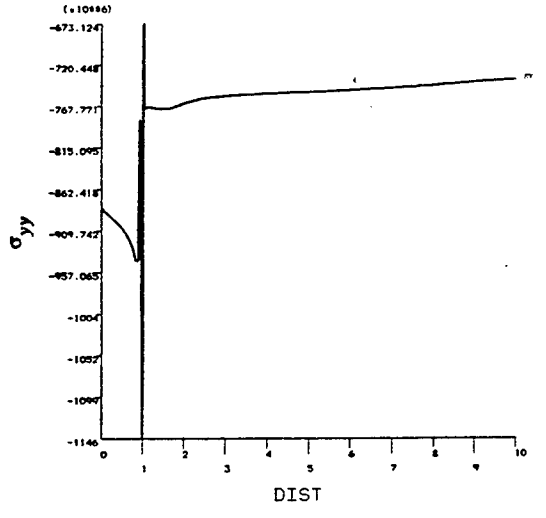
(a)



(b)

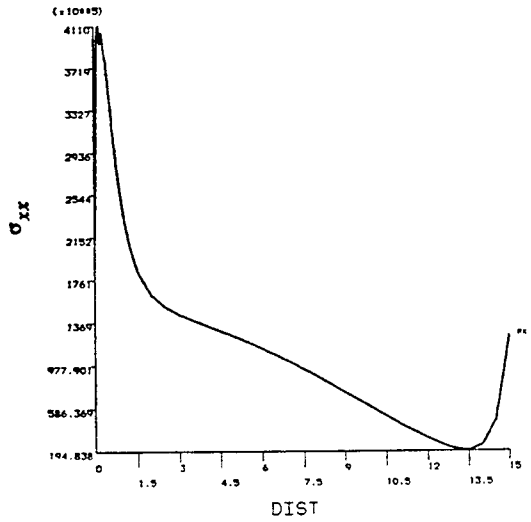


(c)

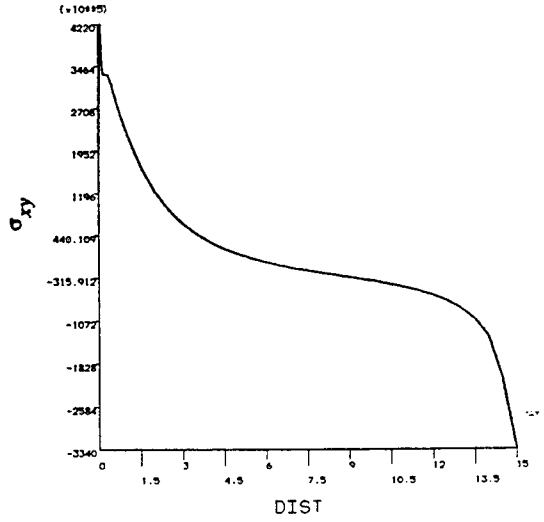


(d)

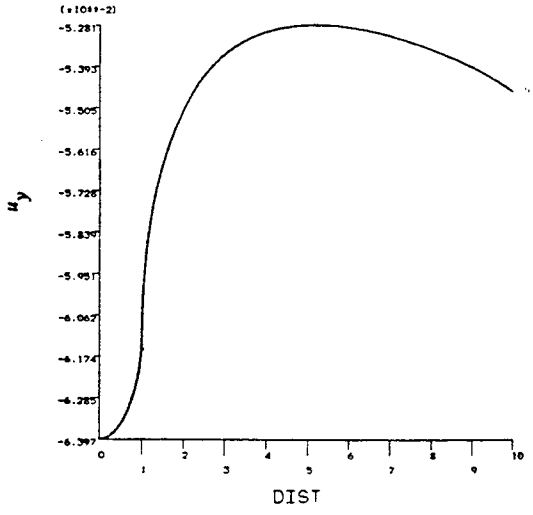
Fig. 9 (a) Normal stress, (b) shear stress along the fiber/matrix interface; and (c) normal deflection, (d) normal stress along the composite/brass interface in an extended model.



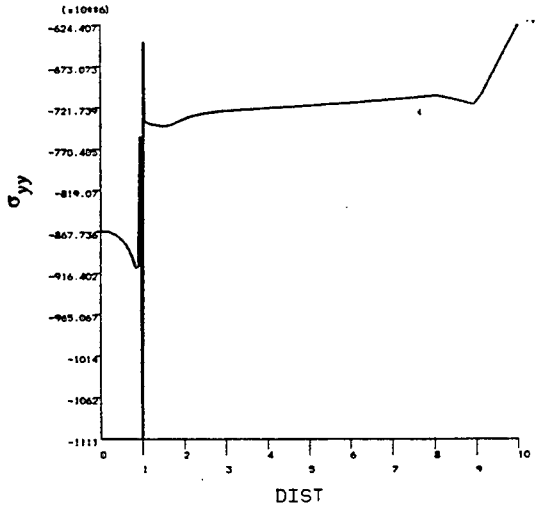
(a)



(b)



(c)



(d)

Fig. 10 (a) Normal stress, (b) shear stress along the fiber/matrix interface; and (c) normal deflection, (d) normal stress along the composite/brass interface in the basic model with imposed displacement boundary conditions.

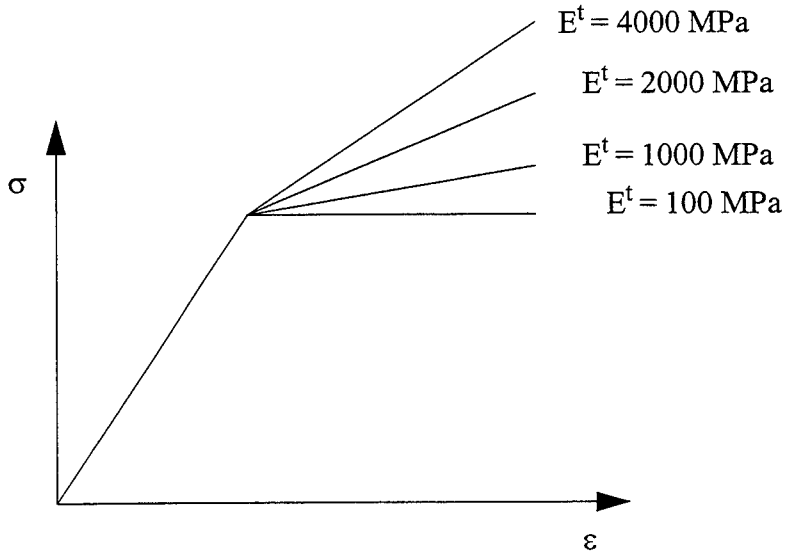
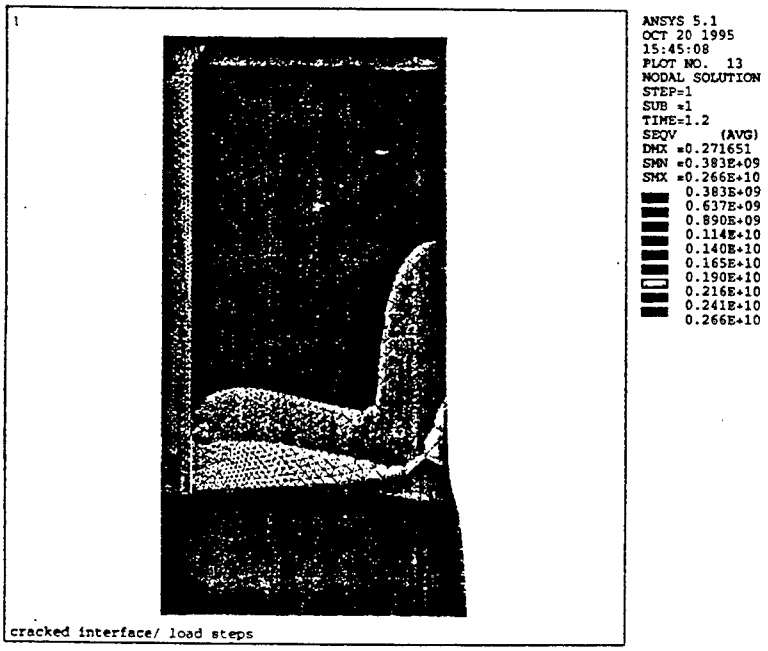
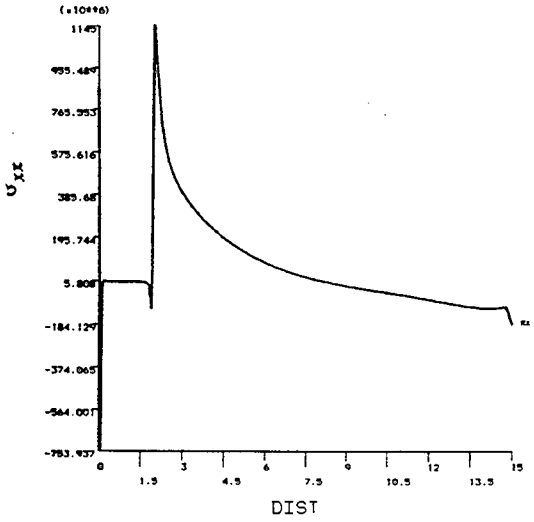


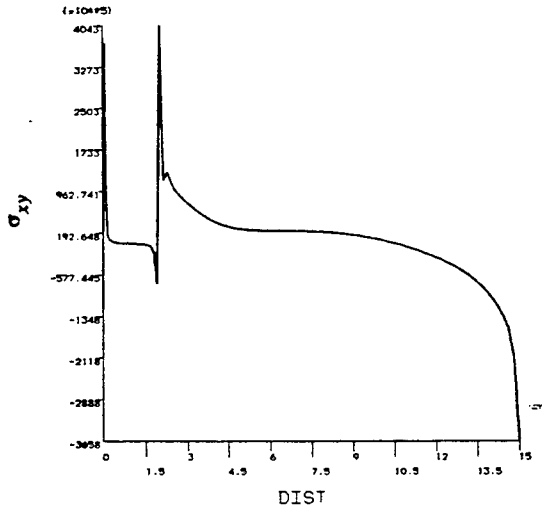
Fig. 11 A schematic graph showing the different E^t which have been tried.



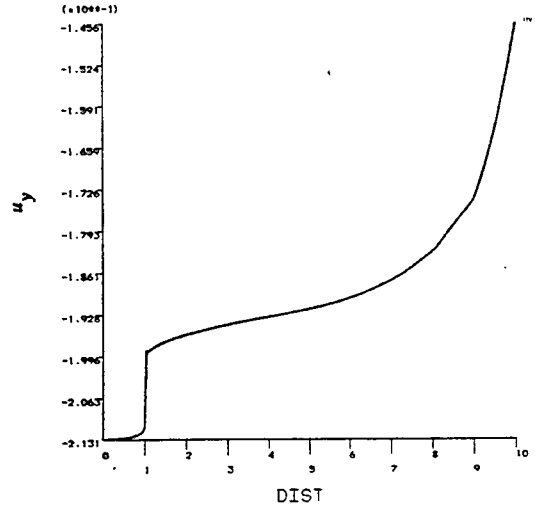
(a)



(b)

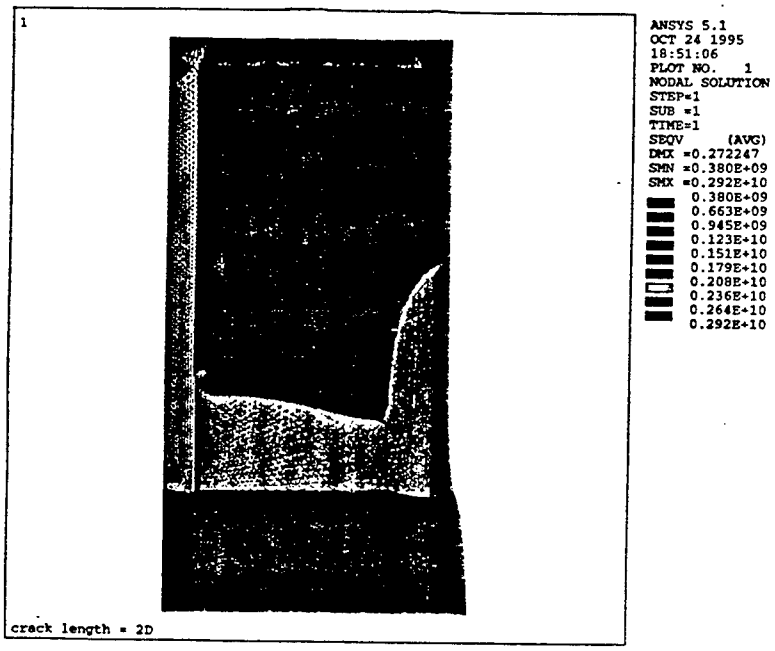


(c)

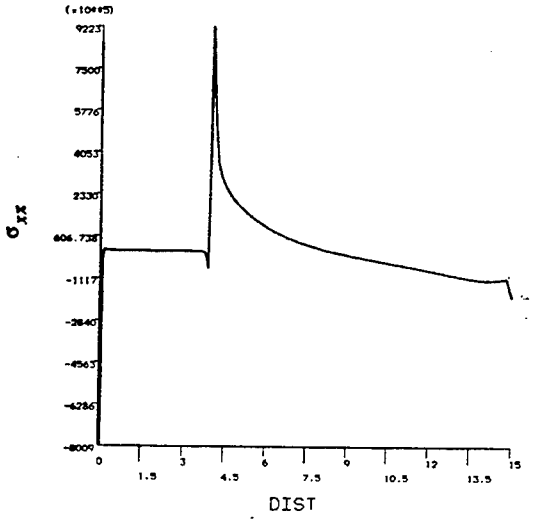


(d)

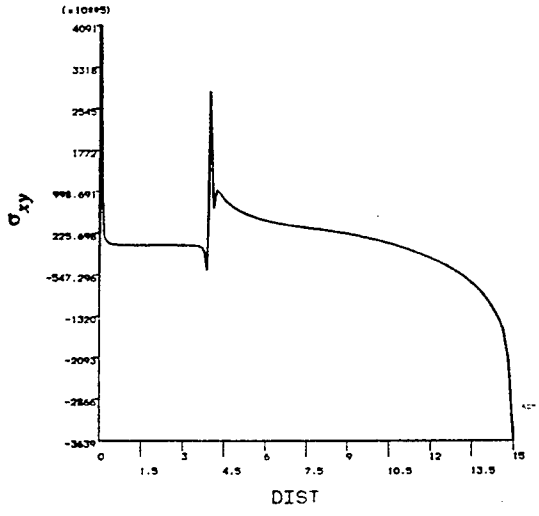
Fig. 12 (a) Contour plot showing the effective stress field and a crack of length equals D along the fiber/matrix interface, (b) normal stress, (c) shear stress along the fiber matrix/interface; and (d) normal deflection along the composite/brass interface.



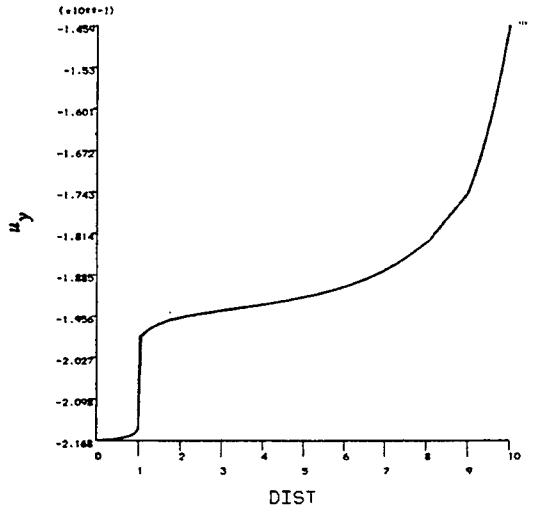
(a)



(b)



(c)



(d)

Fig. 13 (a) Contour plot showing the effective stress field and a crack of length equals 2D along the fiber/matrix interface, (b) normal stress, (c) shear stress along the fiber matrix/interface; and (d) normal deflection along the composite/brass interface.

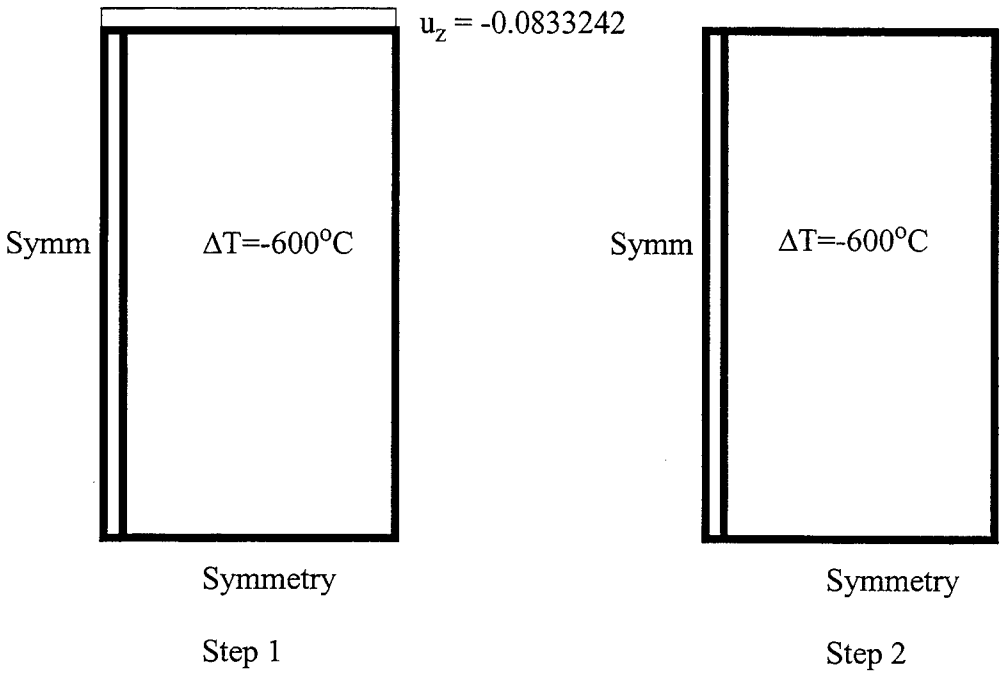
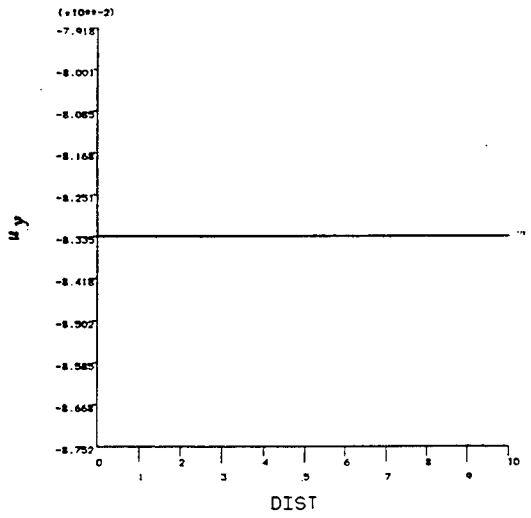
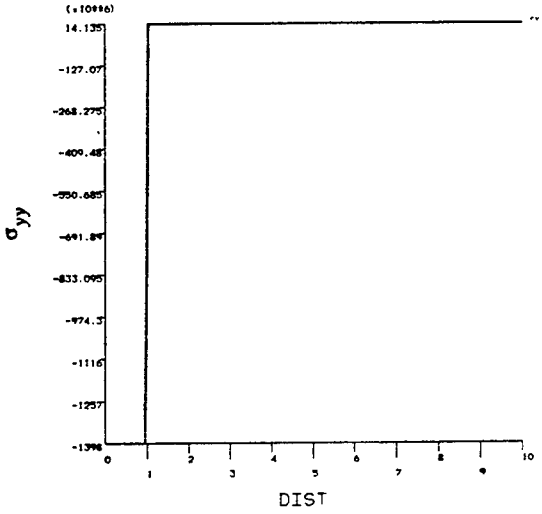


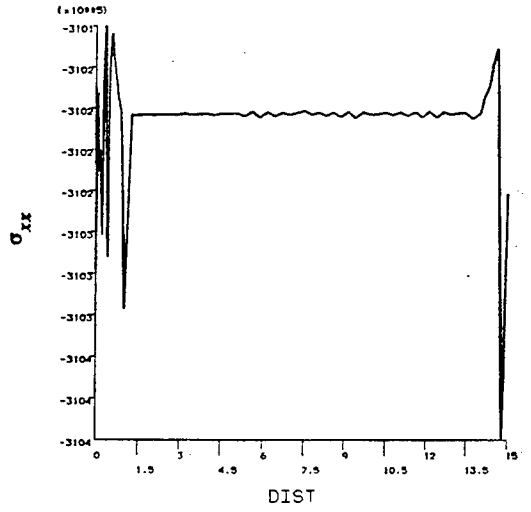
Fig. 14 Two load steps in the thermal analysis.



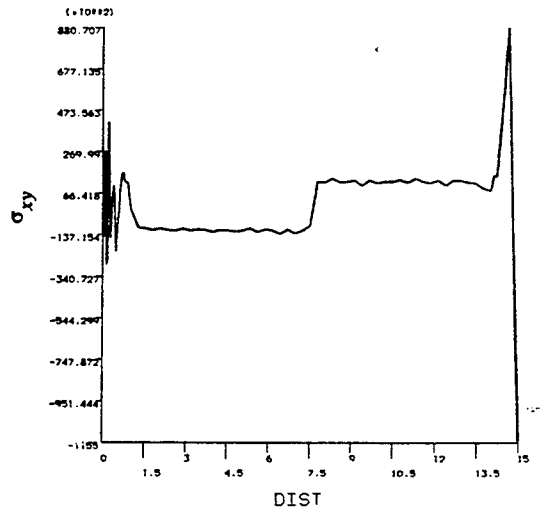
(a)



(b)

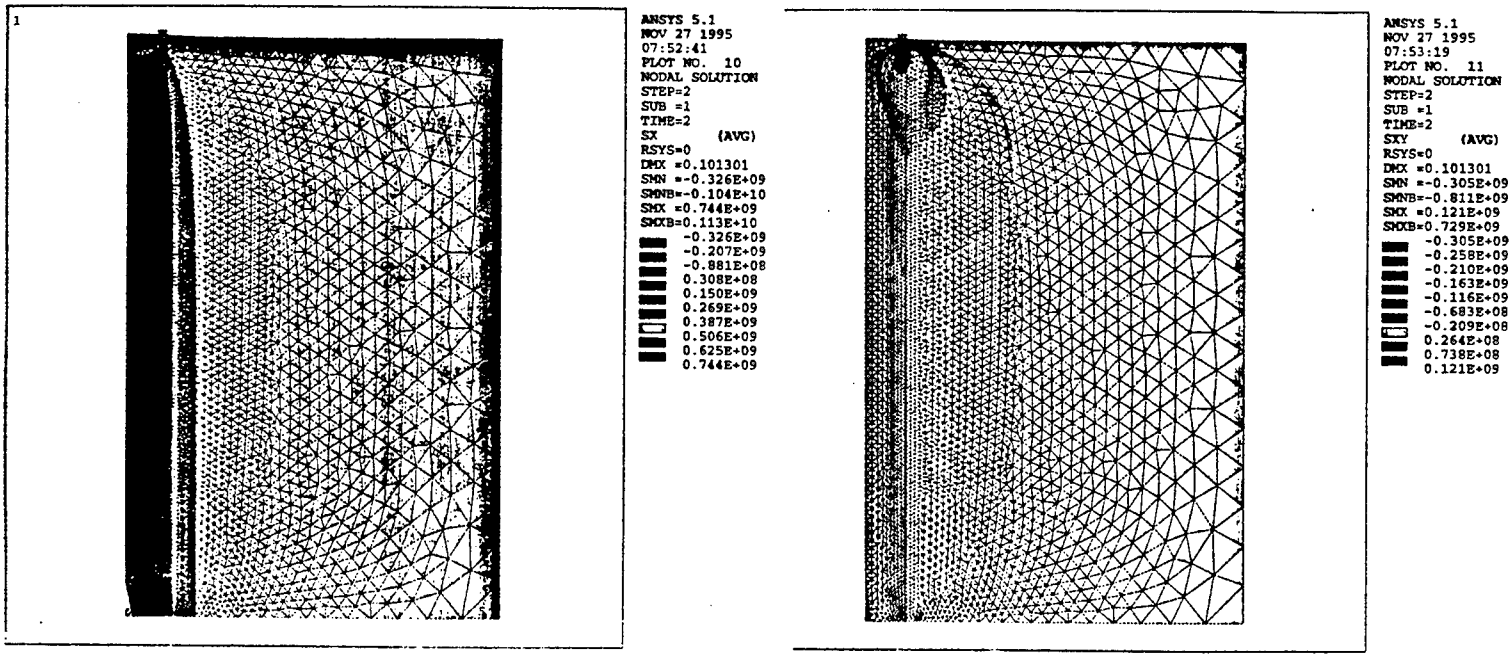


(c)

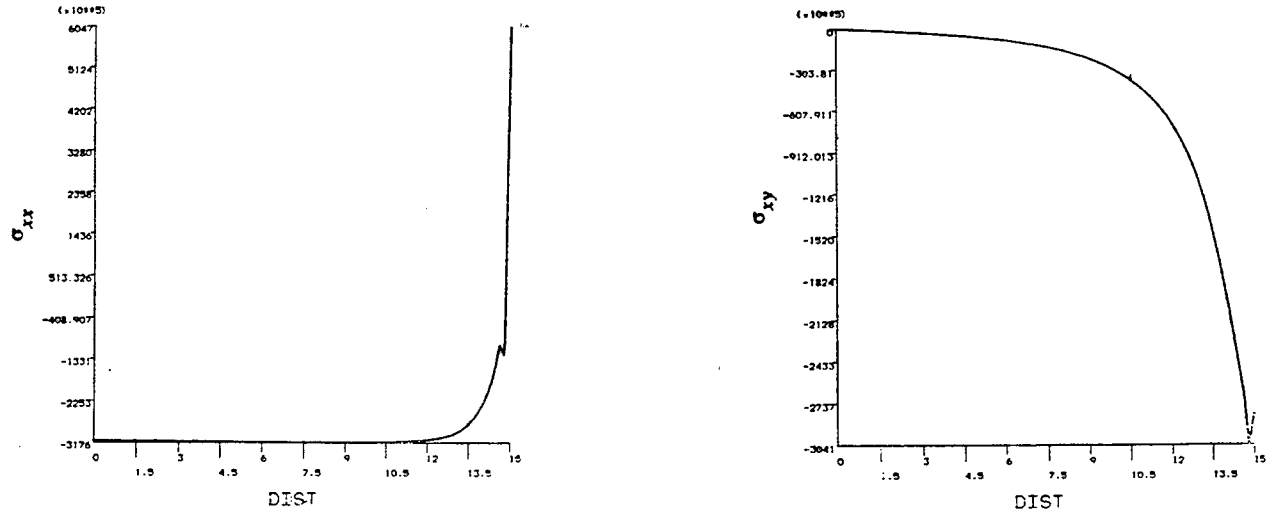


(d)

Fig. 15 (a) The constant applied displacement, (b) normal stress along the composite edge; and (c) the normal stress, (d) the shear stress along the fiber/matrix interface.



(a) (b)



(c) (d)

Fig. 16 Contour plots for (a) σ_{xx}, (b) σ_{xy} in thermal problem; and (c) normal stress (σ_{xx}), (d) shear stress (σ_{xy}) along the fiber/matrix interface.

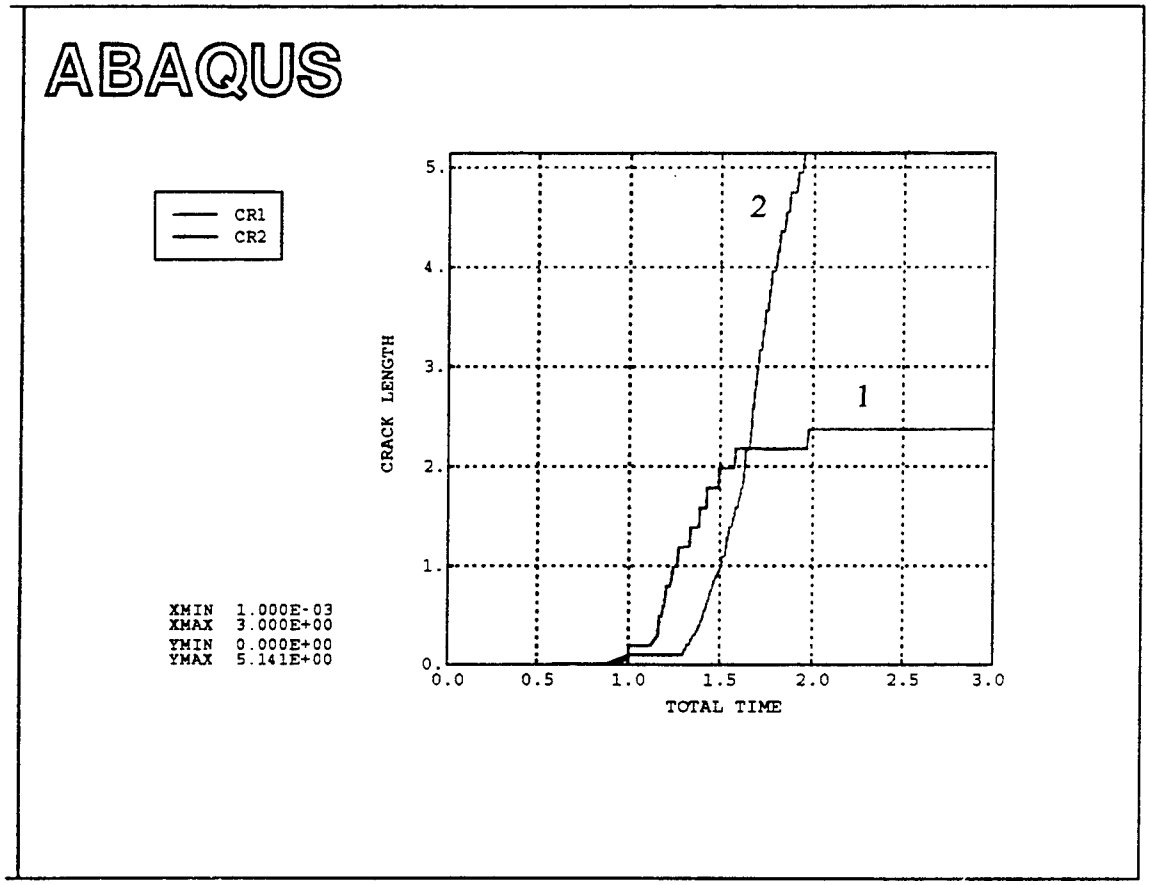


Fig. 17 A typical crack growth graph.

Appendix

Table 5: Properties of SiC (fiber)

T (°C)	E (MPa)	ν	a (x10^{-6}/°C)
21	393000	0.25	3.9907
93	390000	0.25	4.0289
204	386000	0.25	4.0989
315	382000	0.25	4.1801
426	378000	0.25	4.2655
538	374000	0.25	4.3510
648	370000	0.25	4.4324
760	365000	0.25	4.5074
871	361000	0.25	4.5718
1093	354000	0.25	4.5723

Table 6: Properties of Titanium (Matrix)

T (°C)	E (MPa)	ν	σ_y	E_t	a (x10^{-6}/°C)
21	113700	0.3	900	4600	9.44
149	107500	0.3	730	4700	9.62
315	97900	0.3	517	5400	9.78
482	81300	0.3	482	4800	9.83
649	49600	0.3	303	1700	9.72
900	20700	0.3	35	1200	9.81

Table 7: Properties of brass (base)

T (°C)	E (MPa)	ν	σ_y	E_t	a (x10^{-6}/°C)
	103000	0.34	365	4000	20

TSI MITIGATION: A MOUNTAINTOP DATABASE STUDY

Ismail Jouny
Assistant Professor
Department of Electrical Engineering

Lafayette College
Markle Hall, High Street
Easton, PA, 18042

Final Report for:
Summer Research Extension Program
Wright Patterson Air Force Base

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

Wright Patterson AFB

December 1995

TSI MITIGATION: A MOUNTAINTOP DATABASE STUDY

Ismail Jouny
Assistant Professor
Department of Electrical Engineering
Lafayette College

Abstract

This study addresses the feasibility of adaptive array processing in a multipath environment using space-time adaptive processing concepts. The focus is on mitigation of terrain scattered interference (TSI) entering the mainlobe of a receiving antenna array and its impact on the detection performance of that array. The study also includes an investigation into the statistical features of real TSI and how do they compare to perceived models. The Mountaintop database is used as a testbed for time-domain and transform-domain adaptive processing algorithms developed by the author and many researchers in the field. The statistical analysis is based on a major portion of the available database, however due to excessive computing requirements, the data used in examining the performance of adaptive processing architectures is limited to representative samples of the Mountaintop data, each exemplifying a certain TSI scenario. Issues in adaptive processing in a real multipath environment raised by this study are also summarized.

TSI MITIGATION: A MOUNTAINTOP DATABASE STUDY

Ismail Jouny

I. Introduction

Hot clutter returns (ground-terrain scattered interference) due to grounded and airborne jammers are available under the name (Mountaintop database) to researchers in the field of adaptive processing for radar applications. The Mountaintop data is collected for the purpose of determining the feasibility of TSI mitigation in a multipath environment using adaptive antenna arrays.

This study addresses three major elements of the Mountaintop data: 1) statistical features and how do they correspond to presumed models, 2) specific TSI parameters such as doppler, spatial delays, number of multipath scatterers, angle of arrival, intensity, scattering region, etc, and 3) mitigation of Mountaintop TSI using efficient adaptive processing techniques with particular emphasis on number of delay elements, signal decorrelation, number of antennas, and doppler effects.

A critical evaluation of the Mountaintop database main features is presented in the following section. The results of the statistical analysis study are considered in Section III. The TSI parameters of the Mountaintop database and what could be inferred from the data are discussed in Section IV. Section V focuses on the mitigation aspect in a Mountaintop environment and addresses crucial adaptive processing issues such as number of delay elements, number of auxiliaries, and weight lifetime. Section V also presents the results of an investigation into using transform-domain adaptive processing in a multipath environment. Conclusions and recommendations for future work are presented in Section VI.

II. The Mountaintop Database: Practical Issues

The early stages of this study relied on using about 260 Mbytes data available on a magnetic tape from Rome Lab and supplemented by a few briefings from MIT/Lincoln Lab. About halfway through the work, the author gained access to the World Wide Web Mountaintop homepage with approval from Rome Lab. Representative samples of the Mountaintop data were then loaded from the Internet as needed and then analyzed. Although both data sources are primarily the same; Rome Lab with collaboration from MIT/LL, the Internet data had better calibration, sufficient documentation, and was relatively easier to handle particularly when loaded in a Matlab format. The results presented in this report are based on data retrieved from both sources. The reader is however advised that the Internet source is better manageable, easily accessible and is, therefore, more attractive.

The Mountaintop database includes a collection of backscattered radar return recorded in real-time representing various target/clutter/jammer scenarios. Bistatic scattering due to ground and airborne jamming with known noise background constitutes the major portion of this database. Monostatic clutter data with enough CPI's to cover 360 degrees azimuth span is also available. In addition to these two components, Lincoln Lab has made available data representing the results of a mountaintop jamming experiment using an airborne jammer. Other files include samples of an injected test target in the presence of a ground jammer. Data files containing different jammer scenarios, injected test target and involving IDPCA processing are also part of the disseminated portion of the Mountaintop database. To examine the performance of adaptive architectures in a TSI environment, this study has focused on using bistatic data corresponding to different jamming scenarios. For details concerning each file format and associated calibration files, the reader is referred to the documentation provided with each data file.

The data is collected using RSTER (Radar Surveillance Technology Experimental Radar which is a stationary receiver that can be configured either as a 14×24 array (RSTER-0) or as a 24×14 array (RSTER-90). Some important features are summarized hereafter because they contribute significantly to the understanding of the performance of TSI mitigation. RSTER operates in UHF range (about 435 MHz) with a bandwidth of 0.5 MHz, therefore it is a narrowband receiver that does not necessarily reflect some of the TSI mitigation problems associated with wideband receivers. Furthermore, a UHF radar may not be the most desired type of radars particularly considering that future air surveillance radars and fighters have more stringent size and weight limitations. RSTER is also a stationary receiver which does not necessarily show exact doppler effects and other losses encountered during an airborne receiver. It is important to mention, however, that TSI received by an airborne radar is relatively decorrelated compared to a stationary receiver. This decorrelation of TSI received with a moving array is an important phenomena that has not been addressed extensively. Lincoln Lab solution to the stationary receiver problem by including an inverse displaced phase center array (IDPCA) in the vicinity of RSTER may be sufficient to generate the various TSI with various doppler components but is missing the decorrelation of TSI components that results from relative motion of the receiver. Therefore, the requirements for TSI mitigation depicted in this study may be more conservative than required in a real airborne receiver scenario.

Details concerning the position of each jammer and the azimuth position of RSTER are included in the documentation of each file as well as the actual ".MAT" data file as Matlab parameters. Because of space limitations and the large number of data files loaded to perform the required analysis, this report will not specify all jammer and radar parameters such as altitude, latitude, azimuth, power, and many others. The data was loaded on three SPARC-2, and one SPARC-20 SUN workstations all running Matlab including the Signal Processing and Image Processing toolboxes. Some

of the tasks required days of computing in a batch mode.

Not all the data provided by ROME Lab is calibrated, particularly those available on tape. Calibration files are provided and can be used to obtain calibrated data. The data available on the Internet, however, is all calibrated and ready to process. Finally, there are some discrepancies between the number of pulse repetition intervals of the WWW files and those of the tape. The number of range gates may also be different. It is therefore not possible to combine data from both sources for processing with the same mitigation experiment. Instead, separate mitigation and statistical analysis experiments were run and the results compared.

III. Statistical Analysis

Statistical features of terrain scattered interference are important parameters that significantly affect the performance of adaptive processing algorithms. Studies involving TSI mitigation often model hot clutter as a random process with well defined probability density function. Few of those distributions have become the norm in studies involving detection in a clutter environment. Clutter due to weather conditions, ground terrain, urban structures, sea environment, etc, differ in statistical distribution, skewness, and level of stationarity. TSI is in some cases regarded as a nonlinear combination of various forms of clutter. The Mountaintop database includes clutter due to mountainous terrain, normal weather conditions, and insignificant presence of flocks of bird, and no types of sea clutter. A statistical analysis of Mountaintop bistatic clutter is therefore desirable particularly to those interested in developing X-band TSI mitigation using computer generated clutter models. In our investigation we attempted to limit our models to four famous probability distributions. The first is a characteristic of clutter caused by weather conditions and is modeled as zero-mean

Gaussian.

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{x^2}{\sigma^2} \tag{1}$$

The second type which is due mainly to mountainous terrain is Weibull distributed.

$$f_X(x) = \frac{\eta}{a}\left(\frac{x}{a}\right)^{\eta-1} \exp left[-\left(\frac{x^{\eta}}{a}\right] \quad x \geq 0 \tag{2}$$

The third type is that due to ground terrain with foliage and has received considerable attention in the last few years is K- distributed.

$$f_X(x) = \frac{2}{a\Gamma(\nu+1)}\left(\frac{x}{2a}\right)^{\nu+1} K_{\nu}\left(\frac{x}{a}\right) \quad \nu > -1, \quad x \geq 0 \tag{3}$$

In some cases, clutter is better described using the Gamma distribution particularly that due to urban structures and ground terrain.

$$f_X(x) = \frac{x^{\beta-1}}{b^{\beta}\Gamma(\beta)} \exp\left(-\frac{x}{b}\right) \quad x \geq 0 \tag{4}$$

The Mountaintop TSI was first fitted to match any of the above distributions. If a fit was not possible other possibilities were examined. The results were generalized to include a vast number of bistatic data, samples from different pulse repetition intervals (PRI's), different coherent processing intervals (CPI's) and

received by various antenna elements. Samples of the fitted distribution to various Mountaintop data segments are shown in Figures 3-8. Extensive statistical data analysis revealed that the Mountaintop TSI is in most cases Weibull distributed when ground jammers are used. Gaussian clutter becomes more apparent when airborne jammers are used. In some cases, a Cauchy or a log-normal distribution provided better fit than either of the above probability density function. Furthermore, it seems that the distribution parameters did not change significantly from one pulse to the other, or from one CPI to the other, which is an indication that the data is in general stationary and this may be primarily because doppler effects were not included in

the collected TSI signatures (no IDPCA). When IDPCA was used, the data became mostly Gaussian distributed with wider range of parameter variation over several PRI's and CPI's. These argument support the fact that weight adaptation should be more frequent when TSI has significant doppler components. Another surprising result is the lack of K-distributed clutter which may perhaps be attributed to the lack of foliage in the RSTER experimentation area. It is not possible, however, to conclude that ground TSI is in general Weibull distributed. It seems however, that in a terrain such as the RSTER experimentation area in New Mexico a Weibull distributed TSI is a fairly accurate model when ground jammers are used, and a Gaussian distributed TSI is an acceptable model for TSI due to airborne jammers. Finally, in some PRI's and for certain RSTER look angles, no known distribution seems to fit the backscattered TSI. These cases were however rare and do not change the conclusions of this phase of the study.

IV. Mountaintop TSI Parameters

In an earlier study (AFOSR, Summer 1994) the author has proposed a model for TSI as a linear combination of discrete scattering components located within the glistening surface between the jammer and the radar. Models of similar nature have been proposed by Fante [15] and Compton [33]. TSI received at the m^{th} element and the k^{th} tap of an antenna array is modeled as

$$i_{mk}(t) = \sum_{q=0}^{Q} \beta_l i(t - mT_{il} - (k-1)\Delta - T_{sl}) \exp j\,[t - mT_{il} - (k-1)\Delta) + \psi_l] \qquad (5)$$

where Δ is the tap spacing, Q is the number of TSI components, T_{il} is the interelement delay of the l^{th} multipath element, with amplitude β_l, and electric phase ψ_l, and spatial delay T_{sl}. The doppler frequency of each component is denoted by ω_l. Both

specular and diffuse multipath components are included in the above model and differ only in terms of their spatial delay T_{sl} and magnitude β_l.

The effect of each of the parameters included in the above equation on the performance of a TSI mitigation system have been examined using a Monte-Carlo simulation. It was found that the spatial delay T_{sl} which is a function of the arrival angle of each TSI component had little impact on the performance of the receiving array particularly for narrowband systems. The parameter β_l is certainly important because it reflects the power of the TSI component relative to the direct jamming signal. The electric phase ψ_l had little impact on TSI mitigation and is generally presumed zero. The doppler frequency ω_l is certainly an important parameter that determines the weight lifetime and the duration of the interval within which an SINR of 3 db below the optimal is maintained.

In this study we attempted to derive estimates of some of the above parameters from the Mountaintop bistatic clutter database. Issues concerning tap spacing, number of array elements, and weight lifetime are addressed in the following section.

First, the number of TSI scattering components Q was estimated using the range profile of each illuminated area for different RSTER look angles and different jamming scenarios. Such range profiles of radar backscattering data depend on jammer azimuth position, elevation, as well as RSTER elevation and azimuth. In some cases, such range profiles are distinctly different for consecutive PRI's and different CPI's. It is possible to identify about five to six relatively significant scattering components in a particular range profile. Some range profiles did not include any significant TSI components. The results did not depend significantly on whether the jammer is ground-based or airborne. It is unlikely that a range profile that covers hundreds of Kilometers of mountainous terrain contains very few scattering components, we therefore conclude that TSI is due to a large number of low level scattering components, thus a statistical approach such as those described in the previous section

is better suited for modeling TSI. Because of the limited number of identifiable TSI components, it is not possible to estimate the spatial delay parameters.

Doppler filtering helped identify the range of the doppler components of TSI ω_l. Our doppler processing showed that Mountaintop TSI had a doppler frequency range of about 150 Hz which is consistent with platform motion simulation applied to RSTER using IDPCA. This doppler range is acceptable for an airborne radar receiver traveling at a speed near that depicted by IDPCA simulation.

V. Adaptive Processing in a Multipath Environment

Mountaintop TSI mitigation using the adaptive architecture shown in Figure 9 is the focus of this section. TSI cancellation experiments were conducted using the bistatic signatures of Mountaintop, with emphasis on number of delay elements, number of auxiliaries, and weight lifetime for consecutive PRI's and different CPI's. Data files representing TSI due to ground and airborne jammers were both used in this phase of the study. Furthermore, data files corresponding to IDPCA platform simulation effect were also used. The reader is reminded that IDPCA simulated platform motion is designed to yield a constant receiver velocity which results in a relatively constant doppler range. In real TSI situations where the receiver platform is accelerating or decelerating, one should expect a relatively different TSI doppler signature and consecutively a slightly different weight lifetime.

The architecture of RSTER, particularly the spacing between antennas, the size of the array, and the configuration of the array limits to some extent the number of options available for TSI mitigation. RSTER architecture, however, is acceptable as the results do indicate mainly because the receiver operates in a narrowband mode. In a wideband mode, where TSI is even a more serious threat, RSTER may not be

sufficient for mitigating the effects of hot clutter and a more complicated structure may be desirable such as that proposed by Fante [15]. Furthermore, it is possible to achieve a certain cancellation ratio in an adaptive architecture if the number of auxiliaries, the gain of the main channel and the auxiliaries are chosen properly assuming a certain noise level σ^2 and assuming that the number of mainlobe jammers (discrete TSI components is known N_j). The desired residue r can be approximated as [15]

$$r \approx \sigma^2 \left(1 + \frac{G_M N_j}{N G_a}\right) \tag{6}$$

where N is the total number of antennas and G_M, G_a are the gain of the main and auxiliary channels respectively. The above expression indicates that the residue is mainly due to thermal noise and noise carry-over from the auxiliaries. The above design formula determines the number of auxiliaries needed and the required gain on each auxiliary to achieve a certain cancellation ratio. Although RSTER has a limited number of antenna elements (dependent on how it is configured), the above formula was used (when possible) to obtain estimates of the number of auxiliary channels and their gains for Mountaintop TSI mitigation experiments.

In addition to the narrowband mode of RSTER, this study assumes that the transmitted pulses are insensitive to expected doppler shifts, and that all RSTER antennas have identical patterns. The platform velocity vector is assumed to lie exactly in the same plane as the array axis. Furthermore, the jamming source of TSI is a continuous barrage noise jamming which may be assumed stationary over a CPI but is decorrelated for different PRI's. TSI is assumed to have the same stationarity and decorrelation features as its jamming source.

The theory of adaptive array processing will not be detailed in this report because of space limitations, and the reader is referred to [19, 33, 16] for specific adaptive processing architecture and algorithms. Instead, a brief review of basic space-time

adaptive processing is provided. In principle a space-time adaptive processor generates a weighted linear combination of the spatial samples of the array antennas and the temporal samples of several pulses.

Each coherent processing interval corresponds to a cube of data of dimensions $M \times N \times L$ where M is the number of PRI's (varies between 12 and 16), N is the number of antennas (main and auxiliaries), and L is the number of samples per PRI (ranges from about 400 to about 1800) depending on which file is being examined. Physically a CPI represents the radar backscatter signal over a certain range and within the angular sector of the transmit beam. At the range gate of interest, the space-time processor weights the output of all antenna elements at all PRI's with a weight vector of length NM, which is the equivalent of beamforming and doppler filtering. A space-time processor is therefore a two-dimensional filtering scheme placing nulls in the azimuth-doppler space to cancel jamming and clutter. The weights are of course the product of the inverse of the covariance matrix Φ and the steering vector X.

$$W = \Phi^{-1} X \tag{7}$$

which requires the solution of MN linear equations which could be in the order of hundreds and requires on the order of $(MN)^3$ operations. Therefore, space-time adaptive processing may require prohibitive computing power particularly when considering airborne surveillance radars with high speed processors on board. Partially adaptive space-time processing schemes with significant reduction in computing power are available and may be used with minimal loss in signal-to-interference plus noise ratio (SINR). In this study, TSI mitigation is not performed in real time but using powerful processors, and our emphasis is on the number of tap-delay elements, the number of auxiliaries, and weight lifetime, hence a fully adaptive STAP architecture was adopted.

The adaptive architecture shown in Figure 9 was used to determine the feasibility of Mountaintop TSI mitigation. Data files representing various jammer, target, and RSTER scenarios were loaded and mitigation experiments conducted. The cancellation ratio was chosen as the performance measure for TSI mitigation experiments which is a better suited measure particularly when the target is absent. Figures 10-15 show samples of cancellation ratios as a function of the number of tap delay elements, tap spacing, and number of array elements. The results of this extensive TSI mitigation testing applied to both Tape and Internet Mountaintop data can be analyzed and compiled as follows:

- A cancellation loss of about 5 db may be incurred if tap spacing exceeded $1/2(B)$ where B is the bandwidth of the received backscatter signal.

- Using additional auxiliaries with proper gain values does improve the cancellation ratio in almost all TSI scenarios. When proper tap spacing is used, using more than five auxiliaries produces little effect on the cancellation ratio.

- Regardless of whether the jamming source is airborne or grounded, the number of tap-delay elements is in the order of hundreds (about 200-250) for an acceptable cancellation performance with less than five auxiliaries used.

- Cancellation of TSI due to airborne jammers requires more tap delays and auxiliary antennas than its counterpart due to ground jamming. This indicates that TSI with non-zero doppler remains more serious than TSI generated by a stationary jammer.

- Tap spacing is not only a factor of the bandwidth of the received signal but also dependent on the source of TSI. Hot clutter generated by an airborne jammer requires closer tap spacing than that due to a ground jammer. This

result is consistent with the assumption that doppler associated with each TSI component limits the lifetime of the weights of the array.

- Weights computed during PRI1 can be used with little loss in cancellation to compute the residue using successive PRI's for TSI generated with a ground jammer. When an airborne jammer is used, the weights must be updated at each PRI independently of target presence. A weight lifetime of about 2.2 ms maintains the cancellation ratio within 4 db of its optimum value.

- Increasing the number of auxiliaries has little effect on weights lifetime. Increasing tap spacing prolongs the duration of a weight estimate, but results in suboptimum cancellation performance.

The above conclusions are based on general observations made over three hundred mitigation experiments. There are few scenarios representing a particular RSTER look angle, and a specific jammer illumination sector, where a cancellation performance that is relatively robust in terms of tap spacing and number of auxiliaries is attainable.

The above TSI mitigation studies were conducted using conventional space-time adaptive processing techniques applied directly to the Mountaintop bistatic database. Practical concerns and limitations of the ARPA Mountaintop data were also addressed earlier. Some of the features that are lacking either in the Mountaintop database or the mitigation process itself are briefly addressed hereafter.

The issue of TSI decorrelation due to radar platform motion is considered next followed by an attempt to model multipath interference due to rotating engine propeller blades which results in additional mainbeam interference. Finally a transform-domain approach to TSI mitigation and an assessment of its performance is investigated.

A. *TSI Decorrelation Due to a Moving Radar Platform*

As mentioned earlier RSTER is a stationary radar receiver which is incapable of capturing the effect of an airborne radar on multipath interference particularly those attributed to doppler and TSI decorrelation. The intent here is to alert the reader to the issue that the above estimates of tap delay elements and auxiliaries needed to mitigate TSI may be conservative due to the fact that TSI undergoes a decorrelation process by virtue of the motion of the radar platform. Studies have shown [20] that signals received by a moving array decorrelate at a rate that depends on the spacing and the direction of the antennas. Paulraj et al [20] proposed a weighted covariance averaging technique that maximizes the decorrelation between TSI components for any translational dither of the array. The idea is to divide the data into overlapping segments, estimate the covariance matrix for each particular segment and Φ_l then the covariance matrix of the array is estimated using

$$\Phi = \sum_{l=1}^{P} f_l \Phi_l \tag{8}$$

where f_l is chosen so that all off-diagonal entries in Φ are forced to zero. For details of the above algorithm, the reader is referred to [20]. The algorithm requires the arrival angles of all multipath components to achieve complete off-diagonal zeroing. Clearly, such an approach is computationally intense that increases the complexity of an already complicated adaptive processing algorithm. Preliminary results obtained using simulated TSI data, however, do indicate that a better cancellation ration with about 10% less tap delays is possible. The above algorithm was not applied to the Mountaintop data simply because of the inherently lengthy and costly computation time needed. The purpose of this approach is to assert that it is possible to make use of the relative radar platform motion to decorrelate mainbeam TSI and improve the performance of TSI mitigation.

B. Mitigation of propeller generated multipath effects

. This study has focused mainly on hot clutter (TSI) or multipath interference scattered from the ground into the mainbeam of a radar receiver. Because RSTER is a stationary receiver, the effects of multipath interference generated due to the high rotational speed of the propellers of the surveillance radar aircraft were not included in the Mountaintop database. It is well known that propeller motion modulates the frequency and the amplitude of all received signals including TSI, to produce a larger set of correlated multipath components entering the mainbeam of the radar receiver. Figure 16 shows a typical modulated signal due to the motion of a propeller or a helicopter rotor. The spectrum of that signal is shown in Figure 17 and it resembles that of an FM signal with spectral lines spaced at a rate proportional to the number Q and the angular velocity of the propeller blades ω_r.

$$\omega_d = Q\omega_r \tag{9}$$

This added doppler effect ω_d leads to even further constraints on weight lifetime and increases the number of delay elements needed to mitigate TSI. The effects of multipath interference on TSI mitigation was also simulated using computer generated TSI samples which were the modulated by the rotating blades according to a model described in [17]. The multipath signal received at the mth antenna and kth tap delay is then

$$\begin{aligned} v_m k(t) \;=\; & \sum_{q=0}^{Q-1} A_r sinc\left(\frac{4\pi}{\lambda}\frac{L_2 - L_1}{2}\right)\cos(\theta)\sin\left(\omega_r t + \frac{2\pi q}{Q}\right) \\ & \exp\left[j\left(\omega_c t - \frac{4\pi}{\lambda}\left(R + vt + \frac{L_1 - L_2}{2}\cos(\theta)\sin\left(\omega_r t + \frac{2\pi q}{Q}\right)\right)\right)\right] \end{aligned} \tag{10}$$

where L_1 and L_2 denote the distance of the edges of the blades from the propeller center. The range to ground is R, λ is the wavelength of the transmitted signal, and θ is the angle between the plane of propeller rotation, the line of sight from the array to

the center of the propeller. A_r is a scale factor that determines the power of the multipath interference. For a typical aircraft propeller, the above modulation contributes about 200 Hz in additional doppler frequencies. This may cut the weight lifetime by a factor of two. Preliminary results obtained using simulated TSI signatures indicate that the number of required tap delay elements does not increase significantly but the weight lifetime is reduced by about 30%.

C. Transform-Domain Adaptive Processing

In an earlier study on TSI mitigation [19], the author examined the impact of subbanding prior to adaptive filtering on the performance of an adaptive processing architecture for hot clutter mitigation. Although simulated TSI was used in the experimental phase of that study, it was determined that uniform subbanding (via DFT) or nonuniform subbanding (via wavelet transform) prior to adaptive processing would improve the cancellation ratio only when the number of subbands is large.

Both transforms were attempted during the course of this study. No improvement in cancellation ratio was detected because the number of subbands was smaller than what is needed for any noticeable improvement in array performance. Increasing the number of subbands increases the computational burden and is therefore impractical. An alternative approach to TSI mitigation was then attempted where joint adaptation of all array weights was replaced by independent adaptations (LMS algorithm) applied to each frequency bin in the DFT case, or to each scale bin in the wavelet case. Figure 18 shows a block representation of this algorithm where the incoming backscatter data is subbanded prior to applying the LMS algorithm independently in each subband. The advantage is a reduction in computing demand and consequently in hardware requirements, and the disadvantage is that the performance of the TSI mitigation system is suboptimal.

Both types of subbanding produce better nulling of correlated jamming than the

no-transform case regardless of the number of TSI components. The results indicate that subbanding followed by adaptation in each subband independently of the other subbands may reduce the computational burden of TSI mitigation and produce better nulling of correlated interference. Weight lifetime is not an issue in this case because all weights are adaptively estimated for each subband. Convergence may be an issue in this case, but previous studies [19] have indicated that transform-domain adaptation actually improves the convergence rate of the array. Further work is needed to better understand the full potential of this approach.

VI. Conclusions and Future Work

TSI mitigation using conventional space-time adaptive architectures is attainable with a high computational cost made even higher by the need to frequently update the array weights (about 2.2 ms weight lifetime). The hardware requirements are, in this author's opinion, vast and almost impractical. The number of arithmetic operations grows at a rate equal to the cube of the number of degrees of freedom. Such a prohibitive computational requirement renders the inclusion of further pre-processing via subbanding less attractive.

The Mountaintop database lacks two significant features; TSI decorrelation due to radar platform motion is missing, and the multipath contribution of the propellers of the radar aircraft. Both missing information are due to the stationarity of RSTER, and are not recovered by the utilization of IDPCA. Preliminary results used simulated TSI indicate that signal modulation due to receiver propeller increases the vulnerability of adaptive processing architectures to TSI threat.

Limiting the number of antenna elements may be possible (about six auxiliaries), but the temporal processing requirements are in the order of hundreds. Further

velopment in this field should focus on generating a better representative database for hot clutter that is recorded using an airborne receiver as ultimately intended, and on generating efficient array processing algorithms with computational requirements that are within the capabilities of modern digital signal processors.

Acknowledgements: The author wishes to thank the Air Force Office of Scientific Research for supporting this project. The author acknowledges and appreciates the support of fellow researchers at Wright Patterson Air Force Base, OH.

List of Figures

Figure 1: A block diagram of RSTER

Figure 2: Schematic of Mountaintop Data

Figures 3-8: Samples of the statistical fitting of the Mountaintop data

Figure 9: Schematic of a TSI mitigation system

Figures 10-15: Samples of the performance of TSI mitigation. Cancellation ratio is shown versus number of channels for different number of delays.

Figure 16: Modulation effects of an aircraft propeller rotating at high speed.

Figure 17: The spectrum of the returned radar signal after being modulated by a propeller.

Figure 18: Schematic of an alternative TSI mitigation system where RSTER data is subbanded using wavelets or DFT then LMS is applied in each subband.

References

[1] M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael eds., *Wavelets and their applications*, Jones and Bartlett, Boston, 1992.

[2] S. G. Mallat, " A Theory for Multiresolution Signal Decomposition: The wavelet Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, pp 674-693, July 1989.

[3] I. Daubechies, "Orthonormal Basis of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, Vol. XLI, pp 909-996, 1988.

[4] W. D. White, "Wideband Interference Cancellation in Adaptive Sidelobe Cancellers," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 19, No. 6, pp 915-924, November, 1993.

[5] F. W. Vook, R. T. Compton, "Bandwidth Performance of Linear Adaptive Arrays with Tapped Delay-Line Processing," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 28, No. 3, pp 901-908, July 1992.

[6] W. F. Gabriel, "Adaptive digital processing investigation of DFT subbanding vs transversal filter canceler," *Naval Research Laboratory technical report*, NRL report 8981, July 1986.

[7] W. F. Gabriel, "Adaptive Processing Array Systems," *Proceedings of The IEEE*, Vol. 80, No. 1, pp 152-162, January 1992.

[8] J. T. Mayhan, A. J. Simmons, and W. C. Cummings, "Wide-Band Adaptive Nulling Using Tapped Delay Lines," *IEEE Transactions on Antennas and Propagation*, Vol. 29, No. 6, pp 923-936, November 1981.

[9] L. E. Brennan and I. S. Reed, "Adaptive Cancellation of Scattered Interference," *Adaptive Sensors, Inc.*, final report, December 1982.

[10] R. T. Compton, "The bandwidth performance of a two-element adaptive array tapped delay-line processing," *IEEE Transactions on Antennas and Propagation*, Vol. 36, No. 1, pp 5-13, January 1988.

[11] D. R. Morgan and A. Aridgides, "Adaptive Sidelobe Cancellation of Wide-Band Multipath Interference," *IEEE transactions on Antennas and Propagation*, Vol. 33, No. 8, pp 908-917, August 1985.

[12] R. T. Compton, "The Relationship Between Tapped Delay-Line and FFT Processing in Adaptive Arrays," *IEEE Transactions on Antennas and Propagation*, Vol. 36, No. 1, pp 15-26, January 1988.

[13] R. L. Fante, "Cancellation of Specular and Diffuse Jammer Multipath Using a Hybrid Array," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 823-836, Vol. 27, No. 5, September, 1991.

[14] R. L. Fante and J. A. Torres, "Cancellation of Diffuse Jammer Multipath by an Airborne Adaptive Radar", *IEEE Transactions on Aerospace and Electronic Systems*, Vol.31, No. 2, pp. 805-820, April 1995.

[15] R. L. Fante, R. M. Davis, and T. P. Guella, " Wideband Cancellation of Multiple Mainbeam Jammers," submitted to *IEEE Transactions on Antennas and Propagation*, 1995.

[16] Proceedings of 1995 ASAP Workshop, MIT Lincoln Lab.

[17] J. Martin and B. Mulgrew, "Analysis of the theoretical radar return signal from aircraft propeller blades", *Proceedings of 1990 IEEE International Radar Conference (RADAR-90)*, pp. 569-572, Washington, D.C., May 1990.

[18] D. R. Morgan and A. Aridgides, "Comments on; Cancellation of Specular and Diffuse Jammer Multipath Using a Hybrid Adaptive Array," *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 30, No. 3, pp. 932-933, July, 1994.

[19] I. Jouny, "Modeling and mitigation of Terrain Scattered Interference," AFSOR Summer Faculty Report, August 1994.

[20] A. Paulraj, V. Reddy, and T. Kailath, "Analysis of Signal Cancellation Due to Multipath in Optimum Beamformers for Moving Arrays," *IEEE Journal of Oceanic Engineering*, Vol. OE-12, No. 1, pp. 163-171, January 1987.

[21] R. S. Raghavan, "A Model for Spatially Correlated Radar Clutter", *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 27, No. 2, pp.268-275, March 1991.

[22] S. Applebaum, "Adaptive Arrays", *IEEE Transactions on Antennas and Propagation*, pp. 585-598, Vol. 585, No. 5, September 1976.

[23] L. E. Brennan, J. D. Mallett and I. S. Reed, "Adaptive Arrays in Airborne MTI Radar," *IEEE Transactions on Antennas and Propagation*, pp. 607-615, Vol. 24, No. 5, September 1976.

[24] R. A. Monzingo and T. W. Miller, "Introduction to Adaptive Arrays," New York: Wiley, 1980.

[25] R. T. Compton, "Adaptive Antennas," Englewood Cliffs, NJ: Prentice Hall, 1988.

[26] K. Gerlach, "A Numerically Efficient Band-Partitioned Noise Canceller," NRL Report #9050, 1987.

[27] W. Harrison and D. Tufts, "Rapidly Adaptive Mainbeam Jamming Nulling," *Proceedings of ASAP Workshop*, MIT/LL, March 1994.

[28] J. K. Jao, "Application of Two-Scale Surface Scattering to Bistatic Clutter Modeling", *Proceedings of ASAP Workshop*, MIT/LL, pp. 427-447, March 1994.

[29] S. D. Coutts, "Mountaintop Jammer Multipath Mitigation Experiment," *Proceedings of ASAP Workshop*, MIT/LL, pp. 595-625, March 1994.

[30] T. W. Miller and J. Ortiz, "Jammer Multipath Phenomenology and Mitigation," *Proceedings of ASAP Workshop*, MIT/LL, pp. 665-683, March 1994.

[31] C. L. Basham, "Mitigation of Multipath Jamming by Space- Time Adaptive Processing," *Proceedings of ASAP Workshop*, MIT/LL, pp. 665-683, March 1994.

[32] O. Brovko, T. T. Nguyen, and B. J. Heiman, "High PRF TSI Mitigation in Fighter Radars," *Proceeding of ASAP Workshop*, MIT/LL, pp. 683-714, March 1994.

[33] R. T. Compton, "The Effects of Multipath Jamming on an Adaptive Array," *Proceedings of ASAP Workshop*, MIT/LL, pp. 833-861, March 1994.

[34] G. Beylkin, R. Coifman, and V. Rokhlin, *Communications on Pure and Applied Mathematics*, Vol. 44, pp. 141-183.

[35] B. Z. Steinberg and Y. Leviatan, "On the Use of Wavelet Expansion in the Method of Moments," *IEEE Transactions on Antennas and Propagation*, Vol. 41, pp. 610-619, May 1993.

[36] "Adaptive Processing Radar Study", Final Report, Westinghouse Electric Corporation, Electronic Systems Group, Baltimore, MD, 1994.

[37] P. Beckmann and A. Spizzichino, "The Scattering of Electromagnetic Waves From Rough Surfaces," Dedham, MA: Artech House, 1987.

[38] R. S. Raghavan, "A Model for Spatially Correlated Radar Clutter," *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 38, No. 7, pp. 268-275, March 1991.

EXPERIMENTAL RADAR CONFIGURATION

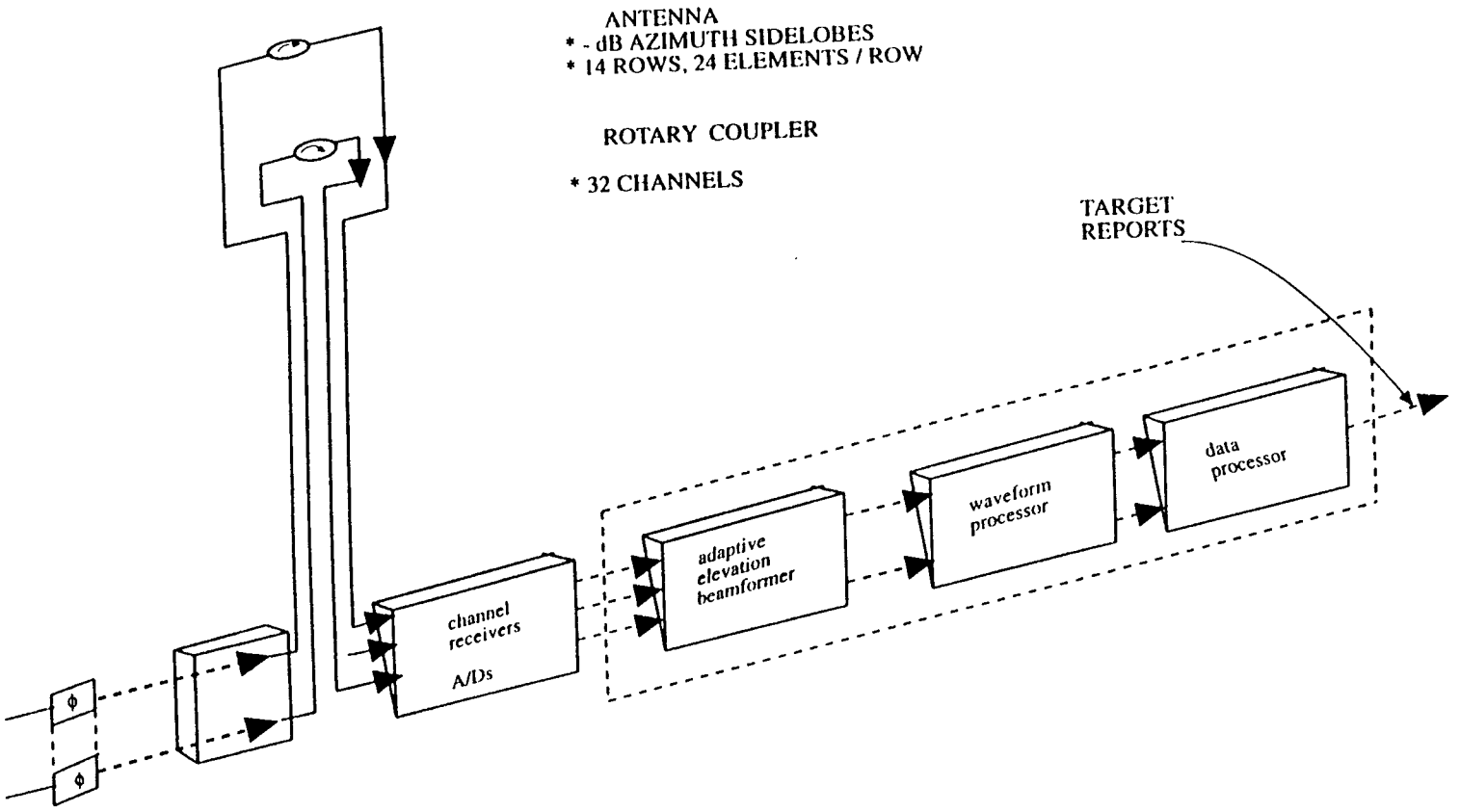


Figure 1

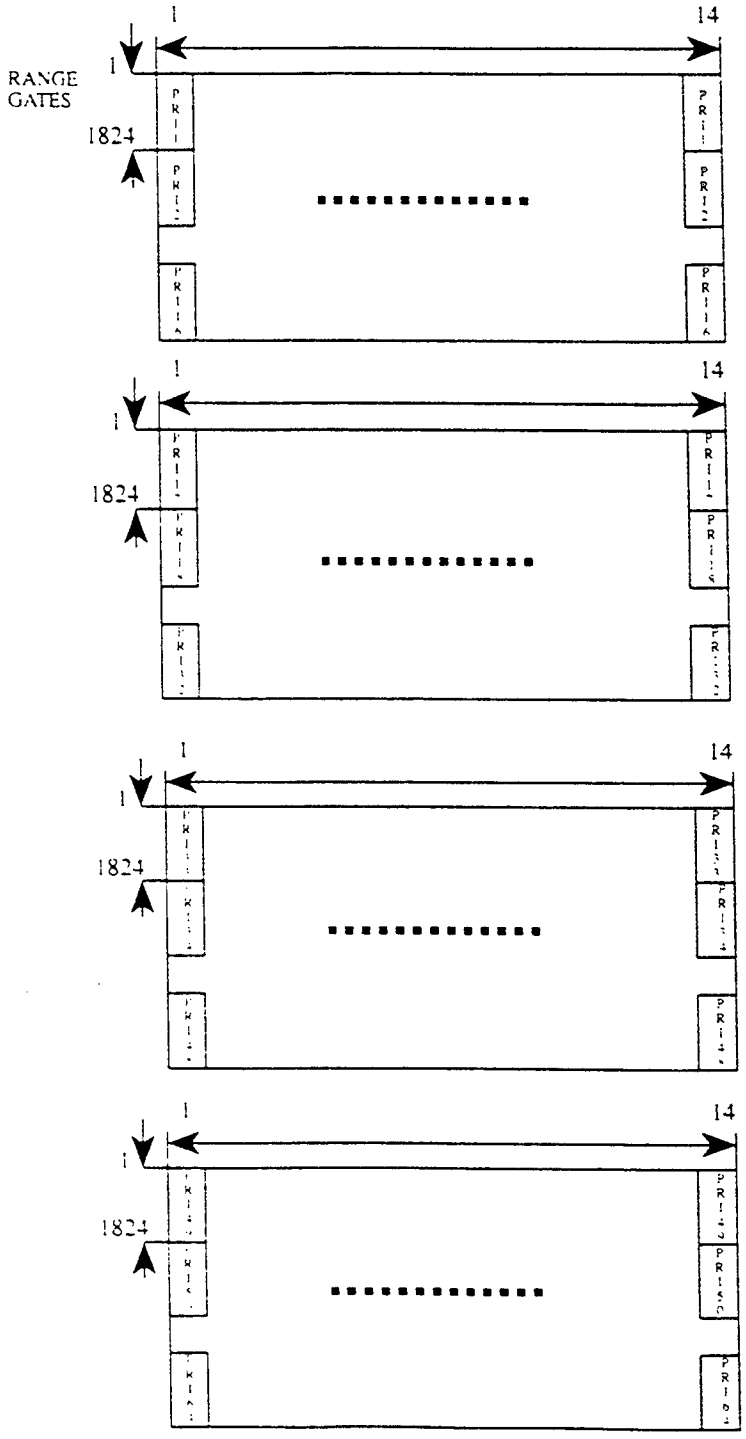


Figure 2

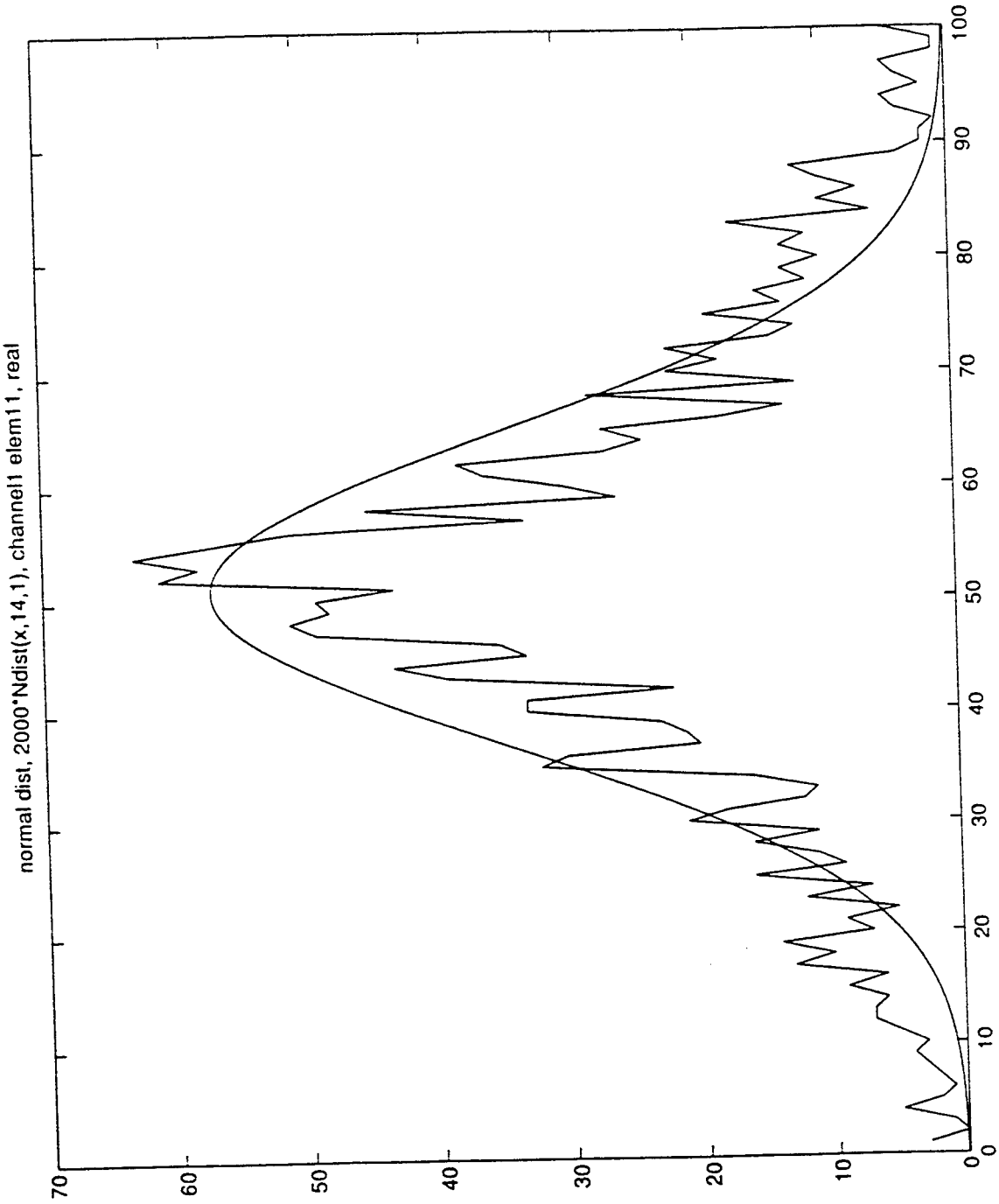

normal dist, 2000*Ndist(x,14,1), channel1 elem11, real
0
10
20
30
40
50
60
70
0
10
20
30
40
50
60
70
80
90
100

Figure 3

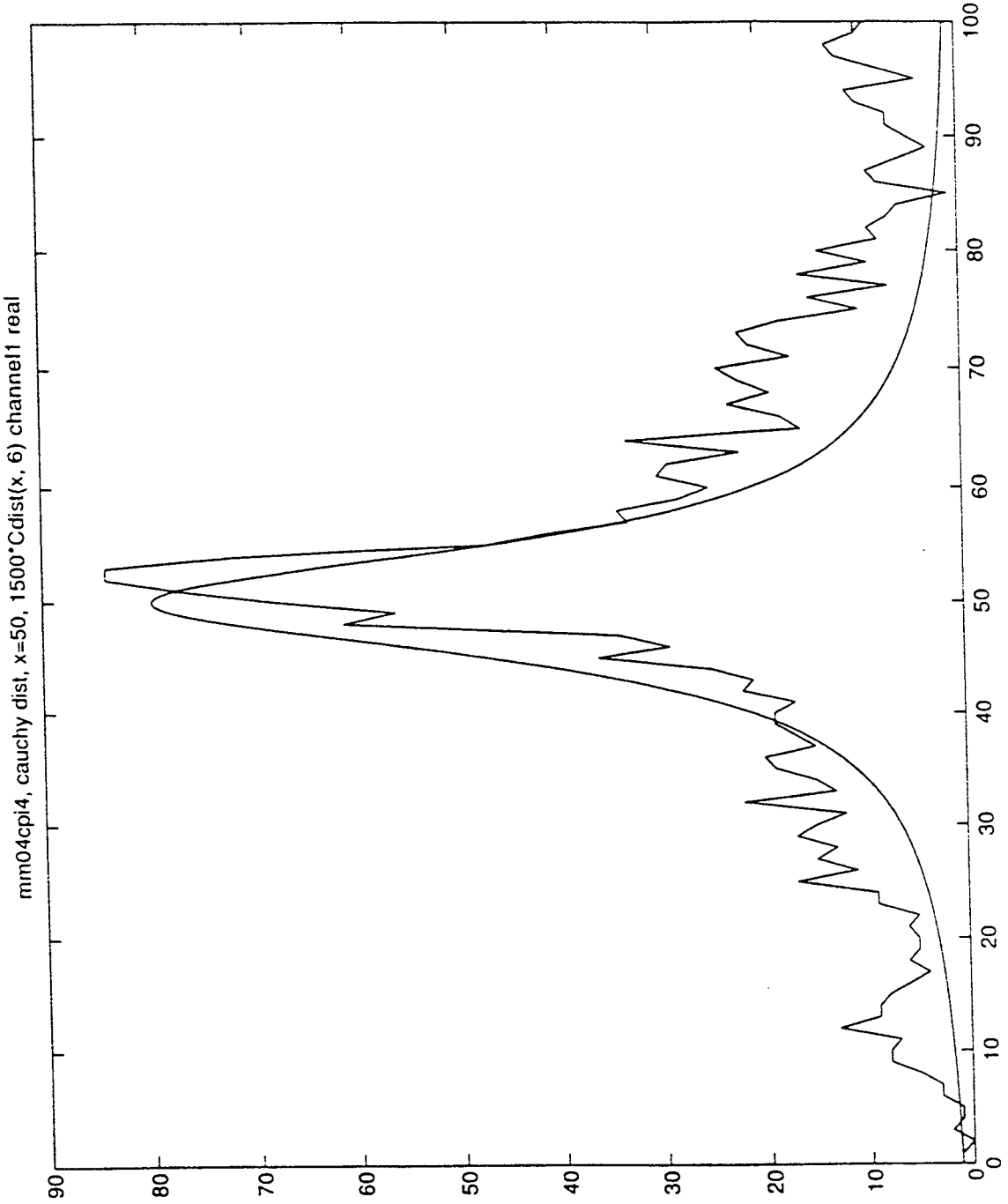

mm04cpi4, cauchy dist, x=50, 1500*Cdist(x, 6) channel1 real
0
10
20
30
40
50
60
70
80
90
100

Figure 4

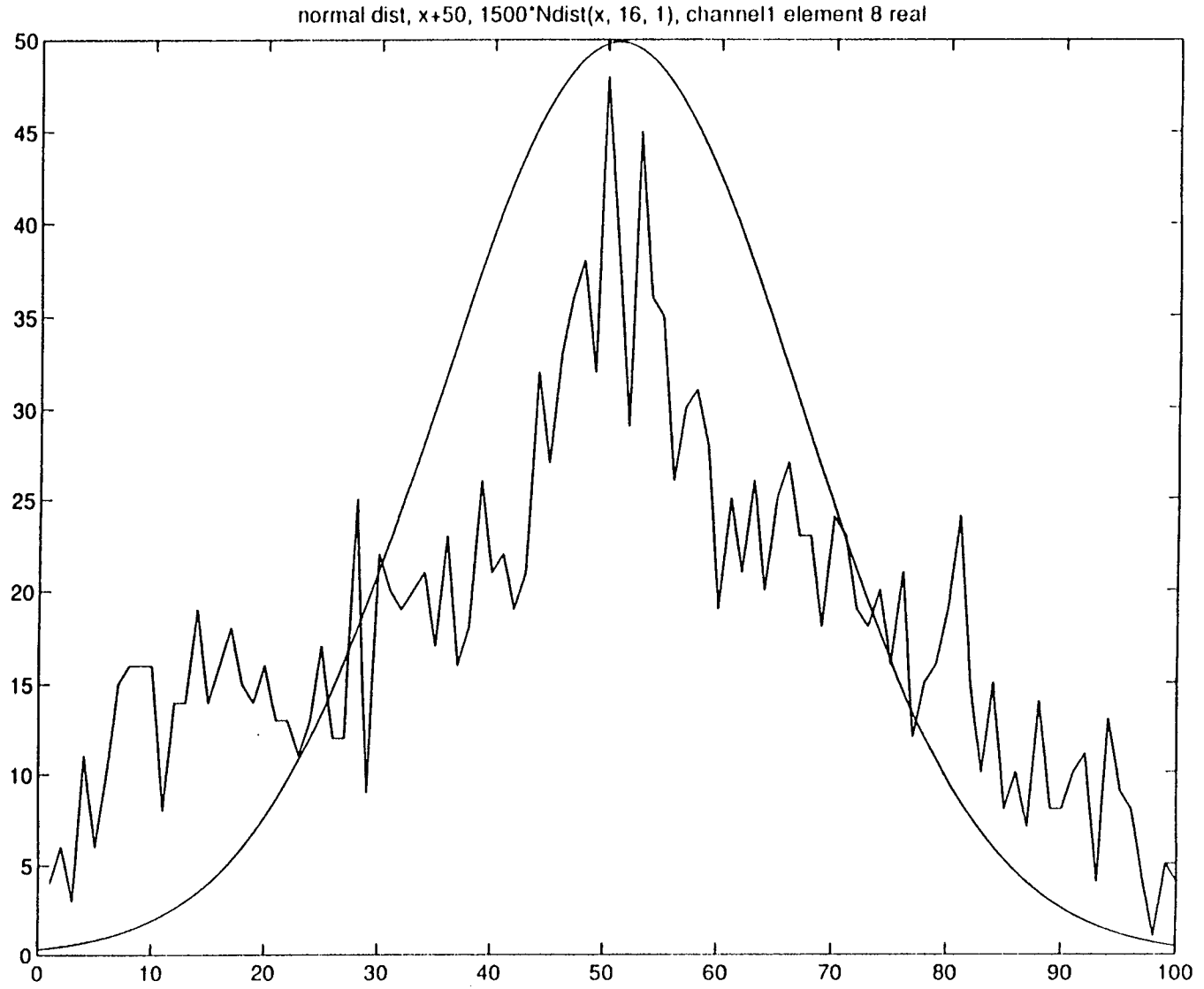

normal dist, x+50, 1500*Ndist(x, 16, 1), channel1 element 8 real
0
5
10
15
20
25
30
35
40
45
50
0
10
20
30
40
50
60
70
80
90
100

Figure 5

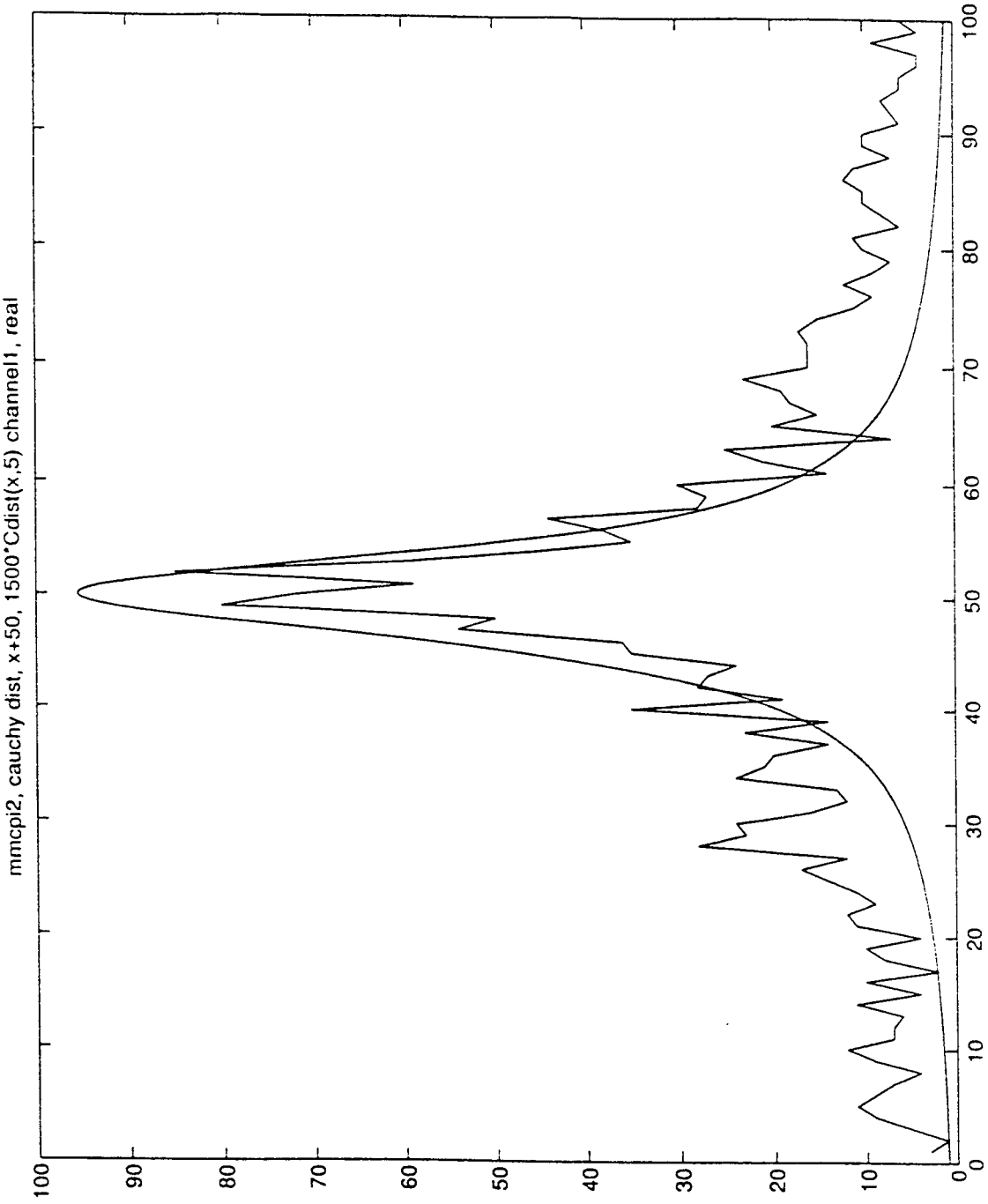

mmcpi2, cauchy dist, x+50, 1500*Cdist(x,5) channel1, real
0
10
20
30
40
50
60
70
80
90
100

Figure 6

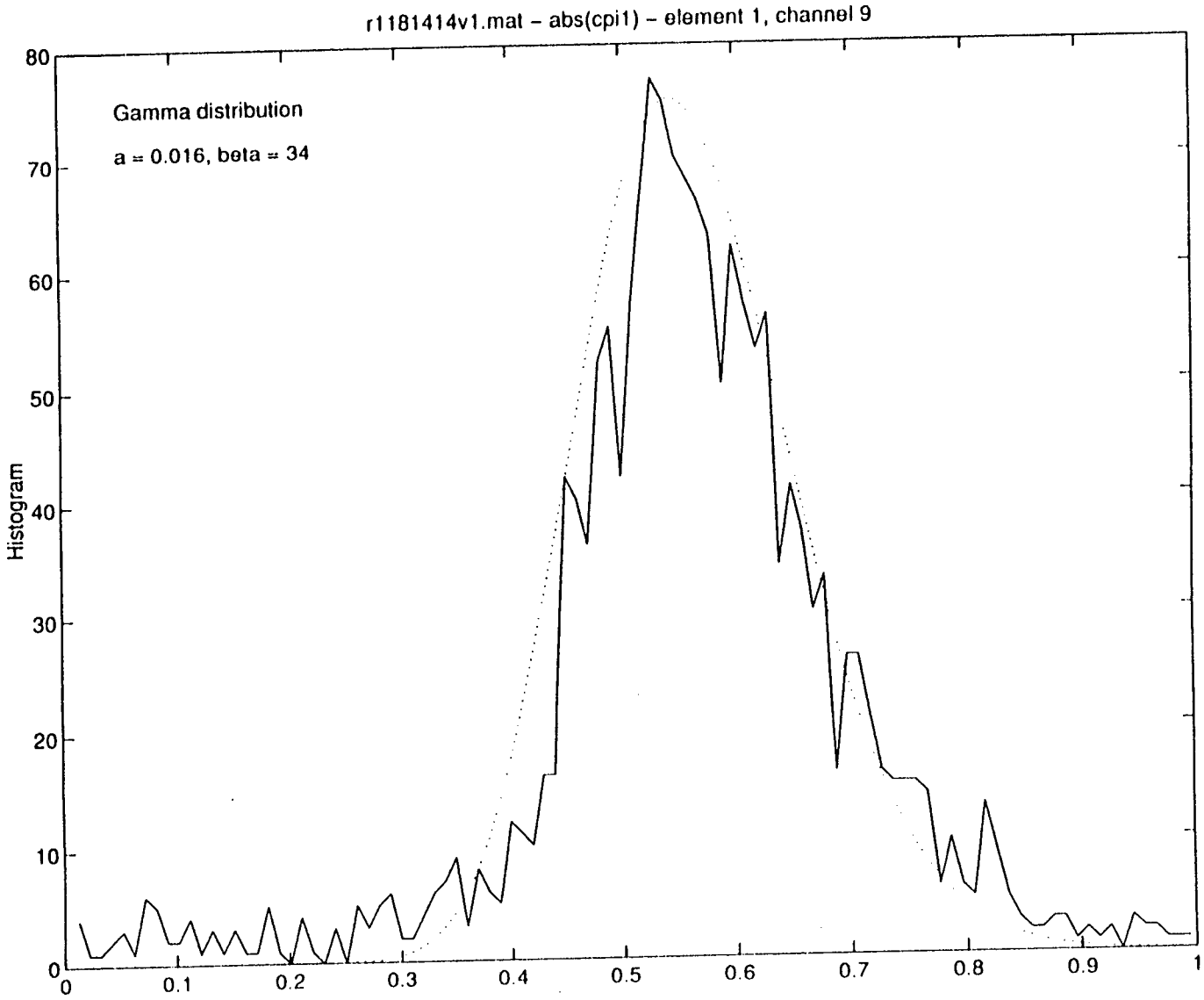

r1181414v1.mat – abs(cpi1) – element 1, channel 9
Gamma distribution
a = 0.016, beta = 34
Histogram
0
10
20
30
40
50
60
70
80
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 7

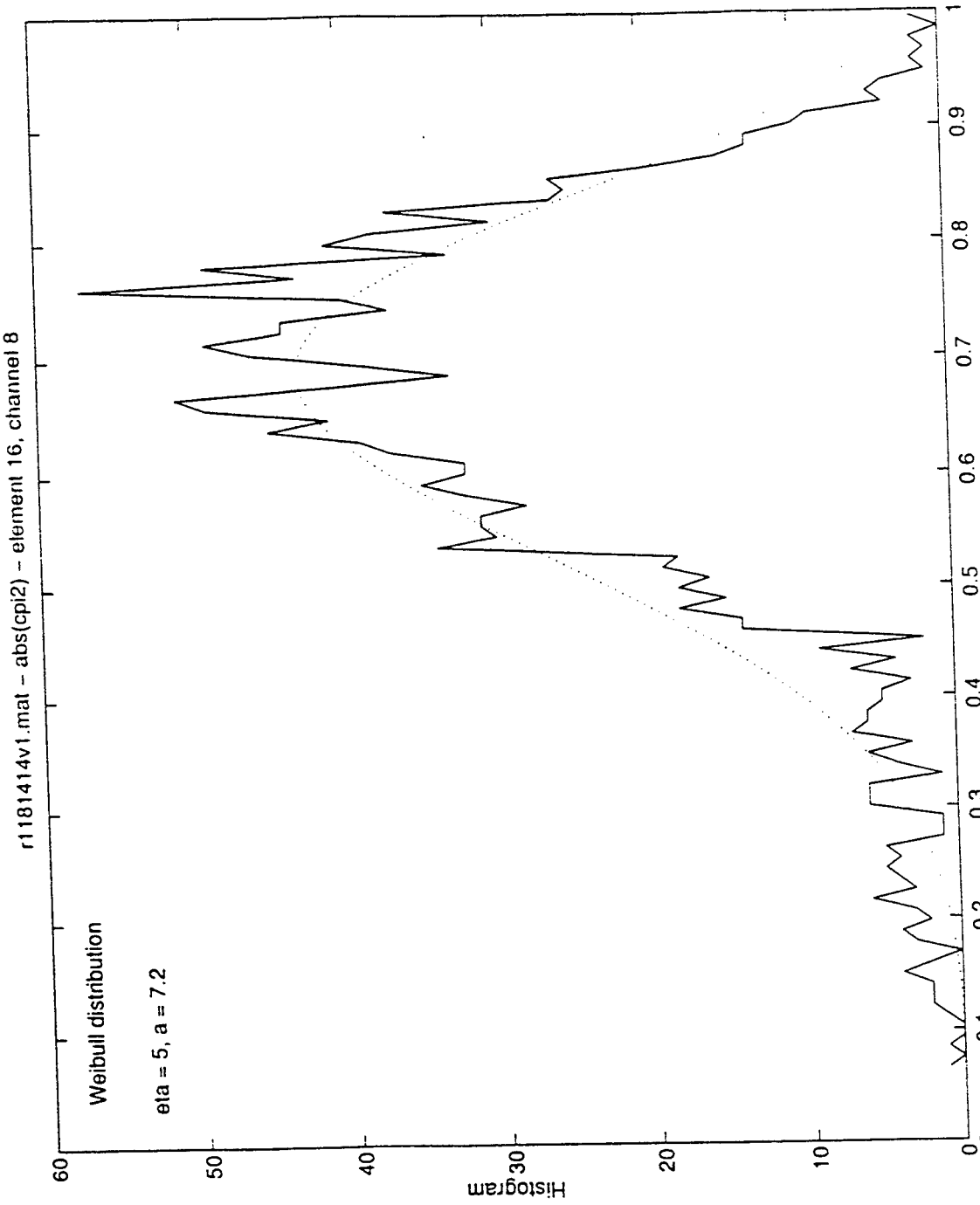

r1181414v1.mat – abs(cpi2) – element 16, channel 8
Weibull distribution
eta = 5, a = 7.2
Histogram
0
10
20
30
40
50
60
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 8

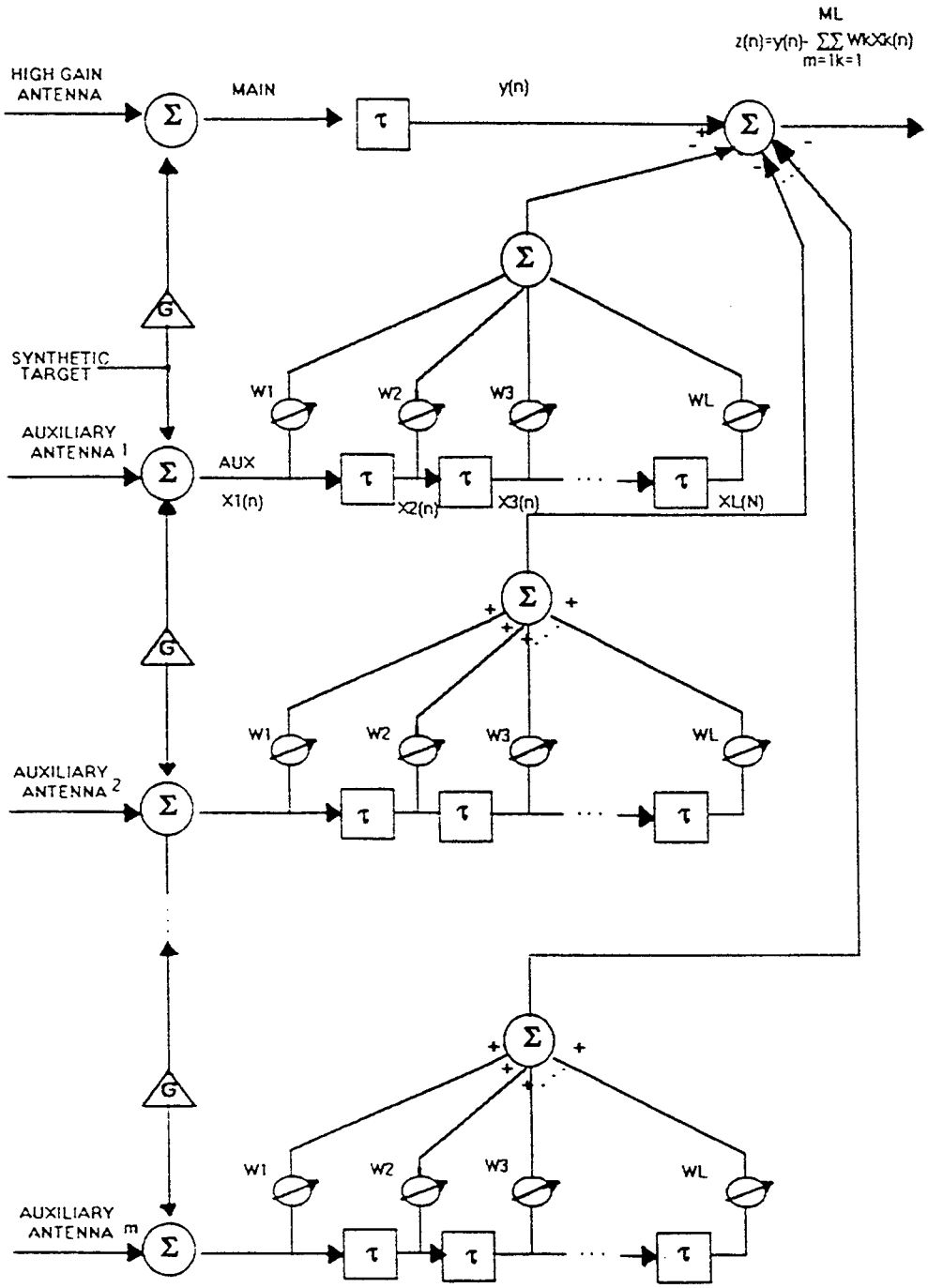

ML
z(n)=y(n)- ∑∑ WkXk(n)
m=1k=1
HIGH GAIN ANTENNA
MAIN
y(n)
SYNTHETIC TARGET
AUXILIARY ANTENNA 1
AUXILIARY ANTENNA 2
AUXILIARY ANTENNA m
AUX
X1(n)
X2(n)
X3(n)
XL(N)
W1
W2
W3
WL
G
Σ
τ

Figure 9

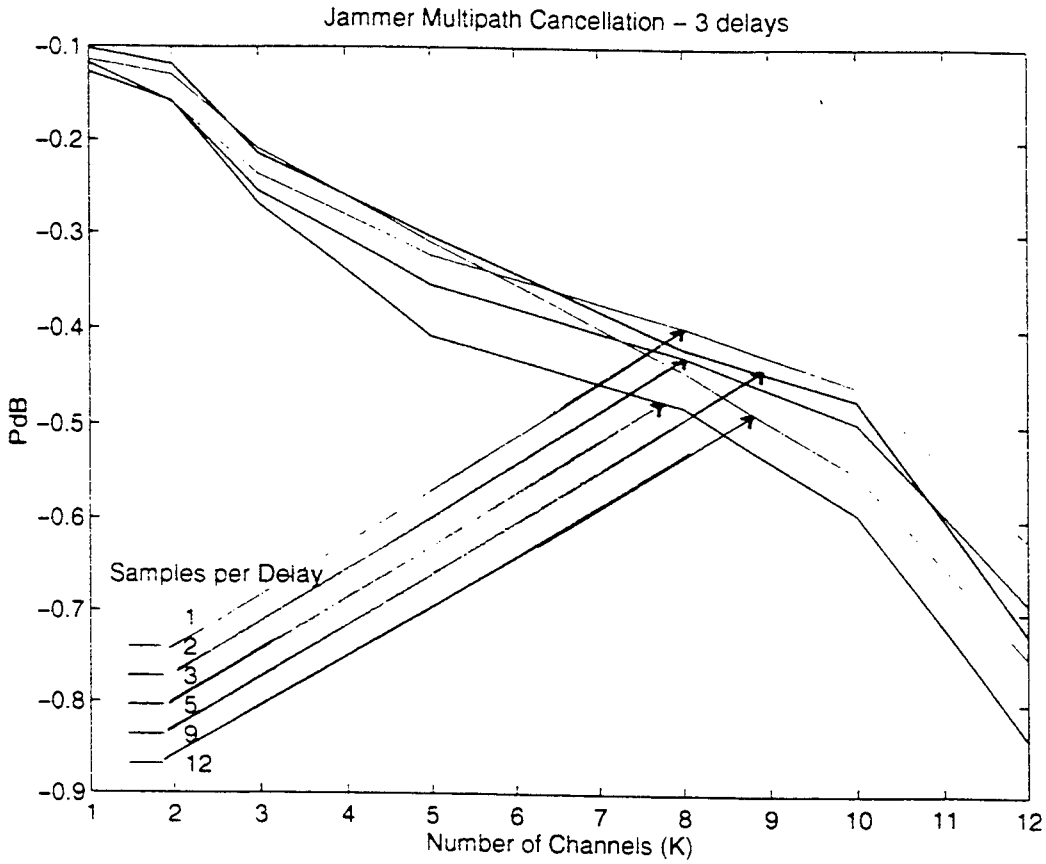

Jammer Multipath Cancellation – 3 delays
PdB
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
1
2
3
4
5
6
7
8
9
10
11
12
Number of Channels (K)
Samples per Delay
1
2
3
5
9
12

Figure 10

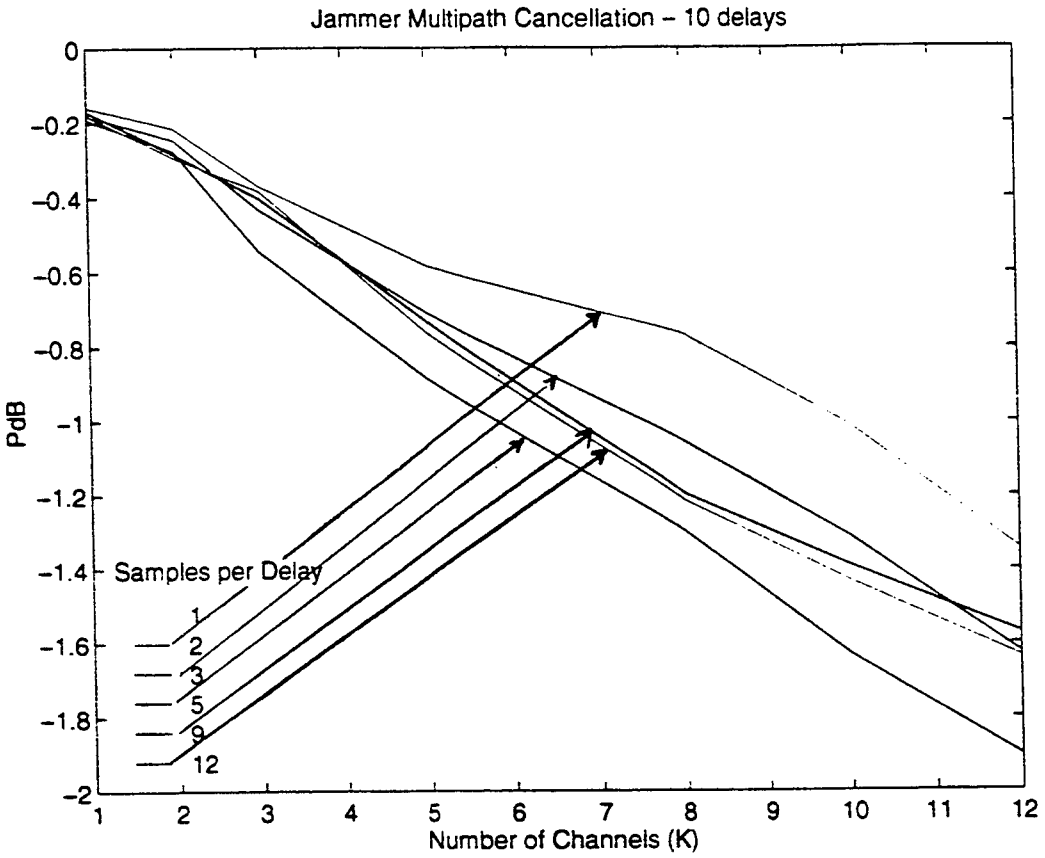

Jammer Multipath Cancellation – 10 delays
PdB
Number of Channels (K)
Samples per Delay
1
2
3
5
9
12
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
1
2
3
4
5
6
7
8
9
10
11
12

Figure 11

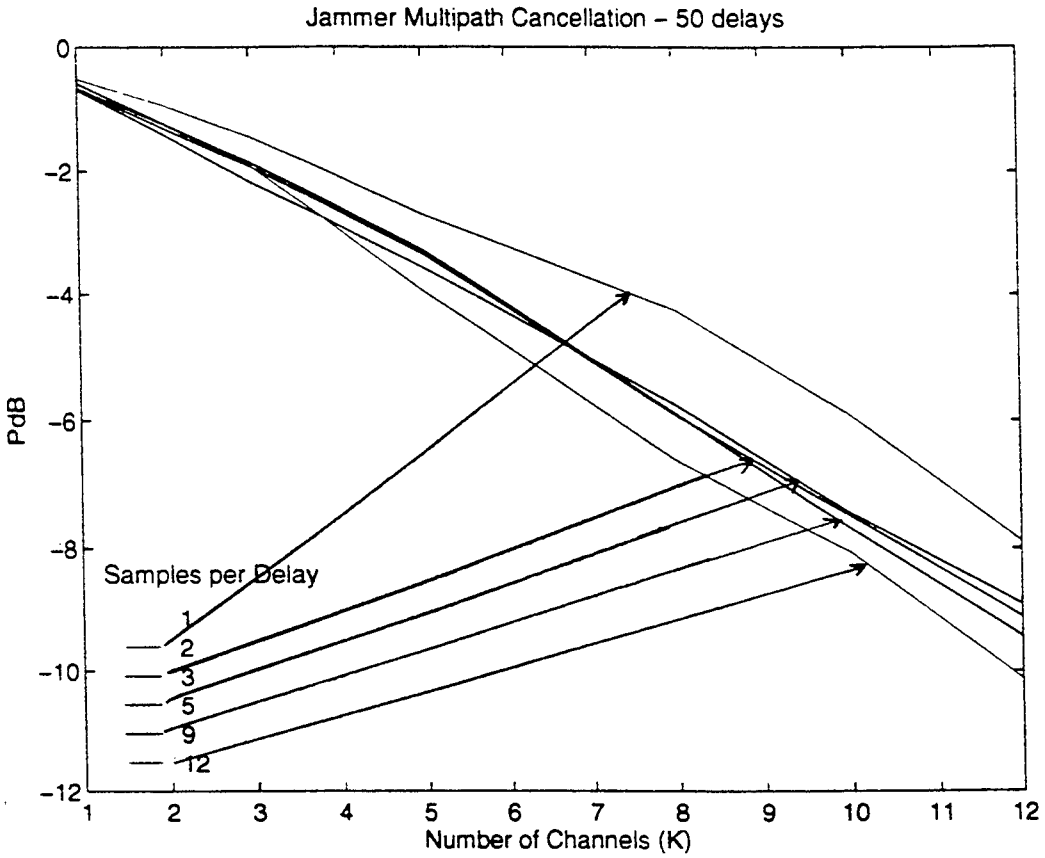

Jammer Multipath Cancellation – 50 delays
PdB
0
-2
-4
-6
-8
-10
-12
1
2
3
4
5
6
7
8
9
10
11
12
Number of Channels (K)
Samples per Delay
1
2
3
5
9
12

Figure 12

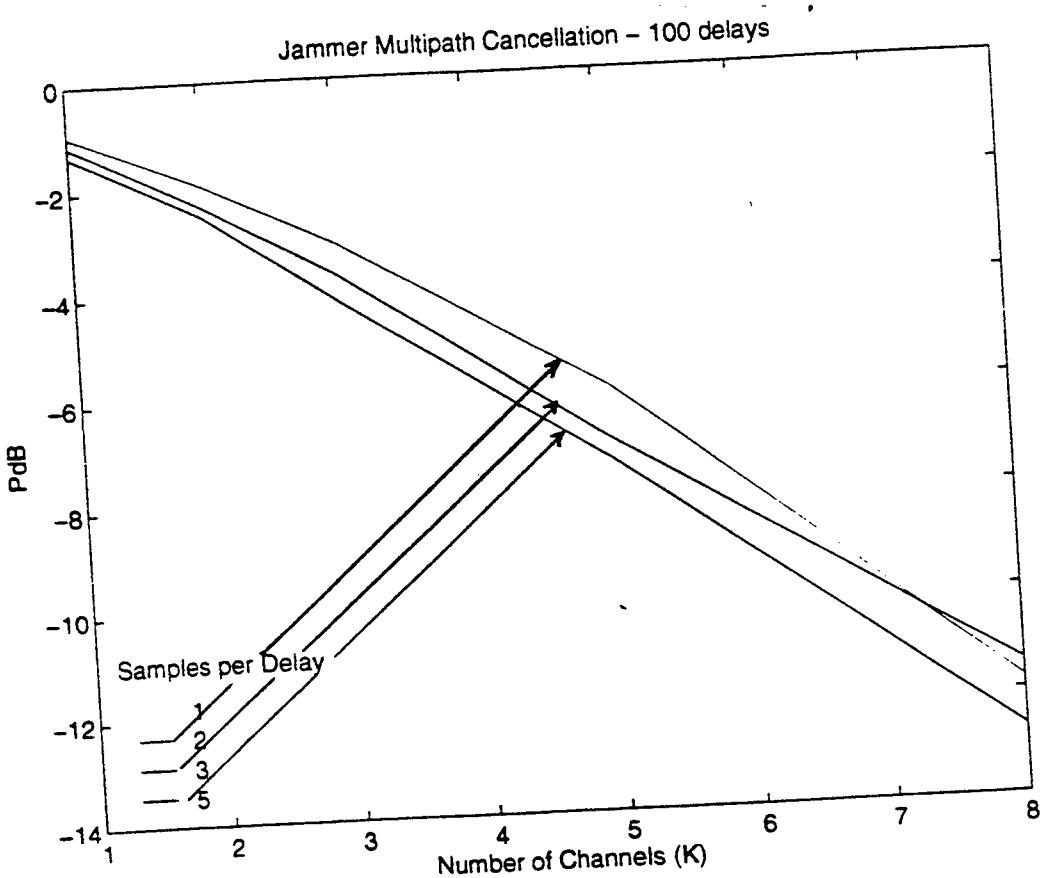

Jammer Multipath Cancellation – 100 delays
PdB
0
-2
-4
-6
-8
-10
-12
-14
1
2
3
4
5
6
7
8
Number of Channels (K)
Samples per Delay
1
2
3
5

Figure 13

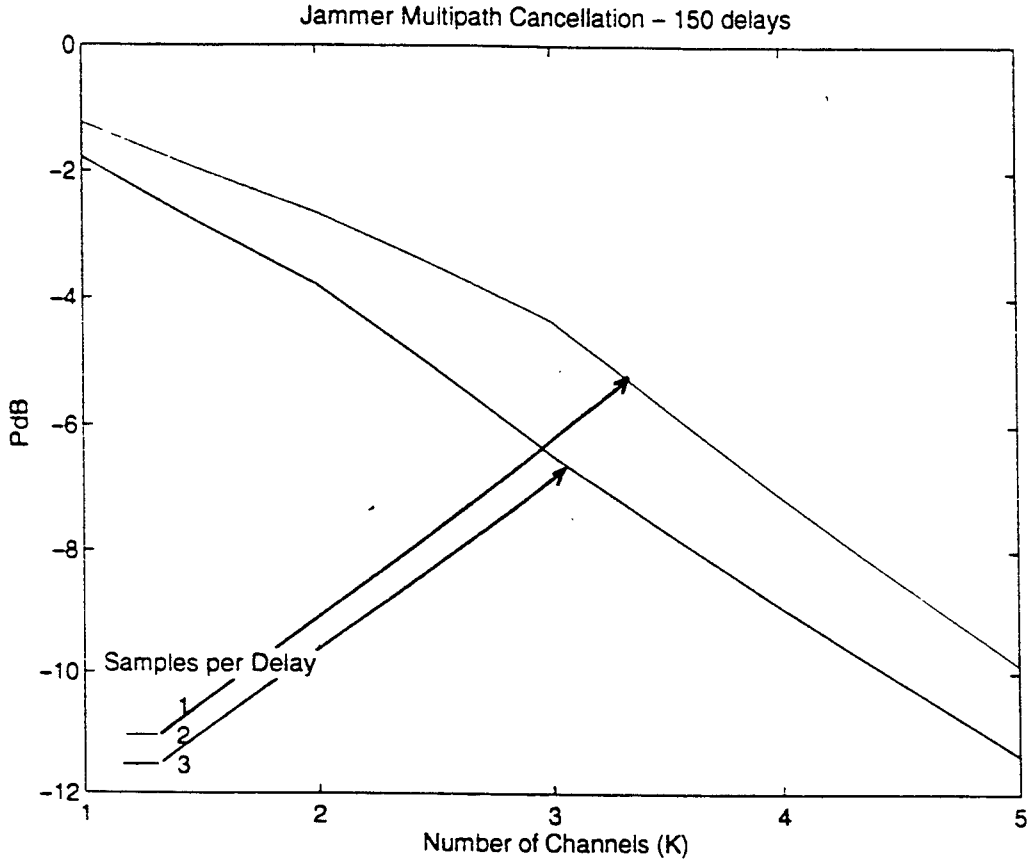

Jammer Multipath Cancellation – 150 delays
0
-2
-4
-6
-8
-10
-12
PdB
Samples per Delay
1
2
3
1
2
3
4
5
Number of Channels (K)

Figure 14

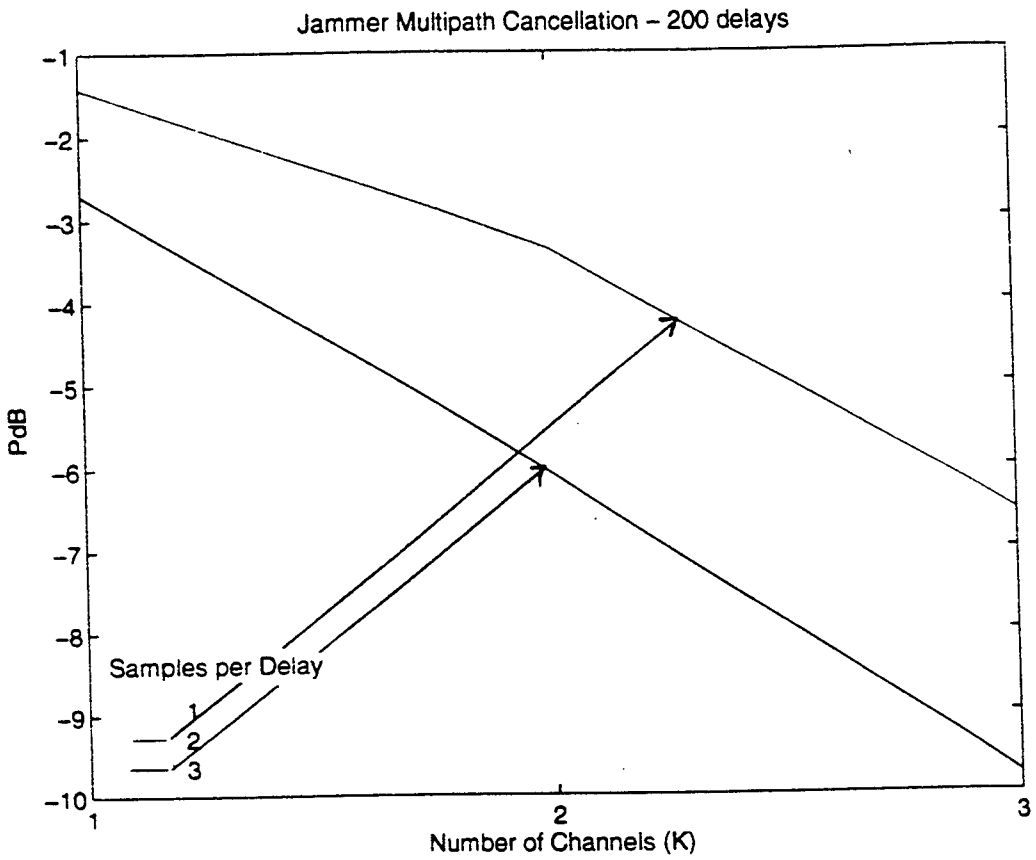

Jammer Multipath Cancellation – 200 delays
PdB
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
Samples per Delay
1
2
3
1
2
3
Number of Channels (K)

Figure 15

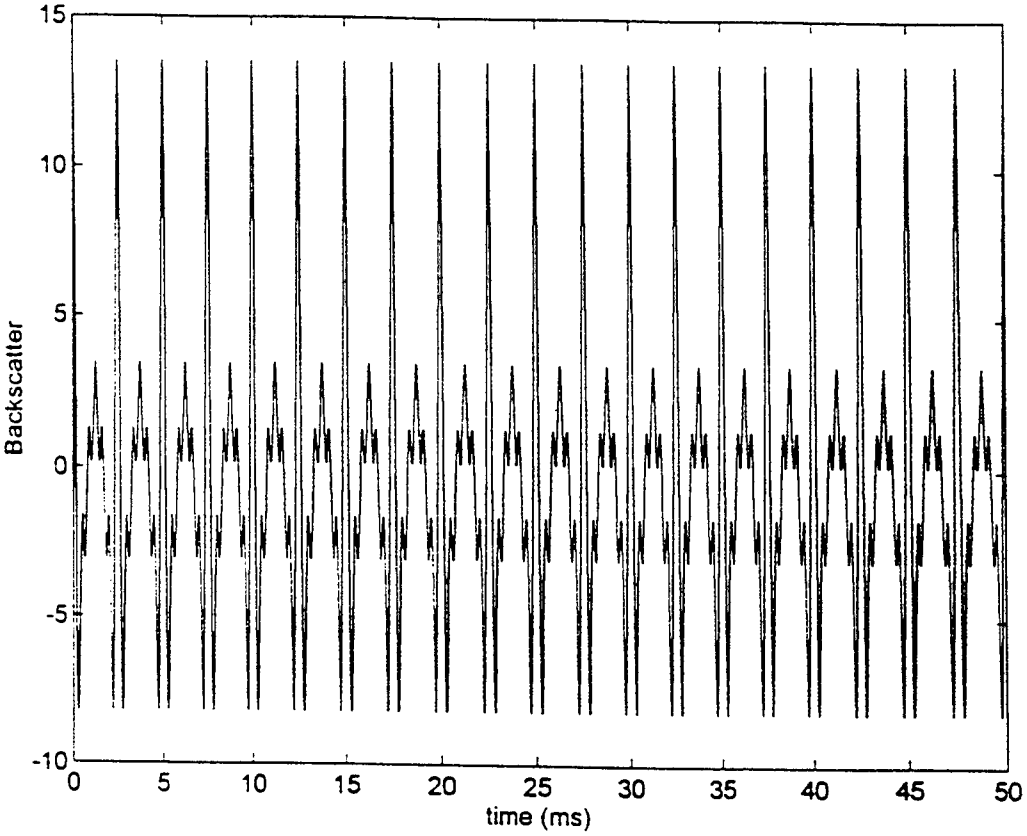


Figure 16

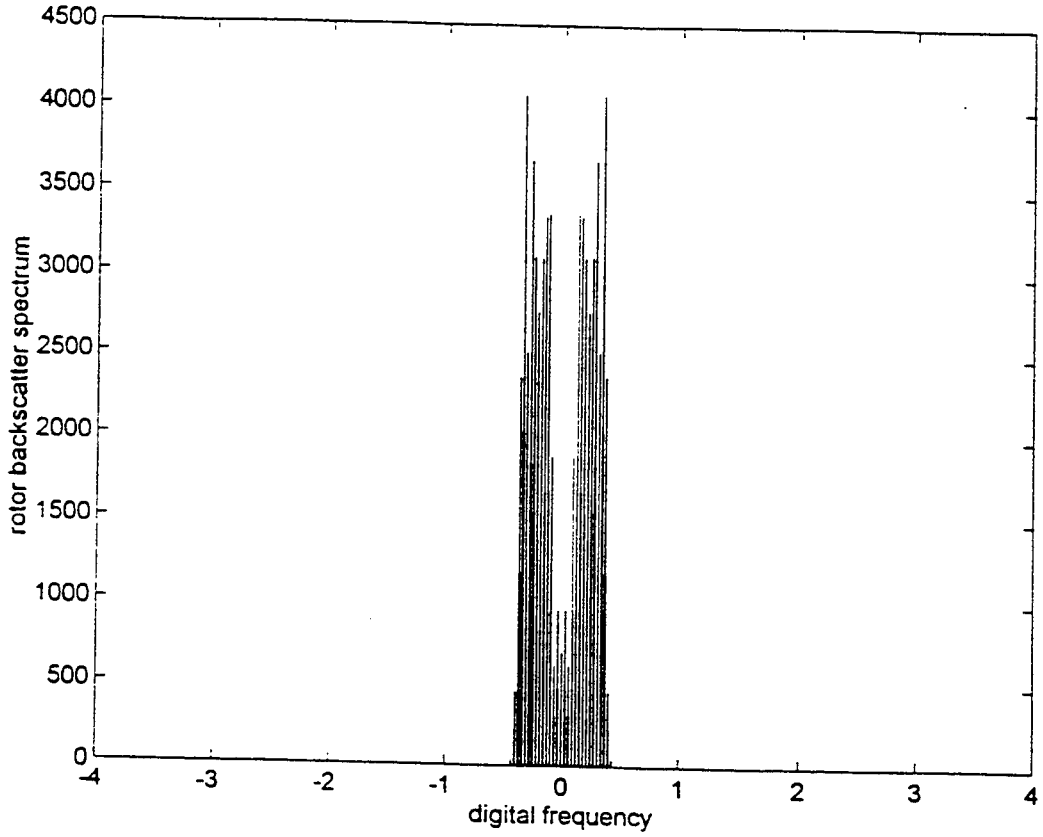

4500
4000
3500
3000
2500
2000
1500
1000
500
0
rotor backscatter spectrum
-4
-3
-2
-1
0
1
2
3
4
digital frequency

Figure 17

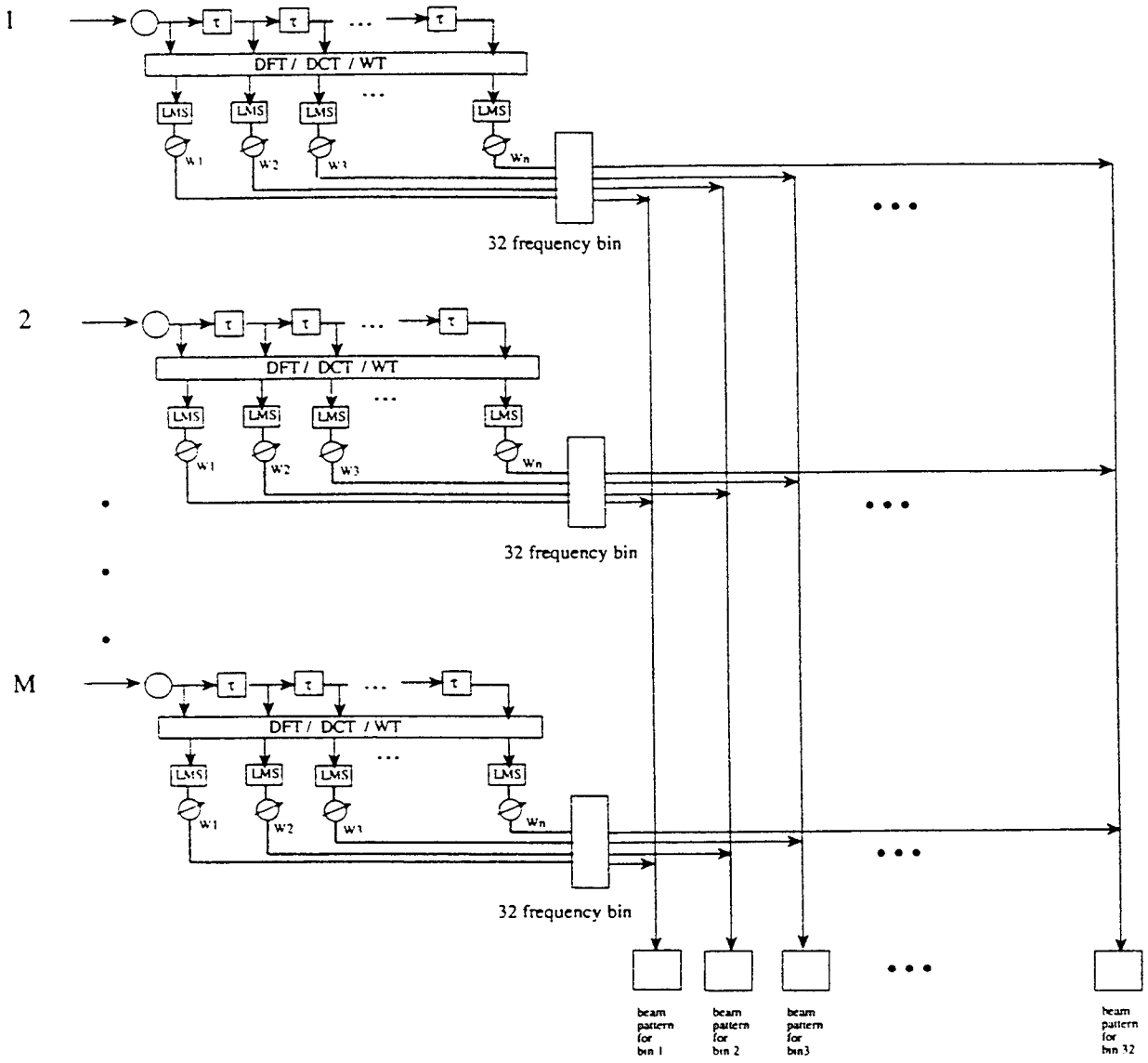

1
2
M
τ
DFT / DCT / WT
LMS
W1
W2
W3
Wn
32 frequency bin
beam pattern for bin 1
beam pattern for bin 2
beam pattern for bin3
beam pattern for bin 32

Figure 18

USING APES FOR INTERFEROMETRIC SAR IMAGING

Jian Li
Assistant Professor
Department of Electrical Engineering

405 CSE, P. O. Box 116130
University of Florida
Gainesville, FL 32611

Final Report for:
Summer Research Extension Program

Sponsored by
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.
and
University of Florida

November 1995

USING APES FOR INTERFEROMETRIC SAR IMAGING

Jian Li

Assistant Professor

Department of Electrical Engineering

University of Florida

Abstract

We present an adaptive FIR filtering approach, which is referred to as the APES (Amplitude and Phase Estimation of a Sinusoid), for interferometric SAR imaging. We compare the APES algorithm with other FIR filtering approaches including the Capon and FFT methods. We show via both numerical and experimental examples that the adaptive FIR filtering approaches such as Capon and APES can yield more accurate spectral estimates with much lower sidelobes and narrower spectral peaks than the FFT method. We show that although the APES algorithm yields somewhat wider spectral peaks than the Capon method, the former gives more accurate overall spectral estimates and SAR images than the latter and the FFT method.

USING APES FOR INTERFEROMETRIC SAR IMAGING

Jian Li

I. Introduction

In this paper, we show how the APES (Amplitude and Phase Estimation of a Sinusoid with known frequency in unknown colored noise) algorithm [1] can be used for interferometric SAR imaging. We compare the APES algorithm with other approaches including Capon [2] and FFT methods. We show by means of both numerical and experimental examples that the APES approach can yield more accurate spectral estimates with much lower sidelobes and more narrow spectral peaks than the FFT (fast Fourier transform) method. We show that although the APES algorithm gives slightly wider spectral peaks than the Capon method, the former yields more accurate overall spectral estimates than the latter and the FFT method.

In Section 2, we formulate the problem of interest. In Section 3, we describe how the APES algorithm is used for interferometric SAR imaging and compare it with other approaches including the FFT and Capon methods. In Section 4, we present both numerical and experimental examples showing the performance of the APES algorithm when used for interferometric SAR imaging. Finally, Section 5 contains our conclusions.

II. Problem Formulation

In interferometric SAR systems, we have two vertically displaced apertures to collect registered and phase coherent signals. For one-dimensional sequences, at a frequency ω (which is proportional to the range) of interest, we model the data sequences collected by the two apertures as

$$\left.\begin{aligned} z_{1_n} &= \alpha(\omega)e^{jn\omega} + e_{1_n}(\omega), \\ z_{2_n} &= \alpha(\omega)e^{j\gamma(\omega)}e^{jn\omega} + e_{2_n}(\omega), \end{aligned}\right\} \tag{1}$$

where $e_{1_n}(\omega)$ and $e_{2_n}(\omega)$ denote the unmodeled noise and interference at frequency ω, $\alpha(\omega)$ is the complex amplitude at ω and is proportional to the radar cross section (RCS) of the scatterer at ω, and $\gamma(\omega)$ is the

phase difference between the two channels at ω and is proportional to the height of the scatterer. The problem of interest herein is to estimate $\alpha(\omega)$ and $\gamma(\omega)$ from $\{z_{1_n}\}_{n=0}^{N-1}$ and $\{z_{2_n}\}_{n=0}^{N-1}$ for all ω of interest.

In a similar way, we can model the two-dimensional data matrices $\{z_{1_{n_1,n_2}}\}$ and $\{z_{2_{n_1,n_2}}\}$, $n_1 = 0,1,\cdots,N_1-1$, $n_2 = 0,1,\cdots,N_2-1$, collected by the two apertures and at a frequency pair (ω_1,ω_2) (which are proportional to the range and cross-range, respectively) of interest as

$$\left.\begin{aligned} z_{1_{n_1,n_2}} &= \alpha(\omega_1,\omega_2)e^{j(n_1\omega_1+n_2\omega_2)} + e_{1_{n_1,n_2}}(\omega_1,\omega_2), \\ z_{2_{n_1,n_2}} &= \alpha(\omega_1,\omega_2)e^{j\gamma(\omega_1,\omega_2)}e^{j(n_1\omega_1+n_2\omega_2)} + e_{2_{n_1,n_2}}(\omega_1,\omega_2), \end{aligned}\right\} \tag{2}$$

where $\alpha(\omega_1,\omega_2)$ and $\gamma(\omega_1,\omega_2)$ are proportional to the RCS and height of the scatterer at (ω_1,ω_2). The unmodeled noise and interference at frequency (ω_1,ω_2) is denoted by $e_{1_{n_1,n_2}}(\omega_1,\omega_2)$ and $e_{2_{n_1,n_2}}(\omega_1,\omega_2)$. The problem now is to determine $\alpha(\omega_1,\omega_2)$ and $\gamma(\omega_1,\omega_2)$ from the 2-D data for all (ω_1,ω_2) of interest.

A simple method of estimating $\alpha(\omega)$ and $\gamma(\omega)$ or $\alpha(\omega_1,\omega_2)$ and $\gamma(\omega_1,\omega_2)$ is to use 1-D or 2-D FFT (Fast Fourier Transform), which is computationally efficient. For 1-D data sequences, for example, let $\mathcal{Z}_1(\omega)$ and $\mathcal{Z}_2(\omega)$ denote the normalized Fourier transforms of $\{z_{1_n}\}$ and $\{z_{2_n}\}$ at frequency ω, i.e., the Fourier transforms of $\{z_{1_n}\}$ and $\{z_{2_n}\}$ at frequency ω divided by N. Then the FFT estimates $\hat{\gamma}(\omega)$ and $\hat{\alpha}(\omega)$, of $\gamma(\omega)$ and $\alpha(\omega)$, respectively, are given by

$$\hat{\gamma}(\omega) = \text{phase of} \quad [\mathcal{Z}_1^*(\omega)\mathcal{Z}_2(\omega)]\,, \tag{3}$$

and

$$\hat{\alpha}(\omega) = \frac{1}{2}\left[\mathcal{Z}_1(\omega) + \mathcal{Z}_2(\omega)e^{-j\hat{\gamma}(\omega)}\right], \tag{4}$$

where $(\cdot)^*$ denotes the complex conjugate. If the phase of $\alpha(\omega)$ is not of interest, $|\hat{\alpha}(\omega)|^2$ may be obtained by

$$|\hat{\alpha}(\omega)|^2 = |\mathcal{Z}_1^*(\omega)\mathcal{Z}_2(\omega)|\,. \tag{5}$$

(It appears that $|\hat{\alpha}(\omega)|$ obtained from Equation (4) will be more accurate as it is the result of averaging more data.) However, Fourier transform methods are known to suffer from high sidelobe effects and poor accuracy. Many different types of windows can be applied to the data before FFT processing to reduce the sidelobes. This is achieved, however, at the cost of widening the estimated spectral peaks and hence worsening the resolution.

We show below how the APES algorithm [1] can be used for the estimation of $\alpha(\omega)$ and $\gamma(\omega)$ or $\alpha(\omega_1,\omega_2)$ and $\gamma(\omega_1,\omega_2)$. The APES algorithm is an adaptive FIR filtering approach that yields significantly reduced sidelobes and narrower spectral peaks than the FFT-based methods.

III. The APES Algorithm for Interferometric SAR Imaging

We present our results in detail for 1-D data sequences and then extend them to the 2-D case. We begin by briefly describing the use of FIR filters for the estimation of $\alpha(\omega)$ and $\gamma(\omega)$ or $\alpha(\omega_1,\omega_2)$ and $\gamma(\omega_1,\omega_2)$. We next present the specific adaptive FIR filter used in the APES algorithm. Finally, we compare the APES algorithm with the Capon [2] and FFT approaches.

A. Interferometric SAR Imaging with FIR Filters

We first consider 1-D data sequences. Let $\mathbf{h}(\omega)$ denote the impulse response of an M-tap FIR filter, where

$$\mathbf{h}(\omega) = \left[\begin{array}{cccc} h_1(\omega) & h_2(\omega) & \cdots & h_M(\omega) \end{array} \right]^T, \tag{6}$$

with $(\cdot)^T$ denoting the transpose. The dependence of ((6)) on ω is introduced for later use; it signifies the fact that ((6)) is a narrowband filter with center frequency equal to ω. Let $\mathbf{z}_1$ and $\mathbf{z}_2$ denote the data vectors from the two channels. Let $\tilde{\mathbf{z}}_1$ and $\tilde{\mathbf{z}}_2$ denote the data vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ complex conjugated and in backward order. Then we have

$$\left.\begin{array}{l} \mathbf{z}_1 = \left[\begin{array}{cccc} z_{1_0} & z_{1_1} & \cdots & z_{1_{N-1}} \end{array} \right]^T, \\ \tilde{\mathbf{z}}_1 = \left[\begin{array}{cccc} z^*_{1_{N-1}} & z^*_{1_{N-2}} & \cdots & z^*_{1_0} \end{array} \right]^T, \\ \mathbf{z}_2 = \left[\begin{array}{cccc} z_{2_0} & z_{2_1} & \cdots & z_{2_{N-1}} \end{array} \right]^T, \\ \tilde{\mathbf{z}}_2 = \left[\begin{array}{cccc} z^*_{2_{N-1}} & z^*_{2_{N-2}} & \cdots & z^*_{2_0} \end{array} \right]^T. \end{array}\right\} \tag{7}$$

We make use of combined forward and backward approaches as they often yield better estimates as compared to forward only approaches [3]. Let

$$\left.\begin{aligned}
\bar{\mathbf{z}}_{1_i} &= \begin{bmatrix} z_{1_i} & z_{1_{i+1}} & \cdots & z_{1_{i+M-1}} \end{bmatrix}^T, \\
\bar{\bar{\mathbf{z}}}_{1_i} &= \begin{bmatrix} \tilde{z}_{1_i} & \tilde{z}_{1_{i+1}} & \cdots & \tilde{z}_{1_{i+M-1}} \end{bmatrix}^T, \\
\bar{\mathbf{z}}_{2_i} &= \begin{bmatrix} z_{2_i} & z_{2_{i+1}} & \cdots & z_{2_{i+M-1}} \end{bmatrix}^T, \\
\bar{\bar{\mathbf{z}}}_{2_i} &= \begin{bmatrix} \tilde{z}_{2_i} & \tilde{z}_{2_{i+1}} & \cdots & \tilde{z}_{2_{i+M-1}} \end{bmatrix}^T,
\end{aligned}\right\} \quad i = 0, 1, \cdots, N-M, \tag{8}$$

be the overlapping subvectors of the data vectors in ((7)). From Equation ((7)), we note that the output samples obtained by passing $\mathbf{z}_1$ and $\mathbf{z}_2$, through the FIR filter $\mathbf{h}(\omega)$ are given by

$$\left.\begin{aligned}
\mathbf{h}^H(\omega)\bar{\mathbf{z}}_{1_i} &= \alpha(\omega)\left[\mathbf{h}^H(\omega)\mathbf{a}(\omega)\right] e^{ji\omega} + \bar{w}_{1_i}(\omega), \\
\mathbf{h}^H(\omega)\bar{\mathbf{z}}_{2_i} &= \alpha(\omega)e^{j\gamma(\omega)}\left[\mathbf{h}^H(\omega)\mathbf{a}(\omega)\right] e^{ji\omega} + \bar{w}_{2_i}(\omega),
\end{aligned}\right\} i = 0, 1, \cdots, N-M, \tag{9}$$

where $(\cdot)^H$ denotes the complex conjugate transpose, $\bar{w}_{1_i}(\omega)$ and $\bar{w}_{2_i}(\omega)$ denote the perturbations at the filter outputs, and

$$\mathbf{a}(\omega) = \begin{bmatrix} 1 & e^{j\omega} & \cdots & e^{j(M-1)\omega} \end{bmatrix}^T, \tag{10}$$

For convenience of what follows, we normalize $\mathbf{h}(\omega)$ so that

$$\mathbf{h}^H(\omega)\mathbf{a}(\omega) = 1. \tag{11}$$

Thus, using the first equation in ((9)), we obtain the intermediate least squares estimate of $\alpha(\omega)$ [1] as

$$\hat{\alpha}_1(\omega) = \frac{1}{(N-M+1)}\left[\mathbf{h}^H(\omega)\left(\sum_{i=0}^{N-M} \bar{\mathbf{z}}_{1_i} \exp\{-ji\omega\}\right)\right]. \tag{12}$$

Let $\bar{\mathbf{Z}}_1(\omega)$ denote the normalized row-wise Fourier transform of $\{\bar{\mathbf{z}}_{1_i}\}_{i=0}^{N-M}$, i.e., the Fourier transform of $\{\bar{\mathbf{z}}_{1_i}\}_{i=0}^{N-M}$ divided by $(N-M+1)$, which can be computed efficiently via FFT (note that padding with zeros is often necessary). Then we have

$$\hat{\alpha}_1(\omega) = \mathbf{h}^H(\omega)\bar{\mathbf{Z}}_1(\omega). \tag{13}$$

Similarly, using the second equation in ((9)), we obtain the intermediate least squares estimate of $\alpha(\omega)e^{j\gamma(\omega)}$ as

$$\hat{\alpha}_2(\omega) = \mathbf{h}^H(\omega)\bar{\boldsymbol{z}}_2(\omega). \tag{14}$$

where $\bar{\boldsymbol{z}}_2(\omega)$ denotes the normalized row-wise Fourier transform of $\{\bar{z}_{2_i}\}_{i=0}^{N-M}$. Thus the estimates of $\gamma(\omega)$ and $\alpha(\omega)$ can be obtained, respectively, from Equations ((13)) and ((14)) as

$$\hat{\gamma}(\omega) = \text{phase of } \left[\hat{\alpha}_1^*(\omega)\hat{\alpha}_2(\omega)\right], \tag{15}$$

and

$$\hat{\alpha}(\omega) = \frac{1}{2}\left[\hat{\alpha}_1(\omega) + \hat{\alpha}_2(\omega)e^{-j\hat{\gamma}(\omega)}\right], \tag{16}$$

We remark that the FIR filtering approach based on ((16)) and ((15)) can be used to compute $\hat{\alpha}(\omega = \omega_j)$ and $\hat{\gamma}(\omega = \omega_j)$ in parallel for all ω_j of interest. We also remark that when $M = 1$, $\mathbf{h}(\omega)$ becomes a scalar, (in fact $\mathbf{h}(\omega) = 1$, owing to ((11))), and the FIR filtering approaches become the same as the FFT (without windowing) approach (comparing Equations ((15)) and ((16)) with Equations ((3)) and ((4)), respectively) described at the end of Section 2. For $N \gg M$, the FIR filtering approach approximately reduces to FFT. Therefore, significant differences between FFT and the FIR filtering approaches occur only when M is sufficiently large as compared to N.

B. The APES Algorithm

The adaptive FIR filter $\mathbf{h}_{\text{APES}}(\omega)$ corresponding to the APES algorithm is given in Appendix A of [1]. Since now the data sequences are collected by two apertures, we modify $\mathbf{h}_{\text{APES}}(\omega)$ as follows. Let

$$\hat{\mathbf{R}} = \frac{1}{4}\left(\hat{\bar{\mathbf{R}}}_1 + \hat{\bar{\mathbf{R}}}_2 + \hat{\bar{\bar{\mathbf{R}}}}_1 + \hat{\bar{\bar{\mathbf{R}}}}_2\right), \tag{17}$$

where

$$\hat{\bar{\mathbf{R}}}_1 = \frac{1}{N-M+1}\sum_{i=0}^{N-M} \bar{\mathbf{z}}_{1_i}\bar{\mathbf{z}}_{1_i}^H, \tag{18}$$

and $\hat{\bar{\mathbf{R}}}_2$, $\hat{\bar{\bar{\mathbf{R}}}}_1$, and $\hat{\bar{\bar{\mathbf{R}}}}_2$ are obtained similarly from $\bar{\mathbf{z}}_{2_i}$, $\bar{\bar{\mathbf{z}}}_{1_i}$, and $\bar{\bar{\mathbf{z}}}_{2_i}$, respectively. Let

$$\boldsymbol{\mathcal{Z}}(\omega) = \frac{1}{2}\left[\begin{array}{cccc} \bar{\boldsymbol{z}}_1(\omega) & \bar{\bar{\boldsymbol{z}}}_1(\omega) & \bar{\boldsymbol{z}}_2(\omega) & \bar{\bar{\boldsymbol{z}}}_2(\omega) \end{array}\right], \tag{19}$$

where $\bar{\bar{\mathcal{Z}}}_1(\omega)$ and $\bar{\bar{\mathcal{Z}}}_2(\omega)$ denote the normalized Fourier transforms of $\{\bar{\bar{\mathbf{z}}}_{1,i}\}_{i=0}^{N-M}$ and $\{\bar{\bar{\mathbf{z}}}_{2,i}\}_{i=0}^{N-M}$, respectively. Let

$$\hat{\mathbf{Q}}(\omega) = \hat{\mathbf{R}} - \mathcal{Z}(\omega)\mathcal{Z}^H(\omega). \tag{20}$$

Then $\mathbf{h}_{\mathrm{APES}}(\omega)$ can be written as

$$\mathbf{h}_{\mathrm{APES}}(\omega) = \frac{\hat{\mathbf{Q}}^{-1}(\omega)\mathbf{a}(\omega)}{\mathbf{a}^H(\omega)\hat{\mathbf{Q}}^{-1}(\omega)\mathbf{a}(\omega)}. \tag{21}$$

Note that $\mathbf{h}_{\mathrm{APES}}(\omega)$ satisfies the constraint in ((11)). We also note that since only $\mathcal{Z}(\omega)$ in the right-hand side of ((20)) depends on ω, the $\hat{\mathbf{Q}}^{-1}(\omega)$ in ((21)) can be computed efficiently by using the matrix inversion lemma:

$$\hat{\mathbf{Q}}^{-1}(\omega) = \hat{\mathbf{R}}^{-1} - \hat{\mathbf{R}}^{-1}\mathcal{Z}(\omega)\left[\mathcal{Z}^H(\omega)\hat{\mathbf{R}}^{-1}\mathcal{Z}(\omega) - \mathbf{I}\right]^{-1}\mathcal{Z}^H(\omega)\hat{\mathbf{R}}^{-1}, \tag{22}$$

where $\mathbf{I}$ denotes the 4×4 identity matrix.

It is readily verified that too large an M can cause $\hat{\mathbf{Q}}(\omega)$ in ((20)) to be singular. In [1] it has been shown that $M = N/2$ gives the best spectral estimates, i.e., estimates with good accuracy, low sidelobes, and narrow spectral peaks for peaky spectra. For our case, since we have data from two channels to average, we find that the best estimates are obtained when $M \approx 2(N - M + 1)$, that is $M \approx \frac{2}{3}(N + 1)$.

Next, we compare the APES algorithm with the Capon approach. When $\hat{\mathbf{Q}}^{-1}(\omega)$ in ((21)) is replaced by $\hat{\mathbf{R}}^{-1}$, we obtain the Capon approach [2], i.e., $\mathbf{h}_{\mathrm{APES}}(\omega)$ with the previous replacement becomes $\mathbf{h}_{\mathrm{Capon}}(\omega)$. Note that both $\mathbf{h}_{\mathrm{APES}}(\omega)$ and $\mathbf{h}_{\mathrm{Capon}}(\omega)$ are data dependent, and hence APES and Capon are so-called "adaptive methods".

It can be seen from Appendix A of [1] that $\hat{\mathbf{Q}}(\omega)$ is an estimate of the covariance matrix of the noise and interference in $\bar{\mathbf{z}}_i$ and $\bar{\mathbf{z}}_i$. Thus $\mathbf{h}_{\mathrm{APES}}(\omega)$ is a *matched* filter, while $\mathbf{h}_{\mathrm{Capon}}(\omega)$ is *not*. This may partially explain the poorer performance of Capon, as compared with APES, observed in the examples given in the next section.

C. Extensions

The 2-D complex spectral estimation by means of a general 2-D FIR filtering approach is briefly explained in what follows. Let $\mathbf{H}(\omega_1, \omega_2)$ denote an $M_1 \times M_2$ 2-D impulse response whose (m_1, m_2)th element is $h_{m_1,m_2}(\omega_1, \omega_2)$. Once again, we stress by means of notation the fact that the time-domain response $\mathbf{H}(\omega_1, \omega_2)$ corresponds to a narrowband filter which is to be associated, by design, with a certain 2-D center frequency

(ω_1, ω_2). Let

$$\mathbf{h}(\omega_1, \omega_2) = \text{vec}[\mathbf{H}(\omega_1, \omega_2)], \tag{23}$$

where vec$(\cdot)$ denotes the operation consisting of stacking the columns of a matrix on top of each other. Let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ denote the 2-D data matrices whose (n_1, n_2)th elements are given by $z_{1_{n_1,n_2}}$ and $z_{2_{n_1,n_2}}$, respectively. Let $\bar{\mathbf{Z}}_{1_i}$ and $\bar{\mathbf{Z}}_{2_i}$, $i = 0, 1, \cdots, (N_1 - M_1 + 1)(N_2 - M_2 + 1) - 1$, denote the $M_1 \times M_2$ overlapping submatrices of $\mathbf{Z}_1$ and $\mathbf{Z}_2$, respectively. Let $\tilde{\mathbf{Z}}_1$ and $\tilde{\mathbf{Z}}_2$ denote the complex conjugate of the data matrix $\mathbf{Z}_1$ and $\mathbf{Z}_2$, with its rows and columns in backward order, respectively. Let $\bar{\tilde{\mathbf{Z}}}_{1_i}$ and $\bar{\tilde{\mathbf{Z}}}_{2_i}$ be formed from $\tilde{\mathbf{Z}}_1$ and $\tilde{\mathbf{Z}}_2$, respectively, in the same way as $\bar{\mathbf{Z}}_{1_i}$ and $\bar{\mathbf{Z}}_{2_i}$ are formed from $\mathbf{Z}_1$ and $\mathbf{Z}_2$, respectively. Let $\bar{\boldsymbol{\mathcal{Z}}}_1(\omega_1, \omega_2)$ be a vector obtained by stacking the columns of the 2-D normalized Fourier transform of $\{\bar{\mathbf{Z}}_{1_i}\}$, $i = 0, 1, \cdots, (N_1 - M_1 + 1)(N_2 - M_2 + 1) - 1$, i.e., the Fourier transform of $\{\bar{\mathbf{Z}}_{1_i}\}$ divided by $(N_1 - M_1 + 1)(N_2 - M_2 + 1)$. Similarly, let $\bar{\boldsymbol{\mathcal{Z}}}_2(\omega_1, \omega_2)$ be obtained from $\{\bar{\mathbf{Z}}_{2_i}\}$, $i = 0, 1, \cdots, (N_1 - M_1 + 1)(N_2 - M_2 + 1) - 1$, respectively. Let

$$\mathbf{a}(\omega_1, \omega_2) \quad = \quad \mathbf{a}_1(\omega_1) \otimes \mathbf{a}_2(\omega_2), \tag{24}$$

where $\otimes$ denotes the Kronecker matrix product and

$$\mathbf{a}_i(\omega_i) = \begin{bmatrix} 1 & e^{j\omega_i} & \cdots & e^{j(M_i-1)\omega_i} \end{bmatrix}^T, \quad i = 1, 2. \tag{25}$$

Then the least squares estimates of $\gamma(\omega_1, \omega_2)$ and $\alpha(\omega_1, \omega_2)$ are obtained as

$$\hat{\gamma}(\omega_1, \omega_2) = \text{phase of } \left[\hat{\alpha}_1^*(\omega_1, \omega_2)\hat{\alpha}_2(\omega_1, \omega_2)\right], \tag{26}$$

and

$$\hat{\alpha}(\omega_1, \omega_2) = \frac{1}{2}\left[\hat{\alpha}_1(\omega_1, \omega_2) + \hat{\alpha}_2(\omega_1, \omega_2)e^{-j\hat{\gamma}(\omega_1,\omega_2)}\right], \tag{27}$$

where

$$\hat{\alpha}_1(\omega_1, \omega_2) = \mathbf{h}^H(\omega_1, \omega_2)\bar{\boldsymbol{\mathcal{Z}}}_1(\omega_1, \omega_2) \tag{28}$$

and

$$\hat{\alpha}_2(\omega_1, \omega_2) = \mathbf{h}^H(\omega_1, \omega_2)\bar{\boldsymbol{\mathcal{Z}}}_2(\omega_1, \omega_2) \tag{29}$$

To derive the 2-D FIR filter $\mathbf{h}_{\text{APES}}(\omega_1, \omega_2)$ used by the APES algorithm, let $\hat{\bar{\mathbf{R}}}_1$, $\hat{\bar{\mathbf{R}}}_2$, $\hat{\bar{\tilde{\mathbf{R}}}}_1$ and $\hat{\bar{\tilde{\mathbf{R}}}}_2$ denote the sample covariance matrices associated with vec $\left[\bar{\mathbf{Z}}_{1_i}\right]$, vec $\left[\bar{\mathbf{Z}}_{2_i}\right]$, vec $\left[\bar{\tilde{\mathbf{Z}}}_{1_i}\right]$ and vec $\left[\bar{\tilde{\mathbf{Z}}}_{2_i}\right]$, respectively.

Let $\hat{\mathbf{R}}$ denote the average of $\hat{\bar{\mathbf{R}}}_1$, $\hat{\bar{\mathbf{R}}}_2$, $\hat{\bar{\bar{\mathbf{R}}}}_1$ and $\hat{\bar{\bar{\mathbf{R}}}}_2$ as in ((17)), and let

$$\mathcal{Z}(\omega_1,\omega_2) = \frac{1}{2}\left[\begin{array}{cccc} \bar{\mathcal{Z}}_1(\omega_1,\omega_2) & \bar{\bar{\mathcal{Z}}}_1(\omega_1,\omega_2) & \bar{\mathcal{Z}}_2(\omega_1,\omega_2) & \bar{\bar{\mathcal{Z}}}_2(\omega_1,\omega_2) \end{array}\right], \tag{30}$$

where $\bar{\bar{\mathcal{Z}}}_1(\omega_1,\omega_2)$ and $\bar{\bar{\mathcal{Z}}}_2(\omega_1,\omega_2)$ denote the vectors obtained by stacking the columns of the 2-D normalized Fourier transforms of $\{\bar{\bar{\mathbf{Z}}}_{1,i}\}$ and $\{\bar{\bar{\mathbf{Z}}}_{2,i}\}$, $i = 0, 1, \cdots, (N_1 - M_1 + 1)(N_2 - M_2 + 1) - 1$, respectively. Then $\mathbf{h}_{\mathrm{APES}}(\omega_1,\omega_2)$ is given by

$$\mathbf{h}_{\mathrm{APES}}(\omega_1,\omega_2) = \frac{\hat{\mathbf{Q}}^{-1}(\omega_1,\omega_2)\mathbf{a}(\omega_1,\omega_2)}{\mathbf{a}^H(\omega_1,\omega_2)\hat{\mathbf{Q}}^{-1}(\omega_1,\omega_2)\mathbf{a}(\omega_1,\omega_2)}, \tag{31}$$

where

$$\hat{\mathbf{Q}}^{-1}(\omega_1,\omega_2) = \hat{\mathbf{R}}^{-1} - \hat{\mathbf{R}}^{-1}\mathcal{Z}(\omega_1,\omega_2)\left[\mathcal{Z}^H(\omega_1,\omega_2)\hat{\mathbf{R}}^{-1}\mathcal{Z}(\omega_1,\omega_2) - \mathbf{I}\right]^{-1}\mathcal{Z}^H(\omega_1,\omega_2)\hat{\mathbf{R}}^{-1}. \tag{32}$$

We remark that for too large M_1 and M_2, the matrix $\hat{\mathbf{Q}}(\omega_1,\omega_2)$ may be singular. In general, using $M_1 \approx \frac{2}{3}(N_1 + 1)$ and $M_2 \approx \frac{2}{3}(N_2 + 1)$ in the 2-D APES algorithm appears to yield the best spectral estimates.

Since the APES algorithm for the two dimensional case requires the inversion of $(M_1M_2) \times (M_1M_2)$ matrix, it is computationally prohibitive to apply the algorithm directly to data matrices with large dimensions, as is the case in some of the examples that we shall consider in the next section. Instead, we proceed in a different way. We break up each frequency domain image (obtained by taking the 2-D FFT of the time domain data) into overlapping chips of size $N_1 \times N_2$ (N_1 and N_2 are smaller than the original data dimensions), as shown in Figure 1. We choose the overlap to be 50%. We take the 2-D inverse FFT (IFFT) of the frequency domain chips to obtain the time domain chips and apply the APES algorithm to the time domain chips. Since the APES algorithm can be used to compute $\hat{\alpha}(\omega_1,\omega_2)$ and $\hat{\gamma}(\omega_1,\omega_2)$ in parallel for each frequency pair (ω_1,ω_2) of interest, we evaluate $\hat{\alpha}(\omega_1,\omega_2)$ and $\hat{\gamma}(\omega_1,\omega_2)$ only over the frequencies given by $2\pi(0.25) \leq \omega_1 < 2\pi(0.75)$ and $2\pi(0.25) \leq \omega_2 < 2\pi(0.75)$. The results obtained by applying APES to the time domain chips are put together to form the final image. We will show with examples in the next section that this procedure does not produce the mosaicing or tiling effect, which is noticible for the images shown in [4].

IV. Numerical and Experimental Results

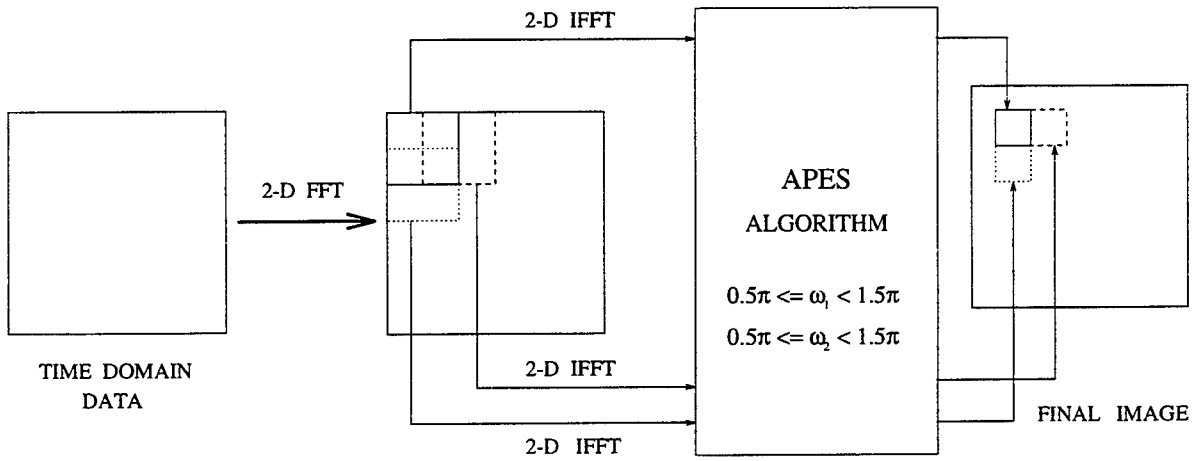


Figure 1: Applying APES to data matrix with large dimensions

We present both numerical and experimental examples showing the performance of the APES algorithm in interferometric SAR imaging. We compare the performance of APES with those of FFT, windowed FFT, and Capon methods.

A. 1-D Complex Spectral Estimation

We first consider estimating the complex amplitude and phase estimates for data sequences (as modeled in (1)) with $N = 64$ samples. The modulus of the true complex spectrum is shown in Figure 2(a). Note that the data consists of 4 dominant lines and many small randomly spaced scatterers. The true phase differences associated with the 4 dominant lines are shown in Figure 3(a). We are most interested in estimating the complex amplitudes and phase differences associated with the dominant lines. The data is corrupted by additive zero-mean white Gaussian noise with variance 1. Figure 2(b) shows the modulus of the complex amplitude estimate obtained with the FFT method. We note that the use of the FFT method results in high sidelobes, which makes it hard to distinguish the peak caused by the 4th dominant line in Figure 2(a) from the sidelobe peaks. Note also that the peak heights given by the FFT method are not very accurate estimates of the true ones shown in Figure 2(a). Figure 3(b) shows the phase difference estimates at the frequencies corresponding to the 4 dominant lines obtained with the FFT method. We see that the phase difference estimate for the 4th dominant line is quite different from the corresponding true value, and the estimates for the remaining lines are not very accurate either. Figure 2(c) shows the modulus of the complex amplitude estimate obtained by using FFT with Kaiser window and shape parameter 2. We note that the windowed FFT method reduces the sidelobes of the spectral peaks. However, the windowed FFT method widens the spectral peaks and as a result, the third and fourth peaks in Figure 2(c) are smeared together. Figure 3(c) shows the phase difference estimates obtained by the windowed FFT method. We observe that the estimates are no better than those obtained by the FFT method. Figure 2(d) shows the modulus of the complex amplitude estimate obtained by using the Capon method with $M = 16$. Although the amplitude estimates are accurate, the corresponding spectral peaks are not very narrow. As a result the 4th peak appears to be smeared with the 3rd peak. Figure 3(d) shows the phase difference estimates obtained by the Capon method with $M = 16$. We find that the phase difference estimates at the frequencies corresponding to the 4 dominant lines are fairly accurate. Figure 2(e) shows the modulus of the complex amplitude estimate obtained by using the Capon method with $M = 32$. Although the spectral peaks given by the

Capon method are very narrow, the corresponding estimated peak heights are not accurate. Figure 3(e) shows the phase difference estimates obtained by the Capon method with $M = 32$. We note that the phase difference estimates are not very accurate either. Figure 2(f) show the modulus of the complex amplitude estimate obtained by using the APES algorithm with $M = 42$. Note that using APES with $M = 42$ can better resolve the 4th spectral peak than using Capon with M =16, and gives more accurate estimates than using Capon with $M = 32$. Figure 3(f) shows the phase difference estimates obtained by using the APES algorithm with $M = 42$. We observe that the phase difference estimates given by the APES algorithm are the most accurate among all of the methods considered.

We now consider the effect of M, the length of the FIR filter, on the accuracy of the complex amplitude and phase difference estimates at frequencies that correspond to the four dominant lines in Figure 2(a). The empirical root mean-squared errors (RMSEs) of the estimates were obtained from 100 independent trials. Figure 4 shows the RMSEs of the complex amplitude estimates as a function of M. For some carefully picked values of M, such as $M = 8$ and $M = 16$, the RMSEs obtained with Capon are similar to those obtained with APES. Yet for $M = 16$, it can be seen from Figure 2(d) that the Capon method yields wider spectral peaks. Figure 5 shows the RMSEs of the phase difference estimates for the APES and Capon methods. We see that the curves are of a similar nature. The RMSEs of the estimates obtained with the APES algorithm are nearly the same for a wide range of values of M. Hence, in contrast to Capon, the selection of M in APES for accurate complex amplitude and phase difference estimation appears to be straightforward. Since $M = \frac{2}{3}(N + 1) \approx 42$ gives almost the narrowest spectral peaks and the most accurate complex amplitude and phase difference estimates, we recommend using $M = \frac{2}{3}(N + 1)$ with the APES algorithm.

B. 2-D Complex Spectral Estimation

We first consider estimating the mixed-spectrum for a 2-D data sequence with $N_1 = N_2 = 24$. The data sequence is corrupted by additive zero-mean white Gaussian noise with variance 20. The modulus (in dB and scaled to be between 0 and 255) of the true complex spectrum is shown in Figure 6(a). Note that the data consists of 3 spectral lines and two closely-spaced one-dimensional continuous pulses. (The line spectra in Figure 6(a) simulate corner reflectors and the 1-D continuous pulses simulate dihedrals in SAR images.) The true phase difference or height corresponding to the spectral lines and the one-dimensional pulses are shown in Figure 7(a). To give a meaningful representation to the phase difference estimates,

we set the phase difference estimates to be zero at those frequency points where the estimated modulus of complex amplitude is below a certain threshold, since those phase difference estimates are useless. For all of the methods considered, we set the threshold to be some percentage (obtained by trial and error so as to have the best result without loosing phase difference information at the frequencies of interest) of the maximum of the estimated complex amplitudes. Figure 6(b) shows the modulus of the complex amplitude estimate obtained with the 2-D FFT method. The use of the 2-D FFT results in high sidelobes; the two 1-D continuous pulses in the spectrum are barely resolved. Figure 7(b) shows the phase difference estimates obtained with the 2-D FFT method. The effect of the high sidelobes is evident in the estimated phase difference, as can be seen from the smeared phase difference estimates for the two 1-D continuous pulses in the spectrum. Figure 6(c) shows the modulus of the complex amplitude estimate obtained by using 2-D FFT with a circularly symmetric Kaiser window with shape parameter 4. The windowed FFT method reduces the sidelobes. However, it widens the spectral peaks and, as a result, the closely spaced one-dimensional continuous pulses are smeared together. This effect is also visible in the phase difference estimates shown in Figure 7(c), where the smearing for the two 1-D continuous pulses is more pronounced as compared to the FFT method. Figure 6(d) shows the modulus of the complex amplitude estimate obtained by applying the Capon method with $M_1 = M_2 = 6$, that is $M_1 = 0.25N_1$ and with $M_2 = 0.25N_2$. We observe that the amplitude estimates are fairly accurate (M_i being 25% of N_i, $i = 1, 2$, seems to give the most accurate complex amplitude and phase difference estimates for the Capon method, as is evident from the RMSEs of the estimates shown in Figures 4 and 5). The resolution, however, is poor owing to the wide spectral peaks, but the sidelobes are reduced far more than those obtained by using the 2-D FFT and the 2-D windowed FFT. Figure 7(d) shows the phase difference estimates obtained by applying the Capon method with $M_1 = M_2 = 6$. Although the phase difference estimates obtained for the spectral lines are better (owing to the reduced sidelobes), we see that the estimates for the two continuous 1-D pulses are smeared together. Figure 6(e) shows the modulus of the complex amplitude estimate obtained by applying the Capon method with $M_1 = M_2 = 12$, that is $M_i = 0.5N_i$, $i = 1, 2$. Although the spectral peaks are very narrow, we observe that the amplitude estimates are highly inaccurate, especially for the two continuous 1-D pulses. Figure 7(e) shows the phase difference estimates obtained by applying the Capon method with $M_1 = M_2 = 12$. We see that the phase difference estimates are not accurate, with visible fragmentation for the two continuous 1-D pulses. Figure 6(f) shows the modulus of the complex amplitude estimate obtained

by applying the APES algorithm with $M_1 = M_2 = 16$. We observe that although the spectral peaks are not as narrow as those obtained with Capon with $M_1 = M_2 = 12$, the amplitude estimates are much more accurate. Conversely, although the amplitude estimates in Figure 6(f) are as accurate as those obtained with the Capon method with $M_1 = M_2 = 6$, the resulting spectral peaks are much narrower, with further reduced sidelobes. Thus we see that the APES algorithm provides the best complex amplitude estimates for all of the cases that we have considered. A similar conclusion can be drawn for the phase difference estimates obtained by the APES algorithm shown in Figure 7(f).

We now compare the APES algorithm with the FFT and Capon methods for interferometric SAR imaging. First we show that large chip sizes have negligible effect on the resolution and complex amplitude estimates, so that a reasonable chip size can be used for the case of large data dimensions as described at the end of Section 3. Figures 8(a) and (b) show the modulus of complex SAR images (in dB and scaled to be between 0 and 255 and of size 96×96) obtained by applying the APES algorithm to one channel (as described in [1]) of a small portion of the data collected by ERIM's (Environmental Research Institute of Michigan's) DCS IFSAR, with chip sizes $N_1 = N_2 = 16$ and $N_1 = N_2 = 32$, and parameters $M_1 = M_2 = 8$ and $M_1 = M_2 = 16$, respectively. We observe that both images show nearly the same texture and are comparable in resolution as well. Hence we choose the chip size to be $N_1 = N_2 = 16$ in what follows.

We now compare the performance of the methods that we have used in our simulations, when they are applied on the experimental data collected by ERIM's DCS IFSAR. The images in Figure 9 show the modulus of the complex amplitude estimates, obtained by the four different methods, on a portion surrounding and including a part of the Michigan stadium. The modulus of the complex amplitude estimate is evaluated in dB and is scaled to be between 0 and 255. The radar in these images is at the top of the image as is evident from the shadows cast by objects such as trees. Figure 9(a) shows the result of the FFT method. We observe that a lot of detail is lost in the sidelobes and the shadows are also not defined properly. Figure 9(b) shows the result of windowed FFT method when a circularly symmetric Kaiser window with shape parameter 4 is used. We observe that although the sidelobes are significantly reduced, the resolution has become poorer as a result of windowing. Figure 9(c) shows the result of the Capon method with chip size $N_1 = N_2 = 16$ and parameters $M_1 = M_2 = 4$. We note that the sidelobes are reduced more compared to the windowed FFT method and the resolution is better as well. The image has a grainy appearance although the shadows are more properly defined as compared to the FFT and the windowed FFT methods. Figure 9(d) shows

the result of the APES algorithm with chip size $N_1 = N_2 = 16$ and parameters $M_1 = M_2 = 11$. We see that apart from the sidelobes being reduced as compared to the Capon method, the image has the best resolution with very well defined regions, not only where the return is strong (as opposed to the granularity seen with the Capon method), but also for the shadow portions. Among other areas, the result is most pronounced when we compare the stadium rim and the upper left portion of the image (which is the area that we have used in Figure 8). We mention here that although using larger M_1 and M_2 with the Capon method will yield narrower spectral peaks than the APES algorithm, it will give highly inaccurate complex amplitude and phase difference estimates, as is evident from the simulation results for the one-dimensional and two-dimensional cases considered earlier.

Finally, we compare the results of the four methods again, but this time we make a composite color image using both the modulus of the complex amplitude and the phase difference estimates. We have not shown the phase difference estimates in a gray scale image as it is difficult to discern the different heights from the image that we obtain, especially the portions which have very small complex amplitude estimates are more prone to have phase estimates which are incorrect owing to the manner in which the data is modeled in Equation (2). The images shown in Figures 10(a)–(d) are formed using a pseudo color mapping. The modulus of the complex amplitude image is evaluated in dB and scaled to be between 0 and 255. The phase difference image is also scaled to be between 0 and 255. The hue at each pixel is decided by the scaled phase difference estimate, while relative brightness of the color is decided by the scaled modulus of the complex amplitude estimate, with the colors being fully saturated. The use of color now discriminates between the portions of the image having different heights and more details show up which are not visible in the images shown in Figure 8. The conclusions that we have drawn for the previous example are evident once again. The APES algorithm yields the best spectral estimates with very good resolution and highly reduced sidelobes.

V. Conclusions

We have demonstrated how the APES (Amplitude and Phase Estimation of a Sinusoid) algorithm, which is an adaptive FIR filtering approach, can be used for interferometric SAR imaging. We have compared the APES algorithm with other FIR filtering approaches including the FFT and Capon methods. We have

shown, by means of both numerical and experimental examples, that the adaptive FIR filtering approaches such as Capon and APES can yield accurate complex amplitude and phase difference or height estimates with much lower sidelobes and narrower spectral peaks than the FFT method. In particular, we have shown that the APES algorithm gives more accurate complex amplitude and phase difference estimates than the Capon method, although the latter can yield narrower spectral peaks than the former.

Appendix – Performance Prediction of the APES algorithm

In this appendix we very briefly describe an algorithm for the performance prediction of the APES algorithm. Consider first the case of 1-D data sequences. According to Equations (9) in [1], the complex amplitude of a sinusoid with frequency ω is determined by the following over-determined equations:

$$e^{ji\omega}\alpha(\omega) = \mathbf{h}^H(\omega)\bar{\mathbf{z}}_i, \quad i = 0, 1, \cdots, N-M. \tag{33}$$

where the definitions of the notations in (33) can be found in [1]. We estimate the mean-squared errors (MSE) of the estimated $\alpha(\omega)$ by

$$\mathrm{MSE}_{\alpha(\omega)} = \frac{1}{(N-M+1)} \sum_{i=0}^{N-M} \left| e^{-ji\omega}\mathbf{h}^H(\omega)\bar{\mathbf{z}}_i - \hat{\alpha}(\omega) \right|^2. \tag{34}$$

where $\hat{\alpha}(\omega)$ is the estimate obtained with the APES algorithm. The case of 2-D data matrices is similar.

Figures 11-15 are five examples showing the performance of the performance prediction algorithm. These examples are described in detail in [1]. We note from Figures 11 and 12 that the estimated normalized mean-squared errors (NMSE) are close to the NMSE obtained with 100 Monte-Carlo simulations. Figures 13, 14, and 15 are also as expected. Hence our approach can be used to predict the reliability of the estimates obtained with APES.

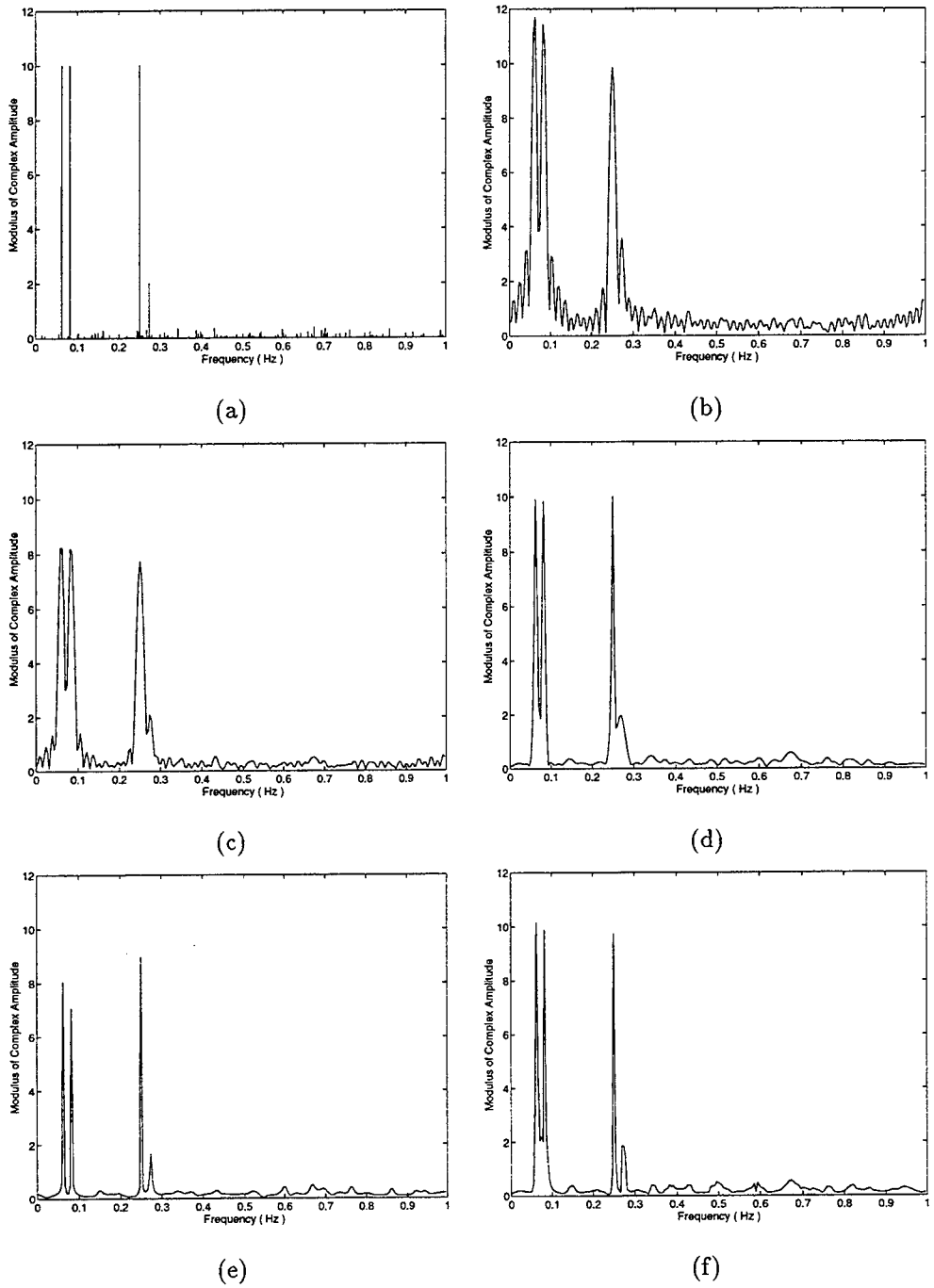


Figure 2: Complex amplitude estimates when noise variance is 1 and $N = 64$. (a) True spectrum. (b) FFT . (c) FFT with Kaiser window and shape parameter 2. (d) Capon with $M = 16$. (e) Capon with $M = 32$. (f) APES with $M = 42$.

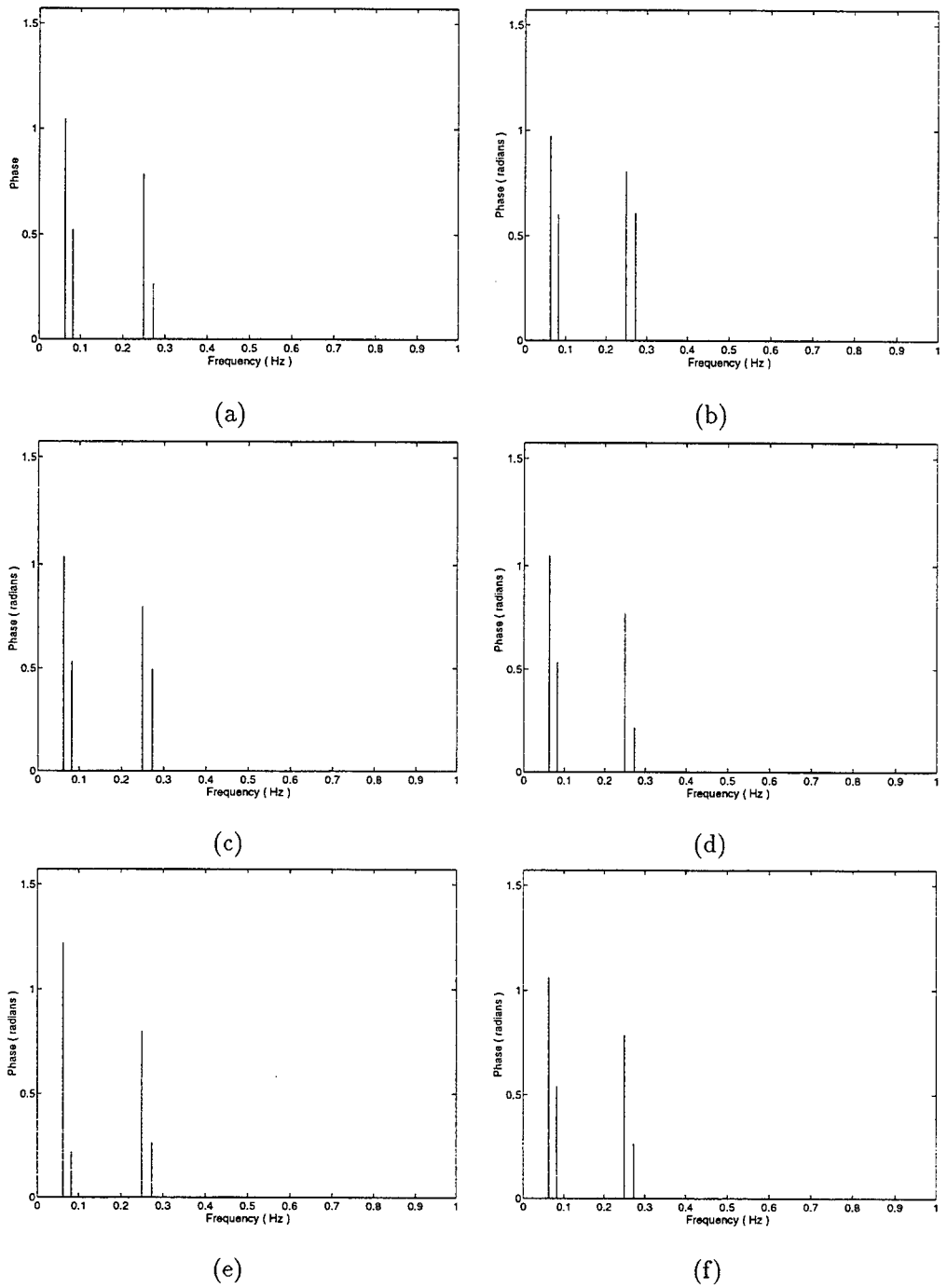
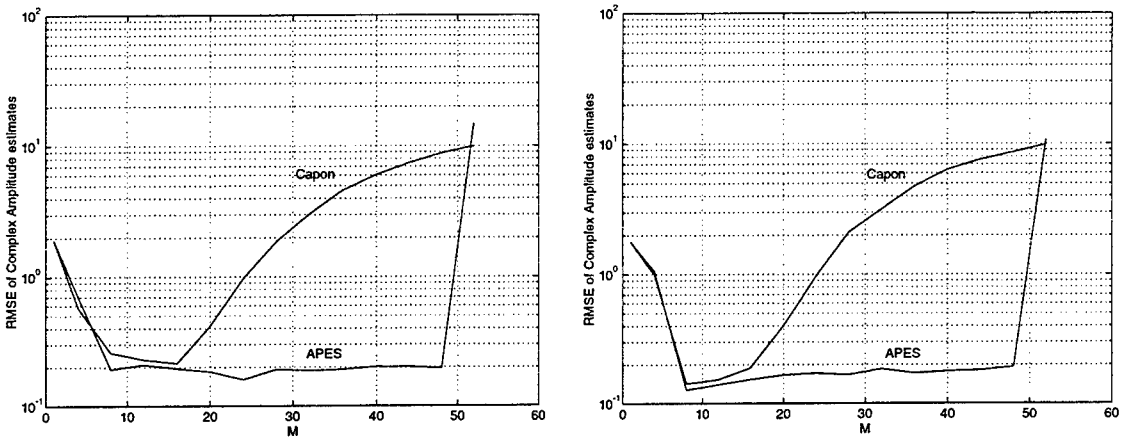
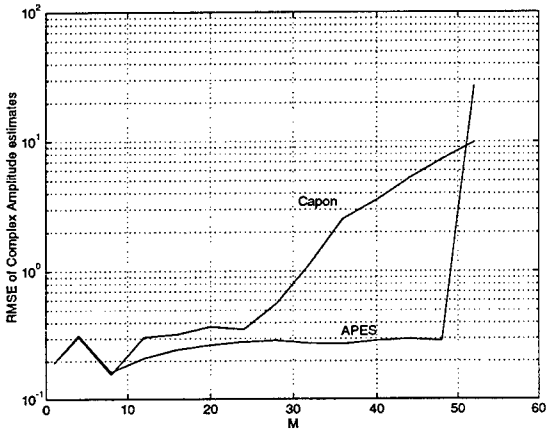


Figure 3: Phase difference or height estimates (in radians) when noise variance is 1 and $N = 64$. (a) True phase difference. (b) FFT. (c) FFT with Kaiser window and shape parameter 2. (d) Capon with $M = 16$. (e) Capon with $M = 32$. (f) APES with $M = 42$.

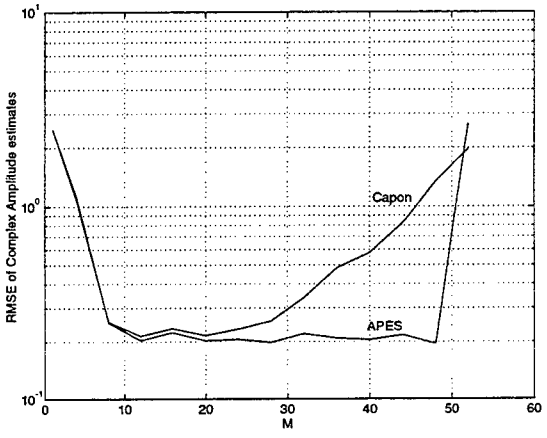


(a)

(b)



(c)



(d)

Figure 4: Root-Mean-squared errors (RMSEs) of the complex amplitude estimates obtained with Capon and APES as a function of M for the example in Figure 1. (a) - (d) are for the 1st - 4th spectral lines in Figure 2(a), respectively.

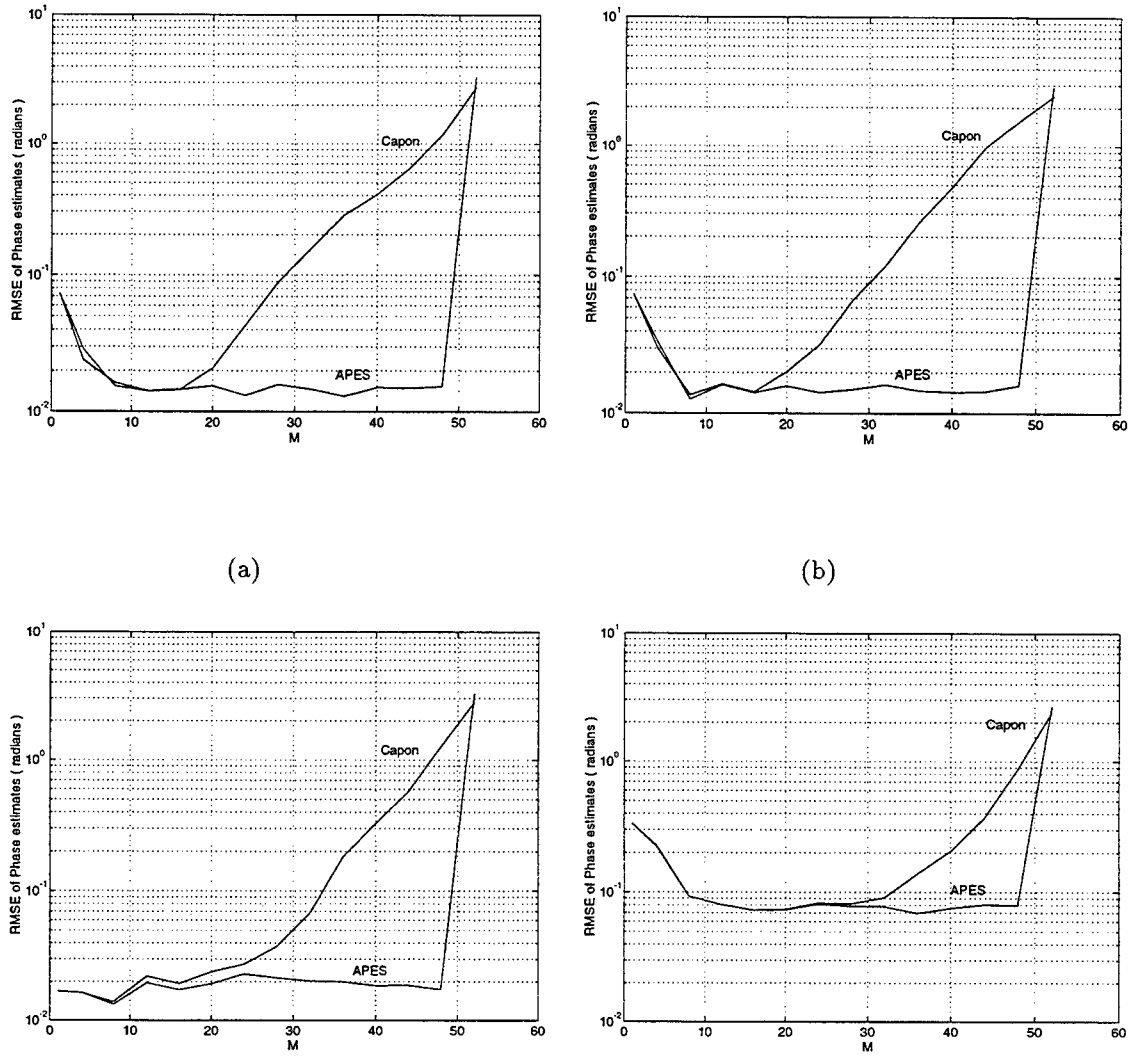


Figure 5: Root-Mean-squared errors (RMSEs) of the phase difference estimates obtained with Capon and APES as a function of M for the example in Figure 2. (a) - (d) are for the 1st - 4th spectral lines in Figure 3(a), respectively.

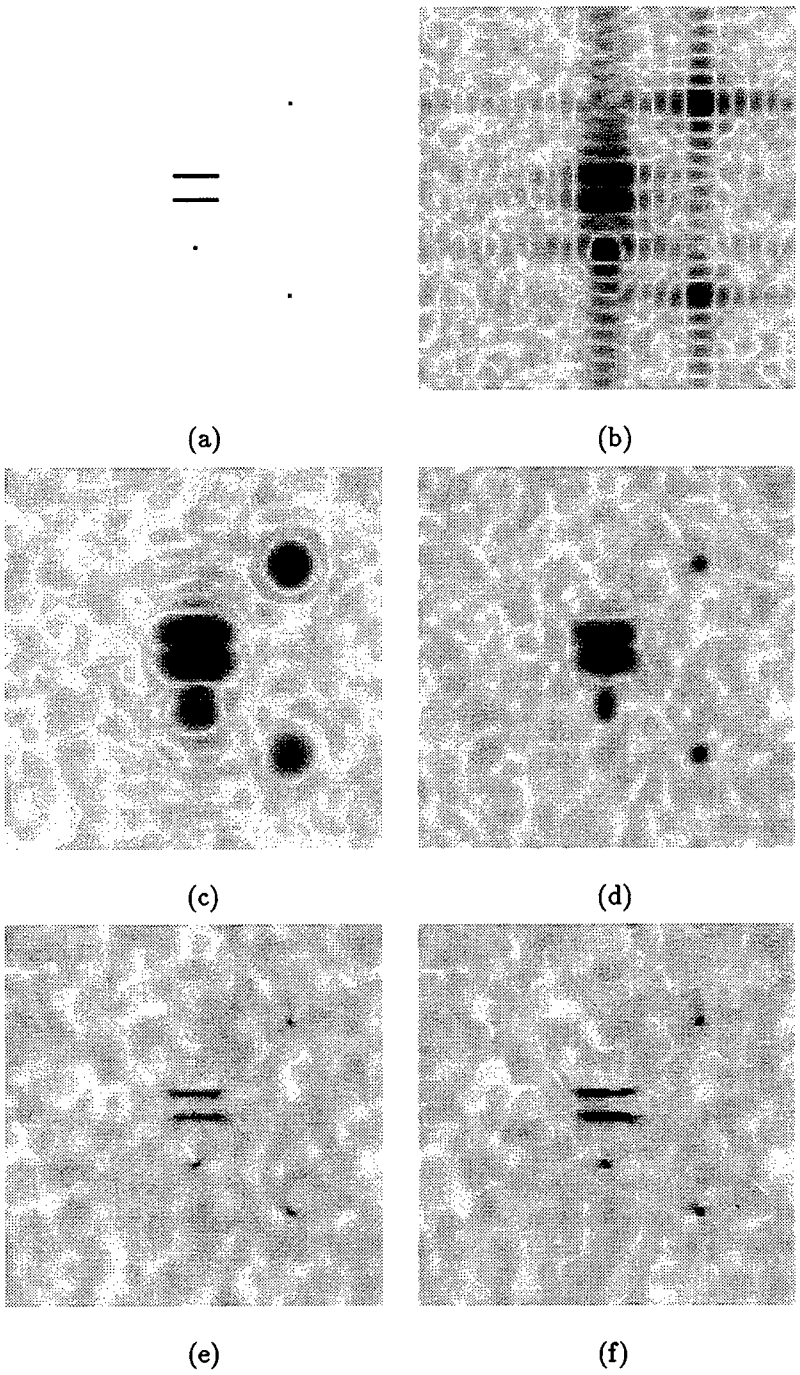


Figure 6: Modulus of complex amplitude estimates when noise variance is 20 and $N_1 = N_2 = 24$. (a) True Spectrum. (b) FFT. (c) FFT with Kaiser window and shape parameter 4. (d) Capon with $M_1 = M_2 = 6$. (e) Capon with $M_1 = M_2 = 12$. (f) APES with $M_1 = M_2 = 16$.

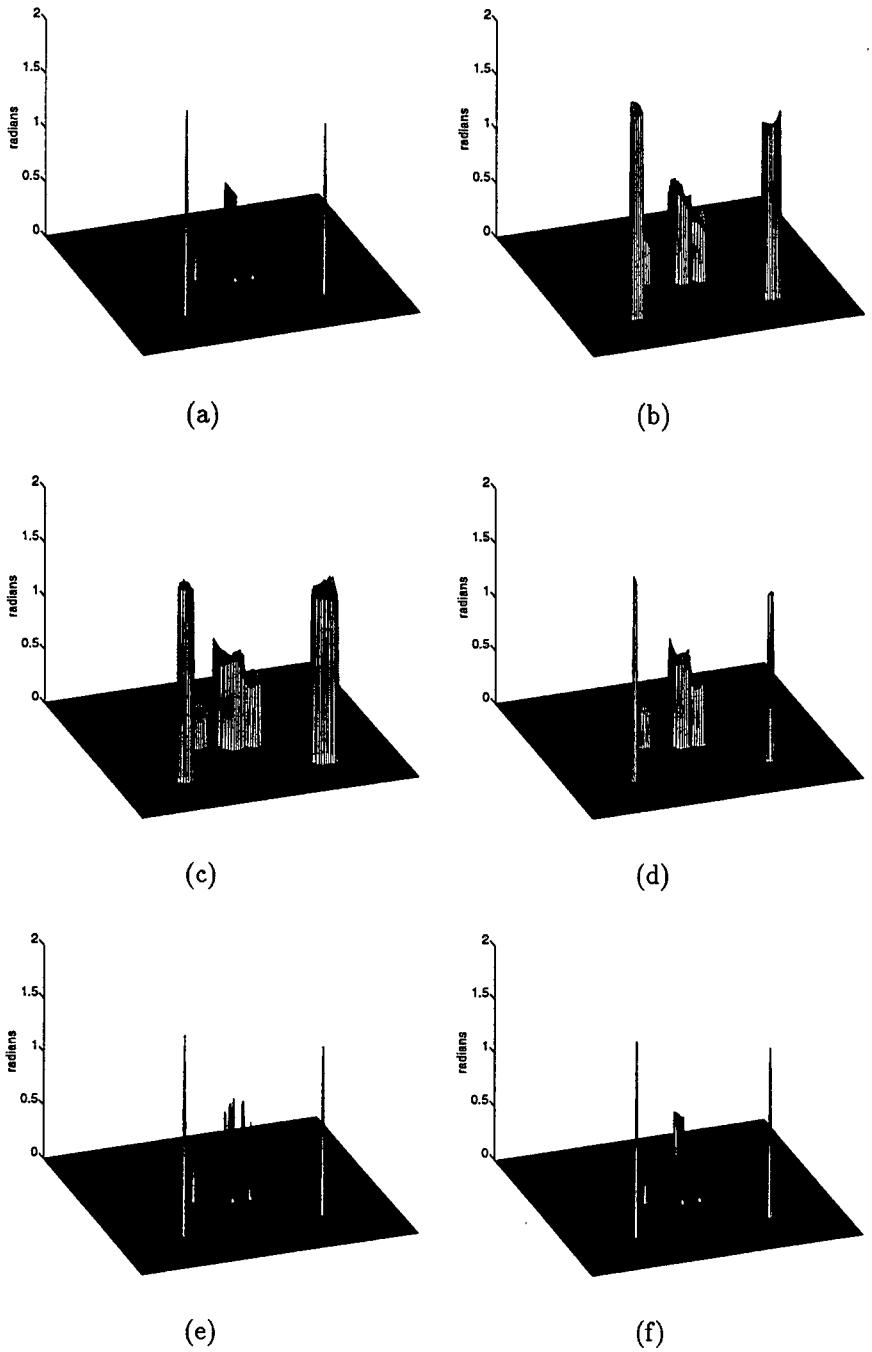
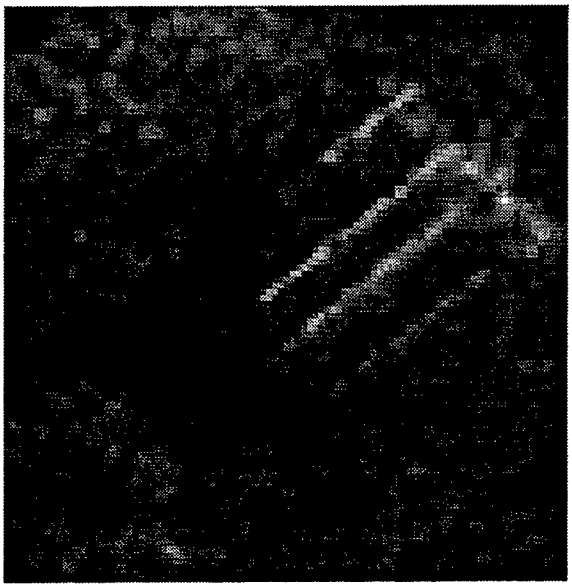


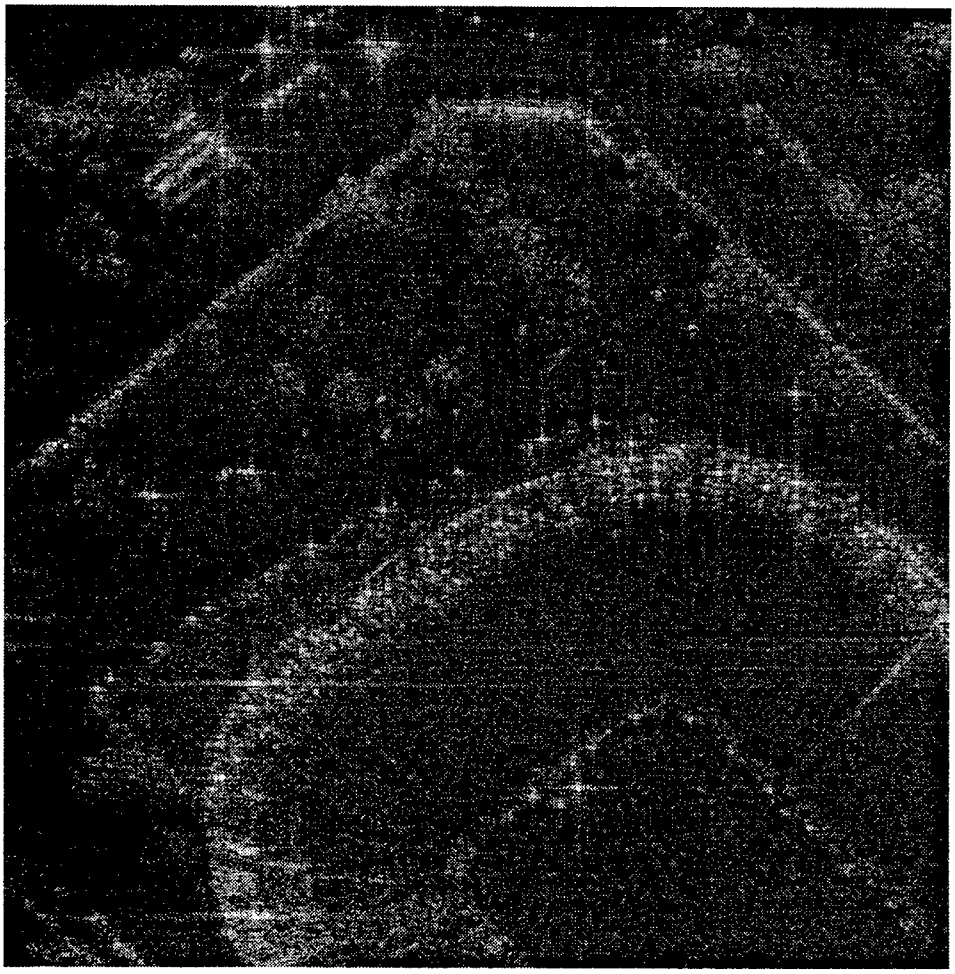
Figure 7: Phase difference or height estimates (in radians) when noise variance is 20 and $N_1 = N_2 = 24$. (a) True phase difference. (b) FFT. (c) FFT with Kaiser window and shape parameter 4. (d) Capon with $M_1 = M_2 = 6$. (e) Capon with $M_1 = M_2 = 12$. (f) APES with $M_1 = M_2 = 16$.



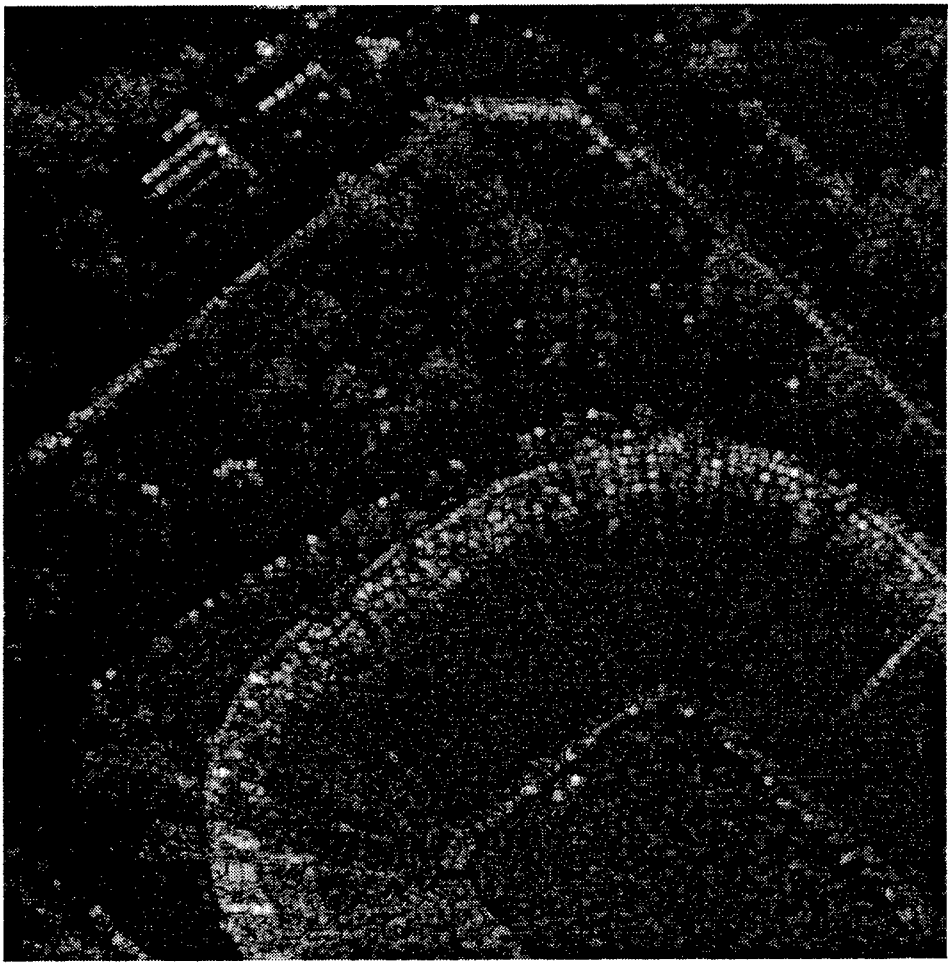
(a)

(b)

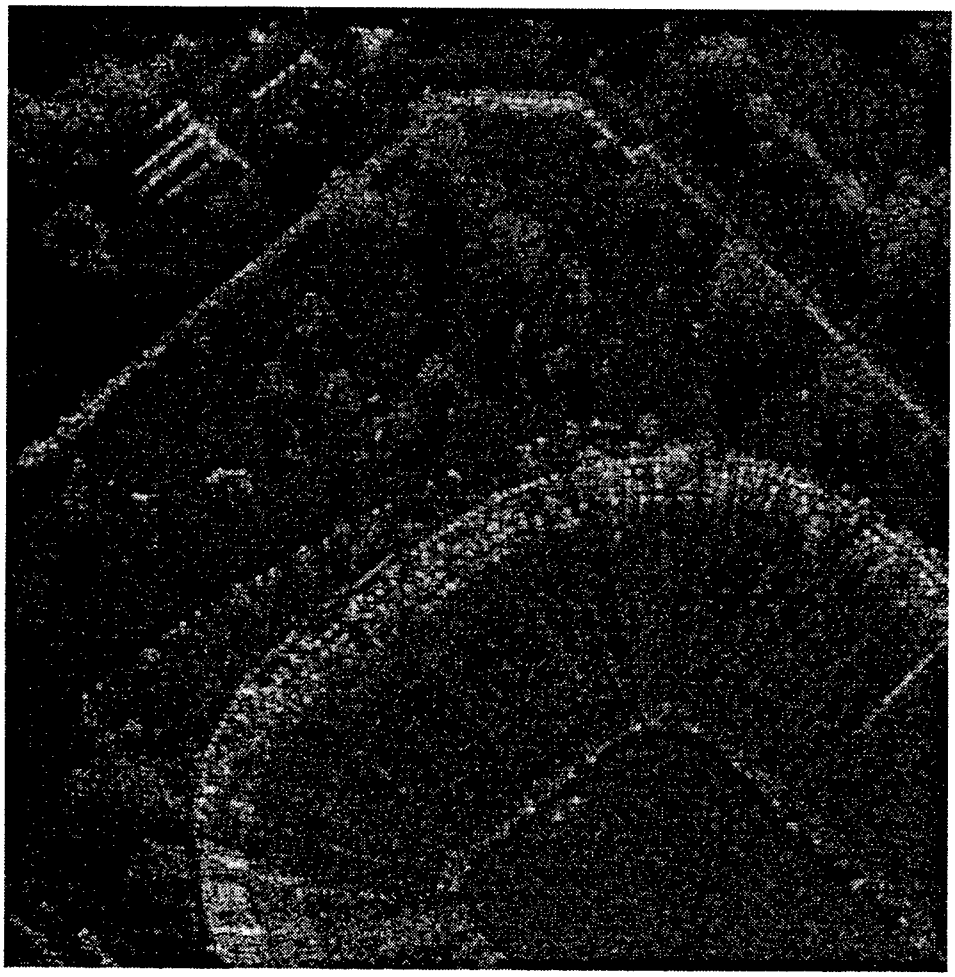
Figure 8: SAR images (96×96) obtained by applying the APES algorithm to one channel of a small portion of the data collected by ERIM's DCS IFSAR. (a) Chip size $N_1 = N_2 = 16$ and overlap $M_1 = M_2 = 8$. (b) Chip size $N_1 = N_2 = 32$ and overlap $M_1 = M_2 = 16$.



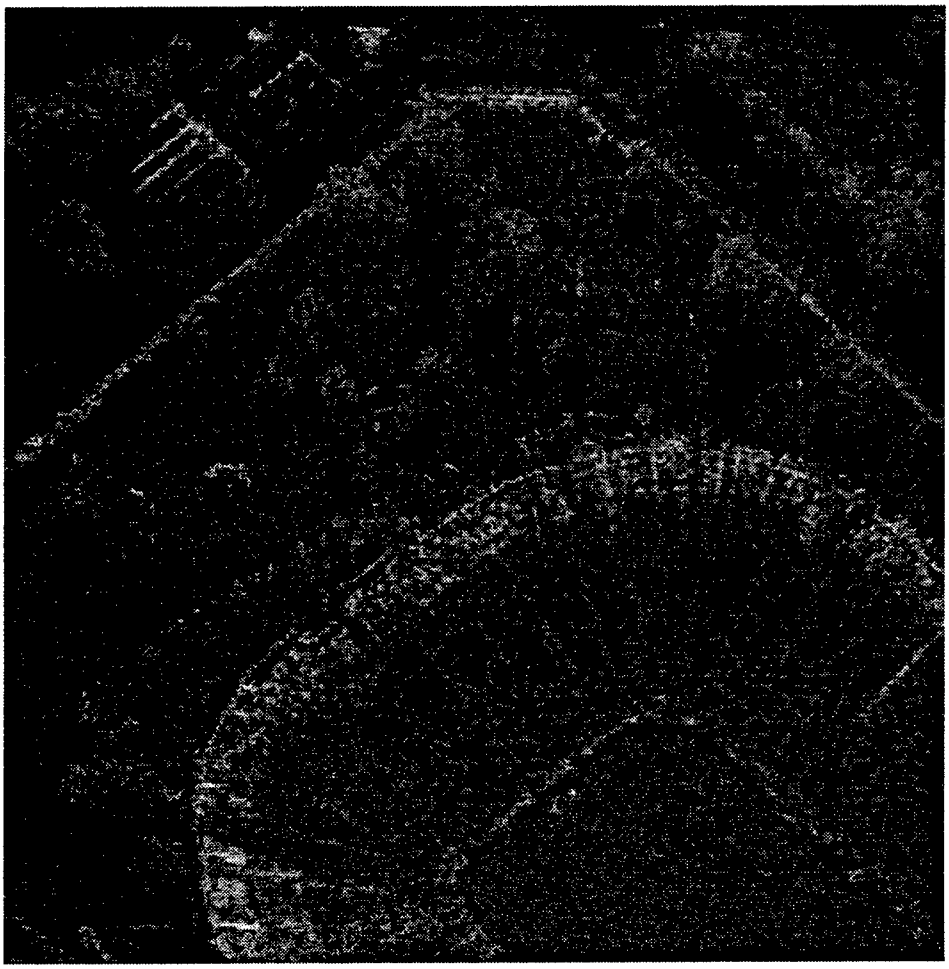
(a)



(b)



(c)

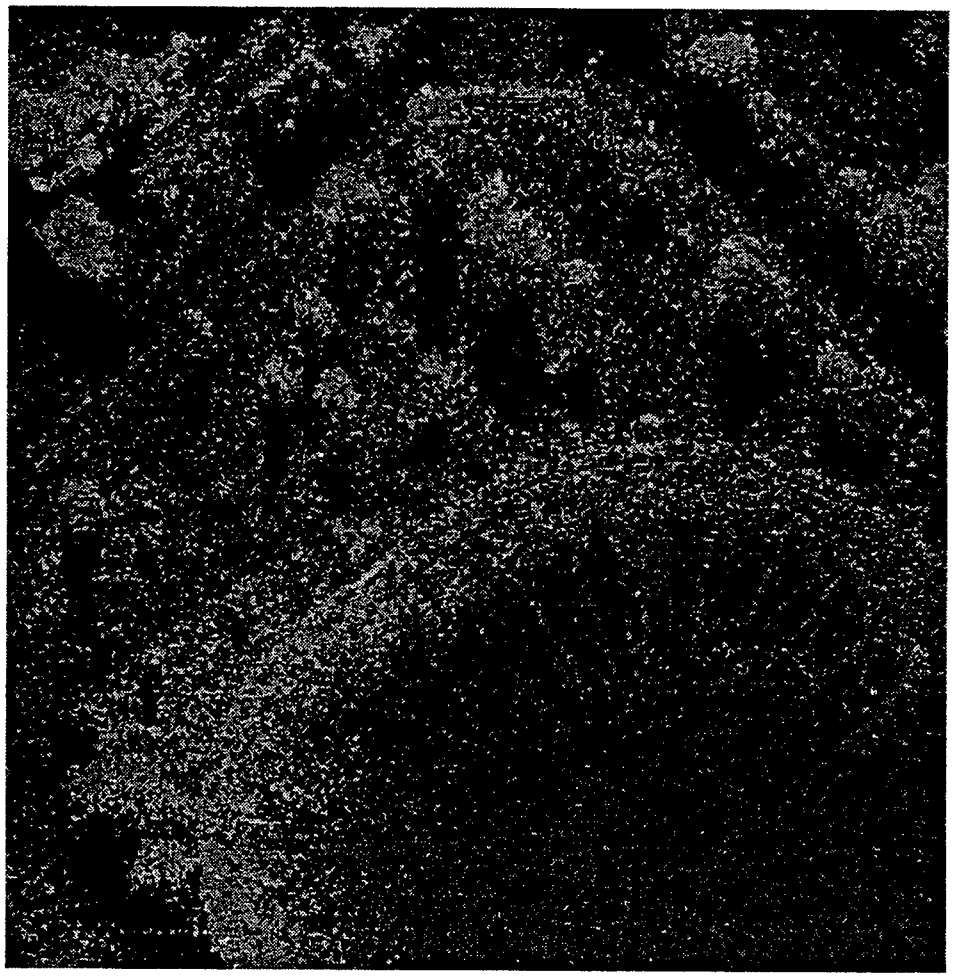


(d)

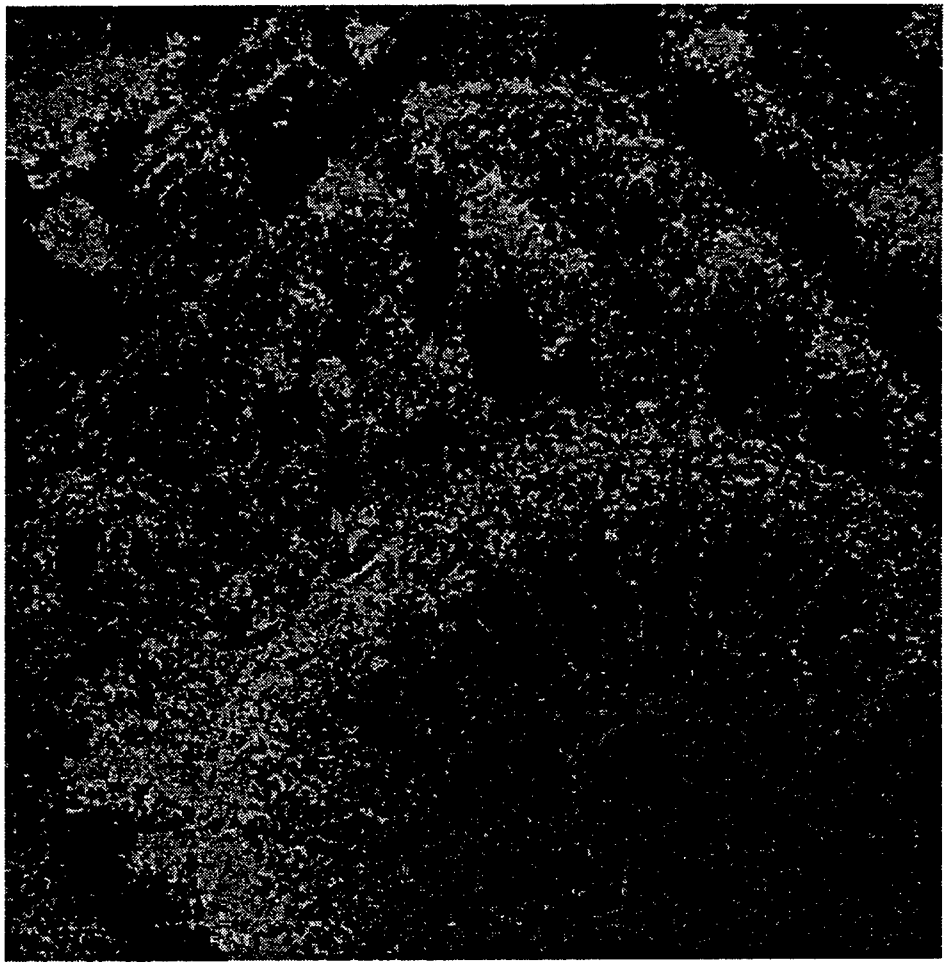
Figure 9: SAR images (496×496) obtained from a portion of the data collected by ERIM's DCS IFSAR. (a) FFT. (b) FFT with Kaiser window and shape parameter 4. (c) Capon with $N_1 = N_2 = 16$ and $M_1 = M_2 = 4$. (d) APES with $N_1 = N_2 = 16$ and $M_1 = M_2 = 11$.

(a)

(b)



(c)



(d)

Figure 10: 3-D SAR images (496 × 496) obtained from a portion of the data collected by ERIM's DCS IFSAR. (a) FFT. (b) FFT with Kaiser window and shape parameter 4. (c) Capon with $N_1 = N_2 = 16$ and $M_1 = M_2 = 4$. (d) APES with $N_1 = N_2 = 16$ and $M_1 = M_2 = 11$.

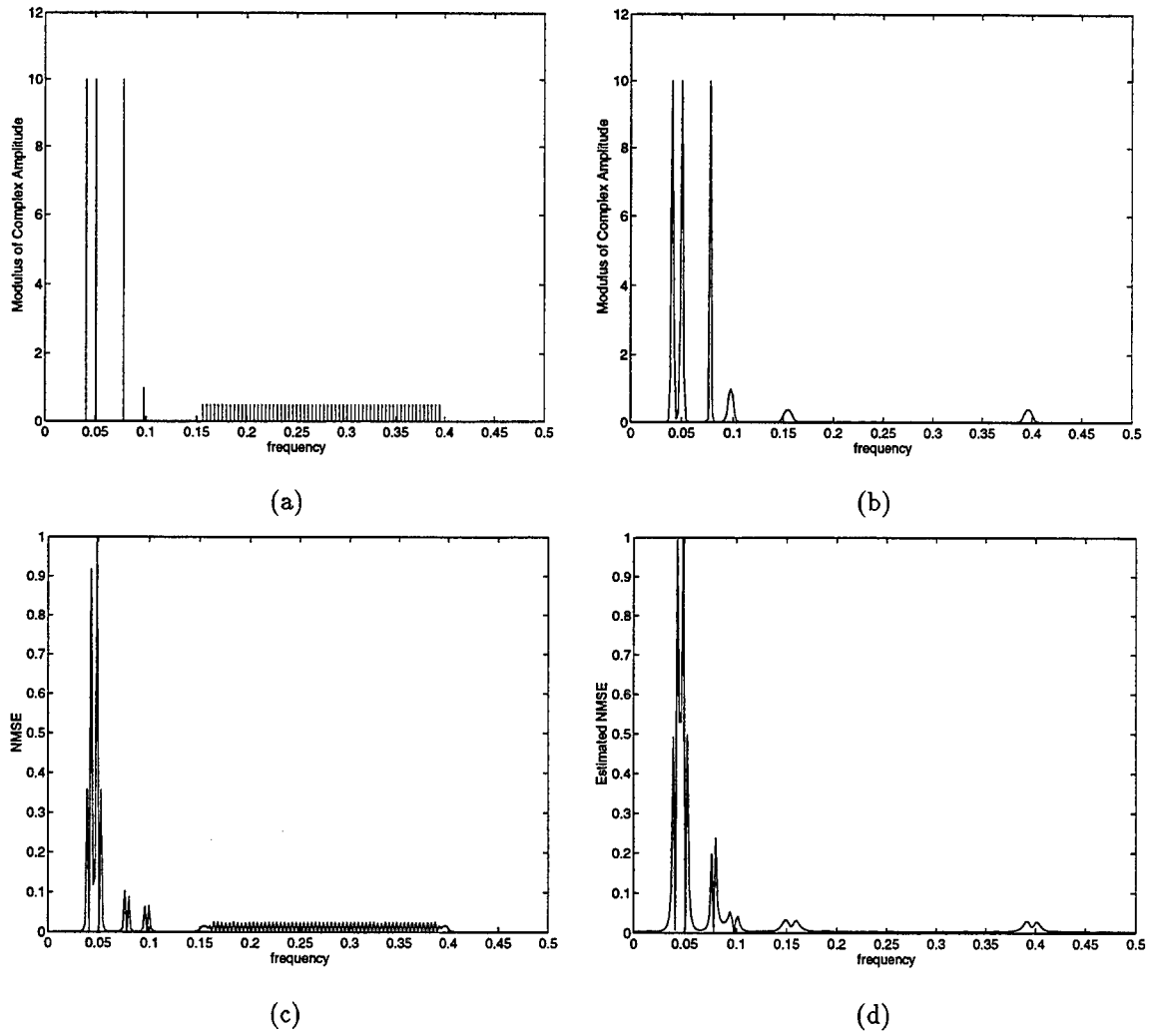


Figure 11: Performance prediction of APES. (a) True spectrum. (b) Estimated spectrum obtained with APES. (c) Normalized mean-squared errors (NMSE) obtained from 100 Monte-Carlo simulations. (d) Estimated NMSE.

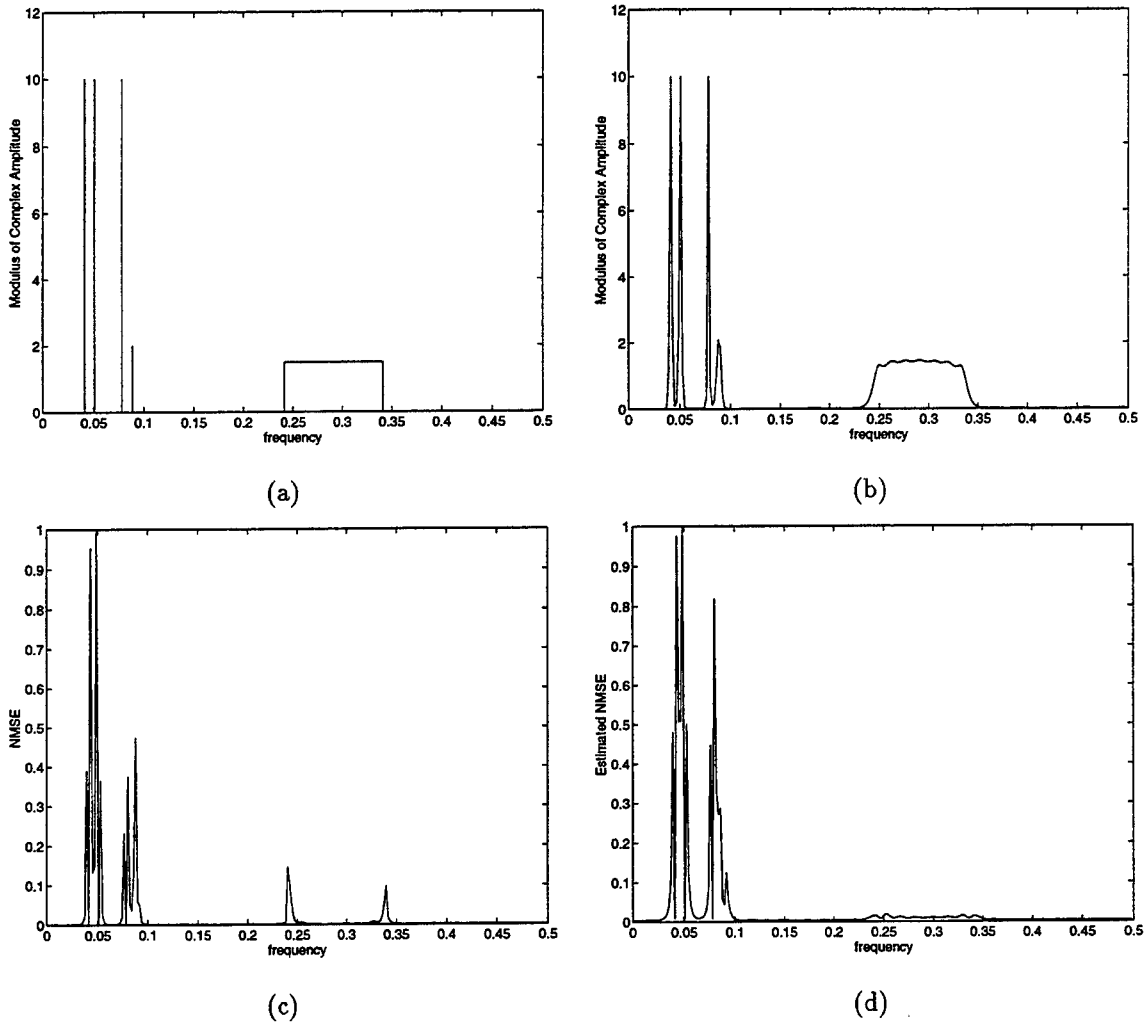
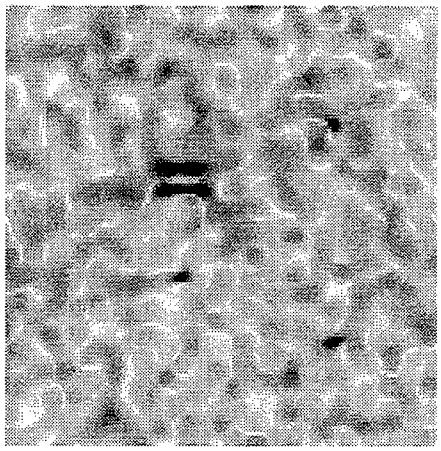
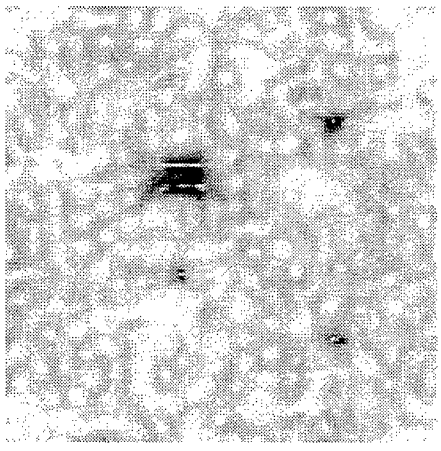


Figure 12: Performance prediction of APES. (a) True spectrum. (b) Estimated spectrum obtained with APES. (c) Normalized mean-squared errors (NMSE) obtained from 100 Monte-Carlo simulations. (d) Estimated NMSE.

(a)



(b)



(c)

Figure 13: Performance prediction of APES. (a) True spectrum. (b) Estimated spectrum obtained with APES. (c) Estimated NMSE.

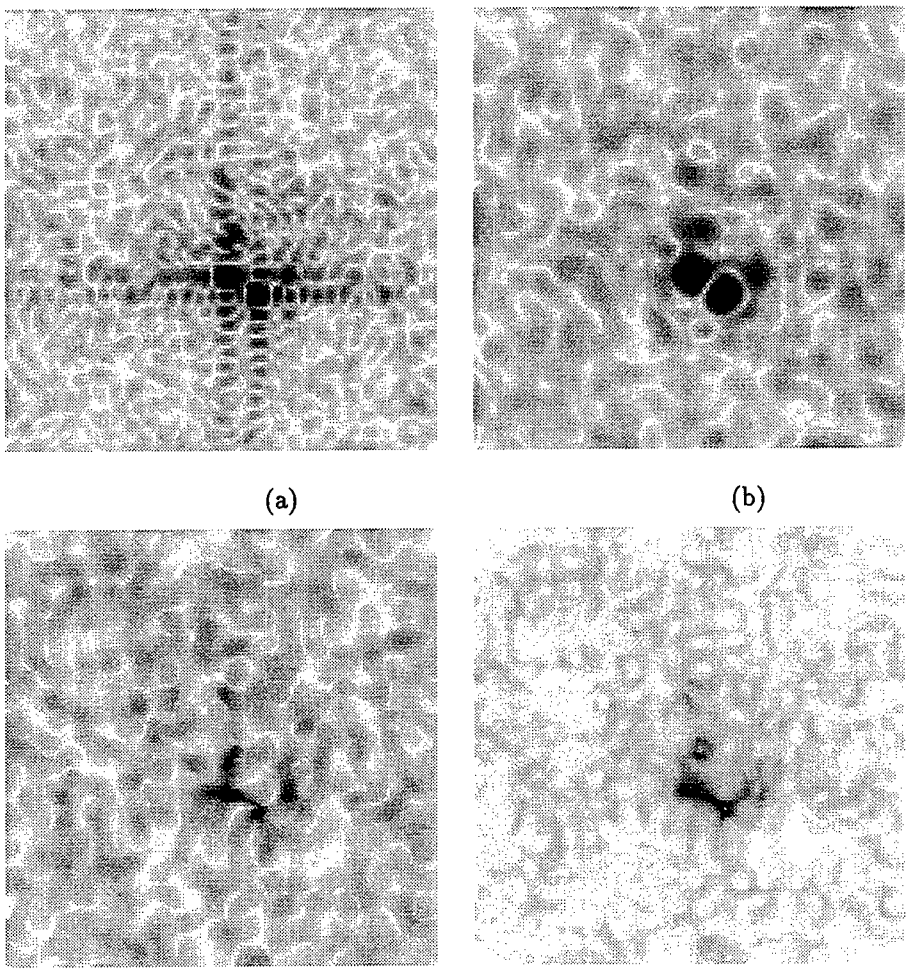


Figure 14: Performance prediction of APES (applied in SAR image estimation of an XPATCH simulated tank with data dimensions $N = \bar{N} = 32$. (a) Estimated with the 2-D FFT method. (b) Estimated with the 2-D FFT with circularly symmetric Kaiser window and shape parameter 4. (c) Estimated with the APES algorithm. (d) Estimated NMSE.

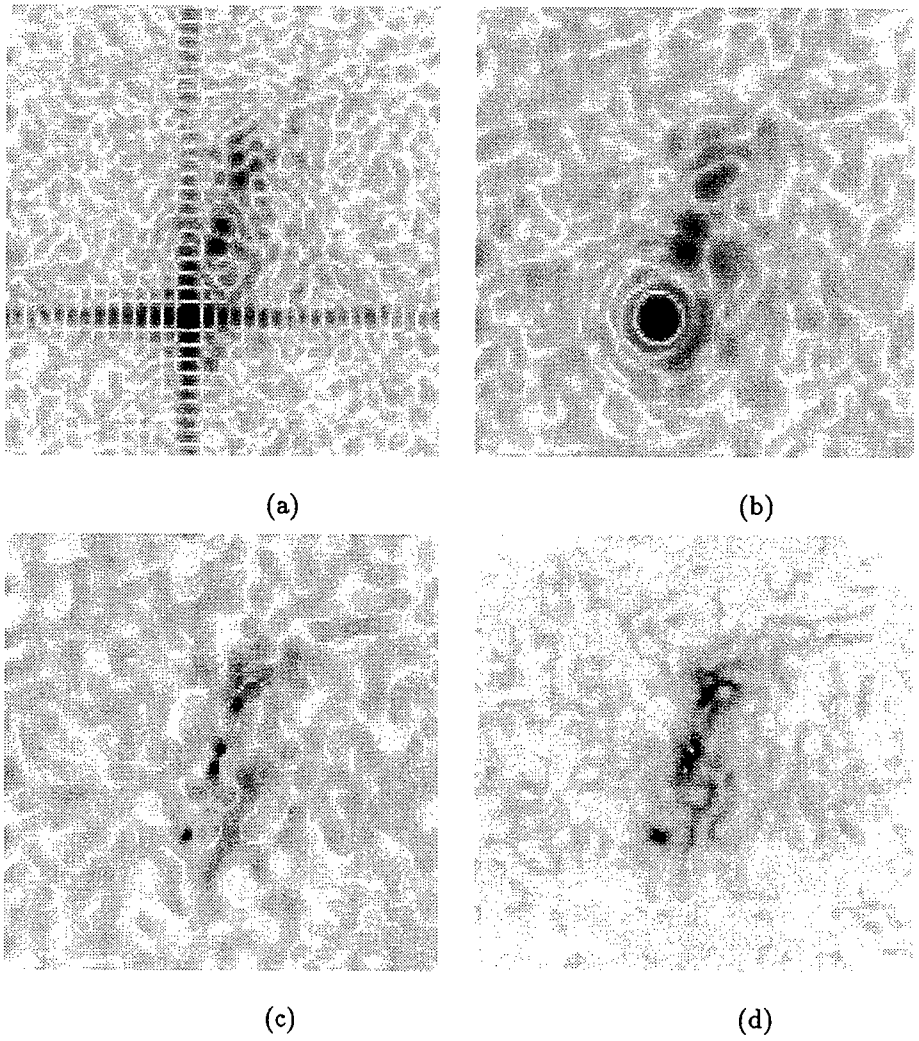


Figure 15: Performance prediction of APES (applied in SAR image estimation of an XPATCH simulated fire truck with data dimensions $N = \bar{N} = 32$. (a) Estimated with the 2-D FFT method. (b) Estimated with the 2-D FFT with circularly symmetric Kaiser window and shape parameter 4. (c) Estimated with the APES algorithm. (d) Estimated NMSE.

References

[1] J. Li and P. Stoica, "Adaptive filtering approach to spectral estimation and SAR imaging," accepted for publication in *IEEE Transactions on Signal Processing.*

[2] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proceedings of IEEE*, vol. 57, pp. 1408–1418, August 1969.

[3] S. M. Kay, *Modern Spectral Estimation: Theory and Application.* Englewood Cliffs, NJ: Prentice-Hall, Inc., 1988.

[4] S. R. DeGraaf, "SAR imaging via modern 2-d spectral estimation methods," *SPIE Proceedings on Optical Engineering in Aerospace Sensing,* Orlando , FL, April 1994.