REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time fc and maintaining the data needed, and completing and reviewing the collection of information. Send comments regard information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for informa 1204, Arlington, VA 22202-4302, and to the Office of management and Budget, Paperwork Reduction Project (0704-0188)

AFRL-SR-BL-TR-98-
0792

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December, 1994	Final

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
USAF Summer Research Program - 1994 Summer Faculty Research Program Final Reports, Volume 5A, Wright Laboratory	
6. AUTHORS Gary Moore	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Research and Development Labs, Culver City, CA	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR/NI 4040 Fairfax Dr, Suite 500 Arlington, VA 22203-1613	10. SPONSORING/MONITORING AGENCY REPORT NUMBER
11. SUPPLEMENTARY NOTES Contract Number: F49620-93-C-0063	
12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for Public Release	12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The United States Air Force Summer Faculty Research Program (USAF- SFRP) is designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by the faculty members being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory Technical Directorates and Air Force Air Logistics Centers. Each participant provided a report of their research, and these reports are consolidated into this annual report.

14. SUBJECT TERMS	15. NUMBER OF PAGES
AIR FORCE RESEARCH, AIR FORCE, ENGINEERING, LABORATORIES, REPORTS, SUMMER, UNIVERSITIES	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

RESEARCH & DEVELOPMENT LABORATORIES
AFOSR PROGRAM OFFICE

March 5, 1999

Lilla Mae Davis,

The following pages are unavailable because of bindery problems. In 1994 the publishing company we used had problems with the bindery.

Volume 5A page 4-5 missing.

Volume 5B page 34-6 missing.
Volume 5B page 58-14 side of page cut-off.

Volume 8 page 15-4 missing.

Johnetta Thompson
Program Administrator

RDL, AFOSR PROGRAM OFFICE, 5800 UPLANDER WAY, CULVER CITY, CALIFORNIA 90230-6608 (800)677-1363 & (310)410-1244
e-mail (INTERNET) ADDRESS: afosr@rdl.com FAX (800)677-1363 & (310)216-6940

UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1994

SUMMER FACULTY RESEARCH PROGRAM FINAL REPORTS

VOLUME 5A

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

Program Administrator, RDL
Johnetta Thompson

19981204 071

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Bolling Air Force Base

Washington, D.C.

December 1994

DTIC QUALITY INSPECTED 4

PREFACE

Reports in this volume are numbered consecutively beginning with number 1. Each report is paginated with the report number followed by consecutive page numbers, e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3.

Due to its length, Volume 5 is bound in two parts, 5A and 5B. Volume 5A contains #1-33. Volume 5B contains reports #34-66. The Table of Contents for Volume 2 is included in both parts.

This document is one of a set of 16 volumes describing the 1994 AFOSR Summer Research Program. The following volumes comprise the set:

VOLUME	TITLE
1	Program Management Report
	Summer Faculty Research Program (SFRP) Reports
2A & 2B	Armstrong Laboratory
3	Phillips Laboratory
4	Rome Laboratory
5A & 5B	Wright Laboratory
6	Arnold Engineering Development Center, Frank J. Seiler Research Laboratory, and Wilford Hall Medical Center
	Graduate Student Research Program (GSRP) Reports
7	Armstrong Laboratory
8	Phillips Laboratory
9	Rome Laboratory
10	Wright Laboratory
11	Arnold Engineering Development Center, Frank J. Seiler Research Laboratory, and Wilford Hall Medical Center
	High School Apprenticeship Program (HSAP) Reports
12A & 12B	Armstrong Laboratory
13	Phillips Laboratory
14	Rome Laboratory
15A&15B	Wright Laboratory
16	Arnold Engineering Development Center

SFRP FINAL REPORT TABLE OF CONTENTS i-xxi

1. INTRODUCTION 1

2. PARTICIPATION IN THE SUMMER RESEARCH PROGRAM 2

3. RECRUITING AND SELECTION 3

4. SITE VISITS 4

5. HBCU/MI PARTICIPATION 4

6. SRP FUNDING SOURCES 5

7. COMPENSATION FOR PARTICIPANTS 5

8. CONTENTS OF THE 1994 REPORT 6

APPENDICIES:

A. PROGRAM STATISTICAL SUMMARY A-1

B. SRP EVALUATION RESPONSES B-1

SFRP FINAL REPORTS

SRP Final Report Table of Contents

Author	University/Institution Report Title	Armstrong Laboratory Directorate	Vol-Page
Dr. James L Anderson	University of Georgia , Athens , GA Determination of the Oxidative Redox Capacity of	AL/EQC	2- 1
Dr. Hashem Ashrafiuon	Villanova University , Villanova , PA ATB Simulation of Deformable Manikin Neck Models	AL/CFBV	2- 2
DR Stephan B Bach	Univ of Texas-San Antonio , San Antonio , TX Pre-Screening of Soil Samples Using a Solids Inser	AL/OEA	2- 3
Dr. Suzanne C Baker	James Madison University , Harrisonburg , VA Rat Pup Ultrasonic Vocalizations: A Sensitive Indi	AL/OER	2- 4
DR Alexander B Bordetsky	Univ of Texas - Dallas , Richardson , TX Knowledge-Based Groupware for Geographically Distr	AL/HRGA	2- 5
DR. Michael J Burke	Tulane University , New Orleans , LA An Empirical Examination of the Effect of Second-O	AL/HRMI	2- 6
DR Yu-Che Chen	University of Tulsa , Tulsa , OK A Study of the Kinematics, Dynamics and Control Al	AL/CFBS	2- 7
DR Shashikala T Das	Wilmington College , Wilmington , OH The Benchmark Dose Approach for Health Risk Assess	AL/OET	2- 8
DR. Donald W DeYoung	University of Arizona , Tucson , AZ Noise as a Stressor: An Assessment of Physiologic	AL/OEBN	2- 9
DR Judy B Dutta	Rice University , Houston , TX Memory for Spatial Position and Temporal Occurence	AL/CFTO	2- 10
DR Paul A Edwards	Edinboro Univ of Pennsylvania , Edinboro , PA Fuel Identification by Neural Network Analysis of	AL/EQC	2- 11

Author	University/Institution Report Title	Armstrong Laboratory Directorate	Vol-Page
Dr. Daniel L Ewert	North Dakota State University , Grand Forks , ND Regional Arterial Compliance and Resistance Change	AL/AOCN	2- 12
Dr. Bernard S Gerstman	Florida International Universi , Miami , FL Laser Induced Bubble Formation in the Retina	AL/OEO	2- 13
DR Robert H Gilkey	Wright State University , Dayton , OH Relation Between Detection and Intelligibility in	AL/CFBA	2- 14
Dr. Kenneth A Graetz	University of Dayton , Dayton , OH Using Electronic Brainstorming Tools to Visually R	AL/HRGA	2- 15
Dr. Donald D Gray	West Virginia Unicersity , Morgantown , WV Improved Numerical Modeling of Groundwater Flow an	AL/EQC	2- 16
Dr. Pushpa L Gupta	University of Maine , Orono , ME Regression to the Mean in Half-Life Studies	AL/AOEP	2- 17
Dr. Thomas E Hancock	Grand Canyon University , Phoenix , AZ An Expanded Version of the Kulhavy/Stock Model of	AL/HR2	2- 18
DR. Alexis G Hernandez	University of Arizona , Tucson , AZ Preliminary Results of the Neuropsychiatrically En	AL/AOCN	2- 19
DR P. A Ikomi	Central State University , Wilberforce , OH A Realistic Multi-Task Assessment of Pilot Aptitud	AL/HRMI	2- 20
Dr. Arthur Koblasz	Georgia State University , Atlanta , GA Distributed Sensory Processing During Graded Hemod	AL/AOCI	2- 21
DR Manfred Koch	Florida State University , Tallahassee , FL Application of the MT3D Solute Transport Model to	AL/EQC	2- 21

SRP Final Report Table of Contents

Author	University/Institution Report Title	Armstrong Laboratory Directorate	Vol-Page
Dr. Donald H Kraft	Louisiana State University , Baton Rouge , LA An Exploratory Study of Weighted Fuzzy Keyword Bo	AL/CFHD	2- 23
Dr. Brother D Lawless	Fordham University , New York , NY Apoptosis Advanced Glycosylated End Products, Auto	AL/OER	2- 24
Dr. Tzesan Lee	Western Illinois University , Macomb , IL A Statistical Method for Testing Compliance	AL/OEM	2- 25
DR Robert G Main	California State Univ-Chico , Chico , CA A Study of Interaction in Distance Learning	AL/HRTT	2- 26
Dr. Augustus Morris	Central State University , Wilberforce , OH A Novel Design Concept for a Small, Force Reflecti	AL/CFBS	2- 27
DR Mark A Novotny	Florida State University , Tallahassee , FL Computer Calculation of Rate Constants for Biomole	AL/EQS	2- 28
Dr. Joseph H Nurre	Ohio University , Athens , OH A Review of Parameter Selection for Processing Cyl	AL/CFHD	2- 29
DR Edward L Parkinson	Univ of Tennessee Space Inst , Tullahoma , TN Improving the United States Air Force Environmenta	AL/EQS	2- 30
DR Malcom R Parks	University of Washington , Seattle , WA Communicative Challenges Facing Integrated Product	AL/AOE	2- 31
DR David R Perrott	California State Univ-Los Ange , Los Angeles , CA Aurally Directed Search: A Comparison Between Syn	AL/CFBA	2- 32
Dr. Edward H Piepmeier	University of South Carolina , Columbia , SC Dose Response Studies for Hyperbaric Oxygenation	AL/AOHP	2- 33

Author	University/Institution Report Title	Armstrong Laboratory Directorate	Vol-Page
DR Miguel A Quinones	Rice University , Houston , TX The Role of Experience in Training Effectiveness	AL/HRTE	2- 34
Dr. Ramaswamy Ramesh	SUNY, Buffalo , Buffalo , NY AETMS: Analysis, Design and Development	AL/HRAU	2- 35
DR Gary E Riccio	Univ of IL Urbana-Champaign , Urbana , IL REPORT NOT AVAILABLE AT PRESS TIME	AL/CFHP	2- 36
DR Kandasamy Selvavel	Claflin College , Orangeburg , SC Sequential Estimation of Parameters of Truncation	AL/AOEP	2- 37
DR David M Senseman	Univ of Texas-San Antonio , San Antonio , TX Multisite Optical Recording of Evoked Activity in	AL/CFTO	2- 38
DR Wayne L Shebilske	Texas A&M University , College Station , TX Linking Laboratory Research and Field Applications	AL/HRTI	2- 39
Dr. Larry R Sherman	University of Scranton , Scranton , PA Using The Sem-EDXA System at AL/OEA for Analysis o	AL/OEA	2- 40
Dr. Richard D Swope	Trinity University , San Antonio , TX Regional Arterial Complicance and Resistance Chang	AL/AOCI	2- 41
DR Steven D Tripp	The University of Kansas , Lawrence , KS Representing and Teaching a Discrete Machine: An	AL/HRTC	2- 42
DR Ryan D Tweney	Bowling Green State University , Bowling Green , OH Automated Detection of Individual Response Charact	AL/CFHP	2- 43
Dr. Brian S Vogt	Bob Jones University , Greenville , SC A Multiplexed Fiber-Optic Laser Fluorescence Spect	AL/EQW	2- 44

SRP Final Report Table of Contents

Author	University/Institution Report Title	Armstrong Laboratory Directorate	Vol-Page
DR Janet M Weisenberger	Ohio State University , Columbus , OH Investigation of the Role of Haptic Movement in Ta	AL/CFBA	2- 45

Author	University/Institution Report Title	Phillips Laboratory Directorate	Vol-Pag
DR Behnaam Aazhang	Rice University , Houston , TX High Capacity Optical Communication Networks	PL/VTPT	3- 1
DR Nasser Ashgriz	SUNY-Buffalo , Buffalo , NY On The Mixing Mechanisms in a Pair of Impinging Je	PL/RKFA	3- 2
Dr. Raymond D Bellem	Embry-Riddle Aeronautical Univ , Prescott , AZ Radiation Characterization of Commerically Process	PL/VTET	3- 3
DR Gajanan S Bhat	Tennessee , Knoxville , TN Polyetherimide Fibers: Production Processing and	PL/RKFE	3- 4
DR Ronald J Bieniek	University of Missouri-Rolla , Rolla , MO Practical Semiquantal Modelling of Collisional Vib	PL/GPOS	3-
DR Jan S Brzosko	Stevens Institute of Tech , Hoboken , NJ Conceptual Study of the Marauder Operation in the	PL/WSP	3-
DR Ping Cheng	Hawaii at Manoa , Honolulu , HI Determination of the Interfacial Heat Transfer Coe	PL/VTPT	3-
DR Meledath Damodaran	University of Houston-Victoria , Victoria , TX Concurrent Computation of Aberration Coefficients	PL/LIMI	3-
Dr. Ronald R DeLyser	University of Denver , Denver , CO Analysis to Determine the Quality Factor of a Comp	PL/WSA	3-
DR Jean-Claude M Diels	University of New Mexico , Albuquerque , NM Unidirectional Ring Lasers and Laser Gyros with Mu	PL/LIDA	3-
Dr. David M Elliott	Arkansas Technology University , Russellville , AR REPORT NOT AVAILABLE AT PRESS TIME	PL/RKFE	3-

SRP Final Report Table of Contents

Author	University/Institution Report Title	Phillips Laboratory Directorate	Vol-Page
DR Vincent P Giannamore	Xavier University of Louisiana , New Orleans , LA An Investigation of Hydroxylammonium Dinitramide:	PL/RKA	3- 12
DR James E Harvey	University of Central Florida , Orlando , FL A New Mission for the Air Force Phillips Laborator	PL/LIM	3- 13
DR Stan Heckman	Massachusettes Inst of technol , Cambridge , MA REPORT NOT AVAILABLE AT PRESS TIME	PL/GPAA	3- 14
DR. James M Henson	University of Nevada , Reno , NV High Resolution Range Doppler Data and Imagery for	PL/WSAT	3- 15
Dr. San-Mou Jeng	University of Cincinnati , Cincinnati , OH Can Design for Cogging of Titanium Aluminide Alloy	PL/RKFA	3- 16
MR. Gerald Kaiser	University of Mass/Lowell , Lowell , MA Physical Wavelets fo Radar and Sonar	Pl/GPOS	3- 17
MR Dikshitulu K Kalluri	University of Mass/Lowell , Lowell , MA Backscatter From a Plasma Plume Due to Excitation	PL/GP	3- 18
Lucia M Kimball	Worcester Polytechnic Inst. , Worcester , MA Investigation of Atmospheric Heating and Cooling B	PL/GPOS	3- 19
MR. Albert D Kowalak	University of Massachusetts/Lo , Lowell , MA Investigations of Electron Interactions with Molec	PL/GPID	3- 20
MR. Walter S Kuklinski	University of Mass/Lowell , Lowell , MA Ionspheric Tomography Using a Model Based Transfor	PL/GP	3- 21
Dr. Min-Chang Lee	Massachusetts Institute , Cambridge , MA Studies of Plasma Turbulence with Versatile Toroid	PL/GPSG	3- 22

Author	University/Institution Report Title	Phillips Laboratory Directorate	Vol-Page
DR Kevin J Malloy	University of New Mexico , Albuquerque , NM REPORT NOT AVAILABLE AT PRESS TIME	PL/VTRP	3- 23
Dr. Charles J Noel	Ohio State University , Columbus , OH Preparation and Characterization of Blends of Orga	PL/RKA	3- 24
DR Hayrani A Oz	Ohio State University , Columbus , OH A Hybrid Algebraic Equation of Motion-Neural Estim	PL/VTSS	3- 25
DR Sudhakar Prasad	University of New Mexico , Albuquerque , NM Focusing Light into a Multiple-Core Fiber: Theory	PL/LIMI	3- 26
DR Mark R Purtill	Texas A&M Univ-Kingsville , Kingsville , TX Static and Dynamic Graph Embedding for Parallel Pr	PL/WSP	3- 27
DR Krishnaswamy Ravi-Chandar	University of Houston , Houston , TX On the Constitutive Behavior of Solid Propellants	PL/RKAP	3- 28
Dr. Wolfgang G Rudolph	University of New Mexico , Albuquerque , NM Relaxation Processes In Gain Switched Iodine Laser	PL/LIDB	3- 29
DR Gary S Sales	Univof Massachusetes-Lowell , Lowell , MA Characterization of Polar Patches: Comparison of	PL/GPIA	3- 30
DR I-Yeu Shen	University of Washington , Seattle , WA A Study of Active Constrained Layer Damping Treatm	PL/VTSS	3- 31
DR Melani I Shoemaker	Seattle Pacific University , Seattle , WA Frequency Domain Analysis of Short Exposure, Photo	PL/LIMI	3- 32
DR Yuri B Shtessel	University of Alabama-Huntsvil , Huntsville , AL Topaz II Reactor Control Law Improvement	PL/VTPC	3- 33

SRP Final Report Table of Contents

Author	University/Institution Report Title	Phillips Laboratory Directorate	Vol-Page
Dr. Alexander P Stone	University of New Mexico , Alburquerque , NM Impedances of Coplanar Conical Plates in a Uniform	PL/WSR	3- 34
DR Charles M Swenson	Utah State University , Logan , UT Reflected Laser Communication System	PL/VTRA	3- 35
Dr. Y. C Thio	University of Miami , Coral Gables , FL A Mathematical Model of Self Compression of Compac	PL/WSP	3- 36
DR Jane M Van Doren	College of the Holy Cross , Worcester , MA Investigations of Electron Interactions with Molec	PL/GPID	3- 37
DR Daniel W Watson	Utah State University , Logan , UT A Heterogeneous Parallel Architecture for High-Spe	PL/VTEE	3- 38
Dr. Wayne J Zimmermann	Texas Woman's University , Denton , TX Determination of Space Debris Flux Based on a Fini	PL/WS	3- 39

Author	University/Institution Report Title	Rome Laboratory Directorate	Vol-Page
DR Valentine A Aalo	Florida Atlantic University , Boca Raton , FL A Program Plan for Transmitting High-Data-Rate ATM	RL/C3BA	4- 1
DR Moeness G Amin	Villanova University , Villanova , PA Interference Excision in Spread Spectrum Using Ti	RL/C3BB	4- 2
Richard G Barakat	Tufts University , Medford , MA REPORT NOT AVAILABLE AT PRESS TIME	RL/EROP	4- 3
DR David P Benjamin	Oklahoma State University , Stillwater , OK Designing Software by Reformulation Using Kids	RL/C3CA	4- 4
DR Frank T Berkey	Utah State University , Logan , UT The Application of Quadratic Phase Coding to OTH R	RL/OCDS	4- 5
DR Joseph Chaiken	Syracuse University , Syracuse , NY A Study of the Application of Fractuals and Kineti	RL/ERDR	4- 6
Dr. Pinyuen Chen	Syracuse University , Syracuse , NY On Testing the Equality of Covariance Matrices Use	RL/OCTS	4- 7
DR. Julian Cheung	New York Inst. of Technology , New York, , NY On Classification of Multispectral Infrared Image	RL/OCTM	4- 8
DR Ajit K Choudhury	Howard University , Washington , DC Detection Performance of Over Resolved Targets wit	RL/OCTS	4- 9
Dr. Eric Donkor	University of Connecticut , Stroes , CT Experimental Measurement of Nonlinear Effects in	RL/OCPA	4- 10
DR. Frances J Harackiewicz	So. Illinois Univ-Carbondale , Carbondale , IL Circular Waveguide to Microstrip Line Transition	RL/ERA	4- 11

SRP Final Report Table of Contents

Author	University/Institution Report Title	Rome Laboratory Directorate	Vol-Page
DR Joseph W Haus	Rensselaer Polytechnic Inst , Troy , NY Simulation of Erbium-doped Fiber Lasers	RL/OCP	4- 12
DR Yolanda J Kime	SUNY College-Cortland , Cortland , NY A Macroscopic Model of Electromigration: Comparis	RL/ERDR	4- 13
DR. Phillip G Kornreich	Syracuse University , Syracuse , NY Semiconductor Cylinder Fibers for Fiber Light Ampl	RL/OCP	4- 14
DR Guifang Li	Rochester Institute of Tech , Rochester , NY Self-Pulsation and Optoelectronic Feedback-Sustain	RL/OCP	4- 15
Dr. Beth L Losiewicz	Colorado State University , Fort Collins , CO Preliminary Report on the Feasibility of Machine S	RL/IR	4- 16
DR. Mohamad T Musavi	University of Maine , Orono , ME Automatic Extraction of Drainage Network from Di	RL/IR	4- 17
DR John D Norgard	Univ of Colorado-Colorado Sprg , Colorado Springs , CO Infrared Images of Electromagnetic Fields	RL/ERPT	4- 18
DR Michael A Pittarelli	SUNY Institute of Technology , Utica , NY Anytime Inference and Decision Methods	RL/C3CA	4- 19
DR Dean Richardson	SUNY Institute of Technology , Utica , NY Ultrafast Spectroscopy of Quantum Heterostructures	RL/OCP	4- 20
DR. Daniel F Ryder, Jr.	Tufts University , Medford , MA Synthesis and Properties of B-Diketonate-Modified	RL/ERX	4- 21
DR Gregory J Salamo	University of Arkansas , Fayetteville , AR Photorefractive Development and Application of InP	RL/ERX	4- 22

Author	University/Institution Report Title	Rome Laboratory Directorate	Vol-Page
Dr. Scott E Spetka	SUNY, Institute of Technology , Utica , NY The TkWWW Robot: Beyond Browsing	RL/IR	4- 23
DR James C West	Oklahoma State University , Stillwater , OK Polarimetric Radar Scattering from a Vegation Can	RL/ERC	4- 24
DR Rolf T Wigand	Syracuse University , Syracuse , NY Transferring Technology Via the Internet	RL/XP	4- 25
Dr. Xi-Cheng Zhang	Rensselaer Polytechnic Institu , Troy , NY Temperature Dependence of THz Emission for <111> G	RL/ERX	4- 26

SRP Final Report Table of Contents

Author	University/Institution Report Title	Wright Laboratory Directorate	Vol-Page
DR Sunil K Agrawal	Ohio Univeristy , Athens , OH A Study of Preform Design Problem for Metal Deform	WL/MLIM	5- 1
DR Michael E Baginski	Auburn University , Auburn , AL Calculation of Heating and Temperature Distributio	WL/MNMF	5- 2
Dr. William W Bannister	Univ of Massachusetts-Lowell , Lowell , MA Anomalous Effects of Water in Fire Firefighting:	WL/FIVC	5- 3
Mr. Larry A Beardsley	Athens State College , Athens , AL RFSIG Target Model Intergrated With the Joint Mode	WL/MNSH	5- 4
DR Thomas L Beck	McMicken Coll of Arts & Sci , , OH Multigrid Method for Large Scale Electronic Struct	WL/MLPJ	5- 5
DR Victor L Berdichevsky	Wayne State University , Detroit , MI Diffusional Creep in Metals and Ceramics at High T	WL/FIB	5- 6
DR. Steven W Buckner	Colullmbus College , Columbus , GA Quantitation of Dissolved O2 in Aviation Fuels by	WL/POSF	5- 7
DR. James J Carroll	Clarkson University , Potsdam , NY Development of an Active Dynamometer System	WL/POOC-	5- 8
Dr. Ching L Chang	Cleveland State University , Cleveland , OH Least-Squares Finite Element Methods for Incompres	WL/FIMM	5- 9
Dr. David B Choate	Transylvania University , Lexington , KY A New Superposition	WL/AAWP	5- 10
DR Stephen J Clarson	University of Cincinnati , Cincinnati , OH Synthesis of Novel Second and Third Order Nonlinea	WL/MLBP	5- 11

Author	University/Institution Report Title	Wright Laboratory Directorate	Vol-Page
Dr. Milton L Cone	Embry-Riddel Aeronautical Univ , Prescott , AZ The Sensor Manager Puzzle	WL/AAAS-	5- 12
DR Robert W Courter	Louisiana State University , Baton Rouge , LA A Research Plan for Evaluating Wavegun as a Low-Lo	WL/MNAA	5- 13
DR Vinay Dayal	Iowa State University , Ames , IA Longitudinal Waves in Fluid Loaded Composite Fiber	WL/MLLP	5- 14
DR Jeffrey C Dill	Ohio University , Athens , OH Discrete Wavelet Transforms for Communication Sign	WL/AAW	5- 15
DR Vincent G Dominic	University of Dayton , Dayton , OH Electro-Optic Characterization of Poled-Polymer Fi	WL/MLPO	5- 16
DR Franklin E Eastep	University of Dayton , Dayton , OH Influence of Mode Complexity and Aeroseleastci Con	WL/FIBR	5- 17
DR Georges M Fadel	Clemson University , Clemson , SC A Methodology for Affordability in the Design Proc	WL/MTR	5- 18
Dr. Joel R Fried	University of Cincinnati , Cincinnati , OH Computer Modeling of Electrolytes for Battery Appl	WL/POOS-	5- 19
DR Paul D Gader	University of Missouri-Columbi , Columbia , MO Scanning Image Algebra Networks for Vehicle Identi	WL/MNGA	5- 20
DR Philip Gatt	University of Central Florida , Orlando , FL Laser Radar Performance Modelling and Analysis wit	WL/MNGS	5- 21
Dr. Richard D Gould	North Carolina State Univ , Raleigh , NC Analysis of Laser Doppler Velocimetry Data	WL/POPT	5- 22

SRP Final Report Table of Contents

Author	University/Institution Report Title	Wright Laboratory Directorate	Vol-Page
Dr. Raghava G Gowda	University of Dayton , Dayton , OH Issues Involved in Developing an Object-oriented S	WL/AAAS-	5- 23
DR Guoxiang Gu	Louisiana State University , Baton Rouge , LA Gain Scheduled Missile Autopilot Design Using Obse	WL/MNAG	5- 24
Dr Venkata S Gudimetla	OGI , Portland , OR Thermal Modeling of Heterojunction Bipolar Transis	WL/ELMT	5- 25
Dr. Raimo J Hakkinen	Washington University , St. Louis , MO Further Development of Surface-Obstacle Instrument	WL/FIMN	5- 26
DR Russell C Hardie	Univsity of Dayton , Dayton , OH Adaptive Quadratic Classifiers for Multispectral T	WL/AARA	5- 27
DR Larry S Helmick	Cedarville College , Cedarville , OH Effect of Humidity on Friction and Wear for Fombli	WL/MLBT	5- 28
DR Alan S Hodel	Auburn University , Auburn , AL Automatic Control Issues in the Development of an	WL/MNAG	5- 29
DR Vinod K Jain	University of Dayton , Dayton , OH Can Design for Cogging of Titanium Aluminide Alloy	WL/MLLN	5- 30
DR Jonathan M Janus	Mississippi State University , Mississippi State , MS Multidemensional Algorithm Development and Analysi	WL/MNAA	5- 31
DR Iwona M Jasiuk	Michigan State University , East Lansing , MI Characterization of Interfaces in Metal Matrix Com	WL/WLL	5- 32
Dr. Jack S Jean	Wright State University , Dayton , OH Reed-Solomon Decoding on Champ Architecture	WL/AAAT-	5- 33

Author	University/Institution Report Title	Wright Laboratory Directorate	Vol-Page
Dr. Ismail I Jouny	Lafayette College , Easton , PA Modeling and Mitigation of Terrain Scattered Inter	WL/AARM	5- 34
DR Tribikram Kundu	University of Arizona , Tucson , AZ Lamb Wave Scanning of a Multilayed Composite Plate	WL/MLLP	5- 35
DR. Jian Li	University of Florida , Gainesville , FL HIgh Resolution Range Signature Estimation	WL/AARA	5- 36
DR. Chun-Shin Lin	University of Missouri-Columbi , Columbia , MO Prediction of Missile Trajectory	WL/FIPA	5- 37
Dr. Paul P Lin	Cleveland State University , Cleveland , OH Three Dimensional Geometry Measurement of Tire Def	WL/FIVM	5- 38
Dr. Juin J Liou	University of Central Florida , Orlando , FL A Model to Monitor the Current Gain Long-Term Inst	WL/ELRD	5- 39
Dr. James S Marsh	University of West Florida , Pensacola , FL Numerical Reconstruction of Holograms in Advanced	WL/MNSI	5- 40
DR Rajiv Mehrotra	Univ. of Missouri-St. Louis , St. Louis , MO Integrated Information Management for ATR Research	WL/AARA	5- 4
DR Douglas J Miller	Cedarville College , Cedarville , OH A Review of Nonfilled Intrinsically Conductive Ela	WL/MLBP	5- 4
DR Nagaraj Nandhakumar	University of Virginia , Charlottesville , VA Thermophysical Affine Invariants from IR Imagery	WL/AARA	5- 4
Dr. M. G Norton	Washington State University , Pullman , WA Surface Outgrowths on Laser-Deposited YBa2Cu3O7 Th	WL/MLPO	5- 4

Author	University/Institution Report Title	Wright Laboratory Directorate	Vol-Page
DR. James F O'Brien	Southwest Missouri State Univ. , Springfield , MO The Importance of Lower Orbital Relaxations in Po	WL/MLBP	5- 45
DR Krishna M Pasala	University of Dayton , Dayton , OH Performance of Music and Monopulse Algorithms in t	WL/AARM	5- 46
DR Robert P Penno	University of Dayton , Dayton , OH An Assessment of the WL/AAAI-4 Antenna Wavefront S	WL/AAAI-	5- 47
DR Marek A Perkowski	Portland State University , Portland , OR A Survey of Literature on Function Decomposition	WL/AAAT-	5- 48
DR Ramachandran Radharamanan	Marquette University , Milwaukee , WI A Study on Virtual Manufacturing	WL/MTI	5- 49
DR Ramu V Ramaswamy	University of Florida , Gainesville , FL Annealed Proton Exchanged (APE) Waveguides in LiTa	WL/MNG	5- 50
DR Stanley J Reeves	Auburn University , Auburn , AL Superresolution of Passive Millimeter-Wave Imaging	WL/MNGS	5- 51
Dr. William K Rule	University of Alabama , Tuscaloosa , AL <RESTRICTED DISTRIBUTION - CONTACT LABORATORY>	WL/MNM	5- 52
DR Arindam Saha	Mississippi State University , Mississippi State , MS Evaluation of Network Routers in Real-Time Paralle	WL/AAAT-	5- 53
DR John J Schauer	University of Dayton , Dayton , OH Turbine Blade Film Jet Cooling with Free Stream Tu	WL/POTT	5- 54
DR Carla A Schwartz	University of Florida , Gainesville , FL Neural Networks Identification and Control in Meta	WL/FIGC	5- 55

Author	University/Institution Report Title	Wright Laboratory Directorate	Vol-Page
DR. James P Seaba	University of Missouri-Columbi , Columbia , MO Multiple Jet Mixing and Atomization in Reacting an	WL/POSF	5- 56
DR Sivanand Simanapalli	University of NC-Charlotte , Charlotte , NC HRR Radar Based Target Identification	WL/AARA	5- 57
DR. Terrence W Simon	University of Minnesota , Minneapolis , MN Documentation of Boundary Layer Characteristics Fo	WL/POTT	5- 58
DR Marek Skowronski	Carnegie Melon University , Pittsburgh , PA Mechanism for Indium Segregation In InxGa1-xAs Str	WL/ELRA	5- 59
DR Joseph C Slater	Wright State Univesity , Dayton , OH QFT Control of an Advanced Tactical Fighter Aeroel	WL/FIGS	5- 60
DR John A Tague	Ohio University , Athens , OH Performance Analysis of Quadratic Classifiers for	WL/AARA	5- 61
Dr. Barney E Taylor	Miami Univ. - Hamilton , Hamilton , OH Electroluminescence Studies of the Rigid Rod Polym	WL/MLBP	5- 62
DR Krishnaprasad Thirunarayan	Wright State University , Dayton , OH VHDL-93 Paser in Prolog	WL/ELED	5- 63
DR Robert B Trelease	University of California , Los Angeles , CA Developing Qualitative Process Control Discovery S	WL/MLIM	5- 64
DR. Chi-Tay Tsai	Florida Atlantic University , Boca Raton , FL A Study of Massively Parallel Computing on Epic Hy	WL/MNM	5- 65
DR James M Whitney	University of Dayton , Dayton , OH Stress Analysis of the V-Notch (Iosipescu) Shear T	WL/MLBM	5- 66

SRP Final Report Table of Contents

Author	University/Institution Report Title	Arnold Engineering Development Center Directorate	Vol-Page
DR Ben A Abbott	Vanderbilt University , Nashville , TN The Application Challenge	Sverdrup	6- 1
DR Theodore A Bapty	Vanderbilt University , Nashville , TN Development of Large Parallel Instrumentation Syst	Sverdrup	6- 2
Dr. Csaba A Biegl	Vanderbilt University , Nashville , TN Univeral Graphic User Inteface for Turbine Engine	Sverdrup	6- 3
DR Steven H Frankel	Purdue University , West Lafayette , IN Towards The Computational Modeling of Postall Gas	Sverdrup	6- 4
Dr. Peter R Massopust	Sam Houston State University , Huntsville , TX A Wavelet-Multigrid Approach To Solving Partial Di	Calspan	6- 5
DR Randolph S Peterson	University of the South , Sewanee , TN Infrared Imaging Fourier Transform Spectometer	Sverdrup	6- 6
DR Roy J Schulz	Univ of Tennessee Space Inst , Tullahoma , TN Design of Soot Capturing Sample Probe	Sverdrup	6- 7
DR S A Sherif	College of Eng-Univ of Florida , Gainesville , FL A Model For Local Heat Transfer & Ice Accretion In	Sverdrup	6- 8
DR. Michael Sydor	University of Minnesota-Duluth , Duluth , MN Dimensional Analysis of ARC Heaters	Calspan	6- 9
Dr. John T Tarvin	Samford University , Birmingham , AL Ultraviolet Flat-Field Response of an Intensified	CALSPAN	6- 10

Author	University/Institution Report Title	Frank J Seiler Research Laboratory Directorate	Vol-Page
Dr. Gene O Carlisle	West Texas State University , Canyon , TX REPORT NOT AVAILABLE AT PRESS TIME	FJSRL/ NC	6- 11
DR John R Dorgan	Colorado School of Mines , Golden , CO Fundamental Studies on the Solution and Adsorption	FJSRL/NE	6- 12
DR Mary Ann Jungbauer	Barry University , Miami , FL Non-Linear Optical Properties of a Series of Linea	FJSRL/NC	6- 13
DR. Lawrence L Murrell	Pennsylvania State University , University Park , PA Catalytic Gasification of Pitch Carbon Fibers with	FJSRL/NE	6- 14
DR David E Statman	Allegheny College , Meadville , PA Charge Transport and Second Harmonic Generation in	FJSRL/NP	6- 15

Author	University/Institution Report Title	Wilford Hall Medical Center Directorate	Vol-Page
DR Walter Drost-Hansen	University of Miami , Coral Gables , FL Effects of Temperature on Various Hematological Pa	WHMC/RD	6- 16

1. INTRODUCTION

The Summer Research Program (SRP), sponsored by the Air Force Office of Scientific Research (AFOSR), offers paid opportunities for university faculty, graduate students, and high school students to conduct research in U.S. Air Force research laboratories nationwide during the summer.

Introduced by AFOSR in 1978, this innovative program is based on the concept of teaming academic researchers with Air Force scientists in the same disciplines using laboratory facilities and equipment not often available at associates' institutions.

AFOSR also offers its research associates 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. In 1994 the maximum amount of each grant was increased from $20,000 to $25,000, and the number of AFOSR-sponsored grants decreased from 75 to 60. A separate annual report is compiled on the SREP.

The Summer Faculty Research Program (SFRP) is open annually to approximately 150 faculty members with at least two years of teaching and/or research experience in accredited U.S. colleges, universities, or technical institutions. SFRP associates must be either U.S. citizens or permanent residents.

The Graduate Student Research Program (GSRP) is open annually to approximately 100 graduate students holding a bachelor's or a master's degree; GSRP associates must be U.S. citizens enrolled full time at an accredited institution.

The High School Apprentice Program (HSAP) annually selects about 125 high school students located within a twenty mile commuting distance of participating Air Force laboratories.

The numbers of projected summer research participants in each of the three categories are usually increased through direct sponsorship by participating laboratories.

AFOSR's SRP has well served its objectives of building critical links between Air Force research laboratories and the academic community, opening avenues of communications and forging new research relationships between Air Force and academic technical experts in areas of national interest; and strengthening the nation's efforts to sustain careers in science and engineering. The success of the SRP can be gauged from its growth from inception (see Table 1) and from the favorable responses the 1994 participants expressed in end-of-tour SRP evaluations (Appendix B).

AFOSR contracts for administration of the SRP by civilian contractors. The contract was first awarded to Research & Development Laboratories (RDL) in September 1990. After completion of the 1990 contract, RDL won the recompetition for the basic year and four 1-year options.

2. PARTICIPATION IN THE SUMMER RESEARCH PROGRAM

The SRP began with faculty associates in 1979; graduate students were added in 1982 and high school students in 1986. The following table shows the number of associates in the program each year.

Table 1: SRP Participation, by Year

YEAR	Number of Participants			TOTAL
	SFRP	GSRP	HSAP	
1979	70			70
1980	87			87
1981	87			87
1982	91	17		108
1983	101	53		154
1984	152	84		236
1985	154	92		246
1986	158	100	42	300
1987	159	101	73	333
1988	153	107	101	361
1989	168	102	103	373
1990	165	121	132	418
1991	170	142	132	444
1992	185	121	159	464
1993	187	117	136	440
1994	192	117	133	442

Beginning in 1993, due to budget cuts, some of the laboratories weren't able to afford to fund as many associates as in previous years; in one case a laboratory did not fund any additional associates. However, the table shows that, overall, the number of participating associates increased this year because two laboratories funded more associates than they had in previous years.

3. RECRUITING AND SELECTION

The SRP is conducted on a nationally advertised and competitive-selection basis. The advertising for faculty and graduate students consisted primarily of the mailing of 8,000 44-page SRP brochures to chairpersons of departments relevant to AFOSR research and to administrators of grants in accredited universities, colleges, and technical institutions. Historically Black Colleges and Universities (HBCUs) and Minority Institutions (MIs) were included. Brochures also went to all participating USAF laboratories, the previous year's participants, and numerous (over 600 annually) individual requesters.

Due to a delay in awarding the new contract, RDL was not able to place advertisements in any of the following publications in which the SRP is normally advertised: *Black Issues in Higher Education, Chemical & Engineering News, IEEE Spectrum* and *Physics Today*.

High school applicants can participate only in laboratories located no more than 20 miles from their residence. Tailored brochures on the HSAP were sent to the head counselors of 180 high schools in the vicinity of participating laboratories, with instructions for publicizing the program in their schools. High school students selected to serve at Wright Laboratory's Armament Directorate (Eglin Air Force Base, Florida) serve eleven weeks as opposed to the eight weeks normally worked by high school students at all other participating laboratories.

Each SFRP or GSRP applicant is given a first, second, and third choice of laboratory. High school students who have more than one laboratory or directorate near their homes are also given first, second, and third choices.

Laboratories make their selections and prioritize their nominees. AFOSR then determines the number to be funded at each laboratory and approves laboratories' selections.

Subsequently, laboratories use their own funds to sponsor additional candidates. Some selectees do not accept the appointment, so alternate candidates are chosen. This multi-step selection procedure results in some candidates being notified of their acceptance after scheduled deadlines. The total applicants and participants for 1994 are shown in this table.

Table 2: 1994 Applicants and Participants

PARTICIPANT CATEGORY	TOTAL APPLICANTS	SELECTEES	DECLINING SELECTEES
SFRP	600	192	30
(HBCU/MI)	(90)	(16)	(7)
GSRP	322	117	11
(HBCU/MI)	(11)	(6)	(0)
HSAP	562	133	14
TOTAL	1484	442	55

4. SITE VISITS

During June and July of 1994, representatives of both AFOSR/NI and RDL visited each participating laboratory to provide briefings, answer questions, and resolve problems for both laboratory personnel and participants. The objective was to ensure that the SRP would be as constructive as possible for all participants. Both SRP participants and RDL representatives found these visits beneficial. At many of the laboratories, this was the only opportunity for all participants to meet at one time to share their experiences and exchange ideas.

5. HISTORICALLY BLACK COLLEGES AND UNIVERSITIES AND MINORITY INSTITUTIONS (HBCU/MIs)

In previous years, an RDL program representative visited from seven to ten different HBCU/MIs to promote interest in the SRP among the faculty and graduate students. Due to the late contract award date (January 1994) no time was available to visit HBCU/MIs this past year.

In addition to RDL's special recruiting efforts, AFOSR attempts each year to obtain additional funding or use leftover funding from cancellations the past year to fund HBCU/MI associates. This year, seven HBCU/MI SFRPs declined after they were selected. The following table records HBCU/MI participation in this program.

Table 3: SRP HBCU/MI Participation, by Year

YEAR	SFRP		GSRP	
	Applicants	Participants	Applicants	Participants
1985	76	23	15	11
1986	70	18	20	10
1987	82	32	32	10
1988	53	17	23	14
1989	39	15	13	4
1990	43	14	17	3
1991	42	13	8	5
1992	70	13	9	5
1993	60	13	6	2
1994	90	16	11	6

6. SRP FUNDING SOURCES

Funding sources for the 1994 SRP were the AFOSR-provided slots for the basic contract and laboratory funds. Funding sources by category for the 1994 SRP selected participants are shown here.

Table 4: 1994 SRP Associate Funding

FUNDING CATEGORY	SFRP	GSRP	HSAP
AFOSR Basic Allocation Funds	150	98*1	121*2
USAF Laboratory Funds	37	19	12
HBCU/MI By AFOSR (Using Procured Addn'l Funds)	5	0	0
TOTAL	192	117	133

*1 - 100 were selected, but two canceled too late to be replaced.
*2 - 125 were selected, but four canceled too late to be replaced.

7. COMPENSATION FOR PARTICIPANTS

Compensation for SRP participants, per five-day work week, is shown in this table.

Table 5: 1994 SRP Associate Compensation

PARTICIPANT CATEGORY	1991	1992	1993	1994
Faculty Members	$690	$718	$740	$740
Graduate Student (Master's Degree)	$425	$442	$455	$455
Graduate Student (Bachelor's Degree)	$365	$380	$391	$391
High School Student (First Year)	$200	$200	$200	$200
High School Student (Subsequent Years)	$240	$240	$240	$240

The program also offered associates whose homes were more than 50 miles from the laboratory an expense allowance (seven days per week) of $50/day for faculty and $37/day for graduate students. Transportation to the laboratory at the beginning of their tour and back to their home destinations at the end was also reimbursed for these participants. Of the combined SFRP and GSRP associates, 58% (178 out of 309) claimed travel reimbursements at an average round-trip cost of $860.

Faculty members were encouraged to visit their laboratories before their summer tour began. All costs of these orientation visits were reimbursed. Forty-one percent (78 out of 192) of faculty associates took orientation trips at an average cost of $498. Many faculty associates noted on their evaluation forms that due to the late notice of acceptance into the 1994 SRP (caused by the late award in January 1994 of the contract) there wasn't enough time to attend an orientation visit prior to their tour start date. In 1993, 58 % of SFRP associates took orientation visits at an average cost of $685.

Program participants submitted biweekly vouchers countersigned by their laboratory research focal point, and RDL issued paychecks so as to arrive in associates' hands two weeks later.

HSAP program participants were considered actual RDL employees, and their respective state and federal income tax and Social Security were withheld from their paychecks. By the nature of their independent research, SFRP and GSRP program participants were considered to be consultants or independent contractors. As such, SFRP and GSRP associates were responsible for their own income taxes, Social Security, and insurance.

8. CONTENTS OF THE 1994 REPORT

The complete set of reports for the 1994 SRP includes this program management report augmented by fifteen volumes of final research reports by the 1994 associates as indicated below:

Table 6: 1994 SRP Final Report Volume Assignments

	VOLUME		
LABORATORY	SFRP	GSRP	HSAP
Armstrong	2	7	12
Phillips	3	8	13
Rome	4	9	14
Wright	5A, 5B	10	15
AEDC, FJSRL, WHMC	6	11	16

AEDC = Arnold Engineering Development Center
FJSRL = Frank J. Seiler Research Laboratory
WHMC = Wilford Hall Medical Center

APPENDIX A -- PROGRAM STATISTICAL SUMMARY

A. Colleges/Universities Represented

Selected SFRP and GSRP associates represent 158 different colleges, universities, and institutions.

B. States Represented

SFRP -Applicants came from 46 states plus Washington D.C. and Puerto Rico. Selectees represent 40 states.

GSRP - Applicants came from 46 states and Puerto Rico. Selectees represent 34 states.

HSAP - Applicants came from fifteen states. Selectees represent ten states.

C. Academic Disciplines Represented

The academic disciplines of the combined 192 SFRP associates are as follows:

Electrical Engineering	22.4%
Mechanical Engineering	14.0%
Physics: General, Nuclear & Plasma	12.2%
Chemistry & Chemical Engineering	11.2%
Mathematics & Statistics	8.1%
Psychology	7.0%
Computer Science	6.4%
Aerospace & Aeronautical Engineering	4.8%
Engineering Science	2.7%
Biology & Inorganic Chemistry	2.2%
Physics: Electro-Optics & Photonics	2.2%
Communication	1.6%
Industrial & Civil Engineering	1.6%
Physiology	1.1%
Polymer Science	1.1%
Education	0.5%
Pharmaceutics	0.5%
Veterinary Medicine	0.5%
TOTAL	100%

Table A-1. Total Participants

Number of Participants	
SFRP	192
GSRP	117
HSAP	133
TOTAL	442

Table A-2. Degrees Represented

Degrees Represented			
	SFRP	GSRP	TOTAL
Doctoral	189	0	189
Master's	3	47	50
Bachelor's	0	70	70
TOTAL	192	117	309

Table A-3. SFRP Academic Titles

Academic Titles	
Assistant Professor	74
Associate Professor	63
Professor	44
Instructor	5
Chairman	1
Visiting Professor	1
Visiting Assoc. Prof.	1
Research Associate	3
TOTAL	192

Table A-4. Source of Learning About SRP

SOURCE	SFRP		GSRP	
	Applicants	Selectees	Applicants	Selectees
Applied/participated in prior years	26%	37%	10%	13%
Colleague familiar with SRP	19%	17%	12%	12%
Brochure mailed to institution	32%	18%	19%	12%
Contact with Air Force laboratory	15%	24%	9%	12%
Faculty Advisor (GSRPs Only)	--	--	39%	43%
Other source	8%	4%	11%	8%
TOTAL	100%	100%	100%	100%

Table A-5. Ethnic Background of Applicants and Selectees

	SFRP		GSRP		HSAP	
	Applicants	Selectees	Applicants	Selectees	Applicants	Selectees
American Indian or Native Alaskan	0.2%	0%	1%	0%	0.4%	0%
Asian/Pacific Islander	30%	20%	6%	8%	7%	10%
Black	4%	1.5%	3%	3%	7%	2%
Hispanic	3%	1.9%	4%	4.5%	11%	8%
Caucasian	51%	63%	77%	77%	70%	75%
Preferred not to answer	12%	14%	9%	7%	4%	5%
TOTAL	100%	100%	100%	100%	99%	100%

Table A-6. Percentages of Selectees receiving their 1st, 2nd, or 3rd Choices of Directorate

	1st Choice	2nd Choice	3rd Choice	Other Than Their Choice
SFRP	70%	7%	3%	20%
GSRP	76%	2%	2%	20%

APPENDIX B -- SRP EVALUATION RESPONSES

1. OVERVIEW

Evaluations were completed and returned to RDL by four groups at the completion of the SRP. The number of respondents in each group is shown below.

Table B-1. Total SRP Evaluations Received

Evaluation Group	Responses
SFRP & GSRPs	275
HSAPs	116
USAF Laboratory Focal Points	109
USAF Laboratory HSAP Mentors	54

All groups indicate near-unanimous enthusiasm for the SRP experience.

Typical comments from 1994 SRP associates are:

> "[The SRP was an] excellent opportunity to work in state-of-the-art facility with top-notch people."
>
> "[The SRP experience] enabled exposure to interesting scientific application problems; enhancement of knowledge and insight into 'real-world' problems."
>
> "[The SRP] was a great opportunity for resourceful and independent faculty [members] from small colleges to obtain research credentials."
>
> "The laboratory personnel I worked with are tremendous, both personally and scientifically. I cannot emphasize how wonderful they are."
>
> "The one-on-one relationship with my mentor and the hands on research experience improved [my] understanding of physics in addition to improving my library research skills. Very valuable for [both] college and career!"

Typical comments from laboratory focal points and mentors are:

"This program [AFOSR - SFRP] has been a 'God Send' for us. Ties established with summer faculty have proven invaluable."

"Program was excellent from our perspective. So much was accomplished that new options became viable "

"This program managed to get around most of the red tape and 'BS' associated with most Air Force programs. Good Job!"

'Great program for high school students to be introduced to the research environment. Highly educational for others [at laboratory]."

'This is an excellent program to introduce students to technology and give them a feel for [science/engineering] career fields. I view any return benefit to the government to be 'icing on the cake' and have usually benefitted."

The summarized recommendations for program improvement from both associates and laboratory personnel are listed below (Note: basically the same as in previous years.)

A. Better preparation on the labs' part prior to associates' arrival (i.e., office space, computer assets, clearly defined scope of work).

B. Laboratory sponsor seminar presentations of work conducted by associates, and/or organized social functions for associates to collectively meet and share SRP experiences.

C. Laboratory focal points collectively suggest more AFOSR allocated associate positions, so that more people may share in the experience.

D. Associates collectively suggest higher stipends for SRP associates.

E. Both HSAP Air Force laboratory mentors and associates would like the summer tour extended from the current 8 weeks to either 10 or 11 weeks; the groups state it takes 4-6 weeks just to get high school students up-to-speed on what's going on at laboratory. (Note: this same arguement was used to raise the faculty and graduate student participation time a few years ago.)

2. 1994 USAF LABORATORY FOCAL POINT (LFP) EVALUATION RESPONSES

The summarized results listed below are from the 109 LFP evaluations received.

1. LFP evaluations received and associate preferences:

Table B-2. Air Force LFP Evaluation Responses (By Type)

		How Many Associates Would You Prefer To Get ? (% Response)											
		SFRP				GSRP (w/Univ Professor)				GSRP (w/o Univ Professor)			
Lab	Evals Recv'd	0	1	2	3+	0	1	2	3+	0	1	2	3+
AEDC	10	30	50	0	20	50	40	0	10	40	60	0	0
AL	44	34	50	6	9	54	34	12	0	56	31	12	0
FJSRL	3	33	33	33	0	67	33	0	0	33	67	0	0
PL	14	28	43	28	0	57	21	21	0	71	28	0	0
RL	3	33	67	0	0	67	0	33	0	100	0	0	0
WHMC	1	0	0	100	0	0	100	0	0	0	100	0	0
WL	46	15	61	24	0	56	30	13	0	76	17	6	0
Total	121	25%	43%	27%	4%	50%	37%	11%	1%	54%	43%	3%	0%

LFP Evaluation Summary. The summarized repsonses, by laboratory, are listed on the following page. LFPs were asked to rate the following questions on a scale from 1 (below average) to 5 (above average).

2. LFPs involved in SRP associate application evaluation process:
    a. Time available for evaluation of applications:
    b. Adequacy of applications for selection process:
3. Value of orientation trips:
4. Length of research tour:
5. a. Benefits of associate's work to laboratory:
    b. Benefits of associate's work to Air Force:
6. a. Enhancement of research qualifications for LFP and staff:
    b. Enhancement of research qualifications for SFRP associate:
    c. Enhancement of research qualifications for GSRP associate:
7. a. Enhancement of knowledge for LFP and staff:
    b. Enhancement of knowledge for SFRP associate:
    c. Enhancement of knowledge for GSRP associate:
8. Value of Air Force and university links:
9. Potential for future collaboration:
10. a. Your working relationship with SFRP:
    b. Your working relationship with GSRP:
11. Expenditure of your time worthwhile:

(Continued on next page)

12. Quality of program literature for associate:
13. a. Quality of RDL's communications with you:
    b. Quality of RDL's communications with associates:
14. Overall assessment of SRP:

Laboratory Focal Point Reponses to above questions

	AEDC	*AL*	*FJSRL*	*PL*	*RL*	*WHMC*	*WL*
# Evals Recv'd	10	32	3	14	3	1	46
Question #							
2	90 %	62 %	100 %	64 %	100 %	100 %	83 %
2a	**3.5**	**3.5**	4.7	4.4	4.0	4.0	**3.7**
2b	4.0	3.8	4.0	4.3	4.3	4.0	3.9
3	4.2	3.6	4.3	3.8	4.7	4.0	4.0
4	3.8	3.9	4.0	4.2	4.3	NO ENTRY	4.0
5a	4.1	4.4	4.7	4.9	4.3	3.0	4.6
5b	4.0	4.2	4.7	4.7	4.3	3.0	4.5
6a	**3.6**	4.1	**3.7**	4.5	4.3	**3.0**	4.1
6b	3.6	4.0	4.0	4.4	4.7	3.0	4.2
6c	3.3	4.2	4.0	4.5	4.5	3.0	4.2
7a	3.9	4.3	4.0	4.6	4.0	3.0	4.2
7b	4.1	4.3	4.3	4.6	4.7	3.0	4.3
7c	3.3	4.1	4.5	4.5	4.5	5.0	4.3
8	4.2	4.3	5.0	4.9	4.3	5.0	4.7
9	3.8	4.1	4.7	5.0	4.7	5.0	4.6
10a	4.6	4.5	5.0	4.9	4.7	5.0	4.7
10b	4.3	4.2	5.0	4.3	5.0	5.0	4.5
11	4.1	4.5	4.3	4.9	4.7	4.0	4.4
12	4.1	3.9	4.0	4.4	4.7	3.0	4.1
13a	**3.8**	**2.9**	4.0	4.0	4.7	**3.0**	**3.6**
13b	**3.8**	**2.9**	4.0	4.3	4.7	**3.0**	**3.8**
14	4.5	4.4	5.0	4.9	4.7	4.0	4.5

3. 1994 SFRP & GSRP EVALUATION RESPONSES

The summarized results listed below are from the 275 SFRP/GSRP evaluations received.

Associates were asked to rate the following questions on a scale from 1 (below average) to 5 (above average)

1. The match between the laboratories research and your field:		4.6
2. Your working relationship with your LFP:		4.8
3. Enhancement of your academic qualifications:		4.4
4. Enhancement of your research qualifications:		4.5
5. Lab readiness for you: LFP, task, plan:		4.3
6. Lab readiness for you: equipment, supplies, facilities:		4.1
7. Lab resources:		4.3
8. Lab research and administrative support:		4.5
9. Adequacy of brochure and associate handbook:		4.3
10. RDL communications with you:		4.3
11. Overall payment procedures:		**3.8**
12. Overall assessment of the SRP:		4.7
13. a. Would you apply again?	Yes:	85%
b. Will you continue this or related research?	Yes:	95%
14. Was length of your tour satisfactory?	Yes:	86%
15. Percentage of associates who engaged in:		
a. Seminar presentation:		52%
b. Technical meetings:		32%
c. Social functions:		03%
d. Other		01%

16. Percentage of associates who experienced difficulties in:

a. Finding housing:	12%
b. Check Cashing:	03%

17. Where did you stay during your SRP tour?

a. At Home:	20%
b. With Friend:	06%
c. On Local Economy:	47%
d. Base Quarters:	10%

THIS SECTION FACULTY ONLY:

18. Were graduate students working with you? Yes: 23%

19. Would you bring graduate students next year? Yes: 56%

20. Value of orientation visit:

Essential:	29%
Convenient:	20%
Not Worth Cost:	01%
Not Used:	34%

THIS SECTION GRADUATE STUDENTS ONLY:

21. Who did you work with:

University Professor:	18%
Laboratory Scientist:	54%

4. 1994 USAF LABORATORY HSAP MENTOR EVALUATION RESPONSES

The summarized results listed below are from the 54 mentor evaluations received.

1. Mentor apprentice preferences:

Table B-3. Air Force Mentor Responses

		How Many Apprentices Would You Prefer To Get ?			
		HSAP Apprentices Preferred			
Laboratory	***# Evals Recv'd***	***0***	**1**	**2**	**3+**
AEDC	6	0	100	0	0
AL	17	29	47	6	18
PL	9	22	78	0	0
RL	4	25	75	0	0
WL	18	22	55	17	6
Total	**54**	**20%**	**71%**	**5%**	**5%**

Mentors were asked to rate the following questions on a scale from 1 (below average) to 5 (above average)

2. Mentors involved in SRP apprentice application evaluation process:
    a. Time available for evaluation of applications:
    b. Adequacy of applications for selection process:
3. Laboratory's preparation for apprentice:
4. Mentor's preparation for apprentice:
5. Length of research tour:
6. Benefits of apprentice's work to U.S. Air force:
7. Enhancement of academic qualifications for apprentice:
8. Enhancement of research skills for apprentice:
9. Value of U.S. Air Force/high school links:
10. Mentor's working relationship with apprentice:
11. Expenditure of mentor's time worthwhile:
12. Quality of program literature for apprentice:
13. a. Quality of RDL's communications with mentors:
    b. Quality of RDL's communication with apprentices:
14. Overall assessment of SRP:

	AEDC	*AL*	*PL*	*RL*	*WL*
# Evals Recv'd	**6**	**17**	**9**	**4**	**18**
Question #					
2	100 %	76 %	56 %	75 %	61 %
2a	4.2	4.0	**3.1**	**3.7**	**3.5**
2b	4.0	4.5	4.0	4.0	3.8
3	4.3	**3.8**	**3.9**	**3.8**	**3.8**
4	4.5	**3.7**	**3.4**	4.2	**3.9**
5	**3.5**	4.1	**3.1**	**3.7**	**3.6**
6	4.3	3.9	4.0	4.0	4.2
7	4.0	4.4	4.3	4.2	3.9
8	4.7	4.4	4.4	4.2	4.0
9	4.7	4.2	3.7	4.5	4.0
10	4.7	4.5	4.4	4.5	4.2
11	4.8	4.3	4.0	4.5	4.1
12	4.2	4.1	4.1	4.8	3.4
13a	**3.5**	**3.9**	**3.7**	4.0	**3.1**
13b	4.0	4.1	3.4	4.0	3.5
14	4.3	4.5	3.8	4.5	4.1

5. 1994 HSAP EVALUATION RESPONSES

The summarized results listed below are from the 116 HSAP evaluations received.

HSAP apprentices were asked to rate the following questions on a scale from 1 (below average) to 5 (above average)

1. Match of lab research to you interest:		3.9
2. Apprentices working relationship with their mentor and other lab scientists:		4.6
3. Enhancement of your academic qualifications:		4.4
4. Enhancement of your research qualifications:		4.1
5. Lab readiness for you: mentor, task, work plan		**3.7**
6. Lab readiness for you: equipment supplies facilities		4.3
7. Lab resources: availability		4.3
8. Lab research and administrative support:		4.4
9. Adequacy of RDL's apprentice handbook and administrative materials:		4.0
10. Responsiveness of RDL's communications:		**3.5**
11. Overall payment procedures:		**3.3**
12. Overall assessment of SRP value to you:		4.5
13. Would you apply again next year?	Yes:	88%
14. Was length of SRP tour satisfactory?	Yes:	78%
15. Percentages of apprentices who engaged in:		
a. Seminar presentation:		48%
b. Technical meetings:		23%
c. Social functions:		18%

A Study of 'Preform Design Problem' for Metal Deformation Processes

Sunil Kumar Agrawal

Assistant Professor
Department of Mechanical Engineering
Ohio University, Athens, OH 45701.

Final Report for:
Summer Faculty Research Program
Materials Laboratory (MLIM)

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

and
Materials Laboratory (MLIM)

August, 1994

A Study of Preform Design Problem for Metal Deformation Processes

Sunil Kumar Agrawal

Assistant Professor
Department of Mechanical Engineering
Ohio University, Athens, OH 45701.

1 Abstract

Metal deformation is a complex phenomenon where externally applied forces on the boundary change the external shape and internal material properties. Metal deformation processes such as forging and extrusion are used very widely in industries to fabricate new and complex parts. From the point of view of design for near-net shape manufacturing of parts, the following question becomes very important: *What is the starting metal shape and the time history of externally applied boundary forces/velocities that will transform a given volume of metal to a desired final shape with desired material properties?*

My study focuses on a subset of the above mentioned problem, which is to find out the starting shape of the metal which for a given time history of the externally applied boundary forces/velocities will transform the metal to a desired final shape. This problem is also refered to as the 'Preform Design Problem' by the metal working community. Presently, this problem is addressed using a 'trial and error' approach. An initial shape is assumed by an experienced designer. It is then either modeled within an FEM simulation code or experimented upon in the laboratory to study the resulting final shape. This final shape is used to alter the initial geometry and the process is repeated until the designer is satisfied with the outcome. The following points about this 'trial and error' approach must be noted: (a) the process requires multiple iterations which could cost the designer many man-hours, (b) substitution of the actual experiment by FEM simulation could substantially reduce the design time, however, it is not uncommon for a single FEM run to take more than an hour, (c) for every new part to be fabricated, the 'trial and error' approach must be repeated.

In this study, a new framework is being suggested to address the Preform Design Problem which has the potential to reduce the design cycle time at least ten-fold when compared to FEM in the loop. In this framework, Boundary Element Method (BEM) is used for analysis. This analysis technique is coupled to a gradient based search algorithm for optimization. The BEM is naturally suited for studying the preform design problem since in this method, the discretizations must be done primarily at the boundary while in FEM, the entire domain must be discretized. Due to the need for fewer number of nodes in BEM as compared to FEM, the author believes that BEM has the potential to be a very efficient numerical tool for solving preform design problems.

A summary of my recommendations based on this study are as follows: (1) to develop BEM analysis codes for simulation of planar and axisymmetric deformation processes, (2) to compare the results of BEM code against the existing FEM codes, and (3) to develop a gradient based optimization module for studying and testing preform designs.

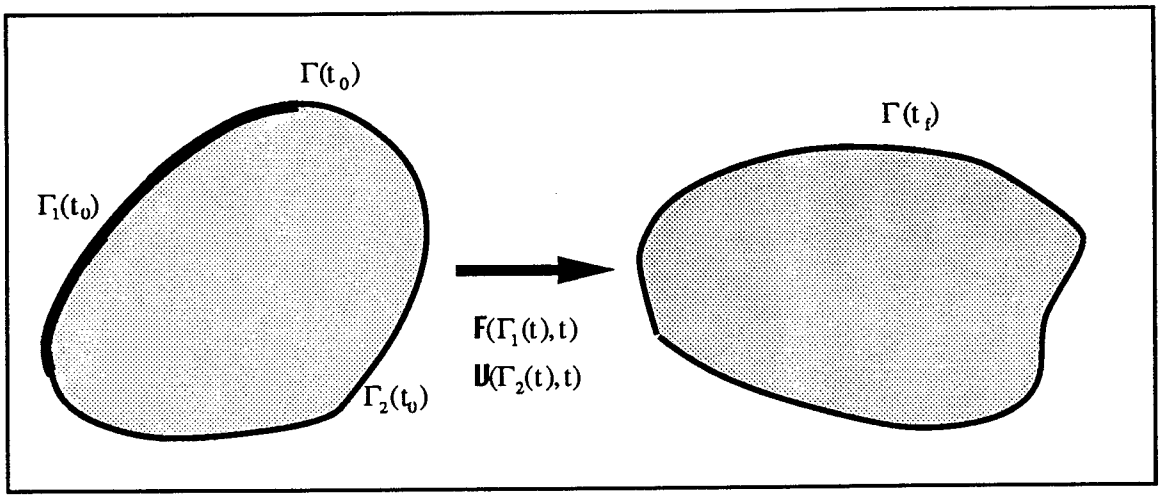


Figure 1: A metal forming process is a transformation on $(\mathcal{D}(t), \Gamma(t))$ due to the inputs $\mathbf{F}(\Gamma_1(t), t)$ and $\mathbf{V}(\Gamma_2(t), t)$.

2 Introduction

The organization of this report is as follows: Section 3 outlines the preform design problem. Section 4 describes the mathematical model of metal deformation processes and points out their salient features compared to elastic deformation processes. Section 5 motivates the study of Boundary Element Method for preform design problems. The details of the gradient based optimization scheme are listed in Section 6. An overview of boundary element method is presented in Section 7. These are followed by conclusion and recommendations for future work.

3 Preform Design Problem

Consider a piece of metal which is described by its internal domain $\mathcal{D}(t)$ and external boundary $\Gamma(t)$ at a time instance t. This piece of metal is subjected to prescribed external forces $\mathbf{F}(\Gamma_1(t), t)$ over a part of the boundary $\Gamma_1(t)$ and prescribed velocities $\mathbf{V}(\Gamma_2(t), t)$ over the the boundary $\Gamma_2(t)$. As a result of the time histories of these externally applied boundary forces and velocities, the metal changes shape and internal properties. A graphical representation of this transformation is shown in Figure 1.

From the perspective of the quality of a finished product, the following points are important: (i) the final shape of the metal must be very close to the desired final shape, i.e., $\Gamma(t_f) \simeq \Gamma_d(t_f)$, where t_f is the final time and Γ_d is the desired final shape, (ii) the internal material properties such as microstructural grain size and volume fraction must be close to the desired final properties.

From the perspective of control or regulation of the metal deformation process, in principle, the following quantities can be controlled: (1) the starting shape, i.e., $\Gamma(t_0)$, (2) the time history of the prescribed boundary forces $\mathbf{F}(\Gamma_1(t), t)$, and (3) the time history of the prescribed boundary velocities $\mathbf{V}(\Gamma_2(t), t)$. However, in reality, regulation of a process may become difficult due to the following reasons: (i) the boundary of the metal at an instance of time can not be predicted apriori and is dependent on the time history of the applied boundary conditions prior to that time, (ii) due to geometric limitations of the metal forming equipment and the instantaneous shape of the metal undergoing deformation, arbitrary choices of $\Gamma_1(t)$ and $\Gamma_2(t)$ may not be feasible for applying boundary force/velocity conditions, (iii) due to equipment limitations, it may not be possible to apply force/velocity conditions which are varying over the lengths of $\Gamma_1(t)$ and $\Gamma_2(t)$.

This study focuses on a subset of the above mentioned problem, i.e., to find out the starting shape of the metal which for a given time history of the externally applied boundary forces/velocities will transform the metal to a desired final shape. This problem is often refered to as the 'Preform Design Problem'. Mathematically, it can be stated as: 'Find $(\mathcal{D}(t_0), \Gamma(t_0))$ which on the application of $\mathbf{F}^*(\Gamma_1(t), t)$ and $\mathbf{V}^*(\Gamma_2(t), t)$ transforms to $(\mathcal{D}(t_f), \Gamma(t_f))$', where $\mathbf{F}^*(\Gamma_1(t), t)$ and $\mathbf{V}^*(\Gamma_2(t), t)$ are the given time histories of the externally applied boundary forces and velocities.

4 Mathematical Model of Metal Deformation

Let the stress and strain rate tensors for a point in the domain at any time t be respectively σ_{ij} and $\dot{\epsilon}_{ij}$. The velocity components of this point are v_i. These three quantities satisfy the following relationships: (i) equilibrium equations, (ii) stress/strain-rate equations, (iii) strain-rate velocity equations, and (iv) the equations derived for the material constitutive models ([7],[8]). Mathematically, these are:

$$\sigma_{ij,j} + f_i = 0, \quad i = 1, ..., 3 \tag{1}$$

$$s_{ij} = \frac{2}{3}\frac{\bar{\sigma}}{\dot{\bar{\epsilon}}}\dot{\epsilon}_{ij} \tag{2}$$

$$\dot{\epsilon}_{ij} = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) \tag{3}$$

where f_i are the components of the body forces, s_{ij} is the deviatoric stress, $\bar{\sigma}$ is the flow stress, and $\dot{\bar{\epsilon}}$ is the flow strain rate defined as

$$s_{ij} = \sigma_{ij} - \frac{1}{3}\sigma_{pp}\delta_{ij} \tag{4}$$

$$\bar{\sigma} = \sqrt{\frac{3}{2}S_{ij}S_{ij}} \tag{5}$$

$$\dot{\bar{\epsilon}} = \sqrt{\frac{2}{3}\dot{\epsilon}_{ij}\dot{\epsilon}_{ij}} \tag{6}$$

The above equations involve a total of nine unknowns: (i) the six independent components of $\sigma_{ij}(x_1, x_2, x_3)$ because of its symmetry, and (ii) three independent components of velocity $v_i(x_1, x_2, x_3)$. These nine variables over the domain $\mathcal{D}(t)$ are obtained by solving a boundary value problem with the boundary conditions prescribed over $\Gamma(t)$. As mentioned earlier, force boundary conditions are specified over $\Gamma_1(t)$ and velocity boundary conditions over $\Gamma_2(t)$. The material constitutive models are of the following general form:

$$\bar{\sigma} = f(\bar{\epsilon}, \dot{\bar{\epsilon}}, T) \tag{7}$$

where T is the temperature and $f(\ )$ is a nonlinear function of its arguments.

Some salient features of the metal deformation model are: (i) $\dot{\epsilon}_{kk} = 0$, i.e., the flow is incompressible, (ii) σ_{ij} can not be expressed as partial derivatives of velocity components, (iii) unlike elastic deformation processes, it is not possible to obtain 'Navier-like' flow equation for the metal deformation systems. All these features can be easily verified by simple manipulation of Eqs. (1)-(3). It turns out that these features arise due to the particular form of the governing stress, strain-rate relationship (2).

Once the boundary value problem is solved and the velocities are determined over the domain $\mathcal{D}(t)$, the boundary $\Gamma(t + \Delta t)$ is found by solving an initial-value problem in time by well known integration schemes.

4.1 Comparison with Elastic Systems

The metal deformation models described in the last paragraph can be contrasted with elastic deformation models to point out the similarities and differences. In *elastic deformation*, the variables describing the

chacteristics of a point are the stress σ_{ij}, the strain ϵ_{ij}, and displacement u_i. Eqs. (1)- (3) for elastic systems are:

$$\sigma_{ij,j} + f_i = 0 \tag{8}$$

$$\sigma_{ij} = \lambda \epsilon_{kk} \delta_{ij} + 2G\epsilon_{ij} \tag{9}$$

$$\epsilon_{ij} = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}) \tag{10}$$

where λ and G are constants for the elastic material. On substituting, (10) in (9) and then later substituting the result in (8), the resulting Navier's equation is:

$$u_{i,jj} + \frac{1}{1-2\nu} u_{k,ki} = -\frac{f_i}{G} \tag{11}$$

The solution of $u_i(x_1, x_2, x_3)$ over the domain $\mathcal{D}(t)$ is obtained by solving the boundary value problem with force/velocity boundary conditions specified over $\Gamma(t)$.

If one attempted to write a Navier-like flow equation for the metal deformation systems, one can show that the resulting equation will involve not only the partial derivatives of the three flow velocities v_i but also another variable $\sigma_h = \frac{1}{3}\sigma_{pp}$, which is commonly refered to as the hydrostatic stress. In the boundary value solution of this Navier-like flow equation for metal deformation problems, boundary conditions on σ_h must be used to find the solution of the variables inside the domain.

5 Boundary Value Problem: FEM vs. BEM

A single forward simulation of a metal flow problem over the time t_0 to t_f can be broken down into n discrete time instances t_0, $t_0 + \Delta t$, $t_0 + 2\,\Delta\, t$,..., $t_0 + (n-1)\,\Delta\, t = t_f$. At every time instance, e.g., $t_0 + (i-1)\,\Delta\, t$, the boundary value problem must first be solved to determine the velocity variables over the domain $\mathcal{D}\{t_0 + (i-1)\,\Delta\, t\}$, followed by solution of an initial value problem to update the boundary $\Gamma\{t_0 + i\,\Delta\, t\}$. As a result of this discretization over time, a series of n boundary value problems followed by intial value problems must be solved during a single forward simulation run of the metal deformation process.

The boundary value problem can be solved by a number of different methods using weighted residual schemes. The 'Finite Element Method (FEM)' is one of the more popular methods to solve this problem. In the finite element method, the entire domain is discretized into nodes and the variables of interest are computed at each one of the nodes. The force/velocity boundary conditions are imposed on the nodes that fall on the boundary of the domain. Due to discretization of the entire domain, the solution becomes highly computation intensive. As a result, the solution takes a large execution time.

Another method based on weighted residual scheme, though slightly less popular, is the Boundary Element Method (BEM). In this method, the weighting solution is taken to be of a special form so that it satisfies the governing equations exactly within the domain. As a result, discretizations are necessary primarily at the boundary. In a typical run of the boundary element method, the variables are solved only at the boundary. Once the boundary solutions are obtained, if desired, the variables within the domain could also be computed.

With this relatively brief discussion of the FEM and BEM methods, it is evident that if one only requires solution of the variables on the boundary, then BEM is a more computationally efficient tool compared to FEM. This conclusion is based on the observation that the number of variables required in BEM is usually of an order of magnitude lower than the number of variables required in FEM. Since the preform design problem, as stated in Section 3, requires study of the evolution of the domain boundary $\Gamma(t)$ during t_0 and t_f, it seems more natural to use BEM instead of FEM for the solution of the boundary value problem. Since a single forward simulation of a metal deformation process requires n solutions of the boundary value problem, the computational benefit of BEM over FEM could be tremendous in just a single simulation run.

6 Preform Design Problem: A Solution Approach

As described in Section 3, the preform design problem is to find the initial shape of the metal, $\Gamma(t_0)$, that will transform the metal to the desired final shape, $\Gamma(t_f)$, for a given time history of force and velocity boundary conditions. In order to pose this problem as an optimization problem, one possible approach is to parametrize the initial and final shapes. The optimization procedure, then, finds the parametric description of the initial shape which after simulation over t_0 and t_f results in a final shape parametrically close to the desired description ([1], [6]). In principle, one can describe the initial and final shapes by the same class of shape functions. However, starting shapes are often limited to generalized cylinders with circular or rectangular cross sections. Hence, from a practical point of view, one can describe the initial shape by restricting to circular or rectangular cylinders. For example, if circular cylinders are selected as starting shapes, they can be described by only two parameters such as height h and radius r. A similar method can be adopted for rectangular cylinders. In general, one can describe allowable starting shapes by a set of m parameters $(a_1, a_2, ..., a_m)$. The intermediate shapes and the final shape, in general, can be quite arbitrary. Hence, these are represented as a linear sum of n basis functions $f_1(x_1, x_2), f_2(x_1, x_2), ..., f_n(x_1, x_2)$:

$$x_3^k(t) = \sum_{i=1}^{n} b_i^k(t) f_i(x_1, x_2) \tag{12}$$

where $b_i^k(t)$ are the coefficients at a time t of the kth simulation run, $t_0 < t \leq t_f$. With this description of input and output shapes, a simulation run $\mathcal{S}^k$ can be looked upon as a mapping between the shape parameters $(a_1^k, a_2^k ..., a_m^k)$ and $(b_1^k(t_f), b_2^k(t_f) ..., b_n^k(t_f))$.

$$(a_1^k, a_2^k, ..., a_m^k) \xrightarrow{\mathcal{S}^k} (b_1^k(t_f), b_2^k(t_f) ..., b_n^k(t_f)) \tag{13}$$

As quite expected, when starting out from an arbitrary set of $(a_1, a_2, ..., a_m)$, there is a very remote possibility that $(b_1(t_f), b_2(t_f), ..., b_n(t_f)) = (b_1^*(t_f), b_2^*(t_f), ..., b_n^*(t_f))$, where * denotes the desired values. As a result, the starting parameters must be altered so that $(b_1(t_f), b_2(t_f), ..., b_n(t_f))$ becomes closer to the desired value. This alteration of the starting parameters at a step k can be done by computing the local Jacobian matrix between the input parameters and the the output parameters defined as:

$$J^k = \begin{bmatrix} \frac{\partial b_1(t_f)}{\partial a_1} & \frac{\partial b_1(t_f)}{\partial a_2} & \cdots & \frac{\partial b_1(t_f)}{\partial a_m} \\ \frac{\partial b_2(t_f)}{\partial a_1} & \frac{\partial b_2(t_f)}{\partial a_2} & \cdots & \frac{\partial b_2(t_f)}{\partial a_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial b_n(t_f)}{\partial a_1} & \frac{\partial b_n(t_f)}{\partial a_2} & \cdots & \frac{\partial b_n(t_f)}{\partial a_m} \end{bmatrix}_k \tag{14}$$

where J_k is the $(n \times m)$ Jacobian matrix computed at the kth step. Once the Jacobian matrix is computed, a perturbation in the $(a_1^k, ..., a_m^k)$ can be easily found using the properties of this matrix that results in a change of the parameters $(b_1(t_f), b_2(t_f), ..., b_n(t_f))$ closest to $(b_1^*(t_f), b_2^*(t_f), ..., b_n^*(t_f))$. The metric for describing 'closeness' can be taken as the two norm, $|\ |_2$.

It must be noted that each computation of the Jacobian matrix requires $m + 1$ forward simulation runs with a_1^k,..., a_m^k perturbed, one at a time. A flowchart of the optimization algorithm is shown in Figure 2.

7 Boundary Element Method

Over the last two decades, the boundary element method has been used for solving a large number of problems in elasticity, electromagnetics, and fluids ([5], [3], [2], [4]). In recent years, this method is being applied to plasticity and metal deformation problems. One of the tricks [9] that has been applied in the analysis of plasticity problems is to write its flow equations similar to the Navier elastic flow equation. This

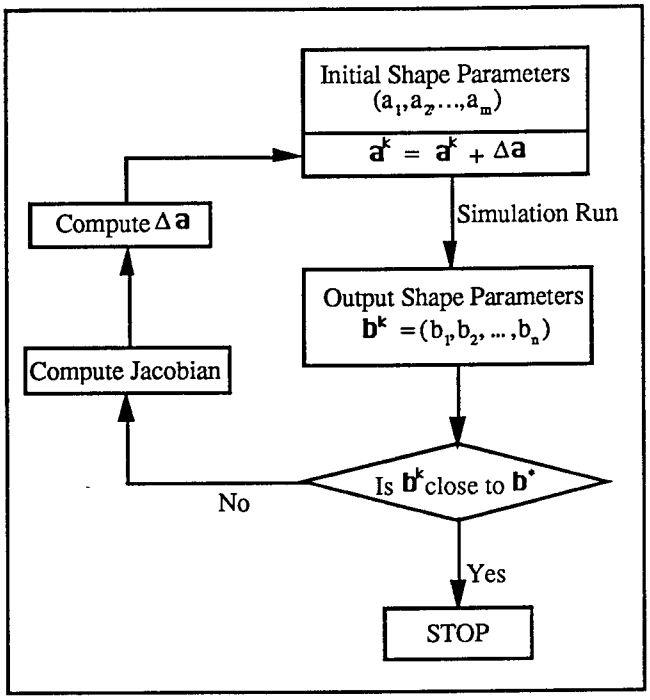


Figure 2: A flowchart of the gradient-based algorithm for solution of the preform design problem.

is achieved by writing the elastic strain as a difference between the total strain and the plastic strain:

$$\dot{\epsilon}^e_{ij} = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) - \dot{\epsilon}^p_{ij} \tag{15}$$

The constitutive models for the plastic and elastic parts of the strain rate are then taken from Section 4 and Section 4.1 respectively. These are:

$$\dot{\epsilon}^p_{ij} = \frac{3}{2}\frac{\dot{\bar{\epsilon}}}{\bar{\sigma}} s_{ij} \tag{16}$$

$$\sigma_{ij} = \lambda \epsilon^e_{kk}\delta_{ij} + 2G\epsilon^e_{ij} \tag{17}$$

The expression for $\dot{\epsilon}^p_{ij}$ satisfies the incompressibility constraint, $\dot{\epsilon}^p_{kk} = 0$. On substituting the above expressions in the equilibrium Eq. (1), the flow equation can be written as:

$$v_{i,jj} + \frac{1}{1-2\nu} v_{k,ki} = -\frac{\dot{f}_i}{G} + 2\dot{\epsilon}^p_{ij,j} \tag{18}$$

which is similar to the Navier flow equation (11) except for an additional term $\dot{\epsilon}^p_{ij,j}$.

Once the plastic flow equation is written in this elastic-like form, the details of the boundary element method are the same as those for elasticity problems. In the absence of the domain forces f_i, the general solution of the velocity at any point P on the boundary can be written as:

$$c_{ij}(P)v_i(P) = \int_\Gamma [V_{ij}(P,Q)\dot{\tau}_i(Q) - T_{ij}(P,Q)v_i(Q)]ds_Q + \int_{\mathcal{D}} [2GU_{ij,k}(P,q)\dot{\epsilon}^p_{ik}(q)]d\mathcal{D} \tag{19}$$

where c_{ij} depends on the local geometry at P, Q is a second point on the boundary, and q is a point within the domain. The two-point function $V_{ij}(P,Q)$ is the velocity at Q in the i direction due to a unit point load

at P in the j direction. The function $T_{ij}(P,Q)$ has a similar physical meaning, but in terms of traction rates. These functions are computed from Kelvin's singular solution due to the point load in an infinite elastic solid

$$V_{ij} = \frac{1}{16\pi(1-\nu)Gr}\{(3-4\nu)\delta_{ij} + r_{,i}r_{,j}\}$$
$$T_{ij} = -\frac{1}{8\pi(1-\nu)r^2}[\{(1-2\nu)\delta_{ij} + 3r_{,i}r_{,j}\}\frac{\partial r}{\partial n} + (1-2\nu)(r_{,i}n_j - r_{,j}n_i)] \quad (20)$$

where $r(p,q)$ is the distance from a source point p to a field point q and n_i are the components of the unit outward normal to Γ at a point Q on it. The convention used here is that lowercase letters p and q denote points inside the domain $\mathcal{D}$ and the capital letters denote points on the boundary Γ. A comma denotes a derivative with respect to a field point, i.e.,

$$r_{,i} = \frac{\partial r}{\partial x_{oi}} = \frac{x_{oi} - x_i}{r} \quad (21)$$

where x and x_0 are the source and field points, respectively. The traction rate $\dot{\tau}_i$ at on the boundary Γ is given by

$$\dot{\tau}_i = \dot{\sigma}_{ij}n_j = G[(v_{i,j} + v_{j,i})n_j + \frac{2\nu}{1-2\nu}v_{k,k}n_i - 2\dot{\epsilon}^p_{ij}n_j] \quad (22)$$

One of the salient features that must be pointed out from this section is that if $\dot{\epsilon}^p_{ij} = 0$, Eq. (19) simplifies to an integral on the boundary. However, due to the presence of plastic strain in metal deformation processes, the boundary velocities are dependent not only on velocities of other points on the boundary but also on plastic strain of points within the domain. The current research on the use of boundary element method to metal deformation processes has shown that the solution of the metal deformation processes are reasonably accurate if a very coarse discretization is carried out within the domain. As a result, the number of nodes needed in the boundary element analysis can be substantially less than those needed in the finite element analysis for a reasonably accurate solution.

8 Conclusions and Future Work

In order to run a single iteration of the gradient based optimization procedure, the boundary value problem must be solved $n(m+1)$ times where n is the number of discretizations of the time duration t_0 and t_f and m is the number of parameters that describe the initial shape of the preform. If it takes k iterations for the optimizer to converge to the solution, the total number of boundary value problems that must be solved is $kn(m+1)$. Then, it is quite clear that for solving preform design problems, boundary element technique could be computationally several orders of magnitude faster than finite element technique in the loop. Morover, the variables inside the domain are of no particular interest for the preform design problem.

My recommendations based on this study are: (1) to develop boundary element analysis codes for planar and axisymmetric metal deformation problems based on the theory of Section 7, (2) to benchmark the results of boundary element and finite element methods, (3) to develop a gradient based optimizer for the preform design problem based on Section 6, (4) to use this optimizer as a benchmark to compare the results of other optimizers built on approximate models of the plastic deformation process such as just with volume conservation or with linearized constitutive models.

The author believes that this preform design module could be written for a PC 486 machine which are relatively cheap and affordable to most companies involved in metal forming work. The results of this analysis could be interfaced with graphics for visual display.

9 Acknowledgments

I sincerely thank Dr. James Malas and Dr. William G. Frazier of WL/MLIM and Dr. Anil Chaudhary of UES, Inc. for numerous discussions during the course of this work. The support of RDL through the AFOSR Summer Program is gratefully appreciated.

10 Bibliography

References

[1] Agrawal, S.K. and Fabien, B.C., *Optimization of Dynamic Systems*, book manuscript (under preparation), 1994.

[2] Bakr, A.A., *The Boundary Integral Equation Method in Axisymmetric Stress Analysis Problems*, Springer-Verlag Book Company, 1986.

[3] Banerjee, P.K. and Butterfied, R., *Boundary Element Methods in Engineering Science*, McGraw-Hill Book Company, 1981.

[4] Becker, A.A., *The Boundary Element Method for Engineers*, McGraw-Hill Book Company, 1992.

[5] Brebbia, C.A., *The Boundary Element Method in Engineering*, John Wiley and Sons, 1978.

[6] Haug, E.J., Choi, K.K., and Komkov, V., *Design Sensitivity Analysis of Structural Systems*, Academic Press Inc., 1986.

[7] Kobayashi, S., Oh, S., and Altan, T., *Metal Forming and Finite Element Method*, Oxford University Press, 1989.

[8] Mase, G.E., *Continuum Mechanics*, Schaum's Outline Series in Engineering, McGraw Hill Book Company, 1970.

[9] Mukherjee, S., *Boundary Element Methods in Creep and Fracture*, Applied Science Publishers, 1982.

CALCULATION OF HEATING AND TEMPERATURE DISTRIBUTIONS IN ELECTRICALLY EXCITED FOILS

Michael E. Baginski
Associate Professor
Department of Electrical Engineering
Auburn University

Auburn University
200 Broun Hall
Auburn University, Alabama 36849-5201

Final Report for:
Summer Faculty Research Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, D.C.
and
Auburn University

September 1994

CALCULATION OF HEATING AND TEMPERATURE DISTRIBUTIONS IN ELECTRICALLY EXCITED FOILS

Michael E. Baginski
Associate Professor
Department of Electrical Engineering
Auburn University

ABSTRACT

A finite element analysis of the transient thermal and electrical distributions in electrically exploded thin copper foils to the point of melt is presented. The research focuses on an analysis of a novel system that is currently under development for use in future experiments. All simulations are based on the intrinsic characteristics of copper and require only a two dimensional solution due to the planar nature of the foil's geometry. The simulated behavior shows trends observed in measurements of similar configurations. Specifically, the thermal enhancement observed at abrupt changes in the foils edge geometry.

CALCULATION OF HEATING AND TEMPERATURE DISTRIBUTIONS IN ELECTRICALLY EXCITED FOILS

Michael E. Baginski

I. INTRODUCTION

The principle focus of this report is to present a method for the calculation of temperature profiles in electrically thin foils to the point of melt. A major reason that this topic is of importance to the military is that thin foils are often used as detonation devices in explosives (i.e., slappers). After a literature survey was conducted, it was found that if the pre-burst temperature profile in electrically thin foils is known, the prediction of how the foil will expand during the explosion is highly predictable [1]. This is of extreme importance in the application of electrically exploded foils used for the purpose of device detonation.

Before discussing the code and methodology in depth, a brief history of the previous work in this area most related to the topic will be discussed.

All of the models proposed in the unclassified literature so far have been limited by the problems complexity. One of the more complete models is that of the Lawrence Livermore National Laboratory (LLNL, EBF1) in which the simulated fireset is cast in terms of simple RLC circuit and switch. This is used in conjunction with a finite difference technique to simulate the thermodynamics of the bridge foil to the point of burst. LLNL have also presented a model (FUSE) with a more simplified treatment of the circuit and bridge, with special emphasis on post-burst behavior.

Sandia National Laboratory (SNL) has developed a model (CAPRES) assuming a lumped resistance model of the entire bridge. They use an empirical formulation of the exploding bridge in place of basing their model on the governing physics. The rest of the circuit is described in a similar manner to that of LLNL in EBF1. SNL researchers (Kennedy, Stanton, McGlaun, Tucker) used the concept of specific action integrals in their formulation of the bridge resistance where A = cross-sectional area of the bridge, I is total current flowing in the bridge as time t, and g is the specific action integral defined to the time of burst as follows:

$$g = (1/A)\int I^2 dt \qquad (1)$$

The use of the universality theorem was then applied to the specific action to extrapolate to the bridge geometry of interest and desired current density. There is a large amount of information contained in this type of data acquisition, but the models developed from this data have serious limitations (the limits of applicability to many data bases are not well defined).

All three DOE labs have used hydrodynamic simulation codes to model the burst phenomenon in more detail. The major problem appears to be deriving equations based on first principles that described the material's transient behavior.

The research presented here will be based in part on some of the previous work with the addition of several constraining equations derived from the material parameters. Variables in the analysis are the geometric configuration of the slapper, specifically the slope of the connecting pad to the explosive member of the slapper, and the applied voltage. The thermal conductivity of the foil is included in the numerical model and derived from catalogued data. Copper will be used as the metal foil in a necked-down geometric configuration.

Although the research is targeted at understanding realizable device behavior, an effort is made to investigate any interesting behavior that is unforeseen. The research is also conducted with the assumption that the device modeled will be fabricated and, in the future, experimentally analyzed.

II. DIFFERENTIAL EQUATIONS

The differential equations that govern the behavior of the exploding foil system are formulated in terms of Maxwell's equations and the heat equation. Several first order assumptions are made. The copper foils being studied are considered to be approximately 1 micron in thickness and therefore it is assumed that the magnetic and electric flux can diffuse through the foil fast enough to track any changes in current so that skin effects can be ignored. Logan [1] has investigated exploding foils and found this assumption to be true for much larger systems, reinforcing the validity of the assumption. The governing thermodynamic equation is based on the assumption that heat loss in the area surrounding the slapper is negligible and therefore can be ignored as has been in previous studies [2]. This is reinforced by the fact that the results of previous modeling studies show good agreement with measured data using this assumption.

The differential equations are solved in two dimensions using the finite element method. They differ form work done by Logan et. al., in that they include the thermal heat flow in the copper foil and use measured values of the electrical conductivity as a function of temperature.

The temperature rise at a point on the foil is calculated from the equation

$$dT/dt = (\sigma_e\,(T)*\mathbf{E}^2)/(C_v(T)*\rho) + \nabla(K\bullet\nabla T) \qquad (2)$$

where T is temperature in degrees Kelvin, σ_e (T) is the electrical conductivity, ρ is the mass density, K is the thermal conductivity, **E** is the electric field and C_v(T) is the specific heat as a function of temperature.

The electrical behavior in the region is described by assuming the magnetic field can be neglected and therefore **E** is defined as $\mathbf{E} = -\nabla V$ where V is the voltage. By using the previous assumptions made by Logan et.al., (localized charge accumulation is set to 0) the electrical description of the slapper's behavior is given as

$$\nabla\bullet(\sigma_e\,(T)*\mathbf{E}) = 0 \qquad (3)$$

III. GEOMETRY OF THE REGION

The region selected for the investigation is referred to as a "BOW-TIE" configuration and shown in Fig. 1. Three principal reasons were involved in the selection of this foil layout as opposed to other possibilities geometries. Firstly, the geometry was sufficiently large that the device could be patterned and realized without difficulty. The second reason is that certain expected effects (e.g., edge effect current and thermal enhancement) would most likely be observed in the finite element modeling of this geometry, since no radius of curvature is assigned to the necked-down portion of the circuit. This geometry may overestimate the non-linear behavior of the system. However, it allows the researcher to glean an understanding of the likely trends that will appear when a device is fabricated and, in the future, monitor the experiment appropriately. The final reason for the use of the "BOW-TIE" geometry is that it is in the process of being fabricated and used in high explosive testing. This will allow the simulated behavior to be compared against measured data for possible additional model refinement.

IV. ELECTRICAL SOURCES

The electrical source that is used as the forcing function for the system is modeled after a typical fireset. A fireset circuit allows a low inductance path for the external electrical energy to be coupled to the exploding foil. After considering Richardson's report [2], a forced voltage described by $V(t) = 1000(1-\exp(-t/\tau))$ volts was selected as the electrical source for the circuit where $\tau =$ 100 nanoseconds. A voltage of $V(t) = 100(1-\exp(-t/\tau))$ was initially used to ensure the code's correct operation with negligible non-linear behavior observed. The empirical formulation of this voltage was based on cataloged data from several different sources cited by Richardson. This voltage is used to energize ARCS -1 (+V(t)) and ARCS -5 (-V(t)) in the simulations (Fig. 1). The remainder of the boundary ARCS allow no electrical or thermal flux to flow normal to their surfaces (i.e., they appear as perfect thermal and electrical insulators).

V. THE CONDUCTIVITIES

The electrical conductivity used in the simulation was derived from experimental data [3] that was obtained for static conditions. This approach of describing the electrical conductivity used in the modeling differs from much previous research that investigated exploding foils [1]. In many reports cited by Richardson, the electrical conductivity used for modeling purposes was obtained from an exploding foil experiment. Data used to approximate constitutive parameters acquired by this method may lead to a model that simulates the correct behavior, but, only by coincidence [2].

The electrical conductivity used in this model is shown in Fig. 2 . It was obtained by using a third order polynomial fit of 10 data points and given as

$$\sigma_e(T) = 8.7589*10^7 - 2.0711*10^5*T + 23576*T^2 - 9.9964E\text{-}2*T^3 \quad (4)$$

where T is the temperature in degrees Kelvin, and $\sigma_e(T)$ is the electrical conductivity in $(\text{ohm-meters})^{-1}$.

The thermal conductivity K was derived in a similar manner with one important exception. During the early stages of the model's development, it was observed that allowing the conductivity to assume its maximum value (value at room temperature) had no effect on the solutions characteristics for the time frames of interest. It was

therefore set to the largest value in the simulations discussed.

There was one major reason for not removing the thermal conductivity entirely from the analysis. The inclusion of the thermal conductivity provided a small amount of damping on the numerical solutions, making the code slightly more efficient.

VI. FINITE ELEMENT CODE

The finite element code used is the standard six-node triangle with first order elements (the element degree can vary from 1 to 4), with one edge curved when adjacent to a curved boundary, according to the isoparametric method.

In the problem, the algebraic equations are solved by Newton's method. The linear system which must be solved to do a Newton iteration is solved by Gaussian elimination. The Reverse Cuthill-McKee algorithm and a special bandwidth reduction algorithm are used to number the nodes and give this system a banded structure. In some cases simulated, symmetry is also taken advantage of in the elimination process. If the matrix is too large to keep in core, the frontal method is used to efficiently organize its storage out of core. In the virtual memory environment, the in-core option operates efficiently, with a minimum of page faults. the frequency of updating of the Jacobian matrix is determined adaptively.

The research relies on several subroutines that allow an initial triangulations to be input with a minimum number of triangles to define the region, and allows the user the ability to specify where the largest number of triangles is to be located in the final triangulation. This would be in a location where the solution is most likely to experience the greatest change. Optimal convergence is is possible if the final triangle density function is specified according to the criteria given by Sewell [4].

Each time a triangle is divided, it is divided by a line from the midpoint of its longest side to the opposite vertex. If this side is not on the boundary, the triangle which shares that side must also be divided to avoid nonconforming elements and discontinuous basis functions. The initial triangulation is shown in Fig. 1 and the final graded triangulations in Fig. 3 (2000 triangles). A time step of 0.1 nanoseconds was used and the duration of the simulations allowed to progress until the copper foil's melting point was observed.

A complete discussion of the entire code is beyond the scope of this research and may be found in references [4]. The finite element model has the additional benefit of allowing for any scaling to be introduced without major code revisions.

VII. SIMULATIONS

Before describing the simulations, consider again the phenomenon of interest: an electrically thin foil (slapper) is energized in order to cause a selected portion of the surface to explode. Because of the complexity of the system, the results of the modeling should be comparable to future experimental work so, if required, a model refinement could take place.

An additional trait in the device behavior observed in previous research is that in order to achieve optimal device operation, the exploding region of the foil should be uniformly heated to the point of melt. Therefore, the model constructed must meet several criteria:

1) If not obviously constrained, the model will focus on device behavior that is likely to be measured in future experiments.

2) As alluded to earlier, the simulations should clearly identify the device behavior that is unexpected.

The simulations will be presented in sections that show the temporal progression of the exploding member of the foil heating, the associated current density, and the normalized power absorption of the device.

Due to the obvious limits on the amount of information that can be presented, representative scheme will be shown that best depicts the device's behavior. Since the feature size of importance in many of the figures is small, it was necessary to allow full size figures to be shown to best demonstrate the more important trends. The graphical output will be extracted by interpolation from a 50x50 evenly spaced grid.

The first set of data (Fig. 4) illustrates how the material's non-linear properties effect the contours of equal potential (if this was a linear set of partial differential equations no change would occur in the contour's shape but only in magnitude). Fig. 5 shows the thermal heating that takes place with the most notable feature being significant temperature increases at the corners. The current density plots shown in Fig. 6 also indicate this same behavior. The plot of normalized power versus time (Fig.7) indicates that the joule heating taking place is in no way proportional to the voltage applied.

VIII. DISCUSSION OF RESULTS

There are a number of interesting phenomena occurring in the simulations that imply that a type of optimization could be made in the exploding foil's geometry and that suggest further research in this area is necessary. Probably the two most obvious traits in the foil's simulated transient behavior are the thermal heating near a corner and the decrease in the late time power absorption for the selected electrical source.

Richardson [2] was the first to note that the necked down portion of the foil has a significant effect on the joule heating due predominantly to the radius of curvature (the smaller the radius of curvature the more dominant the field fringing at the corner becomes) . Since no radius of curvature was included in the modeling, we may assume that this geometry overestimates the non-linear behavior of the system. This, however, will be addressed in future studies.

REFERENCES

[1] J. D. Logan, R. S. Lee, R. C. Weingart, and K. S. Yee, "Calculation of heating and burst phenomena in electrically exploded foils," Journal of *Applied Physics, Vol. 48,* No. 2, February 1977.

[2] D. D. Richardson, Technical Report on Theoretical modelling of Slapper Detonators, AFATL-TR-86-63.

[3] Y. S. Touloukian, Thermophysical Properties of High Temperature Solid Materials, Macmillian, New York, 1960, Vol. 1.

[4] G. Sewell, Analysis of a Finite Element Method, Springer-Vergal, New York 1985.

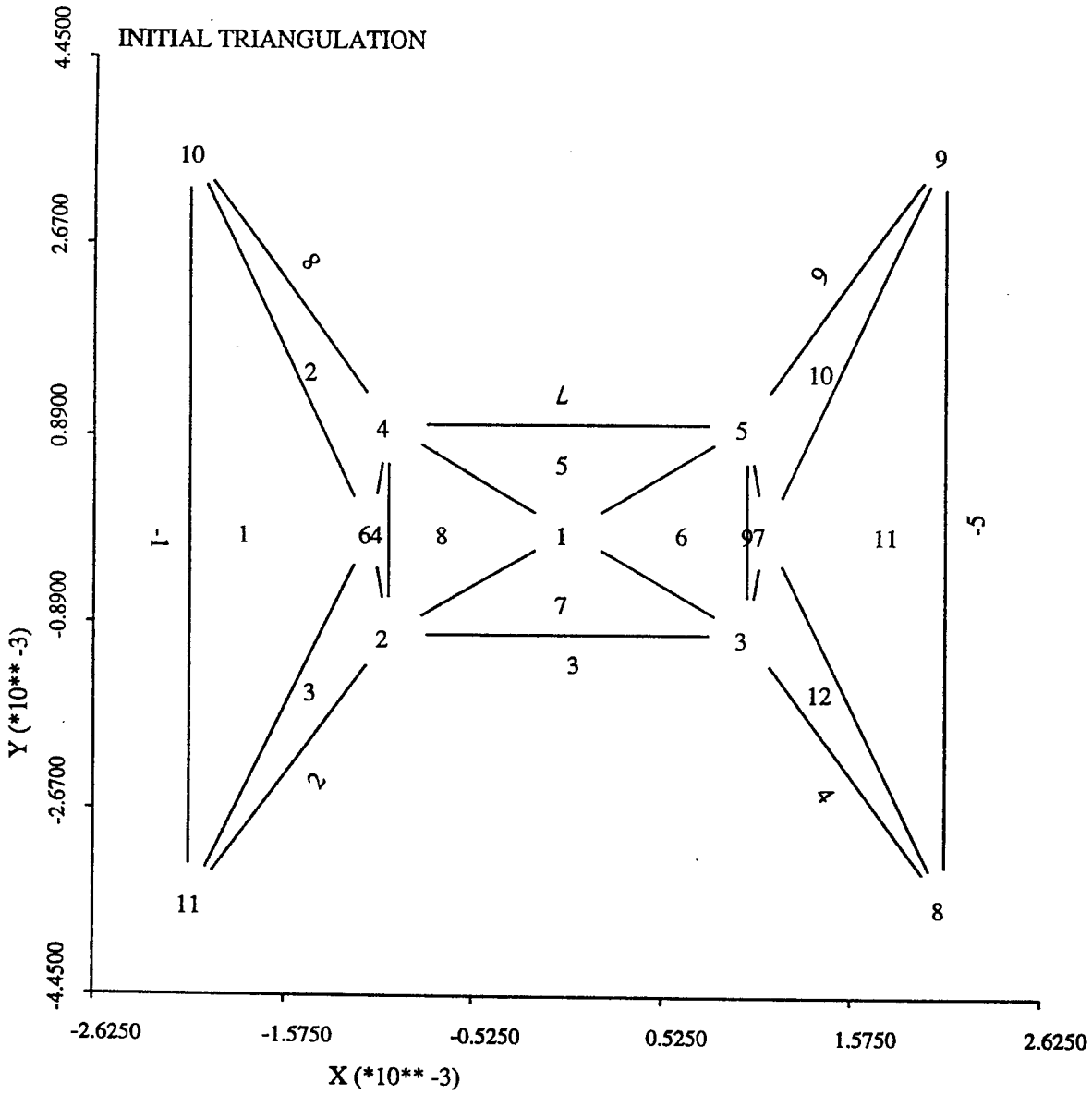


FIGURE 1. INITIAL TRIANGULATION

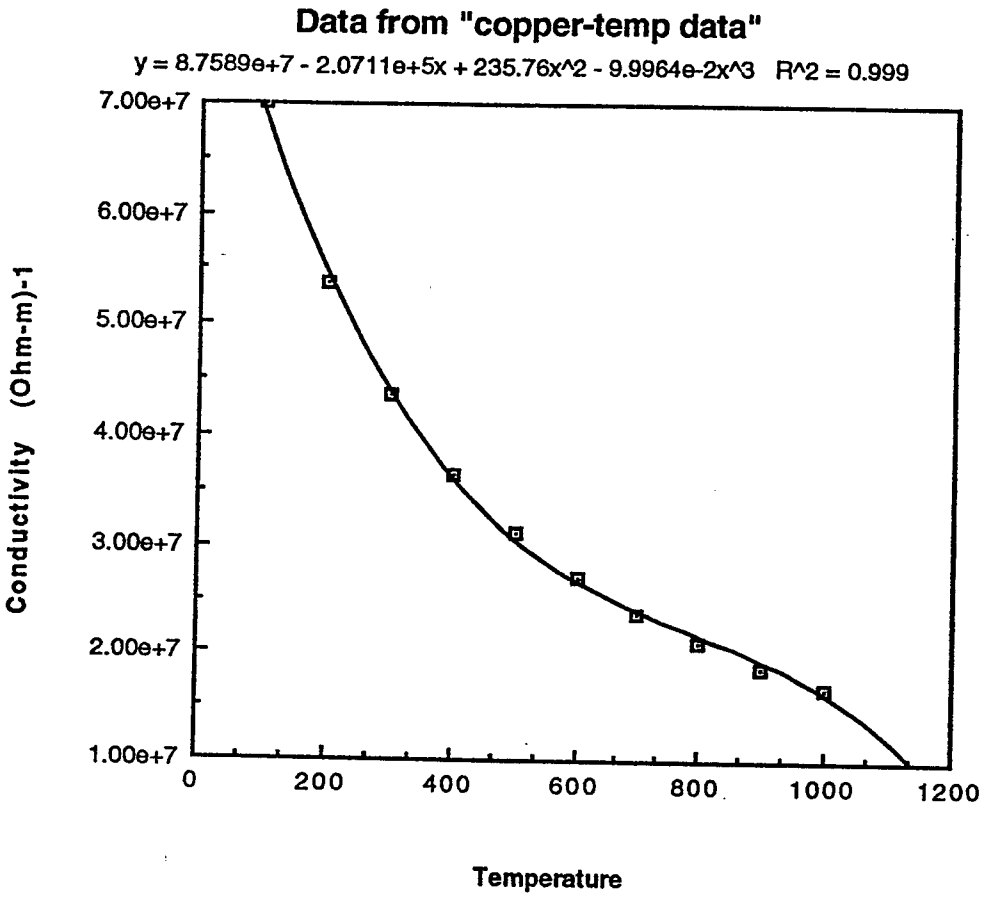


FIGURE 2. ELECTRICAL CONDUCTIVITY

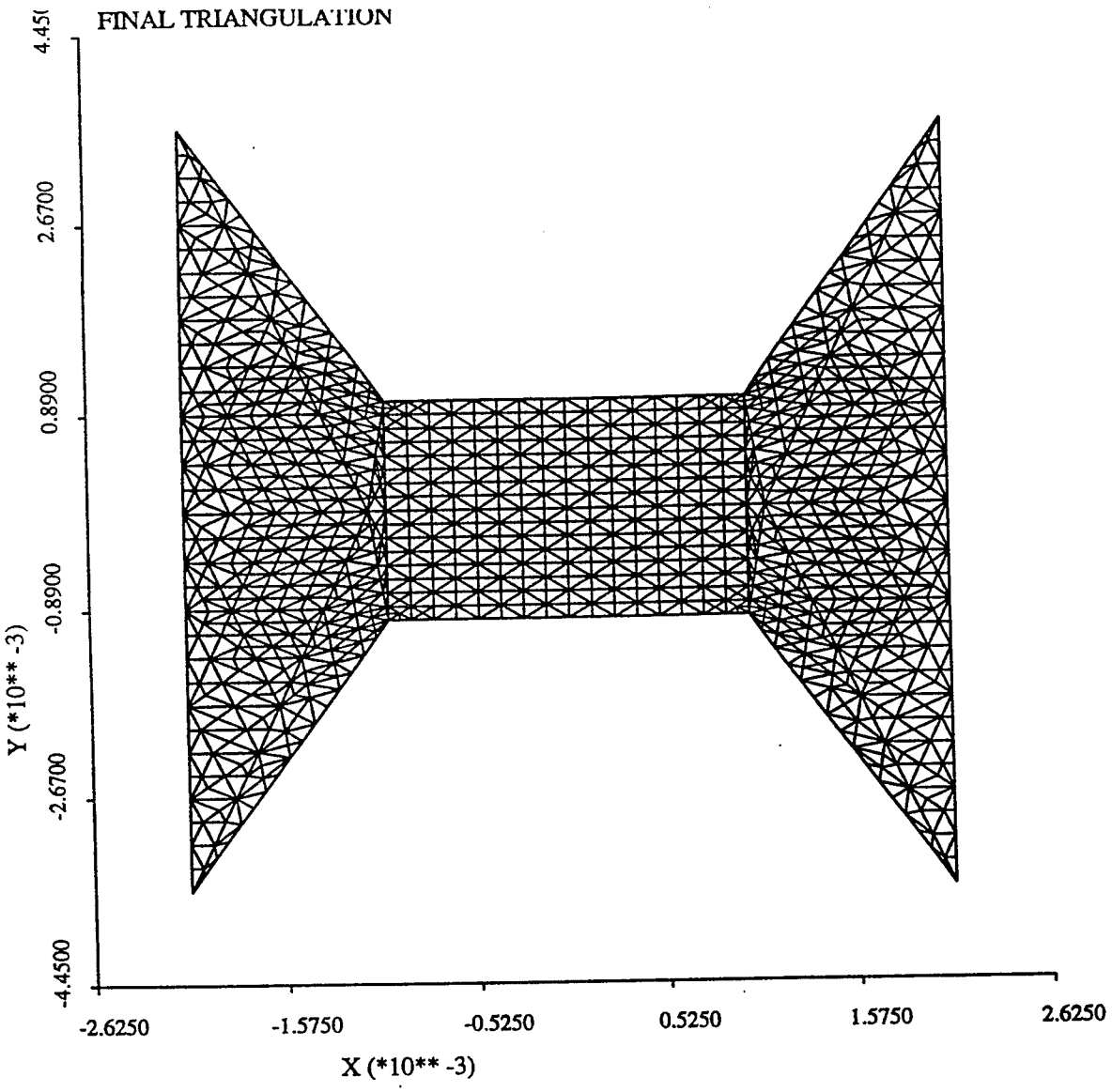


FIGURE 3. FINAL TRIANGULATION

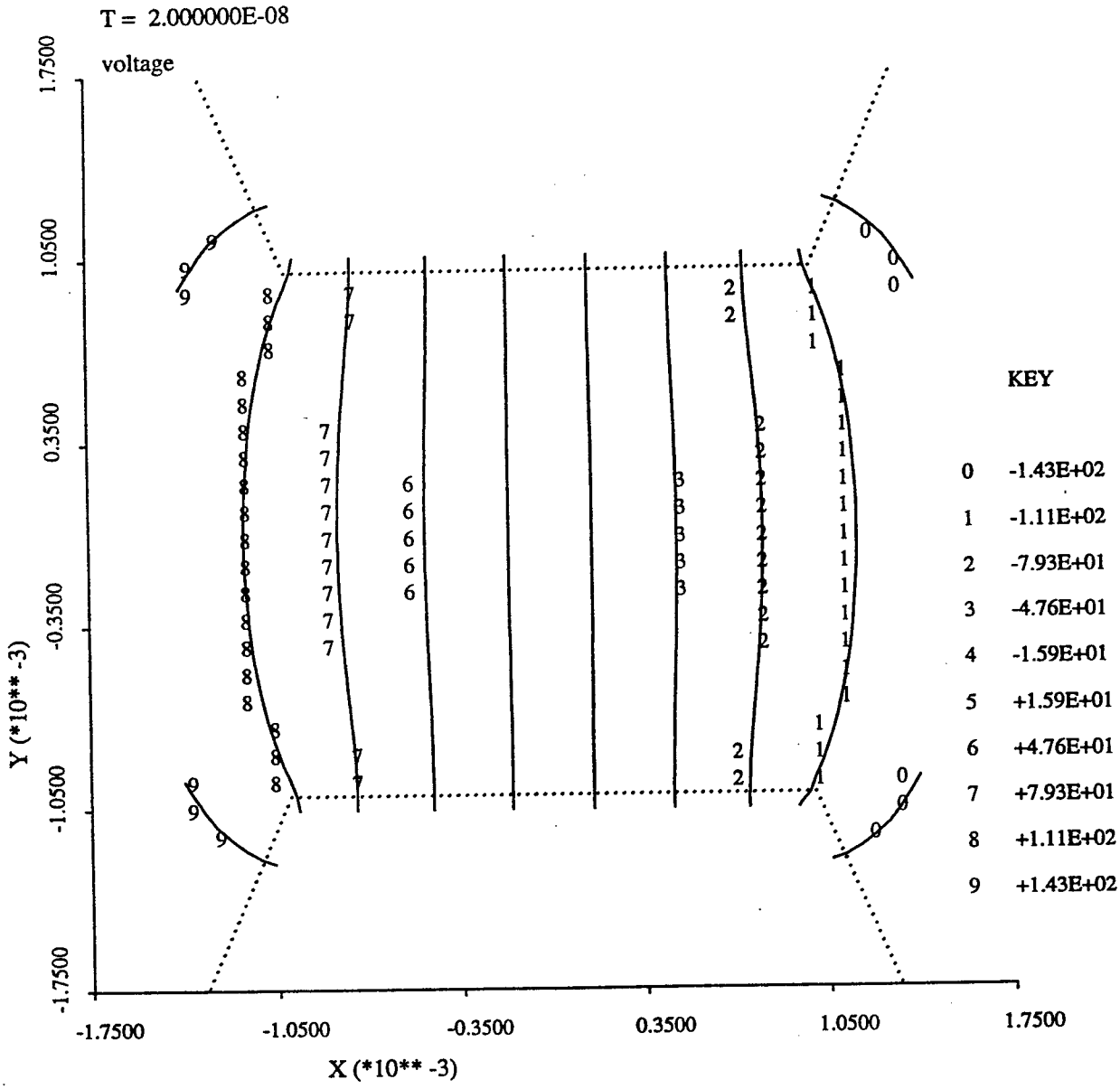


FIGURE 4-A. CONTOURS OF EQUAL POTENTIAL

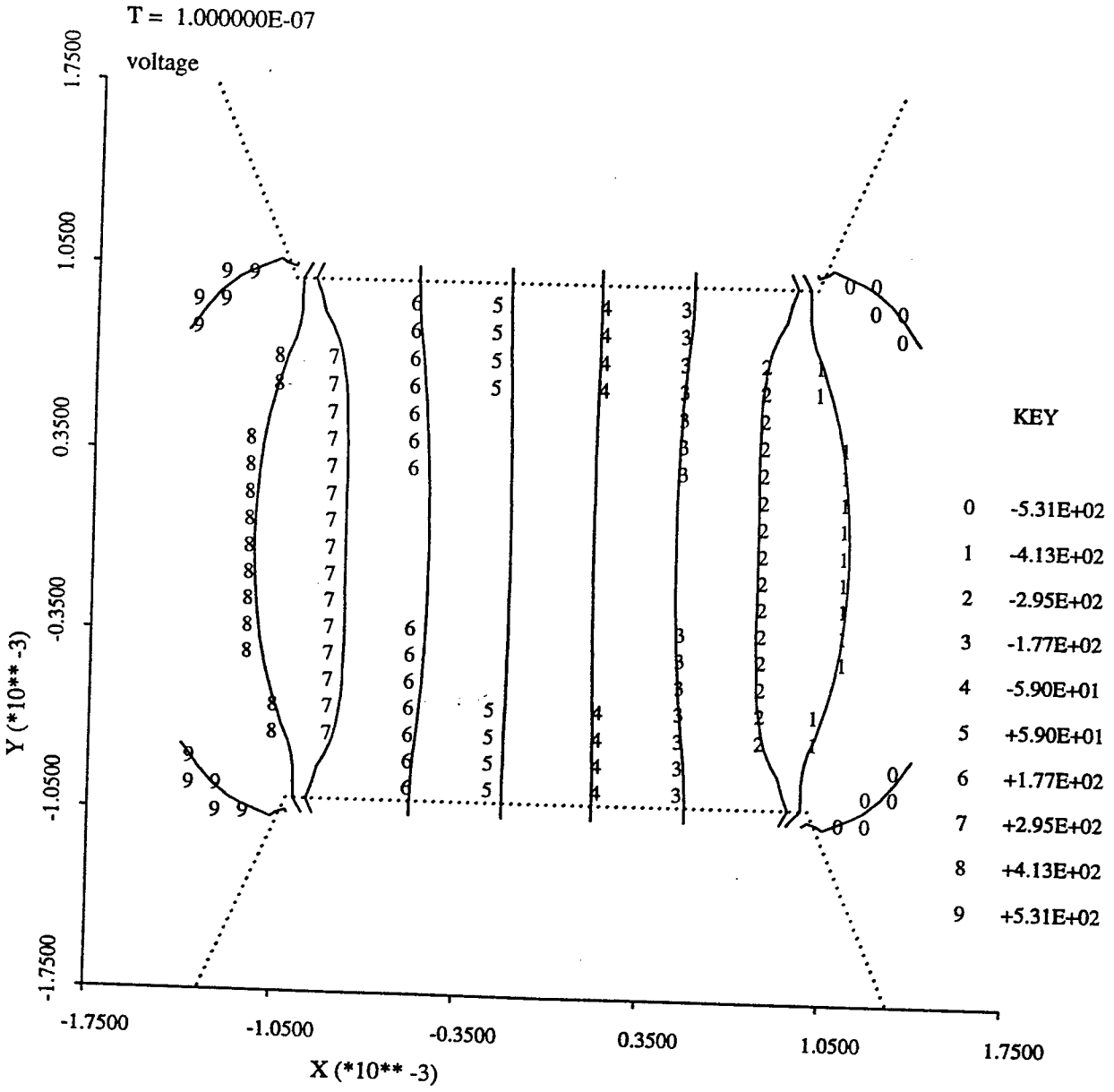


FIGURE 4-B. CONTOURS OF EQUAL POTENTIAL

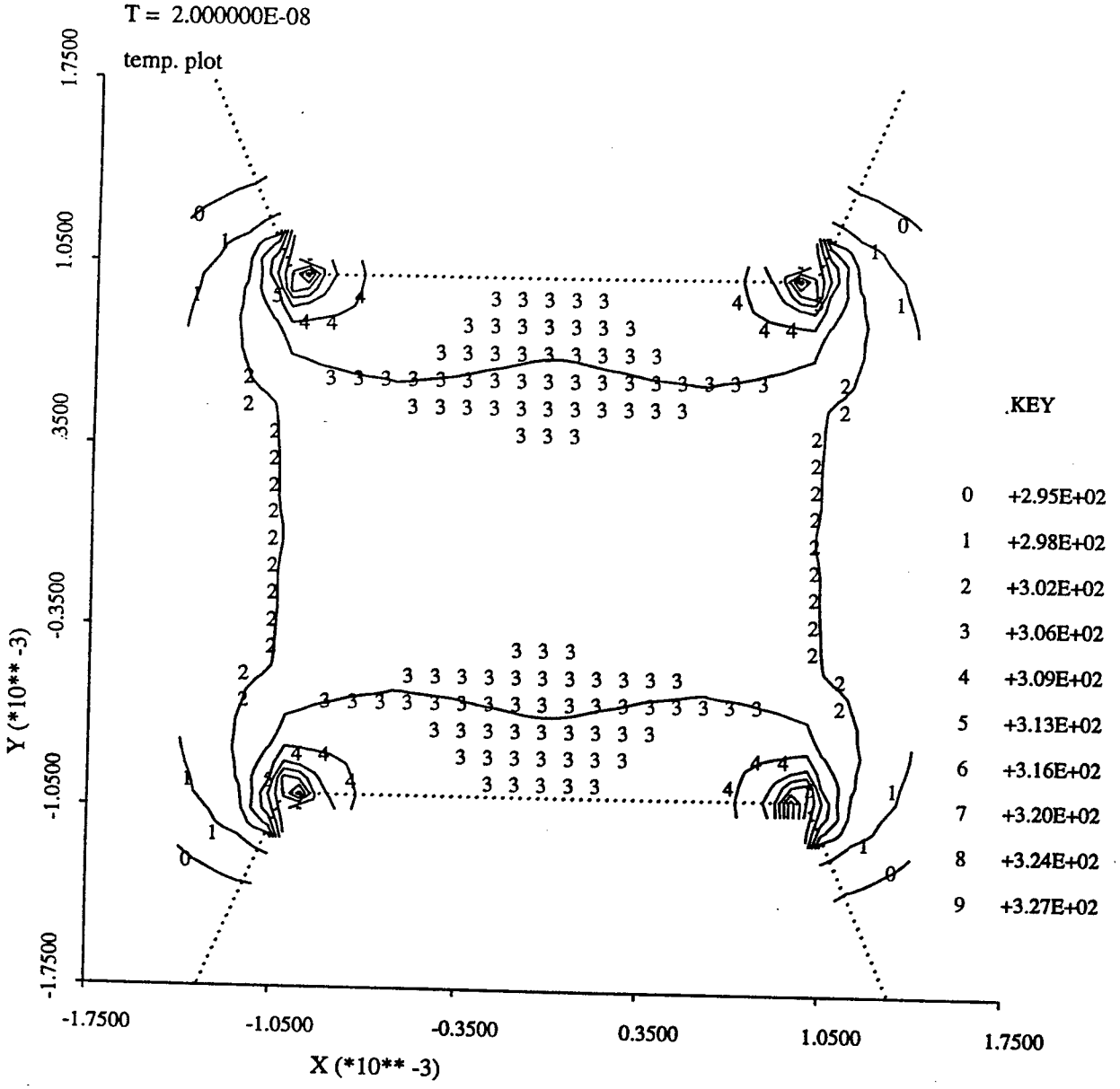


FIGURE 5-A. TEMPERATURE PLOT

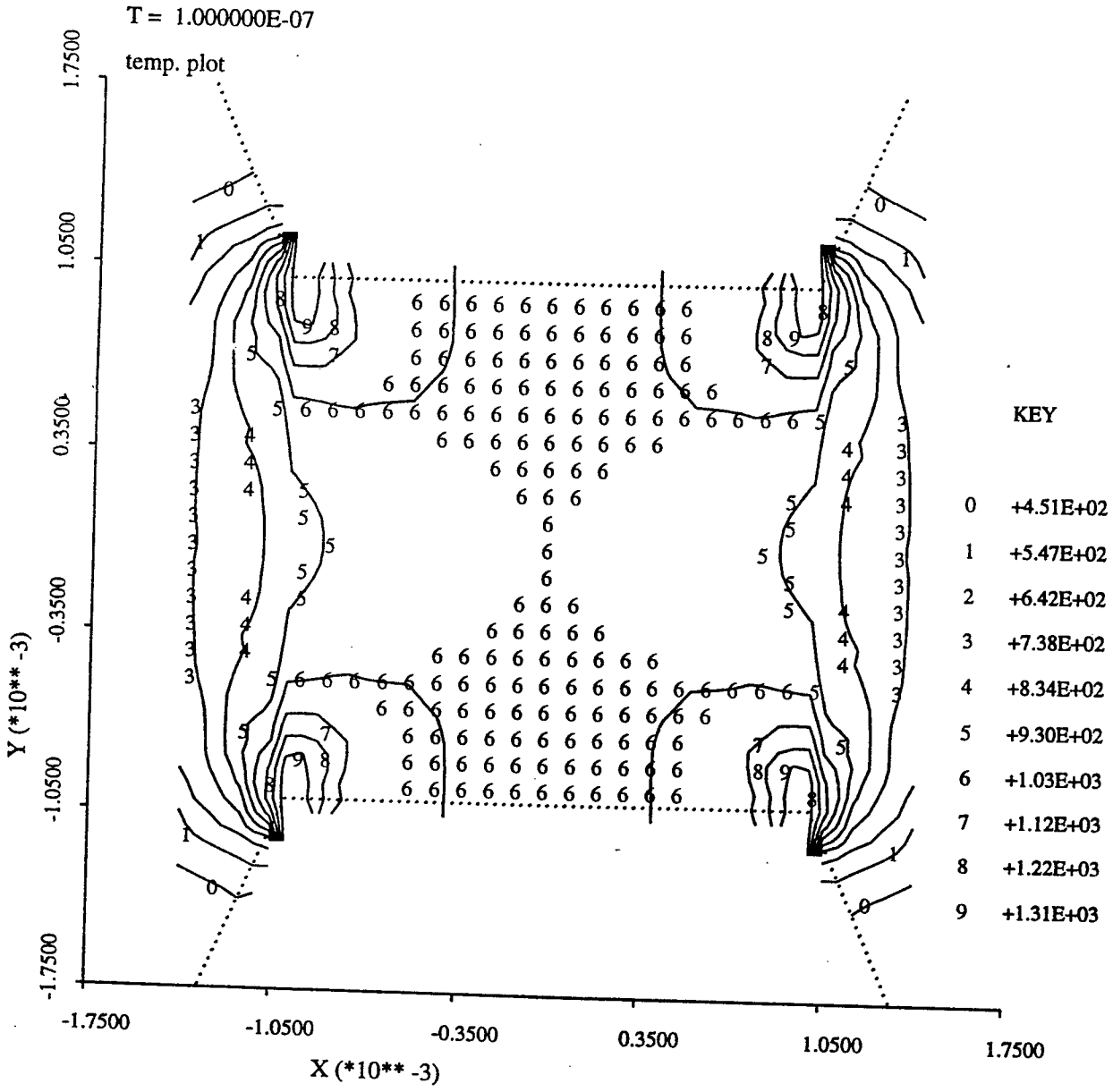


FIGURE 5-B. TEMPERATURE PLOT

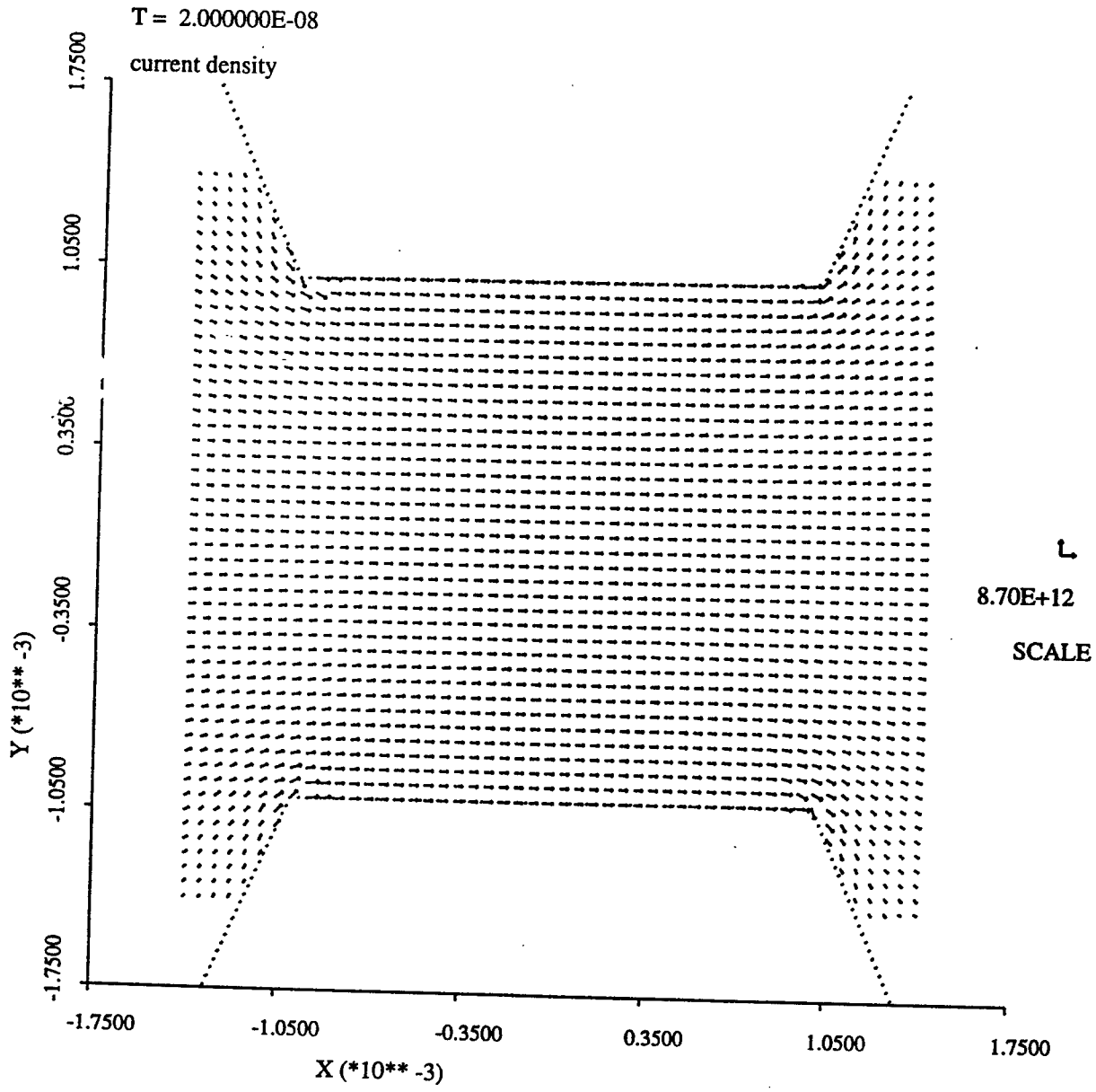


FIGURE 6-A. CURRENT DENSITY (microamperes/m^2)

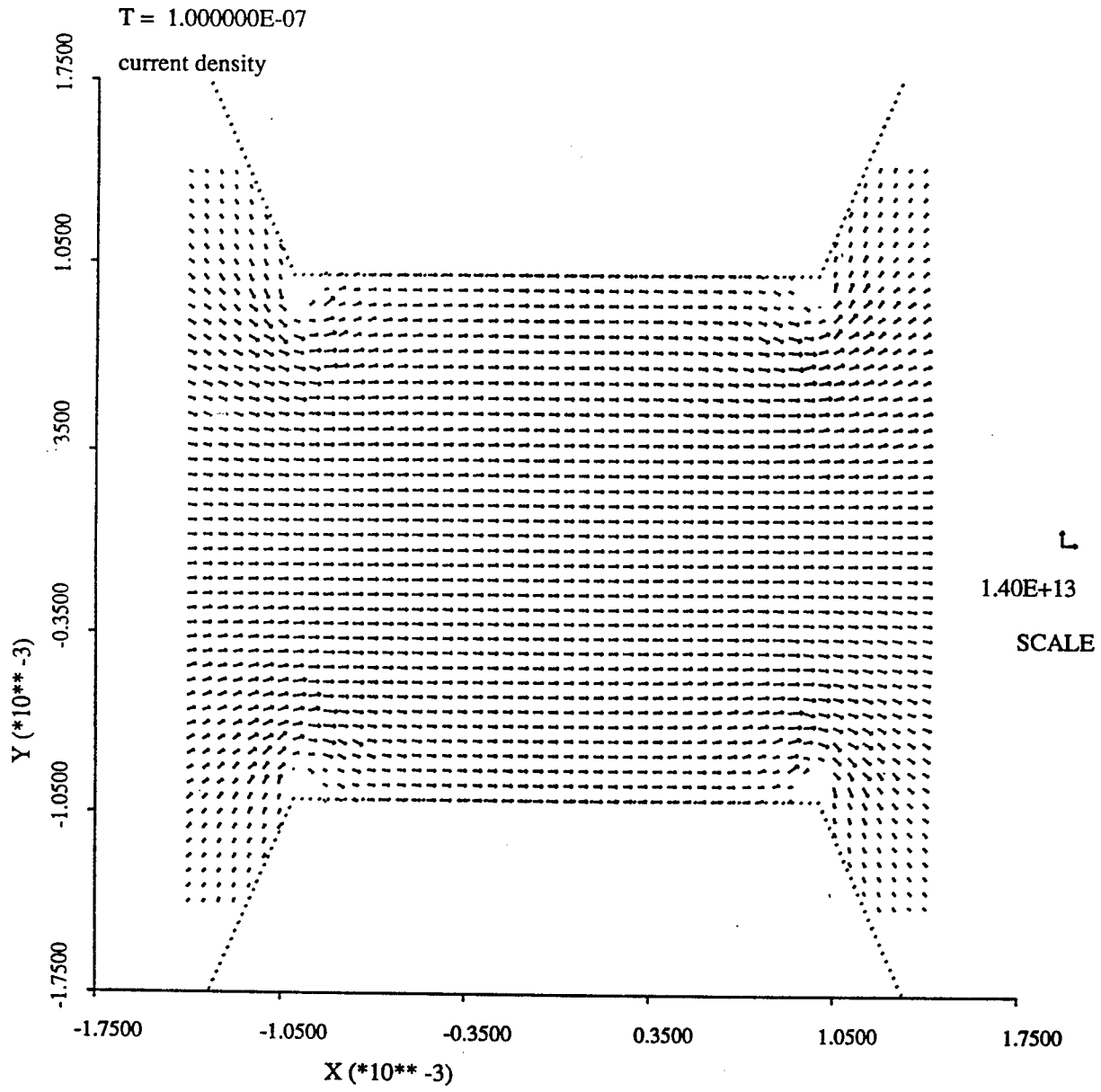
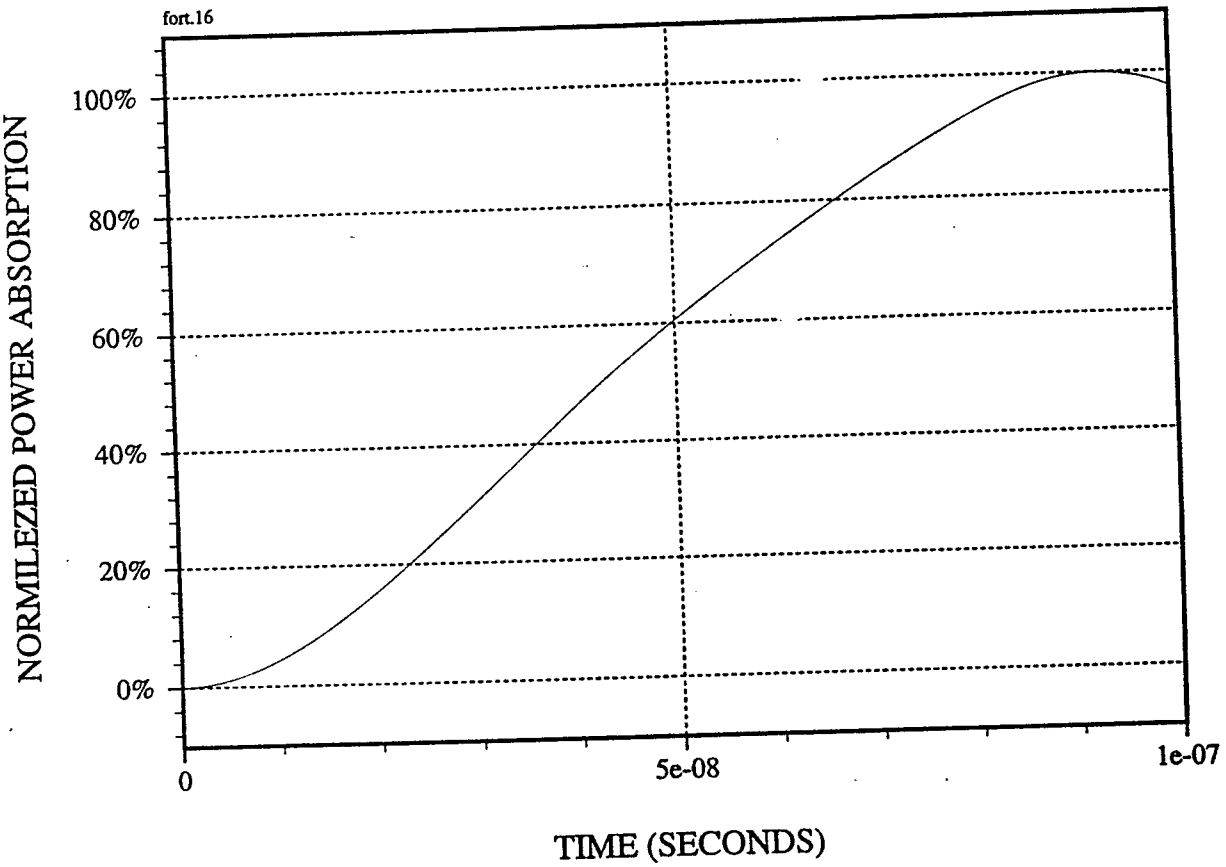


FIGURE 6-B. CURRENT DENSITY (microamperes/m^2)

NORMILIZED POWER ABSORPTION



NORMALIZED POWER ABSORPTION

ANOMALOUS EFFECTS OF WATER IN FIRE FIGHTING: INTENSIFICATION OF HYDROCARBON FIRES BY AZEOTROPIC DISTILLATION AND FREE RADICAL EFFECTS

William W. Bannister
Professor, Department of Chemistry
University of Massachusetts at Lowell
Lowell, Massachusetts 01854

Final Report for:
Summer Faculty Research Program
Wright Laboratories

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

and

Wright Laboratories

September 1994

ANOMALOUS EFFECTS OF WATER IN FIRE FIGHTING: INTENSIFICATION OF HYDROCARBON FIRES BY AZEOTROPIC DISTILLATION AND FREE RADICAL EFFECTS

William W. Bannister, Professor, Department of Chemistry
Unioversity of Massachusetts/Lowell, Lowell, MA 01854

ABSTRACT

We have shown that water, when applied to burning fuels, substantially increases fuel vaporization rates as a result of azeotropic "steam distillation" effects. Water-induced hot fuel volatility effects are particularly enormous for low volatility (high boiling point) fuels such as JP-8, JP-5 and Jet A-1. Correspondingly severe problems in fire fighting efforts could thus result for fully developed fires involving JP-8 type fuels being extinguished by water fog, AFFF, or other water based extinguishing agents or systems. The effect is not significant for fires involving fuel floating on significant volumes of water. Since almost all large scale training and research fires are conducted in fire pit facilities using tanks of water on which the fuel is floated, this effect has not heretofore been observed in such exercises. Evidence has been found, however, in at least one large serious fire, for very pronounced increases in fire intensities which rationally could have been ascribed to azeotroping effects.

Other, chemical, effects may possibly be operational in these instances. Since the only likely chemical candidates would involve free radical intermediacies, spectroscopic experiments were performed to examine possible free radical pathways.

With increasing emphasis on use of low-volatility JP-8, an importance exists for assessing magnititude of the effect for large scale real-life fires, and for developing countermeasures for obviating this effect; and for developing realistic training exercises to demonstrate the effect and appropriate countermeasures.

ANOMALOUS EFFECTS OF WATER IN FIRE FIGHTING: INTENSIFICATION OF HYDROCARBON FIRES BY AZEOTROPIC DISTILLATION AND FREE RADICAL EFFECTS

William W. Bannister

I. INTRODUCTION

Water has always been used as a fire extinguishing agent, either alone or as the main component of agent compositions such as AFFF or other water based compositions. The greatest single effect of the water in such applications is its great cooling capacity, lowering the temperature of a burning fuel below its flash point, thus removing one of the four essential conditions for fire maintenance. (In addition to heat, the other essential bases of the so-called fire tetrahedron are oxygen, the fuel itself, and existance of propagating free radical pathways in the flame system.)[1]

There are several well-known situations in which application of water actually serves to intensify a fire:

1. Water applied to hot grease fires can flash into steam, causing spattering which can greatly intensify the fire.
2. Water reacts violently with active metals such as sodium.
3. Direction of a vigorous jet of water from a fire hose into burning liquid fuel can result in mechanical "digging", scattering the burning liquid over a wider area and increasing the size of the blaze.
4. Air entrained in water jets, sprays or mists can enhance fires by feeding oxygen to the system.
5. "Boil over" can result from a heat wave moving down through burning fuel floating on water. On reaching the water this comes to a rapid boil with forcible ejection of burning fuel upward from the surface. "Boil over" occurs only with burning fuel mixtures comprised of both high and low density components (the effect is not observed for pure liquids); the fuel must be floating on

water; and the effect requires several hours to build-up before it is observed.[2]

6. Previous work by us[3] (since confirmed by others[4]) has shown that high humidity facilitates spontaneous ignition by lowering fuel hotsurface ignition temperatures.

None of these are involved in any way with azeotroping or free radical effects as will be described in this report.

The following is a discussion of azeotropic and possible free radical effects which may result on application of water to fires, and which appear to be much more important in fire intensifications than items #1 - #6 above. This is a follow-on of work previously accomplished on this project.[5]

II. AZEOTROPIC AND FREE RADICAL EFFECTS OF WATER ON FIRES

A. AZEOTROPIC EFFECTS

Azeotropy is a well-known and phenomenon, sometimes called "steam distillation", or immiscible phase azeotropy, whereby distillations can be performed at relatively low temperatures for what would ordinarily be very low volatility, high boiling point liquids. A brief description of steam distillation is provided below. (See also reference [6].)

As shown in Figures 1 and 2 for benzene, xylene and water, the boiling point of any liquid or mixture of liquids is that temperature at which the vapor pressure of the liquid system exactly equals the atmospheric pressure (the standard atmospheric pressure being 760 mm Hg). (Benzene and xylene will be discussed in detail in this paper, since these have boiling points which are analogous to the boiling points of volatile JP-4 and less volatile JP-8 fuels, respectively.)

In Figure 1, benzene boils at 80^{o} C at a vapor pressure

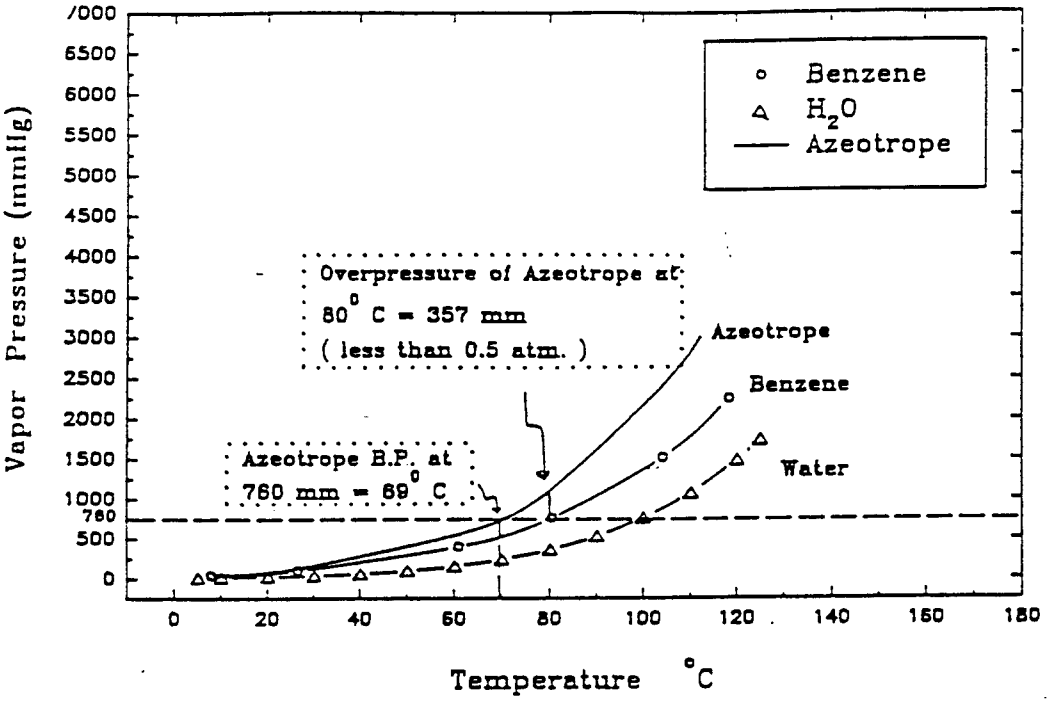


Figure 1.

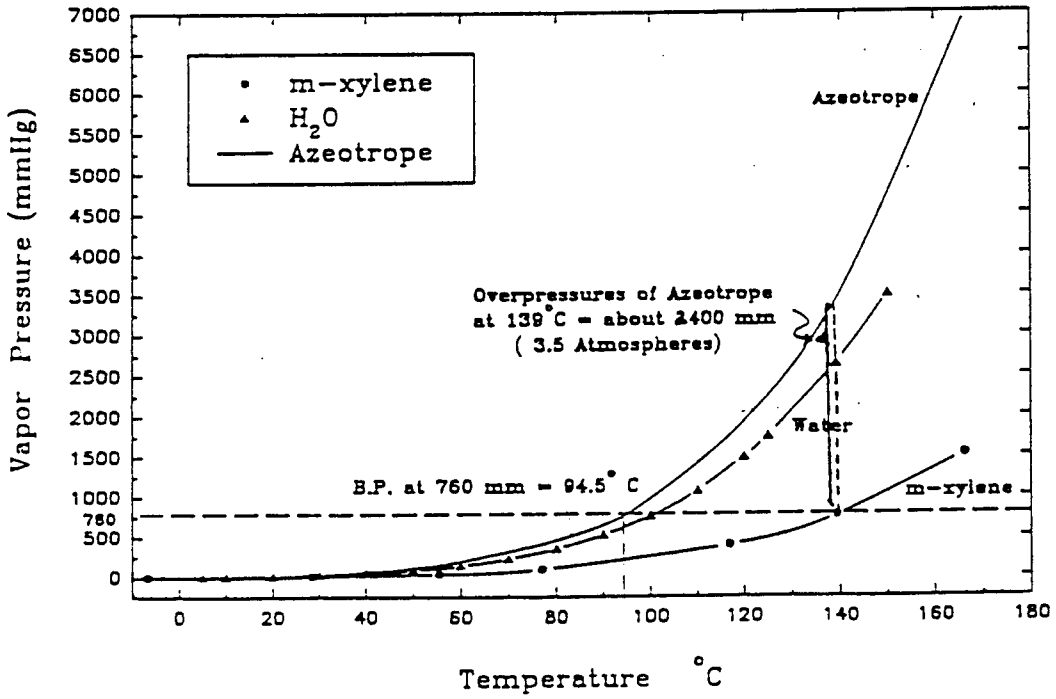


Figure 2.

of 760 mm (one atmosphere), and water has a vapor pressure of one atmosphere at 100° C. The two liquids are insoluble in each other, and each phase exerts its own vapor pressure at a given temperature. The total pressure is then the sum of the vapor pressures for each liquid at that temperature. Thus, at 69° C the vapor pressure of benzene is 533 mm Hg, and of water, 227 mm Hg. Since the total pressure is 760 mm, the mixture will boil at this lower temperature, 11° C lower than the boiling point of pure benzene.

The effect is more pronounced for higher boiling liquids, as seen in Figure 2 for the insoluble mixture of water and xylene (xylene alone boils at 139°C at 760 mm pressure). At 94.5° C the vapor pressures of xylene and water total one atmosphere -- some 45°C cooler than for xylene alone.

These azeotropic effects have serious implications for the flammability of hydrocarbon fuels in contact with water -- e.,g., when fuel fires are being extinguished by water or water based extinguishing agents such as AFFF.

Fuel flammability, and intensity of fire for the fuel, is typically regarded in terms of the fuel's flash point -- i.e., temperature of the liquid at which its vapors are sufficiently present over the fuel to sustain a fire.[7] The more flammable fuels are those with the lower flash points, and fuels with higher flash points are typically regarded as being more safe from the standpoint of such ignitions. Due to these volatility considerations, aviation fuels have undergone dramatic changes since World War II. In 1951 the US Air Force and Army changed from a highly volatile blend of gasoline and kerosene to the less volatile JP-4 formulation which had been widely used by these services until very recently. In 1952 the US Navy adopted a much less volatile blend (JP-5) as a result of the extreme fire fighting constraints peculiar to Navy carrier operations. In 1958 an intermediate blend (less

volatile than JP-4, but more volatile than JP-5) was adopted as Jet A-1 fuel for use in commercial aviation; and since 1968 a slightly modified version of Jet A-1, designated as JP-8, has been gradually implemented for general military use.

The matter of "azeotropic overpressures", referred to in Figures 1 and 2, was early regarded in this work as being an area of prime concern.

In a liquid fuel fire, the surface of the burning liquid is at its boiling point. Although there will be a temperature gradiant in the liquid fuel below the surface, there will be a significant fraction of the liquid fuel beneath the burning surface which will be at a considerably elevated temperature.

If a water-based extinguishing agent (e.g., fog, AFFF, or even a solid stream) is applied to this fire, the incoming water will also be significantly heated as it passes through the flame and into the burning liquid.

If a low boiling point fuel such as benzene (Figure 1), with a high volatility representative of JP-4, is heated to its boiling point (80^{o}), and water is added at a rate such as to allow it to be heated to about this same temperature, the vapor pressures of the water (357 mm) and of benzene (760 mm) total now to 1117 mm. This is an overpressure of 357 mm (about 0.5 atmosphere) in vapor pressure which has suddenly been installed in what had been a gently boiling liquid. The effect will be similar to that which we would observe if we heated the benzene to $92_{o}C$ in a closed pressure cooker, which would now show a pressure of about 7 psi or 0.5 atmosphere on its dial. If we suddenly open the pressure cooker, the contents will erupt in vigorous boiling. This same effect will be observed on addition of water to boiling benzene (or JP-4 type of fuel), without a pressure cooker; and a somewhat increased intensity in the fire will be observed.

If xylene (a low volatility fuel representative of JP-8 components) is heated to its boiling point (139°C) and water is added at a rate to allow it to be heated to the same temperature, the vapor pressures of water and of xylene total now to 3,400 mm -- an overpressure of 2,640 mm (see Figure 2). This is an overpressure of 3.5 atmospheres, again, suddenly unleashed in what had been a gently boiling liquid. This is what we would observe if we heated the xylene to 200° C in a closed pressure cooker, which would now show a pressure of 52 psi or 3.5 atmosphere on its dial. If we suddenly open the lid, the contents will erupt in very violent boiling. This same effect will be observed on addition of water to burning xylene (or JP-8 type of fuel); and an extremely greatly pronounced increased intensity in the fire can be anticipated. Thus, low volatility JP-8 type fuels might be subject to hazardous, sudden and unexpected increases of vaporization of burning fuel during extinguishing operations involving use of AFFF, water fog, or other water based agents, with concomitant increases in flash back, fireballing and similar unexpected flame flare-ups. Such situations could be particularly hazardous for large scale firefighting operations.)

Experiments described on the next two pages were performed to verify the anticipated great increase in fire intensity for non-volatile JP-8 fuel fires on application of water.

No previous attention appears to have been directed to the possibility of increased flammability hazards arising from azeotroping effects from application of water systems to hydrocarbon fuel fires. Statements to the contrary have been encountered in responsible fire manuals:

> " ... water ... entrained in fuel ... is not particularly significant from a fire hazard viewpoint ... ."[8] This is valid for firefighting implications for

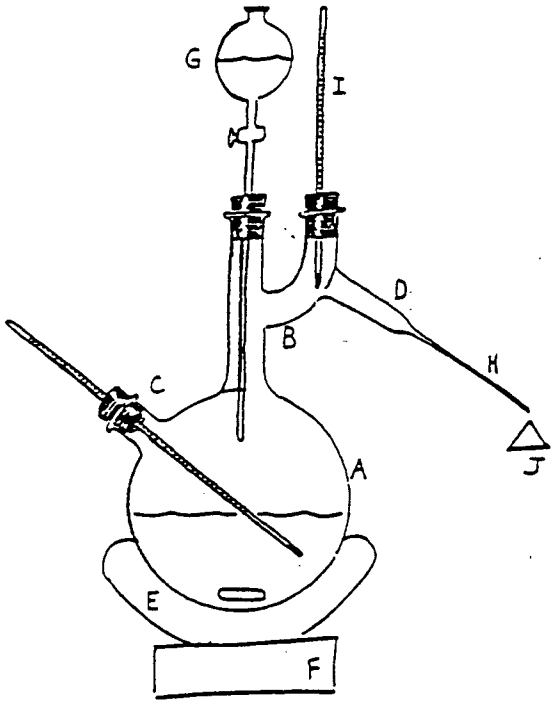


Figure 3. Ignition Flask

A. 500-ml round bottom, with
B. Two-necked Claisen head;
C. Pot thermometer well; and
D. Side arm extending from Claisen head.
E. Electric heating mantle.
F. Magnetic stirrer.
G. Addition funnel for adding water.
H. Capillary extension tube from side arm tube.
I. Head, pot temperature thermometers.
J. Igniter

Not shown: Aluminum foil insulation around assembly; nitrogen tank for purging air from assembly; heating tape for side arm tube; emergency fire extinguishers.

Procedures:

1. Add 100 ml fuel to flask, with magnetic stirring bar.
2. Purge air from assembly with nitrogen tank.
3. Set controls for heating mantle and heating tape to about 20° C above boiling point of fuel.
4. When fuel begins to distill from capillary extension, light distillate with igniter.
5. Add 1 ml water from addition funnel; observe flame growth (if any) at capillary extension.

RESULTS OF EXPERIMENTS WITH XYLENE, AND XYLENE AND WATER;
AND WITH BENZENE, AND BENZENE AND WATER

1. For both xylene and benzene experiments, distillation rates were achieved (prior to adding water) which provided just enough fuel at the capillary to sustain a small flame at the capillary extension tube.
2. For distilling benzene, addition of water did not materially increase the magnitude of the flame.
3. For distilling xylene, addition of water resulted in a huge increase in flame size; see Figures 7 and 8 below.

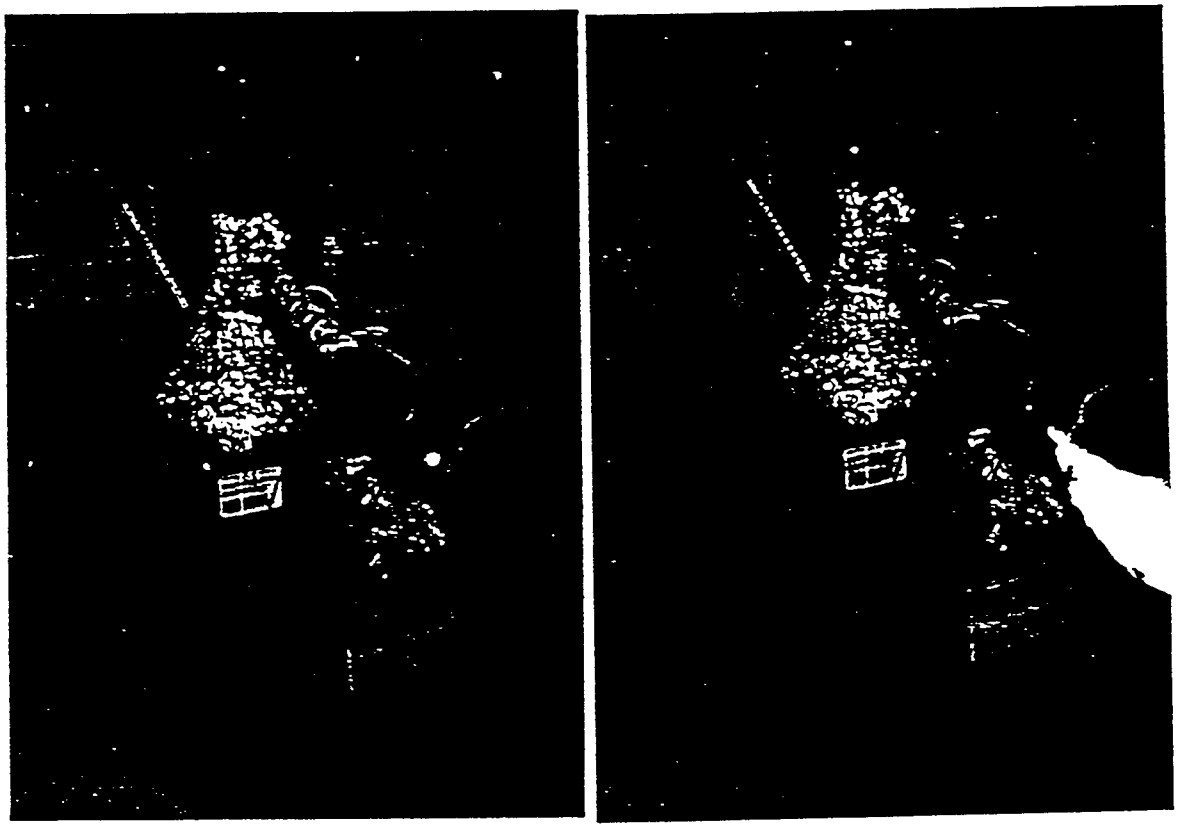


Figure 4. Xylene flame prior to adding water.

Xylene flame is greatly intensified by adding of water.

preexisting entrainments of water in fuels; for high volatility fuels such as JP-4 or AVGAS; or any type of fuel floating on water (except for boil-over effects as previously described). Work accomplished in this project demonstrates, however, that this is not valid for water being applied to large scale fires involving low volatility JP-8, JP-5 or Jet A-1 type fuels.

There is at least one instance in the fire fighting literature which can now possibly be reinterpreted in the light of a possible water/fuel azeotroping effect.

> On 26 May 1981 an EA-6B crashed into several F-14's while landing on the US Navy carrier NIMITZ (CVAN 68). In the ensuing fire 14 men were killed and 42 injured, with $60 million damages to the carrier and its planes. Firefighting efforts commenced immediately, using water hoses and AFFF washdown systems (although AFFF systems were not deployed until well into the fire fighting effort). It was subsequently suggested that possibly there had been contamination of JP-5 fuel in the Navy aircraft by JP-4 fuel as a result of refuelling from an Air Force tanker; and that there had been a reduction in flash point of the Navy jet fuel as a result of the possible admixture with the more volatile JP-4.[9]

A possibility also exists, however, that greatly increased volatilization occurred when the water based extinguishing agents (fog or AFFF) contacted the hot fuel. It is now suggested that this should be investigated from the standpoint of future fire fighting technologies. It should be noted that Halon extinguishing agents will be increasingly unavailable in the future, with an increased reliance on water based extinguishing systems. From the standpoint of Air Force interests, with conversion from more volatile JP-4 to less

volatile JP-8 fuel: since JP-8 is prone to increased vaporization rates in the presence of water, due to azeotropic effects, the need for an in depth evaluation of this effect assumes even greater dimensions of importance. Moreover, the Navy is now using low-volatility JP-5 fuel, and that commercial aircraft are now exclusively fueled with low volatility Jet A-1 (essentially identical to JP-8). Thus, need can be established for examination of the azeotroping effects from the standpoint of Navy and commercial aviation interests, as well.

In summary, the following implications pertain for azeotropic water effects in operational firefighting:

(1) Application of water onto burning fuels can result in an increase rate of volatilization of the fuel, and a correspondingly increased fire intensity will result.

(2) The effect is particularly pronounced for "fire-safe" low volatility fuels such as JP-8, JP-5 and Jet A-1.

(3) Due to high increases in rates of volatilization which can result with low volatility fuels on application of AAAF, water fog or other water-based firefighting agent, it may be best to use halon or alternative halons for supplementary extinguishment.

(4) A need exists for increased firefighter awareness of unanticipated high increases in rates of volatilization for low volatility fuel fires, when using water-based extinguishing agents.

(5) Water suspended in the fuel before the fire will not materially affect the flash point.

In typical firefighting training exercises, a large fire pit is partially filled with water to provide a flat surface for fuel layered to a depth of an inch or less over the water.

(The flat water surface minimizes fuel volume requirements, and serves to cool the fire pit thus minimizing maintenance and pit replacement costs). Until recently JP-4, then more available, was used for training fires. Currently, almost all training fires are conducted with JP-8, reflecting the operational change-over to this less volatile fuel. A diagram of a typical fire pit assembly, using JP-8 fuel, is shown in Figure 5.

As shown in this project, for high volatility fuel fires such as JP-4, there is little effect on volatility when water is added. We have also shown that if water is added to hot non-volatile fuels (JP-8, JP-5 or Jet A-1), there is a serious increase in volatilization rate and a corresponding serious increase in fire intensity for burning fuel.

It is therefore not surprising that azeotroping effects are not observed for the countless number of JP-4 Air Force training fires conducted annually, since for such high volatility fuels water has little impact on volatilization rates. It needs to be emphasized here, however, that the effect will not be observable for training fires involving the low volatility JP-8, JP-5 and Jet A-1 fuel fires, either.

Thus, as shown in Figure 5, for high boiling point (low volatility) fuel training fires there will be a very sharp temperature gradiant in the very thin layer of burning fuel floating on the fire pit's pool of water. At the burning surface, the fuel temperature will be at its boiling point (226^{o}C, or 440^{o}F for JP-8); but an inch or less below this, at the interface of the fuel layer with the underlying water, the temperature drops to ambient water temperature (typically no more than 30^{o}C, or 80^{o} or 90^{o} F). Therefore, almost all of the fuel will be at a temperature which is far below its azeotropic boiling point (in the case of JP-8, about 94^{o}C or 200^{o} F).

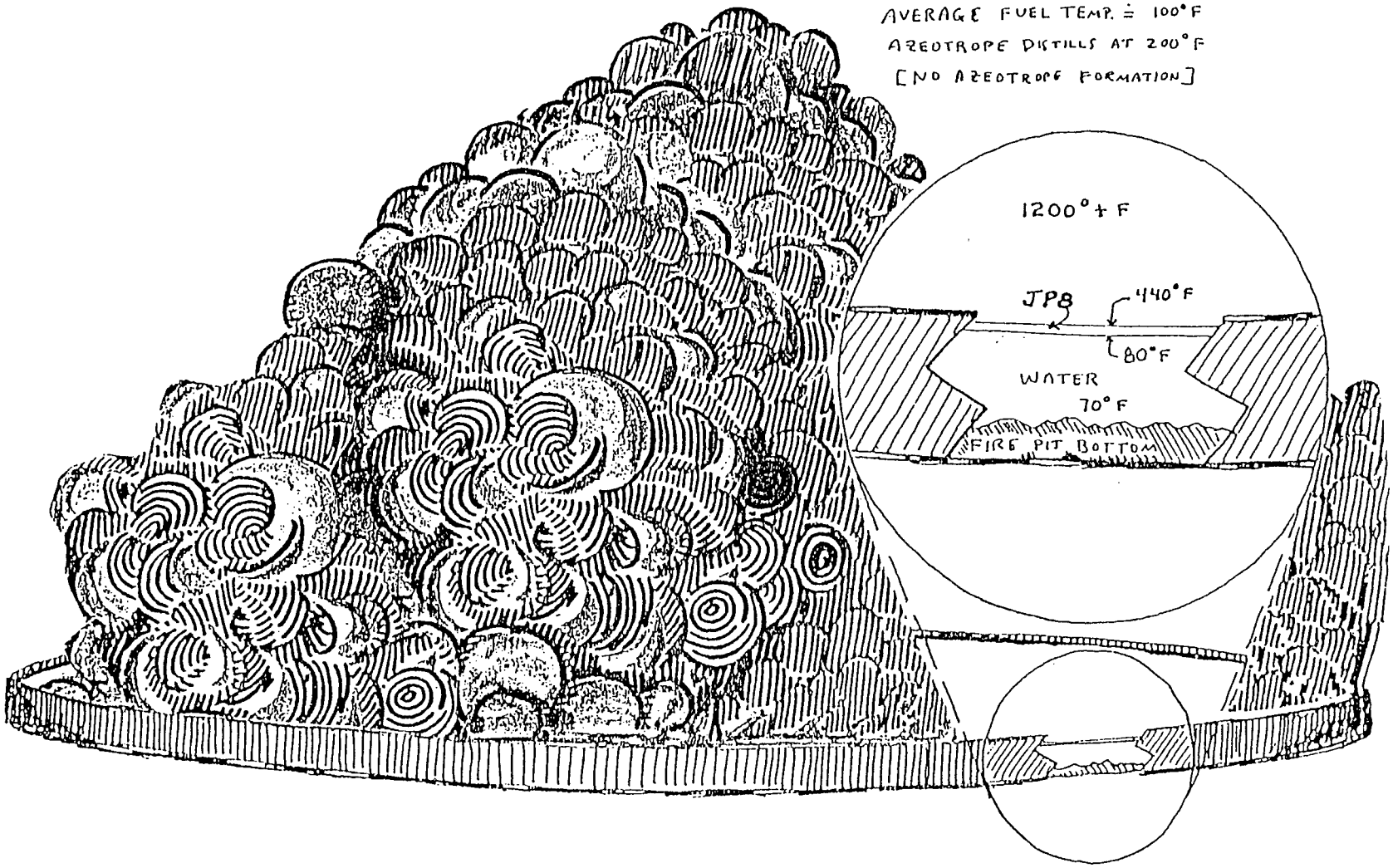


Figure 5. Typical firefighting fire pit assembly. (No significant increase in rate of volatilization will be observed for fuel floating on a water surface.)

Thus, for even the most non-volatile hydrocarbon fuel such as JP-5, there will be no observed increase in rate of volatilization of the burning fuel when water-based extinguishing agents are applied to the fire! The implications of this fact are that current fire research and firefighting facilities can provide no capabilities for:

(1) Demonstration of azeotroping effects which can promote very serious increases in fire intensities when water-based extinguishing agents are applied to low volatility fuel fires.

(2) Investigation of azeotroping effects on a practical real-life research scale, with a view to:
   a. Evaluation of situations which are most conducive to development of the effect;
   b. Development of technologies to miminize formation of the effect in firefighting operations; and
   c. Development of firefighting technologies to minimize the effect after it has occurred.

(3) Demonstration of the azeotroping effect and of appropriate firefighting techniques to prevent it, and to minimize it if it established an important component of the fire.

B. POSSIBLE ROLE OF FREE RADICAL EFFECTS IN ANOMALOUS FLAME INTENSIFICATION BY APPLICATION OF WATER

Subsequent to the completion of this group's preliminary experimental work on azeotropic intensificaton of hydrocarbon fires by application of water, another group reported its preliminary findings on this same phenomenon.[10] Although no mechanistic details were discussed, mention was made in their report of possible "chemical enhancement" of flames by interactions with water.

The only likely candidates for chemical involvements would be free radicals. A strong argument against free radicals might exist in the fact that, as this group has shown, water is a fire intensifying factor most significantly for high molecular weight (high boiling point) fuels, although fuel molecular weight is not usually considered important to ease of formation of free radicals. However, we will be proposing catalyzed production of hydroxyl free radicals when water is applied to the flame, these radicals greatly facilitating flame formation by attack on the fuel molecules. High molecular weight fuel molecules are slower at given temperatures than low molecular weight molecules; and high molecular weight molecules obviously have greater profiles. Thus, slower and larger high molecular weight molecules present better targets for free radical attacks.

Another argument against free radical intermediacies can be mounted in terms of overall energetic considerations. Thus, much energy certainly has to be absorbed to cleave water molecules into free radical components, and even though most of this energy would be redelivered to the fire system, there would be a considerable entropy diversion with which to contend. However, as will be detailed in the discussion to follow, other workers have shown that transition element oxides (found in appreciable abundance in all typical flames as a result of oxidation of metallic objects in the flame environment) can catalyze formation of hydroxyl and other oxygen-containing radicals by ultraviolet radiation (produced in great abundance in all flames).[11]

Thus, there are cogent arguments for free radical participation in the anomalous intensification of flames by application of water, and it therefore it is urged that this possibility should be included in this investigation.

The basis for dissociative processes into free radicals or free radical negative ions was first presented in 1971 by one of the investigators in this project, Dr. Alex Green.[12] Of particular interest are the low lying dissociative states [H + OH and O + H_2], requiring about 5 eV excitation. It is not likely that these can be excited directly in single photon processes from light available in typical fires since the wavelength required:

$$\text{wavelength} = 1240 \text{ ev/E} = 1240/5 = 248 \text{ nm}$$

is too short for radiation that provides most of the light available in smoking fires.

However, as is also shown in Figure 6, possible processes involving vaporized water negative ions may be more promising, since dissociation of H_2O into hydroxide ion (OH^-) and hydrogen radical (H) requires only 3.2 eV radiation, or wavelengths shortward of 388 nm. Here the strong OH peak in hydrocarbon flames at the 306 - 309 nm range[25] can convert $H2O_-$ ions into OH^- ion and H radical; and the electron can be photo-detached from hydroxide to form the hydroxyl (HO) radical near the peak of the Planck spectrum.

We thus come to the question as to how the negative water ion (H_2O^-) might be formed. It is well known that fine sprays of most liquids are frictionally charged[8]. Moreover, prominent electrical field effects have been well demonstrated to be important characteristics of flame chemistry. Thus, negatively charged water molecules can certainly form and survive in mist and vapor form in fire environments.

Moreover, other workers[11] have noted that ultraviolet (UV) light is capable of dissociating water containing catalytic traces of ions of such transition elements as titanium into hydroxyl and other free radical species. As noted above,

hydrocarbon fuel flames are strong emitters of UV radiation, and typically many transition metal oxides are present in most fire environments, and are found in significant concentrations in smoke particles. These metal oxide and other ionic species can generate electron holes under exposure to the Planck spectrum; as electrons tend to migrate to the surface in such particulates, these can readily be captured by water molecules which collide with semi-conducting species, thus forming the negative water ions, which then proceed to form the oxygen and oxygen containing radicals which are well known to facilitate combusion processes. The heat of combustion could be also considered to provide pronounced enhancement of such dissociative processes.

Thus, although there is no overall gain in energy due to the interaction of water, there can be very significant conversion of ultraviolet energy within the flame, under the catalytic effect of trace quantities of transition metal species in interaction with water to form hydroxyl and other oxygen containing free radicals, thereby facilitating the fire. In effect, large amounts of ultraviolet energy which would otherwise be radiated away from the fire zone can be converted to thermal energy for enhanced propagation of the flame front. As an important phase of this investigation, we propose to investigate the possibility that water mists can thus enhance the combustion process.

Preliminary spectroscopic results recently observed at the University of Florida tend to confirm that the hydroxyl free radical population is substantially increased when steam is gently introduced into a lab-scale heptane pool fire. Soot formation was also substantially decreased, and the flame size was substantially increased.

ACKNOWLEDGMENTS:

I wish to acknowledge with great gratitude the assistance provided to me in this work by the Air Force Office of Scientific Research at Bolling AFB and the Wright Laboratories at Wright-Patterson AFB, who provided the necessary funding for the work; to RDL Corporation at Culver City, California which administered the Summer Faculty Research Program; to Mr. Richard Vickers and Dr. Charles Kibert at Tyndall AFB, who supervised my efforts; and to Dr. Alex Greene at the University of Florida who assisted me in all aspects of the work and who provided original concepts regarding free radical effects which are important in the effects of water on fire extinguishment.

REFERENCES

1. Hucknall, D. J. "Chemistry of Hydrocarbon Combustion", New York, Chapman and Hall, 1957.

2. Verhvalin, C. H. "Fire Protection Manual", 3rd ed. (Gulf Publishing Co., Houston), Vol. 1, pp. 139-143.

3. Bannister, W.; Floden, J. "Autoignition. I. Thermo-electric Effects of Hot Surfaces: An Ionic Mechanism for Spontaneous Ignition." Proc., Am. Chem. Soc. nat'l meeting, Dallas, TX, April 1989; Proc., Combustion Inst., Eastern States Meeting, Tampa, FL, Dec. 1988.

4. Smyth, K. C.; Bryner, N. P. "Short-Duration Autoignition Temperature Measurements for Hydrocarbon Fuels." Nat'l Inst. of Standards and Technology Report to US Air Force Engineering and Services Center, Tyndall AFB [JON 2104-3039], May 1989 - September 1990; Proceedings, Combustion Institute Eastern States Meeting, December 1990.

5. Bannister, W. W. "Anomalous Effects of Water in Firefighting: Facilitation of JP Fires by Azeotropic Distillation Effects". Final Report to Universal Energy Systems, Inc. [Dayton, OH] for 1989-1990 AFOSR Research Initiation Grant; June 1991.

6. Hunt, H. "Physical Chemistry", New York, Thomas Y. Crowell Co., 1947; pp. 233-234.

7. "Fire Hazard Properties: Flammable Liquids, Gases, Volatile Solids", Nat'l Fire Prot. Assn., Boston, MA, 1960.

8. Appendix A, Standard 407, NFPA Fire Codes (National Fire Prevention Association, Boston, MA); p. 407-23.

9. Carhart, H. W., et al. "Aircraft Carrier Flight Deck Fire Fighting Tactics and Equipment Evaluation Tests". NRL Memorandum Report 5952, Feb. 1987.

10. Atreya, A.; Crompton, T.; Agrawal, S. "Chemical Enhancement of Combustion During Fire Suppression by Water". Nat'l Inst. of Standards and Technology conference on fire research, Rockville, Md., Oct. 18-20, 1993.

11. (a) Matthews, R. W. "Photo-Oxidation of Organic Material in Aqueous Suspensions of Titanium Dioxide'. Water Research, Vol.20 No. 5, pp. 569-578 (1986).

    (b) Turchi, C. S.; Ollis, D. F. "Photocatalytic Degradation of Organic Water Contaminants: Mechanisms Involving Hydroxyl Radical Attack". J. of Catalysis, Vol. 122, pp. 178-192 (1990).

12. Green, A.E.S., et al. "Micro-Dosimetry of Low-Energy Electrons". Proc., symp. on Biophysical Aspects of Radiation Quality, pp. 79-98, Int'l Atomic Energy Agency [Vienna, Austria], March 1971.

The Statistical Target Model in The J-Mass Environment

Larry A. Beardsley
Assistant Professor of Mathematics
Department of Mathematics
Waters Hall
Athens State College
Athens, AL 35816

Final Report for
Summer Faculty Research Program
Wright Laboratories/MNSH
Eglin Air Force Base, FL 32542-5434

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

September, 1994

RFSIG Target Model Integrated With the Joint Modeling and Simulation System (J-MASS) Environment

Larry A. Beardsley
Assistant Professor
Department of Mathematics
Athens State College
Athens, Alabama

Abstract

As attested to in last summer's report, the J-MASS architecture is a relatively new modeling system designed to support engineers, model developers, analysts and decision makers. J-MASS is written in the object-oriented DOD-standard Ada language. It is designed to be transportable between different J-MASS compliant hardware configurations and to operate on workstations using a Posix-compliant Unix operation system. The current beta test site 2.0 J-MASS release provides almost total functionality through the system environment, allowing a user to log onto J-MASS and develop components, assemble them into models, configure a simulation scenario and place players within the scenario, execute the simulation, and analyze the results through post-processing. Currently, the WL/MNSH and WL/MNMF branches have tested the 3DOF missile code under the J-MASS architecture, and plans have begun for creation of the 6DOF code into the recommended architecture.

My tasks this summer were to rewrite the statistical target model currently written in Fortran into Ada. Also, I worked on an Ada shell which allows for the passing of data from a fortran program to an Ada program. This involved such considerations as reading the Ada boolean "True" or "False" and converting it to the Fortran boolean "1" or "0" respectively. The reason for the writing of the shell was to provide a means for an already fortran program to function in conjunction with other programs written in Ada. If time permitted, I was to perform Monte Carlo analyses on data on a sun workstation. The analyses had been performed on a VAX system, but had not been attempted on a SUN workstation.

The above tasks consisted of understanding of the statistical target model, programming in the Ada language, and a fairly good understanding of J-MASS. The first few weeks I spent studying the model, the next couple working on the shell, and the remaining weeks I devoted to writing, compiling, and debugging of the newly written Ada version of the statistical target model. The last task was carried out in two parts. First, the programs were written and compiled on a VAX system, and then the code was transported to a Sun workstation for compilation.

ACKNOWLEDGMENTS

I wish to thank the Air Force Material Command and the Air Force Office of Scientific Research, whose sponsorship provided me with another opportunity to participate in a challenging and stimulating research project. I also thank the Armament personnel for providing the enriching environment in which to work. Also just as important is the friendliness and helpfulness which has been widely manifest throughout the Armament Directorate both summers that I have conducted research there.

Several individuals in the Armament laboratory at Eglin Air Force Base have contributed greatly to making my summer tenure both summers enjoyable and productive. In particular in the Simulation and Assessment branch, Capt. Patti Egleston was always available to answer questions about a project that I was working on, and has been encouraging both summers. Tim Pounds, also in the same branch, has consistently been helpful in answering questions related to both hardware and software and has an uncanny knack of solving immediate problems quickly. Over the two summers that I have been here, I not only learned quite a bit about the simulations in operation: the Ada DOD-standard language and J-MASS; but also I have a strong appreciation for the depth of talent of both the civilian and Air Force personnel who are employed in the Armament Directorate of Wright Laboratory. My focal point both summers was Larry Lewis who has the ability to challenge personnel in a very positive way. It was a very positive experience working in his branch both summers.

This summer, as was true with last summer, I found that the environment that I worked in was very conducive to my work. The books, hardware, and software were readily available. Last, but not least, the palm tree outside the window next to where I sat, was inspirational. I am not sure that was planned, but I appreciate it.

I am very thankful to PhiPhi McGrath for undertaking the Ada conversion of the target model on the VAX system before I worked on the conversion for the SUN and IBM systems. I am grateful for all of the friends that I made in the branch including Dr. Martin Moorman, Ed Moorman, and Yves Mallette.

The Statistical Target Model in the J-Mass Environment

Larry A. Beardsley

INTRODUCTION

My primary tasks as a researcher this summer were to gain a through understanding of the RF (radio frequency) statistical target model which will be referred hereafter as RFSTAT, and to create an Ada wrap around shell so that RFSTAT can be used in the J-MASS environment, and to rewrite the fortran target model into Ada. The purpose of a shell is to provide a means for programs currently written in Fortran to be useable in the J-MASS architecture which is written in Ada, the DOD standard language, without having to rewrite the Fortran program into Ada.

There is a trade-off in having a Fortran program interface in a J-MASS environment (for a description of J-MASS see "The Integration of MOMS with MSTARS in the J-MASS Environment" by Beardsley[1]) via a shell which allows data to be passed back and forth from the Fortran program to the J-MASS environment. With the use of a shell, programs will usually run more slowly than they would if written in Ada. However, with the processor speeds of today, the additional time for a program to run using an shell interface may pale in comparison to the time it may take to reengineer a lengthy forrtran program into Ada. Therefore, careful thought should be given before deciding to rewrite current operational code into a new language.

RFSTAT that I was requested to rewrite is not a lengthy program. It contains seven subroutines, which altogether rewritten will comprise about twenty pages of code, or about twice the length of the Fortran written version. The main hurdle I crossed was that of fully understanding the earlier fortran version and rewriting it with my minimal understanding of Ada. However, I am grateful to PhiPhi McGrath, who after already having done a conversion from Fortran to Ada, was willing to help rewrite the statistical target model into Ada. In reality, she did the majority of the writing on the Vax system, and my subsequent task became making it run on the SUN/IBM RS/6000 was to make it run in a Unix

1 What are the features of the model?

(1) It can be used as a generic target with a specified radar cross section (RCS).

(2) It can be used with measured data, suitably prepared, as an input.

(3) The output results are dependent on actual antennae size.

(4) The model correctly accounts for correlation between RCS and glint and also for correlation between vertical and horizontal glint.

(4) The model accounts for signal time correlation based on target aspect angle rate.

(5) The model is based on a model developed by TSI, Inc for the AFATL RFTS facility and modified Dynetics, Inc for a digital seeker simulation.

2 Functional Description of the Statistical Target Subprogram

The main driver of the statistical target subprogram determines the azimuth and elevation cross-correlation terms, glint bandwidths, and a cholesky decomposed covariance matrix. Also, the systems and control matrices for a second order butterworth filter are developed. The main routine initializes and settles the filter at startup and whenever the transmit frequency is changed. The two glints, mean azimuth and mean elevation, and the radar amplitude are calculated upon receipt of the filtered azimuth and elevation signals.

(1) What is the function of the butterworth filter update subprogram?

This subroutine generates 3 independent, complex random variates with user specified mean raleigh magnitude and uniform phase. The random variates are then modified by the elements of a previously computed matrix which introduces the desired correlation between the variates. The resultant complex variable represent the signals received from a far-field target by a 4 port-antenna interferometer. The necessary fourth signal is synthesized from the other three. A 2nd-order low-pass butterworth filter provides the spectral shaping of the correlated signal components. The glints and RCS bandwidths are controlled by making the filter cutoff frequency a function of aspect angle rates. In order to minimize filter drift this routine is called at the basic simulation update rate. When the cutoff frequency violates the nyquist interval, the filter is bypassed. This program calls the program "GAUSS", a gaussian random number generator.

(2) How does Gauss work?

Except on the first call, "GAUSS" returns a pseudo random number having a gaussian (that is, a normal) distribution with zero mean and "sig" standard deviation. Therefore, the density is:

F(x) = exp(-5.0*x**2)/sqrt(2.0*PI) if "sig" equals 1.0.

The first call initializes "GAUSS" and returns zero. "GAUSS", in turn calls "RANU". It is assumed that successive calls to "RANU()" give independent pseudo random numbers distributed uniformly on (0,1) , possibly including 0 but not 1. The method used was suggested by Von Neumannn, and improved by Forsythe, Ahrens, Dieter, and Brent. On the average there 1.3777 calls of "RANU" for each call of "GAUSS". The original name of "GAUSS" was "GRAND" and it was published in algorithm "488" in the collected algorithms from "CACM".

(3) A description of RANU

This function generates a uniform distribution of random numbers between 0.0 and 1.0. Before its use, the generator should be seeded. Any integer in the range from 1 to 2147483646 will serve as a good seed. The function is portable to any system that has a maximum integer value of 2**31 - 1 or greater.

Conclusions

Several weeks of the summer were spent learning the basic theory of the statistical target model. I devoted two weeks considering how to write a shell for the fortran statistical target subprogram that would allow for the passing of data from an Ada routine to a Fortran routine. The remaining five to six weeks of the summer, I channeled my efforts toward rewriting the fortran code, main driver and all subroutines into Ada. This task was began by Ms. McGrath on a digital VAX system. Her prior experience with the target model and Fortran to Ada reengineering gave me helpful support. As she wrote the routines, with some input from me, I compiled and debugged the programs on a UNIX based system. At the time of this writing, all of the routines but one had compiled on the VAX system. However, a supplementary report will be issued yielding the results of this task as well as the functionality of the Ada shell.

A great deal has been learned in the time that I devoted to both studying to understand the target model, as well as understanding the Fortran version and time spent in its conversion.

As previously stated, a shell may be useful for interaction between Ada and lengthy Fortran programs. However, for short programs reengineering Fortran to Ada is reasonable. Of course, all of this is taking into account that the current DOD standard language is Ada; therefore, many of the newer programs will be written in Ada from the start.

REFERENCES

1. "The Integration of MOMS with MSTARS in the J-MASS Environment", Larry A. Beardsley, AFOSR Summer Research Program Final Report, September, 1993.

1. "The Technical Cooperation Program, TCCP" , U.S. National Report, Larry E. Lewis and Josephine Tran", August, 1987.

MULTIGRID METHOD FOR LARGE SCALE ELECTRONIC STRUCTURE OF MATERIALS

Thomas L. Beck
Associate Professor
Department of Chemistry

University of Cincinnati
P. O. Box 210172
Cincinnati, OH 45221

Final Report for:
Summer Faculty Research Program
Wright Patterson Air Force Base

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

September 1994

MULTIGRID METHOD FOR LARGE SCALE ELECTRONIC STRUCTURE OF MATERIALS

Thomas L. Beck
Associate Professor
Department of Chemistry
University of Cincinnati

Abstract

A novel method for density functional theory calculations was developed. The Kohn-Sham equations were solved entirely in coordinate space using a finite difference algorithm. The method employed the recently developed multigrid algorithm for solving both the Poisson equation and the electronic variational problem. Order of magnitude accelerations were obtained relative to solution on the finest grid alone. Numerical examples are presented for atomic problems. If the orbitals are localized, the method scales linearly with the number of electrons. Therefore, it holds promise for large scale *ab initio* simulations of materials. Future applications to computation of nonlinear optical properties are discussed.

MULTIGRID METHOD FOR LARGE SCALE ELECTRONIC STRUCTURE OF MATERIALS

Thomas L. Beck

Introduction

Quantum chemical methods have become tools of routine use in both theoretical and experimental laboratories. One can now obtain software packages which perform accurate computations on relatively large molecules, and the computations can be carried out on desktop workstations in reasonable amounts of time. The methods and software required to carry out these calculations have been developed with thousands of years of total human labor, and the codes are now relatively efficient and user friendly.

Several questions can be asked. First, why develop more efficient methods? Traditional quantum chemical methods are founded on basis set expansions of the electronic wavefunctions and solution of the Schrödinger equation via matrix methods. A fundamental difficulty is that the computer time required scales at least as severely as N^3, and often (for the more accurate methods) as N^4 or higher. Major advances have been made which allow computations for systems of 50 to 100 electrons on supercomputers (vector or parallel), but the scaling problem will ultimately prevent calculations on systems with thousands of electrons even accounting for any foreseeable advances in computer technology.

Second, why do we need to do *ab initio* calculations on systems with thousands of electrons? Many problems can be modeled quite accurately using effective potentials obtained from extensive

theoretical calculations and refinements based on experiment. However, there are a wide range of phenomena which require explicit inclusion of the electrons. Notable examples are: chemical reactions in liquids, electronic structure of disorderd solids, electron transfer reactions, and nonlinear optical properties of polymers. Each example requires explicit inclusion of a large number of electrons for an accurate treatment which transcends traditional use of model potentials. Even larger systems which will be treated fully quantum mechanically one day are interactions of solvent and solute molecules with segments of DNA strands and the fracture of metallic solids and alloys under stress. These kinds of problems are hopeless with existing electronic structure methods.

My research this summer focused on the development of more efficient means of carrying out large electronic structure calculations. In the last twenty years, applied mathemeticians (led by Professor Achi Brandt of the Weizmann Institute) have developed a new approach to solving partial differential equations called *multigrid* which dramatically increases the convergence properties of iterative methods.[1,2] I worked on the inclusion of multigrid for two aspects of the electronic structure problem. The first is the solution of the Poisson equation for an arbitrary distribution of charges. The second is for the solution of the quantum variational problem itself. I obtained preliminary promising results in my lab at the University of Cincinnati for one dimensional N electron problems. During my eight weeks at Wright Patterson Air Force Base, I wrote an extensive computer code to carry out a multigrid calculation which located the ground state electron density for many electron atoms. The results suggest that the multigrid method gives at least an order of magnitude acceleration of the calculation in relation to iteration on the finest scale. My findings imply that the multigrid approach holds promise for large scale *ab initio* simulations of materials.

The research was carried out in collaboration with Dr. Ruth Pachter. An interest of her research group is to calculate nonlinear optical properties of polymeric materials. It is our hope that the multigrid approach will lead to realistic calculations on large systems such as polymers in the presence of strong electromagnetic fields.

Theory

The underlying theoretical foundation to the electronic structure calculations described here is electron density functional theory(DFT).[3-5] This theory has a long history dating back to the work of Thomas and Fermi on the electron gas. The theory was formalized by Hohenberg and Kohn and then turned into a viable computational method by Kohn and Sham. DFT has been the predominant computational method in solid state physics, and recently has become more popular in quantum chemical applications (in relation to say Hartree-Fock theory). One major advantage of DFT over Hartree-Fock (HF) is that the effective potential operator is *local*, although approximate, whereas in HF the exchange operator is *nonlocal*. Hence, the DFT calculations generally require much less computational effort while giving comparable accuracies for many molecular properties. Several direct comparisons have been made in recent years, and certainly there is room for large improvements in the computation of the important exchange-correlation energies in DFT. The important theme of DFT is that the ground state energy can *in principle* be computed from a knowledge of the one electron density, $\rho(\mathbf{r})$.

Kohn and Sham[5] developed a useful numerical method by introducing one electron orbitals into the calculation; this procedure yields much more accurate kinetic energies than the relatively crude

Thomas-Fermi theory. Then, by solving a set of self consistent one electron equations, one can obtain the ground state electron density and hence the total electronic energy. A term in the one electron effective potential was added which includes contributions of both electron exchange and correlation, in an approximate way. Typically, it is assumed that this term of the potential is that for a uniform electron gas at density $\rho(\mathbf{r})$. Exact numerical Monte Carlo results have been obtained for the uniform electron gas.[5] Surprisingly, this approximation works well for a wide range of solid state and chemical applications.

The Kohn-Sham total electronic energy can be represented as (we consider only doubly occupied states here):

$$E[\{\psi_i\}] = 2\sum_{i=1}^{N/2} \int \psi_i^* \left[-\frac{1}{2}\nabla^2\right] \psi_i d\mathbf{r} + \int \mathrm{v}_{eff}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} \tag{1}$$

where the ψ_i are the Kohn-Sham orbitals, the effective potential is:

$$\mathrm{v}_{eff}(\mathbf{r}) = \mathrm{v}_{ext}(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{\mid \mathbf{r} - \mathbf{r}' \mid} d\mathbf{r}' + \mathrm{v}_{xc}(\mathbf{r}). \tag{2}$$

and the electron density is:

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{N/2} \mid \psi_i(\mathbf{r}) \mid^2 . \tag{3}$$

Typically, the external potential is that due to the nuclei and the exchange-correlation potential is

computed at the LDA level.

Iterative minimization of the energy is carried out by solving the following steepest descent equation:

$$\dot{\psi}_i(x) = -\frac{\delta E[\{\psi_i(\mathbf{r})\}]}{\delta \psi_i^*(\mathbf{r})} = 0, \tag{4}$$

subject to the constraints of orbital orthonormality:

$$\sigma_{ij} = \int \psi_i(\mathbf{r})^* \psi_j(\mathbf{r}) d\mathbf{r} - \delta_{ij} = 0. \tag{5}$$

These constraints are enforced iteratively at each propagation step using the SHAKE method developed for simulations of rigid molecules. (We are currently introducing more efficient and accurate methods for the minimization and orthogonalization steps, namely conjugate gradients and Gram-Schmidt, respectively.[4])

If one takes a functional derivative of Eqn. 1 with respect to $\psi_i^*(\mathbf{r})$ and sets it equal to zero, the traditional self consistent one electron equations are obtained. These equations are typically solved by matrix methods. In our work, we rather minimize the total energy directly as described above. This procedure is the basis of the well known Car-Parrinello method,[6] which incorporates plane wave basis states and uses repeated application of FFT methods to obtain rapid convergence. The basic problem with their method is that it still suffers from the N^3 scaling problem since they represent the orbitals with completely delocalized basis functions. This causes the orthogonalization

step to scale as the volume cubed.

The point of our work is to do all of the minimization directly in coordinate space, in which case we can exploit any possible localization of the orbitals. If the orbitals are localized, then we can surmount the N^3 barrier. In addition, we can exploit the substantial efficiency gains and scaling properties of multigrid methods. With the inclusion of multigrid, the method scales in principle *linearly* with N. There have been several developments in recent years which essentially lead to propagation in coordinate space.[7] To my knowledge, only two attempts have been made at a multigrid process for electronic structure.[8,9] The first employed multigrid for the Poisson equation and the authors solved several *one orbital* problems.[8] In the second paper,[9] a multigrid approach was mentioned at the end of a conference proceedings paper; no details of the method were given, and results were presented for one orbital problems only. I believe the results presented below are the first application to a multi-orbital problem (the Ne atom, 10 electrons).

Details

First, we must represent the relevant equations (Poisson and Schrödinger) in coordinate space. The electrostatic potentials and electron orbitals are then represented simply by their numerical values on the cartesian grid. Then, the multigrid method is utilized to accelerate solution of both problems. The Poisson equation must be solved at each update of the orbitals to generate a new effective potential. Once the global minimum has been reached to within some tolerance the minimization process is terminated.

In our work, we have used a nine point finite difference formula (in one dimension) to represent

the kinetic energy operator. Relatively high accuracy is required for this operator due to the strong and singular potentials near an atomic nucleus.

The iterative process then proceeds as follows: 1) make initial guess at orbitals (orthogonal) 2) solve Poisson equation and generate effective potential 3) compute orbital 'forces' by applying the Hamiltonian in coordinate space 4) move the orbitals 5) reorthogonalize the orbitals 6) compute total energy 7) return to step 2.

This process is solved first on a coarse scale. The solution is then passed to the next finer scale (grid spacing halved) by interpolation. These interpolated functions are then used as the initial state on the finer scale where a new set of iterations are begun. This whole sequence is called nested iteration, and is not a full multigrid cycle. The full multigrid is well-described in Ref. 1. We have written a full multigrid code for the Poisson equation and are currently adapting Brandt's FAS scheme for solving our minimization problem (with helpful advice from Prof. Brandt).[2] We have observed linear scaling and rapid convergence for solution of large Poisson problems in 3-d periodic boundary conditions. The present results are nested interation results for atomic structure. Even with this limited form of multigrid, dramatic accelerations are observed; the full multigrid cycle should perform substantially faster.

Numerical Results

The majority of my research this summer was directed at writing a large-scale computer code to do calculations on realistic atomic systems. Then, I spent several weeks testing the code on several atomic problems which have been solved by other means in the literature. Here I will present only

numerical results which illustrate the method for the He and Ne atoms. The He atom is a one orbital problem and the Ne atom is a five orbital problem with all the complexity of a general *ab initio* calculation. The calculations were carried out on various Silicon Graphics workstations at Wright Patterson Air Force Base in collaboration with Dr. Pachter.

Figure 1 presents the convergence behavior of nested iteration vs. iteration on the finest scale alone for the He and Ne atoms. Clearly the nested iteration yields large scale speedups on the convergence properties for both atoms. In addition, the grid method yields physically reasonable results for the total energy, namely within a couple of percent of accepted literature values. The absolute value of the total energy is likely off since we represent the nucleus as a slightly distributed charge. The largest concentration of charge is at the nucleus of course so errors due to the grid representation are largest there. However, for most quantities of interest (except perhaps NMR) the exact behavior at the nucleus is not crucial. For example, with Ne the 2 1s electrons carry on the order of at least 50% of the total atomic energy. The representation on a uniform grid is poorest in the core where much electron density is localized. Therefore, small errors can be expected in that region, while the second shell is accurately represented. Since this is where chemical interactions take place, the numerical results can be deemed acceptable.

Figure 2 shows the radial density profile for electron density away from the Ne nucleus ($4\pi r^2 P(r)$). The grid method captures the shell effect as the 1s core is clearly visible. Visual inspection of individual orbitals shows that the 1s, 2s, and 2p orbitals are obtained (using Gram-Schmidt orthogonalization).

Future Plans

Currently we are improving the minimization procedures by employing Gram-Schmidt orthogonalization and conjugate gradients minimization.[4] Both lead to substantial improvements over the current method. The next step is incorporation of the FAS scheme of Brandt *et al.*.[2] We are now adapting our method for this full multigrid cycle. This method will be similarly tested on atomic many electron problems. Then a full scale simulation of a large system will be attempted. Our current plans are to minimize the total electronic energy of bulk silicon with an all electron calculation. An advantage of our method is that it is very simple to adapt to periodic boundary conditions. The wavefunctions and electrostatic potential must only match on the boundaries. Therefore it is trivial to employ in one, two, or three dimensions. This circumvents the need for difficult Ewald summation methods say for surface problems.

A second area concerns development of higher accuracy kinetic energy representations. To this end we are examining a new method of functional representation called Distributed Approximating Functionals(DAFs) developed in quantum dynamics.[10] We are also exploring the use of adaptive grid methods for electronic structure.[2] Since the electron density is often quite low in large regions of configuration space (for example in silicon the packing density is low), the grid should be adaptable to do work only where necessary. Multigrid is ideally suited for this problem, and corresponding methods have been developed in computational fluid dynamics.

Our computational interests at this time are 1) nonlinear optical properties of large systems such as polymers. DFT has been used with the derivative methods to compute polarizabilities and hyperpolarizabilities of atoms and molecules.[11] Our methods should extend the size range for

which these calculations are possible. 2) metal-solution and metal-polymer interphases. We are collaborating with Prof. F. James Boerio in our Materials Science and Engineering Department. His group makes SERS and IR measurements on polymer and self assembled monolayer ordering at metal surfaces. The metal-molecule interface is inherently quantum mechanical. To my knowledge, very few fully quantum mechanical simulations have been performed on these highly complex systems.

Acknowledgments

I would like to thank Dr. Ruth Pachter for many stimulating discussions and for her support during the summer. In addition I thank the Air Force for the Summer Fellowship which allowed me to tackle a difficult computational problem in very good conditions!

References

[1]W. L. Briggs, *A Multigrid Tutorial* (Society for Industrial and Applied Mathematics, Philadelphia, 1987); P. Wessiling, *An Introduction to Multigrid Methods* (Wiley, New York, 1992).

[2]A. Brandt, Math. of Comput. **31**, 333 (1977); A. Brandt, S. McCormick, and J. Ruge, SIAM J. Sci. Stat. Comput. **4**, 244 (1983); A. Brandt, Nucl. Phys. B **26**, 137 (1992).

[3]D. K. Remler and P. A. Madden, Mol. Phys. **70**, 921 (1990).

[4]M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).

[5]R. G. Parr and W. Yang, *Density Functional Theory of Atoms and Molecules* (Oxford, New York, 1989).

[6]R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).

[7]See, for example, E. B. Stechel, A. R. Williams, and P. J. Feibelman, Phys. Rev. B **49**, 10088 (1994), for a clear review and discussion of the problem.

[8]S. R. White, J. W. Wilkins, and M. P. Teter, Phys. Rev. B **39**, 5819 (1989).

[9]J. Bernholc, J.-L. Yi, and D. J. Sullivan, Far. Disc. **92**, 217 (1991).

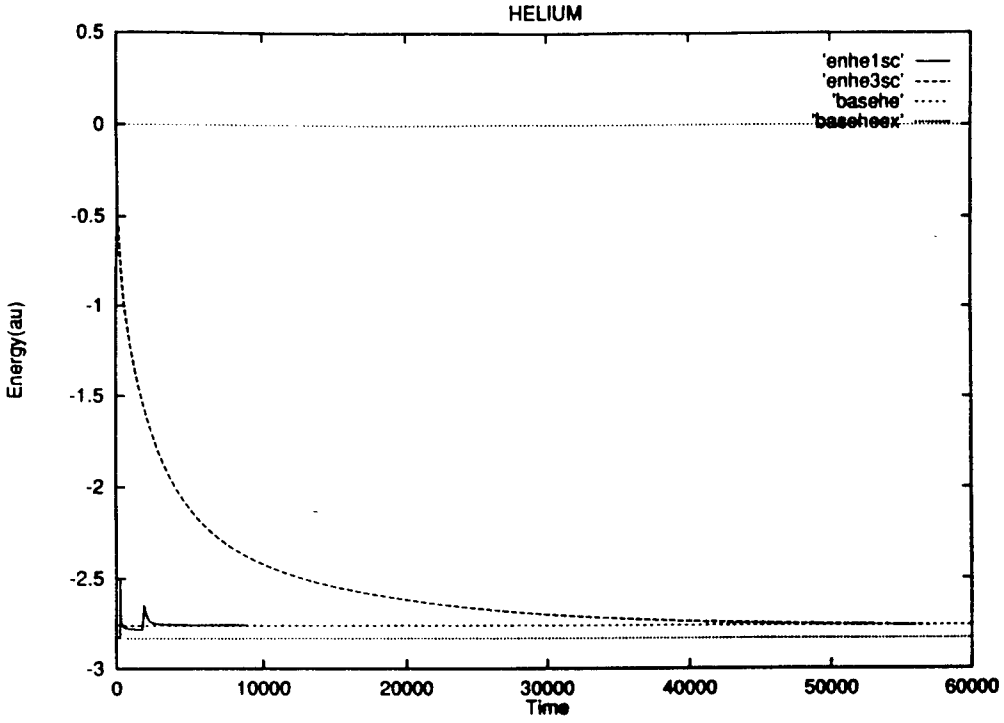
[10]D. K. Hoffman, M. Arnold, and D. J. Kouri, J. Phys. Chem. **96**, 6539 (1992).

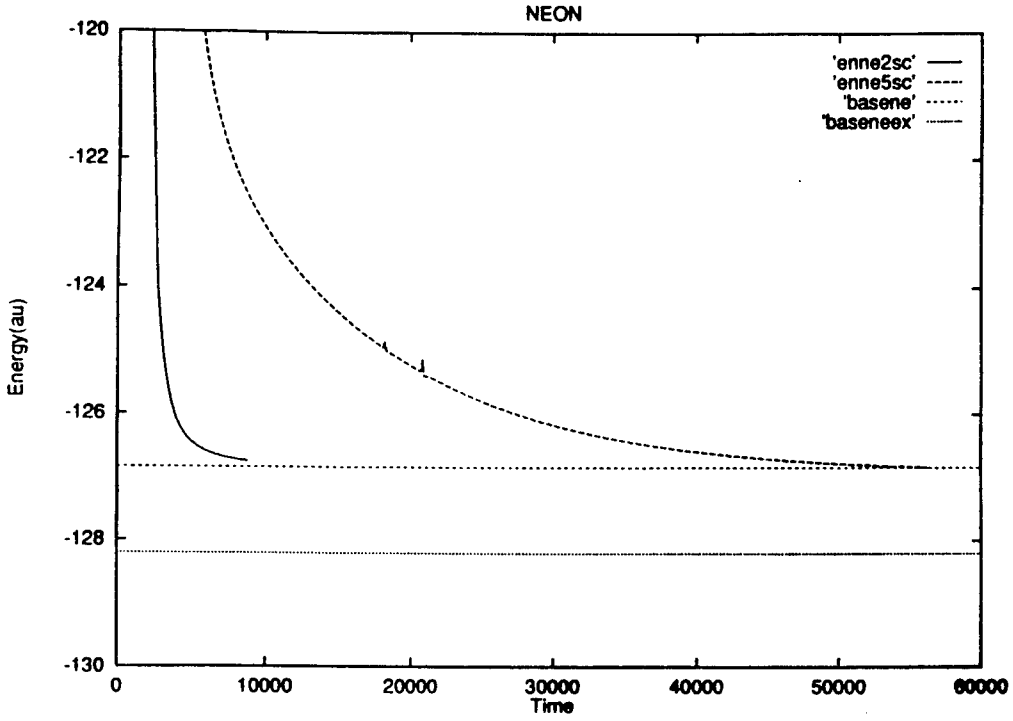
[11]J. Guan, P. Duffy, J. T. Carter, D. P. Chong, K. C. Casida, M. E. Casida, and M. Wrinn, J. Chem. Phys. **98**, 4753 (1993).

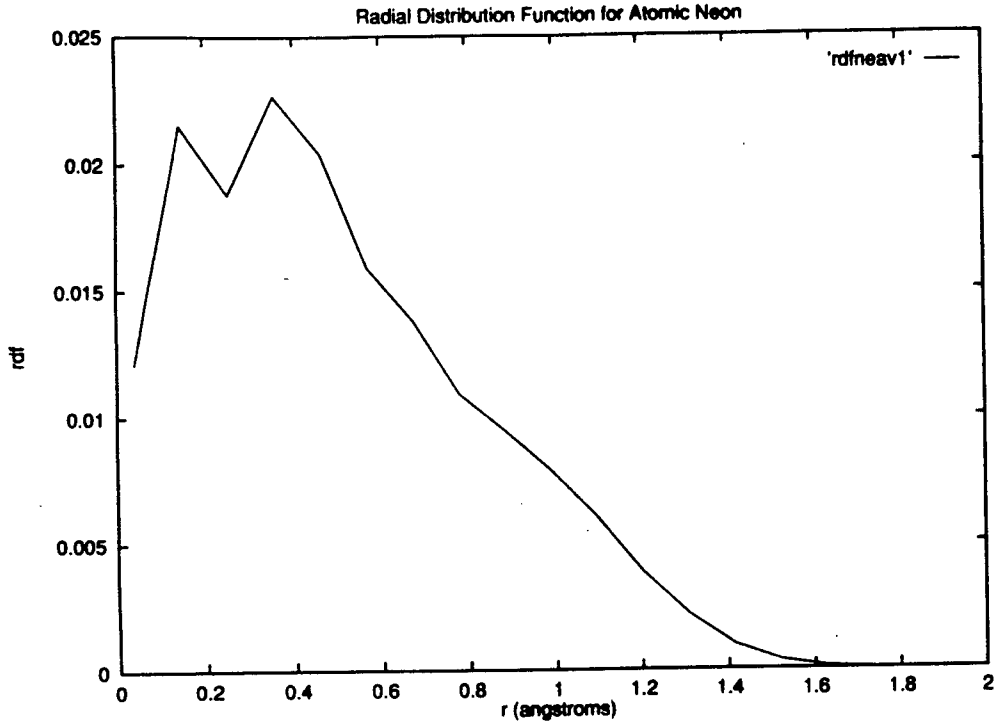
Figure Captions

Figure 1. Convergence behavior for computation of the total energies of the He and Ne atoms. The energies are in atomic units and the 'time' merely gives the relative times for nested iteration vs. fine scale iteration.

Figure 2. The radial distribution function for the Ne atom. Distance is in Å.


HELIUM
'enhe1sc'
'enhe3sc'
'basehe'
'baseheex'
Energy(au)
Time
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
0
10000
20000
30000
40000
50000
60000


NEON
'enne2sc'
'enne5sc'
'basene'
'baseneex'
Energy(au)
Time
-120
-122
-124
-126
-128
-130
0
10000
20000
30000
40000
50000
60000


Radial Distribution Function for Atomic Neon
'rdfneav1'
rdf
r (angstroms)
0.025
0.02
0.015
0.01
0.005
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

Diffusional Creep in Metals and Ceramics at High Temperatures

Victor L. Berdichevsky
Professor
School of Aerospace Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0150

Final Report for:
Summer Faculty Research Program
Wright Labs

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

August, 1994

Diffusional Creep in Metals and Ceramics at High Temperatures

Victor L. 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 could be elastic or plastic dependingly on the level of stresses. For sufficiently low stresses solids behave elasticity. However, during a long time even for low stresses solids develop irreversible deformations. This phenonenom is called creep. Actually, solids creep even at zero external load. This is due to the fact that practically none of polycrystallin bodies is in thermodynamical equilibrium. Energy of a polycrystal can be decreased, for example, moving grain boundaries. This occurs in reality, but very slowly, by means of thermodynamical fluctuations. The rate of changes magnifize significatly by elevating of temperature and applying an external load.

Two major mechanisms of creep are known: deformation created by motion of dislocations and by diffusion of vacancies. The typical deformation mechanism map is shown in stress-temperature plane on Fig. 1 [1]. Above the curve γ (high stresses) the dominating mechanism is dislocation motion, below γ (low stresses)deformations occur due to diffusion of vacancies. It is believed that for low temperatures, vacancies move mostly over the grain boundaries (Cobble creep) while for high temperature motion vacancies through the lattice dominates (Herring-Nabarro creep). Diffusional creep is the leading phenomenom in many technical processes at high temperatures. Superplasticity, sintering, void formation, occur mostly due to diffusional creep. The foundation of the theory of duffusional crepp was laid down by Nabarro [2], Herring [3], Cobble [4], and Lifshitz [5]. Extensive reviews of various aspects of the creep theory can be found in [1, 6-24]. At present, to the best of the author knowledge, only a linear version of the theory of diffusional creep exists, and even the linear theory has some gaps which does not allow one to attack such

problems as quantative theory of superplasticity. The aim of this paper is to develop a thermodynamically consistent theory of diffusional creep which incorporate nonlinear phenomena such as grain rotation in the course of superplastic deformation. The contents of the report is as follows. In Section 2 a logic scheme of the theory is presented, in section 3 the basic kinematical relations are discussed, in section 4 time derivative of free energy is calculated and it's negativeness is used to obtain the closed system of equations. Closed system of equations is presented in Section 5. Future developments are discussed in Section 6.

2. Logical Skeleton of the Theory.

The closed system of equations of theory of elastic bodies consist of equilibrium equations for the stress tensor σ_{ij} (Latin indices run values 1,2,3; summation over repeated indices is implied)

$$\frac{\partial \sigma_{ij}}{\partial x_j} = 0 \qquad (2.1)$$

relation between stress tensor σ_{ij} and tensor of elastic strains $\varepsilon_{ij}^{(e)}$

$$\sigma_{ij} = A_{ijke}\,\varepsilon_{ke}^{(e)} \qquad (2.2)$$

relation between tensor of total strains ε_{ij} and tensor of elastic strains $\varepsilon_{ij}^{(e)}$ and tensor of elastic strains

$$\varepsilon_{ij} = \varepsilon_{ij}^{(e)} \qquad (2.3)$$

stating that all deformations in the body are pure elastic, and kinematical relations between total strains and displacement veetor w_i

$$\varepsilon_{ij} = \left(\frac{\partial w_i}{\partial x_j} + \frac{\partial w_i}{\partial x_j} \right) \qquad (2.4)$$

If some nonelastic deformation occurs, equation (2.3) is no longer valid. It should be substituted by the equation

$$\varepsilon_{ij} = \varepsilon_{ij}^{(e)} + \varepsilon_{ij}^{(p)} \qquad (2.5)$$

where $\varepsilon_{ij}^{(p)}$ is tensor of plastic strains. Since six new unknown characterisitcs $\varepsilon_{ij}^{(p)}$ appear one needs six additional equations.

In theory of plasticity and theory of creep these six additional equations usually have the form

$$\dot{\varepsilon}_{ij}^{(p)} = f_{ij}\left(\varepsilon_{km}^{(p)}, \sigma_{km}\right) \qquad (2.6)$$

where dot denotes time derivative.

In diffusional creep, kinematics is quite special (it will be discussed in more details below). In diffusional creep, tensor of plastic deformation is compatible, i.e. it can be expressed in terms of some "plastic displacements" $w_i^{(p)}$

$$\varepsilon_{ij}^{(p)} = \left(\frac{\partial w_i}{\partial x_j} + \frac{\partial w_j^{(p)}}{\partial x_j} \right) \qquad (2.7)$$

So, one needs in three additional equations for $w_i^{(p)}$. Kinematical reasonings lead to the relation between "plastic velocity" $\dot{w}_i^{(p)}$ and flux of vacancies

$$\dot{w}_i^{(p)} = D\frac{\partial c}{\partial x_i} \qquad (2.8)$$

where c is volume fraction of vacancies and D is diffusion coefficient. So, the number of additional unknowns is reduced now to one: concentration of vacancies c. To close the system of equations one needs an equation for c. In the simplest situation this is just classical diffusion equaion

$$\frac{\partial c}{\partial t} = D\Delta c \qquad (2.9)$$

where Δ is Laplace's operator. Equations (2.1), (2.2), (2.4), (2.5), (2.7), (2.8) and (2.9) form a closed system of equations of diffusional creep.

3. Kinematics

Consider a monocrystal with a perfect lattice. Let it occupy some region V at zero stresses and some temperature T_0. If external load is applied and temperature is elevated then the crystal is deformed and occupy some region V. Region V, the actual state of the crystals depends on time because, as we assume the crystal creeps. If the crystal is unloaded, it occupies a region V^*. Region V^* does not coincide with V_0 becasue the crystal gained some plastic deformation. Region V^* can be defined at each instant by the thought process of unloading. It is assumed that in the unloading

process temperature is also returned to the initial value T_0. Region V^* depends on time t.

Three states, initial, V_0, unloaded, $V^*(t)$, and actual, V(t), can be defined for any deformation of the crystal. Deformation cuased by the diffusion of vacancies has some specifics. In discussion of these specifics we assume that vacancies are not created inside the crystal and can come into the crystal only from its boundary. In reality, it is possilbe the bulk nucleation of vacancies by simultaneous production of a vacancy and an intersticial atom or by dislocation climb, but these processes are not considered here. They can be taken into account by a complication of the presented theory.

Consider a bulk flux of vacancies (Fig. 3a), which goes from one piece of boundary to anthter one. It corresponds to the flux of matter in opposite direction. Therefore, after some time one observes a deformed state shown on Fig. 3b. If the flux keeps going one would see after some time the deformed state of Fig. 3c. Fig. 3 suggests that the transition from initial state V_0 to an unloaded state V^* can be described by some displacement vector field: each point of initial state of the material moves to some new position and there is one-to-one correspondence between points of V_0 and V^*.

Quite different situation we have for boundary flux of vacancies (or, that is the same, boundary flux of matter). In this case material from one piece of the boundary moves to another one along the boundary. The motion occurs gradually: first the upper layer of V_0 moves to the right side of the body, then the second one, and so on (see Fig. 4). the major differences from the bulk diffusion is that material is highly mixed and particles which were close at the initial state may be far away from each other in the deformed state. If description of the bulk diffusion we may keep the notion of Lagrangian particle, for boundary diffusion it seems impossible. So, we have to construct a theory which does not use the major assumption of continuum

mechanics: existence of material (Lagrangian) particles. To this end we consider the following kinematical scheme for a monocrystal.

Let all three region, V_0, V^* and V are referred to some Cartesian coordinates x_i Coordinates of points in V_0, V^* and are denoted by $\overset{\circ}{x}_i$, y_i and x_i correspondingly. If boundary diffusion does not occur, then plastic deformation can be described as one-to-one map $V_0 \leftrightarrow V^*$. It is characterized by the functions

$$y_i = y_i(\overset{\circ}{x}_k, t) \tag{3.1}$$

We have a usual continuum law of motion. We may introduce plastic velocity, the velocity of unloaded state,

$$v_i^{(p)} = \frac{\partial y_i(\overset{\circ}{x}_k, t)}{\partial t} \tag{3.2}$$

Plastic velocity may be considered as a function of coordinates y_i of the unloaded state due to the mapping (3.1).

$$v_i^{(p)} = v_i^{(p)}(y_k, t) \tag{3.3}$$

Let now the boundary diffusion takes place. That means that the region V^* is deformed not only due to the bulk motion (3.1) but also by the mass transfer at the vicinity of boundary. Such notion like velocity $V_i^{(p)}$ of a "particle" $\overset{\circ}{x}_k$, defined by (3.2), no longer exists in the entire region V_0 because some particles disappear in the course of deformation. However, at each moment one can define velocity $v_i^{(p)}(y_k, t)$ (3.3) in the region $V^*(t)$. This velocity becomes a primary kinematical characteristics instead of the law of motion (3.1) of classics continuum mechanics.

To describe the evolution of the region $V^*(t)$ due to the boundary diffusion we introduced the velocity, u, which is the rate of migration of boundary in normal direction. Note, that $u \neq 0$ for a rigid motion of the region V*. Therefore, to bind u with the physical process of the boundary diffusion, we put the constraints on plastic velocity $v_i^{(p)}$, eliminating rigid motion:

$$< v_i^{(p)} > \equiv \frac{1}{V^*(t)} \int_{V^*(t)} v_i^{(p)}(y_k, t) d^3y = 0 \tag{3.4}$$

$$< \frac{\partial v_i^{(p)}}{\partial y_i} - \frac{\partial v_j^{(p)}}{\partial y_i} > \equiv \frac{1}{V^*} \int_{V^*(t)} \left[\frac{\partial v_i^{(p)}(y_k,t)}{\partial y_j} - \frac{\partial v_{ji}^{(p)}(y_k,t)}{\partial y_j} \right] d^3y = 0 \qquad (3.5)$$

In the process of boundary diffusion boundary velocity u is not arbitrary. It should obey the law of conservation of mass

$$\rho_o(1-c)u = \nabla_\alpha J^\alpha \qquad (3.6)$$

where ρ_o is the mass density of the perfect crystal, c is volume vacancy concentration, and J^α are the components of the surface mass flux. Although in applications c<<1, we keep c in all relations in order to underline the physical origination of various terms.

It is assumed that boundary velocity u is the velocity of the boundary points when no bulk diffusion occurs. Therefore, the total normal velocity of the boundary poinst is:

$$u_{tot} = v_i^{(p)} n^i + u \qquad (3.7)$$

Here n_i are the components of the unit normal vector on the boundary ∂V^* of V* directed outward of the region V*. In accordance with (3.7), u is positive at some point A on ∂V^* if material arrives at A and negative in the opposite case.

Let us establish now the kinematical realtions between plastic velocity and the flux of vacancies. We assume that he crystal is the "mixture" of two substances: atoms and vacancies. Each one has its own velocity. Velocity of matter (atoms) is plastic velocity $v_i^{(p)}$, velocity of vacancies is denoted by u_i. One might consider a piece of crystal lattice, a "representative volume of material," and think of $v_i^{(p)}$ as average velocity of all atoms of this piece

$$v_i^{(p)} = \frac{1}{N_a} \sum_\alpha v_i^\alpha \qquad (3.8)$$

Where N_a is the number of atoms, v_i^α is the velocity of α-th atom and summation is taken over all of atoms of the piece. Similarly,

$$u_i = \frac{1}{N_v} \sum_\alpha u_i^\alpha \qquad (3.9)$$

where N_a is the number of vacancies and u_i^α is the velocity of α-th vacancy.

Volume average velocity is, by definition,

$$v_i = \frac{1}{N}\left(\sum_\alpha v_i^\alpha + \sum_\alpha u_i^\alpha\right) \tag{3.10}$$

where N is the toal number of lattice sites

$$N = N_a + N_v \tag{3.11}$$

It follows from (3.8) - (3.11) that

$$v_i = (1-c)v_i^{(p)} + cu_i \tag{3.12}$$

where the volume fraction of vacancies is

$$c = \frac{N_v}{N} \tag{3.13}$$

Relation (3.12) holds for mixture of any two substances. Now we have to express in some way that we deal with diffusion of vacancies. We may assume that in the process of the position exchange of an atom and a vacancy the velocities of the atom and the vacancy are equal in the magnitude and opposite in the sign. Therefore, $v_i = 0$ and (3.12) links the velocities of atoms and vacancies. It is clear that it is not necessary to put v_i=0; one might add the motion of the considered piece of material as a rigid body. As it is known [25] this means that the corresponding strain rate tensor is equal to zero

$$\frac{\partial v_i(y_k,t)}{\partial y_j} + \frac{\partial v_j(y_k,t)}{\partial y_i} = 0 \tag{3.14}$$

for each y_i and t. The general solution of (3.14) is the velocity field of rigid motion

$$v_i = a_i + b_{ij}y_j \tag{3.15}$$

where a_i and b_{ij} are some constants, and $b_{ij} = -b_{ij}$.

The relation (3.12) can be rewritten in the form

$$V_i^{(p)} = v_i - \frac{1}{1-c}J_i$$

$$u_i = v_i + \frac{1}{c}J_i \tag{3.16}$$

where J_i is an arbitrary vector field. Finally, (3.15) and (3.16) yield

$$v_i^{(p)} = a_i + b_{ij} y_j - \frac{1}{1-c} J_i$$

$$u_i = a_i + b_{ij} y_j + \frac{1}{1-c} J \qquad (3.17)$$

The constants a_i and b_{ij} can be found in terms of vector J_i by means of the constraints (3.4), (3.5). These constraints can be rewritten as

$$a_i + b_{ij} < y > = \frac{1}{V^*} \int_{V^*} \frac{1}{1-c} J_i d^3 y$$

$$b_{ij} = \frac{1}{2V^*} \int_{V^*} \left[\frac{\partial}{\partial x_j} \frac{1}{1-c} J_i - \frac{\partial}{\partial x_{ij}} \frac{1}{1-c} J_i \right] d^3 y \qquad (3.18)$$

There is another way to fix rigid motion of the unload state. One might put $v_i = 0$ in (3.16). Then

$$v_i^{(p)} = \frac{-1}{1-c} J_i$$

$$u_i = \frac{1}{c} J_i \qquad (3.18')$$

In this case region V* will move in space. The motion of V* is determined by the diffusion flux J_i. In particular, the translational velocity of this motion is equal to

$$\frac{-1}{V^*} \int_{V^*} \frac{1}{1-c} J_i d^3 y$$

We choose the second option, because it simplifies the following relations.

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

$$\frac{\partial c}{\partial t} + \frac{\partial}{\partial y_k} c u_k = 0 \qquad (3.19)$$

Similarly, for the flux of matter

$$\frac{\partial}{\partial t}(1-c) + \frac{\partial}{\partial y_k}(1-c) v_k^{(p)} = 0 \qquad (3.20)$$

Equation (3.20) is a consequence of (3.19) because the sum of (3.19) and (3.20) is the identity due to (3.12) and (3.15). In accordance with (3.17), equation (3.19) can be rewritten as

$$\frac{\partial c}{\partial t} + \frac{\partial J_k}{\partial y_k} = 0 \qquad (3.21)$$

It is seen that J_k has the sense of the diffusion flux of vacancies.

The deformed state is obtained as a result of elastic displacement from unloaded state V* to the actual state V(t). We have

$$x_i = y_i + w_i \tag{3.22}$$

It is natural to consider the solution of all problems in the coordinates of the actual state x_i. Therefore, w_i in (3.22) are assumed to be some functions of x_i and t. From (3.22) we obtain the relation between coordinates of the unloaded state y_i and the actual state x_i:

$$y_i(x_k, t) = x_i - w_i(x_k, t) \tag{3.23}$$

By assumption, functions (3.23) determine a one-to-one correspondence between V*(t) and V(t).

Kinematical relations have been written above in terms of y-coordinatees. We need to have one of them, the diffusion equation for vacancies (3.15) also in terms of x-coordinates. The transformation is based on the identity [25]

$$\frac{\partial}{\partial y^k}\left(\frac{\partial y^k}{\partial x^i}\det\left\|\frac{\partial x_i}{\partial y_m}\right\|\right) = 0 \tag{3.23}$$

Note that

$$\det\left\|\frac{\partial x_i}{\partial y_m}\right\| = \frac{1}{\Delta}$$

where

$$\Delta = \det\left\|\frac{\partial y_m}{\partial x_j}\right\| = \det\left\|\delta_l^m - \frac{\partial w^m}{\partial x_j}\right\| \tag{3.24}$$

It is assumed that $\Delta \neq 0$. Derivatives $\frac{\partial x^j}{\partial y^k}$ will be denoted also by s_k^j. Matrix $\|s_k^j\|$ is the inverse matrix to the matrix $\left\|\frac{\partial y^k}{\partial x^j}\right\|$ because

$$\frac{\partial x^j}{\partial y^k}\frac{\partial y^k}{\partial x^i} = \delta_m^j \Leftrightarrow s_k^j\left(\delta_m^k - \frac{\partial w^k}{\partial x^m}\right) = \delta_m^j \tag{3.25}$$

Thus s^j_k are the certain functions of the displacement gradient which can be found from (3.25). For small displacement gradients in the first approximation $s^j_k = \delta^j_k$, in the second approximation

$$s^j_k = \delta^j_k + \frac{\partial w^j}{\partial x^k} \tag{3.26}$$

Lagrangian elastic velocity of particles caused by elastic deformation is defined in terms of displacement vector by the relations

$$\frac{\partial w^j}{\partial t} + v^k_{(e)} \frac{\partial w^j}{\partial x^k} = v^i_{(e)} \tag{3.27}$$

Equations (3.27) can be considered as the system of linear equations with respect of $v^i_{(e)}$. Solution of this system has the form

$$v^i_{(e)} = s^i_k \frac{\partial w^i(x_m, t)}{\partial t} \tag{3.28}$$

Velocity $v^i_{(e)}$ is well-defined in the internal points on region V(t) but at the boundary the expression (3.28) it should be rectified because partial derivatives at the boundary points do not make sense since if point x belongs to the boundary at moment t, it might not be in V at moment $t + \Delta t$. We assume that for $x \to \partial V$ the derivatives in (3.28) are understood as limit values on ∂V derivatives found inside V.

From (3.23) we have that for each vector J_k

$$\frac{\partial J^k}{\partial y^k} = \frac{\partial}{\partial y^k} \frac{\partial y^k}{\partial x^i} \mathcal{J}^i = \frac{1}{\Delta} \frac{\partial y^k}{\partial x^i} \frac{\partial}{\partial y^k} \Delta \mathcal{J}^i = \frac{1}{\Delta} \frac{\partial}{\partial x^i} \Delta \mathcal{J}^i \tag{3.29}$$

where

$$J^k = \frac{\partial y^k}{\partial x^i} \mathcal{J}^i$$

or

$$\mathcal{J}^i = s^i_k J^k \tag{3.30}$$

Finally, we have for the diffusion equation of vacancies

$$\frac{\partial c}{\partial t} + \frac{1}{\Delta} \frac{\partial}{\partial x^i} [\Delta J^i] = 0 \tag{3.31}$$

Now we proceed to dynamics.

4. Free Energy Rate of Monocrystal

We assume that temperature T is kept constant. Therefore, the relevant thermodynamical potential is free energy of the crystal F. We assume that free energy has a volume density F per unit mass of the perfect lattice.

$$F = \int_{V(t)} \rho_0 F d^3x \tag{4.1}$$

Energy density F is supposed to be a function of the gradient of elastic displacements, vacancy concentration and temperature

$$F = F(w_{i,j}, c, T), \qquad w_{i,j} \equiv \frac{\partial w_i}{\partial x^j} \tag{4.2}$$

Let us find time derivative of free energy. We have

$$\frac{dF}{dt} = \int_V \left(\rho_0 \frac{\partial F}{\partial w^i{}_{,t}} \frac{\partial}{\partial x_j} w^i{}_{,t} + \rho_0 \frac{\partial F}{\partial c} \frac{\partial c}{\partial t} \right) d^3x + \int_{\partial V} \rho_0 F \left(v_i^{(e)} n_i + v_m^{(p)} s^i{}_m n^i + u \right) d^3x \tag{4.3}$$

Here it is implied that the velocity of the boundary ∂V of the region V is the sum of (3.7) and velocity caused by elastic motion. After plugging in (4.3) the expression for $\frac{\partial c}{\partial t}$ from (3.7) and integration by part we obtain

$$\frac{dF}{dt} = \int \left[\left(-\frac{\partial}{\partial x_j} \rho_0 \frac{\partial F}{\partial w^i{}_{,j}} \right) \frac{\partial w^i}{\partial t} + \Delta J^i \frac{\partial}{\partial x^i} \frac{\rho_0}{\Delta} \frac{\partial F}{\partial c} \right] d^3x$$

$$+ \int \left\{ \rho_0 \frac{\partial F}{\partial w^i{}_{,j}} n_j \frac{\partial w^i}{\partial t} - \mathfrak{J}^i n_i \rho_0 \frac{\partial F}{\partial c} + \rho_0 F \left(v_i^{(e)} n_i + v_m^{(p)} s^i{}_m n^i + u \right) \right\} d^2x \tag{4.4}$$

To consider the constraints put by thermodynamics we have also to describe the power of external forces dA/dt. We assume that external forces act out on the boundary ∂V of body V and have the surface density f^i. We accept allso the assumption that noraml external surface forces works on the total displacement while tangent surface forces work only on elastic displacement.

$$\frac{dA}{dt} = \int_{\partial V} \left\{ f^i v_i^{(e)} + f^i n_i (s_k^m v_m^{(p)} n^k + u) \right\} d^2x \tag{4.5}$$

The factors s_k^m relates to the fact that plastic velocity is taken in y-space while normal n_k is in x-space.

It is known from statistical mechanics that

$$\frac{dF}{dt} - \frac{dA}{dt} \leq 0 \qquad (4.6)$$

Inequality (4.6) applied to the expressions for $\frac{dF}{dt}$ and $\frac{dA}{dt}$ yields the equilibrium equation

$$\frac{\partial}{\partial x_j} \rho_0 \frac{\partial F}{\partial w^i{}_{,j}} = 0 \qquad (4.7)$$

(otherwise, w^i, t can be chosen in such a way that (4.6) is violated).

To comply with (4.6) the diffusion flux can be chosen as

$$J^i = - D^{ij} \Delta \frac{\partial}{\partial x^i} \frac{\rho_0}{\Delta} \frac{\partial F}{\partial c} \qquad (4.8)$$

where D^{ij} is a positive tensor of diffusivity.

The surface terms in the inequality (4.6) can be written in the form

$$\text{Surface terms} = \int_{\partial V} \left\{ \begin{array}{l} \rho_0 \frac{\partial F}{\partial w^i{}_{,j}} \left(\delta^i_k - w^i{}_{,k} \right) n_j v^{(e)k} - n_i \rho_0 \frac{\partial F}{\partial c} + \rho_0 F \left(v^{(e)}_i n_i + s^m_j v^{(p)}_m i + u \right) - \\ f^i \left(v^{(e)}_i + v^{(p)}_m s^i_m \right) - f^i n_i u \end{array} \right\} d^2x =$$

$$\int \left\{ \begin{array}{l} \rho_0 \frac{\partial F}{\partial w^i{}_{,j}} \left(\delta^i_k - w^i{}_{,k} \right) n_j - f_k r^k_d v^{(e)\infty} + \left(\rho_0 \frac{\partial F}{\partial w^i{}_{,j}} \left(\delta^i_k - w^i{}_{,k} \right) n_j n_k + \rho_0 F - f^i n_i \right) \\ \left(v^{(e)}_k n^k + s^m_k v^{(p)}_m n^{\alpha k} + u \right) - \rho_0 \frac{\partial F}{\partial w^i{}_{,j}} \left(\delta^i_k - w^i{}_{,k} \right) n_j n_k \left(s^m_j v^{(p)}_m n^i + u \right) - \mathcal{J}^i n_i \frac{\partial F}{\partial c} \end{array} \right\} d^2x \qquad (4.9)$$

To warrant the inequality (4.6) one needs to put

$$\sigma^i_j n_j = f_i \qquad (4.10)$$

where σ^j_i is given by the constitutive equation

$$\sigma^j_k = \rho_0 \frac{\partial F}{\partial w^i{}_{,j}} \left(\delta^i_k - w^i{}_{,k} \right) + \rho_0 F \delta^i_k \qquad (4.11)$$

It is seen from (4.10) that s^i_j play the role of the component of stress tensor. It can be shown that equilibrium equations (4.7) are equivalent to the equations [25].

$$\frac{\partial \sigma_i^j}{\partial x^j} = 0 \qquad (4.12)$$

So, using (3.18') and (3.6) we obtain

$$\text{Surface terms} = \int_{\partial V} \left\{ \left(\frac{\sigma_{nn} - \rho_0 F}{1-c} - \rho_0 \frac{\partial F}{\partial c} \right) J^i n_i - \frac{\sigma_{nn} - \rho_0 F}{1-c} \nabla_\alpha J^\alpha \right\} d^2x \qquad (4.13)$$

After integration by parts the second term the surface integral takes the form

$$\text{Surface terms} = \int_{\partial V} \left\{ \left(\frac{\sigma_{nn} - \rho_0 F}{1-c} - \rho_0 \frac{\partial F}{\partial c} \right) J^i n_i + \nabla_\alpha J^\alpha \left(\frac{\sigma_{nn} - \rho_0 F}{1-c} \right) \right\} d^2x \qquad (4.14)$$

To provide the negativeness of the surface integral we may accept the following boundary conditions

$$\frac{\sigma_{nn} - \rho_0 F}{1-c} - \rho_0 \frac{\partial F}{\partial c} = -\lambda J^i n_i - \lambda_\alpha J^i n_i \qquad (4.15)$$

where $\lambda \geq 0, \lambda_{\alpha\beta}$ is a positive tensor, and λ_α obey the inequalities following from the positiveness of the quadratic form

$$\lambda x^2 + 2\lambda_\alpha x x^\alpha + \lambda_{\alpha\beta} x^\alpha x^\beta$$

for all x, x_α.

Equations (3.6), (3.18'), (3.30), (3.31), (4.10), (4.11), (4.12), (4.15) form a closed system of equations determining the evolution of the stress state and plastic deformation in case of diffusional creep of a loaded monocrystal.

5. Linearization

Usually, elastic deformation $\varepsilon_{ij}^{(e)}$ and vacancy concetraion c are of order 10^{-4} and can be neglected compared to the unity, while for free energy density one can use the quadratic expression

$$\rho_0 F = \frac{1}{2} A^{ijkl} \varepsilon_{ij}^{(e)} \varepsilon_{kl}^{(e)} + \frac{1}{2} A (c - c_o)^2 + \text{function of } T \qquad (5.1)$$

where A^{ijkl} are the Young moduli and c_o is the equilibrium value of vacancy concentration. The material constant A can be found from elementary statistical considerations [1]

$$A = \frac{\rho_0 T}{m c_o} \tag{5.2}$$

where m is the mass of one atom of the crystal.

Kinematical relations (3.6) and (3.18) take the form

$$\rho_0 u = \nabla_\alpha J^\alpha$$
$$v_i^{(p)} = -J_i \tag{5.3}$$

Matrix $\|s^i_k\|$ may be taken equal to the unit matrix if rotation from loaded state to unloaded state is small. We write the following equations under this assumption. Since $s^i_k = \delta^i_k$, we have $\mathfrak{J}_i = J_i$.

$$\frac{\partial c}{\partial t} + \frac{\partial J^i}{\partial x^i} = 0 \tag{5.4}$$

Besides

$$\frac{\partial \sigma^j_i}{\partial x^j} = 0, \qquad \sigma^{ij} = A^{ijkl}\varepsilon^{(e)}_{kl}, \qquad \varepsilon^{(e)}_{kl} = \frac{1}{2}\left(\frac{\partial w_k}{\partial x^l} + \frac{\partial w_l}{\partial x^k}\right) \tag{5.5}$$

In accordance with (4.8), diffusion flux is given by

$$J^i = -D^{ij} A \frac{\partial c}{\partial x^j} \tag{5.6}$$

In the boundary conditions (4.15) F may be neglected compared to σ_{nn}. Coefficient λD^{ij} has the dimension of length. The only parameter with this dimension in (almost) perfect lattice is the interatomic distance d. Since the characteristic length of diffusion process is supposed to be much larger then d, the term $lJ^i n_i$ in (4.15) can be neglected compared to $\rho_0 \frac{\partial F}{\partial c}$. Note that at the singular points like the points of high curvature of the boundary surface this term might be essential.

Coefficients λ_α describe the appearance of surface diffusion due to bulk diffusion and the inversed effect. In first approximation this effect, probably, can be neglected. We obtain

$$\sigma_{nn} = A(c - c_o)$$
$$J^\alpha = -D^{\alpha\beta}\nabla_\beta \sigma_{nn} \tag{5.7}$$

where $D^{\alpha\beta}$ is the tensor inversed to $\lambda_{\alpha\beta}$.

If $D^{\alpha\beta}=0$ and $D^{ij} = D\delta^{ij}$ we come to the system of equaitons which has been written down and studied for the first time by I. Lifshitz [5].

6. Diffusional Creep in Polycrystals.

In case of polycrystals the differential equations remain the same, and one needs to establish the boundary conditions on the grain boundaries.

We assume that the surface forces are continuous on the grain boundaries

$$\left[\sigma_i^j\right] n_j = 0 \qquad (6.1)$$

Here and in the following [A] means the difference of A on two sides of the surface. Denoting the values of A on each side by indices + and - correspondingly, one can write

$$[A] = A_+ - A_-$$

The normal vector n_i is directed, by condition, from the side + to the side -.

We assume that the total normal velocity on both sides of the grain boundary conicide

$$\left[v_i^{(e)} + v_i^{(p)}\right] n_i + [u] = 0 \qquad (6.2)$$

As to tangent velocity, it can be sliding along grain boundaries, and tangent velocities may have a jump.

In accordance with (4.9) surface terms on the grain boundary Σ have the form

Surface terms =

$$\int \left\{ \begin{array}{l} \left(\sigma_i^j - \rho_0 F\delta_i^j\right)_+ n_j v_+^{(e)i} - \left(\sigma_i^j - \rho_0 F\delta_i^j\right)_- N_j v^{(e)i} - \left(J^i{}_+ \rho_{0+} \dfrac{\partial F}{\partial c}\Big|_+ - J^i \rho_{0-} \dfrac{\partial F}{\partial c}\Big|_-\right) \\ n^i + [\rho_0 F]\left(v_i^{(e)} n^i + v_m^{(p)} s_i^m n^i + u\right) \end{array} \right\} d^2x \qquad (6.3)$$

This expression can be rewritten in the form

Surface terms =

$$\int_{\Sigma}\left\{\begin{array}{l}\left(\sigma_i^j n_j\left[v^{(e)\alpha}\right]+\sigma_{nn}\left[v_n^{(e)}\right]-\left[\rho_0 F v_n^{(e)}\right]-\left[J\rho_0\frac{\partial F}{\partial c}\right]\right)\\ +\left[\rho_0 F\left(v_n^{(e)}+v_n^{(p)}+u\right)\right]\end{array}\right\}d^2x =$$

$$=\int_{\Sigma}\left\{\sigma_\alpha^j n_j\left[v^{(e)\alpha}\right]-\sigma_{nn}\left[v_n^{(p)}+u\right]-\left[J_n\rho_0\frac{\partial F}{\partial c}\right]-\left[\rho_0 F\left(v_n^{(p)}+u\right)\right]\right\}d^2x \qquad (6.4)$$

Different boundary conditions can be consistent with the negativeness of (6.4). For the law of grain sliding one may assume that

$$\sigma_\alpha^j n_j = -\mu\left[V_\alpha^{(e)}\right] \qquad (6.5)$$

Boundary conditions for the bulk diffusion depends significantly on the properties of the boundary. If diffusion occurs "indepently" on each side of the boundary, one may put

$$\frac{\sigma_{nn}-\rho_0 F}{1-c}-\rho_0\frac{\partial F}{\partial c}=0 \qquad (6.6)$$

$$J^\alpha = -D^{\alpha\beta}\nabla_\beta\left(\frac{\sigma_{nn}-\rho_0 F}{1-c}\right) \qquad (6.7)$$

If vacancy flux J_n is continuous on the grain boundary then (6.7) should be replaced by one condition

$$\left[\frac{\sigma_{nn}-\rho_0 F}{1-c}-\rho_0\frac{\partial F}{\partial c}\right]=0 \qquad (6.8)$$

7. Future Developments

The constructed equations seem describe adequately the diffusional creep in polycrystals. Using these equations one may attack such problems like study of superplastic deformation, void formation, constitutive equations of primary and secondary creep, etc. These problems are supposed to be considered during the Summer Extension Program.

References

1. J.-P. Poirier, Creep of Crystals, Cambridge Univ. Press, 1985.
2. Natarro,F.R.N., Deformation of Crystals by the motion of single ions, Report of a Conference on strength of Solids, The Physical Soc., 75-90, 1948.
3. Herring C. Diffusional Viscosity of a Polycrystalline Solid, J. Appe. Phys, 21, 437-45, 1950.
4. Coble, R.L., A model for boundary-diffusion controlled creep in polycrystalline materials, J. Appl. Phys, 34, 1679-82, 1963.
5. Lifshitz, I.M., On the theory of diffusion-viscous flow of polycrystalline bodies, Soviet Physics JETP, 17, 909-20, 1963.
6. Yu. Rabotnov, Creep Problems in Structural Members, John Wiley & Sons, 1969.
7. C.P. Flynn, Point Defects and Diffusion, Clarendon Press, 1972.
8. F.A. Gulo-Lopez, C.R.A. Catlow, P.D. Townsend, Point Defects in Materials, Academic Press, 1988.
9. S. Shesterikov, A. Lokotshenko, Creep and Long Term Strength of Materials, Itogi Nauki, V. 13, 1980.
10. J. H. Crawford, L. M. Slifkin, Points Defects in Solids, Plenum Press, 1975.
11. Theory of Crystal Defects, Proe. Summer School in Hrarany, 1964, Ed. B. Gruber, Academic Press 1966.
12. Ashby, M.F. & Verral, R.A. Diffusion accomodated Flow and Superplaticity, Acta Metall., 21, 149-63, 1973.
13. Diffusion in Crystalline Solids, Ed. G.E. Murch and A.S. Nowick, Academic Press, 1984.
14. J.W. Christian, The Theory of Transformations in Metals and Alloys, Perganom Press, 1975.
15. P.A. Varotsos, K.D. Alexopoulos, Thermodyanmics of Point Defects and Their Relation with Bulk Properties, North-Holland, 1986.
16. J.Friedel, Dislocations, Perganom Press, 1964.
17. A.H. Cottrell, Dislocations and Plastic flow in Crystals, Oxford, 1958.
18. L.A. Girifalco, Statistical Physics of Materials, J.Wiley & Sons, 1973.
19. C. Kittel, Introduction to Solid State Physics, J.Wiley 7 Sons, 1976.
20. H.E. Evans, Mechanisms of Creep Fracture, Elsevier Applied Sciences, 1984.
21. J.W. Martin, R.D. Doherty, Stability of Micorstructure in Metallic Systems, Cambridge Univ. Press, 1976.
22. Structure and Properties of Solid Surfaces, Ed. R. Gomer and C. S. Smith, Univ. of Chicago Press, 1953.
23. Yu. Rabotnov, Mechanics of Deformable Solids.
24. J. Gittus, Creep, Viscoelasticity and Creep Fracture in Solids, J. Wiley & Sons, 1975.
25. V. Berdichevsky, Variational Principles of Continuum Mechanics, Nauka, Moscow, 1983.

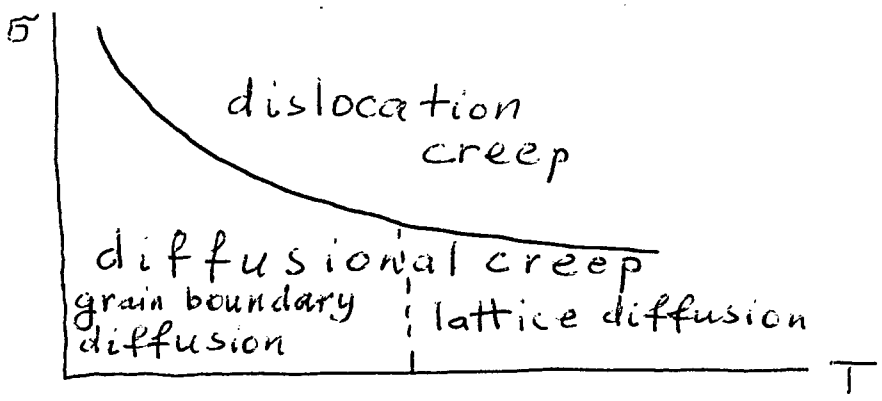


Fig. 1

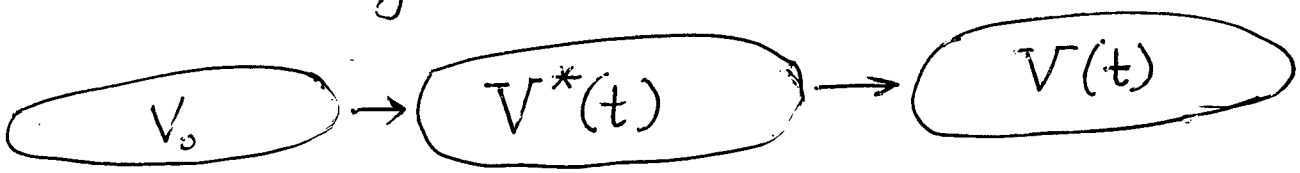
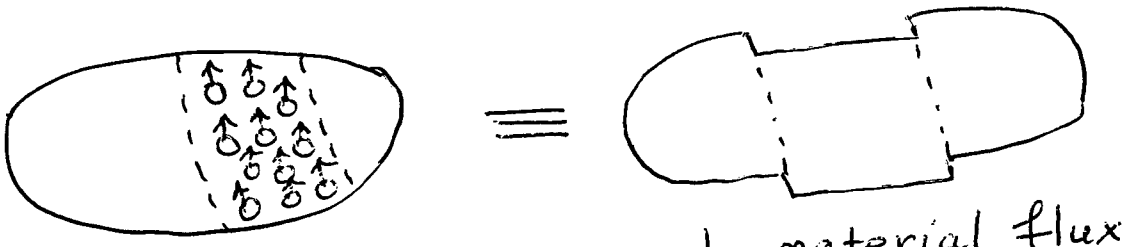


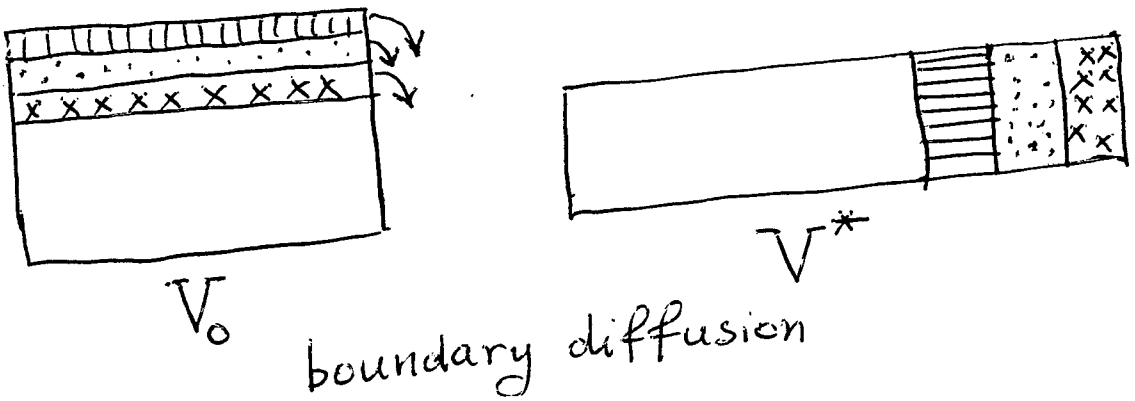
Fig. 2



a. vacancy flux

b. material flux

Fig. 3



boundary diffusion

Fig. 4

Quantitation of Dissolved O_2 in Aviation Fuels by Fluorescence Lifetime Quenching.

Steven W. Buckner
Assistant Professor
Department of Chemistry and Geology

Columbus College
Columbus, GA 31907

Final Report for:
Summer Faculty Research Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling AFB, Washington D.C.
and Wright Laboratory

September, 1994

Quantitation of Dissolved O_2 in Aviation Fuels by Fluorescence Lifetime Quenching.

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

Abstract

A new method for quantitation of dissolved molecular oxygen in aviation fuels is described. The approach is based on determination of the fluorescence lifetime of pyrene doped in the fuel at the ppm level. Oxygen quenches the pyrene fluorescence lifetime permitting the generation of linear calibration curves based on Stern-Volmer kinetics. The method is rapid, sensitive, less expensive than current methods, insensitive to thermal stressing of the fuel, and capable of on-line analysis and spatial profiling of oxygen concentration in fuel lines. Application to flowing fuel oxygen consumption tests demonstrates the technique.

Quantitation of Dissolved O_2 in Aviation Fuels by Fluorescence Lifetime Quenching.

Steven W. Buckner

Introduction

Advanced aircraft use on-board fuel as a coolant. This induces reaction between the fuel and dissolved oxygen. Oxidation leads to formation of insoluble products and deposits within the aircraft fuel system. In order to develop rational solutions to the problem of deposit formation, an understanding of the oxidation of these fuels under conditions of high temperature and low oxygen concentration is necessary[1]. One of the difficulties in this area is the determination of oxygen concentration in the aviation fuel. The concentration of oxygen in air saturated fuels (the starting point for the oxidation reactions) is on the order of 70 parts per million[2]. Thus, limit of detection is an issue. The techniques of choice currently employ gas chromatography (GC), with gas chromatography/mass spectrometry (GC/MS) often used due to its combination of selectivity and sensitivity[2,3]. However, there are drawbacks to this approach. First, the oxygen must be separated from the fuel prior to its introduction to the column. Any fuel reaching the column destroys its efficiency. Second, GC is slow and does not allow study of rapidly time-varying signals. Third, GC is necessarily performed off-line which prevents in-situ and spatially resolved experiments. Finally, GC methods, and GC/MS in particular, are relatively high cost techniques. It would be most desirable to supplement the GC technique with an optical spectroscopic probe of molecular oxygen[1].

Spectroscopically, molecular oxygen is difficult to study in solution[4]. It does not absorb in the infrared and its electronic transitions are far in the UV where organic solutions absorb strongly. O_2 has a Raman allowed transition, but this is not typically the technique of choice for trace analysis. O_2 also has a unique ESR signature, but the expense and low sensitivity of this technique

prohibit its use for this application. An alternative optical approach is to use the oxygen concentration dependence of the fluorescence of pyrene.

Oxygen efficiently quenches the fluorescence of pyrene due to the energy match between the singlet-triplet gap in O_2 and the energy of the first excited singlet state of pyrene[5]. Variations in oxygen concentration result in variations in the oxygen-pyrene collision frequency which change the total fluorescence quantum yield and the lifetime of the excited state. Thus, both the quantum yield and the lifetime exhibit an inverse oxygen concentration dependence. Instrumentally, it is a simpler task to measure the total fluorescence intensity, which may be converted to a quantum yield. Previous work has shown this approach to oxygen determination to be intractable in fuels. The amount of pyrene added to the solution must be precisely controlled. Also, if pyrene is consumed (or produced) during the oxidation of the fuel, the total fluorescence intensity will show variations which are not related to oxygen concentration. However, the lifetime of the excited state of pyrene (within certain limits) is independent of the amount of pyrene present, circumventing the above problems. Here we present a summary of the application of pyrene fluorescence lifetime quenching to the determination of oxygen concentration in aviation fuel.

Experimental

All fluorescence measurements were made on a home-built fluorimeter described below. A block diagram of the instrument is shown in Figure 1. Excitation of samples was accomplished using a N_2 laser (Laser Sciences VSL-337ND) with an output power of 300 μJ/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[6]. The output from the photodiode initiates data acquisition by a LeCroy 9354 digital storage oscilloscope. Typical fluorescence decays have time constants of 20 to 300 nsec. After the glass flat the pump beam is focussed into a sample cuvette. The fluorescence emission from the sample is collected and passed through an ND=3 neutral density filter due to

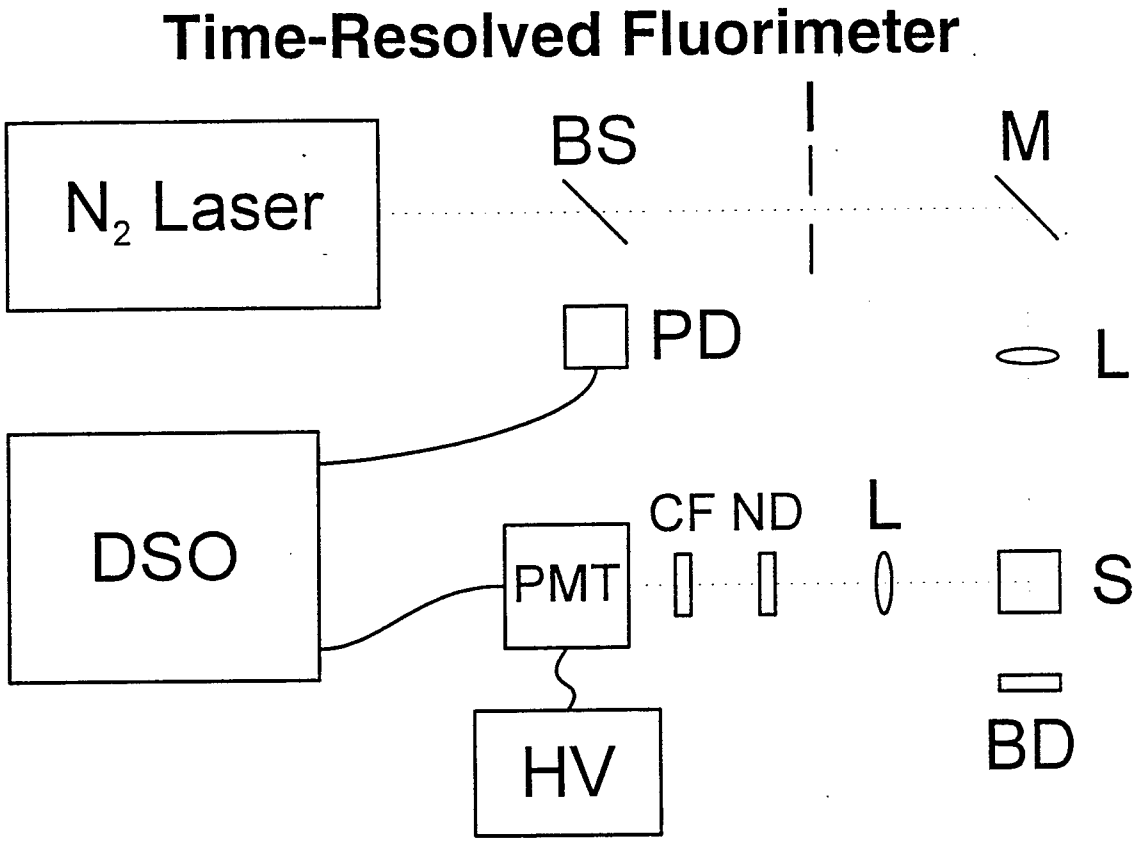

Time-Resolved Fluorimeter
N2 Laser
BS
M
PD
L
DSO
CF ND
L
PMT
S
BD
HV

Figure 1

the intense fluorescence from the pyrene. A glass filter is used to reject direct scatter of the laser beam. The fluorescent photons are directed onto an RCA 931A photomultiplier tube which has been fast-wired to yield a rise-time of ~1.3 nsec[7]. One hundred fluorescence decay curves were averaged for each point on the calibration curves shown in this work.

The fuels used in this work include POSF-2827, POSF-2980, and POSF-2926 (all Jet A aviation fuels) and 2818 (a JP-7 fuel). Isooctane and cyclohexane were obtained from Fisher Scientific and Mallinckrodt, respectively, and used without further purification. The pure hydrocarbons did not produce any detectable fluorescence. The fuels and hydrocarbons were doped with 10 to 23 ppm (w/w) pyrene, which was purchased from Kodak and also used without further purification. Fuels and hydrocarbons with varying amounts of dissolved oxygen were prepared by sparging the solution with mixtures of air and nitrogen. The oxygen concentrations in these solutions were determined using GC/MS (Hewlett-Packard Model 5988 with a Model 5980 GC) with selected ion monitoring of the m/z =32 ion. The determination of O_2 in fuel by GC/MS has previously been described in detail[3].

The on-line studies were performed with the near-isothermal flowing test rig (NIFTR) in which air-saturated fuel is pumped through a fuel line maintained at or near the Jet operating temperature and pressure (185 C for this study). The fuel lines in this study included both stainless steel and coated stainless steel. After the heated section, the fuel flows through one meter of stainless steel tubing in which it rapidly returns to room temperature. No further oxygen is consumed in this region. The oxygen concentration is then obtained by gas chromatography (Hewlitt Packard 5980) using a previously described technique[3]. For the fluorescence lifetime determination, the fuel then passes through a quartz flow cell with flat faces. The output from the nitrogen laser is focussed into this cell for excitation. The fluorescence intensity decay is monitored as described above in the static flow cell experiments.

Results

kinetic model

Before discussing the results it will be illustrative to consider the O_2 quenching kinetics. After photoexcitation of a molecule A (A = pyrene in this study), the excited state A^* may decay by fluorescence with a rate constant k_f or be quenched by oxygen with a rate constant $k_q[Q]$ (where Q = O_2 in this study). The overall first order rate constant k_s for reaction (1) is given in equation (2). The integrated rate expression for the decay of excited states of A is given in equation (3). A single excited state will show a simple exponential decay. From equation (2) it is clear that determination of the rate constant for the decay of the excited state at a series of oxygen concentrations should yield a straight line. In all of the work presented here we will use the lifetime (τ) for the decay of the excited state ($\tau = 1/k$). Equation (2) can be expressed using the lifetime of the unquenched excited state (τ_0) and the lifetime of the excited state at a concentration [Q] of quencher (τ_q) as:

$$A + \text{photon} \xrightarrow{} A^* \quad \begin{cases} \xrightarrow{(k_f)} A + \text{photon} \\ \xrightarrow{(k_q[Q])} A + Q^* \end{cases} \qquad (1)$$

$$k_s = k_f + k_q[Q] \qquad (2)$$

$$[A]_t = [A]_0 \exp(-k_s t) \qquad (3)$$

$$[Q]\ (\tau_0)/(\tau_q) + 1 = (\tau_0)/(\tau_s) \qquad (4)$$

where

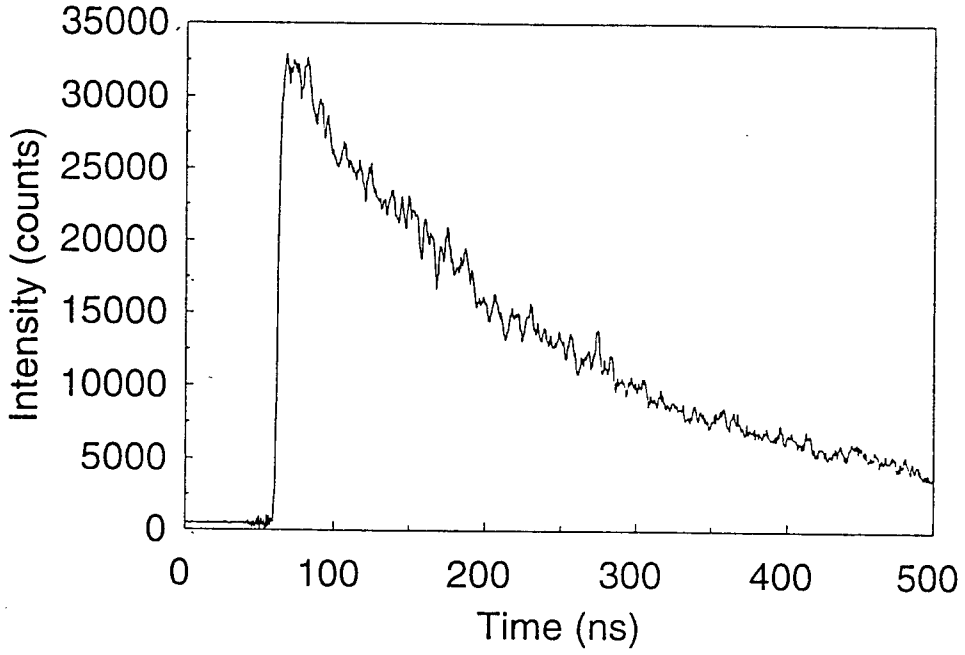
$$1/(\tau_s) = k_s \qquad (5)$$

Equation (4) is termed the Stern-Volmer equation[5]. All calibration curves will be presented as in equation (4) with $(\tau_0)/(\tau_s)$ plotted versus $[O_2]$.

model systems

Cyclohexane and isooctane were studied first as model systems. Figure 2(a) shows the decay of fluorescence intensity as a function of time for a 9 ppm solution of pyrene in cyclohexane after sparging with nitrogen. Figure 2(b) shows a linearized version in which the natural logarithm of the fluorescence intensity is plotted. The lifetime may be obtained by directly fitting the exponential decays; however, we found it most useful to linearize the data and

Pyrene Fluorescence Decay in Cyclohexane



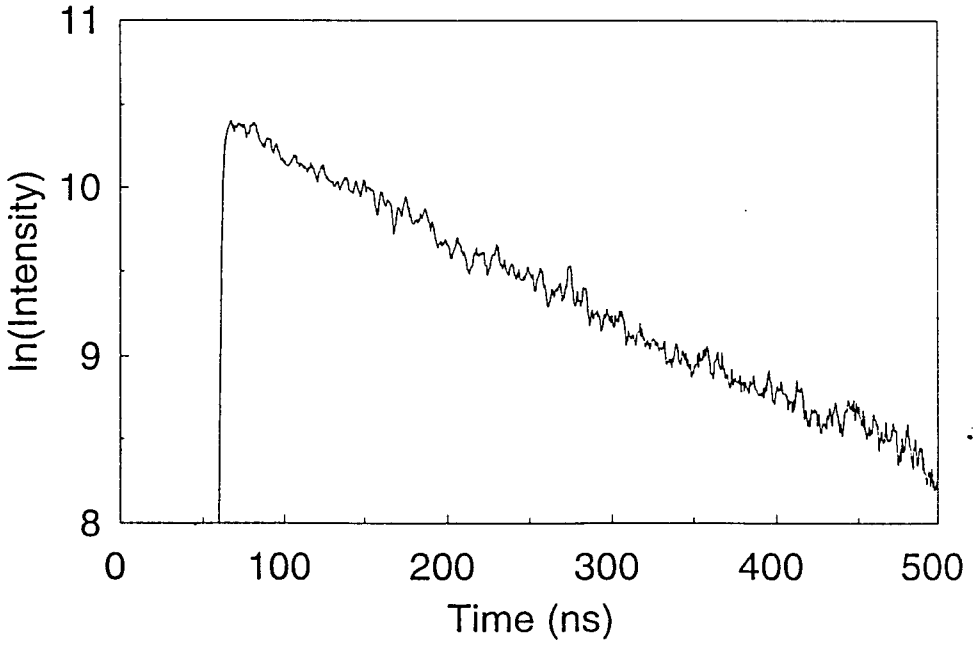
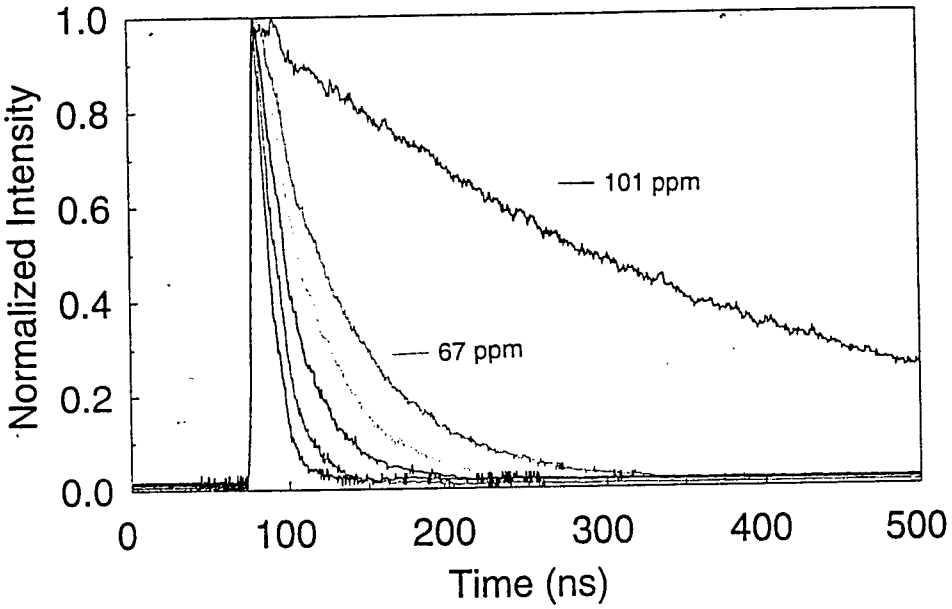
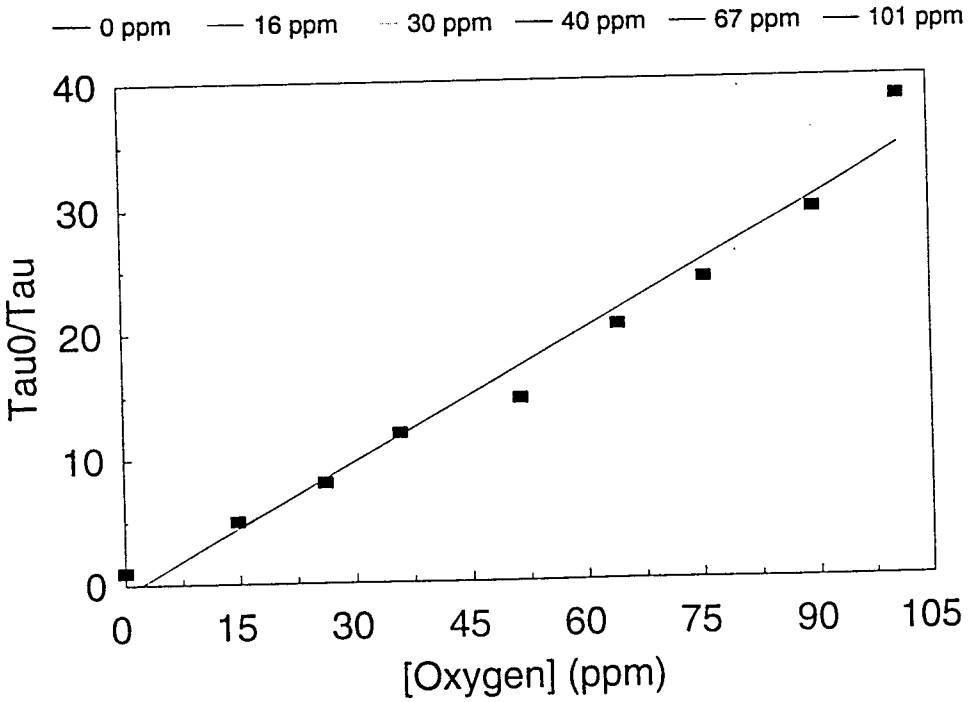


Figure 2 (a) and (b)

Oxygen Quenching of Pyrene Fluorescence in Isooctane





Figures 3 (a) and (b)

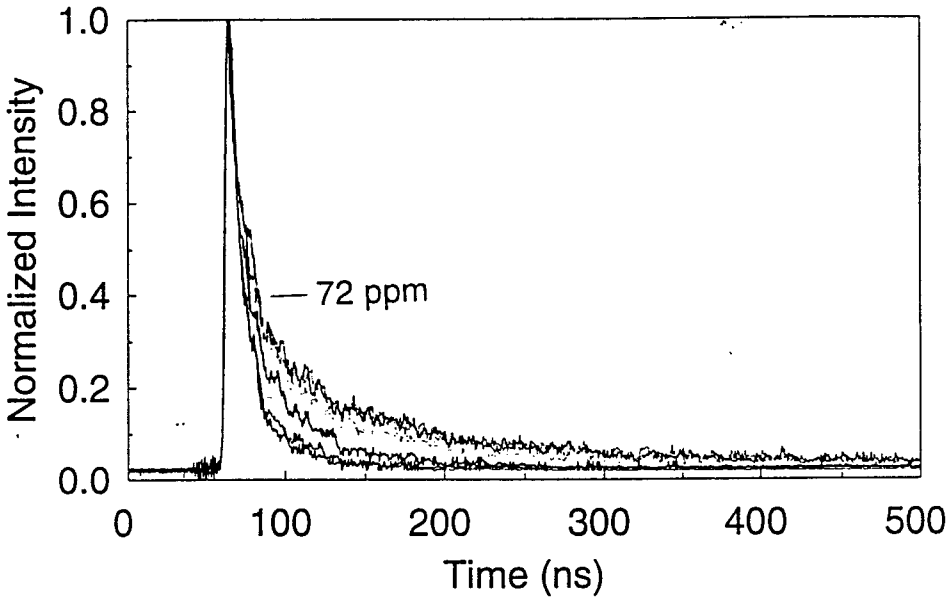
use a least squares approach to determine the lifetime. Figure 2(b) clearly yields a single exponential over at least two orders of magnitude in fluorescence intensity. This was typical for all of the pyrene decays we observed. The effect of oxygen quenching is shown in Figure 3(a). The lifetime of pyrene in isooctane clearly decreases with increasing $[O_2]$. Figure 3(b) is a Stern-Volmer plot for oxygen quenching in the pyrene-isooctane system. Good linear behavior is observed. Figure 3 can function as a calibration curve for the determination of $[O_2]$ in isooctane.

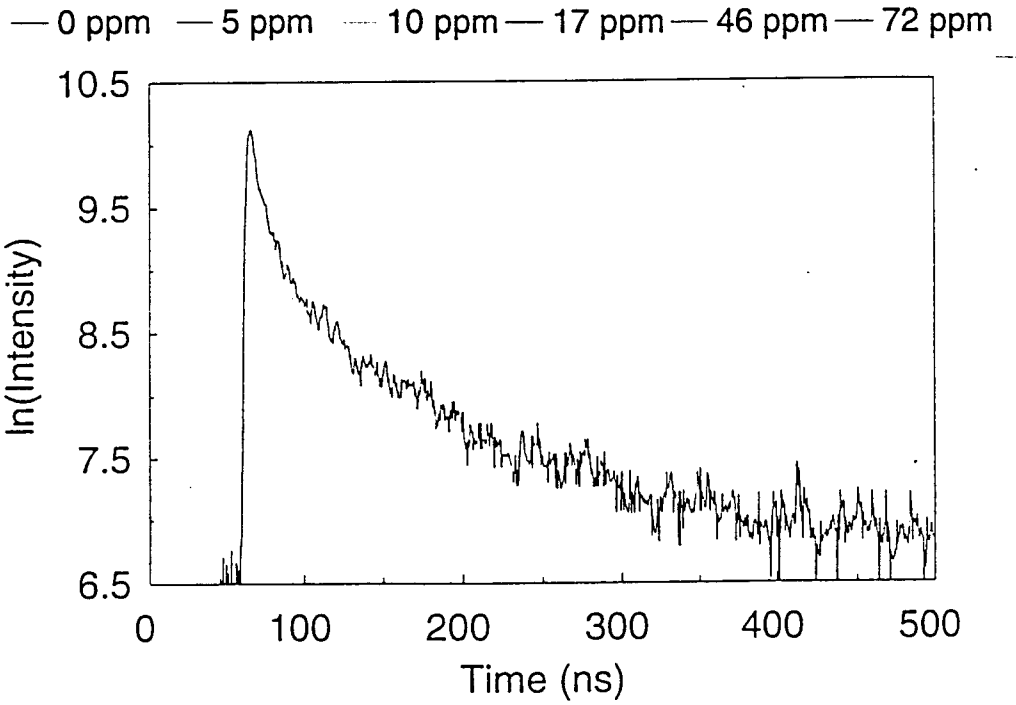
pure aviation fuel

Before considering the lifetime quenching of pyrene in aviation fuel, it is useful to consider the intrinsic fluorescence of the fuel. There are many native fluorophores in the fuel, some of which will be quenched by the presence of oxygen. Using the native fluorescence of the fuel would circumvent the need for pyrene addition to the fuel.

The time-resolved fluorescence of pure nitrogen sparged POSF-2827 appears in Figure 4(a). The decay is a **logarithmic** plot of the fluorescence intensity. This sharply contrasts with the results in Figure 2(b) for the pyrene-cyclohexane system. The POSF-2827 "linearized" decay is clearly not linear. In fact, there appears to be a virtual continuum of fluorescence lifetimes (each linear portion of the decay corresponds to a different lifetime). This arises from the large number of fluorescent species in the native fuel. Each fluorophore has a different lifetime. Thus, it is impossible to assign a single lifetime to the decay, and it is extremely difficult to fit the data even qualitatively with less than three lifetimes. A second difficulty with the fluorescence properties of the pure fuel is illustrated in the oxygen quenching plot of the fluorescence of pure POSF-2827 shown in Figure 4(b). Note the very short lifetimes of the unquenched system. This results in a compression of the range of lifetimes and a correspondingly narrow dynamic range. With the variance associated with data for each decay, this would result in a large uncertainty for determination of the oxygen concentration (assuming a meaningful set of lifetimes could be obtained).

Oxygen Quenching of POSF-2827 Intrinsic Fluorescence





Figures 4 (a) and (b)

From this it is apparent that the intrinsic time-resolved fluorescence of the fuels is not useful for oxygen quantitation.

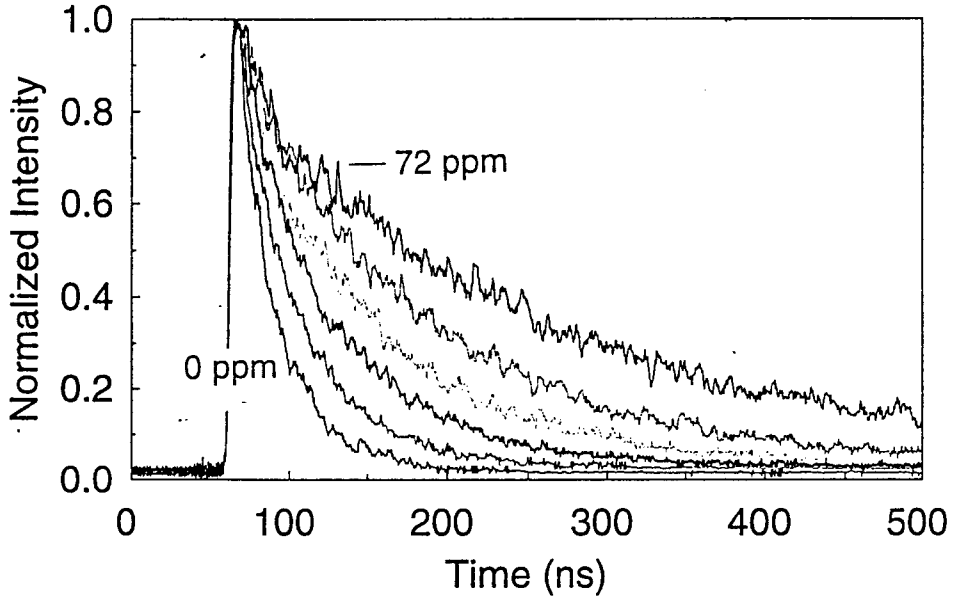
pyrene-doped fuels

Now we return to pyrene addition to the fuel. The pyrene fluorescence intensity at a level of 10 - 20 ppm in cyclohexane and isooctane is greater than the intensity of the native fuel fluorescence. This, coupled with the much longer fluorescence lifetime of pyrene relative to the pure fuel, should result in pyrene dominated fluorescence at long lifetimes. This is born out in the results in Figures 5(a) and 5(b) for the time-resolved fluorescence of the POSF-2827 fuel containing ~ 20 ppm pyrene. The natural logarithm of the decay of fluorescence intensity with time is clearly linear. The lifetimes show a broad range from air saturated to fully unquenched giving good precision to the results. It is also visible to the eye that fluorescence of the pyrene containing fuel is more intense. The pyrene could probably be decreased in concentration to the 2-3 ppm level without diminishing the quality of results significantly. All these data were collected with a neutral density filter which reflects ~99% of the fluorescence intensity away from the detector. The pyrene addition solves both problems associated with the lifetime behavior of the native fuels: the narrow dynamic range and the multi-exponential behavior.

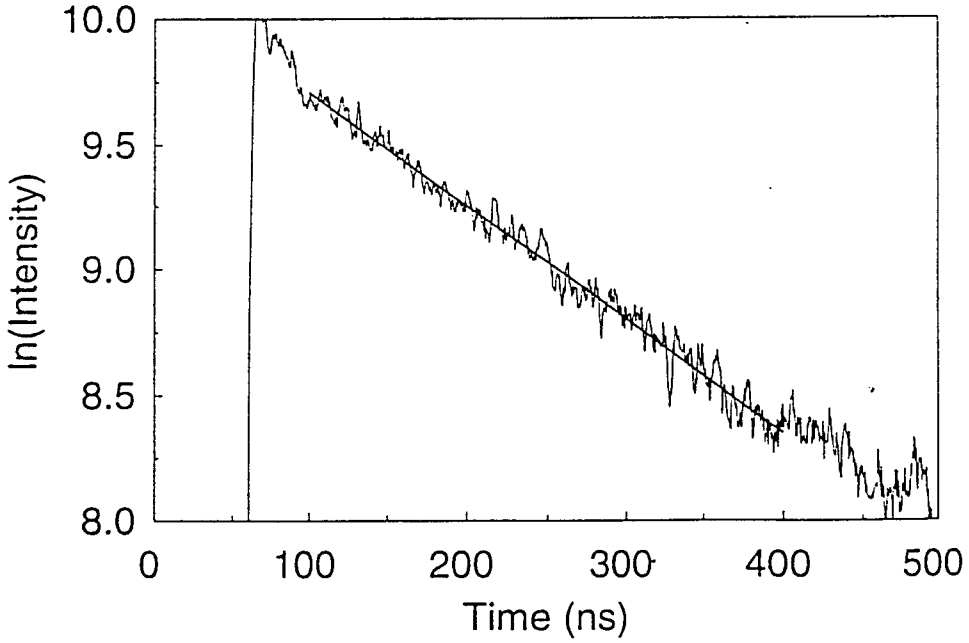
A series of Stern-Volmer plots for the oxygen lifetime quenching of pyrene-doped fuels (all doped at 17-23 ppm pyrene) is shown in Figure 6. Good linear behavior is observed for all four fuels. These calibration curves were used in the on-line flowing fuel studies discussed below.

A final consideration is the effect of thermal stressing on the lifetime behavior of the pyrene-doped fuel. As discussed in the introduction, thermal stressing of fuel results in oxidation and consumption of some of the pyrene so that steady-state fluorescence intensity measurements cannot be employed for oxygen quantitation. From equation (3) it is apparent that the rate constant for decay of the excited state is independent of the initial concentration of the excited state. (At much higher concentrations self-quenching and exciplex formation complicate the simple model shown in reaction 1.) This

Oxygen Quenching of Pyrene Fluorescence in POSF-2827



— 0 ppm — 10 ppm --- 17 ppm — 25 ppm — 46 ppm — 72 ppm



Figures 5 (a) and (b)

Pyrene in Aviation Fuel Stern-Volmer Plots

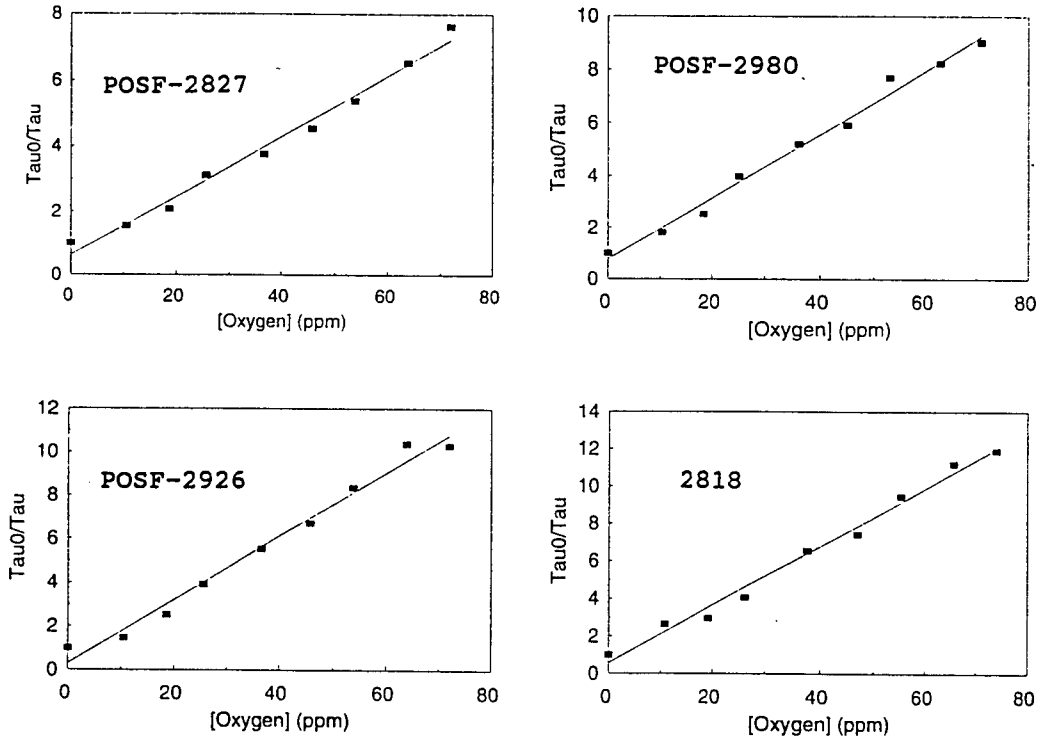


Figure 6

Effect of Thermal Stressing on Pyrene Fluorescence in POSF-2827

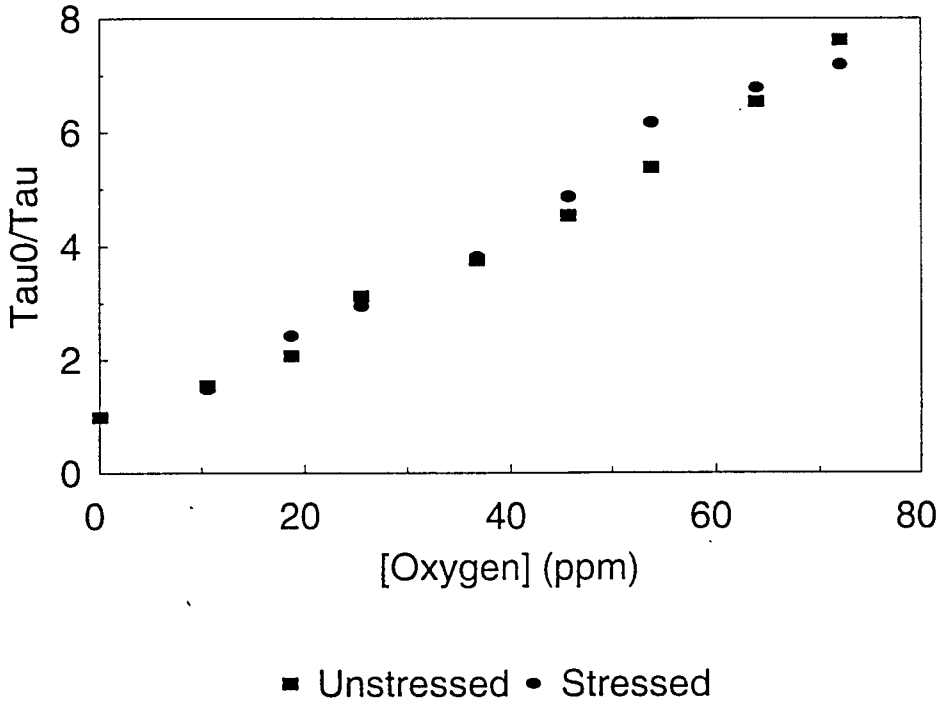


Figure 7

is demonstrated by the data in Figure 7. The stressed fuel has been doped with 20 ppm pyrene and has passed through the NIFTR consuming of all the oxygen for the initially air-saturated fuel. The stressed and unstressed systems show identical oxygen quenching behavior when reoxygenated.

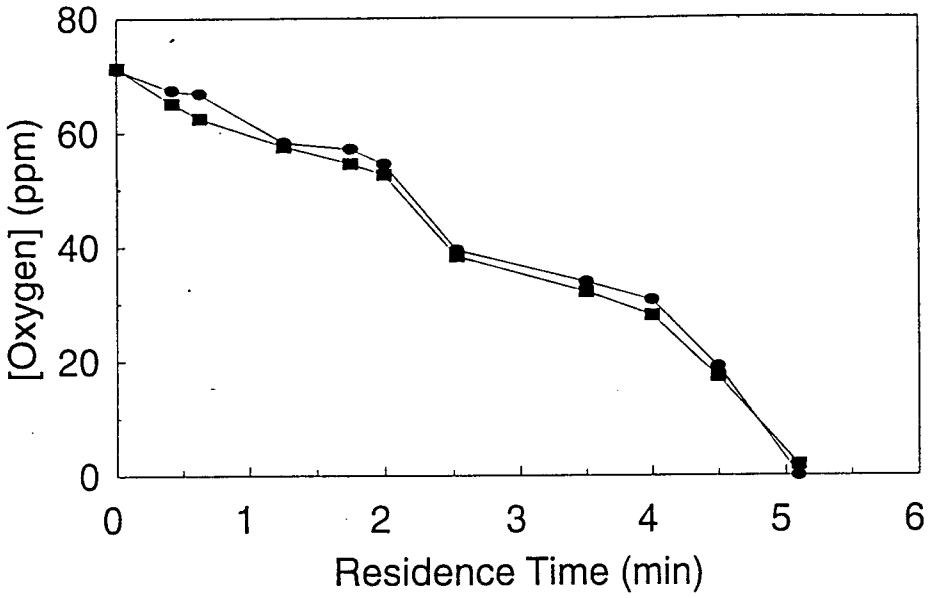
flowing rig studies

After demonstration of the viability of this approach to oxygen concentration measurements we applied this method to flowing fuel simulations. In the flowing fuel studies the fuel is pumped through a heated tube which simulates a jet fuel line. The pyrene is added to the fuel prior to the reaction. The fuel reacts with the initially dissolved oxygen. There are no headspaces in the system so further absorption of oxygen cannot occur. By varying the flowrate of the fuel through the tube it is possible to vary the residence time of the fuel in the tube and, hence, the reaction time at high temperature. The concentration of oxygen as a function of flowrate is measured to yield an oxygen consumption plot. The current method of oxygen measurement is gas chromatography (GC). The difficulties associated with this technique were discussed earlier. In order to test the validity of the fluorescence method we placed an optical cell in-line with the GC. It is important to emphasize that the fuel has returned to room temperature prior to the fluorescence and GC measurements. The fluorescence lifetime is dependent on temperature, which must be controlled.

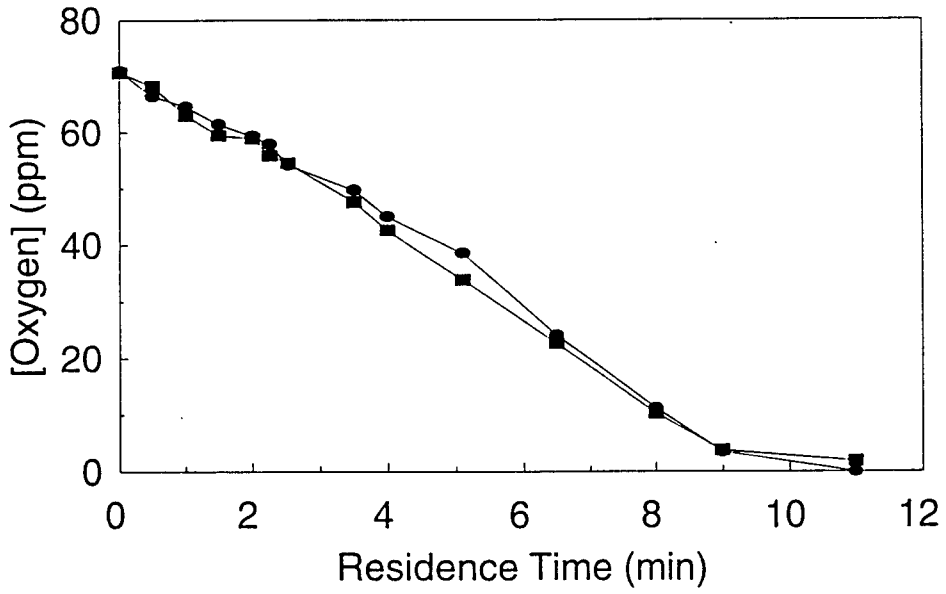
The results of the oxygen consumption experiment are shown in Figure 8. The agreement between the fluorescence and GC results is excellent. Even the structure on the decay curves (which reflects passivation effects of the reactor tube surfaces) is accurately reproduced. Also, the addition of pyrene does not appear to affect the rate of oxygen consumption. This suggests that the addition of pyrene to the fuel in trace amounts for diagnostic purposes does not affect the oxidation characteristics of the fuel. This point deserves further study.

One final point concerns the need for an external calibration curve. The Stern-Volmer plots of $(\tau_0)/(\tau_q)$ shown earlier were used as calibration curves for the data in Figure 8. However, a two point on-the-fly calibration curve can

Absolute Oxygen Consumption Curve: POSF-2980 with MDA in SS Tube



POSF-2980 with MDA in Restek Tube



Fluorescence GC

Figure 8

be generated. All the Stern-Volmer plots we obtained are linear, and accurate measurement of the air-saturated and fully unquenched lifetimes would be sufficient to define the calibration curves. By measuring the fluorescence lifetime of the fuel at the slowest flowrate, at which all oxygen is consumed, and then measuring the lifetime of the air-saturated fuel, a two-point calibration curve may be obtained. A comparison of the two-point fluorescence calibration method with GC is shown in Figure 9. The agreement between the gc and fluorescence data, though not as good as for the full calibration curve, is still good. Thus, under conditions in which an external calibration curve cannot be generated, this technique is still extremely useful. It is important to note that the GC technique also relies on a one-point calibration curve and the uncertainty in each of the data points in Figures 8 and 9 are $\pm$ 5 % absolute.

Summary

Fluorescence lifetime quenching of pyrene by oxygen has been demonstrated as a viable technique for determination of oxygen concentration in aviation fuel. Though the intrinsic fluorescence of the fuel shows oxygen quenching effects, the time-resolved behavior of the pure fuel fluorescence is sufficiently poor to prevent its use as a diagnostic. Addition of trace amounts of pyrene overcomes this problem, and a Stern-Volmer kinetic approach gives good results in both static and flowing rig studies. The advantages of lower cost, non-destructive in-situ monitoring, shorter measurement time (by a factor of 100), and capabilities for spatially resolved and rapidly time-varying measurements make this new technique a very attractive alternative to current chromatographic approaches.

Relative Oxygen Consumption Curve: POSF-2980 with MDA in SS Tube

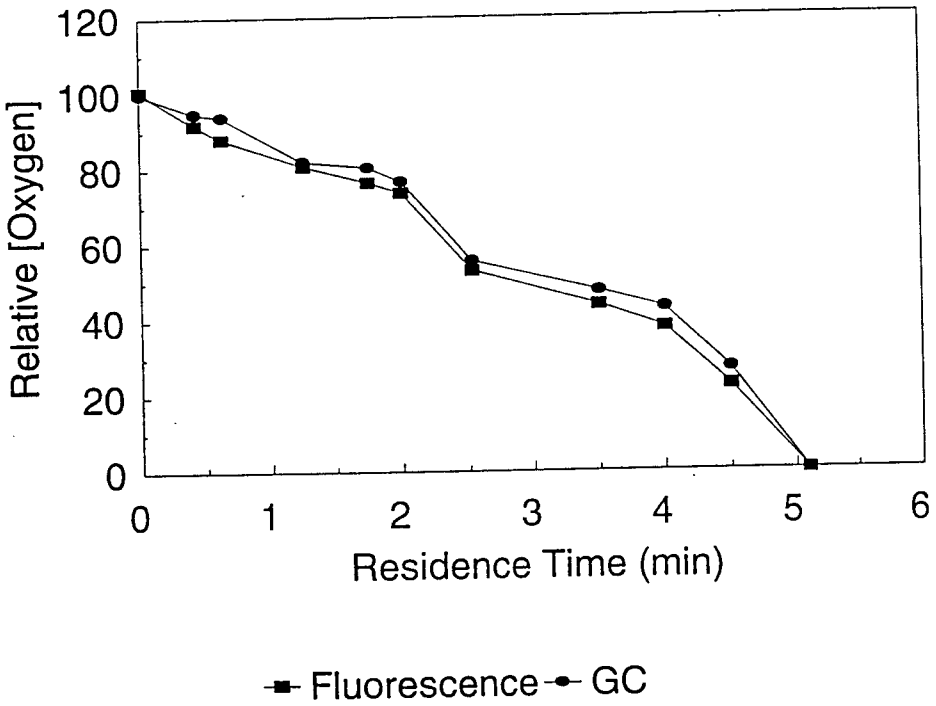


Figure 9

References

1) (a) Roquemore, W.M.; Pearce, J.A.; Harrison, W.E.; Krazinski, J.L.; Vanka, S.P.; Prepr.-Am. Chem. Soc. Div. Pet. Chem., 1989, 34, 841. (b) Parker, T.E.; Foutter, R.R.; Rawlins, W.T.; Ind. Eng. Chem. Res., 1992, 31, 2243. (c) Zabarnick, S.; Ind. Eng. Chem. Res., 1994, 33, 1348.

2) (a) Battino, R.; Rettich, T.R.; Tominaga, T.; J. Phys. Chem. Ref. Data, 1983, 12, 163. (b) "Oxygen and Ozone", IUPAC Solubility Data Series, Battino, R., Ed.; Pergammon:Oxford, 1981, Vol. 7.

3) Striebich, R.C.; Rubey, W.A.; Prepr.-Am. Chem. Soc. Div. Pet. Chem., 1994, 47-50.

4) Herzberg, G.; "Infrared and Raman Spectra of Polyatomic Molecules", Van Nostrand:New York, 1945.

5) Lakowicz, J.R.; "Principles of Fluorescencce Spectroscopy", Plenum:New York, 1983.

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

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

DEVELOPMENT OF AN ACTIVE DYNAMOMETER SYSTEM

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

Clarkson University
Box 5720
Potsdam, NY 13676

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

August 1994

DEVELOPMENT OF AN ACTIVE DYNAMOMETER SYSTEM

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

Abstract

In this paper, we describe the experimental development of a prototype computer controlled dynamometer system (i.e., an active load) which can produce arbitrary desired load torques for machines and drives testing. The dynamometer system consists of an arbitrary motor under test which is rigidly coupled to a load dynamometer motor. An advanced motion controller is then designed for the dynamometer motor such that it presents desired load dynamics to the motor under test. The control algorithm is implemented using a digital signal processor based data acquisition and control system. The nonlinear control approach is based on an integrator backstepping technique and facilitates the application of a broad class of high-performance dynamometer motion controllers.

DEVELOPMENT OF AN ACTIVE DYNAMOMETER SYSTEM

James J. Carroll

I. Introduction

The More Electric Aircraft (MEA) concept emphasizes the utilization of electrical power as opposed to hydraulic, pneumatic, and mechanical power for optimizing aircraft performance and life cycle cost [1]. Studies have shown that the MEA concept, with its long term goal of producing an All Electric Aircraft, yields a significant increase in aircraft reliability, maintainability, and supportability. A significant challenge to the MEA concept is its increased dependence on electric motors for servo and variable speed drive applications. MEA applications, such as (i) flight control and utility actuation, (ii) compressors for cooling aircraft subsystems, and (iii) pumping fuel and lubrication, require reliable high power density machines and drives with ratings from a few horsepower to a few hundred horsepower. For example, hydraulically driven flight control actuators could be replaced by electric motor driven actuators, gearbox driven lubrication and fuel pumps could be replaced by electric driven pumps, and pneumatically driven compressors for environmental control systems could be replaced by electric driven compressors. These MEA applications motivate the development of high performance motor drive systems with advanced motor controls for induction, permanent magnet, and switched reluctance machines. In addition, it motivates the development of sophisticated test equipment which can be used to verify the performance of proposed MEA systems. One such piece of equipment is an active dynamometer.

An active dynamometer (i.e., a computer controlled user-definable load), is an attractive concept for many reasons. Such a device could be used by motor drive manufacturers for the rapid design and implementation of motor controllers for a wide variety of machines and drive applications, such as MEA. The computer controlled dynamometer would eliminate the need to connect the motor/drive under test (MUT) to an actual load for performance test purposes. This would greatly reduce the motor drive manufacturer's expense and allow the motor drive to be tested under a wide variety of anticipated conditions. The potential "dual use" market for this type of device is clearly indicated by a recent Small Motors Manufacturers Association report of over one billion dollars in annual motor sales [2], a figure which represents only 20% of the North American fractional horsepower motor market.

Although the concept of a user defined load is not new, the type of dynamometers currently available on the market are only capable of simulating desired steady-state load torques. These devices typically consist of simple pony brakes, generators, or eddy current/magnetic particle brakes; therefore, they are not capable of accurately simulating some of the most commonly encountered industrial loads. Excluding the inertial components, typical dynamic loads can be divided into three main categories: (1) torques which are a function of time, (2) torques which are a function of load position, and (3) torques which are a function of load speed. The first class of torque profile is commonly produced by devices such as mixers, rolling mills, flying shears, and conveyors. The second class of torque profile is produced by piston based devices such as compressors or pumps. The latter class

of torque profile is commonly produced by devices such as fans, blowers, and centrifugal type compressors and pumps. MEA applications encompass all of the categories noted above.

The proposed active dynamometer consists of two rigidly coupled machines: a motor under test (MUT) and a dynamometer motor, which combine to form a common inertial load. In general, the MUT is driven by some form of servo control algorithm which is designed to meet specific performance objectives given an anticipated (i.e. desired) load torque, and is completely independent of the proposed dynamometer control system. As such, the dynamometer control algorithm must actively servo the dynamometer motor in such a way that it presents a desired load torque to the MUT without affecting its servo performance. This implies that the closed-loop response (i.e., position, velocity, acceleration, and current), of the MUT-dynamometer system will be the same as the response obtained when MUT drives the actual desired load. The proposed device would facilitate accurate performance testing of even complex mechanical loads and load profiles, such as those encountered in MEA applications.

This paper describes the experimental development of a prototype active dynamometer system which was constructed at the Wright Laboratory. The prototype construction uses a permanent magnet brush dc motor for the dynamometer actuator which is controlled by a state-of-the-art digital signal processor (DSP) based data acquisition and control (DAC) system, as shown in the block diagram of Figure 1. The completed dynamometer system will facilitate two levels of motor drive testing: (1) accurate computer simulation of the proposed MUT

controllers under desired load torques and (2) the real-time control of the MUT servo control algorithm under a *simulated* desired load which is provided by the dynamometer system. The word *simulated* is used to stress that for all intents and purposes, the MUT is experiencing the desired load torque profile, but that this load torque is being provided by the dynamometer actuator under some form of advanced computer control.

II. The Prototype Dynamometer System

The dynamometer system consists of two individual machines (i.e., a MUT and the dynamometer motor), which are rigidly coupled together as shown in the upper portion of the dynamometer system block diagram of Figure 2. In order to simplify the analytical controller development for this prototype system, two machines with linear electrical dynamics [3] were selected. The machine on the left-hand side of Figure 2 is a separately excited brush direct-current (SEBDC) motor, and it serves as the MUT. The machine on the right-hand side of Figure 2 is a permanent magnet brush dc (BDC) motor, and it serves as the dynamometer actuator. For notational convenience, we shall subscript all references to the MUT with a 1, and all references to the dynamometer motor (DYNA) with a 2. For modeling purposes, the MUT is assumed to be excited with a constant field current. Given this, the dynamic model of the dynamometer system can be written as shown [4]

$$\bar{J}\ddot{q} + \bar{B}\dot{q} = \tau_1 + \tau_2, \qquad (2.1)$$

$$\tau_1 = K_1 I_1, \tag{2.2}$$

$$\tau_2 = - K_2 I_2, \tag{2.3}$$

$$L_1 \dot{I}_1 + R_1 I_1 + K_1 \dot{q} = V_1, \tag{2.4}$$

$$L_2 \dot{I}_2 + R_2 I_2 + K_2 \dot{q} = V_2, \tag{2.5}$$

with the auxiliary parameter definitions

$$\bar{J} = [J_1 + J_2], \quad \text{and} \quad \bar{B} = [B_1 + B_2], \tag{2.6}$$

where J_i, B_i, K_i, L_i, and R_i represent the coefficients of rotor inertia, viscous damping, electromechanical torque coupling, winding inductance, and winding resistance associated with each motor, respectively. The variables q, τ_i, I_i, and V_i refer to the position, torque, current, and voltage associated with each motor, respectively. Note, a "dot" is used throughout the development to designate a differentiation with respect to time.

III. Problem Definition

To be a practical device, a computer controlled dynamometer must present a desired load torque $\tau_d(q,\dot{q},\ddot{q})$ to the MUT in such a way that it ensures an accurate servo response (i.e., position, velocity, acceleration, and current), given an arbitrary class of MUT controllers (i.e., voltages V_2). The MUT should therefore reproduce the servo

response obtained when driving an actual load torque $\tau_d(q,\dot{q},\ddot{q})$. This implies that the control action taken by the dynamometer motor to produce $\tau_d(q,\dot{q},\ddot{q})$ should not affect the servo performance of the MUT. It also implies that the MUT does not require any information about the dynamometer system in order to drive the desired load torque $\tau_d(q,\dot{q},\ddot{q})$. Given these criteria, the control objectives can be stated as follows: (1) design a voltage level controller V_2 such that the dynamometer motor presents a desired load torque $\tau_d(q,\dot{q},\ddot{q})$ to the MUT, and (2) structure the control design such that the MUT controller (i.e., voltage V_1), is independent of the dynamometer dynamics and controller.

IV. Dynamometer Controller Development

To develop the dynamometer control algorithm, we use an *integrator backstepping* technique [5,6] which allows us to directly specify the dynamometer motor voltage. The approach facilitates the development of a wide variety of dynamometer controllers, such as embedded computed torque, robust, or adaptive controllers, depending on the amount of information available on the dynamometer system (i.e., MUT and dynamometer motor). For purposes of this discussion, we assume exact knowledge of the entire dynamometer system and construct embedded computed torque controllers for both the dynamometer motor and the MUT. These controllers will theoretically yield a globally uniform asymptotic stability result for the MUT trajectory tracking error and the dynamometer load torque tracking error given full-state measurements. If we select desired load torque dynamics to simulate windage as shown

in [4], then we can specify an embedded voltage controller V_2 of the form

$$V_2 = R_2 I_2 + K_2 \dot{q} \qquad (4.1)$$
$$- \left[L_2 K_1 [1 - \phi] [- R_1 I_1 - K_1 \dot{q} + V_1] - L_1 L_2 \dot{w}_2 + \Gamma_2 \eta_2 + \Gamma_3^{-1} \eta_2 \right] / (L_1 K_2 \phi),$$

where all terms are explained fully in [4].

V. MUT Controller Development

The control design is structured such that the MUT control voltage V_1, can be specified independently of the dynamometer controller. In fact, we are free to design V_1 as if we were actually driving the desired load torque $\tau_d(q,\dot{q},t)$, in this case simulated windage. Assuming exact knowledge of the entire dynamometer system as noted above, we can specify an embedded computed torque controller (i.e., trajectory tracking) for the MUT, as shown

$$V_1 = R_1 I_1 + K_1 \dot{q} + \Gamma_1 \eta_1 + K_1 r \qquad (5.1)$$
$$+ L_1 \left[\bar{J}_1 \left(\frac{d}{dt} \{ \ddot{q}_d \} + a \ddot{q}_d \right) + [2 C_d \dot{q} + \bar{B}_1 - \bar{J}_1 a] \ddot{q} + \Gamma_0 [(\ddot{q}_d + a \dot{e}) - \ddot{q}] \right] / K_1,$$

where all terms are explained fully in [4].

Remark 5.1

Since the proposed MUT controller requires knowledge of the system acceleration $\ddot{q}$, the signal must be generated on-line by the dynamometer controller (see the lower block of Figure 2), using the procedure described in [4].

VI. System Simulation

The proposed prototype controllers of (4.1) and (5.1) were simulated using the measured motor parameters values given in Appendix B of [8], and a commercially available numerical integration software package for PC compatible computers called Simnon [7]. Since these results were qualitatively similar to those obtained in [4], they are not repeated here for brevity.

VII. The Experimental Setup

The prototype dynamometer system consists of the following key components: (1) an IBM AT compatible 33MHz 386PC, (2) a digital signal processor (DSP) board, (3) an analog to digital input board, (4) an encoder interface board, (5) a digital to analog output board, (6) two Hall-effect current sensors, (7) a pulse width modulated (PWM) power amplifier, (8) a linear power amplifier, (9) a separately excited (SE) brush dc motor (which serves as the MUT), (10) a permanent magnet (PM) brush dc motor (which serves as the dynamometer actuator), and (11) assorted electronics interfacing and hardware.

As noted in Section I, the prototype dynamometer system can be broken up into two distinct parts: (i) the dynamometer teststand, which consists of a MUT and a dynamometer actuator rigidly coupled together, and (ii) the digital signal processor (DSP) based data acquisition and control (DAC) system. A close-up photograph of the prototype dynamometer teststand is shown in Figure 3(a), and the complete prototype dynamometer system, including the PI, is pictured in Figure 3(b).

The MUT, pictured on the left-hand side of Figure 3(a), is a salvaged SE dc motor with ratings of 24 Vdc at 9.0 A. The dynamometer actuator, pictured on the right-hand side of Figure 3(a), is a PM Baldor Model M4070 dc motor with ratings of 100 Vdc at 9.2 A. The dynamometer actuator is also equipped with an encoder and a tachometer for position and velocity measurements, respectively. The encoder has 500 counts per revolution and the tachometer outputs 7 Vdc per 1 KRPM. Both the MUT and the dynamometer motor were fully parameterized for control purposes using a standard Magtrol Absorption Dynamometer. The results of these tests are summarized in Appendix B of [8].

Power is supplied to the dynamometer actuator (i.e., the PM motor), by a Copley Control Model 261V PWM power amplifier. This sophisticated device has an 81 KHz PWM carrier frequency and is capable of true four quadrant operation (i.e., ±50 A at ±350 Vdc). The effects of PWM switching noise are minimized using a 13 KHz low pass LC filter at the amplifier output. As a result, the device looks for all practical purposes, as if it were a linear amplifier. The amplifier was adjusted to provide a noninverting voltage gain of 4 V/V with 0.0 Vdc offset, and an output voltage limit of ±32 Vdc. The amplifier is supplied via a Hewlett Packard HP6477C adjustable dc supply with the bus voltage set at 100 Vdc, the crowbar overvoltage protection set at 150 Vdc, and the current over limit set at 20 Adc. The bus is fused with a 20 A solid state fuse. For user safety, the bus has an analog volt meter for visual inspection of the bus voltage and a user selectable dump network. The dump network consists of a 16 Ω resistive load which can be switched across the PWM supply while simultaneously opening the supply bus. This

insures that all energy is completely drained from the PWM amplifier's internal 3,000 μF capacitor bank. Power is supplied to the MUT (i.e., the SE motor), by a Kepco linear power amplifier (model number not available), which is capable of supplying +6 A at ±40 Vdc at bandwidth of approximately 2 KHz. The amplifier was adjusted to provide an inverting voltage gain of -4 V/V with 0.0 Vdc offset. The amplifier input/output cabling and connections are completely described in Appendix B of [8]

The motor currents are measured indirectly using Microswitch Model CSSLB1AH Hall-effect sensors. These devices output a voltage which is linear with respect to the measured current, at frequencies from DC to 125 KHz. The current and tachometer signals must be properly conditioned so that they can be interfaced with the DAC hardware. The signal conditioning circuitry was built using an internally supplied ±12 Vdc protoboard with added screw terminal connectors for input and output connections. The control signal inputs are via cabling described in Appendix B of [8]. These circuits were designed to scale, offset, and when necessary clip, the control measurements (i.e., MUT current, DYNA velocity, and DYNA current), to the ±2.5 Vdc signal level required by the Analog Input Board. This is accomplished for each channel using the opamp based circuit shown in Figure 4 (see Appendix B of [8] for a component listing). Note, the solid state devices in the opamp feedback path are precision adjustable shunt regulators which function as zener diodes. These diodes act to protect the Analog Input Board from possible overvoltage by clipping the opamp output at ±3 Vdc.

The DAC's central nervous system is a DSP Processor Board (Spectrum

Signal Processing, Inc.), which also serves its computational engine. This board consists of the following key components: a central processing unit (CPU), local memory, a PC interface, and a parallel expansion bus. The processor is a TMS320C30 DSP chip which contains both integer and floating-point arithmetic units, 2048 × 32 bit words (8K bytes) of on-chip RAM, 4096 × 32 bit words of on-chip ROM, a control unit and parallel/serial interfaces. The CPU operates from a 33.3 MHz clock and can achieve 16.7 million instructions per second (MIPS) with a peak arithmetic performance of 33.3 million floating-point operations per second (FLOPS). Two memory areas are provided off-chip two supplement the 2K word RAM on-chip. These memory areas are divided up into 64K words of zero wait state memory and 64K words of one wait state memory. The board is compatible with IBM AT class computers using a full 16-bit ISA interface. Access to memory passes through dual porting hardware on the TMS320C30, and interface throughput is limited only by the speed of the PC and its software. The TMS320C30 dual port interface includes an address counter for block transfers and hardware to transfer between the 16-bit AT bus and the 32-bit DSP bus. Interrupts from the PC to the TMS320C30, and vice-versa, are also supported. A parallel expansion system is provided as a memory-mapped peripheral area via a 50-pin connector. It has a 16-bit width and follows a standard DSPLINK arrangement. All of the DAC boards communicate with each other via this high speed, parallel link which is independent of the PC's ISA bus. Transfers over this link use 2 wait-states to achieve a 180 nsec transfer cycle which is suitable for ribbon-cable connection to the peripheral DAC boards described below.

A 32 Channel Analog Input Board (Spectrum Signal Processing, Inc.), provides the DAC system with 32 analog input channels multiplexed to a fast (3 μsec), 12-bit analog to digital converter (ADC). All 32 channels have input buffering and a first order loss pass filter to reduce unwanted high frequency noise. The channels are arranged in four groups, with all channels in a group being sampled simultaneously. Each channel has an input voltage range of ±2.5 Vdc. A 16-bit counter can be programmed to provide a regular sampling rates up to a maximum of 100 KHz per channel. All control and data transfer is via the 50-pin DSPLINK connector to/from the DSP Processor Board. External connections to the 32 channels of ADC input are made via a 37-pin Type D connector.

A DS-2 I/O Board (Integrated Motions Inc.), is a two axis data acquisition and control module consisting of 2 channels of DACs, 2 channels of ADCs, four bits of digital input and output, and two channels of quadrature decoding (i.e., shaft encoder interface). The quadrature decoders are based on the Hewlett Packard HCTL-2016 chip. The Phase A and Phase B encoder inputs are at TTL logic levels. The maximum frequency on either channel is 2 MHz (i.e., if both channels are switching at 2 MHz then the position is changing at 8 MHz, giving a maximum axis velocity of 8,000,000 encoder counts per second). Since the HCTL-2016 has a 16-bit internal counter, a maximum of ±32,767 encoder counts can be stored before over/under flow occurs (note, this determines the minimum required sampling period for position updates). All control and data transfer is via the 50-pin DSPLINK connector to/from the DSP Processor Board. External connections to the two channels of quadrature decoding are made via a 37-pin Type D connector.

A 16 Channel Analog Output Board (Spectrum Signal Processing, Inc.), provides 16 digital to analog channels (DAC), each consisting of a double buffered 12-bit DAC, a 2nd order programmable analog output filter, and a buffer amplifier. The output voltage range for each channel is +8.188 to -8.192 volts using an internal reference voltage. All control and data transfer is via the 50-pin DSPLINK connector to/from the DSP Processor Board. External connections to the 16 channels of DAC are made via a 37-pin Type D connector. See Appendix B of [8] for a complete listing of all DAC hardware settings and connections.

The software used to implement the proposed DAC system consists of the following key components: (i) Matlab, a windows based analysis environment from MathWorks, Inc., (ii) user developed C-code which executes the desired control algorithms on the DSP board, and (iii) a user developed C++-code PC control environment which provides an interface between Matlab and the DSP board. The control environment allows the user to write, load, and run the C-code programs on the DAC system, as well as, perform data analysis within Matlab.

VIII. Experimental Results/Conclusions

We have initiated the development of a prototype active dynamometer system at Wright Laboratory. The prototype system will be capable of producing arbitrary desired loads for machines and drives testing. A control algorithm has been designed for the prototype dynamometer system which can presents desired load dynamics to MUTs. A test servo control algorithm has been implemented on the prototype teststand, to

demonstrate basic DAC functionality. Once software development for the DAC system is completed, we will be able to implement the proposed dynamometer controller and verify the simulation results experimentally.

References

[1] J. Weimer, "Electrical Power Technology for the More Electric Aircraft," Proc. of the AIAA/IEEE Digital Avionics Systems Conference, Orlando, FL, July 1993, Vol. 1, pp. 445-450.

[2] "SMMA Motor Market Survey Confirms Sluggish Recovery", PCIM, November 1992, pp. 8.

[3] P. Krause, Analysis of Electric Machinery, NY: McGraw-Hill, 1986.

[4] J. Carroll, D. Dawson and E. Collins, "A Nonlinear Control Technique for the Development of a Computer Controlled Dynamometer,"Proc. of the ASME Winter Annual Meeting, New Orleans, LA, November 1993, DSC-Vol. 53, pp. 31-36.

[5] P. Kokotovic, "The Joy of Feedback: Nonlinear and Adaptive", IEEE Control Systems Magazine, Vol. 12, pp. 7-17, June 1992.

[6] J. Carroll and D. Dawson, "A Backstepping Technique for the Tracking Control of Permanent Magnet Brush DC Motors Using Full State Feedback: Theory, Simulation, and Experimental Verification", IEEE Industry Applications Society Annual Meeting, Toronto, Canada, October 1993, submitted.

[7] Simnon User's Guide for MS-DOS Computers Version 3.2, Goteborg, Sweeden:SSPA Systems, 1993.

[8] J. Carroll, "Development of a Prototype Active Dynamometer System," Wright Laboratory Technical Report, to appear.

Appendix A

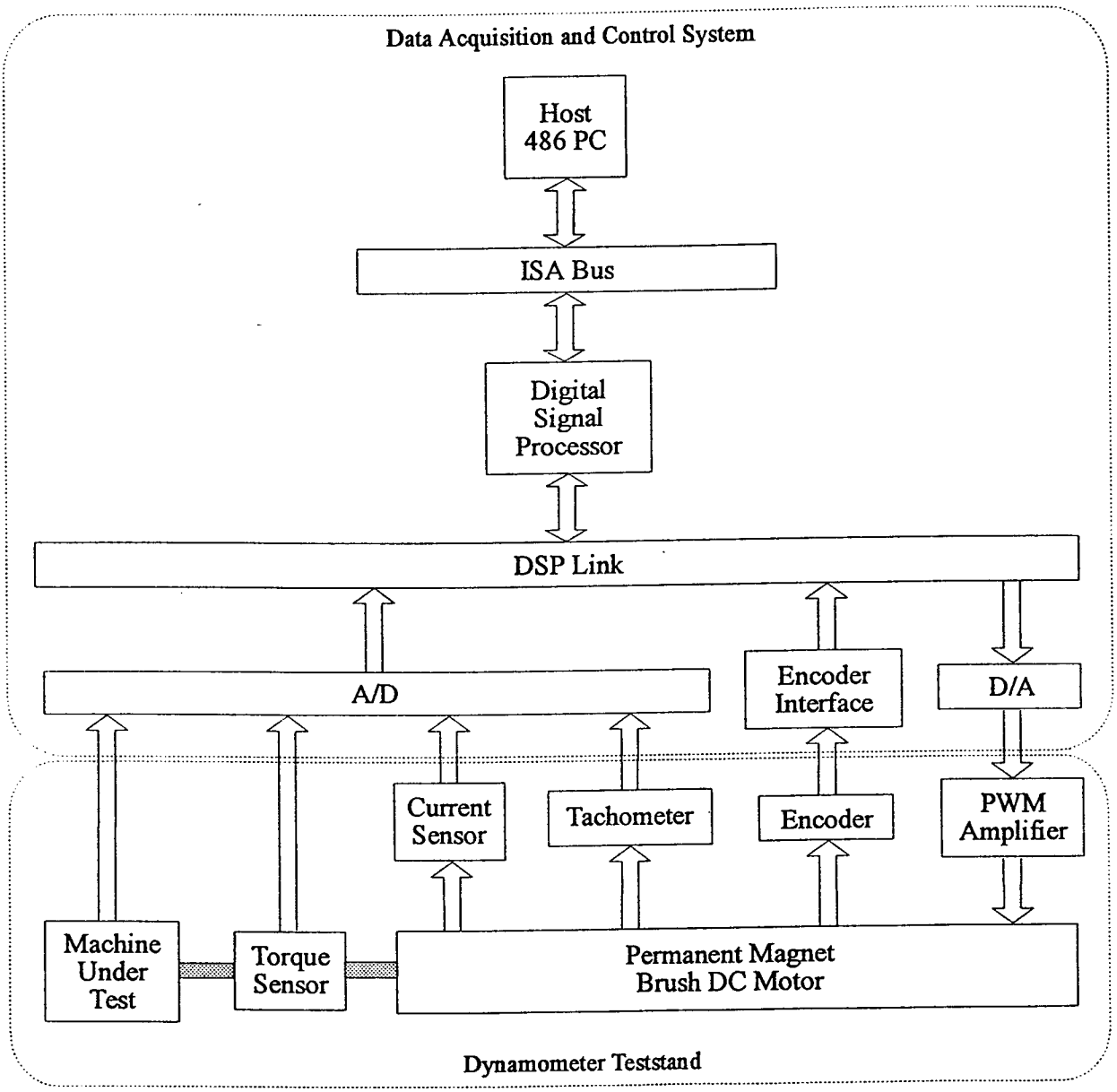


Figure 1: A Functional Block Diagram
of the Prototype Dynamometer System.

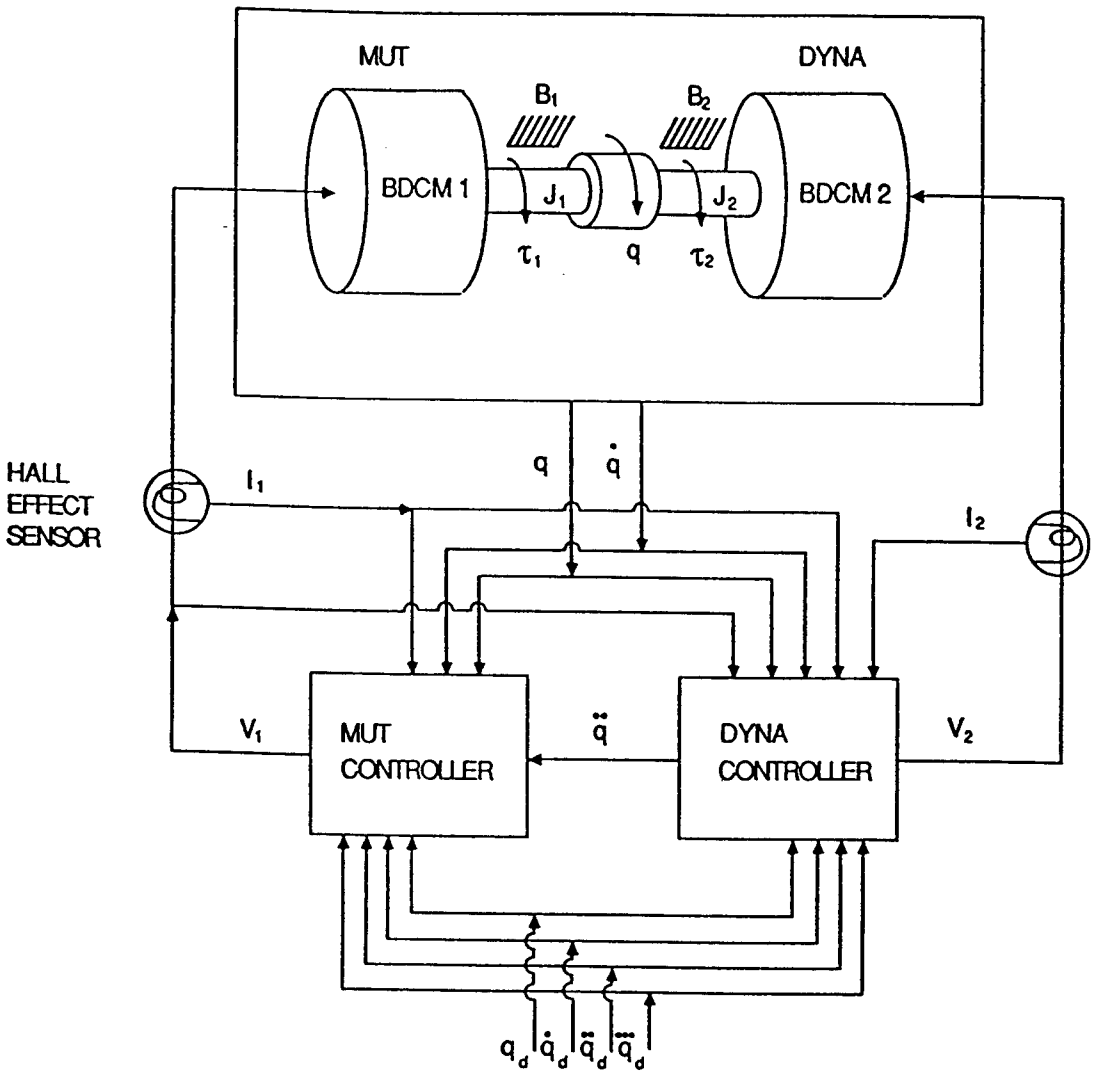


Figure 2: A Control Block Diagram
of the Prototype Dynamometer System.

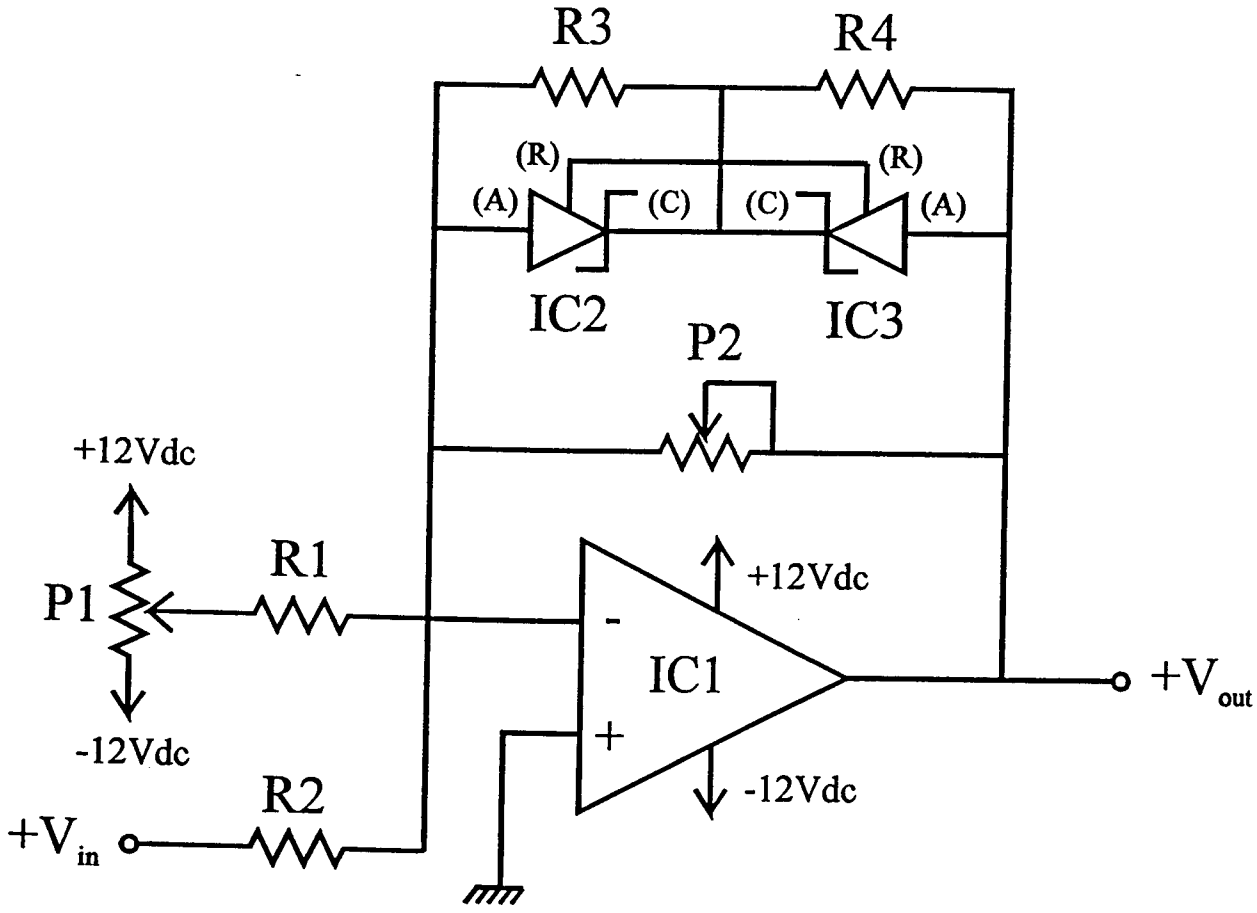


Figure 3: (a-top) A Close-up of the Prototype Dynamometer Teststand, (b-bottom) the Complete Prototype Dynamometer System.

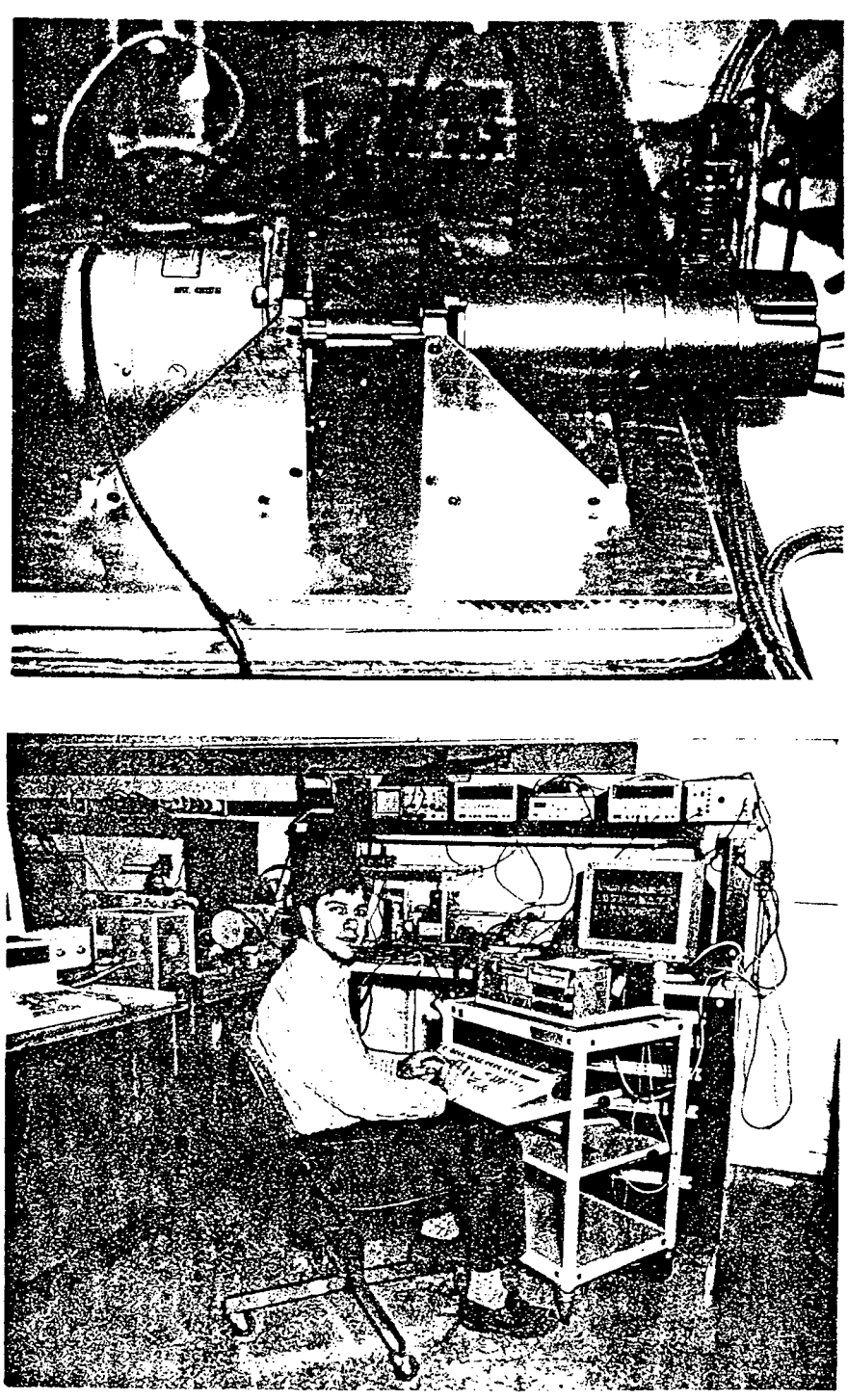


Figure 4: A Schematic Diagram of the Signal Conditioning **Circuit.**

The Least-squares Finite Element Methods for incompressible Flow with Zero residual for Mass Conservative Law

Ching Lung Chang
Associate Professor
Department of Mathematics

Cleveland State University
Euclid Avenue at East 24th Street
Cleveland, OH 44115

Final Report for:
Summer Faculty Research Program
Wright Laboratory
Wright-Patterson Air Force Base

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

and

Wright Laboratory

August 1994

Least-Squares Finite Element Methods for Incompressible Flow with Zero Residual for Mass Conservative Law

Ching Lung Chang
Associate Professor
Department of Mathematics
Cleveland State University

Abstract

This report contains two parts. In part one, a numerical method for least-squares finite element method (LSFEM) which enforces mass conservation, and the corresponding mathematical analysis is presented. In part two, a LSFEM for Stokes flows with multiple fluids is developed.

During the last few years, people have tried to find a new method for simulating incompressible flow without being subjected to the inf-sup condition, to which end LSFEM has been developed. In this work it was found that in simulating the flow about a cylinder moving along the axis of a narrow channel using the LSFEM in the vorticity-velocity-pressure form, the mass conservation law was not satisfied everywhere in the domain. During the Summer of 1994, Captain John Nelson and I developed a restricted LSFEM using the Lagrange multiplier which insure that the mass is conserved everywhere.

In the second part, a LSFEM is developed to simulate flows involving multiple fluids. For multiple fluid flows, not only the body equations governing the flow, but also conditions of continuity of velocity and stress across the interface separating the two fluids must be satisfied. Unlike the Galerkin method, the conditions for continuity of stress must be explicitly added to the LSFEM. In the mutiple fluid LSFEM, the condition for continuity of stress are viewed as restrictions and are added to the standard LSFEM in the stress-velocity-pressure form by using Lagrange multiplers. We present the results of using this method for a test case simulation.

This work was performed with Captain John J. Nelson in the Wright Laboratory WP AFB, 1994

PART ONE

Least–Squares Finite Element Method for the Stokes Equations with Zero Residual of Mass Conservation

Ching Lung Chang

In this research the simulation of incompressible flow in 2 dimensions by the least-squares finite element method (LSFEM) in the voticity-velocity-pressure version is studied. In the LSFEM, the equation for continuity of mass, equations of momentum and a vorticity equation are minimized on a descritization of the domain of interest. A problem is these equations are minimized in a global sense. Thus this method may not enforce that $div\underline{u} = 0$ at every point of the discretization. In this research a modified LSFEM is developed which insures near zero residual of mass conservation, i.e. $div\underline{u}^h$, is nearly zero at everywhere of this discretization. This is accomplished by adding an extra restriction in the divergence free equation through the Lagrange multiplier strategy. In this numerical method the inf-sup, or say Babuska-Brezzi, condition is no longer neccesary and the matrix resulting from applying the method on a discretization is symmetric. The uniqueness of the solution and the application of the conjugate gradient method is also valid. Numerical experience is given by simulating the flow of a cylinder with diameter 1 moving in a narrow channel of width 1.5. Results obtained by the LSFEM show that mass is created or destroyed at different points in the interior of discretization. The results obtained by the modified LSFEM show the mass is nearly conserved everywhere.

1 Introduction

During the last decade , many mathematicians and engineers have studied the least-squares finite element methods (LSFEM) for the incompressible Navier-Stokes equations, *e.g.* [2], [4], [6], [7], [10], [13], [14], [16], [17]. In these methods, a functional is defined which measures the error between any solution which may exist in the defined space, and the continuous solution to governing equations of motion. For example, the functional defined for the general LSFEM in the vorticity-velocity-pressure formulation for Stokes flow is

$$J^h(\underline{U}) = \| \omega_y + p_x - f_1 \|_0^2 + \| -\omega_x + p_y - f_2 \|_0^2 + \| \text{curl}\underline{u} - \omega \|_0^2 + \| \text{div}\underline{u} \|_0^2, \qquad (1.1)$$

where ω is the vorticity and is defined in §2. The member of the space which minimizes this functional in the space gives the approximated solution to the governing equations. These methods release the divergence free restriction. Therefore equal order finite element spaces can be applied for the test and trial function spaces. The advantages of the LSFEM are that the continuous piecewise polynomials can be used for test and trial functions without being subjected to the saddle point condition; and the corresponding matrix of the linear systems is symmetric and positive definite. This allows the use of efficient schemes to solve large systems.

In many cases the application of velocity boundary conditions plays the crucial role in the successful simulation of incompressible flows. Recently Pavel Bochev and Max Gunzburger [6] presented a mesh-dependent least-squares finite element method. In this method they formulate a weighted least squares functional:

$$J^h(\underline{U}) = \| \omega_y + p_x - f_1 \|_0^2 + \| -\omega_x + p_y - f_2 \|_0^2 + h^{-2}(\| \text{curl}\underline{u} - \omega \|_0^2 + \| \text{div}\underline{u} \|_0^2). \qquad (1.2)$$

A theoretical analysis by the theory of ADN [1] shows the mesh dependent LSFEM is optimal for the simulation of flows with velocity boundary conditions. For example, after one defines a finite element space of piecewise quadratic polynomial functions for the velocity $\underline{u}$, and piecewise linear polynomial functions for ω and p, the approximated solution $\underline{U}^h$ which minimizes (1.2) is the approximated solution of the Stokes problem which is of optimal order.

In order to test the practical applicability of the LSFEM, several authors have used this method to simulate the flow in a driven cavity [3], [5], [13], [14]. All calculations present reasonably good results. In this research, we are going to test the general and mesh dependent LSFEM in the velocity-vorticity-pressure formulation [6], [10], [14], [17] by simulating a cylinder of diameter 1 moving along the centerline of a narrow channel of width 1.5. The centerline of the channel is along the x coordinate axis. If the cylinder is moving with speed 1, by mass conservation the average value of u_1 (x-component of velocity) along a vertical (x = constant) line connecting the top of the cylinder and the nearest wall should be 3. Our calculations using the above methods give an average value of about 0.8, i.e. neither of the above methods ensures that mass is conserved in each element in our calculation. The cause of this problem is felt to be in the LSFEM the error is minimized on a global scale, allowing errors of significant size to remain on a local scale, especially in areas which the gradients of the variables are of significant size. These areas are the places of most interest. In this paper we modify the general LSFEM for the Stokes problem so that the method nearly conserves mass at every point. This is done by adding an extra restriction to this method (a restricted LSFEM) which ensures that the equation for conservation of mass is satisfied in every element.

2 The Least-Squares Method in the Vorticity-Velocity-Pressure Formulation

In this section we present an overview of the LSFEM in the vorticity-velocity-pressure formulation. We assume Ω is a bounded and connected domain in 2-D with a polygon boundary Γ. Let $\underline{f} \in [L^2(\Omega)]^2$ be a given function of body force. The Navier-Stokes problem can be presented as:

$$\begin{cases} -\nu\Delta\underline{u} + \underline{u}\cdot\text{grad}\underline{u} + \text{grad}p = \underline{f} & \text{in } \Omega \\ \text{div}\underline{u} = 0 & \text{in } \Omega \\ \underline{u} = \underline{u}_0 & \text{on } \Gamma, \end{cases} \tag{2.1}$$

where $\underline{u}$, p with $\int_\Omega p = 0$ are velocity and pressure, all of which are assumed to be nondimensionalized, and $\underline{u}_0$ is a given function on Γ. The parameter ν is the inverse of the Reynolds number $\Re$. The velocity-vorticity-pressure version has the following form if we introduce the vorticity, $\omega = \text{curl}\ \underline{u}$,

$$\begin{cases} \nu\text{curl}\omega + \underline{u}\cdot\text{grad}\underline{u} + \text{grad}p = \underline{f} & \text{in } \Omega \\ \text{curl}\underline{u} - \omega = 0 & \text{in } \Omega \\ \text{div}\underline{u} = 0 & \text{in } \Omega \\ \underline{u} = \underline{u}_0 & \text{on } \Gamma. \end{cases} \tag{2.2}$$

In this paper we restrict our attention to the Stokes problem with velocity boundary conditions. Without loss of generality we assume we have a homogeneous boundary condition as:

$$\begin{cases} \text{curl}\omega + \text{grad}p = \underline{f} & \text{in } \Omega \\ \text{curl}\underline{u} - \omega = 0 & \text{in } \Omega \\ \text{div}\underline{u} = 0 & \text{in } \Omega \\ \underline{u} = 0 & \text{on } \Gamma. \end{cases} \tag{2.3}$$

This system can be written in a matrix form as

$$L\underline{U} = A\underline{U}_x + B\underline{U}_y + C\underline{U} = F, \tag{2.4}$$

where $\underline{U} = (\underline{u}, \omega, p)^T, \underline{F} = (f_1, f_2, 0, 0)^T$ and

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2.5}$$

We calculate $\det(\xi A + \eta B) = (\xi^2 + \eta^2)^2$, which is always positive for any non-vanished pair (ξ, η), so that the linear system is elliptic. In [6] the authors prove that if we apply piecewise quadratic polynomials for $\underline{u}$, and piecewise linear polynomials for p and ω the mesh dependent, weighted LSFEM achieves optimal rates of convergence for all of the 4 unknowns in H_1 *and* L_2 norms.

We require some spaces on Ω and Γ. The standard notations of the Sobolev spaces and their associated norms will be employed throughout this paper. We let $H^m(\Omega)$ denote the Sobolev space of functions having square integrable derivatives of order up to m over Ω. We define the norms by $\|u\|_m^2 = (u,u)_m$. and the inner product in $H^m(\Omega)$ is defined as

$$(u,v)_m = \sum_{|\alpha|\le m} \int_\Omega \partial^\alpha u \cdot \partial^\alpha v\ . \tag{2.6}$$

We also define the space for our problem,

$$S = \{\underline{V} \in [H^1(\Omega)]^4; \quad u_1, u_2 = 0 \ \text{ on } \Gamma \text{ and } \int_\Omega p = 0\}, \tag{2.7}$$

where $\underline{V} = (u_1, u_2, \omega, p)^T = (u_1, u_2, u_3, u_4)^T$. We will use finite dimensional subspace $S^h \subset S$ of functions to approximate our solutions. The parameter h, which represents a mesh spacing, is used to indicate the approximation property of S^h. For example, if we define S^h to be the space consisting of continuous piecewise quadratic functions in Ω, the approximation property shows: For every $\underline{V} \in S \cap [H^2(\Omega)]^4$, there exists $\underline{V}^h \in S^h$ such that

$$h\|\underline{V} - \underline{V}^h\|_1 + \|\underline{V} - \underline{V}^h\|_0 \leq Ch^3\|\underline{V}\|_2, \tag{2.8}$$

where the positive constant C is independent of $\underline{V}$ and h.

We construct the least–squares quadratic functional:

$$J(\underline{V}) = \int_\Omega (L\underline{V} - \underline{F}) \cdot (L\underline{V} - \underline{F}) \qquad \text{for } \underline{V} \in S. \tag{2.9}$$

The least-squares method reads: Find $\underline{U} \in S$, such that

$$J(\underline{U}) \leq J(\underline{V}) \qquad \text{for any } \underline{V} \in S. \tag{2.10}$$

If there is $\underline{U} \in S$ which minimizes $J(\underline{V})$ for any $\underline{V} \in S$, or say $J(\underline{U} + \epsilon\underline{V}) \geq J(\underline{U})$ for any $\underline{V} \in S$ we can easily to have:

$$\int_\Omega L\underline{U} \cdot L\underline{V} = \int_\Omega \underline{F} \cdot L\underline{V} \qquad \text{for any } \underline{V} \in S. \tag{2.11}$$

Similar to (2.11), if $\underline{U}^h$ minimizes (2.9) in the space S^h, we have the corresponding finite algebraic equations

$$\int_\Omega L\underline{U}^h \cdot L\underline{V}^h = \int_\Omega \underline{F} \cdot L\underline{V}^h \qquad \text{for any } \underline{V}^h \in S^h. \tag{2.12}$$

If the basis for S^h is chosen to be the piecewise quadratic polynomial, we can see that (2.12) is equivalent to a symmetric and positive definite linear algebraic system.

3 Numerical Test Case for General LSFEM

In order to test the practical applicability of the LSFEM, we ran a numerical test case using the functional (1.1) with piecewise quadratic elements. The same test case was run using the mesh dependent weighted LSFEM of the functional (1.2) using piecewise quadratic elements for u_1 and u_2, and piecewise linear elements for ω and p. For our test case we chose the phenomenon of a solid cylinder with diameter 1 moving with constant speed 1 parallel to the wall of a channel with width 1.5. (Fig.1) The domain is defined as a rectangle with corners (3, 0.75),(-1.5, 0.75),(-1.5, -0.75) and (3, -0.75). The center of the cylinder is on the origin.

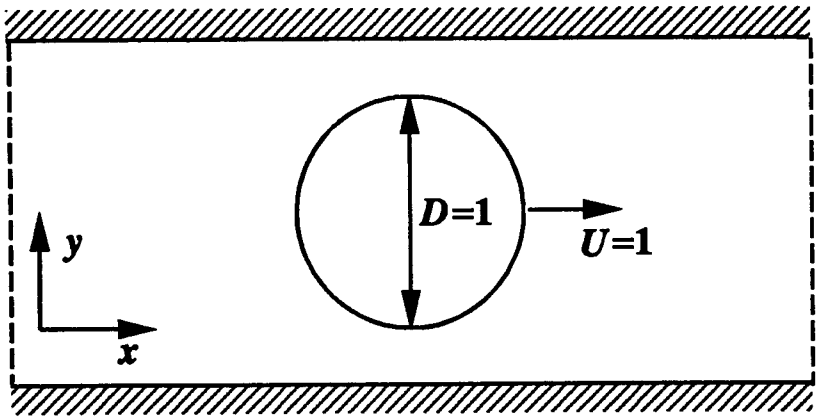


Figure 1. Problem setup.

We chose triangles as our elements. The triangle vertices and the midpoints of the triangle faces were chosen as the nodes for the quadratic basis functions. In our test case the domain was subdivided into 2,262 triangles with 3,485 faces and 4,708 nodes. There are four unknowns at each point except at the boundary points and the points on the cylinder where there are 2 unknowns since u_1 and u_2 are given. Instead of setting $\int_\Omega = 0$, we set $p = 0$ at the point (3,0). Figure 2 shows the grid which was used for the test case.

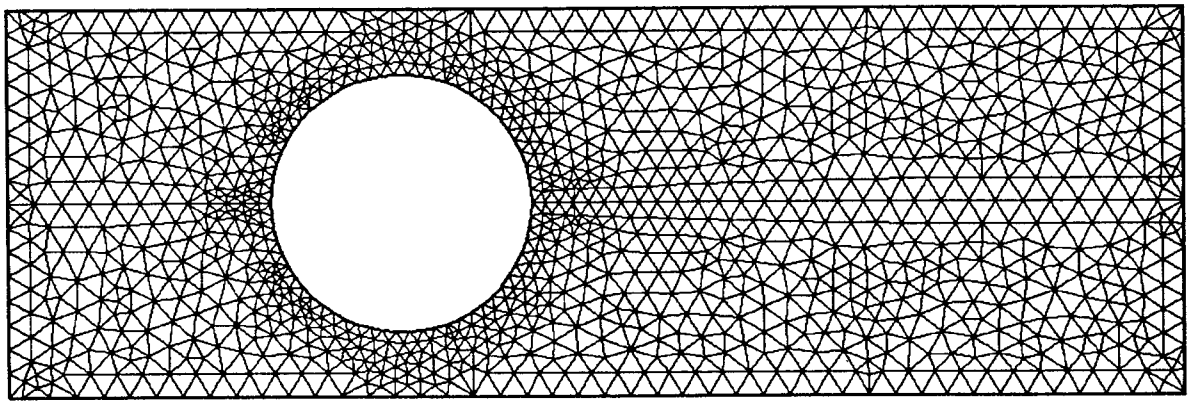


Figure 2. Grid used in test case simulation.

3.1 Test case simulation using general LSFEM

We set $u_1 = 1$ and $u_2 = 0$ at each point on the outer boundary, and set $u_1 = u_2 = 0$ on the surface of the cylinder. We also set $p = 0$ at the point (3,0) to ensure there is a unique solution for the pressure (equivalent to $\int_\Omega p = 0$). Piecewise quadratic functions were applied for all 4 unknowns at each node. Equation (1.1) was minimized in the finite element space S^h, which had 18,095 elements.

We assembled the linear system by the formulation (2.11). A sparse square matrix with dimension 18,095 $\times$ 18,095 was generated. Only the non-trivial entries of this matrix (with tolerance of 10^{-5}) were stored using upper storage by row. A double precision, smoothly converging variant of the conjugate gradient squared method was used to solve the system. Figures 3 and 4 show velocity vectors and level contours of u_1 for the calculated solution using the LSFEM.

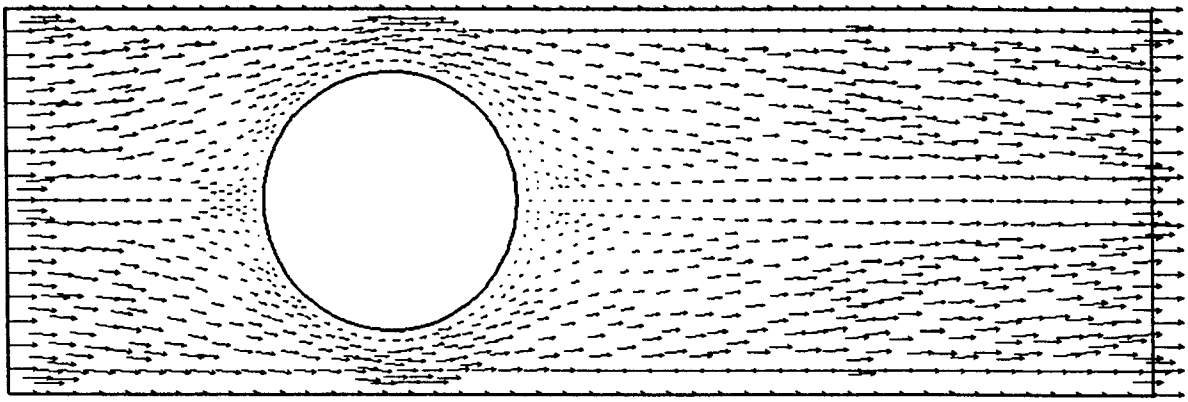


Figure 3. Velocity vectors for solution of test case using LSFEM.

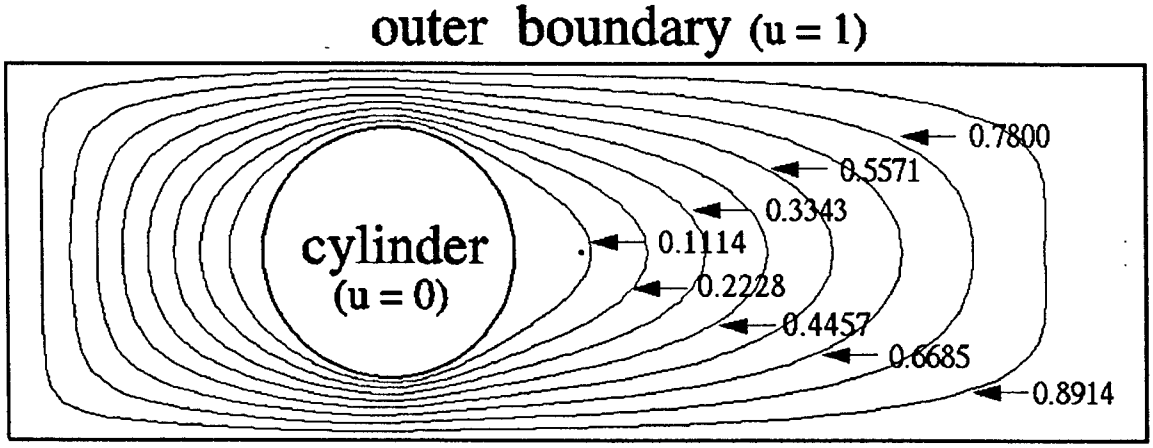


Figure 4. Level curves of u_1 for solution of test case using LSFEM.

3.2 Test case simulation using mesh dependent LSFEM

In our test case simulation with the mesh dependent LSFEM of [6], we used the space defined in [6], i.e. S^h was defined to be $u_1, u_2 \in H^1$ with piecewise quadratic polynomials in each element, and $\omega, p \in H^1$ were represented by piecewise linear polynomials is each element. The same boundary conditions and grid that were used in the simulation using the general LSFEM were used for this simulation. There were a total of 11,125 elements in the finite element space S^h.

After minimizing (1.2) with $h = 0.1$ in the space S^h, the resulting linear system was solved using the conjugate gradient method. After 4,000 iterations, the relative error $\| A\underline{x} - \underline{b} \|_0 / \| \underline{x} \|_0$ was less that 10^{-8}. Figures 5 and 6 show velocity vectors and level contours of u_1 for the calculated solution for our test case using the mesh dependent LSFEM.

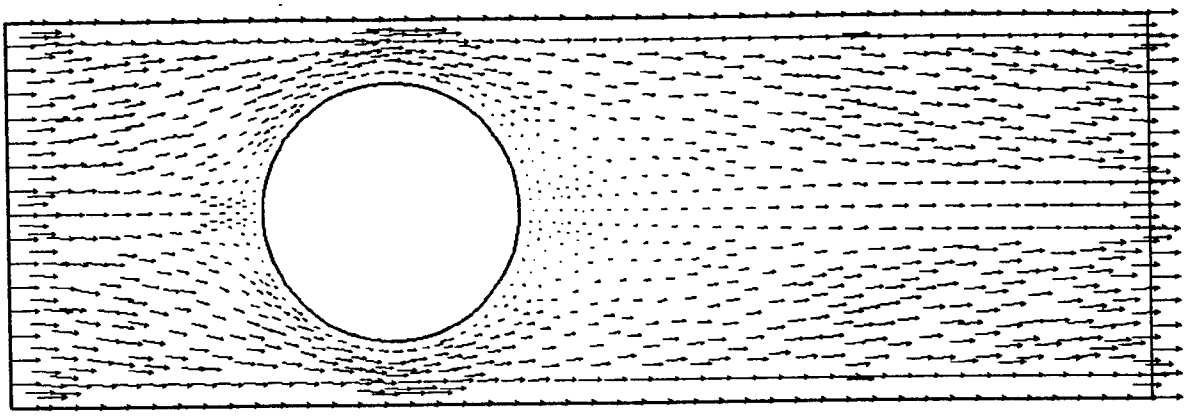


Figure 5. Velocity vectors for solution of test case using the mesh dependent LSFEM.

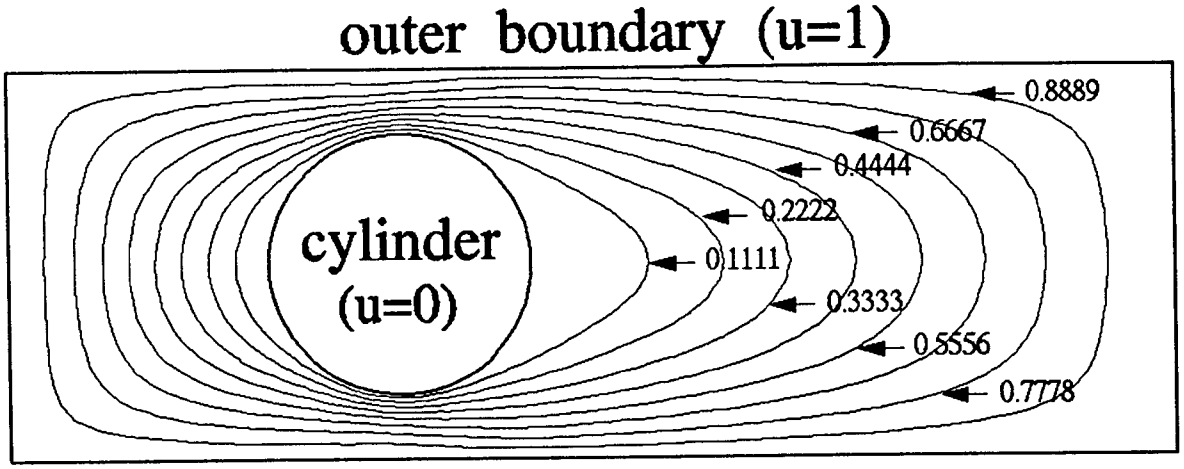


Figure 6. Level curves of u_1 for solution of test case using the mesh dependent LSFEM.

Upon inspection the numerical results for both simulations look fine. All of the dynamic equations, the vorticity relation equation and the mass conservative equation are minimized globally. But if one pays attention at some special points, for example, at the points between the solid wall and the top point of the cylinder, one finds that the value of u_1 is 1 at the solid wall and reduces steadily to 0 at the cylinder. Since the flow at the entrance, $x = -1.5$, and the outlet, $x = 3.0$ has speed 1, the average speed between (0, 0.5) and (0, 0.75) should be 3, but the numerical results from both simulations show that the maximum value of u_1 is 1.0, and the average value of u_1 along the above mentioned line is about 0.8 (see figure 9). At this region the mass conservative law has broken down totally and this region could be the very interesting part of our application. The result shows that the simple LSFEM can not be applied to similar problem directly.

4 Zero Residual for Mass Conservative Law

In [3] and [4], boundary conditions were considered as constraints and Lagrange multipliers were used to introduce the boundary conditions into the formulation of the problems they were studying. For the LSFEM, the mass conservation property is critical, so that we wish to enforce $\int_{\Omega_i}(\frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y}) = 0$ in each triangle. We will view the mass continuity equation as a constraint [3], and use the Lagrange multiplier method as a natural means of incorporating the constraint within the LSFEM statement of the problem without any special weighting. The restricted LSFEM is then:

Find $\underline{U}_e^h \in S_h$, such that

$$J(\underline{U}_e^h) \leq J(\underline{V}^h) \qquad \text{for any } \underline{V}^h \in S^h, \tag{4.1}$$

subject to the condition:

$$\int_{\Omega_i} \left(\frac{\partial u_1^h}{\partial x} + \frac{\partial u_2^h}{\partial y}\right) = 0 \qquad \text{for } i = 1, \cdots, \ell, \tag{4.2}$$

where ℓ is the total number of elements and for our test case $\ell = 2,262$ (the number of triangles). Here S_h is defined the same as was in §3.2, i.e. $u_1, u_2, \omega, p \in H^1(\Omega)$ with piecewise quadratic polynomial representations for u_1 and u_2 and piecewise linear polynomial representations for ω and p.

An equivalent formulation of this problem is: find the vector function $\underline{U}_e^h \in S^h$ which minimizes the expression

$$J_e(\underline{V}^h, \underline{\mu}) = \frac{1}{2}\int_\Omega (L\underline{V}^h - \underline{F}) \cdot (L\underline{V}^h - \underline{F}) \; + \; \underline{\mu}^T \cdot \Lambda \underline{V}^h, \tag{4.3}$$

for all $\underline{V}^h \in S^h$, and $\underline{\mu}$. Here $\underline{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T$, and $\Lambda \underline{V}^h$ is a linear vector functional defined in S^h with ℓ elements, the ith element of which represents the numerical integration of $\int_{\Omega_i}(\frac{du_1^h}{dx} + \frac{du_2^h}{dy})$ in the ith triangle.

Taking the first order variation of J_e with respect to $\underline{V}^h$ and $\underline{\mu}$ respectively, and setting $\delta J_e = 0$ leads to the weak statment: Find $\underline{U}_e^h \in S^h$ and $\underline{\lambda}$ such that

$$\int_\Omega L\underline{U}_e^h \cdot L\underline{V}^h + \underline{\lambda}^T \cdot \Lambda \underline{V}^h + \underline{\mu}^T \cdot \Lambda \underline{U}_e^h = \int_\Omega L\underline{V}^h \cdot \underline{F} \qquad \text{for any } \underline{V}^h \in S^h \text{ and any } \underline{\mu}, \tag{4.4}$$

where $\underline{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_\ell)^T$. The restricted LSFEM then has n more unknowns than the standard LSFEM for the same discretization, which for our test case means there are 11,125 + 2,262 = 13,387 unknowns using the restricted method if we let $\underline{U}_e^h$ and $\underline{V}^h$ be represented by piecewise quadratic polynomials for u_1 and u_2 and piecewise linear polynomials for ω and p.

One can check that the linear algebraic system resulting from the restricted method is symmetric, as is the case when the standard LSFEM method is used. Furthermore we can prove that this linear algebraic system is so-called pseudo-positive definite, which we explain in the following. The restricted LSFEM generates an extended linear algebraic system as:

$$A_e \begin{bmatrix} \underline{x} \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} \underline{b} \\ \underline{b}_\ell \end{bmatrix}, \tag{4.5}$$

where A is a symmetric positive definite matrix with dimension of $n \times n$, B is a matrix with dimension of $n \times \ell$, $\underline{b}$, $\underline{\lambda}$, $\underline{b}_\ell$ are vectors with dimensions of n, ℓ and ℓ respectively, and $\underline{b}_\ell$ is formed by the velocity boundary conditions $u_1 = 1$ at the outer boundary when the matrix B^T is assembled.

From the first n equations in (4.5) we have

$$A\underline{x} + B\underline{\lambda} = \underline{b}, \tag{4.6}$$

so that

$$\underline{x} = A^{-1}(\underline{b} - B\underline{\lambda}). \tag{4.7}$$

Also from the last ℓ equations we have

$$B^T\underline{x} = \underline{b}_\ell. \tag{4.8}$$

Combining the above two expressions, we obtain

$$B^T A^{-1} B\underline{\lambda} = B^T A^{-1}\underline{b} - \underline{b}_\ell. \tag{4.9}$$

We can prove the matrix of $B^T A^{-1} B$ is also symmetric and positive definite.

Lemma 1: If A is positive definite of dimension $n \times n$ and B is a matrix with dimension of $n \times \ell$, and if $B\underline{y}$ is non-trivial for any non-trivial vector $\underline{y}$ with dimension ℓ; then $B^T A^{-1} B$ is positive definite.

[Proof] For any equation $A\underline{x} = \underline{b}$, if $\underline{b}$ is a non-trivial vector, then $\underline{x}$ is also non-trivial, and vice versa. Since A is positive definite, $\underline{x}^T A\underline{x} = \underline{x}^T\underline{b} > 0$ for any non-trivial $\underline{x}$. We then have $\underline{b}^T A^{-1}\underline{b} = \underline{b}^T\underline{x} > 0$, so that A^{-1} is also positive definite.

For any given non-trivial vector $\underline{y}$ with dimension ℓ,

$$\underline{y}^T(B^T A^{-1} B)\underline{y} = (B\underline{y})^T A^{-1}(B\underline{y}) > 0. \tag{4.10}$$

Therefore the linear algebraic system of (4.5) has a unique solution if the matrix A is positive definite and if B is defined as in (4.2). □

Theorem 1: The linear algebraic system (4.5) in which A and B are defined by (4.4) and (4.2) has a unique solution.

[Proof] From the definition of matrix B in (4.4), each column represents the numerical integration of $\int_{\Omega_i} \mathrm{div}(\underline{u})$ in each triangle Ω_i, so the column vectors are linearly independent. Therefore, the rank of B is ℓ, since $\ell < n$. Then for any non-trivial vector $\underline{y}$, $B\underline{y}$ is non-trivial. Combining the above lemma, the proof follows. □

It is easy to check that the extended matrix A_e defined by (4.5) is still symmetric, but no longer positive definite. Since A is symmetric and positive definite, the solution of (4.5) is equivalent to locating the minimum of the quadratic problem

$$\min_{(\underline{\xi},\underline{\mu})} \left(\frac{1}{2}\underline{\xi}^T A\underline{\xi} + \frac{1}{2}\underline{\mu}^T B^T\underline{\xi} + \frac{1}{2}\underline{\xi}^T B\underline{\mu} - \underline{\xi}^T\underline{b} - \underline{\mu}^T\underline{b_\ell} \right), \tag{4.11}$$

where $\underline{\xi}$ and $\underline{\mu}$ are vectors of dimension $n \times \ell$. The above equation is equivalent to

$$\min \left(\frac{1}{2} \begin{bmatrix} \underline{\xi} \\ \underline{\mu} \end{bmatrix}^T A_e \begin{bmatrix} \underline{\xi} \\ \underline{\mu} \end{bmatrix} - \begin{bmatrix} \underline{\xi} \\ \underline{\mu} \end{bmatrix}^T \begin{bmatrix} \underline{b} \\ \underline{b_\ell} \end{bmatrix} \right). \tag{4.12}$$

Therefore the common conjugate gradient method can still be used to solve the system (4.5). Figures 7 and 8 show the solution of our test case simulation using the restricted LSFEM.

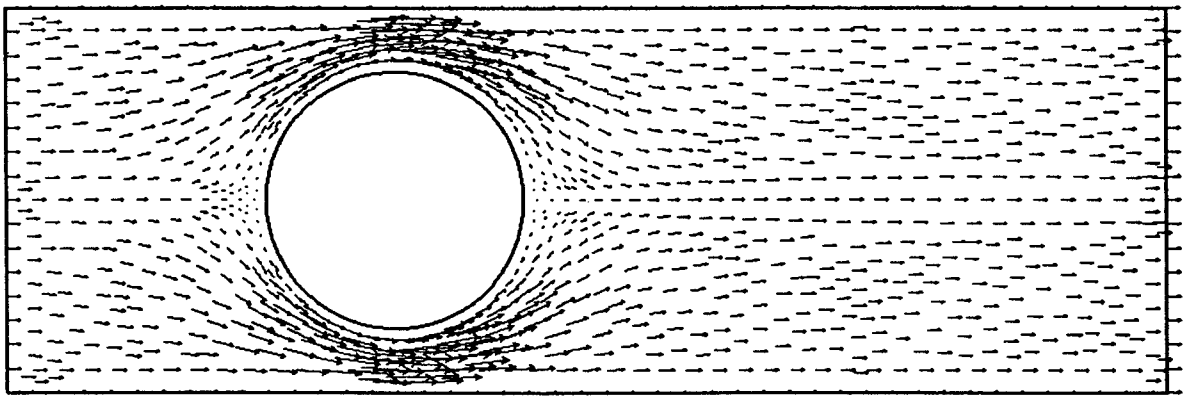


Figure 7. Velocity vectors for solution of test case using restricted LSFEM.

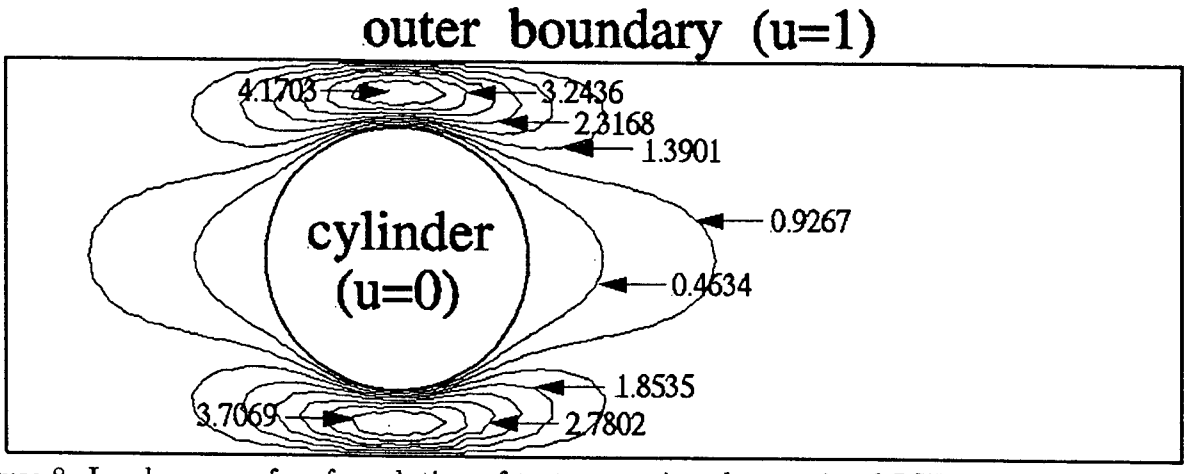


Figure 8. Level curves of u_1 for solution of test case using the restricted LSFEM.

The numerical results after solving this linear system show that the divergence of the velocity $\int_{\Omega_i}(\frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y})$ is less than 10^{-4} in each triangle. In checking the mass conservation in the region we examined earlier, we draw a line from (0, 0.5) and (0, 0.75). The average velocity normal to the line is 2.96. Figure 9 shows profiles of u_1 calculated along the above mentioned line using the mesh dependent weighted LSFEM and the restricted LSFEM.

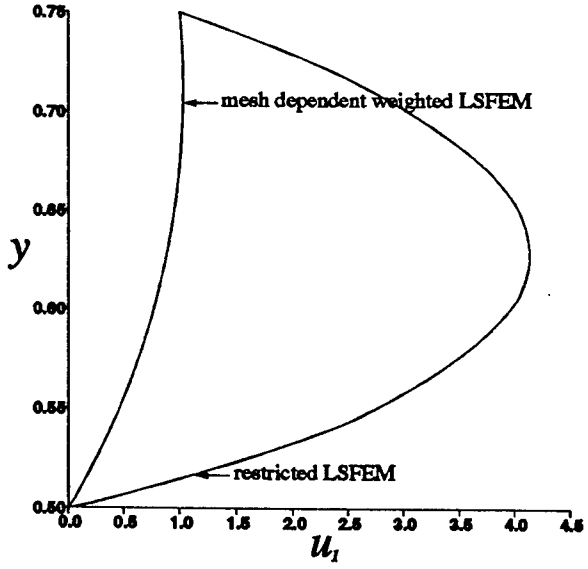


Figure 9. Profile of u_1 along a vertical line connecting the top of the cylinder and the nearest wall using the mesh dependent weighted LSFEM and the restricted LSFEM.

The equation $\int_{\Omega_i}(\frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y}) = 0$, for $i = 1, \cdots, \ell$ does not mean the flow is divergence free at each point but at the center of each triangle. As we mentioned in §3.1, there are 11,125 degrees of freedom totally for the general LSFEM. After adding the divergence free restriction at each center of each triangle, the number of degrees of freedom will reduce, but the reduced freedom will not influence the global results since a divergence free solution at each point is the result as $h \to 0$.

References

[1] S. Agmon, A. Douglis, and L. Nirenberg, "Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions II, Comm. Pure Appl. Math., 17, 35-92, 1964

[2] A. K. Aziz, R. B. Kellogg and A. B. Stephens, "Least squares methods for elliptic systems", Math. of Computation, Vol. 44, No. 169, 53–70, Jan. 1985.

[3] P. B. Bochev and M. D. Gunzburger, "Accuracy of least-squares methods for the Navier-Stokes Equations",NASA Technical Memorandum 106209, ICOMP-93-19,1993.

[4] P.B. Bochev and M. D. Gunzburger, "Analysis of Least-Squares finite element methods for the Stokes Equations ", Math. of Computation. to appear.

[5] J. H. Bramble and A. H. Schatz, "Least squares for 2mth order elliptic boundary–value problems", Math. of Computation. Vol. 25, 1–32, 1971.

[6] C. L. Chang, "An error estimate of the least squares finite element method for the Stokes problem in three dimensions", Math. Comp., Vol. 63, No. 207, 41–50, 1994.

[7] C. L. Chang, "A mixed finite element method for Stokes problem: acceleration–pressure formulation", Appl. Math. and Computation, Vol. 36, 135–146, 1990.

[8] C. L. Chang, "Finite element approximation for Grid-Div type systems in the plane", SIAM Numer. Anal., Vol. 29, No. 2, 452-461, 1992.

[9] C. L. Chang and B. N. Jiang, "An error analysis of least–squares finite element method of velocity–pressure–vorticity formulation for Stokes problem", Comput. Methods Appl. Mech. Engrg., Vol. 84, 247–255, 1990.

[10] P. Ciarlet, "The finite element method for elliptic problems", North–Holland, 1977.

[11] G. J. Fix, M. D. Gunzburger and R. A. Nicolaides, "On finite element methods of the least squares type", Comput. Math. Appl., Vol. 5, 87–98, 1979.

[12] V. Girault and P. A. Raviart, "Finite element methods for Navier–Stokes equations" Springer–Verlag, Berlin, 1986.

[13] B. N. Jiang and C. L.Chang, "Least–squares finite elements for Stokes problem", Comput. Methods Appl. Mech. Engrg, Vol. 78, 297–311, 1990.

[14] B. N. Jiang, and L. Povinelli, " Least-squares finite element method for fluid dynamics", Comput. Meth. Appl. Mech. Engrg., 81, 13-37, 1990.

[15] J. T. Oden and G. F. Carey, "Finite Elements", Prentice–Hall,Inc., Englewood Cliffs, NJ, 1984.

[16] D. Lefebvre, J. Peraire, and K. Morgen, " Least-squares finite element solution of compressible and incompressible flows, Int. J. Num. Meth. Heat Fluid Flow 2, 99-113, 1992.

[17] L. Tang and T. Tsang, "A least-squares finite element method for time-dependent incompressible flows with thermal convection", to appear in Int. J. Numer. Meth. Fluids.

[18] W. L. Wendland, "Elliptic system in the plane", Prentice–Hall,Inc., Englewood Cliffs, NJ, 1984. London, 1979.

Part 2: Least–Squares Finite Element Method for the Stokes Problem with Multiple Fluids

In this part of the report, the simulation of incompressible flow with multiple fluids in 2-D by the least-squares finite element method (LSFEM) in the stress-velocity-pressure version is studied. Unlike the Galerkin finite element method, all the conditions for continuity of stress must be explicitly added to the method. This is accomplished by viewing the equations for continuity of stress as restrictions to the standard LSFEM. These restrictions are added to the standard method by use of the Lagrange multiplier strategy. In this numerical method the inf-sup (LBB) condition is not necessary and the matrix resulting from applying the method on a discretization is symmetric, the uniqueness of the solution and the application of the conjugate gradient method is also valid. The method is used to simulate a test case flows.

1 Restricted LSFEM for Flows with Multiple Fluids

In part 1, it was found that the LSFEM in the vorticity-velocity-pressure formulation did not conserve mass at all points of the domain in a test case simulation. In order to make the method conservative, extra conditions were explicitly added to the linear system using Lagrange multipliers which ensured the flow was divergence free in each element of the domain. In part 2 a restricted LSFEM for flows with multiple fluids is developed by viewing the conditions for continuity of stress as constraints and using Lagrange multipliers as a natural means of incorporating these constraints into the LSFEM statement of the problem.

In the ensuing discussion, we assume the computational domain Ω is comprised of two subdomains Ω_1 and Ω_2 which have a common boundary Γ_i. In physical terms, one fluid is contained in Ω_1, a second fluid is contained in Ω_2 and Γ_i represents the interface. The equations for continuity of stress are

$$2(t_1 n_1 - t_2 n_2)\,[\mu\phi_1] + (t_1 n_2 + t_2 n_1)\,[\mu\,(\phi_2 + \phi_3)] = 0 \qquad \text{on } \Gamma_i \tag{1.1}$$

$$[p] - 2(n_1^2 - n_2^2)\,[\mu\phi_1] - 2n_1 n_2\,[\mu\,(\phi_2 + \phi_3)] = \frac{T}{R} \qquad \text{on } \Gamma_i, \tag{1.2}$$

where T is a nondimensional surface tension parameter (Webber number), R is the radius of curvature, $\phi_1 = \frac{\partial u}{\partial x}$, $\phi_2 = \frac{\partial u}{\partial y}$, $\phi_3 = \frac{\partial v}{\partial x}$ and p is the pressure. Here n_1 and n_2 are the x and y components of $\underline{n}$, where $\underline{n}$ is the unit normal on the interface between two fluids contained in the discretized domain Ω. Also t_1 and t_2 are the x and y components of the tangent vector to the interface, which is oriented so that $\underline{t}$, $\underline{n}$, $\underline{t} \times \underline{n}$ forms a right handed system.

Equations (1.1) and (1.2) are considered as constraints which are applied at all points in the discretized domain in S^h which lie on the interface. These constraints are applied to the standard LSFEM in the stress-velocity-pressure formulation [1]. Four variables are allowed to be discontinuous across the interface, but the conditions for continuity of stress only provide two restrictions. Thus kinematical conditions are used to provide two more restrictions which are applied at the

same points the stress condition restrictions are applied. Since the velocities are continuous across the interface, it can be written that

$$t_1\,[\phi_1] + t_2\,[\phi_2] = 0 \tag{1.3}$$

$$t_1\,[\phi_3] - t_2\,[\phi_1] = 0. \tag{1.4}$$

The Lagrange multiplier method is used to add the above restrictions to the linear system which results from the standard LSFEM. The restricted LSFEM statement of the problem then is: find the vector function $\underline{U_e^h} \in S^h$ which minimizes the expression

$$\begin{aligned} J_e(\underline{V}^h, \underline{\mu}_1, \underline{\mu}_2, \underline{\mu}_3, \underline{\mu}_4) &= \frac{1}{2}\int_\Omega \Big((L\underline{V}^h - \underline{F})\cdot(L\underline{V}^h - \underline{F})\Big) + \underline{\mu}_1^T \cdot \Delta\underline{V}^h + \underline{\mu}_2^T \cdot \Delta\underline{V}^h \\ &+ \underline{\mu}_3^T \cdot \Delta\underline{V}^h + \underline{\mu}_4^T \cdot \Delta\underline{V}^h, \end{aligned} \tag{1.5}$$

for all $\underline{V}^h \in S^h$, and $\underline{\mu}_1$,$\underline{\mu}_2$,$\underline{\mu}_3$ and $\underline{\mu}_4$. Here the $\underline{\mu}_i = (\mu_{i_1}, \mu_{i_2}, \cdots, \mu_{i_n})^T$, where n is the number of nodes in Ω. The $\Delta\underline{V}_i^h$ are linear vector functional defined in S^h with ℓ elements, ℓ being the number of nodes for each of the restricted variables with lie on Γ_i in each domain Ω_1 and Ω_2. These vector functionals represents the application of the restrictions (1.1), (1.2), (1.3) and (1.4) to all points in the descretized domain which lie on Γ_i.

Taking the first order variation of J_e with respect to $\underline{V}^h$, $\underline{\mu}_1$, $\underline{\mu}_2$, $\underline{\mu}_3$ and $\underline{\mu}_4$ respectively, and setting $\delta J_e = 0$ leads to the weak statement: find $\underline{U_e^h} \in S^h$ and $\underline{\lambda}_1$, $\underline{\lambda}_2$, $\underline{\lambda}_3$, $\underline{\lambda}_4$ such that

$$\begin{aligned} &\int_\Omega \Big(L\underline{U_e^h}\cdot L\underline{V}^h\Big) + \underline{\lambda}_1^T \cdot \Delta\underline{V}^h + \underline{\mu}_1^T \cdot \Delta\underline{U_e^h} + \underline{\lambda}_2^T \cdot \Delta\underline{V}^h + \underline{\mu}_2^T \cdot \Delta\underline{U_e^h} + \underline{\lambda}_3^T \cdot \Delta\underline{V}^h \\ &+\underline{\mu}_3^T \cdot \Delta\underline{U_e^h} + \underline{\lambda}_4^T \cdot \Delta\underline{V}^h + \underline{\mu}_4^T \cdot \Delta\underline{U_e^h} = \int_\Omega L\underline{V}^h \cdot \underline{F} \qquad \forall \underline{V}^h \in S^h, \forall \underline{\mu}_1, \underline{\mu}_2, \underline{\mu}_3, \underline{\mu}_4, \end{aligned} \tag{1.6}$$

where $\underline{\lambda}_i = (\lambda_{i_1}, \lambda_{i_2}, \cdots, \lambda_{i_n})^T$. The restricted LSFEM has 4ℓ more unknowns that the standard LSFEM for the same discretization.

The linear system resulting from the restricted method is symmetric. Furthermore we can prove that this linear algebraic system is so-called pseudo-positive definite, which we explain in the following. The restricted LSFEM generates an extended linear algebraic system as

$$A_e \begin{bmatrix} \underline{x} \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} \underline{b} \\ \underline{b}_\ell \end{bmatrix}, \tag{1.7}$$

where $\underline{\lambda} = (\underline{\lambda}_1, \underline{\lambda}_2, \underline{\lambda}_3, \underline{\lambda}_4)^T$. Here A is a symmetric positive definite matrix with dimension $n \times n$, B is a matrix with dimension of $n \times 4\ell$, and $\underline{b}$, $\underline{\lambda}$, $\underline{b}_\ell$ are vectors with dimensions of n, 4ℓ and 4ℓ respectively. The vector $\underline{b}_\ell$ is formed by the velocity boundary conditions when the matrix B^T is assembled.

From (1.7) it can be stated that

$$\underline{x} = A^{-1}(\underline{b} - B\underline{\lambda}),\ B^T\underline{x} = \underline{b}_\ell. \tag{1.8}$$

Combining the these two expressions leads to the following:

$$B^T A^{-1} B\underline{\lambda} = B^T A^{-1}\underline{b} - \underline{b}_\ell. \tag{1.9}$$

In part 1 is was proved that the matrix $B^T A^{-1} B$ is symmetric and positive definite.

Theorem 1: The linear algebraic system (1.7) in which A and B are defined by (1.6) has a unique solution.

[Proof] From the definition of matrix B in (1.6), the columns of B represent the application of four linearly independent restrictions ((1.1), (1.2), (1.3) and (1.4)) at the points of the discretized domain which lie on Γ_i. Since each restriction is applied at each point only once, the column vectors of B are linearly independent. Therefore, the rank of B is 4ℓ. Then for any non-trivial vector $\underline{y}$, $B\underline{y}$ is nontrivial. Combine this with the fact that $B^T A^{-1} B$ is symmetric and positive definite and the proof follows. □

Since A is symmetric and positive definite, the solution of (1.7) is equivalent to locating the minimum of the quadratic problem

$$\min_{(\underline{\xi},\underline{\mu})} \left(\frac{1}{2}\underline{\xi}^T A \underline{\xi} + \frac{1}{2}\underline{\mu}^T B^T \underline{\xi} + \frac{1}{2}\underline{\xi}^T B \underline{\mu} - \underline{\xi}^T \underline{b} - \underline{\mu}^T \underline{b_\ell} \right), \tag{1.10}$$

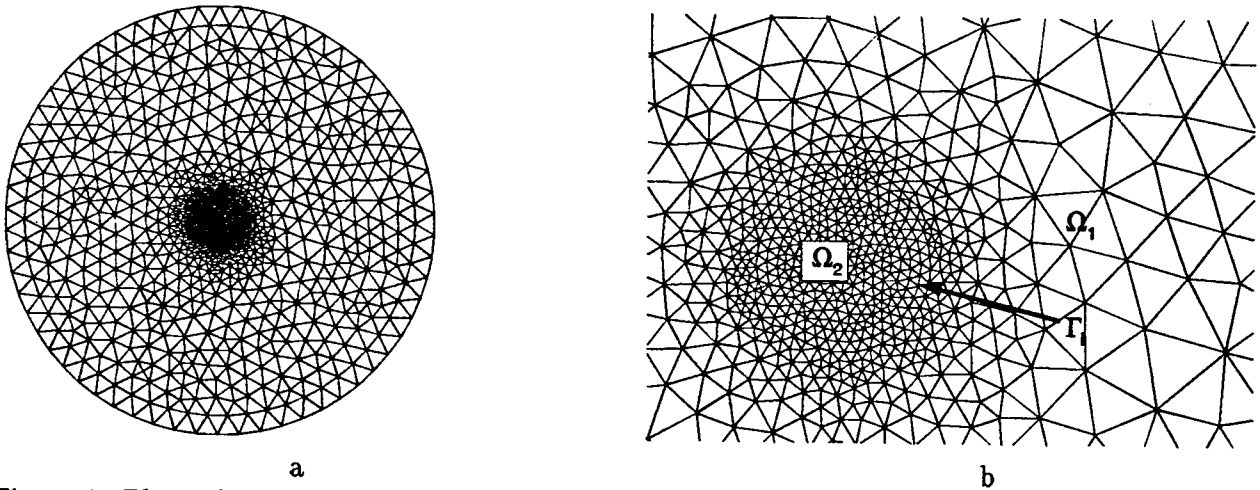
where $\underline{\xi}$ and $\underline{\mu}$ are vectors of dimension $n \times 4\ell$. The above equation is equivalent to

$$\min \left(\frac{1}{2} \begin{bmatrix} \underline{\xi} \\ \underline{\mu} \end{bmatrix}^T A_e \begin{bmatrix} \underline{\xi} \\ \underline{\mu} \end{bmatrix} - \begin{bmatrix} \underline{\xi} \\ \underline{\mu} \end{bmatrix}^T \begin{bmatrix} \underline{b} \\ \underline{b_\ell} \end{bmatrix} \right). \tag{1.11}$$

Therefore the common conjugate gradient method can still be used to solve the system (1.7).

2 Numerical Example

The restricted LSFEM developed above was used to simulate the Hele-Shaw flow of a two-dimensional bubble of fluid falling in a slightly less dense fluid under the action of gravity. In this flow the flow in the lighter fluid should resemble the potential flow around a sphere, while the flow inside the bubble should consist of two counter-rotating vorticies. The discretized domain used in this simulation is shown in figure 1.

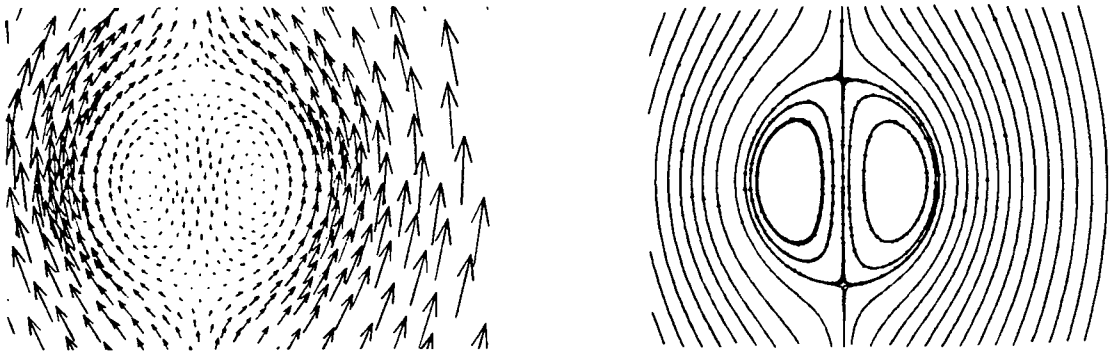


a b

Figure 1. Plots of discretized domain used in simulation of two fluid Hele-Shaw flow. (a) Plot of whole domain. (b) Detailed plot showing Ω_1, Ω_2 and Γ_i.

The domain Ω_1 is assumed to be a circular shape with diameter of 1 and center at the point $x = 0$, $y = 0$. The domain Ω_2 is also assumed to be a circular shape with diameter of 10 and center at the point $x = 0$, $y = 0$ with an annulus of diameter 1 cut out of the middle. The viscosity and density of the fluid in Ω_2 are both assumed to be twice as large as the viscosity and density of the

fluid in Ω_1. The simulated solution exhibits the flow characteristics stated above as shown in figure 2.



a b

Figure 2. Plots of simulated solution of two fluid Hele-Shaw flow. (a) Velocity vector plot. (b) Streamline plot.

References

[1] C. L. Chang, S. Y. Yang and J. S. Hsu,, "A least-squares finite element method for incompressible flow in stress-velocity-pressure version", submitted to Comput. Methods Appl. Mech. Engrg.

A NEW SUPERPOSITION

David B. Choate
Associate Professor
Mathematics Program

Transylvania University
Lexington, Kentucky
40508

Final Report for
Summer Faculty Research Program
Wright Patterson AFB

Sponsored by
Air Force Office of Scientific Research
Dayton, Ohio

and

Wright Patterson AFB

August, 1994

A NEW SUPERPOSITION

David Choate
Associate Professor
Department of Mathematics
Transylvania University

Abstract

A channel's fading can be modeled as the product of a slowly varying component and the transmitted signal. An amplitude-modulated signal is also represented by a product of a carrier signal and envelope function. In these systems homomorphic signal processing for multiplication can be used to give impressive results. Superposition is a generalized principle of homomorphic signal processing.

The logarithmic function will transform a system modeled on a product to a conventional linear system that will yield to a classical attack. It is shown here that the logarithm, as a generalized superposition, will also transform a conventional linear system into another linear system and therefore nothing need be known about the original system before applying a logarithmic transformation.

I. Subaddition

Definition 1. Let $S = \{z = x + yj \mid -\pi < y \leq \pi\}$ or equivalently, after an appropriate adjustment of the residue of y, $S = \{z = x + y(\mod 2\pi)j \mid x \text{ and } y \text{ real}\}$, a horizontal strip.

Definition 2. Let $C^* = S \cup \{-\infty\}$

Definition 3. Let $z = x_1 + y_1 j$ and $w = x_2 + y_2 j \in C^*$.

$$z \oplus w = (x_1 + x_2) + [(y_1 + y_2) \mod (2\pi)]j$$

Definition 4. If $z = re^{j\theta}$, then define $\ln(z) = \ln|r| + [\theta(\mod(2\pi)]j$ as usual.

Definition 5. Let $z, w \in C^*$. Then we define new operation called *subaddition*, denoted by $\sqcap$, the southeast corner of addition, by

$$z \sqcap w = \ln(e^z + e^w) .$$

Note 1. Clearly $\ln(z + w) = \ln z \sqcap \ln w$.

We intend to show that $(C^*, \oplus, \sqcap)$ is a field that is isomorphic to $(C, \cdot, +)$, the complex field under multiplication and addition.

II. The Field $(C^*, \oplus, \sqcap)$

Lemma 1. C* is closed under $\oplus$ and $\sqcap$.

Proof. If $z, w \in S$, then the Lemma is immediate from definitions 3,4 & 5. If $z = -\infty$ and $w \in C^*$, then $-\infty \oplus w = -\infty$ and $-\infty \sqcap w = w$.

Lemma 2. The operations $\oplus$ and $\sqcap$ are commutative.

Proof. Let $z = x_1 + y_1 j$ and $w = x_2 + y_2 j$ be elements of C^*. Then

$$\begin{aligned} z \oplus w &= (x_1 + x_2) + (y_1 + y_2) \pmod{2\pi} j \\ &= (x_2 + x_1) + (y_2 + y_1) \pmod{2\pi} j \\ &= w \oplus z . \end{aligned}$$

$$z \sqcap w = \ln(e^z + e^w) = \ln(e^w + e^z) = w \sqcap z$$

Lemma 3. The operations $\oplus$ and $\sqcap$ are associative.

Proof. It is clear that the operation $\oplus$ is associative since both ordinary addition and modular addition are associative.

To show that $\sqcap$ is associative let u, v and w $\in$ C*. Then

$$u \sqcap (v \sqcap w) = u \sqcap [\ln(e^v + e^w]$$

$$= \ln(e^u) \sqcap \ln(e^v + e^w)$$

{since u $\in$ C* and by p.76 [1]}

$$= \ln[e^u + (e^v + e^w)]$$

$$= \ln[(e^u + e^v) + e^w]$$

$$= [\ln(e^u + e^v)] \sqcap \ln(e^w)$$

$$= [\ln(e^u) \sqcap \ln(e^v)] \sqcap \ln(e^w) \quad \text{Note 1}$$

$$= (u \sqcap v) \sqcap w \quad \text{since } u,v,w \in C^* .$$

Lemma 4. The operation $\oplus$ distributes over $\sqcap$.

Proof. Let $u, v, w \in C^*$.

Then $u \oplus (v \sqcap w) = \ln(e^u) \oplus \ln(e^v + e^w)$

$= \ln[e^u(e^v + e^w)]$

$= \ln(e^{u+v} + e^{u+w})$

$= (u \oplus v) \sqcap (u \oplus w)$

Lemma 5. The $\sqcap$ identity is $-\infty$.

Proof. $z \sqcap -\infty = \ln(e^z + e^{-\infty}) = z$.

Lemma 6. If $z \in S$, then the inverse of z under $\sqcap$ is $z \oplus \pi j$.

Proof. $z \sqcap (z \oplus \pi j) = \ln(e^z + e^{z+\pi j})$

$= \ln[e^z(1 + e^{\pi j})]$

$= \ln(0)$

$= -\infty$.

Lemma 7. $C^*/\{-\infty\} = S$ is a group under $\oplus$.

Proof. Lemmas 1, 2, 3 and 6 .

Note 2. Clearly $-\infty$ has no $\oplus$ inverse in C . This is analogous to 0's having no multiplicative inverse in C. And the equation $-\infty \oplus z = -\infty$ is the "*" equivalent to (0)z = 0 in C .*

All of the above proves

Theorem 1. (C*, $\oplus$, Γ) is a field.

We can now prove that this field is isomorphic to the field of complex numbers.

Theorem 2. (C, $\circ$, +) $\cong$ (C*, $\oplus$, Γ) .

Proof. Define φ: C -> C* by $\varphi(z) = \ln(z)$.

$$\varphi(z_1 z_2) = \ln(z_1 z_2)$$
$$= \ln(z_1) \oplus \ln(z_2)$$
$$= \varphi(z_1) \oplus \varphi(z_2) .$$
$$\varphi(z_1 + z_2) = \ln(z_1 + z_2)$$
$$= \ln(z_1) \Gamma \ln(z_2) \qquad \text{Note 1}$$
$$= \varphi(z_1) \Gamma \varphi(z_2)$$

We now show φ is onto.

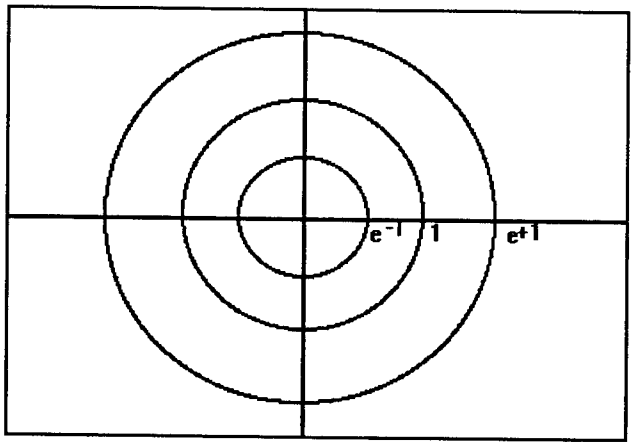
If $w \in C^*$, then $w \in S$ or $w = -\infty$. If $w \in S$, then e^w is the preimage of φ. If $w = -\infty$, then 0 is the preimage of w.

We now show φ is one-to-one.

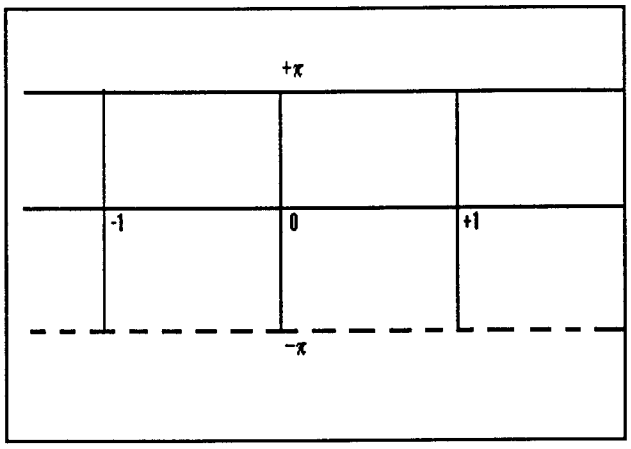
If $\ln(z) = \varphi(z) = 0_{C^*} = -\infty$, then $z = 0$.

III. <u>The Complex Cylinder</u>

A simple geometrical interpretation of Theorem 2 can be given as follows. Map the complex plane

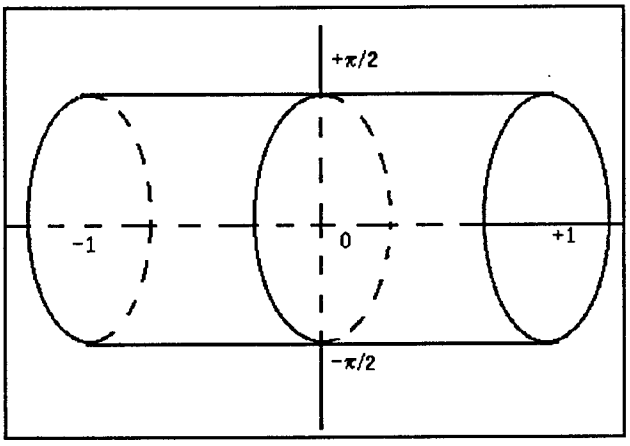


onto the horizontal strip



by $f(z = re^{j\theta}) = \ln|r| + [\theta \pmod{2\pi}]j$.

The three circles in the complex plane with radii e^{-1}, 1 and e are mapped into the three vertical lines in the figure above. Since the top and lower line of the horizontal strip have been identified in the congruence class, we really have a cylinder.



If we look down the right end of the cylinder, then the circle with center at $+1 = \ln(e^{+1})$ appears largest, the circle with center at $-1 = \ln(e^{-1})$ appears smallest, and the end of the cylinder, $-\infty$, is a dot.

This is exactly what we saw in the original complex plane.

IV. Linear to linear superposition

On p. 481 of [2] we have a definition of *generalized superposition*:

$$H[x_1(n) \boxdot x_2(n)] = H[x_1(n)] \bigcirc H[x_2(n)] \qquad (10.2a)$$

$$H[c : x(n)] = c \lrcorner H[x(n)] \qquad (10.2b)$$

Define $H: C \to C^*$ by $H(z) = \ln(z)$.

If we let $\boxdot$ be $+$, ordinary addition in C

$\bigcirc$ be $\sqsubset$, or subaddition in C*

$:$ be scalar multiplication in C

and $\lrcorner$ be a scalar operation in C*

defined by $c \lrcorner H[x] = \ln(c) \oplus H(x)$

, then we have a generalized superposition H (where H stands for *homomorphism*.)

But we have more than that. We know that the homomorphic system can be written as a cascade of three systems by p.482 of [2].

We have this since Γ satisfies the conditions he gives:

(i.) Γ is commutative and associative by Lemma 3. {See (10.3) and (10.4) on p. 482 of [2].}

(ii.) The C* under Γ is a vector space over the field C with scalar multiplication $\lfloor$ as we will see.

To establish (ii) we must show that C* under Γ is a vector space over the field C. We will have done so if we establish the following four properties for every $\alpha \in C$ and every $v, w \in C^*$

1. $\alpha \lfloor (v \Gamma w) = (\alpha \lfloor v) \Gamma (\alpha \lfloor w)$

Proof. $\alpha \lfloor (v \Gamma w) = \ln(\alpha) \oplus (v \Gamma w)$ def. of $\lfloor$

$= [\ln(\alpha) \oplus v] \Gamma [\ln(\alpha) \oplus w]$ Lemma 4

$= (\alpha \lfloor v) \Gamma (\alpha \lfloor w)$ def. of $\lfloor$

2. $(\alpha \oplus \beta) \backslash v = (\alpha \backslash v) \oplus (\beta \backslash v)$

Proof. $(\alpha \oplus \beta) \backslash v = \ln(\alpha + \beta) \oplus v$ def. of $\backslash$

$= [\ln(\alpha) \sqcap \ln(\beta)] \oplus v$ Note 1

$= [\ln(\alpha) \oplus v] \sqcap [\ln(\beta) \oplus v]$ Lemma 4

$= (\alpha \backslash v) \sqcap (\beta \backslash v)$ def. of $\backslash$

3. $\alpha \backslash (\beta \backslash v) = (\alpha\beta) \backslash v$

Proof. $\alpha \backslash (\beta \backslash v) = \ln(\alpha) \oplus [\ln(\beta) \oplus v]$ def. of $\backslash$

$= \ln(\alpha\beta) \oplus v$

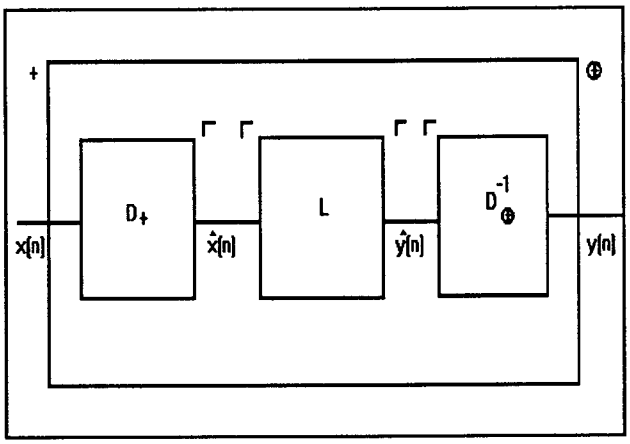
$= (\alpha\beta) \backslash v$

4. $1 \backslash v = v$

Proof. $1 \backslash v = \ln(1) \oplus v$ def. of $\backslash$

$= v$

By p. 482 of [2] we know that since the system imputs constitute a vector space of complex numbers under addition and ordinary scalar multiplication and the system outputs constitute a vector space under Γ , the subaddition, and $\wr$, the scalar multiplication, then all systems of this class can be represented as a cascade of three systems



The effect of system D_+ is to transform the combination of signals $x_1(n) + x_2(n)$ into another convention linear system under subaddition of corresponding signals $D_+[x_1(n)]$ and $D_+[x_2(n)]$.

4. The Complex Pencil

We will briefly mention that the complex cylinder is by no means the end of it.

Definition 6. Attach a new element $\ln(-\infty)$ to $C^* = C^{1*} = C \cup \{-\infty\}$

to get $C^{2*} = C^{1*} \cup \{\ln(-\infty)\}$.

Definition 7. Let $a, b \in C^{2*}$. Then define a new operation $\square$ on

C^{2*} by $a \square b = \ln\{\ln[\exp(e^a) + \exp(e^b)]\}$.

Note 3. Closure of the operation $\square$ on C^{2*} can be guaranteed by a proper suturing of the logarithm of the strip S at the top of page 10-9. See chapter 6 below.

Lemma 8. i. $a \square b = \ln(e^a \sqcap e^b)$ ***and***

ii. $\ln(e^a \sqcap e^b) = \ln(a) \square \ln(b)$.

Proof. Equation (8i) follows directly from the definition of $\sqcap$.
Equation (8ii) is obtained by replacing a and b with $\ln(a)$ and $\ln(b)$.

Lemma 9. The operation $\square$ is commutative.

Proof.

$a \square b = \ln(e^a \sqcap e^b)$		Lemma 8i
$= \ln(e^b \sqcap e^a)$		Lemma 2
$= b \square a$		Lemma 8i

Lemma 10. The operation $\square$ is associative.

Proof. $a \square (b \square c) = a \square \ln(e^b \oplus e^c)$ Lemma 8i

$= \ln[e^a \oplus (e^b \oplus e^c)]$ Lemma 8i

$= \ln[(e^a \oplus e^b) \oplus e^c]$ Lemma 3

$= [\ln(e^a \oplus e^b)] \square c$ Lemma 8ii

$= (a \square b) \square c$ Lemma 8i

Lemma 11. The operation $\oplus$ is distributive over $\square$.

Proof. $a \oplus (b \square c) = \ln(e^a + e^{b \square c})$ Definition 5

$= \ln[e^a + (e^b \oplus e^c)]$ Lemma 8i

$= \ln[(e^a + e^b) \oplus (e^a + e^c)]$ Lemma 4

$= \ln\{[e^{(a \oplus b)}] \oplus [e^{(a \oplus c)}]\}$ Note 2

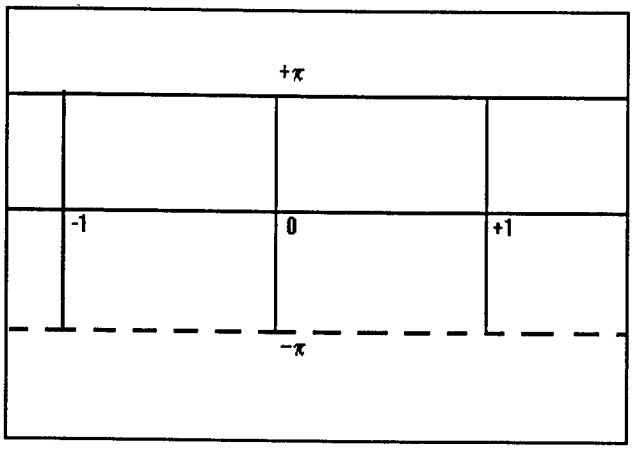
$= (a \oplus b) \square (a \oplus c)$ Lemma 8ii

Theorem 3. (C**, ⌐, □) is a field that is isomorphic to (C, ∘, +).

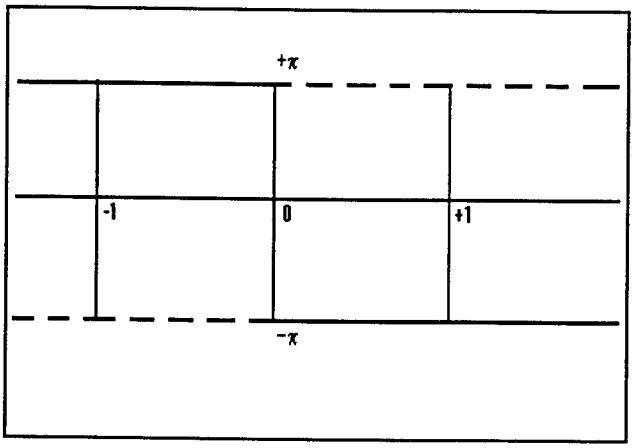
Sketch of proof. The map ϕ:C -> C** defined by

$\phi(z) = \ln[\ln(z)]$ is an isomorphism.

Even without attending the equivalence classes too closely, we can still get a good intuitive idea of the C** surface by just examining the strip S first shown on the top of p. 10-9.



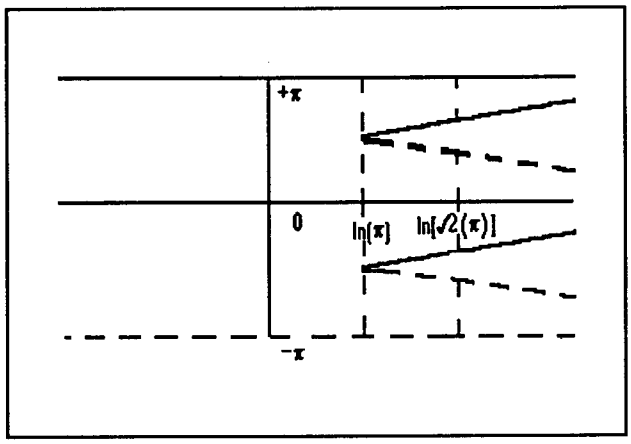
After adjusting the residues we can obtain a slightly altered strip S'.



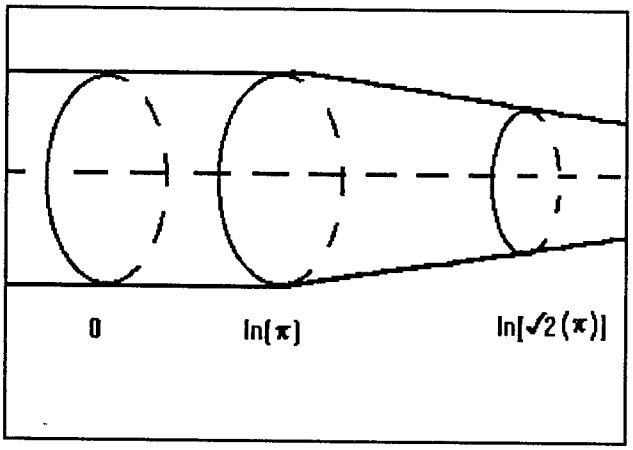
The strip S' represents the logarithm of the complex plane. The logarithm of S' is then image of the isomorphism defined in Theorem 3.

A circle lying in the strip S' and having center at the origin may have a radius no greater than π. And the logarithm will conformally map a solid circle of radius π into an half infinite strip extending from $-\infty$ to $\ln(\pi)$.

But a circle with center at the origin and radius $\pi/2$ will *not* be completely contained in the strip if its central angle lies in $(\pi/4, 3\pi/4]$ or in $(-3\pi/4, -\pi/4]$. It clear now that the image of our isomorphism has the form



In order to insure the closure of the new operation $\square$ we must identify equivalence classes and fold up the figure above to obtain the Complex Pencil.



5. Aftermath

If we continue this process, then we will have generated an infinite number of superpositions formed by a repeated application of the logarithm. As we have seen, a second application of the log will shred the strip S. A third will shred the shredded strip S.

In order to insure the closure of the necessarily new "addition" operation, we must, after an adjustment of residues, suture the shredded S back into a new surface that is, after attaching an identity, isomorphic to the complex plane.

References

1. Churchill, R.V., and J. W. Brown: "Complex Variables and Applications," 5th ed., McGraw-Hill Publishing Company, 1990.

2. Oppenheim, A.V., and R. W. Schafer: "Digital Signal Processing," Prentice-Hall, Inc, 1975.

SYNTHESIS OF NOVEL SECOND AND THIRD ORDER NONLINEAR OPTICAL MATERIALS

Stephen J. Clarson
Associate Professor
Department of Materials Science and Engineering
497 Rhodes Hall
University of Cincinnati
OH 45221-0012
RDL SRP Faculty Associate 94-0138

Lawrence L. Brott
Graduate Student
Department of Materials Science and Engineering
497 Rhodes Hall
University of Cincinnati
OH 45221-0012

Final Report for:
Summer Faculty Research Program
WL/MLBP
Wright Patterson AFB

September 1994

SYNTHESIS OF NOVEL SECOND AND THIRD ORDER NONLINEAR OPTICAL MATERIALS

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

ABSTRACT

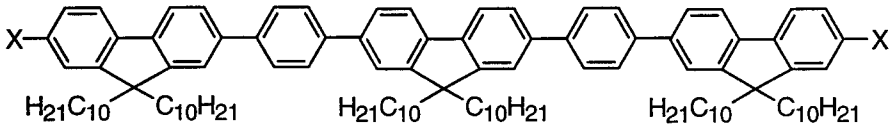
Polyparaphenylene based systems have interesting electrooptical properties but major solubility problems and hence require side chains to assist in subsequent characterization or use. In this work, we have developed both second and third order nonlinear optical materials based on incorporating fluorene FL groups into the backbone of the desired materials along with paraphenyl groups. Here the bridging carbon was alkylated to ensure the solubility of the resulting materials in common organic solvents. The symmetric A-FL-A type systems were designed to have novel third order properties, whereas the non-symmetrically substituted systems A-FL-B (in this case A being a thiophene group and B being a pyridine group) have interesting second order properties.

SYNTHESIS OF NOVEL SECOND AND THIRD ORDER NONLINEAR OPTICAL MATERIALS

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

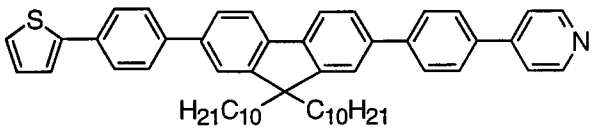
INTRODUCTION

Third order nonlinear optical materials typically are compounds consisting of extended π-electron conjugation and multiple aromatic rings [1]. Likewise, second order NLO compounds have multiple double bonds between electron donor and acceptor groups. This research involves the design and synthesis of new fluorene-containing para-polyphenylene compounds for a monomer (**1**) and chromophore (**2**).



X = Br, CO_2H

1 Fluorene-containing Monomer



2 Fluorene-containing Chromophore

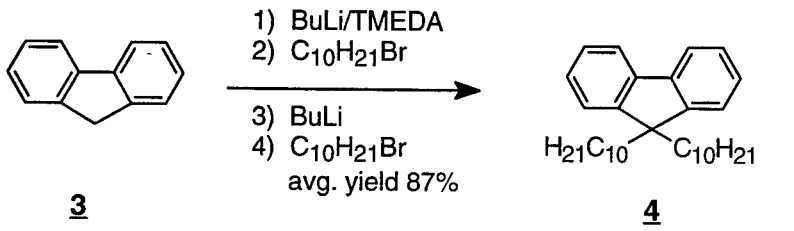
EXPERIMENTAL DETAILS

COMMON SYNTHESES FOR THE MONOMER AND CHROMOPHORE

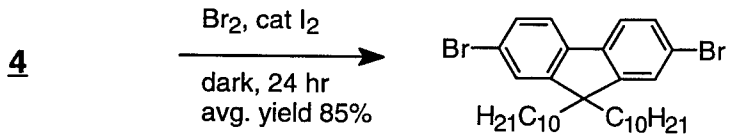
Both substances are similar and their syntheses share several steps. The central building block of each product is the fluorene molecule. To make it more soluble and easier to handle,

two long alkyl chains were added to the C-9 position of the fluorene. This approach allowed the C-2 and C-7 carbons to still be sterically unhindered and reactive. As seen in Scheme I, fluorene was first treated with BuLi, complexed with TMEDA [2,3], and then treated with bromodecane to obtain monoalkylated fluorene. The monoalkylated fluorene, without isolation and purification, was further reacted with a second equivalent of BuLi and bromodecane to obtain dialkyl compound **4**. The product was purified by column chromatography followed by distillation under reduced pressure to remove any residual bromodecane.

SCHEME I



3 **4**

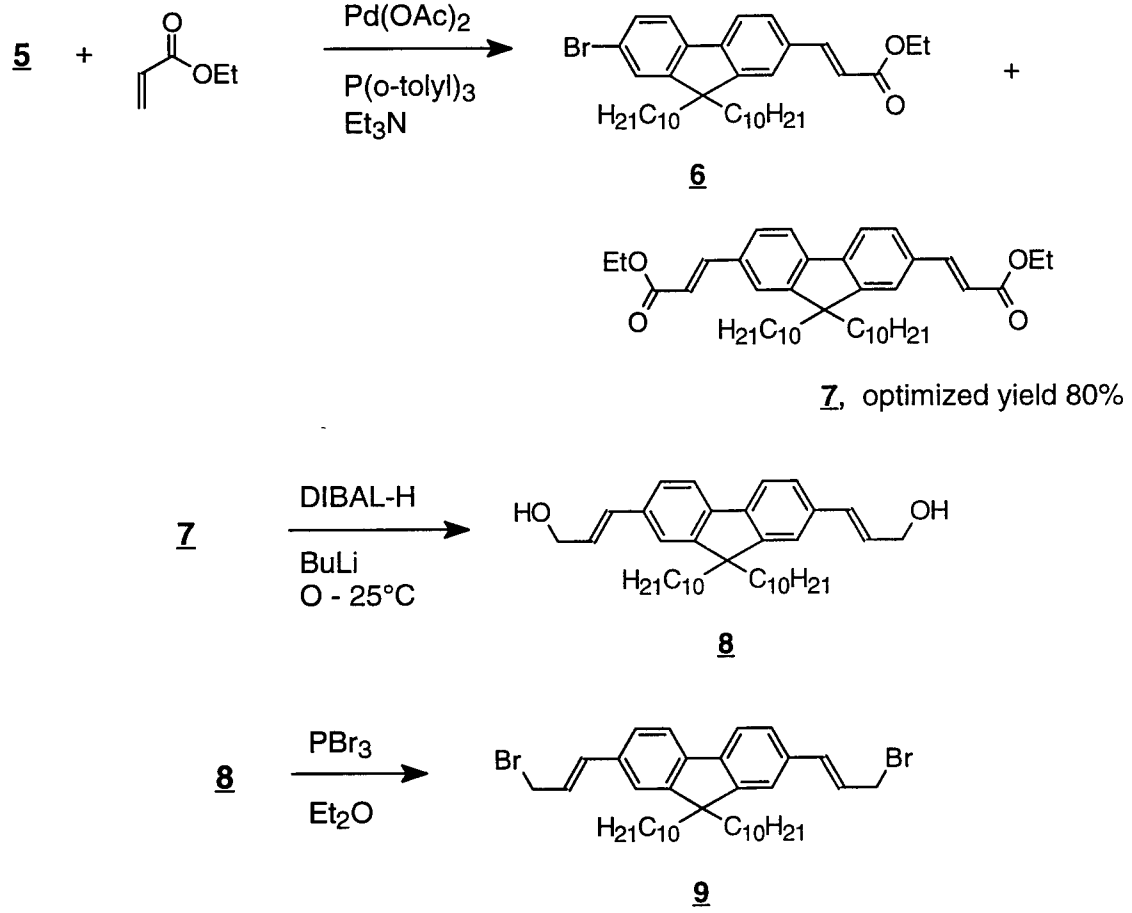


4 **5**

The next key step was to synthesize 2,7-dibromofluorene **5**. Compound **4** was reacted with bromine in the presence of a small amount of iodine and the complete absence of light. The dibromofluorene **5** was then purified by column chromatography.

The dibromofluorene was further reacted with ethylacrylate in the presence of $Pd(OAc)_2$ as a catalyst to produce a diacrylate as shown in Scheme II. Compound **7** was purified by recrystallization in ethanol. The diacrylate was then selectively reduced to the bis(allyl alcohol) **8** by using diisobutylaluminum hydride (DIBAL-H) and BuLi in an inert atmosphere [4]. The alcohol was purified by column chromatography with a yield of 63%. Finally, the diol **8** was converted to the corresponding dibromo compound **9** by treatment with PBr_3 at 0°C [5]. Unfortunately, the oil product was found to be too reactive to purify by chromatography and therefore was used without further purification.

SCHEME II



ADDITION OF PHENYL GROUPS THROUGH RING CLOSURE

The synthesis of polyphenylene compounds **1** and **2** was completed by Stevens rearrangement / thermal cyclization reactions as shown in Scheme III. Ar^2 represents the dialkylfluorene **4** in both cases of the monomer **1** and chromophore **2**, while the Ar^1 and Ar^3 can portray one of three different compounds. When making the monomer, the aromatic groups Ar^1 and Ar^3 are brominated dialkylfluorenes, but when making a chromophore, one uses a thiophene ring as Ar^1 and a pyridine as Ar^3 (see Table 1).

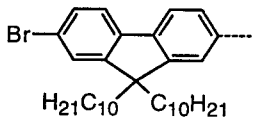
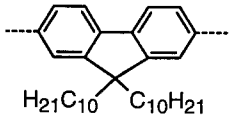
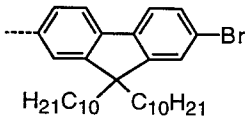
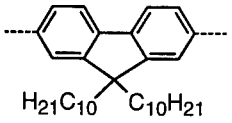
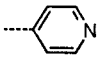
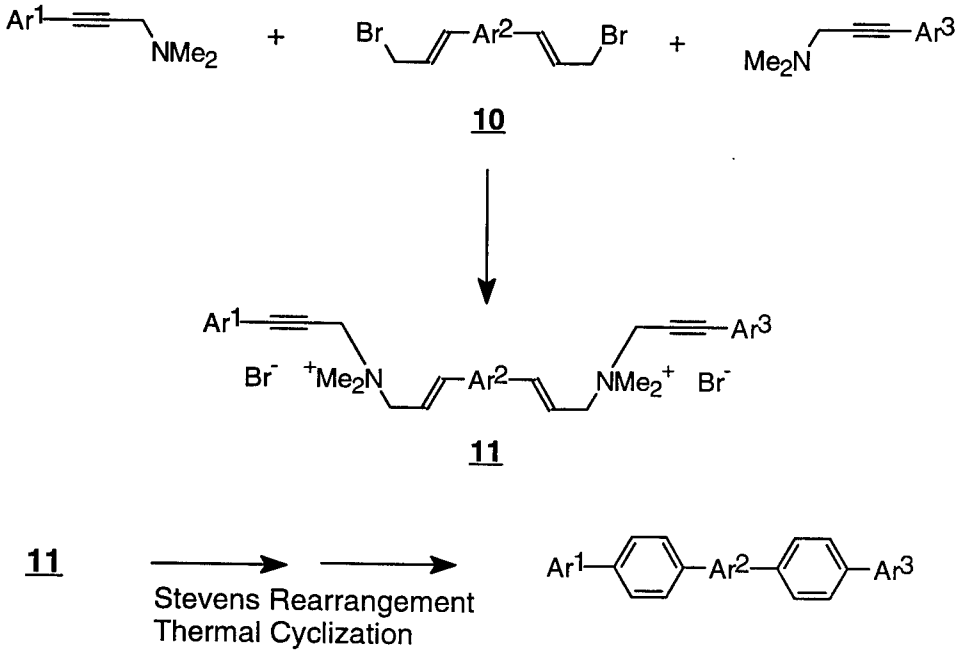
Compound Number	Ar^1	Ar^2	Ar^3
1			
2			

Table 1. Aromatic groups used in Stevens rearrangement.

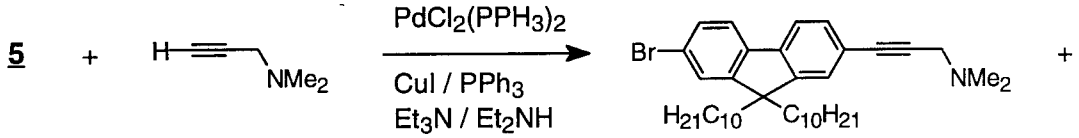
SCHEME III



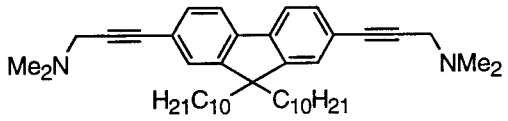
MONOMER PREPARATION

It is now necessary to synthesize a monopropargyl amine used for the preparation of the monomer. The amine is derived from the dibromo compound **5** using $PdCl_2(PPh_3)_2$ and CuI as catalysts (see Scheme IV). Due to the difficulty in suppressing the reaction to form the bis(propargyl amine) **13**, only moderate yields of the desired product were obtained. In addition, purification was difficult since the amine tends to "smear" along the whole chromatography column.

SCHEME IV



12, avg. yield 40%

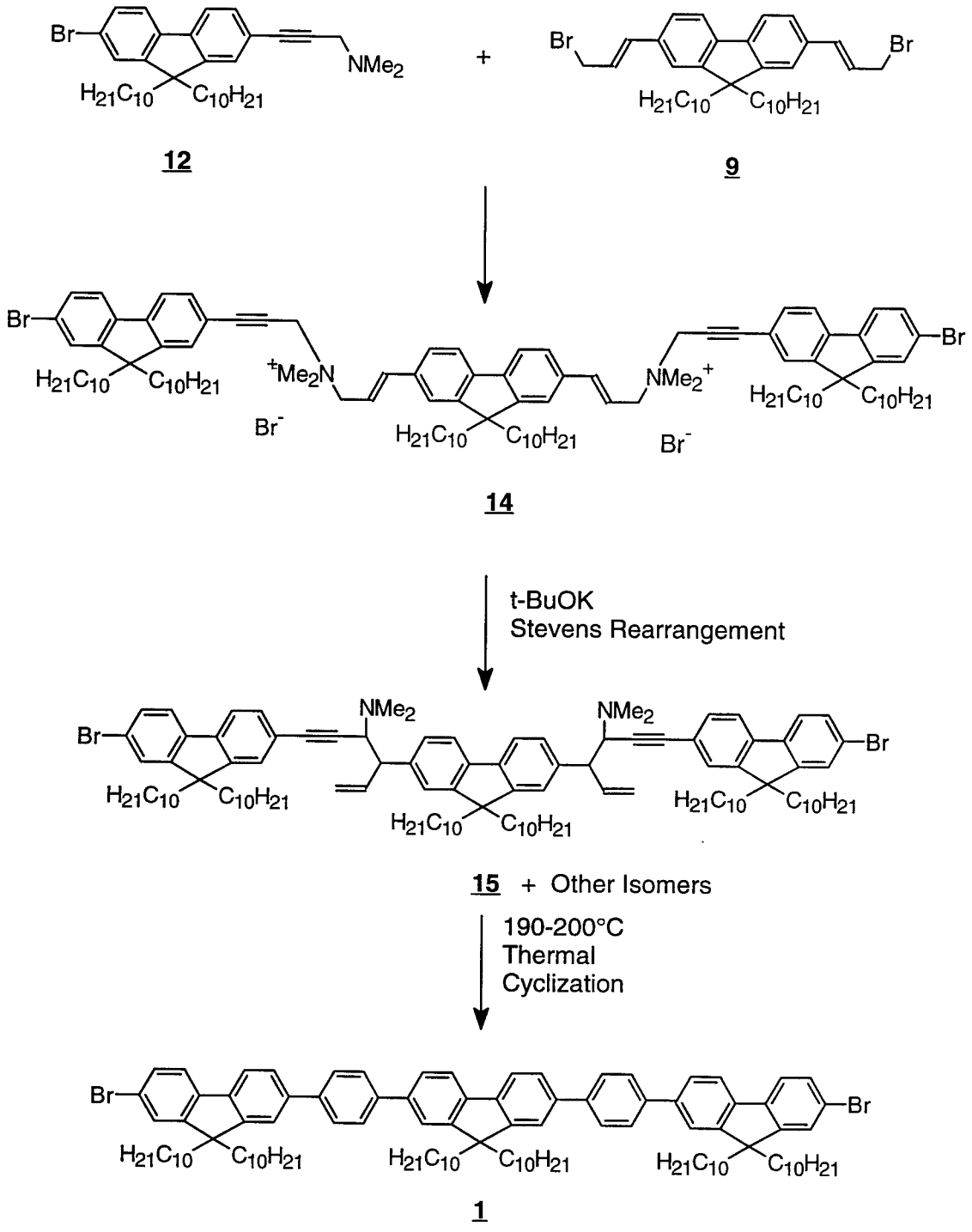


13, avg. yield 20%

Once completed, all that is needed to complete the monomer synthesis is to combine the bis(allyl bromide) **9** with amine **12**, as shown in Scheme V. After dissolving compound **9** in chloroform, the solution was cooled to 0°C with an ice bath. Compound **12** was then slowly added. The solution was allowed to stir for 24 hours and then rotavaped. The salt **14** was recovered by precipitating in hexane.

The product **14** was then reacted with t-BuOK at room temperature for one day in toluene. Once completed, the solvent was removed, and without further purification, the resulting dark viscous oil was heated to 190°C for an hour. The final product **1** was then purified by chromatography.

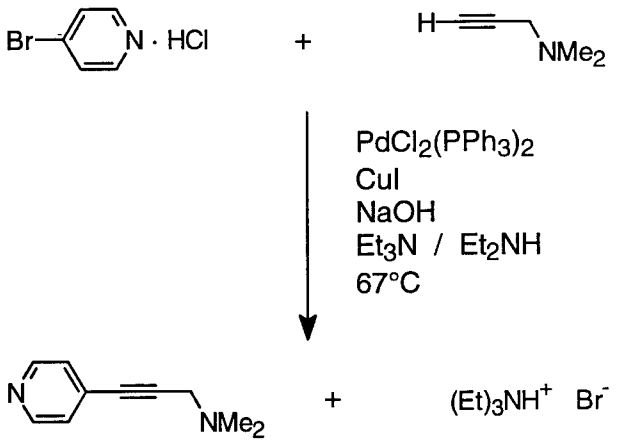
SCHEME V


Br
NMe2
H21C10
C10H21
12
+
Br
Br
H21C10
C10H21
9
Br
+Me2N
Br-
H21C10
C10H21
NMe2+
Br-
Br
H21C10
C10H21
14
t-BuOK
Stevens Rearrangement
NMe2
NMe2
Br
Br
H21C10
C10H21
H21C10
C10H21
H21C10
C10H21
15 + Other Isomers
190-200°C
Thermal
Cyclization
Br
Br
H21C10
C10H21
H21C10
C10H21
H21C10
C10H21
1

CHROMOPHORE PREPARATION

The chromophore uses two different propargyl amine compounds to surround the brominated fluorene **9** - one for the electron donor group, and one as the acceptor. In this research, a pyridine ring will act as the acceptor while a thiophene ring will be the donor. Consequently, two new products must be synthesized. Scheme VI describes the preparation of the pyridine compound.

SCHEME VI



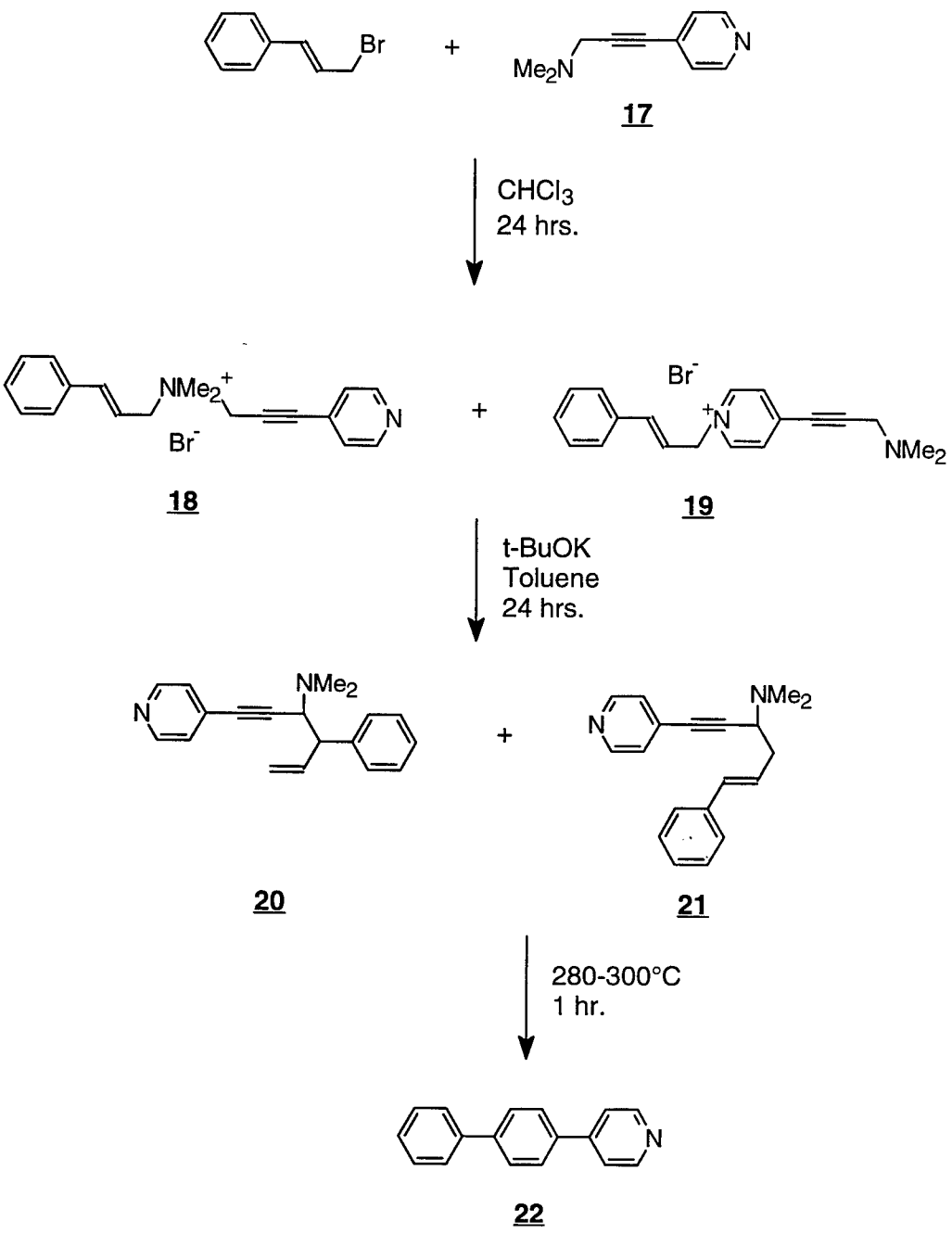
17

Although the 4-bromopyridine hydrochloride did not dissolve well in the solvent, once the catalyst and amine were added, the solution turned green and everything appeared to dissolve well as the reaction progressed. Upon heating, the mixture turned black. The product was purified by distillation under high vacuum.

A model reaction of the proposed Stevens rearrangement was carried out next using cinnamyl bromide (instead of compound **9**) and compound **17**. Originally the two materials were added at room temperature; enough heat was released to boil some of the chloroform and the solution quickly turned black. It is thought that the heat liberated was enough to activate the formation of the undesired product **19**.

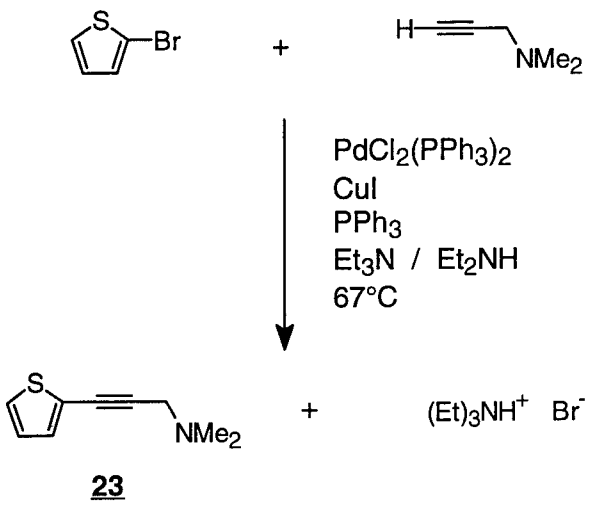
Since the procedure results in a black viscous oil, the product could not be recrystallized in any solvent. Therefore the Stevens rearrangement was carried out with the impure salt in toluene. The desired product **20** was purified by column chromatography. The final step involved the thermal cyclization in which **20** was heated in a sand bath at 280°C for an hour. The product **22** was purified by recrystallization in hexane.

SCHEME VII



The thiophene / amine (**23**) preparation is nearly identical to the pyridine work-up except that NaOH is no longer necessary to neutralize the acid. The product can be purified either by chromatography, or more simply, distillation under high vacuum.

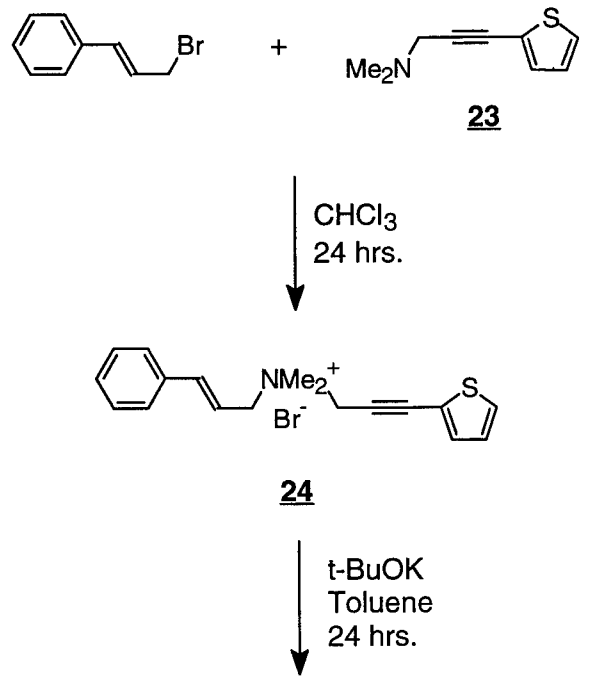
SCHEME VIII



23

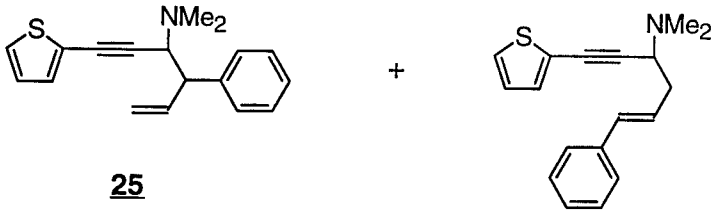
The work-up for the model thiophene ring-closure reaction is similar to the pyridine, although it was found to be cleaner. The reaction is described in Scheme IX. This time only one salt formed **24** which was recrystallizable in THF. The product of the Stevens rearrangement **25** was purified by chromatography. Once the ring-closure reaction was completed, the product **26** was recrystallized in hexane.

SCHEME IX

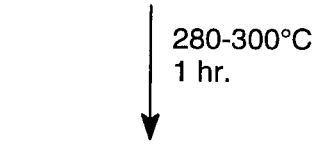


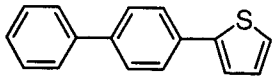
23

24



25

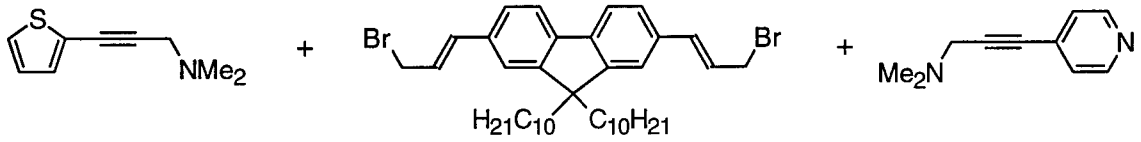




26

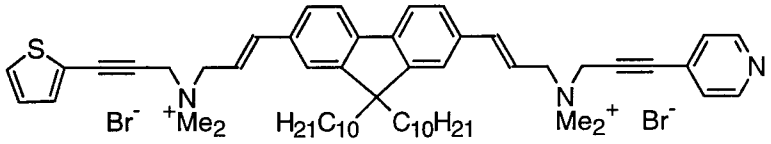
Compounds **23**, **9**, and **17** have been combined as shown in Scheme X. As the end groups were slowly dripped into a chilled brominated-fluorene solution **9**, the mixture turned a clear yellow-orange color. However once the ice bath was removed, the solution turned black. It is proposed that the reaction should be kept at 0°C for the entire 24 hours. Since the mixture became a black tar once the solvent was removed, the salt could not be isolated. Therefore the viscous oil was dissolved in toluene and t-BuOK directly added. To be sure the reaction had gone to completion, the solution was allowed to stir for 48 hours. The toluene was removed, and the resulting tar was heated to 180°C for 10 minutes. The products were then separated by column chromatography.

SCHEME X



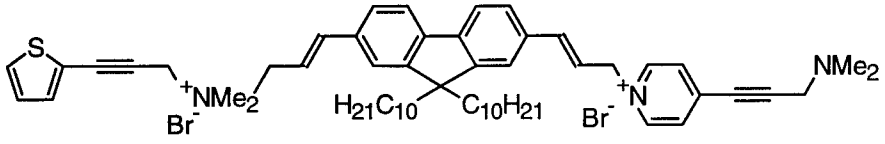
23 **9** **17**

$CHCl_3$
O°C - rt



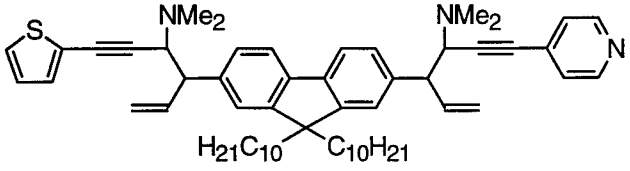
27

+



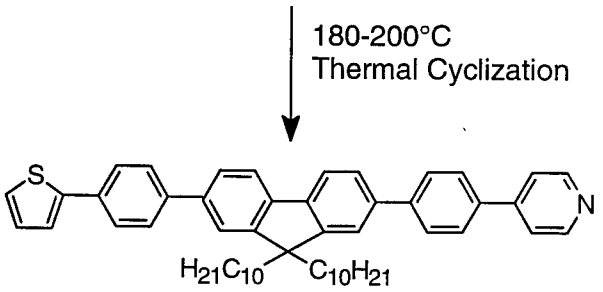
28

t-BuOK
Toluene



+ Other Isomers

29



2

CONCLUDING REMARKS

The synthetic methods to yield both novel second and third order NLO materials containing fluorene based groups have been successfully developed. Under a University of Cincinnati / WL/WPAFB educational partnership agreement, the synthesis, purification and characterization of these compounds will be carried out. Dr. Steve Clarson, Mr. Lawrence Brott and Ms. Lora Cintavey (RDL graduate student 94-0402) [6] will continue these successful investigations on NLO materials both at the University and WL/WPAFB under this joint agreement. Second and third order NLO measurements of the materials will be carried out in the coming months.

ACKNOWLEDGMENTS

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

We would like to thank RDL for financial support to Dr. Steve Clarson and the University of Cincinnati for support to Mr. Lawrence L. Brott.

BIBLIOGRAPHY

1. Prasad, P.N. and Williams, D. J., *Introduction to Nonlinear Optical Effects in Molecules and Polymers*, Wiley-Interscience, New York, 1991.
2. Yasuda, H.; Walczak, M.; Rhine, W. and Stucky, G., *J. Organomett. Chem.*, **1975**, 90, 123-31.
3. Zerger, R.; Rhine, W. and Stucky, G., *J. Am. Chem. Soc.*, **1974**, 96, 5441-8.
4. Kim, S. and Ahn, K.H., *J. Org. Chem.*, **1984**, 49, 1717-24.
5. Unroe, M. and Reinhardt, B., *Synthesis*, **1987**, 11, 981-986.
6. Cintavey, L. A., *Processing and Characterization of Nonlinear Optical PBZT Films*, RDL Summer Graduate Student 94-0402, Final Report 1994.

THE SENSOR MANAGER PUZZLE

Milton L. Cone
Assistant 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 Program
Wright Laboratory

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

and

Wright Laboratory

August 1994

THE SENSOR MANAGER PUZZLE

Milton L. Cone
Assistant Professor
Department of Computer Science/Electrical Engineering
Embry-Riddle Aeronautical University

Abstract

The task of a sensor manager is to improve the performance of the individual avionics sensors by coordinating their activities based on the sensor manager's best estimate of the future. This paper reviews planning and scheduling literature to identify developments that might apply to the design of a sensor manager. Applications examined primarily come from the planning and scheduling of manufacturing plants. An interpretation of the sensor manager as a manufacturing plant is included. Most of the reviewed work is from the artificial intelligence community.

THE SENSOR MANAGER PUZZLE

Milton L. Cone

Introduction

There is a myth of a superhuman pilot, one that can fly a complex fighter airplane the first day like a veteran, drop bombs on a dime, fly mach 1 at 50 feet through mountains, manage the electronic surveillance measure, radar and infrared sensors and not forget to pick up milk on the way home. Such a pilot never existed and probably never will. In a modern fighter aircraft that is engaged in air combat there are too many demands, many of them conflicting, on the pilot's time. A sensor manager is one way of reducing the workload.

What is a sensor manager? Literally the sensor manager manages the sensors on board the aircraft. The Data Fusion Group of the Joint Directors of Laboratories (Waltz and Llinas (20)) describe the sensor manager as one part of a data fusion subsystem in the avionics system of a modern fighter aircraft. By combining data from many sensors, a data fusion subsystem tries to derive more information about the environment than can be gathered from the individual sensor's outputs. Figure 1 from Musick and

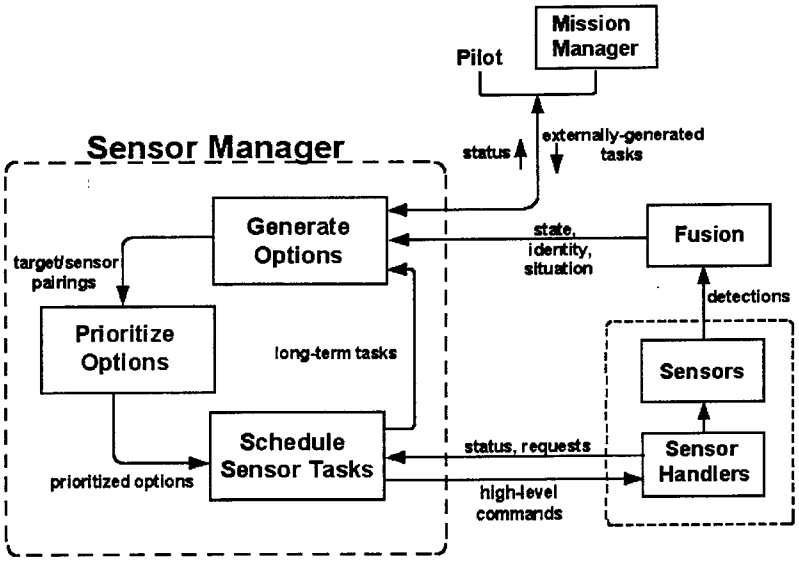


Figure 1. Sensor manager task flow

Malhotra (12) shows a diagram of a typical sensor manager system embedded in an avionics system composed of a mission manager and a fusion subsystem. Musick and Malhotra list several functions that a sensor manager should have. These include:

- Generate options (sensor/task pairings) for action
- Prioritize options
- Formulate sensor schedule to execute desired actions

- Communicate desired actions to sensor handlers
- Monitor sensor health and performance, respond to sensor feedback, account for sensor availability

The design of a sensor manager is similar to putting a puzzle together, see Figure 2. In this puzzle there are

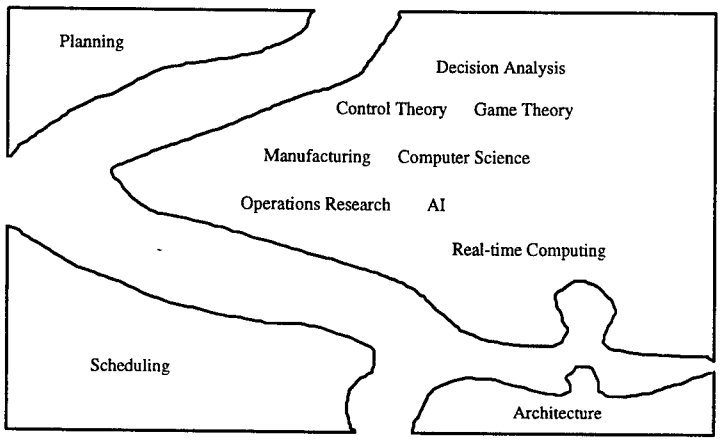


Figure 2. The sensor manager puzzle

many pieces. There is a puzzle piece called planning. Sycara (18) defines planning as the selection and sequencing of a set of actions (or plans) that can be expected to allow a system to reach one or more desired goals. Planning is a long term decision process. The planning subsystem of the sensor manager looks ahead to see if there are any combinations of sensor modes or target configurations that it can take advantage of to increase data throughput.

The next piece is scheduling. Scheduling deals with the allocation of resources to planning actions and the assignment of time intervals during which actions could be executed (Sycara (18)). Scheduling refers to the short term response of the sensors and includes the detailed temporal sensor sequencing assignments. Architecture is another puzzle piece. An architecture explains how components interact to solve complex problems. There are many architectures proposed that would be applicable to the sensor manager. Several will be explored in this paper.

The last puzzle piece lists the approaches that might help put the sensor manager puzzle together. It turns out that many disciplines have something to offer. Developing long range plans and short term schedules is a ubiquitous problem. Operations research looks at the problem as a dynamic programming program or a queuing problem. Control theory would like to recast the problem in terms that allow the body of classical and optimal control theory to be applied to solve the problem. Game theory brings a sense of one's action influencing the reaction of the opponent which in turn influences one's next action. Real-time computing tries to guarantee that tasks are completed on time so that a schedule can be made. Generally real-time

systems analysts come from a computer science background. Planning and scheduling is just one of many problems artificial intelligence (AI) has tried to solve. Manufacturing has spent many years of research trying to solve the planning and scheduling problem. Decision analysis can bring its power to bear on the planning problem, for example, using influence diagrams. The robotics literature has many examples which could guide the design of the sensor manager.

The problem for the sensor manager design is to put all of the pieces together. This paper examines four areas of research that might be important to the design of a sensor manager. These four areas are: manufacturing technology, real-time systems theory, control theory and AI/planning. In some cases the examples show considerable overlap between areas.

Manufacturing

Manufacturing has always been concerned with scheduling operations in a factory. As factories have become more automated and competition has demanded more efficient operation, computer programs were written to help develop schedules. This section starts with an overview of scheduling terminology for the manufacturing sector. Graves (7) developed a classification to help discuss models for production scheduling. He lists three classification categories which are requirements generation, processing complexity and scheduling criteria.

The first category, requirements generation, refers to the way customer orders are generated and filled. In an open shop all production orders are generated by customers. No inventory is held. In a closed shop all customer requests are filled from inventory. Here production tasks are based on inventory levels.

The second aspect of scheduling classification is processing complexity. Processing complexity refers to the number of steps involved with a production task. Graves differentiates four types of shops. They are:

- One stage, one-processor
- One stage, parallel processors
- Multistage, flow shop
- Multistage, job shop.

The one stage, one processor facility has one machine. The one machine is capable of several different operations but only one can be performed at a time. All tasks are completed on this one machine and only require one step to complete. An example is a plant with one machine that slits steel bands into narrower widths. Scheduling is a matter of prioritizing tasks and minimizing setup time to improve throughput. The one stage, parallel processors facility is like the one stage, one processor except that there are now several identical machines on which a task can be completed. The steel processing plant now has grown to include several slitters capable of producing the desired output. Multistage means that there are several machines

that must sequentially process the task before it is completed. Generally the order of processing is fixed and is called a process routing or precedence. In a flow shop all of the tasks go through the same sequence of processors. A steel processing plant operating under the flow shop model might have several slitters. The first cuts the big rolls into smaller rolls for subsequent slitters to process. A flow shop might also include other machines to complete the processing of the product. The important characteristic of the flow shop is that there is only one process routing leading to a finished product. A job shop is the most general category of manufacturing plant. There are no restrictions on processing steps and alternate process routings are allowed. A job shop steel processing plant will have many machines capable of performing different tasks. The scheduling problem is to assign the machines to get the most product through the plant subject to scheduling constraints created by customer demand.

The third way of classifying production scheduling problems is by the scheduling criteria. There are many ways to judge the success of a schedule but most either reduce product cost or improve schedule performance. Typical items that increase product costs are machine setup times, overtime costs and inventory holding cost while schedule performance measures include percentage of late tasks, time to move a task through all of the processing steps (flow time) and the number of tardy jobs.

Graves introduces two other ways that distinguish production scheduling problems although he didn't use them in the cited article. In addition to regarding requirements generation as occurring in a closed or open shop, requirements can be generated by a deterministic or stochastic process. Another way to distinguish scheduling models is to determine whether the environment in which they operate is static or dynamic. Graves groups these into two basic models, deterministic or stochastic requirements generation, and static or dynamic environment.

What type of model is most appropriate for the sensor manager? The answer to this question can simplify the literature search. In terms of open or closed shop the sensor manager is more like an open shop. While a detailed interpretation of the sensor manager in terms of a manufacturing example is given in the next section for this discussion the sensor manager can be viewed as a facility serving a number of customers. These customers include the pilot, mission manager, and fusion process. While some products can be viewed as coming from inventory, i.e. data that is stored in a data base, most requests result in new sensor assignments.

The multiple sensors, each of which possess multiple modes, compose the plant. Many of these modes can provide similar information for a job such as tracking or identifying a target. This allows a task to take multiple routes through the array of avionics sensors and implies that the plant can best be modeled as a multistage, job shop (usually just called a job shop). There are some cases in which a flow shop might be

appropriate as well as either single stage categories but generally the job shop classification will be most useful. Thus the sensor manager is analogous to the most complex form of production scheduler.

For scheduling criteria cost is not as important as schedule performance. In a manufacturing plant it takes some time to set up a machine for a new operation but in modern sensors, reconfiguration is comparatively efficient. What is important is that the sensor taskings be completed on time and that the sensors be effectively used.

The two categories that Graves does not use are important to the sensor manager. Generally the requirements generation process is stochastic. New job generation is a random process. The environment is very dynamic. Hence stochastic and dynamic scheduling process models are more appropriate than deterministic and static ones.

The next section develops the sensor manager/manufacturing analogy in more detail.

The Manufacturing Plant Analogy

This section develops the sensor manager as a type of manufacturing plant in order to apply the results of the planning and scheduling literature for manufacturing to the sensor manager. The sensors are the machines of the manufacturing plant. They convert raw material into finished products. The input is the environment detected by the sensors while the primary finished products are the data on targets or emitters which are tracked, IDed or subjected to other information gathering activities.

Marketing, whose function is to bring in jobs, is at the highest organizational level. It, in consultation with the customer, sets a job's priority and due-date. The pilot, the mission and fusion managers, as well as the sensor manager itself are typical customers. Other customers such as displays and navigation are certainly possible. The marketing arm of the sensor manager maintains a menu of products that the plant produces. The customer selects a job from this menu. Typical menu items (products) include track, ID, target, search, align, etc. These could be modified depending on the type of customer. For instance, if the customer is the mission manager and the mission is defensive counter-air, then the sensor manager's menu would be different than if the mission is air-to-ground.

Each item on the menu is a product that the plant produces. For each product a process routing is prepared that describes how the product is to be produced. A process routing is the sequence of operations that describes the path through the plant that leads to a finished product. The knowledge-base is contained in the routing. A routing associates sensors with items ordered from the menu. It lists the sensors, alternate sensors, sequencing and any special information necessary to produce the product. Process routings can be

prepared off-line during the development of the sensor manager, on-line with a mission planning system that modifies the routings in the field or in real-time with an onboard planning system. In manufacturing terminology, preparing the process routing is called planning. Here planning is used in a broader context that also includes the prioritization of the tasks to be accomplished. Planning is a long range activity that tries to make the scheduling algorithm more efficient over the long run.

The sensor manager's scheduler assigns sensors and times in order to fill the jobs that marketing has identified. This is sometimes called timetabling in manufacturing terminology. Timetabling is generally made so as to minimize some cost function. Typical cost functions include tardiness (the amount by which a job is completed late), sensor idleness, and makespan. The scheduler has to resolve any bottlenecks that might occur for the limited sensor resources. It also must resolve conflicts between sensors by relaxing constraints associated with each job. Relaxation may include changing sensor modes, changing sensors or moving the collection times.

In any plant, pop-up (new orders) and pop-out (canceled) orders can create havoc with a plan and with a schedule. Real plants can handle pop-up orders either in the planning system where marketing in conjunction with the customer works the job into the existing plan or in the scheduling system where a total or partial rescheduling effort includes the new job. Pop-out jobs can be handled similarly. The intervals assigned to the removed operation can be left unused or a complete new schedule can be completed. Presently the preferred method is to minimize the impact to the current schedule for both pop-up and pop-out jobs. This allows a seamless integration into the existing schedule and is accomplished at the planning level for pop-up jobs by inserting the new job into the prioritized list. All lower priority jobs move down one position. At the scheduling interval the new job can be integrated into the job flow. With electronically steerable arrays and interruptable processors, inserting a new task is made easier (at least is not difficult). Still there is some overhead that slows throughput even for these devices. Mechanically slewed arrays and non-interruptable processors present more of a problem. The solution is to select a scheduling technique that supports rescheduling with a minimum change in the schedule. Another requirement is that timetabling occur frequently enough to keep the schedule current. A goal is to reschedule at least every 0.2 seconds. Most events change on the order of seconds rather than tenths of seconds. This procedure makes sure that pop-ups and pop-outs are quickly integrated into the schedule.

The next section examines one popular approach to planning and scheduling in job shops.

ISIS

ISIS is the Intelligent Scheduling and Information System (Fox and Smith (6)). It has been described by Smith, et al. (16) (not the same Smith) as the most popular intelligent scheduling system and by Turksen, et al (19) as one of the first successful applications of artificial intelligence to a production planning problem.

ISIS is a program that constructs schedules for production of orders in a job shop. In a job shop several jobs may exist at once. These jobs have different levels of importance and different due dates. They generally require a sequence of operations for completion with each operation requiring a specific type of machine for a specified length of time. The next operation in the sequence only begins when all of its preceding operations have been completed.

Fox and Smith view schedule construction as a constraint-directed activity that should be influenced by all relevant scheduling knowledge. The goal is to produce a schedule that reflects the current state of the factory and external environment. The problem-solving strategy is to find a solution that best satisfies the constraints. Fox and Smith use constraints to discriminate between alternative schedules as well as to reduce the number of potential schedules generated. Because constraints can conflict with each other, some constraints may have to be relaxed. This requires that detailed knowledge of all aspects of the job shop scheduling domain must be available to the scheduling system.

ISIS constructs a schedule by conducting a hierarchical, constraint directed search over a subset of all possible schedules. Its architecture consists of four levels of abstraction of the scheduling problem. Each level is characterized by the types of constraints considered at that level. Communication between levels is accomplished by constraint propagation. Control generally flows down but may go up in order to resolve conflicts. Processing at any one level consists of a pre-analysis, search and post-analysis. The pre-analysis bounds the level's search space; the search phase performs the actual search; and the post-analysis evaluates the quality of the partial schedules produced. If post analysis determines that the partial schedules are valid, then the results are passed as constraints to the next level. If there are no valid partial schedules at this level, then one or more constraints are relaxed in order to expand the search space. If no valid schedules can be generated, then control will be passed to a higher level where new constraints can be relaxed.

ISIS Architecture

The four levels of the ISIS architecture are order selection, capacity-based scheduling, resource scheduling and reservation selection.

At the highest level, order selection (level 1), ISIS prioritizes the orders based on the category of the order (forced outage, critical replacement, etc.) and its due date. This establishes the system's global strategy for integrating unscheduled orders into the existing job shop schedule. There are two types of orders: new orders received since the last planning and those whose schedules have been invalidated. Invalidation may occur because of a change in the plant status, changes to the order's description, or decisions imposed by the user. Once prioritized, orders are scheduled one at a time.

At the next level, capacity-based scheduling, ISIS determines the availability of the machines required by the selected order of the tasks. For the highest priority order ISIS will list all of the machines available to process the order. The purpose of level two is to detect bottlenecks in the plant. These bottlenecks become constraints which are passed onto level three to resolve.

Level three, resource scheduling, performs the detailed scheduling of all the resources necessary to produce an order. It extends the level two scheduling by considering more detailed information about operation resource requirements (in addition to the machines considered in level two) and by considering additional constraints. Level three's output is a list of possible schedules which are examined to see if one is acceptable. If there are no acceptable schedules, then a diagnostic program is invoked to determine the source of the problem and to fix it. This may require backtracking to a previous level and relaxing some of the assumptions there.

Once level three is finished the schedule is nearly complete. In level three a specific process routing has been selected for the order under consideration, resources (generally machines) have been selected for each operation in the routing and resource time bound constraints (e.g. this resource is reserved for a portion of this time period) have been associated with each selected resource. Level four uses the resource time bound constraints determined in level three to try to minimize the order's work-in-process time. The resulting specific resource reservations are added to the existing shop schedule and act as additional constraints for use in scheduling subsequent orders.

ISIS Design and the Sensor Manager

Can Fox and Smith's techniques be applied to the sensor manager design? Yes. The strength of their technique is its ability to model virtually any kind of constraint imaginable. It is easy to see how the sensor manager is a constrained process. In an air-to-ground scenario the sensor manager may have to resolve conflicting requests. For example, during weapons delivery the sensor manager may have to respond to requests from the fire control system for sensor time to acquire accurate coordinates of a target and from the pilot to determine if a surface to air missile or antiaircraft artillery is about to engage ownship. Constraint propagation is an ideal way to model such an environment. An ISIS style sensor manager takes tasks of

different levels of importance and time sensitivity (due dates) and develops a schedule that assigns various sensors to various tasks in a way that maximizes the amount of information available to other systems on the aircraft and other cooperative systems outside the aircraft. Constraint propagation is a way of synergistically tying the sensor systems together.

Many details of the sensor manager have duals in the manufacturing plant modeled by ISIS. The sensors are the machines that ISIS scheduled. Products on the sensor manager menu might include: associate, identify, track, target, search area, alignment, etc.

- associate—try to tie objects in the data base together by collecting more information on one or the other objects
- identify—determine whether an object in the data file is a friend, foe or neutral
- track—continue to track or develop a track on an object in the data file
- locate object—develop coordinates on an object in the data base sufficient to support targeting
- search area—scan an area for potential targets
- alignment—test alignment between sensors or develop an estimate of sensor misalignment

A process routing is a description of the steps that a product goes through in the manufacturing process. For the sensor manager it is a list of the sensors that must acquire data and the order in which they must operate. This scheme is a natural for handling all queuing operations. Assume the following sensors are available on the avionics suite: a radar with NCTR (noncooperative target recognizer), an omnidirectional ESM (electronic support measure), a directional ESM, an IFF (identification friend foe) interrogator and an IR (infrared) search and track. For these sensors a process routing for ***identify*** might be:

- task omnidirectional ESM to search frequency spectrum for object
- task directional ESM to look at object
- task IFF interrogator to challenge object
- task NCTR to examine object.

Cone (3) described in more detail how ISIS could be applied to a scenario taken from Waltz and Llinas (20). Following are some of the observations from that memo. There are sensor manager tasks such as ***identify*** that don't have clear completion times. ISIS does not have an easy way of handling nonending jobs. In order to schedule sensors the sensor manager has to know how long a task takes at the mode level. Scheduling at a higher level may force so much padding into the schedule that the sensor manager actually hurts the performance. Thus the routings have to be made at the lower, mode level where the sensor manager can handle sequential processes whose start and stop times are better defined. Another problem is that some products may age. For example a track's error generally increases over time. It may be necessary to warranty a product which means bringing it back into the plant and updating the information.

The hierarchical structure of ISIS does not support quick reaction by the sensors to a pop-up target. The information has to flow down through the hierarchy to affect the sensor activity. This is a problem shared by all vertical architectures (Dean and Wellman (4), page 463). This may prohibit an ISIS like system from operating in real-time at least for the time scales of the sensor manager.

ISIS does not provide a method of generating the optimal schedule. It only tries to generate an acceptable schedule. The first schedule it finds that works is the one it uses. ISIS also does not provide a way of estimating how far from the optimal schedule the schedule it suggests might be. These drawbacks led some researchers to look for more quantitative techniques. One such technique is Lagrangian Relaxation.

Lagrangian Relaxation Techniques

Fisher (5) used Lagrangian multipliers to find a lower bound on the cost of an optimal solution to a resource constrained network scheduling problem of which the job shop is a special case. He then used a branch and bound algorithm to find a solution to the scheduling problem. While a branch and bound solution provides an optimal solution to the scheduling problem, it is impracticable to use for realistically sized problems. References (9), (10) and (11) present a series of results that extend Fisher's work by using an augmented Lagrangian relaxation technique to find approximately optimal solutions to realistically sized scheduling problems. Luh and his coworkers break the problem into three paths. Each path addresses an increasingly difficult problem. The three paths are: (1) single operation (SO) or multiple operation (MO) jobs; (2) no precedence (NP) constraints, simple fork/join (SP) precedence constraints, or generic precedence (GP) constraints; and (3) identical machines (IM) or nonidentical machines(NM). The solution to the general job shop problem (MO/GP/NM) is given by Hoitomt, et al. (9). Assumptions made in this article are that the precedence constraints (precedence means assigning a particular type of machine to each operation of a job) can be represented as a directed acyclic graph and that each job ends with a single operation. Other assumptions include:

- job processing is nonpreemptive
- time horizon is long enough to complete all jobs
- the number of jobs, their weights, the processing time requirements, due dates, number of machine types, machine capacity, operation to machine mapping and required time-outs are known.

The result of the scheduling process determines the beginning times of all operations (a job is made up of a series of operations) and the machine types to process a particular operation.

What does the Lagrangian relaxation technique offer for the sensor manager puzzle? It provides an algorithm that can be executed to provide a schedule that is near optimal and it provides an estimate of how

near optimal the schedule is. There are shortfalls. Lagrangian relaxation does not address how the jobs should be prioritized but does provide a mechanism for including prioritization as weights on jobs. It was not designed to run on a real-time system such as the sensor manager. The Lagrangian relaxation technique may not be able to converge to a schedule within the planning cycle. In a related work Chang and Liao (2) show that a combination of Lagrangian relaxation with a rolling horizon scheme can be used to speed convergence to a near optimal schedule for a flexible flow shop. (A flexible flow shop is an extension of a traditional flow shop that allows multiple identical machines to be assigned at each operation in the process routing.) Since the sensor manager does not fit the flexible flow shop model this work needs to be extended to apply to the general job shop considered by Hoitomt, et al. (9). Finally, it is not clear that the scheduling workload is large enough to require such a sophisticated scheduler. The simple scheduler proposed by Popoli (14) may be sufficient.

Both of the techniques for scheduling examined so far suffer from not being designed to operate in real-time. The next section discusses real-time system problems and a real-time technique that might fit into the sensor manager puzzle.

Real-Time Systems

Stankovic (17) defines real-time systems as ones that depend on the time at which the results are produced as well as on the correctness of the result. He then lists the following ***misconceptions*** that people who are outside of the real-time community have about real-time systems:

- There is no science in real-time system design.
- Advances in supercomputer hardware will take care of real-time requirements.
- Real-time computing is equivalent to fast computing.
- Real-time programming is assembly coding, priority interrupt programming, and device driver writing.
- Real-time systems research is performance engineering.
- The problems in real-time system design have all been solved in other areas of computer science or operations research.
- It is not meaningful to talk about guaranteeing real-time performance because we cannot guarantee that the hardware will not fail and the software is bug free or that the actual operating conditions will not violate the specified design limits.
- Real-time systems function in a static environment.

This list could also be titled misconceptions about the sensor manager. Stankovic's main points are that (1) the time dimension must be elevated to a central principle of the system, (2) a new, more deterministic paradigm is needed that possesses well understood, bounded and predictable operating systems and application tasks and (3) a highly integrated and time-constrained resource allocation approach is necessary

to adequately address timing constraints, predictability, adaptability, and fault tolerance. The sensor manager, which is really like an operating system for a group of distributed processors, must also posses these same characteristics and operate in the same deterministic environment.

It is as important for an algorithm to execute on time as it is for the result to be correct. An algorithm has to do both to be useful. Many AI techniques and heuristics are not suited to analysis that provide guaranteed response times (Shin and Ramanathan (15)). Even when AI techniques can be shown to have predictable response times, the variance in those times is so large that providing timeliness guarantees based on the worst case performance result in severe under utilization of the computational resources during normal operations. Musliner et al. (13) proposed a Cooperative Intelligent Real-time Control Architecture (CIRCA) that uses an AI subsystem to reason about task-level problems that can afford to have unpredictable response times while a separate real-time subsystem deals with control-level problems that require predictable response times. This architecture may be appropriate for the sensor manager.

Rate Monotonic Scheduling

Rate-Monotonic Scheduling (RMS) is a theory for managing system concurrency and timing constraints at the level of tasking and message passing. It ensures that as long as the system utilization of all tasks lies below a certain bound and appropriate scheduling algorithms are used, all tasks meet their deadlines. RMS was developed to schedule tasks that are periodic but of differing periods. Originally the results required all tasks to be periodic, perfectly preemptable and independent. Under these conditions the tasks could be scheduled by assigning the task with the highest rate first and then assigning the remaining tasks monotonically in order of decreasing rates if they meet the requirements of the Liu and Layland theorem (Liu and Layland (10)). RMS has been extended to handle aperiodic tasks by making them appear to be periodic. It might be able to handle the sensor manager scheduling problem but most of the tasks the sensor manager handles are aperiodic. This does not play to the strength of RMS.

Real-time issues are a piece of the sensor manager puzzle. Another piece may be discrete event dynamic systems (DEDS). The next section looks at this possibility.

Control Theory

Classical control theory consists of a controller, a plant to be controlled and a feedback path. It has two main branches, analog or digital, depending on whether the controller is implemented with analog (resistors, capacitors and op amps) or discrete (generally a digital computer) components. In a DEDS both the controller and plant are discrete systems. The notion of time in classical control theory is replaced by an event sequence in a DEDS. *IEEE Control Systems Magazine* ran a special issue on DEDS in 1990 while the *Proceedings of the IEEE* did one in 1989. Applications for DEDS analysis include software

verification, database management, performance evaluation of manufacturing systems and optimization of distributed processing systems. DEDS's theorists hope to be able to apply the large body of control theory to the discrete event problem. There is some promise that a mathematical basis for the design of a sensor manager system could be developed based on the DEDS's work. The problem is that not all of the theory exists to be applied to such a large problem.

The next section reviews an article by Bonissone, Dutta and Wood (1). Their approach specifically includes planning as well as scheduling and does so in dynamic and uncertain environments.

AI/Planning

Bonissone, Dutta and Wood (1) (hereafter referred to as BDW) present a new approach to planning in dynamic and uncertain environments. Planning is defined as a sequence of actions designed to achieve certain goals. Because the environment can change while a plan is being executed it is by no means certain that the goals can be reached. BDW call planners that cannot react to a changing environment, static planners. Static planners often develop plans that will not execute in the real world. In response, many approaches have been developed that plan in uncertain and dynamic environments. Hendler, Tate and Drummond (8) summarize many of these ways. Generally the approaches either replan when the environment changes, expect the environment to change and plan ahead for it, interleave planning and execution by developing and executing partial plans sequentially, or develop highly reactive planners that minimize the look ahead. BDW approach the problem by trying to balance long term strategic planning and short term tactical planning. They introduce the concepts of goals, plans, and strategies. A goal is defined as an objective that an agent tries to achieve. Goals can be long or short term. A plan is a sequence of actions that when executed achieve the agent's goals, at least partially. Long range plans are called strategic while short range plans are called tactical. Tactical plans react quickly to changes in the environment while strategic plans change when required by a drastic change in the environment or modified long term goals. Strategies are general principles that help the planner generate and select goals and plans. Strategies as used by BDW are similar to military strategies guiding the development of a plan to attack the enemy.

BDW create a strategy hierarchy to direct planning. Planning means deciding which path to traverse down the strategy tree. Goals and plan scripts (plan scripts are lists of actions to be taken) are associated with each node in the strategy tree. Executing the plan is to execute each of the scripts at each node. Moving down the strategy tree one finds scripts dealing with shorter term events and new goals to which the planner must react. At each node decisions have to be made. In the real world, decisions have to be made under conditions of uncertainty. BDW use the uncertainty calculi of RUM/PRIMO (explained in the next paragraph) to aggregate the uncertainties associated with the multiple proof paths contributing to each decision. They also introduce the role of prior experience in the context of planning. Prior cases are treated

as additional sources of information influencing the decision process. RUM/PRIMO is used to combine multiple cases into the planning process.

RUM is a software program for reasoning with uncertainty. Its approach consists of decomposing the problem into three layers: representation, inference and control. The representation layer is the interface used to capture information for the inference and control layers. The inference layer provides the uncertainty calculi with which conclusions can be drawn. Five uncertainty calculi are provided. The third layer, control, selects which calculus is right for the context of the problem.

The architecture proposed by BDW has many strengths. It works in a constantly changing dynamic world. The strategic and tactical planning modes allow for a coherent combination of long range goals and overall objectives with short range responses to dynamic changes. It admits uncertainty and allows experience to be included into the reasoning process. The drawback is that there is no experience with this architecture in real-time control systems.

Putting the Puzzle Together

Each piece of the puzzle has something to offer to completing the picture. ISIS introduces constraints into the scheduling of jobs. Its model is very flexible and can include most every constraint conceivable. Lagrange relaxation provides a computational vehicle that provides an estimate of how close its schedule is to an optimal schedule. Real-time systems point out the problems associated with operating any embedded controller in real time. Any sensor manager design will have to tackle and solve those problems. DEDS has the potential to apply the whole body of literature developed for control systems to the scheduling problem. AI planning techniques merge strategic and tactical planning in dynamic and uncertain environments. These are the pieces. The challenge is fitting the pieces together.

References

1. Bonissone, Piero P., Soumitra Dutta and Nancy C. Wood. Merging strategic and tactical planning in dynamic and uncertain environments. *IEEE Trans. Systems, Man and Cybern.*, vol. 24(6), Jun 1994, pp. 841-862.

2. Chang, Shi-Chung and Da-Yin Liao. Scheduling Flexible flow shops with no setup effects. *IEEE Transactions on Robotics and Automation*, vol. 10(2), Apr 1994, pp. 112-122.

3. Cone, Milton L. RDL week 7. Internal memo. Jul 1994.

4. Dean, Thomas L. and Michael P. Wellman. *Planning and Control.* Morgan Kaufman Publishers, San Mateo, Ca, 1991.

5. Fisher, M. L. Optimal solution of scheduling problems using lagrangian multipliers. *Operations Research*, vol. 21(5), 1973, pp. 1114-1127.

6. Fox, Mark S. and Stephen F. Smith. ISIS - a knowledge-based system for factory scheduling. *Expert Systems*, vol. 1(1), Jul 1984, pp. 25-49.

7. Graves, Stephen C. A review of production scheduling. *Operations Research*, vol. 29(6), Jul/Aug 1981, pp. 646-675.

8. Hendler, James, Austin Tate and Mark Drummond. AI planning: systems and techniques. *AI Magazine*, vol. 11(2), Summer 1990, pp. 61-77.

9. Hoitomt, Debra J., Peter B. Luh and Krishna R. Pattipati. A practical approach to job-shop scheduling problems. *IEEE Transactions on Robotics and Automation*, vol. 9(1), Feb 1993, pp. 1-13.

10. Liu, C. L. and James W. Layland. Scheduling algorithms for multiprogramming in hard real-time environment. *Journal of the Association for Computing Machinery*, vol. 20(1), Jan 1973, pp. 46-61.

11. Luh, Peter B., Debra J. Hoitomt, Eric Max, and Krishna R. Pattipati. Schedule generation and reconfiguration for parallel machines. *IEEE Transaction on Robotics and Automation*, vol. 6(6), Dec 1990, pp. 687-696.

12. Musick, Stan and Raj Malhotra. Chasing the Elusive Sensor Manager. *Proceedings of the NAECON, May 1994.* Dayton, Ohio, 1994.

13. Musliner, D. J., E. H. Durfee and K. G. Shin. CIRCA: a cooperative intelligent real-time control architecture. *IEEE Trans. Systems, Man, Cybern.*, vol. 23(6), Nov/Dec 1993, pp. 1561-1574.

14. Popoli, Robert. The sensor management imperative. In Yaakov Bar-Shalom, editor, *Multitarget-Multisensor Tracking: Applications and Advances Volume II.* Artec House, Norwood, MA, 1992.

15. Shin, Kang G. and Parameswaran Ramanathan. Real-time computing: a new discipline of computer science and engineering. *Proc. IEEE*, vol. 82(1), Jan 1994, pp. 6-24.

16. Smith, Allen E., Thomas D. Fry, Patrick R. Philipoom, and James R. Sweigart. A comparison of two intelligent scheduling systems for flexible manufacturing systems. *Expert Systems with Applications*, vol. 6(3), Jul/Sep 1993, pp. 299-308.

17. Stankovic, John A. A serious problem for next-generation systems. *Computer*, Oct 1988, pp. 10-19.

18. Sycara, Katia P. Introduction to the special edition on planning, scheduling, and control. *IEEE Trans. Systems, Man, Cybern.*, vol. 23(6), Nov/Dec 1993, pp. 1489-1490.

19. Turksen, I. B., D. Ulguray, and Q. Wang. Hierarchical scheduling based on approximate reasoning - A comparison with ISIS. *Fuzzy Sets and Systems*, vol. 46(3), Mar 1992, pp. 349-371.

20. Waltz, Edward and James Llinas. *Multisensor Data Fusion.* Artech House, Norwood, Ma 1990.

A RESEARCH PLAN FOR EVALUATING WAVE GUN AS A LOW-LOADING MODEL LAUNCHER FOR HIGH SPEED AEROBALLISTIC TESTS

Robert W. Courter
Associate Professor
Department of Mechanical Engineering

and

Jason J. Hugenroth
Graduate Student
Department of Mechanical Engineering

Louisiana State University
Baton Rouge, LA 70803

Final Report for:
Summer Faculty Research Program
Wright Laboratory - Armament Directorate
Eglin Air Force Base, FL

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

and

Wright Laboratory - Armament Directorate

August, 1994

A RESEARCH PLAN FOR EVALUATING WAVE GUN AS A LOW-LOADING MODEL LAUNCHER FOR HIGH SPEED AEROBALLISTIC TESTS

Robert W. Courter
Associate Professor
Department of Mechanical Engineering
Louisiana State University

and

Jason J. Hugenroth
Graduate Student
Department of Mechanical Engineering
Louisiana State University

Abstract

A specialized light gas gun firing cycle, developed by Thomas Dahm of Astron Research and Engineering and named by him the Wave Gun, is investigated as a candidate for launching models in a ballistic range to high speed with relatively low model loading. The Wave Gun firing cycle features a very light piston which oscillates during the shot and produces a series of shock impulses on the model. A light gas gun interior ballistics code that simulates the Wave Gun firing cycle was used to evaluate launcher performance for a matrix of launcher geometric and launch parameters. A Wave Gun test facility, designed and constructed by Astron, was used to provide data with which to verify the fidelity of the simulation code. Pressure histories were recorded in the combustion chamber, the pump tube exit, the nozzle exit and at three axial stations along the launch tube. In addition the first-pass piston velocity and the model muzzle velocity were determined. Two test shots were fired. During the second shot a nozzle structural failure occurred and further testing was suspended pending fabrication of a new nozzle. The data acquired from the tests were not sufficient to verify the numerical model. However, the tests did provide experience in operation of the gun and data acquisition, and they provided insight into the status of the numerical model and the direction that future testing should take. A plan is presented for numerical and experimental studies to identify parameter sets that produce high velocity with moderate model loading. Initial testing and analysis will be devoted to validation of the gun cycle simulation code. Then parametric studies, supported by appropriate tests, will be carried out. Six parameters identified for consideration in these studies are propellant type and weight, helium charge pressure, pump tube volume, piston start pressure and model start pressure. Launch tube and model configurations will be held constant.

Acknowledgements

The authors would like to thank Mr. Gerald Winchenbach for selecting them to participate in this summer research program. His enthusiasm and continuing encouragement are sincerely appreciated. We are also indebted to Mr. Charles McClenahan and his staff for setting up the test facility, fabricating apparatus components and carrying out the experimental program. We are grateful to Messrs. John Huntington and John Case of Astron Research and Engineering for providing us with much useful information on the history and technical details of the Wave Gun program. Finally, we acknowledge the support of the Air Force Office of Scientific Research and the Wright Laboratory-Armament Directorate and the program management of Research and Development Laboratories.

A RESEARCH PLAN FOR EVALUATING WAVE GUN AS A LOW-LOADING MODEL LAUNCHER FOR HIGH SPEED AEROBALLISTIC TESTS

Robert W. Courter

Jason J. Hugenroth

Introduction

In 1981 Thomas Dahm of Astron Research and Engineering invented a unique firing cycle that provided the potential for weaponization of the light-gas gun. He called the device Wave Gun[1]. Essentially, it employed a very light piston in conjunction with a long propellant burning time and a high light-gas charge pressure to produce an oscillatory piston motion that, in turn, caused propellant burning rate fluctuations and multiple pressure pulses on the projectile. Analytical studies in conjunction with some crude experiments led to the suggestion that with appropriate tailoring of the gun and shot parameters, a high muzzle velocity could be achieved within the bounds imposed by weapon design. An interesting by-product of that study was the possibility of designing a cycle that could achieve a high muzzle velocity with a relatively low loading on the projectile[1]. It is advantageous in free-flight aeroballistic testing to have the capability of launching fragile models to high speeds without imposing destructive loads on the model. It is this prospect that Wave Gun might be used as a low-loading model launcher that motivates the present study.

In 1992 the Astron Wave Gun test apparatus was acquired by the Ballistics Branch of the Armament Directorate of Wright Laboratory, Eglin AFB, Florida. This facility has been activated for experiments to support the present study. The experimental results achieved will aid in validating a numerical simulation of the Wave Gun firing cycle. The objectives of the present study are to develop a simulation code, initiate the validating experimental program and provide a plan for future research that will produce an evaluation of Wave Gun as a "soft launch" aeroballistic model accelerator.

The Wave Gun Concept

A conventional gun uses an explosive propellant to accelerate a projectile in a launch tube. For a given configuration the muzzle velocity of such a gun is limited because some of the energy from the propellant must be expended to accelerate the heavy propelling gas. This difficulty is circumvented in the two-stage light gas gun. This gun features a chamber of light gas (the pump tube) between the propellant chamber and the projectile, sealed from the propellant gas by a moveable piston and from the projectile by a frangible diaphragm. Here the

propellant gas drives the piston which, in turn, compresses the light gas. Ultimately, the frangible diaphragm ruptures, and the compressed light gas accelerates the projectile through the launch tube. The high velocities attainable by this type of launcher make it attractive for use in free flight aeroballistic testing. It is standard practice in aeroballistic testing to use an "isentropic compression" firing cycle for the light gas gun. This cycle employs a heavy piston to produce a continuous, almost isentropic, pressure rise which eventually propels the model smoothly without large pressure spikes (Figure 1). The Wave Gun cycle, on the other hand, uses a very light piston that simply acts as a barrier separating the propellant gas from the light gas. This firing cycle routinely features an oscillating piston which alternately compresses and expands the propellant gas, producing fluctuations in burning rate and driving pressure. Some characteristics of a typical Wave Gun cycle are shown in Figure 1. It has been shown that this type of cycle can be optimized to produce very high muzzle velocities within the constraints of gun design[2]. It is also believed that high velocity shots with low model loading are possible through judicious selection of launcher geometric and shot parameters.

Firing Cycle Simulation

To investigate the capabilities of Wave Gun as a low-loading launcher a firing cycle simulation program has been constructed. The light gas gun code currently used at the Arnold Engineering Development Center (the "AEDC code") is used as the basic simulation engine. The code was originally developed by Piacesi, Gates and Seigel[3], and it has been extensively modified by DeWitt[4]. The code uses a von Neumann-Richtmyer time-stepping procedure with artificial viscosity for integration of the fundamental equations of motion, a power law relationship for propellant burning rate and a virial-type real gas model for the light gas (in this case, helium). The present authors have modified the code to be compatible with the requirements of the experimental program. In addition provisions are made to alter treatment of propellant conservation laws, piston and model release and friction and pump tube heat transfer . These adjustments will be guided by the results of the experimental program.

Experimental Facility Description

The Astron Wave Gun 30mm test facility was assembled at Eglin AFB to provide experimental support for the present research program. The gun was originally designed to investigate potential firing cycles for light gas gun weaponization, a program requiring flexibility of configuration. This flexibility was achieved by using a massive steel tube to contain the internal parts of the launcher where very high pressures are generated. A schematic of the gun is shown in Figure 2. The main components of the gun are the internal breech plug and ignition system, the propellant chamber, the polypropylux piston, the pump tube liner, the nozzle, the aluminum model and the launch tube. The parts subjected to high pressure are contained under compression in the outer tube by a breech plug at one and and a barrel nut at the other. A spacer is used to permit adequate compression of the internal parts. The original facility had three different sets of components so that the internal volume relationships could be

parametrically investigated. However, only the configuration shown in the figure is possible at the present time. Detail drawings of the gun components are shown in the Appendix.

Helium gas is supplied to the pump tube through a fill valve which is not shown in the figure. A black powder primer is used in the spit tube, and the main propellant is bagged and wrapped around the spit tube. Ignition is with a 50 volt electrical pulse. The piston is actually screwed into the pump tube end of the propellant chamber. Piston motion begins when the combustion pressure is sufficient to shear the polypropylux piston threads. Thus, the number of piston threads engaged determines the piston start pressure. The model is simply an aluminum cylinder with integral flange. Model motion begins when the driving pressure is sufficient to shear the flange. The shot start pressure is therefore determined by the flange thickness and material.

There are thirteen instrumentation ports in the gun, nine in the high pressure tube and four in the launch tube. It is important to note that the positioning of the internal parts must be precise so that these ports are open. In this regard it is essential that when the gun is being sealed prior to firing, the breech plug and barrel nut must be tightened simultaneously so as not to disturb this instrumentatin port alignment. The position of the ports and the designation of those used for the tests of this program are indicated in Figure 3. Transducers 3, 8 and 9 provide, respectively, the propellant chamber pressure, the nozzle entrance pressure and the nozzle exit pressure. Transducers LT1, LT2 and LT3 provide pressures at the repective launch tube locations. Ports 4, 5 and 7 are used for breakwires that signal passage of the piston during its first travel down the pump tube.

Figure 4 is a schematic of the overall setup for the Wave Gun tests. Each quartz pressure transducer was connected through a charge amplifier to a digital oscilloscope. The transducers were set to trigger simultaneously from the signal of the first transducer. The breakwires were connected across two gated digital timers to provide the elapsed time for piston passage between the breakwire stations. Two infrared sky screens were used in the same way to determine the approximate muzzle velocity of the projectile. Finally, a Doppler radar unit was used to determine the projectile trajectory from launch tube exit to the model-catching bunker. It is intended that this unit be used for downbore velocity measurements in later tests. In addition a VISAR unit (Velocity Interferometer System for Any Reflector) will also be available for down-bore measurements in later tests.

Experimental Results

The Astron Wave Gun was assembled and commissioned at Eglin AFB during the present summer research program. In Reference 2 some data from previous firings of the gun are provided. Since the gun had not been fired in about seven years, it was deemed adviseable to initiate the new firing program by duplicating a low performance shot from the previous program. However, even this was not exactly possible because a supply of the

same propellant and a supply of high pressure helium were not available. The propellant deficiency was not of major importance for the first shot, and M30/19 MP propellant was used. The helium deficiency was important. The Wave Gun uses a low volume pump tube and light piston. The low volume necessitates using helium at unusually high pressure in the pump tube. Standard light gas guns use helium pressures of about 200 psi. The Wave Gun pressures are between 2500 and 4000 psi. This high pressure plays an important role in the piston behavior, particularly with regard to piston speed during the cycle. The parameters for the two shots that made up the present experimental program are shown in the table below.

Shot number	Propellant weight (gm)	Primer weight (gm)	Model weight (gm)	Piston start pressure (psi)	Model start pressure (psi)	Helium charge pressure (psi)
1	1304	18.2	111.6	3100	34800	900
2	1304	18.2	111.6	3100	34800	1600

The radar was not available for Shot 1 so the muzzle velocity was estimated from sky screen measurements (see Figure 4). The radar was used instead of the sky screens for Shot 2.

The results of the two shots were disappointing. No data were acquired from Shot 1. The instrumentation trigger was activated by the firing switch. One possible cause for the failure was a delayed ignition which caused the scope to sweep prior to the main part of the firing cycle. Another was a possible short in the trigger circuit. The transducer traces indicated negligible activity, so it is not possible to determine the exact cause. The sky screens did not give an indication of projectile passage, so it seems likely that their circuit suffered a short during the early stages of the shot. The piston, the model and the sheared model flange were all recovered. The piston was partially deformed and wedged in the nozzle throat. The model separated cleanly from the flange in a pure shear failure, indicating failure at a pressure that was near the design value. There was no damage to the steel backing washer or to any of the gun components. Some very mild erosion, probably from blow-by, was detected on the upstream nozzle face. The physical evidence after the shot and the relatively low helium charge pressure suggest that the Wave Gun operated in a fashion close to that of a standard light gas gun. The numerical simulation of this shot indicates something else. A discussion of the numerical simulations and the experimental results for both shots is given in the next section.

The second shot was triggered with the propellant chamber pressure transducer, so ignition delay was removed as a factor in data acquisition. Some experimental data were acquired for this shot. Pressure traces from the nozzle entrance, the nozzle exit and the first launch tube station were recorded. Also, the elapsed time for piston travel between ports 4 and 5 and the radar track of the projectile (including the muzzle velocity) were acquired. The

other sensors failed to provide any data. The data acquired are shown in Figure 5. The radar track, which is very reliable, indicates a muzzle velocity of 5800 fps. The pressure traces, however, are suspect. Since the instrumentation was to have been triggered from the pressure sensor in port 3 (at the propellant chamber), it would be expected that the transducers at the pump tube and nozzle exits would initially indicate a low pressure followed by a sudden rise in pressure as the pump tube gas is compressed by the advancing piston. Each of the actual traces shows a high value at the triggering time and an uncharacteristic trace. These traces do provide an indication of the pressure levels in the gun, but it is surmised from the previous arguments that these are probably not maximum values. The pressure trace from the launch tube transducer has the appropriate characteristics. However, it is impossible to assess the timing of the pulse in light of the triggering difficulties with the other transducers. It is obvious that this shot experienced at least a double compression by the piston. The piston was extruded through the nozzle and launch tube and propelled down range. The model was not recovered, but the model flange had a jagged edge indicating a combined stress failure. This would seem to indicate that the pressure was high enough to disturb the seal between nozzle and launch tube. The steel backing washer was also eroded. Most importantly, however, the upstream nozzle face was found to be severely eroded by blow-by and several stress cracks, one quite severe, were in evidence. In addition, the inside surface of the steel gun shell was eroded at the location of the pump tube-nozzle joint. All of this evidence points to a very powerful shot that was particularly hard on the nozzle structure, probably because the charge pressure of the helium was below that used by Astron in their initial tests with this gun.

Simulation Results

The results of the two simulations are summarized in the following table. In addition the behavior of piston and model for the two cases is show in Figures 6 and 7.

Shot Number	Muzzle Velocity (fps)	Maximum Base Pressure (psi)	Maximum System Pressure (psi)	Total Cycle Time (ms)	Model Release Time (ms)
1	4916	50000	170000	7.27	4.39
2	5805	46000	92000	8.26	5.92

Comparison of these simulations with the experimental results prompts some interesting observations. The simulated results for Shot 2 compare very favorably with the experimental data. The system pressure indicated by the simulations is not consistent with what can be gleaned from the experimental results in that it would be expected for the higher pressures to occur for Shot 2 where some structural damage and complete piston extrusion occurred. Having made these observations, it is now appropriate to consider each simulation separately.

The plots of Figure 6 indicate that the model in Shot 1 was released at the second compression by the piston. The relatively large amplitude of the piston oscillation would suggest a pressure fluctuation that would yield a lower average driving pressure than would occur with smaller piston oscillations. Thus, the muzzle velocity would suffer accordingly. This would also indicate that the release load would be softer. It is apparent from the model trajectory that a third compression reaches the model before it leaves the launch tube.

The simulated results of the second shot tell a different story. The model is launched just at the third compression by the piston. Note that the piston osicillates at a lower amplitude than occurs for the first shot. Thus, a larger load is imposed on the gun structure and the model. The calculated muzzle velocity and maximum base pressure are surprisingly close to the experimental results for this shot. At this point of the investigation the descrepancy that occurs in the maximum base and system pressure for the two shots cannot be adequately explained. It is obvious that we have a long way to go to have a validated simulation code. However, we have made a beginning, and the results achieved so far are not drastically unreasonable.

Future Research

The present research effort has been disappointing in that no measurements have been made that can be used directly in verifying the simulation code. However, the experience gained in assembling and firing the Wave Gun has been valuable, and the lessons learned regarding data acquisition have considerably raised the probability of success with future tests. It is important, then, to plan carefully for the next series of tests so that the maximum benefit can be gained. The following general observations are important:

1. All instrumentation mounted on the gun should be activated with a common trigger. The propellant chamber pressure transducer is a good choice for this trigger.
2. It would be advantageous to connect the breakwires into the system such that the elapsed time from trigger to wire disconnect can be determined for each wire.
3. All sensors should be recorded on a unit with a common time base. It is also advantageous for the data to be easily transferable to a computer. Consideration should be given to activating the 12-channel Soltec recording unit.
4. High pressure helium (over 3000 psi) should be available for the tests
5. If possible, in-bore measurements of velocity with a VISAR or Doppler radar should be made.

At least the first three tests in the new series should be aimed at duplicating results reported by Astron in Reference 2 for three different test conditions. These tests will provide definitive information on the piston and model start conditions, the performance of the propellant and the behavior of the piston. Also, careful attention to the wave

timing and pressure levels should provide some insight into the heat transfer and friction models in the code. Prior to the initiation of new tests the simulation code will be modified to account for spatial property changes in the propellant chamber and mass transfer from the gun liner into the surrounding space. Guidance for these revisions will be provided by the experimental results reported by Astron. Since the objectives of our program are somewhat different from those reported by Astron, it is not possible to specify the testing matrix for the additional tests required to validate the simulation code. As with any investigative effort, the direction taken will depend on the results of the most recent tests and analysis. However, six parameters will be considered initially.during the program. These are propellant type and weight, helium charge pressure, pump tube volume, piston start pressure and model start pressure

References

1. Dahm, T. J. and D. S. Randall, "The Wave Gun Concept for a Hypervelocity Rapid Fire Weapon," presented at the 1984 JANNAF Propulsion Meeting, New Orleans, LA, 8 February 1984.
2. Mawhinney, R. C., D. S. Randall and R. P. Heydt, "Wave Gun Charge Development," Astron Research and Engineering Report Number 7110-01, 29 September 1989.
3. Piacesi, R., D. F. Gates and A. E. Seigel, "Computer Analysis of Two-Stage Hypervelocity model Launchers," Naval Ordnance Laboratory, NOLTR 62-87, February, 1963.
4. Cable, A. J. and J. R. DeWitt, "Optimizing and Scaling of Hypervelocity Launchers and Comparison with Measured Data," Arnold Engineering Development Center, AEDC-TR-67-82, April, 1967.

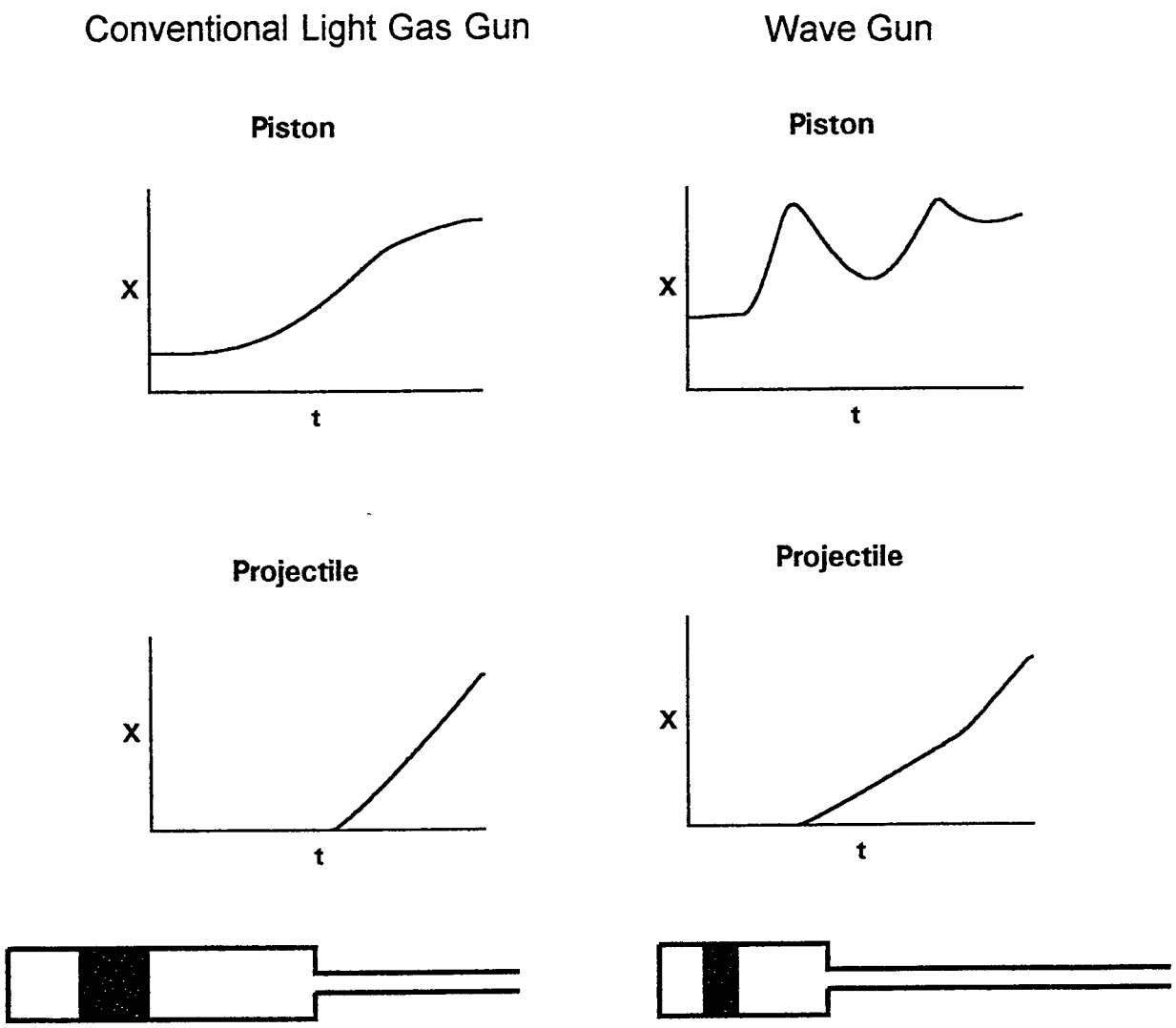
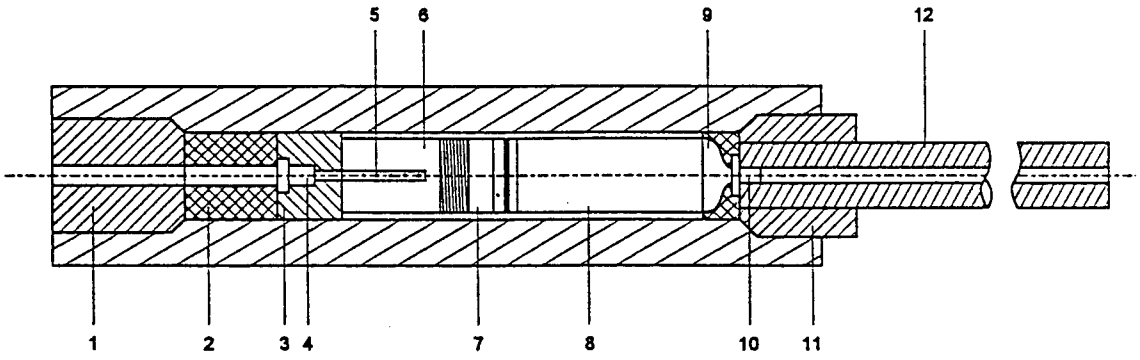
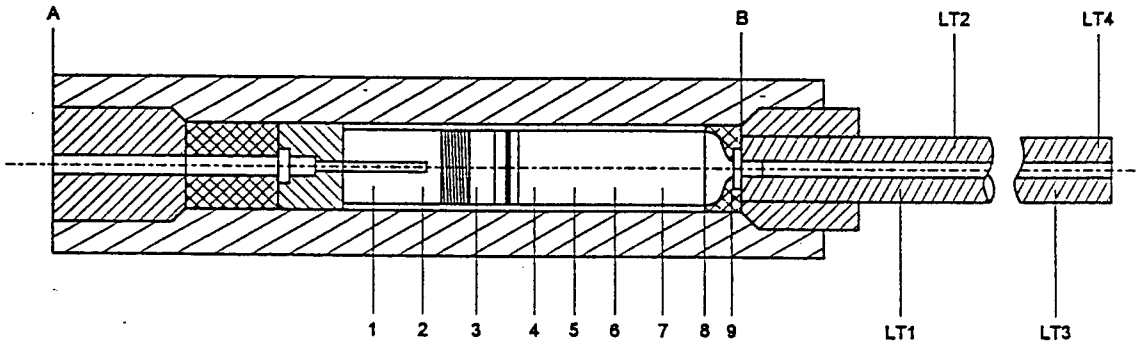


Figure 1. Comparison of Wave Gun and Conventional Light Gas Gun firing cycles.



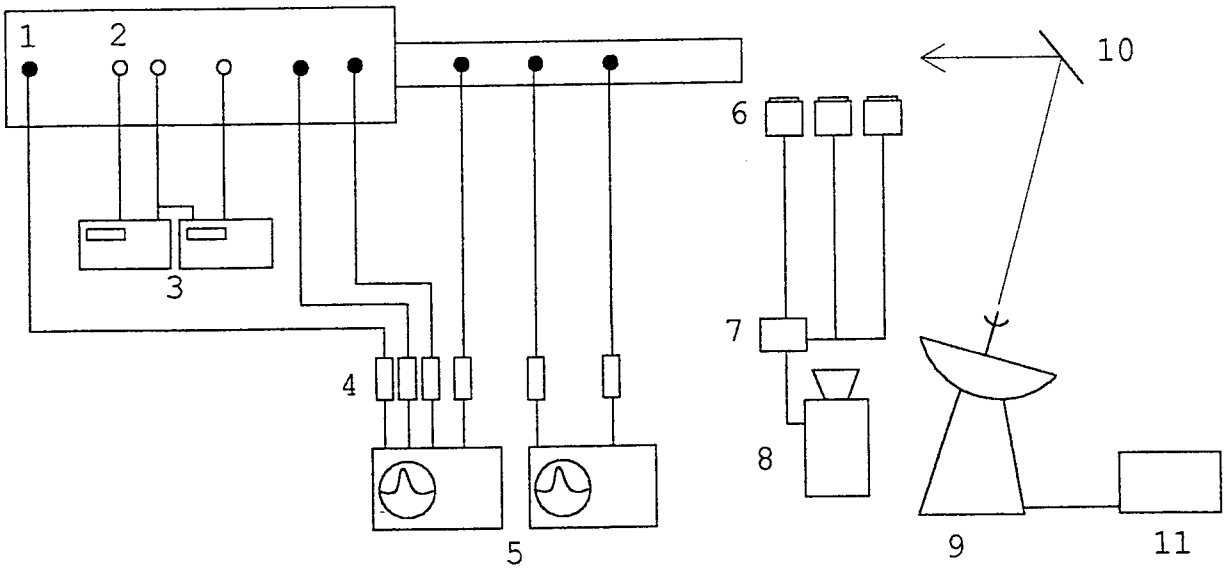
No.	Part	Length (cm.)	Diameter (cm.)	No.	Part	Length (cm.)	Diameter (cm.)
1	Breech plug (outer)	-	-	7	Piston	10.29	11.43
2	Spacers	11.47	13.29	8	Pump tube	49.43	11.43
		16.04	13.29	9	Nozzle	7.67	11.43
		22.91	13.29	10	Projectiles	5.715	3.00
3	Breech plug (inner)	12.07	13.29			6.033	3.00
4	Igniter	-	-			6.350	3.00
5	Spit tube	-	-	11	Barrel nut	-	-
6	Propellant chambers	18.67	11.43	12	Launch tube	243.84	3.00
		23.25	11.43				
		28.19	11.43				

Figure 2. Astron Wave Gun Test Apparatus



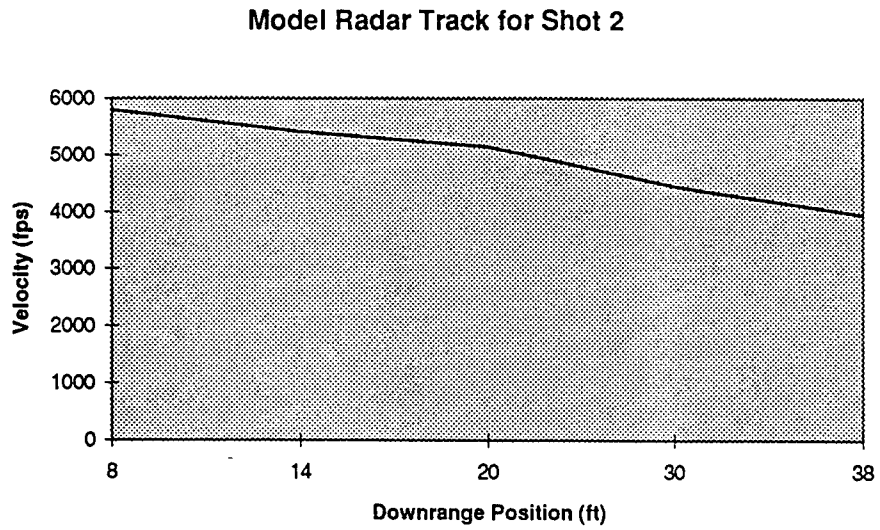
No.	Location (cm.)*	Use	No.	Location (cm.)**	Use
1	45.72	Not active	LT1	45.72	Pressure transducer
2	60.96	Pressure transducer	LT2	76.20	Pressure transducer
3	71.12	Not active	LT3	137.16	Pressure transducer
4	81.28	Break wire	LT4	198.12	Not active
5	91.44	Break wire			
6	101.60	Not active			
7	111.76	Break wire			
8	121.92	Pressure transducer	* Measured from A		
9	128.11	Pressure transducer	** Measured from B		

Figure 3. Instrumentation ports on Astron Wave Gun



No.	Instrument	Model
1	Piezeoelectric pressure transducers	Kistler 60704
2	Breakwires	-
3	Universal counters	HP 5315B
4	Charge amplifiers (1-4)	Kistler 504E4
	Charge amplifiers (5-6)	PCB 463A
5	Digital oscilloscopes	Nicolet 4094B
		Hi Techniques HT-600
6	Flashers	Hadland Photonics
7	Flash control unit	Hadland Photonics CU-2
8	Camera	Hadland Photontics SV-553BR
9	Radar	Opos Electronics
10	Sacrificial mirror	-
11	Radar analyser	Terma DR-5000

Figure 4. Experimental test setup



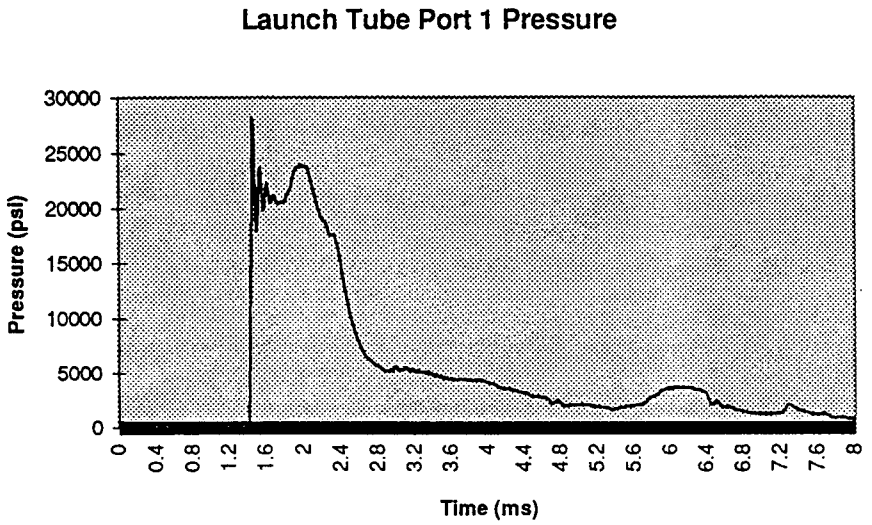
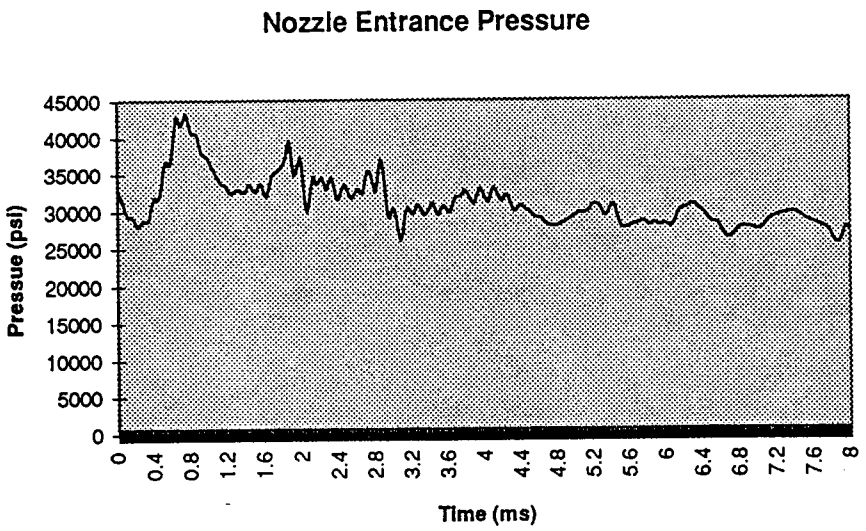


Figure 5. Experimental Results



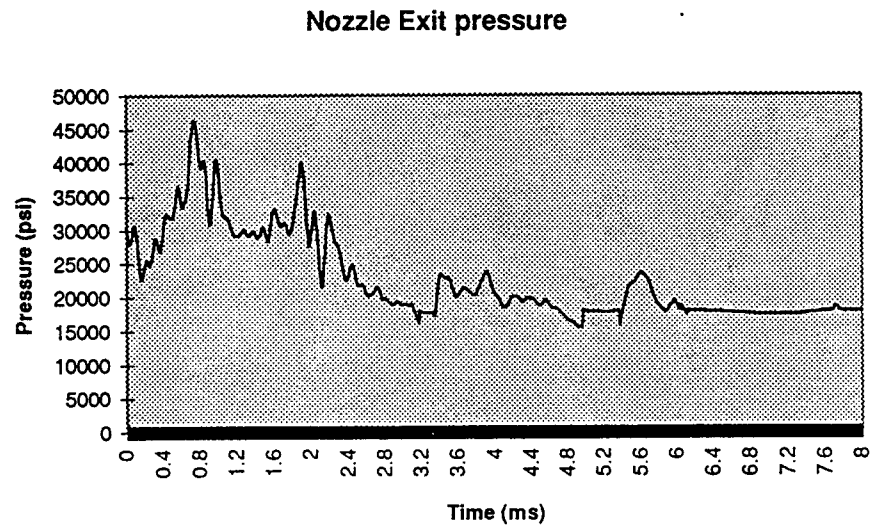
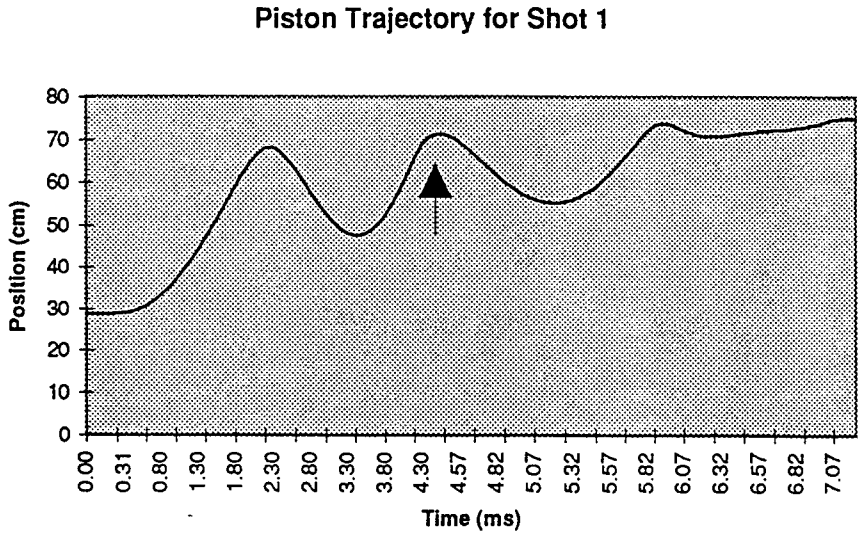


Figure 5. Experimental Results (Concluded)



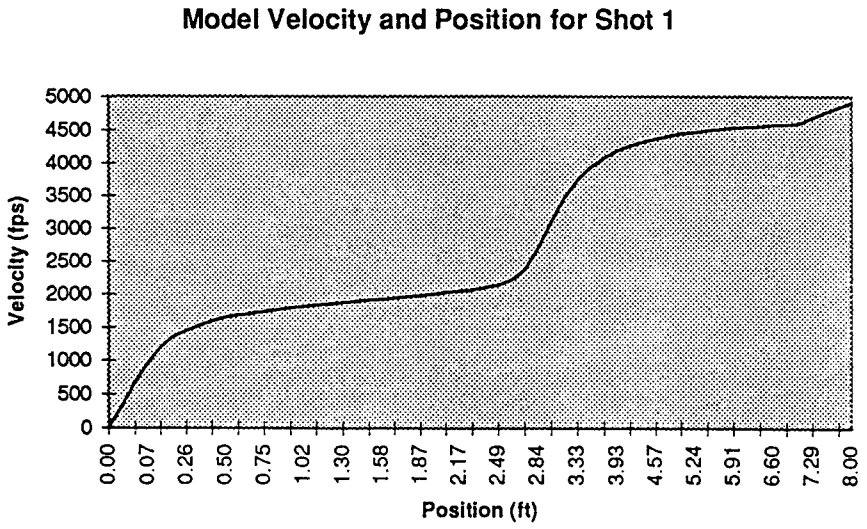
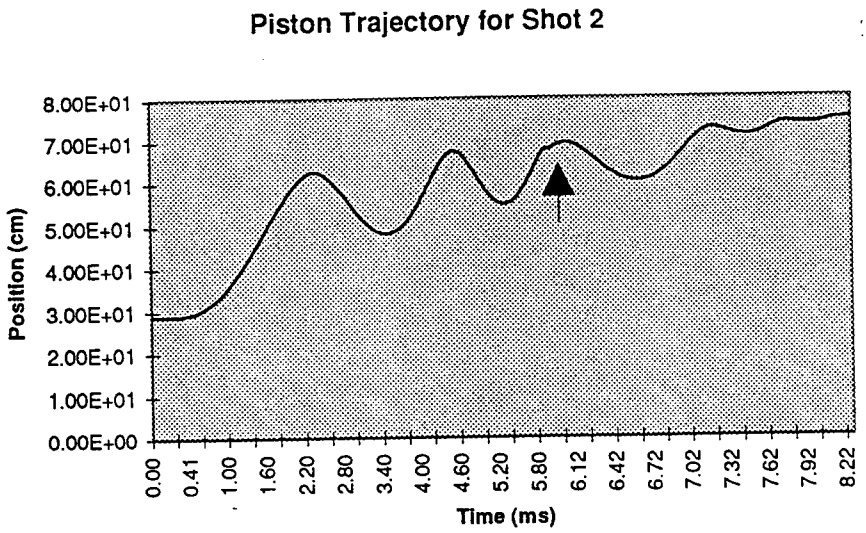


Figure 6. Numerical Results for Shot 1



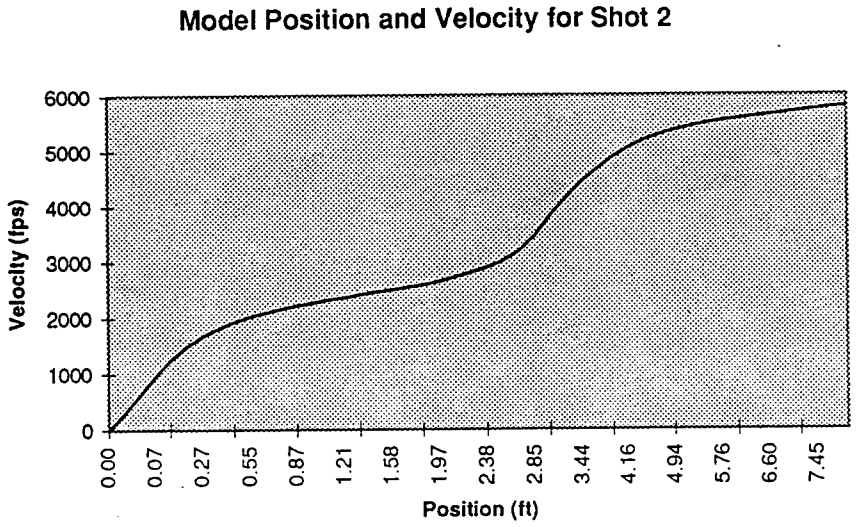
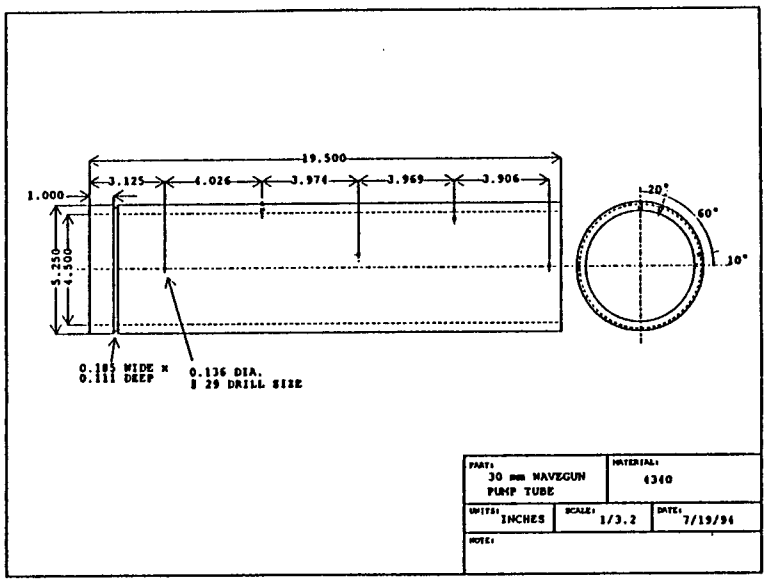
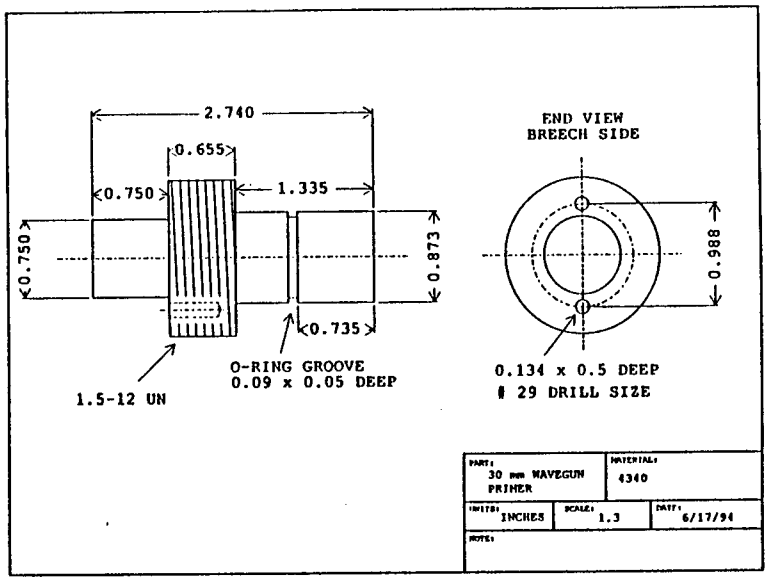
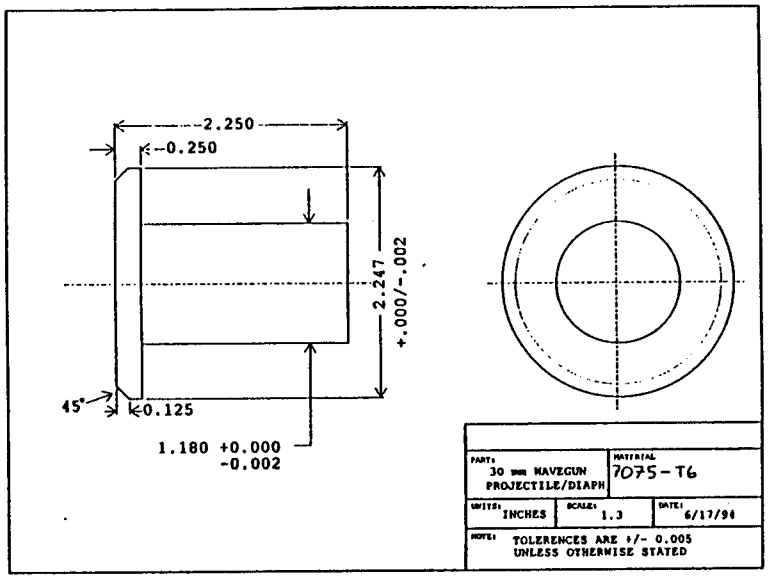


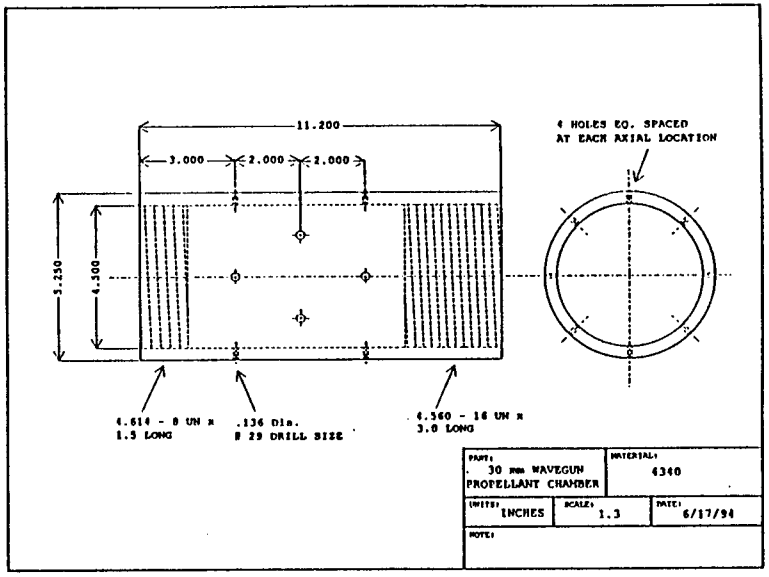
Figure 7. Numerical Results for Shot 2

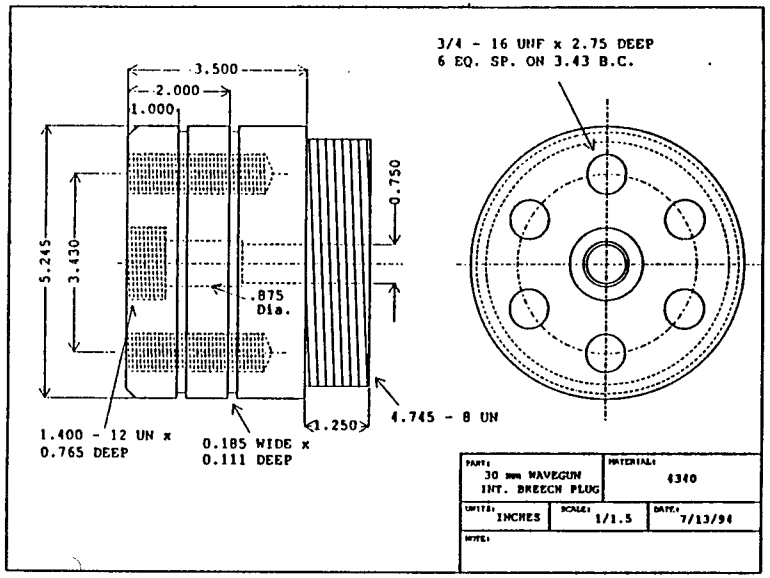
Appendix

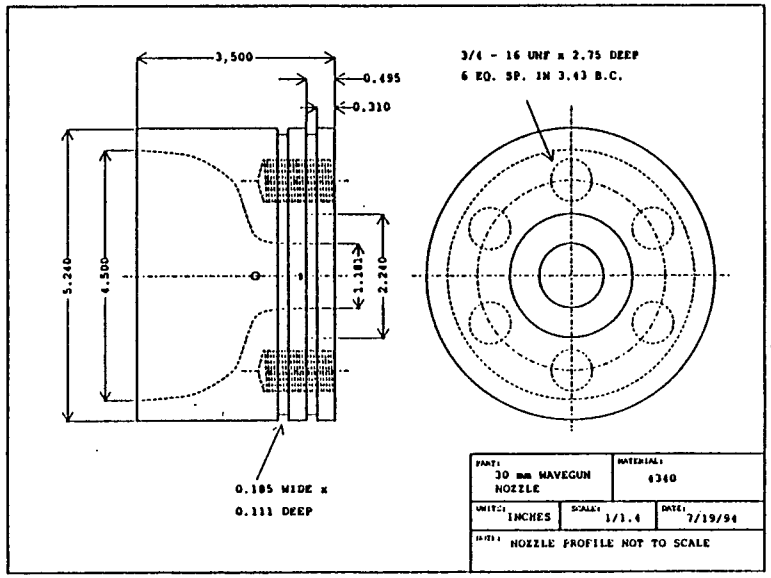

19.500
1.000
3.125
4.026
3.974
3.969
3.906
5.250
4.500
20°
60°
10°
0.185 WIDE x
0.111 DEEP
0.136 DIA.
29 DRILL SIZE
PART: 30 mm MAVEGUN PUMP TUBE
MATERIAL: 4340
UNITS: INCHES
SCALE: 1/3.2
DATE: 7/19/94
NOTE:

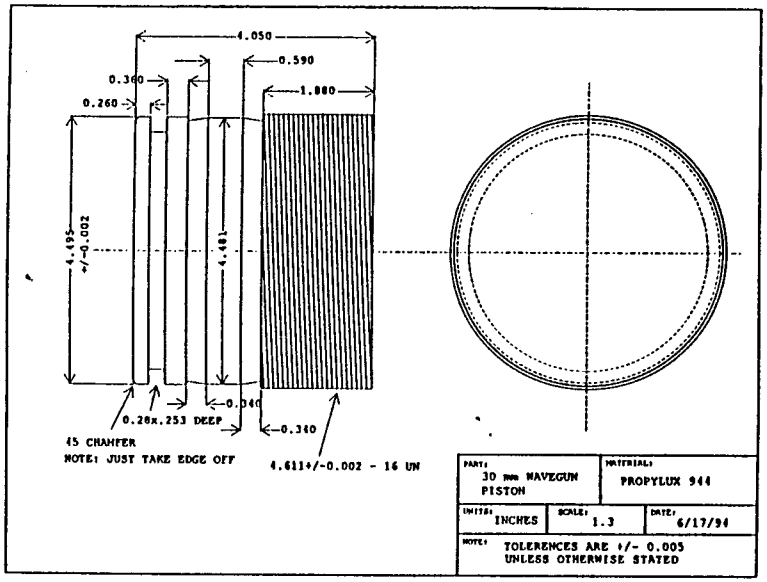

2.740
0.655
0.750
1.335
0.750
0.873
0.735
1.5-12 UN
O-RING GROOVE
0.09 x 0.05 DEEP
END VIEW
BREECH SIDE
0.988
0.134 x 0.5 DEEP
29 DRILL SIZE
PART: 30 mm MAVEGUN PRIMER
MATERIAL: 4340
UNITS: INCHES
SCALE: 1.3
DATE: 6/17/94
NOTE:

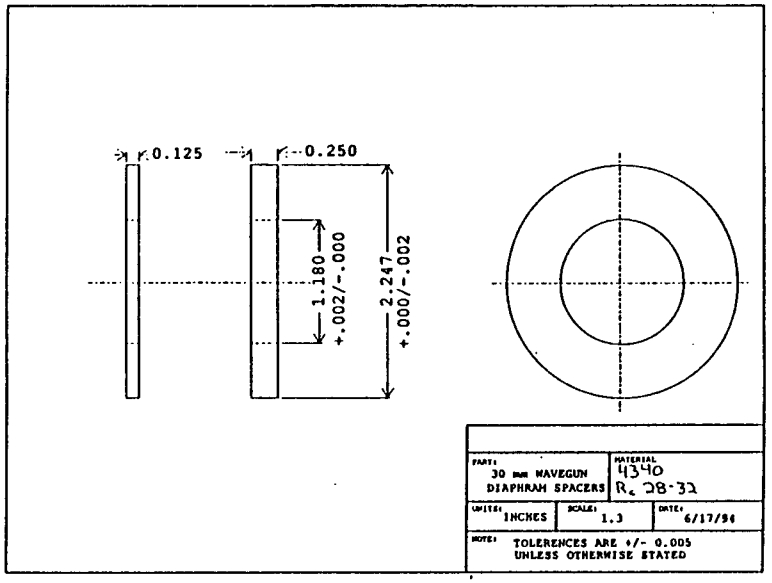

2.250
0.250
2.247
+.000/-.002
45°
0.125
1.180 +0.000
-0.002
PART: 30 mm MAVEGUN PROJECTILE/DIAPH
MATERIAL: 7075-T6
UNITS: INCHES
SCALE: 1.3
DATE: 6/17/94
NOTE: TOLERENCES ARE +/- 0.005 UNLESS OTHERWISE STATED


11.200
3.000
2.000
2.000
5.250
4.500
4 HOLES EQ. SPACED
AT EACH AXIAL LOCATION
4.614 - 8 UN x
1.5 LONG
.136 Dia.
29 DRILL SIZE
4.560 - 16 UN x
3.0 LONG
PART: 30 mm MAVEGUN PROPELLANT CHAMBER
MATERIAL: 4340
UNITS: INCHES
SCALE: 1.3
DATE: 6/17/94
NOTE:


3/4 - 16 UNF x 2.75 DEEP
6 EQ. SP. ON 3.43 B.C.
3.500
2.000
1.000
5.245
3.430
.875
Dia.
0.750
1.400 - 12 UN x
0.765 DEEP
0.185 WIDE x
0.111 DEEP
1.250
4.745 - 8 UN
PART: 30 mm MAVEGUN INT. BREECH PLUG
MATERIAL: 4340
UNITS: INCHES
SCALE: 1/1.5
DATE: 7/13/94
NOTE:


3,500
0.495
0.310
5.240
4.500
1.181
2.240
0.185 WIDE x
0.111 DEEP
3/4 - 16 UNF x 2.75 DEEP
6 EQ. SP. IN 3.43 B.C.
PART: 30 mm MAVEGUN NOZZLE
MATERIAL: 4340
UNITS: INCHES
SCALE: 1/1.4
DATE: 7/19/94
NOZZLE PROFILE NOT TO SCALE


4.050
0.590
0.360
0.260
1.880
4.495
+/-0.002
4.481
0.28x.253 DEEP
0.340
0.340
45 CHAMFER
NOTE: JUST TAKE EDGE OFF
4.611+/-0.002 - 16 UN
PART: 30 mm MAVEGUN PISTON
MATERIAL: PROPYLUX 944
UNITS: INCHES
SCALE: 1.3
DATE: 6/17/94
NOTE: TOLERENCES ARE +/- 0.005 UNLESS OTHERWISE STATED


r.0.125
r.0.250
1.180
+.002/-.000
2.247
+.000/-.002
PART: 30 mm MAVEGUN DIAPHRAM SPACERS
MATERIAL: 4340 Rc 28-32
UNITS: INCHES
SCALE: 1.3
DATE: 6/17/94
NOTE: TOLERENCES ARE +/- 0.005 UNLESS OTHERWISE STATED

LONGITUDINAL WAVES IN FLUID LOADED COMPOSITE FIBERS AND FIBERS EMBEDDED IN A SOLID MATRIX

VINAY DAYAL

Assistant Professor

Aerospace Eng. and Eng. Mechanics Dept.

304 Town Eng. Bldg.

Iowa State University

Ames, IA 50011

Final Report for

Summer Faculty Research Program

Wright–Patterson AFB

Sponsored by

Air Force Office of Scientific Research

Bolling Air Force Base, DC

and

Wright–Patterson AFB

August 1994

LONGITUDINAL WAVES IN FLUID LOADED COMPOSITE FIBERS AND FIBERS EMBEDDED IN A SOLID MATRIX

VINAY DAYAL
Assistant Professor
Aerospace Eng. and Eng. Mechanics Dept.
304 Town Eng. Bldg.
Iowa State University
Ames, IA 50011

Abstract

The theoretical model for longitudinal waves traveling in a transversely isotropic fiber in a transversely isotropic matrix has been developed. The fiber first is studied in a fluid and then in the solid matrix. Dispersion curves for various modes of wave propagation in fiber, a hollow tunnel and then, the fiber in a matrix, have been obtained. The governing equations for a damage zone around the fiber have been derived. The damage is modelled as a thin layer of material as well as a massless spring.

LONGITUDINAL WAVES IN FLUID LOADED COMPOSITE FIBERS AND FIBERS EMBEDDED IN A SOLID MATRIX

VINAY DAYAL

INTRODUCTION

A major damage mode in reinforced composite is the inability of the fiber to form a good bond with the matrix. This results in the damage accumulation on the surface of the fiber, which on subsequent loading can lead to major failure. During my work at the Wright–Patterson AFB I have been involved in the development of a theoretical model for the characterization of the fiber–matrix debonding in the continuous fiber composites. The major focus of this work has been to study the ultrasonic wave propagation in fibers embedded in a matrix. The waves will be travelling in the axial direction of the fiber and as they progress, they will leak energy into the surrounding medium (or matrix). Now, if the bonding between the fiber and matrix is good then we will observe good bonding and if not, then the bonding will be weak. Matikas and Karpur[1] have studied the phenomenon of the interface debonding by the use of reflected shear waves. Here they produce shear waves in the matrix which are reflected from the fiber matrix interface and the interface characteristics can be measured. They have modelled the interface as a massless spring of stiffness which characterizes the interface. One disadvantage of the method is that if there are fibers very close together then experimentally it is difficult to focus on a single fiber. If the waves can be propagated along the fibers then this limitation can be resolved. The disadvantage will be that the fiber ends have to be accessible. Also, as will be shown later both the normal stiffness and shear stiffness of the interface can be modelled in this mode.

DETAILS OF THE WORK DONE

The first task during this work was to start with a fiber in a fluid and then introduce the constraining effect of the fluid on the fiber. This is analogous to the pressure due to the residual stresses on the fiber. Dayal(2) has analyzed the effect of fluid on the wave propagation in a fiber and the derivation could be applied here. These equations were modified to include the pressure effects and an analytical analysis shows that the fluid pressure will not produce any effect on the wave propagation in the fiber. This can be explained easily as the wave propagation is a transient phenomenon while the static loads will not effect the transient phenomenon. A more mathematical explanation is in order and is as follows.

The coordinate system used here is as shown in Fig. 1

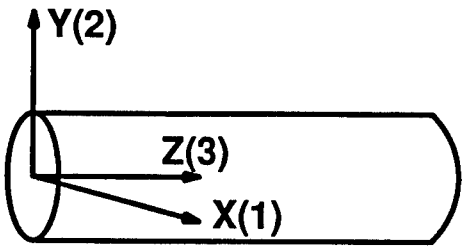


Fig. 1 The coordinate system

The waves in the fiber can be represented by,

$$\begin{aligned} U_r^w &= [-A\gamma J_1(\gamma r) - BJ_1(\eta r)ik]\ exp[i(kz - \omega t)] \\ W^w &= [Aik\gamma J_0(\gamma r) + B\eta J_0(\eta r)ik]\ exp[i(kz - \omega t)] \end{aligned} \tag{1}$$

where U is the radial displacement, W is the axial displacement, A and B are arbitrary magnitudes, r is the radius of the fiber, k is the wave number, ω the circular frequency, t is time, J_0 and J_1 are the Bessel's functions of first kind and zero and first order, γ and η depend on the elastic constants of the materials and are defined as

$$\gamma = \pm \sqrt{\frac{\varrho\omega^2 - k^2(C_{13} + 2C_{55})}{C_{11}}} \qquad \eta = \pm \sqrt{\frac{\varrho\omega^2 - C_{55}}{C_{11} - C_{13} - C_{55}}}$$

From elasticity we can write the relation between stress and strain,

$$\begin{aligned} \varepsilon_x &= S_{11}\sigma_x + S_{12}\sigma_y + S_{13}\sigma_z \\ \varepsilon_y &= S_{12}\sigma_x + S_{22}\sigma_y + S_{23}\sigma_z \\ \varepsilon_z &= S_{13}\sigma_x + S_{23}\sigma_y + S_{33}\sigma_z \end{aligned} \tag{2}$$

For the case of hydrostatic pressure on the fiber the stresses will all be equal to the applied pressure, p_0

$$\sigma_x = \sigma_y = -p_o\ ; \quad \sigma_z = 0 \tag{3}$$

Thus, the strains now can be written in terms of the external pressure,

$$\varepsilon_x = -(S_{11} + S_{12})p_o;\ \varepsilon_y = -(S_{12} + S_{22})p_o\ ;\ \varepsilon_z = -(S_{13} + S_{23})p_o \tag{4}$$

or in terms of displacement,

$$\frac{\partial U_r^o}{\partial r} = -(S_{11} + S_{12})p_o \qquad \frac{\partial W^o}{\partial z} = -(S_{13} + S_{23})p_o \tag{5}$$

Integrating this equation we obtain the displacement field given by,

$$U_r^o = -(S_{11} + S_{12})p_o r + C \;;\; C = 0 \; as \; U^o = 0 \; at \; r = 0 \tag{6}$$

The constant of integration vanishes due to the boundary condition. Going back to the wave equation in fibers, the boundary condition for the fiber under fluid loading denoted by p (this is the inertial loading in the form of pressure) and a static pressure p_0 are given by,

$$\begin{aligned} \sigma_{rr} &= -p - p_o \\ U_r &= U_r^w + U_r^o \\ W &= W^w + W^o \end{aligned} \tag{7}$$

Here superscript w denotes the water loading and 0 denotes the static pressure. Now the boundary conditions can be written in terms of the displacement,

$$\begin{aligned} &C_{11}\frac{\partial U_r}{\partial r} + C_{12}\frac{U_r}{r} + C_{13}\frac{\partial W}{\partial z} = -p - p_o \\ &C_{11}\frac{\partial U_r^w}{\partial r} + C_{12}\frac{U_r^w}{r} + C_{13}\frac{\partial W^w}{\partial z} + C_{11}\frac{\partial U_r^o}{\partial r} + C_{12}\frac{U_r^o}{r} + C_{13}\frac{\partial W^o}{\partial z} = -p - p_o \end{aligned} \tag{8}$$

In the above equation we take just the static pressure terms and write them in terms of the compliance coefficients, S,as

$$C_{11}[-(S_{11} + S_{12})p_o] + C_{13}[-(S_{11} + S_{12})p_o]C_{13}[-(S_{13} + S_{23})p_o] \tag{9}$$

The compliance terms can be replaced by the stiffness terms,C, using the following relations

$$\begin{aligned} S_{11} &= \frac{C_{22}C_{33} - C_{23}^2}{C} & S_{13} &= \frac{C_{12}C_{23} - C_{13}C_{22}}{C} \\ S_{12} &= \frac{C_{13}C_{23} - C_{12}C_{33}}{C} & S_{23} &= \frac{C_{12}C_{13} - C_{23}C_{11}}{C} \end{aligned}$$

Finally, for a transversely isotropic case ($C_{11} = C_{22}$) the static pressure on the right hand side is equal

to the stresses developed due to the pressure and cancels on the two sides of Eq. 8

$$= -p_o[\frac{C - C_{12}C_{33}(C_{11} - C_{22})}{C}]$$
$$= -p_o$$

It is readily seen that this analysis is valid under the following conditions,

1. Principal of superposition holds
2. Small displacement,
3. within elastic range,
4. Transversely isotropic.

Another effect of the fluid pressure is the change in the density of fluid. The pressure and fluid density are related by the relation where Ev is the Bulk Modulus of the fluid, defined by the relation.

$$E_v = \frac{dp}{d\varrho/\varrho} \tag{10}$$

The change in density of water (E_v = 2.15 GPa.) to the applied pressure is presented in Table I. This shows that very little change in fluid pressure takes place under the hydrostatic loads applied in the experiments. Now if the effect of fluid on the governing relation of wave propagation is studied it is observed that the fluid density and the longitudinal and shear wave speeds in the fiber are parts of the fluid loading terms. Hence theoretically it is possible to measure the attenuation and from it deduct the changes in the wave speed in the fiber. But experimentally this is a formidable task as the attenuation measurement is difficult and the accuracy of measurement is generally low.

σ	% change
1 MPa	0.05%
100MPa	4.65%
500MPa	23.26%

Table I Effect of Pressure on the density of water.

Now, Dr. Renee Kent has made measurement of the wave speed in a SiSc–6 fibers with the application of pressure and she has observed a reduction in wave speed. The reason for the observation is not clear. The analysis of my work shows that if we treat the material linear and transversely isotropic then there should

be no change in wave speed. The observed change can be conjectured due to the fact that the fiber is not an isotropic material but comprises of many layers. If the outer layers are not uniformly bonded to the inner core and as the pressure is applied the debonding is reduced, even though the contact if mechanical the reduction in the wave speed can be real. Another reason for this change in wave speed could be the closing of the micro–cracks in the fiber and thus increasing the apparent density of the material. Third reason for the observed change could be due to the fact that we assume that the fiber material is linearly elastic and we use the principle of superposition. Any nonlinearity in the fiber properties, or the second order nonlinear material properties could be picked up which show up as the change in velocity. The observations should be further investigated. Some suggested experiments are: 1. Change the frequency of experiments, 2. use different lengths of fiber under compression, 3. Polish the ends of the fiber so that the speckle effect of transverse displacement can be removed and pure longitudinal modes can be observed.

FIBER IN A SOLID MATRIX

We will now study the wave propagation in a fiber in a solid matrix. Wave propagation in a cylindrical fiber was first solved by Pochammer[3] and Chree [4,6]. Since then a significant work has been done in this field, see Thruston[7]. The cladded fiber problem has been solved by a few researchers, the most recent one being Simmons et al. [8]. They have all assumed that the fiber is an isotropic material. In reality the fiber is transversely isotropic and hence I have modified these equation so that the anisotropy of the fiber, and of the matrix, if need arises, can be incorporated. A brief description of the derivation is is now presented.The stress displacement relation for a transversely isotropic fiber is given by the relation,

$$\begin{Bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{yz} \\ \sigma_{xz} \\ \sigma_{xy} \end{Bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & 0 & 0 \\ C_{12} & C_{22} & C_{23} & 0 & 0 & 0 \\ C_{13} & C_{23} & C_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & C_{66} \end{bmatrix} \begin{Bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial w}{\partial z} \\ \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \\ \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \end{Bmatrix} \tag{11}$$

The Equation of motion, assuming no body forces is given by,

$$\sum_{j=1}^{3} \sigma_{ijj} = \varrho \ddot{U}_i \tag{12}$$

Combining Eq.s (11) and (12) we can obtain the differential equations for the displacement field of the motion. These equations are in the form of Bessel's equations and their solutions are very well known in terms of Bessel's functions.

The displacement equations for such a problem can be obtained from the potential theory. The displacement and stresses in the fiber can be written in terms of the Bessel's function as,

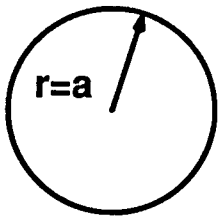
$$\begin{aligned}
U_r^r &= [-A\gamma J_1(\gamma r) - BJ_1(\eta r)ik]\ exp[i(kz - \omega t)] \\
W^r &= [AikJ_0(\gamma r) + B\eta J_0(\eta r)]\ exp[i(kz - \omega t)] \\
\tau^r &= AC_{55}[-2ik\gamma J_1(\gamma r)] + BC_{55}[(k^2 - \eta^2)J_1(\eta r)] \\
\sigma^r &= A[\tfrac{\gamma}{r}(C_{11} - C_{12})J_1(\gamma r) - (C_{13}k^2 + C_{11}\gamma^2)J_0(\gamma r)] \\
&\quad + Bik[(C_{11} - C_{12})\frac{J_1(\eta r)}{r} - (C_{11} - C_{13})\eta J_0(\eta r)]
\end{aligned} \tag{13}$$

The displacement field and stresses in the matrix can be written as

$$\begin{aligned}
U_r^c &= [-C\gamma K_1(\gamma r) - DK_1(\eta r)ik]\ exp[i(kz - \omega t)] \\
W^c &= [Ck\ K_0(\gamma r) - D\eta K_0(\eta r)]\ exp[i(kz - \omega t)] \\
\tau^c &= CC_{55}[-2ik\gamma K_1(\gamma r)] + DC_{55}[(k^2 + \eta^2)K_1(\eta r)] \\
\sigma^c &= C[\tfrac{\gamma}{r}(C_{11} - C_{12})K_1(\gamma r) + (-C_{13}k^2 + C_{11}\gamma^2)K_0(\gamma r)] \\
&\quad + Dik[(C_{11} - C_{12})\frac{K_1(\eta r)}{r} + (C_{11} - C_{13})\eta K_0(\eta r)]
\end{aligned} \tag{14}$$

here K_0 and K_1 are the modified Bessel's functions of the second kind.
The problem now has four boundary conditions,

$$\begin{aligned}
\sigma_{rr}^r + \sigma_{rr}^c + \sigma_o &= 0 \\
\sigma_{rz}^r + \sigma_{rz}^c &= 0 \\
U_r^r &= U_r^c \\
W^r &= W^c
\end{aligned}$$



When the boundary conditions are applied, a set of four equations are obtained. The terms of the 4X4 matrix are;

$$T^r_{11} = \frac{\gamma}{a}(C_{11} - C_{12})J_1(\gamma a) - (C_{13}k^2 + C_{11}\gamma^2)J_0(\gamma a)$$

$$T^r_{12} = ik[(C_{11} - C_{12})\frac{J_1(\eta a)}{a} - (C_{11} - C_{13})\eta J_0(\eta a)]$$

$$T^c_{13} = \frac{\gamma}{a}(C_{11} - C_{12})K_1(\gamma a) + (-C_{13}k^2 + C_{11}\gamma^2)K_0(\gamma a)$$

$$T^c_{14} = ik[(C_{11} - C_{12})\frac{K_1(\eta a)}{a} + (C_{11} - C_{13})\eta K_0(\eta a)]$$

$$T^r_{21} = C_{55}[-2ik\gamma J_1(\gamma a)]$$

$$T^r_{22} = C_{55}[(k^2 - \eta^2)J_1(\eta a)]$$

$$T^c_{23} = C_{55}[-2ik\gamma K_1(\gamma a)]$$

$$T^c_{24} = C_{55}[(k^2 + \eta^2)K_1(\eta a)]$$

$$T^r_{31} = -\gamma J_1(\gamma a)$$

$$T^r_{32} = -J_1(\eta a)ik$$

$$T^c_{33} = \gamma K_1(\gamma a)$$

$$T^c_{34} = K_1(\eta a)ik$$

$$T^r_{14} = ikJ_0(\gamma a)$$

$$T^r_{24} = \eta J_0(\eta r)ik$$

$$T^c_{34} = -ikK_0(\gamma a)$$

$$T^c_{44} = \eta K_0(\eta a)$$

and of course Det [T] =O is used to obtain the dispersion curves.

RESULTS AND DISCUSSION

The matrix [T] has been solved numerically. It will be noticed in these equations that the wave number k can be complex, where the imaginary part provides us with the attenuation. The Bessel's functions J and K can also be complex and this can be seen from Eq. 1. The value of the longitudinal and shear wave velocities will determine if γ and η are real or imaginary. Hence the programming is done for all terms in complex

plane.

To check the validity of the program the 4x4 matrix was used but smaller portions of it were actually used in the calculations. First, it was assumed that the cladding did not exist and so the density of the cladding was made zero and the dispersion curves were obtained for a fiber in vacuum. The results are shown in Fig. 2 by triangles. Next, it was assumed that the cladding had infinite stiffness, which simulates the condition that the fiber has the Dirichlet type boundary conditions, ie, the outer surface had zero displacement and the dispersion curves are shown as circles in Fig. 2. It is interesting to note that the fiber mode of low frequency has totally vanished in the Dirichlet fiber. Also, note that there can be no generalization made about the change in wave velocity between the two modes. Depending on the location at the dispersion curve the velocities may be higher or lower.

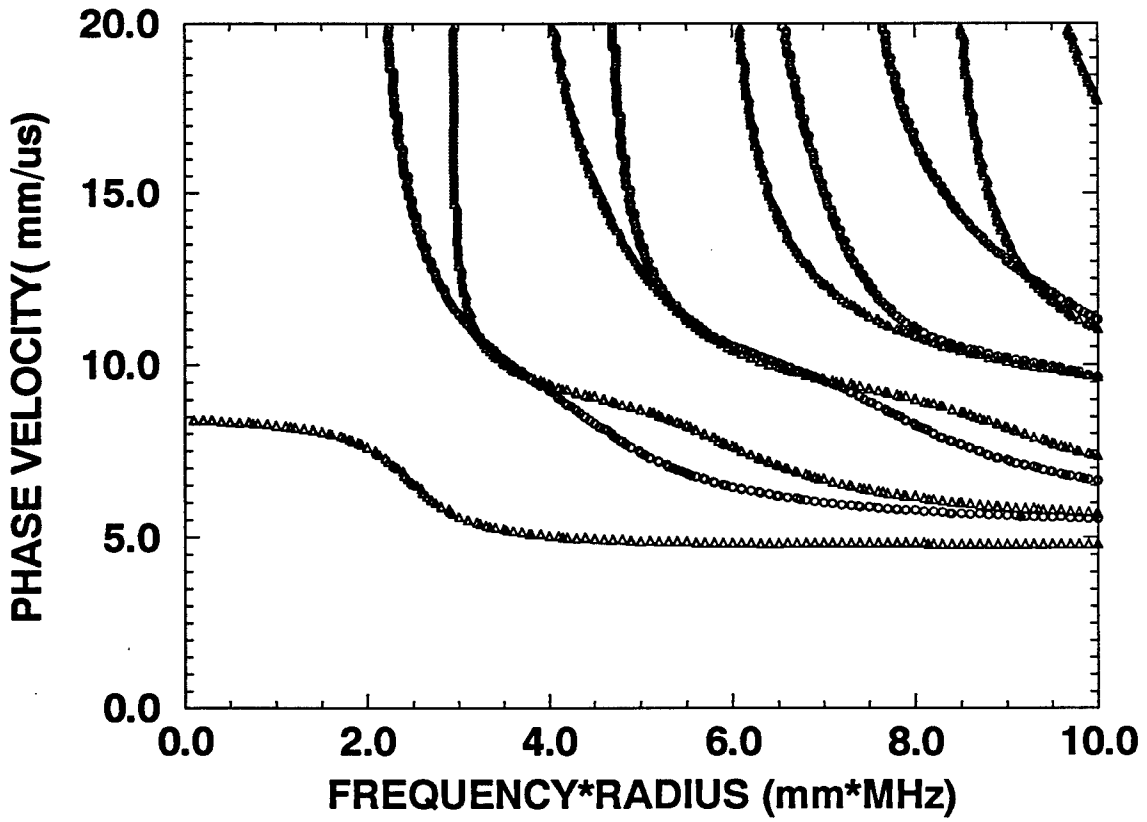


Fig.2 Dispersion curves for a free fiber (zero traction on surface) shown by triangles, and fiber under Dirichlet boundary condition (zero displacement on surface) shown by circles.

In the second set of tests it is assumed that the fiber does not exist and there is a tunnel in an infinite block of aluminum. The wave propagation along the surface of the tunnels is shown by four different modes, as shown in Fig. 3. These modes correspond to the + or – sign of the γ and η as defined in Eq. 1.

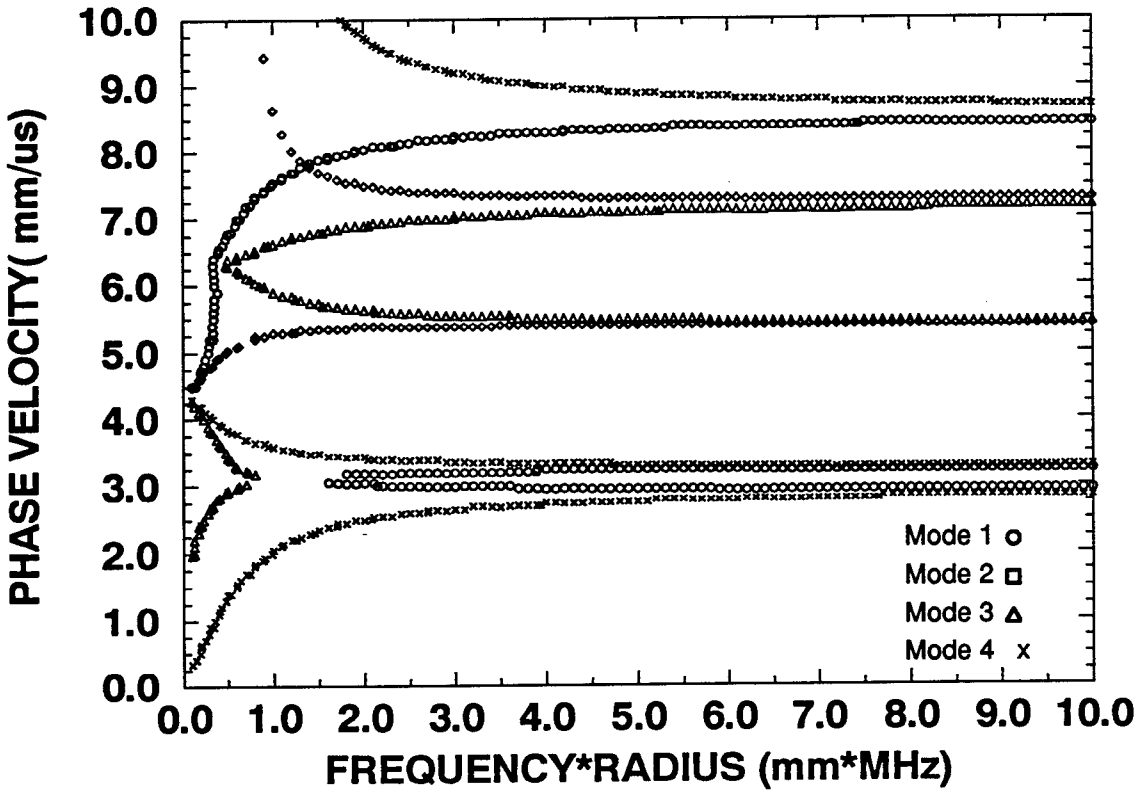


Fig.3 Dispersion curves for a tunnel in aluminum, mode1 (γ +ve and η +ve), mode 2 (γ –ve and η +ve), mode3 (γ +ve and η –ve), mode 4(γ –ve and η –ve).

Figure 4 shows the case where the surface of this tunnel is considered rigid, i.e. Dirichlet type boundary condition. In this case the velocity has a cutoff point at the longitudinal wave velocity in aluminum and rises slowly and becomes invariant with frequency. This is analogous to the Raleigh wave velocity on a plane surface. Thus, it is seen that a various interesting modes of wave travel can be generated from these equations.

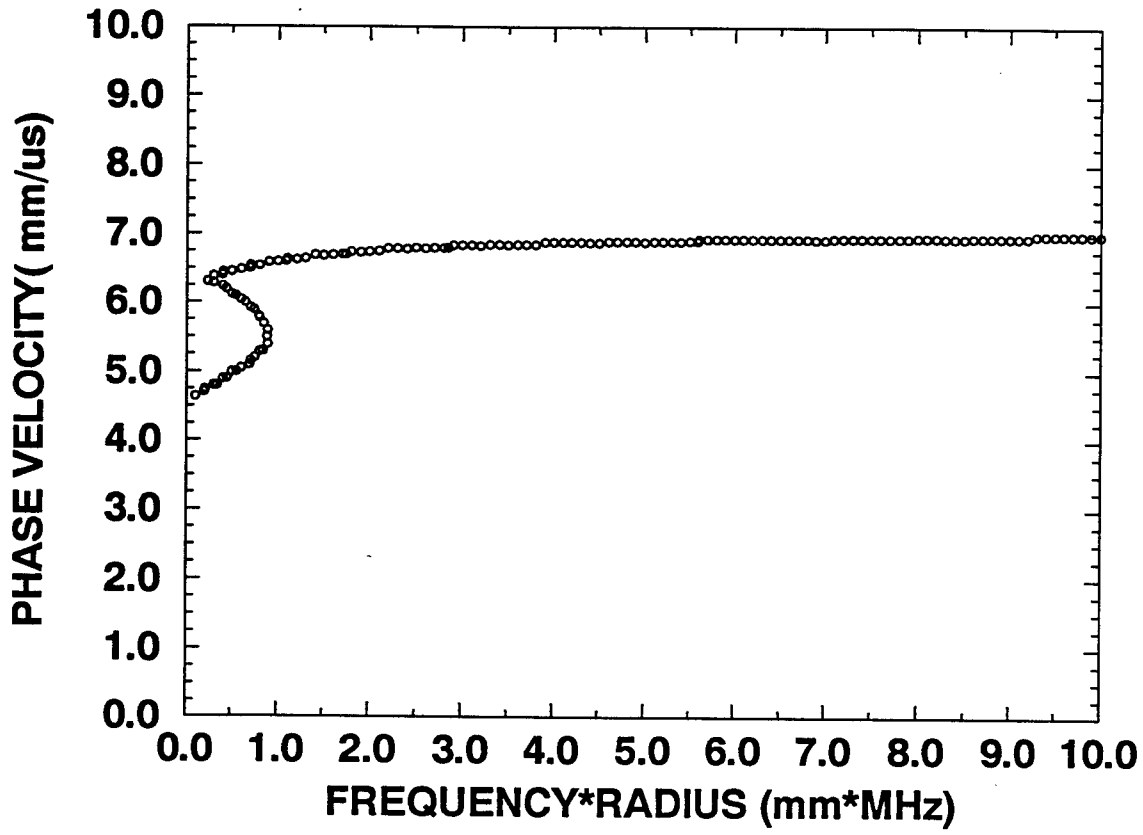


Fig.4 Dirichlet mode dispersion curve for a tunnel in aluminum.

Finally, in Fig. 5 the dispersion curves for a free fiber and a fiber embedded in the matrix are presented. It will be observed that at the low frequency range the data is scattered all over the place. Careful observation in this region shows that there are some lines emanating from the origin. It is not very difficult to show that if we take the limit of frequency tending to zero then the roots converge to the origin. Hence we can confidently state that all the dispersion curves for the embedded fiber will originate from the origin. The problem of obtaining dispersion curves is complicated due to the fact that there are many modes present close together and the convergence methods used in the software are not able to discern them. This tells us the need for some sophisticated numerical techniques and very fine search for the modes. This is not impossible and will take some hard work and careful search for the roots. At this stage of work I have not been able to obtain good convergence of the complex part of the roots and hence the trends can not be shown for the attenuation curves.

Now, as mentioned earlier, the wave propagation in the fiber enclosed in a matrix will leak energy into the matrix. This is the attenuation part of the wave number. In reality there will be two components of the attenuation, one which is due to the natural absorption of the energy by the material and the other due to the leaky part. We assume that the natural absorption component is small in comparison to the leaky part and

hence in any measurement the only the leaky part will be considered.

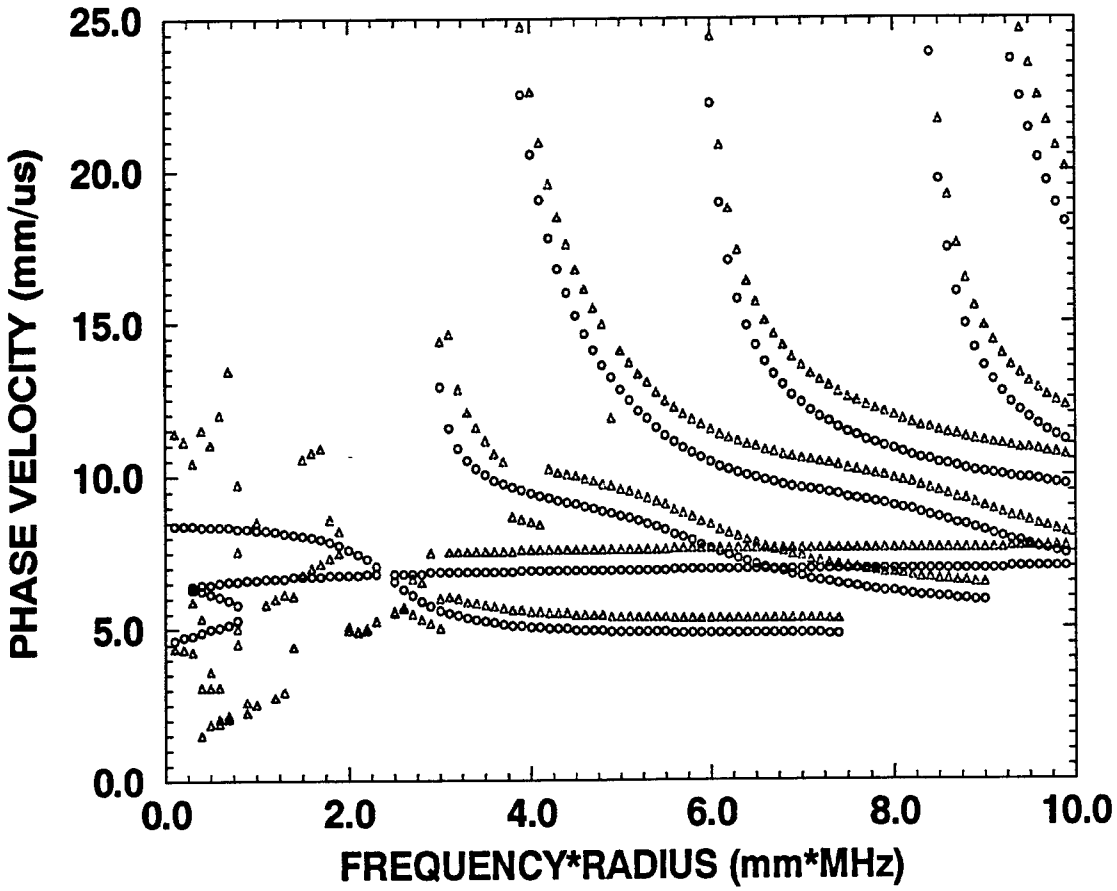


Fig. 5 Dispersion curves for a free fiber, circles, and fiber in matrix triangles.

MODELLING OF THE FIBER-MATRIX INTERFACE DAMAGE ZONE

This problem can be attempted in two different ways.

DAMAGE AS CYLINDER MODEL In this model the interface layer is assumed to be a thin layer of material whose properties are known. In this case the governing equations for the fiber and the cladding remain same as shown earlier but the interface displacement and stresses can be represented by

$$U_r^d = -[C\gamma J_1(\gamma r) + D\gamma Y_1(\gamma r) + EJ_1(\eta r)ik + FY_1(\eta r)ik]$$

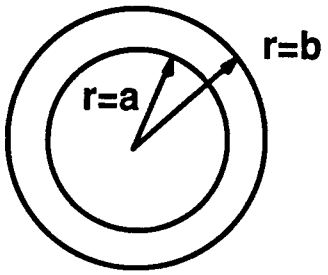
$$W^d = [CikJ_0(\gamma r) + DikY_0(\gamma r) + E\eta J_0(\eta r) + F\eta Y_0(\eta r)]$$

$$\tau_{rz}^d = C_{55}\{C[-2ik\gamma J_1(\gamma r)] + D[-2ik\gamma Y_1(\gamma r)]\} + C_{55}\left\{E[(k^2-\eta^2)J_1(\eta r)] + F[(k^2-\eta^2)Y_1(\eta r)]\right\}$$

$$\sigma_{rr}^{d} = C[\frac{\gamma}{r}(c_{11} - c_{12})J_1(\gamma r) - (C_{13}k^2 + C_{11}\gamma^2)J_0(\gamma r)]$$
$$+ D[\frac{\gamma}{r}(c_{11} - c_{12})Y_1(\gamma r) - (C_{13}k^2 + C_{11}\gamma^2)Y_0(\gamma r)]$$
$$+ Eik[(c_{11} - C_{12})\frac{J_1(\eta r)}{r} - (C_{11} - C_{13})\eta J_0(\eta r)]$$
$$+ Fik[(c_{11} - C_{12})\frac{Y_1(\eta r)}{r} - (C_{11} - C_{13})\eta Y_0(\eta r)]$$

The set of boundary conditions can now be imposed to determine the unknown constants,

1. $\sigma_r^r + \sigma_r^d = 0 \quad @r = a$
2. $\tau_{rz}^r + \tau_{rz}^d = 0 \quad @r = a$
3. $u_r^r = u_r^d \quad @r = a$
4. $w^r = w^d \quad @r = a$
5. $\sigma_r^d + \sigma_r^c = 0 \quad @r = b$
6. $\tau_{rz}^d + \tau_{rz}^c = 0 \quad @r = b$
7. $u_r^d = u_r^c \quad @r = b$
8. $w^d = w^c \quad @r = b$



Limitations of this model are,

1. The properties of the damage zone, ie its longitudinal and shear moduli should be known,
2. The thickness of the damage zone should be known, and
3. The governing equations are now 8X8 matrix and understanding the equations to get a good solution will be even more difficult.

Even then it is not impossible to obtain the dispersion curves and it will be very useful to setup these equations and study the effects of various elastic properties on the dispersion curves.

DAMAGE AS MASSLESS SPRING MODEL In this model the interface is considered as a massless spring. This formulation is similar to the work of Matikas and Prasanna [1] for the modelling of the interface. It is assumed that the interface is a massless spring of stiffness K_T and K_N. Here K_T is the shear stiffness spring and K_N is the linear stiffness spring.

We also assume the the interface is very thin layer and hence we can write the boundary conditions as

$$\sigma_{rr}^r + \sigma_{rr}^c = 0$$
$$\tau_{rz}^r + \tau_{rz}^c = 0$$
$$\sigma_{rr}^r = K_N\,[U_r^r - U_r^c]$$
$$\tau_{rz}^r = K_T\,[w_r^r - w_r^c]$$

The first two are merely the equilibrium equations but the next two relate the stiffness of the interface to the displacement jump across the layer. This formulation has an added advantage that the residual stress are included in the boundary conditions. Residual stresses, will alter the stiffness coefficients and they are also written as the displacement jump across the thin layer. The square brackets show the jump across the interface.

Based on this governing set of equations and the resulting matrix looks like;

$$\begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ T_{11} - K_N T_{31} & T_{12} - K_N T_{32} & -K_N T_{33} & -K_N T_{34} \\ T_{21} - K_T T_{41} & T_{22} - K_T T_{42} & -K_T T_{44} & -K_T T_{44} \end{bmatrix} \begin{Bmatrix} A \\ B \\ C \\ D \end{Bmatrix} = \{o\}$$

This equation can be studied in some details now. Let K_T and K_N become very small which means that the interface has no stiffness or the fiber is in the air. In this case last two equations will become the same as the first two and give us the governing matrix for the fiber in air. On the other hand when K_T and K_N are very large then we can assume that T_{11}, T_{21}, T_{12} , and T_{22} are small and hence the equations reduce to,

$$\begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ -K_N T_{31} & -K_N T_{32} & -K_N T_{33} & -K_N T_{34} \\ -K_T T_{41} & -K_T T_{42} & -K_T T_{44} & -K_T T_{44} \end{bmatrix} \begin{Bmatrix} A \\ B \\ C \\ D \end{Bmatrix} = \{o\}$$

As can now be seen that the last two equations have terms which cancel out and we are reduced to the original set of equations.

CONCLUSIONS

The governing equations for a transversely isotropic fiber in a transversely isotropic matrix are derived. The solutions of these equations is obtained and the dispersion curves can be drawn. The computer code can obtain the complex wave number and hence the attenuation of the waves, or the leakage of the waves from the fiber into the matrix can be estimated. The damage around the fiber is modelled by two methods. First, it is assumed as a thin layer of finite width with elastic properties and the dispersion equations are obtained. Second, the damage zone is modelled as a infinitesimal thin layer represented by massless longitudinal and shear springs, and the governing equations are obtained. Computational solutions of these equations is being worked out.

WORK FOR THE FUTURE

The numerical solution of these various formulations will be obtained and the complex parts of the wave equations will be calculated. These values will help us understand the wave propagation in damaged zone around the fibers and a quantitative measure of the stiffness reduction will be possible. The attenuation, or the leakage of the waves into the surrounding will tell the experimentalist which modes are useful and measurable and which are not. Damage zone stiffness reduction will be calculated and measured by ultrasonic methods and will help improve the interface to produce better composites. The work will be done in close collaboration with the researchers working at the base so that realistic values for the interface conditions can be put into the model and then experimental check of the model could be performed.

REFERENCES

1. T.E. Matikas, and P. Karpur, "Ultrasonic reflectivity technique for the characterization of fiber–matrix interface in metal matrix composites," J. Appl. Phys. **74**(1), 228–236 (1930).
2. V. Dayal, "Longitudinal waves in homogeneous anisotropic cylindrical bars immersed in fluid," J. Acoust. Soc. Am., **93**(3),1249–1255 (1993).
3. L. Pochhammer, "Uber die Fortpflanzungsgeschwindigkeinten Schwingungen in einemunbegrenzten isotropen Kreiscylinder," Z. Math. **81**, 326–336 (1986).
4. C. Chree, "Longitudinal Vibrations of a Circular Bar," Q.J.Math. **21**, 287–298 (1886).
5. C. Chree, "On Longitudinal Vibrations," Q.J.Math., **23**, 317–342 (1889).
6. C. Chree, "On Longitudinal Vibrations of Aelotropic Bars with one axis of Material Symmetry," Q.J. Math., **24**, 340–358 (1890).
7. R.N. Thruston, "Elastic Waves in Rods and Clad Rods," J. Acoust. Soc. Am., **64**, 1–37, (1978).
8. J.A. Simmons, E. Drescher–Krasicka, and H.N.G. Wadley, "Leaky Axisymmetric modes in infinite clad rods. I", J. Acoust. Soc. Am. **92**(2), Pt.1 1061–1090, (1993).

ACKNOWLEDGEMENT

I thank Dr. Tom Moran for the support and providing me the opportunity to work at the Base. Thanks are due to Dr. Prasanna Karpur, Dr. Rene Kent, Dr. Theo Matikas and Mr. Mark Ruddel for discussions and help during my work. I am grateful for the computation facilities and support by Mr. Jeff Fox, Ms. Laura Mann.

DISCRETE WAVELET TRANSFORMS FOR COMMUNICATION SIGNAL DETECTION

Jeffrey C. Dill
Associate Professor
Department of Electrical and Computer Engineering

Ohio University
329 Stocker Center
Athens, OH 45701

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

and

Armstrong Laboratory

September, 1994

DISCRETE WAVELET TRANSFORMS FOR COMMUNICATION SIGNAL DETECTION

Jeffrey C. Dill
Associate Professor
Department of Electrical and Computer Engineering
Ohio University

Abstract

This report describes the possible use of wavelet transforms for detection of communication signals. An initial study was performed to assess the possible use of the discrete wavelet transform for detection of LPI spread-spectrum communication signals. A set of MATLAB M-files were developed for this purpose. They include a "fast wavelet transform" (FWT) and its inverse, as well as a number of graphical display routines which serve to present the wavelet transform in a number of different formats, as well as to be used as a tutorial by someone who is unfamiliar with this area. The FWT is implemented in C code as a MEX file, and executes very quickly on the Sparc 10 workstation. Timing tests show it to be faster than the MATLAB implementation of the FFT for vector lengths greater than 2048. (Note that the FWT is a k N algorithm, while the FFT is an N log N algorithm.) The results of this study indicate that the discrete wavelet transform can potentially be effective in this application.

DISCRETE WAVELET TRANSFORMS
FOR COMMUNICATION SIGNAL DETECTION

Jeffrey C. Dill

Introduction

This report describes the possible use of wavelet transforms for detection of communication signals. An initial study was performed to assess the possible use of the discrete wavelet transform for detection of LPI spread-spectrum communication signals. A set of MATLAB M-files were developed for this purpose. They include a "fast wavelet transform" (FWT) and its inverse, as well as a number of graphical display routines which serve to present the wavelet transform in a number of different formats, as well as to be used as a tutorial by someone who is unfamiliar with this area. The FWT is implemented in C code as a MEX file, and executes very quickly on the Sparc 10 workstation. Timing tests show it to be faster than the MATLAB implementation of the FFT for vector lengths greater than 2048. (Note that the FWT is a k N algorithm, while the FFT is an N log N algorithm.)

Problem Statement

The purpose of this work is to evaluate the use of the discrete wavelet transform in the particular application of LPI signal detection. Toward this end, the discrete wavelet transform and its inverse were implemented using MATLAB. An additional purpose of this study was to develop tutorial materials to present the concepts of wavelet analysis to individuals who are new to the this field. The software which was developed is presented below, followed by examples of its operation in detecting traditional communication signal formats, and recommendations for further research.

Software Operation

Several MATLAB routines have been developed to evaluate the performance of the discrete wavelet transform in this signal detection application. This section provides the detailed instructions for operation of this software. The source code of these routines is included in the appendix.

FWT

To use the MATLAB implementation of the discrete wavelet transforms, perform the following steps:

1. Initialize the wavelet filter banks.

The first step is to initialize the wavelet filter bank coefficients. There are a large number of possible sets of wavelet basis functions which can be used for analysis, as described in the literature. A major area of future research will be to evaluate different wavelet bases for their appropriateness in particular applications, such as signal detection. All of the wavelet transform programs described here operate by specifying the particular wavelet filter coefficients as an input vector, in addition to the signal being analyzed. Thus the programs are general and can easily accommodate any wavelet chosen.

There are four families of wavelets currently implemented in this package. Others can be easily added by writing a small m-file to generate the filter coefficients. The four currently implemented are: Haar, Daubechies, Lemarie, and Square-root Raised Cosine.
They are invoked by entering the command:

```
[h,g]=haar(n);
[h,g]=daub(n);
[h,g]=lemarie(n);
[h,g]=sqrtrc(n,a);
```

Input arguments:

n = the length (number of taps) in the filter. Note that the order of the filter is n-1.

The following restrictions apply for n:
haar: n is always 2, and the input argument is ignored.
daub: n must be even and >=4.
lemarie: n must be even
sqrtrc: n must be odd and >4
a is the roll-off parameter, and must be in the range 0<a<=1/3. [Jones]

Output arguments:

h = a row vector of length n, containing the coefficients of the "h" (i.e. lowpass) filter.

g = a row vector of length n, containing the coefficients of the "g" (i.e. highpass) filter.

2. Perform the Discrete Wavelet Transform.

Once the wavelet filters have been chosen and initialized, the wavelet transform is invoked using the command:

y=fwt(x,h);

Input arguments:

x = the sampled data signal you wish to transform. x must be a row vector whose length is an integer power of 2.

h = the filter coefficients initialized in step 1.

Output arguments:

y = the discrete wavelet transform of x. y is returned as a row vector, of the same length as x.

The components of y represent the coefficients of the wavelet basis functions, and must be interpreted dyadically, as follows:
y(1) = the coefficient of the scaling function
y(2) = the coefficient of the mother wavelet function
y(3)-y(4) = the coefficients of the half-scale wavelet functions
y(5)-y(8) = the coefficients of the quarter-scale wavelet functions
y(9)-y(16) = the coefficients of the 1/8th-scale wavelet functions
etc.

3. Plot the DWT on a time-scale plane.

One of the major advantages of the DWT is its time-scale representation, which is not obvious from the vector output obtained from step 2 above. In order to plot the wavelet transform for visual analysis, the routine waveplot has been implemented with several useful outputs. Waveplot is invoked using the command:

y=waveplot(x,h,sf,p1,p2,p3,p4,'title')

Input arguments:

x = the input signal.
h = the h filter coefficients.
sf = the sampling frequency of x in Hz. This is used to set the time scale in the plots.
p1 = turn plot 1 on or off (p=0 turns plot off; p>=1 turns plot on)
p2 = turn plot 2 on or off (p=0 turns plot off; p>=1 turns plot on)
p3 = turn plot 3 on or off (p=0 turns plot off; p>=1 turns plot on)
p4 = turn plot 4 on or off (p=0 turns plot off; p>=1 turns plot on)
'title' = a title for marking plots (quotes included in input command)

Output arguments:

y = the DWT of x, identical to that in step 2.

Plots:

plot 1 - this plots the coefficients at each scale, with dyadic spacing so that time alignment with the input signal is achieved. It is equivalent to a series of cross-sections of the time-scale plane, one taken at each scale. This plot is the most useful, and is produced quickly.

plot 2 - this plots the actual basis functions, scaled by their coefficients, which sum to produce the input, x. It is instructive initially, to develop a basic understanding of wavelet transforms, but has limited value as an analysis tool. It is also slow, since it must compute numerous inverse transforms to produce the basis functions.

plot 3 - this plots the basis functions summed by channel. Again, it is initially instructive, but is of limited use in actual analysis, and slows down the computation significantly.

plot 4 - this is a 3-dimensional surface plot of the discrete wavelet transform, as a function of time and scale. The curves in plot 1 represent cross sections of this surface.

4. Perform the Inverse Discrete Wavelet Transform.

Once the wavelet filters have been chosen and initialized, the inverse wavelet transform can be invoked using the command:

x=ifwt(y,h);

Input arguments:

y = the wavelet coefficients which represent the discrete wavelet transform of the signal. This representation of y is input as a row vector, of the same length as x (the sampled time signal). The components of y represent the coefficients of the wavelet basis functions, and must be interpreted dyadically, as follows:

y(1) = the coefficient of the scaling function
y(2) = the coefficient of the mother wavelet function
y(3)-y(4) = the coefficients of the half-scale wavelet functions
y(5)-y(8) = the coefficients of the quarter-scale wavelet functions
y(9)-y(16) = the coefficients of the 1/8th-scale wavelet functions
etc.

h = the filter coefficients initialized in step 1.

Output arguments:

x = the sampled time signal resulting from the inverse transform of y.

Results

The following figures show the results of performing the FWT on various test signals. The routine waveplot produces a sequence of wavelet plots, one for each scale of the transform. These represent the wavelet coefficients for each scale, which are presented on a common time axis so that time alignment can readily be determined, as shown in figure 1. The wavelet transform is also shown in two dimensions as a time-scale plot, as in figure 2. Figures 1 and 2 present a simple sine wave; figures 3 and 4 present an impulsive signal; figures 5 and 6 present a chirp signal; and figures 7 and 8 present a direct sequence binary phase shift keying (BPSK) communication signal. Note in all cases that the discrete wavelet transform performs a "time-frequency" analysis, and thus preserves both time and frequency information (with the necessary compromises in resolution required by the uncertainty principle). In particular, note in figure 7 that the DWT tracks both the carrier signal (at scale 1/64) and the bit transitions (at scale 1/128) of the BPSK signal.

Conclusions

The discrete wavelet transform can be implemented such that its speed compares favorably with the fast Fourier transform. In addition, the DWT performs a time-frequency analysis, so that the time varying nature of the input signal is preserved. Thus, it is the authors belief that the DWT has potential as an LPI signal detection tool. Unanswered questions remain, however, as to the appropriate choice of the wavelet basis functions to be used for this application, and also as to the sensivity which can be achieved by this transform in a noisy environment, since the signals in question typically operate at a very low signal to noise ratio. Further research in this area is required.

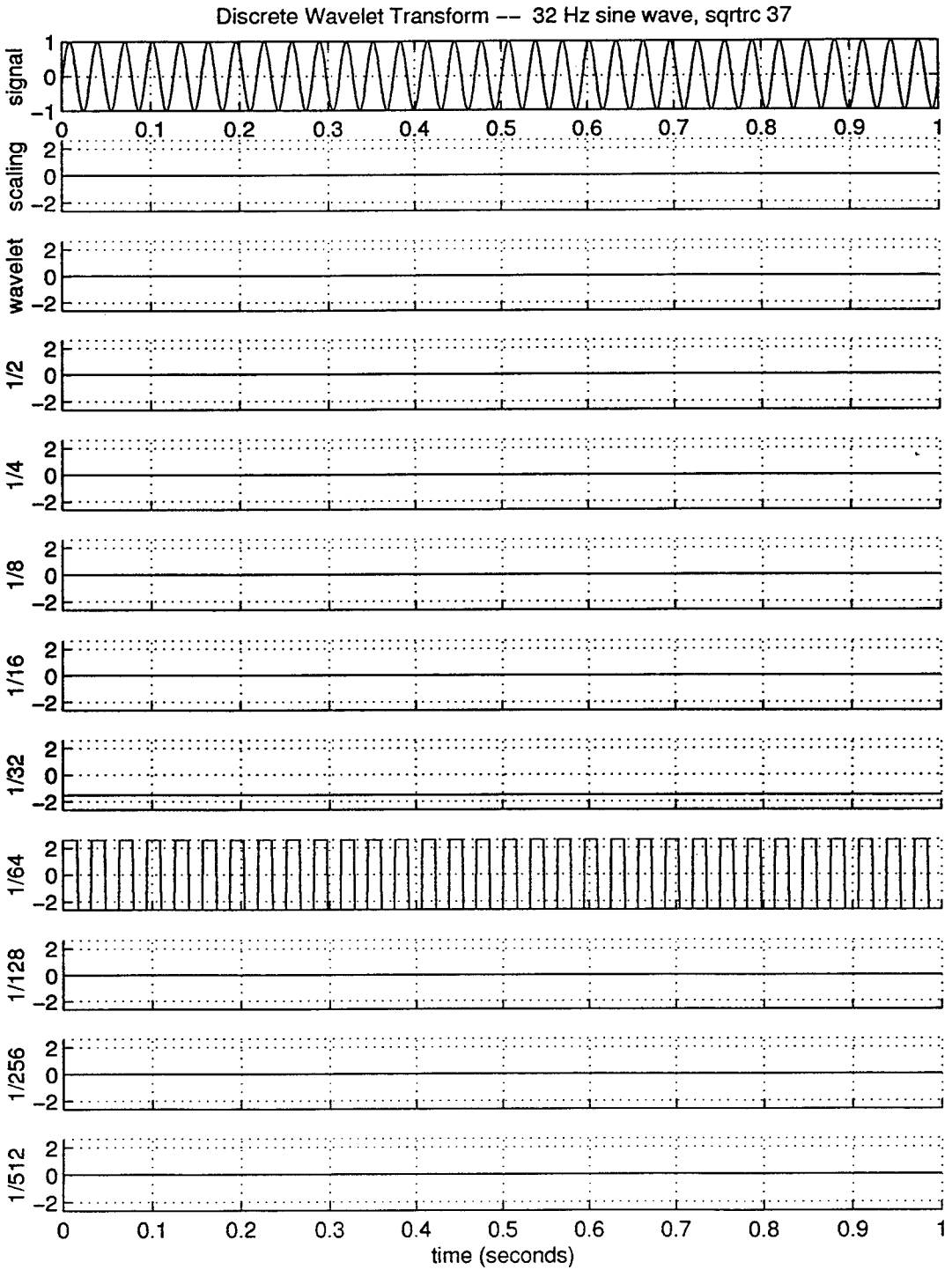

Discrete Wavelet Transform -- 32 Hz sine wave, sqrtrc 37
signal
scaling
wavelet
1/2
1/4
1/8
1/16
1/32
1/64
1/128
1/256
1/512
time (seconds)

Figure 1.

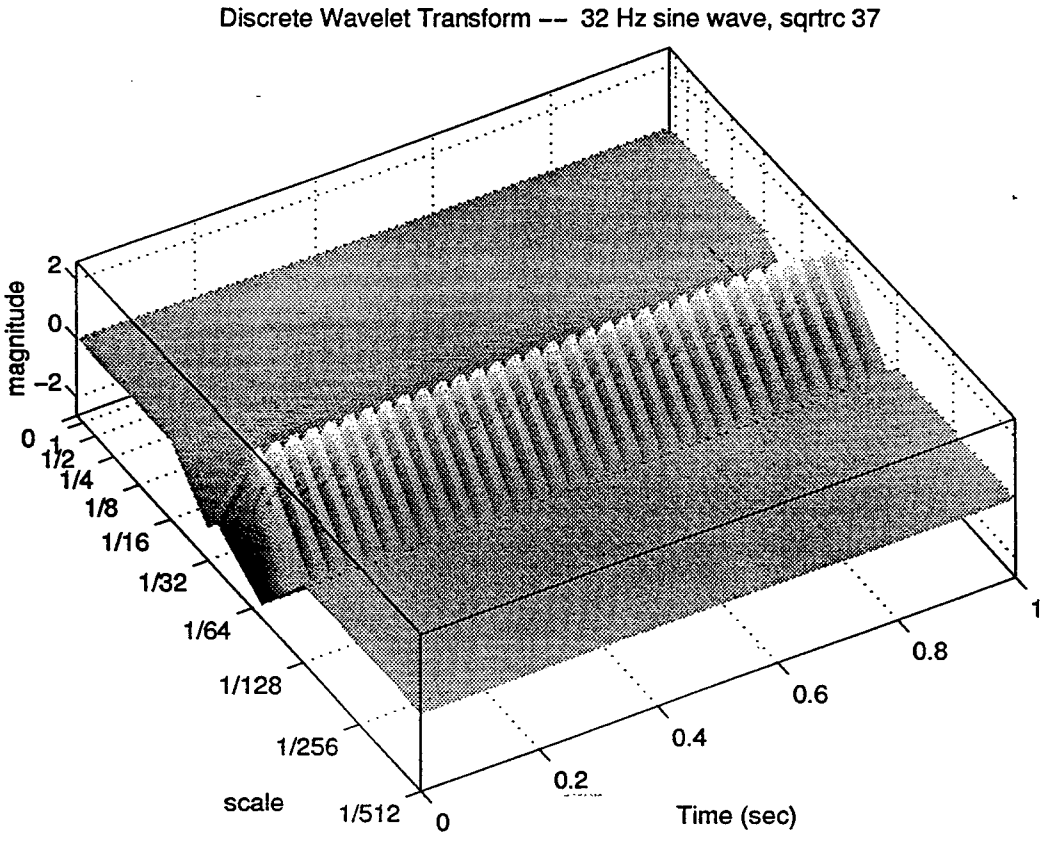

Discrete Wavelet Transform -- 32 Hz sine wave, sqrtrc 37
magnitude
2
0
-2
0
1
1/2
1/4
1/8
1/16
1/32
1/64
1/128
1/256
1/512
scale
0
0.2
0.4
0.6
0.8
1
Time (sec)

Figure 2.

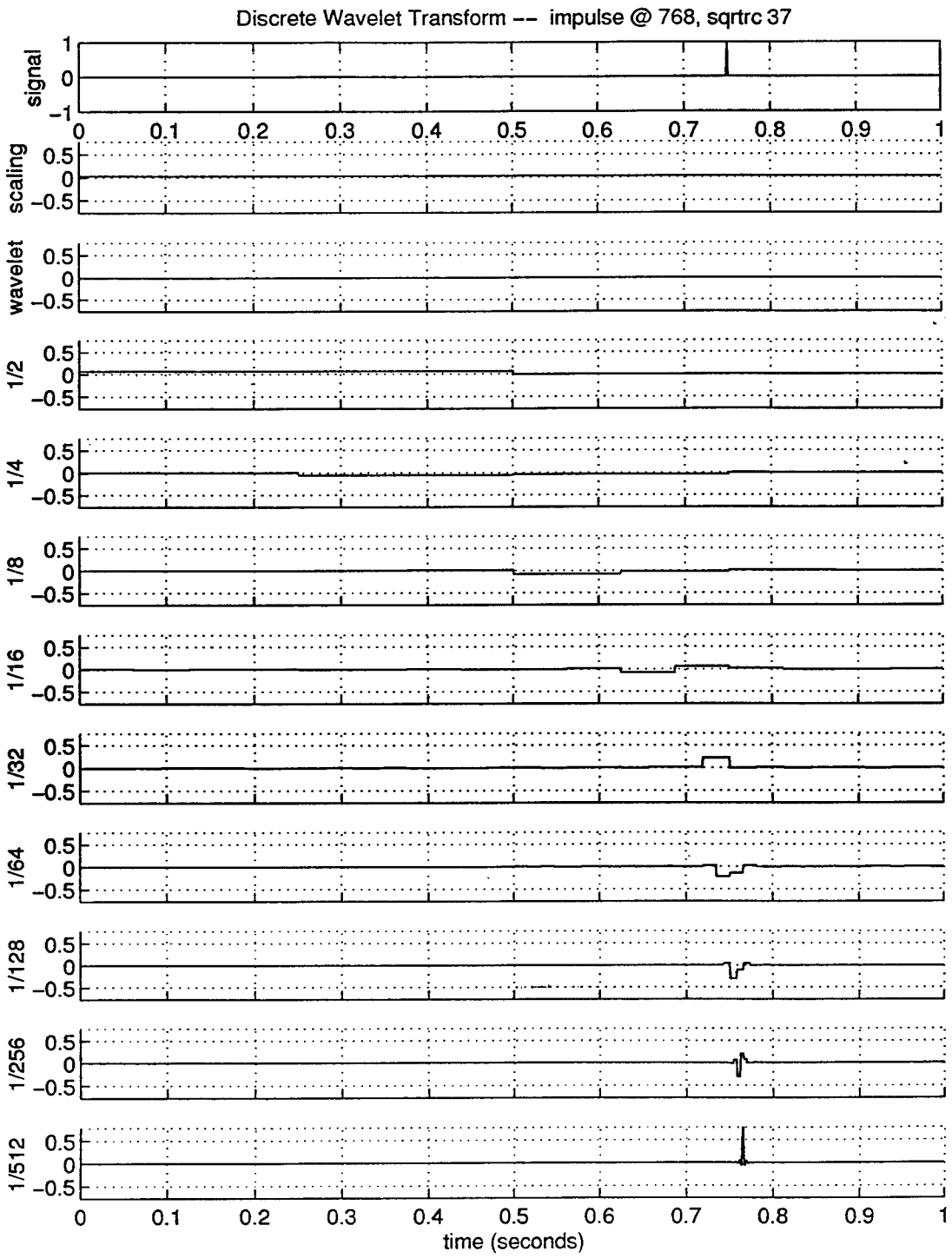

Discrete Wavelet Transform -- impulse @ 768, sqrtrc 37
signal
scaling
wavelet
1/2
1/4
1/8
1/16
1/32
1/64
1/128
1/256
1/512
time (seconds)

Figure 3.

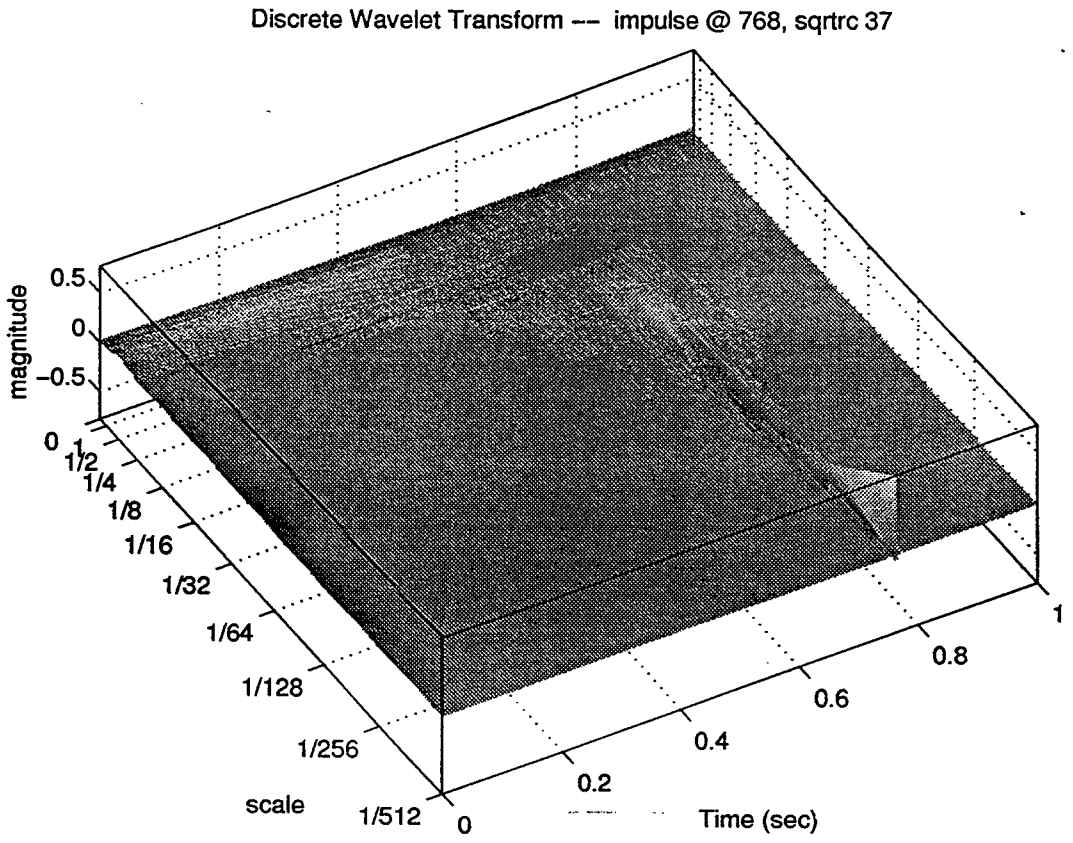

Discrete Wavelet Transform -- impulse @ 768, sqrtrc 37
magnitude
0.5
0
−0.5
0
1/2
1/4
1/8
1/16
1/32
1/64
1/128
1/256
1/512
scale
0
0.2
0.4
0.6
0.8
1
Time (sec)

Figure 4.

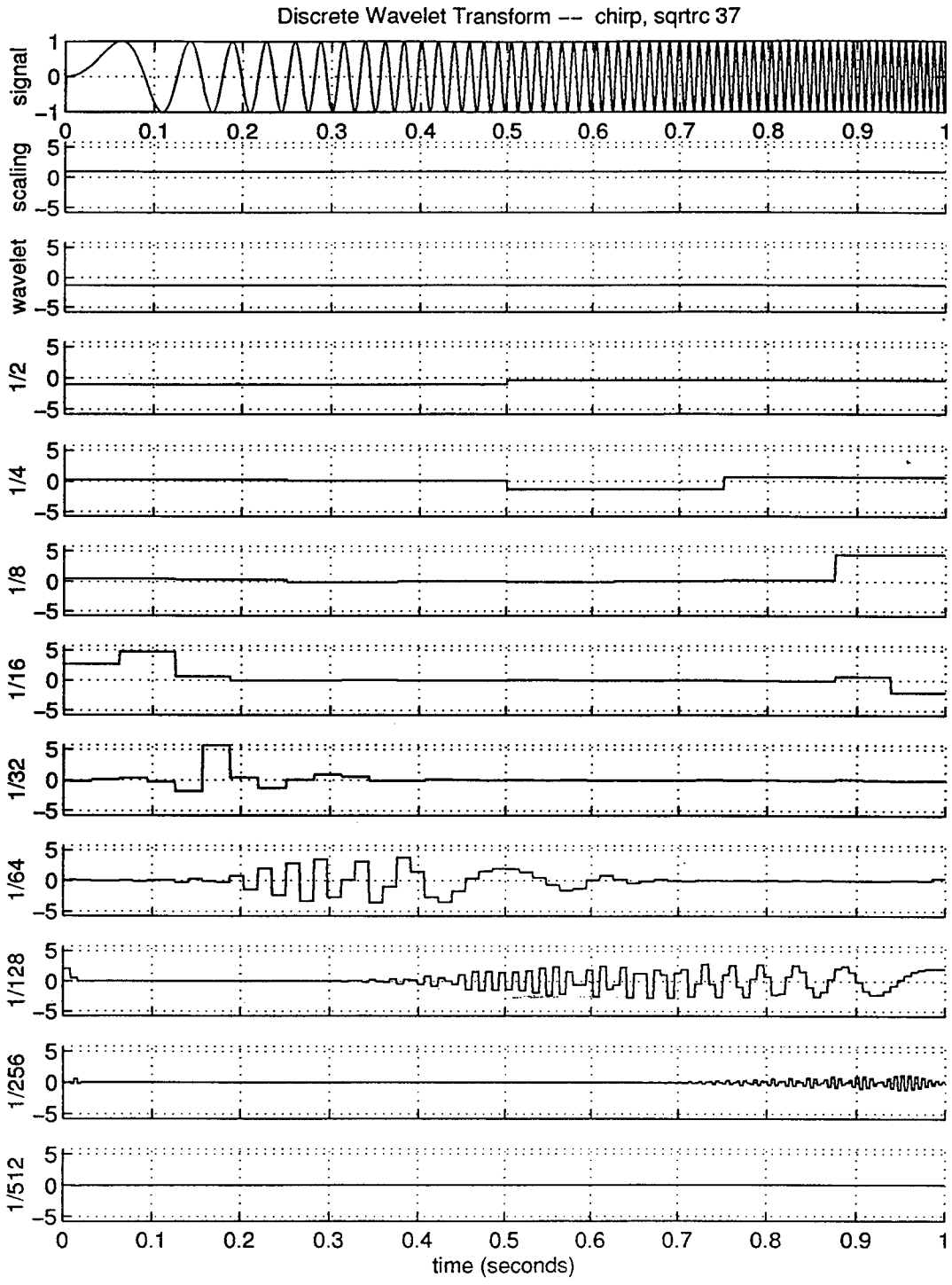

Discrete Wavelet Transform -- chirp, sqrtrc 37
signal
scaling
wavelet
1/2
1/4
1/8
1/16
1/32
1/64
1/128
1/256
1/512
time (seconds)

Figure 5.

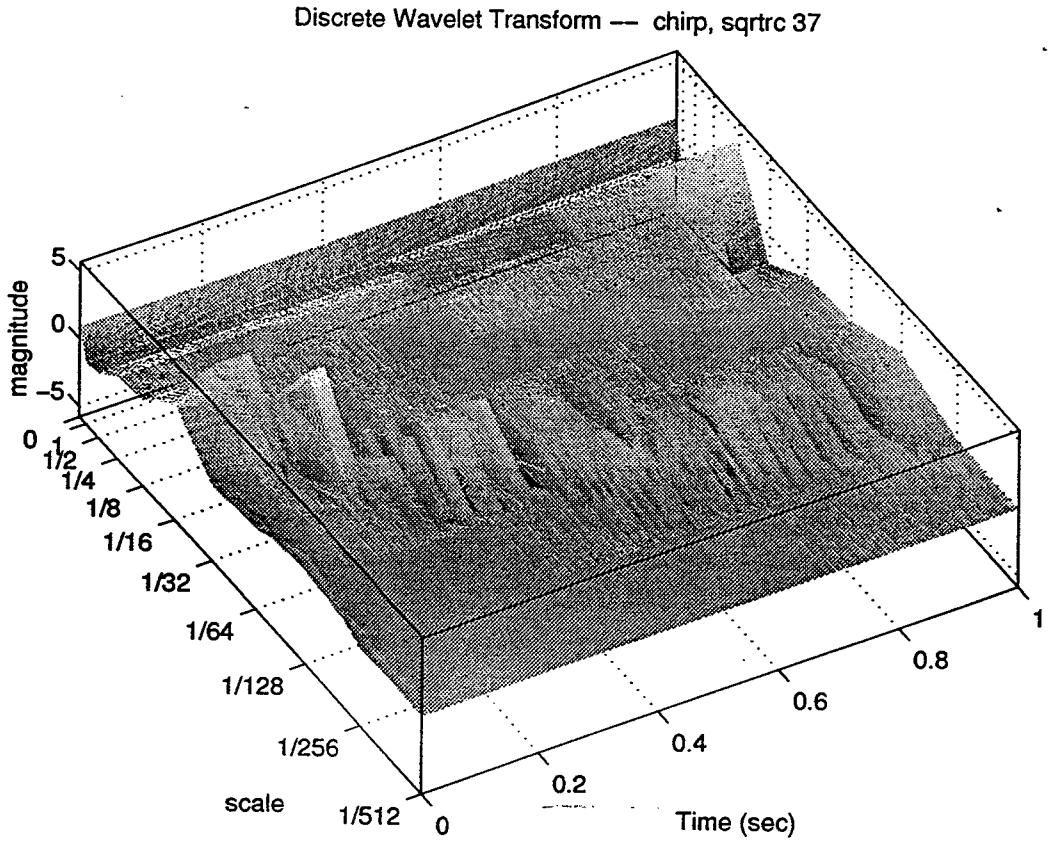

Discrete Wavelet Transform — chirp, sqrtrc 37
magnitude
5
0
−5
0
1
1/2
1/4
1/8
1/16
1/32
1/64
1/128
1/256
1/512
scale
0
0.2
0.4
0.6
0.8
1
Time (sec)

Figure 6.

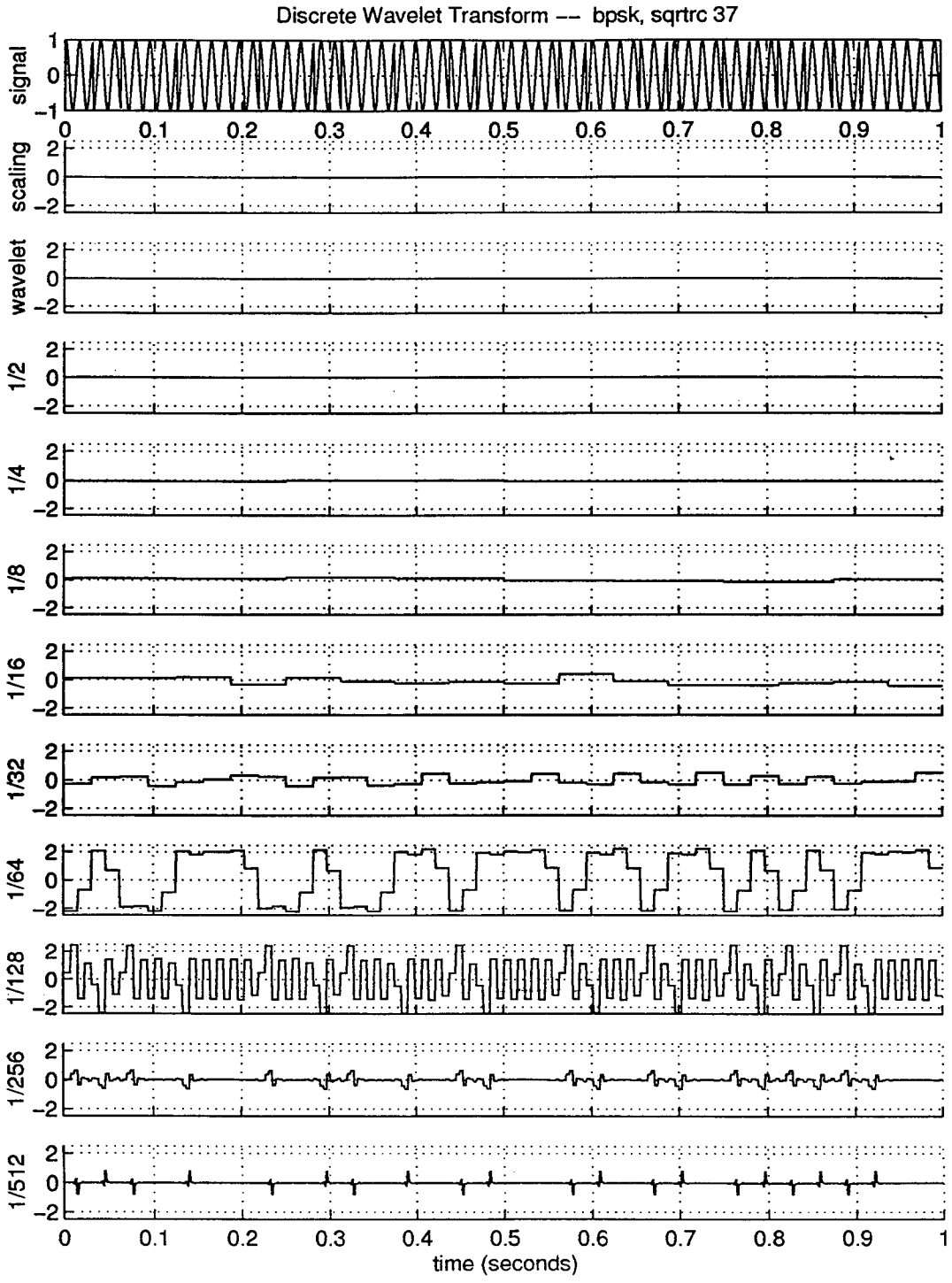

Discrete Wavelet Transform -- bpsk, sqrtrc 37
signal
scaling
wavelet
1/2
1/4
1/8
1/16
1/32
1/64
1/128
1/256
1/512
time (seconds)

Figure 7.

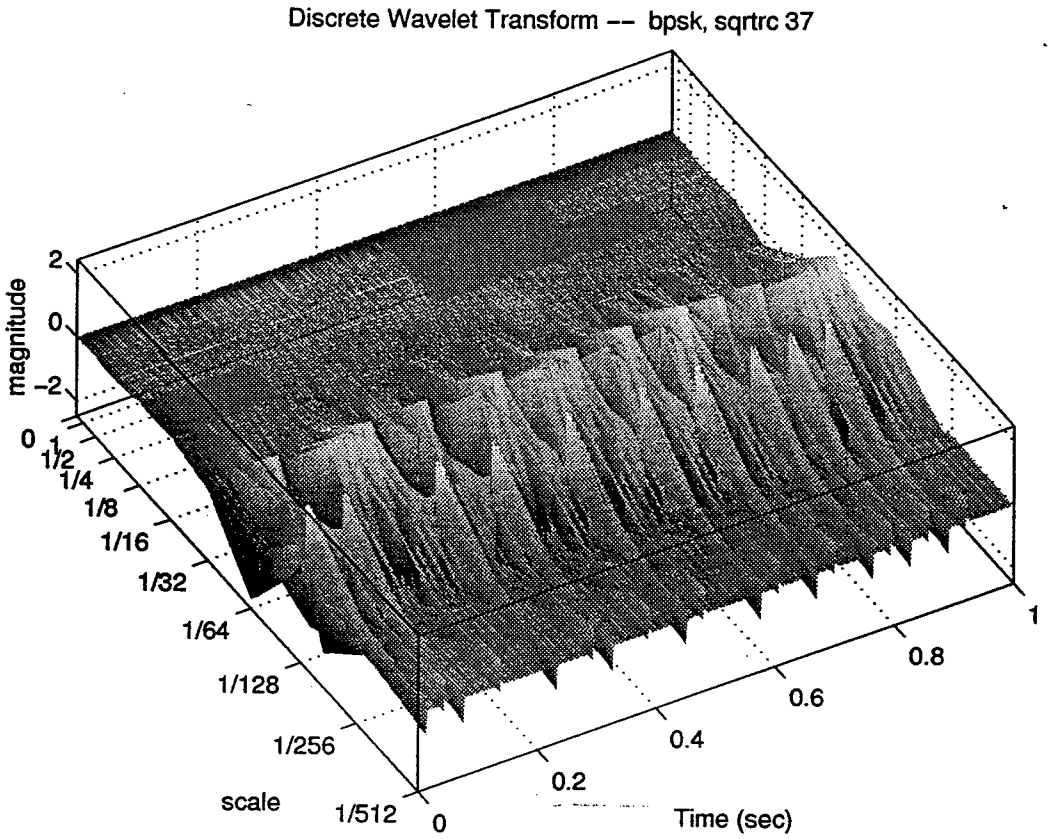

Discrete Wavelet Transform -- bpsk, sqrtrc 37
magnitude
2
0
-2
scale
1
1/2
1/4
1/8
1/16
1/32
1/64
1/128
1/256
1/512
0
0.2
0.4
0.6
0.8
1
Time (sec)

Figure 8.

References

C. Chui, An Introduction to Wavelets, Academic Press, 1992.

O. Rioul, M. Vetterli, "Wavelets and signal processing," IEEE Signal Processing Magazine, October, 1991, pp14-38.

C. Herley, M. Vetterli, "Wavelets and recursive filter banks," IEEE Transactions on Signal Processing, August, 1993, Vol 41, pp 2356-556.

C. Burrus, R. Gopinath, Introduction to Wavelets and Wavelet Transforms, ICCC ICASSP-93 Tutorial, to appear as a monograph.

W. Jones, "A Unified Approach to Orthogonally Multiplexed Communication Using Wavelet Bases and Digital Filter Banks," Ph.D. Dissertation, Ohio University, August, 1994.

Appendix

```
#include <math.h>
#include "mex.h"

#define X_INPUT         prhs[0]
#define H_FILTER        prhs[1]

#define Y_OUTPUT        plhs[0]

mexFunction(nlhs, plhs, nrhs, prhs)
        int nlhs;
        Matrix *plhs[];
        int nrhs;
        Matrix *prhs[];
{
        double *h, *g, *a, *wksp, *xin, *yout, junk;
        int ncof, nn, n, ioff, joff, nmod, n1, nh;
        int j, ii, i, ni, nj, jf, jr, sig, k;

h = mxGetPr(H_FILTER);
xin = mxGetPr(X_INPUT);
ncof = mxGetN(H_FILTER);
g = mxCalloc(ncof,sizeof(junk));
nn = mxGetN(X_INPUT);
wksp=mxCalloc(nn,sizeof(junk));
a=mxCalloc(nn,sizeof(junk));

Y_OUTPUT = mxCreateFull(1,nn,0);

yout=mxGetPr(Y_OUTPUT);
sig=(2*(ncof%2))-1;
for(j=0; j<=ncof-1; j++) {
  g[ncof-j-1]=sig*h[j];
  sig=-sig;
}
for(j=0; j<=nn-1; j++) a[j]=xin[j];
ioff=-2;
joff=-ncof+2;

for (n=nn; n>=2; n>>=1) {
  nmod=ncof*n;
  n1=n-1;
  nh=n>>1;
  for(j=0; j<=n-1; j++) wksp[j]=0.0;
  for(ii=1,i=1; i<=n; i+=2,ii++) {
    ni=i+nmod+ioff;
    nj=i+nmod+joff;
    for (k=1; k<=ncof; k++) {
      jf=n1 & (ni+k);
      jr=n1 & (nj+k);
      wksp[ii-1] += h[k-1]*a[jf];
      wksp[ii+nh-1] += g[k-1]*a[jr];
/*printf("%d  %d  %d  %f  %f  %d  %d   \n",n,ii,i,h[k-1],g[k-1],jf,jr);*/
    }
  }
  for (j=0; j<=n-1; j++)      a[j]=wksp[j];
  for (j=n>>1; j<=n-1; j++) yout[j]=wksp[j];

}
yout[0]=wksp[0];

}
```

```
#include <math.h>
#include "mex.h"

#define X_INPUT         prhs[0]
#define H_FILTER        prhs[1]

#define Y_OUTPUT        plhs[0]

mexFunction(nlhs, plhs, nrhs, prhs)
        int nlhs;
        Matrix *plhs[];
        int nrhs;
        Matrix *prhs[];
{
        double *h, *g, *a, *wksp, *xin, *yout, junk, ai, ai1;
        int ncof, nn, n, ioff, joff, nmod, n1, nh;
        int j, ii, i, ni, nj, jf, jr, sig, k;

h = mxGetPr(H_FILTER);
xin = mxGetPr(X_INPUT);
ncof = mxGetN(H_FILTER);
g = mxCalloc(ncof,sizeof(junk));
nn = mxGetN(X_INPUT);
wksp=mxCalloc(nn,sizeof(junk));
a=mxCalloc(nn,sizeof(junk));

Y_OUTPUT = mxCreateFull(1,nn,0);

yout=mxGetPr(Y_OUTPUT);
sig=(2*(ncof%2))-1;
for(j=0; j<=ncof-1; j++) {
  g[ncof-j-1]=sig*h[j];
  sig=-sig;
}
for(j=0; j<=nn-1; j++) a[j]=xin[j];
ioff=-2;
joff=-ncof+2;

for (n=2; n<=nn; n<<=1) {
  nmod=ncof*n;
  n1=n-1;
  nh=n>>1;
  for(j=0; j<=n-1; j++) wksp[j]=0.0;
  for(ii=1,i=1; i<=n; i+=2,ii++) {
    ai=a[ii-1];
    ai1=a[ii+nh-1];
    ni=i+nmod+ioff;
    nj=i+nmod+joff;
    for (k=1; k<=ncof; k++) {
      jf=n1 & (ni+k);
      jr=n1 & (nj+k);
      wksp[jf] += h[k-1]*ai;
      wksp[jr] += g[k-1]*ai1;
    }
  }
  for (j=0; j<=n-1; j++)      a[j]=wksp[j];
}
for(j=0; j<=nn-1; j++) yout[j]=a[j];
mxFree(g);
mxFree(a);
mxFree(wksp);
}
```

ELECTRO-OPTIC CHARACTERIZATION OF POLED-POLYMER FILMS

Vincent G. Dominic
Assistant Professor
Electro-Optics Program

Center for Electro-Optics
University of Dayton
300 College Park
Dayton, OH 45469-0245

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

and

Wright Laboratory

September 1994

ELECTRO-OPTIC CHARACTERIZATION OF POLED-POLYMER FILMS

Vincent G. Dominic
Assistant Professor
Electro-Optics Program
University of Dayton

Abstract

We investigated methods for measuring the electro-optic activity of poled-polymer films. In particular, we scrutinized the well-known reflection technique of Teng and Man[1] and corrected the expressions for the electro-optic coefficient as determined by that method. We show that the original expressions in ref. [1] overestimate the effective coefficient r_{33} by a factor of at least 1.32. Because of limitations in the technique of Teng and Man[1], we pursued independent electro-optic activity measurements using an interferometer. To achieve this goal, we developed a computer algorithm that utilizes a controllable, motorized Babinet-Soleil compensator in one arm of an interferometer to stabilize the fringe pattern. With a Michelson interferometer we measured the r_{13} electro-optic coefficient as well as a resonant piezoelectric and electrostrictive effect. We show that in general the ratio $r_{33}/r_{13} \neq 3$, and we discuss how the piezoelectric effect might pollute electro-optic measurements made with the reflection technique.

ELECTRO-OPTIC CHARACTERIZATION OF POLED-POLYMER FILMS

Vincent G. Dominic

Introduction

Poled-polymers are quite promising materials for applications requiring high-speed electro-optic switching or modulation. Recently, Lockheed demonstrated switches with 3 dB bandwidths of at least 20 GHz using Mach-Zender electro-optic polymer waveguides.[2] In addition to their speed, poled-polymer systems are much less expensive than lithium niobate ($LiNbO_3$) which is the current electro-optic material-of-choice. Poled-polymers are also compatible with integrated circuit processing and manufacturing technology. The great promise of this technology has spurred tremendous interest in the recent past[1-8] and active support of this research by the Air Force.

To create a poled-polymer system, we first mix a chromophore molecule (one that possesses a permanent dipole moment and whose electrons respond in a strongly nonlinear manner to an applied optical field) into a polymeric material. This mixture is heated to the glass transition temperature T_g and subjected to a large dc electric field of ~100 V/μm or more. Near the glass transition temperature the chromophore molecules acquire some freedom to rotate and consequently move into alignment with the applied electric field to relieve the torque that they experience in the non-aligned orientation. Upon cooling the sample in the presence of the field the polar alignment of the chromophore molecules is retained. Such an alignment has only one unique direction - the direction of the applied field (along which the chromophores try to align). The result is an optically uniaxial system with point group symmetry ∞mm that is electro-optic, displaying the well-known Pockels effect. This process of transforming the isotropic chromophore/polymer mixture into an aligned system is called poling.

After poling a polymer waveguide or bulk sample we wish to determine the amount of electro-optic activity that poling has imposed by measuring the electro-optic coefficients. Before the poling regimen, applying an electric field to the sample makes no linear change in the refractive index. However, the polar-aligned chromophores display Pockels effect:[9]

$$\Delta n = -\frac{1}{2} n^3 r E_{applied} \qquad (1)$$

where Δn is the change in the refractive index caused by the applied field $E_{applied}$ and r denotes the appropriate electro-optic coefficient, which determines the activity of the sample. In general the electro-

optic coefficients form a third-rank tensor with three independent elements in an ∞mm symmetry material: $r_{xxz} = r_{yyz}$, r_{zzz}, and $r_{xzx} = r_{yzy} = r_{zxx} = r_{zyy}$, where z denotes the direction of the unique axis. In the standard contracted notation form[9] we say that $r_{13} = r_{23}$, r_{33} and $r_{51} = r_{42}$. If Kleinman symmetry holds[10], then additionally $r_{13} = r_{51}$. When the chromophore molecules are considered to be solely one-dimensional and completely free to rotate then the ratio[4,7,8] $r_{33}/r_{13} = 3$ and r_{33} is the largest coefficient.

We need to accurately measure the r_{33} coefficient for a poled-polymer system. With an efficient and reliable measurement technique we may vary the parameters of sample preparation and determine the resultant effect on the electro-optic activity. Some of the important issues to address involve monitoring the magnitude of r_{33} versus: 1) poling temperature, 2) poling field, 3) readout wavelength, 4) chromophore concentration, 5) polymer/chromophore structure and 6) time (after poling) at elevated temperature. These questions must be addressed for several reasons. Answering questions 1 & 2 will determine the optimum poling conditions for a given sample. Question 5 indicates that there is still much room for clever and innovative chemistry; new system designs are emerging rapidly. Question 6 addresses the crucial requirement that the electro-optic system have sufficient thermal and temporal stability for reasonable device lifetime. Unfortunately, nature seems to impose a trade-off between the strength and the thermal stability of the nonlinearity in poled polymer systems. Stringent Air Force requirements for thermal stability of electro-optic materials imply that researchers must strive to accurately characterize the electro-optic lifetime of these materials at various temperatures. For this reason, we designed[11] a temperature-controlled environment for determining thermal stability. We conducted an initial test of the electro-optic signal decay at elevated temperatures with this apparatus and we will utilize this setup in the future for *in situ* monitoring during poling as well.

Experimental Arrangement

Our electro-optic thin film samples are not designed for waveguiding, but rather for bulk measurements of the induced activity. We start with a glass microscope slide that is coated on one side with ITO (indium tin oxide) to form a transparent conducting electrode. The ITO is masked and patterned so that it extends only about two-thirds the total length of the slide. The polymer/chromophore layer is then spin-coated onto the slide to a thickness of ~1 μm. After the polymer layer has dried gold electrodes are evaporated on top. The gold serves the dual role of both electrode and mirror. The ITO and gold electrodes overlap in a small 25 mm^2 rectangular region and the polymer is only poled in this region between the electrodes.

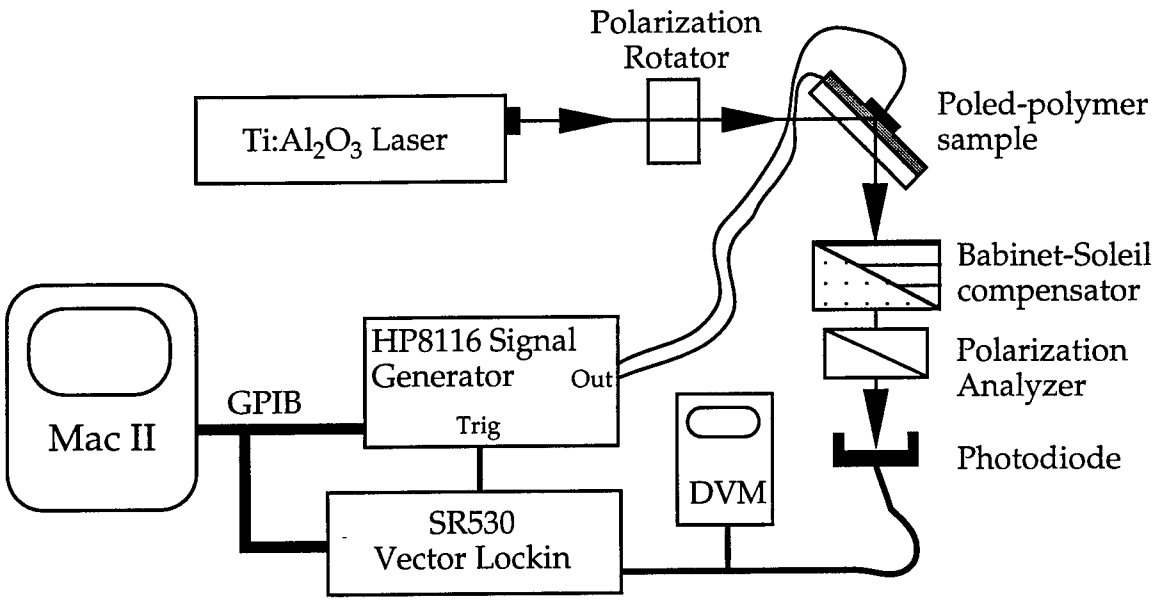


Figure 1. Reflection technique for measuring the electro-optic coefficient of a poled-polymer sample. After reflection from the sample, the Babinet-Soleil compensator changes the polarization to circular (in the absence of an applied field). The signal generator applies a sinusoidal voltage to the sample that produces small polarization changes so that the light power passing through the analyzer and detected by the photodiode/lockin is modulated.

After poling the samples, we measure the strength of the electro-optic activity by monitoring the change in the reflected polarization state caused by an applied modulation field.[1] We perform this measurement as shown schematically in Fig. 1. The Hewlett-Packard signal generator (HP8116) supplies a sinusoidal voltage (0-16V peak-to-peak) across the poled region of the electro-optic sample. We monitor the change in polarization with an analyzer/photodetector combination that feeds its signal into a Stanford Research dual channel lockin (SR530). This vector lockin amplifier expedites the measurement process since we may simultaneously determine the magnitude and phase of the modulation signal. We can use the phase of the signal to determine several important aspects of the measured effect. The digital voltmeter (monitoring the photodetector in Fig. 1) allows us to set the appropriate retardance on the compensator and also to measure the average optical power. The entire experiment is computer controlled so that after initial optical alignment of the system, we simply insert new samples into the holder to rapidly determine their EO activity.

The input polarization is initially at 45° with respect vertical and horizontal so that we have equal $\hat{s}$ and $\hat{p}$ input components. The angle of incidence is also set at ~45° and the ITO/polymer/gold layer is on the back surface of the microscope slide sample. Application of a modulating electric field causes the reflected polarization state to vary slightly about its 45°-linear steady state value. We align the analyzer by first omitting the Babinet-Soleil compensator, turning off the modulation voltage, and rotating the

analyzer to null the detector signal. We then insert the compensator and adjust its retardance to give circularly polarized light exiting the compensator. At this setting the signal measured by the photodetector should be the average of the minimum and maximum photodiode readings observed as the compensator is adjusted over a range greater than one wave of retardance. Variation of the compensator retardance Γ causes the photodiode signal I to trace out:

$$I(\Gamma) = (Max - Min)\sin^2(\Gamma/2) + Min \tag{2}$$

We set the bias retardance Γ_{bias} of the compensator to give $I(\Gamma_{bias}) = 0.5*(Max + Min)$. This should not be confused with $I_{1/2} = 0.5*(Max - Min)$ used in the calculations below, although $I_{1/2} = I(\Gamma_{bias})$ in the ideal case where $Min = 0$. Notice that we may choose Γ_{bias} so that we are either on the upslope of the $\sin^2$ curve in Eqn. (2) where $dI/d\Gamma > 0$ or on the downslope where $dI/d\Gamma < 0$.

Let us now carefully consider the beam paths followed inside the polymer sample. Because the poled polymer is optically uniaxial, the $\hat{s}$ (ordinary) and $\hat{p}$ (extraordinary) components separate slightly inside the sample. To determine the polarization state of the light that exits the sample we must keep track of the phase accumulation of each eigenpolarization component ($\hat{s}$ and $\hat{p}$) as it traverses the sample. The exiting polarization state depends on the applied voltage signal because the $\hat{s}$ and $\hat{p}$ components experience different changes in their respective refractive indices. The details of the calculation of the phase accumulation in the presence of the electro-optic effect are shown in the Appendix. Here we briefly note that the results presented in the original paper by Teng and Man[1] are erroneous. Their result was:

$$r_{33} = \frac{3\lambda}{4\pi n^2 V_{applied}} \frac{\left(n^2 - \sin^2\theta\right)^{3/2}}{\left(n^2 - 2\sin^2\theta\right)\sin^2\theta} \frac{\Delta I}{I_{1/2}} \tag{3}$$

where λ is the wavelength, n is the average refractive index, θ is the external angle of incidence, and ΔI is the peak change in the power incident on the photodetector (peak lockin signal). This answer, however, is not quite correct. Teng and Man[1] missed a portion of the path over which the optical eigenpolarizations accumulate a phase difference. We show in the Appendix that the correct determination of r_{33} is:

$$r_{33} = \frac{3\lambda}{4\pi n^2 V_{applied}} \frac{\sqrt{n^2 - \sin^2\theta}}{\sin^2\theta} \frac{\Delta I}{I_{1/2}} \frac{1}{\frac{3}{2}\left(1 - r_{13}/r_{33}\right)} \tag{4}$$

where in addition to the correction discussed above we show how deviation of the ratio r_{13}/r_{33} from the ideal value 1/3 alters the derived value for r_{33}. Of course, one must conduct a separate experiment to determine the r_{13}/r_{33} ratio.

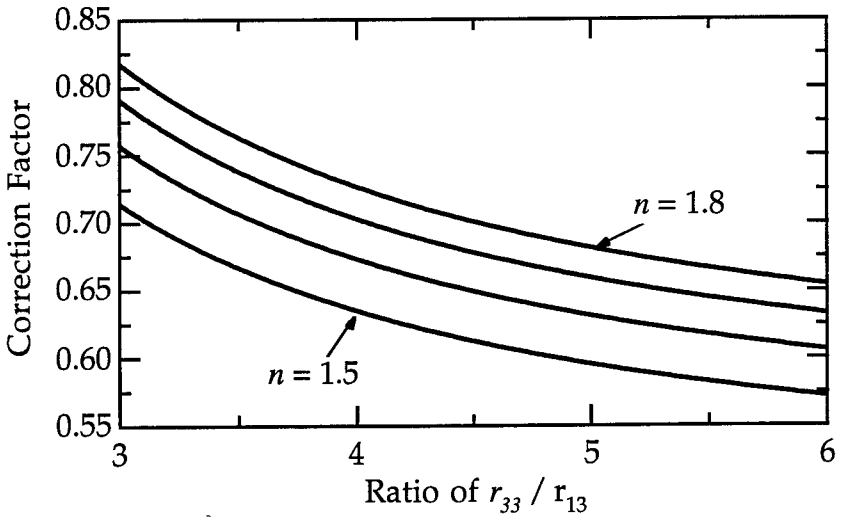


Figure 2. Correction factor that must be applied to the literature expression (ref. [1]) for r_{33} at 45° incidence.

Figure 2 shows how these two corrections affect the reported value of r_{33}, where the correction factor is defined as the ratio of the correct value {Eqn. (4)} to the literature value {Eqn. (3) and ref. [1]}. We see that in the range of typical refractive indices for doped-polymer systems ($1.5 < n < 1.7$) the correction factor lies between 0.55 and 0.8 depending on the ratio r_{13}/r_{33}. Herminghaus *et al.*[7] showed that for a one dimensional chromophore molecule the ratio r_{33}/r_{13} lies anywhere in the range $3 \leq r_{33}/r_{13} \leq 6$. The ratio $r_{33}/r_{13} = 3$ indicates that the molecule is completely free to rotate into alignment with the electric field at the poling temperature. A ratio greater than three occurs if the molecules are constrained so that they can only rotate about one easy axis (which points in a different direction for each molecule) and if there is a minimum torque required before rotation proceeds. The ratio $r_{33}/r_{13} = 6$ indicates that only those molecules whose easy rotation axis is perpendicular to the applied poling field direction can rotate.

An example of the measurement of r_{33} using the reflection technique is shown in Fig. 3, where we show the results for two different samples: #1 was made at WPAFB by doping the Lockheed chromophore DADC into 12F-PBO and #2 is a Dow Corp. sample. The Dow sample (labeled TP83A) is a thermoplastic polymer heavily doped with chromophore to give a dark red appearance. The sample shows a rather strong electro-optic effect, but unfortunately possesses too much absorption in the near infrared region to be useful. As shown in Fig. 3, we take many measurements of the electro-optic signal at different applied voltages. This provides a measure of the accuracy of our determinations and also confirms the linear dependence of Pockels effect on the applied field. The lockin signal is in phase with the applied modulation voltage if we choose to bias the Babinet compensator so that $dI/d\Gamma > 0$ and out of phase by 180° if we bias at $dI/d\Gamma < 0$. The fact that the phase flips by π depending on the

compensator bias is an important verification that the applied field is altering the refractive index (and *not* the absorption) of the polymer layer. Electroabsorption[13,14] also yields a signal that varies linearly with the applied field. However, the electroabsorption signal will not depend on the bias setting of the Babinet compensator. Indeed, we observed electroabsorption when we probed the Dow TP83A sample with a visible (Ar^+, $\lambda = 5145$ Å) laser source.

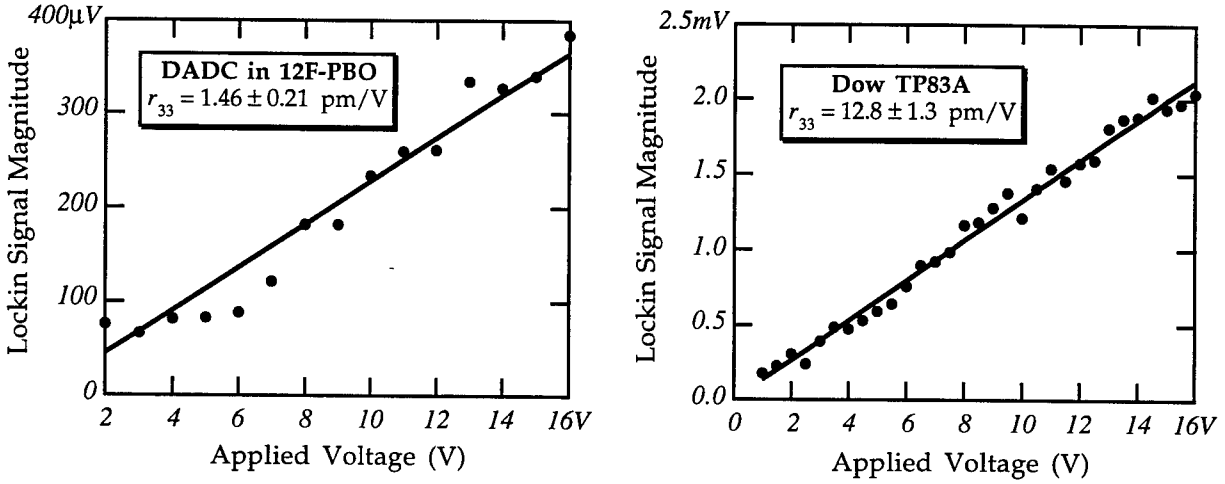


Figure 3. Electro-optic reflection technique measurement in two samples. 1) Lockheed DADC chromophore doped into 12F-PBO. This sample was prepared and poled at WPAFB. 2) Dow sample TP83A.

Interferometric Measurements

The reflection technique of Teng and Man[1] cannot independently determine the two important electro-optic coefficients r_{33} and r_{13}. As we saw above, deviations of the ratio r_{33}/r_{13} from the frequently assumed value of 3 leads to significant error in the reported value of r_{33}. A separate means of measuring these coefficients is then very useful. The standard technique involves constructing an interferometric arrangement and placing the electro-optic sample in one arm of the interferometer. Typically, a Mach-Zender configuration is utilized because one may then switch from measuring r_{33} to measuring r_{13} simply by rotating the probing polarization. We chose a Michelson interferometric setup to simplify additional measurements, described below, of the piezoelectric and electrostrictive effects. The bane of interferometric measurement setups is the problem of maintaining fringe stability for the duration of the experiment. We took all the standard precautions: the number of degrees of freedom (translation stages, rotation stages, *etc.*) were kept to a minimum, the optics were mounted low to the table with rigid mounts, the entire apparatus was enclosed, and the optical table was vibration isolated. In spite of our precautions, the interferometric fringes still tended to drift over the course of our measurements (1/2-1 hrs.). To combat this problem we utilized a motorized, computer-controlled Babinet-Soleil compensator in one arm of our interferometer and we developed a computer algorithm to stabilize the long term drift

in the output of the interferometer. Our algorithm works as follows: First we vary the retardance of the Babinet-Soleil compensator so that the photodiode voltage measured by the computer covers several cycles of the fringe visibility function. We then fit the variation of the photodiode voltage to the function:

$$V_{PD}(x) = \frac{1}{2}(V_{\max} + V_{\min}) + \frac{1}{2}(V_{\max} - V_{\min})\sin\left\{2\pi\frac{x}{\Lambda} + \phi\right\} \tag{5}$$

where x is the position of the compensator actuator, Λ is the spatial periodicity of the visibility function, V_{max} and V_{min} are the maximum and minimum photodiode voltages, and ϕ is an arbitrary phase factor. Once we have fit the data to Eqn. (5) we can easily determine the compensator actuator position x_{bias} that yields $V_{PD}(x_{bias}) = 1/2\ (V_{\max} + V_{\min})$. This is the optimum bias location for measuring fringe shifts due to the electro-optic effect because the slope of the visibility function is a maximum here. We want to insure that the interferometer remains at this half-visibility bias point throughout the duration of the experimental measurement. To do this we constantly monitor the average voltage on the photodiode and when it wanders away from $1/2\ (V_{\max} + V_{\min})$ we adjust the compensator retardance by moving the actuator a distance:

$$\Delta x = \frac{\Delta V_{PD}\,\Lambda}{\pi\,(V_{\max} - V_{\min})} \tag{6}$$

By making small, continual adjustments to the compensator retardance we can insure that the interferometer remains biased at the half-visibility point. We assume, of course, that the laser power incident on the interferometer is relatively unchanging over the course of the measurement. This requirement is reasonably well satisfied so long as the laser is warmed up before starting our drift compensation algorithm. We tested our program over several hour-long periods and found that we could completely eliminate drift in the interferometer and that the average deviation of the photodiode signal from the desired bias point was less than ±4%.

If we utilize the stabilized Michelson interferometer to measure the electro-optic coefficient, a normally incident optical beam experiences a perturbation due to r_{13} alone, with no contribution from r_{33}. When we apply a modulating voltage with a peak of $V_{applied}$ to the sample, the optical phase accumulation through the polymer layer changes thus causing the fringes to shift. The shifting fringes consequently produce a signal variation on the apertured photodiode monitoring the fringes. (We modulate the applied voltage at a high enough rate (typically 1-10 kHz) so that the fringe stabilization algorithm ignores this perturbation.) The peak (**not** RMS) lockin signal ΔV is related to r_{13} according to:

$$r_{13} = \frac{\Delta V\,\lambda}{\pi\,(V_{\max} - V_{\min})\,n^3\,V_{applied}}. \tag{7}$$

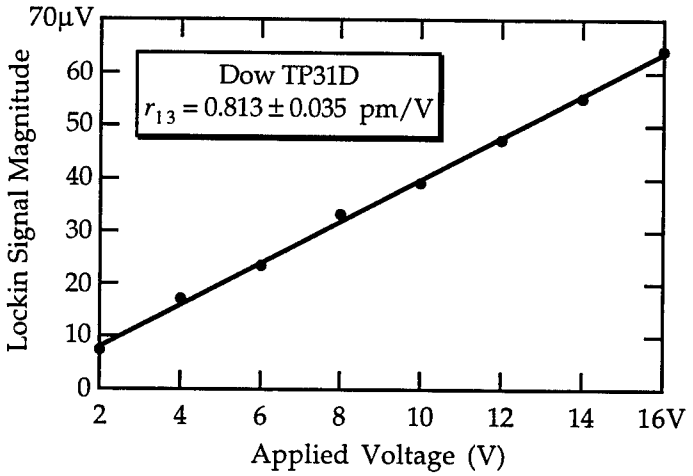


Figure 4. Electro-optic measurement using a Michelson interferometer to measure r_{13} the Dow sample TP31D.

Figure 4 shows the lockin signal variation with the peak amplitude of the modulation voltage. As in the reflection measurement, we take several readings of the electro-optic signal so that we can verify the linear dependence and accurately determine the r_{13} coefficient. The sample used for the data in Fig. 4 was a Dow sample, labeled TP31D, in which we previously measured $r_{33} = 4.37$ pm/V with the reflection technique assuming $r_{33}/r_{13} = 3$. Thus, the assumption used to determine the original value for r_{33} was wrong since $r_{33}/r_{13} = 5.38$. The correct value for r_{33} is:

$$r_{33}(correct) = \tfrac{2}{3} r_{33}(assuming\ ratio = 3) + r_{13} \qquad (8)$$

which gives $r_{33} = 3.73$ pm/V and $r_{33}/r_{13} = 4.59$. Our results are in accordance with the recent observations by Norwood *et al.*[8] that the assumption $r_{33}/r_{13} = 3$ frequently encountered in the literature is not necessarily well founded.

Piezoelectricity and Electrostriction

One other nice feature of our setup is that we can readily measure the piezoelectric displacement of the poled polymer layer, as well as any electrode attraction or electrostrictive effects that may be present.[14,15] We accomplish this by simply reversing the orientation of the sample so that the light impinges on the gold reflector/electrode first and thus does not traverse the polymer layer itself. The interferometric fringe shifts are then only sensitive to changes in the thickness of the sample and not the refractive index of the poled polymer. The piezoelectric effect causes the thickness of the polymer layer to vary linearly with the applied electric field. Thus the application of a modulating electric field at frequency ω gives rise to a lockin signal also at ω. The electrode attraction or electrostrictive effect is

independent of the absolute sign of the applied field, and thus scales quadratically with the applied voltage. This signal appears at 2ω under the same excitation conditions and is thus readily separated on the lockin amplifier.

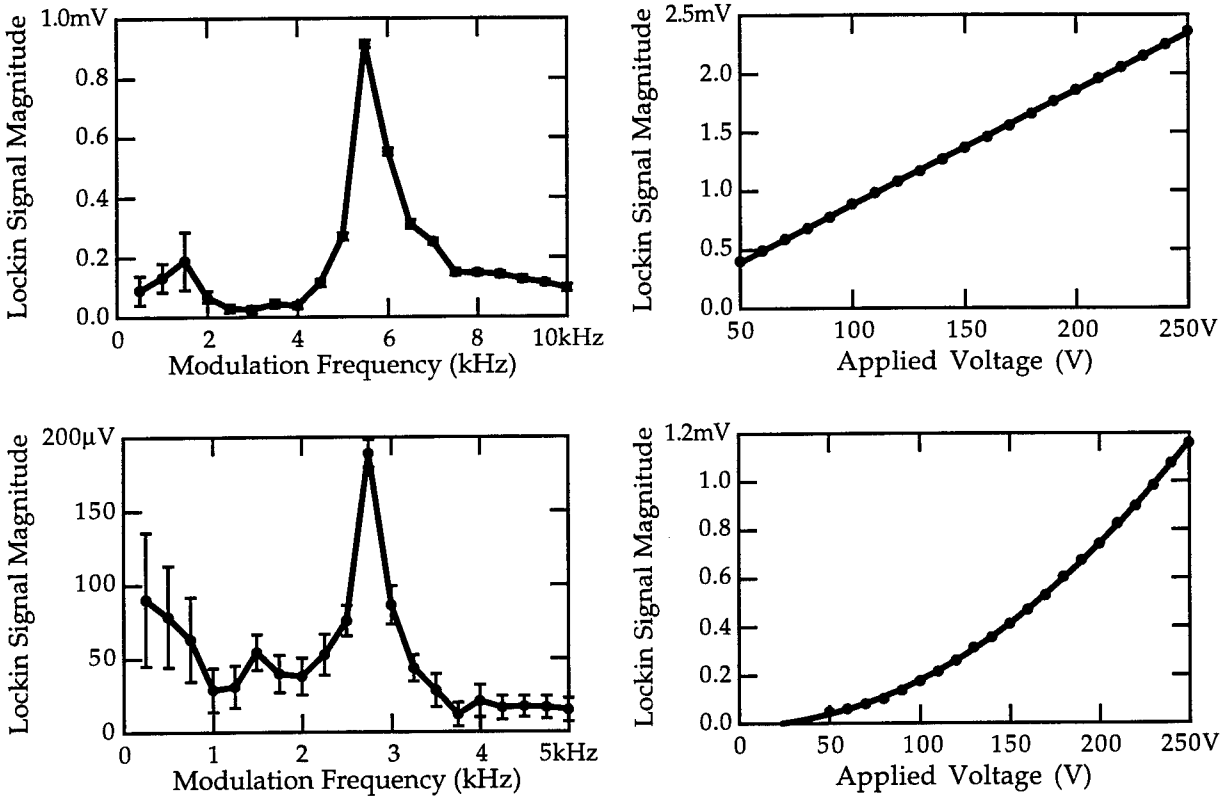


Figure 5. Piezoelectric (top row) and electrostrictive measurement in Dow TP83A using a stabilized Michelson interferometer. For the piezoelectric measurement, the lockin detects signal variations at the same frequency as the applied voltage modulation. For the electrostrictive measurements we set the lockin to monitor signals at twice the modulation frequency.

We show in Fig. 5 a measurement of the piezoelectric and electrostrictive signals from the Dow sample labeled TP83A. The sample was placed in an actively stabilized Michelson interferometer and probed with a $\lambda = 5145$ Å beam. We show both the variation of the signal as the modulation frequency is changed (from 1-10 kHz) at a fixed modulation amplitude as well as the signal variation with applied voltage for an excitation frequency at the maximum of the resonance behavior (~5.5 kHz and ~2.75 kHz for the piezoelectric and electrostrictive measurements, respectively). As expected the piezoelectric signal varies linearly with the applied field while the electrostrictive signal varies quadratically. It is important to monitor these effects, especially the piezoelectric effect, since in the normal reflection measurement geometry the reflected polarization state will be altered not only by the electro-optic effect, but also by the retardance change produced by the altered sample thickness in collaboration with the birefringence of the poled polymer. Fortunately, the piezoelectric contribution for this sample is readily avoided simply by

choosing to operate the reflection measurement at frequencies away from the resonance shown in Fig. 5. This data shows that it is prudent to make measurements at several different frequencies in order to avoid any possible contribution from resonant piezoelectric effects.

In the presence of a piezoelectric effect, how will the electro-optic measurements be affected? One can show, following the strategy discussed in the Appendix, that the piezoelectric effect will introduce a phase shift between the $\hat{s}$- and $\hat{p}$-polarized reflected component:

$$\Delta\psi_{ps} = \Delta\psi_p - \Delta\psi_s = \frac{2\pi}{\lambda} 2 \left\{ \frac{\delta n}{n} \sin^2\theta \frac{\left(n^2 - 2\sin^2\theta\right)}{\left(n^2 - \sin^2\theta\right)^{3/2}} \right\} d_{33} V_{applied} \tag{9}$$

where $\delta n = n_e - n_o$ is the birefringence, d_{33} is the piezoelectric coefficient (such that $\frac{\Delta L}{L} = d_{33} E_{applied}$), and the remainder of the symbols have the same significance as those in the Appendix. This means that the presence of a piezoelectric effect alone will give the appearance of an electro-optic effect and the apparent EO coefficient r_{33} will be:

$$r_{33}^{eff} = -3 \frac{\delta n}{n^3} \frac{\left(n^2 - 2\sin^2\theta\right)}{\left(n^2 - \sin^2\theta\right)^2} d_{33} \tag{10}$$

under the assumption that $r_{33}/r_{13} = 3$. Typical numbers[7] for poling-induced birefringences fall in the range $\delta n = 0.01$ to 0.1. If $\delta n = 0.1$, the refractive index is $n = 1.67$, and the angle of incidence is 45°, then a piezoelectric coefficient of $d_{33} = 20$ pm/V will give a signal that might be misinterpreted as an electro-optic signal with $r_{33}^{eff} = 1$ pm/V. If the birefingence is 10 times smaller then the piezoelectric coefficient must be 10 times larger to give the same effective signal. In the Dow sample TP83A the piezoelectric coefficient measured with the stabilized Michelson interferometer was $d_{33} = 8.52$ pm/V at 5.5 kHz. The reflection measurement of r_{33} gave the same value $r_{33} = 12.9$ pm/V to within 0.5 pm/V at all frequencies between 500 Hz and 10 kHz. We conclude that the poled sample birefringence δn is less than 0.12.

Conclusion

We carefully investigated different techniques to measure the electro-optic activity of a poled polymer layer. The reflection technique originally proposed by Teng and Man[1] is simple to implement and allows one to quickly process many samples. However, two warnings must be kept in mind. First, the derivation in the original paper contains a minor flaw that we have corrected here. If one uses the original equations then the derived value of r_{33} is an overestimate by ~25% (for $n \approx 1.6$). Second, the usual assumption that $r_{33}/r_{13} = 3$ is frequently violated in practice and one should independently measure both pertinent electro-optic coefficients. A Mach-Zender interferometer readily lends itself to such an

independent measurement: we simply flip the polarization from $\hat{p}$ (measures a combination of r_{33} and r_{13}) to $\hat{s}$ (measures r_{13} only). Unfortunately, interferometric measurements are susceptible to drift and noise from the interferometer itself. To this end, we built a system to stabilize an interferometer using a motorized Babinet-Soleil compensator and a personal computer to run the computer algorithm. We made use of our interferometer stabilizer to measure the r_{13} electro-optic coefficient, and the piezoelectric and electrostrictive effects in the polymer film. We found that piezoelectric and electrostrictive pollutions of the electro-optic signal are of minor importance unless the poling-induced birefringence is quite large.

Our efforts have resulted in a well defined set of procedures in place at WPAFB to investigate poled polymer samples. The system is essentially turn-key so that one may simply insert a poled-polymer sample into the sample holder and start a computer package to measure r_{33}. Measurement of the r_{13} electro-optic coefficient, and the piezoelectric and electrostrictive effects, requires alignment of an interferometer and two computers, one to stabilize the interferometer and one to acquire the appropriate signals. We have also developed an apparatus to measure the thermal stability of the aligned state and also to perform *in situ* monitoring of the poling process using second-harmonic generation. We are now in a strong position to provide most of the optical measurement capabilities necessary to advance the development of poled-polymer films with higher nonlinearities and better thermal stability.

Acknowledgments:

Jar-Wha Lee (WL/MLBP) skillfully prepared the WPAFB samples and measured the polymer layer thickness. Paul von Richter provided excellent design assistance and machining support. Jerry Landis (UDRI) cheerfully evaporated the gold electrodes onto the samples. Bob Gulotty was kind enough to send many Dow samples, and provided assistance via telephone. We also thank Bruce Reinhardt, John Zetts, Dave Zelmon, Steve Caracci, Uma Ramabadran, and Joe Binford for suggestions, support, and fellowship over the course of the summer. Finally, we thank Pat Hemenger for making this summer research possible.

References:

1. 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).
2. Results presented at WPAFB mini-symposium on Electro-Optic Polymers, August 19, 1994.
3. K. D. Singer, J. E. Sohn, & S. J. Lalama, "Second-harmonic generation in poled polymer films," Appl. Phys. Lett. **49** (5), 248-250 (1986).
4. K. D. Singer, M. G. Kuzyk, & 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).
5. K. D. Singer, *et al.*, "Electro-optic phase modulation and optical second-harmonic generation in corona-poled polymer films," Appl. Phys. Lett. **53** (19), 1800-1802 (1988).
6. M. Eich, B. Reck, D. Y. Yoon, C. G. Willson, & G. C. Bjorklund, "Novel second-order nonlinear optical polymers via chemical cross-linking-induced vitrification under electric field," J. Appl. Phys. **66** (7), 3241-3247 (1989).
7. S. Herminghaus, B. A. Smith, & J. D. Swalen, "Electro-optic coefficients in electri-field-poled polymer waveguides," J. Opt. Soc. Am. B **8** (11), 2311-2317 (1991).
8. 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).
9. A. Yariv and P. Yeh, Optical Waves in Crystals, John Wiley & Sons (New York, 1984), pp. 220-270.
10. D. A. Kleinman, "Nonlinear dielectric polarization in optical media," Phys. Rev. **126**, 1977 (1962).
11. J. L. Binford, III, "Thermal stability apparatus design and error analysis for measurements of electro-optic poled polymers," AFOSR Graduate Student Research Program Final Report, (August, 1994).
12. R. H. Page, M. C. Jurich, R. 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).
13. K. Clays, and J. S. Schildkraut, "Dispersion of the complex electro-optic coefficient and electrochromic effects in poled polymer films," J. Opt. Soc. Am. B **9** (12), 2274-2282 (1992).
14. M. G. Kuzyk, J. E. Sohn, and C. W. Dirk, "Mechanisma of quadratic electro-optic modulation of dye-doped polymer systems," J. Opt. Soc. Am. B **7** (5), 842-858 (1990).
15. H.-J. Winkelhan, H. H. winter, and D. Neher, "Piezoelectricity and electrostriction of dye-doped polymer electrets," Appl. Phys. Lett. **64** (11), 1347-1349 (1994).

Appendix: Derivation of r_{33} in the Teng & Man[1] experimental setup.

Poling a polymer film that has been doped with chromophore molecules transforms the once isotropic layer into an optically uniaxial structure. The chromophore molecules are spatially asymmetric, and the long-axis direction of the molecules tends to align with the applied electric field. The electric field is usually applied across the thin dimension of the film so that the uniaxial $\hat{c}$-axis direction is also across the thin dimension. Because the chromophores are aligned with their most polarizable direction along the $\hat{c}$-axis, we expect that $n_e > n_o$. The imposed anisotropic nature of the poled polymer film means that the extraordinary and ordinary polarization components travel different paths through the polymer layer. We show in the Fig. 6 below how the two components propagate through the sample. Note that the optical phase accumulation is:

$$\frac{2\pi}{\lambda} n \frac{2L}{\cos\alpha} \tag{i}$$

where n is the refractive index, L is the layer thickness, and α is the angle of propagation in the layer. Applying an electric field to the polymer causes a small change in the refractive index of the layer. This perturbation to the refractive index changes the direction of propagation in the polymer layer, as required by Snell's law:

$$\begin{aligned} \sin\theta &= n\sin\alpha \\ \therefore\ \sin(\theta+\Delta\theta) &= (n+\Delta n)\sin(\alpha+\Delta\alpha) \end{aligned} \tag{ii}$$

where θ is the angle of incidence of the light onto the sample. Thus, we must carefully account for both the refractive index change as well as the propagation direction change to track the phase accumulation of the optical waves traversing our sample.

Another critical point is the fact that the beam separation inside the material implies that there is a path length difference external to the polymer layer itself. The original derivation by Teng and Man[1] missed this part of the pathlength difference between the $\hat{s}$- and $\hat{p}$-polarized components. This portion of the pathlength difference occurs in a non-electrooptic region and thus applying an electric field does not alter the refractive index there. One might thus assume that it is safe to ignore this external portion of the pathlength difference. ***However***, the altered refractive index in the polymer layer changes the direction of propagation inside the polymer and this in turn affects the differential path length external to the polymer.

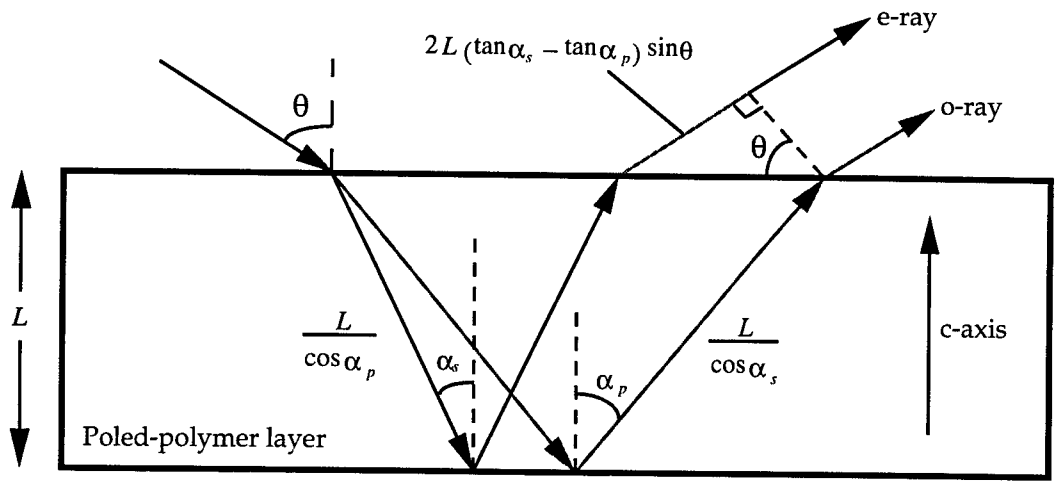


Figure 6. Ray paths for the extraordinary and ordinary components of the polarization. The total pathlength difference between the two rays includes a portion that lies external to the polymer layer itself.

The incident optical beam (45°-polarized) has equal amounts of $\hat{s}$- and $\hat{p}$-polarized components. We denote the refractive indices of these components as n_s and n_p; their corresponding propagation angles are α_s and α_p (see Fig. 6). The refractive index change of the $\hat{s}$ component is:

$$\Delta n_s = \Delta n_o = -\frac{1}{2} n_o^3 r_{13} E_{applied} \qquad \text{(iii)}$$

and the resulting change in propagation angle imposed by Snell's law is:

$$\Delta\alpha_s = -\frac{\sin\theta}{\sqrt{n_s^2 - \sin^2\theta}} \frac{\Delta n_s}{n_s}. \qquad \text{(iv)}$$

We find that the electro-optic effect alters the phase accumulation of the $\hat{s}$-polarized component by:

$$\Delta\psi_s(\text{inside}) = \frac{2\pi}{\lambda} n_s \frac{2L}{\cos\alpha_s}\left(\frac{\Delta n_s}{n_s} + \Delta\alpha_s \tan\alpha_s\right) \qquad \text{(v)}$$

where we have emphasized that this differential phase occurs entirely inside the polymer layer.

The $\hat{p}$-polarized component has an index of refraction that depends on the direction of propagation according to:

$$\frac{1}{n_p^2} = \frac{\cos^2\alpha_p}{n_o^2} + \frac{\sin^2\alpha_p}{n_e^2}. \qquad \text{(vi)}$$

This makes it a little tricky to find the change in refractive index for the $\hat{p}$ component, since n_o, n_e, **and** α_p depend on the applied electric field. One may show that:

$$\Delta n_p \;=\; n_p^3\left(1-2\,\frac{\delta n}{n_o^3}\sin^2\theta\right)\left\{\frac{\Delta n_o}{n_o^3}\cos^2\alpha_p \;+\; \frac{\Delta n_e}{n_e^3}\sin^2\alpha_p\right\} \qquad \text{(vii)}$$

where the birefringence is: $$\delta n \;=\; n_e - n_o, \qquad \text{(viii)}$$

the change in n_e is: $$\Delta n_e \;=\; -\frac{1}{2}\,n_e^3\,r_{33}\,E_{applied}, \qquad \text{(ix)}$$

and we substituted for the change in propagation angle according to:

$$\Delta\alpha_p \;=\; -\frac{\sin\theta}{\sqrt{n_p^2-\sin^2\theta}}\,\frac{\Delta n_p}{n_p}. \qquad \text{(x)}$$

The change in propagation phase for the $\hat{p}$-polarized component includes a portion from inside the polymer layer as well as some from outside the layer. We write this:

$$\Delta\psi_p(\text{total}) \;=\; \left[\frac{2\pi}{\lambda}\,n_p\,\frac{2L}{\cos\alpha_p}\left(\frac{\Delta n_p}{n_p} \;+\; \Delta\alpha_p\tan\alpha_p\right)\right]_{inside} \;+\; \left[\frac{2\pi}{\lambda}\sin\theta\left(\Delta W_s-\Delta W_p\right)\right]_{outside} \qquad \text{(xi)}$$

where the first term is the internally-produced phase change and the second term denotes the external phase change. The change in the external path is:

$$\Delta W_s \;-\; \Delta W_p \;=\; -2\,L\sin\theta\;\left\{\frac{n_s\,\Delta n_s}{\left[n_s^2-\sin^2\theta\right]^{3/2}} \;-\; \frac{n_p\,\Delta n_p}{\left[n_p^2-\sin^2\theta\right]^{3/2}}\right\} \qquad \text{(xii)}$$

Now we must put all this information together to find the differential phase shift $\Delta\psi_{ps}$:

$$\Delta\psi_{ps} \;=\; \Delta\left(\psi_p-\psi_s\right) \;=\; \Delta\psi_p-\Delta\psi_s \qquad \text{(xiii)}$$

Application of a modulating electric field to the sample changes the state of polarization of the beam reflected from the sample. This electro-optically produced change in the reflected polarization is directly related to this phase shift $\Delta\psi_{ps}$.

Now let's make several simplifying assumptions. First assume that the total birefringence is small (typical measured δn = 0.01 to 0.1):

$$\frac{2\,\delta n}{n_o^3}\sin^2\theta \;<<\; 1 \qquad \text{(xiv)}$$

where δn was defined in Eq. (viii) above. Now that we have calculated the differential phase shift to first order in the electro-optical perturbation we are free to make the additional simplification:

$$\begin{aligned} n_e \approx n_o \approx n_s \approx n_p &\Rightarrow n \\ \alpha_s \approx \alpha_p &\Rightarrow \alpha \end{aligned} \tag{xv}$$

After crunching around with the equations for a while we finally end up with:

$$\Delta\psi_{ps} = 2\frac{2\pi n^2 V_{applied}}{\lambda}\frac{\sin^2\theta}{\sqrt{n^2-\sin^2\theta}}\{r_{33}-r_{13}\} \tag{xvi}$$

where we have defined:

$$E_{applied} = \frac{V_{applied}}{L} \tag{xvii}$$

One can readily show that the differential phase shift in Eqn. (xvi) causes a change in the irradiance ΔI (peak **not** RMS) at the detector following the polarization analyzer (see Fig. 1) that is proportional to the 1/2 visibility irradiance $I_{1/2}$ according to:

$$\Delta\psi_{ps} = \frac{\Delta I}{I_{1/2}} \tag{xviii}$$

In conclusion, we determine the electro-optic coefficient from the measurement according to:

$$r_{33} = \frac{\lambda}{2\pi n^2 V_{applied}}\frac{\sqrt{n^2-\sin^2\theta}}{\sin^2\theta}\frac{\Delta I}{I_{1/2}}\frac{1}{1-r_{13}/r_{33}}. \tag{xix}$$

Please note that we have not made the assumption that $r_{33} = 3\,r_{13}$ here. An accurate determination of r_{33} depends on knowledge of the ratio r_{33}/r_{13}.

INFLUENCE OF MODEL COMPLEXITY AND AEROELASTIC CONSTRAINTS ON THE MULTIDISCIPLINARY OPTIMIZATION OF FLIGHT VEHICLE STRUCTURES

Franklin E. Eastep
Professor
Department of Mechanical and Aerospace Engineering

The University of Dayton
103 College Park Dr.
Dayton, OH 45469-0227

Jonathan A. Bishop
NSF Graduate Research Fellow
School of Aerospace and Mechanical Engineering

The University of Oklahoma
865 Asp Avenue, FH 206
Norman, OK 73019-0601

Final Report for:
Summer Faculty Research Program
Summer Graduate Research Program
Wright Laboratory

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

and

Wright Laboratory

August 1994

INFLUENCE OF MODEL COMPLEXITY AND AEROELASTIC CONSTRAINTS ON THE MULTIDISCIPLINARY OPTIMIZATION OF FLIGHT VEHICLE STRUCTURES

Franklin E. Eastep
Professor
Department of Mechanical and Aerospace Engineering
The University of Dayton

Jonathan A. Bishop
NSF Graduate Research Fellow
School of Aerospace and Mechanical Engineering
University of Oklahoma

Abstract

This investigation focused upon the structural weight optimized design of two finite element models of a fighter-type wing of low aspect ratio using ASTROS. The optimal redesign of a fighter wing with the wing structure represented by a coarse and a complex finite element model is obtained with constraints imposed on strength, control reversal, and flutter using both subsonic and supersonic aerodynamic theories. The results from the two wings are comparable for flutter analysis; however, the results differ somewhat for control reversal. The reasons for this difference are investigated. Further study of both wings using different design variable schemes is also conducted.

INFLUENCE OF MODEL COMPLEXITY AND AEROELASTIC CONSTRAINTS ON THE MULTIDISCIPLINARY OPTIMIZATION OF FLIGHT VEHICLE STRUCTURES

Franklin E. Eastep
Jonathan A. Bishop

Introduction

An aircraft structural designer must consider aeroelastic instabilities (i.e. flutter, divergence, and control reversal) in addition to the strength requirements for the structural design of a high performance aircraft. In particular, he must design a structure such that the critical aeroelastic instability velocity is at least 15% above the maximum operational flight velocity while still insuring satisfactory strength at the velocity of the critical aeroelastic instability. The critical aeroelastic instability is defined to be the lowest velocity of the flutter, divergence, or control reversal velocities. At the same time, the structural designer desires to adjust the structural sizes to minimize the structural weight.

In recent years, structural optimization as needed and used by the aerospace industry has expanded in scope to include such additional disciplines as static and dynamic aeroelasticity, composite materials, aeroelastic tailoring, etc. One of the more promising multidisciplinary codes presently under development is the Automated Structural Optimization System (ASTROS)[1-3]. In this computer code, static, dynamic, and frequency response finite element structural modules, subsonic and supersonic steady and unsteady aerodynamic modules, and an optimization module are combined and allow for either analysis or optimized design of given aircraft configurations. Interfering surface aerodynamics are incorporated to handle the aerodynamic modeling of combinations of wigs, tails, canards, fuselages, and stores. Structures are represented by fully built-up finite element models, constructed from rod, membrane, shear, plate and other elements. Static and dynamic aeroelastic capabilities include trim, lift effectiveness, aileron effectiveness, gust response, and flutter analysis. The optimization and aeroelasticity modules of this code were used as a tool for the structural optimization of fully built-up finite element wing mdels in subsonic and supersonic flow with strength as well as static and dynamic aeroelastic constraints.

This project draws heavily on a previous study by Striz, Eastep, and Venkayya[4], which studied the behavior of a coarse finite element fighter wing model under strength, flutter, and aileron effectiveness constraints. The model used in that study is shown in Figure 1a.

For this study, a geometrically similar but more detailed fighter wing model with 550 nodal points was investigated to test the design capabilities of ASTROS when applied to complex structural representations (Figure 1b). This

model was a major modification of a previous finite element model by Love[5]. The same aeroelastic properties were determined as for the coarse model using the same number of design variables to allow for direct comparison of the results. Finally, the complex model was optimized using a larger number of design variables for a few sample cases.

Background and Objectives

The importance of this investigation can be stated as follows: in a modal-type flutter analysis of fully built-up finite element wing models as used in ASTROS, the structural behavior, which depends on the sizes of the structural members, influences the flutter behavior, i.e., the flutter speed and the flutter mode shapes, as asserted by Striz and Venkayya[6]. Also, optimization is very sensitive to the types of analyses used and their assumptions. Sometimes, even minor deviations can be compounded and exaggerated in the results as pointed out by Stria and Venkayya[7]. As the wing model is optimized, the thicknesses of the structural members are adjusted in each iteration, changing the normal modes and flutter behavior. The same essentially holds for such constraints as roll effectiveness, aileron effectiveness, and strength. Therefore, an understanding of the behavior of the optimization under the action of these imposed constraints will help determine such factors as move limits, upper and lower constraint bounds, etc. Furthermore, the comparison between the two models will illustrate the effects of model complexity on behavior. This will be of value in determining whether a simplified model is sufficient for a given task, or whether a more time consuming complex model is necessary.

Numerical Results and Discussion

The two fighter-type wing structural models were selected to be representative of a wing of low aspect ratio. Here, the theory is based on an idealized wing planform with an aileron located near the wing tip. The underlying structure of each wing is represented using finite elements, which is typical of built-up structures. These models were selected for weight optimization under strength as well as static and dynamic aeroelastic constraints; the first (Figure 1a), created for the previous study, was reasonably sized to allow for parametric investigations as performed in conceptual design. The more complex second model (Figure 1b), used in this project, would likely be used in the later phases of the preliminary structural design process. The geometry and the dimensions of the models are given in Figure 2. This figure shows the coarse structural model and the aerodyn@mic model used for flutter analysis. The complex model has identical exterior dimensions. The aerodynamic model used for steady aeroelastic analysis, i.e. the 9g pullup and roll cases, is similar to the flutter model except that it extends to the aircraft centerline.

In Reference 4, the sizes and locations of the structural elements of the 10-spar, 4-rib coarse model were selected and used as a nominal structural model. The structural mass of this wing was 497 lbs. Additionally, concentrated

weights were placed at the structural nodal points to simulate non-modelled structural and non-structural masses representing fuel, actuators, and stores. The aerodynamic modeling for the nominal and optimized structure was selected to be 36 aerodynamic boxes with 6 chordwise divisions and 6 spanwise division.

The 20-spar, 18-rib complex model was extensively modified from the original model by Love so that a valid comparison between it and the coarse model could be made. The wing box thickness profile was changed to that of the coarse model, and cross bracing at hard-point locations was removed. Each type of element was sized to obtain the same mass as in the coarse design, e.g., the total mass of the ribs in the complex model is equal to the total mass of the ribs in the coarse model, etc. Within this limitation, the thicknesses of the members were tapered from root to tip, as with the coarse model. Due to rounding error in the ASTROS weight generator, which was used to find the weight of each structural group, the final structural mass of the complex wing was slightly lower than that of the coarse wing at 488 lbs. The deflection of each wing when subjected to a concentrated tip load was then found to insure that the two wings were comparable. Finally, the non-structural masses were placed on the complex wing in a pattern similar to that on the coarse wing and then moved to "tune" the complex wing so that its first few natural frequencies and mode shapes were reasonably close to those of the coarse wing.

Nominal Wing Structures

During the previous project, ASTROS was used to determine the individual stresses in the structural elements resulting from a 9-g pull-up at a Mach number of M = 0.85 at sea level. Additionally, using the same input Mach number, a flutter speed of approximately 30,100 in/sec for the nominal coarse model was found.

Finally, the roll effectiveness for a roll control system was determined for the nominal wing structure at a dynamic pressure near the control reversal dynamic pressure. The variation of aileron effectiveness of the nominal structure is displayed in Figure 3. As indicated there, the reversal dynamic pressure for the nominal coarse model at an input Mach number of M = 0.85 was approximately 45 psi.

The previous analysis was conducted in the subsonic Mach number regime (M = 0.85) using USSAERO for the steady flow and the Double Lattice aerodynamic formulation for the unsteady flow as incorporated in ASTROS. The application of these subsonic aerodynamic theories resulted in the prediction of aeroelastic instability velocities in the supersonic regime. To remove this inconsistency in problem formulation, USSAERO for the steady flow and one of the supersonic aerodynamic formulations in ASTROS, the Constant Pressure Method, were utilized in the continuation of this study. The Mach number was selected to be M = 1.2. When the nominal model was analyzed at this supersonic Mach number, a slightly lower flutter speed of 29,600 in/sec and a slightly lower reversal pressure

of 41 psi were found.

The splining method used to transfer aerodynamic loads to structural grid points which was used in this section for roll reversal analysis is considered suspect. It is believed that too many structural grid points were selected. Therefore, this portion of the project is currently being repeated with a more appropriate set of splining points. This revised set of splining points, which is similar to that used in Reference 5, has performed well in analysis of the nominal wing, but has not yet been included in the optimization runs. The new data will not all be available until after the deadline for this final report. The current data does illustrate the basic trends of the optimization.

Results for the nominal complex wing model differed, but agreed within expected limits. The flutter speed of 32,400 in/sec was 7.5% higher, indicating that the dynamic bending characteristics of the two wings are comparable.

However, the roll reversal dynamic pressure for the complex model was considerably lower than for the coarse model, at 31 psi. Since control reversal is caused by torsion due to an increased pitching moment when a control surface is deflected, it appears that the coarse wing has different torsional characteristics than the complex wing. Determining the cause of this difference occupied much of the summer research period. The conclusions are summarized later in this paper. Using the revised spline set, the roll reversal pressure was found to be 37 psi, which agrees much more closely with the coarse model results.

When the nominal complex model was analyzed at M = 1.2, two effects were observed: first, the reversal pressure decreased to 21 psi. This indicates that, unlike for the coarse wing, the pressure obtained using subsonic analysis seems to be greatly in error, since it was 50% higher than that obtained using supersonic analysis. Second, the flutter speed actually increased to 37,700 in/sec for the subsonic analysis. This does not necessarily represent an increase in the actual flutter speed of the wing; rather, it is an effect of the aerodynamic model used to calculate the aerodynamic matrices for the flutter model. Whether the flutter speed goes down or up when the formulation is changed from subsonic to supersonic cannot be predicted in advance of running the analysis. To determine the actual flutter speed of the wing, a matched-point iteration would need to be conducted.

Optimized Wing Structures

Coarse Wing

The fighter wing structural model was resized and optimized using ASTROS with both single and multiple constraints active at any given time as shown in Table 2a. These optimizations were originally performed in the

previous study, but different design variable minima were used for the strength optimization. Since the strength optimization results are used as constraints for all subsequent optimizations, the results are in most cases slightly different. Five design variables were used: one each for the ribs, spar webs, spar caps, quadrilateral skins, and triangular skins. The initial optimal structural model was obtained for a 9-g symmetric pull-up maneuver at M = 0.85 with a Von Mises stress constraint and prescribed stress yield values. The resulting optimum structure weighs 102 lbs, has a flutter speed of 17,900 in/sec, and a roll reversal pressure of 11 psi. The following optimization studies were conducted using this optimum as minimum allowable sizes for the individual structural members.

First, at M = 0.85, the structural model was resized and optimized using a constraint of an improvement of the reversal dynamic pressure from 45 psi to 52 psi. The increase in the reversal dynamic pressure was accomplished while the structural weight of the optimized wing was reduced from a nominal weight of 497 lbs to 480 lbs, as shown in Table 2a. As a by-product, the flutter speed increased to 29,500 in/sec.

As a comparison problem, with the reversal velocity used as a constraint, the structure was resized to yield the same reversal dynamic pressure as the nominal structure. This reduced the flutter velocity which, with the constraint on reversal dynamic pressure, dropped to 29,500 in/sec, as also indicated in Table 2a.

Using an increased flutter speed of 32,500 in/sec as a constraint resulted in a structure which was both lighter than the nominal wing and had a higher control reversal pressure (51 psi). This optimized wing had a weight of 440 lbs.

Increasing the flutter speed still further, to 33,000 in/sec, and requiring that the reversal dynamic pressure be at least that of the nominal model resulted in a wing which, paradoxically, is lighter than that of the previous case (424 lbs), while having a higher flutter speed. Obviously, the design ASTROS produced for the previous case is not a global optimum. This illustrates that adding constraints, even though they may not be active, can change the course of the optimization. It also reinforces the fact that numerical optimization schemes can only find local optima. The user must check these results, usually by using different initial conditions.

Finally, the structural model was resized and optimized using multiple constraints. In this case, it was required that the reversal dynamic pressure be the same as that for reversal of the nominal structure and that the flutter speed be increased to 31,000 in/sec. This increase of flutter speed is beyond that obtained when only a single constraint on reversal pressure was imposed. In this manner, the structural designer has the added advantage of precisely placing the velocity of certain aeroelastic instabilities relative to other aeroelastic instability velocities. Here, it was desired to make improvements in the flutter velocity while the structure is being weight optimized. In this particular case, the weight of the optimized structure was obtained as 382 lbs.

As with the analysis of the nominal coarse wing, the subsonic formulation used in the initial optimization predicted aeroelastic instabilities in the supersonic regime. Therefore, the optimizations were also evaluated for a supersonic input Mach number. The selected constraints for M = 1.2 were similar to those for M = 0.85. The respective reversal pressures, flutter speeds, and weight reductions for the optimized structure are again shown in Table 2a. They essentially confirm the trends found for M = 0.85 except that the minimum weights tended to be higher than for the subsonic cases. The exception to this was the optimization for nominal reversal pressure and increased flutter speed. Here, the stricter flutter constraint was the active constraint for both M = 0.85 and M = 1.2. Since the only effect of the supersonic aerodynamics is to shift the center of pressure from the quarter-chord line to the half-chord line, the flutter behavior of the wing should be nearly identical in both cases, causing the optimized designs to be essentially equal in weight.

Complex Wing

In order to provide a good comparison, the complex wing was optimized with the same types of constraints as the coarse wing. First of all, since the nominal performances of the wings were different, equivalent constraints were developed based on the percentage increases. Alternatively, the constraints from the coarse wing were applied directly to the complex wing. Both of these methods were used in this optimization study.

As with the analysis results, the spline set used for roll reversal analysis is suspect. Thus, the roll reversal results should be considered tentative.

When optimized for a 9-g pullup, the optimal complex wing weighed 151 lbs vs. 102 lbs for the coarse wing. The same number of design variables and the same linking scheme were used for both cases. The weight difference is probably due to the large element sizes in the coarse model, which "smear" the stresses and do not capture localized concentrations. The complex model, which does capture these concentrations, requires thicker elements to keep the stresses below material limits. Despite weighing more than the coarse wing, the flutter speed is slightly lower (17,500 in/sec). The roll reversal pressure for this case was 11 psi.

Next, the wing was optimized for an increase in reversal pressure to 36 psi. This is an increase of 15.5% over the reversal pressure for the nominal complex wing, the same percentage increase that was used for the coarse wing. The weight of this optimized wing was 392 lbs.

During this optimization, and all subsequent optimizations for control effectiveness, ASTROS had difficulty converging to the optimum. It often oscillated between a feasible and an infeasible design on successive iterations

rather than smoothly converging. To minimize this effect, the design variable move limits were manually restricted. Also, the program never actually identified the optimum; it simply oscillated between two nearly identical weights.

The complex wing was then optimized for the same (rather high) absolute reversal pressure as the coarse wing (52 psi). For this case, the structural weight increased greatly, to 950 lbs. As a by-product of the stiffer structure, the flutter speed increased to 54,700 in/sec. The high final weight seems to indicate that ASTROS was not able to converge to a reasonable minimum.

When the complex wing was optimized for a reversal pressure of 31 psi (the reversal pressure for the nominal wing), the optimized structure had a weight of 325 lbs. This is slightly better than the 349 lbs optimum for the coarse wing. At the same time, this optimization raised the flutter speed to 32,900 in/sec, whereas, for the coarse wing, the flutter speed dropped to 29,700 in/sec. Optimizing the complex wing for the same reversal pressure as the coarse wing, 45 psi, resulted in a design weighing 590 lbs with a flutter speed of 43,800 in/sec.

The complex wing was then optimized for a flutter speed of 34,900 in/sec, corresponding to an 8% increase, the same as for the coarse model. This optimization reduced the structural weight of the wing to 378 lbs, better than the 414 lbs optimum for the coarse design. Optimizing the complex wing for the same flutter speed as the coarse wing, 32,500 in/sec, produced a design with a weight of 334 lbs and a reversal pressure of 31.6 psi.

When the complex wing was optimized for a 9.6% increase in flutter speed with the same control reversal pressure as for the nominal wing, an optimal structural weight of 369 lbs was obtained. Flutter was the driving constraint in this design.

A different result was obtained when the complex model was optimized for a reversal pressure of 45 psi (nominal for the coarse wing) and an increased flutter speed of 33,000 in/sec. For this case, roll reversal was the driving constraint. The resulting design weighed 668 lbs and had a flutter speed of 43,300 in/sec, which is not consistent with the earlier case where a reversal pressure constraint of 45 psi was used without a flutter speed constraint, resulting in a structural weight of 590 lbs.

As with the corresponding coarse wing case, the inclusion of the flutter speed constraint (although not active) directs the optimization to a different optimum. This phenomenon also highlights the fact that numerical optimization cannot guarantee that the absolute minimum structural weight will be found; it only finds local minima. The designer must check the results, usually by optimizing beginning with different initial designs.

Finally, the wing was optimized for a 3% increase in flutter speed and for the nominal reversal speed. The optimum structure had a weight of 331 lbs. This represents substantial weight savings over the 466 lbs optimum for the coarse wing. When optimized for the same reversal pressure and flutter speed as the coarse wing (45 psi and 31,000 in/sec), the optimum structural weight was 614 lbs. For this case, the flutter constraint was not active, so the increase in weight seems to be due to the increased reversal pressure.

As with the coarse wing, the complex wing was also optimized using a supersonic aerodynamic formulation. The results of these analyses are given in Table 2b. In general, the weights were similar to those for the subsonic case when flutter was the driving constraint and higher than those for the subsonic case when roll reversal dominated. The exception to this was the case where a reversal pressure of 45 psi was used with no flutter constraint. The supersonic design weighed 306 lbs, while the subsonic design weighed 590 lbs.

A major deviation from the general trend was observed for the last case in Table 2b. For this case, ASTROS was not able to converge to a reasonable structural minimum weight but asked for ever-increasing member sizes to satisfy the constraints. The flutter constraint was not active, but, again, an inactive constraint influenced the optimization.

Further Study

To this point, all results for the coarse and complex wings have the simple, five-design variable linking scheme. A few additional cases, using both wings, were tested to determine the effects of using a slightly more complex scheme. For the coarse wing, the substructure (ribs and spars) was not designed; the optimum thicknesses from the corresponding heavily linked results are used. For the skins, each top element was linked with its corresponding bottom element. Thus, there are 31 skin design variables.

When optimized for strength during a 9g pullup, the optimum structural weight was found to be 101 lbs. This is a trivial overall weight savings over the heavy linking scheme. However, the actual structure of this design is different; some skin panels are thicker than those in the heavily linked optimum, and some are thinner.

Optimizing for a flutter speed of 32,500 in/sec at M=0.85, a design with a weight of only 303 lbs was found. The corresponding heavily linked design weighs 440 lbs. Thus, there is a definite benefit to allowing each skin panel to be optimized separately. With the heavy linking scheme, all skin panels are sized based on the most critical element; here, each element is sized based only on local conditions. At M=1.2, the optimum design weighs 400 lbs; again, this is substantially lighter than that obtained with heavy linking.

For the complex model, optimizing every skin panel would result in an excessively large number of design variables. Therefore, more extensive linking was used, and the substructure was also designed. As with the simple scheme, the linking was along structural function lines, i.e., ribs, spars, spar caps, and skins. Unlike with the previous scheme, not all elements of a specific type were linked to one variable. Rather, elements were linked to variables in three-bay increments. Since there are 18 bays on the wing, there are 6 design variables of each type, for a total of 24 design variables (quadrilateral skins and triangular skins were not separated).

It would be expected that this scheme would result in lighter optimum structures than the five-variable scheme, which forces overdesign of all elements linked to a particular variable if the most critical element is inadequate in the nominal model. The more complex linking schem$ limits this tendency toward overdesign to only the particular bay containing the critical element.

When the 24-variable scheme was used to optimize the complex wing for a 9-g pull-up maneuver at M = 0.85, an optimum structural weight of 126 lbs was obtained. This weight was 16.6% lighter than the previous optimum design. However, as might be expected, the performance in flutter and roll reversal was worse than for the five-variable scheme. The control reversal pressure was slightly lower, at 10 psi. The flutter speed is 16,800 in/sec, which is also slightly less than the flutter speed for the wing optimized with five design variables.

Finally, the 24-variable wing was optimized for its flutter behavior. A flutter speed of 34,900 in/sec was chosen as the constraint. The resulting wing weighed 307 lbs (vs. 377 lbs for the five-variable wing), with a control reversal pressure of 30 psi (compared to 35 psi).

From these basic results, it can be seen that using a more complex design variable scheme seems to result in lighter optimum structures. However, the performance of these structures in areas other than those for which they were optimized may be poor. Therefore, all aeroelastic constraints should be included to ensure that none violate safety limits. For the given cases, the number of design variables used did not have a great effect on the computational expense of an analysis. Therefore, it may often be worthwhile to use more design variables. On the other hand, the variables should be linked along physical lines so that the resulting design can actually be manufactured.

Conclusions and Recommendations

The examples presented in thisinvestigation demonstrate that the optimization capabilities of the ASTROS procedure are well suited for the preliminary design environment. Any number of constraints of strength, divergence, control reversal, and flutter can be imposed on general finite element structural models of flight vehicles. The ability to

simultaneously consider many constraints from each of several disciplines allows the structural designer to develop non-intuitive solutions to the complex design problem placed on modern flight vehicle structures.

This investigation focused upon the structural weight optimal design of two models of a fighter-type wing of low aspect ratio. The optimal weight redesign of the wing structure was obtained with imposed constraints on strength, control reversal, and flutter, using both subsonic and supersonic aerodynamic theories. In general, the weight savings (at least for strength and flutter) were greater for the complex model than for the coarse model, when both used five design variables, and greatest for the complex wing using 24 design variables.

REFERENCES

1. Neill, D.J., and Herendeen, D.L., "ASTROS Enhancements, Volume I -- ASTROS User's Manual", **WL-TR-93-3025**, Flight Dynamics Directorate, Wright Laboratory, March 1993.

2. Johnson, E.H., and Venkayya, V.B., "Automated Structural Optimization System (ASTROS), Volume I - Theoretical Manual", **AFWAL-TR-88-3028/I**, Air Force Wright Aeronautical Laboratories, December 1988.

3. Neill, D.J., Johnson, E.H., and Herendeen, D.L., "Automated Structural Optimization System (ASTROS), Volume II - User's Manual", **AFWAL-TR-88-3028/III**, Air Force Wright Aeronautical Laboratories, December 1988.

4. Striz, A.G., Eastep, F.E., and Venkayya, V.B., "Influence of Static and Dynamic Aeroelastic Constraints on the Optimal Structural Design of Flight Vehicle Structures", **Proceedings, 32nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference**, Baltimore, Maryland, 1991, pp. 470-476.

5. Love, M.H., Barker, D.K., and Bohlmann, J.D., "An Aircraft Design Application Using ASTROS", **WL-TR-93-3037**, Flight Dynamics Directorate, Wright Laboratory, June 1993.

6. Striz, A.G. and Venkayya, V.B., "Influence of Structural and Aerodynamic Modelling on Flutter Analysis", **Proceedings, 31st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference**, Long Beach, California, 1990, pp. 110-118.

7. Striz, A.G. and Venkayya, V.B., "Influence of Structural and Aerodynamic Modelling on Optimization with Flutter Constraint", **Proceedings, 3rd USAF/NASA Symposium on Recent Advances in Multidisciplinary Analysis and Optimization**, San Francisco, California, 1990.

TABLE 1a. Geometrical, Material, and Environmental Property Model Data

	COARSE DESIGN LOW ASPECT RATIO WING M = 0.85 and M = 1.2, Sea Level	
Constraints:	Strength (Von Mises), Reversal, Flutter	
Input:	Shear panel thickness:	0.08" in ribs 0.075" to 0.03" in spars
	Membrane thickness:	0.25" to 0.04" in skins
	Spar cap cross-sectional area:	1.0 to 0.5 in^2
	Spar stiffener cross-sectional area	0.05 in^2 (not designed)
	All values decreasing from root to tip	
Material:	E = 1.0*E7 lb/in^2, $\nu = 0.33$, $\rho = 0.1$ lb/in^3	
	Allowable stresses: 60.0*E3 lb/in^2 (tension and compression), 40.0*E3 lb/in^2 (shear)	

TABLE 1b. Geometrical, Material, and Environmental Property Model Data

	COMPLEX DESIGN LOW ASPECT RATIO WING M = 0.85 and M = 1.2, Sea Level	
Constraints:	Strength (Von Mises), Reversal, Flutter	
Input:	Shear panel thickness:	0.015" in ribs 0.062" to 0.005" in spars
	Membrane thickness:	0.25" to 0.046" in skins
	Spar cap cross-sectional area:	1.6 to 0.92 in^2
	Spar stiffener cross-sectional area	0.006 in^2 (not designed)
	All values decreasing from root to tip	
Material:	E = 1.0*E7 lb/in^2, $\nu = 0.33$, $\rho = 0.1$ lb/in^3	
	Allowable stresses: 60.0*E3 lb/in^2 (tension and compression), 40.0*E3 lb/in^2 (shear)	

TABLE 2a. WEIGHT OPTIMIZED WING WITH VARIOUS CONSTRAINTS
Coarse Wing, M = 0.85 and M = 1.2, Sea Level

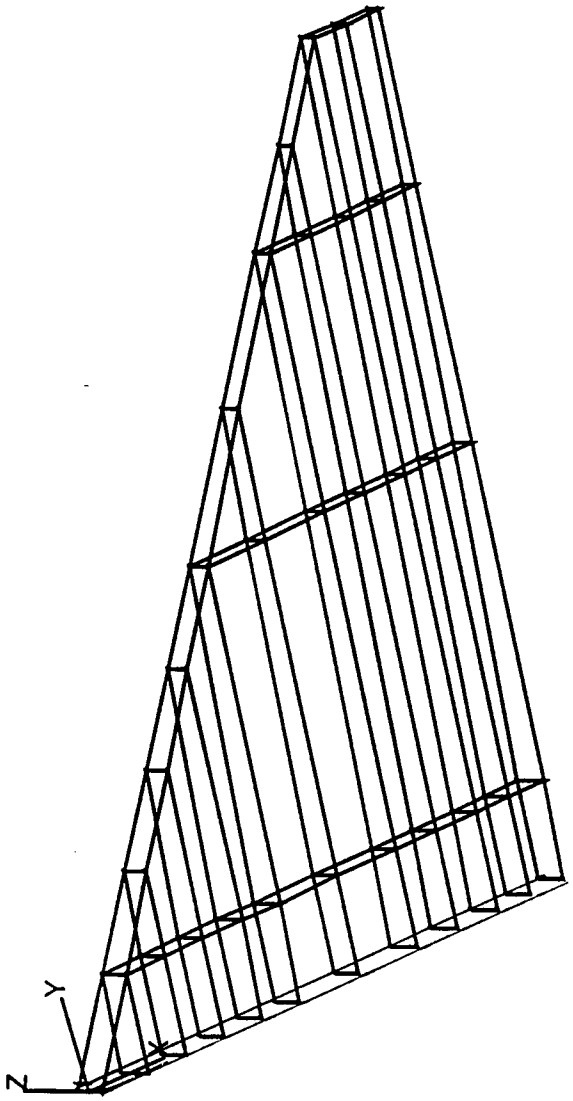
MODEL	MACH NUMBER	REVERSAL-q (#/in^2)	FLUTTER SPEED (in/sec)	STRUCTURAL WEIGHT (#)
Nominal	M = 0.85	45	30,100	497
	M = 1.2	41	29,600	497
Optimized (Strength)	M = 0.85	11	17,900	102
	M = 1.2	11	30,400	102
Optimized	M = 0.85	52*	32,400	480
	M = 1.2	52*	27,304	381
Optimized	M = 0.85	45*	29,500	398
	M = 1.2	41*	25,600	323
Optimized	M = 0.85	51	32,500*	440
	M = 1.2	53	32,500*	510
Optimized	M = 0.85	45*†	33,000*	424
	M = 1.2	41*†	33,000*	493
Optimized	M = 0.85	45*†	31,000*	382
	M = 1.2	41*†	31,000*	453

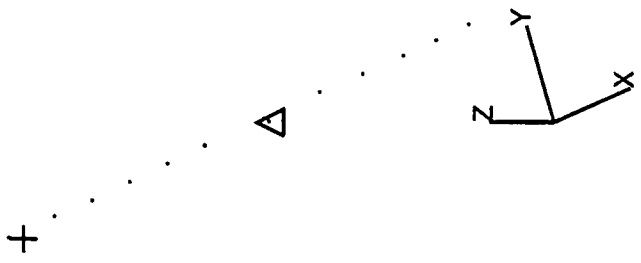
* indicates quantity was a constraint
† indicates constraint was not active in the optimized design

TABLE 2b. WEIGHT OPTIMIZED WING WITH VARIOUS CONSTRAINTS
Complex Wing, M = 0.85 and M = 1.2, Sea Level

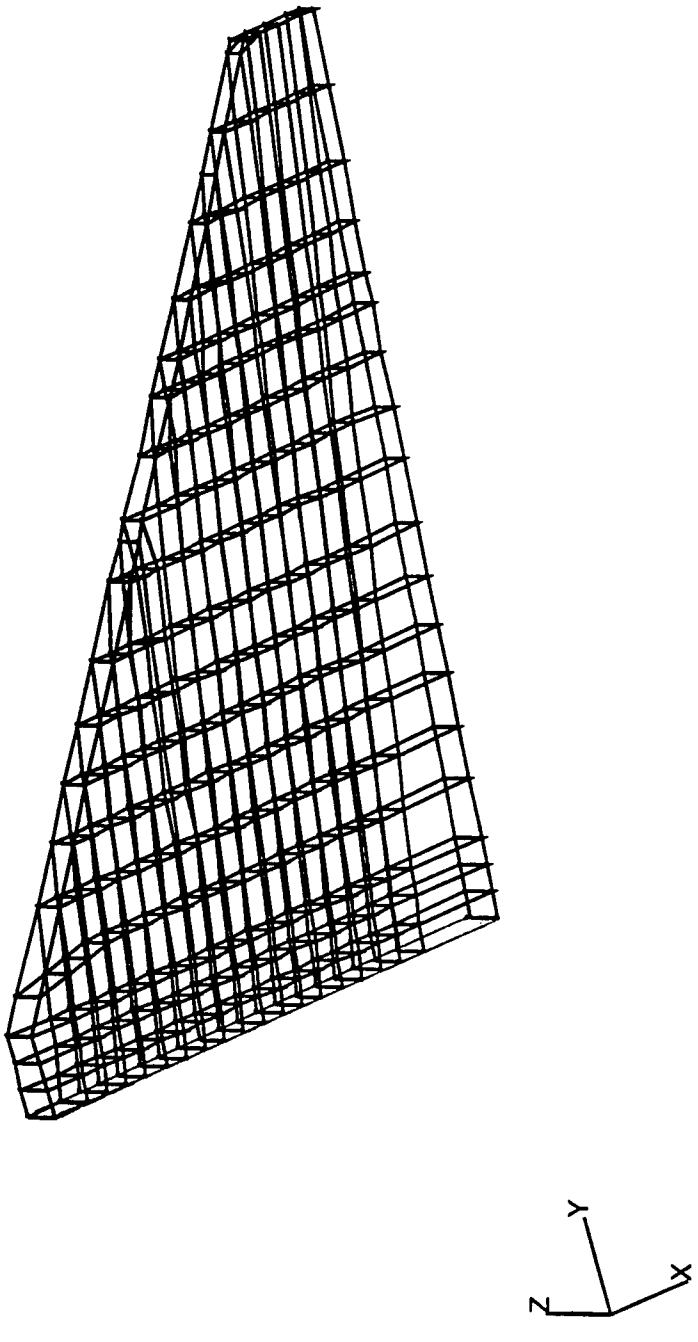
MODEL	FLIGHT CONDITION	REVERSAL-q (#/in^2)	FLUTTER SPEED (in/sec)	STRUCTURAL WEIGHT (#)
Nominal	M = 0.85	31	32,400	488
	M = 1.2	21	37,700	488
Optimized (Strength)	M = 0.85	11	17,500	151
	M = 1.2	9	23,200	151
Optimized	M = 0.85	36*	36,700	392
	M = 0.85	52*	47,700	950
Optimized	M = 0.85	31*	32,900	325
	M = 1.2	31*	53,000	799
	M = 0.85	45*	43,800	590
	M = 1.2	45*	47,900	306
Optimized	M = 0.85	35	34,900*	377
	M = 1.2	16	34,900*	377
	M = 0.85	32	32,500*	334
	M = 1.2	14	32,500*	321
Optimized	M = 0.85, M = 1.2	31*†	35,500*	369
	M = 0.85	45*	33,000*†	668
Optimized	M = 0.85	31*†	33,300*	331
	M = 1.2	21*	33,300*†	304
	M = 0.85	45*	31,000*†	614
	M = 1.2	45*	31,000*†	(see text)

* indicated quantity was a constraint
† indicates constraint was not active in the optimized design

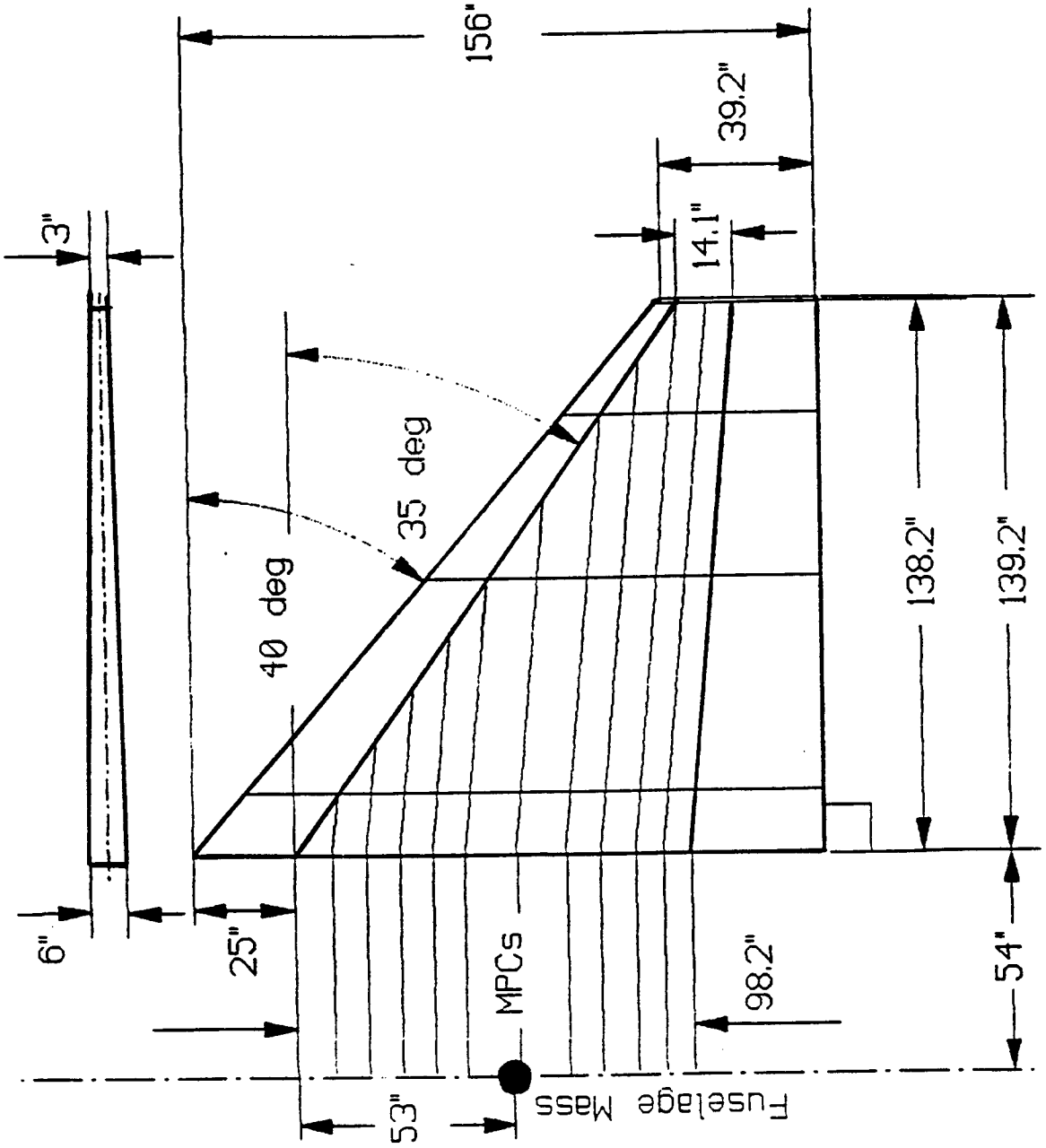




Coarse Model
Figure 1a



Fine Model
Figure 1b



Aerodynamic and Structural Geometry
Figure 2
(From Ref. 4)

A METHODOLOGY FOR AFFORDABILITY IN THE DESIGN PROCESS

Georges M. Fadel
Assistant Professor
Mechanical Engineering Department

Clemson University
317B Riggs 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

August, 1994

A METHODOLOGY FOR AFFORDABILITY IN THE DESIGN PROCESS

Georges M. Fadel
Assistant Professor
Mechanical Engineering Department
Clemson University

Abstract

Performance has always been the driver of weapon systems design. Today, with dwindling financial resources, the consideration of life-cycle cost has become another major driver that profoundly affects the design. In order to access affordability, tools and techniques need to be developed. It is well known that over 80% of the cost of a design is committed at the conceptual level, therefore, in order to have the greatest leverage, affordability has to be considered before the form of the design is firmed up. Once the concept is established, tools such as Cognition's cost modeler and Boothroyd Dewhurst's DFA can predict to a certain level the cost of manufacture and assembly. Other tools such as Taguchi's and the Six Sigma approach allow producibility to be quantified. No such tools exist at the conceptual stage. This report summarizes the published work that deals with costing and producibility tools that apply to different stages of the design process. It then proposes a methodology to deal with affordability early at the conceptual stage. The methodology is based on the use of the QFD house of quality and the identification of the non-linear engineering characteristics that affect the performance drivers. Such a tool aims at identifying where a design lies on the performance - affordability or performance - cost curves to help in prioritizing research needs. The report concludes with recommendations to the Air Force aimed at dealing with the affordability issue in educational institutions, industry and the government.

A METHODOLOGY FOR AFFORDABILITY IN THE DESIGN PROCESS

Georges M. Fadel

Introduction

The issue of *affordability* in weapons system design will lead to a significant redirection of the design process in defense oriented industries. It will also affect any manufacturing concern since the drive to cut costs, albeit at the expense of neither performance nor functionality, will separate successful and globally competitive industries from the others. Designs no longer depend on performance and performance only, rather, on *life-cycle costs, performance,* and *environmental impact.* Measures such as producibility, maintainability, reliability, supportability contribute to defining the *affordability* issue since these measures ultimately affect the life-cycle cost of a system. Methods to evaluate these "ilities" need to be defined. Metrics have to be established and used to monitor the design of existing weapons systems, their modifications, and of proposed new systems. The consideration of affordability at the conceptual design stage provides a high leverage early in the process since design modifications become more and more costly as the design evolves and the design freedom is reduced. Unfortunately, this consideration is difficult, since at the conceptual stage, design details are not available for trade-off studies, and designers involved in detail design do not have tools nor techniques to perform trade-off exercises. Databases are practically non existent, and individual manufacturers typically rely on rules of thumb derived by experience. Affordability needs also to be considered later in the design process, when the embodiment of the design is firmed up. The issues of ease of manufacture and ease of assembly are two of the main cost drivers at this stage. This work addresses the definition of a methodology to address *affordability* at the conceptual level of the design process.

Issues and Previous Work

The Air Force has defined affordability in it's 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. These equations are listed in Appendix A at the end of the report. 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 affordability, we should obtain curves such as displayed in Figure 1. One such performance-affordability curve should show when a new technology is needed. The figure describes how a technology becomes less and less affordable when more performance is needed. (Performance - Cost curves would show the curve exponentially increasing as cost increases when performance increases.) At a certain point, there is a need to move to another technology that might be more or less affordable, but that allows a significant 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, 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.

The affordability region is displayed as the shaded area in Figure 1. The region extends above the minimum performance required, and below some prescribed cost or above the affordability max1. Should more performance be required, does the affordability region change? The performance expectations are raised ("higher" on performance axis), this usually translates in

higher cost, less affordability, (max2) so the overall affordability could be some area above the curve displayed in dashed lines. The whole area itself does not represent the affordability region, but, for some performance level, the curve defines the affordability level above which the technology is affordable, or the cost level below which the technology is affordable.

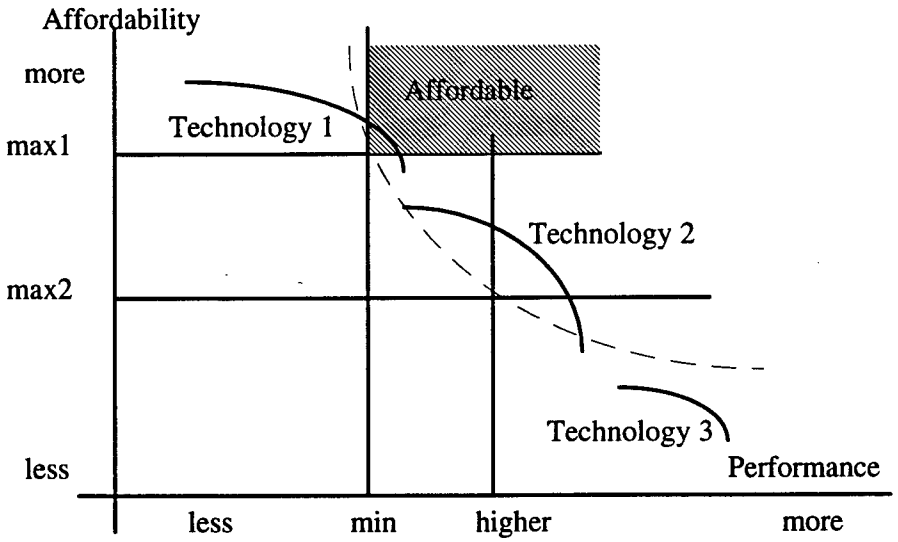


Figure 1. Affordability - Performance Hypothetical Graph

The issue of *costing* later in the *design* stage has been implemented in many methods. The Boothroyd-Dewhurst method (DFA) [Booth, 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 [Booth, 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.

Other costing related research at the process level is abundant. Many textbooks exist that deal with the cost of manufacturing parts [Trucks, 74], [Gunth, 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 [Gutow, 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 if data has been gathered and made available through some 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 [Booth, 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.

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 firmed up, 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, 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 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, 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 [Taguc, 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. In the next section, a proposed methodology to deal with affordability at the design stage is detailed and applied to sample mechanical problems.

Methodology

As explained in the previous section, there is a need to bring the decision making process affecting cost and producibility all the way back to the conceptual design stage. This means that performance and cost are part of the design parameters, specifically customer and designer's inputs into the design process which will result in some engineering characteristics related to performance, resource allocation and quality. Figure 2. illustrates this process.

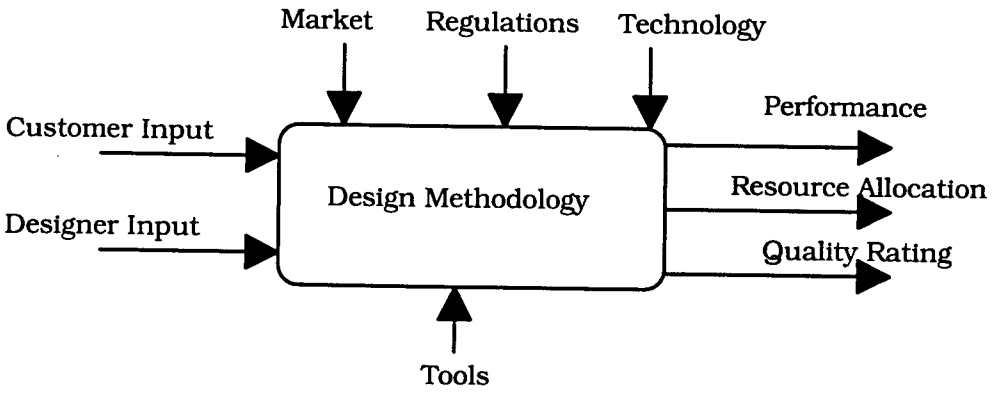


Figure 2. Generic Design Method [Staples, 94]

For a weapons system for instance, the customer (DoD) specifies mission goals in terms of performance characteristics. These include mission, range, velocity goals, thrust/weight characteristics, load, stealth capabilities, stability, etc. All these performance characteristics affect the cost equation to some extent. Typically, the higher the performance expectations are raised, the higher the overall cost is. The question is how to allocate cost, volume and weight to the different sub-components of a design to achieve the optimal affordable system. The designer at this point rarely, if ever, considers the life cycle cost as a whole. The design decisions are based on flyaway costs in the case of a weapon system, and any performance increase request during the production phase can significantly affect overall cost.

What are we faced with? The customer has some performance objectives in mind when deciding to develop a new weapon system. These objectives can be listed with some minimum acceptable value and some target value. These are similar to the aspiration levels of Staples.[Staples, 94.] For instance, for a supersonic attack fighter, a maximum velocity of Mach 1.5 might be the minimum acceptable, and a Mach value of 2 is desirable or targeted. Staples uses Fuzzy sets and satisfaction levels to develop an algorithm to maximize the sum of the satisfaction levels associated with each performance measure. Furthermore, an importance level derived from a QFD house alters the slope of the curve [Staples, 94.] Figure 3. shows a Satisfaction versus performance function profile where the performance and satisfaction levels are normalized. What this graph provides is a numeric satisfaction level corresponding to the performance and to the importance of this performance objective in the design. For a performance of 50% for instance, the customer will have a satisfaction level of around 0.85 if that performance factor is not very important, a satisfaction level of 0.5 if the factor is important, and a level of 0.15 if the factor is very important.

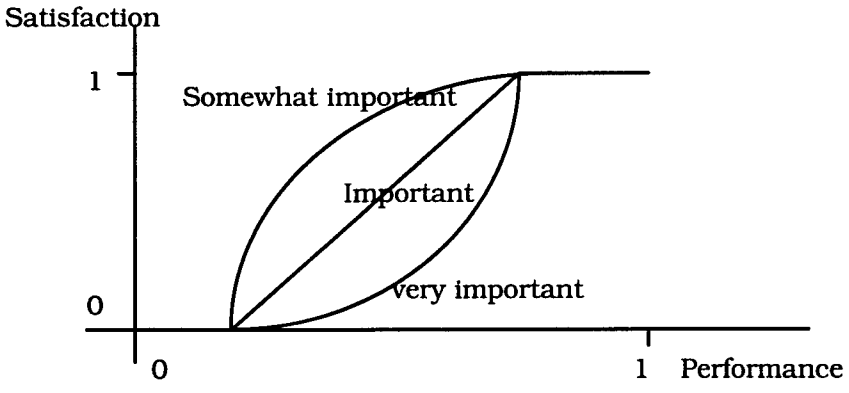


Figure 3. Satisfaction versus performance [Staples, 94]

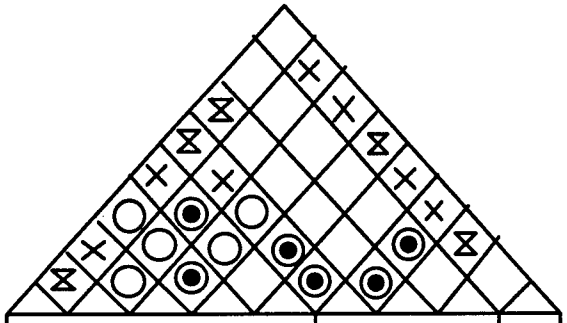
Staples uses this information to optimize the sum of the individual satisfactions and to generate a customer performance level.

Once the objectives are listed, they need to be related to some engineering characteristics that can help define the design at this early stage. What are these characteristics? Typically, designers break down their design into sub-components and associate some characteristic such as horsepower for an engine, structural integrity for a body structure, to derive the first House of Quality of the design. For an airplane for instance, body, wings, avionics, cockpit, engines are some of the form descriptors that could help define the engineering characteristics. A functional decomposition along the lines of *motion, control, power* and *enclosure* (for mechanical designs) provides a convenient and systematic method to generate all the functional components of the design.(For electronic designs, the motion function would not exist) For instance, for the airplane, the functions required are lift (upward) and forward motion, these are created by a combination of engine(s) and wings. Control is provided by the tail, the flaps, their control mechanism and the pilot interface. Power is generated through combustion of fuel which has to be stored in the plane. Finally, enclosure provides means to connect all these "functional features" together, to incorporate the cockpit, avionics, etc.. Breaking down the functions to a lower level before focusing on the functional features allows a further decomposition of the plane. For instance, Motion has to be generated on the ground and in the air. Ground motion can then be translated into landing gear, its control means steering and brakes, etc..

After decomposing the design, engineering characteristics have to be associated to these functional features. For instance, the wings have structural integrity, area, volume, weight and drag coefficient as critical measures, the engine has horsepower, maximum operating temperature, weight, the fuselage has weight, structural integrity, and so on.

No decision has been made so far in the design. The construction of the first QFD house can proceed before decisions are made. Figure 4. shows an example of a QFD house. A subset of the functional features and the associated engineering characteristics are displayed with some representative relations. For instance, the top of the house shows that the wing area typically increases with an increase in wet volume, in weight, the area strongly affects the drag coefficient of the wing, and negatively affects the size of the engine. This means that for a larger wing area, a smaller engine could be used. The matrix itself shows relationships between performance goals and engineering characteristics. The range for instance, is strongly related to the wing area, wing wet volume, drag coefficient, and weakly related to the weights of the engine and wing. Assigning numeric values to strong, medium and weak (9, 3 and 1), we

can add up the importance levels and come up with an absolute importance and relative importance. In this particular example, the most important engineering characteristic seems to be the wing area, followed by the engine horsepower and the drag coefficient of the wings.



	WING					ENGINE			FU
	Str Int	Area	Wet Vol	Wgt	Cd	Hp	Wgt	Tm	Str Int
Velocity	○	◉			◉	◉		○	△
Range		◉	◉	△	◉		△		
Thrust/Weight		○		◉		◉	◉	◉	
Load	◉	◉		△	○	◉	△		◉
Stealth		○						△	△
Stability	△	◉				○			△
Tot Importance	13	42	9	11	21	30	11	13	12
Rel Importance	8	26	6	7	13	19	7	8	7

Legend

Roof
- ◉ Strong +
- ○ Positive
- × Negative
- ⧗ Strong -

Matrix
- ◉ Strong 9
- ○ Medium 3
- △ Weak 1

Figure 4. Example QFD House of Quality for an Aircraft

What are the next steps in the design process? How can we use this importance rating? When do we make decisions? Supposing we seek a much higher thrust to weight ratio, the house of quality shows us that reducing the weight of the engine, of the wings and increasing the power of the engine(s) are the modification to the engineering characteristics we should aim for. Weight is related to material selection and the design problem becomes the selection of a material that has acceptable strength, fatigue, thermal and stealth capabilities among others,

but is lighter. Also, how difficult is this material to machine, to produce to the specifications required?

Thurston and Essington [Thurs, 93] describe a spreadsheet based tool to help in the "Optimal Manufacturing Design Decision making" process. The methodology criticizes the fact that most costing procedures use weighted methods that presume that the different parameters that affect the cost (performance parameter) and other objectives are related linearly. Instead, the authors describe a non-linear method based on the "Utility Theory" to come up with more appropriate weighting means to deal with pareto optimal solutions of multi-criteria design problems. The utility theory is based on a heuristic method to access trade-off decisions using lottery type questions (Do you prefer a design that weights 100 lbs and costs $10,000 with a probability of 0.6 or one that weights 400 lbs and costs $90,000 with a probability of 0.4? These probabilities are changed until the user is indifferent to both choices. These probabilities then become the weighting factors for the Utility function). This method becomes untractable when several variables have to be considered.

What is proposed is to consider each engineering characteristic that is significant for the performance characteristic considered, and study the relationship between cost and that characteristic. In the example proposed, we would consider wing area versus cost, and derive a relationship which is probably relatively linear since the cost here is mainly a function of the amount of material used. The same type of relationship would be apparent for the wing weight, engine weight, but a different type of relationship would surface for cost versus maximum temperature or cost versus thrust. These two engineering characteristics are related in a highly non linear way to cost. It is known that the material properties would not allow a sustained maximum operating temperature above some limit, and the power produced is related to that temperature as displayed in the hat of the house of quality. Thus, taking into account the non-linearity and importance level, we can deduce that these engineering characteristics are cost drivers when operating close to the upper limits.

What about the characteristic that is very costly, but not at the upper limit of the performance? How can we evaluate an existing process, material or design and decide whether its cost is justified? The example of the airplane is not detailed enough from the point of view of engineering characteristics, lets consider a simpler example and try deduce a method.

Example Problem: Additional Fuel Tank for Airplane

Let's consider the conceptual design of a fuel tank whose mission is to extend the flight range of an aircraft. Such a tank has one basic function, that of storing fuel. It has additional

functions which are not explicit: fill, drain tank, relieve pressure. We can give the tank some concrete measurements: let's suppose its capacity has to be >=25 ft^3. Naturally, the individual dimensions cannot be negative nor zero, and we want to manufacture a tank that does not exceed a certain cost. The problem could be restated as: manufacture the most cost effective (affordable??) external fuel tank for an airplane. Now looking at the constraints, such a tank has to be easily attached under the body or wing of a plane. It must have some means of filling it and draining it, and it must definitely be aerodynamic. Also, safety has to be considered (pressure relief, for instance). We can therefore describe the problem using a functional hierarchy that we extend until the functional feature level, i.e., a level at which some component can be listed that accomplishes the desired function [Figure 5.] This functional tree captures one of the three main topics of interest of the customer, namely that the device should be functional, the others being pleasing and safe [Figure 6.] The functional attribute is also one of the drivers of the designer who aims at satisfying the customer, but also at pleasing or satisfying himself or herself.

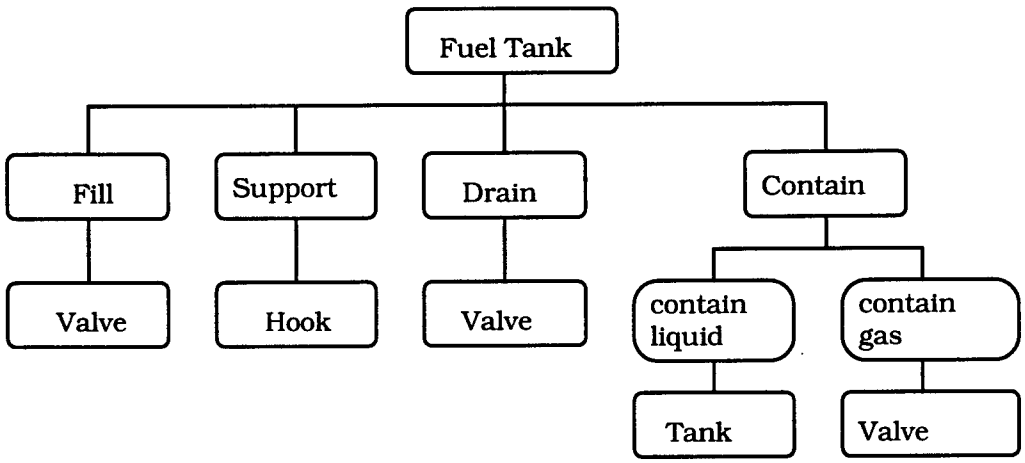


Figure 5. Functional Description of Fuel Tank

The designer needs to consider all the criteria of the interested or participating parties in the process, and it is at the conceptual design level that the decisions that have the greatest leverage on producibility and cost are made. After detailing the functional description, the designer translates the requirements into forms. This translation assumes some manufacturing technology, and a set of alternative means or techniques to accomplish the design objective. For instance, for our fuel tank model, a casting process might not differentiate between the tank walls and components of the valves or support structure,

whereas alternative methods using metal rolling and welding dictate an assembly of parts or forms. Independently of the manufacturing method chosen, the forms must be described using features, dimensions, manufacturing process, tooling requirements, dunnage, gauges, cost, fixtures and jigs, center of gravity, weight, material, etc. At this point however, the designer has to make decisions. What material must be used for the tank? What manufacturing process is more appropriate? How will the valves and attachment mechanism be incorporated in the design of the tank? These decisions directly affect producibility and affordability. The experienced designer will go to past designs and use them to guide him or her in the decision making process. Should the new design constraints be significantly different from those of previous problems, novel approaches have to be considered. This is where cost becomes more difficult to derive early on in the design.

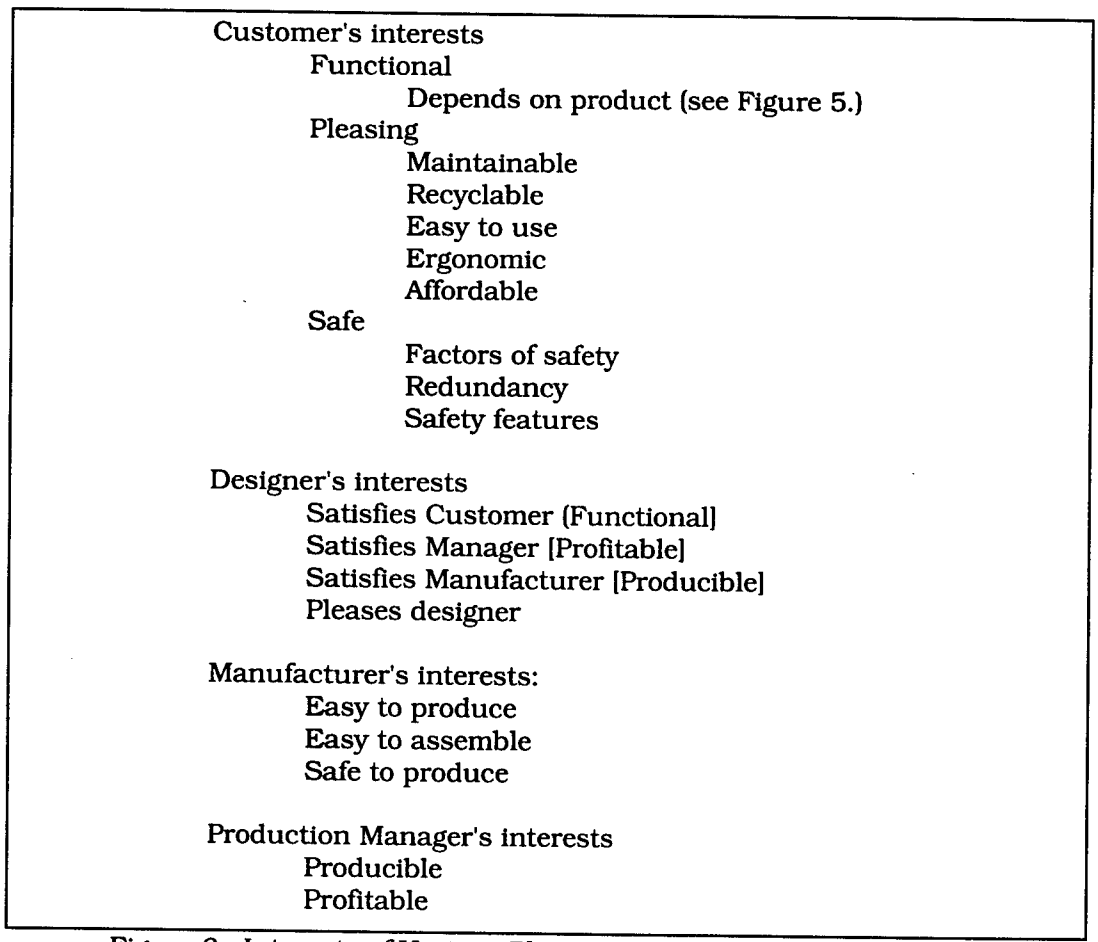
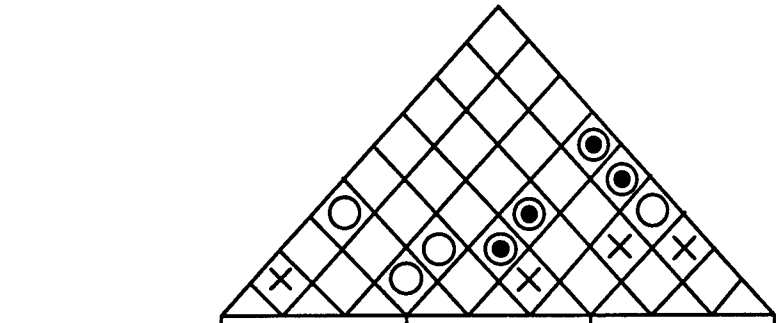


Figure 6. Interests of Various Players in the Design Process

Identifying the performance characteristics and engineering characteristics as previously described, the QFD House of quality displayed in Figure 7 can be constructed.



	Valves			Tank			Connector		
	Mat	Type	Size	Mat	shap	thck	Mat	Type	Size
Contain Fuel	△	○	△	○	◉				
Light	△		△	◉		◉	△		△
Safe		◉	○	○	△	○		◉	○
Easy fill/drain		◉	◉		△				
Easy connect.	△	◉	○					◉	○
Maintainable	○	◉	△	◉		△	○	◉	○
Recyclable	◉		△	◉			◉		△
	15	39	19	33	11	13	13	27	11

Legend

Roof
- ◉ Strong +
- ○ Positive
- × Negative
- ⧗ Strong -

Matrix
- ◉ Strong
- ○ Medium
- △ Weak

Figure 7. QFD House of Quality for Fuel Tank

What contributes to the cost at this point? First material, second, process, and third, labor. At the design stage, the material issue is paramount since it affects the process used, the properties of the part, and the static, dynamic, thermal and other responses of the system. Typically, analyses are conducted in an iterative fashion to reach some optimal design that satisfies the constraints and minimizes some objective function. This objective function should be related in some way to cost. (Typically, weight)

How can we proceed with the design of the fuel tank from the functional description including cost as a design parameter? How can we factor cost in before deciding on the material, process and final conceptual design? At the functional level, the flexibility still exists, and various translations of these functions into forms are possible, each with different affordability and producibility values. The Germans Pahl and Beitz [Pahl, 76] describe the ***Use-Value Analysis*** to create a functional tree with some weights that indicate the relative importance of design characteristics or evaluation criteria. If we select as performance driver the weight of the fuel tank, the material and thickness of the tank are the most important engineering characteristics and the material and size of the connector and valve(s) are the other much less important characteristics. Note that in this particular example, material was used as an engineering characteristic whereas in the previous example, weight was considered. This is done to illustrate the variability that will occur when two designers consider a problem. However, whether material is included as a characteristic, or weight and structural integrity, by evaluating the engineering characteristic that is related in a non-linear way to the cost, the thickness of the tank wall surfaces as the cost driver.

Considering previous designs, some baseline parameter can be identified. For instance, fuel tanks made out of some Aluminum alloy have a minimum thickness that has to be maintained for fabrication and structural integrity. What other material can be substituted for that alloy that would reduce the performance driver (weight) without reducing strength, structural integrity, maintainability, ease of fabrication, and still be affordable? If a composite is selected, the material characteristics can show that density of composites is in general lower, that strength can be higher, but that cost is also higher. What about fabrication, assemblability questions with the valves and connector? What about maintainability, recyclability? Which is more affordable in the long run? The better question is what is the additional cost incurred in reducing weight versus the additional cost in increasing thrust since in an airplane, the thrust to weight ratio is the performance driver.

Assuming the additional cost in increasing thrust is much more important than the one to decrease weight, how then can we define the affordability of Aluminum versus composites designs? The cost of the fuel tank body can be broken down as follows:

Cost = material cost + engineering cost + tooling cost + manufacturing cost + test and Quality control costs + maintenance cost + recycling cost

Manufacturing cost includes labor, fabrication and assembly.

We should then assume a design with a base thickness made of an Aluminum alloy, and compute the cost differential in achieving a different thickness. This cost differential should

then be compared to the cost differential for a composite design at the two thicknesses. These two differentials provide a measure of sensitivity of cost to performance (thickness) that can be used to assess affordability. The two slopes and the cost differential between the metallic and composite alternatives have to be used in tandem. The ability to generate numbers that can be compared for the two material alternatives is still questionable and merits further investigation.

Conclusion

A method to deal with affordability at the conceptual level is presented. This method is intended to help the designer make decisions very early in the design process. The method 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. 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.

These issues highlight the lack of proper training of student engineers that are rarely if ever exposed to economic considerations in their curriculum. Additionally, the needs for costing databases, standards, and extensive process documentation surface as extremely urgent endeavors for government driven projects. The maturation and spread of design for manufacturing, the development of tools such as Virtual Manufacturing can only help in the long run, since designers educated with such a mind set are aware of the producibility issues and ultimately will produce more affordable designs.

Proposed research projects:

Develop costing rules for processes to be included in Cognition's database

Study the development of design features that are not process specific.

Develop an expert system tool to evaluate cost differentials for specific materials and processes

Bibliography

[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.

[Chow, 78] Chow, W.C. Cost Reduction in Product Design, Van Nostrand Reinhold, 1978

[Gutow, 94] Gutowski, T., Hoult, D., Dillon, G. et al, "Development of a Cost Model for Advanced Composite Fabrication" Submitted to Composite Manufacturing, May 1994.

[Gunth, 71] Gunther, W., Die Grundlagen der Wertanalyse, Z. VDI 113, 238-241, 1971.

[Harry, 92] Harry, M. and Lawson, J.R., "Six Sigma Producibility Analysis and Process Characterization" Motorola University Press, 1992.

[Knight, 85] Knight, W.A. and Poli, C., "A Systematic Approach to Forging Design", Machine Design, 57, January 24, 1985

[Micha, 89] Michaels, J.V., Wood, W.P. Design to Cost, Wiley Interscience, 1989.

[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

[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.

[Staples, 94] Staples, J. W., "Optimizing the Allocation of Resources during Preliminary Automobile Design," M.S. Thesis, Georgia Tech, 1994.

[Taguc, 93] Taguchi, G., Introduction to Quality Engineering - Designing Quality into Products and Processes, New York, UNIPUB/Quality Resources, 1986.

[Thurs, 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.

[VDI, 72] VDI-Taschenbuch 135, Wertanalyse - Idee, Methode, System 4. Dusseldorf, VDI Verlag, 1972.

Appendix A

Modified DAPCA IV Cost model according to Raymer [Raymer, 92]

The cost of aircrafts in constant 1986 Dollars is listed as a function of the following parameters:

W_e	= empty weight (lb)
V	= maximum velocity (knots)
Q	= production quantity
FTA	= number of flight test aircraft (typically 2- 6)
N_{eng}	= total production quantity times number of engines per aircraft
T_{max}	= engine maximum thrust (lb)
M_{max}	= engine maximum Mach number
T_{ti}	Turbine inlet temperature (Rankine)
C_{av}	avionics cost

The equations are the following:

Eng hours	$= 4.86\ W_e^{0.777}\ V^{0.894}\ Q^{0.163}$	$= H_E$
Tooling hours	$= 5.99\ W_e^{0.777}\ V^{0.696}\ Q^{0.263}$	$= H_T$
Mfg hours	$= 7.37\ W_e^{0.82}\ V^{0.484}\ Q^{0.641}$	$= H_M$
QC hours	= 0.076 (mfg hours) if cargo airplane	$= H_Q$
	= 0.133 (mfg hours) otherwise	$= H_Q$
Devlp. support cost	$= 45.42\ W_e^{0.630}\ V^{1.3}$	$= C_D$
Flt test cost	$= 1243.03\ W_e^{0.325}\ V^{0.822}\ FTA^{1.21}$	$= C_F$
Mfg materials cost	$= 11.0\ W_e^{0.921}\ V^{0.621}\ Q^{0.799}$	$= C_M$
Eng prodt cost	$= 1548[0.043\ T_{max} + 243.25\ M_{max} + 0.969\ T_{ti} - 2228] = C_{eng}$	

$$\text{RDT\&E + flyaway} = H_E R_E + H_T R_T + H_M R_M + H_Q R_Q + C_D + C_F + C_M + C_{eng}\ N_{eng} + C_{av}$$

They are applicable to the design and fabrication of an aluminum aircraft. For other materials, the following fudge factors are listed:

aluminum	1.0
graphite-epoxy	1.5 - 2.0
fiberglass	1.1 - 1.2
steel	1.5 - 2.0
titanium	1.7 - 2.2

The hours estimated above are multiplied by hourly rates to calculate labor costs. Some average 1986 wrap rates (including salaries, overhead, benefits and adinistrative costs) are:

engineering	$59.10	$= R_E$
tooling	$60.70	$= R_T$
quality control	$55.40	$= R_Q$
manufacturing	$50.10	$= R_M$

Avionics costs are not estimated using similar formulae. A rough range of 5 - 25% of flyaway cost is given, or approximately $2000 per pound in 1986 Dollars.

COMPUTER MODELING OF ELECTROLYTES FOR BATTERY APPLICATIONS

Joel R. Fried
Professor
Department of Chemical Engineering

University of Cincinnati
Cincinnati, OH 45221-0171

Final Report for:
Summer Faculty Research Program
Wright Laboratory–WL/PPOS-2
Battery Electrochemistry Section

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

September 1994

COMPUTER MODELING OF ELECTROLYTES FOR BATTERY APPLICATIONS

Joel R. Fried
Professor
Department of Chemical Engineering
University of Cincinnati

Abstract

An objective of this effort was to evaluate high-end software packages for computational chemistry for their ability to model salt–electrolyte systems in a variety of circumstances. The goal was to provide an understanding of the molecular basis for efficient electrolyte systems and to develop a methodology for follow-up studies. Of initial interest was the simulation of salt dissolution and electrolyte association.

The program selected for use in this study was Cerius^2 (Molecular Simulations). The UNIVERSAL force field (UFF) combined with the Charge Equilibration (QEq) method was evaluated for its reliability in predicting atomic charges and geometric parameters for a range of structures. These include several ether structures such as tetrahydrofuran, model compounds for the polymeric electrolyte poly(ethylene oxide) (PEO), crown ethers, PEO, and the anions of several lithium salts. Comparison of UFF/QEq results was made with experimental data and predictions from ab initio and semiempirical molecular orbital methods where available. In general, UFF/QEq gave good results, comparable to ab initio values in many cases; however, there appears to be a systematic problem with the UFF or QEq parameterization for sulfur-containing compounds, such as tetrahydrothiophene (THS) and triflate anions. Recommendations for future studies are given.

Within the limitations of the UFF/QEq method cited above, the molecular-dynamics simulation capability of Cerius^2 was used to investigate the association of Li cations in 1,2-dimethoxyethane (DME) — a model compound for PEO. In the first approach, a single Li cation and a molecule of DME were minimized (UFF/QEq) in a nonperiodic box. The resulting structure showed the association of the cation with one oxygen of DME which remained in its low energy conformation. Next, constant volume and energy (NVE) anneal dynamics was used to investigate other minimum energy structures. The result was a change in the DME conformation to accomodate the association of both oxygen atoms with the cation. In a higher level simulation, 4 Li cations were added to a (nonperiodic) solvent box containing 64 DME molecules at 300 K. This system was then minimized (UFF without QEq). The Li cations were well dispersed within the box but no specific associations were identified. Finally, quench dynamics was used. All cations were associated with at least one oxygen. Two cations were each associated with two oxygens from different DME molecules.

COMPUTER MODELING OF ELECTROLYTES FOR BATTERY APPLICATIONS

Joel R. Fried

Introduction

There is significant interest to develop superior electrolyte materials for lithium battery applications with high ionic conductivity (in the order of 10^{-3} S/cm) over a broad temperature range (e.g., -40° to 70°C) and with good electrode/electrolyte interfacial stability. The objective of this current research effort is to explore the applicability of computational chemistry and simulation methods to help identify factors contributing to high conductivity of lithium electrolytes.

There are several major commercial programs available on workstation platforms that provide a broad range of computational tools including ab initio, semiempirical molecular orbital, and molecular mechanics methods and provide the ability to simulate solvent systems. These include programs from Molecular Simulation (Cerius^2) and from Biosym (INSIGHT and DISCOVER) which provide molecular-dynamics (MD) simulations as well as user interfaces to semiempirical (e.g., MOPAC) and to ab initio (e.g., GAUSSIAN) methods. In addition, SPARTAN (WaveFunction) particularly excels at the ab initio and semiempirical level but also provides access to common molecular mechanics (MM) methods. Although all three programs have been obtained for review at the University of Cincinnati, the efforts during the summer focused on learning to use Cerius^2 and evaluating its ability to study lithium electrolytes of current interest for secondary battery applications. Work is continuing with both Cerius^2 and SPARTAN.

Discussion of Problem

An objective for the summer effort was to evaluate the ability of Cerius^2 to properly simulate salt/electrolyte systems. An important consideration is the suitability of the available force fields to model electrolyte systems that contain a variety of atom types encountered in electrolyte systems including Li, B, F, and S in addition to the common atoms C, H, and O. Only with an appropriate, validated force field, is it possible to explore the ability of MM methods to address two issues of importance to polymeric battery applications. One is the relationship of the conformation of poly(ethylene oxide) (PEO) to crystallinity and conductivity. The other is the mechanism of salt dissolution and association in a variety of electrolyte environments. These issues are discussed below.

Force Field Evaluation. An important feature of Cerius^2 is its open force field capability for MM calculations and MD simulations. Principal force fields (FF) include the UNIVERSAL[1] or UFF (UFF 1.01 the default FF), DREIDING II[2], Burchart–Dreiding, AMBER, and MM2 (85) force fields (FF). An advantage of UFF is that it is parameterized for the full periodic table; the DREIDING FF is another all-purpose FF suitable for organic, biological, and main-group inorganic molecules; the Burchart–Dreiding FF is suitable for organic molecules inside a

silica/aluminophosphate framework; AMBER (2.01), which was developed for the simulation of proteins and nucleic acids, is limited to H, C, O, S, N, and P but contains atom types for biologically important ions (Br, Na, Cl, and Ca); MM2 is the 1985 implmentation of Allinger's FF.

Other than its parameterization for all elements, an important advantage of the UFF is that it can be used with the Charge Equilibration (QEq) method of Rappé and Goddard[3] to calculate charges for *nonperiodic* structures.† This approach was developed using experimental atomic ionization potentials, electron affinities, and atomic radii from which an atomic chemical potential is constructed. This method is claimed to lead to charges in excellent agreement with experimental dipole moments and with atomic charges obtained from the electrostatic potentials of ab initio methods. An advatnage of the QEq method is its ability to *recalculate* charges periodically at equilibrium during minimization or dynamics simulation. This capability, which is important in the simulation of electrolyte systems where charge interactions can be affected by conformation and association, is not available for the other force fields.

It is clear that UFF/QEq would be the preferred approach for simulation studies of electrolyte systems providing that these methods provide a reasonable estimate of geometry and charge assignments. For this reason, the predictions of charge and geometry provided by UFF/QEq has been compared to those predicted by ab initio methods and MOPAC methods (as described in the section on Methodology) for several model compounds which have relevance to electrolyte studies.

Conformational Studies. It has been reported that higher conductivity can be achieved by decreasing PEO crystallinity. There is evidence that this higher conductivity may be associated with increased molecular mobility. For example, Johansson and Tegenfeldt[4] have reported NMR measurements that show that the rate of molecular motion is about 100 times higher in the amorphous phase than in the crystalline phase of PEO.

One approach to reduce PEO crystallinity is through copolymerization. An example is the copolymerization of PEO and polyoxymethylene (POM). At WPAFB, a recent effort has focused on the copolymerization of dimethylene chloride and diethylene glycol to give an amorphous polymeric electrolyte having an approximate 8:1 ratio of $-CH_2-CH_2-O-$ to $-CH_2-O-$ repeat units. One objective of the current computational effort is to evaluate to what extent copolymerization can affect the conformation of PEO and its ability to form a helix structure that can form a crystalline structure.

Simulation of Electrolyte Systems. Another objective of this work is to evaluate the ability of MD simulations to investigate how salts dissociate and how cations and anions may associate with themselves and with the nonaqeous medium forming the electrolyte. Current understanding of electrolyte associations has been deduced from spectroscopy data and, in some cases, from ab initio calculations as discussed briefly below.

Frech and Huang[5] have used infrared and Raman spectroscopy to study salt association of lithium and tetrabutylammonium trifluoromethanesulfonate (triflate) solutions in a variety of solvents. Results indicate that

† QEq for periodic structures should become available in version 1.6 of Cerius[2] expected for release at the end of 1994.

contact ion pairing is the only interaction in solvents with relatively weak electron acceptor properties such as tetrahydrofuran, triglyme, acetone, and acetonitrile while interactions of the anion with the solvent increase in strength with increasing electrophilic nature of the solvent. All modes are shifted to higher frequencies compared to free ion.

In a related study, Huang et al.[6] have employed ab initio (GAMESS) calculations to study the nature of lithium–triflate associations (triflate anion, lithium ion pair, and aggregate). Analysis of IR spectra shows that asymmetric vibrations are split only for SO_3 which indicates that interaction of the lithium cation should be primarily with the SO_3 end rather than with the CF_3 end of the triflate anion. The bidentate structure (Fig. 1) was found to have the lowest total energy for isolated ion pairs. Similar behavior may be deduced from studies of lithium tetrafluoroborate and lithium perchlorate.

Fig. 1

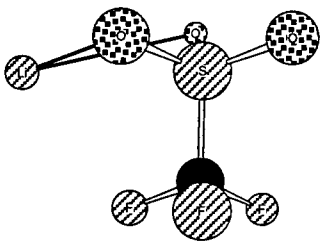


Fig. 2

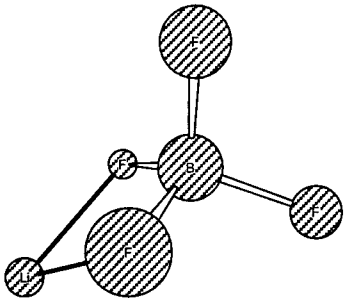
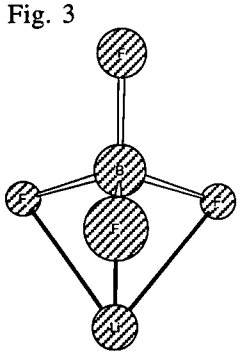


Fig. 3



Huang et al.[7] have employed vibrational spectroscopy to study the association of $LiCF_3SO_3$ in diglyme, $CH_3O(CH_2CH_2O)_2CH_3$, and concluded that diglyme forms a 1:1 complex with Li^+ at room temperature. At low temperatures (-80°C), diglyme forms only a 2:1 complex with Li^+ and all triflate ions existing as "free" ions when the O:Li ratio is >6:1.

Lightfoot et al.[8] have used powder X-ray diffraction to investigate the crystal structure of $PEO_3{\bullet}LiCF_3SO_3$. Analysis indicated that the Li^+ cation is coordinated by five oxygen atoms — three ether oxygens and one from each of two adjacent $CF_3SO_3^-$ groups. Each $CF_3SO_3^-$ in turn bridges two Li^+ ions to form chains running parallel to and intertwined with the PEO chain. There are no interchain links between PEO chains, and the electrolyte can be regarded as an infinite columnar coordination complex.

In an earlier study, Lightfoot et al.[9] used X-ray powder data (Rietveld method) to deduce the crystal structure of $PEO/NaClO_4$ for which was composed of a helical polymer chain wrapped around the Na^+ cation. Perchlorate ions associated with the Na^+ cation in a zigzag manner, resulting in five-fold coordination around Na^+ comprising three oxygens from the polymer chain and one each from the electrolyte complex.

Francisco and Williams[10] have used ab initio to study the energies of different structures of various salts of the tetrafluoroborate anion (BF_4^-) including $LiBF_4$. Their results indicated that the bidentate (C_{2v}) structure (Fig. 2)

was preferred for Li^+. In general, tridentate coordination is preferred for larger cations, such as Na^+. The ionic binding energy of $BF_4^-Li^+$ relative to BF_4^- and Li^+ was estimated to be -601 kJ/mol.

In a subsequent study by Spoliti et al.[11], ab initio calculations suggested that $LiBF_4$ prefers tridentate (Fig. 3) binding; however, the stability of the bidentate structure of $LiBF_4$ was found to be comparable to that of the tridentate isomer (the energy separation between C_{2v} and C_{3v} is only 1.6 kJ/mol).

Farber et al.[12] have used electrical conductivity measurements, Raman spectroscopy, ultrasonic absorption, and dielectric measurements to investigate the molecular relaxation of $LiAsF_6$ in 1,2-dimethoxyethane (DME). Results indicated that the AsF_6^- ion was spectroscopically free, namely, either unpaired or solvent separated from the cation.

Methodology

Initially, Cerius[2] (version 1.0) operating off a DEC 3000 workstation (mdl. 600, 128 MB RAM, OSF/1, version 2.0) on loan from Molecular Simulations, Inc., was used for molecular computations and simulations. In early September, version 1.5 was obtained and installed on a IBM PowerStation 320 (RS 6000) equipped with ca. 1.8 GB disk and 48 MB RAM at the University of Cincinnati. Version 1.5 provides internal MOPAC capability that was not available in the earlier release.

This provision for MOPAC (MOPAC 5 and 6) is provided through the MOPAC User's Interface (MOPACUI) by which charges can be assigned (although not recalculated during minimization or dynamics simulation). Included within MOPAC is MINDO (Modified Intermediate Neglect of Differential Overlap) and MINDO/3.[13] Also included in MOPAC are two MNDO (Modified Neglect of Diatomic Overlap) programs[14] — AM1 (Austin Model 1)[15] and PM3 (Parametric Method Number 3).[15-17] Parameters that can be obtained include the heat of formation, electronic energy, core–core repulsion, dipole moment, and ionization potential. The Cerius[2] implementation of MOPAC is that of QCPE (MOPAC 6.0 QCPE 455). Although MOPACIU does not provide energy minimization, some or all of the geometric variables (e.g., lengths, angles, and torsion angles) can be set for optimization during a run.

In the present studies, the MOPAC 6 implementation was used. MOPAC 6 has the advantage over MOPAC 5 in its treatment of electrostatic potential (ESP) and an improved PM3 Hamiltonian. For MOPAC calculations, structures were first minimized using UFF with QEq recalculation (UFF/QEq). For comparisons with UFF/QEq and ab initio results, the MOPAC programs AM1 and PM3 were used. Of these two programs, PM3 is considered to be the better.[18] Atoms parameterized at the PM3 level of interest to electrolyte studies include H, C, N, O, P, S, and Cl. Boron is included in AM1 but sulfur is not. Neither are parameterized for Li which is a distinct problem for direct application of MOPAC such as the minimization of Li^+ electrolyte structures but, nonetheless, MOPAC (as well as ab initio) results can be a useful base with which to compare the suitability of MM force fields for model compounds with parameterized atoms.

Cerius^2 provides powerful capability for molecular dynamics (MD) simulation of both periodic and nonperiodic structures. Principal algorithms for coordinate and lattice (periodic structure) minimization are the conjugate gradient and steepest descent methods. For structure minimization, the default UFF (1.01) was used with conjugate gradient 200 minimization and 15 Å cutoff combined with QEq charge recalculation. Methods available for dynamics include constant volume and constant energy (default), constant volume/constant temperature, and constant pressure/constant temperature. Other options include Quench Dynamics (dynamics with structure minimization) and Anneal Dynamics. A solvent box can be created using the Amorphous Builder module of Cerius^2.

Results and Discussion

A. Force Field Validation and Structure Determination

To evaluate whether the UFF/QEq force field, or other available force fields chosen for dynamic simulation of selected electrolyte systems (see previous section on Methodology), gives a reasonable representation of charge distributions and molecular geometry, results of UFF/QEq and MOPAC (AM1 and PM3) calculations were made for

- Hydrofurans and other simple organic compounds such as tetrahydrofuran, tetrahydrothiophene, dimethylformamide, and propylene carbonate
- Model compounds for poly(ethylene oxide) (PEO)
- PEO and polyoxymethylene (POM)
- Crown ethers and related macrocyclic structures
- Anions of several important lithium salts (BF_4^-, AsF_6^-, and $CF_3SO_3^-$).

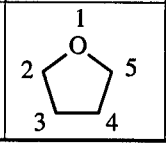
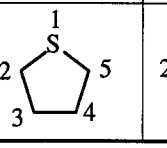
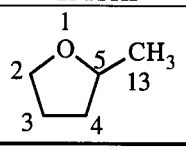
Results of these calculations are then compared to available experimental (e.g., X-ray) and ab initio results in the following sections.

Hydrofurans and Related Structures. Results of atomic charge predictions from ab initio, semiempirical molecular orbital (MOPAC), and molecular mechanics (UFF/QEq) calculations on tetrahydrofuran (THF), tetrahydrothiophene (THS), and 2-methyl tetrahydrofuran (2MTHF) are summarized in Table I. In addition, MOPAC 6 (AM1 and PM3) parameters (obtained from Cerius^2 MOPACUI) for THF are given in Table II. Finally, predictions of energies and dipole moments for THF, THS, and 2MTHF from MOPAC and ab initio studies are given in Table III. Unfortunately, MOPACUI of Cerius^2 is unable to calculate single point energies; these can be calculated from minimized structure using SPARTAN in future work.

In general, the combination of the UFF and the Charge Equilibration (QEq) method (see section on Methodology) gives a good representation of the atomic charge assignments for THF and 2MTHF as shown by comparison with ab initio results in Table I. In fact, agreement is better than afforded by MOPAC 6 in most cases, especially in the assignment of charge for oxygen which is critical for meaningful simulation studies. One difficulty of the UFF/QEq method appears to be the handling of charge on the sulfur atom in THS. UFF/QEq predicts a large negative charge (-0.403) not indicated by the ab initio or MOPAC results. Whether UFF improperly parameterizes

sulfur will be discussed again when the predicted structure of the the triflate anion ($CF_3SO_3^-$) is discussed in a subsequent section. MOPAC (AM1 and PM3) does appear to handle sulfur well in the case of THS but appears to significantly underpredict the negative charge on oxygen in THF and 2MTHF. For purposes of comparison, MINDO/3 was also used to get bond, angle, and charge information for THS as values for geometry of THS have been reported by Dewar et al.[19] using this method.

TABLE I. Calculation of Atomic Charges

Molecule	THF	THS	2MTHF
Structure			

Molecule	Atom	Ab Initio Charges UHF[a]	UMP2[b]	UFF QEq Charge	Energy[c] kcal/mol	MOPAC 6 AM1	PM3	MINDO/3
THF	O1	-0.643	-0.494	-0.540	10.87	-0.253	-0.249	
	C2	0.0346	0.00767	0.0397		-0.0291	0.0542	
	C3	-0.373	-0.265	-0.237		-0.1945	-0.138	
	C4	-0.376	-0.264	-0.239		-0.1938	-0.138	
	C5	-0.0358	0.00516	0.0398		-0.0294	0.0540	
THS	S1	0.0943	-0.149	-0.403	7.71	-0.007	-0.038	-0.591
	C2	-0.471	-0.275	-0.156		-0.248	-0.181	0.295
	C3	-0.322	-0.247	-0.349		-0.164	-0.104	0.070
	C4	-0.322	-0.247	-0.350		-0.164	0.0664	0.0154
	C5	-0.471	-0.275	-0.155		-0.247	-0.181	0.445
2MTHF	O1	-0.649	-0.482	-0.542	-16.3	-0.259	-0.253	
	C2	0.0254	0.0671	0.0253		-0.0274	0.055	
	C3	-0.357	-0.346	-0.266		-0.195	-0.138	
	C4	-0.385	-0.261	-0.221		-0.193	-0.140	
	C5	0.197	0.267	0.158		0.231	0.067	
	C13	-0.481	-0.509	-0.362		-0.217	-0.130	

[a] Gaussian 90; UHF/6–31 G(d) FOPT (courtesty L. Scanlon); [b] Gaussian 90; UMP2/6–31+G(d,p) SPE calc. (courtesty L. Scanlon);[c] Energy of minimized structure (hartree = 627.51 kcal/mol).

Of the two MNDO methods, the prediction of PM3 for the heat of formation of THF more closely approximates the experimental value as shown by data given in Table II but AM1 appears to give slightly better charge assignment and prediction of dipole moment for THF and 2MTHF (see Table III); both AM1 and PM3 give very good estimation of dipole moment. Values of charge assignments are given in Table I. Values of bond distances obtained by the Cerius[2] implementation of MINDO/3 are in reasonably good agreement with the literature values[19] as follows: 1.518 Å for C–C (1.502 lit.) and 1.815 Å for C–S (1.789 lit.); for bond angles, these are 92.9° for C–S–C (97.2° lit.) and 107.8° for S–C–C (108.5° lit.). Interestingly, our results for MINDO/3 charges indicate

that MINDO/3 also seriously overpredicts a negative charge for sulfur (-0.591) compared to ab initio results as shown by data given in Table I.

TABLE II. MOPAC 6 Parameters for THF

Parameter	AM1	PM3	exp[a]
Heat of formation (kcal)	-54.7	-47.9	-44.0
Electronic energy (EV)	-3461	-3391	NA
Core–core repulsion (EV)	2518	2499	NA
Dipole (D)	1.73	1.55	NA
Ionization potential (EV)	10.4	104	NA
C–C bond length (Å)	1.505	—	—
C–O bond length (Å)	1.405	—	—
C–H bond length (Å)	1.113	—	—
O–C–C bond angle (deg)	108.3	—	—
C–O–C bond angle (deg)	106.3	—	—
H–C–H bond angle (deg)	109.6	—	—

[a] J. J. P. Stewart.[18]

TABLE III. Energy and Dipole Moment of THF, THS, and 2MTHF Calculated from Ab Initio and MOPAC Methods

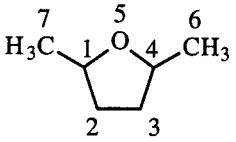
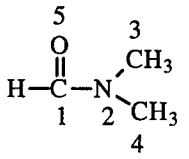
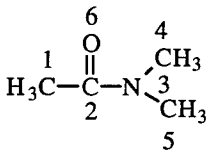
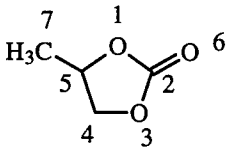
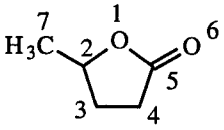
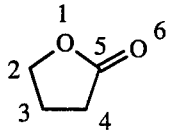
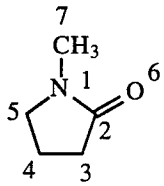
Compound	Ab Initio[a] Single Point Energy (Hartree)	Dipole Moment (D) ab initio[a]	AM1	PM3
THF	-231.7 -231.0	2.05	1.73	1.55
THS	-554.4 -553.7	2.37	2.24	2.40
2MTHF	-270.9 -270.0	1.87	1.76	1.57

[a] UMP2/6-31+G(d,p); second entry UHF. Data presented by L. G. Scanlon, Nov. 18, 1993.

Cerius[2] also provides the ability to draw Connolly surfaces[20,21]. This approach can be used the assess the availabiliy of molecular sites to an ion for association to occur. To probe the access of Li^+, a probe of 0.6 Å was used to represent the ionic radius of $Li^{+\dagger}$ (radii of anions cited by Gray[23] are 2.32 Å for BF_4^- and 2.36 Å for ClO_4^-). The aid in the visualization of the Connolly surface, a dot density 100 dots per $Å^2$ was used. From this visualization, it appears that the accessibility of Li^+ to the oxygen sites of THF, 2MTHF, and 2,5MTHF decreases in the order THF > 2MTHF > 2,5MTHF.

† Bekturov[22] reports the crystallographic radius of Li^+ as 0.68 Å.

TABLE IV. Results of UFF/QEq Calculations of Selected Organic Compounds

Electrolyte	Structure	Energy kcal/mol	Atomic Charges	
2,5THF		-46.24	O5	-0.539
			C1	+0.150
			C2	-0.240
			C3	-0.240
			C4	+0.150
			C6	-0.367
DMF		27.03	O5	-0.497
			C1	+0.416
			C3*	-0.201
			C4	-0.290
			N2	-0.369
DMAC		2.04	O6	-0.502
			C1	-0.312
			C2	0.506
			C4*	-0.212
			C5	-0.284
			N3	-0.355
PC		5.51	O1	-0.549
			O3	-0.554
			O6	-0.431
			C2	+0.960
			C4	+0.001
			C5	+0.073
			C7	-0.360
VL		-16.65	O1	-0.525
			O6	-0.464
			C2	+0.099
			C3	-0.282
			C4	-0.201
			C5	+0.662
			C7	-0.347
BL		24.07	O1	-0.523
			O6	-0.456
			C2	-0.007
			C3	-0.304
			C4	-0.193
			C5	+0.676
NMP		38.45	O6	-0.507
			C2	+0.486
			C3	-0.173
			C4	-0.296
			C5	-0.182
			C7	-0.289
			N1	-0.496

* C3 refers to methyl carbon closest to carbonyl oxygen; C4 in DMAC same as C3 in DMF.

The results of UFF/QEq calculations for other organic molecules of interest to electrolyte studies including 2,5-dimethyl tetrahydrofuran (2,5MTHF), dimethylformamide (DMF), dimethylacetamide (DMAC), propylene carbonate (PC), γ-valerolactone (VL), γ-butyrolactone (BL), and *N*-methylpyrrolidone (NMP) are summarized in Table IV.

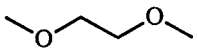
PEO Model Compounds. Several investigators have performed ab initio studies of low molecular weight model compounds for poly(ethylene oxide) (PEO). A commonly studied model compound has been 1,2-dimethoxyethane (DME) ($C_4H_{10}O_2$) $\overset{1}{H_3C}—\overset{2}{O}—\overset{3}{CH_2}—\overset{4}{CH_2}—\overset{5}{O}—\overset{6}{CH_3}$. Using the 6-31+G* basis set, Murcko and DiPaola[24] reported that DME slightly prefers a *trans* O–C–C–O orientation (*trans–gauche* energy difference of 0.4 to 0.5 kcal/mol or less). They also noted that that Allinger molecular mechanics method MM2 provided a reasonable value for the *trans–gauche* energy difference but the more recent implementation MM3 provided better values of the barrier heights to rotation. Other large basis set ab initio calculations performed by Jaffe et al.[25] indicated that the energy of the $tg^{\pm}g^{\mp}$ conformation lies only ca. 0.2 kcal/mol above that of the *ttt* conformation apprarently due to strong O···H attractions. Barzaghi et al.[26] also have performed ab initio calculations (3-21G and 6-31G*) of DME with similar results.

TABLE V. Atomic Charge Assignments for DME

Atom	AM1 Charge	PM3 Charge	UFF/QEq	ab initio[a]
C1	-0.075	0.051	-0.131	
O2	-0.260	-0.249	-0.582	-0.205
C3	-0.034	0.031	0.056	
C4	-0.034	0.031	0.056	
O5	-0.260	-0.249	-0.582	-0.205
C6	-0.075	0.051	-0.131	

[a] Partial atomic charges (esu) were obained from standard Mulliken anlaysis of D95+(2df,p) SCF wave functions by Smith et al.[27]

Use of UFF/QEq to minimize the structure of DME resulted in the *ttt* conformation illustrated below.



Results of Cerius[2] UFF/QEq and MOPAC 6 (AM1 and PM3) calculations on DME are summarized in Tables V and VI. Wherever possible, comparison is made with published ab initio results. It is noted that UFF/QEq appears to overpredict the negative charge on oxygen compared to ab initio and MOPAC results (Table V); however, bond lengths and angles are adequately represented by UFF/QEq (Table VI). This contrasts with rather good prediction of the oxygen charge for THF as discussed earlier.

TABLE VI. MOPAC 6 and UFF/QEq Parameters for DME

Parameter	AM1	PM3	UFF/QEq	3-21G[a]
Heat of formation (kcal)	-96.16	-83.16	NA	NA
Electronic energy (EV)	-4775.6	-4688.5	NA	NA
Core–core repulsion (EV)	3484.6	3473.2	NA	NA
Dipole (D)	0.1077	0.0956	NA	NA
Ionization potential (EV)	10.54	10.62	NA	NA
C–C bond length (Å)	—	—	1.517	1.516
C–O bond length (Å)	—	—	1.415	1.434
C–H bond length (Å)	—	—	1.114	
O–C–C bond angle (deg)	—	—	108.9	105.9
C–O–C bond angle (deg)	—	—	109.4	114.9
H–C–H bond angle (deg)	—	—	109.2	
COCC torsional angle (deg)	—	—	178.4	180.0
OCCO torsional angle (deg)	—	—	179.9	180.0

[a] Barzaghi et al.[26]

Crown Ethers and other Macrocyclic Compounds. Crown ethers are known to bind certain cations.[28] Due to their structural similarity with PEO and other polyethers, crown ethers and other related macrocyclic compounds are thereby appropriate models to study in order to better understand the association between cation and polyether solvent and to evaluate computational methods. Of available crown structures, the compound 12-crown-4 (1.2–1.5 Å cavity diameter), Fig. 4, is the smallest of the common crown ethers and effectively binds to the lithium cation (1.36 Å cation diameter) but not to larger alkali metal cations.

Fig. 4

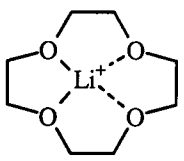
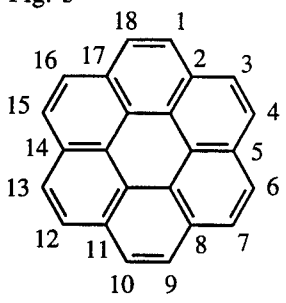


Fig. 5



The structure of 12-crown-4 was drawn using the 3D sketcher of Cerius[2]. Using UFF/QEq, the total (minimized) energy was found to be 119.66 kcal/mol with significant negatives charges on oxygen atoms (-0.57, –0.60). Next, a Li cation was added at random and the structure was minimized once again. This time, the energy reduced to 70.453 kcal/mol, showing energy stabilization of the associated structure, with the Li^+ bound to a single oxygen. Next, the crown molecule was rebuilt and Li^+ was positioned in the center of the cavity and the structure was minimized once again. This time, the energy reduced to -17.44 kcal/mol with Li^+ bound to four oxygen atoms

in center of the crown with very tight association. This is a good demonstration that UFF/QEq can be used to evaluate energetically preferred cation association.

Very recently, Ayalon et al.[29] have reported that the (molecular sandwich) corannulene (Fig. 5) and two alkyl-substituted derivatives exist in bowl-shaped form. MNDO calculations on the corannulene/4 Li^+ dimer indicated a minimum-energy geometry of "stacked bowl" (convex face to concave face) as the global minimum with four Li^+ cations sandwiched between the two corannulene molecules and two cations above and two below. Using Cerius[2], a molecule of corannulene was sketched and its structure was minimized using UFF/QEq. The minimized energy was found to be 147.7 kcal/mol and the bowl conformation was confirmed. The charge calculations indicated that the exterior carbons (carbons 1, 3, 4, 6, 7, 9, 10, 12, 13, 15, 16, and 18) to which hydrogens are attached had a weak negative charge (-0.132) while other carbons were nearly neutral (ca. +0.02). It is reasonable to assume that these carbons are sites for cation association. Next, the amorphous Builder was used to clone and build two corannulene molecules. The geometry of the nonperiodic structure was minimized (295.0 kcal/mol) and confirmed the stacked bowl model.

Poly(ethylene oxide) and Polyoxymethylene. Poly(ethylene oxide) (PEO) is a highly (~85%) crystalline polymer with an amorphous phase T_g of ca. 206 K. The crystal structure of PEO has been investigated by Takahashi and Tadokoro.[30,31] The more recent interpretation of X-ray diffraction data[31] supports a (monoclinic) unit cell with parameters given in Table VII. This structure represents a slightly distorted (7/2) helical structure compared to the often reported D_7 symmetric structure (T_2G type conformation) that had been proposed in the earlier study.[30] In this model, each unit cell (1666.8 $Å^3$ in volume) contains two right-handed and two left-handed chains. In comparison, the earlier study of Tadokoro[30] indicated that the crystalline phase of POM consists of one helical chain providing 9 repeating units per 5 turns. It is noted that this structure of POM results in a distribution of the oxygen atoms at the outside of the helix compared to the internal location of the oxygen atoms in either model of PEO.

TABLE VII. Unit Cell and Molecular Parameters of Crystalline PEO[31]

Parameter	Value
a	8.05 Å
b	13.04 Å
c (fiber axis)	19.48 Å
β	125.4°
space group	$P2_1/a$-C_{2h}^5
unit cell ($Å^3$)	1666.8
C–C bond length (Å)	1.54 Å
C–O bond length (Å)	1.43 Å
C–H bond length (Å)	1.09 Å
O–C–C bond angle (deg)	110.0
C–O–C bond angle (deg)	112.0
H–C–H bond angle (deg)	109.5

Very recently, Neyertz et al.[32] have employed MD simulations (program FUNGUS) of the PEO crystal structure using the X-ray data of Takahashi and Tadokoro[31] given in Table VII as the starting point and ab initio data for diglyme

$$CH_3—O\left(CH_2—CH_2—O\right)_2CH_3$$

to set torsional constants. A periodic box consisting of 8 crystallographic unit cells ($2 \times 2 \times 2$) containing 1568 atoms in 16 chains. Molecular dynamics at 300 K with 10 ps equilibration and 100 ps simulation indicated that the OCCO dihedral angles ($71.3 \pm 3°$) remain in their *gauche* wells and the COCC dihedral angles ($183.\pm.3\ 5°$) similarly remain in their initial *trans* wells eventhough the torsional barriers for OCCO (2.26 kcal/mol) and COCC (1.75 kcal/mol) are low. The average value of the dihedral angles is 185.4° (T) for the C–O bonds; the corresponding averages for the C–C bonds are 68.4° (G^-) in the two left-handed chain helices and 291.6° (G^+) in the two right-handed chain helices in the unit cell, giving TTG^- and TTG^+ conformations, respectively. The Rietveld method was used to generate X-ray diffraction profiles which agreed well with experimental powder diffraction data.

A PEO chain of 18 repeat units (PEO18) $\left[CH_2—CH_2—O\right]_{18}$ was generated using the Polymer Builder module of Cerius[2]. Geometric parameters are given in Table VIII. UFF/QEq charges on the interior oxygens were -0.4975 while the charge on the terminal hydroxyl group was -0.7011. It is noted that some studies have indicated that the solubility of many inorganic salts in PEO is directly related to the fraction of terminal hydroxyl groups.[33] Bernson and Lindgren[34] have used FTIR analysis of the OH stretching-vibration to study end-group coordination of PPG(3000) terminated with hydroxyl groups and concluded that there are two dominating coordination structures involving the OH groups — one is the solvent separated ion pair where the anion and the cation is separated by an OH end group and the other is an OH group coordinated both to the cation of the salt and to an ether oxygen from the polymer. For the PPG/$LiCF_3SO_3$ system, the lithium ions prefer to coordinate to OH groups rather than to ether oxygens.[35]

TABLE VIII. Geometry of PEO Obtained by Cerius[2]

Parameter	Value
C–O bond length (Å)	1.430
C–C bond length (Å)	1.540
C–O–C dihedral angle (deg)	109.0
O–C–C–O torsional angle (deg)	65.0
C–C–O–C torsional angle (deg)	-171.7

It should be noted that the PEO structure generated by the Polymer Builder of Cerius[2] gives the original D_7 symmetric structure. Use of UFF/QEq indicates that the energy is not minimum at 345 kcal/mol. Minimization reduces this to ~292 kcal/mol (alters helix) and the charge on the interior oxygens increases to -0.5080.

Atomic charges were also determined by use of the MOPAC Polymer Facility provided by Cerius[2]. This feature allows MOPAC calculations on a repeat unit from the monomer file or one constructed with special end-tagged groups for use by the Polymer Builder module. Use of MOPAC 6 (AM1) gave a dipole moment of 1.98 D and oxygen charge of -0.1575. This is substantially smaller than that obtained by UFF/QEq.

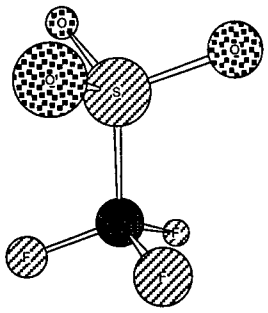
Polyoxymethylene has been reported by Sorensen et al.[36] to exist as helical chains in two crystalline forms — hexagonal and orthorhombic. Their MD simulations indicated that the packing energy of the orthorhombic form (2/1 helix and 2 chains per cell) is slightly more favorable (1.5 kJ/mol) but the free energies of the two crystalline forms are comparable at room temperature.

A polyoxymethylene chain of 18 repeat units (POM18) was built using the Polymer Builder of Cerius[2]. The resulting structure was planar zig zag with QEq charges of -0.4253 on the interior oxygens and -0.6260 on the terminal hydroxyl group. This structure contrasts with the known tight 9/5 or 2/1 helix structure reported for POM[30] and, therefore, torsion angles must be specified to get the correct conformation for this polymer. It is noted that in their paper on the QEq method, Rappe et al.[3] reported a -0.534 to -0.541 charge on interior oxygens (with no preferred conformation).

Finally, a random copolymer containing 24 PEO units and 3 POM units was constructed using the copolymer builder module of Cerius[2]. The resulting structure was a distorted D_7 symmetric helix which would certainly interfere with crystal packing. In future studies, it should be possible to investigate the possible crystal structure of this copolymer using the Crystal Packing module (not available at this time).

Salts. The lithium salts that are most important for electrolyte applications are those with low lattice energies including lithium tetrafluoroborate ($LiBF_4$), lithium hexafluoroarsenate ($LiAsF_6$), lithium perchlorate ($LiClO_4$), and lithium trifluoromethanesulfonate or lithium triflate ($LiCF_3SO_3$). Structures for the anions BF_4^-, AsF_6^-, and $CF_3SO_3^-$ were sketched and a total charge of -1.0 was assigned. The structures were then minimized using UFF/QEq and final atomic charges and geometry determined.

Optimized geometries and total energies of the triflate anion ($CF_3SO_3^-$)



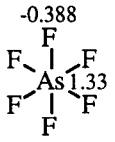
have been obtained by Huang et al.[6,37] using ab initio molecular orbital calculations (GAMESS) as shown in Table IX. Best agreement with experiment (crystal structures of triflate hydrates) was found using UHF/6-31+G*. Results of Mulliken charge calculation gave -0.928 for each oxygen atom and -0.272 for each fluorine atom in the anion. It is noted that X-ray crystal structures suggest that cations interact only with the oxygen atoms of the triflate anion rather than with the CF_3 end[37] and that the $CF_3SO_3^-$ anion has a C_{3v} point group symmetry (staggered conformation). The $-CF_3$ group is close to a regular tetrahedral geometry with a C–S–F angle of 111.8° and a F–C–F angle of 107.0° while the SO_3 group is flatter than the CF_3 groups with a C–S–O angle of 102.2° and a O–S–O angle of 115.4° (6-31+G*).

Charges determined by UFF/QEq for $CF_3SO_3^-$ are 1.285 for C, -0.540 for F, 0.194 for S, and -0.236 for O. The charges for F and O are distinctly different from those obtained by ab initio calculations (F, -0.272; O, -0.928) cited above. Results from Gaussian 92 calculations by L. Scanlon gave charges of 0.938 for C, -0.414 for F, and +1.849 for S, and -0.848 for O with bond lengths of 1.324 Å for C–F, 1.832 for C–S, and 1.443 for S–O. In comparison to the ab initio results, the UFF/QEq charge assignments for C and F are not too bad while those for S and O do not compare well. Geometric parameters using UFF/QEq are given in Table IX with comparison to experimental and ab initio values. Values for bond lengths and dihedral angles obtained by UFF/QEq are in excellent agreement with experimental and ab initio results except for those involving sulfur — the S–O bond length is about 20% larger than the experimental and the O–S–O dihedral angle is about 10% too small while the S–C–F angle agrees well with ab initio results. The results for C and S (charges and bond lengths) question the UFF/QEq parameterization for sulfur as discussed earlier for THS. It is noted that although the original publication reporting the QEq method[3] lists atomic parameters for sulfur, no examples of charges calculation with sulfur compounds were reported. Recent discussions with Molecular Simulations indicates that they are aware of the problem of S parameterization and that it is being addressed.

TABLE IX. Geometry and Energies of the Triflate ($CF_3SO_3^-$) Anion Obtained by Cerius[2]

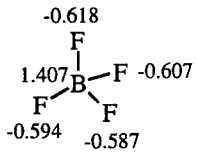
Parameter	Experimental	6-31+G*	UFF/QEq
C–S bond length (Å)	1.827	1.827	1.866
S–O bond length (Å)	1.445	1.422	1.726
C–F bond length (Å)	1.326	1.321	1.389
C–S–O dihedral angle (deg)	103.97	102.6	114.8
S–C–F dihedral angle (deg)		111.8	110.9
O–S–O dihedral angle (deg)	114.6	115.4	103.6
F–C–F dihedral angle (deg)	108.6	107.0	108.0

Next, the AsF_6 anion (octahedral)



was constructed and its structure minimized using UFF/QEq. Calculated charge for F and As were -0.388 and +1.33, respectively while the calculated As–F bond length was 1.817 Å with a F–As–F dihedral angle of 90°. Although direct ab initio results for the anion are not available, ab initio results (GAMESS) for $LiAsF_6$ by L. Scanlon indicate comparable charge assignments for fluorine (-0.301 to -0.358) and for As (1.66) and a As–F bond length of 1.759 Å.

Finally, the structure of BF_4^- (tetrahedral) was drawn, total charge set to -1.0, and minimized by UFF/QEq. Results indicate a B–F bond length of 1.411 Å and dihedral angle of 109.5°. Charge assignments ranged from -0.587 to -0.618 for F and +1.407 for B. Francisco and Williams[10] report the experimental B–F bond length as 1.389 ± 0.002 Å with ab initio values ranging from 1.395 to 1.463 Å. Scanlon reports a B–F bond length of 1.397 Å (UHF 6-31+gd) with a -0.7044 charge on F. Spoliti et al.[11] cite similar values. The UFF/QEq value of 1.411 Å for the B–F bond length and the fluorine charge are therefore reasonable.



B. Simulation Studies

Conformation of PEO and PEO/POM Copolymers. Rotational isomeric state (RIS) models for the PEO chain have been proposed by several groups.[27,38-41] Statistical weights were adjusted and RIS predictions compared to experimental values for the unperturbed chain dimensions and dipole moment and their temperature dependence.

Very recently, Smith et al.[27] have proposed that, in order to get the most accurate model, it is necessary to include *third-order* interactions (those depending upon *three* consecutive torsional angles resulting in a 9×9 statistical weight matrix) to obtain the most realistic model that can properly account for strong O···H attractions. The model parameters were based upon ab initio electronic structure analysis of the model compound 1,2-dimethoxyethane (DME) discussed earlier and diethyl ether (DEE) $H_3C—CH_2—O—CH_2—CH_3$.[†] Analysis reveals that the strength of the oxygen gauche effect for the C–C bond (the stabilization of the O–C–C–X *gauche*

[†] It is noted that, in the development of their RIS model, Miyasaka et al.[40] have used results of molecular mechanics (MM) calculations on the model compound 3,6-dioxaoctane (DOO)

$$CH_3—CH_2—O—CH_2—CH_2—O—CH_2—CH_3$$

conformation relative to the *trans* conformation correlates with the degree of Coulombic repulsion or attraction between the oxygen and the "X" moiety). In the case of DME, ab initio calculations indicate that the O–C–C–O *gauche* conformation is about 0.1 kcal/mol higher in energy than the *trans* conformation while calculations for 1,3-dimethoxypropane and methyl ether indicate that the O–C–C–CH_3 gauche conformation is significantly *lower* in energy than their respective *trans* conformation. Geometric parameters used in the RIS model were obtained from the conformational averages of ab initio optimized conformational geometries of DME. These are given in Table X. For reasons discussed in the section on Conclusions and Recommendations, no attempt was made to employ simulations methods to study PEO or copolymer conformations.

TABLE X. RIS Geometric Parameters[27]

Parameter	Value
C–C bond length (Å)	1.52
C–O bond length (Å)	1.40
C–C–O angle (deg)	111
C–O–C angle (deg)	115
C–O–C–C gauche angle (deg)	100
O–C–C–O (deg)	105

Dissociation and Association. The ability of Cerius[2] to simulate the solvation of a Li cation in DME has been investigated. As an initial approach, the Amorphous Builder module was used to add a single Li cation to DME (previously minimized structure). The structure of DME/$Li^+$was then minimized using UFF/QEq . The resulting structure of DME was the original *ttt* (see earlier discussion) with Li^+ associated with one oxygen as illustrated below (Fig. 6, Chem3D Plus representation).

Fig. 6

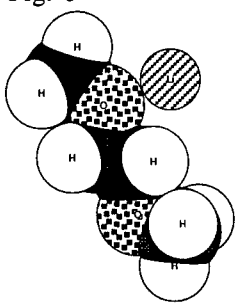
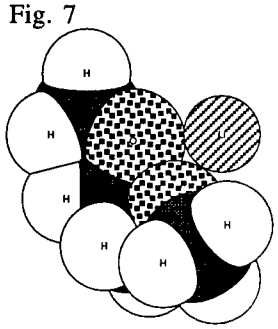


Fig. 7



Next, NVE anneal dynamics was used (5 annealing cycles, 300 K mid-cycle temperature, 50 K temperature increase, 50 steps of dynamics per increment).† The nonperiodic structure showing the conformational change of DME allowing coordination of Li^+ with the two oxygens and energy minimization is shown in Fig. 7.

In the next approach to evaluate the dynamics simulation potential of Cerius[2] for future studies, a solvent box (periodic structure) was constructed using 64 molecules of DME and 4 Li cations at a density of 0.863 g cm^{-3} (the density of DME at 20°C). This ratio was selected because the conductivity of salt-in-polymer solutions is often maximum near a molar ratio of 16 polymer ether repeat units to 1 salt cation.[42] Due to the current restriction of UFF/QEq to nonperiodic structures, the default UNIVERSAL force field was used with no Charge Equilibration. The minimized structure indicated a random DME system with dispersed Li cations. No specific associations were observed at this level. Next, a NVE quench dynamic simulation method‡ were used to investigate low energy associations. At the conclusion of the simulation, all four Li cations were coordinated to one or two oxygens. In each case, multiple associations were between Li^+ and oxygen atoms on two different DME molecules suggesting that such an association is energetically more favored compared to the conformation change required for complexation with both DME oxygens as reported above for the minimization of a single DME molecule and Li cation.

Conclusions and Recommendations

The results of this study indicate that dynamic simulations using Cerius[2] has potential for the investigation of salt dissociation/association in polymer electrolytes and provides an approach for continuing work. Due to the limitations of direct use of the UFF/QEq method for assigning correct charges (and to a lesser extent correct geometry) for the triflate anion and for polyethers as discussed earlier, the following approach is recommended. Ab initio methods (through SPARTAN or GAUSSIAN) should be used to determine charges and all molecular parameters for three anions (AsF_6^-, $CF_3SO_3^-$, and BF_4^-). Similarly, ab initio results for DME, diglyme, or 3,6-dioxaoctane should be used to assign charges to PEO. Charges and all geometry parameters can then be assigned to each molecule using Cerius[2] for the simulation work. Next, dynamic simulations of a nonperiodic box of DME/Li^+/AsF_6^- should be performed to investigate ion pairing and solvent–cation association and to compare the simulation results with the experimental work of Farber et al.[12] cited earlier. The approach developed in this study will be used to simulate the PEO/$LiAsF_6$, PEO/CF_3SO_3, and PEO/$LiBF_4$ systems. For this work, torsion angles for PEO can be set using the rotational isomeric state (RIS) parameters of Smith et al.[27] as discussed earlier. This combination of ab initio and molecular dynamics results should give a realistic simulation of polymeric electrolyte systems.

† In anneal dynamics, the temperature is altered in time increments from an initial temperature to a final temperature and back again.
‡ During quench dynamics, periods of dynamics are followed by a quench period in which the structure is minimized.

References

(1) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; III, W. A. G.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
(2) Mayo, S. L.; Olafson, B. D.; III, W. A. G. *J. Phys. Chem.* **1990**, *94*, 8897.
(3) Rappe, A. K.; III, W. A. G. *J. Phys. Chem.* **1991**, *95*, 3358.
(4) Johansson, A.; Tegenfeldt, J. *Macromolecules* **1992**, *25*, 4712.
(5) Frech, R.; Huang, W. *J. Solution Chem.* **1994**, *23*, 469.
(6) Huang, W.; Frech, R.; Wheeler, R. A. *J. Phys. Chem.* **1994**, *98*, 98.
(7) Huang, W.; Frech, R.; Johansson, P.; Lindgren, J., in press.
(8) Lightfoot, P.; Mehta, M. A.; Bruce, P. G. *Science* **1993**, *262*, 883.
(9) Lightfoot, P.; Mehta, M. A.; Bruce, P. G. *J. Mater. Chem.* **1992**, *2*, 379.
(10) Francisco, J. S.; Williams, I. H. *J. Phys. Chem.* **1990**, *94*, 8522.
(11) Spoliti, M.; Sanna, N.; Martino, V. D. **1992**, *258*, 83.
(12) Farber, H.; Irish, D. E.; Petrucci, S. *J. Phys. Chem.* **1983**, *87*, 3515.
(13) Bingham, R. C.; Dewer, M. J. S.; Lo, D. H. *J. Am. Chem. Soc.* **1975**, *97*, 1294.
(14) Dewer, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
(15) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(16) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
(17) Stewart, J. J. P. *J. Am. Chem. Soc.* **1989**, *10*, 221.
(18) Stewart, J. J. P. *J. Computer-Aided Mol. Des.* **1990**, *4*, 1.
(19) Dewar, M. J. S.; Lo, D. H.; Ramsden, C. A. *J. Am. Chem. Soc.* **1975**, *97*, 1311.
(20) Connolly, M. L. *J. App. Cryst.* **1983**, *16*, 548.
(21) Connolly, M. L. *Science* **1983**, *221*, 709.
(22) Bekturov, E. A.; Dzhumadilov, T. K. *Makromol. Chemie, Macromol. Symp.* **1992**, *58*, 233.
(23) Gray, F. M. *Solid Polymer Electrolytes*; VCH: New York, 1991.
(24) Murcko, M. A.; DiPaola, R. A. *J. Am. Chem. Soc.* **1992**, *114*, 10010.
(25) Jaffe, R. L.; Smith, G. D.; Yoon, D. Y. *J. Phys. Chem.* **1993**, *97*, 12745.
(26) Barzaghi, M.; Gamba, A.; Morosi, G. *J. Mol. Struct. (Theochem)* **1988**, *170*, 69.
(27) Smith, G. D.; Yoon, D. Y.; Jaffe, R. L. *Macromolecules* **1993**, *26*, 5213.
(28) Unruh, G. R.; Cumbest, J. H. In *Chemical Engineering Progress*; 1994; pp 54.
(29) Ayalon, A.; Sygula, A.; Cheng, P.-C.; Rabinowitz, M.; Rabideau, P. W.; Scott, L. T. *Science* **1994**, *265*, 1065.
(30) Tadokoro, H. *J. Polym. Sci., Polym. Symp.* **1966**, *15*, 1.
(31) Takahashi, Y.; Tadokoro, H. *Macromolecules* **1973**, *6*, 672.
(32) Neyertz, S.; Brown, D.; Thomas, J. O. *J. Chem. Phys.* **1994**, *in press*,
(33) Ohno, H.; Ito, K.; Ikeda, H. *Solid State Ionics* **1994**, *68*, 227.
(34) Bernson, A.; Lindgen, J., in press.
(35) Bernson, A.; Lindgren, J., in press.
(36) Sorensen, R. A.; Liau, W. B.; Kesner, L.; Boyd, R. H. *Macromolecules* **1988**, *21*, 200.
(37) Huang, W.; Wheeler, R. A.; Frech, R. *Spectrochimica Acta* **1994**, *50A*, 985.
(38) Mark, J. E.; Flory, P. J. *J. Am. Chem. Soc.* **1965**, *87*, 1415.
(39) Abe, A.; Mark, J. E. *J. Am. Chem. Soc.* **1976**, *98*, 6468.
(40) Miyasaka, T.; Yoshida, T.; Imamura, Y. *Makromol. Chem.* **1983**, *184*, 1285.
(41) Abe, A.; Tasaki, K.; Mark, J. E. *Polym. J.* **1985**, *17*, 883.
(42) Angell, C. A.; Liu, C.; Sanchez, E. *Nature* **1993**, *362*, 137.
(43) Dautzenberg, G.; Croce, F.; Passerini, S.; Scrosati, B. *Chem. Mater.* **1994**, *6*, 538.
(44) Boden, N.; Leng, S. A.; Ward, I. M. *Solid State Ionics* **1991**, *45*, 261.

SCANNING IMAGE ALGEBRA NETWORKS FOR VEHICLE IDENTIFICATION

Paul Gader, Assistant Professor
Joseph R. Miramonti, Graduate Research Assistant
Department of Electrical and Computer Engineering
University of Missouri - Columbia
Columbia, MO 65211

Final Report for:
Summer Faculty Research Program
Armament Directorate, Wright Laboratory
Eglin Air Force Base

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

and

Armament Directorate, Wright Laboratory
Eglin Air Force Base

August 1994

Scanning Image Algebra Networks for Vehicle Identification

Paul Gader, Assistant Professor
Joseph R. Miramonti, Graduate Research Assistant
Department of Electrical and Computer Engineering
University of Missouri - Columbia
Columbia, MO 65211

ABSTRACT

Digital image analysis techniques for identifying vehicles in complex scenes were studied. Neural networks that learn image algebra operations for feature extraction and classification simultaneously were applied to the problems of detecting tanks in Infrared (IR) imagery and Chevrolet Blazers in visible imagery. Results on the tanks reconfirmed earlier results with different networks that show networks are capable of generalizing from a much smaller set of examples than matched filters. The Blazers were in parking lots filled with a variety of vehicles. Several test Blazers were in the lot at a variety of ranges, aspects, and depression angles. Empirical results show that the image algebra networks can store a variety of representations of Blazers, including range, aspect and plane rotation angles. In addition, the networks exhibited the capability of generalizing to Blazers with different paint and options in some cases and could detect partially occluded Blazers. Further research is required to suppress network output on complex backgrounds.

Scanning Image Algebra Networks for Vehicle Identification

Paul Gader and Joseph R. Miramonti

I. INTRODUCTION

This report describes the results of experiments in which Image Algebra networks were applied to the problem of vehicle identification in digital images. The general problem can be stated as follows: Given a digital image of a scene that may or may not contain one or more vehicles of a certain type, indicate the location of each vehicle of that type in the scene. The vehicles in the scenes can be at a variety of aspect, depression, and plane rotation angles and ranges, can be occluded, can have a variety of paints, and can have various accessories attached. In addition, there may be vehicles in the scenes that are very similar to the vehicles sought but which are different.

We considered two types of vehicles: tanks and Chevrolet Blazers. In the first set of experiments, infrared images of tanks were used. Each image contained one vehicle. Image algebra networks were trained on a subset of the images and applied to all images with 100% success in confidently locating the tanks in the images. In the second set of experiments, much more complex visible images were used. The images consisted of scenes of a parking lot which was filled with vehicles. There were three different Blazers that appeared in the image set. One particular Blazer was used to train a variety of Image Algebra networks. The Image Algebra networks were able to find the particular Blazer at a variety of ranges that differed from the training set, at aspect angles that differed by as much as approximately 45° from those found in the training set, and which were partially occluded. The networks had difficulty generalizing to other Blazers but exhibited some capability of doing so. The difficulty lies in the fact that several vehicle types look extremely similar to Blazers and suppression of output on these vehicle types leads to suppression of output on Blazers that look different from the training Blazer.

An Image Algebra network is a variation of a multilayer feedforward neural network that learns feature extraction and classification operations simultaneously. The feature extraction operations are represented using generalized image algebra operations. The classification is performed using standard, feedforward multilayer neural networks. The Image Algebra network structure has roots in many places[1-11]. Le Cun et al at ATT [1] used the shared weight concept, also discussed in [12], to design feedforward neural networks that combine linear feature extraction and classification. In the case of linear feature extraction, our networks are also shared weight networks. Krishnapuram and Lee have used operations that range from below min to above max to learn generalized fuzzy set operations for pattern classification[13]. Yuille et. al. used a similar idea to design linear or morphological operations depending on the choice of parameters[14]. Our networks incorporate all of these notions into a network structure that can learn both linear and morphological operations [15 - 17].

Many researchers are considering the problems of vehicle identification for automatic target recognition. One popular approach is the matched filter. Several impressive studies have been done utilizing a variety of design methodologies [18 - 21]. Theoretically, a multilayer feedforward neural network can form more complex decision boundaries than matched filters and can therefore solve more complex pattern recognition problems. We have

previously shown empirically that neural networks can successfully generalize from a significantly smaller set of examples than a MACE type matched filter [20, 22]. Thus, neural network have several good properties with respect to vehicle identification. It is possible that several matched filters can achieve the same level of performance as the multilayer neural network for a lower computational cost, either due to optical implementations or to efficient digital implementations. A thorough comparison of well developed methodologies utilizing each approach on standard data sets would be useful.

Several authors have applied neural networks to target recognition, cf. [23]. Image Algebra networks have been successfully applied to character recognition and chromosome recognition problems. The results have been good in both cases. In the case of handwritten digit recognition, the classification rates computed from several thousands of test images were in the 96% - 98% range [17]. These rates are at the state of the art. In the case of chromosome recognition, the classification rates were also high, around 89%. The approach taken for vehicle identification in this study differs from these other applications in some respects. In both chromosome and character recognition, there are a large number of classes (e.g. 24 chromosome classes or anywhere from 10 to 62 character classes). Also, the characters and chromosomes were segmented from the background whereas we are not segmenting the vehicles from the background.

In section II, we describe Image Algebra networks. Following that, we describe our experimental approach. Finally, we present some results and observations from the experiments.

II. IMAGE ALGEBRA NETWORKS

Operations and Properties

Let $X \subseteq Z$ or $Z \times Z$ be a coordinate set, where Z = set of integers and F be a value set (such as the integers or reals). An image on X with values in F is a function $a : X \to F$. We denote the set of such images as F^X. A template on X is a function $t : X \to F^X$. For $x,y \in X$, we write t_y for t(y). Thus, t_y denotes an image on X and $t_y(x)$ denotes the value of the image at the location x. A template is called translation invariant if $t_y(x) = t_{\phi(y)}(\phi(x))$ for all translations ϕ with $\phi(y)$, $\phi(x)$, $x, y \in X$. The supports S_0, S_∞ , $S_{-\infty}$ are defined by $S_i(t_y) = \{ x \in X : t_y(x) \neq i \}$ $i = 0, \infty, -\infty$. If t is a template, then t* is the template defined by $t^*_y(x) = -t_y(-x)$.

The standard operations of image algebra, $\oplus$, $\boxed{\vee}$, and $\boxed{\wedge}$ are defined in [15 - 17, 24]. The $\oplus$ operation represents all linear operations. In the translation invariant case, the other operations are related to the standard gray-level morphological operations as follows: $b = a \boxed{\vee} t$ represents the dilation of a by the structuring element t and $b = a \boxed{\wedge} t$ represents the erosion of a by the structuring element t^*.

We define a hit-or-miss transform for gray-scale morphological operations as follows: Let a be an image on X and let h and m be structuring elements on X. Let b = (h,m). The hit-or-miss transform is

$$a \otimes b = a \otimes (h, m) \equiv a \boxed{\wedge} h - a \boxed{\vee} m^* .$$

This definition is motivated by the umbra transform and is fully described in [17].

We also define three generalized operations used in an Image Algebra network: weighted erosion, weighted dilation, and weighted hit-or-miss transform. For these operations, we require that F = [0,1]. Let r(x) be a non-decreasing, real-valued function defined for all real numbers with values in F = [0,1]. We define operations $\boxed{\vee}_r$ and $\boxed{\wedge}_r$ by b = a $\boxed{\wedge}_r$ t where

$$b(y) = \bigwedge_{S_1(t_y)} r[a(x) - t_y(x)]$$

and b = a $\boxed{\vee}_r$ t where

$$b(y) = \bigvee_{S_1(t_y)} r[\, a(x) + t_y(x)\,].$$

Given A = {(x, a(x)): x ∈ X} and T = {w, t; p} where w and t are templates on X and p is a real, nonzero value, we define the weighted erosion operation by:

$$E: \quad (A \boxed{\Omega} T)(y) = [\sum_{x} w_y(x) r[a(x) - t_y(x)\,]^p)]^{1/p})$$

The properties below follow from those for the generalized mean [13].

$$\text{If } w_y(x) = \frac{1}{N} \text{ for all y and x and } p \to -\infty, \text{ then } E = a \boxed{\wedge}_r t$$

and that

$$\text{If } p = 1 \text{ and } t_y(x) = 0 \text{ for all y and x, then } E = a \oplus t.$$

We define the weighted dilation operation similarly [17].

Finally, we define the **weighted hit-or-miss transform**, denoted by A ⊗ B, by letting H = {w^h, t^h; p^h} and M = {w^m, t^m; p^m} and setting

$$A \otimes B = (A \boxed{\Omega} H) - (A \boxed{\Omega} M^*)$$

where M* = {$-w^m(-x)$, $-t^m(-x)$; p^m}. An example is shown [17]. These operations can all be used and optimized as the feature extraction portion of an Image Algebra network as described in the next section.

.Network Structure

The image algebra network is composed of two parts: a feature extraction network followed by the feedforward network. The feature extraction network can consist of one or more layers, and each layer can also consist of one or more feature maps. Each layer performs feature extraction by template operations over the input to that layer. The sizes of the feature maps are determined by the undersampling rate for the convolution over their input. The feedforward network performs classification based on the outputs from the feature extraction network. Figure 1 shows the whole network structure for two-dimensional inputs.

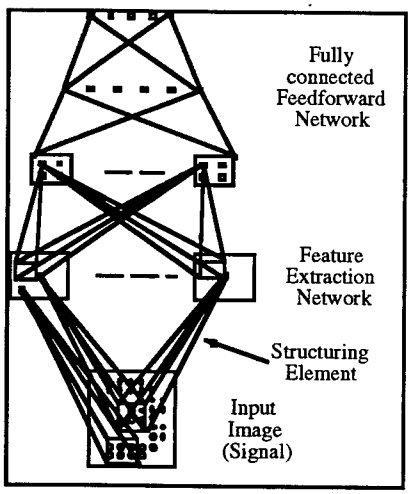


Figure 1. Network Structure.

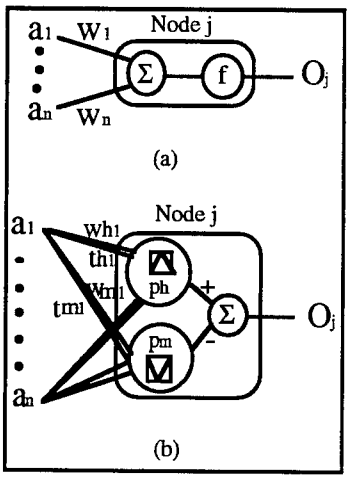


Figure 2 Node operations for the feature extraction network. Linear operation (a) and Morphological Hit-or-Miss operation (b).

The nodes in the feedforward network, except those in the input layer, compute a net input by a weighted sum of its inputs from others and produces the output by a sigmoid activation function. The input nodes for this network are simple buffers which just bypass the input (the output from the feature extraction layer) to their output.

By choosing appropriate values of p in the weighted generalized erosion operation, the nodes in the feature extraction network can perform two different operations: a weighted sum operation and an (approximate) morphological operation (We have also extended this approach to perform exact morphological operations, but this has not yet been published and was not utilized in this work). The weighted sum operation is the usual one for neural networks described above. The morphological operation node performs the weighted Hit-or-Miss transformation which is defined in the previous section. Figure 2 (a) and (b) show the node operation for those two different networks.

In our previous work (and most other neural network pattern classifiers of this type), the standard mode of operation for these networks has been that the input image or signal represents an isolated pattern (such as an image of an isolated character). The outputs of the network are class-coded. That is, there is one output for each pattern class and if the input pattern is from class i, then the i^{th} output is required to be high (typically 1.0) and all other outputs are required to be low (typically 0). In this study, we extended the standard mode of operation to allow the Image Algebra networks to operate on images which can contain one or more of the vehicles of interest in relatively small subregions of the input image. We describe the standard mode of operation more precisely here. We then describe the extension to scanning mode.

Assume we have a network with inputs that are images on a rectangular coordinate set X. Assume the network has L feature map layers and that the i^{th} feature map layer has M_i feature maps, i = 1, 2, ..., L. Each feature map is a coordinate set. Denote the j^{th} feature map of the i^{th} layer by F_{ij}. We let $F_{00} = X$. Each feature map has several

templates associated with it. In the case that the feature extraction is linear, there is one for each feature map in the previous feature map layer. In the morphological case, there are two templates for each feature map in the previous feature map layer. Assume for concreteness that the network is performing linear feature extraction. Denote by t_{ijk} the template associated with the j^{th} feature map in feature map layer i connected to the k^{th} in feature map layer i-1. Given an input image a on X, the output of the feature extraction network is computed as follows:

Let $a_{00} = a$ and $s(x) = \frac{1}{1 + e^{-x}}$ be the standard unipolar sigmoid function.

FOR $i = 1, \ldots, L$

FOR $j = 1, \ldots, M_i$

$a_{ij} = s\left(\sum_{k=1}^{M_{i-1}} a_{i-1\ k} \oplus t_{ijk} \right)$ (a_{ij} is an image on the feature map Fij)

$a_{out} = \bigcup_{j=1}^{M_L} a_{Lj}$ (a_{out} is an image on the union of feature maps $\bigcup_{j=1}^{M_L} F_{Lj}$)

The image aout is the output of the feature extraction network and is also the input to the classification network. Ritter et al have described the operation of the classification network in terms of image algebra [25]. In the morphological feature extraction case, we replace the templates t_{ijk} by pairs of templates h_{ijk} and m_{ijk} and replace the linear operation $\oplus$ by the generalized hit-or-miss operation $\boxed{\circledcirc}$.

For example, one architecture used in this study had one feature map layer with two feature maps, three hidden units in the classification network, and two output units. The values of the outputs range between 0 and 1 and represent the confidences that the input represents a Blazer or a non-Blazer. The output confidences are computed by the feedforward propagation as follows:

(AL1) Feedforward Propagation in Class Coded Mode:

a = input pattern image, (e. g. a 50 x 80 subimage extracted from a scene).

$a_{11} = s(a \oplus t_{110})$.

$a_{12} = s(a \oplus t_{120})$.

$a_F = a_{11} \cup a_{12}$.

$h_1 = s(\sum (a_F \cdot W_1))$.

$h_2 = s(\sum (a_F \cdot W_2))$.

$h_3 = s(\sum (a_F \cdot W_3))$.

$C_{target} = s(c_{t1}h_1 + c_{t2}h_2 + c_{t3}h_3)$.

$C_{background} = s(c_{b1}h_1 + c_{b2}h_2 + c_{b3}h_3)$.

where t_{110}, t_{120}, are templates from F_{11} to X and F_{12} to X respectively, W_1, W_2, and W_3, are images on $F_{11} \cup F_{12}$, c_{ij}, i = t,b; j = 1,2,3 are scalars, and $\cdot$ denotes the pointwise (Hadamard)

image product. The weights of all the templates, images, and scalars involved are learned by a backpropagation type algorithm for the generalized image algebra operations. The algorithm is derived in [17].

An Image Algebra network operating in scanning mode is an image-to-image transformation. Only the target output confidence is used to a form the output image, which can be thought of as a (nonlinear) correlation plane. A typical example is depicted in Figure 3; an analogous depiction of a class-coded network is shown in Figure 4.

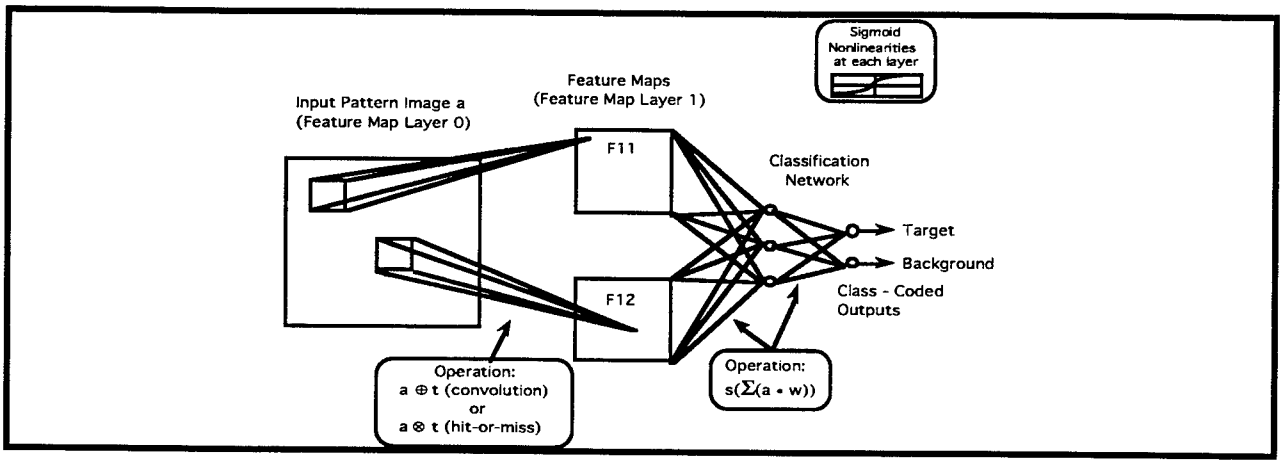


Figure 3. Typical Network Architecture in Class-Coded Mode.

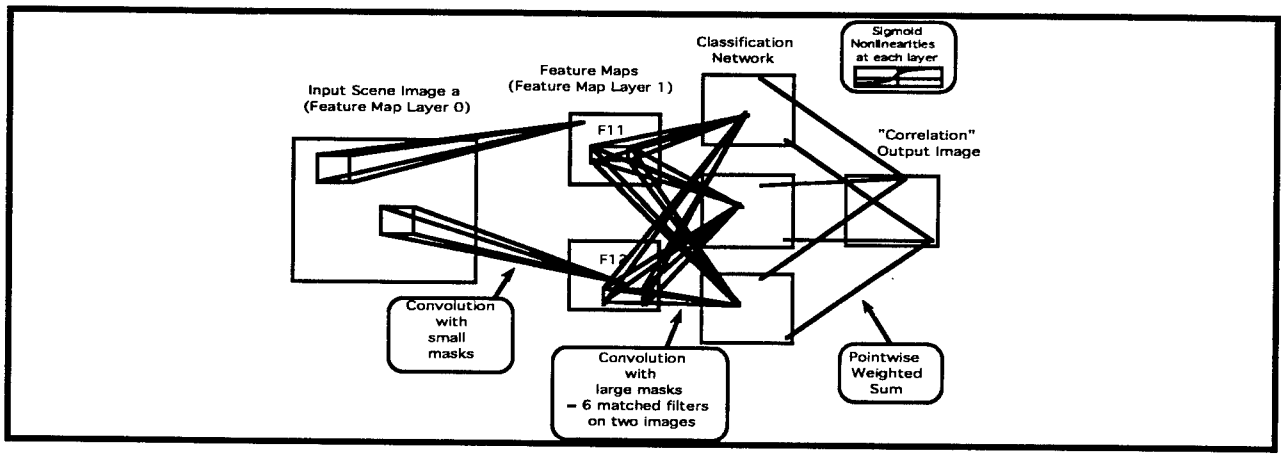


Figure 4. The Same Architecture as in Figure 3 but extended to Scanning Mode.

An Image Algebra representation of the feedforward propagation in scanning mode for the example with input image a, two feature maps, three hidden units, and two outputs is the following:

$$a_{11} = s(a \oplus t_{110}).$$

$$a_{12} = s(a \oplus t_{120}).$$

$$h1 = s((a11 \oplus W_{11}) + (a12 \oplus W_{12}))$$

$$h2 = s((a11 \oplus W_{21}) + (a12 \oplus W_{22}))$$

$$h3 = s((a11 \oplus W_{31}) + (a12 \oplus W_{32}))$$

$$C_{target} = s(c_{t1}h_1 + c_{t2}h_2 + c_{t3}h_3).$$

where t_{110}, t_{120}, are templates from F_{11} to X and F_{12} to X respectively; W_{11}, W_{21}, and W_{31}, are translation invariant templates obtained by restricting W_1, W_2, and W_3 to F_{11} and W_{12}, W_{22}, and W_{32}, are translation invariant templates obtained by restricting W_1, W_2, and W_3 to F_{12}. The supports of the Wij are called the scanning windows. In scanning mode, h_1, h_2, h_3, and C_{target} are images whereas in class-coded mode they are scalars.

It is interesting to note that the Image Algebra networks in scanning mode are feature extraction networks. One can now consider the possibility of training the network using techniques similar to those used to train matched filters. We will discuss how this observation may help to develop better methods for training the networks using constrained optimization in the analysis and conclusions section.

III EXPERIMENTAL APPROACH

Two data sets were used in our experiments: a Tank data set and a Blazer data set. The tanks data set consists of 35 infrared images of a tank in a field and was obtained from MICOM [22]. We divided the data set into two subsets: a training subset and a testing subset. These are shown in Figure 5(a) and (b)

The Blazer data set was collected during the study at Eglin. A camera mounted on a roof was focused on the parking lot. We drove a particular Blazer, which we refer to as the training Blazer, around the parking lot and captured images on video tape. The parking lot contained other Blazers and related vehicles such as Jeep Cherokees. Eighty eight 512 x 512 images were digitized from the video tape. They were categorized into the following subcategories:

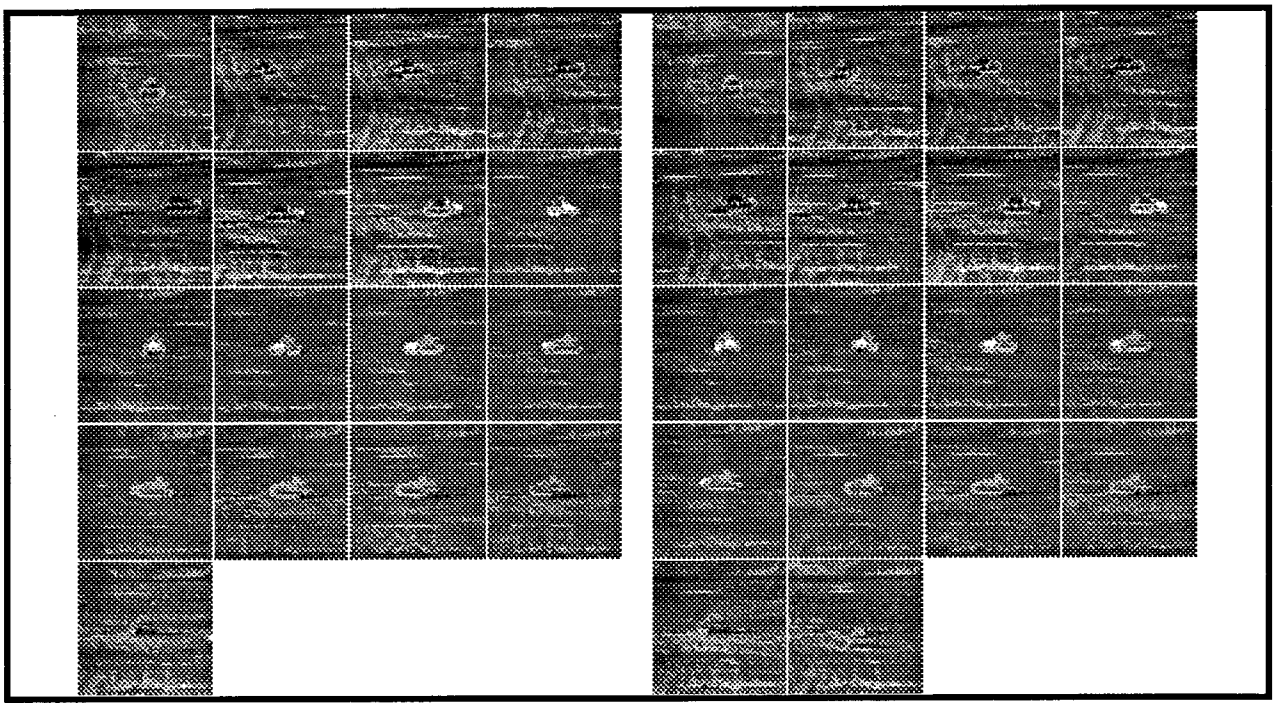


Figure 5. Tank training set (left) testing set (right).

Rotation Sequence A: A set of sixteen images containing the training Blazer at approximately the same location in the front of the lot but at intervals of approximately 20^0 in aspect Half these images are shown in Figure 6. The training Blazer is the vehicle moving between frames. The first vehicle in the row to the left of the training Blazer is also a Blazer and further back in the same row is a Jeep . All or parts of this sequence are used in training.

Rotation Sequence B: A set of twenty images containing the training Blazer at approximately the same location in the back of the lot with different views in aspect angle. Half these images are shown in Figure 6. All or parts of this sequence are used in training.

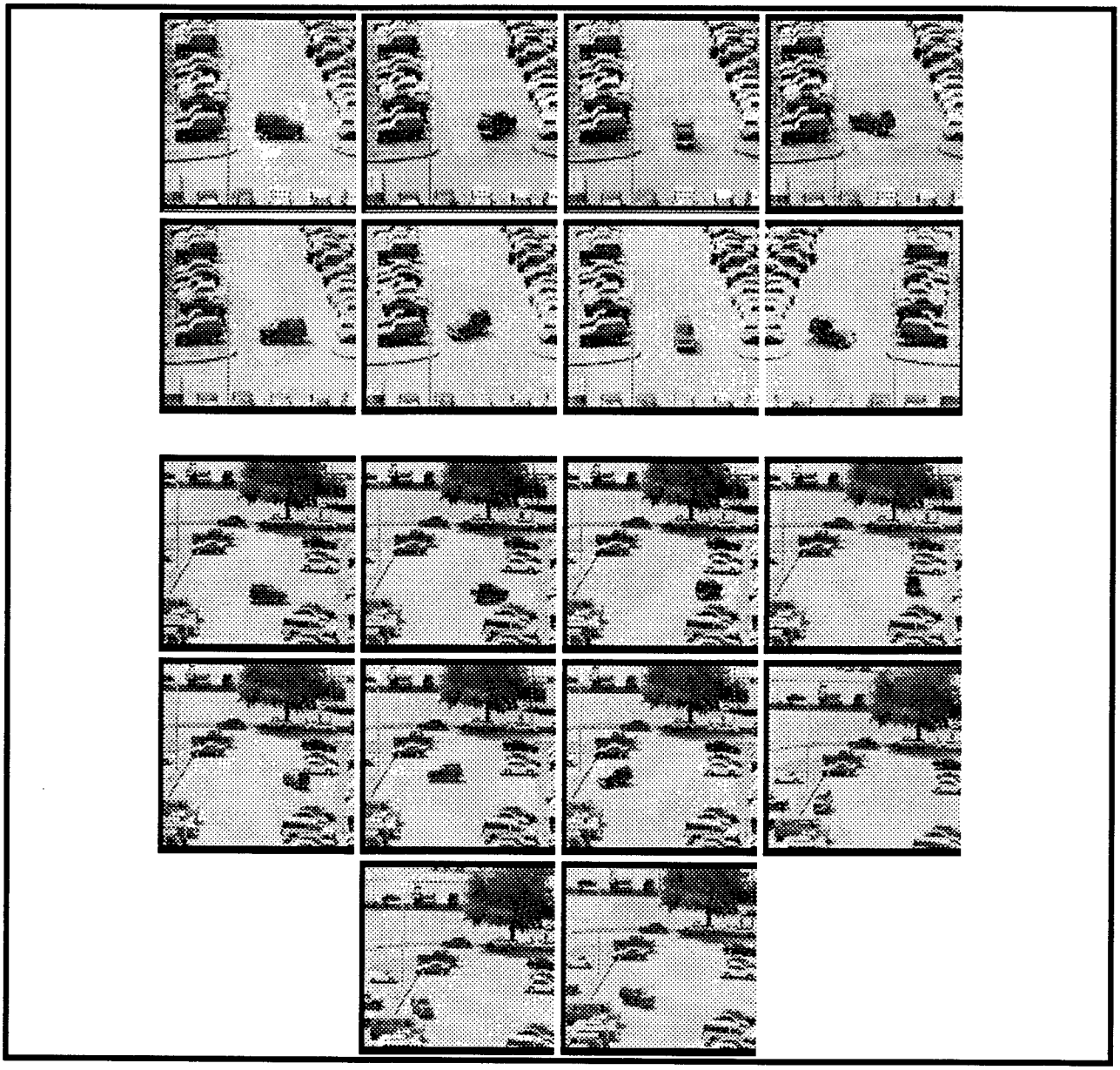


Figure 6. The Blazer training set. The Blazer driving around in the lot is the only one used to train the network.

Parking Sequence: A set of forty two images consisting of eight sequences. Each sequence consists of five or six images with the training Blazer pulling into or out of a parking space. The training Blazer ranges from no occlusion to almost full occlusion in each sequence. This sequence is used for testing only.

Jeep Sequence: A set of four images. Each image contains the training Blazer, a different Blazer, and two different styles of Jeep Cherokee. This sequence is used for testing only.

Extra Sequence: A sequence of six images of the training Blazer driving along the front of the parking lot. The training Blazer is partially occluded in some of the views. This sequence is used for testing only.

Some of the images in the testing sequences will be shown in the Analysis and Conclusion section of this report.

Training

Training consists of three basic steps: (1) Creation of a set of patterns (the training set) that is used to form the network, (2) choosing network parameters, and (3) iteratively presenting patterns to the network and applying modified backpropagation rule to update the weights. The implementation of each step requires resolution of several issues which have some generality across problem domains but which are also somewhat problem dependent.

The creation of the training set is an important function which can significantly affect the features that are learned by a neural network. As noted by Kallman and Goldstein [19], if we embed a vehicle in a constant background image, the overwhelming feature of the image is the background. It is possible that a classifier can learn to classify objects based on the background rather than the object. On the other hand, embedding vehicles in noisy backgrounds can create difficulties in training and can result in the networks learning to recognize the vehicle in specific backgrounds.

In our experiments, we tried a variety of strategies of preparing training sets and training networks. In each strategy, we used "cutout" images. Cutout images are images of vehicles with no background. Our strategies fell into two major categories: Preselected Backgrounds and Dynamically Selected Backgrounds. They are described as follows:

Preselected Background Training Algorithm:

```
Inputs:  List of subimages containing vehicles to be detected, T.
         List of randomly selected background subimages, B
Epoch = 1;
WHILE (Epoch < MaxIter AND Error > ErrorThresh) DO
      FOR i = 1 to number of patterns in T
            Select background image, a,  from B using random selection;
            Perform Forward and Backward propagation with a;
            Select target image, a,  from T using random selection;
            Perform Forward and Backward propagation with a;
            Update Error
      ENDFOR
      Epoch = Epoch + 1;
ENDWHILE
```

The lists of targets and backgrounds are created before training and held fixed. Some target images may be cutouts and some extracted directly from the scene image so that they are embedded in the background they appear in. Shifted

versions of each target are collected into the list. In our experiments, we considered shifts by as much as three pixels in each direction. Thus, 49 x 2 = 98 versions of the training vehicle from each scene were included in the training set, 49 shifts of images in backgrounds and cutouts The number of backgrounds chosen is equal to the number of targets chosen. This algorithm converges to a low RMS error and consistently trains to 100% classification accuracy. Networks trained with this algorithm do not always generate low output on test backgrounds.

Dynamically Selected Background Training Algorithm:

```
Inputs:     List of images of scenes containing training vehicles, S
            List of bounding boxes of training vehicles in scenes.
Epoch = 1
WHILE (Epoch < MaxIter AND Error > ErrorThresh) DO
      FOR i = 1 to number of scenes in T
            Select a subimage, a, containing the training vehicle subimage;
            Perform Forward and Backward propagation with a;
            Randomly select a background subimage, a;
            Perform Forward and Backward propagation with a;
            Update Error
      ENDFOR
      Epoch = Epoch + 1;
ENDWHILE
```

This algorithm is significantly different. It is much more memory efficient than the first algorithm. The backgrounds that are shown to the network are different each Epoch. This algorithm tends not to converge, although it does achieve a low error and networks trained with it perform as well as the first algorithm in scanning mode.

IV. RESULTS

We first discuss tank results. The tanks were trained using the preselected background training algorithm. No preprocessing was performed on the imagery. Half the data set was used for training and half the data set was used for testing. The images were of size 256 x 256. A network was training with two feature maps of size 5 x 5 and had two hidden units in the classification network. The scanning window was of size 50 x 80.

The scanning network was 100% accurate at detecting the tanks in both the training and testing sets. Detection was performed by selecting the maximum output of the "correlation" plane. Sample outputs on testing images are shown in Figure 7. Four examples are shown. For each example, the image in the upper left is the input. The image in the lower left is the nonlinear correlation output. The graph in the lower right is the peak obtained by setting all output values to zero except the maximum output value. The image in the upper right depicts the pixel(s) that contains that maximum output (black pixels) overlaid on the original image; the aimpoint.

These results show that a single network trained with views of a vehicle every 20^0 can represent a vehicle at virtually all rotations in aspect angle (at the same depression angle). Matched filters do not seem to be able to do this.

The Blazer networks were trained using the dynamically selected background algorithm. The images were 512 x 512. The images were preprocessed using the Prewitt edge operator. Thus, all network operations were performed on edge map images. Several different network architectures were tried. We show results from a network that had one feature map of size 5 x 5, and five hidden units in the classification network. The scanning window was of size 100 x 160.

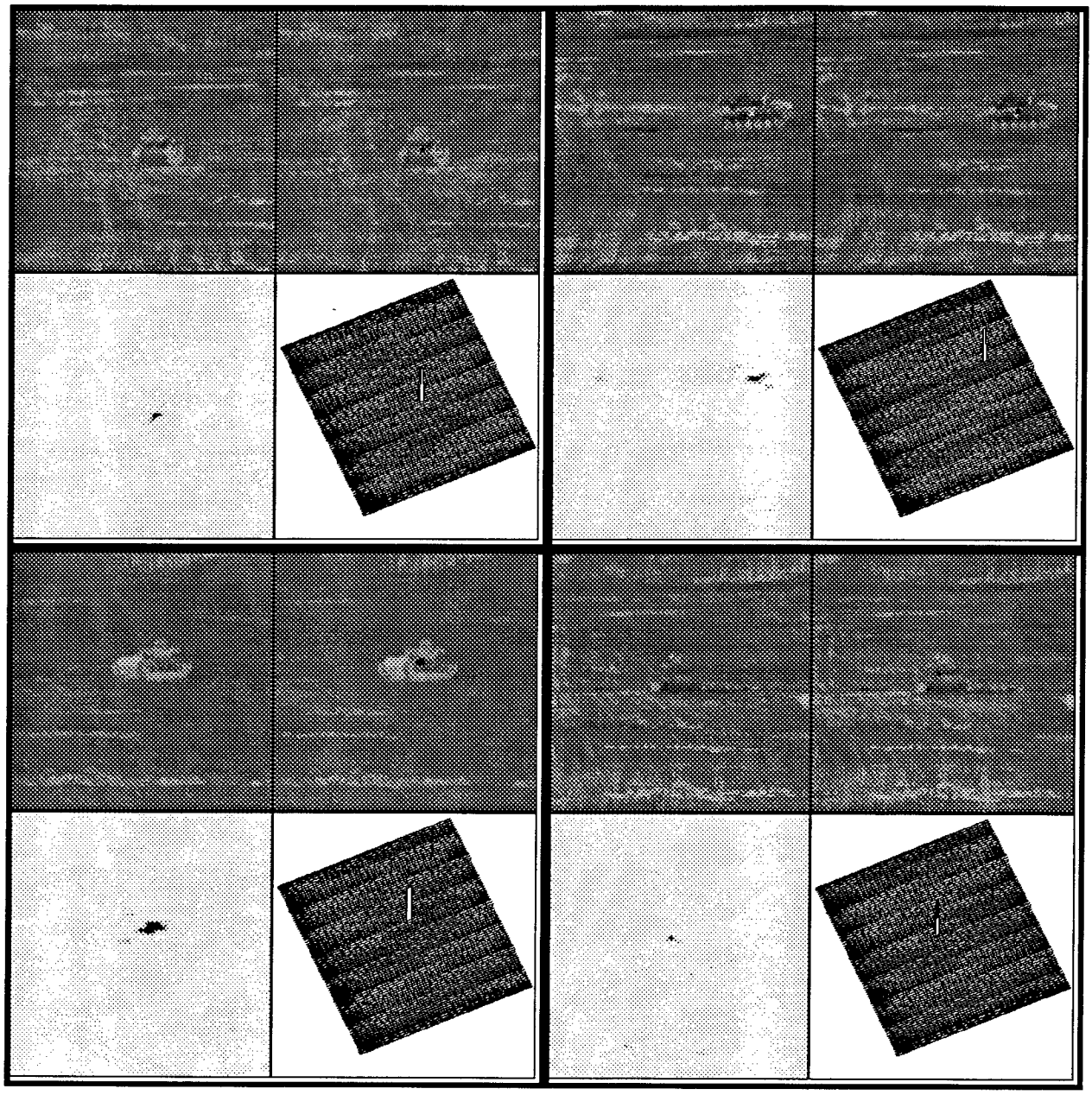


Figure 7. Sample testing results for the tank detection network.

The networks detected the training Blazer image in a variety of testing images. Precise performance measurements were not performed because of time constraints. However, inspection showed that in most images in which the training Blazer was not occluded, the network had high confidence output on the training Blazer The network sometimes detected the Blazer under partial occlusion. The network generally had medium outputs on different

Blazers, but also had medium outputs on other objects such as vans, cars, and non-vehicles. Occasionally, false alarms would have high outputs. Thus, further research is needed to fully exploit the potential of these networks.

Figures 8-10 show some of the results. Each figure shows one or more examples of results on testing images. Each example is a set of four images. The upper left image is the edge map of the input. The lower left image is the thresholded, nonlinear correlation plane (the threshold was set on the training set). The lower right graph shows all values that were above the threshold. The upper right image shows the pixels above the threshold (white pixels).

Figure 8 shows the ability of the network to discriminate between a Blazer and a Jeep in a scene. The Jeep is pulling into a parking place and is not detected by the network whereas the training Blazer, which is at a different range than the training images, is detected. Another Jeep which is parked is not detected. However, the vehicle in the front of the left row of cars is a Blazer and is not detected. In Figure 9, the training Blazer undergoes various levels of occlusion. The Blazer is detected with partial occlusion, but not with a large amount of occlusion. Finally, in Figure 10 an example of a false alarm is shown.

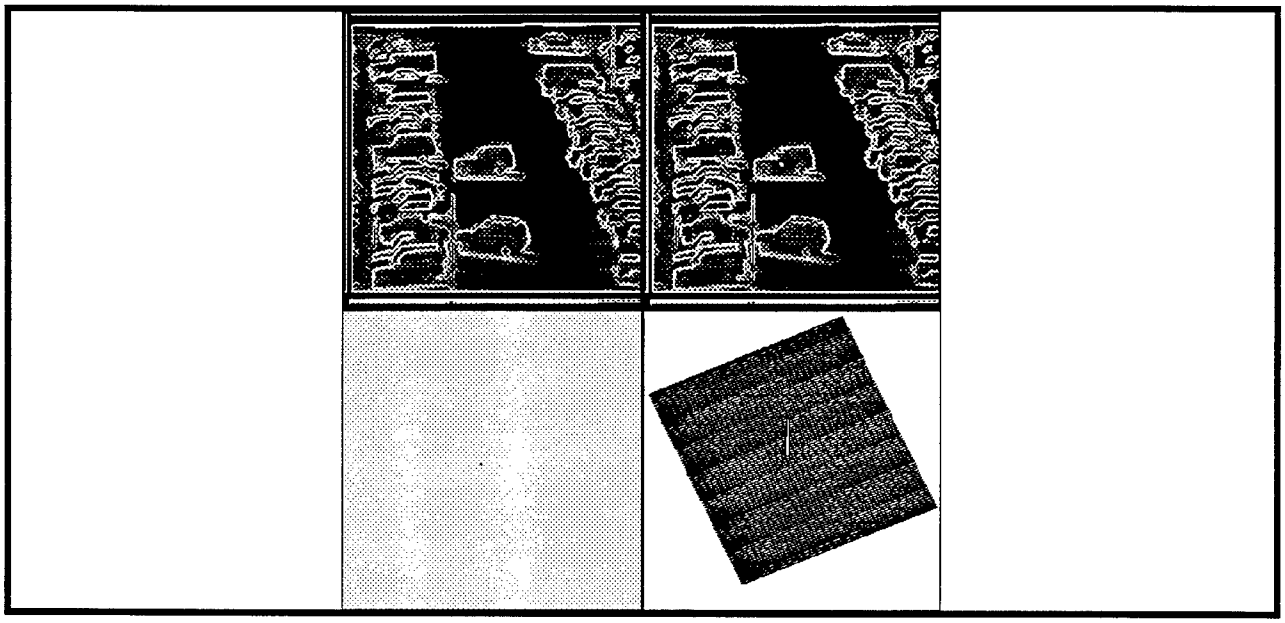


Figure 8. Testing results of Blazer network showing ability to disciminate between Blazer and Jeep but inability to generalize from one Blazer to another

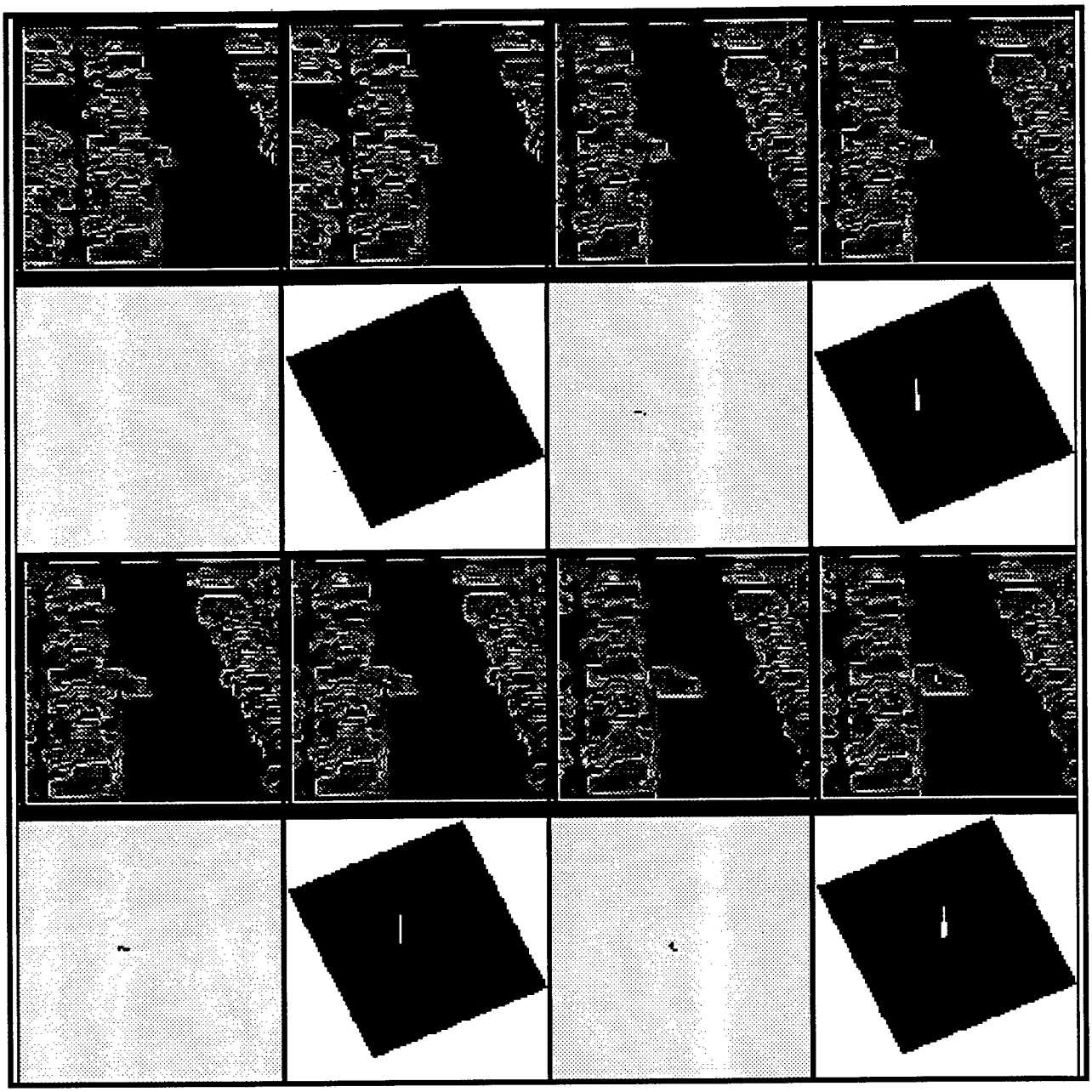


Figure 9. Testing results of Blazer network showing ability to detect Blazer under partial occlusion but not large amounts of occlusion .

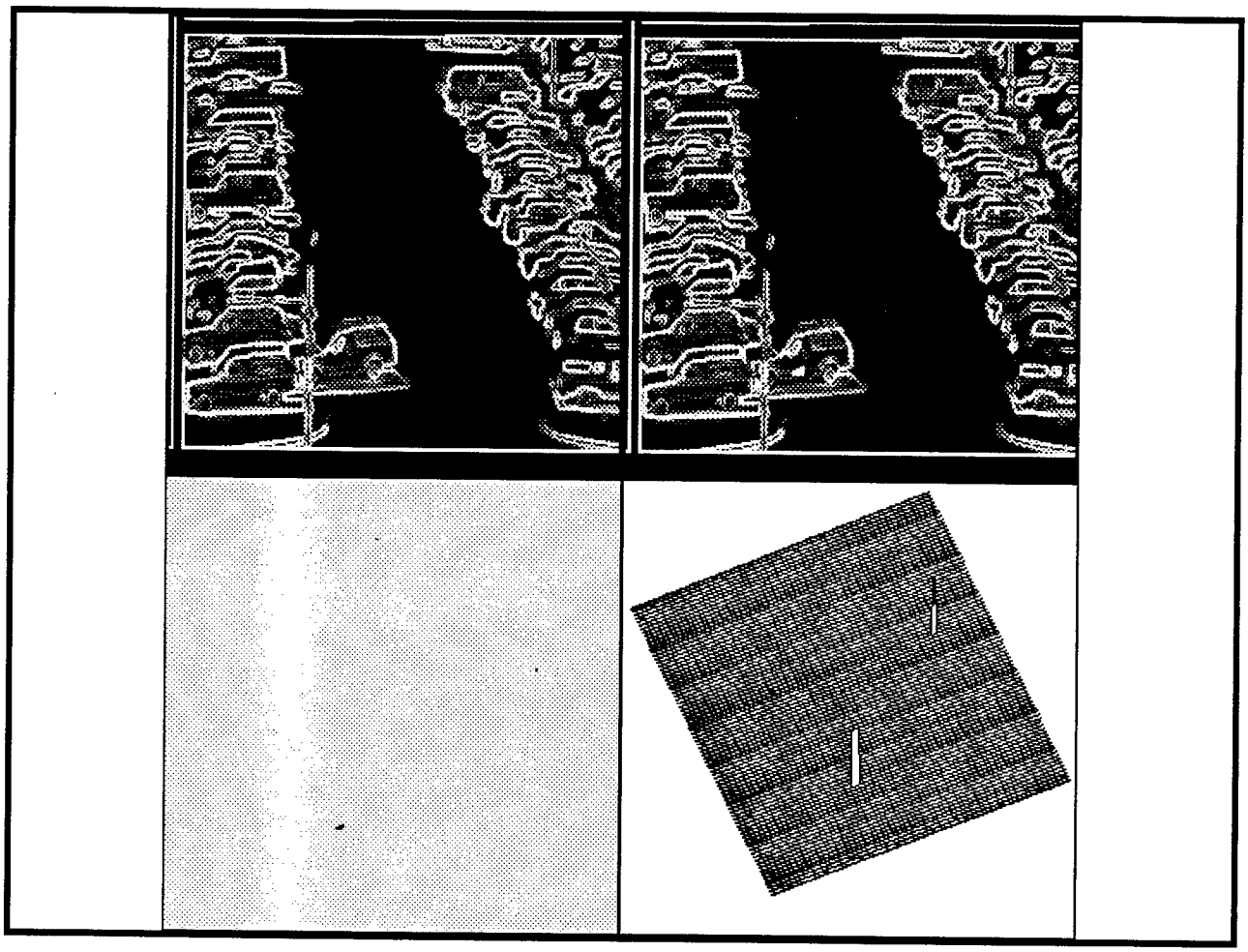


Figure 10. Testing results on Blazer showing false alarm in cluttered region consisting of cars, detection of training Blazer, and missed detection of non-training Blazer.

V. ANALYSIS AND CONCLUSION

The Image Algebra scanning network approach to detecting vehicles in complex scenes is a promising approach that needs further investigation. The universal approximation property of neural networks provides a theoretical basis for believing that the networks can provide improved detection rates. Training methodologies need to be developed that can minimize spurious outputs and computational requirements must be fully understood. Once mature, the networks should be compared to existing techniques on standard data sets using meaningful measures of comparison. Detection and false alarm rates, computational complexity, training or design requirements should be considered.

An important conceptual result obtained in this study was the idea that the Image Algebra networks in scanning mode should be trained as image-to-image transformations rather than image-to-class-coded outputs transformations. This idea allows us to consider better design criteria. Feedforward neural networks are usually trained using unconstrained optimization. This is analogous to designing a Synthetic Discriminant Function (SDF). However, SDF's do not perform well, producing many spurious outputs. Thus, much research in the design of matched filters

has focused on the development of constrained optimization criteria [19,20]. These constrained optimization criteria produce extremely significant enhancements of the matched filter performance, producing sharp correlation peaks without the spurious outputs associated with the SDF's. It is reasonable to assume that using such criteria will lead to equally significant enhancements of the scanning Image Algebra network performance.

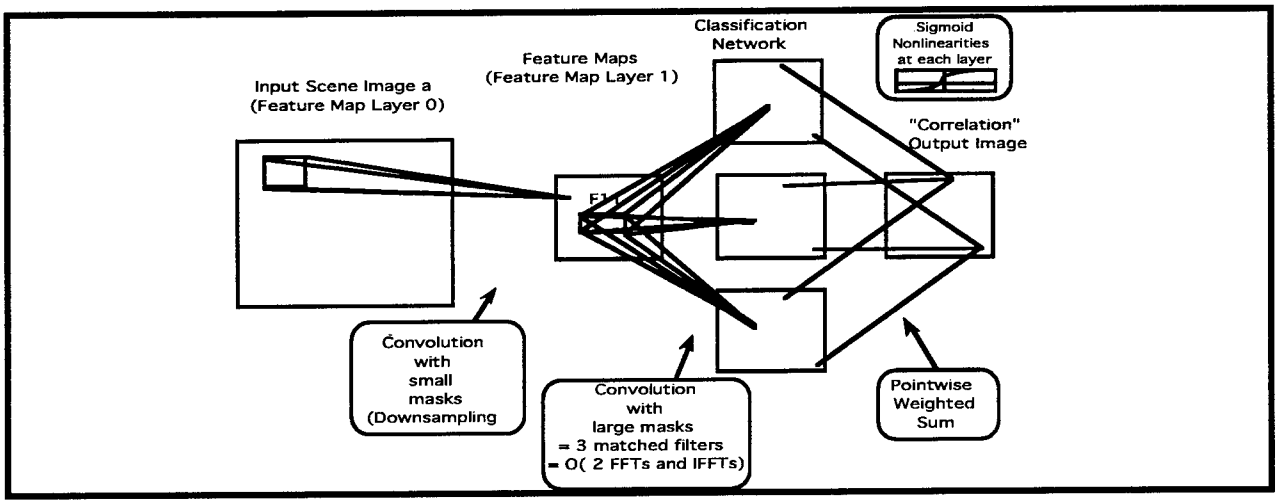


Figure 11. A Network Architecture in Scanning Mode with only one Feature Map and Feature Map Layer. In this case, the Feature Extraction Network performs a Downsampling Operation which is optimized

Frequency domain techniques in the design process have helped to produce better filter designs. We can modify our training procedure in the scanning mode to operate in the frequency domain. For example, consider the network shown in Figure 11. This network has one feature map, say for concreteness of size 5 x 5, which essentially performs downsampling. In the current mode of operation, this stage is followed by scanning with a window large enough to contain all the targets. This scanning window could be replaced by a frequency domain filtering approach. Training could be performed in the frequency domain using a modification of backpropagation.

Let us be more precise. Assume there is no downsampling layer. The forward propagation of an image, a, through a network with n_h hidden units is given by:

$$h_i = s\,(F^{-1}(F(a) \cdot Hi)) \qquad i = 1, 2, \ldots, n_h$$

$$C(a) = s\Big(\sum_{i=1}^{n_h} w_i\, h_i\Big)$$

where F (a) denotes the Fourier transform of a and F^{-1} denotes the inverse Fourier transform. Note that each h_i is an image and can be thought of as the output of a matched filtering operation with filter H_i. In this view, the scanning Image Algebra network is a collection of matched filters that are trained simultaneously. The outputs of the matched filters are combined to produce an output image. Thus, the network servers as a generalization of a matched filter and is identical to a matched filter if the number of hidden units is 1. The beauty of this view of the scanning network, as mentioned previously, is that we can now modify the weights, or filter coefficients, according to a

constrained optimization criteria. For example, a Minimum Average Correlation Energy type constraint could take the following form:

Given training images $a_1, a_2, ..., a_s$, each with associated target points $\{x_{i1}, x_{i2}, ..., x_{ini}\}$

Minimize $$\sum_{i=1}^{s} \left(\Sigma\, C(a_i)^2 \right)$$

subject to the constraints that $C(a_i)(x_{ij}) = d_{ij}$.

One approach to solving this would be to use Lagrange multipliers, i.e.,

Minimize $$E = \sum_{i=1}^{s} \left(\Sigma\, C(a_i)^2 \right) + \lambda \left(\sum_{i} \sum_{j} \left(C(a_i)(x_{ij}) - d_{ij} \right) \right).$$

The approach to minimizing this expression would be to mimic the backpropagation procedure using the forward propagation equations provided above. This requires differentiating the output of the matched filtering process with respect to the filter coefficients in the frequency domain, which is not difficult. The use of constrained optimization in training neural networks has received a small amount of attention recently in the neural network community and a review of proposed techniques should be helpful for this application [26-28].

Furthermore, in this new scheme, the initial weights of the frequency domain filters could be picked in more intelligent ways than purely random. For example, Kallman and Goldstein use the notion of a Basic False Target to initialize several frequencies to zero in the matched filter and they require that the frequencies remain zero during the constrained optimization process. This process can be utilized as well to enhance the network training.

In addition to investigating the implications of the idea of training the scanning networks in image-to-image mode, many other experiments can be performed. These include the following:

Training the network with morphological feature extractors rather than linear feature extractors. This approach could result in faster processing and more appropriate features in some cases.

Train the network with morphological classification networks. This could result in faster processing. These networks have been investigated by Ritter et el.

Investigate the potential of a single network to produce outputs for several target types. More generally, characterize the extent of the variance that can be included in a single network.

Investigate the affects of small changes in the input on the state of the network; an analysis of variance. Characterize the outputs of the network in a more predictable fashion.

REFERENCES

1. Y.L Cun, B. Boser, J. Denker, D. Henderson, R. Howard, W Hubbvard, L. Jackel, and H. Baird. "Constrained Neural Network for Unconstrained Handwritten Digit Recognition, Proc. Frontiers in Handwriting Recognition, CENPARMI, Concordia University, Montreal CA, April 1990.
2. P.D. Gader, and M.A. Khabou, "Automated Feature Generation for Handwritten Digit Recognition," submitted to *IEEE Trans. Pattern Analysis and Machine Intelligence.*, 1993.
3. A. M. Gillies, "Automatic generation of morphological template features", SPIE Image Algebra and Morphological Image Processing I, vol-1350, pp 252 - 262, San Diego CA, July 1990.
4. Wm. F. Pont and P.D. Gader,"Gradient Descent Techniques for Feature Detection Template Generation," SPIE Image Algebra and Morphological Image Processing II, vol-1568, San Diego CA, July 1991.
5. M. Ritzki, L. Tamburino, M. Zmuda, "Adaptive search for morphological feature detectors", SPIE Image Algebra and Morphological Image Processing I, vol-1350, pp 150 - 160, San Diego CA, July 1990.
6. S. Wilson,"Unsupervised Training of Structuring Elements", SPIE Image Algebra and Morphological Image Processing II, vol-1568, San Diego CA, July 1991.
7. F. Stentiford, "Automatic Feature Design For Optical Character Recognition Using An Evolutionary Search Procedure." IEEE Trans. Pattern Anal. Machine Intell., Vol. PAMI-7, pp. 349-355, May 1985.
8. E. Dougherty, "Optimal Mean-Square N-Observation Digital Morphological Filters I Optimal Binary Filters II Optimal Gray-Scale Filters", CVGIP vol-55, pp 36-72, Jan 1992.
9. E. Dougherty and R. Loce, "Optimal Mean-Absolute-Error Hit-or-Miss filters: Morphological Representation and Esitmation of the Binary Conditional Expectation", Optical Engineering, Vol 32. No. 4, April 1993, pp.815-827.
10. R. Loce, *Morphological Filter Mean-Absolute-Error Representation Theorems and Their Application to Optimal Morphological Filter Design*, Ph.D. Thesis, Rochester Institute of Technology, May 1993.
11. J. Davidson and K. Sun,"Template Learning in morpohological nets,"SPIE Image Algebra and Morphological Image Processing II, vol-1568, pp 176-187, San Diego CA, July 1991.
12. D. Rumelhart and J. McClelland, Parallel Distributed Processing,MIT Press, Cambridge Mass, 1986.
13. R. Krishnapuram and J. Lee,"Fuzzy-Set-Based Hierarchical Networks for Information Fusion in Computer Vison", Neural Networks, Vol. 5, pp 335-350, 1992.
14. A. Yuille, D. Geiger, and L. Vincent, "Statistical Physics Interpretation of Mathematical Morphology", SPIE Image Algebra and Morphological Image Processing II, vol-1568, San Diego CA, July 1991.
15. P.D. Gader, "Template Generation for Pattern Classification," *Proceedings of the SPIE Conference on Image Algebra and Morphological Image Processing III*, Vol. 1796, July 1992.
16. P.D. Gader, "Fuzzy Morphological Networks, " *Proceedings of the First Midwest Electro-Technology Conference*, Ames, Iowa, April 1992.
17. P. D. Gader, Y. Won, M. A. Khabou, "Image Algebra Networks for Pattern Classification", *Proceedings of the SPIE Conference on Image Algebra and Morphological Image Processing V,* Vol. 2300, July 1994.
18. *Optical Engineering*, Special Section on Optical Pattern Recognition, J. Horner and B. Javidi (guest eds.), Vol. 33, No. 6, June 1994.
19. R. R. Kallman and D. H. Goldstein, "Phase-encoding input images for Optical Pattern Recognition", Optical Engineering, Vol. 33, No. 6, June 1994, pp. 1806-1813.
20. A. Mahalanobis, BVK Vijaya Kumar, and D. Casasent, "Minimum average correlation energy filters," *Appl. Opt.* Vol. 26, 1987, pp. 3633-3640.
21. S. R. Sims, J. Epperson, B. V. K. Kumar, A. Mahalanobis, "Synthetic Discriminant Functions Using Relaxed Constraints", Proc. SPIE Automatic Object Recognition IV, Orlando, April 1993.
22. G. Hobson, S. R. Sims, P. Gader, J. Keller, "MACE Prefilter Networks for Automatic Target Recognition", Proc. SPIE Automatic Object Recognition IV, Orlando, April 1994.
23. C. Daniell, D. Kemsley, W. Lincoln, W. Tackett, and G. Baraghimian, "Artificial Neural Networks for Automatic Target Recognition", *Optical Engineering*, Special Section on Automatic Target Recognition, F. Sadjadi (guest ed.), Vol. 31, No. 12, December, 1992, pp. 2521-2531.

24. G. Ritter, J. Wilson, and J. Davidson, "Image Algebra: An overview", *CVGIP*, vol. 49, Mar. 1990, pp 297-331,.

25. G. Ritter, D. Li, and J. Wilson, "Image Algebra and its Relationship to Neural Networks, in Proc SPIE Technical Symposium Southeast on Optics, Electro-Optics, and Sensors, Orlando FL, March 1989.

26. F. Fogelman Soulie, "Integrating Neural Networks for Real World Applications" ,in Computational Intelligence: Imitating Life, Zurada, Marks, Robinson(eds), IEEE Press, 1994.

27 D. Mckay, "A Practical Bayesian Framework for Backpropagation Networks", *Neural Computation*, vol. 4, 1992, pp. 448-472.

28 S. Nowlan, G. Hinton, "Simplifying Neural Networks by Soft Weight Sharing", Neural Computation, vol. 4, 1992, pp. 473 - 493.

LASER RADAR
PERFORMANCE MODELING AND ANALYSIS
WITH EMPHASIS ON BMDO'S ASTP PROGRAM

Philip Gatt
Research Scientist
Center for Research and Education in Optics and Lasers
CREOL

University of Central Florida
12424 Research Prky.
Orlando, FL 32826

Final Report for
Summer Faculty Research Program
Wright Laboratory
WL/MGS

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

and

Ladar Development, Evaluation, and Research Facility (LDERF)
Eglin AFB
WL/MNGS

September 1994

LASER RADAR
PERFORMANCE MODELING AND ANALYSIS
WITH EMPHASIS ON BMDO'S ASTP PROGRAM

Philip Gatt
Research Scientist
CREOL
Center for Research and Education in Optics and Lasers
University of Central Florida

Abstract

This report presents the highlights of the results of Dr. Gatt's AFOSR Summer Faculty Research Program. Particular emphasis is given to those activities which related to performance modeling of laser radar systems. Two specific ladar systems were analyzed. The first was, a solid-state coherent detection ladar system, for BMDO's Advanced Sensor Technology Program (ASTP). The second was a direct detection angle-angle-imaging ladar currently in field use at Eglin AFB (WL/MNGS). Results of numerical analyses of the performance models for these two systems are provided in this report.

In addition to these analyses, a brief description of the other major tasks conducted by Dr. Gatt is also given in this report. These tasks included, (1) the design of a laboratory coherent laser radar test bed, (2) the design of an AFGL FASCODE interface program for a personal computer, and (3) teaching of a laser radar short course which compared and contrasted direct detection ladars with coherent detection ladars.

LASER RADAR
PERFORMANCE MODELING AND ANALYSIS
WITH EMPHASIS ON BMDO'S ASTP PROGRAM

Philip Gatt
Research Scientist
CREOL
Center for Research and Education in Optics and Lasers
University of Central Florida

Introduction

This report describes the research activities conducted by Dr. Gatt as an AFOSR Summer Faculty Researcher, working at the Ladar Development, Evaluation, and Research Facility (LDERF) at Eglin AFB, WL/MNGS. During the 1994 AFOSR SFRP, Dr. Gatt assisted the LDERF staff in the following two major areas, (1) laser radar consulting and (2) laser radar performance modeling.

As a consultant, DR Gatt's primary contributions consisted of (1) teaching a short course on laser radar, titled "Laser Radar - An Overview", (2) conducted an analysis of BMDO's Advanced Sensor Technology Program (ASTP) as it related to eye-safe coherent laser radar subsystems requirements, and (3) designed a laboratory coherent laser radar testbed.

As a ladar theoretician, Dr. Gatt (1) designed and developed a FASCOD3p PC interface program named "RunFas.Exe", (2) conducted ASTP scenario atmospheric transmission simulations using AFGL's FASCOD3p and the PC interface program (RunFas.Exe), and (3) designed several direct detection laser performance models and employed these models to predict range performance of an angle-angle-range imaging ladar, with characteristics similar to the Schwartz Electro-Optics imaging ladar, which is currently in use at LDERF. This report documents these highlights of Dr. Gatt's summer research activities.

Laser radar short course "Laser Radar - An Overview"

A full day short course on the fundamentals of laser radar was taught. This course provided both a fundamentals section and an advanced topics section. In the fundamentals section, the laser radar equation was presented followed by an analysis of the effects of target speckle, atmospheric turbulence, and background light on coherent and incoherent receivers. The advanced topics section included discussions on heterodyne (mixing) efficiency, frequency tracking local oscillators, fiber-optic mixing, pulse compression, laser vibration sensing, Range-Doppler signal processing, and coherent arrays. The notes from this short course will serve as a first draft for a, soon to be announced, SPIE laser radar short course.

Design of a laboratory coherent laser radar testbed

The design down to the component level, including a list of suggested vendors, for a laboratory continuous wave, coherent laser radar testbed was developed. A monostatic offset Homodyne configuration was selected for this testbed, which is depicted in Figure 1. A technical description of this testbed is provided below.

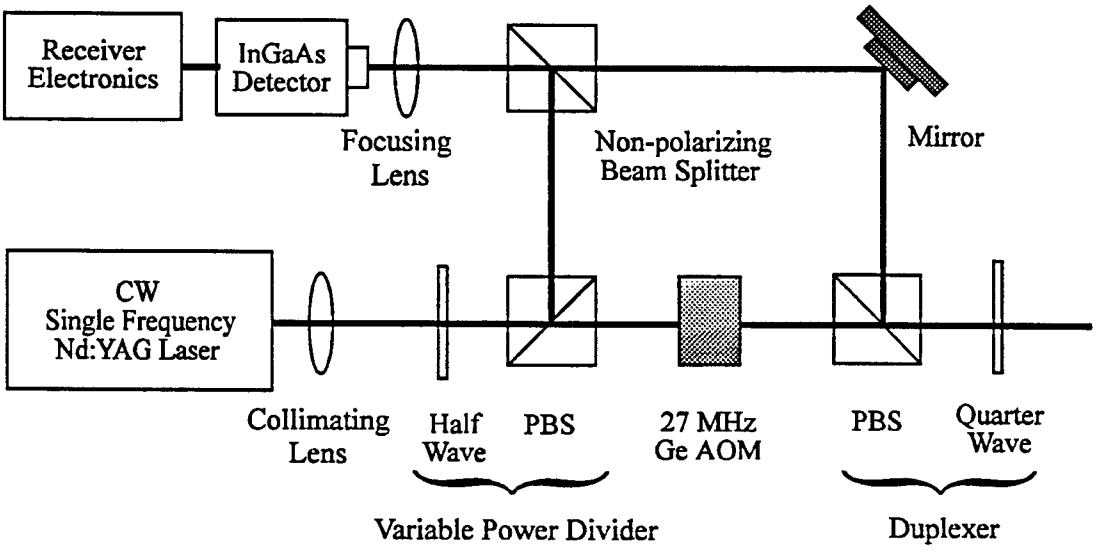


Figure 1. Monostatic offset-homodyne coherent laser radar testbed

Creation of the local oscillator and frequency offset transmitter fields

A small portion of the laser power, on the order of 1 mW, is deflected into the local oscillator beam, via a variable power divider consisting of a half wave plate and polarizing beam splitter. Due the nature in which the PBS reflects and transmits, the polarization of the local oscillator field will be vertical and that of the transmitted beam will be horizontal. The horizontally polarized transmit beam is then passed through an Acousto Optic modulator, which generates the desired offset frequency in the first diffracted order.

Offset frequency related issues

One may be tempted to improve transmitter efficiency by repositioning the AO modulator into the LO beam path. This approach will yield the desired frequency offset and increased transmitter efficiency, at the expense of considerable signal interference. The signal interference is caused by bulk-crystal scattered photons from the zero order beam into the first diffracted order. While the power of these scattered photons is only a minute fraction of the total incident power, it is typically comparable to, if not greater than, the receiver's noise equivalent power (NEP). Thus, maximum sensitivity will be compromised by this configuration. However, when the AO modulator is placed in either the transmit or received path, these unwanted scattered photons are of no

concern, since as mentioned earlier, these scattered photons constitute only a minute fraction of the total incident power.

Often, one will see two modulators employed in a CW offset homodyne ladar, one in the LO path and one in the transmit or receive path. Each of these modulators is driven by a unique RF frequency. Therefore, the static-target signal frequency, which occurs at the difference (or sum, depending upon whether the plus or minus diffracted orders are used) of the two AO drive frequencies, is shifted away from either RF driver frequency. The inclusion of the second modulator is motivated by a desire move the static-target signal frequency away from the RF drive frequency, thereby eliminating RF interference. At CREOL however, we have found that by employing proper RF shielding techniques, the RF interference can be reduced to a level far below the NEP of the receiver, and therefore we have no need for the second modulator.

Optical duplexer

The horizontally polarized frequency shifted transmitter field is then transmitted through the optical duplexer, which consists of a properly oriented quarter wave plate and a second polarizing beam splitter. The transmitted field becomes circularly polarized (right or left hand depending upon wave plate orientation) upon passage through the quarter wave plate. The reflected field, from a non-depolarizing target, will therefore also be circularly polarized, however of the opposite hand, due to a reversal in propagation direction. This polarization of the reflected field will then be converted to a vertical polarization after passage through the half wave plate and deflected into the receiver path by the second polarizing beam splitter.

Mixing of the LO field with the signal field

At the non-polarizing beam splitter, both the local oscillator and the signal beams will be vertically polarized, and will therefore interfere properly. The mixing of these two fields occurs at the beam splitter not on the detector, a subtle point that is often misunderstood. The efficiency with which these two fields mix (interfere) depends upon the respective transverse fields. If the transverse fields are perfectly matched (a difficult condition to achieve) then 100 % mixing efficiency is theoretically possible.

Since the two fields will rarely be matched, careful consideration must be used when specifying the detector lens F-number (F#) and detector size. If too slow of a lens is employed (large signal beam spot relative to the detector diameter) then an excessive number of signal photons will be lost. However, if the lens is too fast (small signal beam spot relative to the detector diameter) then mixing efficiency will be reduced due to destructive interference of adjacent fringes on the detector. In other words, if the fields are not perfectly matched, there will be an optimum detector size for a given lens F#. The exact ratio of the beam spot size and detector size is highly dependent upon the type of fields used. For example if the signal field is assumed to be an Airy pattern on the detector and if the LO field is uniform, the optimum detector diameter turns out to be about 70 % of the diameter of the first null of the Airy pattern. At, this optimum detector size the mixing efficiency is coincidentally on the

order of 70% as well. For a detailed discussion on mixing efficiency, one should consult the journals. In particular David Fink's 1976 Applied Optics paper is considered to be a seminal paper, and is well worth reading.

Advantages of a single mode fiber optic coherent receiver

Single mode fibers and fiber components can replace many of the bulk optic components shown in Figure 1. Advantages of fiber optics include, stability, ruggedness, compactness and increased mixing efficiency. The mixing efficiency is improved, since the fiber counterpart to the bulk-optic non-polarizing beam combiner, a four port fiber coupler, will mix two near perfectly matched fiber fields. The tradeoff then is, of course, between bulk-optic mixing efficiency and free-space to fiber launch efficiency. This tradeoff is currently being studied by several prominent researchers in the field.

Applications

The CW testbed described in the preceding paragraphs, will provide a laboratory testbed for numerous applications which differ by the type of signal processing employed. For example, phase demodulation can be employed, to yield an electronic signal which is proportional to the vibration signature of the target. On the hand, frequency demodulation would yield the target's Doppler signature.

Advantages of the monostatic Offset-Homodyne configuration

The advantage of the monostatic configuration is that, contrary to an incoherent ladar, a coherent ladar's field of view is quite often equal to the diffraction limit of the receiver optics. Therefore, co-alignment of a bistatic receiver's field of view, with the transmitted spot, at the target, is range sensitive and usually requires high pointing accuracy and stability (on the order of 10 to 100 μrad). The disadvantage of the monostatic configuration is "transmitter-feed-through." Transmitter-feed-through, which interferes with signal photons, may or may not be an issue depending upon the specific optical configuration and receiver signal processing. By using high quality optical components and employing proper alignment procedures, transmitter-feed-through effects can usually be reduced to a negligible level.

The advantage of the offset homodyne configuration over a heterodyne configuration is cost and simplicity. The AO modulator is much less expensive and doesn't require the complex frequency locking electronics associated with a two laser heterodyne configuration. A pure Homodyne configuration, on the other hand, suffers from increased noise, due to 1/f noise, in addition to an inability to discriminate positive from negative Doppler frequency shifts.

FASCOD3p PC interface program "RunFas.Bas"

Since the late 1960's the Air Force Geophysics Lab (AFGL) has been developing atmospheric modeling software, which, among others, predicts the intensity transmittance of a field as it propagates through the Earth's atmosphere. These algorithms cover the spectral band from zero cm-1 (infinite wavelength) to 20,000 cm-1 (0.5 μm) . The conversion between wavelength (λ) measured in microns and wavenumber (ν) measured in inverse-cm is simply,

$$\lambda = \frac{10{,}000}{\nu} \tag{1}$$

The AFGL packages of interest are: (1) LOWTRAN, (2) MODTRAN, (3) HITRAN & FASCODE. LOWTRAN is a low spectral resolution algorithm which uses mathematical equations (polynomials) to predict optical properties (transmittance, optical depth, background radiance, etc.) of the atmosphere. Each polynomial is designed to empirically fit measured data for a given wave-band. FASCODE on the other hand uses a high resolution molecular absorption database (HITRAN) and theoretical line-broadening models to predict the same optical parameters, but with a much higher spectral resolution. MODTRAN, is similar to LOWTRAN, however it employs advanced models and as a result has ten times the resolution. Table 1 summarizes the characteristics of these codes.

Table 1. AFGL Software

Algorithm	Calculation Method	Resolution
LOWTRAN	Empirical Fit Band Models	20.0 (1/cm)
MODTRAN	Empirical Fit Band Models	2.0 (1/cm)
FASCODE	Physical Models and HITRAN Database	Variable (very high) Limited by database

To get a better feel for MODTRAN's resolution, we can use the following differential formula,

$$\delta\lambda = \frac{-\lambda^2}{10{,}000}\delta\nu \tag{2}$$

where $\delta\lambda$ is the wavelength resolution measured in microns and $\delta\nu$ is the frequency (wave-number) resolution measured in inverse-cm. Therefore, at $\lambda = 1.0$ micron, a 2 (1/cm) frequency resolution corresponds to a wavelength resolution of two Angstroms, which is broad-band compared to single frequency laser linewidths. Therefore, both LOWTRAN and MODTRAN, are useful for broad-band radiometry and imaging applications but not for narrow laser line calculations. HITRAN, on the other hand, is well suited for laser line transmission calculations.

Recently the HITRAN database has been distributed on a PC compatible CDROM along with a PC executable version of FASCODE, LINEFILE and PLOTFAS. LINEFILE is an algorithm which extracts pertinent data from the large, 70 Mbytes, HITRAN database. FASCODE operates on the output of LINEFILE, and generates a binary output file, which contains the path transmittance data. PLOTFAS plots the FASCODE results, in addition PLOTFAS can be instructed to generate an ASCII output file, which is a convenient data format for subsequent processing.

In these algorithms, one can define an ensemble of atmospheric conditions such as temperature, visibility, rain, clouds, wind speeds, coastal influence, etc. In addition, several standard model atmospheres are predefined in these algorithms. The major difficulty in exercising these algorithms is the difficulty with which one experiences when defining these input parameters. These algorithms were initially designed to run on large mainframes and used primitive techniques for data I/O which simplified the programmer's task, at the expense of user friendliness. These algorithms expect the input data to be in the form of formatted FORTRAN input records. Each parameter must be placed in precisely the correct position (row and column) with no comments. This is further complicated by the fact that the number of records required, depends upon which options are selected.

Dr. Gatt has simplified the use of the AFGL software by designing two programs, PCPLTFAS.EXE and RUNFAS.EXE. PCPLTFAS.EXE is an alternative plotting program to PLOTFAS. This program reads in PLOTFAS output ASCII data and creates a VGA plot on your screen, which can be manipulated with cursor control. The second program, RUNFAS.EXE, provides a user-friendly interface shell for LINEFILE, FASCODE, PLOTFAS and PCPLTFAS. Copies of Dr. Gatt's software, RUNFAS.EXE and PCPLTFAS.EXE, is available to those interested, by sending request to Dr. Gatt at his email address (gatt@laser.creol.ucf.edu)

ASTP eye-safe laser radar analysis

The Advanced Sensor Technology Program (ASTP), which has recently been solicited by the Ballistic Missile Defense Organization (BMDO), is a multi-sensor program. The sensors include passive RF, microwave radar, passive IR imaging, and both coherent and incoherent ladar, with ladar emphasis on solid-state eye-safe transceivers. At the time of this report, system specifications had not yet been hardened. The Air Force anticipates being tasked with coherent eye-safe solid state ladar systems. A brief survey of the preliminary ladar sensor performance goals, indicates technology advancements are required in many areas.

Coherence length related issues

Laser coherence length is an issue worthy of consideration for a coherent ASTP active sensor. The minimum coherence length requirement, which minimizes the sensors frequency resolution, is that the laser be coherent over the length of the transmitted pulse (1 to 100 μsec). A more demanding requirement for a coherent ASTP transceiver, is the laser's frequency stability measured over the round-trip time. The maximum frequency drift should be 10 times less than the frequency resolution of the sensor (ideally limited by the transform width of the pulse). For example, at 500 km (round-trip-time = 3.3 ms) and a 100 μsec pulse (transform width = 10 kHz), the allowed frequency drift is less than 30 kHz per second. Frequency stability, of this order, can be obtained by a variety or techniques, which range from careful cavity design to electro-optic feedback systems, of which the Pound-Drever technique is the most elaborate and yields the highest laser frequency stability.

Range resolution

The range resolution of a conventional pulsed ladar is linearly related to the pulse width,

$$\Delta R = \frac{c\tau}{2}, \tag{3}$$

where c is the speed of light and τ is the pulse width. For example at $\Delta R = 50$ cm the maximum pulse width is 3 ns. As will be discussed next, short pulse ladars have poor velocity (frequency) resolution.

Velocity resolution

From a simple Doppler analysis, the velocity resolution of a conventional transform limited pulsed is linearly related to the pulse width.

$$\Delta V = \frac{k\lambda}{2\tau}, \tag{4}$$

where k is a pulse shape constant near unity. For example, at $\Delta V = 50$ cm/sec and $\lambda = 2$ micron, the minimum pulse width is 2 μs, nearly 1000 times the pulse width which satisfies a 50 cm range resolution.

Velocity resolution range resolution product (Need for pulse compression methods)

The velocity resolution range resolution product is, for a conventional Doppler ladar, linearly related to the wavelength

$$\Delta V\, \Delta R = \frac{k\lambda}{2\tau}\frac{c\tau}{2} = k\frac{\lambda c}{4}. \tag{5}$$

Thus for a given wavelength, there is a tradeoff between frequency resolution and range resolution. There are several unconventional methods designed to beat this resolution product limit. For example range resolution can be improved, without affecting frequency resolution, by using intra-pulse frequency modulation (a.k.a. frequency chirp). On the other hand, frequency resolution can be improved by employing coherent multi-pulse signal processing techniques.

Coherent multi-pulse signal processing techniques, which have been extensively deployed in high resolution microwave radar systems, are equally applicable in pulsed coherent laser radar systems. For these systems, the laser must be phase coherent over the pulse train and frequency stable over the time equal to the round-trip time plus the width of the pulse train.

Frequency chirp methods use a long pulse format to achieve good velocity resolution (10 cm/sec for a 10 μs, 2 micron laser pulse) and pulse compression techniques to achieve good range resolution. Resolution gains in excess of 10,000 have been obtained in high performance chirped microwave radar systems. The coherence length of these systems must be longer than the pulse width.

Transverse Resolution (need for Range-Doppler signal processing)

The transverse resolution specification, for the airborne sensor, was quoted to be 50 cm at 500 km. This resolution corresponds to an angular resolution of 1 μrad. A simple diffraction analysis will show that this

specification requires receiver diameters on the order of $\lambda/\mu m$. For example, a 2 micron receiver would need an aperture diameter on the order of 2 meters, to obtain 1 μrad resolution. Such large diameter receivers may prove to be prohibitive for an airborne platform. An alternative solution to this problem is to employ Range Doppler (RD) signal processing techniques. Unlike a conventional ladar, a RD ladar's resolution is, to a first order, independent of receiver diameter (much like the RF SAR technology). A RD ladar's transverse resolution is defined by the targets spin rate, target diameter, target orientation, wavelength and frequency resolution (or pulse width).

FASCOD3p ASTP simulations

A variety of FASCODE simulations of interest to the LDERF staff were conducted for the two ASTP scenarios at a variety of wavebands. Of particular interest was a comparison of the atmospheric transmission of eye-safe solid state lasers, non-eye-safe solid state lasers, and CO_2 gas lasers. The particular scenarios simulated are defined in Table 2. These scenarios correspond to two ASTP scenarios presented at the March 1994 ASTP Tri-Service Workshop. These two scenarios assume the airborne platform altitude is 12.8 km, a cloud height of 11.3 km and target ranges of 600 km for the ASTP1 scenario and 1000 km for the ASTP2 scenario. The additional data in Table 2, (i.e., Tangent Height and zenith angle) were calculated by FASCODE. In addition to the two ASTP scenarios a 10 km horizontal path scenario was also simulated. Results of these three simulations are provided in the Tables 3 and 4.

Selected FASCODE ASTP1-scenario transmission simulations are provided in Figures 2 through 9. In the 2 to 2.1 micron band, there is significant high resolution absorption structure in addition to a low resolution transmission envelope. The low resolution transmission envelope favors transmission closer to 2.1 microns than 2.0 microns. The high resolution structure, however precludes the use of many of the longer wavelength laser materials. Of all the 2.0 to 2.1 micron materials Er,Tm,Ho:YVO_4 (2.0412 μm), Ho:YAG (2.0975 μm), Ho:LuAG (2.1020 μm), and Er,Tm,Ho:LuAG (2.1020 μm) exhibit high (approximately 90%) ASTP scenario transmission, while Tm:YAG (2.0132 μm), Tm:LuAG (2.0240 μm) and, Ho:YLF (2.0672 μm) exhibit the worst transmittance (4% to 19%).

As can be seen from Figures 6 through 9, common isotope CO_2 lasers suffer from extreme ASTP scenario absorption (transmission less than 1%). This is due to long path integrated CO_2 molecular absorption. For example the most common CO_2 gain medium is the P20 line of the $C^{12}O^{16}{}_2$. isotope, which at standard temp and pressure lases at 10.591 μm. For this common isotope the ASTP1 transmission is only about 1/2%. However, there are a variety of techniques that can be employed to shift a CO2 laser's frequency to increase the ASTP scenario transmission to near 99%. These include pressure and temperature broadening as well as the use of rare CO_2 isotopes. An analysis of these methods should be conducted to determine their viability and relative complexity.

Table 2. Simulation Parameters:
Midlatitude Summer Model; Rural Extinction; Visibility = 23 km, no Clouds/Haze, no Rain

Scenario	Input Parameters		Output Parameters
ASTP1	Slant path; H1 = 12.8 km;	Range = 600 km H2 = 27 km	$\theta_z = 91.192°$; Tangent Z = 11.3 km
ASTP2	Slant path; H1 = 12.8 km;	Range = 1000 km, $\theta_z = 91.192°$	H2 = 70.1 km; Tangent Z = 11.3 km
Horiz	Horiz. path; H1 = 10 m;	Range = 10 km	H2 = 10 m

Table 3. Non-eye-safe Wavelength Solid State Lasers Properties

Common Name	Composition	λ (μm)	ASTP1 Trans (%)	ASTP2 Trans (%)	Horiz Trans (%)
Nd:YAG	$Y_3Al_5O_{12}:Nd^{3+}$(1.3%)	1.06415	83.2	82.8	50.6
Nd:YLF	$LiYF_4:Nd^{3+}$(<1%)	1.0471(π)	82.3	82.0	49.6
Nd:YLF	$LiYF_4:Nd^{3+}$(<1%)	1.0530(σ)	82.6	82.3	50.0
Nd: YVO_4	$YVO_4:Nd^{3+}$(1%)	1.0641(π)	83.2	82.8	50.6
Nd:LuAG	$Lu_3Al_5O_{12}:Nd^{3+}$(0.6%)	1.06425	83.2	82.8	50.6
Nd:YAG	Doubled	0.532075	11.1	9.8	16.8

Table 4. Eye-safe Wavelength Solid State Lasers Properties

Common Name	Composition	λ (μm)	ASTP1 Trans (%)	ASTP2 Trans (%)	Horiz Trans (%)
Ho:YAG	$Y_3Al_5O_{12}:Ho^{3+}$(2%)	2.0975	96.2	96.1	59.3
Ho,Er:YLF	$LiYF_4:Er^{3+}:Ho^{3+}$(2%)	2.066	24.7	24.7	20.8
Ho:YLF	$LiYF_4:Ho^{3+}$(2%)	2.0672	4.4	4.3	19.5
Tm,Ho:YLF	$Li(Y,Er)F_4:Tm^{3+}:Ho^{3+}$(1.7%)	2.0654	53.1	53.0	52.4
Er,Tm,Ho: YVO_4	$YVO_4:Er^{3+}$, $Tm^{3+}:Ho^{3+}$ (1%)	2.0412	88.9	88.6	58.0
Ho:LuAG	$Lu_3Al_5O_{12}:Ho^{3+}$(2%)	2.1020	97.5	96.3	41.5
Ho:LuAG	$Lu_3Al_5O_{12}:Ho^{3+}$ (5%)	2.9460	87.2	87.1	E-6
Er,Tm,Ho:LuAG	$Lu_3Al_5O_{12}:Er^{3+}$, $Tm^{3+}:Ho^{3+}$(2%)	2.1020	97.5	96.3	41.5
Tm:YAG	$Y_3Al_5O_{12}:Tm^{3+}$(?%)	2.0132	9.0	8.9	12.9
Tm:YVO_4	$YVO_4:Tm^{3+}$(?%)	2.0?	46.4	45.7	26.0
Tm:LuAG	$Lu_3Al_5O_{12}:Tm^{3+}$(2%)	2.0240	19.4	19.3	16.3
Er:YAG	$Y_3Al_5O_{12}:Er^{3+}$ (1%)	1.6602	98.4	95.6	65.1
Er:YAG	$Y_3Al_5O_{12}:Er^{3+}$ (0.4%)	1.6449	98.1	95.2	64.9
Er:YAG	$Y_3Al_5O_{12}:Er^{3+}$ (1.2%)	1.7757	97.9	96.0	2.1
Er:LuAG	$Lu_3Al_5O_{12}:Er^{3+}$(2-5%)	1.6525	88.7	86.2	95.5
Er:LuAG	$Lu_3Al_5O_{12}:Er^{3+}$(2-5%)	1.6630	98.2	95.5	64.9
Er:LuAG	$Lu_3Al_5O_{12}:Er^{3+}$(1.5%)	1.7762	71.5	70.0	E-19
Er:LuAG	$Lu_3Al_5O_{12}:Tm^{3+}:Er^{3+}$(9%)	2.6990	0	0	0
Er:LuAG	$Lu_3Al_5O_{12}:Tm^{3+}$, $Ho^{3+}:Er^{3+}$(10%)	2.6990	0	0	0

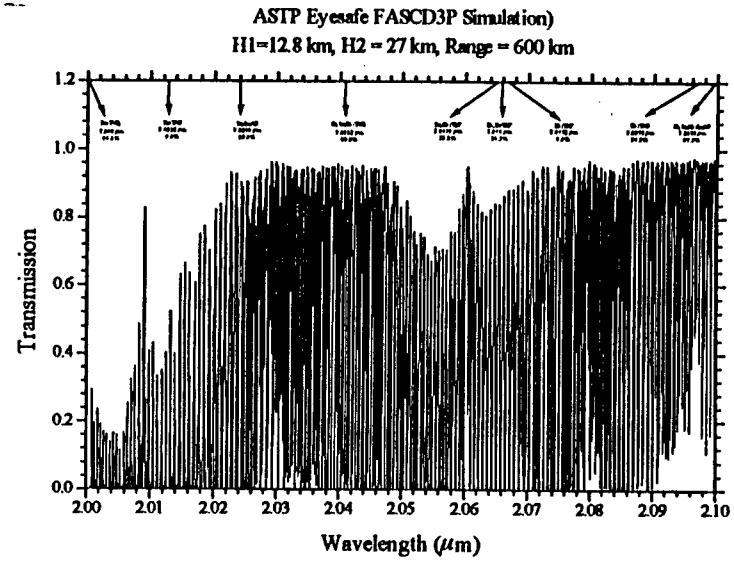


Figure 2. ASTP1 Scenario; 2 to 2.1 microns

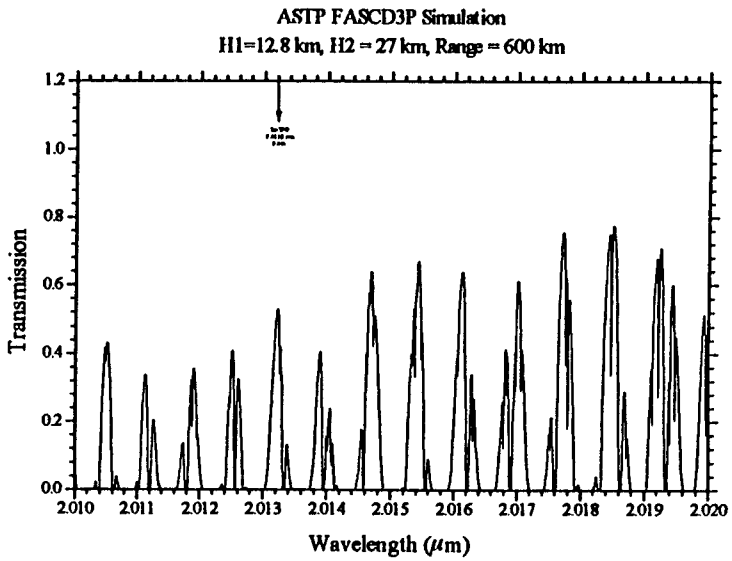


Figure 3. ASTP1 Scenario; 2.01 to 2.02 microns

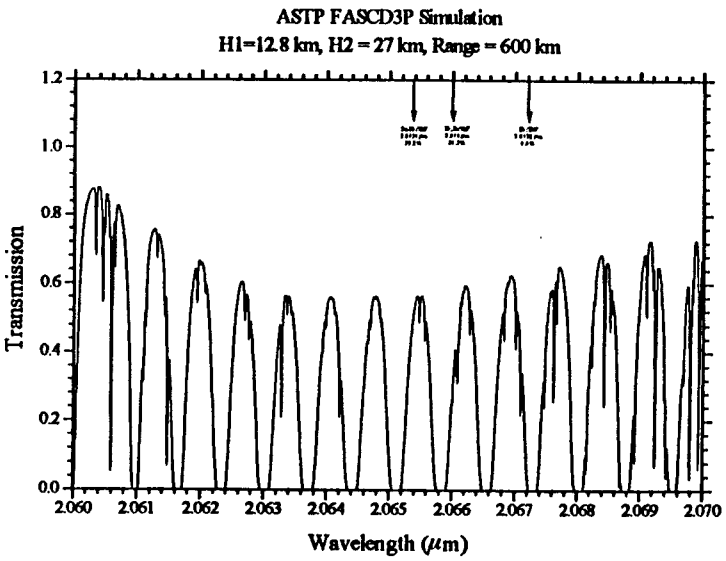


Figure 4. ASTP1 Scenario; 2.06 to 2.07 microns

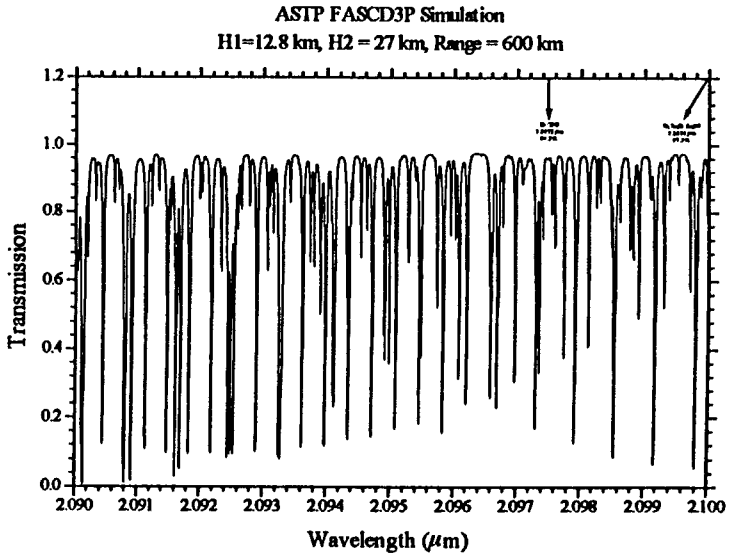


Figure 5. ASTP1 Scenario; 2.09 to 2.10 microns

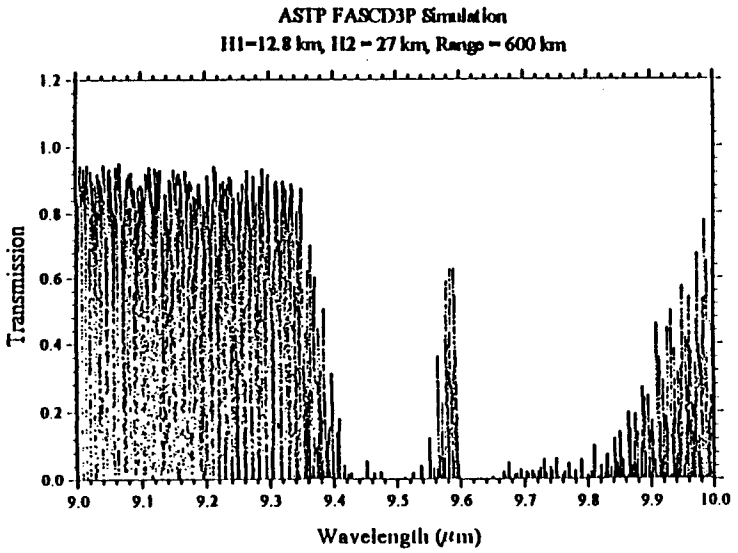


Figure 7. 9 to 10 microns

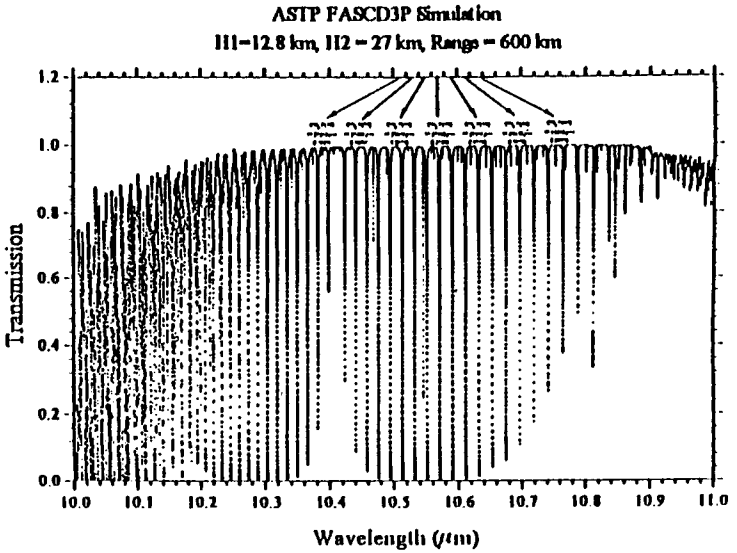


Figure 8. 10 to 11 microns

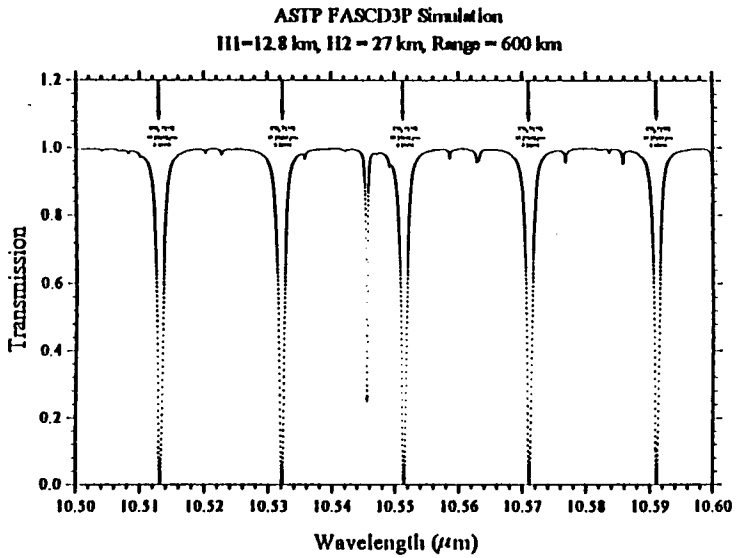


Figure 9. 10.5 to 10.6 microns

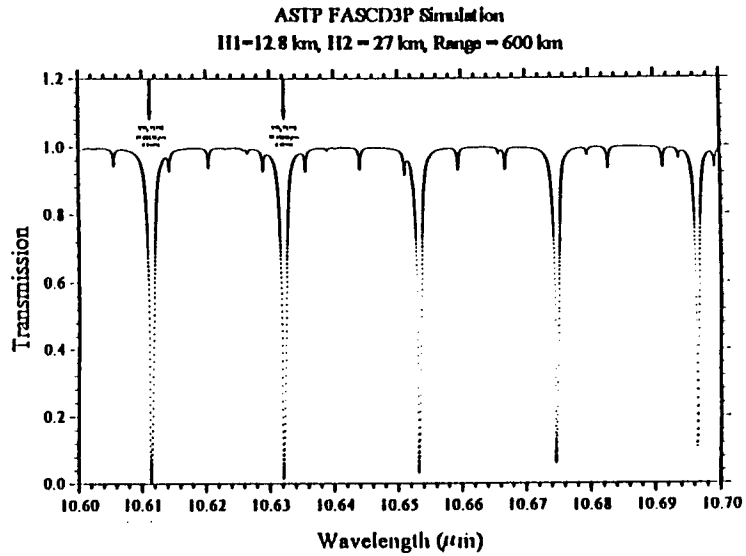


Figure 10. 10.6 to 10.7 microns

Tactical direct detection imaging ladar analysis

The LDERF staff were interested in developing an in-house modeling capability for laser radar range prediction analysis with emphasis on short-pulse high PRF angle-angle-range ladar imagers. To this end, a model for a direct detection laser radar was developed based upon the equations given in this section and the following assumptions: the optical path is horizontal, the target is resolved and Lambertian, the filter is matched to the transmitted signal pulse, the transmitted pulse exhibits a Gaussian temporal profile, and noise sources after the preamplifier are negligible compared to the preamplifier noise.

Maximum range math model

The purpose of the this analysis was to generate plots of the maximum range vs. pulse energy for a direct detection imaging ladar. The maximum range is the range, for which the SNR is equal to some predefined minimum value. The math model used to predict this maximum range is described in this section. Some sample results from this analysis are provided in Figures 10 through 17.

The peak transmitter power, for a Gaussian temporal pulse, can be shown to be given by

$$P_t = \frac{E}{\tau}\sqrt{\frac{4\ln(2)}{\pi}}, \tag{6}$$

where E is the pulse energy and τ is the full-width-half-maximum (FWHM) pulse width. The peak received power follows from the ladar equation for a resolved target, which is given by

$$P_s = \varepsilon\rho P_t T^2 \frac{\pi D_r^2}{4\Omega R^2} \tag{7}$$

where ε is the transceiver optical efficiency, ρ is the target reflectance, D_r is the receiver aperture diameter, T is the one way path transmittance, R is the target range, and Ω is the effective top-hat target reflecting solid angle. The one way path transmittance follows from beers law

$$T = e^{-\alpha R}, \tag{8}$$

where α is the extinction coefficient, which for these simulations was predicted from FASCODE. The power signal-to-noise ratio (SNR) is given by

$$SNR = \frac{(\Re P_s M/\sqrt{2})^2}{2qB_{eff}[(\Re P_s + \Re P_{bg} + i_{db})M^2 F + i_{ds} + i_a]}, \tag{9}$$

where, q is the charge of an electron, B_{eff} is the white noise equivalent bandwidth, $\Re$ is the detector responsivity, P_{bg} is the background light power, i_{db} is the APD bulk dark current, M is the APD gain, F is the APD excess noise factor, i_{ds} is the detector dark surface current, and i_a is the preamplifier noise current density (A/$\sqrt{}$Hz). The signal term factor, 1/$\sqrt{2}$, in the numerator is included to reflect the amplitude reduction through a matched Gaussian

filter. The excess noise factor for a uniformly multiplying APD, which was first derived by McIntyre, is given in terms of the ionization ratio k as

$$F = kM + (1-k)(2+1/M). \tag{10}$$

The detector responsivity is related to the detector quantum efficiency η and the optical frequency ν as follows

$$R = \frac{\eta q}{h\nu}, \tag{11}$$

where h is Plank's constant.

Under signal photon limited conditions and assuming a Gaussian pulse with a matched filter receiver the SNR description given in Eq. (9) reduces to

$$SNR = \frac{RP_s}{4qB_{eff}F}, \tag{12}$$

The noise equivalent bandwidth of a Gaussian filter can be shown to be given by

$$B_{eff} = \frac{1}{\tau}\sqrt{\frac{\ln(2)}{2\pi}}. \tag{13}$$

Substitution of Eqs. (6) and (13) into Eq. (12) results in

$$SNR = \frac{\eta E_s}{\sqrt{2}h\nu F}, \tag{14}$$

where E_s is the received optical pulse energy. The optimum SNR, (i.e., that SNR which would be obtained if a perfectly matched filter were employed; the noise power is dominated by signal shot noise limited; and the APD gain were noiseless) is then easily shown to be given by

$$SNR_{opt} = \frac{\eta E_s}{\sqrt{2}h\nu}. \tag{15}$$

The equations provided in this section, were programmed in Mathematica®. Mathematica was then instructed to solve these equations numerically for (1) the SNR as a function of range and (2) the maximum range which satisfies a given SNR as a function of pulse energy. Results of this Mathematica model are provided in the following section.

Performance comparison between Nd:YAG and Ho:YAG ladars

The following analysis compares the range performance of a direct detection Nd:YAG imaging ladar to that of a Ho:YAG imaging ladar with all parameters begin equal except the wavelength (i.e., photon energy) and the atmospheric transmittance. This analysis assumes the identical detector characteristics, which therefore biases the Ho:YAG results since 2.09 μm InGaAs pin's or Ge APD's typically have higher excess noise factors and dark currents than 1.064 μm InGaAs pin's or YAG enhanced Si detectors APD's.

Effect of haze on atmospheric transmittance

Results of FASCODE simulations which compare the atmospheric transmittance vs. haze condition, quantified by visibility, are listed in Table 5. for the standard Midlatitude Summer Model Atmosphere. As can be seen from the table, a 2.09 μm laser penetrates haze more favorably than a 1.064 μm laser.

Table 5. FASCODE predicted atmospheric transmittance vs. haze condition for a Midlatitude Summer Model 1 km Horizontal path at 10 m elevation

	1 km Horizontal Path Transmittance					
Condition	Standard Clear	Clear	Light Haze	Moderate Haze	Heavy Haze	Extreme Haze
Visibility km	23.5	15	8	5	3	1
Nd:YAG (1.06415μm)	0.934	0.899	0.815	0.720	0.576	0.191
Ho:YAG (2.0975μm)	0.949	0.937	0.905	0.865	0.799	0.539

Effect of rain rate on atmospheric transmittance

Results of FASCODE simulations which compare the atmospheric transmittance vs. rain rate are listed in Table 6. for the standard Midlatitude Summer Model Atmosphere, which has a 23.5 km visibility. As can be seen from the table both wavelengths exhibit near equal transmittance through rain. Therefore both laser wavelengths are equally affected by rain.

Table 6. FASCODE predicted atmospheric transmittance vs. rain rate for a Midlatitude Summer Model 1 km Horizontal path at 10 m elevation

	1 km Horiz Path Transmittance				
Rain Rate mm/Hr	0	1	5	10	25
Nd:YAG	0.934	0.641	0.338	0.195	0.058
Ho:YAG	0.949	0.652	0.343	0.198	0.059

Range vs. pulse energy performance simulations

The model equations described above (i.e., Eqs. (6) through (15)) were implemented in a Mathematica® program. This program solves these transcendental equations for the range which meets a prescribed SNR. Of particular interest to the LDERF staff was a comparison between the performance of a Ho:YAG ladar and a similarly configured Nd:YAG ladar for both APD and pin detector based receivers.

The input parameters for the following simulation results are listed in Tables 5 through 8. These parameters were selected to be a close match to the LDERF solid state imaging ladar system. To normalize the comparison the Ho:YAG detectors was assumed to have identical noise, bandwidth, and quantum efficiency as the Nd:YAG detectors. Although this may not be a realistic assumption, it allows one observe the effects of the differing atmospheric transmittance for the two wavelengths.

Table 7. Transceiver parameters for the simulation shown in Figures 10 through 17

Symbol	τ	D_t	D_r	ε	ρ	i_a
Units	nsec	cm	cm	-	-	pA/√Hz
Value	10	0.5	5	0.7	0.2	1.8

Table 8. Detector parameters for the simulation shown in Figures 10 through 17

	η	M	k	i_{ds} (nA)	i_{db}(pA)
APD	0.35	1	.02	5	50
pin	0.70	120	1	1	0

The results of this analysis are presented in Figures 10 through 17. The curves within each plot correspond to the different atmospheric conditions listed in Table 9 .

Table 9. Atmospheric conditions associated with the curves in Figures 10 through 17 for a Midlatitude Summer Model 1 km Horizontal path at 10 m elevation

Condition	Standard Clear	Clear	Light Haze	Moderate Haze	Heavy Haze	Extreme Haze	25 mm/hr Rain Rate
Nd:YAG	0.934	0.899	0.815	0.720	0.576	0.191	0.058
Ho:YAG	0.949	0.937	0.905	0.865	0.799	0.539	0.059

The data show in these plots are consistent with expectations. For example a pin detector will outperform an APD detector at close ranges (high signal levels). This occurs, since the pin detector typically has a higher quantum efficiency (2 times) and no excess multiplication noise (predicted to be near 4 for a typical APD detector). For the simulation parameters used in this analyze the pin detectors SNR will be on the order of 8 times (9 dB) that of an APD at close range. This becomes apparent when one looks at Eq. (9) in the limit as the signal shot noise dominates the other noise sources. Under this condition, the SNR becomes

$$SNR_{APD} = \frac{\eta_{APD} P_s}{2hvBF} \quad \text{and} \quad SNR_{pin} = \frac{\eta_{pin} P_s}{2hvB}., \tag{16}$$

In addition, since these equations predict that at close ranges the SNR is proportional to the signal power, which is itself proportional to the reciprocal of the square range, one would expect the SNR curves to exhibit a -20 dB per decade slope in the low range regime. This slope is apparent in both APD curves of Figures 11 and 15. The same phenomena occurs in the pin SNR vs. range curves. However, for the parameters used in this simulation, this low range SNR slope occurs at a much lower range regime.

Summary

Dr. Gatt's AFOSR Summer Faculty Research Program was mutually successful for all parties involved. The topic chosen was an interest area to both Dr Gatt and the LDERF staff. The short course presented to the LDERF staff, was highly successful and will be used as the basis for an SPIE short course. The laboratory based

coherent laser radar testbed, designed during this summer research program is a proven technology, that will serve as a valuable educational tool for those interested in gaining a deeper understanding of the more subtle technical issues associated with coherent detection. In addition, this testbed provides a instrument for comparing direct detection and coherent detection ladar systems. The FASCODE interface program, RUNFAS.EXE, was highly successful. It significantly simplifies the user interface to FASCODE, and was used extensively in this summer research program.

The ASTP analysis highlighted some of the more difficult technical issues associated with the system goals as well as potential solutions to these technical problems. In addition, results of the ASTP scenario FASCODE simulations indicate that Er,Tm,Ho:YVO_4 (2.0412 μm), Ho:YAG (2.0975 μm), Ho:LuAG (2.1020 μm), and Er,Tm,Ho:LuAG (2.1020 μm) propagate much better than other eye-safe solid state wavelengths, such as Tm:YAG (2.0132 μm), Tm:LuAG (2.0240 μm) and, Ho:YLF (2.0672 μm). Common isotope CO_2 laser exhibited extremely poor ASTP scenario transmission, however rare isotope CO_2 lasers had exceptionally high (99.5%) ASTP scenario transmission. The comparison widens when one considers the round-trip transmission which is the square of the one-way transmission data presented herein.

Results from the direct detection imaging ladar performance analysis, indicate that in general a Ho:YAG ladar will outperform an equivalent Nd:YAG laser (where equivalent assumes equal laser power, and equal detector specifications). In addition, it was shown that for pulse energies on the order of 25 μJ, along with the other simulation parameters, that a p-i-n detector's maximum range can be as high as 2 km.

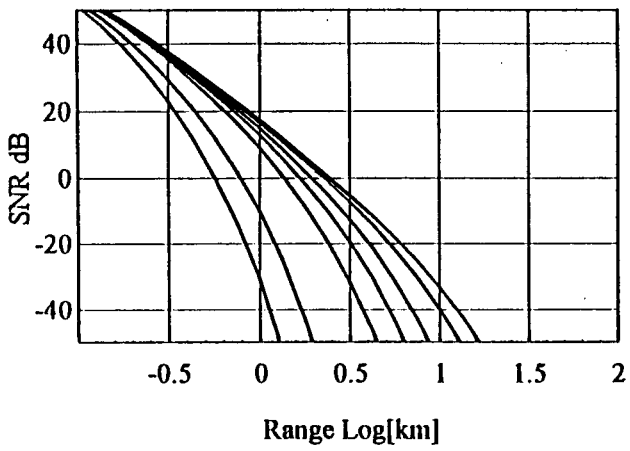


Figure 10. Nd:YAG pin SNR

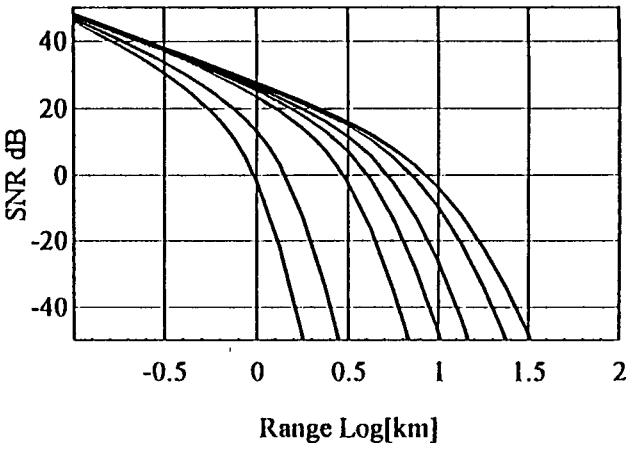


Figure 11. Nd:YAG APD SNR

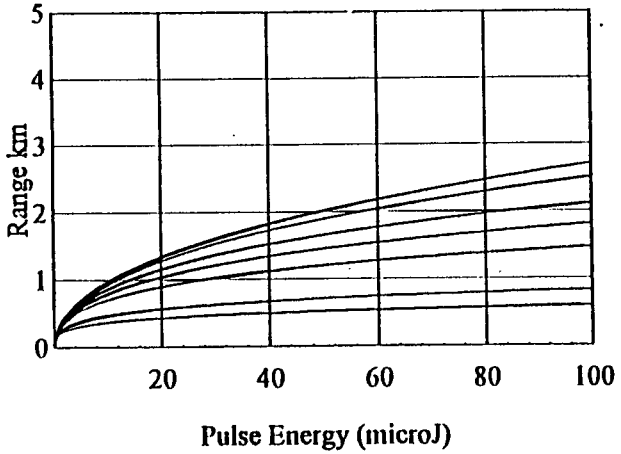


Figure 12. Nd:YAG pin Range

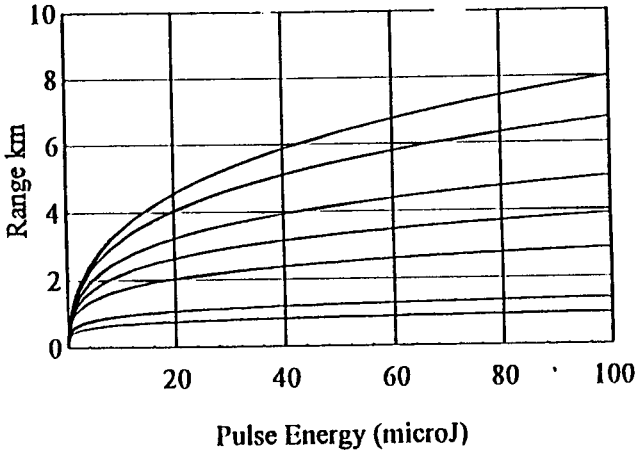


Figure 13. Nd:YAG APD Range

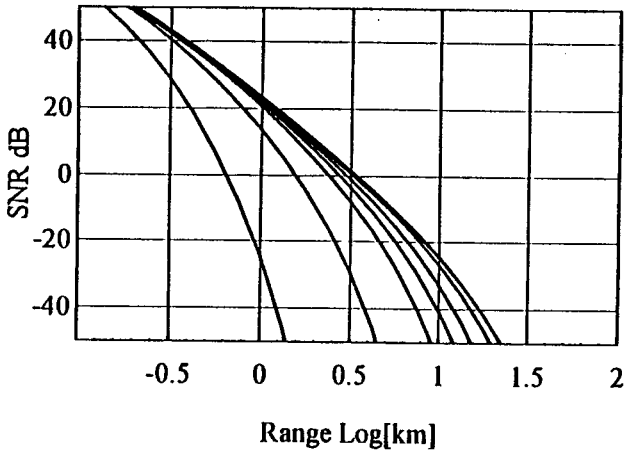


Figure 14. Ho:YAG pin SNR

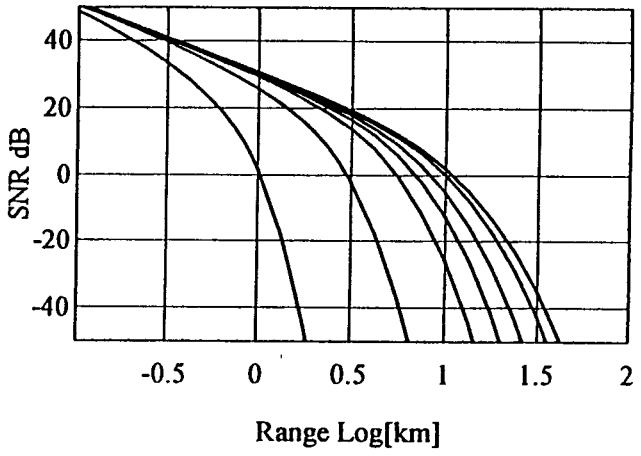


Figure 15. Ho:YAG APD SNR

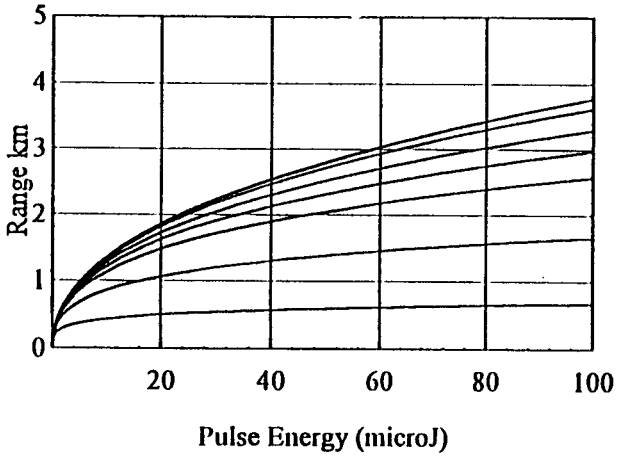


Figure 16. Ho:YAG pin Range

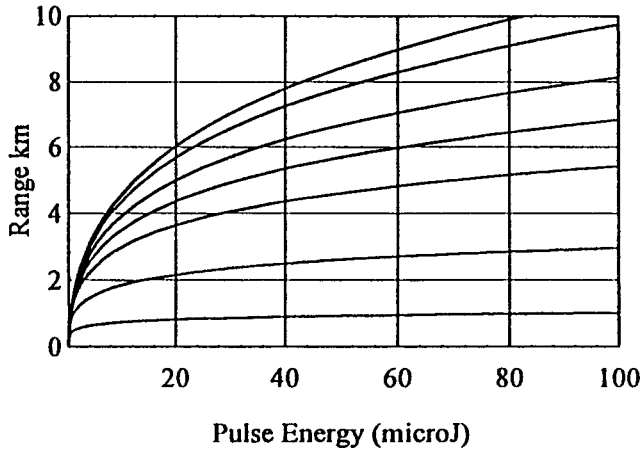


Figure 17. Ho:YAG APD Range

ANALYSIS OF LASER DOPPLER VELOCIMETRY DATA

Dr. Richard D. Gould

Mechanical and Aerospace Engineering
North Carolina State University
Raleigh, NC 27695

Final Report for:
Summer Research Program
Wright Laboratory

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

September 1994

ANALYSIS OF LASER DOPPLER VELOCIMETRY DATA

Dr. Richard D. Gould
Mechanical and Aerospace Engineering
North Carolina State University
Raleigh, NC 27695

ABSTRACT

The work accomplished at Wright Laboratory this past summer can be subdivided into three major tasks. The first task was the development of a computer based data smoothing algorithm so that measured profiles of turbulence statistics (i.e. mean, Reynolds stresses, and turbulent triple products) can be smoothed automatically. Specifically, laser Doppler velocimeter (LDV) measurements were smoothed in this work, however, the algorithm is general and thus could be applied to smooth any experimentally obtained profiles. The second task involved the determination of the statistical uncertainty of the measured mean velocities and turbulent normal stresses. Lastly, a series of computer programs which smooth, numerically differentiate and plot LDV measurements made in the flow field surrounding an integrated fuel injector (IFI) were developed to help aide in the analysis of this data. In addition, a Wright Laboratory senior engineer was trained well enough to use and modify this software as part of this effort.

ANALYSIS OF LASER DOPPLER VELOCIMETRY DATA

Dr. Richard D. Gould

INTRODUCTION

The laser Doppler Velocimeter (LDV) has become a common and useful tool in the study of turbulent flow phenomena. The main reasons for this include the fact that the instrument is non-intrusive and has relatively high spatial and temporal resolution. The study of turbulent transport phenomena and the validation of turbulence modeling requires that measurements of all turbulence quantities be made at a large number of locations in the flow field under study. A number of factors contribute to make this task less than ideal. These include finite data sample size at each measurement location, a compromise between available flow facility run time and the number of measurement locations that can be probed and, noise entering the measurement system. The first item limits the accuracy of the turbulent statistic at each measurement location. Statistical theory states that the mean estimator is the actual mean only if an infinite number of samples are obtained at each measurement point, which of course, is not possible. The second item can sometimes be addressed by curve fitting the experimental data points. If measurement locations are closely spaced enough to resolve the gradients in the flow field curve fitting allows for interpolation between the measurement points. Since no measurement system is perfect, noise from spurious reflections, scratches on test section windows, signal processor electronics, photomultiplier tubes, poor seeding and other undetermined sources may cause erroneous measurements. One simple way to identify if this has occurred is by plotting the mean statistic of interest giving a profile of the measurement variable. This plotted data, in absence of discrete flow phenomena like shocks, should appear smooth and continuous. Two methods for smoothing the experimental data were investigated during this study. The first method defined the smoothed data to be a least squares polynomial curve fit to the original experimental data while the second method smoothed the experimental data by convoluting each data point and a user selected number of neighboring points with a convolution kernel. The least squares polynomial curve fitting method worked well so long as the profile did not have sharp corners or inflections whereas the second smoothing method was adopted for profiles which had sharp corners or inflections. A major objective of this study was to automate this process and to create a smoothed data set which could then be used to analysis turbulence transport and turbulence modeling assumptions. A description of the smoothing algorithm is given below. The least squares polynomial fitting method is not described here since it is well known and is discussed in many numerical methods textbooks(see Gerald (1978) for example).

SAMPLE STATISTICS AND STATISTICAL UNCERTAINTY

The sample mean and sample variance (Dougherty, (1990), p. 327) of a random variable X of sample size N are given by

$$\overline{X} = \frac{1}{N-1}\sum_{i=1}^{N} X_i \tag{1}$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2. \tag{2}$$

The sample standard deviation is defined as $S = \sqrt{S^2}$. In the limit as N goes to infinity these sample estimates given by equations (1) and (2) approach the true mean, μ, and the true variance, σ^2, or true standard deviation, σ, respectively. Since sample sizes are always finite one is interested in knowing how well the sample statistics given by equations (1) and (2) above agree with the true sample statistics. Confidence intervals - sometimes called statistical uncertainty - for the mean and variance of normally distributed random variables can be used to address this question. If X is a normally distributed random variable with unknown mean μ and known variance σ^2, and if $\overline{X}$ is an empirical mean given by equation (1) resulting from a random sample of size N, then

$$\left[\overline{X} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{N}},\ \ \overline{X} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{N}}\right] \tag{3}$$

is a $(1-\alpha)\cdot 100\%$ confidence interval for μ where z is the standard normal random variable evaluated where the area of the tail of the normal distribution curve accounts for $\alpha/2$ % of the cumulative area under the normal distribution curve (Dougherty, (1990), pp. 349-352). For example, standard normal distribution tables give $z(\alpha/2 = 0.025) = 1.96$ and $z(\alpha/2 = 0.005) = 2.326$. The first value would be used if one is interested in determining the statistical uncertainty in the mean for a *95%* confidence interval while the second valued would be used if a confidence interval of *99%* is required. It is important to recognize that equation (3) gives the uncertainty in the calculated sample mean for a prescribed confidence interval. Note that this interval decreases as the standard deviation of the sample decreases, as the sample size increases (actually as $1/\sqrt{N}$) and, as the confidence interval decreases.

The application of equation (3) to a sample of turbulence measurements obtain using LDV goes as follows. Consider a sample containing 5000 velocity measurements with a sample mean and variance of $\overline{X}$ = 10 m/s and σ^2 = 100 m^2/s^2 (or σ = 10 m/s), respectively. If the sample population is normally distributed (*i.e.* near Gaussian probability distribution) equation (3) is valid. This analysis gives that the mean of this sample lies in the range between [9.723, 10.277] with 95% confidence. Thus, the best we can say is that the mean velocity of our sample is *10 m/s* ± *0.277 m/s* with 95% confidence.

A similar procedure can be used to determine the statistical uncertainty or precision of the variance of a normal distribution as given by equation (2). In this case the random variable is X^2 which processes a chi-squared distribution with *N-1* degrees of freedom, as opposed to a normal distribution like the mean. The uncertainty in the variance (Dougherty, (1990), p. 329-331) can be determined using equation (4) below

$$\left[\frac{N}{\chi^2_{N-1,\alpha/2}}\cdot S^2\ ,\ \frac{N}{\chi^2_{N-1,1-\alpha/2}}\cdot S^2\right] \tag{4}$$

where N is the sample population, S^2 is the sample variance and $\chi^2_{N-1,\alpha/2}$ is the value the chi-square variable with N-1 degrees of freedom evaluated where the area of the tail of this distribution curve accounts for$\alpha/2$ % (or *1* - $\alpha/2$ %) of the cumulative area under the chi-square distribution curve. The values of $\chi^2_{N-1,\alpha/2}$ can be found in standard statistics books if the population size is small(typically N-1 < 100). It is important to note that the chi-square distribution, unlike the normal distribution, is non-symmetric. For example, $\chi^2_{30,0.005} = 13.79$ whereas, $\chi^2_{30,0.995} = 53.67$. This distribution does, however, approach symmetry when the sample size becomes large, as is the case for LDV samples. An approximate formula (see Bendat and Piersol (1986), see footnote of chi-square table) for the chi-square variable when the sample size is large is

$$\chi^2_{n,\alpha/2} \approx n\left[1-\frac{2}{9n}+z_{\alpha/2}\sqrt{\frac{2}{9n}}\right]^3 \tag{5}$$

where n is the number of degrees of freedom (n = *N-1*) and z is the standard normal random variable evaluated where the area of the tail of the normal distribution curve accounts for $\alpha/2$ % of the cumulative area under the normal distribution curve.

The application of equation (4) to the same example ($N = 5000$, $\overline{X} = 10$ *m/s* and $\sigma^2 = 100$ m^2/s^2) given above goes as follows. If the sample population is normally distributed (*i.e.* near Gaussian probability distribution) equation (4) is valid. Equation (5) gives $\chi^2_{4999,0.005} = 5196.9$ whereas, $\chi^2_{4999,0.995} = 4804.9$ when a 95% confidence level is used (*i.e.* $z_{0.005} = 1.96$ and $z_{0.995} = -1.96$). This analysis gives that the variance of this sample lies in the range between [96.21, 104.06] with 95% confidence. Thus, the best we can say is that the variance of our sample is $\sigma^2 = 100\ m^2/s^2\ +4.06\ -3.79\ m^2/s^2$ with 95% confidence. Note that the error becomes reasonably symmetric about the estimated value due to the large sample size.

SMOOTHING EXPERIMENTAL DATA

Smoothing and curve fitting experimental measurements is a common practice in engineering and science since one is usually interested in obtaining an empirical expression which describes the measured phenomena. It is especially important to smooth experimental measurements if one is interested in obtaining the derivative of the measured phenomena. This is because numerical differentiation is known to be a "noisy" process since small errors in the measurements can lead to large errors in the finite differenced approximation to the derivative. This can be described by considering a first order one sided difference, $dU/dy = (U(y+\Delta y) - U(y))/\Delta y$. Because the two measurements are close to one another and thus of about the same magnitude, any noise added to the measurements can be of the same order or larger than the difference in the two measurements. Dividing this difference by a small distance, Δy, can further amplify this error.

Probably the most common curve fitting algorithm used is based on the least squares minimization procedure. In this method a general curve, of prescribed type, is fitted through the experimental measurements. The coefficients of the curve are found by minimizing the square of the error between the values given by the curve fit expression and the measurement values. Various types of curves, such as n^{th} order polynomial curves, exponential curves, power law curves, and others can be specified. It is important to note that this curve may not pass through any of the measurement values. Once this curve is found it can be differentiated in closed form since these curves are analytic.

Another common curve fitting method, the cubic spline procedure, is based on fitting a cubic polynomial between each of the measurement points (see Gerald (1978)). The curves are

defined by requiring the cubic polynomial to pass through the two measurement points at the end of each interval, and to have the same slopes at the ends of the interval as the cubic splines of each adjacent interval. This method forces piecewise cubic polynomial curves through each measurement point.

A compromise between the least squares polynomial fitting method and the cubic spline method has been suggested by Savitzky and Golay (1964). Their method, which employs a convolution procedure, can be described as a moving segment least squares polynomial curve fitting algorithm. Consider a segment of experimental measurements made up of 5 points. A quadratic or cubic polynomial is fit to these 5 points using the least squares criteria. This curve fit is then used to find the best value for the central point of this segment. Thus, a new "smoothed" measurement point at the center of this segment is produced. This procedure is repeated for each segment of 5 measurement points, dropping one at the left and picking up one at the right each time. The result of this process is a new set of "smoothed" measurement points(excluding the two points on each end of the experimental measurement set if 5 point segments are used). Savitzky and Golay discovered that a convolution procedure could be used to obtain the smoothed central point value which is exactly equivalent to the least squares method. They tabulated the convolution integers for the cases where from 5 to 25 measurement points are used in each segment. This algorithm was coded as a subroutine in the FORTRAN programming language for cases where 5 to 13 measurement points are used in each segment to calculate the smoothed data point at the center of each segment. A computer listing of this algorithm, with complete comment statements, is given in the Appendix.

Three experimentally obtained profiles are presented here to demonstrate how the smoothing and curve fitting algorithms performed. In all the figures the square symbol represents the experimental measurements. Up to three curves are "fitted" to the experimental measurements shown here. The solid line is obtained by applying the 5 point Savitzky-Golay smoothing algorithm (smoothing=1 legend) to the experimental measurements. New "smoothed" data points are then calculated during the convolution procedure as mentioned above. Next a cubic-spline is used to interpolate additional points between these new smoothed points so that the solid line can be drawn. The dotted line is similar to the solid line except that 13 measurement points (smoothing=5 legend) are used in the convolution procedure. The new data points are therefore influenced by more adjacent points when 13 points are used as opposed to when 5 points are used. The dash-dot-dot pattern is a least squares 8^{th} order polynomial curve fit to the experimental data.

Figure 1 shows experimental LDV measurements of the mean axial velocity downstream (x/H =1) and in the wake of a bluff body. The profile is symmetric and has three rather sharp corners where the velocity changes rapidly with position. It can be seen that the smoothed data using 5 points in each segment(solid line) fits the experimental measurements extremely well, even around the three sharp corners. It should also be mentioned that the solid line does not pass through the center of the experimental measurement points. Recall that the smoothing procedure finds a new value for the data point at the center of each segment. The dash-dot-dot line shows the curve fit to the smoothed data when the number of points in the smoothing interval is increased to 13. It is clear that sharp corner features cannot be resolved when too many points are included in the smoothing procedure. Thus, there is a trade-off between too much smoothing and accurately following the experimental measurements. Lastly, the dotted line shows a least squares 8th order polynomial curve fit to the experimental data. This polynomial curve shows over-shoot and under-shoot on the flat parts of the experimental data and a poor fit in the central region. It turns out that all polynomial curves have difficulty fitting the type of data presented in this figure.

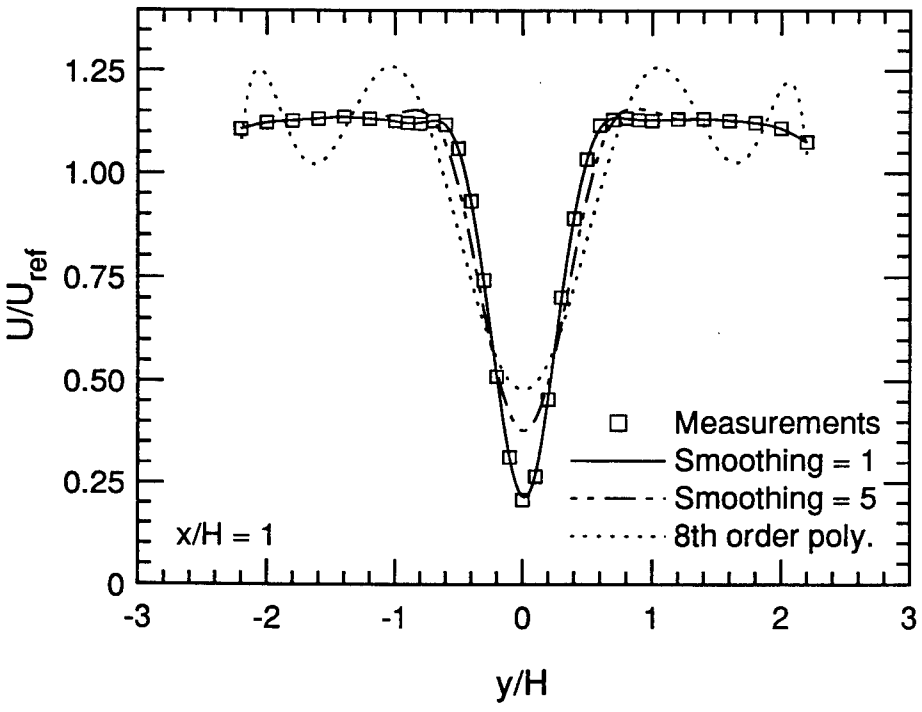


Figure 1. Smoothing and curve fitting experimental data.

Savitzky and Golay also included convolution integers to find the derivative of the smoothed data point at the center of each segment. This approach was not used here. Instead, the first derivative of the experimental data was found using second order accurate finite difference

equations. The second order accurate finite difference algorithm for the first derivative of unequally spaced central points is

$$\left.\frac{\partial u}{\partial y}\right)_{i,j} = \frac{u_{i,j+1} + (\alpha^2 - 1)u_{i,j} - \alpha^2 u_{i,j-1}}{\alpha(\alpha+1)\Delta y_-} \tag{6}$$

where $\Delta y_+ = y_{i,j+1} - y_{i,j}$, $\Delta y_- = y_{i,j} - y_{i,j-1}$ and $\alpha = \Delta y_+ / \Delta y_-$. Note that this equation can be used for equally spaced (α = 1) points also. For end points, the following second order finite difference algorithms were used to calculate the first derivatives.

$$\left.\frac{\partial u}{\partial y}\right)_{i,j} = \frac{-u_{i,j+2} + 4u_{i,j+1} - 3u_{i,j}}{2\Delta y} \qquad \left.\frac{\partial u}{\partial y}\right)_{i,j} = \frac{u_{i,j-2} - 4u_{i,j-1} + 3u_{i,j}}{2\Delta y} \tag{7}$$

Equations (6) and (7) were used to calculate the first derivative of the experimental measurements with respect of y. Figure 2 shows the first derivative of the data presented in Figure 1 and curves from the same three curve fitting methods discussed above. Note that the 5

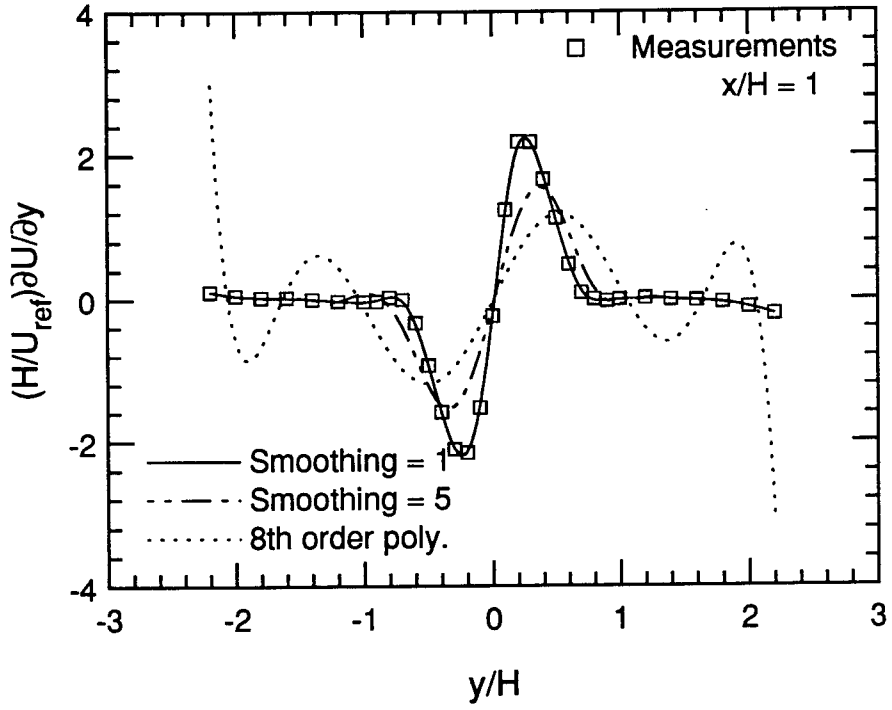


Figure 2. Smoothing and curve fitting the first derivative of experimental data.

point smoothing algorithm fits the first derivative data extremely well. The 13 point smoothing algorithm has difficulty in fitting the sharp corners while the least squares 8th order polynomial curve oscillates about the measured points.

Figures 3 and 4 show experimental measurements and the first derivative of experimental measurements, respectively, of the transverse mean velocity behind (x/H = 1) and in the wake of a bluff body. The shape of this profile is quite different from that shown in Figure 1. However, the 5 point smoothing algorithm does a good job smoothing the data while maintaining the proper shape. In particular, note how the solid line passes through the three points towards the top of Figure 3. The 13 point smoothing algorithm does a reasonable job smoothing the data but tends to over-smooth the sharp corners. The 8th order polynomial curve fit cannot follow the shape of this profile. The Figure 4 shows the first derivative of these experimental measurements. The same comments about the curve fits made for Figure 3 apply to Figure 4 also.

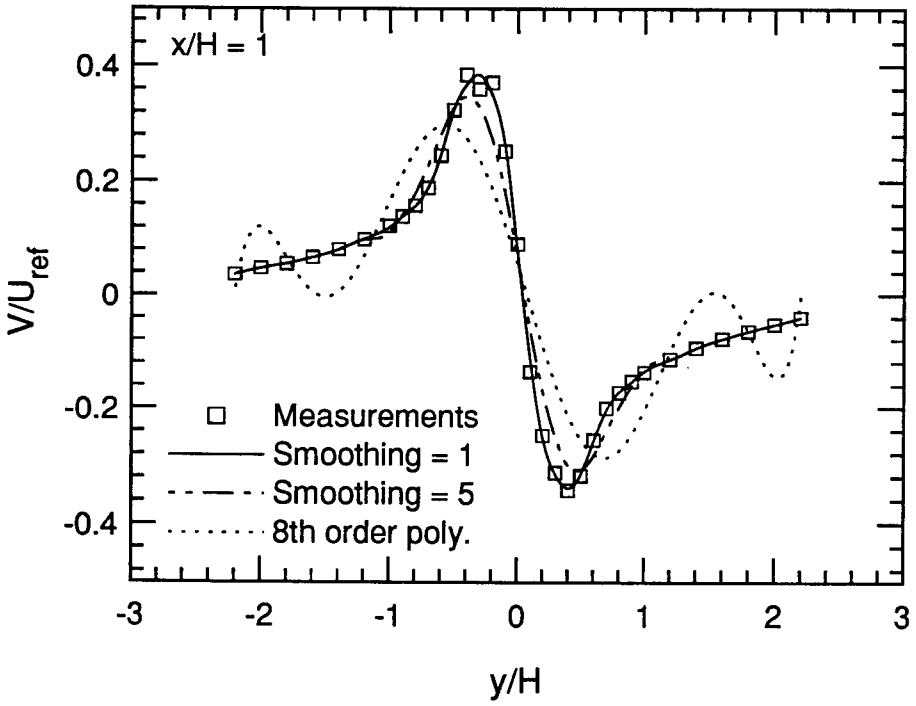


Figure 3. Smoothing and curve fitting experimental data.

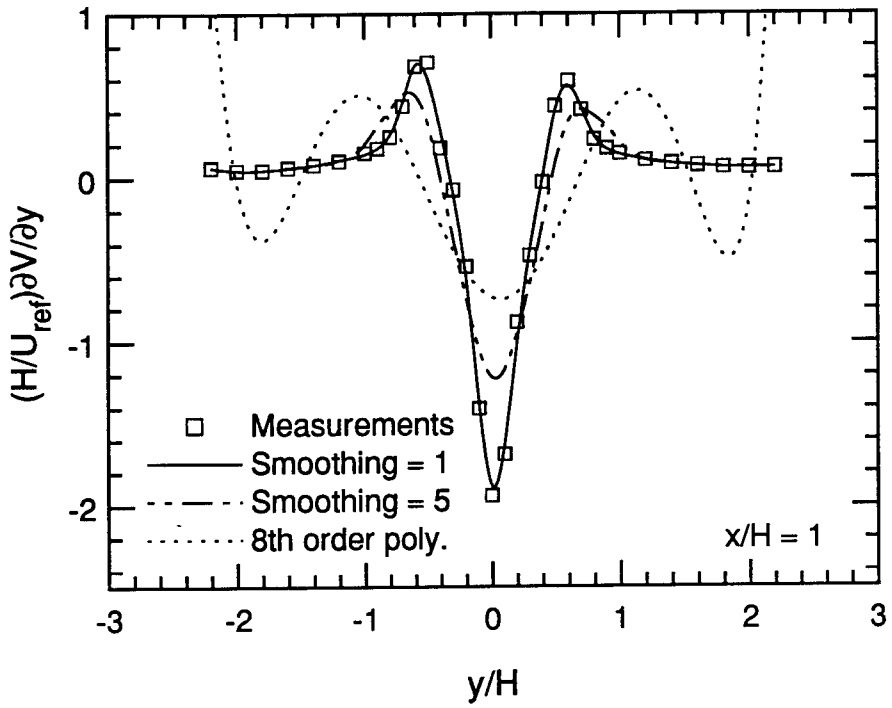


Figure 4. Smoothing and curve fitting the first derivative of experimental data.

Lastly, examples of experimental data which can be fit well to an 8th order polynomial curve are given in Figures 5 and 6. Figure 5 shows the axial turbulent normal stress profile downstream (x/H = 7) of a bluff body. Only two curve fits, the 5 point smoothed and the least squares polynomial fit are shown on this figure. This is because the 13 point smoothed curve lies somewhere between these two curves and cannot be seen well. Both curve fits shown do a good job fitting the experimental measurements, however, the polynomial fit is smoother and thus would be a better choice. Also, notice that both curves fit between the measurement points toward the top of the figure. Figure 6 shows the first derivative of the data shown in Figure 5. The points in this figure oscillate more than those shown in the previous figures. This is because the measurements (see Figure 5) oscillate somewhat and the differentiation process amplifies this. The polynomial curve fit to the data shown in Figure 6 is much smoother than the 5 point smoothing algorithm and would be a better choice of representing the first derivative of the axial turbulent normal stress.

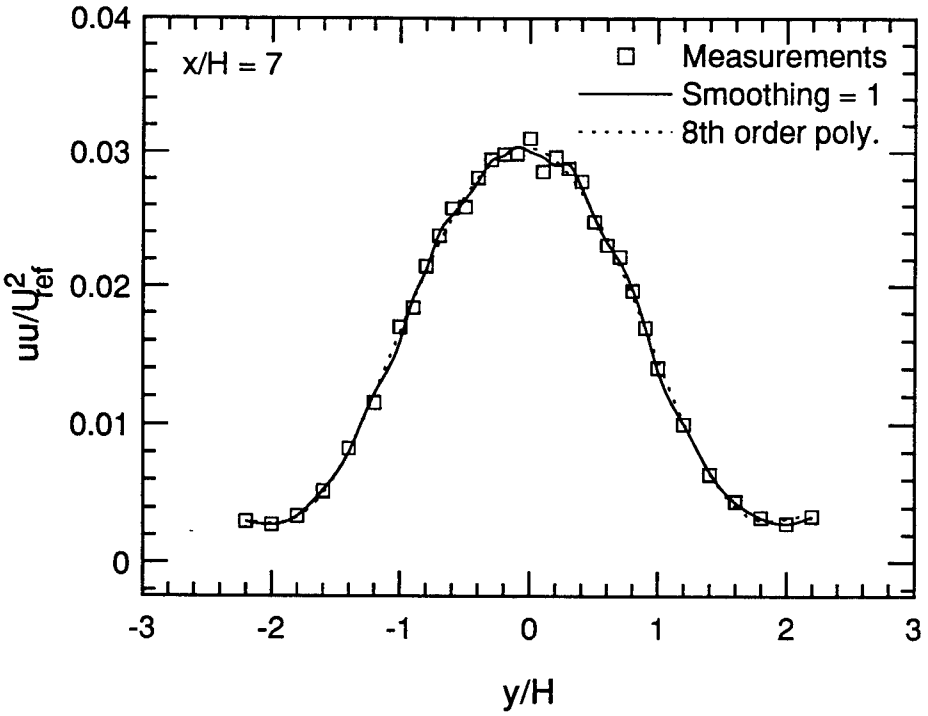


Figure 5. Smoothing and curve fitting experimental data.

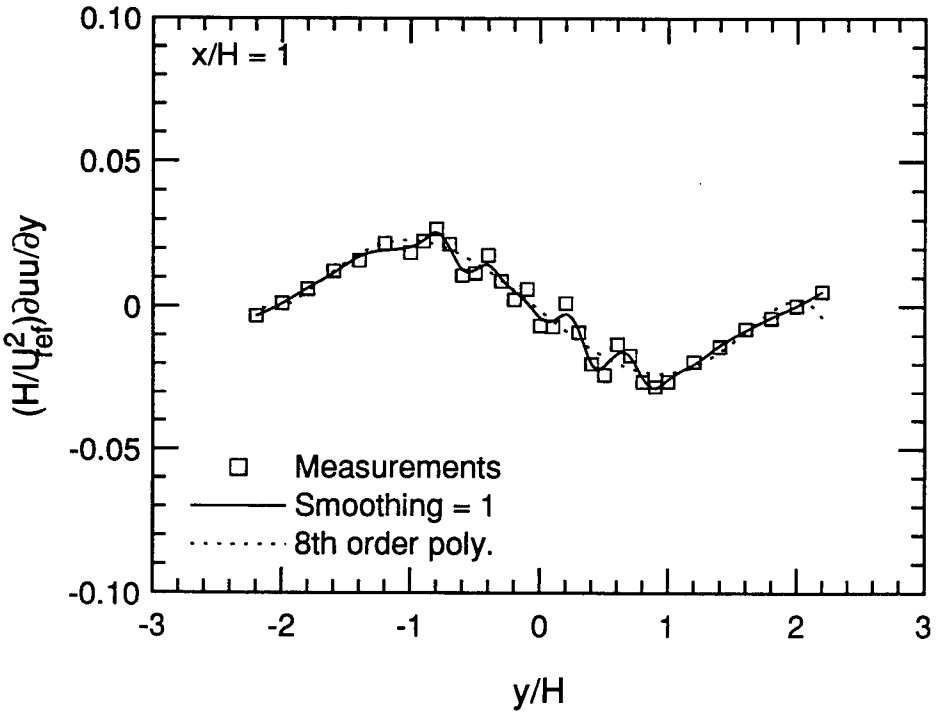


Figure 6. Smoothing and curve fitting the first derivative of experimental data.

CONCLUSIONS

Computer software was developed and delivered to WL/POPT for, 1.) determining the statistical uncertainty of laser Doppler velocimetry (LDV) data, 2.) numerically differentiating experimental measurements, 3.) automated smoothing and curve fitting experimental measurements and, 4.) plotting this data. In addition, a Wright Laboratory senior engineer was trained well enough to use and modify this software as part of this effort.

ACKNOWLEDGMENTS

The author would like to thank Dr. A. S. Nejad for his continued support in the general area of experimental diagnostics and for all the help he provided to make this summer program a success. Thanks are also due to Mr. C. Raffoul for helping define the project and openly sharing his personal and professional ideas. This investigation was performed at the Aeropropulsion and Power Directorate, Wright Laboratory (WL/POPT) under the Summer Faculty Research Program supported by AFOSR.

REFERENCES

Bendat, J. S. and Piersol, A. G. (1986), *Random Data: Analysis and Measurement Procedures*, 2nd ed., John Wiley and Sons.

Dougherty, E. R. (1990), *Probability and Statistics for Engineering, Computing, and Physical Sciences*, Prentice Hall, Inc.

Gerald, C. F. (1978), *Applied Numerical Analysis*, 2nd ed., Addison-Wesley Publishing Co.

Savitzky, A. and Golay, J. (1964), "Smoothing and Differentiation of Data by Simplified Least Squares Subroutine," *Analytical Chemistry*, Vol. **36**, p. 1627.

APPENDIX I

```
      SUBROUTINE SmoothSG (data_in, num_pts, smooth_num, data_out)
c
c*********************************************************************
c
c     The function of this subroutine is to reduce noise(i.e. smooth)
c     in a sampled data set. This algorithm employs the Savitzky-Golay
c     technique of simplified least squares smoothing. The technique
c     uses convolution where each data point is recalculated as a
c     weighted average of its original value and the surrounding data
c     points. The degree of smoothing is a function of the number of
c     surrounding data points used in the convolution. The more data
c     (i.e. the larger the convolution kernel) the larger the degree
c     of smoothing. The weights used in the convolution kernels are
c     initialized in data statements.
c
c     See:
c
c     Savitzky, A. and Golay, J.(1964), "Smoothing and Differentiation
c     of Data by Simplified Least Squares Subroutine,"
c     Analytical Chemistry, Vol. 36, p. 1627.
c
c     Variable definitions:
c
c     data_in    : input data vector to be smoothed (0:num_pts).
c     num_pts    : number of data points in input data set.
c     smooth_num : selects number of points used in convolution.
c     data_out   : smoothed output data vector (0:num_pts).
c     coef(i,j)  : convolution coefficients used to smooth data.
c                  i is smooth_num, j is the coefficient number(0-12).
c     norm(i)    : normalization parameter. i is smooth_num.
c     num_coef   : number of coefficients(and points) used in smoothing.
c     order      : number of points at beginning & end of data set
c                  which cannot be included in the convolution.
c     start      : first point in data set where convolution can occur.
c     stop       : last point in data set where convolution can occur.
c*********************************************************************
c
      REAL data_in(0:num_pts),data_out(0:num_pts)
      INTEGER coef(0:4,0:12),norm(0:4),num_pts,smooth_num,num_coef,
     &order,start,stop,i,j,k
c
c*********************************************************************
c     Define convolution coefficients used in smoothing algorithm
c*********************************************************************
c
      data (coef(0,j), j=0,12) /-3,12,17,12,-3,0,0,0,0,0,0,0,0/
      data (coef(1,j), j=0,12) /-2,3,6,7,6,3,-2,0,0,0,0,0,0/
      data (coef(2,j), j=0,12) /-21,14,39,54,59,54,39,14,-21,0,0,0,0/
      data (coef(3,j), j=0,12) /-36,9,44,69,84,89,84,69,-44,9,-36,0,0/
      data (coef(4,j), j=0,12) /-11,0,9,16,21,24,25,24,21,16,9,0,-11/
      data norm /35,21,231,429,143/
c
      IF(smooth_num .GE. 1 .AND. smooth_num .LE. 5) THEN
c
```

```
c*********************************************************************
c     Calculate the number of convolution coefficients(and points) to be
c     used in the smoothing process. Also, find the beginning and ending
c     data points in the input data set where the convolution process
c     takes place. The following parameters result for given smooth_num.
c
c     smooth_num          num_coef(# points in smoothing)         order
c
c        1                    5                                     2
c        2                    7                                     3
c        3                    9                                     4
c        4                   11                                     5
c        5                   13                                     6
c
c*********************************************************************
c
      num_coef=2*smooth_num+3
      order=(num_coef-1)/2
      start=order
      stop=num_pts-order
c
c*********************************************************************
c     Initialize output data to zero
c*********************************************************************
c
      DO i=0,num_pts-1
      data_out(i)=0.0
      END DO
c
c*********************************************************************
c     Performs convolution on input data set. This algorithm moves
c     through the data set and smooths each point based on the value of
c     the point and the values of its neighboring points and the values
c     of the convolution coefficients.
c*********************************************************************
c
      DO i=start,stop
c
       DO j=0,num_coef-1
       data_out(i)=data_out(i)+data_in(i-order+j)*coef(smooth_num-1,j)
       END DO
c
c*********************************************************************
c     Normalize smoothed output data points
c*********************************************************************
c
       data_out(i)=data_out(i)/REAL(norm(smooth_num-1))
      END DO
c
c*********************************************************************
c     Set end points of smoothed output data to input data values
c*********************************************************************
c
      j=order-1
      k=num_pts+1-order
      DO i=0,order-1
       data_out(j)=data_in(j)
       data_out(k)=data_in(k)
       j=j-1
       k=k+1
      END DO
c
      END IF
      END !Subroutine SmoothSG
```

Issues Involved in Developing an Object-Oriented System by Reengineering an Existing System

Raghava G. Gowda, Ph.D.
Assistant Professor
Department of Computer Science

University of Dayton
300 College Park
Dayton, OH 45469

Final Report for:
Summer Faculty Research Program
Systems Group (WL/AAAS-3)
Avionics Directorate
Wright-Patterson Air Force Base
USAF Focal Point: Michael Bohler
Project Coordinator: Barbara Eldridge

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

September 1994

Issues Involved in Developing an Object-oriented System by Reengineering an Existing System

Raghava G. Gowda, Ph.D.
Assistant Professor
Department of Computer Science
University of Dayton

Abstract

This report describes the issue and process of reengineering an existing system to develop an object-oriented model for the system. A systematic process of reengineering was applied to a subsystem of the Integrated Test Bed (ITB) facility. The ITB is used to support the development, testing, and evaluation of advanced avionics systems in the Avionics Directorate at Wright-Patterson Air Force Base. The subsystem being reengineered was the Guidance subsystem of the Operational Flight Program (OFP). The initial effort concentrated on the Terrain-Aided Flight Algorithm (TFA), or the Vertical Steering subsystem, a component of the Guidance subsystem. Then a bottom-up approach was used to derive an object model for the OFP. The objective of the reengineering task was to develop an object-oriented model for the current system . The system was initially implemented in JOVIAL and later in Ada. The first phase of the effort was to understand the existing system. This was accomplished by studying the related documents and code, and constructing design specifications for the existing system. Structure Charts were used to document the design of an earlier version of the system programmed in JOVIAL. Visibility Diagrams and notations similar to Object-Oriented Structured Design (OOSD) were used to document the design features of the current system implemented in Ada. The processes and corresponding data were grouped together to encapsulate them into objects/classes. The second phase was to develop an object-oriented model for the system. An object-oriented model was developed using the Object Modeling Technique. Some of the CASE tools used for this effort include Software Through Pictures™, Rational Rose™, and OMTool™. The report includes some of the experiences and lessons learned.

Issues Involved in Developing an Object-oriented System by Reengineering an Existing System

Raghava G. Gowda, Ph.D.

Introduction

Reengineering software refers to the task of capturing the design details and algorithms of an existing system by studying its code and modifying the system for future use. The efforts involved in reengineering can be substantial. Sometimes it may be cheaper to abandon the existing system and develop a new system altogether. Reengineering is a time-consuming task as one has to go through old code and relate it to the system's requirements. The task becomes complicated due to poor design, lack of documentation, and use of global variables. In spite of the difficulties encountered in the reengineering process, it may be the only alternative available to modify some existing systems. Reengineering may be needed for:

a. modifying an existing system
b. extracting algorithms or process logic from the existing system
c. building an object-oriented system based on the existing system
d. validating an existing system

One of the most compelling reasons for reengineering is to develop object-oriented systems. This is an important step in enhancing reusability of software. The processes and the data being manipulated have to be thoroughly understood before encapsulating them into objects.

Reengineering Process

A subsystem of the Operational Flight Program (OFP), namely, the Terrain-Aided Flight Algorithm (TFA) subsystem, was selected for the purpose of reengineering. It was selected mainly for the following reasons:

a. The document "Computer Program Product Specification for the Terrain-Aided Flight Algorithm (TFA) System" was available. It documents the subsystem as originally implemented in the JOVIAL language.
b. The system was reimplemented in Ada. Most of the functions of the TFA system were implemented in the package VERTICAL_STEERING which was a component of the Guidance subsystem in the OFP.

The following tasks were performed for reengineering the TFA (or Vertical Steering) subsystem:

a. All the documents related to the Integrated Test Bed (ITB) were studied.

b. The TFA subsystem was documented using the Structure Chart Editor (SCE) of the CASE tool Software Through Pictures™. Data dictionary entries were recorded in the CASE tool for various modules. Documentation was done with the view that these specifications would be useful for checking the consistency between the old and new systems.

c. The Ada packages were studied. A listing of various packages was obtained.

d. Visibility Diagrams for Ada packages were drawn using the Rational Rose™ CASE tool

e. Detailed diagrams were drawn for the package specification and body of VERTICAL_STEERING.

f. Data associated with the procedures and tasks of VERTICAL_STEERING were tracked in various packages and documented separately.

g. Other Ada packages of the Guidance subsystem such as HORIZONTAL_STEERING were documented.

h. An overall Object Model for the OFP was developed using the Object Modeling Technique (OMT). The OMTool™ from the Advanced Systems Concepts was used for developing the Object Model.

An overview of the Integrated Test Bed (ITB) Facility

The purpose of the Integrated Test Bed (ITB) facility is to support the development, test, and evaluation of advanced avionics systems and subsystems. It provides a real-time simulation of military aircraft performing an operational mission. The simulation generates the interface signals between the aircraft sensor suite and the avionics system so that the avionics equipment are subjected to a data signal environment which is nearly identical to actual flight. A simulated cockpit is also a part of the simulation for realistic evaluation of the avionics system. The ITB facility consists of the following components:

a. Avionics equipment: Mission Processing Clusters, Integrated Terrain Access/Retrieval System (ITARS), Quiet Knight Data Processor (QKDP), Constant Source Operator's Terminal (CSOT), MIL-STD-1553B Multiplex Buses, High Speed Data Buses (HSDB), Equipment Under Test (e.g., Common Modules), and Mission Software

b. Simulated Avionics and Environment: Real-Time Simulation Computers, Simulation Software, Out-The-Window Scene/Sensor Generation, Crewstation, and High Speed Simulation Network

c. Support Elements: ITB Support Hardware and Software and System Test Software

The subsystem selected for object-oriented modeling was the Guidance subsystem which consists of the Terrain-Aided Flight Algorithm (TFA). This subsystem has been modified/expanded and implemented in the Ada package VERTICAL_STEERING. The subsystem we selected was a part of the OFP software. A brief description of OFP components is given below to characterize the environment in which the Guidance subsystem operates.

OFP Implementation in Ada

The OFP implemented in Ada supports a number of mission modes. Each mission mode has an alternate mode with minimum functions. The mission modes are supported by avionics functions, which are bundled into Loadable Program Units (LPU) for implementation purposes. The Table 1 gives an overview of the mission modes, avionics functions, and LPUs.

Mission Modes	Avionics Functions	Available LPUs
Null_Master _Mode	Initial_Guidance_Function	Guidance
Preflight_Master_Mode	Enhanced_Guidance_Function	Horizontal_Guidance
Degraded_Preflight_Master_Mode	Guidance_Function	Vertical_Guidance
Takeoff_and_Climb_Master_Mode	Degraded_Guidance_Function_1	Navigation
Degraded_Takeoff_and_Climb_MM_1	Degraded_Guidance_Function_2	SITAN_Main
Degraded_Takeoff_and_Climb_MM_2	Enhanced_Navigation_Function	SITAN_Acq1
Degraded_Takeoff_and_Climb_MM_3	Navigation_Function	Inflight_Route_Repair
Approach_Land_Master_Mode	Degraded_Navigation_Function	Sensors
Degraded_Approach_Land_MM_1	Inflight_Route_Repair_Function	Cockpit
Degraded_Approach_Land_MM_2	Sensor_Control_Function	IMFK_Control
Cruise_Master_Mode	PVI_Function	DGS_Interface
Degraded_Cruise_MM_1	Degraded_PVI_Function_1	IMFK_Partial
Degraded_Cruise_MM_2	Degraded_PVI_Function_2	Mission_Manager
Degraded_Cruise_MM_3	Mission_Manager_Function	
TF_TA_Master_Mode		
Degraded_TF_TA_MM_1		
Degraded_TF_TA_MM_2		
Degraded_TF_TA_MM_3		
Up_And_Away_Minimum_Mode		

Table 1. Mission Modes, Avionics Functions and LPUs

The OFP is supported by the Ada Avionics Real-Time Software (AARTS) operating system. It performs task loading, scheduling, running, reconfiguration, and unloading of tasks. The services provided by AARTS include I/O via PI Bus/HSDB/MIL-STD-1553B, file services, "Wait" for messages, semaphore services, and event services.

Subsystem being modeled:

Terrain-Aided Flight Algorithm (TFA) or Vertical Steering implemented in JOVIAL

The purpose of the TFA is to implement a Terrain Following/Terrain Avoidance (TF/TA) algorithm using digital terrain elevation data (DTED) from the Integrated Terrain Access/Retrieval System (ITARS). It also implements the Sandia Inertial Terrain Aided Navigation (SITAN) algorithm using terrain elevation data from ITARS. Using these

algorithms the TFA generates in real-time lateral and vertical flight commands suitable for use in directing the flight cue on the Head-Up Display (HUD) or as input to an autopilot system. The system also generates a corrected aircraft position using the SITAN algorithm. Figure 1 shows the functional relationships among the TFA and other subsystems implemented in JOVIAL. The TFA performs the following major functions:

a. Controls algorithm activity as directed by incoming MIL-STD-1553B messages.
b. Receives and maintains terrain elevation data from the ITARS across the High Speed Data Bus (HSDB).
c. Generates heading and pitch commands as directed by the TF/TA algorithms.
d. Generates accurate latitude and longitude position estimates using the slopes of the terrain beneath the aircraft.
e. Outputs TF/TA/SITAN results over the MIL-STD-1553B Bus.
f. Coordinates processor-to-processor communications over the Shared Memory Architecture Real-time Network (SMARTNet).

Algorithm Control

When MIL-STD-1553B messages are received, the TFA subsystem generates appropriate TF, TA, SITAN, and aircraft steering results. The system is driven by a MIL-STD-1553B minor cycle interrupt. The interrupt acts as a heartbeat for the system and all internal events occur at specified minor cycle intervals.

ITARS Control

ITARS is an avionics terrain database subsystem that stores terrain, image, and/or threat information. It generates three video display channels and outputs pertinent terrain and feature data to other avionics functions over an HSDB. The three video channels include two raster channels and one stroke channel.

ITARS data is received by the TFA subsystem across a HSDB. Two separate streams of data, one for each algorithm are sent to TFA periodically and buffered internally in system memory. The data consists of an Area Load of the terrain elevation points surrounding the aircraft. ITARS continuously sends new area loads of terrain data centered about the current aircraft position.

Terrain Following (TF) Processing

When activated by a MIL-STD-1553B message, the TF algorithm scans the ITARS terrain data in front of the aircraft and computes a flight vector command that will insure terrain clearance over approaching critical terrain features. The TF algorithm will also generate negative flight vector commands in an attempt to maintain a set clearance altitude over the terrain. Final TF results are sent to the Steering process for later use.

TERRAIN-AIDED FLIGHT ALGORITHM
(JOVIAL SYSTEM)

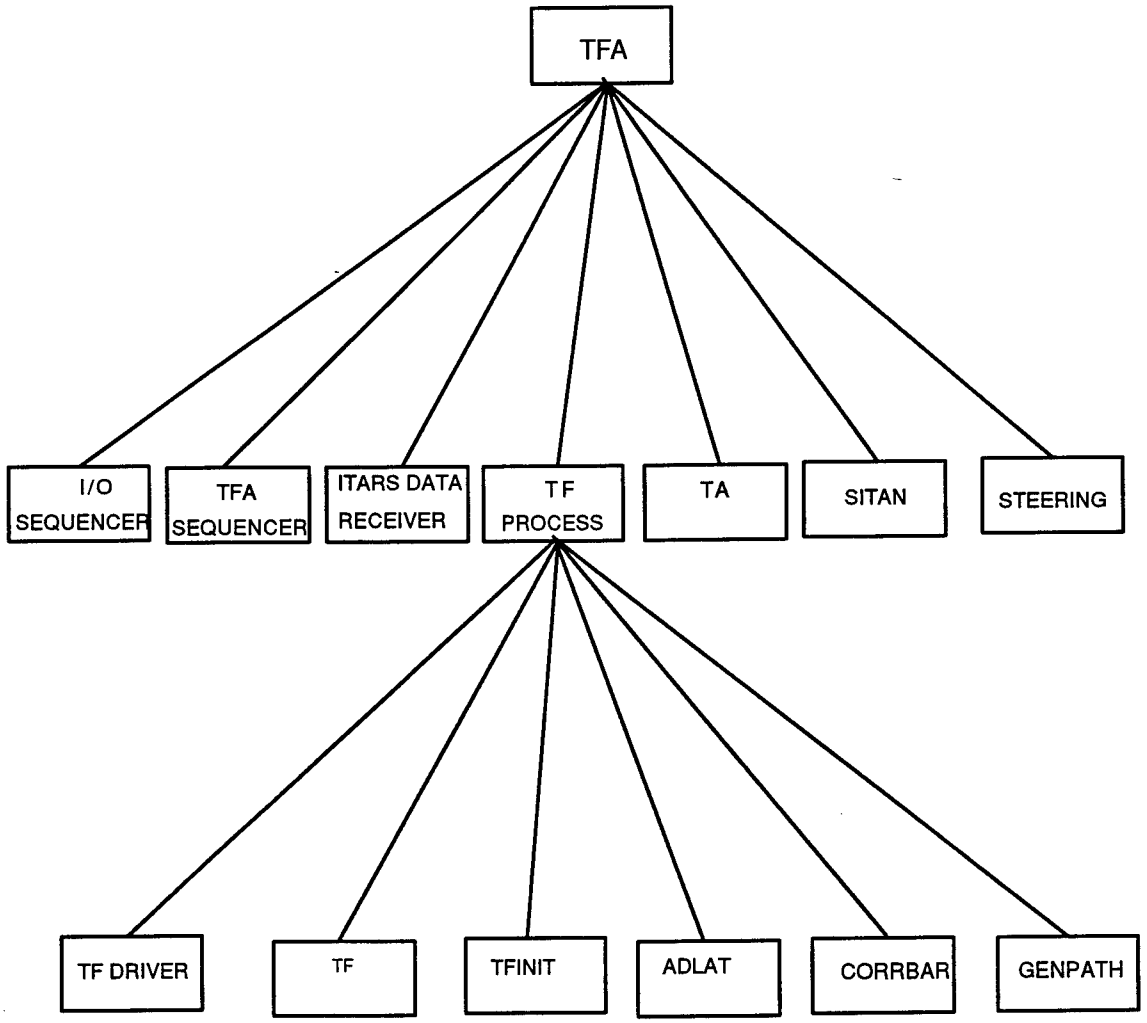


Figure 1

Terrain Avoidance (TA) Processing

The TA algorithm is activated periodically by a MIL-STD-1553B message and generates lateral heading commands to guide the aircraft around high elevation terrain features. TA uses resident ITARS terrain data as the major input into a cost function that weighs elevation height with other cost parameters during lateral path generation. This lateral path is sent to the Steering process for use in steering computations.

SITAN Processing

Incoming Inertial Navigation System (INS) data is sent periodically to the SITAN algorithm which computes a corrected aircraft position based upon computed slope values from the underlying ITARS terrain data. SITAN uses the initial INS position as a best-guess estimate of the true aircraft position. Using a series of Kalman filters, SITAN refines this estimate and then tracks the computed position using terrain data from ITARS. SITAN results are sent across the MIL-STD-1553B Bus for use by the mission software.

Steering Process

Vertical and lateral guidance commands from the TF and TA processes are stored internally by the steering process. Each time a MIL-STD-1553B steering command message is received, the steering process builds a steering output message from the most current vertical and lateral guidance commands and outputs this message across the MIL-STD-1553B Bus.

VERTICAL_STEERING Implementation in Ada

After documenting the TFA algorithm as implemented in the JOVIAL language, the current system implemented in the Ada language was studied. A detailed listing of Ada files was obtained and the TFA algorithm implemented in the Guidance subsystem was documented. It was found that the Ada version had modified the JOVIAL version substantially. The VERTICAL_STEERING package implements some aspects of the Terrain Following algorithm, a component of TFA. The next step was to document essential aspects of the Ada packages and their relationships with other packages. The details of the package body were documented using Structured English and detailed diagrams.

Structured English for Vertical Steering

Process details were documented using Structured English from the body of the VERTICAL_STTERING package. It could be documented in a CASE tool. Indentation helps to highlight control structures. Action Diagrams may used with the Structured English. The package specification should emphasize the major processes and algorithms and should avoid language features.

File: [ITBC.PMF.OFP.GUIDANCE]VERTICAL_STEERING_. ADA

package spec VERTICAL_STEERING

-- Vertical Steering Manger. Generates flight path guidance command.

-- Includes specification of VERTICAL_STEERING_TASK

File: [ITBC.PMF.OFP.GUIDANCE]VERTICAL_STEERING ADA

package body VERTICAL_STEERING

procedure BROADCAST_VERTICAL_STEERING_RESULTS

Send the steering results message out on the PI Bus.

Maintain the Vertical Steering 'deadman counter', incrementing the value and resetting it to prevent its getting too big.

Set 'no nav data' status.

procedure CREATE _STRAIGHT

Create a lateral path in lieu of horizontal steering. The path is straight ahead (i.e. along the current heading).

procedure INIT_VERTICAL_STEERING

Initialize Vertical Steering's status to "UNCONFIGURED" status

task body VERTICAL_STEERING_TASK

INIT_VERTICAL-STEERING

Loop for ever. Any configuration messages and aircraft sensor messages will be available whenever they come in. No need to 'wait' for them.

Wait until it is time to run. In the meanwhile, monitor the reception of nav data. Regardless of what timer does, don't run until a new nav solution is received.

```
loop
        begin
        TIME_TO_RUN := false;
        while not (TIME_TO_RUN) loop
                select
                        accept TIMER_EVENT do
                                TIME_TO_RUN := TRUE;
                                ...
                        end TIMER_EVENT;
                or
                        accept NAV_MSG (....)
                                ...
                        end NAV_MSG;
                or
```

```
                    accept HOR_STEER_MSG (... )
                        ...
                    end  HOR_STEER_MSG;
            end select;
    end loop;
    if  not (NAV_SOLUTION_RECEIVED) then
            BROADCAST_VERTICAL_STEERING_RESULTS
            Send invalid message
    elseif not (AIRPLANE.WEIGHT_OFF_WHEELS)
                        or
            (PILOT_COMMAND.VERTICAL_OPT /= NO_VERTICAL_STEERING) then
             set values
            BROADCAST_VERTICAL_STEERING_RESULTS
    else
            Transfer cumulative map errors into output message;
            Extract message info
            if  (PILOT_COMMAND.HOR_STEER_OPT =  NO_HORIZONTAL_STEERING) then
                    No horizontal steering
                    Build our  path
                    CREATE_STRAIGHT(CURRENT_AC_STATE, LATERAL_FLIGHTPATH);
            else
                    If path from Horizontal Steering is valid, copy it to a local buffer
            endif;
            if (PILOT_COMMAND.VER_STEER_OPT = STEER_IF.NO_VERTICAL_STEERING ) then
                    Set locally formatted GLOBAL_SENSOR values to SENSOR
                    Generate TF Command
            endif;
            Set the validity bit on and broadcast the result
    endif;
    If an exception is raised during processing, set the validity bit off and continue.
    This will show up in the cockpit as a 'blink' in the Vertical Guidance Cue.
end loop;
```

Visibility Diagrams for Ada Packages

The Structured English captures process details. Ideally it should contain data details also. As the data used in a particular package could be derived from multiple packages, a Visibility Diagram may be drawn using the package

specification. A Visibility Diagram for VERTICAL_STEERING is shown in Figure 2. The objective is to isolate data elements which are essential for a particular process. Once we specify essential process logic and data we may encapsulate them into objects.

The Visibility Diagram shows that sixteen packages are visible for VERTICAL_STEERING. Some of them are generic packages such as por_reals and por_integers which deal with portable reals and portable integers. Attention was given to those packages which have been used (either to access data or a procedure) in the package body of VERTICAL_STEERING.

From the packages shown in the visibility diagram the data can be systematically tracked and documented. Table 2 shows a format for tracking data elements.

Variable Name/Record	Type/Subtype	Size in words	Value	Defined in Package	Used in Package
WAKEUP_REASON	WAKEUP_TYPE	1	...	MESSAGE_IO	VERTICAL_ STEERING ...
HORIZONTAL_STEERING_ OPTIONS	HORIZONTAL_ STEERING_ OPTIONS	1	...	STEER_IF	VERTICAL_ STEERING ...
TIME_TO_RUN	boolean			VERTICAL_STEERING	VERTICAL_ STEERING ...
VERTICAL_STEERING	VERTICAL_ STEERING_ RESULTS	6	...	STEER_IF	VERTICAL_ STEERING
...	...			...	...

Table 2. Tracking Data Elements

This method of tracking data elements can be very laborious. Reverse Engineering tools capable of tracking data in the entire system should be used for this purpose. The basic task is to determine whether the data element is essential for the process or not. Most of the flags used in parameter lists are dependent on programming style and other constraints. They could be simplified or eliminated. The derived data elements, in general, should not be part of the data structure as they can be easily computed from the basic data.

Detailed Diagrams

A detailed diagram for each process can be drawn using notations similar to Object-Oriented Structured Design (OOSD). It includes package interfaces, procedures, functions and tasks of the package, data declared within the package, and possibly its interface with other packages. Figure 3 shows a detailed diagram for VERTICAL_STEERING.

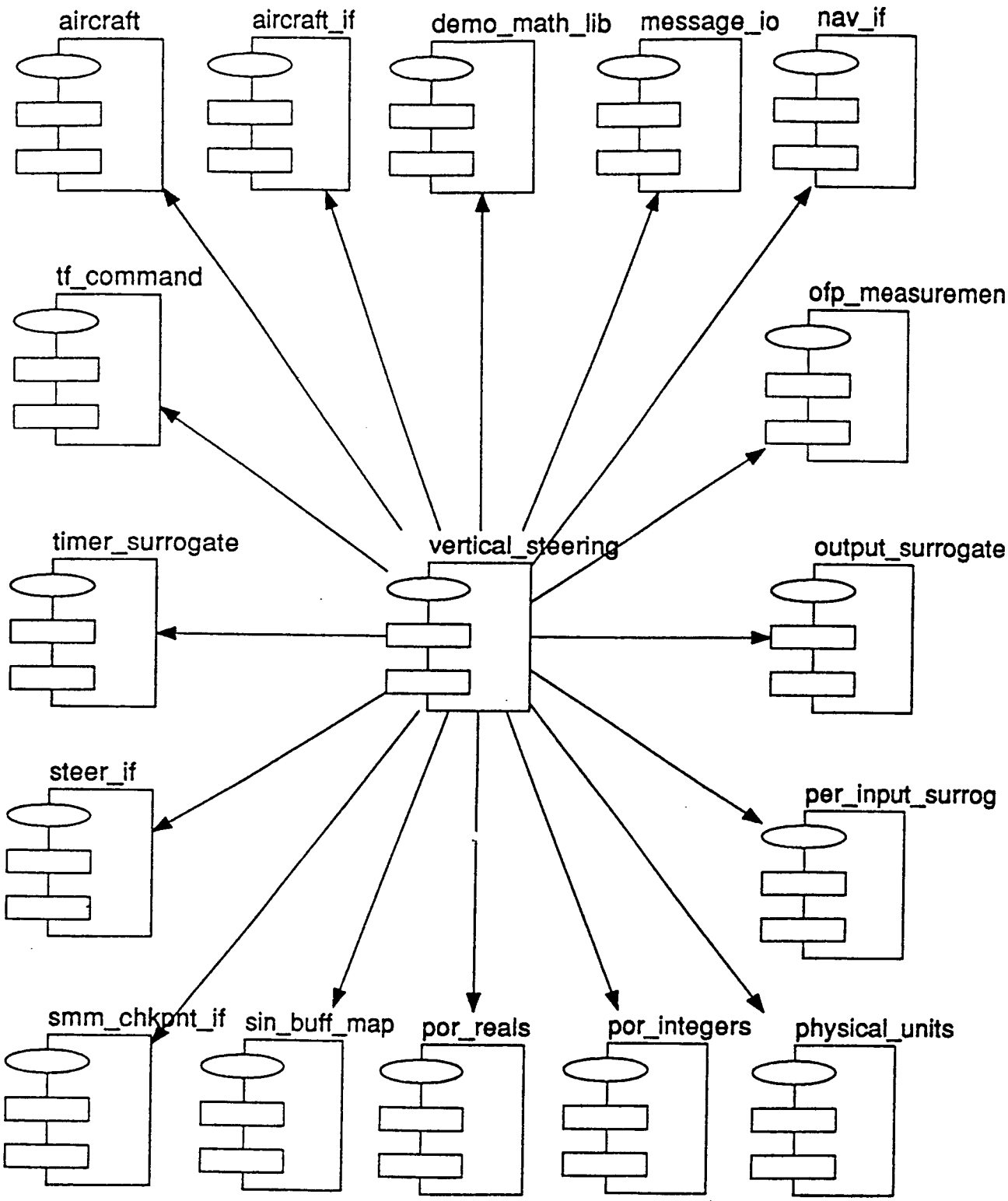


Figure 2. Visibility Diagram for VERTICAL_STEERING

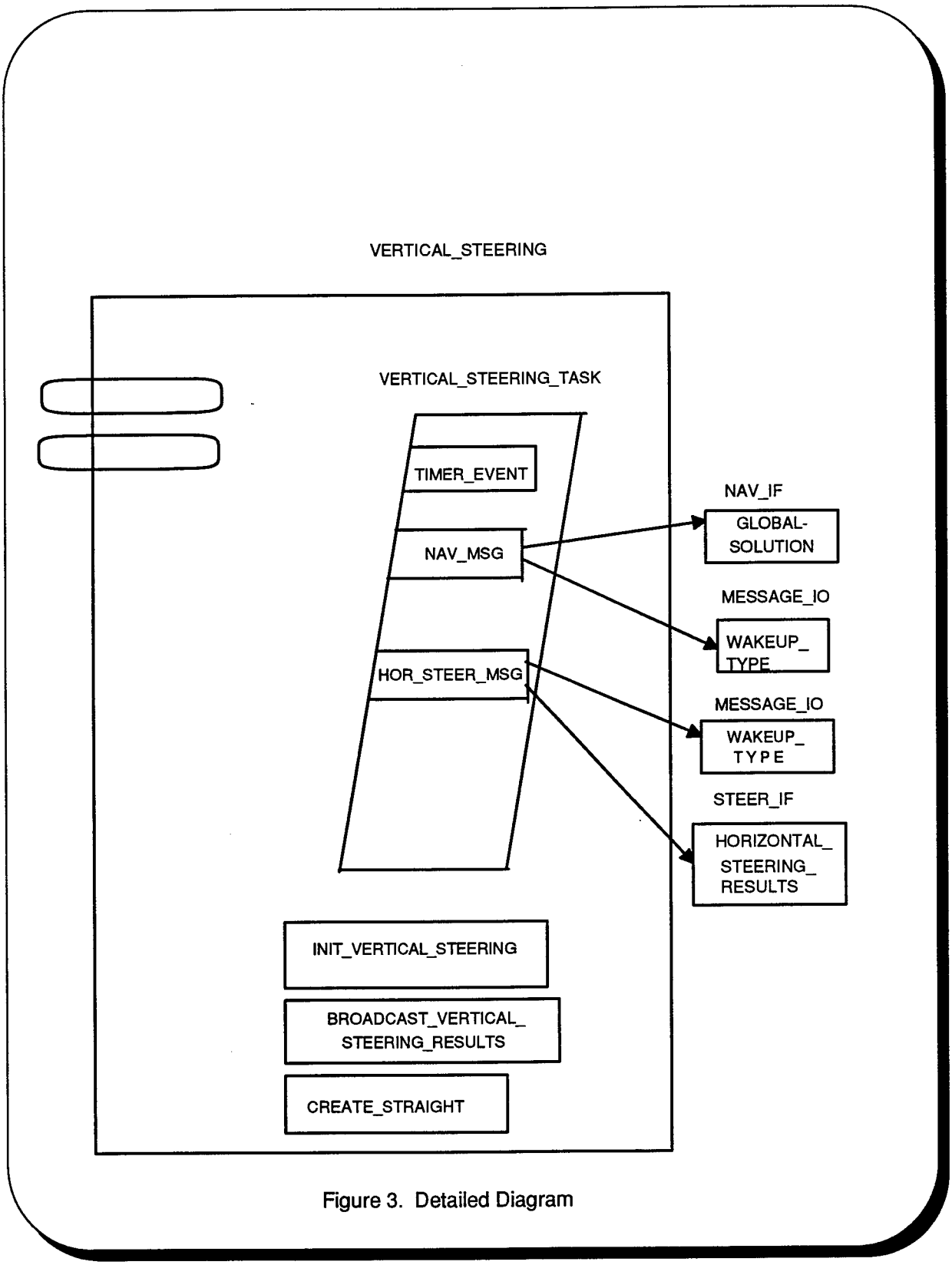


Figure 3. Detailed Diagram

Object Model for OFP

In an ideal case the approach for deriving an overall model for the system will use the bottom-up approach. First, objects and classes are identified, and classes are grouped into subsystems so that cohesion of subsystem components is maximized and the coupling between subsystems is minimized. Sometimes the top-down approach may be needed because of the size of the system and multiple resource requirements. We have used both the top-down and bottom-up approaches in arriving at an overall object model. The system documentation provided hardware and system software requirements, and Ada code provided a bottom-up view of the system . Partitioning the system into subsystems is a very crucial activity. The poor partition will adversely affect the resource requirements, performance, and reusability of subsystems.

After documenting the aspects related to VERTICAL_STEERING, the HORIZONTAL_STEERING and other packages of the Guidance system were examined. Finally an Object Model for the Operational Flight Program was developed using OMT. The designer of the OFP used object-oriented concepts (Abstract Data Types) in the design of Ada packages. The designer was consulted in developing the overall model. The Object Model closely resembles the current architecture of the OFP. The OMTool™ was used for drawing the Object Model shown in Figure 4.

Data dictionary entries are an essential aspect of any development or reengineering effort. The Structured English and data details can be used to define objects/classes in the data dictionary. Object Modeling Technique also provides for a Dynamic Model and a Functional Model. The Functional Model uses Data Flow Diagrams to capture process details. Since we have already documented the process details in the reengineering process, we may not need the Functional Model. The Dynamic Model uses State Transition Diagrams to represent control flow of the system. The Dynamic Model should be used to represent relationships among classes/objects. Collaboration graphs or object-interaction diagrams can also be used to represent interactions among subsystems/classes. The Object Model, Functional Model, and Dynamic Model show three distinct orthogonal views of the system. The Object Model shows inheritance, association and aggregation relationships between classes/objects. The Dynamic Model shows state transitions. The Functional Model shows the transformation of inputs to outputs. Consistency in naming processes and data should be observed across these models as it is essential to provide a coherent view.

The major classes of the Operation Flight Program are the Pilot/Vehicle Interface, Guidance, ITARS, Navigation, Sensor Management, and Steering. Though the current system also has a Mission Manager function, we consider that to be a part of the Avionics Operating System.

The Mission Manager is a subsystem of AARTS which manages the OFP. It is responsible for managing the invocation of appropriate avionics functions for each master (mission) mode invoked by the pilot. The Mission Manager will determine the appropriate functions and associated LPUs for each mission mode and will initiate load/start of these functions through the AARTS System Executive. It also provides the capability to specify functional Hot Backups and

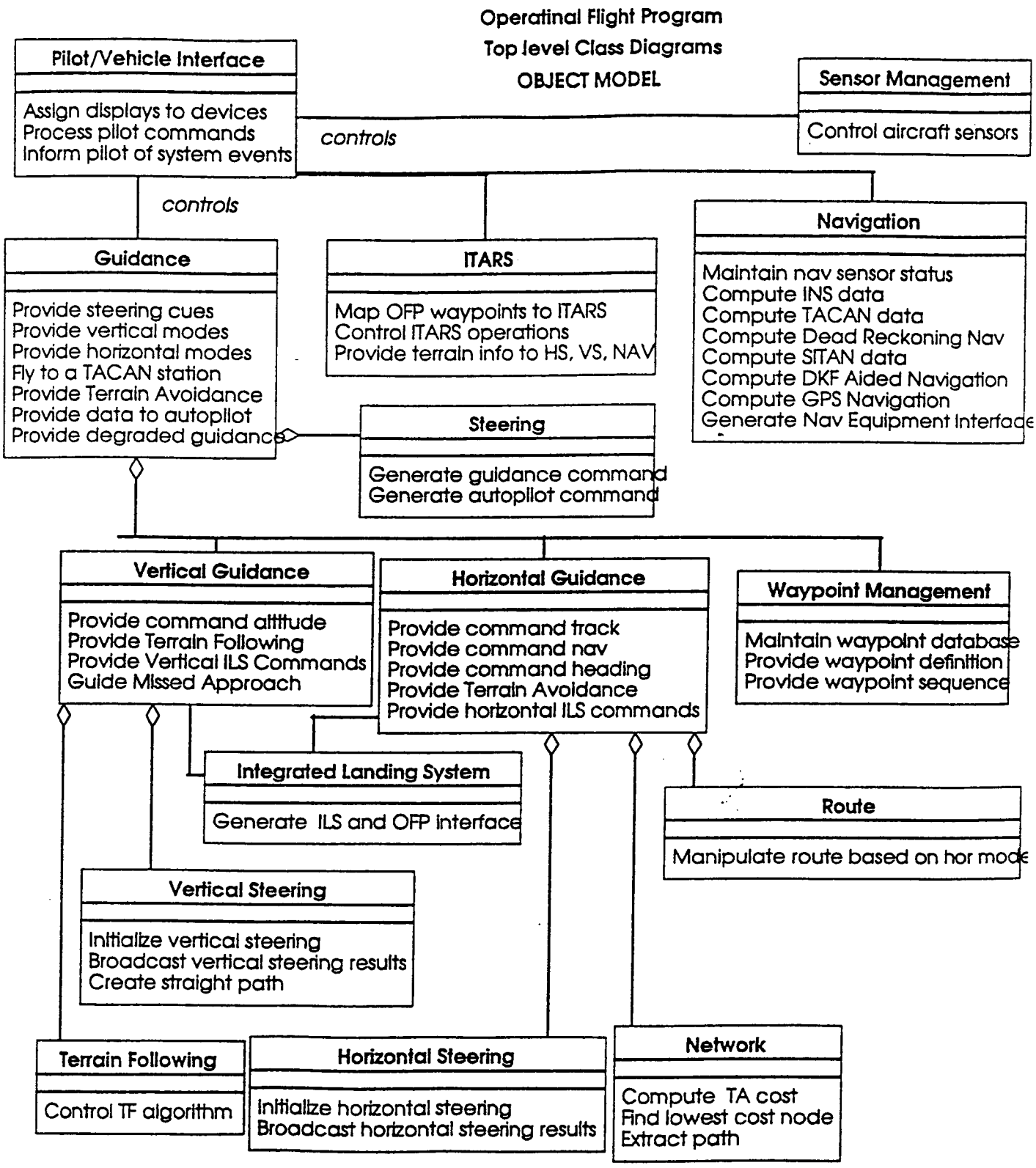

Operatinal Flight Program
Top level Class Diagrams
OBJECT MODEL
Pilot/Vehicle Interface
Assign displays to devices
Process pilot commands
Inform pilot of system events
controls
Sensor Management
Control aircraft sensors
controls
Guidance
Provide steering cues
Provide vertical modes
Provide horizontal modes
Fly to a TACAN station
Provide Terrain Avoidance
Provide data to autopilot
Provide degraded guidance
ITARS
Map OFP waypoints to ITARS
Control ITARS operations
Provide terrain info to HS, VS, NAV
Navigation
Maintain nav sensor status
Compute INS data
Compute TACAN data
Compute Dead Reckoning Nav
Compute SITAN data
Compute DKF Aided Navigation
Compute GPS Navigation
Generate Nav Equipment Interface
Steering
Generate guidance command
Generate autopilot command
Vertical Guidance
Provide command attitude
Provide Terrain Following
Provide Vertical ILS Commands
Guide Missed Approach
Horizontal Guidance
Provide command track
Provide command nav
Provide command heading
Provide Terrain Avoidance
Provide horizontal ILS commands
Waypoint Management
Maintain waypoint database
Provide waypoint definition
Provide waypoint sequence
Integrated Landing System
Generate ILS and OFP interface
Route
Manipulate route based on hor mode
Vertical Steering
Initialize vertical steering
Broadcast vertical steering results
Create straight path
Terrain Following
Control TF algorithm
Horizontal Steering
Initialize horizontal steering
Broadcast horizontal steering results
Network
Compute TA cost
Find lowest cost node
Extract path

Figure 4

alternate (degraded or reduced) mission modes. The basic functions of the Mission Manager are to determine software requirements for each mission phase; to command load of required software; to accept system resource update reports; to determine degraded mission phase requirement; to auto-transition between mission phases; and to inform applications of mission mode changes.

a. Pilot/Vehicle Interface

The Pilot Vehicle Interface consists of the Controls and Display, Position Update, and Autopilot functions. The primary task of the Controls and Displays function is to accept and initiate the pilot's control commands, and maintain and display appropriate information on the aircraft state, avionics subsystems, and mission status. The Position Update function will allow the pilot to designate a location of known position, using the Laser Ranger and aiming reticle on the HUD. Based on the difference between the calculated position of the designated location and the known value, actions to update the INS position will be initiated. The Position Update function will also allow updates based on GPS, SITAN, or DKF outputs. The Autopilot Interface function will format an interface message for the autopilot.

b. Guidance

The Guidance subsystem shall determine deviations from the projected flight path and determine the steering information (steering cues and steering commands) necessary to maintain flight along the projected flight path. The major functions of the Guidance subsystem are to provide steering cues to direct the pilot to the desired destination; to generate Vertical Modes and Horizontal Modes; to fly to a specified TACAN station; to fly over lowest terrain (Terrain Avoidance); to generate commands for autopilot; and to provide degraded modes if processing resources are lost. Some of the features of the guidance/steering modes are given below:

Waypoint Management

Waypoint Management (WPM) will be responsible for maintaining the mission waypoint database. This includes both the mission waypoint definitions and the current waypoint sequence. The WPM will provide inquiry services to provide waypoint definition data as well as waypoint sequence information.

Horizontal Guidance

The Horizontal Guidance subsystem consists of a Horizontal Steering Algorithm which implements Command Track, Command Navigation , Command Heading, Terrain Avoidance, and Instrument Landing System.

Command Track - will provide lateral commands to fly the great circle connecting the last waypoint to the next waypoint.

Command Nav - will provide lateral commands to fly a great circle from the present position to the next waypoint.

Command Heading - will provide lateral commands to fly a heading commanded by the pilot.

Terrain-Avoidance - generates lateral commands to fly a path along a designated route between the current

waypoint and the next, choosing the optimal flight path over the lowest terrain. Only the horizontal flight path is modified.

Instrument Landing System (ILS) - generates lateral commands to achieve the heading commanded by the pilot.

Vertical Guidance

The Vertical Guidance subsystem consists of a Vertical Steering algorithm which implements Terrain Following Command Altitude, and Instrument Landing System.

Command Altitude - generates vertical commands to maintain an altitude commanded by the pilot. It can be used in conjunction with Command Track, Command Nav, or Command Heading modes.

Terrain Following - generates vertical commands to fly a low-level flight path between the current waypoint and the next waypoint, maintaining a specified clearance above the terrain. Only the vertical flight path is modified to maintain the specified clearance. Lateral commands either come from the pilot or are automatically generated. Automatically generated lateral path commands may be any of the following: Command Track, Command Nav, Command Heading, or Terrain Avoidance.

Instrument Landing System (ILS) - generates vertical commands to achieve the vertical glide slope.

c. Navigation

The Navigation subsystem will be responsible for computing position (latitude and longitude), aircraft velocity, and wind information. This data is used throughout the OFP for steering/guidance calculations, weapons delivery guidance, pilot display, and other applications.

d. Sensor Management

Sensor Management controls the operations of the aircraft sensors. It controls the on/off logic and the processing of the sensor messages. Specific sensors controlled include the Radar Altimeter and the Laser Ranger.

e. ITARS

The ITARS Control function will be responsible for sending ITARS mode control messages to the ITARS subsystem and for accepting ITARS status messages. ITARS maintains a database of displayed Point Features, controls user requests, displays ranges, depictions, and assists active viewing.

Conclusions

Developing an object-oriented model for the OFP will enable use of the modules in future systems. It will enhance modifiability, reliability, and testability of the system. In the object-oriented approach the problem is modeled at the high level of abstraction of the real world, whereas in the traditional approach, the abstraction is basically around the processes

or data. The object-oriented approach makes the system easy to understand, modify, and reuse. The efforts should pay off in the long run.

In the current implementation of the OFP, most of the data declarations are done in the interface packages with generic packages. These packages contain detailed record descriptions of major components of the system such as the aircraft and navigation. They are instantiated and manipulated by packages of various subsystems. The record templates of the interface packages provide attributes of object classes. These data attributes can be partitioned and encapsulated into objects/classes by providing appropriate processes. The next important task is to identify the relationships among objects/classes such as inheritance, association, and aggregation.

Ideally, all the declarations of a package should be confined to the specification part of the package. In the current implementation, data elements used in a package are mainly subsets of interface definitions and other packages. This makes it difficult to verify the completeness and correctness of the code. It necessitates an understanding of all the packages which define records in order to understand a specific package. If a package imports a number of data elements which are components of many records it just adds to the complexity. The following guidelines may be kept in mind when redesigning the system:

1. The application package should contain data which are necessary for supporting its procedures, functions, or tasks. Exposure to unnecessary data introduces ambiguity in the process logic.
2. Some of the flags used in packages are intended to synchronize processes of other packages. Defining the role of a package in the overall structure will lead to a better understanding of the system and control over the flags. The nature of control mechanisms will vary in the object-oriented system Instead of hierarchic control, the objects should interact in a client-server relationship or use of inheritance.

The current design might have been motivated by a number of factors. One of the major factors is data communication between various subsystems through the PI Bus which needs a central store of updated information. Another factor is one of the strong feature of the Ada language, the generic package. With generic packages, one set of code can be reused with different sets of data types. The third factor relates to different formats of data needed by different subsystems due to hardware and software constraints. INPUT and OUTPUT SURROGATES packages are of this type. The input and output data are formatted in one package instead of distributing the tasks to different packages.

If we want to visualize packages as independent software "IC"s (Integrated Circuits!) there has to be a compromise made with the above issues. In an object-oriented approach emphasis is placed on "completeness" of classes/objects. We have provided an Object Model for the OFP which closely resembles the current design. In the current system, packages perform specified tasks and are similar to "has a" (aggregation or "consists of") relationship in an object-oriented paradigm.

The inheritance relationships among classes needs to be explored in order to take advantage of the reusability and modifiability aspects of an object-oriented approach. The new design should take advantage of object-oriented features of Ada 9X.

CASE tools have a significant role in the reengineering process. One of the facilities of CASE tools applicable for most of the systems is the Structure Chart facility used in the Structured Design. Structure Charts show hierarchic relationships between the system and subsystems. In, out, and inout parameters can be shown in the diagrams. The subsystem , process logic and data descriptions should be recorded in the data dictionary. Most of the CASE tools have this facility. To document Ada systems, Visibility Diagrams by Booch and Object-oriented Structured Design (OOSD) by Wasserman et al. are quite useful. The notations by Nielsen can also be used for this purpose. Some of these notations are implemented in the CASE tools described earlier. All the techniques used in our study are not available in any one CASE tool. Selection of a methodology and appropriate CASE tools are of vital importance as they dictate software development process.

Forward Engineering is done using an object-oriented approach. There are a number of Object-oriented analysis and design methodologies such as Object-oriented Design by Booch, Responsibility Driven Design by Wirfs-Brock et al., Object-oriented Analysis and Design by Rumbaugh et al., and by Coad and Yourdon. Some of the hybrid methodologies such as FUSION contain features of multiple object-oriented methods. A methodology which is consistent with the reverse engineering approach needs to be used for the new development. We have used the Object Modeling Technique (OMT) by Rumbaugh et al. for developing an object-oriented model. Leading CASE tools such as Software Through Pictures™ and team*work*™ incorporate this methodology.

One of the challenging tasks in representing an object-oriented design is to represent flow of control in the system. In traditional methodologies this is accomplished by the hierarchic diagrams such as Structure Charts. In the object-oriented paradigm the bottom-up approach is used in identifying objects and classes and the hierarchical approach (also bottom-up approach) is used in grouping classes into subsystems. Single inheritance is exercised in hierarchic fashion and multiple inheritance is similar to a network design. Control flows between classes are peer-to-peer collaborations. A set of techniques such as object-interaction diagrams, collaboration graphs, and state diagrams may be used for this purpose. The other major issue encountered in the Object Modeling Technique is the correspondence between the Object Model, Dynamic Model, and Functional Model. This issue may be addressed by following consistency in naming operations and attributes across the three models and cross references.

A systematic reengineering process will enable us to produce a reusable object-oriented system. Documents related to various aspects of the system would be generated during the process (not after-the-fact). The documents generated will address logical aspects. The system can be easily adapted to changing hardware and software configurations, and minimize future maintenance and modification costs and dependency on a specific vendor or contractor. Some of the subsystems such

as the Pilot/Vehicle Interface, Sensor Management, and ITARS are dependent on the type of aircraft, hardware, system software, and communication facilities. Others such as Guidance and Navigation are basically algorithmic and are useful in any environment. If the laboratory decides to reengineer the existing system the algorithm intensive subsystems may be tried first. Adhering to a particular methodology and selection of CASE tools are crucial in the success of such an effort.

General Comments

I had an opportunity to read a number of documents related to various systems during the summer research program. Based on my experience and observations I would like to offer some general comments:

1. Most of the documents provide an overview of the system. The logical and implementation aspects of the system are intermixed in these documents. They offer a good management overview of the system but do not contain the technical details needed As such, they do not offer a lot of help to the personnel who may want to modify the system.

2. Algorithms are crucial in any avionics systems. Documentation of these algorithms or references to the sources would be very beneficial. They may be represented using Structured English or pseudocode. Program Flow Charts should be avoided as they reflect source code. Action Diagrams can be used at the Structured English or pseudocode levels as they enforce indentation and highlight control structures. At present no documents are available which describe algorithms.

3. Adherence to a specific methodology and use of appropriate CASE tools will not only assist the development process, but will also help in future efforts.

REFERENCES

[BAR92] Barnhart, D., Benning, S., Blair, J., Bohler, M., Eldridge, B., Fanning, F.J., Goldman, P., Koenig, W., Morales, R., Powers, P., Rubertus, G., Schelling, E., Wilgus, J., Williams, D., Woodyard, J., PAVE PILLAR IN-HOUSE, Final Report, Systems Section, WL/AAAS-2, WPAFB, Report Number WL-TR-92-1128, October 1992.

[NIE92] Object-Oriented Design with Ada - Maximizing Reusability for Real-Time Systems, Nielsen, Kjell., Bantam Books, New York, 1992.

[WHA92] Whalen, Phillip W., Stoudt, Amy C., Miedler, Michael J., The Integrated Test Bed (ITB) System Integration Project, Final Report for the Period July 1988 to July 1992, TRW Avionics & Surveillance Group, Sponsored by the Avionics Directorate,WL/AAAS-2, Wright Laboratory, WPAFB, Report Number WL-TR-92-1119, September 1992.

Computer Program Product Specification for the Terrain-Aided Flight Algorithm (TFA) System Type C5 (STA0001), Released by Powers, Phillip H., Avionics Directorate, WL/AAAS-2, Wright Laboratory, WPAFB, August, 1991.

ADA Avionics Real-Time Software (AARTS) Program, Final Report for the Period March 1987 - October 1991, TRW Avionics & Surveillance Group, Sponsored by the Avionics Directorate,WL/AAAS-2, Wright Laboratory, WPAFB, Report Number WL-TR-91-1135, February 1992.

Gain Scheduled Missile Autopilot Design Using Observer-Based $\mathcal{H}_\infty$ Control

Guoxiang Gu
Assistant Professor
Department of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803-5901

Final Report for
AFOSR Summer Faculty Research Program
Wright Laboratory/Armament Directorate
Eglin Air Force Base, FL 32542

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

September 1994

Gain Scheduled Missile Autopilot Design Using Observer-Based $\mathcal{H}_\infty$ Control

Guoxiang Gu
Department of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803-5901

September 15, 1994

Abstract

This final report summarizes the research work performed at Wright Laboratory of Eglin Air Force Base by the author during the summer of 1994, sponsored by AFOSR Summer Faculty Research Program. It studies gain scheduled autopilot design in the pitch axis. To achieve the tracking specification with both stability and performance robustness for pitch autopilot design, frequency shaping is adopted and $\mathcal{H}_\infty$ optimization is employed. It is shown that with suitable modification of the design objective in [11], the resulting central $\mathcal{H}_\infty$ controller has an observer form and it can be computed without iteration. This is especially attractive for gain schedule implementation. Moreover, the proposed synthesis procedure has a two-step structure. In the first step, a state estimator gain is synthesized to achieve the desired sensitivity at the plant output. In the second step, a state feedback gain is synthesized to minimize the associated $\mathcal{H}_\infty$ cost of the objective function. This two-step procedure combines both advantages of the LQR/LTR in [16] and $\mathcal{H}_\infty$ loopshaping in [10] and is applicable to pitch autopilot design. The performance of a typical pitch autopilot synthesized by the proposed design method is then evaluated by computer simulations.

1 Introduction

Recently there has been some new developments for control of nonlinear systems using gain schedule approach [12, 13, 14, 7]. Various techniques are employed to formalize the gain schedule with the aim to establish unified framework in which both stability and performance of nonlinear feedback control systems can be studied. Notably, Shamma and Athans [14] convert the control of nonlinear systems into that of linear parameter varying (LPV) systems where slow varying parameters or exogenous signals are used as "scheduling variables". A similar approach is also used by Rugh [13] in conjunction with the local linearization but with different "trim conditions". Packard [12] approaches gain schedule from different perspective using structured singular value where uncertain parameters are chosen as scheduling variables. A common feature in [12, 13, 14] and their corresponding applications to pitch autopilot design in [15, 11] is that robust control of linear systems such as $\mathcal{H}_\infty$ and μ synthesis are employed to design a set of linear feedback controllers and each ensures stability and performance of the feedback system on a range of operating conditions or parameter variations. While the research work reported in [12, 13, 14] seeks new framework for gain scheduling, Hyde and Glover [7] adopts a more traditional approach focusing on applying $\mathcal{H}_\infty$ based loopshaping to schedule flight control systems. Because only a limited number of linear controllers can be stored, uniform structure of the controllers and their continuous dependence of the linearized plant models are emphasized in [7] in order to linearly interpolate the array of controllers smoothly. It is shown in [7, 10] that $\mathcal{H}_\infty$-based loopshaping results in observer-based controller that provides a natural structure of the feedback controller for which only two gains (state feedback and state estimator gains) need to be designed, stored, and scheduled.

In this final report, we study a particular problem arising from pitch autopilot design where the design objective is quite different from that of the $\mathcal{H}_\infty$-based loopshaping. This problem concerns the tracking performance that often dominates the design objective of the feedback control system as configurated in Figure 1. Roughly speaking, the objective is to synthesize feedback compensators $K_1(s)$ and $K_2(s)$ such that the error signals e_1 and e_2 are minimized in presence of the command input v_1 for a given linear plant $P(s)$. This type of design problems is quite common in flight control systems [1] including pitch autopilot design [11]. The classical wisdom converts such design problem into frequency domain where the tracking performance is transformed into the desired frequency shape of the transfer matrices from the command input v_1 to the tracking errors e_1 and e_2. The conventional procedure designs first $K_2(s)$ to achieve the inner-loop performance and then $K_1(s)$ to achieve the outer-loop performance. Although $\mathcal{H}_\infty$ technique as well as μ-synthesis have been used in [1, 11] to treat the design problem represented in Figure 1, the separate design of inner and outer

loops remains the same. While the performances of inner and outer loops are coupled, the effect of the separate design on the overall performance have not been investigated. Moreover the resulting controllers in [1, 11] are not guaranteed to have a uniform structure and thus may not be suitable for gain scheduled control of nonlinear systems.

Our contributions in this final report are as follows. First by appropriate formulation of the design problem in Figure 1, it is ahown that the resulting $\mathcal{H}_\infty$ controller has an observer structure. Indeed with suitable choice of the weighting function, the resulting $\mathcal{H}_\infty$ controller can be computed by solving two standard algebraic Riccati equations without iteration. This is especially attractive for gain schedule design. Second the proposed synthesis procedure has a two-step structure. In the first step, a state estimator gain is synthesized to achieve the tracking performance at the plant output with output injection. In the second step, a state feedback gain is synthesized to minimize the $\mathcal{H}_\infty$ cost of the objective function associated with tracking performance. This two-step procedure combines both advantages of the LQR/LTR in [16] and $\mathcal{H}_\infty$ loopshaping in [10]. Finally the proposed design procedure is applied to pitch autopilot design to show its effectiveness.

Finally the notation in this report is standard. The symbol $\mathbf{R}$ denotes the field of real numbers. For matrices, the symbols $\overline{\sigma}(\cdot)$ and $\underline{\sigma}(\cdot)$ denote maximum and minimum singular values respectively. A square matrix A is said to be stable, if all its eigenvalues are in the open left half plane. A transfer function matrix $T(s) = D + C(sI - A)^{-1}B$ is said to be internally stable, if A is stable, and in this case, its $\mathcal{H}_\infty$ norm is defined by

$$\|T\|_\infty := \sup_{w \in \mathbf{R}} \overline{\sigma}(T(j\omega)), \quad j = \sqrt{-1}.$$

The transpose and conjugate transpose of a matrix M are denoted by M^T and M^* respectively. The para-hermitian of the transfer matrix $T(s)$ is defined by $T^\sim = [T(-s)]^T$.

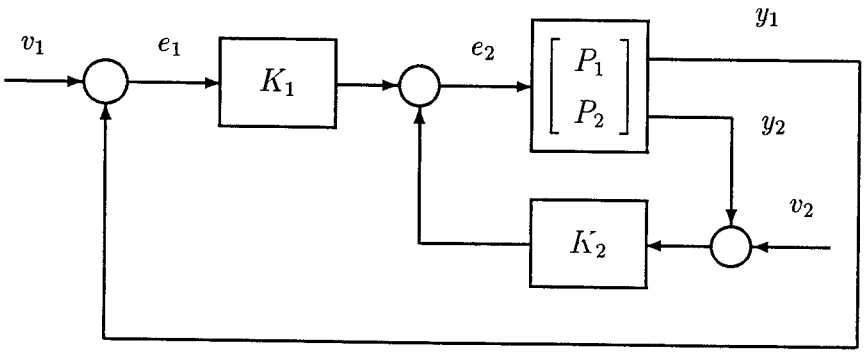


Figure 1: Two-loop feedback systems

2 Observer-based $\mathcal{H}_\infty$ Controller

Consider feedback system configurated in Figure 1 where the plant model is given by

$$P(s) = \begin{bmatrix} P_1(s) \\ P_2(s) \end{bmatrix} = C(sI - A)^{-1}B = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} (sI - A)^{-1} B \tag{1}$$

where (A, B) is stabilizable and (C, A) is detectable. It is assumed that $C_1 \in \mathbf{R}^{p_1 \times n}$, $C_2 \in \mathbf{R}^{p_2 \times n}$, and $B \in \mathbf{R}^{n \times m}$. We thus have that $v_1, y_1, e_1 \in \mathbf{R}^{p_1}$, $v_2, y_2 \in \mathbf{R}^{p_2}$, and $e_2 \in \mathbf{R}^m$. The design objective is to synthesize feedback compensators $K_1(s)$ and $K_2(s)$ such that the error signals e_1 and e_2 are small in presence of the command input v_1 within the operational frequency range. In order for y_1 to track v_1 at the outer loop, it is necessary that $p_1 \leq m$ and in many practical situations, $p_1 = m$. On the other hand, the actuating signal generated by $K_1(s)$ often represents a command input to y_2 that is different from y_1. This gives the rise of the inner loop design. The necessity for the two-loop design is also manifested by the possible two different time scales in plant models $P_1(s)$ and $P_2(s)$. A conventional approach designs first $K_2(s)$ such that e_2 is small and then $K_1(s)$ such that e_1 is small in presence of the command input v_1. To analyze this particular synthesis problem, the measurement noise v_2 is introduced. We consider the case where $p = p_1 + p_2 > m$ that is of engineering significance. Denote

$$K(s) = \begin{bmatrix} K_1(s) & K_2(s) \end{bmatrix}.$$

Then the sensitivity function at the plant output (that measures the tracking ability of the system to both v_1 and v_2) is given by

$$S_o(s) = (I_p - PK)^{-1} = I_p + P(I - KP)^{-1}K. \tag{2}$$

It follows that for $p > m$, $\sigma_{max}(S_o) \geq 1$ at any frequency $s = j\omega$. Hence simple $\mathcal{H}_\infty$ design based on frequency shaping of the sensitivity $S_o(s)$ does not work for the case $p > m$.

Denote $e^T = \begin{bmatrix} e_1^T & e_2^T \end{bmatrix}$ and $v^T = \begin{bmatrix} v_1^T & v_2^T \end{bmatrix}$. The transfer matrix from v to e is given by

$$T_{ev}(s) = \tilde{K}(I - PK)^{-1}, \quad \tilde{K} = \begin{bmatrix} I_{p_1} & 0 \\ K_1 & K_2 \end{bmatrix}.$$

It should be clear that the frequency shape of the command signal can be obtained from the tracking specification. In particular, it can be assumed that v is driven by the white noise η through filter or weighting function $W(s)$:

$$v = W\eta, \quad W = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \Longrightarrow v_1 = W_1\eta, \quad v_2 = W_2\eta \tag{3}$$

where $W_1(s)$ is square. Because our design objective is to synthesize $K(s)$ such that the error signals are minimized in presence of the command signal v_1, $W_2(j\omega)$ should be small (or zero ideally) in comparison with $W_1(j\omega)$ over all frequency. Therefore we are led to study the following $\mathcal{H}_\infty$ optimization problem:

P1: Find a stabilizing controller $K(s)$ such that $\gamma = \|\tilde{K}\,(I_p - PK)^{-1}\,W\|_\infty$ is minimized.

For gain schedule purpose, it is preferred that the resulting $\mathcal{H}_\infty$ controller be observer. It is also preferred that the corresponding $\mathcal{H}_\infty$ controller be synthesized without iteration. These two objectives can be indeed achieved.

Remark 2.1 *Denote* $T_{e\eta} = \tilde{K}\,(I_p - PK)^{-1}\,W$ *and Partition* $T_{e\eta}$ *conformally with that of* $W(s)$. *Then we have*

$$T_{e\eta} = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix},\ T_1 = \begin{bmatrix} I_{p_1} & 0 \end{bmatrix}(I_p - PK)^{-1}\,W.\ T_2 = K(I_p - PK)^{-1}W.$$

Although the problem **P1** *does not take plant uncertainty into consideration, it should be clear that because* $\|T_{e\eta}\|_\infty < \gamma$ *implies that* $\|K(I_p - PK)^{-1}W\|_\infty < \gamma$, *the closed-loop system is robustly stable for all possible additive perturbations of the form*

$$P_\Delta(s) = P(s) + W(s)\Delta(s),\ \ \|\Delta\|_\infty \le \gamma^{-1} \tag{4}$$

provided that $P_\Delta(s)$ *and* $P(s)$ *have the same number of unstable poles [5].*

Several issues are involved with the $\mathcal{H}_\infty$ design problem **P1**. We will discuss each of them next.

A. Synthesis of the Weighting Function

Using the framework in [4], the transfer matrix $T_{e\eta}(s)$ in **P1** can be written into the following linear fractional form:

$$T_{e\eta} = \mathcal{F}_\ell(G, K),\ \ G = \left[\begin{array}{c|c} W_1 & P_1 \\ 0 & I_{p_2} \\ \hline W & P \end{array}\right]. \tag{5}$$

See the block diagram in Figure 2(a). Because the weighting function $W(s)$ in (3) is tall in general, the $\mathcal{H}_\infty$ optimization problem **P1** is not covered in regular $\mathcal{H}_\infty$ control and thus the state space solution in [4, 6] is not readily applicable to problem **P1**. Moreover if $W(s)$ has a different "A" matrix from that of $P(s)$, then the McMillan degree of the generalized plant $G(s)$, thus that of the $\mathcal{H}_\infty$ controller $K(s)$, will increase. This is also true for the case $W_2 = 0$ even if $W(s)$ has the same

"A" matrix as $P(s)$. Finally the tracking performance requires that $\underline{\sigma}(W_1(j\omega))$ be much larger than one in the low (operational) frequency bandwidth, and that $\overline{\sigma}(W_1(j\omega))$ be close to one in the high frequency bandwidth, while keeping $\overline{\sigma}(W_2(j\omega))$ small relative to the value of $\underline{\sigma}(W_1(j\omega))$. The above considerations impose severe restrictions on the weighting functions.

In order to obtain an observer-based $\mathcal{H}_\infty$ controller we adopt a similar approach to the loopshaping in [9, 10] in the synthesis of weighting function $W(s)$. We assume that the plant model $P(s)$ in (1) is obtained from original plant $P_o(s)$ through appropriate shaping:

$$P(s) = \Phi_1(s)P_o(s)\Phi_2(s)$$

where Φ_1 and Φ_2 are square transfer matrices. The corresponding feedback controller $K_o(s)$ for $P_o(s)$ is then given by $K_o(s) = \Phi_2(s)K(s)\Phi_1(s)$. See [10, 7] for more details. It should be clear that the design specification on the steady-state response can be met by choosing Φ_1 and Φ_2 appropriately. In particular, in order for $y_1(t)$, the output of the plant $P_1(s)$, to track step signal of the command input $v_1(t)$ with zero steady-state error, Φ_1 can be chosen such that $P_1(s)$ has an integrator in its each output channel. Furthermore $P(s)$ can be made diagonally dominant with suitable choice of Φ_1 and Φ_2 which implies that in our problem **P1**, the square plant $P_1(s)$ is almost diagonal and $P_2(s)$ is significantly smaller than $P_1(s)$ in terms of the frequency shape. A simple way to achieve this is to choose $\Phi_1 = \text{diag}\,(\Phi_{11}, \Phi_{12})$ such that the resulting C_2 is significantly smaller than C_1. The purpose of Φ_1 and Φ_2 is to shape the plant $P(s)$ such that the weighting function $W(s)$ represents the tracking specification. In fact by internal model principle, Φ_1 and Φ_2 should be chosen such that the shaped plant $P_1(s)$ resembles $W_1(s)$. This results in two simple methods for the synthesis of the weighting functions.

The first method chooses $W(s)$ such that

$$W_1 W_1^{\sim} = I + P_1 P_1^{\sim} \tag{6}$$

and $W_2(s)$ is significantly smaller than $W_1(s)$ in terms of the frequency shape. The next result describes the procedure of synthesizing such a weighting function $W(s)$.

Proposition 2.2 *Let (A, B, C) be the realization of shaped plant model $P(s)$ such that (A, B) is stabilizable and (C_1, A) is detectable. Let Y be the stabilizing solution of the following algebraic Riccati equation (ARE):*

$$AY + YA^T - YC_1^T C_1 Y + BB^T = 0. \tag{7}$$

Then with

$$W(s) = \begin{bmatrix} I \\ 0 \end{bmatrix} - \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} (sI - A)^{-1} H_1, \quad H_1 = -YC_1^T,$$

relation (6) holds and $W_1(s)$ dominates $W_2(s)$ approximately.

Proof: By the stabilizability and detectability of (A, B, C_1), there exists a normalized left coprime factorization $P_1 = \tilde{D}_1^{-1}\tilde{N}_1$ such that $\tilde{D}_1\tilde{D}_1^{\sim} + \tilde{N}_1\tilde{N}_1^{\sim} = I_{p_1}$. In light of [8], a state space realization is given by

$$\begin{bmatrix} \tilde{D}_1 & \tilde{N}_1 \end{bmatrix} = \begin{bmatrix} I_p & 0 \end{bmatrix} + C(sI - A - LC)^{-1}\begin{bmatrix} H_1 & B \end{bmatrix}$$

where $H_1 = -YC_1^T$ and Y is the stabilizing solution of (7). By taking $W_1 = \tilde{D}_1^{-1}$, it is easy to verify that (6) holds. Because C_1 dominates C_2, $P_1(s)$ dominates $P_2(s)$, and $W_1(s)$ resembles $P_1(s)$ in terms of frequency shape, it follows that $W_1(s)$ dominates $W_2(s)$ approximately. ∎

In Proposition 2.2, the detectability of (C_1, A) is required which may not hold while the detectability of (C, A) is a standard assumption. Hence an alternative is sought next. Our second method to the synthesis of $W(s)$ aims to satisfy

$$W_1W_1^{\sim} \approx I + P_1P_1^{\sim}, \quad W_2W_1^{\sim} \approx P_2P_1^{\sim} \tag{8}$$

and thus $W_1(s)$ resembles the frequency shape of $P_1(s)$ and $W_2(s)$ resembles that of $P_2(s)$. We have a similar procedure to synthesize such a weighting function $W(s)$.

Proposition 2.3 *Let the shaped plant model $P(s)$ be as in (1) where (A, B, C) is stabilizable and detectable. Let Y be the stabilizing solution of the following ARE:*

$$AY + YA^T - YC^TCY + BB^T = 0, \quad C^TC = C_1^TC_1 + C_2^TC_2. \tag{9}$$

Then with

$$W(s) = \begin{bmatrix} I \\ 0 \end{bmatrix} - \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}(sI - A)^{-1}H_1, \quad H_1 = -YC_1^T,$$

the approximate relation in (8) holds.

Proof: Since (A, B, C) is stablizable and detectable, there exists a left normalized coprime factorization $P(s) = \tilde{D}^{-1}\tilde{N}$ such that $\tilde{D}$ and $\tilde{N}$ are stable and $\tilde{N}\tilde{N}^{\sim} + \tilde{D}\tilde{D}^{\sim} = I_p$. In light of [8], we have that

$$\begin{bmatrix} \tilde{D} & \tilde{N} \end{bmatrix} = \begin{bmatrix} I_p & 0 \end{bmatrix} + C(sI - A - LC)^{-1}\begin{bmatrix} H & B \end{bmatrix}$$

where $H = \begin{bmatrix} H_1 & H_2 \end{bmatrix} = -YC^T = -Y\begin{bmatrix} C_1 & C_2 \end{bmatrix}^T$ and Y is the stabilizing solution of (9). There thus holds

$$I + PP^{\sim} = \left(\tilde{D}^{\sim}\tilde{D}\right)^{-1} = \tilde{W}\tilde{W}^{\sim} \tag{10}$$

which implies that

$$\tilde{W} = \begin{bmatrix} \tilde{W}_{11} & \tilde{W}_{12} \\ \tilde{W}_{21} & \tilde{W}_{22} \end{bmatrix} = \tilde{D}^{-1} = \begin{bmatrix} I_{p_1} & 0 \\ 0 & I_{p_2} \end{bmatrix} - \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} (sI - A)^{-1} \begin{bmatrix} H_1 & H_2 \end{bmatrix}.$$

Since C_1 dominates C_2, $H_1 = -YC_1^T$ dominates $H_2 = -YC_2^T$. It follows that $\tilde{W}_{11}(s)$ dominates $\tilde{W}_{ik}(s)$ for $i,k = 1,2$ and $ik \neq 1$. Hence with $W_1 = \tilde{W}_{11}$ and $W_2 = \tilde{W}_{21}$, relation (10) yields

$$I + P_1 P_1^{\sim} = W_1 W_1^{\sim} + \tilde{W}_{12}\tilde{W}_{12}^{\sim} \approx W_1 W_1^{\sim}, \;\; P_1 P_2^{\sim} = W_1 W_2^{\sim} + \tilde{W}_{12}\tilde{W}_{22}^{\sim} \approx W_1 W_2^{\sim}. \qquad \blacksquare$$

It is noted that the two methods for the synthesis of $W(s)$ are quite similar and in fact result in similar H_1 provided that C_2 is significantly smaller than C_1. Moreover the corresponding synthesis procedure involves solving a standard Kalman-Bucy filtering ARE that is a routine in many software packages such as MATLAB. We would like to comment that in shaping plant $P(s)$, it is more appropriate to employ dynamic Φ_1 to achieve the dominance of $P_1(s)$ over $P_2(s)$ rather than purely scaling C_1 and C_2. For example, by using integrator in Φ_{11}, $P_1(s)$ can be made dominant at low frequency range. Because $H = -YC_1^T$ where $Y \geq 0$, it follows that $W_1(s) \approx I_{p_1} + C_1 Y C_1^T/s$ dominates $W_2(s) \approx C_2 Y C_1^T/s$ at high frequency range. Thus the objective in shaping the plant can be achieved without changing $P_2(s)$ significantly. Recall that the output of the plant $P_2(s)$ is employed to improve the tracking performance of the feedback system in Figure 1.

B. Observer Form of the Central $\mathcal{H}_\infty$ Controller

With the weighting function $W(s)$ synthesized from Proposition 2.2 and 2.3, we will show that the central $\mathcal{H}_\infty$ controller for problem **P1** has an observer form. It involves solving one ARE of $\mathcal{H}_\infty$ type that can in turn be converted into a Lyapunov type equation or an LQR type ARE. Using the same notation as in [4], a state space realization of the generalized plant $G(s)$ in Figure 2(a) is given by

$$G(s) = \left[\begin{array}{c|c} \begin{matrix} I_{p_1} & 0 \\ 0 & I_{p_2} \end{matrix} & \begin{matrix} 0 \\ I_{p_2} \end{matrix} \\ \hline \begin{matrix} I_{p_1} & 0 \\ 0 & 0 \end{matrix} & \begin{matrix} 0 \\ 0 \end{matrix} \end{array}\right] + \left[\begin{array}{c} \begin{matrix} C_1 \\ 0 \end{matrix} \\ \hline \begin{matrix} 0 \\ C_2 \end{matrix} \end{array}\right] (sI_n - A)^{-1} \left[\begin{array}{c|c} -H_1 & B \end{array}\right]. \tag{11}$$

An observer-based stabilizing controller has the form

$$K_{ob}(s) = -F(sI_n - A - LC - BF)^{-1} L \tag{12}$$

where $A + LC$ and $A + BF$ are both stable. The constant matrices F and L are called state feedback and state estimator gain respectively. The following result converts the $\mathcal{H}_\infty$ problem **P1** into an equivalent $\mathcal{H}_\infty$ state feedback problem.

Theorem 2.4 *Let $T_{e\eta}$ be as in (5) where $W(s)$ is synthesized from either Proposition 2.2 or 2.3. Assume that (A,B) is stabilizable and (C,A) is detectable. Then there exists a stabilizing controller $K(s)$ such that $\|T_{e\eta}\|_\infty < \gamma$, if and only if there exists a stabilizing state feedback gain F such that $\|T_F\|_\infty < \gamma$ where*

$$T_F = \begin{bmatrix} I_{p_1} \\ 0 \end{bmatrix} - \begin{bmatrix} C_1 \\ F \end{bmatrix} (sI_n - A - BF)^{-1} H_1. \tag{13}$$

Proof: For the feedback system in (5), there exists an observer-based stabilizing controller (12). Indeed it is easy to see that with

$$L = \begin{bmatrix} L_1 & L_2 \end{bmatrix} = \begin{bmatrix} H_1 & H_2 \end{bmatrix}, \tag{14}$$

and $H_1 = -YC_1^T$ where Y is the stabilizing solution of (7) or (9), there exists an H_2 such that $A + LC$ is stable. Now all stabilizing controllers can be parameterized by $K(s) = \mathcal{F}_\ell(J,Q)$ where

$$J(s) = \left[\begin{array}{c|c} 0 & I_m \\ \hline I_p & 0 \end{array}\right] + \left[\begin{array}{c} F \\ \hline -C \end{array}\right] (sI_n - A - BF - LC)^{-1} \left[\begin{array}{c|c} -L & B \end{array}\right] \tag{15}$$

and $Q \in \mathcal{H}_\infty$ [3]. With the parameterization of all stabilizing controllers, the block diagram in Figure 2(a) has the form of Figure 2(b). After suitable state space coordinate transformation and eliminating the uncontrollable subsystem corresponding to $A + LC$, Figure 2(b) is equivalent to Figure 3(a) where

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \left[\begin{array}{c|c} I_{p_1} & 0 \\ 0 & I_{p_2} \\ \hline I_{p_1} & 0 \\ 0 & 0 \end{array}\right] + \left[\begin{array}{c} C_1 \\ F \\ \hline 0 \\ 0 \end{array}\right] (sI_n - A - BF)^{-1} \left[\begin{array}{c|c} -H_1 & B \end{array}\right].$$

Then simple calculation shows that $T_{e\eta} = \mathcal{F}_\ell(G, \mathcal{F}_\ell(J,Q)) = \mathcal{F}_\ell(M,Q) = T_F + M_{12}QM_{21}$ due to

$$M_{11} = T_F, \quad M_{12} = \begin{bmatrix} 0 \\ I_{p_2} \end{bmatrix} + \begin{bmatrix} C_1 \\ F \end{bmatrix} (sI_n - A - BF)^{-1} B, \quad M_{21} = \begin{bmatrix} I_{p_1} \\ 0 \end{bmatrix}, \quad M_{22} = 0.$$

It follows that $T_{e\eta} = T_F + M_{12}\tilde{Q} = \mathcal{F}_\ell(\tilde{M}, \tilde{Q})$ where

$$\tilde{M} = \left[\begin{array}{c|c} I_{p_1} & 0 \\ 0 & I_{p_2} \\ \hline I_{p_1} & 0 \end{array}\right] + \left[\begin{array}{c} C_1 \\ F \\ \hline 0 \end{array}\right] (sI_n - A - BF)^{-1} \left[\begin{array}{c|c} -H_1 & B \end{array}\right], \quad \tilde{Q} = Q \begin{bmatrix} I \\ 0 \end{bmatrix}. \tag{16}$$

By using suitable variable substitution, Figure 3(a) is then equivalent to the *full information* problem [4] in Figure 3(b) where

$$\tilde{M}_{FI} = \left[\begin{array}{c|c} I_{p_1} & 0 \\ 0 & I_{p_2} \\ \hline 0 & 0 \\ I_{p_1} & 0 \end{array}\right] + \left[\begin{array}{c} C_1 \\ 0 \\ \hline I_n \\ 0 \end{array}\right] (sI_n - A)^{-1} \left[\begin{array}{c|c} -H_1 & B \end{array}\right].$$

The corresponding result from [4] can then be used to conclude that there exists a stabilizing controller such that $\|T_{e\eta}\|_\infty < \gamma$ if and only if there exists a stabilizing state feedback gain F such that $\|T_F\|_\infty < \gamma$. ∎

Corollary 2.5 *The central controller for problem* **P1** *has an observer form as in (12) where the state estimator gain L can be computed from $L = -YC^T$ and Y is the stabilizing solution of filtering ARE (7) or (9), and the state feedback gain F can be computed from minimization of $\|T_F\|_\infty$.*

Proof: By Theorem 2.4 we need only to show that $A + LC$ is stable where $L = -YC^T$. If Y is the stabilizing solution of (9), then the stability of $A + LC$ follows trivially. On the other hand, if Y is the stabilizing solution of (7), then the hypothesis of Proposition 2.2 implies that $A + L_1C_1$ is stable with $L_1 = -YC_1^T$. Adding and subtracting $YC_2^TC_2Y$ in (7) yields

$$(A + LC)Y + Y(A + LC)^T + LL^T + L_2L_2^T + BB^T = 0.$$

Hence a simple application of Lyapunov stability theorem concludes that $A + LC$ is indeed a stable matrix. ∎

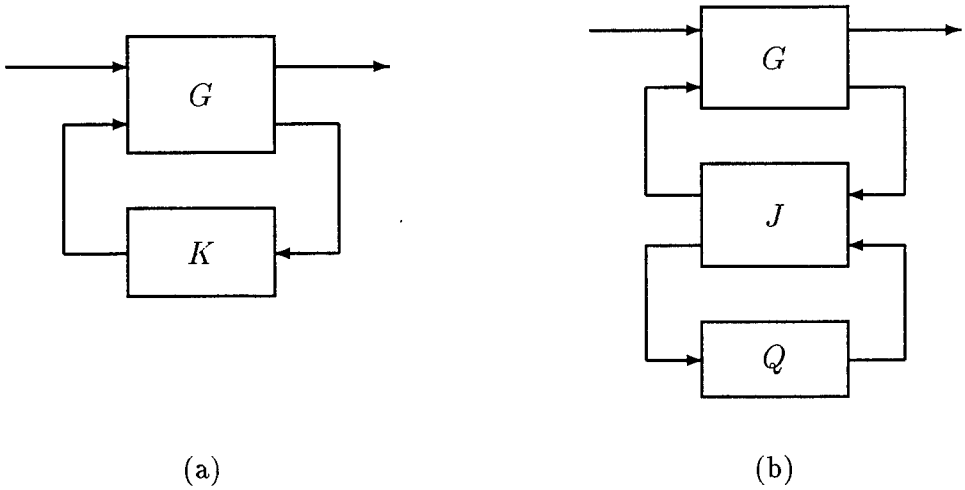
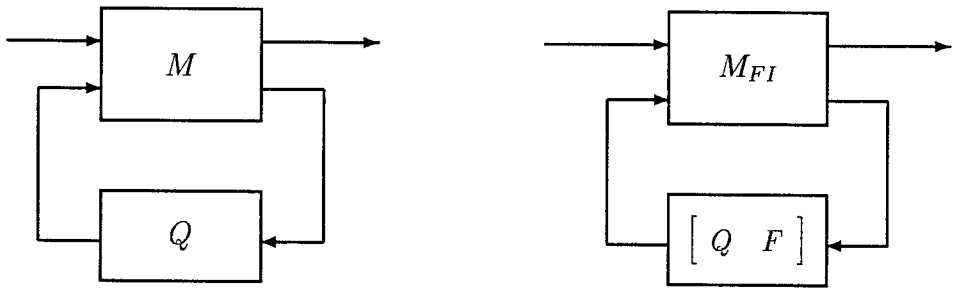


Figure 2: Linear fractional feedback systems



(a) (b)

Figure 3: Equivalence to full information problem

C. Computation of the $\mathcal{H}_\infty$ Controller

Although it has been shown that the central $\mathcal{H}_\infty$ controller has an observer form, it remains unclear for the optimal value

$$\gamma^* := \inf\{\|T_{e\eta}\|_\infty : \ K \text{ stabilizing}\} = \inf\{\|T_F\|_\infty : \ F \text{ stabilizing}\} \tag{17}$$

as well as the computation of the central $\mathcal{H}_\infty$ controller. It turns out that for the case $L = -YC^T$ with Y stabilizing solution of (7), the optimal value of γ^* can be computed without iteration and for each $\gamma > \gamma^*$, the suboptimal state feedback gain F such that $\|T_F\|_\infty < \gamma$ can be computed from a standard Lyapunov equation.

Theorem 2.6 *Suppose (C_1, A) is detectable and $L_1 = H_1 = -YC_1^T$ where Y is the stabilizing solution of (7). Let X be the solution of*

$$(A + L_1C_1)^T X + X(A + L_1C_1) + C_1C_1^T = 0. \tag{18}$$

Then $X \geq 0$ and $\gamma^* = 1/\sqrt{1 - \rho(YX)}$.

Proof: By the state space formulae in [6], there exists a stabilizing F such that $\|T_F\|_\infty < \gamma$ for any $\gamma > \gamma^*$ if and only if there exists a stabilizng solution $X_\infty \geq 0$ for the $\mathcal{H}_\infty$ ARE:

$$\left(A + \alpha YC_1^TC_1\right)^T X_\infty + X_\infty\left(A + \alpha YC_1^TC_1\right) - X_\infty\left(BB^T - \alpha YC_1^TC_1Y\right)X_\infty + (1+\alpha)C_1C_1^T = 0$$

where $\alpha = (\gamma^2 - 1)^{-1}$. The above ARE can be written as

$$0 = \begin{bmatrix} X_\infty & -I_n \end{bmatrix} H_\infty \begin{bmatrix} I_n \\ X_\infty \end{bmatrix}, \quad H_\infty = \begin{bmatrix} A + \alpha YC_1^TC_1 & -BB^T + \alpha YC_1^TC_1Y \\ -(1+\alpha)C_1^TC_1 & -(A + \alpha YC_1^TC_1)^T \end{bmatrix}. \tag{19}$$

Denote similarity transform matrix

$$S = \begin{bmatrix} I_n & -(1+\alpha)Y \\ 0 & (1+\alpha)I_n \end{bmatrix}, \quad S^{-1} = \begin{bmatrix} I_n & Y \\ 0 & (1+\alpha)^{-1}I_n \end{bmatrix}.$$

Then simple matrix manipulation gives

$$H_o = S^{-1}H_\infty S = \begin{bmatrix} A - YC_1^TC_1 & 0 \\ -C_1^TC_1 & -(A - YC_1^TC_1)^T \end{bmatrix} \tag{20}$$

where (7) is used to obtain the zero block in the (1,2) position of H_o. Hence with X the solution of (18), we have that

$$\begin{bmatrix} X & -I_n \end{bmatrix} H_o \begin{bmatrix} I_n \\ X \end{bmatrix} = 0.$$

It follows that

$$\begin{bmatrix} \beta X(I_n - \beta YX)^{-1} & -I_n \end{bmatrix} H_\infty \begin{bmatrix} I_n \\ \beta X(I_n - \beta YX)^{-1} \end{bmatrix} = 0$$

and thus $X_\infty = \beta X(I_n - \beta YX)^{-1}$ where $\beta = 1 + \alpha = \gamma^2(\gamma^2 - 1)^{-1}$. By the fact that $A + L_1C_1$ is stable, $X \geq 0$. Moreover as $s \to \infty$, $\overline{\sigma}(T_F(s)) = 1$. The optimal value $\gamma^* \geq 1$ that implies $\beta \geq 0$. Hence $X_\infty \geq 0$ if and only if $\rho(YX) < \beta^{-1} = 1 - \gamma^{-2}$ that in turn implies that the optimal value γ^* can be computed from $\gamma^* = 1/\sqrt{1 - \rho(YX)}$ for which X_∞ ceases to exist. ∎

Remark 2.7 *The suboptimal state feedback gain F for any $\gamma > \gamma^*$ can be computed from $F = -B^TX_\infty = -B^T\beta X(I_n - \beta YX)^{-1}$ where $\beta = \gamma^2(\gamma^2 - 1)^{-1}$ and X, Y are computed from (18) and (7) respectively. It is interesting to see that*

$$\rho(YX) = \left\| \begin{bmatrix} \tilde{D}_1 & \tilde{N}_1 \end{bmatrix} \right\|_H^2$$

where $\|\cdot\|_H$ is the Hankel norm and $\tilde{D}_1$, $\tilde{N}_1$ are same as in the proof of Proposition 2.2. In comparison with the results in [9, 10], we conclude that the $\mathcal{H}_\infty$ solution for problem **P1** *is similar to the $\mathcal{H}_\infty$-based loopshaping in [9, 10] although our problem* **P1** *is quite different from that of [9, 10]. A consequence is that the $\mathcal{H}_\infty$ solution for* **P1** *has a loopshaping interpretation and the resulting $\mathcal{H}_\infty$ design for* **P1** *admits those stability and performance robustness highlighted in [9, 10].*

For the case that (C_1, A) is not detectable, the state estimator gain needs to be computed from $L = -YC^T$ where Y is the stabilizing solution of (9). In this case the optimal value of γ^* has to be computed iteratively. However comparing to the general case in [4, 6] where three conditions govern the optimal value γ^* that involves solving two $\mathcal{H}_\infty$ type AREs and computing one spectral radius

in each iteration, we have a much simpler computation in searching the optimal value γ^* that is governed by only one condition involves solving only one $\mathcal{H}_\infty$ type ARE (19) in each iteration.

Before concluding this section, we note that in computation of $W(s)$ and the solution that $\|T_F\|_\infty < \gamma$, $L_2 = H_2 = -YC_2^T$ is never used. It implies that any L_2 will give an observer-based controller $K(s)$ that solves **P1** provided that $A + LC = A + L_1C_1 + L_2C_2$ is stable. Therefore we may use this free parameter L_2 to further improve the performance of the system. In particular, robustness at the plant input can be improved by synthesizing an appropriate output injection gain L_2. For instance, we may use L_2 to achieve a better sensitivity of the closed-loop system at the plant input that is different from the sensitivity at the plant output in (2). With the observer controller, the sensitivity at the plant input is given by

$$S_i(s) = (I_m - K(s)P(s))^{-1} = \left(I_m + F(sI - A - BF)^{-1}B\right)\left(I_m - F(sI - A - L_1C_1 - L_2C_2)^{-1}B\right).$$

Because $S_F(s) = I_m + F(sI - A - BF)^{-1}B$ represents the sensitivity at the plant input with state feedback and is fixed after the synthesis of the $\mathcal{H}_\infty$ controller, one may consider $\mathcal{H}_\infty$ minimization of the relative error $\|E_{rel}\|_\infty$ where

$$E_{rel} = I - S_F^{-1}S_i = F(sI - \overline{A} - L_2C_2)^{-1}B, \quad \overline{A} = A + L_1C_1. \tag{21}$$

This is a special singular $\mathcal{H}_\infty$ problem and is related to the loop transfer recovery. In light of [2], the optimal solution L_2 which stabilizes $A + LC$ and minimizes $\|E_{rel}\|_\infty$ can also be computed without iteration.

D. Proposed Algorithm for Synthesis of Observer-based Controller

We summarize our proposed synthesis procedure as in the following algorithm.

- Step 1: Given original plant $P_o(s)$ and tracking performance requirements, synthesize $\Phi_1(s)$ and $\Phi_2(s)$ to obtain shaped plant $P(s) = \Phi_1(s)P_o(s)\Phi_2(s) = \begin{bmatrix} P_1(s) \\ P_2(s) \end{bmatrix}$ such that $P_1(s)$ has desired frequency shape as required by tracking performance that dominates the frequency shape of $P_2(s)$.

- Step 2: If (C_1, A) is detectable, compute $L_1 = -YC_1^T$ where $Y \geq 0$ is solved from ARE (7), and compute $F = -B^T\beta X(I_n - \beta YX)^{-1}$, $\beta = \gamma^2(\gamma^2 - 1)^{-1}$, where $X \geq 0$ is solved from (18) and $\gamma > \gamma^* = 1/\sqrt{1 - \rho(YX)}$.

  Otherwise, compute $L_1 = -YC_1^T$ where $Y \geq 0$ is solved from ARE (9), and compute $F = -B^TX_\infty$ where X_∞ is computed iteratively from ARE (19).

- Step 3: Compute L_2 through $\mathcal{H}_\infty$ minimization of $\|E_{rel}\|_\infty$ as in (21) and set $L = \begin{bmatrix} L_1 & L_2 \end{bmatrix}$.
- Step 4: Set the feedback controller as $K_o(s) = -\Phi_2 F(sI - A - LC - BF)^{-1} L\Phi_1(s)$.

3 An Application to Pitch Autopilot Design

This section studies pitch autopilot design using the $\mathcal{H}_\infty$ synthesis procedure in conjunction with gain scheduling. The pitch-axis model involves angle of attack and pitch rates as described by [11]:

$$\begin{aligned} \dot{\alpha}(t) &= fK_\alpha M(t) C_n[\alpha(t), \delta(t), M(t)] \cos(\alpha(t)/f) + q(t), \\ \dot{q}(t) &= fK_q M^2(t) C_m[\alpha(t), \delta(t), M(t)] \end{aligned}$$

where the aerodynamic coefficients are given by

$$\begin{aligned} C_n[\alpha(t), \delta(t), M(t)] &= \alpha\left(a_n\alpha^2 + b_n|\alpha| + c_n(2 - M/3)\right) + d_n\delta, \\ C_m[\alpha(t), \delta(t), M(t)] &= \alpha\left(a_m\alpha^2 + b_m|\alpha| + c_m(-7 + 8M/3)\right) + d_m\delta. \end{aligned}$$

The controlled output is normal acceleration $\eta(t) = K_z M^2(t) C_n[\alpha(t), \delta(t), M(t)]$. By standard notation, $\alpha(t)$ is the angle of attack in degrees, $q(t)$ is the pitch rate in degrees per second, and $M(t)$ is the Mach number. The variable $\eta_c(t)$ and $\eta(t)$ are commanded and actual normal acceleration in g's respectively. The actuator dynamics is the tail deflection $\delta(t)$ driven by commanded tail deflection $\delta_c(t)$ both in degrees described by $\ddot{\delta}(t) + 2\zeta\omega_a\dot{\delta}(t) + \omega_a^2\delta(t) = \omega_a^2\delta_c(t)$. The following table summarizes the details of the pitch-axis missile model [11]:

K_α	=	$0.7P_oS/(mv_s)$	$K_q = 0.7P_oSd/I_y$
K_z	=	$0.7P_oS/(32.2m)$	$f = 180^o/\pi$
P_o	=	973.3 lbs/ft^2	static pressure at 20000 ft
S	=	0.44 ft^2	surface area
m	=	13.98 slugs	mass
v_s	=	1036.4 ft/s	speed of sound at 20000 ft
d	=	0.75 ft	diameter
I_y	=	182.5 slug $\cdot$ ft^2	pitch moment of inertia
ζ	=	0.707	actuator damping ratio
ω_a	=	150 rad/s	actuator undamped natural frequency
a_n	=	0.000103 deg^{-3}	$a_m = 0.000215$ deg^{-3}
b_n	=	-0.00945 deg^{-2}	$b_m = -0.0195$ deg^{-2}
c_n	=	-0.1696 deg^{-1}	$c_m = 0.051$ deg^{-1}
d_n	=	-0.034 deg^{-1}	$d_m = -0.206$ deg^{-1}

The performance goals for the closed-loop system are described as follows (see [11]):

- Maintain robust stability in presence of $\pm 25\%$ parameter uncertainty in the aerodynamic coefficients $C_m[\alpha, \delta, M]$ and over the operating range specified by $-20^o \leq \alpha(t) \leq 20^o$ and $1.5 \leq M(t) \leq 3$.
- Track step commands in $\eta_c(t)$ with time constant no greater than 0.35 s, maximum overshot no greater than 10%, and with zero steady-state error.
- Maintain at least 40 dB attenuation for the open-loop transfer function at the input to the actuator that ensures that the autopilot avoids exciting unmodeled structural dynamics.
- Maximum tail deflection rate for 1 g step command in $\eta_c(t)$ should not exceed 25 deg/s.

It is noted that the above specifications are more stringent than those studied in [11].

The nonlinear plant to be controlled includes actuator dynamics. The input to the plant is the commanded tail deflection δ_c and the measured outputs are $\eta(t)$ and $q(t)$. Our first step is to obtain a linearized model at constant α^o and M^o that is routine. See [11]. The zero steady-state tracking error implies that an integrator needs to be employed in shaping the plant P_1. We are thus led to the following shaped plant

$$P(s) = \text{diag}\,(\frac{w_1}{s}, w_2)P_o(s) = \begin{bmatrix} P_1(s) \\ P_2(s) \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} (sI - A)^{-1}B$$

where $P_o(s)$ is the linearized plant and w_1 and w_2 are the two constants used to synthesize the weighting functions. Since $\eta(t)$ is also the controlled output, this corresponds to the two-loop feedback system in Figure 1 where P_1 is the shaped plant with output $\eta(t)$ and P_2 the shaped plant with output $q(t)$. We note that the plant $P(s)$ has McMillan degree 5 by taking actuator and integrator into account. Its realization has a controller form. In what follows next, we consider linear controller design at $M = 2.25$ and $\alpha = 10^o$ to illustrate our proposed synthesis procedure.

Because (C_1, A) is observable, ARE (7) is employed to compute the weighting function. After a few trials, we found that with $(w_1, w_2) = (20, 0.0008)$ the frequency shape of $W_1(s)$ dominates that of $W_2(s)$ as shown in Figure 4 Using Theorem 2.6, the corresponding optimal value is found to be $\gamma^* = 3.2894$. An application of Step 2 of the algorithm yields the required state feedback and estimator gains

$$F = \begin{bmatrix} -1.803 \times 10^1 & -4.001 \times 10^3 & -4.447 \times 10^5 & -4.319 \times 10^6 & -4.222 \times 10^7 \end{bmatrix},$$
$$L_1 = \begin{bmatrix} -7.2 \times 10^{-3} & 4.508 \times 10^{-4} & 1.7947 \times 10^{-5} & -3.1304 \times 10^{-6} & -4.4344 \times 10^{-7} \end{bmatrix}.$$

To obtain L_2, we set $\gamma = 8$ and $\epsilon = 7.2 \times 10^{-6}$ to compute the stabilizing solution of the following ARE:

$$\overline{A}Y + Y\overline{A}^T - Y\left(C_2^T C_2 - \frac{\epsilon}{\gamma^2}F^T F\right)Y + \frac{BB^T}{\epsilon} = 0, \quad \overline{A} = A + L_1 C_1, \tag{22}$$

and set $L_2 = -YC_2^T = \begin{bmatrix} -8.007 \times 10^3 & -1.965 \times 10^1 & -1.127 \times 10^{-1} & 5.4 \times 10^{-3} & 2.944 \times 10^{-4} \end{bmatrix}$. The above is resulted by converting singular $\mathcal{H}_\infty$ control into a regular one and by noting that

$$\begin{aligned} \|E_{rel}\|_\infty < \gamma \quad &\Longrightarrow \quad \overline{\sigma}\left(F(sI - \overline{A} - L_2 C_2)^{-1}\begin{bmatrix} B & \sqrt{\epsilon}L_2 \end{bmatrix}\right) < \gamma \ \ \forall s \in j\mathbf{R} \\ &\Longleftrightarrow \quad \overline{\sigma}\left(\frac{\sqrt{\epsilon}}{\gamma}F(sI - \overline{A} - L_2 C_2)^{-1}\begin{bmatrix} \frac{B}{\sqrt{\epsilon}} & L_2 \end{bmatrix}\right) < \gamma \ \ \forall s \in j\mathbf{R} \end{aligned}$$

for some $\epsilon > 0$ where E_{rel} is same as in (21) that yields ARE (22). In the process of tuning γ and ϵ, we found that large γ tends to give smaller overshot although the design of L_2 is not very sensitive to the γ value. On the other hand, L_2 is quite sensitive to the value of ϵ because of the use of small value. The loop gain when broken at the actuator input is plotted in Figure 5 with solid line and the corresponding sensitivity is plotted with dashed line. The roll-off of the loop gain at $\omega = 300$ is 59 dB that satisfies the design specification. Moreover its corresponding (normalized) step response for linearized model is plotted in Figure 6 with solid line that again satisfies the design requirement. The dashed line shows the control signal at the input to the actuator.

Using the proposed algorithm illustrated above, we computed four linear controllers at $(\alpha, M) = (0, 1.5), (0, 3), (20, 1.5), (20, 3)$. The design results are tabulated as follows.

$(\alpha, M) =$	$(0, 1.5)$:				
$F =$	-1.073×10^2	-2.347×10^4	-2.565×10^6	-1.593×10^7	-8.487×10^6
$L_1 w_1 =$	-3.29×10^{-3}	-6.083×10^{-4}	-1.127×10^{-4}	-2.087×10^{-5}	-3.732×10^{-6}
$L_2 w_2 =$	-1.708×10^{-1}	3.485×10^{-3}	4.906×10^{-5}	3.878×10^{-7}	7.426×10^{-9}
$(\alpha, M) =$	$(0, 3)$:				
$F =$	-1.934×10^1	-4.245×10^3	-4.659×10^5	-3.322×10^6	-1.516×10^7
$L_1 w_1 =$	-9.617×10^{-2}	7.065×10^{-3}	-6.951×10^{-5}	-2.5×10^{-4}	-4.756×10^{-5}
$L_2 w_2 =$	-1.502×10^{-3}	-8.817×10^{-6}	4.753×10^{-6}	8.201×10^{-7}	6.933×10^{-8}
$(\alpha, M) =$	$(20, 1.5)$:				
$F =$	-3.048×10^1	-6.934×10^3	-7.902×10^5	-1.159×10^7	-1.276×10^8
$L_1 w_1 =$	-5.316×10^{-1}	2.2954×10^{-2}	5.252×10^{-4}	-1.248×10^{-4}	-1.180×10^{-5}
$L_2 w_2 =$	-3.170×10^{-2}	1.963×10^{-4}	1.567×10^{-5}	4.583×10^{-7}	1.398×10^{-8}

$(\alpha, M) =$	$(20, 3)$:				
$F =$	-2.502×10^{1}	-5.663×10^{3}	-6.433×10^{5}	-9.03×10^{6}	-1.259×10^{8}
$L_1 w_1 =$	-2.512×10^{-1}	1.4×10^{-2}	3.354×10^{-4}	-4.312×10^{-5}	-3.934×10^{-6}
$L_2 w_2 =$	-1.52×10^{-2}	7.184×10^{-5}	7.161×10^{-6}	1.908×10^{-7}	4.281×10^{-9}

The above four controllers do meet performance requirements for the four linearized models respectively. Moreover robust stability is guaranteed in presence of 25% perturbation in the aerodynamic coefficients. The designed linear controllers are then scheduled with the same scheme as in [7] that involves interpolation of state feedback and estimator gains. The gain scheduled controller is implemented as in Figure 7 where the nonlinear plant $\mathcal{N}$ includes the actuator dynamics. However the scheduled control performance does not meet the performance requirement in terms of step response even though each of the four controllers yields excellent performance for each of the four linearized plants respectively. As demonstrated in [7], the linear controllers are continuous functions of scheduling variables as well as the uncertain parameters in the aerodynamic coefficients because state feedback and estimator gains are obtained from solutions of two AREs that are continuous functions of the plant parameters. Unsatisfactory performance implies that four linear controllers are not sufficient to guarantee the smoothness of the scheduled controllers as linearized models have quite different dynamics at different trim conditions. This problem will be studied further in the future.

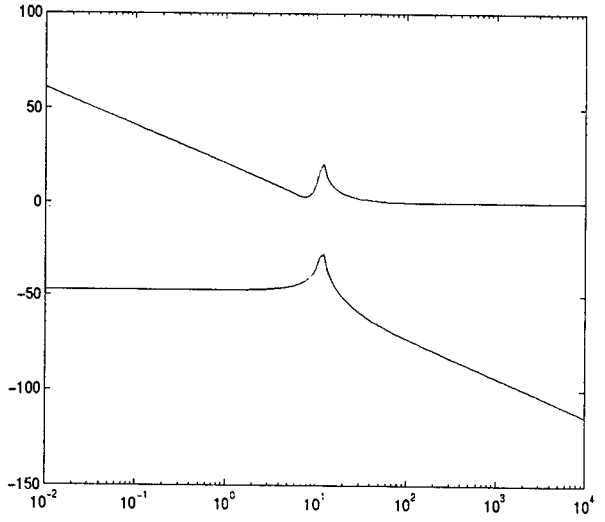


Figure 5: Gain plots of W_1 and W_2

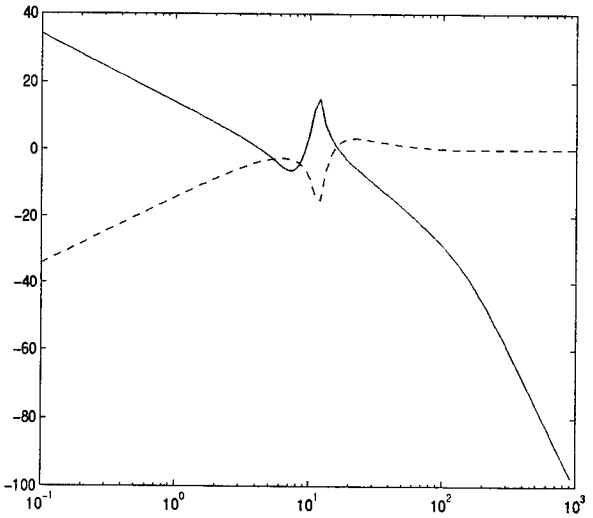


Figure 6: Loop gain and sensitivity

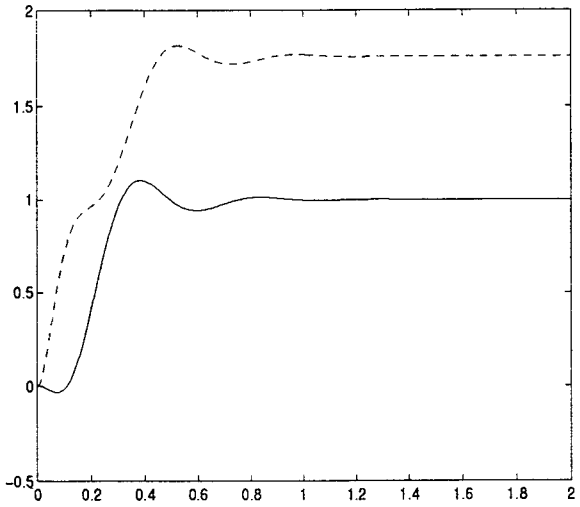


Figure 7: Step responses of normal acceleration and control input

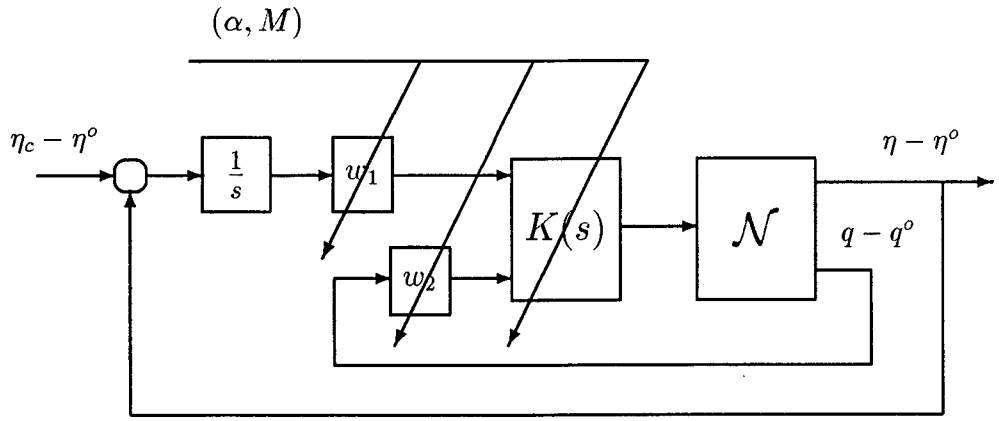


Figure 8: Implementation of gain scheduled control

References

[1] R. J. Adams and S. Banda, "Robust flight control design using dynamic inversion and structured singular value synthesis," *IEEE Trans. Contr. Syst. Tech.*, vol. 1, pp. 80, 1993.

[2] B.M. Chen, A. Saberi and U.-L. Ly, "Exact computation of the infimum in $\mathcal{H}_\infty$-optimization via output feedback," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 70, Jan. 1992.

[3] J. C. Doyle, "Lecture notes in advances in multivariable control," ONR/Honeywell Workshop, Minneapolis, MN, 1984.

[4] J. C. Doyle, K. Glover, P. P. Khargonekar and B. A. Francis. "State space solutions to standard H_2 and $\mathcal{H}_\infty$ control problems," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 831, 1989.

[5] K. Glover, "Robust stabilization of linear multivariable systems: Relations to approximation," *Int. J. Contr.*, vol. 43, pp. 741, 1986.

[6] K. Glover and J. C. Doyle, "State-space formulae for all stabilizing controllers that satisfy an $\mathcal{H}_\infty$-norm bound and relation to risk sensitivity," *Syst. and Contr. Lett.*, vol. 11, pp. 167, 1988.

[7] R. A. Hyde and K. Glover, "The application of scheduled $\mathcal{H}_\infty$ controllers to a VSTOL aircraft," *IEEE Trans. Automat. Contr.*, pp. 1021, 1993.

[8] D. Meyer and G. Franklin, "A connection between normalized coprime factorizations and linear quadratic regulator theory," *IEEE Trans. Automat. Contr.*, vol. 32, pp. 227, 1987.

[9] D. C. McFarlane and K. Glover, "Robust controller design using normalized coprime factor descriptions," *Lecture Notes in Control and Information Sciences*, New York: Springer-Verlag, 1990.

[10] D. McFarlane and K. Glover, "A loopshaping design procedure using $\mathcal{H}_\infty$ synthesis," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 759, 1992.

[11] R. T. Nichols, R. T. Reichert and W. J. Rugh, "Gain scheduling for $\mathcal{H}_\infty$ controllers: a flight control example," *IEEE Trans. Contr. Syst. Tech.*, vol. 1, pp. 69, 1993.

[12] A. Packard, "Gain scheduling via linear fractional transformations," *Syst. and Contr. Lett.*, vol. 22, pp. 79, 1994.

[13] W. Rugh, "Analytical framework for gain scheduling," *IEEE Contr. Syst. Magz.*, pp. 79, 1991.

[14] J. S. Shamma and M. Athans, "Gain scheduling: potential hazards and possible remedies," *IEEE Contr. Syst. Magz.*, pp. 101, 1992.

[15] J. S. Shamma and J. R. Cloutier, "A linear parameter varying approach to gain scheduled missile autopilot design," *Proc. Amer. Contr. Conf.*, pp. 1317, 1992.

[16] G. Stein and M. Athans, "The LQG/LTR procedure for multivariable feedback control design," *IEEE Trans. Automat. Contr.*, vol. 32, pp. 105, 1987.

THERMAL MODELING OF HETEROJUNCTION BIPOLAR TRANSISTORS

V.S.Rao Gudimetla
Assistant Professor
Department of Electrical Engineering and Applied Physics

Oregon Graduate Institute
P. O. Box 91000
Portland, Oregon 97291-1000

Final Report for:
Summer Faculty Research Program
Wright Laboratory

Sponsored by:
Air Force Office Of Scientific Research
Boiling Air Force Base, DC

and

Wright Laboratory

September, 1994

THERMAL MODELING OF HETEROJUNCTION BIPOLAR TRANSISTORS

V.S.Rao Gudimetla
Assistant Professor
Department of Electrical Engineering and Applied Physics
Oregon Graduate Institute

Abstract

Using Gummel-Poon model and a nonlinear simulator, effects of power dissipation on the performance of heterojunction bipolar transistors were studied. Emphasis is given to thermal effects on the output power at 1db gain compression point, output power at several higher order harmonics, intermodulation intercept and large signal S-parameters. Variation of these nonlinear phenomena with thermal impedance, frequency and thermal coefficients of the elements in the equivalent circuit was also studied in detail with an intention to optimize the linear power output.

THERMAL MODELING OF HETEROJUNCTION BIPOLAR TRANSISTORS

V.S.Rao Gudimetla

Introduction

One of the important features of GaAs-based Heterojunction Bipolar Transistors (HBTs) is that they can provide large amounts of power with a high degree of linearity due to their high power density and large base-collector junction voltage breakdown[1-3]. However due to poor thermal conductivity of GaAs, at high power densities thermal effects via self-heating are very pronounced and the device model must take into consideration thermal effects and also be amenable for computer-aided design applications. Because some of the existing SPICE bipolar transistor models do not include important effects such as time constants associated with the transit charge in the base and base-collector space charge regions, current gain dependence on the collector current, and the effects of device self-heating[4,5], there is a need to develop an accurate large signal HBT device model, that includes thermal effects, for computer-aided design. A conventional way to achieve this is to modify the existing Gummel-Poon (GP) models for bipolar transistors and add a thermal subcircuit with proper time constants to the equivalent circuit. The thermal subcircuit takes into consideration the total dissipation and depending on the thermal resistance, the temperature of the device due to self-heating is predicted. Then several equivalent circuit parameters (ECPs) in the GP model are updated and circuit performance at the new temperature is calculated. Though other models have been proposed in the literature, their development has been along the above lines except that they may have emphasized to one or more phenomena. For example, Grossman and Choma propose[5] a very elaborate model for TRW HBTs. Whitefiled [6,7] *et al.* reported a large signal model in the Ebers-Moll topology that can be used in commercial nonlinear simulators for further study. Its important features are simulation of the base collector characteristics by two reverse-biased diodes (these correspond to intrinsic and extrinsic junctions), bias dependent transit time and effects of surface recombination. They are able to obtain excellent match between their experimental and simulation results. Validity of all such models should be verified by comparing the simulated results with experimental data not only for linear performance (such as DC I-V data and small signal s-parameters) but also for the nonlinear performance such as 1 db gain compression output, harmonic power output, power added efficiency and intermodulation intercept. Otherwise, models may not be useful in the computer-aided design of microwave circuits. Thus far, no comprehensive simulations for nonlinear performance were reported in the literature. An effort has been made in this report to address this problem. In this report, first results for an existing GP model for HBT was compared with DC data to check the validity of the model and then thermal and nonlinear performance of the model were studied extensively. It is shown that the predicted results have correct theoretical behavior and the relevant magnitudes compare well with those reported in the literature. Efforts have also been made to compare GP model for another device, reported in the literature. In general, the results are satisfactory though some limitations exist. These are detailed later.

Gummel-Poon Model

A GP model[8,9] for HBTs, which is similar to that of a bipolar transistor but with different underlying equations, is shown in Figure 1. Using measured DC I-V data and small signal S-parameters at different temperatures as input, the GP equivalent circuit for the HBT at various temperatures was extracted using a program, called SPECIAL[8]. Using the extracted ECP values at various temperatures, the program evaluates the corresponding thermal coefficients. The GP model along with thermal coefficients is then used in the microwave harmonica[9] program to study the nonlinear performance of the device such as intermodulation, power added efficiency and power output at 1db, 2db and 3db gain compressions at various temperatures and frequencies. Model extraction using experimental data is a very difficult procedure for HBTs and substantial literature exists on this topic[10-14].

In the following, equations used in SPECIAL program for thermal modeling of HBTs are described and some comments are added with regard to the validity of these equations[1,8]. Although these are the suggested equations, one can use any SPICE model along with temperature coefficients for use in microwave harmonica. For the common emitter (CE) configuration, the base and collector currents are given by

$$I_c = I_{sf}\left[e^{\frac{V_{be}}{n_f kT}} - 1\right] - I_{sr}\left(1 + \frac{1}{\beta_r}\right)\left[e^{\frac{V_{be}}{n_r kT}} - 1\right] - I_{sc}\left[e^{\frac{V_{bc}}{n_c kT}} - 1\right]$$

and

$$I_b = \frac{I_{sf}}{\beta_f}\left[e^{\frac{V_{be}}{n_f kT}} - 1\right] + I_{se}\left[e^{\frac{V_{be}}{n_c kT}} - 1\right] + \frac{I_{sr}}{\beta_r}\left[e^{\frac{V_{bc}}{n_r kT}} - 1\right] + I_{sc}\left[e^{\frac{V_{bc}}{n_c kT}} - 1\right]$$

where the ideality factors n_j, current gains β_y, and the saturation currents I_{sj} are assumed to be temperature dependent. I_{sf} is the forward transport current and I_{sr} is the reverse saturation current. I_{se} and I_{sc} are the base-emitter and base-collector reverse leakage saturation currents respectively. Note different saturation currents are used in the model. Thermal dependence of these parameters is described by

$$n_j(T) = n_x(T_0)\,[1 + A_{nj}(T - T_0)]$$

with j = f, c, r, or e and T_0 = 273 K.

$$\beta_y(T) = \beta_y(T_0)[1+A_{nj}(T-T_0)]$$

with y= r, or f and T_0 = 273 K. Finally,

$$I_{sj} = A_{sj}\left[\frac{T}{RT}\right]^3 e^{\frac{V_{sj}}{k}\left(\frac{1}{RT}-\frac{1}{T}\right)}$$

with RT = 300 K.

In the last equation, the exponent 3 may not be valid for all saturation currents and it is more useful to have it as an input parameter. Because the base is doped typically several order of magnitude higher than the base of an ordinary bipolar device, the conductivity modulation effects are ignored. It was suggested[5] that the ideality factor n_f should be independent of temperature but the saturation current I_{sf} modeled as

$$\ln(I_{sf}) = \ln(I_{sf\infty}) - T_s / T$$

where the parameter T_s is related to the activation energy

$$\Phi_s = \frac{k}{q} T_s$$

It is the ideality factor n_e which models optical recombination term in the base-emitter space charge region that is strongly temperature dependent, given by the equation:

$$n_e(T) = n_e(T_0)\left[1 + \alpha_e \Delta T + \beta_e \Delta T^2\right]$$

For TRW HBTs, n_e increases linearly with temperature and it is found to be so also for HBTs tested at WL/ELMT. For Ti HBTs, the temperature coefficients of all four ideality factors are roughly of the same magnitude and their relative importance should be judged by their contributions to the nonlinear phenomena. The temperature reference point is assumed to be the room temperature. With V_b and V_c as the applied base-emitter and collector-emitter voltages, the intrinsic voltages across the base-emitter junction V_{be} and base-collector junction V_{bc} are given by

$$V_{be} = V_b - (I_b + I_c)R_e - I_b R_b$$

$$V_{bc} = V_b - V_c + I_c R_c - I_b R_b$$

where R_e is the base-emitter junction resistance, R_b is the base spread resistance and R_c is a combination of resistances from different layers of collector. The temperature dependence on power dissipation is given by

$$T = T_0\left[\frac{(-k+1)\,\theta(T_0)\,P}{T_0} + 1\right]^{\frac{1}{-k+1}}$$

where θ is called thermal impedance and P is the power dissipation in the circuit and given by

$$P = I_c V_c + I_b V_b$$

The depletion or junction capacitances[10,14] are given by

$$C_{be} = C_{be,\,pad} + \frac{C_{be0}}{\left(1 - \frac{V_{be}}{V_{be,\,bar}}\right)^{1/2}}$$

$$C_{bc} = C_{bc,\,pad} + \frac{C_{bc0}}{\left(1 - \frac{V_{bc}}{V_{bc,\,bar}}\right)^{1/2}} \qquad V_{bc} > V_0$$

$$C_{bc} = C_{bc,\,pad} + \frac{C_{bc0}}{\left(1 - \frac{V_0}{V_{bc,\,bar}}\right)^{1/2}} \qquad V_{bc} < V_0$$

For both C_{bc} and C_{be}, when V_{bx} is larger than $fV_{bx,bar}$ (f=0.01), the formula

$$C(V) = \frac{C_0}{(1-f)^{1.5}}\left[1 - 1.5f + \frac{0.5V}{V_{bar}}\right]$$

The seven parameters, $X = C_{be,pad}$, C_{be0}, $V_{be,bar}$, $C_{bc,pad}$, C_{bc0}, $V_{bc,bar}$, and C_{ce} vary as

$$X(T) = X(T_0)\,[1 + B_x(T - T_0)]$$

In general, thermal effects on the capacitances are minimal (as their absolute magnitudes are in the order of femto-farads) and can be ignored for performance study in GHz frequency range. The base-emitter diffusion capacitance is given by

$$C_{bet} = \frac{\tau_b + \tau_c}{r_e}$$

where $r_e = n_f kT/ I_e$, τ_c is the collector transit time and τ_b is the base transit time. For small signal models, τ_b is given by

$$\tau = \frac{W_b^2}{\eta D_n} + r_{bX}\left(1 - \frac{I_c^*}{I_c}\right)$$

where W_b is the neutral base width, η is 2.4 or 2 and D_n is the diffusion constant for electrons in the p-base. The term r_{bX} models the base-widening at high injection levels and is applied only when $I_c^* < I_c$. A method to determine I_c^* using experimental data is given by Anholt[8].

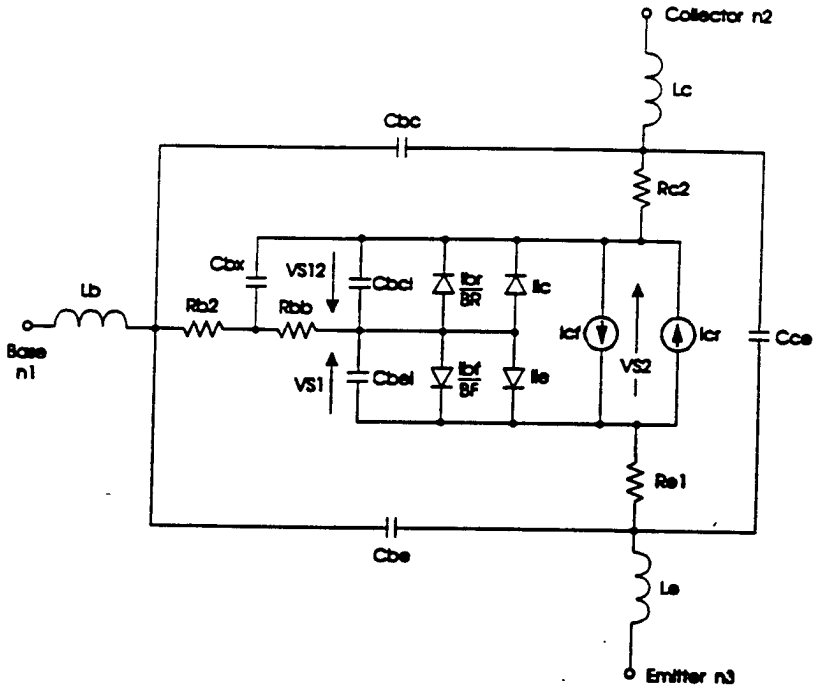


Fig. 1 GP Model for HBTs

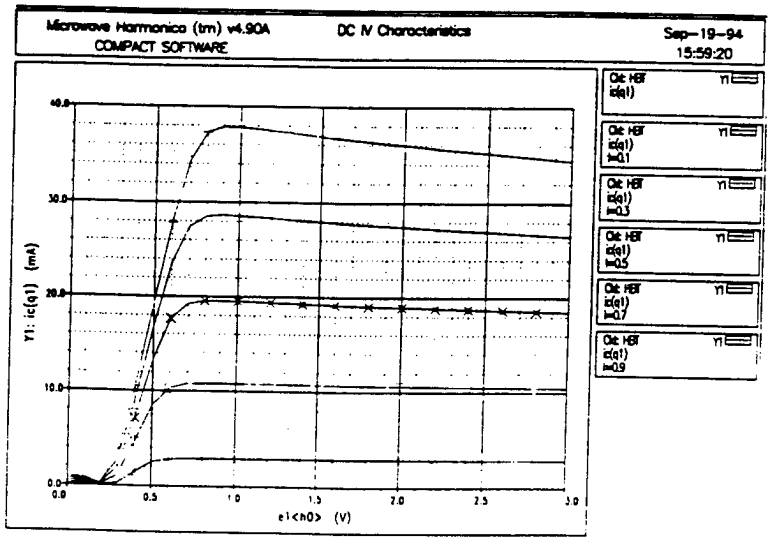


Fig. 2 DC I-V Characteristics

DC Analysis and Validation:

For the Ti HBT tested, the SPECIAL program gave the following values for GP model.

GP Model tiHbt BIP (NHBT VCMX=3 IBMN=0.1ma IBMX=0.9ma NPLT=5 NF=1.5035 ANF=-0.000692 ISF=1.8951E-19 VSF=0.934 NE=1.8732 ANE=-0.000693 ISE=8.2955E-18 VSE=0.7545 NC=2.0214 ANC=-0.000590 ISC=1.073E-13 VSC=0.5580 NR=1.9190 ANR=-0.000764 ISR=3.2635E-16 VSR=0.6868 BF=175.51 ABF=-0.002745 RE1=2.613 ARE1=0.000325 RB=9.8926 RC2=4 ARC2=-0.00467 TNOM=300 CBE=48.2625FF CJE=66.937FF ACJE=0.001257 VJE=1.753 AVJE=-0.001565 CBC=29.549FF CJC=39.970FF ACJC=0.000896 VJC=1.1965 AVJC=-0.001213 CCE=61.723FF TF0=3.5288PS ATF0=0.002742 TF1=12.239PS ATF1=0.000348 ITF=18.526MA AITF=-0.001695 MJE=0.5 MJC=0.5 FCC=0.01)

ANF, ANC,ABF etc. represent the linear temperature coefficients. Harmonica can accept up to quadratic variation with respect to temperature. The thermal effects are modeled by thermal resistance. No thermal capacitance is included to account for thermal time delay effects. The data required to build the GP model of HBT contains the DC I-V data and S-parameter data at several temperature. Though DC I-V data is available at different temperatures, no s-parameter data is available at different temperatures. Hence the model, originally extracted is accepted. A method, described by Dawson[15] *et al.*, is used to determine the thermal resistance as 600 °C/W. This technique uses DC I-V data at various temperatures. Fig.2 shows the simulated DC characteristics, which show correctly the expected drooping feature. For higher thermal impedance, the droop in the characteristics increases.

An Ebers-Moll model for HBT is described by Whitefield[7] *et al.* and they also give the equivalent circuit parameter values for their model. This model is simulated using microwave harmonica and the simulation results appear to agree with their experimental data. More work is being planned to get the raw data and check the validity of the simulator. It is not surprising that the a good match exists between the simulation results and the DC I-V data because the GP or other models are built using DC I-V data and with extensive optimization. Thus true test lies in the comparison of simulated AC results such as large signal S-parameters, 1db gain compression power output and intermodulation intercept with experimental data

Large Signal S-parameters:

Microwave Harmonica is used to generate the large signal s-parameter data over the frequency range 1-18Ghz. Fig. 3 and 4 show S_{11} and S_{22}. It is found that for high input powers, especially near 1 db gain compression, it is important to include the thermal impedance. When the thermal impedance is included, harmonica simulations show that for large input powers, $|S_{11}|$ and $|S_{22}|$ remain constant with respect to frequency though their phases change a lot. Without thermal impedance in the circuit, their magnitudes change with frequency significantly at low

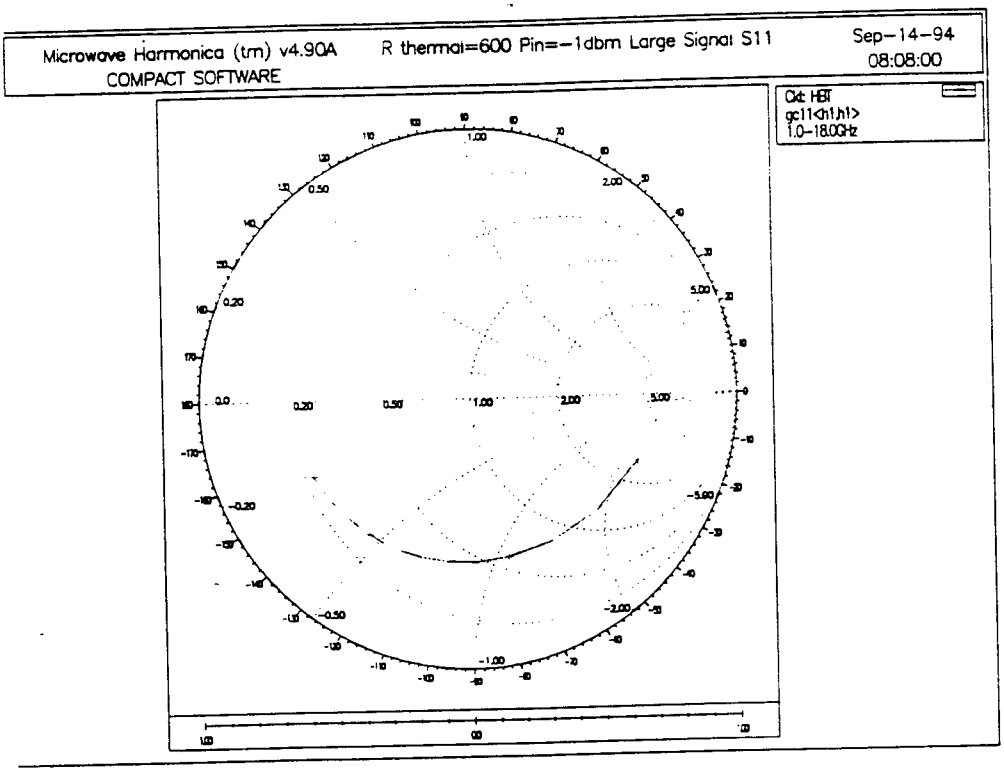


Fig. 3 Large Signal S_{11} Variation with Frequency

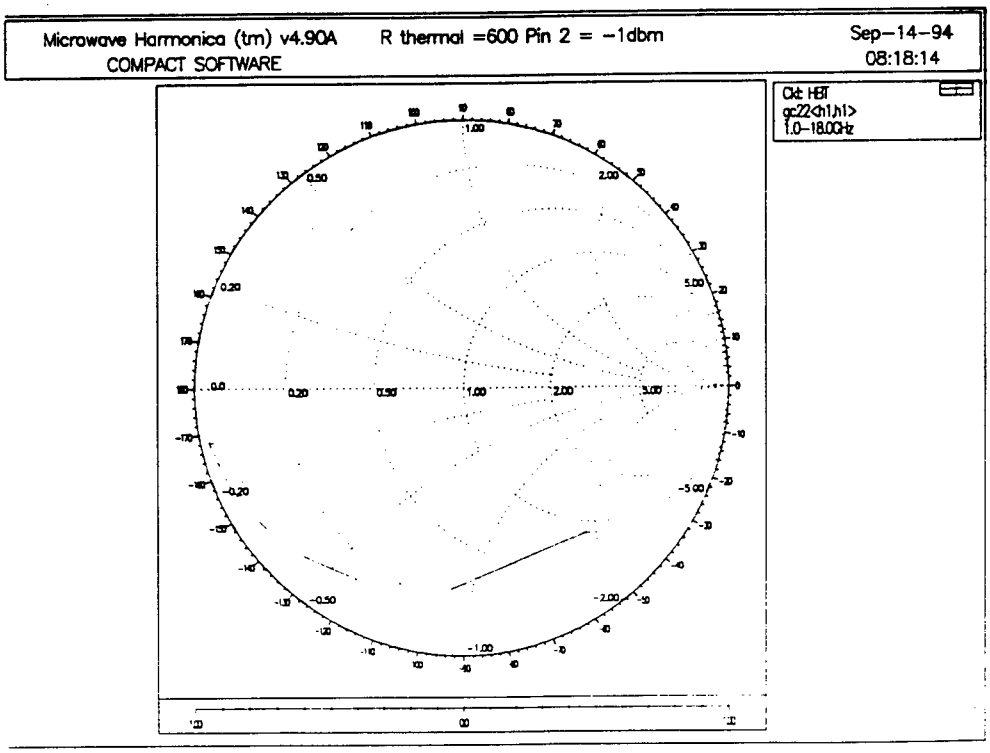


Fig. 4 Large Signal S_{22} Variation with Frequency

Frequency (GHZ)	K	B1	KCSO MAG	KCSO ANG	KCSR	KCLO MAG	KCLO ANG	KCLR	db(MSG)	db(GMAX)
1.0000	0.3108	0.8029	1.4080	29.9018	0.7280	1.2519	41.3936	0.5040	23.7897	23.7897
2.0000	0.2025	0.8267	1.5245	55.7258	0.9659	1.3040	79.9830	0.6586	21.1818	21.1818
3.0000	0.1672	0.8165	1.5933	71.1292	1.0844	1.3262	99.8546	0.7198	20.0007	20.0007
4.0000	0.1677	0.7873	1.6495	83.2509	1.1548	1.3479	111.1043	0.7515	19.2601	19.2601
5.0000	0.1829	0.7575	1.6885	92.3505	1.1899	1.3645	118.6716	0.7634	18.6980	18.6980
6.0000	0.2039	0.7332	1.7059	100.0198	1.1931	1.3722	124.3051	0.7575	18.2075	18.2075
7.0000	0.2283	0.7131	1.7140	106.7788	1.1824	1.3754	128.6864	0.7432	17.7877	17.7877
8.0000	0.2527	0.6991	1.6995	112.5535	1.1445	1.3682	132.5017	0.7147	17.3866	17.3866
9.0000	0.2758	0.6892	1.6747	117.4847	1.0956	1.3564	135.8238	0.6812	17.0030	17.0030
10.0000	0.3037	0.6791	1.6489	122.5502	1.0420	1.3446	139.2040	0.6450	16.5891	16.5891
11.0000	0.3378	0.6689	1.6223	127.7941	0.9835	1.3329	142.6769	0.6060	16.1390	16.1390
12.0000	0.3799	0.6585	1.5951	133.2597	0.9195	1.3215	146.2717	0.5638	15.6448	15.6448
13.0000	0.4328	0.6480	1.5675	138.9916	0.8495	1.3104	150.0146	0.5183	15.0962	15.0962
14.0000	0.5010	0.6376	1.5399	145.0364	0.7727	1.3000	153.9301	0.4691	14.4782	14.4782
15.0000	0.5919	0.6272	1.5129	151.4423	0.6884	1.2904	158.0406	0.4159	13.7691	13.7691
16.0000	0.7190	0.6170	1.4873	158.2573	0.5959	1.2820	162.3668	0.3583	12.9351	12.9351
17.0000	0.9095	0.6070	1.4642	165.5249	0.4944	1.2751	166.9263	0.2960	11.9191	11.9191
18.0000	1.2279	0.5974	1.4452	173.2771	0.3834	1.2703	171.7318	0.2286	10.6129	7.7340

Table 1. Stability and Gain Data Using Large Signal S-Parameters

frequencies. This phenomena was also observed in the numerical simulations by Liou[16,17] *et al.* and in load-pull experiments on similar devices. Since capacitances in the HBT equivalent circuit are in the order of femtofarads, it is assumed that their impedances are very high in the GHz frequency range. However at large signals, capacitance variations are determined by the signal magnitude and frequency. It is found that only at very high frequencies (in this case, > 10 Ghz), the normal 6 db/octave roll-off of $| S_{21} |$ was observed. For frequencies < 10 GHz, the $| S_{21} |$ is not constant but rolls-off at a rate less than 6 db/octave. For the simulated device, $| S_{21} |^2$ has a magnitude of 19.59 db at 2 GHz, 15.51 db at 4 GHz, 9.9 db at 8 GHz and 3.5 db at 16 GHz. Similar results were found in the numerical calculations of Liou[17] *et al.* for both intrinsic and extrinsic HBTs. The large signal S-parameters from Microwave Harmonica can be used for deriving stability factors, stability circles, maximum stable gain and approximate matching networks. Table. 1. shows stability factors and circles for the simulated device. It is found that in general HBTs are more stable at high frequencies. The instability at low frequencies is due to the large low frequency gain of HBT, which even with very little feedback via base-collector resistors can make the device potentially unstable.

Variation of Gain and Fundamental Power Output with Input Power:

Initially the device is simulated at 6 GHz for power output (P_{out}) and power gain with 50 ohm terminations. No efforts to optimize the power output or intermodulation by varying source and load impedances was done because often for such optimum cases, an extremely accurate model of the device is needed. For simulation, six harmonics have been allowed as it is assumed that the harmonic output power will be very low compared with the fundamental power beyond the sixth harmonic. In addition when the input power is very low, use of harmonics beyond the sixth led to poor convergence by Harmonica. One should set the number of harmonics, required for the harmonic balance calculations by examining the power output at several harmonics at each input power level. A discussion on the harmonic power outputs is included in the next section. The device is biased at V_{ce} = 2.7 volts and I_c = 20 mA. Also I_B = 0.4 mA. Near 1 db gain compression, the ac base current is in the order of 10 mA and the device is still in class A operation as the collector current has an ac amplitude of 20 mA superimposed on a 20 mA direct current. The ratio of ac collector and base currents is same of order as $| S_{21} |$ at 6 GHz. Current and voltage waveforms can be monitored both at the base and collector as functions of time and near 1db gain compression, distortion in the waveforms is noticed. Fig. 5 shows variation of P_{out} and transducer gain with P_{in} at 6 GHz at a thermal resistance of 600 °C/watt. It shows an output power of 8.846 dbm at 1 db gain compression. This agrees with the result given in a previous technical report[18]. This result is obtained only when the thermal impedance is close to 600 K/watt. For other thermal impedances, P_{1dbm} changes significantly. In addition, Fig. 6. gives P_{out} at 2 db and 3 db gain compressions for various values of thermal resistance. There is substantial interest in the P_{out} at 2 db and 3 db gain compressions because they provide higher outpower and will be useful operating points if the intermodulation output does not increase significantly from the value at 1db compression. For a thermal impedance of 600 °C/watt,

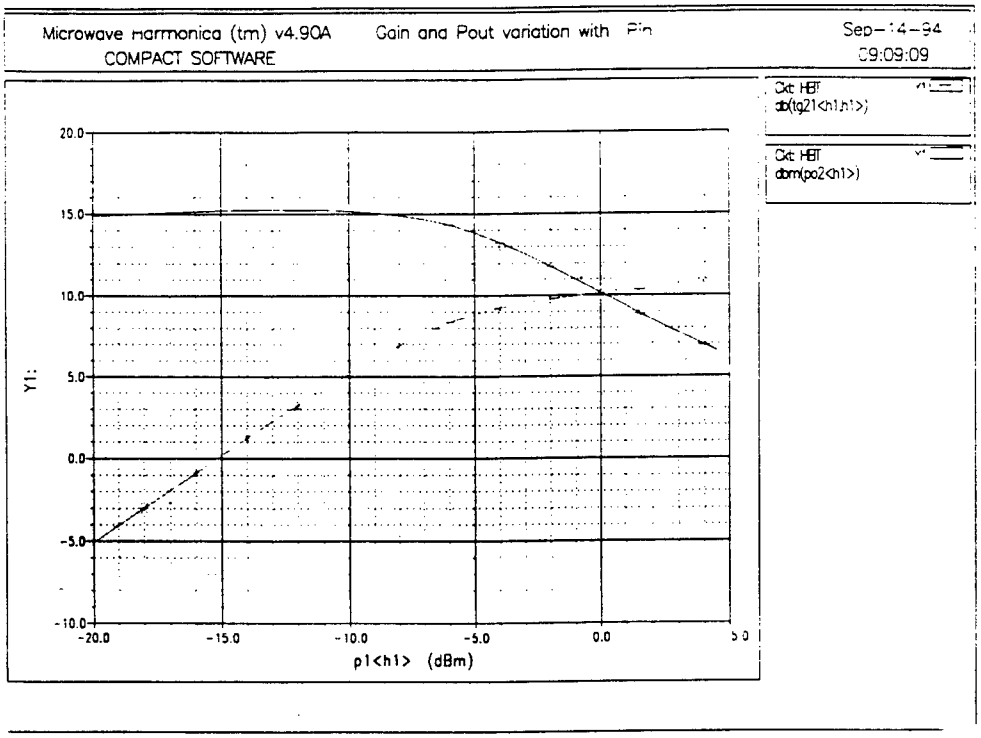


Fig. 5 Gain and P_{out} Variation with P_{in}

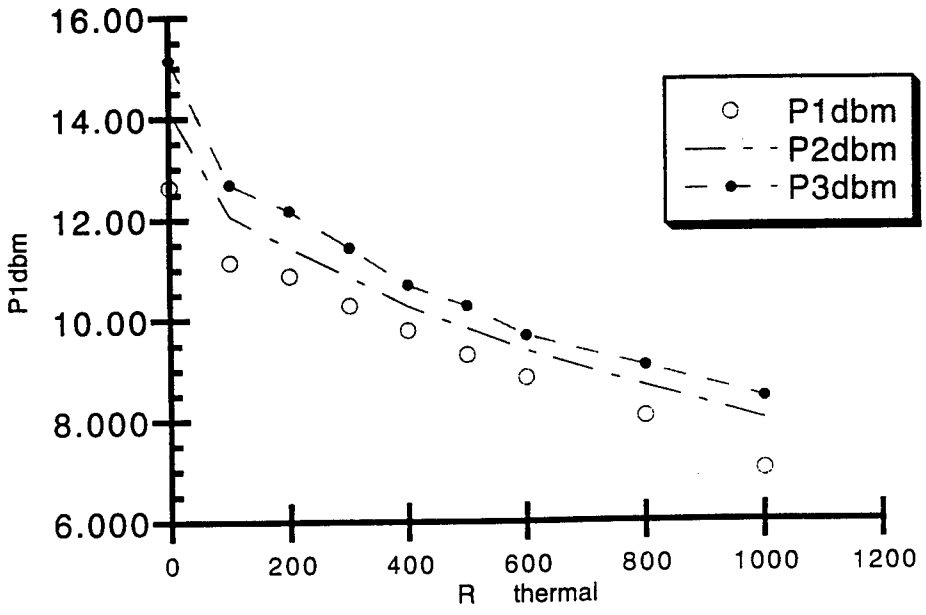


Fig. 6 P_{1dbm}, P_{2dbm} and P_{3dbm} versus $R_{thermal}$

at 6 GHz when the gain compression is increased from 1 db to 3 db, the power output increased by 0.8 dbm where as the intermodulation power output increased by 1.6 dbm. The results remain approximately the same at high frequencies. Note these results are at 50 ohm load and source impedances and further optimization is possible. The results in this section show that if with proper thermal design, thermal resistance is reduced, then P_{out} at various gain compression points can be substantially improved.

Harmonic Power Outputs

Fig. 7. and 8 show the output powers at the fundamental (6 GHz) and harmonics till sixth order. Harmonica has difficulty in converging when the gain is compressed beyond 5 db for the device under consideration. Hence it is assumed that the results till 3db gain compression are reliable. Within this limitation, as input power is increased, the fundamental power increases linearly initially and then eventually saturates or increases very little. On the other hand, second and third order harmonic power outputs increase more rapidly. Near 1db gain compression, the second harmonic power output is smaller than the third harmonic output in the frequency range 3-18 GHz. In any case, the highest harmonic output is about 20db less than the fundamental power output. In general, the results from Harmonica appear to follow the results reported elsewhere for HBTs [19].

Variation of Harmonic Power Output and Gain with Frequency

Fig. 9. shows the variation of gain and power output with respect to frequency at low input power level. Since the device gain in this case is nothing but $| S_{21} |^2$, as remarked earlier, the roll-off depends on the frequency. Thus the gain rolls-off initially slowly and then reaches 6 db/octave at very high frequencies. Till 1 db gain compression, power output follows the gain curve. As remarked earlier, these features agree with numerical simulations, reported elsewhere and experimental data. Fig. 10. shows the P_{1dbm} (Power output at 1db gain compression) with and without thermal impedance in the circuit. When the thermal impedance is assumed as zero, the device P_{1dbm} remains the same (note since gain falls off with increasing frequency, the corresponding power input to achieve 1 db gain compression increases with frequency). But when the thermal resistance is not zero, the P_{1dbm} decreases slowly initially and then falls off rapidly. This behavior is attributed to changes in the ideality factors of the diodes (as used in the model) and short circuit current gain β. It can be shown that for higher thermal resistances (poor thermal designs), P_{1dbm} will decrease further and also at higher frequencies for a given thermal impedance.

Intermodulation

To check the validity of Harmonica with regard to its ability to predict the intermodulation products, the device is simulated with two input frequencies at 6 GHz and 6.05 GHz. An examination of the simulation results show that the intermodulation output power increases with input power three times faster than the carrier power output (all powers are in dbm) as shown in Fig. 11. At a thermal resistance of 600 °C/watt, the third-order intermodulation

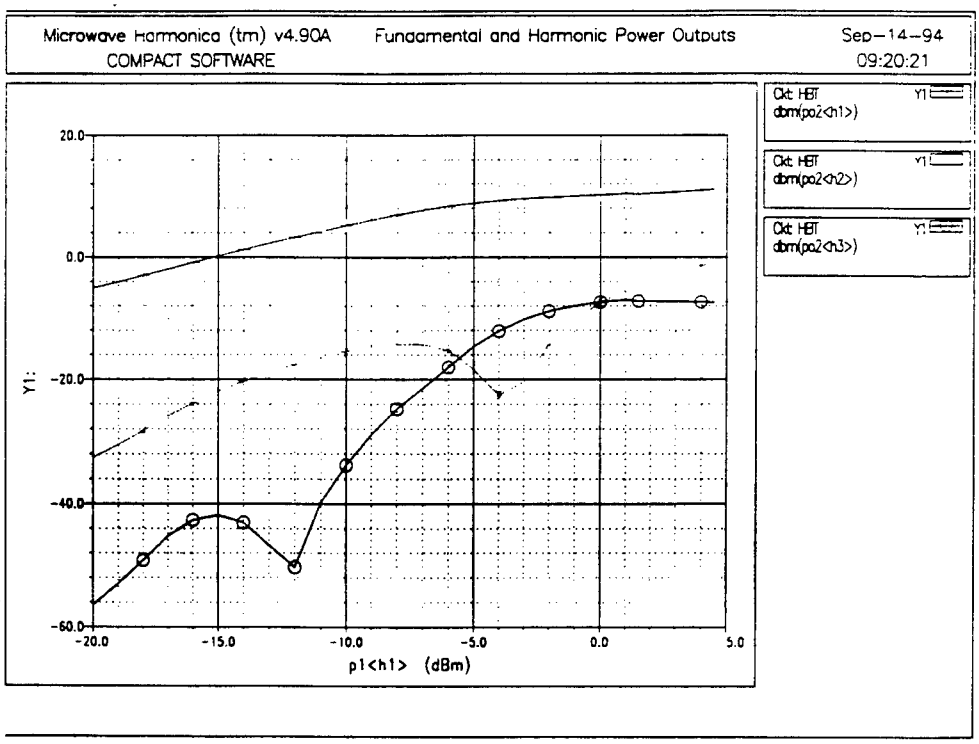


Fig. 7 Fundamental, Second and Third Harmonic Power Outputs: Variation with P_{in}

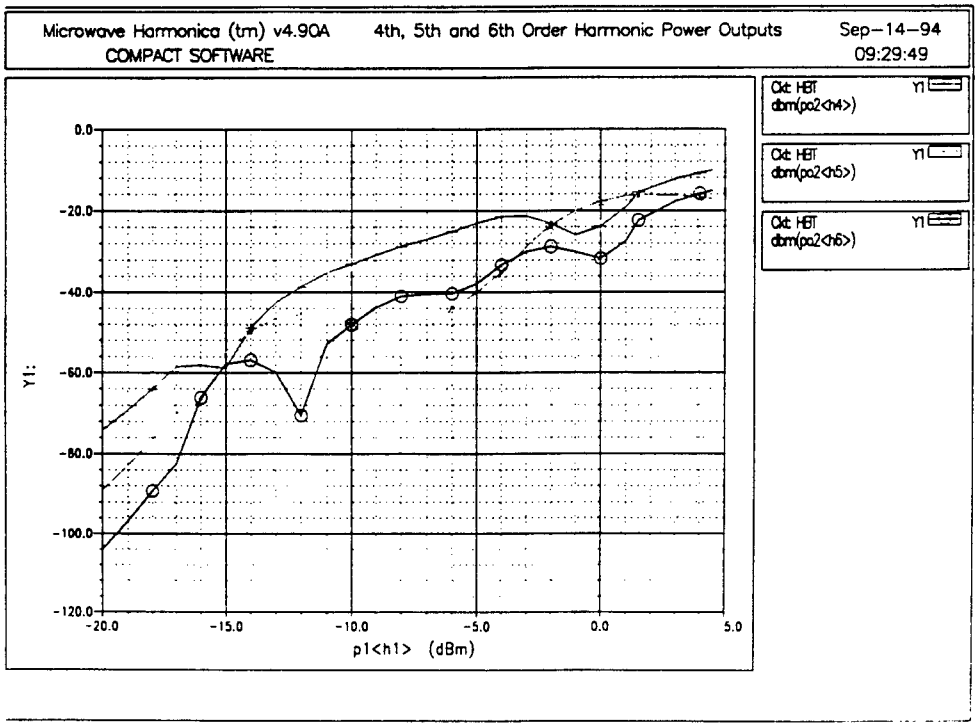


Fig. 8 Fourth, Fifth and Sixth Harmonic Power Outputs :Variation with P_{in}

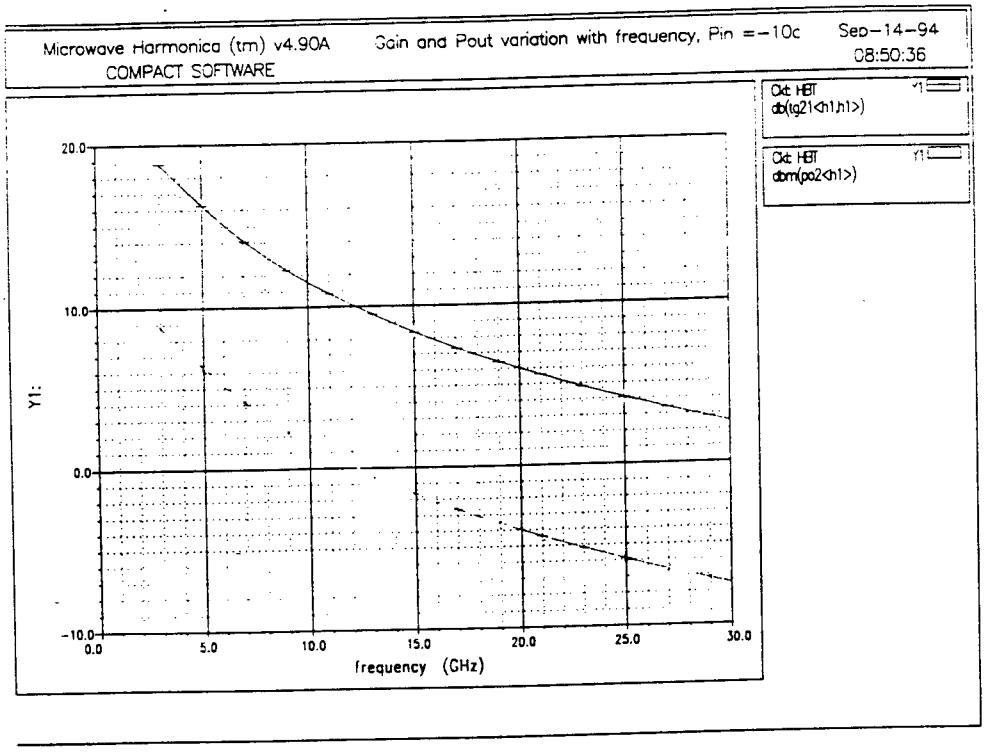


Fig. 9 Gain and P_{out} Variation with Frequency

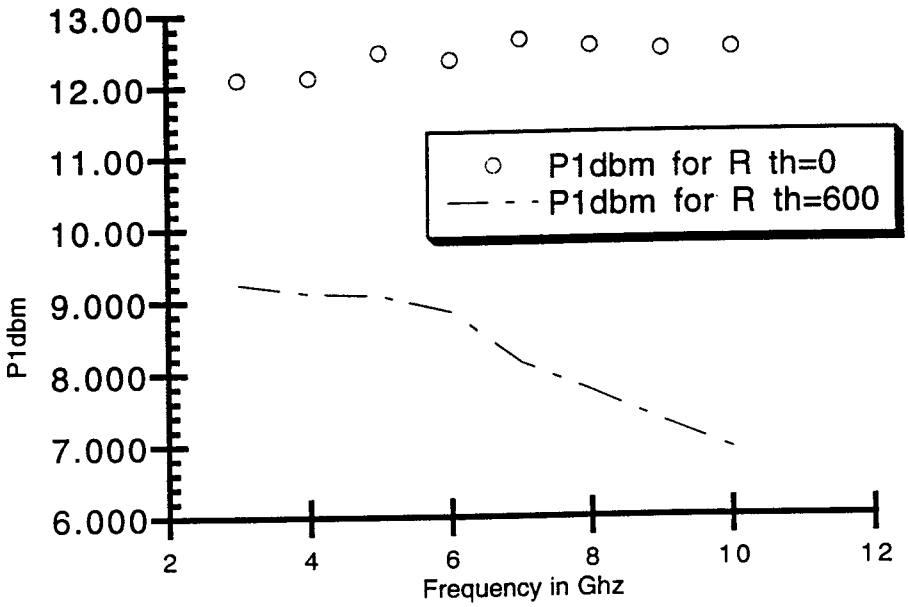


Fig. 10 P_{1dbm} *vs.* Frequency

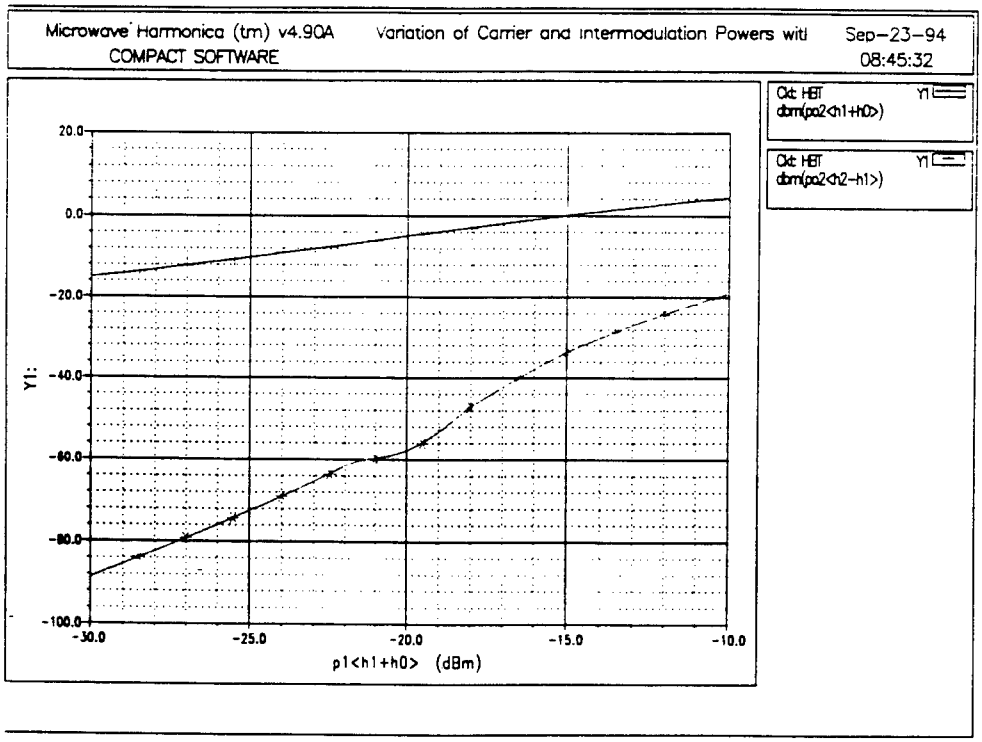


Fig. 11 Variation of Carrier and Intermodulation Powers with P_{in}

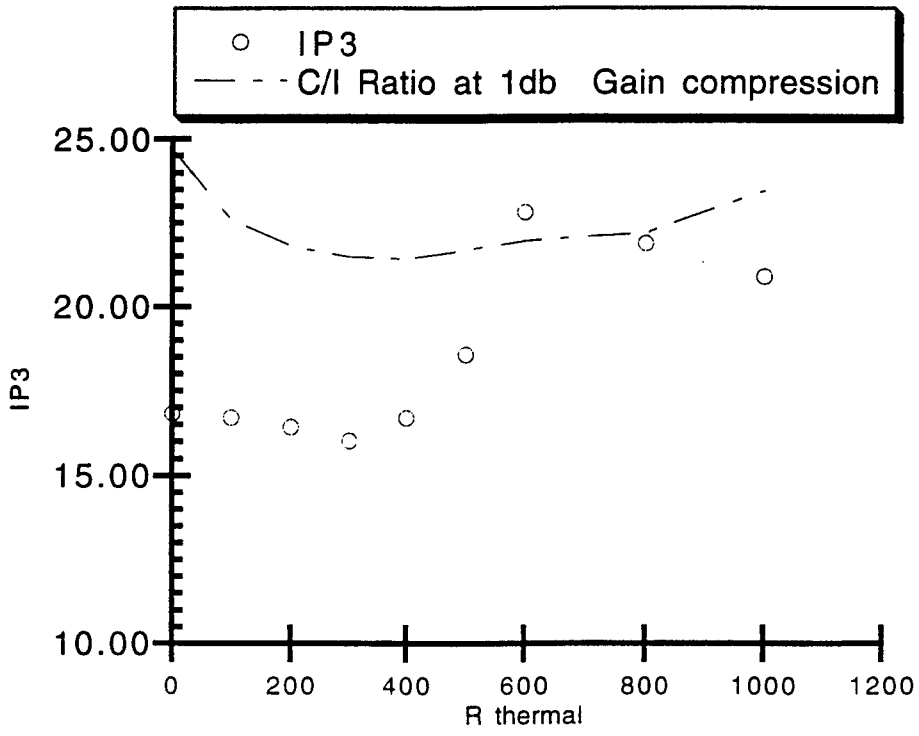


Fig. 12 Variation of IP_3 and C/I (at 1db Gain Compression) with $R_{thermal}$

intercept (IP_3) for this device is 22.83 dbm. Though the device was never experimentally tested for intermodulation results, this value is comparable to the IP_3 values reported in the literature[20,21]. IP_3 falls off for higher thermal impedances (poor thermal designs) as expected. No significant change for C/I (ratio of carrier power to intermodulation power) is observed at P_{1dbm} as thermal impedance is increased but at lower powers, it decreased for higher thermal impedances (Fig. 12). When no thermal effects are considered, it can be shown that the intermodulation current generated by the exponential base-emitter current is in part cancelled by the intermodulation current due to nonlinear capacitance of the junction. Thus the effect of thermal impedance is minimal for low thermal impedance values. For high thermal impedance values, thermal effects on the junction resistance are much more significant than those on the capacitances. Hence they may not cancel each other, there by leading to an increase in the intermodulation power output. For two values of $R_{thermal}$ = 0 and 600 °C/watt, variation of IP_3 with frequency has been studied and the results are shown in Fig. 13. It is noticed that when thermal resistance is incorporated in the simulation, IP_3 increases with frequency, then remains constant for some frequency range and then falls off. For low thermal impedances, the frequency variation has similar results but the range over which IP_3 remains constant is larger. A careful analysis of intermodulation currents, generated by various elements in the equivalent circuit and their thermal variations, is needed to get an insight into these features and to envision methods to optimize the linear power output without increasing the intermodulation power output. Microwave harmonica uses harmonic balance technique and it is possible to check its validity by developing analytical expressions via nonlinear volterra series with thermal impedance incorporated in it. To the best of my knowledge, no thermal effects on IP_3 of HBTs appeared in the literature.

Effects of Source and Load Impedances:

Variation of source and load impedances (generally we vary load impedance and match the source port) affect IP_3 significantly. Time limitations did not permit to develop load-pull contours. It will be useful to do such a study as prelude for the planned load-pull experiments. Load contour plots of constant P_{1dbm} superimposed on the contours of C over I on the Smith chart will be very useful for power amplifier design.

Effects of Thermal Coefficients on Power Compression and Intermodulation

In order to understand the effects of temperature coefficients of various elements in the equivalent circuit, first the temperature coefficients of base-emitter junction capacitance and base-collector junction capacitance have been changed from zero to ten times the experimentally modeled value. No significant deviations (beyond 1.0 dbm) have been noticed either in P_{1dbm} or IP_3 for a given $R_{thermal}$. This simulation has been carried out for four values of $R_{thermal}$ = 0, 300, 600 and 1000 at f=6 GHz and the conclusions remain the same. This is not a surprising result since magnitudes of the capacitances in the equivalent circuits are very small and small deviations due to thermal effects would not have mattered. Similar thermal simulations are carried out for base-emitter junction resistance

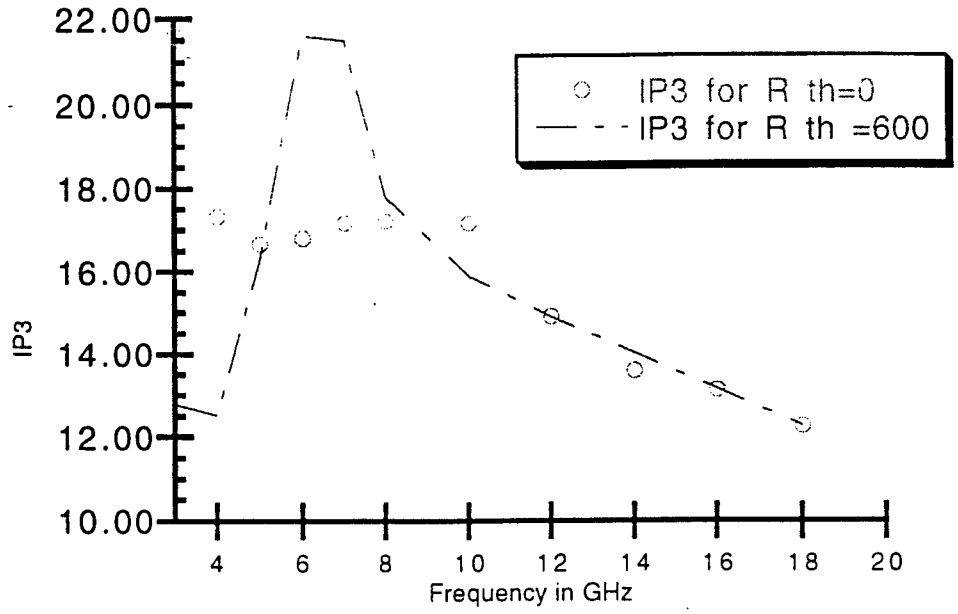


Fig. 13 Variation of IP_3 with Frequency and Thermal Resistance Effects

R_{be}, the base-collector resistance R_{c2} and no significant deviations in P_{1dbm} are noticed. However as one expects, the effect of temperature coefficients of the short circuit current gain β_f is very pronounced. Both IP_3 and P_{1dbm} fell off rapidly with increasing frequency and thermal impedance. Similar studies for temperature variations of ideality factors is under progress.

Conclusions

In this report, we studied and checked the validity of the proposed thermal GP model for HBTs using a nonlinear simulator. Good agreement was obtained with DC I-V data. It is shown that it is possible to generate large signal s-parameters for use in power amplifier design. Only when thermal impedance is included in the simulation, the large signal $| S_{21} |^2$ shows 6 db/octave roll-off at high frequencies and $| S_{11}|$ remains constant in magnitude with high frequencies. These results agree with those from numerical simulations and experimental measurements. The 1 db gain compression output matches with experimental data and for its correct prediction, thermal resistance must be included in the simulation. The intermodulation intercept values are large and comparable to those reported in the literature. Extensive simulations have been done to study the effects of frequency variation, changes in the thermal expansion coefficients and thermal impedance on P_{1dbm}, IP_3 and harmonic power outputs. From these results, it appears that the capacitance thermal expansion coefficients do not affect linear power output and intermodulation intercept. However thermal coefficients of the ideality factors, modeling the base-emitter and base-collector junctions affect both P_{1dbm} and IP_3 and further research on this topic is needed for optimizing the device linear power output. A need to extend this study to evaluate the role of each element in the determination of P_{1dbm} and IP_3 and a further critical examination of effects of thermal coefficients will lead to optimization of device design for the best linear power output. A need also exists to crosscheck the validity of Harmonica results for powers well below P_{1dbm} by developing nonlinear volterra series. In some instances, Harmonica faced converge problems and these cases require further attention.

References:

1. F. Ali, "HBTs and HEMTs", Artech House, Dedham Mass. 1994

2. M. E. Kim, B. Bayraktaroglu and A. Gupta, "HBT devices and circuit applications," Chapter 5 in Ref. 1 above.

3. B. Bayraktaroglu, J. Barrette, L. Kehias, C. I. Hunag, R. Fitch, R. Neidhard and R. Scherer, "Very high power density operation of GaAs/AlGaAs microwave HBTs," submitted to IEEE Elect. Dev. Lett.

4. G. Massobrio in "Semiconductor Modeling with SPICE" ed. by P. Antognetti and G. Massobrio, McGraw Hill, New York, 1988

5. P. C. Grossman and J. Choma, "Large-signal modeling of HBTs including self heating and transit time effects," IEEE Tran. Micro. The. Tech., Vol. 40, p. 449-464, 1992

6. D. S. Whitefiled, C. J. Wei and J. C. M. Hwang, "Temperature-dependent large signal model of heterojunction

bipolar transistors," Technical Digest IEEE GaAs IC Symposium, p. 221-224, October 1992.

7. D. S. Whitefiled, C. J. Wei and J. C. M. Hwang, "Elevated temperature microwave characteristics of heterojunction bipolar transistors," Technical Digest IEEE GaAs IC Symposium, p. 267-270, October 1993.

8. R. Anholt, SPECIAL and Background reports, 1993

9. Microwave Harmonica, User Manual, Compact Software Inc., Patterson, New Jersey, 1993

10. M. B. Das, "HBT device physics and models," chapter 4 in Ref. 1 above .

11. D. R. Pehlke and D. Pavlidis, "Evaluation of the factors determining HBT high-frequency performance by direct analysis of S-parameter data," IEEE Trans. Micro. Theo. Tech., Vol. 40., p. 2367-2373, 1992

12. S. A. Maas and D. Tait, "Parameter extraction method for HBTs," IEEE Microwave and Guided Wave Lett., vol. 2., p. 502-504, 1992

13. R. Anholt, "Investigation of GaAs MESFET equivalent circuits using transient current -continuity equation solutions," Solid State Electronics, vol. 34, 1991

14. D. Costa, W. U. Liu and J. S. Harris, Jr., "Direct extraction of the AlGaAs/GaAs heterojunction bipolar transistor small signal equivalent circuit," IEEE Trans. Elect. Dev., Vol. 38, p.2018-2023, 1991

15. D. E. Dawson, A. K. Gupta and M. L. Sahib, "CW measurement of HBT thermal resistance," IEEE Trans. Elect. Dev., Vol. 39, p. 2235-2240, 1992

16. L. L. Liou, J. L. Ebel and C. I. Huang, "Thermal effects on the characteristics of AlGaAs/GaAs heterojunction bipolar transistors using two-dimensional simulation," IEEE Trans. Elect. Dev., Vol. 40, p. 35-43, 1993

17. Private Discussion with L. L. Liou (WL/ELMD)

18. M. Y. Frankel and D. Pavlidis, "An analysis of the large-signal characteristics of AlGaAs/GaAs heterojunction bipolar transistors," IEEE Trans. Micro. Theo. Tech., Vol. 40, p. 465-474, 1992

19. "Accurate Active Device Models for Computer-aided Design of MMICs" Contractor Report, Compact Software, Inc., Patterson, New Jersey, 1993

20. S. A. Maas, B. L. Nelson and D. L. Tait, "Intermodulation in heterojunction bipolar transistors," IEEE Trans. Micro. Theo. Tech., Vol. 40., p. 442-448, 1992

21. A. Samelis and D. Pavlidis, "Mechanisms determining third order intermodulation distortion in AlGaAs/GaAs heterojunction bipolar transistors," IEEE Trans. Micro. Theo. Tech., Vol. 40., p. 2374-2380, 1992

FURTHER DEVELOPMENT OF SURFACE-OBSTACLE INSTRUMENT FOR SKIN-FRICTION AND FLOW DIRECTION MEASUREMENT

by

Raimo J. Hakkinen
Professor and Director, Fluid Mechanics Laboratory
Department of Mechanical Engineering

Washington University
One Brookings Drive
Saint Louis, Missouri 63130

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

and

Wright Laboratory

September 1994

Raimo J. Hakkinen
Professor and Director, Fluid Mechanics Laboratory
Department of Mechanical Engineering
Washington University

Abstract

The technical background and calibration of surface-obstacle skin friction meters were thoroughly reviewed in the 1993 Summer Program Final Report [1], and a summary was presented in [2]. The objective of the 1994 Summer Program was to complete the detailed design of the specific instrument proposed in [1], to initiate its fabrication, and to define test programs for facilities available at Wright Laboratory (WL) and Washington University (WU). These objectives were completed, fabrication of the instrument has been approved and is underway at Wright Laboratory, and calibration tests are tentatively scheduled in the WL M3 and M6 supersonic wind tunnels during 1995. An exploratory calibration project using a simple proof-of-concept prototype instrument was carried out under private funding in the Washington University Low-Speed Wind Tunnel, in April, 1994, and a thorough incompressible-flow reference calibration of the new WL instrument is planned for 1995.

FURTHER DEVELOPMENT OF SURFACE-OBSTACLE INSTRUMENT FOR SKIN-FRICTION AND FLOW DIRECTION MEASUREMENT

Raimo J. Hakkinen

Introduction

Accurate determination of skin friction drag is of primary importance in efficient aerodynamic design; it may constitute as much as fifty percent of the drag of a cruising aircraft, and knowledge of the shear stress distribution is an essential part of understanding the flow field of any complex flight vehicle in any speed range. Acquisition of precise experimental data has also become essential for validating the emerging computational techniques for the prediction of local skin friction distributions on general, three-dimensional vehicle configurations. While a great variety of experimental techniques exists for the measurement of local skin friction, there is at present no universally applicable method, and a choice must be made according to particular wind tunnel or flight test conditions. Surveys of skin friction measurement technology have been presented by [3], [4], and [5].

As discussed in detail in [1], limitations of the available techniques are especially severe if measurements are desired on general conditions that may include non-planar surfaces, unknown flow direction, and significant pressure gradients. The present project, initiated at Wright Laboratory (WL) under the 1993 Air Force Office of Scientific Research Summer Research Program, is intended to develop and demonstrate an instrument that would overcome some of these limitations and thus become a useful addition to the repertoire of practical skin friction measurement techniques.

The instrument under development is based on the surface-obstacle principle, where the sensed physical quantity is the difference between the pressure on the face of a small obstacle placed on the surface and the local undisturbed static pressure; this pressure differential is calibrated against the shear stress exerted by the boundary layer on the flow in front of the obstacle. The calibration is expressible in terms of dimensionless variables that depend on the physical flow parameters at the wall and the size of the probe. Examples of devices operating on this principle are the surface blocks, sublayer fences, Stanton tubes and Preston tubes, as discussed in the surveys referenced above.

The novel features of the proposed surface-obstacle instrument are twofold: (a) adjustable operation using the principle of minimum protrusion required to sense the pressure differential with satisfactory accuracy and (b) capability of sensing the direction of the flow adjacent to the surface through the angular location of the face pressure pattern given by orifices evenly spaced around the axisymmetric obstacle.

The principle of minimum protrusion was adopted (a) to avoid disturbing the flow in the boundary layer more than absolutely necessary, especially in measurements related to laminar flow control or characterization of turbulent flow structures; (b) to minimize shear-stress measurement errors caused by surface static pressure gradients, as will be discussed in the following; (c) to minimize effects of exposure to hostile environments, with the option of withdrawing the probe before and after the measurement; and (d) to provide a realistic indication of the limiting flow direction at the surface.

The 1994 results of the project presented in this final report consist of the description of the detailed design of the instrument approved for fabrication at WL, and of the test programs proposed for the WL supersonic M3 and M6. In addition, a brief summary is presented of the preliminary test conducted in the Washington University (WU) Low-Speed Wind Tunnel under private funding in April, 1994, as background for the planned incompressible calibration of the WL instrument in the same facility in 1995.

Technical Background

Relative to the well-established techniques for direct local measurement of other fluid-dynamic parameters at the surface, such as pressure and temperature, determination of skin friction presents specific difficulties which have retarded the comparable development of the required experimental technology. As discussed in the survey articles referenced in the Introduction, effective instruments for direct measurement of the surface shear stress do exist and have been used successfully for wind tunnel and flight testing up to hypersonic Mach numbers. However, these tests have generally been limited to simple configurations, such as flat plates, where no significant streamwise pressure gradients are present.

The fundamental problems of direct force measurement under general conditions were discussed in detail in [1], where it was especially emphasized that for the only direct-measurement technique, the floating surface element, the relative magnitude of the pressure-gradient-induced gap-force is independent of the streamwise size of the floating element; hence, because of the dominance of this error under general flow conditions, it must be carefully evaluated even in miniaturized designs where the depth of the surrounding isolation gap is significantly reduced below the dimensions found in traditional floating-element configurations.

In view of this background, it appeared that a directionally sensitive surface-obstacle sensor, adaptable for use in pressure gradients and on curved surfaces, and designed to introduce minimum disturbance to the boundary layer, would be a worthwhile addition to the repertoire of skin friction measurement techniques.

General Calibration of Surface Obstacle

As demonstrated in [1], all practical calibration relationships can be expressed in terms of dimensionless variables containing fluid properties at the wall: density ρ, dynamic viscosity μ, kinematic viscosity ν, and static pressure p_e; the surface shear stress τ; the protrusion height of the obstacle h; and the difference Δp between the pressure on the face of the obstacle and p_e. The surface static pressure p_e is normally assumed equal to the local static pressure outside the boundary layer..

The general calibration relationships can then be expressed as

$$\tilde{p} = f(\tilde{\tau}, \tilde{p}_e) \tag{2}$$

where $\tilde{p} = \Delta p h^2 / \rho \nu^2$, $\tilde{p}_e = p_e h^2 / \rho \nu^2$, and $\tilde{\tau} = \tau h^2 / \rho \nu^2$.

The conversion of available calibration to the form (2) was reviewed and the resulting general relationships presented in [1]. It was expected that the surface-obstacle design developed in this project would be characterized by a similar calibration pattern; indeed, as discussed later in this report, the preliminary results obtained in proof-of-concept prototype tests in the WU Low-Speed Wind Tunnel in a laminar and a turbulent boundary layer fully supported this premise.

For the purpose of selecting the size of a surface obstacle for a given flow situation, the general calibration relationships was expressed [1] as

$$\frac{\Delta p}{p_e} = \frac{\tilde{p}}{\tilde{p}_e} = \frac{1}{\tilde{p}_e} F\left[\frac{\tau}{p_e}\tilde{p}_e, \tilde{p}_e\right] = \bar{F}\left[\frac{\tau}{p_e}, \tilde{p}_e\right] \tag{3}$$

where the probe size, h, appears in only one parameter, $\tilde{p}_e$. The dimensionless parameter $\tilde{p}_e$ is introduced to the calibration relationship (3) by compressibility effects, and disappears from most equations in their absence. However, in form (3) $\tilde{p}_e$ is always present because of the normalization of τ and Δp by p_e. A plot in these variables was also presented in [1], primarily on the basis of the extensive data on Preston-tubes, and it is anticipated that upon completion of the calibration tests a similar guide chart can be prepared for selecting the proper protrusion of the proposed adjustable instrument.

As shown in [1] for Preston-tube calibrations, compressibility effects were in most cases included by introducing the concept of probe Mach number M_p, which is defined by the isentropic face pressure differential to static pressure ratio

$$\frac{\Delta p}{p_e} = \frac{\tilde{p}}{\tilde{p}_e} = \left[1 + \frac{\gamma - 1}{2} M_p^2\right]^{\frac{\gamma}{\gamma - 1}} - 1 \tag{4}$$

or, at supersonic values of M_p, by the normal shock loss combined with isentropic expansion

$$\frac{\Delta p}{p_e} = \frac{\tilde{p}}{\tilde{p}_e} = \frac{\left[\frac{\gamma + 1}{2} M_p^2\right]^{\frac{\gamma}{\gamma - 1}}}{\left[\left[\frac{\gamma - 1}{\gamma + 1}\right]\left[\frac{2\gamma}{\gamma - 1} M_p^2 - 1\right]\right]^{\frac{1}{\gamma - 1}}} - 1 \tag{5}$$

Definition of a "probe dynamic pressure"

$$\tilde{q}_p = \frac{\gamma}{2} \tilde{p}_e M_e^2 \tag{6}$$

allowed the expression of the compressible calibration laws by replacement of $\tilde{p}$ by $\tilde{q}_p$, and a common calibration chart could then be prepared covering both compressible and incompressible regimes (Fig. 1 of [1]). As $M_p \to 0$, $\tilde{q}_p \to \tilde{p}$, and in most cases the parameter $\tilde{p}_e$ disappeared from the equation.

Very few data were available to determine the presence of compressibility effects in the Stanton-range. Such effects have been observed [5] but not consistently in other experiments. As a working hypothesis, the use of the probe dynamic pressure, $\tilde{q}_p$, in place of $\tilde{p}$ appeared reasonable in this regime for compressible boundary layers. It was also anticipated that differences in calibration would probably occur between laminar and turbulent boundary layers.

Characterization of Surface Obstacle Calibration

Using the data collected and analyzed in [1] as the working hypothesis, the following expressions were proposed for general characterization of surface-obstacle calibration:

$12{,}600 < \tilde{\tau} < 100{,}000$:

$$\tilde{q}_p = 35.55\,\tilde{\tau}^{1.13} \tag{7a}$$

$200 < \tilde{\tau} < 12{,}600$:

$$\log_{10}\tilde{\tau} = 2.7741 - 1.1106\log_{10}\tilde{q}_p + 0.3234\left[\log_{10}\tilde{q}_p\right]^2 - 0.0177\left[\log_{10}\tilde{q}_p\right]^3 \tag{7b}$$

$10 < \tilde{\tau} < 200$:

$$\tilde{q}_p = 1.117\,\tilde{\tau}^{\,5/3} \tag{7c}$$

$\tilde{\tau} < 10$:

$$\tilde{p} \approx 1.2\tilde{\tau} \tag{7d}$$

The calibration correlations in the two upper ranges have been established for turbulent boundary layers; in the two lower ranges of $\tilde{\tau}$, both laminar and turbulent boundary layers may be found, and care must be exercised in using calibration relations not established for the particular experimental conditions.

As discussed in (1), the integration of the Preston tube calibration laws with the lower range was based on certain assumptions and extensions of the available data base .In the specific case of round tubes, the 5/3-power relationship has not been directly demonstrated over most of its expected range. For Stanton tubes and sublayer fences, the 5/3-power law rests on firmer ground; however, in neither case is the question of compressibility effects entirely clear.

There exists a vast array of calibration data for various types of surface devices where specific power laws have been identified, the 4/3-power relationship being a common choice. Generally, such experiments can be placed in the transitional regime between the 5/3-power and the 1.13-power regions, and thus these relationships are likely to represent tangent lines on the logarithmic plot of the general calibration laws within the range of a particular experiment. This particular question was part of the motivation for an exploratory series of low-speed measurements with a simplified proof-of-concept prototype instrument under private funding at Washington University in April, 1994.

Summary of Low-Speed Exploratory Measurements at Washington University.

The proof-of-concept prototype instrument designed and fabricated at Washington University was mounted on the existing flat plate installed in the 2 ft-by-2 ft Low-Speed Wind Tunnel in the Fluid Mechanics Laboratory. The 0.5 in-thick Plexiglas plate had a semi-elliptic leading edge with a continuous-curvature spline faired into the flat portion, to provide maximum extent of available laminar boundary layer flow.

The existing boundary-layer pitot probe positioning mechanism was used for adjusting the protrusion of the probe and reading it by means of the calibrated Linear Variable Differential Transformer system connected to the laboratory computer. The probe was a 10 mm-diameter flat-ended plug fitted tightly into a sleeve, which was installed flush with the plate surface. A 0.5 mm square groove extended approximately 5 mm down the side from the end face; the pressure differential between a pressure orifice located in the groove and the flat plate static pressure orifice at the same streamwise location on the plate was measured on a calibrated Validyne 3-in(water) full-range transducer with electronic readout. The configuration of the prototype instrument was similar to the final instrument shown in Figures 1 and 2, with the exception of only one groove being machined into the moving probe plug.

While an absolute measurement of the reference skin friction was not available, its value was estimated from pitot probe velocity-profile survey data matched by the laboratory computer to the theoretical Blasius solution in the laminar case or to the Clauser-chart version of the logarithmic law-region in the turbulent case. The main objective of the experiment was, however, to establish calibration curves at fixed values of shear stress for one laminar boundary layer condition, and for one turbulent boundary layer condition. This objective was achieved for running Reynolds numbers of approximately 750,000 and 1,750,000, respectively, and the functional dependence of the results was completely consistent with the general calibration relationships postulated in [1] and [2]. As expected, the absolute values of the constants in these relationships could not be determined without more accurate knowledge of the actual surface shear stress, but this verification of the functional relationships in incompressible flow is providing essential support to the overall prediction of the expected performance of the planned final instrument. The sensitivity of the instrument to flow direction was also explored in a qualitative sense.

The measured calibration curves, while following similar functional laws, showed a considerable difference between laminar and turbulent cases for the same value of the dimensionless shear stress, as had been already found by several other investigators [7], [8]. Various explanations have been offered, including experimental evidence in [8] on the rectification effect due to non-linear directionally-dependent flow phenomena in pressure tubing, as already mentioned in a more general context in [9]. Some effort was spent during the 1994 Summer Program to collect and analyze available data on the turbulent flow environment around the probe in an attempt to isolate a specific physical mechanism responsible for the observed difference, especially in view of the fact, derivable from close-to-surface turbulence measurements [10], that the surface shear stress in turbulent flow, while locally laminar, in reality fluctuates over a very wide range, possibly producing even momentary excursions to the negative range. However, continuation of this study toward definite conclusions must await resumption of the low-speed experiments under carefully controlled conditions.

Design of instrument for Wright Laboratory:

The specifications determined in [1] for the instrument to be used in the M3 and M6 tunnels included the following requirements: (a) mounting into the existing wall and flat-plate model locations; (b) adjustable-protrusion (maximum 5 mm), circular (10 mm diameter) obstacle to provide adequate face-pressure-differential and directional sensitivity with minimum disturbance to the boundary layer; (c) precision fit of the obstacle cylinder to prevent air leakage but provide for accurate, smooth extension and retraction of the device by direct manual or remote control; (d) evenly spaced pressure orifices (twelve) of minimum practical size located along edge of obstacle opening to provide both maximum face pressure for determination of shear stress and circumferential pattern for determination of flow direction; (e) provision for measurement of static pressure and surface temperature for direct calculation of dimensionless calibration parameters; (f) use of commercially available pressure transducers, and (g) an accurate internal or external means for measurement of the protrusion height of the obstacle. A sketch of the proposed instrument was presented in [1].

During the 1994 Summer Program, a detailed design satisfying these requirements was prepared and documented in a series of computerized (AUTOCAD-12) drawings. Representative sketches of these drawings are included as Figures 1 and 2.

For accurate positioning of the movable sensor element, a remotely controlled actuator with readout was chosen in view of the access environment to be encountered in the WL M3 and M6 wind tunnels. The BM4CC motor-driven micrometer/actuator with 4 mm total motion, manufactured by the Newport/Klinger Corporation, was selected as practically the single alternative consistent with the size and performance restrictions of the planned experiments. The readout resolution is better than 0.05 μm with an axial load capacity of 73 N. The actuator is controlled by a Newport/Klinger Motionmaster 2000 single-axis PC-based control and readout system. These parts have already been acquired by WL and will undergo checkout and calibration before installation into the basic instrument to be fabricated at WL.

The design described above and shown in Figures 1 and 2 is compatible with existing mounting provisions for both the M3 and M6 wind tunnels, and is basically interchangeable with the FIME floating element skin friction meter. A modification of the existing flow shield will also be required in the M6 installation to accommodate the protrusion of the actuator unit.

Calibration test program

After bench calibration of the instrument control/readout system, functioning of the complete unit will be verified for accuracy and repeatability.

In wind tunnel testing, the following signals will be recorded:

- Instrument position
- Seven to twelve differential pressure signals of varying magnitude
- Static pressure at surface of instrument mount
- Surface temperature of instrument mount

The following schedule is planned for each of the test series in the M3 and M6 wind tunnels:

- Installation and functional checkout of drive and sensors
- Three preliminary check runs, each at a fixed instrument deflection, stopping at the following stagnation pressures: 100, 200, 300, 400, and 500 psi. The objectives of this test are to determine time constants of instrument response, and to verify the orders-of-magnitude of the expected probe pressure signals.
- At each of the five stagnation pressures listed above, the instrument protrusion will be varied in a stepwise fashion. The exact number and magnitude of the steps in each run will depend on the time constant of the differential pressure sensing system, and in some cases more than one run may be necessary to obtain at least five data points at each stagnation pressure.

As a preliminary guide for the calibration tests, the predicted pressure differentials vs. probe protrusion are summarized in Figures 3 and 4. These plots, which are based on the extended calibration relationships presented in (7a) through (7d), are anticipated to be valid as functional trends and quantitatively within an order of magnitude.

The considerable data base existing from direct skin friction measurements in the M3 and M6 wind tunnels by means of floating-element gages mounted in the location of the proposed surface-obstacle device [10,11] will be used as the basic reference for the planned measurements.

Future work

Completion of construction of the instrument and the test series at WL are dependent on support for the principal investigator under the Summer Research Extension Program. The forthcoming proposal will also include provisions for resuming the calibration tests in the WU Low-Speed Wind Tunnel.

The ultimate development objective is a remotely adjustable instrument capable of skin friction and surface flow direction measurements in general conditions, including curved surfaces and the presence of arbitrary pressure gradients. Additionally, the simple design of the instrument combined with operation at minimal protrusion and complete retraction after measurement should facilitate measurements at elevated surface temperatures, possibly with the aid of an internal cooling system.

References:

1. Hakkinen, Raimo J., Skin Friction and Flow Direction Measurement by Surface-Obstacle Instruments, AFOSR Summer Faculty Research Program Report, August 1993.

2. Hakkinen, R. J., Calibration of Surface-Obstacle Skin Friction Meters, American Physical Society Fluid Dynamics Division Meeting, Albuquerque, NM, November 1993.

3. Rechenberg, I., Messung der turbulenten Wandschubspannung, Z. Flugwiss. Vol. 1, No. 11, Nov. 1963, pp. 429-438.

4. Winter, K. G., An Outline of the Techniques Available for the Measurement of Skin Friction in Turbulent Boundary Layers, Progress in Aerospace Sciences, Vol. 18, 1977, pp. 1-57.

5. Settles, G. S., Recent Skin Friction Techniques for Compressible Flows, AIAA 86-1099, May 1986.

6. Bradshaw, P., and Gregory, N., The Determination of Local Turbulent Skin Friction from Observations in the Viscous Sub-layer, ARC R & M 3202, H.M.S.O., London, 1961.

7. Fletcher, K. A., and Haritonidis, J. H., A Probe for Measuring Skin Friction in Disturbed Turbulent Boundary Layers, AIAA-92-0265, January 1992.

8. Bryer, D. W., and Pankhurst, R. C., Pressure Probe Methods for Determining Wind Speed and Flow Direction, H.M.S.O., London, 1971.

9. Klebanoff, P. S., Characteristics of Turbulence in a Boundary Layer with Zero Pressure Gradient, NACA Report No. 1247, 1955..

10. Galassi, L., and Scaggs, N., Assessment of CFD Predictions for Mach 6 Heat Transfer and Skin Friction, AIAA-91-5037, AIAA Third International Aerospace Planes Conference, Orlando, FL, 3-5 December 1991.

11. Wagner, M. J., Skin Friction and Heat Transfer Measurements in Mach 6 High Reynolds Number Flows, 15th International Congress on Instrumentation in Aerospace Simulation Facilities, Saint-Louis, France, 20-23 September, 1993.

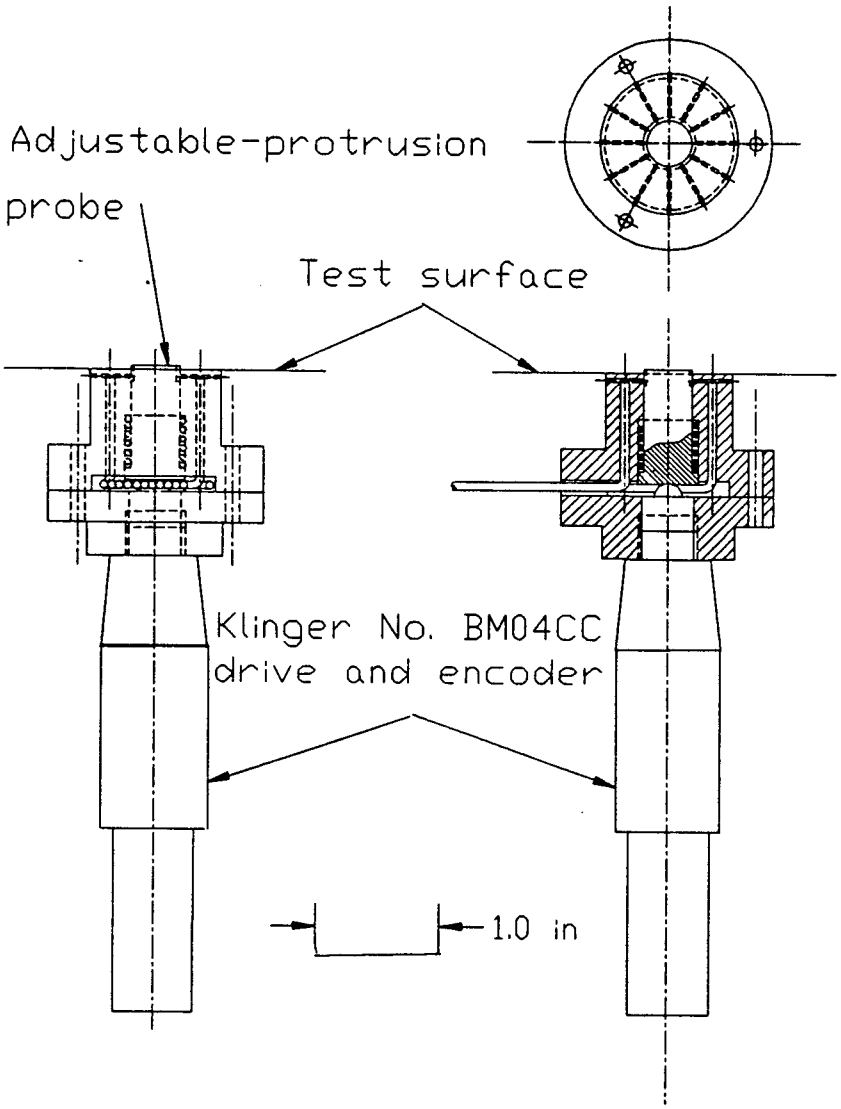


Figure 1. Skin-friction meter assembly for M3 and M6 tests.

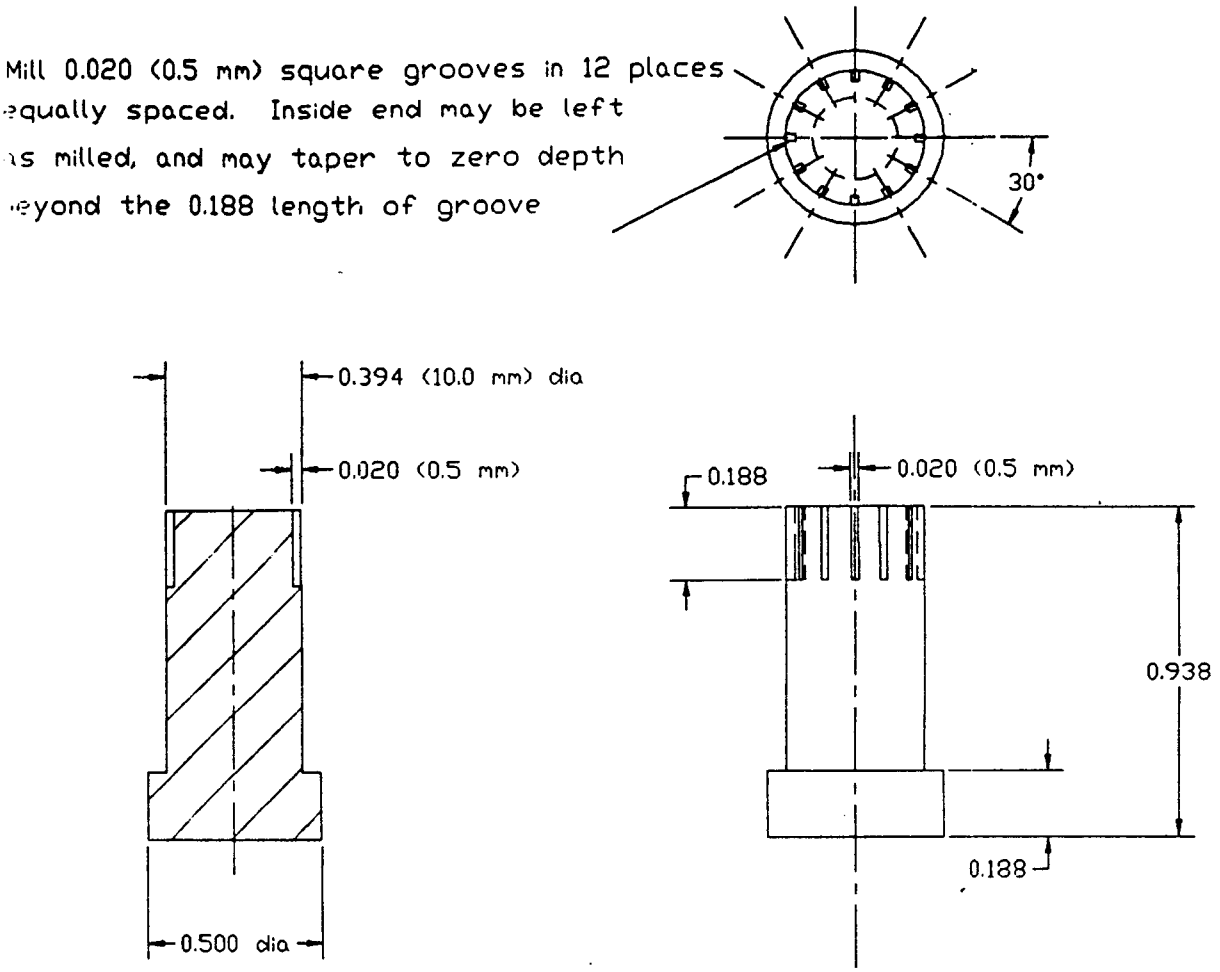


Figure 2. Details of adjustable-protrusion probe

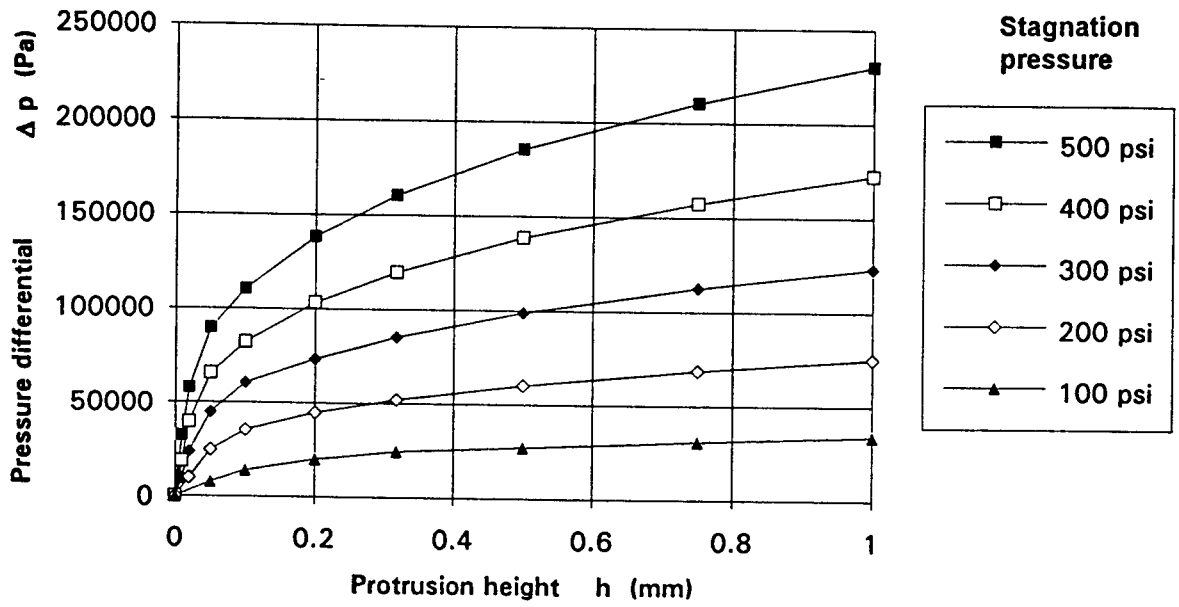


Figure 3. Estimated probe pressure differential vs. protrusion, M3

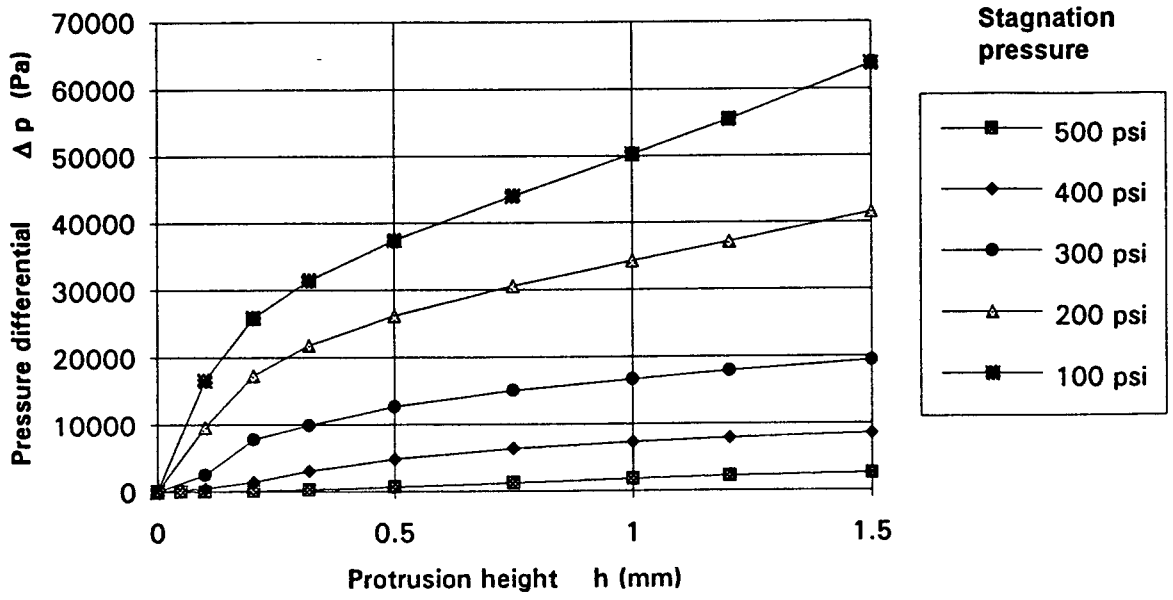


Figure 4. Estimated probe pressure differential vs. protrusion, M6

ADAPTIVE QUADRATIC CLASSIFIERS FOR MULTISPECTRAL TARGET DETECTION

Russell C. Hardie

Assistant Professor
Department of Electrical Engineering
University of Dayton
300 College Park Avenue
Dayton, OH 45469
(513) 229-3178

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

August 1994

ADAPTIVE QUADRATIC CLASSIFIERS FOR MULTISPECTRAL TARGET DETECTION

Russell C. Hardie

Assistant Professor
Department of Electrical Engineering
University of Dayton

Abstract

This paper investigates the use of an adaptive quadratic classifier for multispectral target detection. The system is designed to exploit both mean and covariance differences between the target and background. The detector proposed here is based on a Gaussian Bayes classifier where the local background statistics are estimated adaptively. Thus, these statistics need not be known *apriori*, and the detector will adapt to changing backgrounds. In addition, a forward sequential band selection method using the Bhattacharrya distance criteria is investigated here. This metric measures class separability due to mean and covariance differences. Also, a linear feature extraction technique is examined for data reduction prior to classification. Bomem spectrometer data and Landsat Thematic Mapper (TM) imagery are used to obtain preliminary results.

ADAPTIVE QUADRATIC CLASSIFIERS FOR MULTISPECTRAL TARGET DETECTION

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 [9].

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 long-wave infrared (LWIR) imagery ($8\mu m - 12\mu m$ range). Figure 1 shows a classification of the electro-magnetic spectrum for reference. FLIR systems employ a scanning or starring 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. The first is spectral band selection. That is, determining which spectral bands provide the best performance for target detection. One would also like to know how many bands and what bandwidths are needed for good performance. The problem appears to be that there may exist a unique set of bands which provide optimal discrimination for every possible target/background pair. Therefore, one would like to find a good "compromise" band set which can be used effectively for a large class of targets and backgrounds. Multispectral band selection (feature selection) using statistical distance measures has been studied in [12]. Band selection specifically for classification of soil organic matter content using a Karhunen-Loeve based method is presented in [3]. General feature selection methods are discussed in [2].

With the potential of having many spectral bands for target detection, it is important to perform feature extraction. This is the process of mapping the high dimensional data onto a lower dimensional space while retaining all the relevant information for classification. Band selection itself is a form of multispectral feature extraction. However, band selection and other feature extraction techniques are treated separately in this

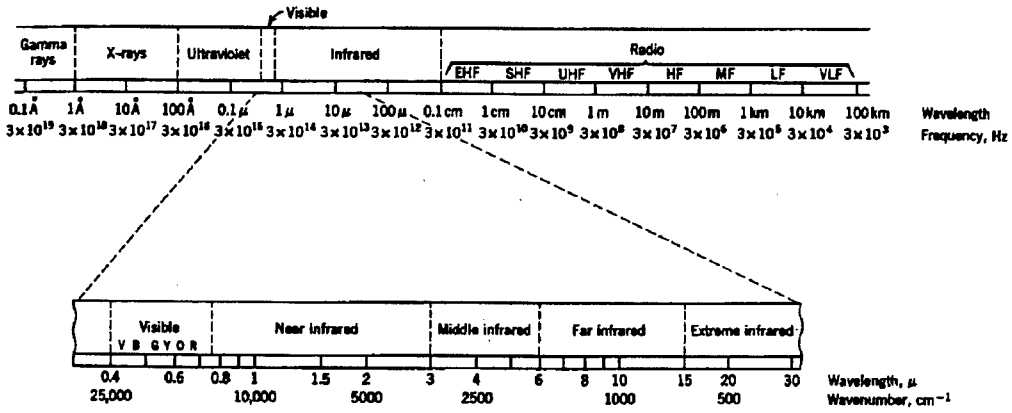


Figure 1: *Electromagnetic spectrum classification.*

paper. Multispectral feature extraction is the primary focus of the work presented in [1, 4, 5, 8]. Furthermore, The type of classifier used will drive the band selection and feature extraction process. The selected bands and the feature space must contain the most relevant information for a given classifier.

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 analyized in [10, 11, 13]. This method, based on a linear Bayes classifier, shows promising results provided that the target's spectral signature is known accurately. The algorithm is designed to detect a target within a single background and is optimal if the two classes (target and background) have Gaussian probability density functions (pdfs) and have equal covariance matrices. These assumptions make the design problem more tractable, however, they are not accurate in many cases. If the covariances of the target and background are different, the adaptive CFAR detector will yield suboptimal results. In this case, a quadratic classifier is needed for optimal performance. There is a trade off in terms of algorithm complexity and performance. If the two distributions are different, the burden of estimating these extra statistics is introduced for a quadratic classifier. However, it has been shown in [6, 7] that second order statistics are a powerful discriminant in high dimensional spectral data.

This paper will investigate the use of an adaptive quadratic classifier for multispectral target detection. The system is designed to exploit both mean and covariance differences. The detector proposed here is based on a Gaussian Bayes classifier where the local background statistics are estimated adaptively. Thus, these statistics need not be known *apriori*, and the detector will adapt to changing backgrounds. In addition, a band selection method using a statistical distance measure which reflects both mean and covariance differences is investigated. Finally, a linear feature extraction technique which preserves mean and covariance information is explored. Bomem spectrometer data and Landsat Thematic Mapper (TM) imagery are used

to provide some preliminary results.

This paper is organized as follows. In Section 2, some mathematical background on statistical pattern recognition is provided. Section 3 addresses the spectral band selection problem using a statistical distance measure approach. Linear feature extraction is discussed in Section 4. Adaptive quadratic classifiers for multispectral target detection are described in Section 5. In section 6, experimental results are presented. Finally, conclusions and discussion of possible future work are given in Section 7.

2 Background

Some mathematical background for statistical multispectral classification is provided in this section. An excellent treatment of statistical pattern recognition is provided in [2]. Before discussing classifiers, some notation must be introduced. Consider the case of a passive multispectral sensor which collects an N-band multispectral image (MSI) 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 and will only be used when necessary for clarity. 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). It will be assumed that the images are co-registered and are properly calibrated.

2.1 Spectral Classifiers

The following discussion 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 $\mathbf{x}$ (a single multispectral pixel) belongs to the background class and let hypothesis H_t be that $\mathbf{x}$ belongs to the target class. Each pixel in the MSI will be classified in this way creating a two dimensional class map.

The Bayes classifier which provides the minimum probability of classification error is defined by the following likelihood ratio:

$$l(\mathbf{x}) = \frac{p(\mathbf{x}|H_t)}{p(\mathbf{x}|H_b)} \overset{H_t}{>} \frac{P_b}{P_t}. \tag{2}$$

The function $p(\mathbf{x}|H_b)$ and $p(\mathbf{x}|H_t)$ are the conditional probability density functions for the observed spectral vector $\mathbf{x}$. The variables P_b and P_t are the *apriori* probabilities for H_b and H_t respectively. Equation (2) states that if the likelihood ratio $l(\mathbf{x})$ is greater than $\frac{P_b}{P_t}$, than one should declare that $\mathbf{x}$ belongs to the target class. Otherwise, $\mathbf{x}$ belongs to the background class. It is often useful to define the the minus-log-likelihood ratio from (2) yielding

$$-\ln\{l(\mathbf{x})\} = -\ln\{p(\mathbf{x}|H_b)\} + \ln\{p(\mathbf{x}|H_t)\} \overset{H_t}{<} \ln\frac{P_t}{P_b}. \tag{3}$$

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

$$h(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mu_b)^T\Sigma_b^{-1}(\mathbf{x}-\mu_b) - \frac{1}{2}(\mathbf{x}-\mu_t)^T\Sigma_t^{-1}(\mathbf{x}-\mu_t) \overset{H_t}{>} \ln\frac{P_t}{P_b} - \frac{1}{2}\ln\frac{|\Sigma_b|}{|\Sigma_t|}. \tag{4}$$

A Gaussian spectral distribution model will be assumed here. Thus, the quadratic classifier in (4) is the basis for the adaptive multispectral detectors defined in Section 5.

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

$$h(\mathbf{x}) = \frac{1}{2}(\mu_t - \mu_b)^T \Sigma^{-1} \mathbf{x} \underset{H_b}{\overset{H_t}{>}} \ln \frac{P_t}{P_b} - \frac{1}{2}(\mu_b^T \Sigma^{-1} \mu_b - \mu_t^T \Sigma^{-1} \mu_t). \tag{5}$$

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 [10, 11, 13]. 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.

2.2 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.

The criteria used for class separability and linear classifier performance in [10, 11, 13], is signal-to-clutter ratio (SCR). The SCR measure is effectively the same as B_1 when $\mu_b = \mathbf{0}$. Thus, the SCR criteria does not take into account class separability based on covariance differences.

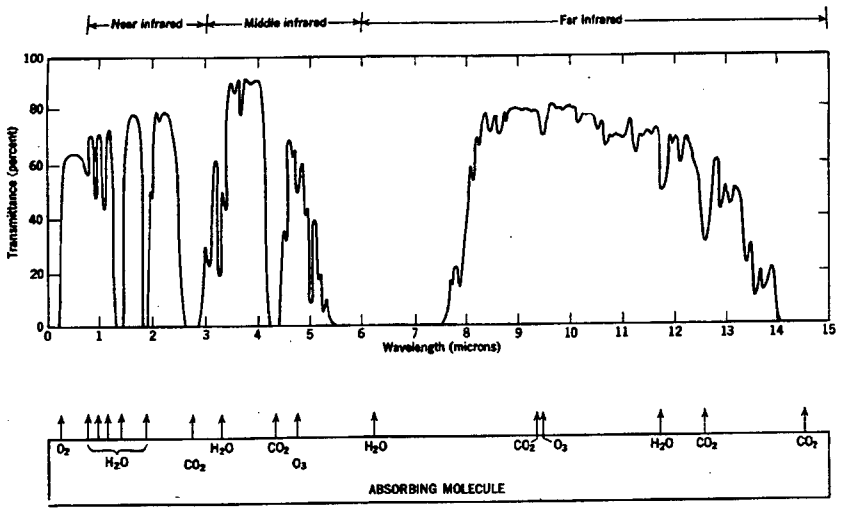


Figure 2: *Atmospheric transmittance curve.*

3 Band Selection Using B-Distance

A critical problem encountered when attempting to exploit multispectral imagery for target detection is wavelength band selection. Some wavelength bands will tend to provide greater class separability than others for specific targets and backgrounds. It may be impractical and expensive to field a hyperspectral sensor for every target detection mission. Furthermore, even if full hyperspectral data is available, it may be impractical to use a large number of bands for classification. Thus, one would like to find a relatively small set of bands which offer good discrimination for a large class of targets and backgrounds.

An empirical method for band selection involves the collection of a large set of finely spaced narrow spectral band data for a desired target and background. These data may be obtained from lab measurements, field measurements, or accurate spectral modeling. The B-distance (or other class separability measures) can be used to identify the most useful bands for class separation. These selected bands can then be used in a more large scale operational system. Bandwidth issues can also be investigated by starting with narrow bands and linearly combining these to simulate wider bands. In most cases, narrow band features offer the best object discrimination because they may be unique to a specific material. Such features get eliminated in wider band systems.

Before applying a statistical band selection technique, it is necessary to perform a preliminary band "screening." The object of the screening is to eliminate bands which do not provide consistent information about intrinsic properties of the target and background. For example, any bands which are severely altered by the atmosphere should be ruled out. Such bands will not tend to supply consistent information about the target and background. A plot of the atmospheric transmittance curve is shown in Fig. 2 for reference. Any bands in or near atmospheric absorbtion features may be unreliable for target detection.

Once the "bad" bands are eliminated from consideration, one would like to find an optimal band set for target detection. Assuming sufficient training data from the target and background are available, then the

Table 1: *Forward sequential spectral band selection method using the B-distance.*

- **Step 1:** Select the band which has the highest individual B-distance of the N candidate bands.
- **Step 2:** Pair each of the remaining $N-1$ candidate bands with the one selected in step 1. Select the band which yields the highest 2-dimensional.
- **Step 3:** Pair the remaining $N-2$ bands with the two selected and choose the band which yields the highest 3-dimensional B-distance.
- **Step 4:** Continue in this fashion until J bands have been selected.

optimal band set can be selected using the B-distance criteria. Let there be N candidate bands of which J are to be selected. The number of possible J band combinations which could be used for detection is given by $N!/(N-J)!$. An exhaustive search over all band combinations using the B-distance measure would yield the optimal J band combination. However, this is impractical for anything but small values of J. A more practical, but sub-optimal, method is outlined in Table 1. This is a forward sequential feature selection algorithm [2]. The number of B-distances to be computed with this method is given by

$$J(N - \frac{1}{2}J + \frac{1}{2}). \tag{10}$$

The forward sequential band selection method in Table 1 can be used with any class separability metric. The particular metric used will depend on the classifier to be employed. The metric need only capture information about class separability that is exploited by the classification algorithm. For a quadratic classifier, the B-distance is a good measure. However, for a linear classifier, the B_1 metric is appropriate.

This forward sequential method does a good job in most cases. However, if the globally optimal band set does not contain the first selected feature, the solution may be far from optimal. One new modification, which can make the algorithm more robust, is to do an exhaustive search to find the optimal K band set, where $1 \leq K \leq J$. The forward sequential method can then be used to find the remaining $J-K$ bands. This compromise approach requires

$$\frac{N!}{(N-K)!} + (J-K)\left[N - \frac{1}{2}(J-K) + \frac{1}{2}\right] \tag{11}$$

distance computations.

To compute the B-distance, the mean and covariance of the target and background are required. In most practical cases, these will have to be estimated from empirical 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

$$\hat{\mu} = \frac{1}{P}\sum_{k=1}^{P} \mathbf{x}_k. \tag{12}$$

The unbiased sample covariance estimate is given by

$$\hat{\Sigma} = \frac{1}{P-1}\sum_{k=1}^{P}(\mathbf{x}_k - \hat{\mu})^T(\mathbf{x}_k - \hat{\mu}). \tag{13}$$

A "rule of thumb" is to have $\mathcal{O}(J^2)$ samples, where J is the number of spectral bands, with which to make the mean and covariance estimates [2].

4 Feature Extraction

Given an N-band multispectral data set, one may wish to reduce the classifier input by performing feature extraction. The goal of feature extraction is to take the R^N observation space and map it to a lower dimensional space R^M *without* losing classifier performance (class separability). This will tend to make the classifiers more robust in practice and simplier to implement. Again, spectral band selection is a form of feature extraction, but will be treated separately here. However, the "bad" bands should be eliminated prior to feature extraction. If a small number of bands selected using the methods described in Section 4 provide sufficient class separability, additional feature extraction may not be necessary. In this case the selected bands can be used directly as the classifier input.

The optimal feature space mapping may, in general, be nonlinear. However, for the sake of tractability, the mapping is often restricted to be linear. That is, the transformation is of the form

$$\mathbf{y} = A^T\mathbf{x}, \tag{14}$$

where A is an $M \times N$ matrix (its column vectors should be linearly independent but not necessarily orthogonal). Note that $\mathbf{x} \in R^N$ and $\mathbf{y} \in R^M$ where $M < N$. Thus, the linear transformation maps $R^N \mapsto R^M$.

If the B-distance is used as the measure of class separability and the target and background covariances are equal ($\Sigma_t = \Sigma_b$), then only the first term of the B-distance, shown in (8), need be considered. Ignoring the second term makes the B-distance essentially the same as the SCR criteria used in [11]. It can be shown that in this case only a one-dimensional feature space is needed for optimal classification [2]. That is, there exists a $1 \times N$ matrix A such that the B-distance for the classes in the transformed space is the same as that for the original N space. This optimal linear transformation is given by

$$y = (\mu_t - \mu_b)^T\Sigma_b^{-1}\mathbf{x}. \tag{15}$$

Note that this is the same result as that obtained by direct implementation of the discriminant function in (5). Unfortunately, however, when $\Sigma_t \neq \Sigma_b$, there is no known way to find an optimal linear transformation using the B-distance criteria. Thus, one must use a sub-optimal method for finding A (or let $A = I$ and rely on good band selection). One suboptimal method for linear feature extraction presented in [2] is investigated here and is described below. This method is most effective when $B_1 >> B_2$.

- **Step 1:** Compute the Eigen vectors ϕ_i and Eigen values λ_i of $\Sigma_{avg}^{-1}(\mu_b - \mu_t)(\mu_b - \mu_t)^T$, where $\Sigma_{avg} = (\Sigma_t + \Sigma_b)/2$ and $i = 1, 2, \ldots, N$. The rank of the matrix is one, so only one Eigen value will be non-zero. Define the Eigen vector corresponding to this non-zero Eigen value as ϕ_1. Note, ϕ_1 is the optimal linear discriminant function in (15). All information of class separability due to mean-difference is retained in the feature produced by the linear transformation $y = \phi_1^T\mathbf{x}$.

- **Step 2:** The remaining Eigen vectors are orthogonal to ϕ_1 and can provide a linear transformation which projects x into a space where there is no separability due to mean difference. Let these remaining Eigen vectors be written as a matrix

$$D = \left[\begin{array}{cccc} \phi_2 & \phi_3 & \cdots & \phi_N \end{array} \right]^T . \tag{16}$$

The covariances for the target and background in this space are given by $\bar{\Sigma}_t = D^T \Sigma_t D$ and $\bar{\Sigma}_b = D^T \Sigma_b D$. Compute the Eigen vectors ψ_i and Eigen values γ_i of $\bar{\Sigma}_b^{-1} \bar{\Sigma}_t$. Select the Eigen vectors which correspond to the $M-1$ largest $(\gamma_i + 1/\gamma_i + 2)$ terms. These Eigen vectors form a linear transformation which provides maximum class separability due to covariance difference in the space orthogonal to the space containing all the mean difference information.

- **Step 3:** The overall linear transformation matrix A is given by

$$A = \left[\begin{array}{cccc} \phi_2 & \phi_3 & \cdots & \phi_N \end{array} \right] \left[\begin{array}{ccccc} 1 & 0 & 0 & \cdots & 0 \\ 0 & & & & \\ 0 & \psi_1 & \psi_2 & \cdots & \psi_{M-1} \\ \vdots & & & & \\ 0 & & & & \end{array} \right] . \tag{17}$$

5 Adaptive Quadratic Classifiers

In this section, an adaptive multispectral classifier for target detection is proposed. Since different levels of *apriori* information may be available, two cases are considered here:

1. Known target and known background.
2. Known target and unknown background.

In the first case, it will be assumed that the mean and covariance of the target and background are known *apriori*. This may be difficult to obtain in a realistic application, but, should offer optimal performance. The second case is where the mean and covariance are known for the target but not for the background. Since the target may be in a variety of backgrounds, even within one geographic area, background statistics may be very difficult to obtain *apriori*. For this case, an adaptive background estimation technique is proposed.

The classifiers discussed below can be implemented in the J selected spectral band space or in some feature space. For simplicity, the following discussion and results focus on implementing the classifiers in J selected spectral band space.

5.1 Known Target with Known Background

If the target and background have normal distributions in the spectral dimension and the mean and covariance for each are known *apriori*, then an optimal spectral quadratic classifier is straightforward to construct. Again, this technique will yield results superior to those of a linear classifier, provided that the target and background have different covariance matrices. A block diagram of an implementation of such a classifier is shown in Fig. 3.

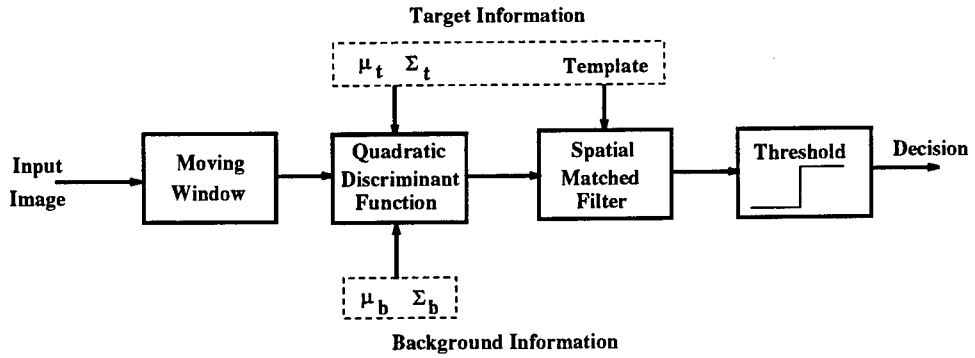


Figure 3: *Block diagram of a target detection system where a full statistical description of the target and background is available apriori.*

The J band multispectral image is the input to the system. A moving spatial window spans a $L_1 \times L_2 = K$ set of pixels. Specifically, for the system in Fig. 3, $L_1 = L_2 = 1$. Therefore, the window only spans a single spatial pixel x, which is a $1 \times J$ spectral vector. The window moves across the image in a raster scan fashion and at each window location two hypotheses are tested:

$$\begin{aligned} H_t: &\quad \mathbf{x} \text{ belongs to target class} \\ H_b: &\quad \mathbf{x} \text{ belongs to background class.} \end{aligned} \tag{18}$$

Assuming that the target and background have normal distributions in the spectral dimension with known second order statistics, the optimal spectral quadratic discriminant function is given by

$$h(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mu_b)^T \Sigma_b^{-1}(\mathbf{x} - \mu_b) - \frac{1}{2}(\mathbf{x} - \mu_t)^T \Sigma_t^{-1}(\mathbf{x} - \mu_t). \tag{19}$$

Given training data, the mean and covariance for the target and background can be computed using the sample estimates in (12) and (13) respectively. Once the output of the quadratic discriminant function is computed, the threshold in (4) could be applied for pure spectral classification. This threshold is given by

$$T = \ln\frac{P_t}{P_b} - \frac{1}{2}\ln\frac{|\Sigma_b|}{|\Sigma_t|}. \tag{20}$$

If *apriori* information about the spatial shape of the target is known, one can capitalize on this. Let the output over the entire image of the quadratic discriminant function be denoted as $\{d(n_1, n_2)\}$. A moving window spatial matched filter can be applied to $\{d(n_1, n_2)\}$. The matched filter coefficients are defined by the projection of the target onto the discriminant space $\{d(n_1, n_2)\}$. If one models the discriminant space as the target with additive white noise, the matched filter will maximize the spatial target signal-to-noise ratio in the discriminant space.

It is important to realize with this detection scheme that it may be difficult to collect full accurate information about every possible background in which one might find a target. Worse yet is that in any reasonably sized geographic area, a target may be found in a number of different backgrounds. For this reason, it would be useful to be able to estimate the background statistics "on-the-fly." The following subsection defines a quadratic detector which estimates local background statistics adaptively. Thus, these statistics need not be known *apriori* and the detector can adapt to changing backgrounds.

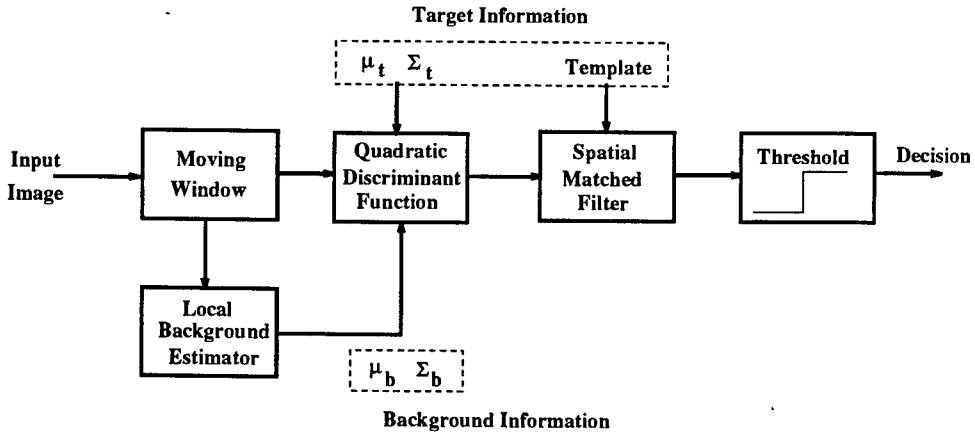


Figure 4: *Block diagram of an adaptive target detection system where the local background statistics are determined "on-the-fly."*

5.2 Known Target with Unknown Background

In many practical applications of target detection, the target size will be small compared to the background. This information can be used in classifier design. Consider a moving observation window which is large with respect to the target. That is, the number of target pixels spanned by the window is small compared with the number of background pixels. Thus, one can use these observed samples to form an initial estimate of the background statistics. If the target occupies a significant portion of the observation window, an iterative processes may be needed for form an accurate estimate of the background statistics. Using this basic idea, an adaptive target detection system can be implemented which does not require *apriori* information about the background. This system is illustrated in Fig. 4.

Specifically, in this system the background statistics are estimated using the $L_1 \times L_2 = K$ samples spanned by a moving window. An iterative classification process, summarized in Table 2, is used to separate background pixels from target pixels. The entire observation window of multispectral pixels can be classified using the procedure in Table 2 on the last iteration. The window can then be moved across the image in a non-overlapping fashion to classify the entire image. An alternative is to compute the discriminant for the center sample and move the window in a overlapping fashion. This would increase the processing required by a significant amount, but may provide a superior local background estimate. As before, a spatial matched filter can be used with this detector given the availability of *apriori* spatial information.

The window size is critical for good performance of the adaptive quadratic detector. The window should span $\mathcal{O}(J^2)$ background samples at all locations to provide a sound background mean and covariance estimate. As with the previous method, the background pixels in any window are assumed to have been derived from a single Gaussian distribution. Any deviation from this in practice will likely degrade detector performance. If an observation window spans multiple background classes, the mixed background class may be highly non-Gaussian and the pdf may not even be unimodal. Treating multiple unknown background classes is a more difficult problem but could possibly be done by extending the method in Table 2. This extension could use multiple initial background estimates in step 1 and multiple classifications in step 2.

Table 2: *Iterative method for estimation of local background statistics.*

- **Step 1:** Compute the sample mean and covariance estimates using all of multispectral pixels in observation window. This is an initial background estimate. If the target is small compared with the window size, one can stop here.
- **Step 2:** Classify all the pixels within the window based on the *apriori* target mean and covariance and the previous estimate of the background statistics. This is done with the quadratic classifier defined in (4).
- **Step 3:** Recompute the background statistics using those multispectral pixels classified as background pixels in step 2.
- **Step 4:** Return to step 2 unless the pixel classification has not changed since the last iteration or a predetermined maximum number of iterations is reached.

6 Experimental Results

In this section some preliminary results are presented. Bomem spectrometer data and Landsat TM data are used. The forward sequential band selection method using B-distance is applied to the Bomem spectrometer data. Band selection, feature extraction, and target detection are tested using the landsat TM data.

6.1 Bomem Spectrometer Data Analysis

High spectral resolution data obtained using a Bomem Infrared Fourier transform spectrometer (FTS) are used to illustrate the band selection method using B-distance. The data collected consist of multiple single-point calibrated radiometric measurements over a spectral range of $2.86\mu m - 14.32\mu m$. The spectrometer provides 728 narrow spectral band measurements within this range for each spatial sample.

Figure 5a shows mean radiance data for a set of samples of a net camouflaged tank and desert scrub over the range $3\mu m - 12\mu m$. The data were acquired at White Sands Missile Range in January, 1994. The difference between the two spectral response patterns is plotted in Fig. 5b. Spectral mean differences provide a basic first order discriminant.

The individual band B-distances are plotted for the LWIR bands ($8\mu m - 12\mu m$) in Fig. 6a. In addition to the overall B-distance, the B_1 and B_2 distances are also plotted. Note that for the individual bands, the B_2 term is dominant. Thus, using most individual bands, variance difference is a more significant discriminant than mean difference. However, these individual band B-distances are relatively low.

A scatter plot showing camouflaged tank and desert scrub samples in the two selected band space is shown in Fig. 6b. The first selected band is at $11.69\mu m$. This can clearly be seen as the peak in Fig. 6a. Note that this band has a small B_1 distance (i.e., little mean difference). The second band selected using the forward sequential band selection method is the 11.29μm. This band does have a relatively large B_1 distance. The optimal two band pair determined with an exhaustive search has only a slightly increased B-distance. The optimal two band pair is 11.73μm and 11.29μm and the B-distance is 1.7481.

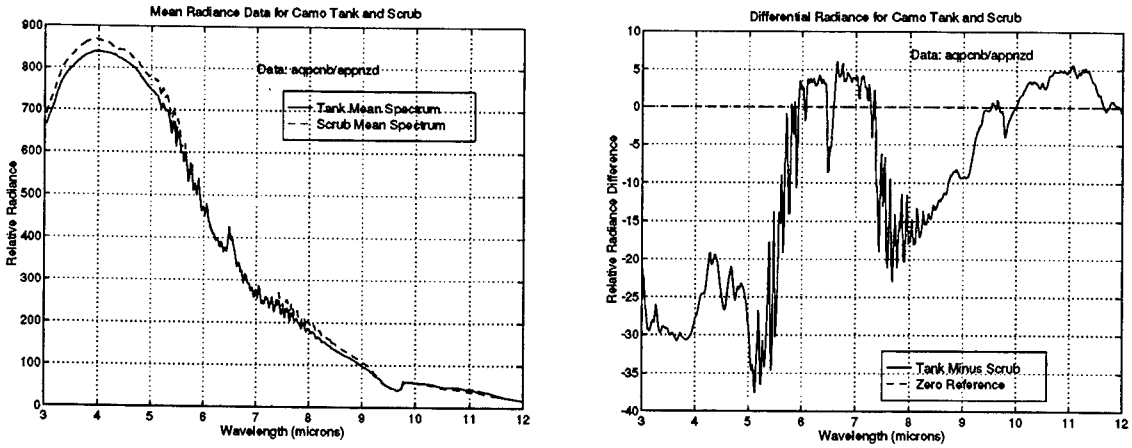


Figure 5: *(left)[a] Mean Bomem radiance data for a camouflaged tank and desert scrub over the range* $3\mu m - 12\mu m$. *(right)[b] The tank mean spectrum minus the desert scrub mean spectrum.*

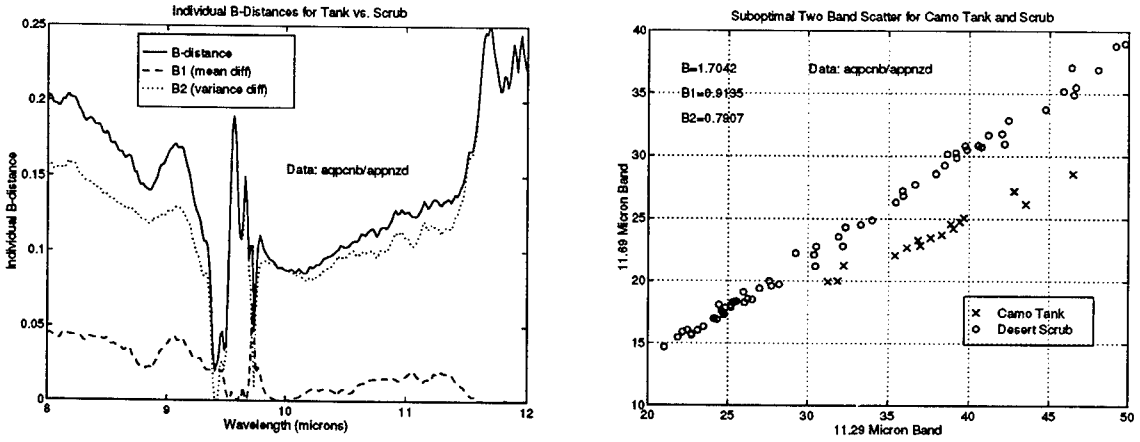


Figure 6: *(left)[a] Individual* $8\mu m - 12\mu m$ *band B-distances for the camouflaged tank vs. desert scrub. (right)[b] Scatter plot showing the tank and scrub in the two band space determined using the forward sequential band selection method.*

Notice that the two band B-distance is close to seven times as large as that for the best individual band. Thus, the use of two bands dramatically increases the class separability of the camouflaged tank and desert scrub samples. Also note that approximately 46% of the total B-distance comes from B_2, the covariance difference. This result supports the notion that exploiting second order statistics may be valuable for target discrimination. Because of the limited number of Bomem multispectral samples available for the target and background, accurate mean and covariance estimates can not be made using more than two bands (recall $\mathcal{O}(J^2)$ samples are required).

6.2 Landsat TM Data Analysis

Landsat TM imagery was used to obtain some preliminary results using the adaptive quadratic detector. Band selection and feature extraction are also explored with this data set. Landsat TM imagery has seven spectral bands and information about these bands is provided in Table 3. The image was acquired over a

Table 3: *Landsat TM band information*

TM Band	Wavelength (μm)	Bandwidth	GSD (meters)
1	0.47	0.075	30
2	0.57	0.075	30
3	0.66	0.06	30
4	0.82	0.13	30
5	1.65	0.20	30
6	11.50	2.1	120
7	2.20	0.25	30

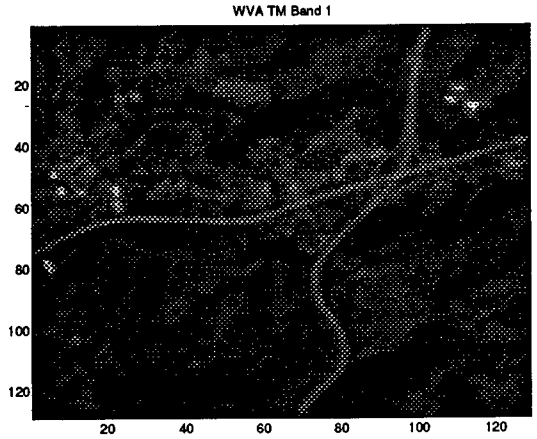


Figure 7: *TM band 1 for a region in West Virginia. Note that this band provides the best individual band discriminant for road vs. vegetation.*

heavily vegetated area in West Virginia containing a major highway intersection. Band 1 (blue radiance) is shown in Fig. 7.

6.2.1 Band Selection

The mean radiances for identified vegetation and road samples versus the TM band number are plotted in Fig. 8a. The road and vegetation samples were "hand picked" based on subjective interpretation of the scene. One standard deviation above and below the mean is also plotted to show the spread in each band. Note that TM band 1 shows significant mean difference for the two classes and the standard deviation for both classes is relatively small. This fact allows for good class separability in band 1. On the other hand, band 7 shows significant mean difference but the standard deviation for each class is large, which hurts class separability. In bands 4, 5 and 6, the vegetation has a slightly larger variance than the road. This provides a second order discriminant where little mean difference exists.

Individual TM band B-distances for road versus vegetation are plotted in Fig. 8b. This figure shows that TM band 1 is the best single band for discriminating road and vegetation according to the B-distance criteria. Note that when a big mean difference exists, B_1 is large. In addition, a small variance amplifies B_1 (e.g., compare band 1 and 7). Also note that $B_1 \approx 0$ for bands 4, 5 and 6, while $B_2 > 0$ for these bands.

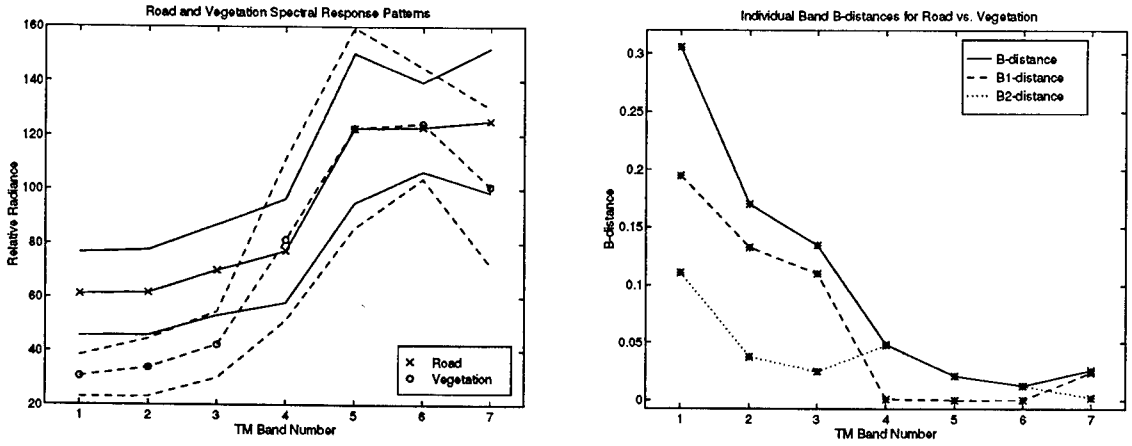


Figure 8: *(left)[a] Mean radiance vs. TM band number for road and vegetation. One standard deviation above and below the mean is also plotted to show the spread in each band. (right)[b] Individual TM band B-distances for road vs. vegetation.*

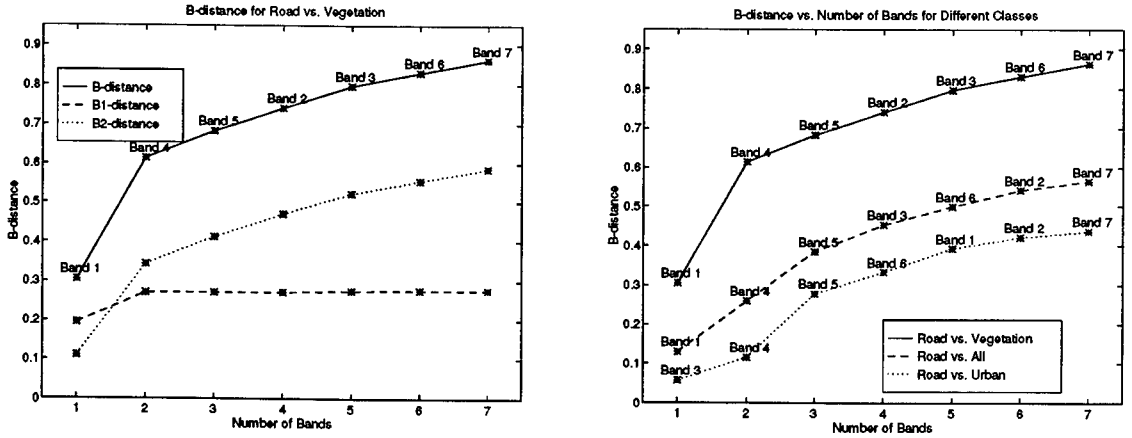


Figure 9: *(left)[a] B-distance as a function of the number of bands used for road vs. vegetation. The bands were added according to the forward sequential band selection method. (right)[b] B-distance for Road vs. three different background classes.*

The B-distance as a function of the number of bands used for road versus vegetation is plotted in Fig. 9a. The bands were added according to the forward sequential band selection method in Table 1. Notice that band 4 is selected for the second band even though the individual B-distances for bands 2 and 3 are greater. This selection is due to the fact that bands 1, 2 and 3 are highly correlated for these two classes. In other words, bands 2 and 3 do not provide much new information for class separability where band 4 does. Also, notice that adding the remaining bands does not increase B_1 very much. Thus, a linear classifier will not benefit significantly from these extra bands.

The B-distance for road versus three different background classes is plotted in Fig. 9b. The "vegetation" class is the same as before. The "all" class consists of the entire scene in Fig. 7 excluding the major roads. Finally, the "urban" class comes from a dominantly urban area. Note that road versus vegetation yields the largest B-distance and road versus urban gives the lowest. This is reasonable since the urban area contains

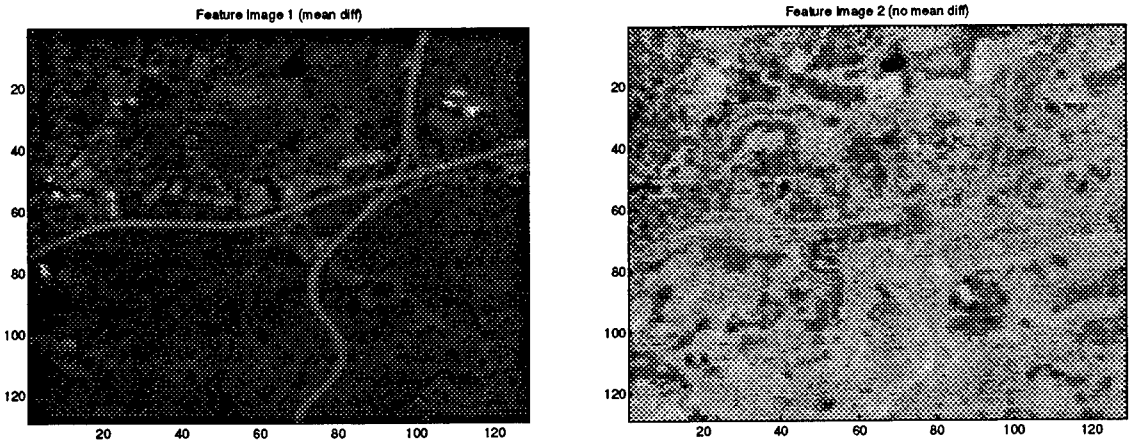


Figure 10: *(left)[a] Feature image 1 which preserves all information about class mean difference. (right)[b] Feature image 2 which contains no information about mean difference.*

smaller roads and concrete. Notice that the best bands for each background differ, particularly for the urban background class.

6.2.2 Feature Extraction

The application of the feature extraction method described in Section 4 to Landsat TM data is considered in this subsection. Using all seven bands as the input, the top two features are extracted for discrimination of road and vegetation. Feature image 1 is shown in Fig. 10a. This feature preserves all information about class mean difference. The major road stands out clearly from the vegetation, which appears flat in this feature. Feature image 2 is shown in Fig. 10b. Here there is no mean difference between the road and vegetation. However, one can detect that there is little variability among the road pixels and significant variability in the background. This affect provides an additional second order discriminant which a quadratic classifier can exploit. A linear classifier would only utilize feature image 1.

Scatter plots showing road and vegetation samples in the optimal two band space and feature space are shown in Figs. 11a and 11b respectively. In Fig. 11b, the horizontal axis shows the first feature which contain all information due to mean difference. The vertical axis shows a feature with covariance difference only. Note that the B-distance for the feature space classes is 0.6616 compared to 0.6140 for the two band space. Thus, using a linear mapping from all the bands does improve the road/vegetation class separability over using the best two bands. However, in this case the increase in modest.

6.2.3 Classification

Application of the multispectral detectors to TM data is considered in this subsection. The classifiers are used in the TM band 1 and 4 space. The target class is defined to be the road class and the background is "all" class used in Fig. 9b. The output of a non-adaptive linear and quadratic classifiers is shown in Fig. 12a and 12b respectively. The *apriori* probabilities have been estimated from the imagery and are set to be $P_t = .025$ and $P_b = .975$ for both classifiers. The optimal thresholds were used and no spatial matched

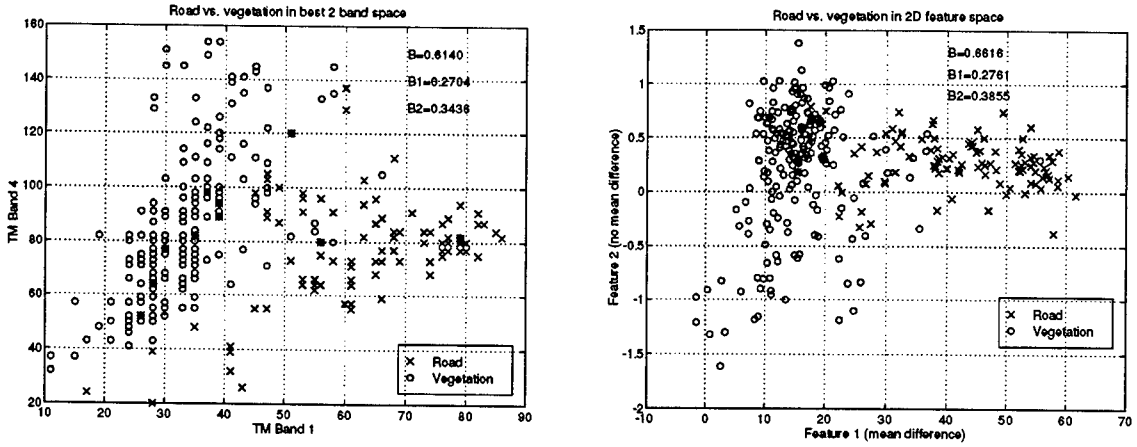


Figure 11: *(left)[a] Scatter plot for road and vegetation in the optimal two band space. (right)[b] Scatter plot showing the classes in a two-dimensional feature space.*

filtering was performed. Both classifiers appear to perform well.

The output of the adaptive quadratic detector for a known target and unknown background (see Fig. 4) is shown in Fig. 13. A window size of 15 × 15 with five iterations of the procedure in Table 2 is used to estimate the background statistics. The *apriori* probabilities are set to be $P_t = .15$ and $P_b = .85$. These *apriori* probabilities differ from those used for the non-adaptive classifiers because of the window size. Over the entire scene the *apriori* are close to $P_t = .025$ and $P_b = .975$. However, within a given 15 × 15 window centered about a road, P_t may be much larger. Thus, the probabilities were adjusted accordingly. No spatial matched filter was employed.

These preliminary results show that the adaptive quadratic classifier can work effectively without *apriori* background statistics. However, some of the advantages of this technique can not be fully illustrated with this data set. In particular, since TM band 1 offers a good discriminant based on mean difference, a quadratic classifier offers little improvement over a linear classifier. Furthermore, the background in the scene used does not vary. Thus, the ability of the adaptive detector to deal with changing backgrounds can not be illustrated.

7 Conclusions and Future Work

An adaptive quadratic classifier for multispectral target detection has been proposed. The system is designed to exploit both mean and covariance differences between the target and background and the local background statistics are estimated adaptively. Thus, these statistics need not be known *apriori* and the detector is able to adapt to changing backgrounds. The preliminary results, using landsat TM data, show that the adaptive detector appears to perform well based on subjective evauation.

Because this detector uses both mean and covariance difference as a discriminant, the band selection and feature extraction process must retain this information. A forward sequential band selection method using B-distance has been explored here in addition to a suboptimal linear feature extraction technique. For the Landsat TM data, the "best bands" have been shown to change for different classes of backgrounds for a

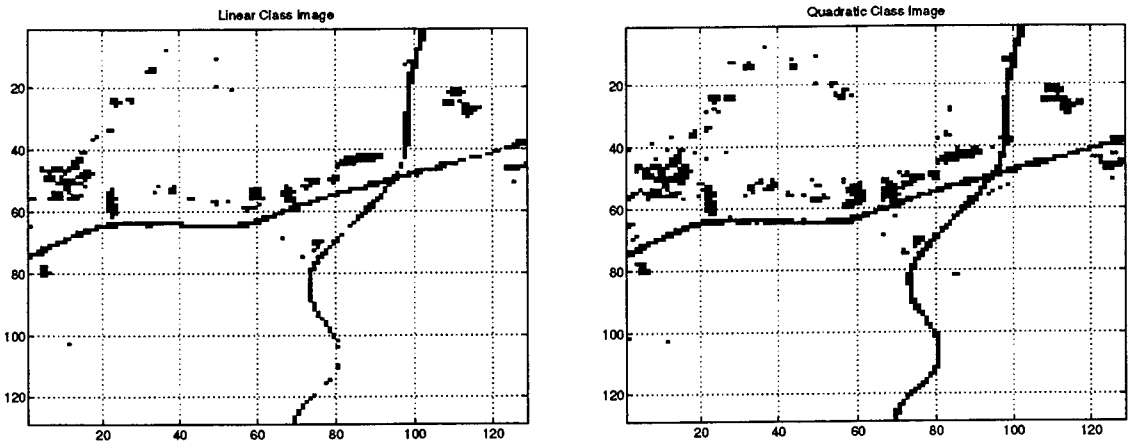


Figure 12: *(left)[a] Linear class map. (right)[b] Quadratic class map. The apriori probabilities are set to be* $P_t = .025$ *and* $P_b = .0975$ *for both. The optimal thresholds were used and no spatial matched filtering was performed.*

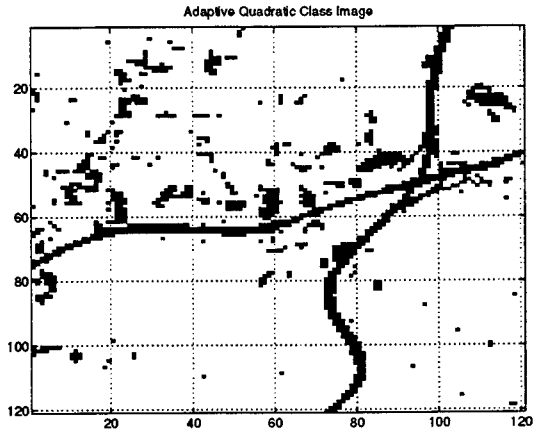


Figure 13: *Adaptive quadratic class image. A* 15×15 *window is used to estimate the background statistics and the apriori probabilities are set to be* $P_t = .15$ *and* $P_b = .85$. *No spatial matched filtering was performed.*

road target. With hyperspectral imagery, this effect will likely be more pronounced.

In future work, hyperspectral data acquired with the Airborne Visual and Infrared Imaging Spectrometer (AVIRIS) will be used. This data contains 224 bands over the range $0.4\mu m - 2.5\mu m$. Such high spectral resolution data will allow for better evaluation of the band selection method and the adaptive quadratic detector. In particular, multiple target and background classes will be identified and used in order to quantifiably study the effect of selecting different spectral bands, different numbers of bands, and different bandwidths. This will be done using the B-distance criteria. In order to evaluate the ability of the new quadratic detector to adapt to changing backgrounds, an image region where a target is found in a variety of backgrounds will be selected and processed. Synthetic multispectral data will also be used to provide an accurate quantitative measure of performance for the adaptive detector. With synthetic data, changing backgrounds can easily be simulated and exact target locations are known which supports error analysis.

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] 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.

[4] 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.

[5] 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.

[6] D. A. Landgrebe, "On the Use of Stochastic Process-Based Methods for the Analysis of Hyperspectral Data," Proceedings of the *International Geoscience and Remote Sensing Symposium (IGARSS)*, Houston, Texas, May 1992.

[7] C. Lee and D. Landgrebe, "Analyzing High Dimensional Data," Proceedings of the *International Geoscience and Remote Sensing Symposium (IGARSS)*, Houston, Texas, May 1992.

[8] 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.

[9] J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*, Springer-Verlag, 1986.

[10] 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.

[11] 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.

[12] 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.

A. D. Stocker, I. S.

[13] 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.

EFFECT OF HUMIDITY ON FRICTION AND WEAR FOR FOMBLIN Z UNDER BOUNDARY LUBRICATION CONDITIONS

Larry S. Helmick
Professor of Chemistry
Department of Science and Mathematics

Cedarville College
Box 601
Cedarville, OH 45314

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

and

Materials Directorate
Wright Laboratory

August 1994

EFFECT OF HUMIDITY ON FRICTION AND WEAR FOR FOMBLIN Z UNDER BOUNDARY LUBRICATION CONDITIONS

Larry S. Helmick
Professor of Chemistry
Department of Science and Mathematics
Cedarville College

Abstract

Using a Cameron-Plint tribometer under controlled environmental conditions, friction and wear were measured for Fomblin Z with M-50 steel under boundary lubrication conditions at 50, 100, and 150 C with relative humidity ranging from 5% to 100%. In general, both friction and wear decrease sharply as humidity is increased from 5 to 20%, then is constant as humidity increases to 100%. Thus, both friction and wear are highly dependent on humidity when relative humidity is less than 20%. Therefore, to improve repeatability of results, humidity should be measured any time Fomblin Z is being tested with steel specimens under boundary conditions, and carefully controlled if it is 20% or less.

EFFECT OF HUMIDITY ON FRICTION AND WEAR FOR FOMBLIN Z UNDER BOUNDARY LUBRICATION CONDITIONS

Larry S. Helmick

Introduction

Perfluoropolyalkyl ethers (PFPAE) are presently being investigated as liquid lubricants for aerospace applications (1,2). PFPAE fluids have been tested with the four-ball tribometer and the newer Cameron-Plint reciprocating tribometer under sliding boundary lubrication conditions with the test cell exposed to the environmental atmosphere (3,4). When testing Fomblin Z fluids with steel specimens with either instrument, erratic results are sometimes obtained. These occasional erratic results make any given individual result less reliable. In the past, this difficulty has been overcome by doing multiple tests and averaging the results. However, with the present need to develop a rapid and reliable screening test for potential antiwear additives for perfluoropolyalkyl ether lubricants (5-7), it is no longer feasible to do multiple tests. Therefore it became necessary to identify and attempt to eliminate the cause of these erratic results.

Since the test cells are exposed to the atmosphere, it was suspected that changing environmental humidity may be the source of the problem. The scientific literature concerning the effect

of humidity on friction and wear for metals has recently been reviewed (8). In general, under lubricated conditions, water vapor can influence friction and wear by modifying the adsorption behavior of long chain fatty acids that act as boundary lubricants, and affect the chemistry of protective film formation by oxygen. However, we are not aware of any specific reports concerning the effects of water vapor on the friction and wear of PFPAE fluids on metals. Therefore, the purpose of this investigation was to measure friction and wear under boundary lubrication conditions with controlled humidity to determine if humidity does indeed have any effect on these parameters for Fomblin Z fluids.

Experimental Procedure

Fomblin Z-04 (1 mL) was tested under sliding boundary lubrication conditions using a Cameron-Plint High Frequency Friction Machine (TE77) with a controlled environment chamber, 6 Hz frequency, 9 mm stroke length, and 250 N load. A short friction integration time constant was used and friction was monitored throughout the 5 hr runs. An average coefficient of friction was calculated for each run by dividing the average friction by the 250 N load. Although this value includes a contribution from the run in period when friction tends to fluctuate erratically, the run in period lasts only a few minutes and consequently has little affect on the magnitude on the coefficient of friction for a 5 hour run. Therefore, this value is comparable to steady state

friction coefficients previously reported (3).

M-50 steel cylinders (6mm length, 6mm dia.) and disks (58 Rockwell hardness, mirror finish) were cleaned in an ultrasonic bath with analytical grade hexane and methanol and dried with a hot air gun immediately before use. Data was obtained with specimen temperatures of 50, 100, and 150 C +/- 2 C and ambient atmospheric temperatures (29 +/- 3 C). Specimen temperatures and electrical resistance between the cylinder and disk were recorded continuously throughout the runs. All tests were done in air.

Relative humidity (RH) was measured with a Vaisala Relative Humidity and Temperature Probe (HMP 13) calibrated to +/- 2% accuracy below 90% RH and +/- 3% above 90% RH, and readable to +/- 0.1%. Humidities below ambient (50-60%) were obtained by controlling the flow rate of dry air into the environment chamber, and were constant to +/- 0.5%. Humidities above ambient were obtained by controlling the flow rate of air bubbled through a water tower, by lining the chamber with wet paper towels, or by adding steam as necessary, and were constant to +/- 3%.

Areas of wear scars were determined by measuring (+/- 1 X 10^{-3}mm) the length and width of the wear scar on the cylinder with a Nikon microscope (model 74514) with 150 power magnification at the end of the 5 hr runs. The visual appearance of the wear scar under the microscope was also noted.

All friction, wear, temperature, and humidity data are recorded in Table 1 and plotted in Figures 1-6. Open data points represent data obtained by a previous researcher in this laboratory, Dr. B. Cavdar, and are consistent with our data.

Results

Figures 1-3 show the average coefficient of friction plotted vs relative humidity. The plot at 50 C (Fig. 1) shows that friction decreases as humidity increases from 5 to 100%. At 100 C (Fig. 2), a sharp decrease in friction is observed as humidity increases from 5 to 20%. Then very little additional decrease occurs as humidity increases to 100%. At 150 C (Fig. 3), the sharp decrease in friction from 5 to 20% RH is even more obvious, but no further decrease occurs as humidity is raised to 100%.

Figures 4-6 show the effect of humidity on wear scar area as temperature is increased from 50 to 100 to 150 C respectively. At 50 C, figure 4 shows a sharp decrease in wear scar area as humidity increases from 5 to 20%. No additional decrease is obvious as humidity increases from 20 to 100%. At 100 and 150 C (Figures 5 and 6), the sharp decrease in wear scar area is even more pronounced. Again, no further decrease in wear is observed at either temperature as the humidity increases to 100%.

The appearance of the wear scar also varied with humidity. At low humidities, it appeared clean and metallic in nature.

However, at high humidities, it was highly colored or even black.

Electrical resistance varied with humidity as well. Although it was somewhat erratic during any given run, in general, it was low during runs at low humidity and high during runs at high humidity.

Discussion

Wear is often a function of friction. Thus one might expect similar trends to be observed if wear scar area and coefficient of friction are each measured and plotted vs humidity for the same series of wear tests. At 100 C as well as 150 C, this is indeed the case. (At 50 C, the correlation is not so obvious.) The sharp decrease in friction as humidity increases from 5 to 20% corresponds to a sharp decrease in wear. Furthermore, the constant friction as humidity increases from 20 to 100% corresponds to constant wear. This may be due to formation of a lubricating film on the surface of the wear scar as humidity increases. The film appears as an increase in electrical resistance between the cylinder and disk, and a visible darkening of the wear scar itself.

Electrical resistance is an indication of the amount of metal-metal contact between the cylinder and disk, and often correlates with the formation of surface films and the degree of wear. Low resistance implies good electrical contact between the cylinder

and disk, which suggests significant metal-metal contact, a metallic wear scar with minimal film formation, and high wear, as were observed. On the other hand, electrical resistance was generally high during runs at high humidity. Therefore, high humidity seems to promote formation of an electrically insulating film and a black wear scar. This film apparently is a good boundary lubricant, preventing metal-metal contact which results in low wear, as observed. The nature of films such as this and possible mechanisms for their formation are being investigated (4,9,10).

Conclusions

In test cells open to the environmental atmosphere, wear rates for Fomblin Z on sliding steel substrates under boundary lubrication conditions depend strongly on relative humidity for humidities below 20%, but are nearly independent of humidity above 20%. Therefore, to obtain the best repeatability for tests of this type, relative humidity should be monitored and tests conducted in the range above 20% where humidity has little effect on wear rates, if possible. If tests must be performed below 20% RH, humidity should be carefully controlled.

References

1. Fusaro, R.L., and Khonsari, M.M., "Liquid Lubrication for Space Applications," NASA TM-105198, 1992.

2. Fusaro, R.L., "Tribology Needs for Future Space and Aeronautical Systems", NASA TM-104525, 1991.

3. Masuko, M., Jones, W.R., Jr., and Helmick, L.S., "Tribological

Characteristics of Perfluoropolyether Liquid Lubricants Under Sliding Conditions in High Vacuum," NASA TM-106257, 1993.

4. John, P.J., Cavdar, B., and Liang, J., "Investigations of Wear Scar Formation on M50 and M50-NIL Substrates in the Presence of a Linear Perfluoropolyether," submitted for publication, 1994.

5. Gschwender, L.J., Snyder, C.E. Jr., and Fultz, G.W., "Soluble Additives for Perfluoropolyalkylether Liquid Lubricants," Lubr. Eng., 49, 9, 702-708 (1993).

6. Srinivasan, P., Corti, C., Montagna, L. and Savelli, P., "Soluble Additives for Perfluorinated Lubricants," JSL, 10, 2, 143-164 (1993).

7. Sharma, S.K., Gschwender, L.J., and Snyder, C.E., Jr., "Development of a Soluble Lubricity Additive for Perfluoropolyalkylether Fluids," JSL, 7, 1, 15-23 (1990).

8. Lancaster, J.K., "A Review of the Influence of Environmental Humidity and Water on Friction, Lubrication and Wear," Tribology International, 23, 6, 371-389 (1990).

9. Cavdar, B., Sharma, S.K., and Gschwender, L.J., "Wear-Reducing Surface Films Formed by a Fluorinated Sulfonamide Additive in a Chlorotrifluoroethylene-based Fluid," accepted for publication, (1994).

10. Vurens, G., Zehringer, R., and Saperstein, D., "The Decomposition Mechanism of Perfluoropolyether Lubricants during Wear," Chapter 10 in Surface Science Investigations in Tribology, Chung, Y.W., Homola, A.M., Sheet, G.B., eds., ACS Symposium Series, American Chemical Society, 1992.

Table 1

Temperature, Humidity, Friction, and Wear Data

Run No.	Temperature (C)	Rel. Hum. (%)	Avg. Coef. of Friction	Wear Area (mm^2)
LH-15	50	5	0.127	3.416
LH-19	50	5	0.111	3.437
LH-22	50	10	0.125	2.389
LH-23	50	15	0.103	1.456
LH-21	50	20	0.101	0.726
BC-17	50	32	0.102	0.622
LH-13	50	45	0.092	0.567
LH-14	50	45	0.092	0.407
LH-7	50	47	0.097	0.513
LH-3	50	57	0.092	0.728
LH-24	50	70	0.075	0.457
LH-17	50	98	0.055	0.627
LH-39	100	5	0.129	3.765
LH-28	100	5	0.130	3.956
LH-41	100	7	0.113	2.933
LH-31	100	10	0.099	0.499
LH-38	100	10	0.106	0.674
LH-29	100	20	0.103	0.478
BC-22	100	27	0.102	0.467
BC-9	100	40	0.106	0.454
BC-1	100	44	0.102	0.475
BC-35	100	48	0.110	0.609
LH-5	100	59	0.099	0.544
LH-30	100	70	0.097	0.492
LH-40	100	100	0.095	0.416
LH-36	150	5	0.148	5.117
LH-32	150	10	0.112	2.495
LH-34	150	10	0.124	2.659
LH-35	150	10	0.114	2.741
LH-37	150	15	0.096	0.506
LH-25	150	20	0.092	0.611
BC-3	150	30	0.102	0.534
BC-21	150	36	0.097	0.860
BC-11	150	41	0.098	0.846
LH-18	150	56	0.101	0.709
LH-20	150	99	0.096	0.427

Figure 1

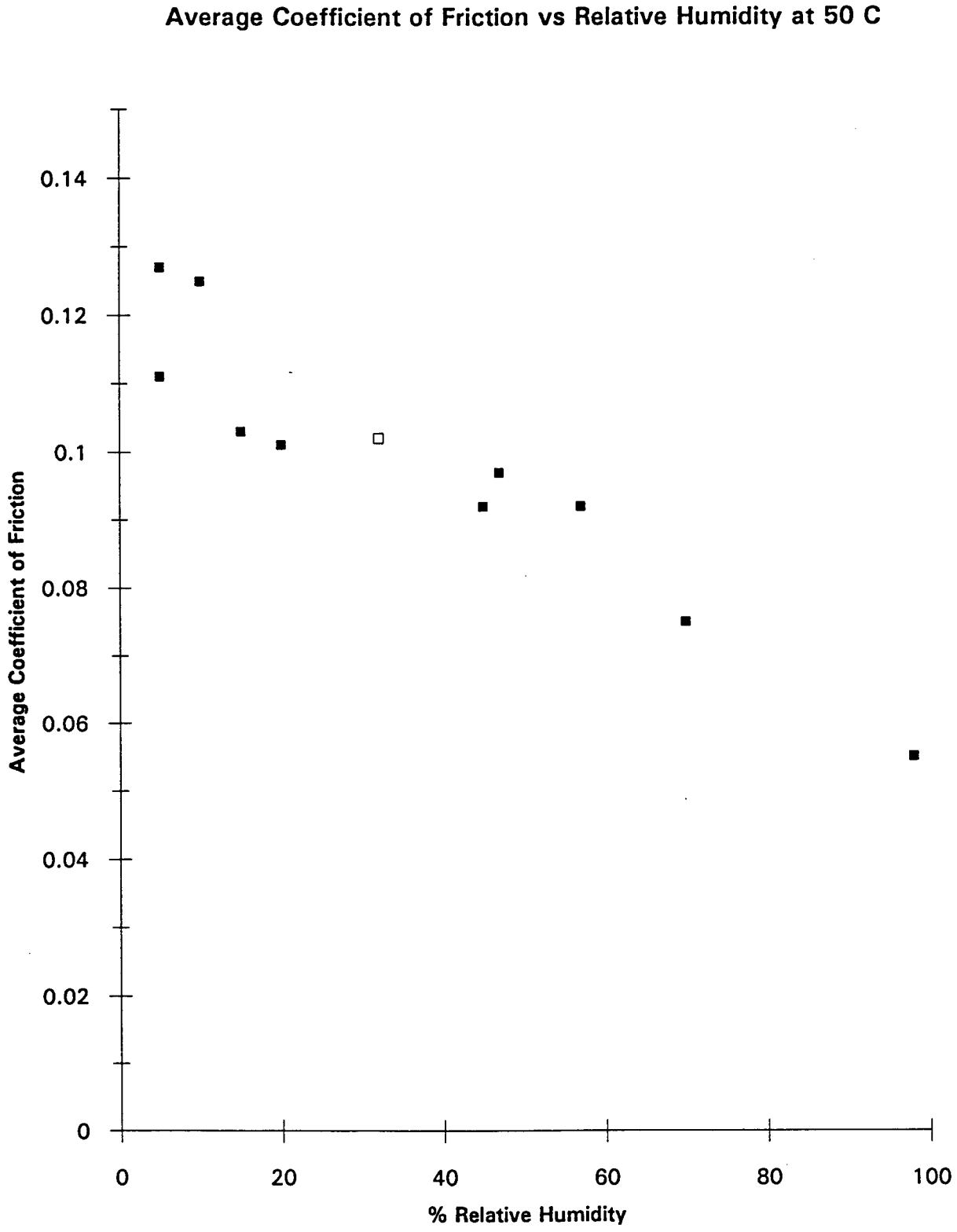

Average Coefficient of Friction vs Relative Humidity at 50 C
Average Coefficient of Friction
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
% Relative Humidity

Figure 2

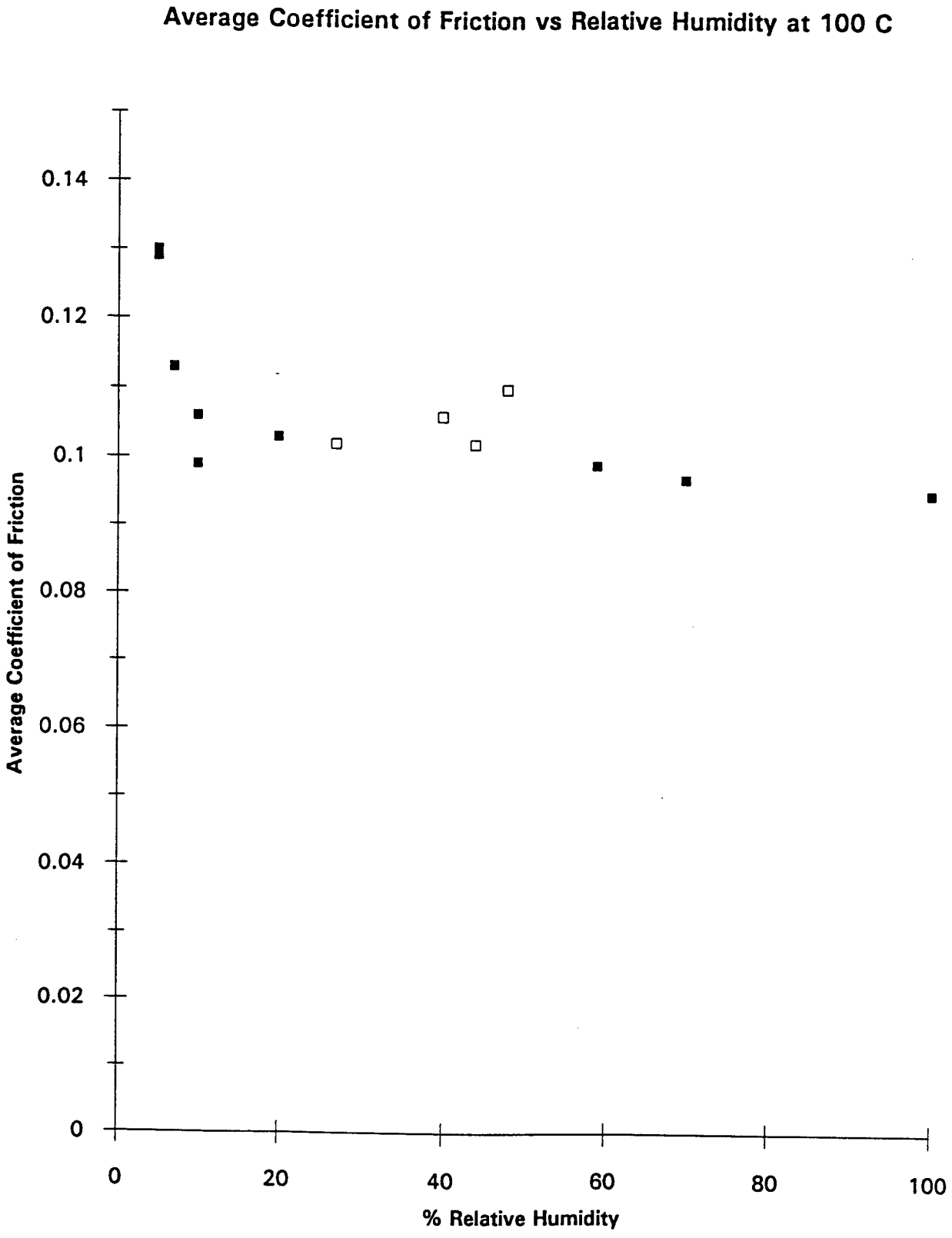

Average Coefficient of Friction vs Relative Humidity at 100 C
Average Coefficient of Friction
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
% Relative Humidity

Figure 3

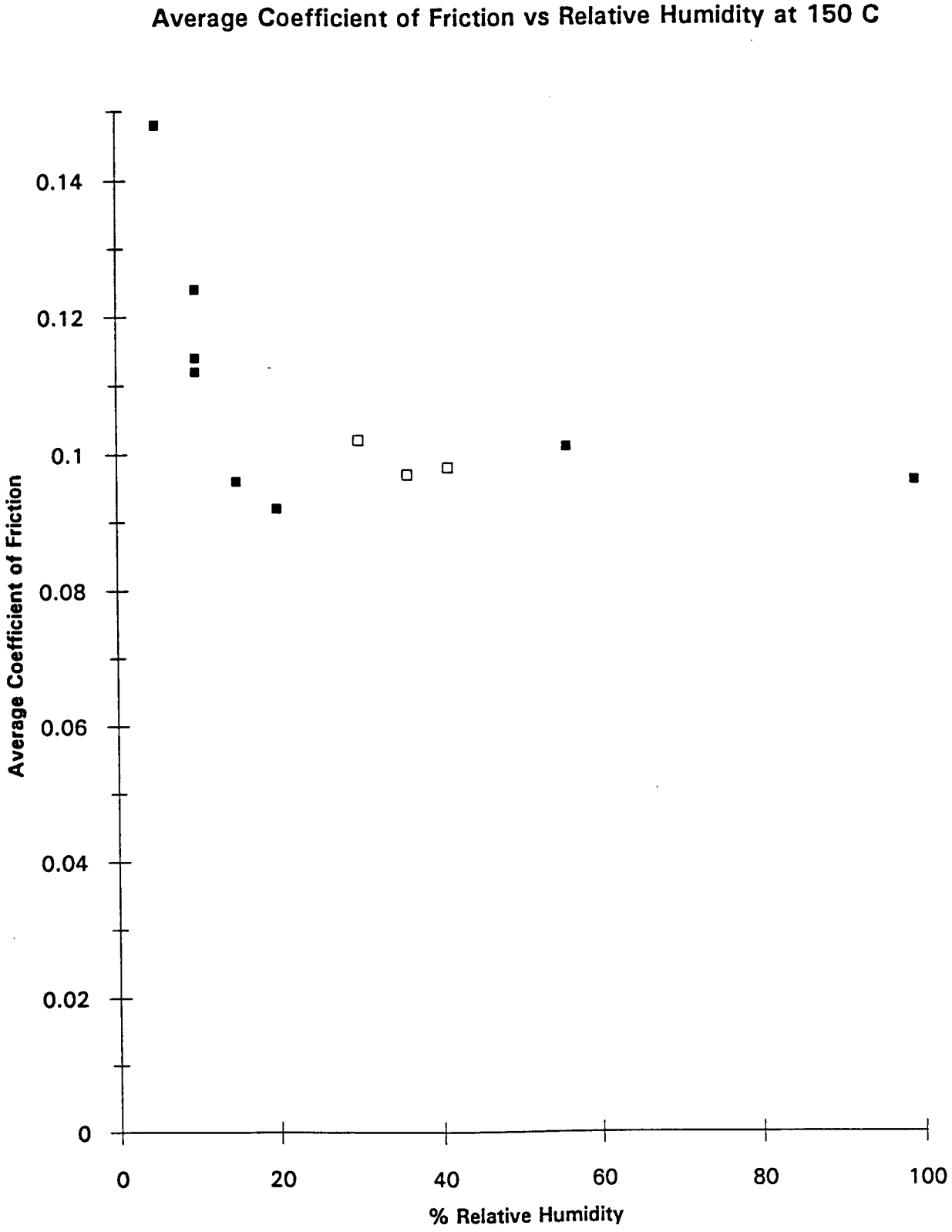

Average Coefficient of Friction vs Relative Humidity at 150 C
Average Coefficient of Friction
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
% Relative Humidity

Figure 4

Wear Scar Area vs Relative Humidity at 50 C

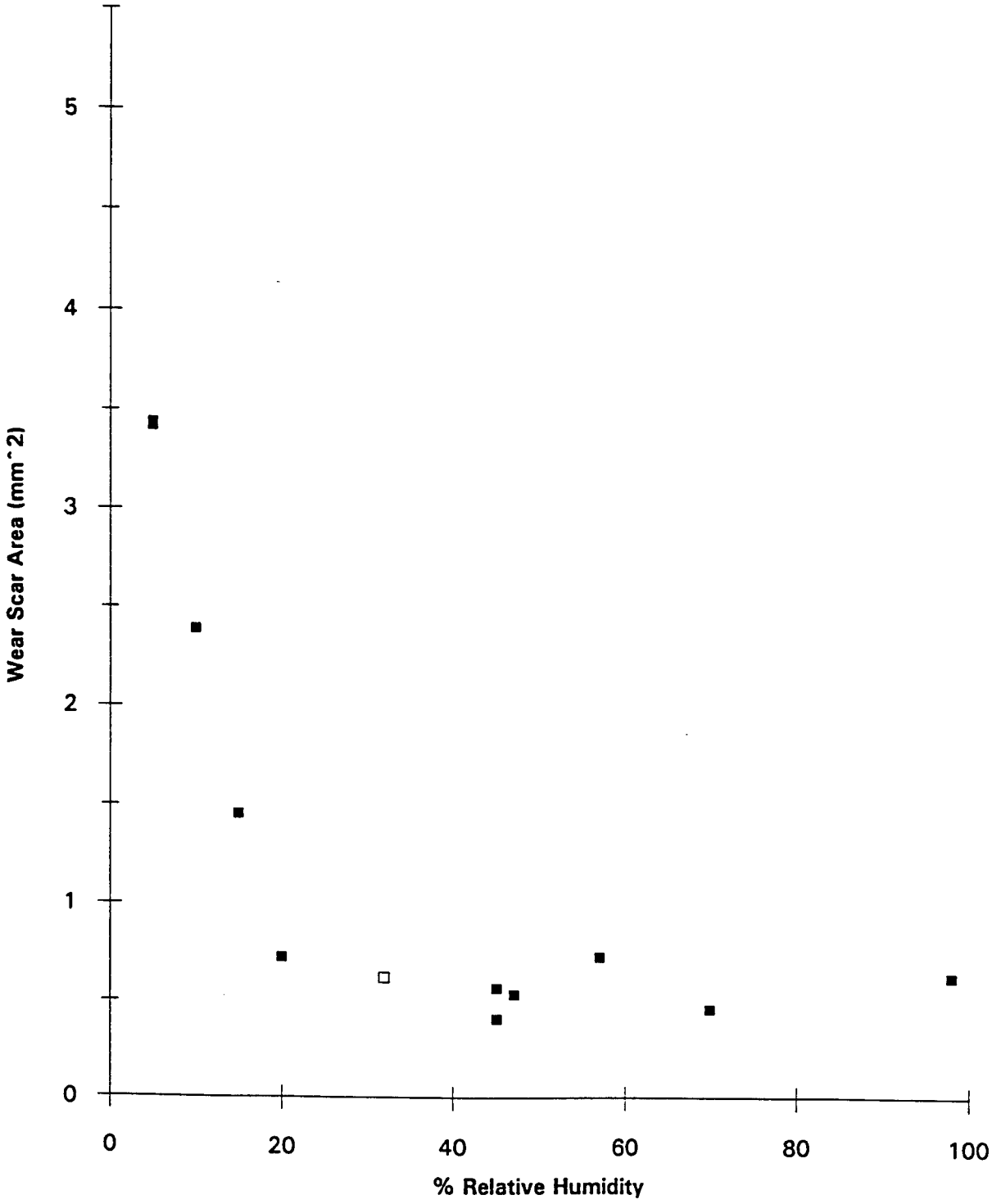


Figure 5

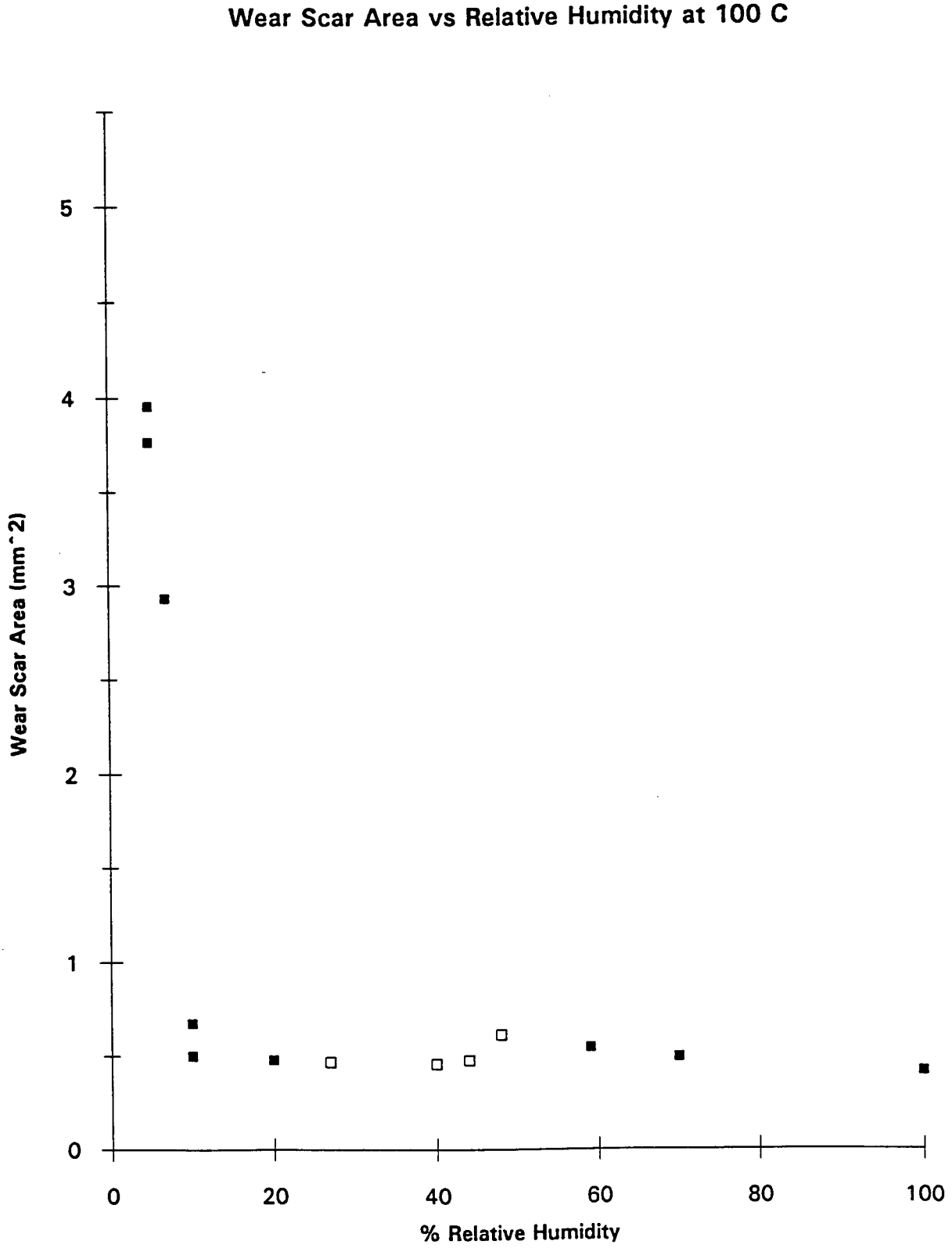
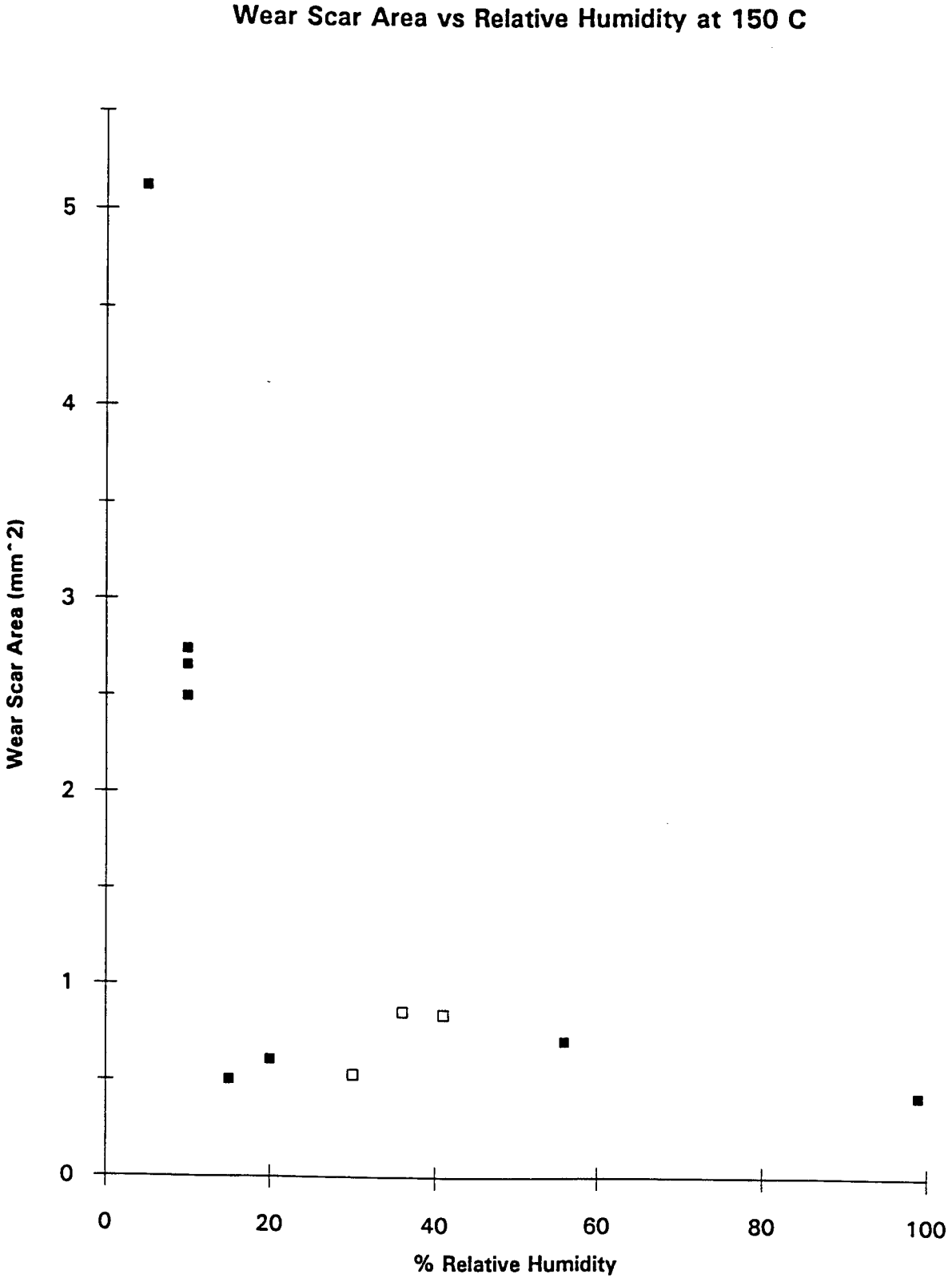

Wear Scar Area vs Relative Humidity at 100 C
Wear Scar Area (mm^2)
5
4
3
2
1
0
0
20
40
60
80
100
% Relative Humidity

Figure 6


Wear Scar Area vs Relative Humidity at 150 C
Wear Scar Area (mm^2)
5
4
3
2
1
0
0
20
40
60
80
100
% Relative Humidity

AUTOMATIC CONTROL ISSUES IN THE DEVELOPMENT OF AN ARTIFICIAL PANCREAS

A. S. Hodel
Assistant Professor
Department of Electrical Engineering
Auburn University

200 Broun Hall
Auburn, AL 36849

Final Report for:

Summer Faculty Research Program
Wright Laboratory/MNAG
Sponsored by:

Air Force Office of Scientific Research
Bolling Air Force Base, DC
and

Armstrong Laboratory

August 1994

AUTOMATIC CONTROL ISSUES IN THE DEVELOPMENT OF AN ARTIFICIAL PANCREAS

A. S. Hodel

Assistant Professor

Department of Electrical Engineering

Auburn University

Abstract

Recent developments in technology for the treatment of diabetes mellitus enable sensing and control of key chemical/hormone species related to the disease. Further, it is expected that non-invasive (infrared) blood glucose sensing techniques will permit continuous on-line sensing of blood glucose levels in insulin-dependent diabetics. It is thus desired to apply modern control systems design techniques in the design of an artificial pancreas in order to provide a robust, fault-tolerant design suitable for clinical and at-home use. A preliminary effort toward this goal was undertaken during the 1994 Summer Faculty/Graduate Student Research program at Eglin Air Force Base; a complementary study is presented by J. S. Naylor in Summer Research Program Report 21. The effort presented in this report comprises the development of (1) a qualitative model of endocrine kinematics related to glucose management and (2) a preliminary approach for system identification to be used in fitting these models to experimental data.

AUTOMATIC CONTROL ISSUES IN THE DEVELOPMENT OF AN ARTIFICIAL PANCREAS

A. S. Hodel

1 Introduction

This report addresses the application of robust multivariable control techniques toward the development of an autonomous control system for an artificial pancreas. The topic is of interest to Wright Lab/MNAG due to the following common properties between the treatment of diabetes and missile guidance:

1. The subject dynamic systems (a missile or a patient's glucose/insulin kinematics) are inherently nonlinear and, often, are poorly modeled relative to required performance levels.

2. The subject dynamic systems are subject to external disturbances that that are not under the authority of the control system (wind gusts, target motion, physical activity or food consumption by the diabetic patient).

3. The subject dynamic systems have limited actuator authority (limited fin deflection, insulin delivery rates)

4. Safe operation of the subject dynamic systems requires that system state variables be kept within a prescribed operating range.

Our effort this summer entails a preliminary attempt to identify prototype control methodologies for the artificial pancreas that, in addition to addressing the four issues above, provides quantitative bounds on the required performance of sensors/actuators to be used in an artificial pancreas system. This work was performed in tandem with J. S. Naylor, a graduate student from Auburn University, and with Johnny Evers and Dr. Darren Schumacher, both of MNAG branch, Wright Lab , Eglin Air Force Base.

2 Background

Althought diabetes, named by Aretaeus of Cappadocia (AD 81-138), has been diagnosed since roughly 1500 BC, it is only recently that its treatment has been made possible, due to the discovery of insulin by Banting and Best in 1921. Nearly fourteen (14) million people in the United States suffer from diabetes mellitus of which approximately ten (10) percent are afflicted with the auto immune illness, Type I diabetes mellitus, also called insulin-dependent diabetes mellitus (IDDM) or juvenile onset diabetes, since these individuals tend to be younger and require daily insulin injections for survival. IDDM patients have essentially no β-cell function, and are thus incapable of producing sufficient quantities of insulin to adequately regulate their blood glucose levels. The majority of afflicted individuals have Type II diabetes, also called non-insulin dependent diabetes (NIDDM)or maturity onset diabetes. While Type II Diabetes involves both β cell malfunction and peripheral resistance of tissues to insulin, the illness is characterized by progressive loss of Beta cell mass and eventual need for insulin administration. At any point in time, approximately fifteen (15) to thirty-five (35) percent of all adult patients being treated in diabetes clinics require insulin injections for management of their diabetes.

In glucoregulation, the alpha and beta cells of the pancreas sense blood glucose (BG) levels and secrete insulin and glucagon as counter-regulatory hormones. Hormones are chemical messengers secreted by the body as required to effect changes in target tissues in order to maintain homeostasis. Insulin promotes cellular uptake of plasma glucose and thus assists in the conversion of glucose to glycogen in the liver and in the muscles and in the storage of fat in adipose cells. The net effect of insulin is to lower BG, whereas glucagon has an antagonistic action on the liver [Fox87], p.595. Diabetic patients have lost the capability either to synthesize or to respond to insulin, or both.

The long term effects of high blood glucose (BG) levels that result from uncontrolled diabetes mellitus of both types are quite severe. In general, the vascular and nervous systems are affected and a large proportion of affected patients is at risk. The patient may in time suffer from peripheral neuropathy (loss of sensations at the periphery), retinopathy (loss of vision), and nephropathy (kidney failure).

In the hope of achieving and maintaining normoglycemia (normal blood glucose level), various formulations of insulin and treatment protocols were tried in an attempt to replicate the natural control process. The advent of miniature sensors [PM87], [Peu92] and pumps coupled with micro-

computers seemed to promise automatic optimal control. However, continued difficulties with insulin instability [SA87] catheter implantation [IKL] and long term glucose sensor drift [Peu92], [IKL] have been problems with this approach. The first two problems have been dealt with by means of the development of new formulations of insulins and the improvement of insulin catheter delivery systems. However the third challenge of a usable glucose sensor is still under development at several private and government locations.

The NIH supported Diabetes Control and Complications Trial (DCCT), a prospective study on the relationship of glycemic control and the development of vascular complications in people with Type I diabetes, conclusively demonstrated and reported in June, 1993 that improved glycemic control can prevent and delay the progression of vascular complications. The improved glucose control is achieved through intensive daily use of multiple injections of insulin. One serious consequence of improved glycemic management is a three-fold increase in severe hypoglycemia in individuals with aggressively managed diabetes. Any acceptable artificial pancreas must avoid this serious complication.

3 AEMG model

Several "minimal" models have been presented in the literature; see J. S. Naylor's report (Fellow #21) in this volume for more details. The compartmental model developed here, the Advanced Endocrine Management of Glucose (AEMG) model, is considerably more complex and allows for further qualitative studies to be made. The model is divided into several subblocks that correspond (roughly) to the liver, the blood, the pancreas, and cell uptake in body tissues. Simulations were done using Simulink (tm) with C-language implementations of the individual modules. C implementation was used for ease of development and for speed in simulation. The main module is shown in Figure 1 While the model appears quite complex, it is fairly straightforward when observed from a compartmental level. The four main compartments are

tissue_uptake Model interaction of adipose, muscle, and nervous tissue with blood chemistry.

blood_chemistry Model concentration levels of 6 relevant chemical species in the blood.

liver Model liver response to blood chemistry and pancreas endocrine production.

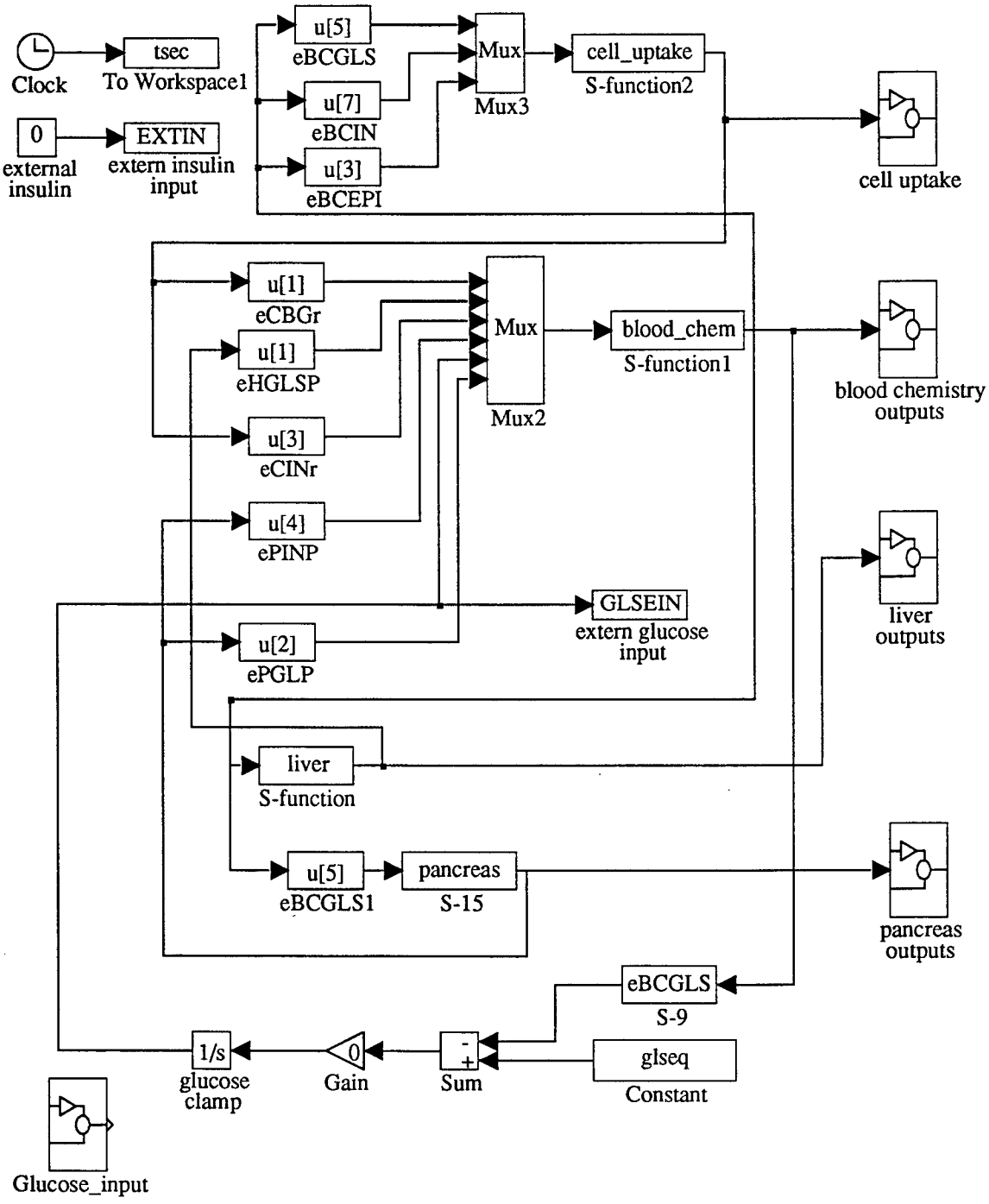
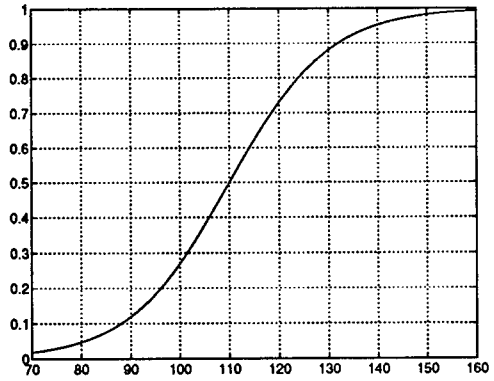


Figure 1: Main block diagram



The trigger function is

$$\text{trig}(v, l, h) = \left(1 + \tanh\left(\frac{2v - l - h}{h - l}\right)\right)/2.$$

The plot shows $\text{trig}(v, 90, 130)$ vs v. Notice that if $h < l$ then the function starts at 1 and descends to 0.

Figure 2: Trigger function behavior and definition

pancreas Model pancreas response to blood chemistry.

Patient health was characterized as the ability to produce insulin and glucagon. On-line adjustable parameters in the range [0,1] were provided in order to simulate different diabetic patients. Similarly, both glucose infusion and external insulin input were provided in order to simulate glucose tolerance tests and insulin infusion by an artificial pancreas.

Model parameters were selected to yield a stable equilibrium point for simulated healthy patient hormone nominal values in "normal" range, based on steady-state (basal) values and qualitative behaviors described in [KNJ82], [BG92] and other references. Not modeled in this simulation is the multi-phasic response of β-cells, production/transport of glycogen, stress hormone production behavior, etc. Multiphasic response of insulin was neglected since, for diabetic patients, insulin response is impaired.

The simulation makes regular use of a hyperbolic-tangent based "trigger" function that represents sensitivity of various compartments to their inputs. The trigger function is shown in 2. The function was selected in order to establish the presence of a lower "cutoff" value, below which the body does not respond to hormone concentrations, and to establish a "saturation point," beyond which the body will not respond to further increases in hormone concentration.

3.1 Modeling issues

IDDM patients β-cell mass is typically reduced to roughly 20% of that of a normal individual. NIDDM patients β-cell mass is typically reduced to 60% of normal. [KGI92]. In IDDM patients

with elevated blood glucose levels above 10 mmol/l, plasma insulin/C-peptide blood concentrations are undetectable. (C-peptide is produced in the reduction of pro-insulin to insulin. Relatively little proinsulin is released in healthy patients. [Mal92]) Impairment of glucoregulation in this case is due to reduced β cell mass, not reduced glucose sensitivity. BG threshhold for proinsulin synthesis is lower than for insulin release into the bloodstream. Excess proinsulin may be degraded via lysosomal action [HW80]. Proinsulin synthesis coupled with ATP generation in islet cells. May be associated with K^+ concentration.

Insulin release in healthy patients in response to a rapid glucose infusion involves a multiphase response as follows, with an early peak after approximately 3 minutes, followed by a decline within 5 minutes. The second phase response reaches its desired value after a total of approximately 16 minutes. This behavior is thought to be due to differing threshholds for insulin release in the β-cells. Widely varying β-cell response as changes in BG are made. Mathematical modeling indicates this may have a role in phasic/oscillatory insulin response [MO89]. Insulin release currently modeled as a "fuel" reponse: D-glucose is consumed in transport of insulin to cell surface ([Mal92] p. 266). Normal β-cells have almost instant equilibrium concentrations of D-glucose; fast response also to other hexoses.

[FWA92] presents several models used to study the effects of insulin and to estimate insulin sensitivity of target tissues. Can use (1) ratio of basal insulin to glucose concentration (not validated), or(2) homeostatic model assessment (HOMA) [THM+79]. Model based on known β-cell response to glucose. HOMA assumptions are

- glucose concentration increases in response to insulin deficiency is governed by shape of normal insulin secretion response to glucose. (Looks like bias diagram for a field effect transistor; see p. 515).

- Basal insulin levels are proportional to insulin resistance.

The model is not used much due to these assumptions and due to a lack of validation. Results are weakly correlated with euglycaemic clamp. [FWA92] indicate that some experiments seem to show it's a good model (p. 514).

Criticism of methods that measure endogenous insulin: measures also measure proinsulin and its products. Proinsulin makes up about 10% of insulin reaction in normal people, but may be 50%

in NIDDM. Thus experimental high insulin levels do not necessarily indicate insulin resistivity. Second criticism: peripheral insulin levels are measured; but insulin is secreted into the portal vein, and is depleted by hepatic extraction (non-constant).

Estimation of insulin sensitivity by simultaneous measurement of glucose and insulin concentrations. Conclusions:

- Mild NIDDM subjects (reduced hyperglycaemia) have elevated postprandial (after-meals) insulin levels relative to normal subjects. Use glucose/insulin ratio to determine resistance.
- Can't extend result to NIDDM patients with elevated glucose but lower insulin levels. Have reduced insulin secretion, but not always increased resistance to insulin.
- Static measures are qualitatively useful.

Tests to determine insulin resistivity and pancreas function include the Oral glucose tolerance (OGTT), the Intravenous glucose tolerance test (IVGTT), and the Insulin Tolerance Test (ITT). The first two tests are used to determine how quickly the body can restore normoglycemia; in the process, insulin resistance can be determined as well. The ITT provides a measure of how quickly insulin causes glucose to be absorbed from the bloodstream (resistivity).

Further data regarding healthy and diabetic dynamic behavior is provided in [WBMK+92], which response to infusion of stress hormones in "natural" and somatostatin-induced diabetes is tested. This paper provides steady state values as follows:

Hormone	**Infusion rate**	**Plasma concentration**	**basal level**
glucagon	0.8μg/(kg-h)	1133 ± 72pg/ml	120 ± 32pg/ml
epinephrine	6μg/(kg-h)	1050 ± 182pg/ml	92 ± 14pg/ml
growth hormone	20 μg/(kg-h)	118 ± 11ng/ml	2.4 ± 0.4ng/ml
cortisol	171 μg/(kg-h)	$38.1 \pm 1/3\mu$g/dl	$11.1 \pm 3.2\mu$g/dl

Basal blood glucose was 85 ± 5mg/dl for healthy subjects. Responses:

Hormone	**peak blood glucose**	**peak time**	**return to basal level**
glucagon	220 ± 8pg/ml	96 ± 6min	2h after infusion
epinephrine	222 ± 16pg/ml	155 ± 14min	4h after infusion
growth hormone	132 ± 5mg/dl	305 ± 24min	13h after infusion
cortisol	143 ± 7mg/dl	527 ± 137min	13h after infusion

Several "minimal" (2 or 3 state) models have been proposed for the prediction of diabetic behavior, e.g. [CAG+78]. These models ignore the multiphasic response profile of the β-cells, and are criticized for their poor long-term predictive behavior. Further, modeling of external insulin control requires knowledge of the type of insulin used and the location of injection. [AS92]. Have changes in duration/absorption rates due to species of origin as well (human, pork, beef/pork).

3.2 Blood chemistry compartment

The inputs to the blood chemistry compartment (see Figure 1) are tissue/membrane blood glucose uptake rate, hepatic (liver) blood glucose production/consumption, pancreas insulin and glucagon secretion rates, and external (digestive/intravenous) glucose input. Six chemical species are monitored in the blood chemistry module: cortisol, epinephrine (adrenaline), glucagon, blood glucose, growth hormone, and insulin (u/ml). That is, three stress hormones are added to the three hormones modeled in the minimal model. (Estimated dynamics associated stress hormone production in response to hypoglycemia are included.) Each of these items is considered in the subsections below. Relevant data used to select constants used in each submodule are presented. Basal values were taken from [BG92] and [KNJ82]; additional references are listed below where relevant.

3.2.1 Cortisol blood chemistry module

Parameters for the design of the cortisol submodule of the blood chemistry module are as follows.

Model parameters

Parameter	Value
Nominal value	10-12μg/dl
Bloodstream half-life	??? (set to 7.5 min)
High blood glucose cutoff	90
Low blood glucose cutoff	60

Cortisol is released in response to reduced blood glucose levels. Derivation of the cortisol module is identical to that of the growth hormone module, with the obvious substitution of relevant parameters; see §3.2.5 for details. The bloodstream half-life of cortisone is **unknown** and assumed to be 5-10 min, and so the decay constant α satisfies $e^{-7.5\alpha} = 0.5$.

3.2.2 Epinephrine

Nominal value is 10 pg/ml. Within the simulation, epinephrine is created in a similar fashion to cortisol. Epinephrine limits cell glucose uptake (see cell uptake module), but this feature is not yet implemented due to a lack of corresponding data.

3.2.3 Glucagon blood chemistry module

The basal value of blood glucagon concentration is 100-150 pg/ml. The process model is based on the following assumptions: (1) glucagon is consumed by natural decay, and (2) Glucagon is produced by the pancreas The module is a linear system; glucagon concentration is modeled as

$$\dot{x} = -\alpha x + bu$$

where x is the concentration of glucagon, α is a bloodstream decay rate constant, and b is a multiplier related to total body fluid volume (not implemented). Glucagon has a bloodstream half-life of approximately 5-10 minutes (7.5 ± 2.5min). Thus α is selected such that $e^{-7.5\alpha} = 0.5$.

3.2.4 Glucose

The assumptions for blood glucose dynamics are as follows.

- Blood glucose is consumed by cells and in the liver by hepatic glycogenolysis glucose→glycogen), and
- blood glucose is produced in the liver by hepatic glycogenesis and hepatic glyconeogenesis.

The current implementation neglects natural decay of glucose in the bloodstream, and so is just an integrator with a summing junction at front:

$$\dot{x} = -b_1 u_1 + b_2 u_2 + b_3 u_3$$

where u_1, u_2, and u_3 are cell blood glucose uptake, hepatic glucose production, and external glucose input, respectively, and b_1, b_2, b_3 are the corresponding constants related to body fluid volume (61).

3.2.5 Growth Hormone Module

While growth hormone does act against insulin in pharmacological doses, it appears to have little effect on glucose regulation. Unstable diabetic subjects have elevated growth hormone levels levels. Relevant model parameters are as follows.

Model parameters

Parameter	Value
Nominal value	2 ng/ml
Bloodstream half-life	1-2hrs (90 min)
High blood glucose cutoff	90
Low blood glucose cutoff	60

Growth hormone is produced in response to severe hypoglycemia; simulation parameters were selected for a trigger function with high/low blood glucose concentration cutoff values of 90mg/dl and 60 mg/dl, respectively for this purpose. The trigger gain value is **unknown** at this time, and is arbitrarily set to 1. The basal concentration of growth hormone is 2ng/ml, and so a constant input value is used to force this concentration of growth hormone under euglycemia. The resulting system model is

$$\dot{x} = -\alpha x + k_1 \text{trig}(u, 90, 60) + k_2 \alpha$$

where x is the concentration of growth hormone, α is a bloodstream decay rate constant selected such that $e^{-90\alpha} = 0.5$, k_1 is a constant to be determined (currently set to 1), and $k_2 = (2 - k_1 \text{trig}(110, 90, 60))$ is selected such that the steady state value of growth hormone is 2ng/ml. Growth hormone promotes glyconeogenesis (see §3.4).

3.2.6 Insulin blood chemistry module

Nominal fasting value for blood insulin concentration is 10-50 μu/ml. The simulation model is based on the assumption that insulin is consumed by natural decay and by tissue/liver membrane consumption. The overall simulation admits insulin production by either the pancreas or by an external input (pump). The resulting (linear) model is

$$\dot{x} = -\alpha x + b(u_1 + u_2)$$

where x is the concentration of blood insulin, α is the blood insulin decay constant u_1 and u_2 are cell insulin absorption rate and pancreas insulin production rate, respectively, and b is the related body fluid volume constant. Insulin has a bloodstream half-life of approximately 5-10 minutes (7.5 ± 2.5min). Thus α is selected such that $e^{-7.5\alpha} = 0.5$.

3.3 Tissue uptake compartment

The initial cell uptake module modeled binding of insulin to cell membrane walls and decay of insulin so bound. However, this model was rejected since bound insulin levels quickly track blood insulin concentration. The final model used is based on the following assumptions [BG92]:

- Insulin dependent glucose uptake occurs at a rate proportional to the product of a blood insulin concentration and blood glucose concentration and comprises 25% of glucose uptake at basal levels.
- Insulin independent glucose uptake occurs at a rate proportional to $\text{trig}(G, 60, 150)$ where G =blood glucose concentration, and comprises 75% of glucose uptake at basal levels. (Threshhold limiting values were selected ad hoc.)
- Basal glucose uptake is 2mg/kg-min (recall that our simulated patient body weight and blood volume are 80 kg and 6l, respectively).

The resulting model is

$$y = c_1 GI + c_2 \text{trig}(G, 60, 150)$$

where G, I are blood glucose and blood insulin levels, respectively, and constants c_1 and c_2 are selected so that the above constraints are met at basal levels.

Experimental measurements indicate that glucose uptake is only 1/3 as sensitive to insulin as is hepatic glucose production levels. This may be due to the fact that 60% of all insulin produced is immediately absorbed as it enters the liver. This feature has not yet been incorporated into our simulation.

3.4 Liver module

The liver submodule models only the production and consumption of glucose and glycogen in response to insulin, glucagon, blood glucose, and stress hormones. Insulin uptake is not yet incorporated into the model. The following data were used in the design of the liver submodule.

Item	Value	promotes
cortisol	11μg/dl	gluconeogenesis
epinephrine	10 pg/ml	glycogenolysis
glucagon level	100-150 pg/ml	glycogenolysis
glucose level	110 mg/dl	inverse effect on glucose production (see below)
growth hormone	2 mg/ml	gluconeogenesis
insulin	25 μu/ml	glycolysis (generate glycogen)

Other relevant data

- body basal glucose uptake: 2mg/(kg-min)
- gluconeogenesis: 30% of production, increases as hypoglycemia persists.
- Glucose has a molecular weight of 119g/M: 11 H, 1 C, and 6 O.
- Increase in glucose level of 2mM/l, or 23.8mg/dl, reduces glucose production by 80%.
- Increment of insulin of 100μu/ml decreases glucose production to less than 10-15% of basal level.
- hepatic glucose production is 3x more sensitive than the glucose uptake rate is to insulin.

As in the simulation of secreted hormones in the body, glucose and glycogen production are simulated via triggering functions. A realistic representation of the responses of the various hormones to changes in blood glucose level is not available; similarly, the way these hormones impact glucose/glycogen dynamics is not known. As a result, most of the values selected herein are quite arbitrary.

Gluconeogenesis is promoted by prolonged hypoglycemia, and occurs through active agents glucagon, cortisol, and growth hormone. Gluconeogenesis occurs on a much slower scale than

glygogenolysis, on the order of an hour, and so a 1st order lag filter $1/(60s+1)$ is used to model response to The liver module takes as inputs the blood chemistry module outputs. The current module does model glycogen stores dynamics. The output is a glucose production rate (positive) or consumption rate (negative). The current block is just a dummy sum of the different hormones related to glucose management. This block is quite weak and needs corrected

3.5 Pancreas compartmnent

The current pancreas model takes blood glucose concentration G as its sole input. This value is passed through the trigger function is $\text{trig}(G, 90, 130)$ in order to establish insulin/glucagon response.

3.5.1 Alpha cells

The alpha cells module models glucagon release in response to *low* blood glucose (BG) levels at a rate proportional to the product of current glucagon stores and $\text{trig}(G, 130, 90) = 1 - \text{trig}(G, 90, 130)$. Glucagon is replenished at a rate proportional to the degree that stores have been depleted, i.e., glucagon production rate is

$$\alpha \frac{\text{GL}_{\text{max}} - \text{GL}}{\text{GL}_{\text{max}}}$$

where α is a maximum production rate constant and GL is the current value of glucagon stores.

The block constants were selected as follows. Euglycemia levels in the simulation are set at 110 mg/dl, yielding a blood glucose trigger value of 0.5. Nominal blood glucagon concentration is 100-150 pg/ml (set to 125 for simulation). Since blood glucagon has a half-life of 5-10 minutes (simulation value set at 7.5min; see §3.2.3), it follows that the glucagon release rate must satisfy $r = 125\alpha V$ where $\alpha = -\ln(0.5)/7.5$ is the glucagon decay constant and V is the blood fluid volume, set to 6l =6000ml for this simulation, yielding a steady state glucagon release rate of 69.315 ng/min. The maximum glucagon production rate was set to twice this amount. For this simulation, the production rate will equal the stead-state release rate when the stores are half full. Storage value was selected (**arbitrarily**) as a factor of 10 larger than the required release rate. Thus, the glucagon stores will be five times the required basal release rate at steady state. Since the euglycemic BG trigger value will be 0.5, the BG trigger constant in the alpha cells block should thus be 2/5.

3.5.2 β-cells

Insulin is released in response to both a high blood glucose trigger level trig$(G, 90, 130)$ and to the current release rate of glucagon. Insulin is released and replenished in a fashion analogous to that of glucagon.

Simulation block constants were determined as follows. While it is known that the pancreas β cells release insulin in response to a release of glucagon from the alpha cells, the degree of this response is unknown. Thus, in the current simulation version, the gain from glucagon secretion rate to insulin production is set to zero. Fasting levels of insulin in the blood are 10-50 μu/ml in the blood (simulation value: 25). We assume that the bloodstream fluid volume is V =6l. Since insulin has a bloodstream half-life of 7.5 min, insulin must be replenished at a rate of $-25 \times 10^{-6} V \ln(1/2)/7.5 \approx$ 13.86mu per minute, or 0.83 u/hr, which is reasonable. As in the alpha cells submodule, the simulated maximum rate of insulin production is set to twice the basal rate of insulin secretion, and maximum stores are set to 10 times the basal secretion per minute. The multiphasic nature of insulin secretion is not yet incorporated into the β-cells module.

4 Simulation results

The AEMG model was excercised first to contrast the behavior of healthy and diabetic patients. Two types of diabetic patients were simulated. Both patients had no β-cell function. The first patient (recent diabetic) had normal α-cell function. The second diabetic patient had minimal (10%) alpha-cell function, representing a patient who has been diabetic for 5–10 years. Glucose tolerance was tested by simulating an insulin infusion of 100 mg/min during the interval $t \in [100, 200]$min. Results of these open-loop simulations are shown in Figure 3.

A variable-structure controller (VSC) was used to administer insulin in a fashion similar to that discussed in Naylor's report. The VSC implemented in this simulation accepted blood glucose (G) as an input; a "dirty derivative" $\dot{G} \approx G_{der} = (s/(10s + 1))G$ was used to approximate blood glucose rate. The switching function was based on the trig(v, l, h) function defined in the previous section. The VSC accepted two parameters: (1) a cutoff value G_{lo} such that no insulin is injected if $G < G_{lo}$, and (2) the slope of the trigger function in its linear region of operation. The desired target value for blood glucose was 100 mg/dl and this was placed at the center of the linear

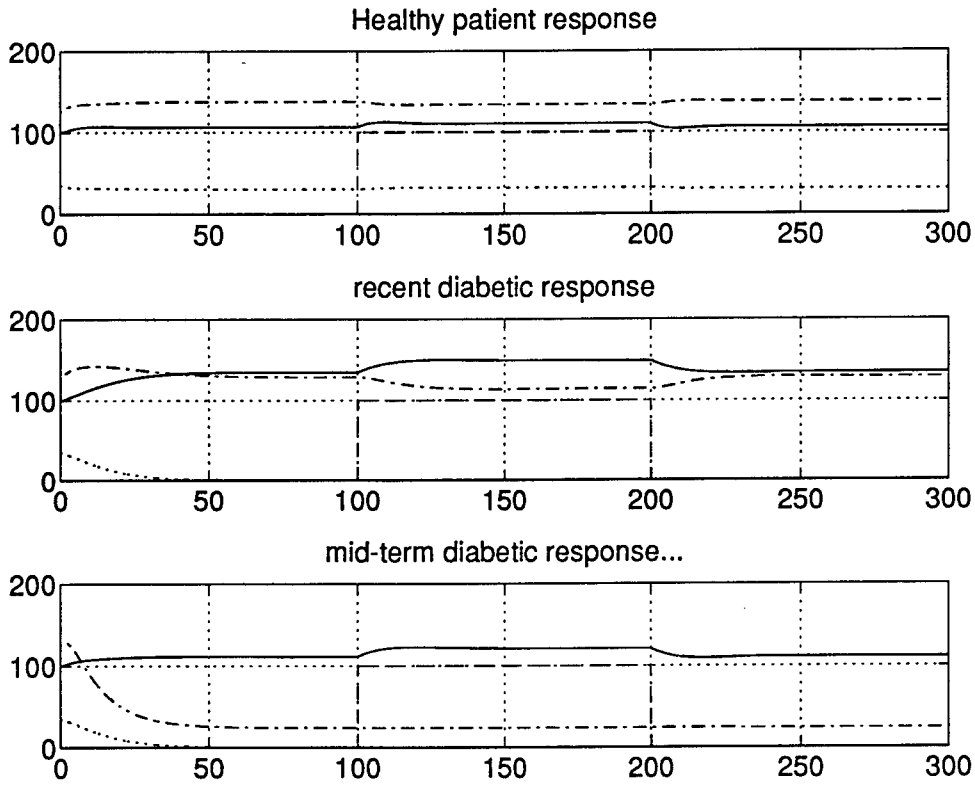


Figure 3: Contrast: healthy/diabetic simulated patients

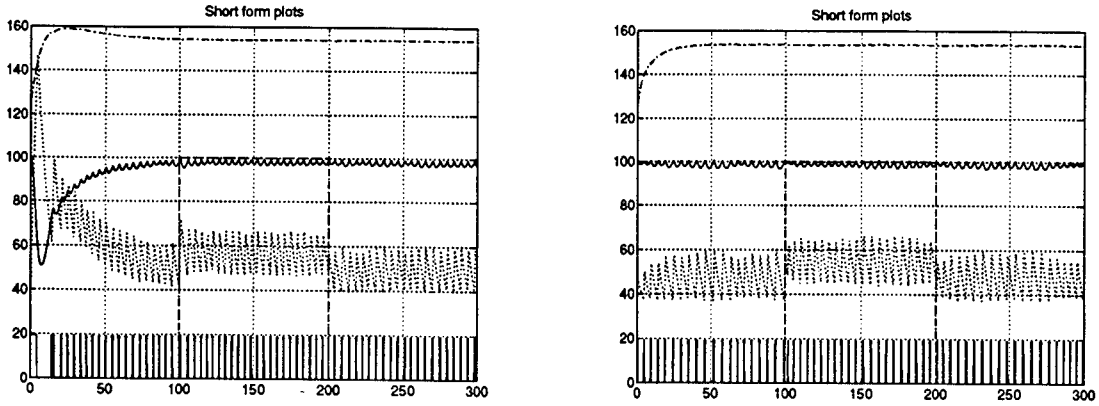


Figure 4: Left: simulated closed loop behaior with $G_{lo} = 0$, slope $= 0.1$. Right: simulated closed loop behavior with $G_{lo} = 100$, slope $= 0.1$

region of the trig function for the selected slope value. Based on these values, a switching function $s = G_{der} - \text{trig}(G, l, h) - \text{trig}(100, l, h)$ was selected. When $s > 0$, insulin was injected at a rate of $10\times$ the basal rate of insulin secretion in a normal human. (This rate is quite high; however, reducing the rate of injection would only reduce the risk of hypoglycemia, and so this simulation provides a proof of concept.) Results for these closed-loop simulations for a recent diabetic patient are shown in Figure 4. Notice that with a cutoff value $G_{lo} = 100$, the regulation is nearly exact.

5 Conclusions

The careful regulation of blood glucose is a crucial element of effective treatment of diabetes. The Diabetes Control and Complications Trial (DCCT) sponsored by the NIH demonstrates that proper regulation of diabetes greatly reduces the lon-term complications associated with insulin-dependent diabetes mellitus (IDDM). Preliminary steps toward the development of a control-oriented model, suitable for the design and testing of a robust, fault-tolerant controller, have been taken under this program. (Further discussion of thsi work can be found in report #21 of this volume). Simulations indicate that a variable structure controller (VSC) can provide effective glucoregulation in simulated diabetic patients; validation of this control strategy against a commercial diabetes simulator remains to be done. It is expected that further work on this topic will yield a beneficial two-way technology

transfer between control strategies developed in Wright Lab/MNAG and the biomedical community.

References

[AS92] A. Michael Albisser and Marianne Sperlich. Adjusting insulins. *The Diabetes Educator*, 18(3):211–222, May/Jun 1992.

[BG92] Geremia B. Bolli and Edwin A. M. Gale. Hypoglycemia. In K.G.M.M. Alberti, R.A. DeFronzo, H. Keen, and P. Zimmet, editors, *International Textbook of Diabetes Mellitus*, pages 1131–1149. John Wiley &Sons, 1992.

[CAG+78] Ruby Celeste, Eugene Ackerman, Laël C. Gatewood, Clayton Reynolds, and George D. Molnar. The role of glucagon in the regulation of blood glucose: model studies. *Bulletin of Mathematical Biology*, 40:59–77, 1978.

[Fox87] I. F. Fox. *Human Physiology, 2nd Edition.* Brown Publishers, 1987.

[FWA92] G. R. Fulcher, M. Walker, and K. G. M. M. Alberti. The assessment of insulin action in vivo. In K.G.M.M. Alberti, R.A. DeFronzo, H. Keen, and P. Zimmet, editors, *International Textbook of Diabetes Mellitus*, pages 513–529. John Wiley & Sons, 1992.

[HW80] P. A. Halban and C. B. Wollheim. Intracellular degradation of insulin stores by pancreatic islets in vitro: an alternativge pathway for homeostatis of pancreatic insulin content. *J. Biol. Chem.*, 255:6003–6006, 1980.

[IKL] K. Irsigler, H. Kritz, and R. G. Lovett. Controlled drug delivery in the treatment of diabetes mellitus. *CRC Critical Reviews in Therapeutic Drug Carrier Systems*, 1(3):189–280.

[KGI92] G. Klŏppel, W. Gepts, and P. A. In't Veld. Morphology of the pancreas in normal and diabetic states. In K.G.M.M. Alberti, R.A. DeFronzo, H. Keen, and P. Zimmet, editors, *International Textbook of Diabetes Mellitus*, pages 224–259. John Wiley & Sons, 1992.

[KNJ82] Keele, Neil, and Joels. *Samson Wright's Applied Physiology.* Oxford University Press, 1982.

[Mal92] W. J. Malaisse. Insulin biosynthesis and secretion in vitro. In K.G.M.M. Alberti, R.A. DeFronzo, H. Keen, and P. Zimmet, editors, *International Textbook of Diabetes Mellitus*, chapter 11, pages 224–259. John Wiley & Sons, 1992.

[MO89] W. J. Malaisse and A. Owen. A matehatical modelling of stimulus-secretion coupling in the pancreatic b-cell. VI cellular heterogeneity and recruitment. *Appl. Math. Mod*, 13:41–6, 1989.

[Peu92] R. A. Peura. Blood glucose biosensors-a review. In H. T. Nagle and W. J. Tomkins, editors, *Case Studies in Medical Instrument Design.* Institute of Electrical and Electronic Engineers (IEEE), 1992.

[PM87] R. A. Peura and Y. Mendelson. Blood glucose sensors: An overview. In *Proc. IEEE/NSF SYmp. Biosensors*, pages 63–68, 1987.

[SA87] D. S. Schade and G. M. Argoud. Implantable insulin pumps. In N. Sakamoto, , K. G. M. M. Alberti, and N. Hotta, editors, *Recent Trends in Management of Diabetes Mellitus*, pages 130–133. Excerpta Medica, 1987.

[THM+79] R. C. Turner, R. R. Holman, D. Matthews, T. D. R. Hockaday, and J. Peto. Insulin deficiency and insulin resistance interaction in diabetes: estimation of their relative contribution by feedback analysis from basal plasma insulin and glucose concentrations. *Metabolism*, 29:1086–1096, 1979.

[WBMK+92] Werner K. Waldhäsl, Paul Bratusch-Marrain, Martin Komjati, Felix Breitenecker, and Inge Troch. Blood glucose response to stress hormone exposuure in helathy man and insulin dependent diabetic patients: Prediction by computer modeling. *IEEE Transactions on Biomedical Engineering*, 39(8):779–790, 1992.

CAN DESIGN FOR COGGING OF TITANIUM ALUMINIDE ALLOYS

Vinod K. Jain

Professor

Mechanical and Aerospace Engineering Department

University of Dayton

Dayton, OH 45469-0210

Final Report for:

Summer Faculty Research Program

Materials Laboratory, WPAFB

Sponsored by:

Air Force Office of Scientific Research

Bolling Air Force Base, DC

and

Materials Laboratory, WPAFB

August 1994

CAN DESIGN FOR COGGING OF TITANIUM ALUMINIDE ALLOYS

Vinod K. Jain
Professor
Mech. & Aero. Eng. Dept.
University of Dayton

Abstract

The design of can to break-down titanium aluminide ingots via cogging process was attempted. A large strain viscoplastic finite elements program DEFORM (Design Environment for Forming) was used to simulate the cogging process for the near-gamma titanium aluminide alloy Ti-45.5Al-2Cr-2Nb in a type 304 stainless steel can. Can and process variables investigated in the FEM simulations included can thickness, can geometry, and ram velocity. It was found that there is an optimum can thickness and ram velocity to obtain moderately uniform flow between can and titanium aluminide workpiece.

CAN DESIGN FOR COGGING OF TITANIUM ALUMINIDE ALLOYS

Vinod K. Jain

Introduction

The development of high temperature and high specific strength alloys has been spurred by the aerospace industry in its efforts to design high performance turbine engines. Among the most promising candidate materials for such applications are gamma titanium aluminide alloys based on the ordered intermetallic phase TiAl. Unfortunately, in the cast condition, these materials are often brittle, due in part to the presence of porosity, macro- and micro-segregation, and large grain size. Thus the first step in producing wrought components with sufficient ductility is to break down the cast microstructure via some sort of thermomechanical processing technique.

Currently, the breakdown of cast gamma-titanium aluminide ingots is performed by isothermal forging or conventional hot extrusion. Considerable success has been realized by the former method, which consists of pancaking cylindrical preforms to high reductions of about 6:1 (ε= 1.8) in the temperature range of 1065 to 1175°C at nominal strain rates between 0.001 and 0.01/s. Under these conditions, the workability of titanium aluminides is fairly high, and sufficient work is imparted to recrystallize the material during hot working. However, the method is tedious and expensive, and cracks may develop at the bulged surface and propagate to the interior of the workpiece (1,2).

Extrusion as a breakdown method involves the use of a can. Canning prevents oxidation of the preform in air at high temperatures and insulates the core from the effects of die chilling during extrusion; the main disadvantage with the technique involves can removal after extrusion, requiring either machining or immersion in an acid bath. Heat losses during extrusion, coupled with tensile stresses (whose magnitudes are a function of reduction ratio, die design, etc.) can greatly affect microstructure uniformity as well as tendency for fracture of the workpiece (3,4). For example, Seetharaman, et al. (3) correlated the microstructures developed during canned extrusion of the gamma titanium aluminide alloy Ti-49.5Al-2.5Nb-1.1Mn (atomic percent) with the temperature distribution across the diameter and showed how temperature nonuniformities depend on die chilling, deformation heating, ram speed, and can design. With this understanding, subsequent work (5) led to the introduction of an insulating layer between the can and workpiece. This design modification considerably reduced the heat loss from the preform in to the can and tooling during extrusion thereby resulting in a relatively uniform microstructure across the cross-section.

Ingot breakdown of titanium aluminides by cogging has not been tried yet. Cogging is an open-die forging operation in which the thickness of a bar is reduced by successive forging steps at certain intervals (Fig.1). Because the contact area per stroke is small, a long section of a bar can be reduced in thickness without requiring large forces or machinery. Since only that part of the surface which is under the bite is being deformed at any one time, there is danger of causing surface laps at the step separating the forged from the unforged portion of the workpiece. For a given geometry of tooling there will be a critical deformation which will

produce laps. Wistreich and Shutt (6) recommend that the squeeze ratio h0/h1 should not exceed 1.3. It is also recommended that the bite ratio b/h should not be less than 1/3 to minimize inhomogeneous deformation. Thus, the only way to employ conventional cogging with titanium aluminides is by insulating and canning the billets to minimize the die chilling effects and eliminate possible preform fracture.

The present work describes the results of FEM simulations which were conducted to model the can and preform deformation to obtain uniform flow of both the materials (can and preform).

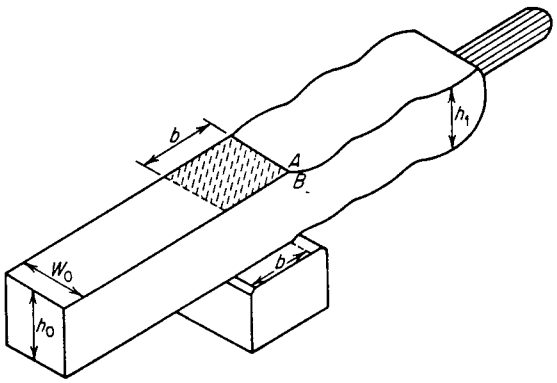


Fig. 1 Schematic of cogging operation in open-die forging. Shaded area shows where contact would occur between workpiece and die.

FEM Modeling

Finite element method (FEM) simulations were conducted to determine the effect of can design and processing parameters on metal flow uniformity and teperature transients during the conventional press forging of canned, cylindrical mults of gamma titanium aluminide alloys. The large strain, thermoviscoplastic FEM program DEFORM (Design Environment for Forming) was used to perform the simulations. Billet cooling during transfer from the furnace to the tooling and during the actual forging operation were modeled. The billet transfer stage incorporated a "controlled dwell" period prior to forging and thus produce more nearly equal flow stresses in the two components.

Input data for the simulations included material flow stress values as a function of strain, strain rate, and temperature, thermophysical properties, interface heat transfer coefficients, and friction factors. The flow stress of the billet material was taken to be that of the cast near-gamma titanium aluminide alloy Ti-45.5Al-2Cr-2Nb and was obtained from constant strain rate, isothermal hot compression tests conducted at 10^{-3}, 10^{-2}, 10^{-1}, and $1s^{-1}$ and temperatures of 1093, 1177, 1260, and 1343°C (7). The can material was assumed to be AISI type 304 stainless steel. Its flow stress dependence on strain rate and temperature were obtained from reference 8. Thermophysical properties (specific heat, thermal diffusivity, and thermal conductivity) were measured for the titanium aluminide alloy (9) or taken from various handbooks (10-12) for the can material and die material (H-12 tool steel). The can emissivity (e) and interface heat transfer coefficients (h's) used in the simulation are summarized in Table 1. The values of the heat transfer coefficients were based on measurements in the literature (13-15) and the authors'

experience; they reflect the influence of pressure (e.g., no pressure during transfer versus high pressure during forging) and the nature of interface (e.g., lubrication between the dies and can versus special insulation between the can and billet). Two values of the friction shear factor, m=0.2 and m=0.6, were used to account for the lubrication conditions at the can-die interface and can billet interface, respectively.

Table 1. Emissivity and Heat Transfer Coefficients Used in FEM Simulations

Stage	Property	Value
Transfer	Emissivity, e	0.65
Transfer	h (can-billet interface)	0.25 kW/m^2°C
Forging	h (can-billet interface)	2.5 kW/m^2°C
Forging	h (can-die interface)	20 kW/m^2°C

In all cases, however, a dwell time, prior to forging , of 40s was assumed. This dwell time comprised a 30s "conrolled dwell" period and the 10s required to place the billet on the dies and start the actual forging sequence. It was assumed that when the billet will be placed on the bottom die, it would be set on a pair of 0.8 mm diameter nichrome wires; hence, only radiation heat loss was condsidered for this period of the forging simulation.

Results and Discussions

The FEM simulations gave a detailed insight into the metal flow patterns, load-stroke curves, temperature transients, etc., during conventional forging of the canned gamma titanium aluminide. For brevity, major attention will be focused on the predicted metal flow patterns and grid distortions. The effects of specific can geometries and process parameters on metal flow are discussed next.

Side Pressing of the Canned Billet. In this series of simulations a 63.5mm diameter preform was canned in a 4.76 mm thick can. The can and preform were heated to 1250°C and cooled outside the furnace for 30 seconds and then placed on the bottem platen of the press. It was assumed that the billet would rest on the platen for 10 seconds before the deformation starts. The simulations were conducted at 10, 15, 25, 5, and 2 mm/s of ram speeds. Fig. 2 shows the undeformed (Fig. 2a) and deformed grids at the various velocities (Fig. 2b-2f). It may be noted, the deformation nonuniformity increases with the ram velocity. However, the can temperature dropped

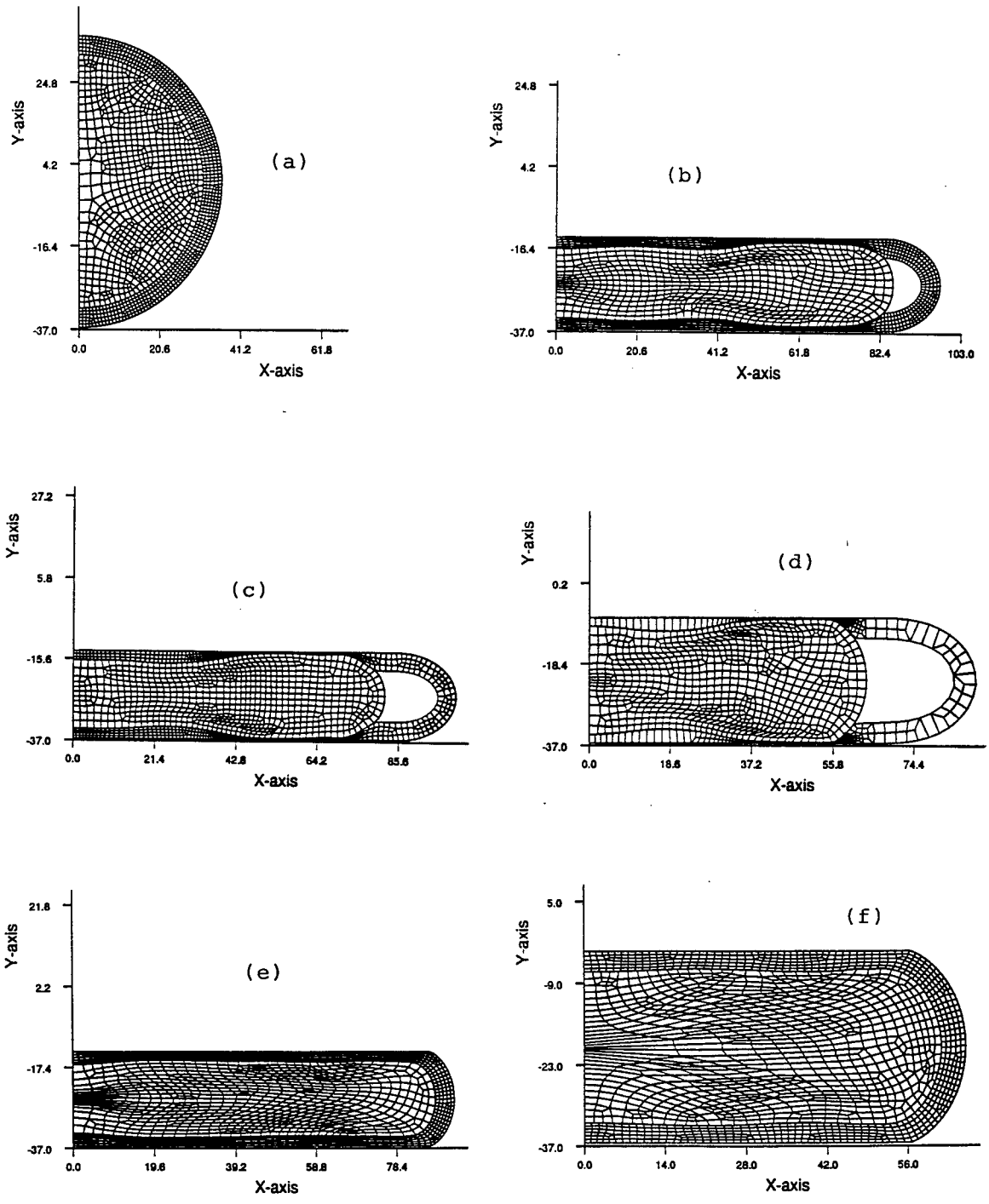


Fig.2 FEM predicted grid distortion for 33% deformation of a billet with 4.76 mm thick can and 63.5 mm diameter preform at various ram velocities: (a) undeformed FEM grid; ram velocity- (b) 10 mm/s, (c) 15 mm/s, (d) 25 mm/s, (e) 5 mm/s, and (f) 2 mm/s. Processing temperature (furnace)- 1250°C.

tremendously at the 2 mm/s ram velocity which might lead to can failure. It was, therefore, decided that the future simulations be conducted at the ram velocities of 5 and 10 mm/s only.

Side Pressing in Two Directions. In this case the simulations were conducted compressing the billet in the vertical and the horizontal directions, to model the billet compression in two directions which is accomplished by rotating the billet 90°. Here, two can thicknesses 2.5 and 1.5 mm and two ram velocities 10 and 5 mm/s were used. First, the billet was deformed along the y-axis and then along the x-axis. Each time the billet was deformed by 5 mm. Undeformed and deformed grids for a can thickness of 2.5 mm are shown in Figs. 3 and 4. Simulations exhibited in Fig. 3 are for a ram velocity of 10 mm/s and those in Fig. 4 for 5 mm/s. It may be noted that can separation occurs at some intermediate steps and that the lower velocity results in more uniform metal flow. Figs. 5 and 6 show the corresponding simulations for a can thickness of 1.5 mm. The smaller can thickness resulted in lower can temperature.

Billet Deformation in Longitudinal Mode

Deformation Between a Full and a Partial Die. Here the simulations were conducted to examine the material flow pattern in the longitudinal direction. A can thickness of 4.76 mm was used (Fig. 7a). The billet was heated to 1250°C and then cooled for 30 seconds to increase the can flow stress. An additional 10 seconds of cooling was incorporated to consider the time period during which the billet rested on the bottom die before the deformation started. Fig. 7(b) shows the simulation where the billet was supported by the lower die and compressed 15 mm with a partial die using a 20 mm bite. The die has sharp corners. In the simulation of Fig. 7(c), the die corner was rounded-off and an excellent material flow was obtained.

Fig. 8 shows the simulations which were conducted for a can thickness of 2.5mm, using two partial dies, at 5 and 10 mm/s ram velocities. Obviously, the lower velocity results in better material flow and smaller can separation.

Multi-Step Simulations. Fig. 9 shows a five-step simulation of the cogging process using two partial dies. Because the deformation is symmetrical, only one-half of the billet was simulated. A ram velocity of 5 mm/s was used. The die location is indicated by the arrow shown. Fig. 10 shows the simulations for a can thickness of 1.5 mm. In both the cases the billet was deformed by 5 mm in each step.

Fig. 11 shows the deformation pattern for a can thickness of 1.5 mm and ram velocity of 10 mm/s. Note the can deformation at the end cap for this case. Here, the can separation is higher than that for 5 mm/s ram velocity.

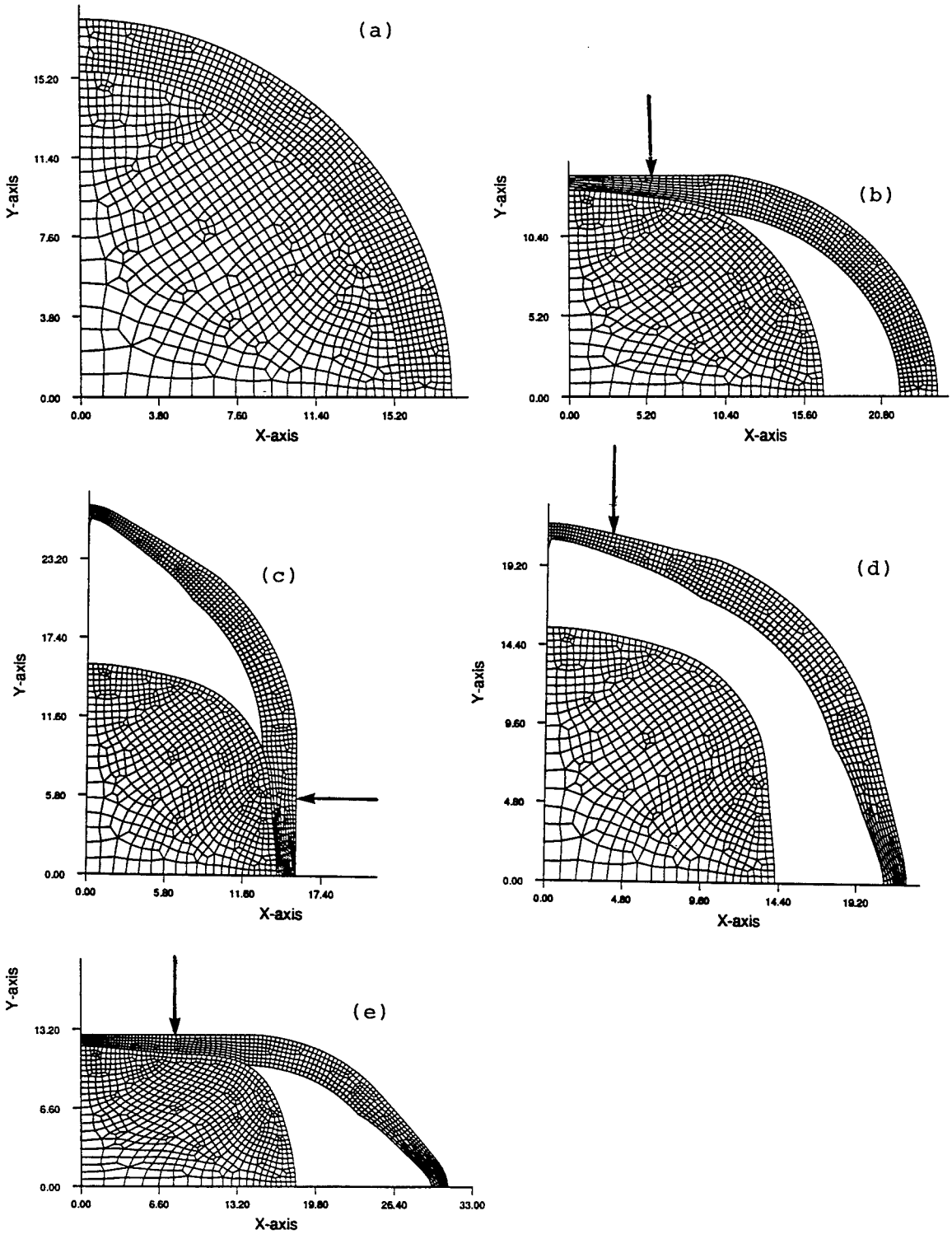


Fig.3 Sequence of FEM predicted grid distortions for two directional deformation of a billet with 2.5mm thick can and 31mm diameter preform at 10 mm/s ram velocity. Arrow indicates the ram movement. Deformation sequence- b to e.

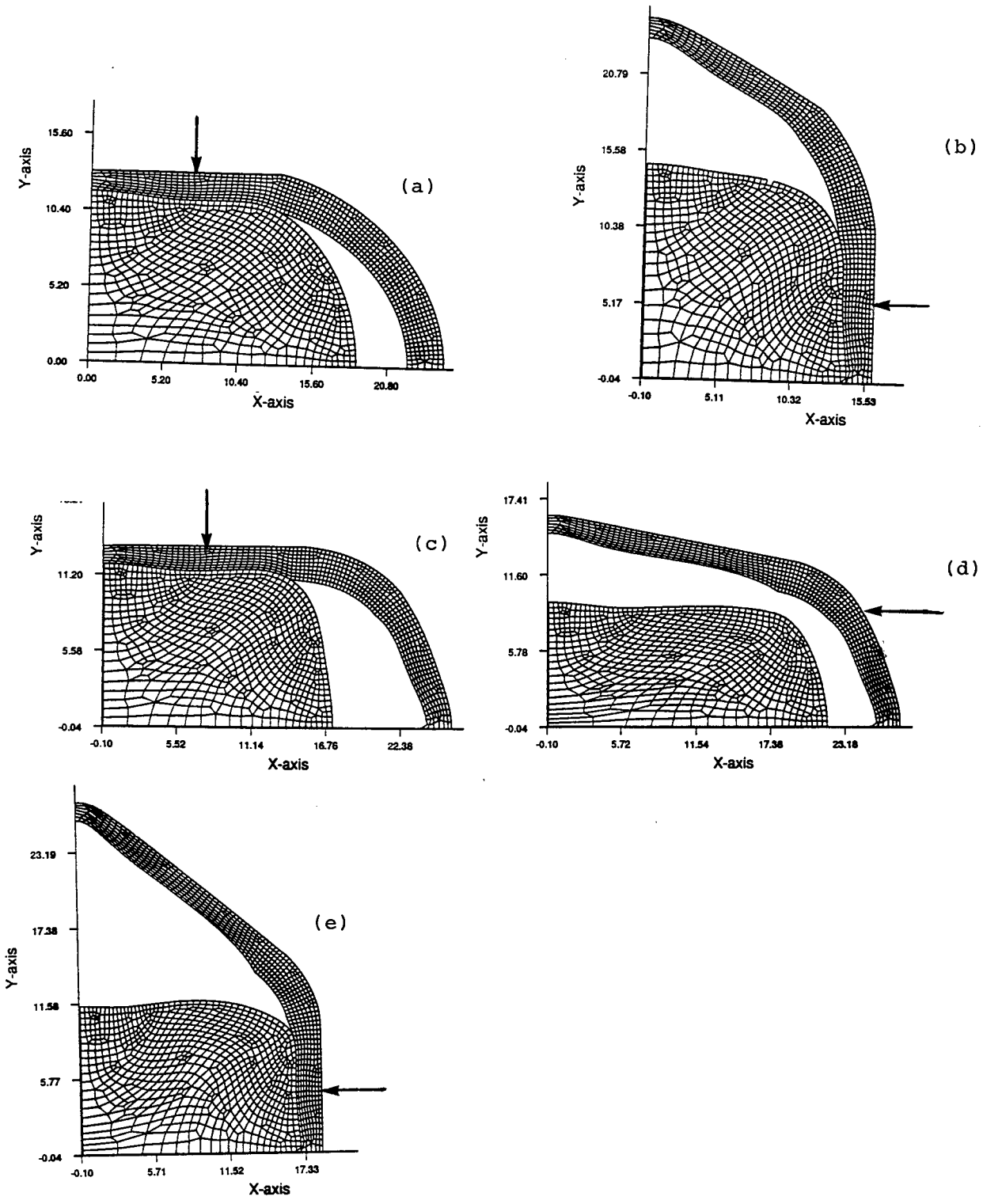


Fig.4 Sequence of FEM predicted grid distortions for two directional deformation of a billet with 2.5mm thick can and 31mm diameter preform at 5 mm/s ram velocity. Arrow indicates the ram movement. Deformation sequence- a to e.

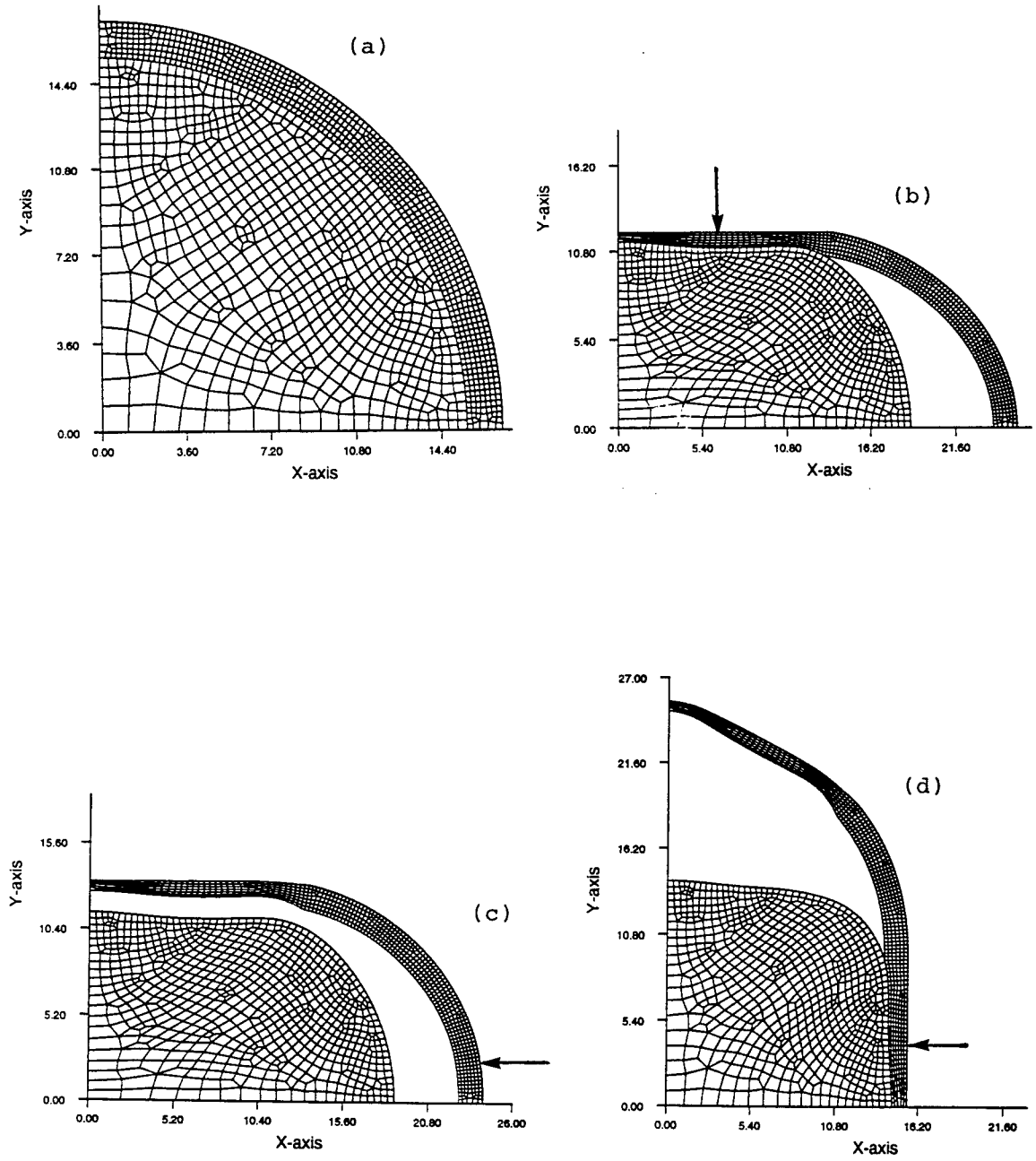


Fig.5 Sequence of FEM predicted grid distortions for two directional deformation of a billet with 1.5mm thick can and 31mm diameter preform at 10 mm/s ram velocity. Arrow indicates the ram movement. Deformation sequence- b to d

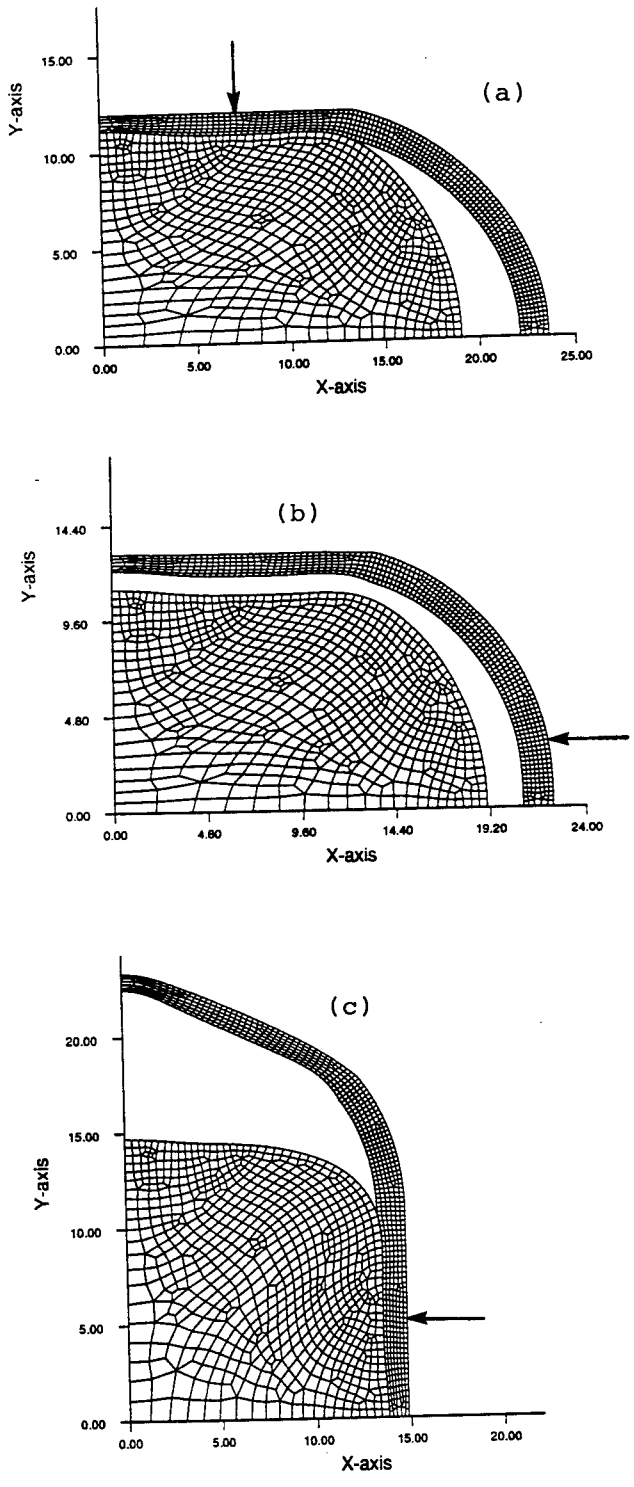


Fig.6 Sequence of FEM predicted grid distortions for two directional deformation of a billet with 1.5mm thick can and 31mm diameter preform at 5mm/s ram velocity. Arrow indicates the ram movement. Deformation sequence- a to c.

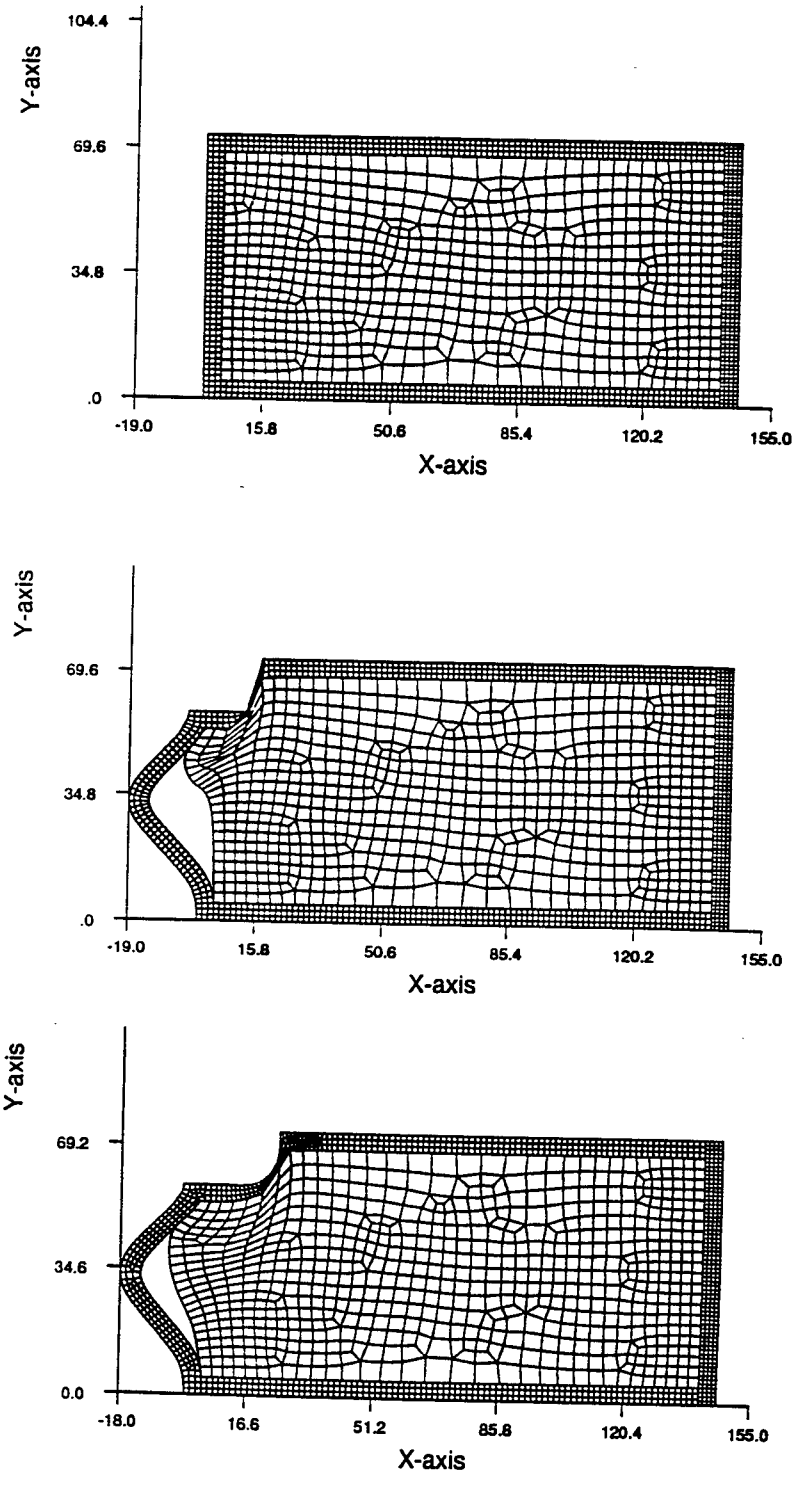


Fig.7 FEM predicted grid distortion of a billet with 4.76mm thick can and 63.5mm diameter preform using a full and a partial die: (a) undeformed FEM grid, (b) grid distortion with a sharp die, and (c) grid distortion with a rounded die. Processing conditions: billet temp. (furnace)- 1250°C, ram velocity-10 mm/s.

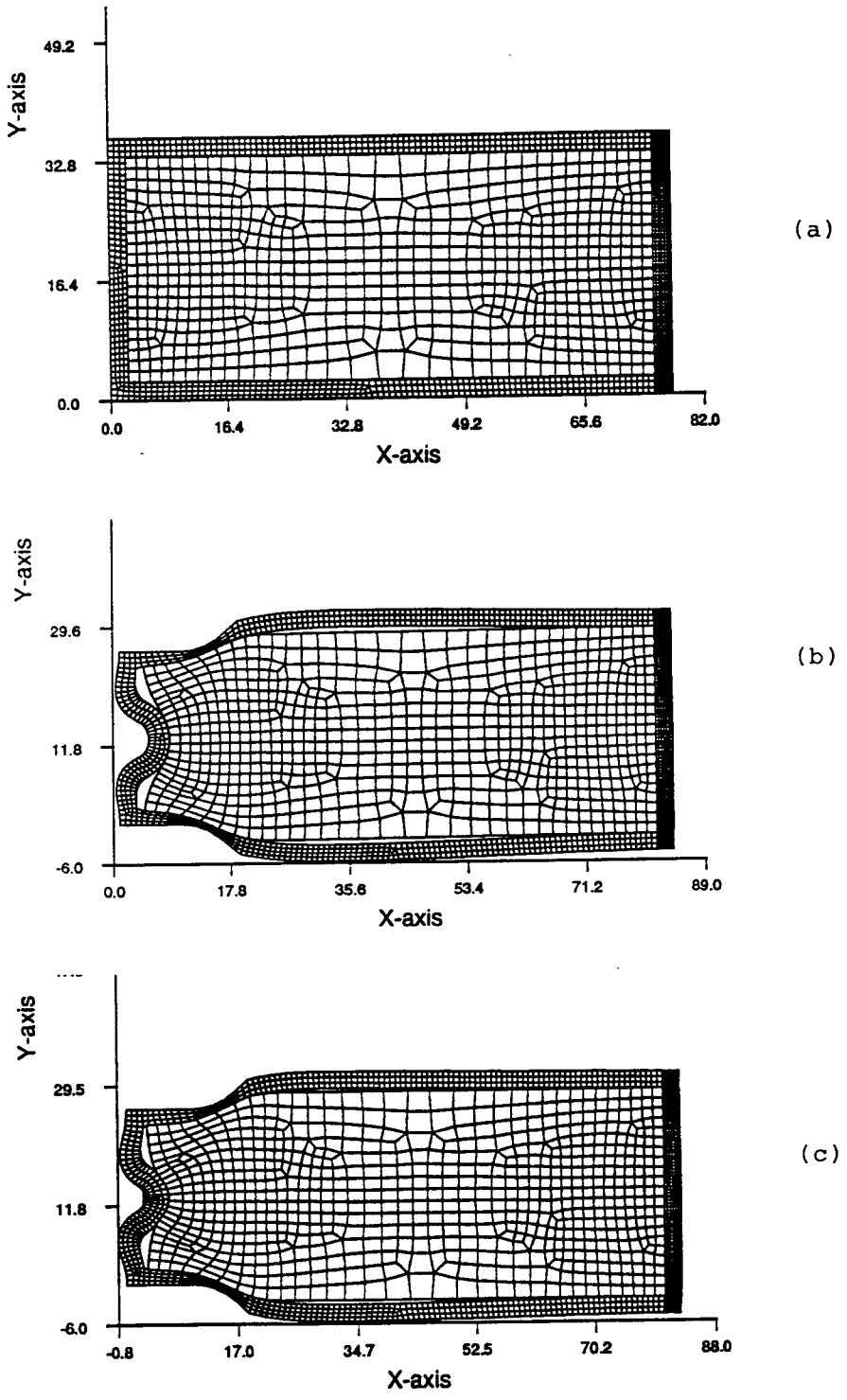


Fig.8 FEM predicted grid distortion for 13% deformation of a billet with 2.5 mm thick can and 31 mm diameter preform using two partial dies; (a) undeformed FEM grid, (b) grid distortion at 10 mm/s, and (c) grid distortion at 5 mm/s ram velocity. Processing conditions: billet temp. (furnace)- 1250°C.

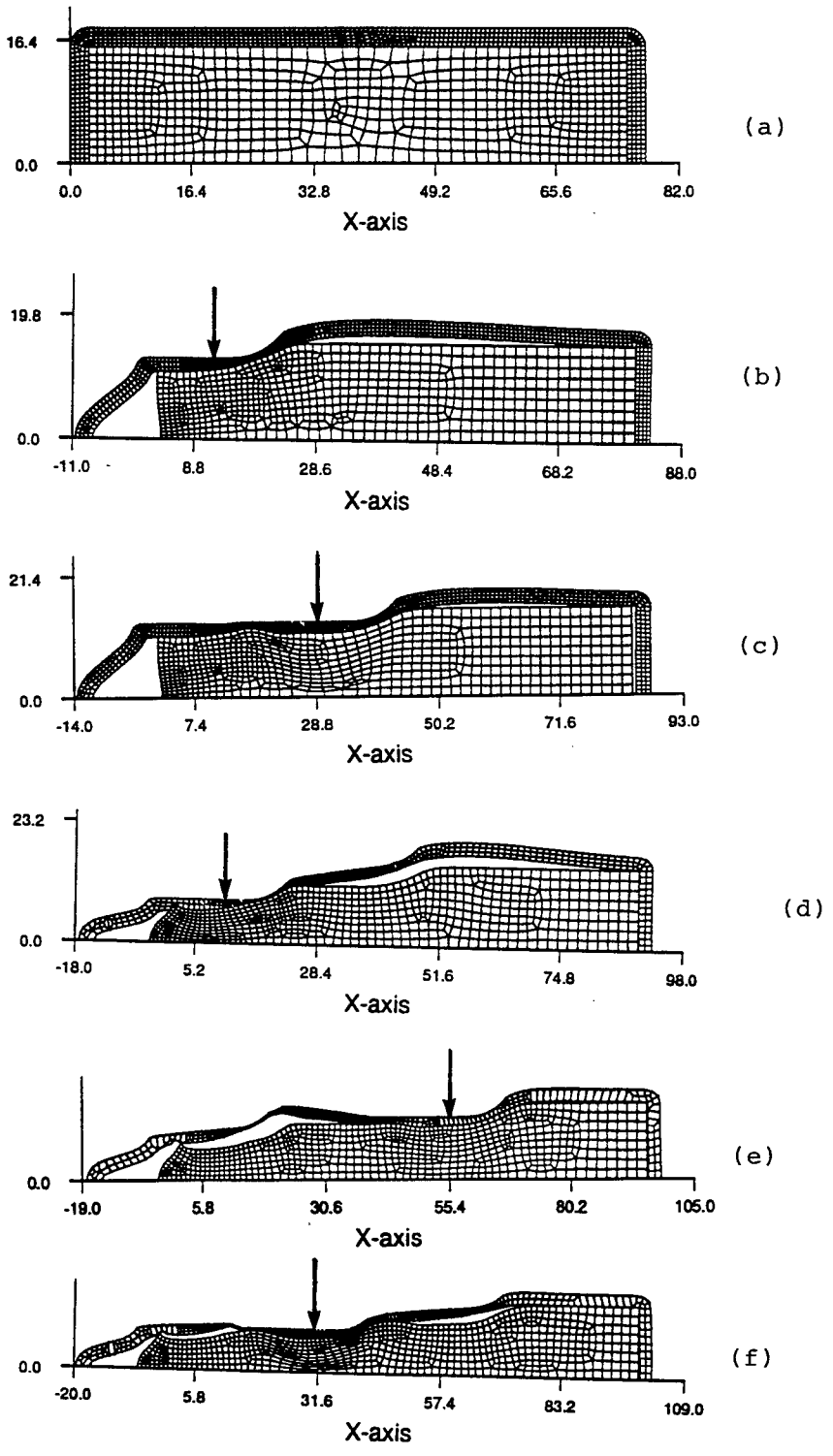


Fig. 9 Sequence of FEM predicted grid distortions of a billet with 2.5 mm thick can and 31 mm diameter preform at 5 mm/sram velocity in longitudinal mode. Arrow indicates the die position. Deformation sequence- b to f. Furnace temp.- 1250°C.

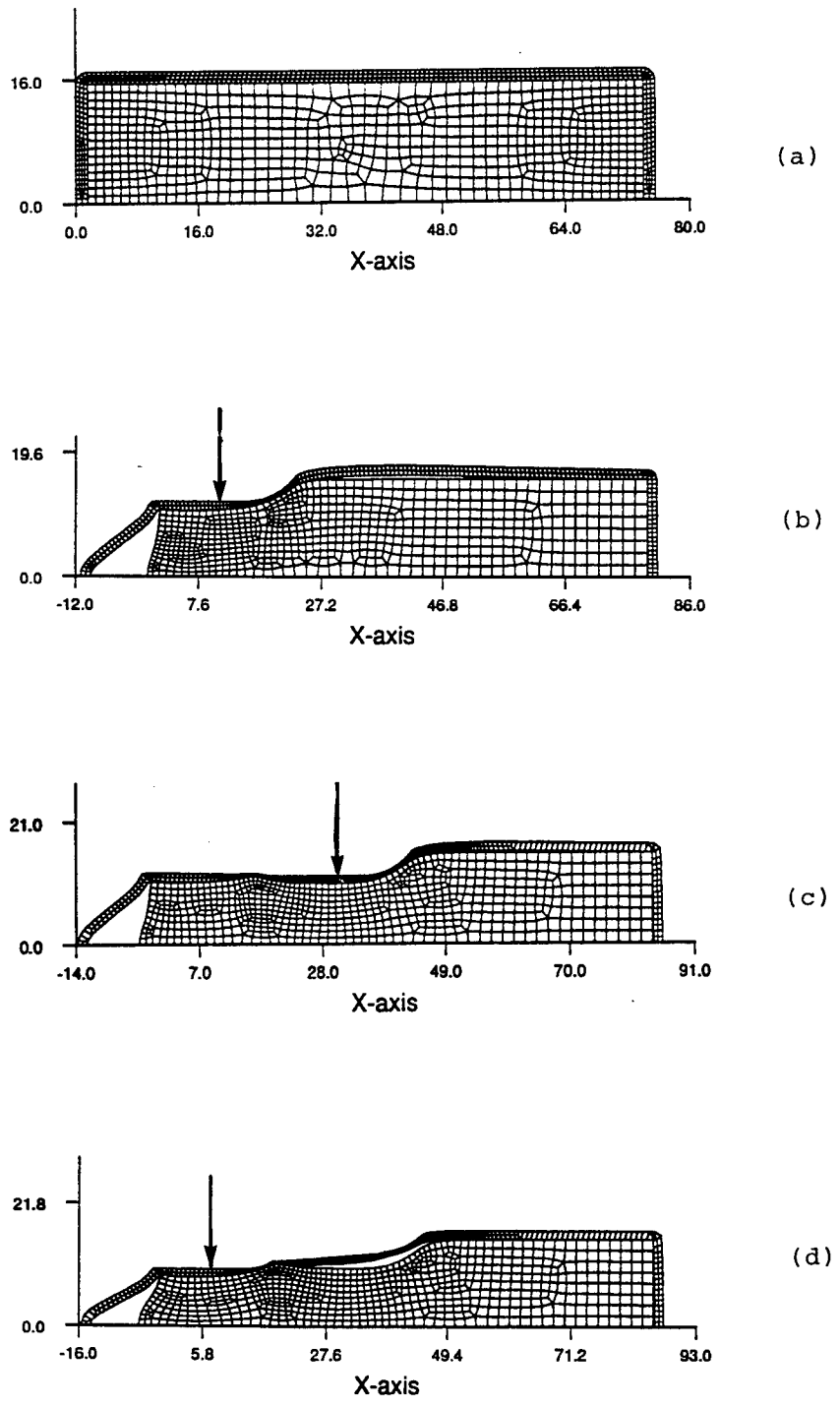
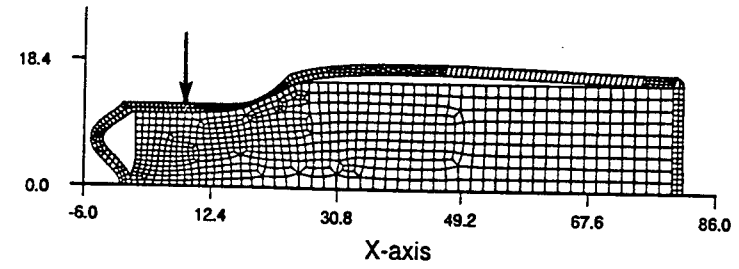
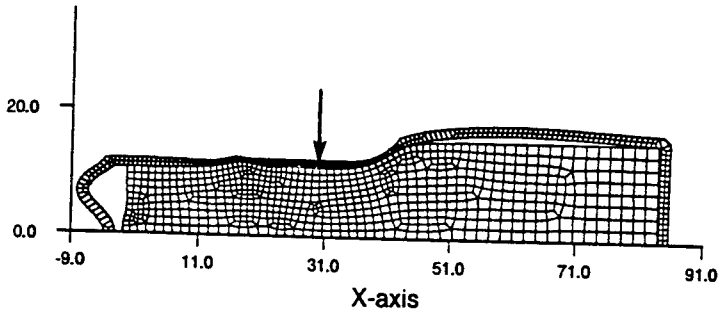


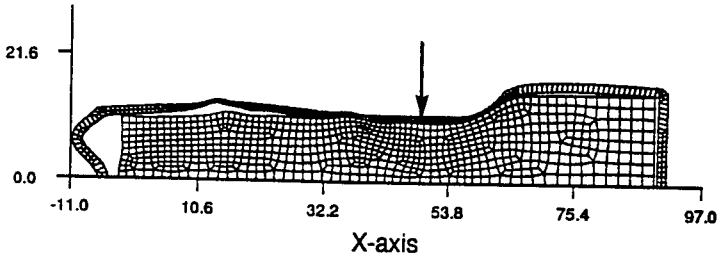
Fig. 10 Sequence of FEM predicted grid distortions of a billet with 1.5mm thick can and 31mm diameter preform at 5 mm/s ram velocity in longitudinal mode. Arrow indicates the die position. Deformation sequence- b to d. Furnace temp.- 1250°C.



(a)



(b)



(c)

Fig.11 Sequence of FEM predicted grid distortions of a billet with 1.5mm thick can and 31mm diameter preform at 10 mm/s ram velocity in longitudinal mode. Arrow indicates the die position. Deformation sequence- a to c. Furnace temp.- 1250°C.

Conclusions

A number of simulations were conducted to examine the material flow in canned Ti-Al preform during cogging process. The simulations indicate that lower velocity and smaller can thickness provide a more uniform flow of the can and the preform.

References

1. Semiatin, S. L., Oh, S.I., Florentino, R.J., Seetharaman, V., and Malas, J.C., 1991, "Hot Working of Titanium Aluminides - An Overview," *Heat-Resistant Materials*, K. Natesau and D.J. Tillack, ed., ASM International, Materials Park, OH, pp. 175-186.

2. Seetharaman, V., Goetz, R.L., and Semiatin, S., 1991, "Tensile Fracture Behavior of a Cast Gamma-Titanium Aluminide," *High Temperature Ordered Intermetallic Alloys IV*, J. Stiegler, et al., ed., Materials Research Society, Pittsburgh, PA, Vol. 213, pp. 895-900.

3. Seetharaman, V., Malas, J.C., and Lombard, C.M., 1991, "Hot Extrusion of a Ti-Al-Nb-Mn Alloy," High Temperature Ordered Intermetallic Alloys IV, J. Stiegler, et al., ed., Materials Research Society, Pittsburgh, PA, Vol. 213, pp. 889-894.

4. Goetz, R.L., 1990, "An Analysis of Canned Extrusion Using Analytical Methods and the Experimental Extrusion of Cast IN 100," Masters Thesis, Ohio University, Athens, OH

5. Goetz, R.L., Jain, V.K., and Lombard, C.M., 1992, "Effect of Core Insulation on the Quality of the Extrudate in Canned Extrusions of Gamma-Titanium Aluminides," *Materials Processing Technology*, Vol. 35, pp. 37-60.

6. Tomlison, A. and Stringer, J.D., 1969, J. Iron Steel Inst. London, Vol 193, pp. 157-162.

7. Srinivasan, R. and Weiss, I., 1992, unpublished research, Wright State University, Dayton, OH.

8. Semiatin, S.L. and Holbrook, J.H., 1983, "Plastic Flow Phenomenology of 304L Stainless Steel," *Metallurgical Transactions A*, Vol. 14A, pp. 1681-1695.

9. Henderson, J.B. and Groot, H., 1993, "Thermophysical Properties of Titanium alloys," Report TPRL 1284, Thermophysical Properties Research Laboratory, Purdue University, Lafayette, IN.

10. Aerospace Structural Metals Handbook, Vol. I (Ferrous Alloys), 1968, AFML-TR-68-115, U.S. Air Force Materials Laboratory, Wright-Patterson AFB, OH.

11. Aerospace Structural Metals Handbook, Vol. II (Non-Ferrous Alloys), 1968, AFML-TR-68-115, U.S. Air Force Materials Laboratory, Wright-Patterson AFB, OH.

12. Aerospace Structural Metals Handbook, Vol. IIA (Non-Ferrous Alloys), 1968, AFML-TR-68-115, U.S. Air Force Materials Laboratory, Wright-Patterson AFB, OH.

13. Jain, V.K. and Goetz, R.L., 1991, "Determination of Contact Heat-Transfer Coefficient for Forging of High-Temperature Materials," presented at the ASME/AIChE National Heat Transfer Conference held in Minneapolis, MN, July 18-20, 1991, ASME 91-HT-34.

14. Semiatin, S.L., Collings, E.W., Wood, V.E., and Altan, T., 1987, "Determination of the Interface Heat Transfer Coefficient for Non-Isothermal Bulk-Forming Processes," *J. Eng. Ind.*, Trans. ASME, Vol.109, pp 49-57.

15. Burte, P.R., Semiatin, S.L., and Altan, T., 1990, "An Investigation of the Heat Transfer and Friction in Hot Forging of 304 Stainless and Ti-6Al-4V," *Proc. of the 18th N. American Mfg. Res. Conf.*, SME, Dearborn, MI, pp.59-66.

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 Faculty Research Program
Wright Laboratory

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

and

Wright Laboratory

September 1994

MULTIDIMENSIONAL ALGORITHM DEVELOPMENT AND ANALYSIS

J. Mark Janus
Assistant Professor
Department of Aerospace Engineering
Mississippi State University

Abstract

Recent experience with multidimensional vortex dominated flows has given an indication that conventional algorithms in general do not live up to their traditional high–resolution capabilities. An interesting example is when field based methods of inferring vehicle loads which involve interrogation of the solution vortical field are used. It has been shown that these techniques can perform quite well yet are highly sensitive to spurious vorticity. Hence even seemingly minor errors in the vortical field can yield inadequate results. Another area of intense interest is computational aeroacoustics in which the degree of accuracy necessary requires extremely close attention to detail (physics) due to pertinent solution gradients being orders of magnitude smaller than that which the "high–resolution" schemes were designed. The effort here has been to develop and analyze a flow model incorporating multidimensional physics with only limited modifications to existing conventional flow software. Around 1986, Professor Charles Hirsch, et.al. presented a rather significant technique to optimally decouple the multidimensional Euler equations (inviscid equations of fluid motion). 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. For this effort the flow domain is restricted to two–dimensions. The base software is 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.

MULTIDIMENSIONAL ALGORITHM DEVELOPMENT AND ANALYSIS

J. Mark Janus

Introduction

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. The theory behind one such approach forms the basis of the algorithm developed in the effort reported on herein.

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.

Recent experience with vortex dominated flows has given indication that conventional algorithms in general do not live up to their traditional high–resolution capabilities. An interesting example is when field–based methods of inferring vehicle loads which involve interrogation of the solution

vortical field are used. It has been shown that these techniques can perform quite well, yet they are highly sensitive to spurious vorticity [11]. Hence even seemingly minor errors in the vortical field can yield inadequate results. Another area of intense interest is computational aeroacoustics in which the degree of accuracy necessary requires extremely close attention to detail (physics) due to pertinent solution gradients being orders of magnitude smaller than that which the "high–resolution" schemes were designed. These are but a couple of reasons to keep interest alive and well in the latest algorithms currently under development.

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 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 [1] and the multidimensional advection algorithm(s) presented in [12]. Two implementations of the multidimensional decoupling procedure were investigated; the first was similar to the characteristic split procedure outlined in [1] and ultimately proved to be unstable, the second utilizes the fluctuation splitting theory outlined in [12] yet does so on cell–centered quadrilateral–based data. This report focuses on the latter of the two implementations.

Methodology

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:

$$\partial \boldsymbol{W}^* = \begin{bmatrix} \partial\varrho - \frac{1}{c^2}\partial p \\ \boldsymbol{l}^{(1)} \cdot \partial \boldsymbol{V} \\ \boldsymbol{k}^{(2)} \cdot \partial \boldsymbol{V} + \frac{\partial p}{\varrho c} \\ -\boldsymbol{k}^{(2)} \cdot \partial \boldsymbol{V} + \frac{\partial p}{\varrho c} \end{bmatrix} \quad , \qquad \boldsymbol{S} = \begin{bmatrix} 0 \\ -\frac{c}{2}\,(\boldsymbol{l}^{(1)} \cdot \nabla)\,(W^3 + W^4) \\ -c\,(\boldsymbol{l}^{(1)} \cdot \nabla)\,W^2 \\ -c\,(\boldsymbol{l}^{(1)} \cdot \nabla)\,W^2 \end{bmatrix} .$$

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 equationsof 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 $\boldsymbol{k}^{(2)} = \boldsymbol{k}^{(1)}$.

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 [13]. 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** [14] given as:

$$\boldsymbol{Z} = \begin{bmatrix} \sqrt{\varrho} \\ \sqrt{\varrho}\,u \\ \sqrt{\varrho}\,v \\ \sqrt{\varrho}\,H \end{bmatrix} \quad , \tag{6}$$

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 [14]. 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 that used in this study.

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

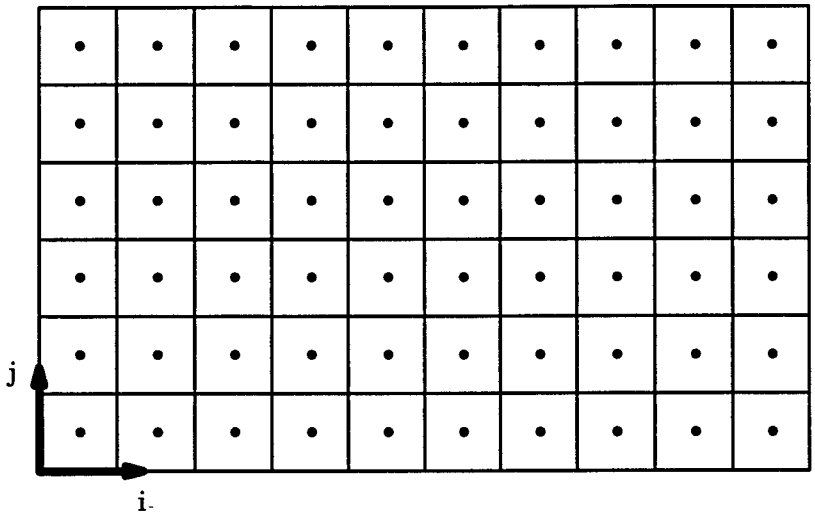


Figure 1. Primary Quadrilateral Grid

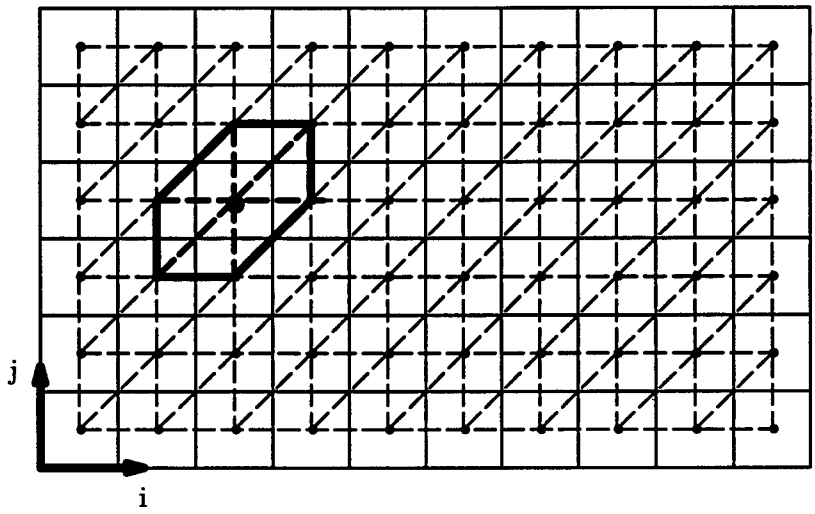


Figure 2. Sub–grid Development

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 [15]. This appears to work yet may be the source of solution difficulties encountered during this effort. A more rigorous treatment of the boundaries is in order.

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).

Results

In order to test the modified software, 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. 3.

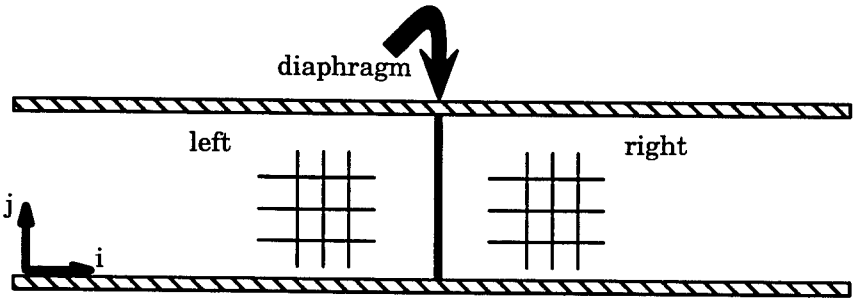


Figure 3. Shock Tube Configuration

The solutions obtained from this test case have been puzzling. For a vertical grid (i.e. the cells of the quadrilateral grid are squares), the solution coming from the fluctuation split code is identical to that produced from a one–dimensional finite–volume approximate Riemann solver using Roe averaging. When the grid is slanted the solution deteriorates as shown in Figs. 4 and 5. The cause for this is presently not known. Many of the ingredients critical to the solution process have been interrogated to determine the source of the solution deterioration, yet nothing has shown significant promise as to being the source. Currently one area of interest is 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 minimizing the source term associated with the second wave vector appear to actually zero the term. An approach is being investigated in this effort which truly zeroes the term when possible. Presently the source term is advected (distributed) in the direction of the wave propagation. This may be de–stabilizing to the solution when the source term is not appropriately minimized (i.e. when the choice of $\boldsymbol{k}^{(2)}$ still yields a source term of significant magnitude). As mentioned previously the dot product Eq. (4) can not be allowed to become too

small, thus it is the $\vec{k}^{(2)}$ vector which gets arbitrarily reset to insure that this does not happen. When this is done a minimal source term for the last two wave equations is not obtained.

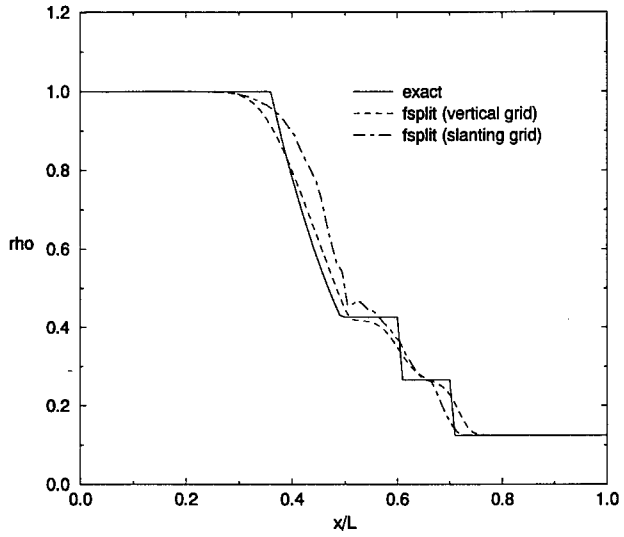


Figure 4. Shock Tube Density Solution (lower wall)

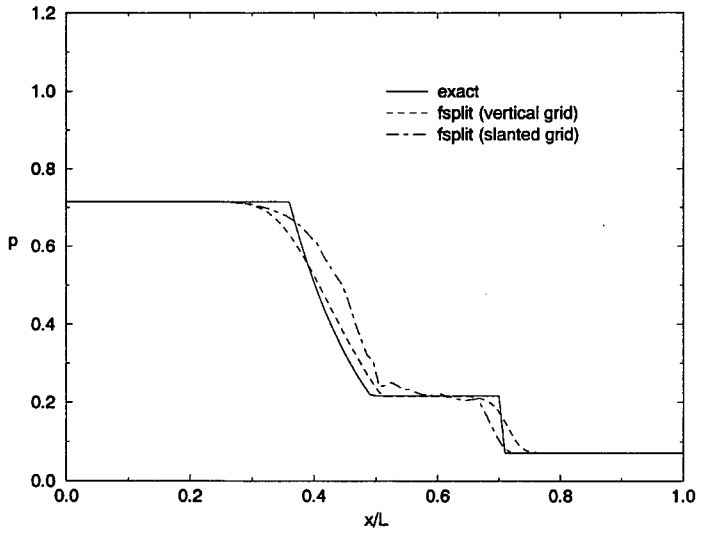


Figure 5. Shock Tube Pressure Solution (lower wall)

The second test case was that of a 15° confined ramp with an inlet Mach number of 1.9. 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 interac-

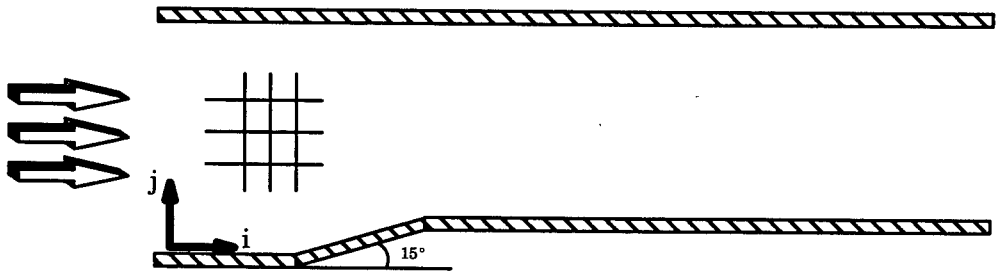


Figure 6. 15° Confined Ramp Configuration

tion with an expansion fan emanating from the ramps downstream corner. Due to the complexity

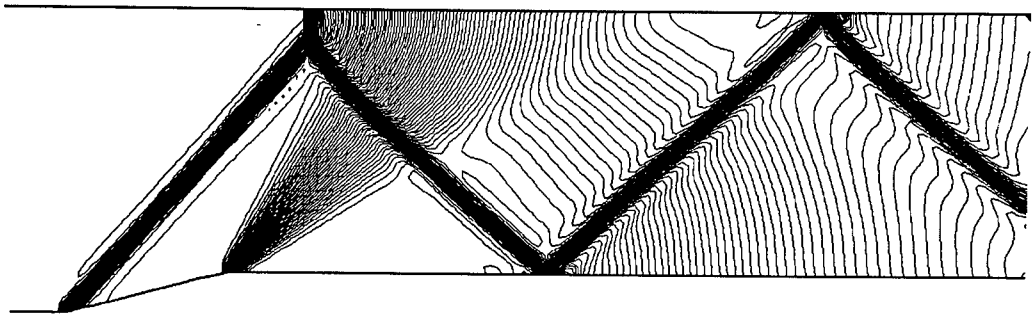


Figure 7. Second–Order Accurate Fine–Grid Solution (conventional algorithm)

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 second–order accuracy is shown in Fig. 7. The figure shows pressure contours making the oblique shocks and expansion fan obvious. 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. 8 and 9.

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 currently being investigated is again 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. This feedback has been noted in the research of others as having hindered convergence. Infrequently updating the wave vectors has not corrected the problem and sometimes results in solution divergence. At present only intermediate solutions have been obtained (and scrutinized), see Fig. 10. It is noted in the intermediate solution that the position of the reflecting shocks does not coincide with that shown in the high–order solution 'standard'. This may be caused by the use of the original phantom cells to main-

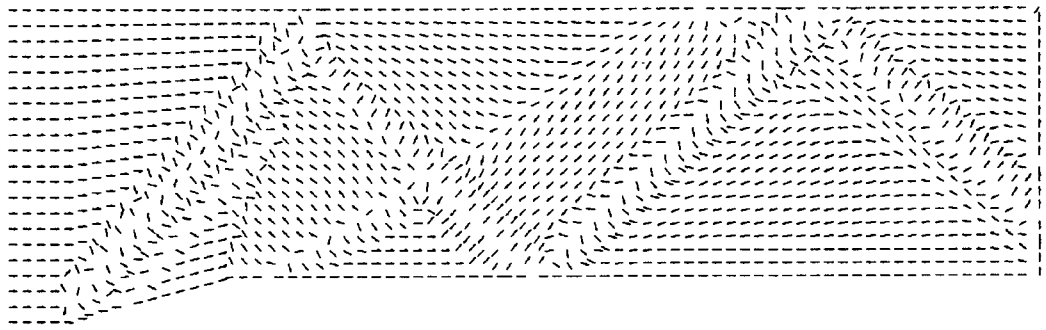


Figure 8. First Wave Vector $\boldsymbol{k}^{(1)}$ (aligned with pressure gradient)

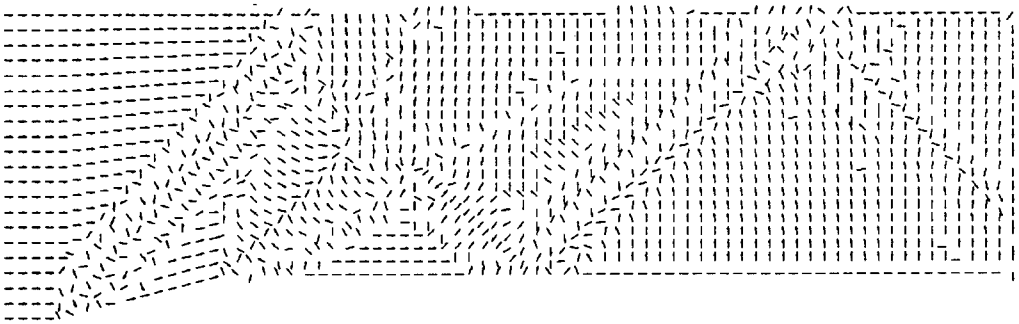


Figure 9. Second Wave Vector $\boldsymbol{k}^{(2)}$ (chosen here to truly zero the source term)

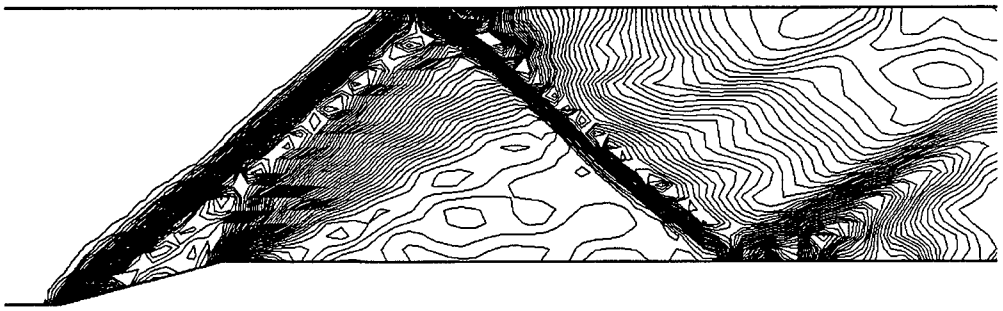


Figure 10. Intermediate Solution from Fluctuation Split Code

tain the boundary conditions. It also tends to make one think that the wave speeds used in the algorithm may be in error in some way. Each of these potential sources are being investigated at the present time. It should be noted that the solution produced using the fluctuation split code has only one–quarter the mesh cells of that shown in Fig. 7. For true comparison purposes, consider the solution shown in Fig. 11. This solution was produced using a conventional algorithm, first–order accuracy and the same primary grid as that used for the fluctuation split code. Note the degradation

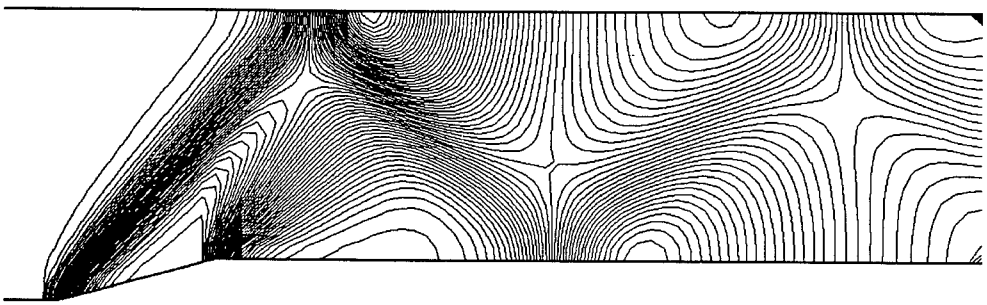


Figure 11. First–Order Accurate Coarse–Grid Solution (conventional algorithm)

in resolution 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. Therefore a serious look at the wave speed and boundary condition treatment is slated for the fluctation split code to determine the problem.

Conclusions

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 development and debugging phase. One goal of this project has been met, in that the modifications to the base software included just three new routines (one replacing an original routine). Thus far the solutions obtained show the implementation has promise, although how they stack up to solutions produced from more conventional software (compared side by side) has yet to be 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] P. L. Roe, R. Struijs and H. Deconinck. A conservative linearization of the multidimensional Euler equations. *to appear, Journal of Computational Physics*, 1993.

[14] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43:357–372, 1981.

[15] 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.

CHARACTERIZATION OF INTERFACES IN METAL MATRIX COMPOSITES

Iwona M. Jasiuk
Associate Professor
Department of Materials Science and Mechanics

Michigan State University
East Lansing, MI 48824-1226

Final Report for:
Summer Faculty Research Program
Wright-Patterson Air Force Base

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

September 1994

CHARACTERIZATION OF INTERFACES IN METAL MATRIX COMPOSITES

Iwona M. Jasiuk
Associate Professor
Department of Materials Science and Mechanics
Michigan State University
East Lansing, MI 48824-1226

Abstract

The characterization of fiber-matrix interfaces in metal-matrix composites is the primary focus of this research program. This research is in collaboration with the Materials Research Division at Wright Patterson Air Force Base (WPAFB) and it is still in progress. There are two components of this research: the Theoretical and Numerical Part, which is described in this report, and the concurrent Experimental Program being conducted in Dayton at WPAFB under the direction of Dr. Dan Miracle in the Materials Research Division. The goal of this research is to develop test methods suitable for analyzing interfaces in metal-matrix composites and to develop the theoretical understanding of these tests so the results can be correctly analyzed and interpreted. This will advance the understanding of these complex materials and will lead to the design and manufacture of superior and more reliable composite materials.

CHARACTERIZATION OF INTERFACES IN METAL MATRIX COMPOSITES

Iwona M. Jasiuk

Introduction

In composite materials the fiber-matrix interface plays a crucial role and it influences both the local stresses and the effective properties of composites (see e.g. Drzal and Madhukar, 1993). For example, stiffness and strength depend of the load transfer across the interface, toughness is affected by the fiber pull-out or crack deflection mechanisms, ductility is influenced by the relaxation of high stresses near the interface (Clyne and Withers, 1993). It is not surprising, therefore, that much of the recent research in the area of the mechanics of composite materials has focused on the interfaces, but due to the complexity of the subject many issues still remain unresolved. A fundamental problem in this area is how to characterize the interfaces.

In this research we focus on the characterization of interfaces in metal-matrix composites.

Background Information

1. Metal-Matrix Composites

Metal-matrix composites (MMC) are becoming very strong candidates as structural materials for high-temperature and aerospace applications (Rosenstein, 1991), and are also being used in automotive and general engineering fields (see e.g. Chawla, 1987; Schoutens, 1989; Taya and Arsenault, 1989). The metal matrix has numerous advantages. These include high tensile strength and Young's modulus, high melting point, small coefficient of thermal expansion, increased creep and wear resistance, good resistance to moisture, dimensional stability, joinability, high ductility, high toughness and impact properties, high surface durability, high electrical and thermal conductivity, and other. In addition, MMCs are resistant to severe environments and retain strength at high temperatures. For high temperature applications MMCs have more desired properties than polymer matrix composites and are more reliable than ceramic matrix composites. From the scientific point of view, however, MMCs are very complex and therefore

still not well understood. The complicating factors are nonlinear behavior of the matrix including plasticity (Doghri and Leckie, 1994), extremely complex fracture and fatigue behaviors, and other factors (see e.g. Majumdar and Newaz, 1992).

There are many factors that influence the properties of MMCs. These include:

- properties of reinforcement, surface finish, wettability, crystalline perfection, shape (aspect ratio) of reinforcement, geometric arrangement of fibers (fiber spacing and orientation), and volume fraction of reinforcing fibers;

- properties of matrix, both elastic and plastic (work hardening), dislocation density, grain structure, ductility, alloying chemistry, toughness, and effects of porosity;

- fiber/matrix interface propeties and structure, interfacial bond strength, thickness of reaction zone, interface constituents, differences in Poisson's ratios and thermal expansion coefficients, reactions at the interface, mechanical aspects of interface, interdiffusion of constituents, and effects of fiber spacing on matrix hardening;

- residual stresses resulting from the thermo-mechanical history of composite;

- effects on constituent properties of processing variables such as pressure, temperature, porosity, fiber breakup and misalignment, degradation of fibers due to high temperature reactions, mechanical damage induced by processing, machining, handling, impact, and other factors (Schoutens, 1989).

Out of these factors the reinforcement/matrix interface and mechanics of fracture are the two most outstanding and least understood problems.

2. Interfaces in Metal-Matrix Composites

In most MMCs a gradient in chemical potential exists at the fiber/matrix interface. This difference in the potential gives rise to diffusion and chemical reaction processes when composites are subjected to high temperatures during processing. Thus the interaction zone, called interphase, develops. If this region is a few Angstroms in thickness, it is desirable for good bonding but its overgrowth is detrimental to the overall properties of composites. The interphase region consists of the fiber and matrix materials that have reacted over some radial distance from the fiber to produce various oxides or carbides (Cantonwine and Wadley, 1994; Das, 1990; Loretto and Koitzer, 1990; Rosenstein, 1991; Schoutens, 1989; Yang and Jeng, 1989). These have in general a greater specific volume than other components and thus introduce internal stresses. In

addition, as expected from the diffusion theory the density of such oxides or carbides will vary radially giving rise to "gradients" in elastic properties in the interphase. These complex processes at the interface may also yield multiple interfaces (Schoutens, 1989). This greatly complicates the study of interfaces in MMCs.

In studying the interfaces it is important to distinguish different types of bonding. These include mechanical and chemical bondings. The mechanical bonding involves mechanical gripping of the fibers by the matrix. The chemical bonding includes wettability bonding which occurs at an electronic scale and reaction bonding involving transport of atoms controlled by diffusion.

3. Tests to Characterize Interfaces

There is a number of experiments that are currently used to measure the properties of interfaces in composite materials. These include the fragmentation test, the pull-out test, the droplet test, the push-through test, the push-in (microindentation) test, the transverse test, the slice compression test, and other.

The fragmentation test involves embedding a single fiber in the matrix, loading the specimen until fiber breaks into segments, and then measuring the mean aspect ratio of the broken fiber pieces. In this test the loading is applied until the fiber lengths become so short that the shear stresses at the interface do not cause enough tensile stress in the fiber to result in any additional breaks. Traditionally, these measurements have been done optically on polymer matrix composites, which are transparent, by using light microscope. An acoustic emission technique, developed recently, does not require a transparent matrix and therefore can be also used for ceramic and metal matrix composites (Waterbury et al., 1994; Wooh and Daniel, 1994). However, the sensitivity of this method requires that fibers need to be large in diameter. One of the criticisms of this test is that it is not clear what interfacial characteristics are being measured there (Clyne and Withers, 1993). This test has been considered either experimentally or analytically by Curtin (1991), Di Anselmo et al. (1992), Feillard et al. (1994), Gent and Liu (1991), Jones and DiBenedetto (1994), Nedele and Wisnom (1994), Wagner and Eitan (1993), Waterbury and Drzal (1991), Yang and Knowles (1992), and others.

The fiber pull-out test usually involves a single fiber embedded in a matrix. A steadily increasing force is applied at the free end of the fiber until either the pull-out occurs or the fiber

breaks, while the load-displacement characteristics are recorded. The papers involving this test include those of Atkinson et al. (1982), Bartos (1980), Betz (1982), Dollar and Steif (1988), Hsueh (1990a,b; 1991a,b; 1993b), Kendall (1975), Kim et al. (1993), Marotzke (1994), Marshall (1992), Penn and Chou (1990), Zhou et al. (1992), Yue and Cheung (1992), and others. This test has been extensively applied to polymer composites, but it has been also used to obtain critical stresses in MMCs. However, there are difficulties in the preparation and handling of the specimen of MMC. Since the matrix is relatively stiff and the bonding is usually good, the fiber often prematurely breaks.

Similar to this method is the microdrop technique (droplet test), which involves the fiber pull-out from the spherical droplet of epoxy. The disadvantages of this method are that the results are very sensitive to the position and type of support used. This factor may be responsible for a high scatter in results obtained from this test.

Another method is the fiber push-in test (also called micro-indentation technique), which has been proposed by Marshall (1984) and Mandell and his coworkers (Grande et al., 1988), in which a standard microindentation hardness tester, such as Vickers pyramid, can be used. The preparation of the specimen involves polishing of a surface of a composite having continuous fibers aligned perpendicular to its surface. Then fibers are individually compressively loaded until debond and/or slip occurs at the interface. The advantage of this method is that it is an *in-situ* test for real composites. This test has been applied mostly to ceramic-matrix composites. In this case the matrix is linearly elastic and the interface debonds easily. When this test is used for testing the interfaces in metal-matrix composites, due to a stronger bond, it often yields to the damage of the indentor before any debond is observed. The fiber push-through method (push-out) is closely related to the push-in test. The difference is in the smaller thickness of the specimen, which allows the fiber to slide through. Theoretical and experimental studies in this area include Bright et al. (1989), Eldridge (1992), Eldridge and Brindley (1989), Ferber et al. (1993), Hsueh (1990b,c,d), Hsueh et al. (1989), Kallas et al. (1992), Koss et al. (1993, 1994), Mackin and Zok (1992), Majumdar et al. (1993), Marshall and Oliver (1987), Marshall et al. (1992), Mital and Chamis (1991), Parthasarathy et al. (1991), Shetty (1988), Wang et al. (1992), Warren et al. (1992), Watson and Clyne (1992a,b), Weihs and Nix (1988, 1991), and others.

The detailed discussion on the above four tests and their comparison and the additional list of references are given in Herrera-Franco and Drzal (1992) and Herrera-Franco et al. (1992).

The transverse test involves observations of the interfacial damage initiation and propagation on the transverse polished surface of a unidirectional fiber reinforced composite, when a remote uniaxial load is applied. The analytical and numerical solutions for this type of loading and geometry have been considered by Folias (1991), Highsmith et al. (1990), Marshall et al. (1994), Nimmer et al. (1991), and others.

Fiber protrusion/intrusion during thermal cycling (Cox, 1990) has also been proposed to study the interfacial characteristics. The measurements of relative displacement of fiber and matrix surfaces during heating and cooling provide estimates of the interfacial shear strength and residual stresses. This test has been applied to titanium and titanium aluminide composites with continuous SiC fibers (Cox et al., 1992) and a shear lag type of analysis (approximate) was used to analyze the results.

The interpretation of the results of all these tests is still an open question as the corresponding mechanics analyses are usually of strength of materials type, i.e. approximate, and involve the shear lag model. Also, the finite element studies often do not account for all the phenomena occuring in the composite.

In case of metal-matrix composites the problem is even more complicated as the currently existing tests may not be fully suitable for these systems and they may need to be modified. Also, the theoretical analyses for the metal-matrix composites are much more complex due to the nonlinear behavior of the matrix, plasticity, residual stresses, complex crack propagation processes, and a complicated microstructure of the interphase. The experimental challenges include the fact that the matrix is non-transparent and due to a stronger bond and the stiffer material properties the tests are more difficult to conduct. The crucial issue is what tests for characterizing the interfaces are suitable for MMCs (Clyne and Withers, 1993).

There is a number of important issues involved in the characterization of interfaces. A common approach has been to determine the stress level at which the damage (inelastic processes) initiate. However, this is a complex issue as different combinations of stresses may initiate the same process. Another problem involves measuring of interfacial toughness and this requires very careful experimental procedures, which are very difficult for metal-matrix composites.

Discussion of research

The experimental characterization of fiber-matrix interface in metal-matrix composites is the subject of current studies in the Materials Research Division at Wright Patterson AFB under the direction of Dr. Dan Miracle. The following four tests involving the characterization of intefaces are currently being explored there: the SLICE COMPRESSION test, the PUSH-IN test, the FRAGMENTATION test, and the TRANSVERSE test. These tests are applied to unidirectional fiber reinforced metal-matrix composites; the composite systems studied are titanium-based metal-matrix composites with continuous SiC fibers. During my eight weeks long visit at WPAFB I have done a throughout review of literature covering the experimental procedures and theoretical analyses of the above tests. Also, I have interacted with the scientists involved in the experimental aspects of these tests and brainstormed on the key issues involved in each of these tests. Since I was most involved in the slice compression test, which is the newest and thus least understood test, I have conducted some preliminary finite element calculations in order to understand better the physics of this test, and have done some preliminary calculations pertaining to this test. This work is currently in progress.

In the following I describe the slice compression test, and the test related to it, the HIDE test.

1. The SLICE COMPRESSION Test

The SLICE COMPRESSION test was introduced by Shafry, Brandon and Terasaki (1989) and has been used so far on ceramic matrix composites. In our research we use it on the metal-matrix composites. This test involves pressing of the polished transverse surface of a unidirectional metal-matrix composite on the brass surface as shown in Fig. 1. As the matrix is more compliant than the fibers, the fibers make imprints on the brass specimen. Accurate measurements of the depths of these indented areas can be made for a given applied load. The challenge here is how to interpret these results and thus there is a need for the mechanics solution. This is a very complex contact problem for several reasons:

- This is an elasto-plastic problem with a complicated geometry, thus intractable analytically; also the measured data accounts for plastic deformation only, not the elasto-plastic deformation which is present under loading. Brass is a material having elastic-strain hardening behavior. The problem of indentation of an elastic block on such a material is a very complex one and the closed form solution does not exist. The additional complication is the inhomogeneous nature of the

indenting material, which is a composite material consisting of the matrix and the fibers. However, if one desires the analytical solution to this problem one needs to know what load is transferred by the brass to the composite. Knowing this one can proceed with the elasticity solution as, based on the preliminary calculations, it is believed that the matrix will not undergo extensive yielding under the loads applied during the test.

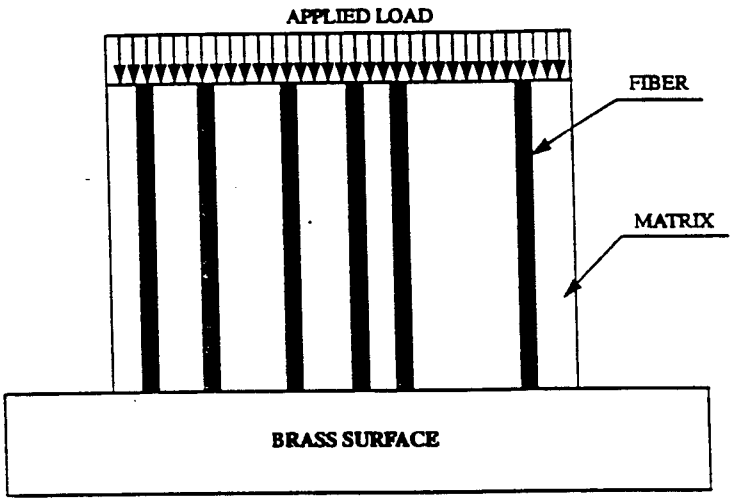


Fig. 1. The SLICE COMPRESSION test.

- From the elasticity point of view there will be singular (infinite) stresses due to the sharp corners (fiber ends) at the surface and at the crack tip at the interface as the interfacial damage occurs. The complication here is that in addition to the three materials in contact there is the interphase layer(s) of yet unknown properties. Thus, it will be very difficult to determine the nature of singularity a priori. Also, when studying the crack initiation and propagation we will need to make a decision about the type of fracture criterion to be used. Based on the experimental observations it is believed that the fracture will be of brittle type and thus the strain energy criterion will be appropriate. However, due to the singular nature of stresses (obtained from an elasticity solution), and thus a local plasticity, this will be very difficult to assess precisely. The subject of singular stresses has been addressed by England (1971), Hein and Erdogan (1971), Kurtz and Pagano (1991), Lee and Jasiuk (1991), Zak and Williams (1963), and others. The subject of fracture at bimaterial interfaces has been considered by Chou and Tetelman (1971), Jensen (1990),

and Shih (1991), among others.

- The effect of residual stresses and in particular the incorporation of the temperature dependence of properties in the calculations of the test is to be used at elevated temperature. This problem is difficult analytically due to the complex geometry involving end effects (Kaw and Goree, 1991; Kurtz and Pagano, 1991). The subject of residual stresses has been addressed by Arnold and Wilt (1992), Ghonem et al. (1994), Nimmer (1990), and Pindera et al. (1993), among others.

Due to the complexity of this problem a numerical solution using finite element method will be required. However, the elasticity solutions will contribute to the understanding of the physical behavior, in particular in the regions where the stresses are singular, and will give guidance in the numerical studies. The challenge here will be to capture the residual state of stress existing prior to loading, the extent of the plastic zone, the crack initiation/damage propagation at the interface, the microstructure information about the interphase zone, the roughness of the interface, the fiber interaction, and other factors.

This test has been studied theoretically and experimentally by Hsueh (1993a; 1994a,b), Lu and Mia (1994), and Kanagawa and Honda (1991). However, the published theoretical analyses of this test involve oversimplifying and in several instances erroneous assumptions. Thus there is a need for the more careful theoretical analysis of this complex test. Also, a number of experimental improvements of this test are currently being explored and incorporated at WPAFB.

In general, a very important issue is how to account for the fiber interaction. In a typical unidirectional fiber reinforced composite the volume fraction is about 60%. This means that the fibers are very close to each other and they interact. Both the residual stresses and those due to the mechanical loads are disturbed due to the presence of neighboring fibers. In fact the stress field will be very non-uniform. There exist approximate analytical treatments of fiber interaction, such as Mori-Tanaka method (Mori-Tanaka, 1973), which involve the single inclusion solution but the loading is modified to include the average stress in the matrix. Another approach is the composite cylinder assemblage concept of Hashin and Rosen (1964), which involves a representative volume element in the form of a composite cylinder. A similar model is a three phase model (Christensen and Lo, 1979), which involves a fiber surrounded by a concentric cylinder having the matrix properties and embedded in the effective medium of yet unknown properties. These approaches are used in the theoretical predictions of the effective properties of composite, while on the local level they provide estimates but only for the average stress or strain in the com-

posite, and not the actual fields.

Currently, in the preliminary stage of this study, the tests are done on model composites which have a dilute volume fraction of fibers i.e. the spacing is large (around 9 diameters). This is done in order to achieve a better control of parameters. Thus, the issue of fiber interaction does not enter here. However, the goal is to use this test on real composites so that a good statistical information can be obtained as a large number of fibers is considered. The interpretation of these statistics will be very difficult due to the non-uniform stresses in the composite as discussed above and due to the statistical variations in interfacial properties. In addition, in these problems the interfacial property is an important input to the numerical or analytical solution, but this is not known a priori. Therefore, an inverse problem type of approach may be needed and here the close experimental and theoretical connection is crucial.

The other important issue is the effect of surface roughness. This factor is usually not incorporated in the analyses. However, several papers on that subject have appeared recently. These include those of Kerans et al. (1994), Mackin et al. (1992), Marshall et al. (1994), Wang and Rack (1992), and others.

The simple analytical solution will be sought to the slice compression test. The approach to be taken will be along the line of papers by Gao et al. (1988), Hutchinson and Jensen (1990), Liang and Hutchinson (1993), McCartney (1989), and Steif (1984).

2. The HIDE Test

The HIDE test is similar to the SLICE COMPRESSION test but it involves the application of a hydrostatic pressure at the polished transverse surface of a unidirectional composite. The advantage of this test is that the applied loading is a uniform traction known a priori but the experimental challenge remains how to increase the magnitude of this load. The advantage of this test is that again it can be used on a real composite and thus a statistical information can be collected. This problem, although more tractable, poses similar challenges from the theoretical viewpoint as again there will be high stresses at the material discontinuity at the surface (elasticity solution may give singular behavior, which is dependent on the mismatch in elastic and thermal constants; see for example Lee and Jasiuk, 1991), local plastic deformation, crack initiation and propagation at the interface or elsewhere, and the end effects. For the above reasons this problem is again intractable analytically and a numerical solution will be required, while analytical approaches will

give guidance in the numerical efforts. The additional challenge here, and in the other two tests, is how to treat the singularity at the crack tip which is at the bimaterial interface and the sharp corners. This test has been explored by the experimentallists at WPAFB but is currently on hold due to the experimental challenge in applying the load high enough to cause the debonding.

3. The Numerical Example

To explore several issues involved in the numerical treatment of the above tests we consider the finite element solution (see Fig. 2) of the composite speciman by using the commercial finite element ANSYS package. In the numerical example to be discussed we consider a two phase composite consisting of titanium matrix and continuous SiC fibers. This is a preliminary study and since we do not have yet available all the experimental parameters we consider an idealized situation. We assume a linear elastic behavior and evaluate the residual stresses. In reality the deformation may be locally elasto-plastic. The fiber/fiber interaction is accounted for (approximately) by a composite cylinder model, and thus the problem becomes axisymmetric. Also, we assume that the fiber/matrix interface is a two dimensional surface and is perfectly bonded. In real systems an interphase region will be present. Fig. 2 gives the effective stress contours everywhere in the cylinder. This will indicate where plasticity will initiate in this idealized system. The next step is to follow this damage initiated at the surface. One can assume that the crack path follows the interface and the new location of the crack tip is now below the surface. The crack direction can be predicted theoretically by evaluating the stress intensity factors. This can be done numerically. However, this is a very complex problem due to the unknown properties of the interface. If we would consider the distributions of stress components such as the radial or the shear stresses, we would observe very high stresses at the free surface at the fiber-matrix interface. The elasticity solution would give the singular stresses at that location.

In conclusion, the finite element modelling must be fully integrated with the experimental results, in particular with the observations on the progress of damage. Also, as seen from this simple example the finite elemet method has its limits and the concurrent analytical studies will be very useful.

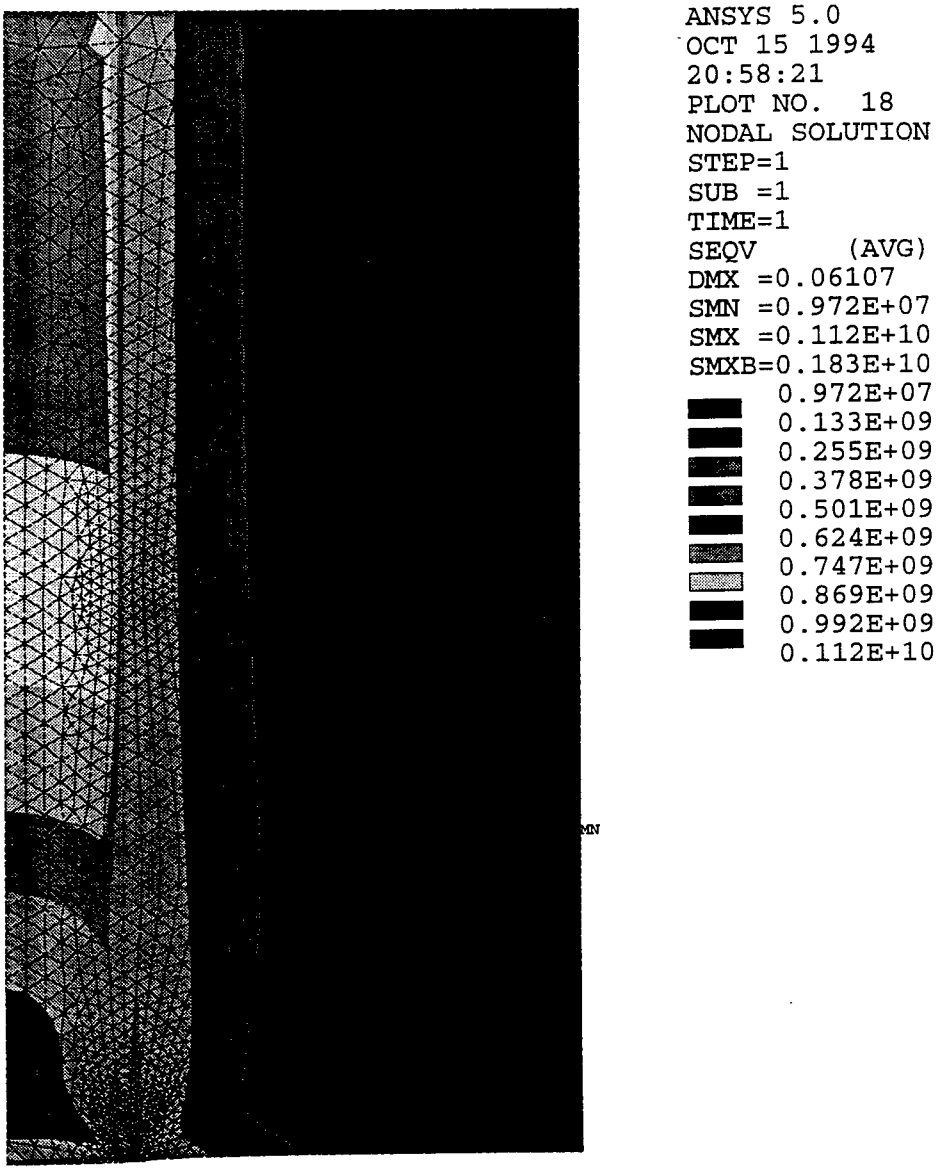


Fig. 2 Residual stresses in a composite cylinder (effective stress).

2.4 REFERENCES

Arnold, S.M. and Wilt, T.E. (1992). "Influence of engineered interfaces on residual stresses and mechanical response in metal matrix composites," *NASA Technical Memorandum 105438,* March, 1-25.

Atkinson, C., Avila, J., Betz, E., and Smelser, R.E. (1982). "The rod pull out problem, theory and experiment," *J. Mech. Phys. Solids*, **30**, 97-120.

Bartos, P. (1980). "Analysis of pull-out tests on fibres embedded in brittle matrices," *J. Mater. Sci.* **15**, 3122-3128.

Betz, E. (1982). "Experimental studies of the fibre pull-out problem," *J. Mater. Sci.* **17**, 691-700.

Bright, J.D., Shetty, D.K., Griffin, C.W., and Limaye, S.Y. (1989). "Interfacial bonding and friction in Silicon Carbide (filament)-reinforced ceramic- and glass-matrix composites," *J. Am. Ceram. Soc.* **72**, 1891-1898.

Cantonwine, P.E. and Wadley, H.N.G. (1994). "The effect of fiber-matrix reactions on the interface properties in a SCS-6/Ti-24Al-11Nb Composite," *Composites Engineering,* **4**, 67-80.

Chawla, K.K. (1987). *Composite Materials: Science and Engineering,* Springer-Verlag, New York.

Chou, T.W. and Tetelman, A.S. (1971). "Elastic cracks and screw dislocation pile-ups crossing a bimaterial interface," *Int. J. Fract. Mech.* **7**, 331-337.

Christensen, R.M. and Lo, K.H. (1979). "Solutions for effective shear properties in three phase sphere and cylinder models," *J. Mech. Phys. Solids,* **27**, 315-330.

Clyne, T.W. and Withers, P.J. (1993). *An Introduction to Metal Matrix Composites,* Cambridge University Press, Cambridge.

Cox, B.N. (1990). "Interfacial sliding near a free surface in a fibrous or layered composite during thermal cycling," *Act metall. mater.* **38**, 2411-2424.

Cox, B.N., Dakhan, M.S., James, M.R., Marshall, D.B., Morris, W.L., and Shaw, M. (1992). "On determining temperature dependent interfacial shear properties and bulk residual stresses in fibrous composites, *Act metall. mater.* **38**, pp. 2425-2433.

Curtin, W.A. (1991). "Exact theory of fibre fragmentation in a single-filament composite," *J. Materials Science,* **26**, 5239-5253.

Das, G. (1990). "A study of the reaction zone in an SiC fiber-reinforced titanium alloy composite," *Mett. Trans. A.* **21A**, 1571-1577.

Di Anselmo, A., Accorsi, M.L., and DiBenedetto, A.T. (1992). "The effect of an interphase on the stress and energy distribution in the embedded single fibre test," *Compos. Sci. Tech.* **48**, 215-225.

Doghri, I. and Leckie, F.A. (1994). "Elasto-plastic analysis of interface layers for fiber-reinforced metal-matrix composites," *Composites Sci. Tech.* **51**, 63-74.

Dollar, A. and Steif, P.S. (1988). "Load transfer in composites with coulomb friction inter-

face," *Int. J. Solids Structures* **24**, 789-803.

Drzal, L.T. and Madhukar, M. (1993). "Fibre-matrix adhesion and its relationship to composite mechanical properties," *J. Mater. Sci.* **28**, 569-610.

Eldridge, J.I. (1992). "Fiber push-out testing of intermetallic matrix composites at elevated temperatures," *Mat. Res. Symp. Proc.* **273**, preprint.

Eldridge, J.I. and Brindley, P.K. (1989). "Investigation of interfacial shear strength in a SiC fibre/Ti-24Al-11Nb composite by a fibre push-out technique," *J. Mat. Sci. Lett.* **8**, 1451-1454.

England, A.H. (1971). "On stress singularities in linear elasticity," *Int. J. Engng. Sci.* **9**, 571-585.

Feillard, P., Desarmot, G., and Favre, J.P. (1994). "Theoretical aspects of the fragmentation test," *Compos. Sci. Tech.* **50**, 265-279.

Ferber, M.K., Wereszczak, A.A., Hansen, D.H., and Homeny, J. (1993). "Evaluation of interfacial mechanical properties in SiC fiber-reinforced macro-defect-free cement composites," *Compos. Sci. Tech.* **49**, 23-32.

Folias, E.S. (1991). "On the prediction of failure at a fiber/matrix interface in a composite subjected to a transverse tensile load," *J. Composite Materials,* **25**, 869-886.

Frocht, M.M. (1948). *Photoelasticity.* Vol. II. John Willey & Sons, New York.

Gao, Y.C., Mai, Y.W. and Cotterell, B. (1988). "Fracture of fiber-reinforced materials," *ZAMP,* **39**, 550-572.

Gent, A.N. and Liu, G.L. (1991) "Pull-out and fragmentation in model fibre composites," *J. Mater. Sci.* **26**, 2467-2476.

Ghonem, H., Wen, Y., and Zheng, D. (1994). "An interactive simulation technique to determine the internal stress states in fiber reinforced metal matrix composites," *Mat. Sci. Engr.* **A177**, 125-134.

Grande, D.H., Mandell, J.F., and Hong, K.C.C. (1988). "Fibre-matrix bond strength studies of glass, ceramic and metal matrix composites," *J. Mater. Sci.* **23**, 311-328.

Hashin, Z. and Rosen, B.W. (1964). "The elastic moduli of fiber-reinforced materials," *J. Appl. Mech.*, **31**, 223-230.

Hein, V.L. and Erdogan, F. (1971). "Stress singularities in a two-material wedge," *Int. J. Fract. Mech.* **7**, 317-330.

Herrera-Franco, P.J. and Drzal, L.T. (1992). "Comparison of methods for the measurement of

fibre/matrix adhesion in composites," *Composites* **23**, 2-27.

Herrera-Franco, P.J., Rao, V, Drzal, L.T., and Chiang, M.Y.M. (1992). "Bond strength measurement in composites- analysis of experimental techniques," *Composites Engineering,* **2**, 31-45.

Highsmith, A.L., Shin, D. and Naik, R.A. (1990). "Local stresses in metal matrix composites subjected to thermal and mechanical loading," *ASTM STP 1080,* J.M. Kennedy, H.H. Moeller, and W.S. Johnson, eds. ASTM, Philadelphia, pp. 3-19.

Hsueh, C.H. (1994a). "Analyses of slice compression tests for aligned ceramic matrix composites - II. Type II boundary condition," *Acta. metall. mater.,* in press.

Hsueh, C.H. (1994b). "Slice compression tests versus fiber push-in tests," *J. Compos. Mater.* **28**, 638-655.

Hsueh, C.H. (1993a). "Analyses of slice compression tests for aligned ceramic matrix composites," *Acta metall. mater.* **41**, 3585-3593.

Hsueh, C.H. (1993b). "Interfacial debonding and fiber pull-out stresses of fiber-reinforced composites IX: A simple treatment of Poisson's effect for frictional interfaces," *Mater. Sci. Engr.* **A161**, L1-L6.

Hsueh, C.H. (1991a). "Interfacial debonding and fiber pull-out stresses of fiber-reinforced composites III: With residual radial and axial stresses," *Mater. Sci. Eng.* **A145**, 135-142.

Hsueh, C.H. (1991b). "Interfacial debonding and fiber pull-out stresses of fiber-reinforced composites IV: Sliding due to residual stresses," *Mater. Sci. Eng.* **A145**, 143-150.

Hsueh, C.H. (1990a). "Interfacial debonding and fiber pull-out stresses of fiber-reinforced composites," *Mater. Sci. Eng.* **A123**, 1-11.

Hsueh, C.H. (1990b). "Fibre pullout against push-down for fibre-reinforced composites with frictional interfaces," *J. Mater. Sci.* **25**, 811-817.

Hsueh, C.H. (1990c). "Interfacial friction analysis for fibre-reinforced composites during fibre push-down (indentation)," *J. Mater. Sci.* **25**, 818-828.

Hsueh, C.H. (1990d). "Effects of interfacial bonding on sliding phenomena during compressive loading of an embedded fibre," *J. Mater. Sci.* **25**, 4080-4086.

Hsueh, C.H., Ferber, M.K., Becher, P.F. (1989). "Stress-displacement relation of fiber for fiber-reinforced ceramic composites during (indentation) loading and unloading," J. Mater. Res. 4, 1529-1537.

Hutchinson, J.W. and Jensen, H.M. (1990). "Models of fiber debonding and pullout in brittle composites with friction," *Mech. Mater.* **9**, 139-163.

Jensen, H.M. (1990). "Mixed mode interface fracture criteria," *Acta metall. mater.* **38**, 2637-2644.

Jones, K.D. and DiBenedetto, A.T. (1994). "Fiber fracture in hybrid composite systems," *Compos. Sci. Tech.* **51**, 53-62.

Kagawa, Y. and Honda, K. (1991). "A protrusion method for measuring fiber/matrix sliding frictional stresses in ceramic matrix composites," *Ceram. Eng. Sci. Proc.* **12**, 1127-1138.

Kallas, M.N., Koss, D.A., Hahn, H.T., and Hellmann, J.R. (1992). "Interfacial stress state present in a "thin-slice" fibre push-out test," *J. Mat. Sci.* **27**, 3821-3826.

Kendall, K. (1975). "Model experiments illustrating fibre pull-out," *J. Mater. Sci.* **10**, 1011-1014.

Kerans, R.J., Jero, P.D., and Parthasarathy, T.A. (1994). "Issues in the control of fiber/matrix interfaces in ceramic composites," *Composites Science Tech.* **51**, 291-296.

Kim, J.K., Zhou, L.M., and Mai, Y.W. (1993). "Interfacial debonding and fibre pull-out stresses: Part III Interfacial properties of cement matrix composites," *J. Mater. Sci.* **28**, 3923-3930.

Koss, D.A., Hellmann, J.R., and Kallas, M.N. (1993). "Fiber pushout and interfacial shear in metal-matrix composites," *J. Metals*, March, 34-37.

Koss, D.A., Petrich, R.R., Kallas, M.N., and Hellmann, J.R. (1994). "Interfacial shear and matrix plasticity during fiber push-out in a metal-matrix composite," *Compos. Sci. Tech.* **51**, 27-33.

Kurtz, R.D. and Pagano, N.J. (1991). "Analysis of the deformation of a symmetrically-loaded fiber embedded in a matrix material," *Composites Engineering,* **1**, 13-27.

Lee, M. and Jasiuk, I. (1991). "Asymptotic expansions for the thermal stresses in bonded semi-infinite bimaterial strips," *J. of Electronic Packaging* **113**, 173-177.

Liang, C. and Hutchinson, J.W. (1993). "Mechanics of the fiber pushout test," *Mech. Mater.* **14**, 207-221.

Loretto, M.H. and Konitzer, D.G. (1990). "The effect of matrix reinforcement reaction on fracture in Ti-6Al-4V-Base Composites," *Metall. Trans. A.* **21A**, 1579-1587.

Lu, G.Y. and Mai, Y.W. (1994). "A theoretical model for the evaluation of interfacial properties of fibre-reinforced ceramics with the slice compression test," *Compos. Sci. Tech.* **51**, 565-574.

Mackin, T.J. and Zok, F.W. (1992). "Fiber bundle pushout: A technique for the measurement of interfacial sliding properties," *J. Am. Ceram. Soc.* **75**, 3169-3171.

Mackin, T.J., Warren, P.D., and Evans, A.G. (1992). "Effects of fiber roughness on interface

sliding in composites," *Acta Metal. Mater.* **40**, 1251-1257.

Majumdar, B.S. and Newaz, G.M. (1992). "Inelastic deformation of metal matrix composites: plasticity and damage mechanisms," *Phil. Mag. A.* **66**, 187-212.

Majumdar, S., Singh, D., and Singh, J.P. (1993). "Analysis of pushout tests on an SiC-fiber-reinforced reaction-bonded Si_3N_4 composite," *Composites Engineering*, **3**, 287-312.

Marotzke, C. (1994). The elastic stress field arising in the single fiber pull-out test," *Compos. Sci. Tech.* **50**, 393-405.

Marshall, D.B. (1984). "An indentation method for measuring matrix-fibre frictional stresses in ceramic composites," *J. Amer. Ceram. Soc.* **67**, C259-C260.

Marshall, D.B. (1992). "Analysis of fiber debonding and sliding experiments in brittle matrix composites," *Acta metall. mater.* **40**, 427-441.

Marshall, D.B. and Oliver, W.C. (1987). "Measurement of interfacial mechanical properties in fiber-reinforced ceramic composites, *J. Amer. Ceram. Soc.* **70**, 542-548.

Marshall, D.B., Shaw, M.C., and Morris, W.L. (1992). "Measurement of interfacial debonding and sliding resistance in fiber reinforced intermetallics," *Acta metall. mater.* **40**, 443-454.

Marshall, D.B., Shaw, M.C., and Morris, W.L. (1994a). "The determination of interfacial properties from fiber sliding experiments: the roles of misfit anisotropy and interfacial roughness," *Acta metall. mater.*, submitted.

Marshall, D.B., Morris, W.L., Cox, B.N., Graves, J., Porter, J.R., Kouris, D., and Everett, R.K. (1994b). "Transverse strengths and failure mechanisms in Ti3Al matrix composites," *Acta metall. mater.* **42**, 2657-2673.

McCartney, L.N. (1989). "New theoretical model of stress transfer between fibre and matrix in a uniaxially fibre-reinforced composite," *Proc. Roy. Soc.* **A425,** 215-244.

Mital, S.K. and Chamis, C.C. (1991). "Fiber pushout test: A three-dimensional finite element computational simulation," *J. Compos. Tech. & Res.* **13,** 14-21.

Mori, T. and Tanaka, K. (1973). "Average stress in matrix and average elastic energy of materials with misfitting inclusions." *Acta metall. mater.* **21**, 571-574.

Nedele, M.R. and Wisnom, M.R. (1994). "Three-dimensional finite element analysis of the stress concentration at a single fibre break," *Compos. Sci. Tech.* **51**, 517-524.

Nimmer, R.P. (1990). "Fiber-matrix interface effects in the presence of thermally induced residual stresses," *J. Compos. Tech. & Res.* **12**, 65-75.

Nimmer, R.P., Blankert, R.J., Russell, E.S., Smith, G.A., and Wright, P.K. (1991). "Micromechanical modeling of fiber/matrix interface effects in transversely loaded SiC/Ti-6-4 metal matrix composites," *J. Compos. Tech. & Res.* **13**, 3-13.

Parthasarathy, T.A., Jero, P.D., and Kerans, R.J. (1991). "Extraction of interface properties from a fiber push-out test," *Scripta Matell. Mater.* **25**, 2457-2462.

Parthasarathy, T.A., Marshall, D.B., and Kerans, R.J. (1994). "Analysis of the effect of interfacial roughness on fiber debonding and sliding in brittle matrix composites," *Acta metall. mater.*, in press.

Penn, L.S. and Chou, C.T. (1990). "Identification of factors affecting single filament pull-out test results," *J. Compos. Tech. & Res.* **12**, 164-171.

Pindera, M.J., Salzar, R.S., and Williams T.O. (1993). "An evaluation of a new approach for the thermoplastic response of metal-matrix composites," *Composites Engineering,* **3**, 1185-1201.

Rosenstein,A.H. (1991). "Overview of research on aerospace metallic structural materials," *Mat. Sci. Engr.* **A143**, 31-41.

Schoutens, J.E. (1989). "Metal matrix composites," in *Reference Book for Composites Technology,* Vol. 1, ed. S.M. Lee, Technomic, Lancaster, PA, pp. 175-269.

Shafry, N., Brandon, D.G., and Terasaki, M. (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, Vol. 3, Applied Science Publishers, UK, pp. 3.453-3.457.

Shetty, D.K. (1988). Shear-lag analysis of fiber push-out (indentation) tests for estimating interfacial friction stress in ceramic-matrix composites," *J. Am. Ceram. Soc.* **71**, C-107-C-109.

Shih, C.F. (1991). "Cracks on bimaterial interfaces: elasticity and plasticity aspects," *Mater. Sci. Engr.* **A143**, 77-90.

Steif, P.S. (1984). "Stiffness reduction due to fiber breakage," *J. Compos. Mater.* **17**, 153-172.

Taya, M. and Arsenault, R.J. (1989). *Metal Matrix Composites: Thermomechanical Behavior,* Pergamon Press, Oxford.

Wagner, H.D. and Eitan, A. (1993). "Stress concentration factors in two-dimensional composites: effects of material and geometrical parameters," *Compos. Sci. Tech.* **46**, 353-362.

Wang, A. and Rack, H.J. (1992). "A statistical model for sliding wear of metals in metal/composite systems," *Acta metall. mater.* **40**, 2301-2305.

Wang, S.W., Khan, A., Sands, R., and Vasudevan, A.K. (1992). "A novel nanoindentor technique for measuring fibre-matrix interfacial strengths in composites," *J. Mater. Sci. Lett.* **11**, 739-

741.

Warren, P.D., Mackin, T.J., and Evans, A.G. (1992). Design, analysis and application of an improved push-through test for the measurement of interface properties in composites. *Acta Metall. Mater.* **40**, 1243-1249.

Waterbury, M.C. and Drzal, L.T. (1991). "On the determination of fiber strengths by in-situ fiber strength testing," *J. Compos. Sci. & Tech.* **13**, 22-28.

Waterbury, M.C., Karpur, P., Matikas, T.E., Krishnamurthy, C., and Mirackle, D.B. (1994). " In situ observation of the single-fiber fragmentation process in metal-matrix composites by ultrasonic imaging," *Compos. Sci. & Tech.*, in press.

Watson, M.C. and Clyne, T.W. (1992a). "The use of single fibre pushout testing to explore interfacial mechanics in SiC monofilament-reinforced Ti - I. A photoelastic study of the test," *Acta Metall. Mater.* **40**, 131-139.

Watson, M.C. and Clyne, T.W. (1992b). "The use of single fibre pushout testing to explore interfacial mechanics in SiC monofilament-reinforced Ti - II. Application of the test to composite material," *Acta Metall. Mater.* **40**, 141-148.

Weihs, T.P. and Nix, W.D. (1991). "Experimental examination of the push-down technique for measuring the sliding resistance of silicon carbide fibers in a ceramic matrix," *J. Am. Ceram. Soc.* **74,** 524-534.

Weihs, T.P. and Nix, W.D. (1988). "Direct measurements of the frictional resistance to sliding of a fiber in a brittle matrix," *Scripta Metall.* **22**, 271-275.

Wooh, S.C. and Daniel, I.M. (1994). "Real-time ultrasonic monitoring of fiber-matrix debonding in ceramic-matrix composite," *Mech. Mater.* **17**, 379-388.

Zak, A.R. and Williams, M.L. (1963). "Crack point stress singularities at a bimaterial interface," *J. Appl. Mech*, , 142-143.

Zhou, L.M., Kim, J.K., and Mai, Y.W. (1992). "On the single fibre pull-out problem: effect of loading method," *Composites Science Tech.* **45**, 153-160.

Yang, J.M. and Jeng, S.M. (1989). "Interfacial reactions in titanium-matrix composites," *J. Metals*, Nov. 1989.

Yang, X.F. and Knowles, K.M. (1992). "The one-dimensional car parking problem and its application to the distribution of spacings between matrix cracks in unidirectional fiber-reinforced brittle materials," *J. Am. Ceram. Soc.* **75**, 141-147.

Yue, C.Y. and Cheung, W.L. (1992). "Interfacial properties of fibre-reinforced composites," *J. Mater. Sci.* **27**, 3843-3855.

REED-SOLOMON DECODING ON CHAMP ARCHITECTURE

Jack S.N. Jean
Associate Professor
Department of Computer Science and Engineering

Wright State University
Dayton, Ohio 45435

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

and

Wright Laboratory

September 1994

REED-SOLOMON DECODING ON CHAMP ARCHITECTURE

Jack S.N. Jean
Associate Professor
Department of Computer Science and Engineering
Wright State University

Abstract

The CHAMP (Configurable-Hardware Algorithm-Mappable Preprocessor) architecture was developed at Wright Laboratory for computation intensive operations across a wide variety of avionic applications. It is oriented to high speed fixed point operations and is cost effective compared to special purpose VLSI chips or general purpose supercomputers. However, it is difficult to "program" the architecture, even when a sequential code is available. Reed-Solomon (R-S) decoding is an important task in many applications. It requires high speed fixed point operations and is therefore considered a good candidate to be implemented on CHAMP architecture. This three-month summer research effort was to study the feasibility of mapping R-S decoding on CHAMP and to generalize the results. The tasks performed include (1) Converted a FORTRAN code for R-S coding to C code, (2) Designed a R-S decoder with VIEWLOGIC schematic editor that results in a hierarchy of 37 schematics, some of them down to gate level, (3) Proposed a partition scheme that requires an utilization ratio of around 70% to 80% for the R-S decoder on CHAMP, and (4) Identified two basic types of nested-loops, namely, loops with shift-invariant dependence graphs and those without, and proposed corresponding solutions.

REED-SOLOMON DECODING ON CHAMP ARCHITECTURE

Jack S.N. Jean

1 Introduction

The CHAMP (Configurable-Hardware Algorithm-Mappable Preprocessor) architecture was developed at Wright Laboratory for computation intensive operations across a wide variety of avionic applications. In the CHAMP architecture, commercial FPGA (Field Programmable Gate Array) chips are arranged into a ring and augmented with a crossbar. The architecture is more flexible than special purpose VLSI chips and is much cheaper than general purpose supercomputers with comparable computing power. A VME board implementatin of the architecture by Lockheed Sanders Inc. has been demonstrated to achieve two billion operations per second in the processing of infra-red images.

Similiar to Von Neumann machines, a CHAMP board can be programmed for various operations due to the existence of on-board FPGA chips. However, the "coding" process currently is very different from traditional computer programming and is more like performing circuit designs. For each application, a sequential code is first developed and simulated on a sequential machine, data flow is then analyzed, circuit design is performed and described in terms of either a HDL (Hardware Description Language) or a circuit schematics, and finally the circuit is partitioned, placed, and routed on multiple FPGA chips. The lengthy process is similar to, or arguably even more difficult than, programming a Von Neumann machine in machine codes. Because of this reason, it is not easy to "program" the CHAMP board, at least not before a CHAMP compiler becomes available.

The 12-week summer research effort is related to the CHAMP coding of a specific application, the Reed-Solomon (R-S) decoding. More specifically, the study focused on

the conversion of sequential C code for R-S decoding to circuit schematics. There were two objectives of the study.

1. To estimate the feasibility of applying the CHAMP board to the R-S decoding which has more sophisticated data flow than the previous application in infra-red image processing.

2. To identify techniques required to derive a circuit schematics (or HDL) from a C code for general applications.

CHAMP Overview As shown in Figure 1, CHAMP contains a "crossbar" to interconnect a ring of PEs (Processing Elements) and some global memory. Each PE consists of two FPGA chips which are run-time reprogrammable (reconfigurable) and some local memory. Several FPGA chips are put together to form the "crossbar" which can also be programmed to perform some computations. The architecture can provide the computing power that commercial microprocessors cannot achieve and is cost effective compared to special purpose VLSI chips or general purpose supercomputers.

The CHAMP board implemented by Lockheed Sanders Inc. has eight PEs, each containing two XILINX XC4013 FPGA chips and 64KB of dual-port local memory. Also contained on the board are a crossbar containing 4 XC4010 chips, a quad-XC4010 controller, a 512KB tri-port global memory, a VME interface, an RS232 serial port interface, and a video I/O interface. The board has been applied to the processing of infra-red images and achieved two billion operations per second.

Reed-Solomon Decoding R-S codes have excellent burst error-correcting capability and have been applied to many areas, including deep-space communication, teletext broadcast, frequency-hop spread spectrum systems, and optical communications. The (31, 15) R-S code is currently employed in the Joint Tactical Information Distributed

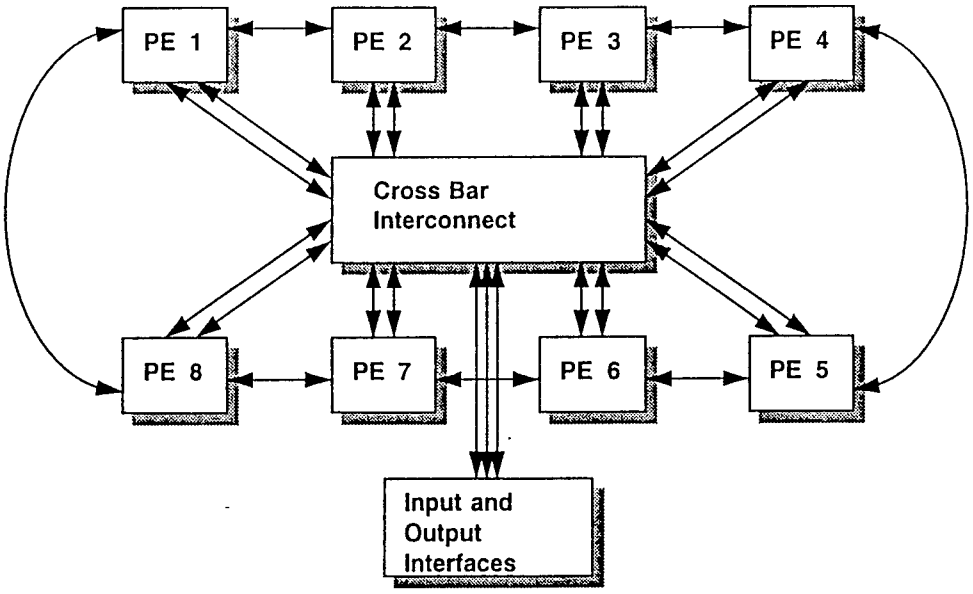


Figure 1: CHAMP architecture overview

System (JTIDS). An R-S codec is asymmetric because the encoding process that attaches error checking codes is much easier than the decoding process that recovers the information from noise-contaminated codes.

The (31, 15) R-S code is defined over a Galois finite field of size 31 and all the elements in the finite field, or symbols, can be encoded in five bits. For every 16 information symbols, or 80 information bits, 15 error checking symbols, i.e., 75 bits, are attached to the symbol string. The decoding process is to recover the 16 information symbols from 31 received (or retrieved) symbols.

There are five basic steps in the RS-decoding of a 31-symbol string. A technical report documenting those steps and a corresponding FORTRAN code was available at the beginning of the research [2]. The code was converted into C code and simulated. The C code clearly defines the decoding algorithm, allows verification of several control flow changes, and provides some testing vectors for each step. In a previous work [4], a VLSI implementation was developed for R-S coding with a more regular algorithm. That algorithm was not selected due to the second research objective, i.e., to study mapping

techniques for algorithms with more general data flow.

After the data dependency in each step was analyzed from the C code, it became clear that all the expensive computations were associated with loops, especially two-level nested loops. It was also discovered that most of the two-level nested loops could be represented in a two-dimensional dependence graph and parallelized with a projection technique.

Report Outline In Section 2, dependence graphs and a projection technique are introduced. With the projection technique, the schematics for R-S decoding is derived and summarized in Section 3. Each of those five steps involved in R-S decoding is described along with the analysis performed. A CLB (Control Logic Block) count and a partition plan are also described in the section. Section 3 concludes the report.

2 Dependence Graphs and Projection

Using dependence graphs (DGs) to describe computations involved in a nested loop and adopting projection to derive a processor array are commonly found in Systolic Array research area [1]. This section briefly introduces the techniques.

2.1 Dependence Graphs

An indexed DG can be considered as the graphical representation of a single assignment algorithm. It is a directed graph, where a node with index $\vec{i}$ represents computations of variables associated with index $\vec{i}$ in the single assignment form, and an arc from node $\vec{i}$ to $\vec{j}$ denotes a data dependency from an $\vec{i}$-indexed variable to a $\vec{j}$-indexed variable.

A DG can be constructed based on the *space-time indices in the recursive algorithm.* For example, a matrix-vector multiplication can be described in the following C code and its corresponding DG is shown in Figure 2(a).

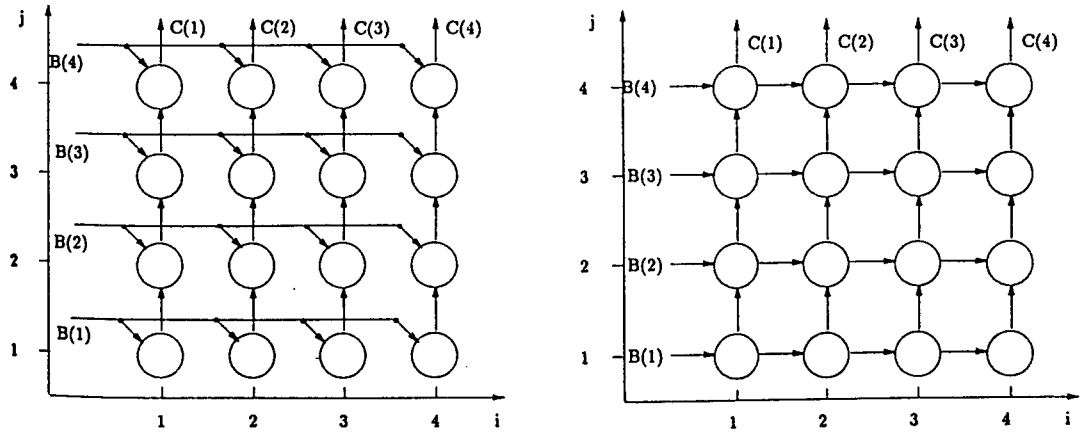


Figure 2: DG for matrix-vector multiplication: (a) with global communication and (b) with only local communication.

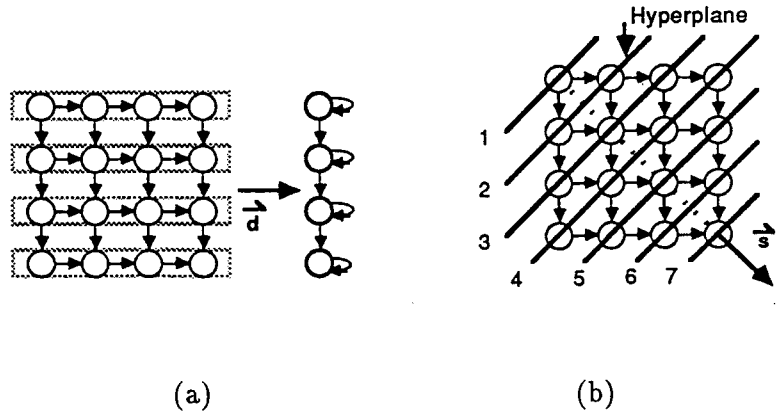


Figure 3: Illustration of (a) a linear projection with projection vector $\vec{d}$; (b) a linear schedule $\vec{s}$ and its hyperplanes.

```
for(i=0; i<4; i++) {
     C[i] = 0;
     for(j=0; j<4; j++)
       C[i] += A[i][j] * B[j];
}
```

Note that the dependency arcs in the above DG are not all local, i.e., the lengths of some arcs are proportional to the size of the matrix. In order to derive an array processor with only local interconnections, it is desirable to modify the DG such that all the arcs are local. This can be accomplished by modifying the indices of the involved variables. The localized DG for the matrix-vector multiplication is shown in Figure 2(b). Note that only the dependencies between nodes are shown in Figure 2. *The operations inside each node are not explicitly shown* in the DG.

Definition: A DG is *shift-invariant* if the dependence arcs corresponding to *all* nodes in the index space remains unchanged w.r.t. their positions. Formally, this means that if a variable at $\mathbf{i}_1$ depends on a variable at $\mathbf{i}_1 - \mathbf{j}$, then a variable at $\mathbf{i}_2$ will depend on a variable at $\mathbf{i}_2 - \mathbf{j}$ in the same manner. Note that the node functions can be different and the border I/O nodes are exempted from such a condition.

2.2 Projection and Scheduling

A straightforward implementation of a DG is to assign each node in the DG to a processor element (PE). This is not efficient since each PE only executes one computation in the algorithm. Therefore, we would like to let each PE execute multiple nodes in the DG and yet achieve maximal parallelism inherent in the DG. This requires a systematic mapping from the DG to the processor array. In mapping shift-invariant DGs onto systolic arrays, we need to specify the node assignment and the schedule for the DG, which are explained as follows.

- The node assignment specifies how nodes in a DG are assigned to the PEs in the array. A *projection* of a DG is a linear mapping of the nodes of the DG to PEs, in which nodes along a straight line are mapped to a PE. The projection direction is denoted by a vector $\vec{d}$ (see Figure 3(a)).

- The schedule specifies the execution time for all the nodes in the DG. The scheduled execution time of a node is represented by a time index. A *linear schedule*, denoted by $\vec{s}$, maps a set of parallel equi-temporal hyperplanes to a set of linearly increased time indices, where $\vec{s}$ is the normal vector of the equi-temporal hyperplanes. That is, the time index of a node can be mathematically represented by $\vec{s}^T\mathbf{i}$, where $\mathbf{i}$ denotes the index of the node.

In order to obtain a systolic design, the projection vector ($\vec{d}$) and the schedule vector ($\vec{s}$) have to satisfy two constraints.

1. $\vec{s}^T\vec{e} > 0$: Here $\vec{e}$ denotes any edge in the DG. That is, if node $\mathbf{i}$ depends on the output of node $\mathbf{j}$, then the node $\mathbf{j}$ is scheduled before the node $\mathbf{i}$. The number $\vec{s}^T\vec{e}$ denotes the number of delays on the edge in the array.

2. $\vec{s}^T\vec{d} > 0$: The projection vector $\vec{d}$ and the schedule vector $\vec{s}$ cannot be orthogonal to each other.

3 R-S Decoder Design

Once the FORTRAN code was converted into a C code, the data dependency in each step was analyzed, a circuit was designed for each step and its schematics was input using the VIEWLOGIC schematic editor on a 486PC. At the end, schematics for all the steps were put together and the number of latches between steps was calculated. Function diagrams were regrouped so to put highly interconnected blocks together. The number of CLBs (Control Logic Block) used was estimated for each small functional unit and then for

the whole schematics so to estimate the feasibility of implementing R-S decoding on the CHAMP board.

In this Section, each step is analyzed and its result summarized. Feasibility is presented after all the steps are discussed and put together.

3.1 Design Description

The algorithmic meaning of the following five steps was clearly explained in Huffman's technical report [2] and is omitted here. The major computation part of each step is described as a C code and its associated DG is analyzed for a proper projection operation.

Step1: Compute Power Sum Given an input stream of symbols, icode[i], Step1 is to evaluate a set of polynomials with the following C code to define the output power_sum[i]. Note that mul_gf(a,b) is the multiplication result of a and b on the Galois field.

```
for(i=0; i<16; i++) {
     power_sum[i] = 0;
     for(j=0; j<31; j++)
       power_sum[i] = mul_gf(power_sum[i], alpha(i+1))^ icode[j];
}
```

The DG is two dimensional with 31 nodes along the j-axis and 16 nodes along the i-axis. Since alpha(i+1) is a constant along the j-axis, a projection was taken along the j-axis so to take advantage of the constant. As a result, 16 copies of hardware were used in the projected array and all the multiplications were replaced with simpler table lookups.

Step2: Compute Elementary Symmetric Function Step2 contains two parts: S2_Front and S2_Loop. S2_front is as follows.

```
erase_sig[0] = 1;
for(i=1; i<=no_erase; i++) erase_sig[i] = 0;
for(k=0; k<no_erase; k++)  /* the first two-level nested loop */
  for(i=0; i<=k; i++) {          /* erase_loc[] is input */
      tmp = erase_sig[k+1] ^ erase_sig[i];
```

```
      erase_sig[k+1] = mul_gf(erase_loc[k], erase_sig[i]);
      erase_sig[i] = tmp;
    }

for(i=n-k-no_erase; i<=n-k; i++) modified_snd[i] = 0;
for(j=0; j<n-k-no_erase; j++)  { /* the second two-level nested loop */
     modified_snd[j] = 0;
     for(i=0; i<= no_erase; i++)
       modified_snd[j] ^= mul_gf(erase_sig[i], power_sum[j+no_erase-i]);
  }
```

It basically contains two two-level nested loops. The first one has a DG of triangular shape and a projection along the k-axis was used. Since the variable no_erase, or equivalently, the range of loop index is run-time dependent, the size of the projected array was chosen to be the largest possible one, i.e., 16. The second two-level nested loop has a rectangular DG and the projection was chosen to go along the i-axis. The variable erase_sig was propagated along the negative j-axis and the variable power_sum was chosen to go along the positive anti-diagonal (i=j line) direction.

As to the S2_Loop part, it is the most difficult part of the whole design. The following is the C code for S2_Loop.

```
m = -1;  dm = 1;  di = power_sum[0]; li = 0;  lm = 0;
for(j=1; j<=(n-k)/2; j++) { sigma[j] = 0; msigma[j] = 0; }
sigma[0] = 1; msigma[0] = 1;

for(i=0; i<n-k-no_erase; I++)  {  /* outer loop */
    di = 0;   /* starting of the first block */
    for(j=0; j<=li; j++)
     di ^= mul_gf(power_sum[i-j], sigma[j]);
    for(j=0; j<=(n-k)/2; j++) nsigma[j] = sigma[j];
    if(di == 0) continue;

    for(j=0; j<=lm; j++)   /* starting of the second block */
      sigma[j+i-m] ^= mul_gf(div_gf(di,dm), msigma[j]);
    ln = li;
    li = MAX(li, lm+i-m);

    if((i-ln) < (m-lm)) continue;
    m = I;   /* starting of the third block */
```

```
    lm = ln;
    dm = di;
    for(j=0; j<=(n-k)/2; j++) msigma[j] = nsigma[j];
  }
```

There are three reasons contributed to the difficulty. Firstly, the loop body of the outer loop has three major blocks which need to be executed in sequence due to data dependence. This data dependence creates bottleneck which governs the whole design. In fact, a double frequency clock was used for this part. Secondly, the existence of conditional branches and a run-time dependent loop range makes extra control necessary. Thirdly, the data dependence in the second block changes with the outer loop index, i.e., the i-index, and the change depends on a non-index variable whose value requires extra effort to predict. More specifically, sigma[$j + i - m$] is dependent on msigma[j] and the value of $i - m$ cannot be determined easily. A worst case senario would be to use a barrel shifter which would severely slow down the system. Fortunately, after carefully analyzing the computation, we found that the problem can be solved by shifting msigma array in a controlled way.

Step3: Determine Error Locations Step 3 computations, as illustrated in the following C code, can be represented as a 2-D DG. Since alpha(j+1) is independent of i, a projection along the i-axis is very useful to replace all the multiplications with table lookups.

```
err = 0;   /* initialization */
for(i=0; i<no_err; i++) a[i] = sigma[i+1];

for(i=n-1; i>=0; i--) {  /* outer loop */
    sum = 0;
    for(j=0; j<no_err; j++) {  /* inner loop */
        a[j] = mul_gf(a[j], alpha(j+1));
        sum ^= a[j];
    }
    if(sum == 1) err_loc[err++] = alpha(i);
}
```

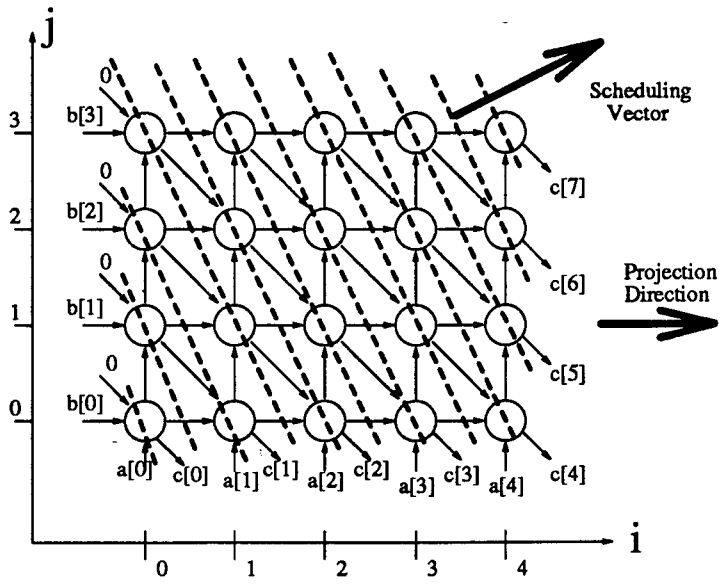


Figure 4: DG for S4_SIG and the corresponding projection direction and the scheduling vector.

Step4: Determine Error Locations Step4 contains two parts, S4_SIG and S4MAG. The S4_SIG part is to evaluate the multiplication of two polynomials, as illustrated in the following C code. Its DG is a 2-dimensional and square one as shown in Figure 4. Projection was chosen to be along the i-index. The DG is linearly scheduled so that the vector (i=2, j=1) becomes the scheduling vector that is normal to all the equi-temporal hyperplanes.

```
for(i=0; i<m+n+1; i++)  c[i] = 0;  /* initialization */

for(i=0; i<=m; i++) { /* outer loop */
    for(j=0; j<n; j++)   /* inner loop */
      c[m+j] = c[m+j+1] ^ mul_gf(b[j], a[i]);

    c[m+n] = mul_gf(b[n], a[i]);
    c[i] = c[m];
}
```

The other part, S4MAG, is simple in terms of dependence analysis. However, it took the largest number of CLBs compared to the other steps because of its complicated loop

body. A projection along the i-axis was used. That took 16 copies of hardware, each as complicated as the loop body.

```
for(j=0; j<no_correction; j++) { /* outer loop */
    num = 0; den = 0; beta = 1;
    for(i=0; i<no_correction; i++) { /* inner loop */
        num ^= mul_gf(beta, power_sum[no_correction-i-1]);
        den = mul_gf(err_loc[j], den) ^ beta;
        beta = sigma[i+1] ^ mul_gf(err_loc[j], beta);
    }
    mag_err[j] = div_gf(num, mul_gf(den, err_loc[j]));
}
```

Step5: Correct Errors Step5 is a simple loop that can be fully parallelized.

```
for(i=0;  i<no_error+no_erase; i++) {
    tmp = n-log_gf(err_loc[i])-1;
    icode[tmp] = icode[tmp] ^ mag_err[i];
}
```

3.2 Feasibility Estimation

The schematics putting all the five steps together is shown in Figure 5. To estimate the feasibility of implementing R-S decoding on the CHAMP board, the CLB count of each step was taken and a global partition plan was proposed.

CLB Counts CLB is the basic functional unit of an FPGA chip. For example, a XC4013 chip contains an array of 24 by 24 CLBs embedded in a lattice of programmable interconnection links. Inside each CLB, which is reconfigurable by itself, there are two filp-flops and a 32-bit memory storage. A CLB count roughly represents the amount of hardware resource to be used for a design. It does not take into account the routability of a design.

Since the decoder design uses XBLOX library, a parameterized library supported by XILIX, it has to be flattened and placed before a real CLB count can be performed. However, the XILINX tool, designed to handle single chip design, was not able to handle

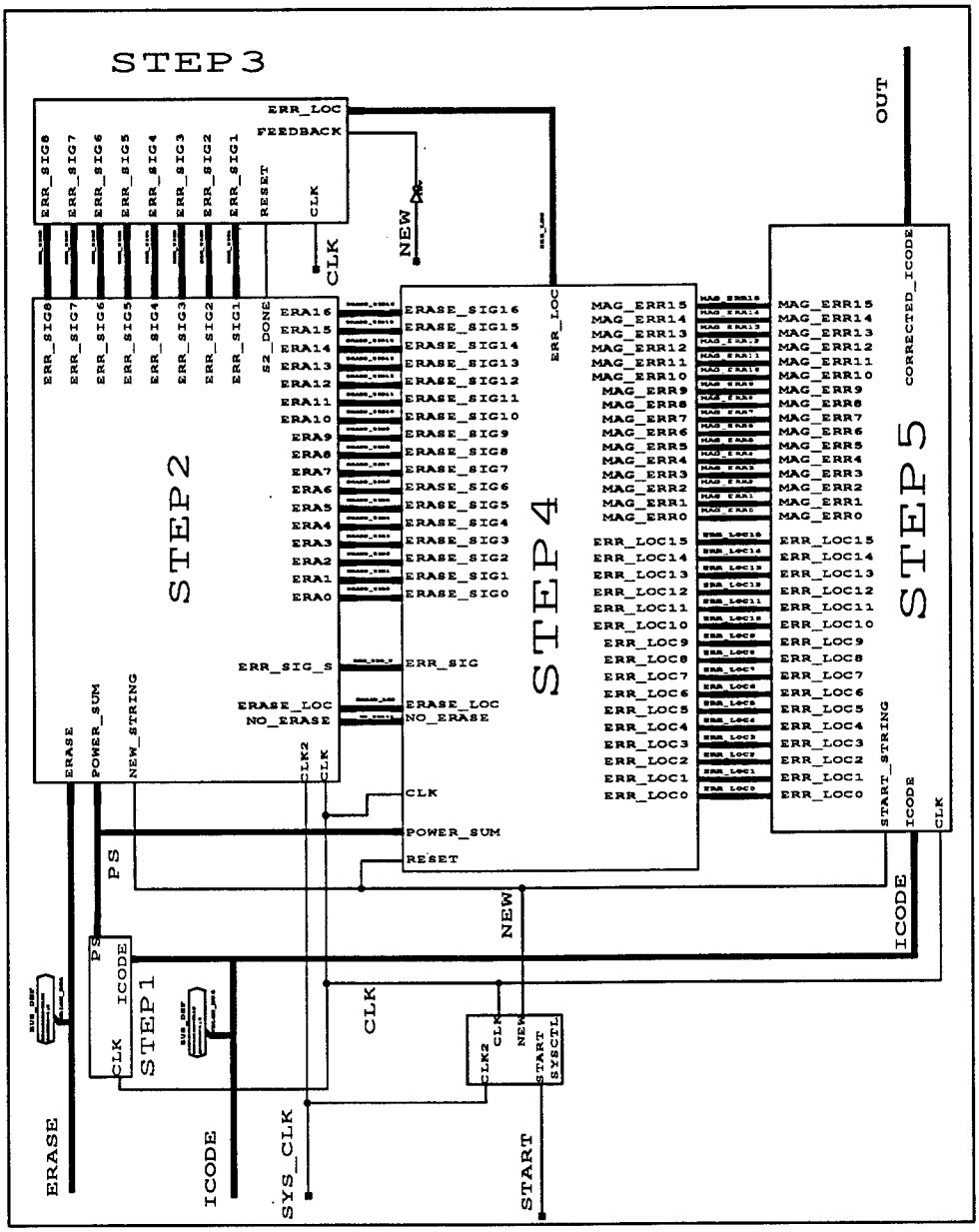


Figure 5: An overview of the Reed-Solomon decoder

the design, which would surely take more than ten XC4013s to accommodate. As a result, a CLB count was performed manually and the results can be summarized as follows.

1. The total number of CLBs counted is 5,245. This number should be considered as a lower bound on the real number of CLBs required,

   (a) Step1 takes 171 CLBs.

   (b) Step2 takes 2,183 CLBs (1,454 for S2_Front and 729 for S2_Loop).

   (c) S3 takes 92 CLBs.

   (d) S4 takes 2,714 CLBs (578 for S4_SIG, 2,086 for S4MAG, 50 for some small parts).

   (e) S5 takes 76 CLBs.

   (f) System clock circuit takes 9 CLBs.

2. Latches are necessary to synchronize the five steps. Most of the latches are used in interfacing Step2 and Step4. It is estimated that around 1,000 CLBs will be sufficient for this purpose.

3. Some parts of the design requires further modifications. Suppose that takes up to 750CLBs. Then the total number of CLBs required is around 7,000. With 16 XC4013's, the CHAMP board has 9,216 CLBs as well as the CLBs in the crossbar. So a successful implementation would have a CLB utilization from 70% to 80% which is a pretty difficult task but not impossible.

A Rough Partition Plan To partition is to group circuits into sets of functions so that two constraints are satisfied: (1) Each set can fit into an FPGA chip (CLB count, Placement, and Routing Constraint) and (2) The communcation channel between chips can sustain the inter-set channel requirement (Communication Constraint). At first glance,

the decoder is complicated and very difficult to partition. However, with the previous analysis and thorough understanding about the overall schematics, two interesting facts relating to the two constraints were observed.

1. *CLB Count, Placement, and Routing Constaint:* This constraint requires the partition of parts that require large number of CLBs. Fortunately, all the parts that have that property have a very modular structure due to the fact that they are for nested loops. This potentially will make partitioning much easier.

2. *Communication Constraint:* The communcation intensive paths are regular and local and can be grouped into three sets: the eight ERR_SIG 5-bit buses between Step2 and Step3, the 17 ERASE_SIG 5-bit buses between Step2 and Step4, and the 30 5-bit buses between Step4 and Step5. Analysis of those three sets indicated that they all can be partitioned so that those paths become on-chip communcation. This will greatly ease partitioning.

With those two observations, an intuitive way to perform partitioning at a very high level is illustrated in Figure 6. The proposed partition uses 16 XC4013's, groups functions by cutting into regular structures, and keeps communication intensive paths on-chip. The detailed partition is not clear at this point and requires a lot of extra effort to figure out.

4 Conclusion

The CHAMP architecture represents a cost effective way to tackle computation intensive applications. However, it is very difficult to "program" the architecture, even when a sequential code is available. This three-month summer research effort was to study the feasibility of mapping Reed-Solomon decoding on CHAMP and to generalize the results. The effort results in a hierarchy of 37 schematics, some of them down to gate level. The CLB count for the whole decoder was estimated and an utilization ratio of around 70% to

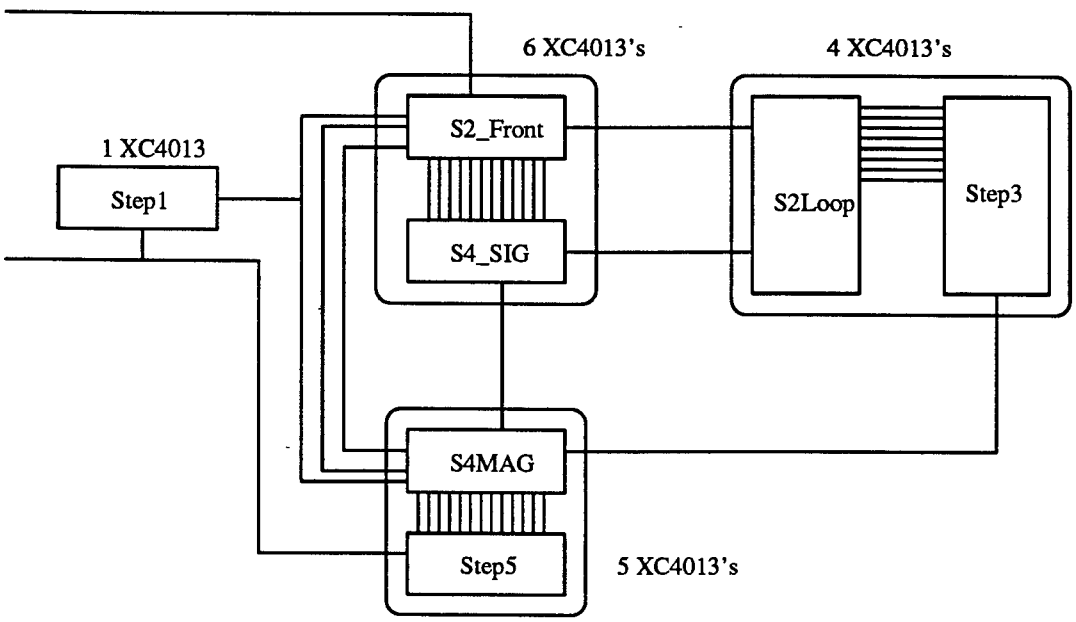


Figure 6: A rough partition plan of the Reed-Solomon decoder: All the interconnection lines are 5-bit buses. Single-bit wires are not shown in the figure.

80% was needed for the design to fit into CHAMP as a result. The high utilization ratio, fortunately, comes with a set of regular structures suitable for partitioning. Therefore the decoder implementation on CHAMP is probably feasible, if enough efforts can be spent in manual partitioning and the later 16 single-chip placement and routing.

To generalize the results, it seems reasonable to assume nested-loops exist for many applications. In R-S decoding, two types of nested-loops were encountered. One can be represented with shift-invariant DGs while the other cannot. For loops with shift-invariant DGs, projection can be used and a modular structure should be expected which is good for partitioning. For the other type of loops, projection cannot be applied and some sequential controllers are to be developed to avoid becoming system bottleneck. Overall, the problem with multiple nested-loops is an optimization problem to balance hardware resource usage and speed, since there exist many design options even for a shift-invariant DG.

Programming the CHAMP board was not easy, is not easy, and will not be easy until some advanced compilers become available. Until then, programming architectures such as CHAMP will stay as a tedious hardware design exercise.

Acknowledgement

I wish to thank Dale Van Cleave of the Avionics Directorate of Wright Laboratory for his introduction to CHAMP, his identifying the project, and his efforts in supporting the study. I also want to thank John Spillane who helped the design works in very much detail. Also appreciated are Kerry Hill, Fred Meyer, and Mark Michael of the Avionics Directorate and Brian Box of Lockheed Sanders Inc. for their cooperative efforts in providing the needed resources during this summer.

References

[1] J.S.N. Jean and S.Y. Kung, "Array Compiler Design for VLSI/WSI Systems," in TRANSFORMATIONAL APPROACHES to SYSTOLIC DESIGN, edited by G.M. Megson, Chapman & Hall, U.K., 1993.

[2] Stephen Huffman, "Reed-Solomon Codes", Technical Memorandum, Research Triangle Institute, 1983.

[3] S.Y. Kung, J.S.N. Jean, S.C. Lo, and P. S. Lewis, "Design Methodologies for Systolic Arrays: Mapping Algorithms to Architectures," in SIGNAL PROCESSING HANDBOOK, pp. 145–191, edited by C.H. Chen, Marcel Dekker Inc., New York, 1988.

[4] S.R. Whittaker, J.A. Canaris, K.B. Cameron, "Reed-Solomon VLSI codec for advanced television," IEEE Transactions on Circuits and Systems, Video Technology Vol.1 No. 2, pp.230-236, June 1991.