UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1993

SUMMER 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
Col. Hal Rhoades

Program Manager, RDL
Scott Licoscos

Program Administrator, RDL
Gwendolyn Smith

Program Administrator, RDL
Johnetta Thompson

Reproduced From
Best Available Copy

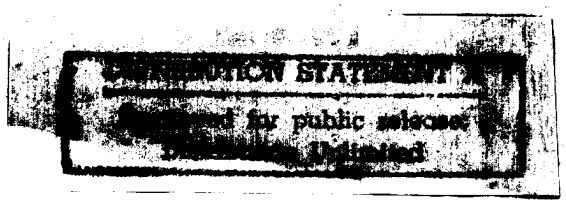
Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Bolling Air Force Base

Washington, D.C.

December 1993



19981127 070

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including th
and maintaining the data needed, and completing and reviewing the collection of information. Send comment
information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate fo
1204, Arlington, VA 22202-4302, and to the Office of management and Budget, Paperwork Reduction Project (070

AFRL-SR-BL-TR-98-

0765

ering
on of
Suite

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPOR
	December, 1993	Final

4. TITLE AND SUBTITLE
USAF Summer Research Program - 1993 Summer Faculty Research Program Final Reports, Volume 5A, Wright Laboratory

5. FUNDING NUMBERS

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: F4962-90-C-0076

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
AIR FORCE RESEARCH, AIR FORCE, ENGINEERING, LABORATORIES, REPORTS, SUMMER, UNIVERSITIES

15. NUMBER OF PAGES

16. PRICE CODE

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

Master Index for Faculty Members

Abbott, Ben
Research, MS
Box 1649 Station B
Vanderbilt University
Nashville, TN 37235-0000

Field: Electrical Engineering
Laboratory: AEDC/

Vol-Page No: 6- 1

Abrate, Serge
Assistant Professor, PhD
Mechanical & Aerospace En
University of Missouri - Rolla
Rolla, MO 65401-0249

Field: Aeronautical Engineering
Laboratory: WL/FI

Vol-Page No: 5-15

Almallahi, Hussein
Instructor, MS
P.O. Box 308
Prairie View A&M University
Prairie View, TX 77446-0000

Field: Electrical Engineering
Laboratory: AL/HR

Vol-Page No: 2-25

Anderson, James
Associate Professor, PhD
Chemistry
University of Georgia
Athens, GA 30602-2556

Field: Analytical Chemistry
Laboratory: AL/EQ

Vol-Page No: 2-18

Anderson, Richard
Professor, PhD
Physics
University of Missouri, Rolla
Rolla, MO 65401-0000

Field: Physics
Laboratory: PL/LI

Vol-Page No: 3- 7

Ashrafiuon, Hashem
Assistant Professor, PhD
Mechanical Engineering
Villanova University
Villanova, PA 19085-0000

Field: Mechanical Engineering
Laboratory: AL/CF

Vol-Page No: 2- 6

Backs, Richard
Assistant Professor, PhD
Dept. of Psychology
Wright State University
Dayton, OH 45435-0001

Field: Experimental Psychology
Laboratory: AL/CF

Vol-Page No: 2- 7

Baginski, Thomas
Assoc Professor, PhD
200 Broun Hall
Auburn University
Auburn, AL 36849-5201

Field: Electrical Engineering
Laboratory: WL/MN

Vol-Page No: 5-40

SFRP Participant Data

Baker, Suzanne
Assistant Professor, PhD
Dept. of Psychology
James Madison University
Harrisonburg, VA 22807-0000

Field:
Laboratory: AL/OE

Vol-Page No: 2-36

Baker, Albert
Assistant Professor, PhD

University of Cincinnati
, - 0

Field: Electrical Engineering
Laboratory: WL/MT

Vol-Page No: 5-53

Balakrishnan, Sivasubramanya
Associate Professor, PhD

University of Missouri, Rolla
, - 0

Field: Aerospace Engineering
Laboratory: WL/MN

Vol-Page No: 5-41

Bannister, William
Professor, PhD

Univ of Mass.-Lowell
Lowell, MA 1854-0000

Field: Organic Chemistry
Laboratory: WL/FI

Vol-Page No: 5-16

Barnard, Kenneth
Assistant Professor, PhD

Memphis State University
, - 0

Field: Electrical Engineering
Laboratory: WL/AA

Vol-Page No: 5- 1

Bayard, Jean-Pierre
Associate Professor, PhD
6000 J Street
California State Univ-Sacramen
Sacramento, CA 95819-6019

Field: Electrical/Electronic Eng
Laboratory: RL/ER

Vol-Page No: 4- 7

Beardsley, Larry
Research Professor, MS

Athens State College
, - 0

Field: Mathematics
Laboratory: WL/MN

Vol-Page No: 5-42

Beecken, Brian
Assoicate Professor, PhD
3900 Bethel Dr.
Bethel College
St. Paul, MN 55112-0000

Field: Dept. of Physics
Laboratory: PL/VT

Vol-Page No: 3-23

SFRP Participant Data

Bellem, Raymond
Dept, CHM. EE cs, PhD
3200 Willow Creek Road
Embry-Riddle Aeronautical Univ
Prescott, AZ 86301-0000

Field: Dept. of Computer Science
Laboratory: PL/VT

Vol-Page No: 3-24

Bellem, Raymond
Dept, CHM. EE cs, PhD
3200 Willow Creek Road
Embry-Riddle Aeronautical Univ
Prescott, AZ 86301-0000

Field: Dept. of Computer Science
Laboratory: /

Vol-Page No: 0- 0

Bhuyan, Jay
Assistant Professor, PhD
Dept. ofComputer Science
Tuskegee University
Tuskegee, AL 36088-0000

Field: Computer Science
Laboratory: PL/WS

Vol-Page No: 3-33

Biegl, Csaba
Assistant Professor, PhD
Box 1649 Station B
Vanderbilt University
Nashville, TN 37235-0000

Field: Electrical Engineering
Laboratory: AEDC/

Vol-Page No: 6- 2

Biggs, Albert
Professor, PhD
Electrical Engineering
Univ. of Alabama, Huntsville
Huntsville, AL 35899-0000

Field:
Laboratory: PL/WS

Vol-Page No: 3-34

Blystone, Robert
Professor, PhD
Trinity University
715 Stadium Drive
San Antonio, TX 78212-7200

Field: Dept of Biology
Laboratory: AL/OE

Vol-Page No: 2-37

Branting, Luther
Assistant Professor, PhD
PO Box 3682
University of Wyoming
Laramie, WY 82071-0000

Field: Dept of Computer Science
Laboratory: AL/HR

Vol-Page No: 2-26

Bryant, Barrett
Associate Professor, PhD
115A Campbell Hall
University of Alabama, Birming
Birmingham, AL 35294-1170

Field: Computer Science
Laboratory: RL/C3

Vol-Page No: 4- 1

Callens, Jr., Eugene
Assocition Professor, PhD
Industrial
Louisiana Technical University
Ruston, LA 71270-0000

Field: Aerospace Engineering
Laboratory: WL/MN

Vol-Page No: 5-43

Cannon, Scott
Associate Professor, PhD
Computer Science
Utah State University
Logan, UT 84322-0000

Field: Computer Science/Biophys.
Laboratory: PL/VT

Vol-Page No: 3-25

Carlisle, Gene
Professor, PhD
Dept. of Physics
West Texas State University
Canyon, TX 79016-0000

Field: Killgore Research Center
Laboratory: PL/LI

Vol-Page No: 3- 8

Catalano, George
Associate Professor, PhD
Mechanical Engineering
United States Military Academy
West Point, NY 10996-1792

Field: Department of Civil &
Laboratory: AEDC/

Vol-Page No: 6- 3

Chang, Ching
Associate Professor, PhD
Euclid Ave at E. 24th St
Cleveland State University
Cleveland, OH 44115-0000

Field: Dept. of Mathematics
Laboratory: WL/FI

Vol-Page No: 5-17

Chattopadhyay, Somnath
Assistant Professor, PhD

University of Vermont
Burlington, VT 5405-0156

Field: Mechanical Engineering
Laboratory: PL/RK

Vol-Page No: 3-14

Chen, C. L. Philip
Assistant Professor, PhD
Computer Science Engineer
Wright State University
Dayton, OH 45435-0000

Field: Electrical Engineering
Laboratory: WL/ML

Vol-Page No: 5-26

Choate, David
Assoc Professor, PhD
Dept. of Mathematics
Transylvania University
Lexington, KY 40505-0000

Field: Mathematics
Laboratory: PL/LI

Vol-Page No: 3- 9

SFRP Participant Data

Chubb, Gerald
Assistant Professor, PhD
164 W. 19th Ave.
Ohio State University
Columbus, OH 43210-0000

Field: Dept. of Aviation
Laboratory: AL/HR

Vol-Page No: 2-27

Chuong, Cheng-Jen
Associtae Professor, PhD
501 W. 1st Street
University of Texas, Arlington
Arlington, TX 76019-0000

Field: Biomedical Engineering
Laboratory: AL/CF

Vol-Page No: 2- 8

Citera, Maryalice
Assistant Professor, PhD
Department of Psychology
Wright State University
Dayton, OH 4-5435

Field: Industrial Psychology
Laboratory: AL/CF

Vol-Page No: 2- 9

Collard, Jr., Sneed
Professor, PhD
Ecology & Evolutionary Bi
University of West Florida
Pensacola, FL 32514-0000

Field: Biology
Laboratory: AL/EQ

Vol-Page No: 2-19

Collier, Geoffrey
Assistant Professor, PhD
300 College St., NE
South Carolina State College
Orangeburg,, SC 29117-0000

Field: Dept of Psychology
Laboratory: AL/CF

Vol-Page No: 2-10

Cone, Milton
Assistat Professor, PhD
3200 Willow Creek Road
Embry-Riddel Aeronautical Univ
Prescott, AZ 86301-3720

Field: Electrical Engineering
Laboratory: WL/AA

Vol-Page No: 5- 2

Cundari, Thomas
Assistant Professor, PhD
Jim Smith Building
Memphis State University
Memphis, TN 38152-0000

Field: Department of Chemistry
Laboratory: PL/RK

Vol-Page No: 3-15

D'Agostino, Alfred
Assistant Professor, PhD
4202 E Fowler Ave/SCA-240
University of South Florida
Tampa, FL 33620-5250

Field: Dept of Chemistry
Laboratory: WL/ML

Vol-Page No: 5-27

Das, Asesh
Assistant Professor, PhD
Research Center
West Virginia University
Morgantown, WV 26505-0000

Field: Concurrent Engineering
Laborat[illegible]: AL/HR

Vol-Page No: 2-28

DeLyser, Ronald
Assistant Professor, PhD
2390 S. York Street
University of Denver
Denver, CO 80208-0177

Field: Electrical Engineering
Laboratory: PL/WS

Vol-Page No: 3-35

DelVecchio, Vito
Professor, PhD
Biology
University of Scranton
Scranton, PA 185[illegible]25

Field: Biochemical Genetics
Laboratory: AL/AO

Vol-Page No: 2- 1

Dey, Pradip
Associate Professor, PhD

Hampton University
, - 0

Field: Computer Science
Laboratory: RL/IR

Vol-Page No: 4-16

Ding, Zhi
Assistant Professor, PhD
200 Broun Hall
Auburn University
Auburn, AL 36849-5201

Field: Electrical En[illegible]ering
Laboratory: WL/MN

Vol-Page No: 5-44

Doherty, John
Assistant Professor, PhD
201 Coover Hall
Iowa State University
Ames, IA 50011-1045

Field: Electrical Engineering
Laboratory: RL/OC

Vol-Page No: 4-21

Dolson, David
Assistant Professor, PhD

Wright State University
, - 0

Field: Chemistry
Laboratory: WL/PO

Vol-Page No: 5-56

Dominic, Vincent
Assustant professor, MS
300 College Park
University of Dayton
Dayton, OH 45469-0227

Field: Electro Optics Program
Laborato[illegible]WL/ML

Vol-Page No: 5-28

Donkor, Eric
Assistant Professor, PhD
Engineering
University of Connecticut
Stroes, CT 6269-1133

Field: Electrical Engineering
Laboratory: RL/OC

Vol-Page No: 4-22

Driscoll, James
Associate Professor, PhD
3004 FXB Bldg 2118
University of Michigan
Ann Arbor, MI 48109-0000

Field: Aerospace Engineering
Laboratory: WL/PO

Vol-Page No: 5-57

Duncan, Bradley
Assistant Professor, PhD
300 College Park
University of Dayton
Dayton, OH 45469-0226

Field: Electrical Engineering
Laboratory: WL/AA

Vol-Page No: 5- 3

Ehrhart, Lee
Instructor, MS
Communications & Intellig
George Mason University
Fairfax, VA 22015-1520

Field: Electrical Engineering
Laboratory: RL/C3

Vol-Page No: 4- 2

Ewert, Daniel
Assistant Professor, PhD
Electrical Engineering
North Dakota State University
Fargo, IN 58105-0000

Field: Physiology
Laboratory: AL/AO

Vol-Page No: 2- 2

Ewing, Mark
Associate Professor, PhD
2004 Learned Hall
University of Kansas
Lawrence, KS 66045-2969

Field: Engineering Mechanics
Laboratory: PL/SX

Vol-Page No: 3-22

Foo, Simon
Assistant Professor, PhD
College of Engineering
Florida State University
Tallahessee, FL 32306-0000

Field: Electrical Engineering
Laboratory: WL/MN

Vol-Page No: 5-45

Frantziskonis, George
Assistant Professor, PhD
Dept of Civil Engrng/Mech
University of Arizona
Tuson, AZ 85721-1334

Field: College of Engrng/Mines
Laboratory: WL/ML

Vol-Page No: 5-29

Frenzel III, James
Assistant Professor, PhD
Dept of Electrical Engnr
University of Idaho
Moscow, ID 83844-1023

Field: Electrical Engineering
Laboratory: WL/AA

Vol-Page No: 5- 4

Fried, Joel
Professor, PhD
Chemical Engineering
University of Cincinnati
Cincinnati, OH 45221-0171

Field: Polymer Science
Laboratory: WL/PO

Vol-Page No: 5-58

Friedman, Jeffrey
Assistant Professor, PhD
Physics
University of Puerto Rico
Mayaguez, PR 681-0000

Field: Physics/Astrophysics
Laboratory: PL/GP

Vol-Page No: 3- 1

Fuller, Daniel
Dept. Chairman, PhD
Chemistry & Physics
Nicholls State University
Thibodaux, LA 70310-0000

Field: Chemistry
Laboratory: PL/RK

Vol-Page No: 3-16

Gao, Zhanjun
Assistant Professor, PhD
203 W. Old Main, Box 5725
Clarkson University
Potsdam, NY 13699-5725

Field: Mechanical/Aeronautical E
Laboratory: WL/ML

Vol-Page No: 5-30

Gavankar, Prasad
Asst Professor, PhD
Campus Box 191
Texas A&I University
Kingsville, TX 78363-0000

Field: Mech & Indust Engineering
Laboratory: WL/MT

Vol-Page No: 5-54

Gebert, Glenn
Assistant Professor, PhD
Mechanical
Utah State University
Logan, UT 84339-0000

Field: Aerospace Engineering
Laboratory: WL/MN

Vol-Page No: 5-46

Gerdom, Larry
Professor, PhD
Natural Science
Mobile College
Mobil, AL 36663-0220

Field: Chemistry
Laboratory: AL/EQ

Vol-Page No: 2-20

Ghajar, Afshin
Professor, PhD
Mech. & Aerospace Enginee
Oklahoma State University
Stillwater, OK 74078-0533

Field: Mechanical Engineering
Laboratory: WL/PO

Vol-Page No: 5-59

Gopalan, Kaliappan
Associate Professor, PhD
Dept of Engineering
Purdue University, Calumet
Hammond, IN 46323-0000

Field:
Laboratory: AL/CF

Vol-Page No: 2-11

Gould, Richard
Assistant Professor, PhD
Mechanical & Aerospace En
N.Carolina State University
Raleigh, NC 27695-7910

Field: Mechanical Engineering
Laboratory: WL/PO

Vol-Page No: 5-60

Gowda, Raghava
Assistant Professor, PhD
Dept of Computer Science
University of Dayton
Dayton, OH 45469-2160

Field: Computer Information Sys.
Laboratory: WL/AA

Vol-Page No: 5- 5

Graetz, Kenneth
Assistant Professor, PhD
300 College Park
University of Dayton
Dayton, OH 45469-1430

Field: Department of Psychology
Laboratory: AL/HR

Vol-Page No: 2-29

Gray, Donald
Associate Professor, PhD
PO Box 6101
West Virginia Unicersity
Morgantown, WV 20506-6101

Field: Dept of Civil Engineering
Laboratory: AL/EQ

Vol-Page No: 2-21

Green, Bobby
Assistant Professor, MS
Box 43107
Texas Tech University
Lubbock, TX 79409-3107

Field: Electrical Engineering
Laboratory: WL/FI

Vol-Page No: 5-18

Grubbs, Elmer
Assistant Professor, MS
Engineering
New Mexico Highland University
Las Vegas, NM 87701-0000

Field: Electrical Engineering
Laboratory: WL/AA

Vol-Page No: 5- 6

SFRP Participant Data

Guest, Joyce
Assoiate, PhD
Department of Chemistry
University of Cincinnati
Cincinnati, OH 45221-0172

Field: Physical ChemIstry
Laboratory: WL/ML

Vol-Page No: 5-31

Gumbs, Godfrey
Professor, PhD
Physics & Astronomy
University New York Hunters Co
New York, NY 10021-0000

Field: Condensed Matter Physics
Laboratory: WL/EL

Vol-Page No: 5-12

Hakkinen, Raimo
Professor, PhD
207 Jolley Hall
Washington University
St. Louis, MO 63130-0000

Field: Mechanical Engineering
Laboratory: WL/FI

Vol-Page No: 5-19

Hall, Jr., Charles
Assistant Professor, PhD
Mech & Aerospace Engr.
North Carolina Univ.
Raleigh, NC 27695-7910

Field:
Laboratory: WL/FI

Vol-Page No: 5-20

Hancock, Thomas
Assistant Professor, PhD

Grand Canyon University
, - 0

Field: Educational Psychology
Laboratory: AL/HR

Vol-Page No: 2-30

Hannafin, Michael
Visiting Professor, PhD
305-D Stone Building,3030
Florida State University
Tallahassee, FL 3-2306

Field: Educational Technology
Laboratory: AL/HR

Vol-Page No: 2-31

Helbig, Herbert
Professor, PhD
Physics
Clarkson University
Potsdam, NY 13699-0000

Field: Physics
Laboratory: RL/ER

Vol-Page No: 4- 8

Henry, Robert
Professor, PhD
Electrical Engineering
University of Southwestern Lou
Lafayette, LA 70504-3890

Field: Electrical Engineering
Laboratory: RL/C3

Vol-Page No: 4- 3

Hong, Lang
Assistant Professor, PhD
Dept of Electrical Engin
Wright State University
Dayton, OH 45435-0000

Field: Electrical Engineering
Laboratory: WL/AA

Vol-Page No: 5- 7

Hsu, Lifang
Assistant Professor, PhD

Field: Mathematical Statistics
Laboratory: RL/ER

Le Moyne College
, - 0

Vol-Page No: 4- 9

Huang, Ming
Assistant Professor, PhD
500 NW 20th Street
Florida Atlantic University
Boca Raton, FL 33431-0991

Field: Mechanical Engineering
Laboratory: AL/CF

Vol-Page No: 2-12

Humi, Mayer
Professor, PhD
Mathematics
Worchester Polytechnic Institu
Worchester, MA 1609-2280

Field: Applied Mathematics
Laboratory: PL/GP

Vol-Page No: 3- 2

Humi, Mayer
Professor, PhD
Mathematics
Worchester Polytechnic Institu
Worchester, MA 1609-2280

Field: Applied Mathematics
Laboratory: /

Vol-Page No: 0- 0

Jabbour, Kamal
Associate Professor, PhD
121 Link hall
Syracuse University
Syracuse, NY 13244-1240

Field: Electrical Engineering
Laboratory: RL/C3

Vol-Page No: 4- 4

Jaszczak, John
Assistant Professor, PhD
Dept. of Physics
Michigan Technological Univers
Houghton, MI 49931-1295

Field:
Laboratory: WL/ML

Vol-Page No: 5-32

Jeng, San-Mou
Associte, PhD
Mail Location #70
University of Cincinnati
Cincinnati, OH 45221-0070

Field: Aerospace Engineering
Laboratory: PL/RK

Vol-Page No: 3-17

Johnson, David
Associate Professor, PhD
Dept of Chemistry
University of Dayton
Dayton, OH 45469-2357

Field: Chemistry
Laboratory: WL/ML

Vol-Page No: 5-33

Karimi, Amir
Associate, PhD
Division Engineering
University of Texas, San Anton
San Antonio, TX 7824-9065

Field: Mechanical Engineering
Laboratory: PL/VT

Vol-Page No: 3-26

Kheyfets, Arkady
Assistant Professor, PhD
Dept. of Mathematics
North Carolina State Univ.
Raleigh, NC 27695-7003

Field:
Laboratory: PL/VT

Vol-Page No: 3-27

Koblasz, Arthur
Associate, PhD
Civil Engineeing
Georgia State University
Atlanta, GA 30332-0000

Field: Engineering Science
Laboratory: AL/AO

Vol-Page No: 2- 3

Kraft, Donald
Professor, PhD
Dept. of Computer Science
Louisiana State University
Baton Rouge, LA 70803-4020

Field:
Laboratory: AL/CF

Vol-Page No: 2-13

Kumar, Rajendra
Professor, PhD
1250 Bellflower Blvd
California State University
Long Beach, CA 90840-0000

Field: Electrical Engineering
Laboratory: RL/C3

Vol-Page No: 4- 5

Kumta, Prashant
Assistant Professor, PhD
Dept of Materials Science
Carnegie-Mellon University
Pittsburgh, PA 15213-3890

Field: Materiels Science
Laboratory: WL/ML

Vol-Page No: 5-34

Kuo, Spencer
Professor, PhD
Route 110
Polytechnic University
Farmingdale, NY 11735-0000

Field: Electrophysics
Laboratory: PL/GP

Vol-Page No: 3- 3

Lakeou, Samuel
Professor, PhD
Electrical Engineering
University of the District of
Washnington, DC 20008-0000

Field: Electrical Engineering
Laboratory: PL/VT

Vol-Page No: 3-28

Langhoff, Peter
Professor, PhD

Indiana University
Bloomington, IN 47405-4001

Field: Dept. of Chemistry
Laboratory: PL/RK

Vol-Page No: 3-18

Lawless, Brother
Assoc Professor, PhD
Dept. Science /Mathematic
Fordham University
New York, NY 10021-0000

Field: Box 280
Laboratory: AL/OE

Vol-Page No: 2-38

Lee, Tzesan
Associate Professor, PhD
Dept. of Mathematics
Western Illinois University
Macomb, IL 61455-0000

Field:
Laboratory: AL/OE

Vol-Page No: 2-39

Lee, Min-Chang
Professor, PhD
167 Albany Street
Massachusetts Institute
Cambridge, MA 2139-0000

Field: Plasma Fusion Center
Laboratory: PL/GP

Vol-Page No: 3- 4

Lee, Byung-Lip
Associate Professor, PhD
Engineering Sci. & Mechan
Pennsylvania State University
University Park, PA 16802-0000

Field: Materials Engineering
Laboratory: WL/ML

Vol-Page No: 5-35

Leigh, Wallace
Assistant Professor, PhD
26 N. Main St.
Alfred University
Alfred, NY 14802-0000

Field: Electrical Engineering
Laboratory: RL/ER

Vol-Page No: 4-10

Levin, Rick
Research Engineer II, MS
EM Effects Laboratory
Georgia Institute of Technolog
Atlanta, GA 30332-0800

Field: Electrical Engineering
Laboratory: RL/ER

Vol-Page No: 4-11

Li, Jian
Asst Professor, PhD
216 Larsen Hall
University of Florida
Gainesville, FL 32611-2044

Field: Electrical Engineering
Laboratory: WL/AA

Vol-Page No: 5- 8

Lilienfield, Lawrence
Professor, PhD
3900 Reservoir Rd., NW
Georgetown University
Washington, DC 20007-0000

Field: Physiology & Biophysics
Laboratory: WHMC/

Vol-Page No: 6-14

Lim, Tae
Assistant Professor, PhD
2004 Learned Hall
University of Kansas
Lawrence, KA 66045-0000

Field: Mechanical/Aerospace Engr
Laboratory: FJSRL/

Vol-Page No: 6- 8

Lin, Paul
Associate Professor, PhD
Mechanical Engineering
Cleveland State University
Cleveland, OH 4-4115

Field: Associate Professor
Laboratory: WL/FI

Vol-Page No: 5-21

Liou, Juin
Associate Professor, PhD
Electrical & Computer Eng
University of Central Florida
Orlando, FL 32816-2450

Field: Electrical Engineering
Laboratory: WL/EL

Vol-Page No: 5-13

Liu, David
Assistant Professor, PhD
100 Institute Rd.
Worcester Polytechnic Inst.
Worcester, MA 1609-0000

Field: Department of Physics
Laboratory: RL/ER

Vol-Page No: 4-12

Losiewicz, Beth
Assistant Professor, PhD
Experimental Psychology
Colorado State University
Fort Collins, CO 80523-0000

Field: Psycholinguistics
Laboratory: RL/IR

Vol-Page No: 4-17

Loth, Eric
Assistant Professor, PhD
104 S. Wright St, 321C
University of Illinois-Urbana
Urbana, IL 61801-0000

Field: Aeronaut/Astronaut Engr
Laboratory: AEDC/

Vol-Page No: 6- 4

Lu, Christopher
Associate Professor, PhD
300 College Park
University of Dayton
Dayton, OH 45469-0246

Field: Dept Chemical Engineering
Laboratory: WL/PO

Vol-Page No: 5-61

Manoranjan, Valipuram
Associate Professor, PhD
Neill Hall
Washington State University
Pullman, WA 99164-3113

Field: Pure & AppliedMathematics
Laboratory: AL/EQ

Vol-Page No: 2-22

Marsh, James
Professor, PhD
Physics
University of West Florida
Pensacola, FL 32514-0000

Field: Physics
Laboratory: WL/MN

Vol-Page No: 5-47

Massopust, Peter
Assistant Professor, PhD

Sam Houston State University
Huntsville, TX 77341-0000

Field: Dept. of Mathematics
Laboratory: AEDC/

Vol-Page No: 6- 5

Miller, Arnold
Senior Instructor, PhD
Chemistry & Geochemistry
Colorado School of Mines
Golden, CO 80401-0000

Field:
Laboratory: FJSRL/

Vol-Page No: 6- 9

Misra, Pradeep
Associate Professor, PhD

University of St. Thomas
, - 0

Field: Electrical Engineering
Laboratory: WL/AA

Vol-Page No: 5- 9

Monsay, Evelyn
Associate Professor, PhD
1419 Salt Springs Rd
Le Moyne College
Syracuse, NY 13214-1399

Field: Physics
Laboratory: RL/OC

Vol-Page No: 4-23

Morris, Augustus
Assistant Professor, PhD

Central State University
, - 0

Field: Biomedical Science
Laboratory: AL/CF

Vol-Page No: 2-14

Mueller, Charles
Professor, PhD
W140 Seashore Hall
University of Iowa
Iowa City, IA 52242-0000

Field: Dept of Sociology
Laboratory: AL/HR

Vol-Page No: 2-32

Murty, Vedula
Associate Professor, MS

Texas Southern University
, - 0

Field: Physics
Laboratory: PL/VT

Vol-Page No: 3-29

Musavi, Mohamad
Assoc Professor, PhD
5708 Barrows Hall
University of Maine
Orono, ME 4469-5708

Field: Elect/Comp. Engineering
Laboratory: RL/IR

Vol-Page No: 4-18

Naishadham, Krishna
Assistant Professor, PhD
Dept. of Electrical Eng.
Wright State University
Dayton, OH 45435-0000

Field: Electrical Engineering
Laboratory: WL/EL

Vol-Page No: 5-14

Noel, Charles
Associate Professor, PhD
151A Campbell Hall
Ohio State University
Columbus, OH 43210-1295

Field: Dept of Textiles & Cloth
Laboratory: PL/RK

Vol-Page No: 3-19

Norton, Grant
Asst Professor, PhD
Mechanical & Materials En
Washington State University
Pullman, WA 99164-2920

Field: Materials Science
Laboratory: WL/ML

Vol-Page No: 5-36

Noyes, James
Professor, PhD
Mathematics & Computer Sc
Wittenberg University
Springfield, OH 45501-0720

Field: Computer Science
Laboratory: WL/FI

Vol-Page No: 5-22

Nurre, Joseph
Assistant Professor, PhD
Elec. & Computer Engineer
Ohio University
Athens, OH 45701-0000

Field: Mechanical Engineering
Laboratory: AL/CF

Vol-Page No: 2-15

Nygren, Thomas
Associate Professor, PhD
1885 Neil Ave. Mail
Ohio State University
Columbus, OH 43210-1222

Field: Department of Psychology
Laboratory: AL/CF

Vol-Page No: 2-16

Osterberg, Ulf
Assistant Professor, PhD
Thayer School of Engrg.
Dartmouth College
Hanover, NH 3755-0000

Field:
Laboratory: FJSRL/

Vol-Page No: 6-10

Pan, Ching-Yan
Associate Professor, PhD
Physics
Utah State University
Logan, UT 84322-4415

Field: Condensed Matter Physics
Laboratory: PL/WS

Vol-Page No: 3-36

Pandey, Ravindra
Assistant Professor, PhD
1400 Townsend Dr
Michigan Technological Univers
Houghton, MI 49931-1295

Field: Physics
Laboratory: FJSRL/

Vol-Page No: 6-11

Patton, Richard
Assistant Professor, PhD
Mechanical&Nuclear Engine
Mississippi State University
Mississippi State, MS 39762-0000

Field: Mechanical Engineering
Laboratory: PL/VT

Vol-Page No: 3-30

Peretti, Steven
Assistant Professor, PhD
Chemical Engineering
North Carolina State Univ.
Raleigh, NC 27695-7905

Field:
Laboratory: AL/EQ

Vol-Page No: 2-23

Petschek, Rolfe
Associate Professor, PhD
Department of Physics
Case Western Reserve Universit
Cleveland, OH 44106-7970

Field: Physics
Laboratory: WL/ML

Vol-Page No: 5-37

Pezeshki, Charles
Assistant Professor, PhD

Washington State University
Pullman, WA 99164-2920

Field: Mechanical Engineering
Laboratory: FJSRL/

Vol-Page No: 6-12

Piepmeier, Edward
Assistant Professor, PhD
College of Pharmacy
University of South Carolina
Columbia, SC 29208-0000

Field:
Laboratory: AL/AO

Vol-Page No: 2- 4

Pittarelli, Michael
Associate Professor, PhD
PO Box 3050, Marcy Campus
SUNY, Institute of Technology
Utica, NY 13504-3050

Field: Information Sys & Engr.
Laboratory: RL/C3

Vol-Page No: 4- 6

Potasek, Mary
Research Professor, PhD

Columbia University
, - 0

Field: Physics
Laboratory: WL/ML

Vol-Page No: 5-38

Prasad, Vishwanath
Professor, PhD

SUNY, Stony Brook
Stony Brook, NY 11794-2300

Field: Mechanical Engineering
Laboratory: RL/ER

Vol-Page No: 4-13

Priestley, Keith
Research Scientist, PhD

University of Nevada, Reno
, - 0

Field: Geophysics
Laboratory: PL/GP

Vol-Page No: 3- 5

Purasinghe, Rupasiri
Professor, PhD
5151 State Univ. Dr.
California State Univ.-LA
Los Angeles, CA 90032-0000

Field: Dept of Civil Engineering
Laboratory: PL/RK

Vol-Page No: 3-20

Raghu, Surya
Assistant Professor, PhD
Mechanical Engineering
SUNY, Stony Brook
Stony Brook, NY 11794-2300

Field: Mechanical Engineering
Laboratory: WL/PO

Vol-Page No: 5-62

Ramesh, Ramaswamy
Associate Professor, PhD
School of Management
SUNY, Buffalo
Buffalo, NY 14260-0000

Field: Magement Science/Systems
Laboratory: AL/HR

Vol-Page No: 2-33

Ramm, Alexander
Professor, PhD
Mathematics
Kansas State University
Manhattan, KS 66506-2602

Field:
Laboratory: AL/CF

Vol-Page No: 2-17

Ray, Paul
Assistant Professor, PhD
Box 870288
University of Alabama
Tuscaloosa, AL 35487-0288

Field: Industrial Engineering
Laboratory: AL/OE

Vol-Page No: 2-40

Reimann, Michael
Assistant Instructor, MS
Information Systems
The University of Texas-Arling
Arlington, TX 76019-0437

Field: Computer Science
Laboratory: WL/MT

Vol-Page No: 5-55

Rodriguez, Armando
Assistant Professor, PhD

Arizona State University
Tempe, AZ 85287-7606

Field: Electrical Engineering
Laboratory: WL/MN

Vol-Page No: 5-48

Rohrbaugh, John
Research Engineer, PhD
347 Ferst St
Georgia Institute of Technolog
Atlanta, GA 30332-0800

Field: Sensors & Applied Electro
Laboratory: RL/ER

Vol-Page No: 4-14

Roppel, Thaddeus
Associate Professor, PhD
200 Broun Hall
Auburn University
Auburn, AL 36849-5201

Field: Electrical Engineering
Laboratory: WL/MN

Vol-Page No: 5-49

Rosenthal, Paul
Professor, PhD
Mathematics
Los Angeles City College
Los Angeles, CA 90027-0000

Field: Mathematics
Laboratory: PL/RK

Vol-Page No: 3-21

Rotz, Christopher
Associate Professor, PhD

Brigham Young University
Provo, UT 84602-0000

Field: Mechanical Engineering
Laboratory: PL/VT

Vol-Page No: 3-31

Rudolph, Wolfgang
Associate Professor, PhD
Dept of Physics and Astro
University of New Mexico
Albuquerque, NM 84131-0000

Field: Physics
Laboratory: PL/LI

Vol-Page No: 3- 0

Rudzinski, Walter
Professor, PhD
Dept. of Chemistry
Southwest Texas State Universi
San Marcos, TX 78610-0000

Field: Professor
Laboratory: AL/OE

Vol-Page No: 2-41

Rule, William
Asst Professor, PhD
Mechanical Engineering
University of Alabama
Tuscaloosa, AL 35487-0278

Field: Engineering Mechanics
Laboratory: WL/MN

Vol-Page No: 5-50

Ryan, Patricia
Research Associate, MS
Georgia Tech Research Ins
Georgia Institute of Tech
Atlanta, GA 30332-0000

Field: Electrical Engineering
Laboratory: WL/AA

Vol-Page No: 5-10

Saiduddin, Syed
Professor, PhD
1900 Coffey Rd
Ohio State University
Columbus, OH 43210-1092

Field: Physiology/Pharmacology
Laboratory: AL/OE

Vol-Page No: 2-42

Schonberg, William
Assoc Professor, PhD
Engineering Dept.
University of Alabama, Huntsvi
Huntsville, AL 35899-0000

Field: Civil & Environmental
Laboratory: WL/MN

Vol-Page No: 5-51

Schulz, Timothy
Assistant Professor, PhD
1400 Townsend Dr
Michigan Technological Univers
Houghton, MI 49931-1295

Field: Electrical Engineering
Laboratory: PL/LI

Vol-Page No: 3-11

Shen, Mo-How
Assistant Professor, PhD
2036 Neil Ave.
Ohio State University
Columbus,, OH 43210-1276

Field: Aerospace Engineering
Laboratory: WL/FI

Vol-Page No: 5-23

Sherman, Larry
Professor, PhD
Dept. of Chemistry
University of Scranton
Scranton, PA 18510-4626

Field: Analytical Chemistry
Laboratory: AL/OE

Vol-Page No: 2-43

Shively, Jon
Professor, PhD
Civil & Industrial Eng.
California State University, N
Northridge, CA 91330-0000

Field: Metallurgy
Laboratory: PL/VT

Vol-Page No: 3-32

Snapp, Robert
Assistant Professor, PhD
Dept of Computer Science
University of Vermont
Burlington, VT 5405-0000

Field: Physics
Laboratory: RL/IR

Vol-Page No: 4-19

Soumekh, Mehrdad
Associate Professor, PhD
201 Bell Hall
SUNY, Buffalo
Amherst, NY 14260-0000

Field: Elec/Computer Engineering
Laboratory: PL/LI

Vol-Page No: 3-12

Spetka, Scott
Assistant Professor, PhD
PO Box 3050, Marcy Campus
SUNY, Institute of Technology
Utica, NY 13504-3050

Field: Information Sys & Engrg
Laboratory: RL/XP

Vol-Page No: 4-26

Springer, John
Associate Professor, PhD

Fisk University
, - 0

Field: Physics
Laboratory: AEDC/

Vol-Page No: 6- 6

Stevenson, Robert
Assistant Professor, PhD
Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556-0000

Field: Electrical Engineering
Laboratory: RL/IR

Vol-Page No: 4-20

Stone, Alexander
Professor, PhD
Mathematics & Statistics
University of New Mexico
Alburquerque, NM 87131-1141

Field:
Laboratory: PL/WS

Vol-Page No: 3-37

Sveum, Myron
Assistant Professor, MS
Electronic Engineering Te
Metropolitan State College
Denver, CO 80217-3362

Field: Electrical Engineering
Laboratory: RL/OC

Vol-Page No: 4-24

Swanson, Paul
Research Associate, PhD
Electrical Engineering
Cornell University
Ithaca, NY 14853-0000

Field: Electrical Engineering
Laboratory: RL/OC

Vol-Page No: 4-25

Swope, Richard
Professor, PhD
Engineering Science
Trinity University
San Antonio, TX 78212-0000

Field: Mechanical Engineering
Laboratory: AL/AO

Vol-Page No: 2- 5

Tan, Arjun
Professor, PhD
Physics
Alabama A&M University
Normal, AL 35762-0000

Field: Physics
Laboratory: PL/WS

Vol-Page No: 3-38

Tarvin, John
Associate Professor, PhD
800 Lakeshore Drive
Samford University
Birmingham, AL 35229-0000

Field: Department of Physics
Laboratory: AEDC/

Vol-Page No: 6- 7

Taylor, Barney
Visiting Assist Professor, PhD
1601 Peck Rd.
Miami Univ. - Hamilton
Hamilton, OH 4-5011

Field: Dept. of Physics
Laboratory: WL/ML

Vol-Page No: 5-39

Thio, Y.
Associate Professor, PhD

University of Miami
Coral Gables, FL 33124-0530

Field: Physics Dept.
Laboratory: PL/WS

Vol-Page No: 3-39

Tong, Carol
Assistant Professor, PhD
Electrical Engineering
Colorado State University
Fort Collins, CO 80523-0000

Field:
Laboratory: WL/AA

Vol-Page No: 5-11

Truhon, Stephen
Associate Professor, PhD
Social Sciences
Winston-Salem State University
Winston-Salem, NC 27110-0000

Field: Psychology
Laboratory: AL/HR

Vol-Page No: 2-34

Tzou, Horn-Sen
Associate Professor, PhD
Mechanical Engineering
University of Kentucky
Lexington, KY 40506-0046

Field: Mechanical Engineering
Laboratory: WL/FI

Vol-Page No: 5-24

Vogt, Brian
Professor, PhD

Bob Jones University
, - 0

Field: Pharmaceutical Sciences
Laboratory: AL/EQ

Vol-Page No: 2-24

Wang, Xingwu
Asst Professor, PhD
Dept. of Electrical Eng.
Alfred University
Alfred, NY 14802-0000

Field: Physics
Laboratory: WL/FI

Vol-Page No: 5-25

Whitefield, Philip
Research Assoc Professor, PhD
Cloud & Aerosol Sciences
University of Missouri-Rolla
Rolla, MO 65401-0000

Field: Chemistry
Laboratory: PL/LI

Vol-Page No: 3-13

Willson, Robert
Research Assoc Professor, PhD
Robinson Hall
Tufts University
Medford, MA 2155-0000

Field: Physics and Astronomy
Laboratory: PL/GP

Vol-Page No: 3- 6

Witanachchi, Sarath
Assistant Professor, PhD
4202 East Fowler Avenue
University of South Florida
Tampa, FL 33620-7900

Field: Department of Physics
Laboratory: FJSRL/

Vol-Page No: 6-13

Woehr, David
Assistant Professor, PhD
Psycholoogy
Texas A&M University
College Station, TX 77845-0000

Field: Psychology
Laboratory: AL/HR

Vol-Page No: 2-35

Xu, Longya
Assistant Professor, PhD
Electrical Engineering
The Ohio State University
Columbus, OH 43210-0000

Field: Electrical Engineering
Laboratory: WL/PO

Vol-Page No: 5-63

Yavuzkurt, Savas
Associate Professor, PhD

Pennsylvania State University
University Park, PA 16802-0000

Field: Mechanical Engineering
Laboratory: WL/PO

Vol-Page No: 5-64

Zhang, Xi-Cheng
Associate Professor, PhD
Physics Department
Rensselaer Polytechnic Institu
Troy, NY 12180-3590

Field: Physics
Laboratory: RL/ER

Vol-Page No: 4-15

Zhou, Kemin
Assistant Professor, PhD
Dept. of Elec & Comp. Eng
Louisiana State University
Baton Rouge, LA 70803-0000

Field:
Laboratory: WL/MN

Vol-Page No: 5-52

Zimmermann, Wayne
Associate Professor, PhD
P.O. Box 22865
Texas Woman's University
Denton, TX 76205-0865

Field: Dept Mathematics/Computer
Laboratory: PL/WS

Vol-Page No: 3-40

Nonmechanical microscanning using optical space-fed phased arrays

Kenneth J. Barnard
Assistant Professor
Department of Electrical Engineering

Memphis State University
Memphis, TN 38152

Final Report for:
Summer Faculty Research Program
Wright Laboratory
WL/AARI

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

August 1993

Nonmechanical microscanning using optical space-fed phased arrays

Kenneth J. Barnard
Assistant Professor
Department of Electrical Engineering

Memphis State University
Memphis, TN 38152

Abstract

A method for microscanning in imaging sensors is developed that allows liquid-crystal beam steerers to be used as nonmechanical microscan devices. This submicroscanning method involves using liquid-crystal beam steerers to shift images on a focal plane array by a fraction of the amount used in typical microscan methods. Interpolation techniques based on interlaced sampling are used to produce images free of aliasing out to twice the Nyquist frequency determined by the focal plane array. Since a continuous phase ramp is produced by the liquid-crystal beam steerer, dispersion effects due to the grating-like nature of the devices are avoided. Simulations for both one- and two-dimensional cases are presented, as well as experimental results using a 3- to 5-μm imaging sensor and a liquid-crystal beam steerer designed for 1.064 μm operation.

Nonmechanical microscanning using optical space-fed phased arrays

Kenneth J. Barnard
Assistant Professor
Department of Electrical Engineering

Memphis State University
Memphis, TN 38152

1. Introduction

Microscanning is a technique for improving the resolution and reducing aliasing effects in staring imaging sensors.[1–3] The microscan process involves recording multiple images of a scene, where each image has been shifted by a fraction of the pixel pitch of the focal-plane array (FPA). The field-of-view (FOV) of the sensor remains fixed during the integration time for each image. Successive image fields are combined to form a single image frame that has a larger effective sampling rate than any of the individual image fields. Currently, microscanning is accomplished using a mechanical scan mirror that steers the FOV of the imaging sensor over sub-pixel distances.

Development of programmable optical space-fed phased arrays for beam steering in laser radar systems has prompted investigation into using these devices to steer broad spectral band radiation in passive sensors.[4] One such application would be as nonmechanical microscan devices. The advantages inherent in a nonmechanical system would be lighter weight, less mechanical complexity, and greater reliability. Costs would eventually decrease as the phased arrays and control electronics become cheaper to manufacture.

Optical space-fed phased arrays steer optical beams in a manner analogous to phased arrays used in microwave radar systems. An array of continuous liquid-crystal phase shifters allow a desired optical path delay (OPD) to be written spatially across an input beam. A linearly increasing OPD is identical to a prism and has the effect of steering the beam into a given direction. The maximum OPD is determined by the product of the birefringence and thickness of the liquid-crystal material. However, the switching speed is proportional to the square of the thickness of the nematic crystal layer.[4] To maximize switching speed, a linear phase ramp is generated by limiting the maximum OPD to one wavelength which produces a phase of 2π, and then periodically resetting the phase to 0 when a phase of 2π is reached in the phase ramp. Each phase ramp itself is also approximated in the liquid crystal by using a quantized number of steps between the 2π phase resets.[4] In this fashion, the liquid-crystal beam steerer still behaves as a prism for a given design wavelength, and the steering angle is determined by the angle of the phase ramp. For wavelengths

other than the design wavelength, the device acts as a blazed-phase grating and will produce dispersion when used as a broadband beam steerer.

In microscanning, small steering angles are required and in some cases it may be possible to use liquid-crystal beam steerers without phase resets, thus eliminating grating dispersion. Also, the liquid crystal material used in these devices is transmissive through the IR wavelength band from 0.8 μm to 12 μm, as well as over the visible wavelength band. For a typical microscan operation, the image on the FPA must be shifted over one half of the pixel pitch. In many cases, with the currently available liquid-crystal beam steerers, it is not possible to steer at these required angles without phase resets. Still, images shifted over a fraction of this distance, or submicroscanned, can be used in conjunction with interlaced sampling techniques to produce the same half-pixel microscan effect.

In this report, I examine the use of interlaced sampling techniques for submicroscanned images as would be produced by a liquid-crystal beam steerer in a staring imaging sensor. Image processing techniques are developed that allow submicroscanned images to be combined to produce an image with a Nyquist frequency that is twice that of the original images. Results of simulations on one- and two-dimensional submicroscanned images will be presented. Also, results of an experimental validation of nonmechanical microscanning using a liquid-crystal beam steerer will be given. These results indicate that interlaced sampling techniques and submicroscanning provide the same image enhancement as typical mechanical microscanning at the expense of added computation time. Use of hardware implementations of FFT algorithms and multiple-processor architectures would provide increased frame rates.

2. Sampling and Interpolation

A typical microscan pattern is a 2×2 microscan where four image fields are sequentially recorded to form one image frame. A standard scenario would have field one correspond to a reference position, field two displaced by half a pixel pitch to the right, field three displaced by half a pixel pitch vertically from field two, and field four displaced by half a pixel pitch to the left of field three. If each image field is composed of $N \times N$ pixels, then the combined image frame will consist of $2N \times 2N$ pixels. This effectively increases the spatial sampling frequency by a factor of two. The Nyquist frequency is likewise increased by a factor of two. Spatial filtering of the image due to the finite size of the detector is not affected by microscanning.

In many cases, nonmechanical liquid-crystal beam steerers cannot produce the required half-pixel pitch image shift for the typical microscan pattern without dispersion. It is possible to use interlaced sampling techniques to compensate for a smaller image shift by interpolating the data values at the half-pixel pitch image locations.[5,6] Thus, interlaced samples can be used to obtain an

image frame consisting of $2N \times 2N$ pixels. This may seem to be in violation of the Whittaker-Shannon sampling theorem. However, it should be recognized that the sampling theorem is a sufficiency condition and not a necessity constraint.[7] The interlaced sampling technique is an alternate reconstruction method that provides image data out to twice the Nyquist frequency without aliasing. Spatial frequencies higher than twice the Nyquist frequency are aliased just as in the 2×2 microscan. Interpolation can provide the same set of sample points as in the 2×2 microscan.

The derivation given here is similar to that contained in Refs. 5 and 6, except that here the sampling function remains fixed and the image is shifted. For simplicity, the one-dimensional case is analyzed first. The technique can be readily extended to two dimensions as will be shown. For the one-dimensional case, two data sets are required, where the image data in one set is shifted by a fraction of the pixel pitch relative to the other. Interpolation is based on using spatial frequency information contained in the zeroth- and ±first-order spectra resulting from the Fourier transform of the sampled data sets. A reconstruction filter modifies the spectra such that the sum of the filtered data sets will be free of aliasing out to twice the Nyquist frequency. We begin with a one-dimensional image function $f(x)$, where $f(x)$ has been blurred by both the optics of the system and by the effects of the finite size of the detectors in the FPA. The image is assumed to be bandlimited such that $F(\xi) = 0$ for $|\xi| > 1/d$, where upper-case letters indicate the Fourier transform, and d is the pixel pitch of the FPA. The image function is sampled to produce one sampled image, then shifted and sampled again to generate the second sampled image. These two sampled images can be written as

$$f_1(x) = f(x)\,\mathrm{samp}(x) \tag{1}$$

and

$$f_2(x) = f(x + a)\,\mathrm{samp}(x), \tag{2}$$

where a is the image shift and $\mathrm{samp}(x)$ is given by

$$\mathrm{samp}(x) = \frac{1}{d}\,\mathrm{comb}\left(\frac{x}{d}\right). \tag{3}$$

The goal of the reconstruction is to find interpolation filters, $m_1(x)$ and $m_2(x)$, such that $f(x)$ can be reconstructed exactly by

$$f(x) = f_1(x) * m_1(x) + f_2(x) * m_2(x). \tag{4}$$

Here, information contained in both sampled images is used to exactly reconstruct $f(x)$ even though $f_1(x)$ and $f_2(x)$ contain aliased information. This may be more intuitively pleasing if it is recognized that the reconstruction of $f(x)$ utilizes twice as many samples as in either of the sampled image functions. Finding $m_1(x)$ and $m_2(x)$ is most readily accomplished in the frequency domain. First, the Fourier transform of each of the sampled images is found. Fourier transforming Eqs. (1) and (2) gives

$$F_1(\xi) = F(\xi) * \mathrm{comb}(d\xi) \tag{5}$$

and

$$F_2(\xi) = F(\xi)\exp(j\,2\pi a\xi) * \mathrm{comb}(d\xi)\ . \tag{6}$$

This expected result indicates that the original spectrum of f(x) is replicated at integer multiples of 1/d in the spatial frequency domain. The only difference between Eqs. (5) and (6) is that the latter includes a linear phase shift. It is this linear phase shift that enables us to recover $F(\xi)$. Restricting our attention to the interval $-1/d < \xi < 1/d$, Eqs. (5) and (6) can be rewritten to include only those terms that are nonzero over this interval, as

$$F_1(\xi) = \frac{1}{d}\left[F\left(\xi + \frac{1}{d}\right) + F(\xi) + F\left(\xi - \frac{1}{d}\right)\right] \tag{7}$$

and

$$F_2(\xi) = \frac{1}{d}\left\{\exp\left[j\,2\pi a\left(\xi + \frac{1}{d}\right)\right] F\left(\xi + \frac{1}{d}\right) + \exp(j\,2\pi a\xi)\,F(\xi) + \exp\left[j\,2\pi a\left(\xi - \frac{1}{d}\right)\right] F\left(\xi - \frac{1}{d}\right)\right\}\ . \tag{8}$$

The phase terms in Eq. (8) remain centered about each of the replicated spectra.

We now have two equations in three unknowns that can be solved for $F(\xi)$ by restricting the domain of the solution. Since it was originally assumed that $F(\xi) = 0$ for $|\xi| > 1/d$, one of the unknowns in Eqs. (7) and (8) will be zero depending on whether $\xi > 0$ or $\xi < 0$. For $\xi > 0$, $F(\xi + 1/d) = 0$ and the set of equations in Eqs. (7) and (8) can be solved for $F(\xi)$ to give

$$F(\xi) = j\frac{d}{2}\frac{\exp\left(-j\,2\pi\frac{a}{d}\right)}{\sin\left(2\pi\frac{a}{d}\right)} F_1(\xi) - j\frac{d}{2}\frac{\exp\left(j\,2\pi\frac{a}{d}\right)}{\sin\left(2\pi\frac{a}{d}\right)} \exp(-j\,2\pi a\xi)\,F_2(\xi)\ . \tag{9}$$

Likewise, for $\xi < 0$, $F(\xi - 1/d) = 0$, and Eqs. (7) and (8) can again be solved for $F(\xi)$, giving

$$F(\xi) = -j\frac{d}{2}\frac{\exp\left(j\,2\pi\frac{a}{d}\right)}{\sin\left(2\pi\frac{a}{d}\right)} F_1(\xi) + j\frac{d}{2}\frac{\exp\left(-j\,2\pi\frac{a}{d}\right)}{\sin\left(2\pi\frac{a}{d}\right)} \exp(-j\,2\pi a\xi)\,F_2(\xi)\ . \tag{10}$$

Comparing the constant multiplying factors of $F_1(\xi)$ and $F_2(\xi)$ in Eqs. (9) and (10), it is evident that they are complex conjugates of each other in each equation and between the equations. This means that on the interval $-1/d < \xi < 1/d$, the two equations can be simplified and combined to form

$$F(\xi) = F_1(\xi)\,M(\xi) + F_2(\xi)\,M^*(\xi)\,e^{-j\,2\pi a\xi}\,, \tag{11}$$

where $M(\xi)$ can be derived from Eqs. (9) and (10) as

$$M(\xi) = \frac{d}{2}\,\mathrm{rect}\left(\frac{\xi}{2/d}\right) - \frac{j}{2}\cot\left(2\pi\frac{a}{2d}\right)\mathrm{tri}'\left(\frac{\xi}{1/d}\right)\ . \tag{12}$$

Here, the prime indicates a derivative operation with respect to ξ.

Thus, f(x) can be reconstructed exactly by filtering the spectra of the two sampled data sets in the spatial frequency domain, summing the results, and then performing an inverse transform. The interpolation function m(x) is given by

$$m(x) = \mathrm{sinc}\left(\frac{2}{d}x\right) - x\frac{\pi}{d}\cot\left(2\pi\frac{a}{d}\right)\mathrm{sinc}^2\left(\frac{x}{d}\right)\ . \tag{13}$$

Referring to Eq. (4), it is evident that $m_1(x)$ is just $m(x)$ and $m_2(x)$ is $m(a-x)$. One special case worth noting is when the image shift is equal to half the pixel pitch. In this case, $m(x)$ reduces to

$$m(x) = \mathrm{sinc}\left(\frac{2}{d}x\right), \tag{14}$$

or in the spatial frequency domain,

$$M(\xi) = \frac{d}{2}\,\mathrm{rect}\left(\frac{d}{2}\xi\right). \tag{15}$$

This corresponds to an ideal reconstruction filter for an image sampled at twice the original sampling frequency of 1/d.

3. Numerical Computation

The most direct computation of the submicroscanned image involves evaluating Eq. (11) in the spatial frequency domain and then inverse Fourier transforming to obtain $f(x)$. Discrete signal processing techniques are used in addition to the reconstruction algorithm to obtain the final image. Producing the submicroscanned image requires the spatial filtering of two N-point data sets to generate one 2N-point data set where every other data point has been correctly interpolated to avoid aliasing.

For the first step in the computation, the number of points in each data set is increased by a factor of two by inserting zeros between the data points. A system for accomplishing this is referred to as a sampling rate expander.[8] Proper discrete filtering in the frequency domain will correctly interpolate the zero-valued data points to the unaliased values. A fast Fourier transform (FFT) of either of the two expanded data sets produces a discrete spectrum similar to the one shown in Fig. 1, where because of the expander, $F_1[k]$ and $F_2[k]$ contain two sampled replications of the continuous zeroth order spectra, $F(\xi)$ and $F(\xi)\exp(j\,2\pi a\xi)$, indicated by the heavy solid lines. Aliasing is indicated by the overlap of the dashed spectra.

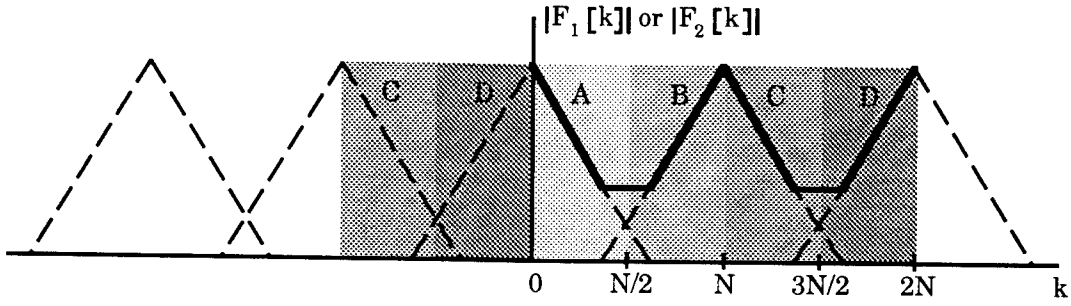


Fig. 1. Conceptual plot of the magnitude of the FFT of each expanded data set.

Recognizing that the portions of the spectrum labeled C and D in Fig. 1 are identical to those on the negative axis, it is evident that the FFT of each of the expanded data sets provides the spectra given in Eqs. (7) and (8) within the spatial frequency range of $-1/d < \xi < 1/d$. Thus, the reconstruction filter can be applied over this range to produce an unaliased spectrum out to twice the original Nyquist frequency. Examination of Eq. (11) indicates that the linear phase terms in Eq. (8) will be eliminated when multiplied by the complex conjugate phase term in Eq. (11) during the spatial filtering step. After applying the reconstruction filter, an inverse FFT of the sum of the two filtered data sets yields the unaliased data set.

4. Error Analysis

The previously described algorithm can exactly reconstruct the bandlimited image function provided that the image shift is precisely known, and round-off and quantization errors do not occur. We now examine the effects of pointing errors, i.e., image shift errors, on the submicroscanned image. For this analysis, a sinusoidal input is assumed that has a spatial frequency such that aliasing occurs when it is sampled. A closed-form solution for the reconstructed image is found and the error in reconstruction is defined as the mean-squared error between the reconstructed signal and the original signal. Round-off and quantization effects are not considered.

Assume the input function is a sinusoidal signal given by

$$f(x) = 2\cos(2\pi\xi_0 x), \tag{16}$$

where $1/(2d) < \xi_0 < 1/d$, so $f(x)$ is aliased when sampled with a sampling period d. The image shift a is now assumed to contain some error Δ such that the two sampled functions become

$$f_1(x) = f(x)\,\mathrm{samp}(x) \tag{17}$$

and

$$f_2(x) = f(x + a + \Delta)\,\mathrm{samp}(x), \tag{18}$$

where $\mathrm{samp}(x)$ is defined in Eq. (3). Using Eqs. (11) and (12) to reconstruct the spectrum of $f(x)$ on the interval $-1/d < \xi < 1/d$ gives

$$\begin{aligned} F_r(\xi) = &\left[\cos(\pi\Delta\xi_0) + \sin(\pi\Delta\xi_0)\cot\left(\frac{\pi a}{d}\right)\right]\left[\delta(\xi-\xi_0)\exp(j\,\pi\Delta\xi_0) + \delta(\xi+\xi_0)\exp(-j\,\pi\Delta\xi_0)\right] \\ &-\left[\frac{\sin(\pi\Delta\xi_0)}{\sin\left(\frac{\pi a}{d}\right)}\right]\left\{\delta\left[\xi-\left(\frac{1}{d}-\xi_0\right)\right]\exp\left[-j\,\pi\left(\frac{a}{d}+\Delta\xi_0\right)\right] + \delta\left[\xi-\left(-\frac{1}{d}+\xi_0\right)\right]\exp\left[j\,\pi\left(\frac{a}{d}+\Delta\xi_0\right)\right]\right\}, \end{aligned} \tag{19}$$

where the r subscript indicates the reconstructed spectrum. The last two terms in Eq. (19) represent aliased spectral components that have not been completely eliminated in the reconstruction. The reconstructed signal is found from the inverse Fourier transform of Eq. (19) as

$$f_r(x) = 2\left[\cos(\pi\Delta\xi_o) + \sin(\pi\Delta\xi_o)\cot\left(\frac{\pi a}{d}\right)\right]\cos(2\pi\xi_o x + \pi\Delta\xi_o)$$
$$-2\left[\frac{\sin(\pi\Delta\xi_o)}{\sin\left(\frac{\pi a}{d}\right)}\right]\cos\left[2\pi\left(\frac{1}{d}-\xi_o\right)x-\left(\frac{\pi a}{d}+\pi\Delta\xi_o\right)\right]. \tag{20}$$

It should be noted that expressions for the reconstructed signal and its spectrum given in Eqs. (19) and (20) are valid for both positive and negative shift errors. For positive shift errors, the term corresponding to the unaliased signal in Eq. (20) increases in amplitude, and the aliased term subtracts. For negative shift errors, the unaliased term decreases in amplitude, and the aliased term undergoes a phase reversal. This suggests that it may be possible to use the phase reversal of the aliased components to determine the exact image shift in real images.

The reconstruction error is defined as the mean-squared error between the reconstructed signal and the original signal and is given by

$$\text{MSE} = \overline{|f_r(x) - f(x)|^2}\,. \tag{21}$$

Using Eqs. (16) and (20), the difference between the reconstructed and original signals can be written as

$$f_r(x) - f(x) = 2\left\{\cos(\pi\Delta\xi_o)\left[\cos(\pi\Delta\xi_o) + \sin(\pi\Delta\xi_o)\cot\left(\frac{\pi a}{d}\right)\right] - 1\right\}\cos(2\pi\xi_o x)$$
$$-2\left\{\sin(\pi\Delta\xi_o)\left[\cos(\pi\Delta\xi_o) + \sin(\pi\Delta\xi_o)\cot\left(\frac{\pi a}{d}\right)\right]\right\}\sin(2\pi\xi_o x)$$
$$-2\left[\frac{\sin(\pi\Delta\xi_o)}{\sin\left(\frac{\pi a}{d}\right)}\right]\cos\left[2\pi\left(\frac{1}{d}-\xi_o\right)x-\left(\frac{\pi a}{d}+\pi\Delta\xi_o\right)\right]. \tag{22}$$

In all cases of interest, the image shift will be less than half the pixel pitch, $a < d/2$, the magnitude of the error will be less than the image shift, $|\Delta| < a$, and spatial frequency of the input signal will be restricted to the recoverable bandwidth, $\xi_o < 1/d$. Assuming small image shifts and small shift errors such that $a/d << 1$, Eq. (22) can be approximated as

$$f_r(x) - f(x) \approx 2\left(\frac{\Delta\xi_o d}{a}\right)\cos(2\pi\xi_o x) - 2(\pi\Delta\xi_o)\left[1+\left(\frac{\Delta\xi_o d}{a}\right)\right]\sin(2\pi\xi_o x)$$
$$-2\left(\frac{\Delta\xi_o d}{a}\right)\cos\left[2\pi\left(\frac{1}{d}-\xi_o\right)x-\left(\frac{\pi a}{d}+\pi\Delta\xi_o\right)\right]. \tag{23}$$

The approximate mean-squared error is found by substituting Eq. (23) into Eq. (21), giving

$$\text{MSE} \approx 4\left(\frac{\Delta\xi_o d}{a}\right)^2 + 2(\pi\Delta\xi_o)^2\left[1+\left(\frac{\Delta\xi_o d}{a}\right)\right]^2. \tag{24}$$

To minimize the mean-squared error in the reconstructed sinusoidal signal, the shift error Δ must be small compared to the submicroscan image shift. The reconstruction error also depends on the frequency of the aliased signal, with smaller frequencies producing less error. In addition, because of the second term in Eq. (24), the MSE will be slightly less for negative shift errors than positive errors. This suggests that when the exact image shift is unknown, it is advantageous to overestimate the image shift when applying the reconstruction algorithm.

5. Two-dimensional Interp

The one-dimensional interpolation technique can be extended into two-dimensions by applying the reconstruction algorithm on each dimension separately. This requires four image data sets shifted on a 2×2 rectangular submicroscan grid, where the x- and y-direction image shifts need not be equal. As will be shown, the separability of the interpolating filters in each direction allows the reconstruction algorithm to be interpreted in terms of two-dimensional filtering. Utilizing two-dimensional FFT routines, the four images can be filtered simultaneously and combined to form the submicroscanned image.

Assuming an image function $f(x, y)$, a 2×2 submicroscan operation will produce four sampled data sets given by

$$f_1(x, y) = f(x, y)\, \text{samp}(x, y), \tag{25}$$

$$f_2(x, y) = f(x + a, y)\, \text{samp}(x, y), \tag{26}$$

$$f_3(x, y) = f(x, y + b)\, \text{samp}(x, y), \tag{27}$$

and

$$f_4(x, y) = f(x + a, y + b)\, \text{samp}(x, y), \tag{28}$$

where a and b are the submicroscan image shifts in the x and y directions, and

$$\text{samp}(x, y) = \frac{1}{d^2}\, \text{comb}\left(\frac{x}{d}, \frac{y}{d}\right). \tag{29}$$

Ignoring the y direction and applying the one-dimensional reconstruction algorithm on the image pairs f_1 and f_2, and f_3 and f_4 along the x direction, generates two images interpolated in the x direction but shifted relative to each other in the y direction. In the spatial frequency domain, the spectra of the two interpolated images can be written according to Eq. (11) as

$$G_1(\xi, y) = F_1(\xi, y)\, M(\xi) + F_2(\xi, y)\, M^*(\xi) \exp(-j\, 2\pi a \xi) \tag{30}$$

and

$$G_2(\xi, y) = F_3(\xi, y)\, M(\xi) + F_4(\xi, y)\, M^*(\xi) \exp(-j\, 2\pi a \xi), \tag{31}$$

where the upper-case letters indicate one-dimensional Fourier transforms. Next, applying the one-dimensional algorithm in the y direction on the image pair in Eqs. (30) and (31) yields the spectrum of the final reconstructed image. Defining an interpolation filter for the y direction based on Eq. (12) as

$$N(\eta) = \frac{d}{2}\, \text{rect}\left(\frac{\eta}{2/d}\right) + \frac{j}{2}\, \cot\left(2\pi\, \frac{b}{2d}\right) \text{tri}\left(\frac{\eta}{1/d}\right), \tag{32}$$

the spectrum of the reconstructed image can be written as

$$F_r(\xi,\eta) = G_1(\xi,\eta)\,N(\eta) + G_2(\xi,\eta)\,N^*(\eta)\,exp(-j\,2\pi b\eta)\,, \tag{33}$$

or

$$\begin{aligned} F_r(\xi,\eta) &= F_1(\xi,\eta)\,M(\xi)\,N(\eta) + F_2(\xi,\eta)\,M^*(\xi)N(\eta)\exp(-j\,2\pi a\xi) \\ &+ F_3(\xi,\eta)\,M(\xi)\,N(\eta)\exp(-j\,2\pi b\eta) + F_4(\xi,\eta)\,M^*(\xi)N^*(\eta)\exp[-j\,2\pi(a\xi + b\eta)]\,. \end{aligned} \tag{34}$$

From Eq. (34), a two-dimensional interpolating filter function can be determined for each of the four sampled image data sets.

Numerically, the four N × N image data sets are first expanded to 2N × 2N points by inserting zeros for every other data point in the x and y directions. Taking the two-dimensional FFT of the images yields the sampled versions of the image spectra in Eq. (34) over the spatial frequency range of $-1/d < \xi < 1/d$ and $-1/d < \eta < 1/d$. Multiplying each image spectra by its associated two-dimensional interpolation filter in Eq. (34), and then summing the results produces the final image spectrum. A final inverse FFT yields the 2N × 2N unaliased submicroscanned image.

6. Simulations

The one-dimensional submicroscan technique was demonstrated using a one-dimensional chirped-frequency signal. This signal was chosen to illustrate aliasing effects in the sampled image. The object function before sampling consisted of 2210 data points and is shown in Fig. 2.

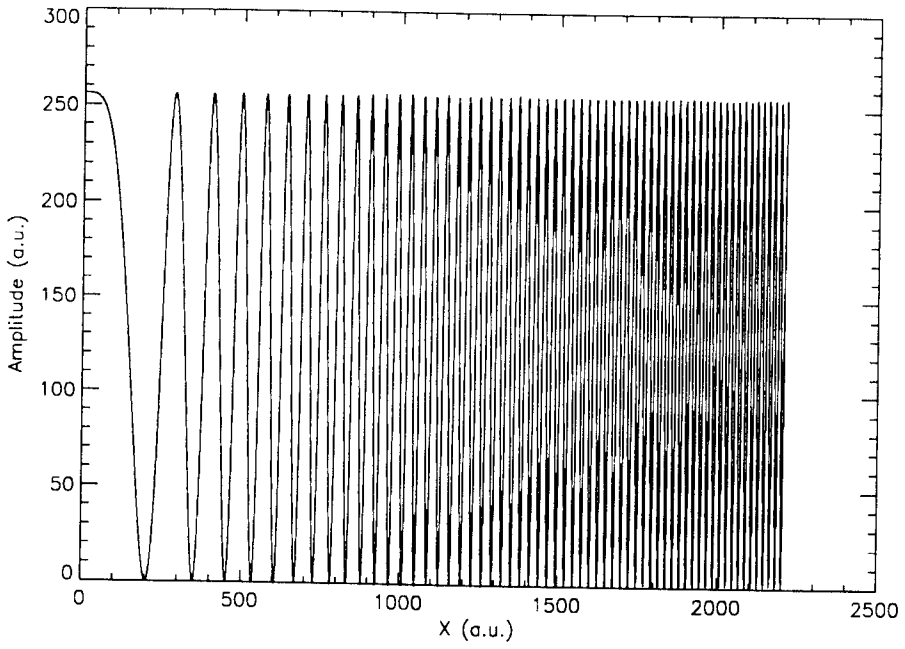


Fig. 2. Chirped-frequency object function.

For this case, the effects of the optics were ignored. The detectors in the FPA were assumed to have a width of 13 points and a pixel pitch of 17 points. Prior to sampling, the object function was convolved

with a detector function to simulate blurring due to the finite detector size. The convolution kernel was assumed to be a rectangle function with a width of 13 points. The blurred image is shown in Fig. 3.

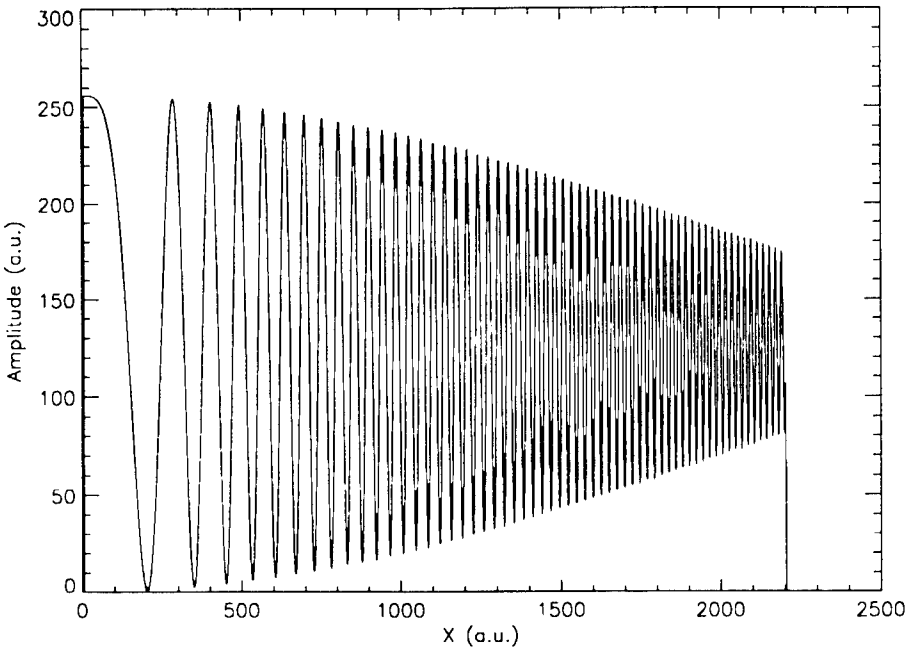


Fig. 3. Blurred object function.

The sampled images were generated by sampling the blurred object function at every 17 points to simulate the detector spacing. This produced images consisting of 128 points. The relative image shift between the two data sets was 2 points, giving an image shift of 12 percent of the pixel pitch. The two 128-point images were then interpolated using the submicroscan technique to produce a 256-point image. Also, one of the 128-point images was interpolated to 256 points using standard techniques for comparison. This standard image and the submicroscanned image are shown in Figs. 4 and 5. Aliasing in the standard image is obvious and it is completely removed in the submicroscanned image.

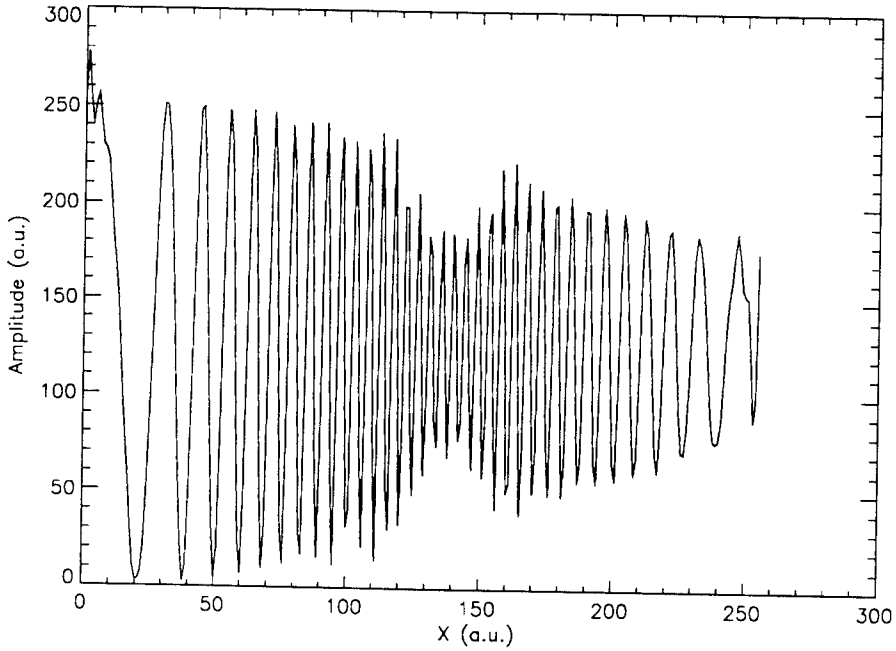


Fig. 4. Image obtained without submicroscanning.

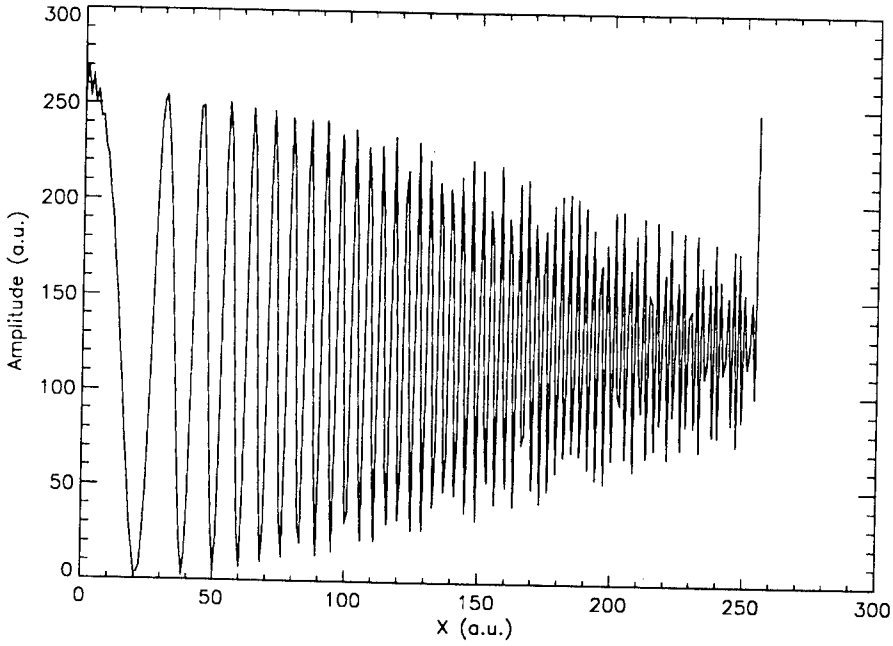


Fig. 5. Image obtained with submicroscanning technique.

Sampling artifacts that cause variations in the amplitude of the signal are apparent in the submicroscanned image . As shown in Fig. 6, these artifacts vanish when the image is interpolated to 1024 points using standard techniques.

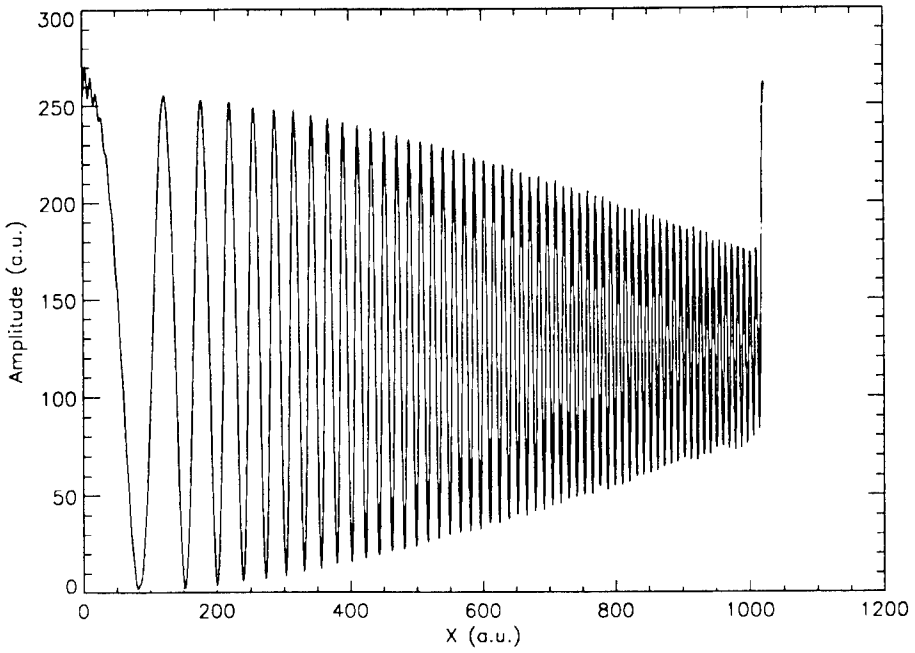


Fig. 6. Submicroscanned image interpolated to 1024 points.

For the two-dimensional case, a Fresnel zone plate was used as the test object. Image blurring was not performed prior to sampling for this case. The images were sampled by taking every 17th point in the object as in the one-dimensional case to produce 128 × 128 images. A rectangular submicroscan grid was used to generate four 128 × 128 images with a relative image shift of 6 percent of the pixel pitch in each shift direction. The four images were interpolated using the two-dimensional submicroscan technique to produce a 256 × 256 image. For comparison, one of the 128 × 128 images was interpolated to 256 × 256 points using standard techniques as was done in the one-dimensional case. This standard image and the submicroscanned image are shown in Figs. 7 and 8. The reduction in aliasing is clearly evident in Fig. 8.

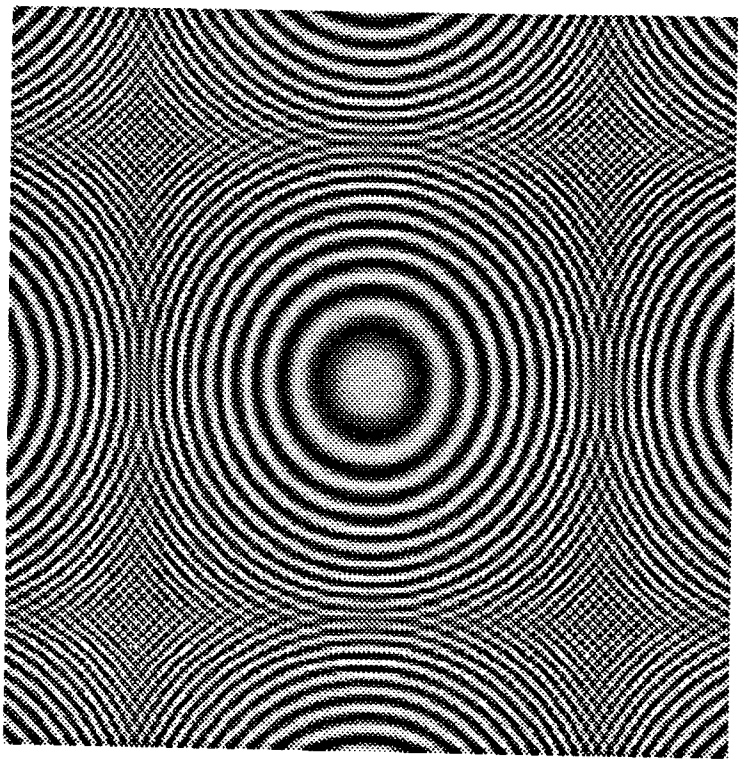


Fig. 7. Two-dimensional image without submicroscanning.

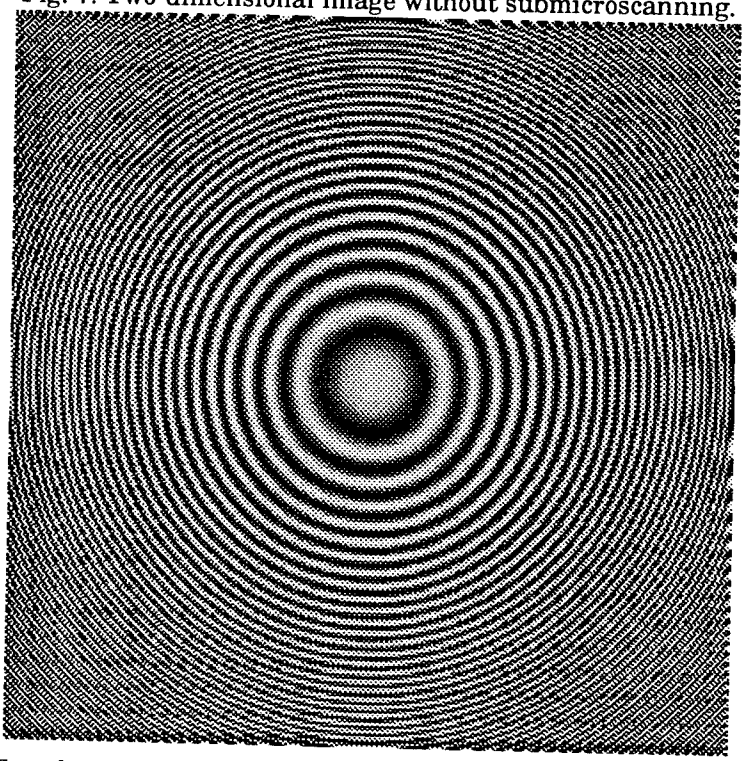


Fig. 8. Two-dimensional image obtained with submicroscanning technique.

7. Experimental Results

A verification of the submicroscan interpolation technique was performed using a liquid-crystal beam steerer designed for operation at 1.064 μm. This device produced a maximum OPD of 1.064 μm over a clear aperture of 2.05 cm, giving a submicroscan angle of 52 μrad. The imaging sensor was an Amber InSb system with a 128 × 128 FPA and associated optics for 3- to 5-μm operation. The detectors in the FPA were 40-μm square with a 50-μm pixel pitch. An F/3 lens system with a 100-mm focal length combined with a submicroscan angle of 52 μrad generated an image shift of 5.2 μm. The experimental setup is shown in Fig. 9.

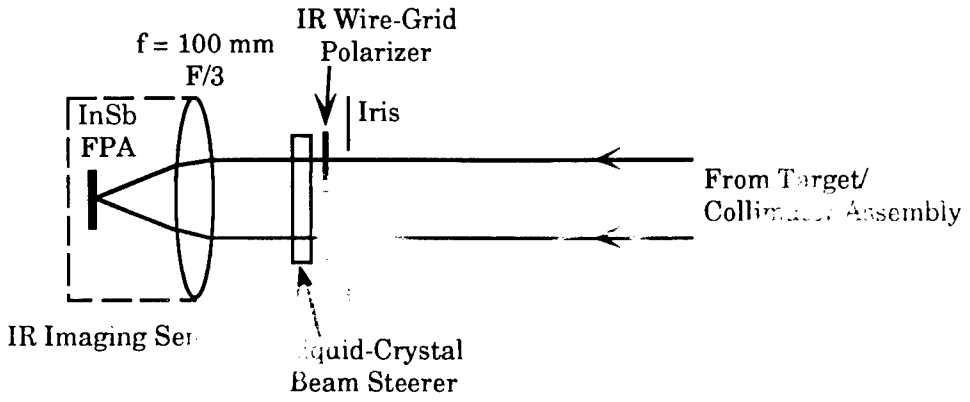


Fig. 9. Nonmechanical microscan experimental setup.

The IR wire-grid polarizer was necessary to polarize the incident radiation along the extraordinary ray direction in the liquid crystal. An adjustable iris positioned in front of the polarizer blocked unwanted stray radiation at the expense of increased background noise. Due to time constraints, the cold shield of the imaging system was not adjusted to reduce this background noise.

Although the beam steerer was designed for 1.064 μm, it was predicted to have approximately 35 percent transmittance from 3 to 3.5 μm. The overall transmittance of the beam steerer-polarizer combination was expected to be below 12 percent. High-temperature blackbody targets were necessary due to the low transmittance and short wavelengths. A target temperature of 61° C with a background temperature of 50° C was chosen to maximize image contrast within the constraints of the blackbody sources. Since the Nyquist frequency of the imaging system was 1.0 cyc/mrad, a four-bar target with a spatial frequency of 1.5 cyc/mrad was chosen so that aliasing would occur. The cutoff spatial frequency due to the detector size was 2.5 cyc/mrad.

Two image fields were acquired; one with the beam steerer off and the other with the beam steerer switched on. In both cases, the image of target was aliased and unresolved. These image fields are shown in Figs. 10 (a) and (b). The two images were then processed using the submicroscan interpolation technique to produce an unaliased submicroscanned image. Contrast in both of the

images was low, but as shown in Fig. 11, the four-bar target is resolvable in the submicroscanned image.

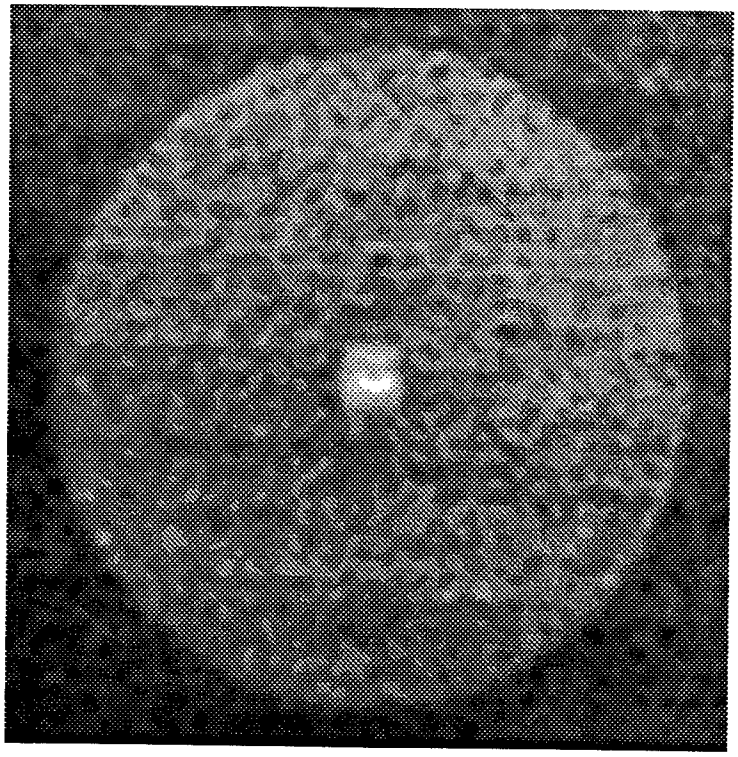


Fig. 10(a). Aliased image of a four-bar target with the beam steerer on.

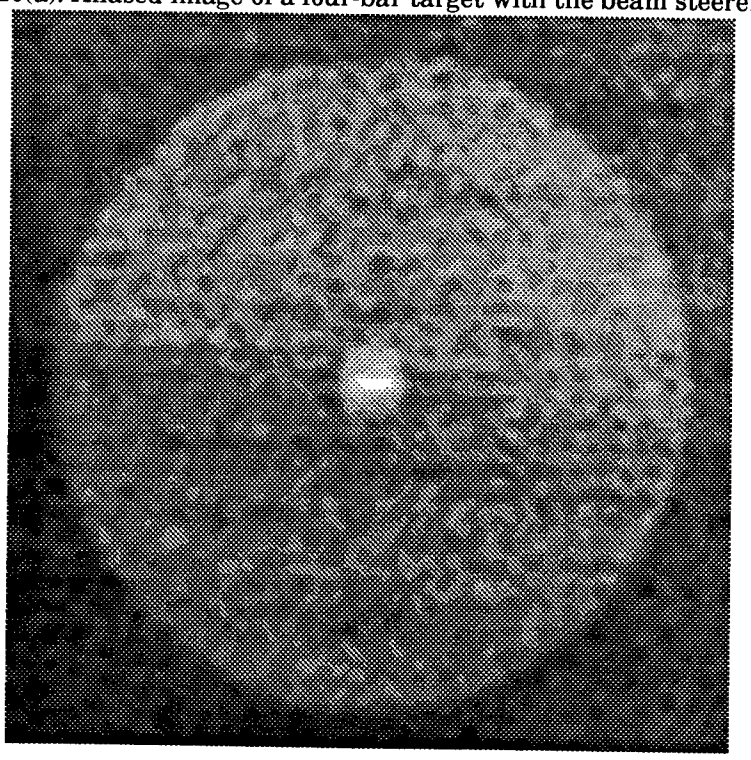


Fig. 10(b). Aliased image of a four-bar target with the beam steerer off.

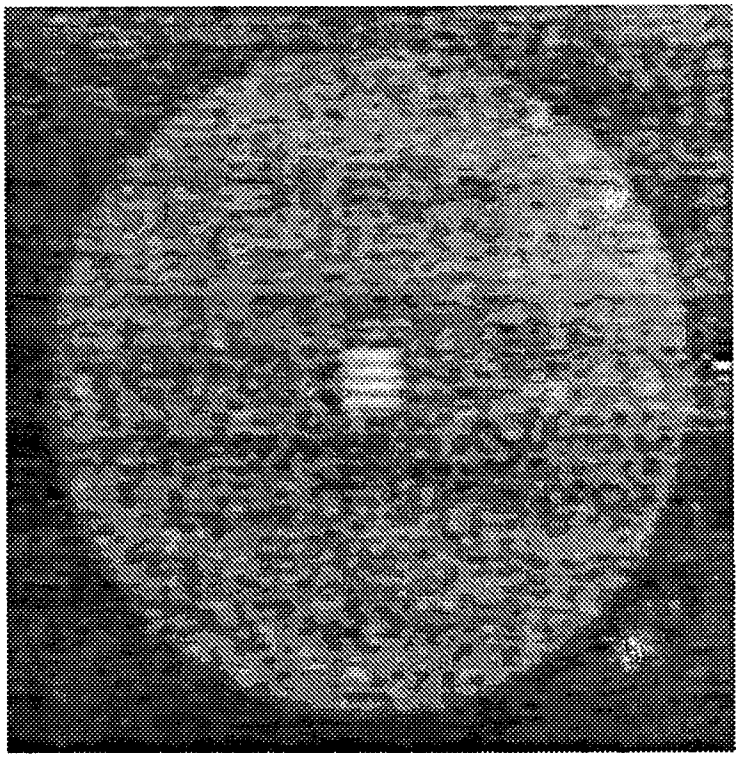


Fig. 11. Submicroscanned image of a four-bar target.

8. Conclusions

Submicroscan interpolation techniques provide a method of incorporating optical space-fed phased arrays as nonmechanical beam steerers in microscanned imaging sensors. While typical microscan patterns require image shifts of a half pixel pitch, the submicroscan patterns used here require image shifts of only a small fraction of the pixel pitch. Small image shifts are necessary because the liquid-crystal beam steerers have a limited steering angle if dispersion effects associated with the grating-like nature of these devices are to be avoided. The interlaced sampled images produced by these devices can be processed using the unique interpolation techniques derived here to obtain results equivalent to those of typical microscanning.

The experimental results demonstrate that liquid-crystal beam steerers can be used to perform nonmechanical microscanning when combined with submicroscan interpolation techniques. The results given here were generated under non optimum operating conditions with a device not designed for operation over 3 to 5 μm. A device specifically designed for this application would have a much higher transmittance over the passband and would produce a dramatic improvement in image contrast.

Currently available liquid-crystal beam steerers can only steer in one direction, but they would be able to provide image enhancement along one direction in the FOV of a staring imaging sensor. Hardware implementation of FFT algorithms and the use of multiprocessor systems should provide increased frame rates for submicroscanned sensors utilizing liquid-crystal beam steerers.

9. Acknowledgments

I would like to thank Paul McManamon for making this research possible, and, along with Ed Watson, Bill Martin, and Bob Muse, for providing valuable discussions. Mike Shelton (EOIR) and Rob Fetner (T/SSI) provided invaluable assistance in obtaining the experimental data. I would also like to acknowledge Melody Maloney for purchasing the IR wire-grid polarizer within the time frame of my summer program through the use of the Fast Acquisition Program. This polarizer was a critical element in performing the experimental verification of the theory.

10. References

1. D.J. Bradley and P.N.J. Dennis, "Sampling effects in CdHgTe focal plane arrays," in *Infrared Technology and Applications*, Proc. SPIE, vol. 590, 53–60 (1985).

2. R.J. Dann, S.R. Carpenter, C. Seamer, P.N.J. Dennis, and D.J. Bradley, "Sampling effects in CdHgTe focal plane arrays — practical results," in *Infrared Technology XII*, Proc SPIE, vol. 685, 123–128 (1986).

3. E.A. Watson, R.A. Muse, and F.P. Blommel, "Aliasing and blurring in microscanned imagery," in *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing III*, Proc. SPIE, vol. 1689, 242–250 (1992).

4. P.F. McManamon, E.A. Watson, T.A. Dorschner, and L.J. Barnes, "Nonmechanical beam steering for active and passive sensors," in *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing IV*, Proc. SPIE (1993).

5. R.N. Bracewell, *The Fourier Transform and Its Applications*, 2nd Ed., pp. 201–202, McGraw-Hill, 1986.

6. J.D. Gaskill, *Linear Systems, Fourier Transforms, and Optics*, pp. 276–278, Wiley, 1978.

7. W.H. Nicholson and A.A. Sakla, "The Sampling and Reconstruction of Band-Limited Signals at Sub-Nyquist Rates," *IEEE International Symposium on Circuits and Systems*, vol. 3, 1796–1800 (1990).

8. A.V. Oppenheim, and R.W. Schafer, *Discrete-Time Signal Processing*, pp. 105–111, Prentice Hall, 1989.

AN APPROACH FOR UNIFIED SENSOR MANAGEMENT DESIGN

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, Washington, D.C.

December 1993

AN APPROACH FOR UNIFIED SENSOR MANAGEMENT DESIGN

Milton L. Cone
Assistant Professor
Department of Computer Science/Electrical Engineering
Embry-Riddle Aeronautical University

Abstract

Modern aircraft are sophisticated machines with increasing numbers of sensors that generate more data and operate in a greater variety of modes than ever before. To improve the performance of these machines a sensor manager, in the words of Stan Musick, is necessary "to direct the right sensor to the right task at the right time." The design of a sensor manager is difficult. The common approach is to use ad hoc methods to develop rules which direct the sensors' operation under various conditions. The goal is to develop a technique that unifies the sensor manager design. This report examines one such technique, the Analytic Hierarchy Process. The author concludes that a modified version of the Analytic Hierarchy Process is a good candidate for the job of design integrator for the development of a sensor manager.

An Approach for Unified Sensor Management Design

Milton L. Cone

Introduction

In future combat missions where high valued assets are used against large numbers of lesser valued assets, a very high exchange rate will be militarily as well as politically necessary. Future sensor systems are expected to produce large quantities of information that can easily overload a pilot thus degrading his performance and reducing the exchange ratio. Several software, pilot interface and hardware improvements are needed to aid the pilot in reacting and controlling these advanced systems. The Formal Mathematical Methods for Sensor Management (FMMSM) is one of these programs. Initiated by Stan Musick of WL/AAAS-3, Wright Labs, Wright-Patterson AFB, OH, this program addresses the design of smart controllers for the on-board sensors. While recent sensor managers for tactical aircraft have been assembled by ad hoc or combinations of methods, this program's approach is to develop rigorous methods for a unified sensor management design.

AHP[1] Overview

This report uses the Analytic Hierarchy Process, AHP, as a method to guide the design of the functions of the sensor manager system. In some cases an AHP design will be used to achieve a sensor manager function. In other cases the AHP design philosophy will be used as a paradigm to guide the development. This approach maintains a consistent design philosophy throughout the sensor management system. The design of one sensor manager function will be completed using two AHP decompositions.

The first principle of AHP is that a system can be broken down into hierarchies. Most techniques subscribe to this type of decomposition. In AHP these hierarchies reflect a natural decomposition of a system into the basic elements of the problem. Saaty[1] calls these basic elements the problem's entities.

By progressing from top to bottom of the AHP hierarchy, the entities change from general or more uncertain to more particular or definite . Entities can be grouped into levels which influence only the entities at the next higher level and are influenced only by entities at the next lower level. Within any one level an entity is influenced by all of the entities at the next lower level. See Figure 1 reproduced from reference 2 for an example. In that figure the entity "Maximize Long Range Knowledge" is affected by all of the level 3 entities but not directly by any of the level 4 requests.

The second AHP principle is comparative judgement. Preferences are developed between each of the entities at a level with each of the entities at the next lower level. These preferences are pair-wise comparative judgements of the relative importance of each of the lower level entities to each one of the entities at the next higher level. For example in Figure 1, "ATR-Classification Requests" is compared to "ATR-Recognition Requests" to determine which entity better satisfies the subgoal of "Maximum Long Range Knowledge" . A pair-wise comparison is made between each of the request characteristics to determine their contribution towards maximizing long range knowledge. The pair-wise comparisons become elements in a comparison matrix which is used to determine overall priorities. Thus "Maximum Long Range Knowledge" produces a 13 x 13 preference matrix of pair-wise comparisons. Additionally, each of the four subgoals also has a 13 x 13 matrix associated with it.

The third principle of AHP is the synthesis of priorities. In going from the bottom to the top of the hierarchy, AHP determines the ability of each of the lowest level actions to contribute to each of the entities in each of the higher levels until finally the best action to satisfy the global priority is chosen. In Figure 1, AHP ranks all requests according to their ability to meet some overall sensing goal. As each request is processed through the hierarchy different request characteristics and subgoal priorities cluster together. The pair-wise preference matrices determined above make sure that the contribution of each request to each of the intervening levels is properly included. At the top level each sensor request has been compared to each other sensor request on the basis of its ability to satisfy the overall goal. Saaty

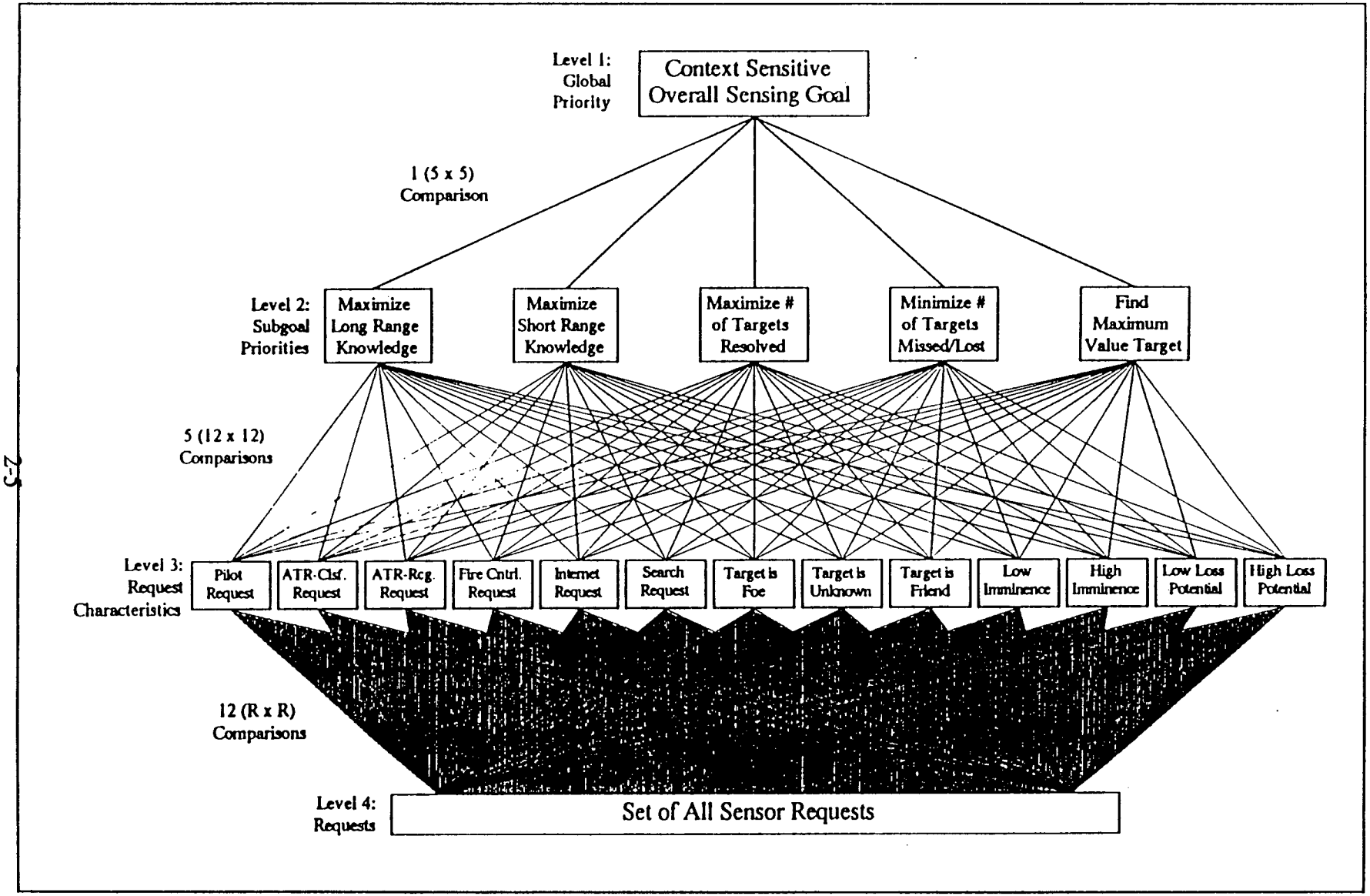


Figure 1 A Sensor Management AHP Hierarchy: Multiple Criteria on the Set of Requests

provides details on the matrix manipulations. They will not be reproduced here as the main interest of this paper is on decomposition.

When the pilot is a sensor manager, he develops a prioritized target list, assigns which sensor he wants to perform each task and then determines when he wants the task to happen. He may have valid reasons for some choices and preferences for others. The difference between a valid reason and a preference is that a valid reason is a preference supported by stronger evidence. AHP is a process which formalizes the pilot's actions into a design for a sensor manager.

Mission Manager Overview

The guidance to contractors for the FMMSM program outlines how a sensor manager system fits intc the larger mission manager system. Generally a mission manager coordinates all on-board operations, tailoring the operation of each of the systems to the particular phase of the mission. In any aircraft there are basically three things to control. These are the flight controls of the plane, the weapon systems that may be on-board the plane and the sensor systems that are available in the plane. Figure 2 shows how

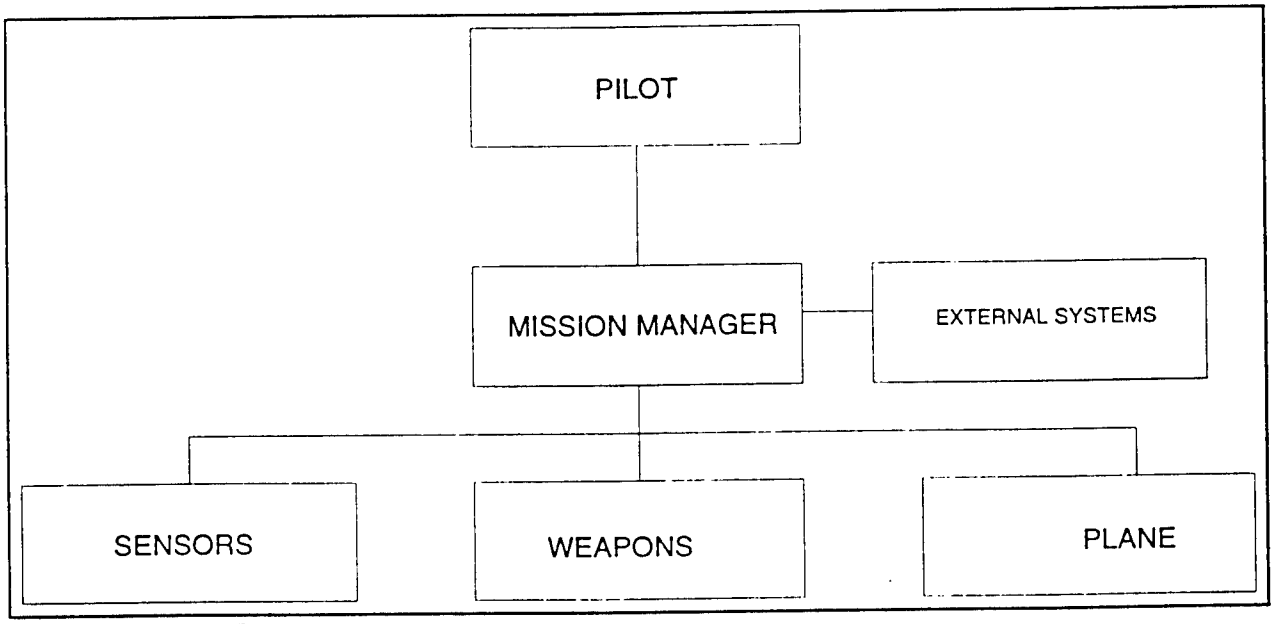


Figure 2 Aircraft System

information flows between the system manager and the pilot, on-board systems and external systems. Modern aircraft will be internetted with other aircraft in its flight as well as with external systems such as AWACS. For the FMMSM study the internetting function was not considered. This approach makes a simpler problem whose solution can be extended when internetting is permitted.

Figure 3 outlines the inner workings of the mission manager. There are seven functions within the mission manager . The Mission Manager Executive directs the operation of the other six. It maintains control of the flow of all information into and out of the mission manager and between functions within the mission manager. While flight control is far more complex than indicated, it is not the focus of this paper and so is represented by a single box. (This will be done again in the description of the pilot-vehicle interface where several functions are grouped together. In AHP terminology this is analogous to

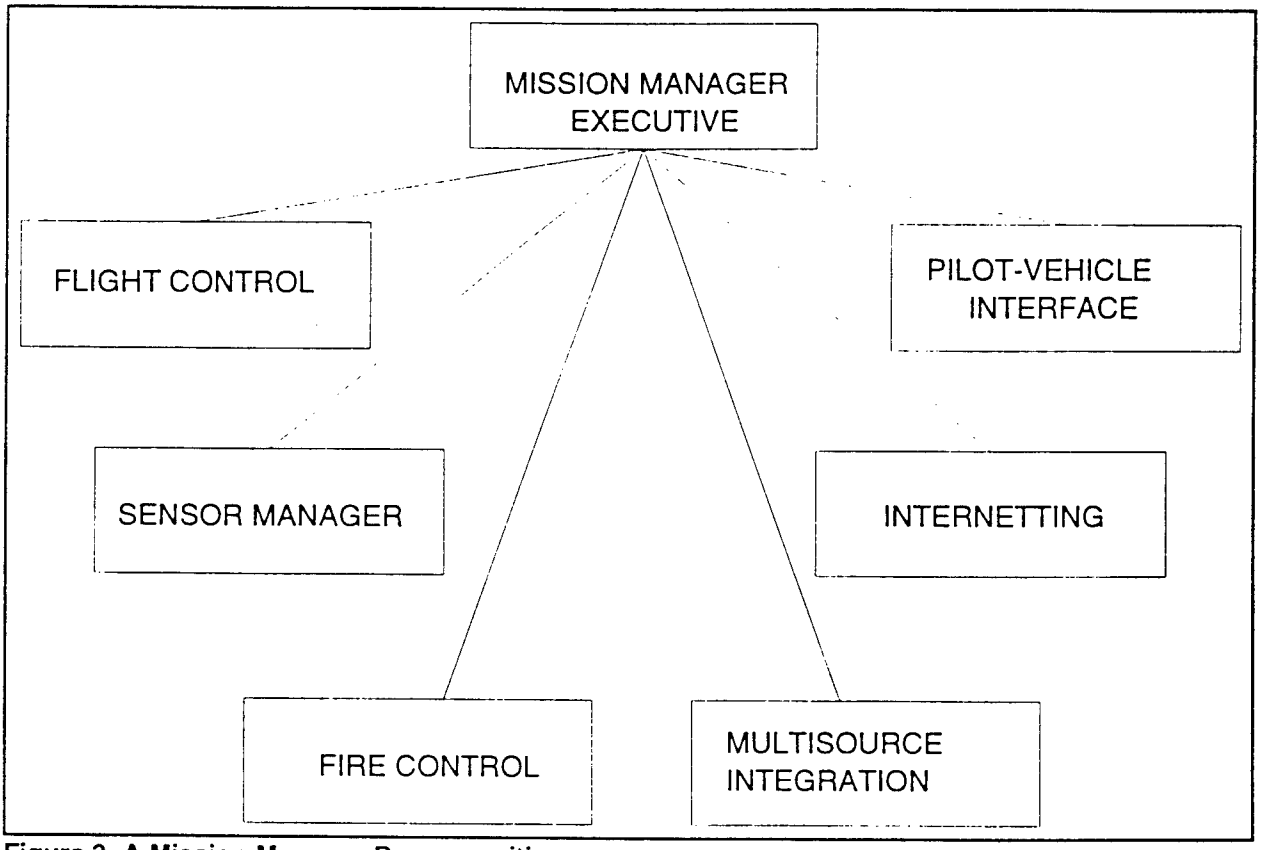


Figure 3 A Mission Manager Decomposition

clustering of entities when moving up the hierarchy.) The internetting function has been disabled and will be referred to only when necessary to show how that function can be incorporated into the current design. Three of the four remaining functions will be briefly described. These descriptions are similar to the ones in the A³M study[3]. The fourth, Sensor Management, will be more fully described using AHP to develop the design.

The function of fire control is to determine prioritization for each track designated by the pilot or the internetting function (ignored here), to select and allocate weapons, and to compute the launch envelopes for ownship's weapons for each target. Missile update guidance and kill assessment can be computed if required.

The pilot-vehicle interface allows the pilot to communicate with the mission manager executive. While the broad view of the mission manager taken here includes all controls with which the pilot interacts, the concentration is on displays and cockpit controls involving the sensors. The status of all sensors and weapons systems, missile launch envelopes and kill assessment may be displayed. The pilot inputs his commands through controls on the stick, cockpit switches or touch-screen menu options. In advanced designs, the pilot-vehicle interface may also include voice actuation. The pilot-vehicle interface includes mission planning and automatic route planning. Many of these functions would be broken out into their own blocks in a complete system description. Here it is convenient to group them together.

Multisource integration, MSI, fuses sensor data and creates and maintains the MSI track files. These files are composite tracks derived from inputs from all on-board sensors. Inputs from the radar and IRST sensors (assumed available) are used to develop composite kinematic tracks. An identification function is also provided which determines classification (fighter, bomber, tank, armored personnel carrier, et cetera), target type (Mig 29, B-52, T-72, BMP-1) and friend/foe/neutral. It is assumed that an individual sensor continues to perform its tracking function if it has one. The multisource integration function fuses

existing correlated tracks or attribute data from each sensor. (In some designs the individual sensor track correlation function is also performed by the MSI.) The MSI then prepares a request for the sensor manager indicating on which targets it wants additional data, what service it requires and why it thinks this data is required. Generally this will be a request for information on all new and active tracks. The architectural implementation of this request generation process may be at the sensor or MSI level as part of that system, at the sensor manager but distributed at the sensor or MSI processor, or at the sensor manager as part of the sensor manager's central processor. Here request generation is assumed to take place at the source of the data. Hence each sensor generates its own requests for the sensor manager as does the MSI. If this assumption is not true then the sensor manager diagram in Figure 4 gains another block for request generation.

Sensor Manager

The sensor manager directs the controllable sensors on the aircraft. The manager intelligently makes sensor assignments based on requests of the mission manager executive. This may include pilot requests, requests from the internetting function for cooperative attack planning, requests from the fire control for weapons launch or attack support and requests for special operations like passive sensing from the mission manager executive. Figure 4 shows the functions that the sensor manager performs and the flow of information into and out of each function. The sensor manager first validates the requests. This includes checking the existing database to see if the request can be fulfilled without any additional sensing. The request then enters onto a list of tracks to be serviced. The sensor manager prepares this list of all tracks requiring sensing from the special requests, the multisensor integrator, the tracking sensors and the mission manager (for search requests). (The basic principle is that every function should prepare its own request. The sensor manager takes the global view and prioritizes these requests.) This process requires inputs from the mission manager executive to establish mission priorities, the navigation system for ownship coordinates, the internetting system and the broad area

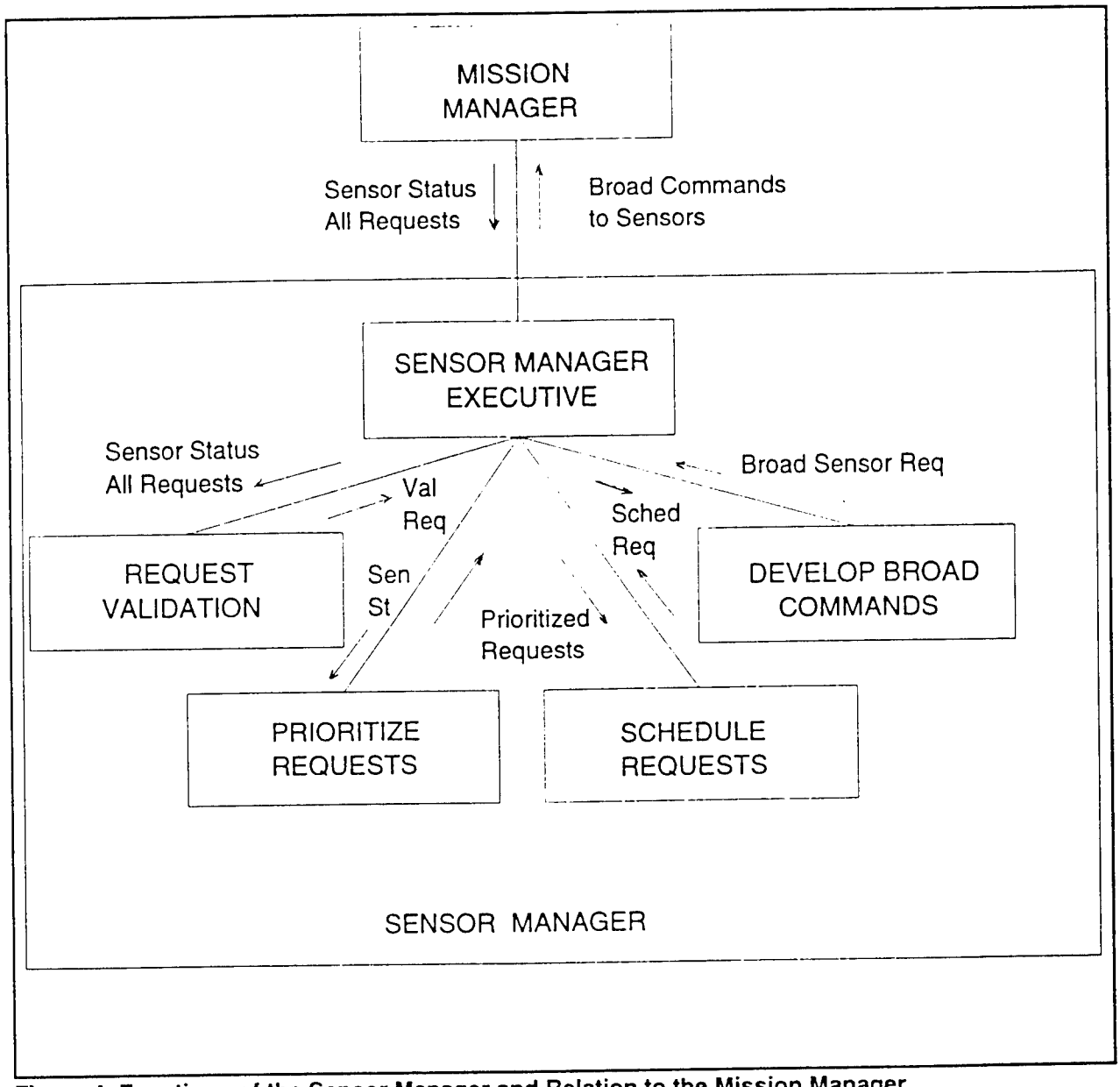


Figure 4 Functions of the Sensor Manager and Relation to the Mission Manager

detection systems like a radar warning receiver for threat areas to search.

Next the sensor manager has to create a prioritized list of tracks.The sensor manager can do four things.

It can:

1. order a track update on an existing track to maintain tracking and/or to increase accuracy,
2. order a sensor to collect more information to try to identify by class, type or friend/foe/neutral,

3. direct a sensor to determine more information about the number of targets in an existing track, and

4. redirect a sensor's search area or technique to satisfy mission priority.

For these four functions the Sensor Manager determines a priority based on a set of predetermined criteria and weights. From the prioritized list of tracks the sensor manager determines which sensor services are to be performed by each controllable sensor and issues the broad commands to each sensor. In the next section more detail about each of the sensor manager functions and an AHP design for the "Prioritize Requests" is given.

Request Validation Function

The Mission Manager communicates the sensor status and all requests for sensor service to the Mission Manager Executive. The Mission Manager Executive coordinates the activities of the other mission manager functions. Requests for sensor service can come from the sensors, the MSI, the pilot, Internetting and Fire Control. A request is composed of a request number assigned by the Sensor Manager Executive, a track number if one exists, a request for the type of service desired (which are track update, track identification, track raid assessment or area search), and a request for a particular sensor and mode if known. These requests are validated according to broad sensor availability rules. These include such criteria as:

1. The sensor is not operational
2. The sensor is under the control of the pilot
3. The mission goals restrict the use of the sensor.

Generally only certain requests will require a specific sensor. These typically come from the pilot or from a lower level cueing where one sensor derives pointing commands for another sensor or predicts the onset of a particular target activity or maneuver. Any interaction between sensors should be coordinated

through the sensor manager. The sensor manager has a global view of sensor requirements. Letting the sensors interact on their own makes it difficult for the sensor manager to effectively control the limited sensor resources.

Prioritize Request Function

The validated requests are passed to the Prioritize Requests function. This function establishes the preference for which request to service first and which sensor mode will fill that request. Figures 5 and 6 show how those preferences are developed using AHP. A request comes in to the sensor manager with a known type of service desired (update track, update ID, update raid assessment or perform search). The type of service picks one of the paths. For example, the preferences for update tracks would be (1, 0, 0, 0) indicating no preference for update ID, update raid assessment or search. (It might be possible for a

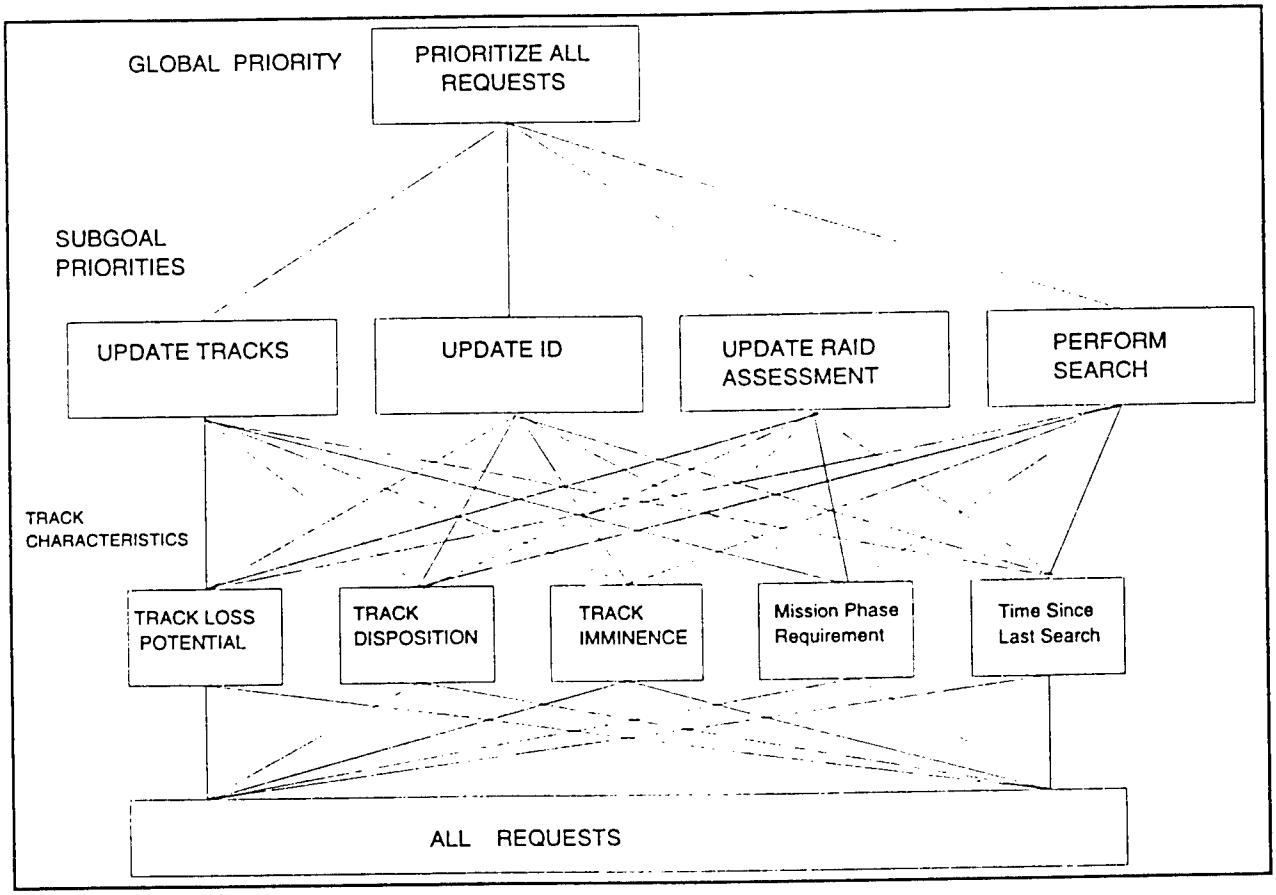


Figure 5 Prioritize All Requests

preference to be mixed for example (1, 1, 0, 0) indicating an equal desire for two services.) The next level lists characteristics that allow discrimination of which type of service best fits the incoming request. At the lowest level are all of the requests. For each request, functions can be created that calculate a value for track loss potential and track imminence[3]. Track loss potential is the likelihood of losing a track; track imminence estimates the time to encounter ownship. For example the function that determines track loss potential might be a function of track speed, altitude, range, track ID, time since last update, variances associated with MSI estimates of track's range, azimuth and elevation. The track imminence function might be a function of distance to ownship, track velocity and their relative variances. The function for track disposition expresses the intent of the track as either friendly, unknown or hostile. This information is available from the attribute fusion function of the MSI. The Mission Phase Requirement is a preference for track or search based on mission phase. The mission phase is established by pilot input or the mission manager. Time Since Last Search shows the preference for search over track as the search data ages. This characteristic forces a search to be done periodically. For this decomposition, AHP's prioritization scheme is similar to that advocated in Popoli[4] using fuzzy set theory. See his "Demonstration 1".

In Figure 5, preferences must be expressed between the various service requests and the various characteristics of the tracks. While there can be a good argument that all are equally important, it seems that in most cases raid assessment is less important than knowing the track's ID or maintaining a good track. The danger in assigning these values is that in the end all requests will be ranked with those highest ranked generally getting the best and fastest service. If update ID is generally given a higher preference than raid assessment, then raid assessment may not ever get done. But if maintaining a good track means doing a raid assessment then the two combine to make this a very high priority task. The strength of AHP is that these preferences are clearly evident and easily modified.

Figure 6 shows how the prioritized requests from Figure 5 get assigned to the desired sensor and sensor

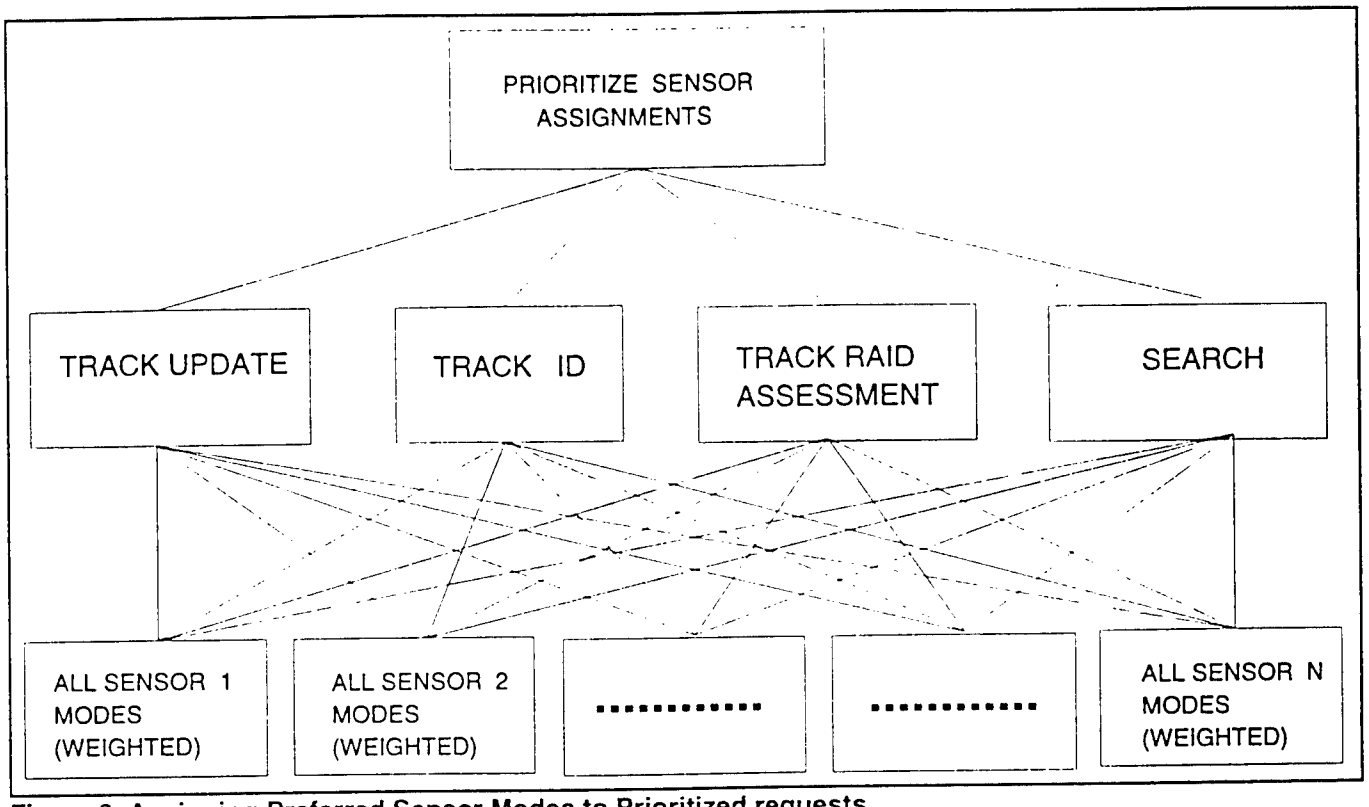


Figure 6 Assigning Preferred Sensor Modes to Prioritized requests

mode. The goal of this process is a list of tracks, along with the associated preference for which tracks to service, and the preference for which sensor mode to assign to that track. The sensor manager exists at two levels[4]. There is the macro-level that is centralized in one processor and a μ-level which is distributed in the individual sensor processors. When the sensors generate the original requests the μ sensor manager calculates whether a track update, ID update or raid assessment is required. This process will be based on a set of criteria such as maximum position error, maximum velocity error, minimum track score determined by some function (there may be several of these), maximum residual allowable or time from last update. The determination to request a track update carries with it the values of these criteria. The macro-manager takes these values and prioritizes them in real time. The preferences for which sensor modes best address a type of request are made off-line. Assignments are made using a weighted AHP[2] technique that reflect a sensor mode's availability. The mode availability will be reduced by a factor k, $0<k<1$, each time that that mode is assigned to a request. The factor k is based on the average expected duration of the tasks to be executed between sensor manager updates. An example

may help clarify how this process works.

Suppose a request for a track update is generated by the radar sensor μmanager. The μmanager has sensed that the track quality is degrading because the maximum time criteria from the last update and the maximum residual has been exceeded. It has been established that the first track quality measure is best corrected by sensor 1, modes a, d or f or sensor 3, modes e or g and the latter track quality measure is corrected by sensor 2, mode b or sensor 3, modes e or h. Depending on the degree by which one sensor mode is preferred over the other, sensor 3, mode e will probably have the highest preference. After picking sensor 3, mode e the weight of that sensor mode is reduced by the factor k. Assume k is 0.25 and the previous weight of sensor 3, mode e is 1.0, then for the next time the preferences are generated that involve sensor 3, mode e the preference for that sensor mode is reduced but not to 0. The reason for not reducing the weight to zero the first time that sensor mode is picked is that the process may be completed in less time than it takes to repeat the sensor manager cycle. This technique allows one sensor in a particular mode to be used more than once. Depending on how the sensor operates it may be necessary to reduce all of the sensor's modes by a factor when any mode for that sensor is picked.

Schedule Requests Function

Leaving the Prioritize Request block are requests that are prioritized according to their importance and which have assigned sensors and sensor modes appropriate for their requirements. The next block on Figure 4 is the Schedule Requests function. The two common approaches to scheduling are the myopic or "best first" and the non-myopic or "brick laying approach[4]." The best first approach takes the highest priority request and fills it first and then works through the list until there are no more requests or no more sensor time to be allocated. The brick laying approach tries to optimize the scheduling so that some criteria is maximized. Frequently, this criteria is to serve the most requests. A search of all combinations of potential sensor assignments finds the combination that serves the most requests. The major

drawback to the bricklaying approach is the amount of time that the exhaustive search takes and the inability to serve pop-up requests[4]. Popoli suggests that the macro-manager has two types of taskings, fluid and hard deadline. Fluid tasks can be scheduled most anytime during the sensor manager cycle but may still have some weak constraint like "can't start before such a time". The hard deadline tasks have to be scheduled within a narrow window of time. An example of a hard deadline task is a missile update request generated from the fire control system. Either approach may be better based upon a specific application but Popoli recommends the myopic approach since the macro-manager generally has very fluid tasking requests. This implies many solutions are nearly optimal so that an exhaustive search is unnecessary. He also feels that a myopic scheduler can best handle pop-up requirements like rescheduling a failed sensor request without waiting until the next macro-manager cycle. The myopic scheduler described by Popoli looks at the hard deadline tasks first. The highest priority task is taken and scheduled. Then as many fluid tasks as can be fit before the first hard deadline occurs will be scheduled. Fluid tasks are taken on a highest priority basis. Anytime a pop-up task is sensed a new fluid schedule is generated and the process repeats.

Develop Broad Commands Function

The next block in figure 4 is develop broad commands. This function generates the high level tasking for the sensor. These commands tell the sensor what tasks to perform. The commands are communicated from the central processor to a part of the sensor manager located in the sensor's processor. This remote sensor manager is called a μ sensor manager. The μ sensor manager develops the low level commands that tell the sensor how a particular task can best be accomplished.

Summary

This report has shown how AHP can be adapted to perform the functions required for a unified sensor design. AHP follows the three principles of problem solving[1]. These are decomposition, comparative judgements and synthesis of priorities. Within the broad framework that AHP provides, many different

techniques for uncertainty management can be incorporated. This makes AHP a good candidate for the job of design integrator for the development of a sensor manager.

References

1. Saaty, T. The Analytic Hierarchy Process, RWS Publications, Pittsburgh, Pa 1990.

2. "An Analytic-Based Sensor Management Software System", AFWAL-TR- , Data Fusion Corporation, December 1992.

3. "Air-to-Air Attack Management System", WRDC-TR-90-1116, Vol I, Part I, Northrop Corporation, October 1990.

4. Popoli, R., "The Sensor Management Imperative" in Multitarget-Multisensor Tracking: Applications and Advances, Vol II, Bar-Shalom, Y. (ed), Artec House, Inc, Norwood, Ma 1992.

SCANNING IMAGE PROCESSING FOR OPTICAL REMOTE SENSING

Bradley D. Duncan, Ph.D.
Assistant Professor
Center for Electro-Optics
Department of Electrical Engineering

University of Dayton
300 College Park Avenue
Dayton, OH 45469-0226

Final Report for:
Summer Faculty Research Program
Wright Laboratory
(WL/AARI-2)

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

July, 1993

SCANNING IMAGE PROCESSING FOR OPTICAL REMOTE SENSING

Bradley D. Duncan, Ph.D.
Assistant Professor
Center for Electro-Optics
Department of Electrical Engineering
University of Dayton

Abstract

An incoherent two pupil scanning image processing technique appropriate for optical remote sensing applications (e.g., LADAR) has been investigated. By scanning simultaneously with two superposed optical beams, of slightly different temporal frequencies, complex (i.e., amplitude and phase) and/or bipolar spatial filtering operations can be applied to intensity representations of remote objects. The technique is similar in a sense to the way by which computer based post processing has traditionally been applied to incoherent optical images in that an appropriate optical convolution kernel, generated by the two pupil interaction process, is convolved by scanning with object intensity records. The all optical technique, however, is capable of real time image processing, and in many instances may be capable of producing higher resolution processed images than are achievable by computer based post processing alone.

SCANNING IMAGE PROCESSING FOR OPTICAL REMOTE SENSING

Bradley D. Duncan, Ph.D.

INTRODUCTION:

In conventional (i.e., single pupil) optically incoherent scanning image processing systems, severe limitations exist on the image processing functions available for implementation due to resulting non-negative intensity spread functions. Such limitations can be avoided, however, by introducing a two pupil system in which both pupils simultaneously and interactively process a desired image through scanning [1,2]. As will be shown, any bipolar and/or complex incoherent impulse response can be synthesized by using two pupil methods, as long as each pupil function can be arbitrarily specified.

Heuristically, optical remote scanning image processing is achieved by convolving a target image with an appropriate point spread function (PSF), or convolution kernel which describes the desired processing function. The convolution kernel is synthesized by choosing appropriate scanner plane pupil functions, such that after diffracting to the target, the desired PSF is generated. The PSF is then moved about the target by scanning (analogous to the sliding process required for mathematical convolution), after which the back scattered light is collected and converted to an electrical voltage or current signal (analogous to the integration process required for mathematical convolution). The processed image can then be viewed (possibly in real time [2,3]) by synchronously mapping (with respect to the scanner position) the signal thus generated to some display device.

Incoherent processing of remote target images is of interest due to the inherent reduction of undesired speckle artifacts which plague coherently processed images. As with most familiar computer based post detection image processing schemes, incoherent optical processing often involves the enhancement or extraction of desired features contained within an intensity based target image. By comparison to computer based processing schemes, though, all optical techniques are capable of producing images processed in real time, as the desired processing function is implemented optically at the target prior to detection. Furthermore, all optical techniques will generally be capable of producing higher resolution processed images than are possible with computer based techniques, since optical processing occurs at the target itself, whereas computer based post processing is performed on already reduced resolution (i.e., spatially low pass filtered) target intensity records [4,5]. Ideally, optical processing techniques are limited in resolution only by the diffraction limited resolution of the receiving optics.

Contained within is a general description of a proposed two pupil superheterodyne scanning image processor appropriate for implementation in optical remote sensing schemes. The accompanying theoretical analysis will detail the process by which bipolar and/or complex incoherent processing functions can be achieved. Comparisons will also be made between the two pupil technique and similar coherent and incoherent single pupil scanning image processing techniques. Finally, some practical trade-off issues will be addressed and directions for continuing and future research will be identified.

TWO PUPIL SUPERHETERODYNE SCANNING IMAGE PROCESSING:

The proposed scanning image processing system based on acousto-optic two pupil interaction is shown schematically in Figure 1. Assuming short intra-processor propagation distances and linear (e.g., horizontal) source polarization, we see that the pupil functions U and V are superposed on an x-y scanning device after passing through a polarizing beam splitter cube (PBS) (e.g., oriented to pass horizontally polarized light) and a quarter wave plate oriented to produce a circularly polarized transmit/scanning beam. The composite scanning beam is then expanded by a telescope and is subsequently used to scan a remote target Γ, located a distance z from the telescope's monostatic transmit/receive aperture. The back scattered energy received from the scanned target then follows a reciprocal path back through the telescope and quarter wave plate to produce a linearly polarized beam which is reflected by the polarizing beam splitter cube in the direction of a photodetector (PIN). Also notice lens ℓ_1. This lens is chosen and placed so as to produce a reduced image of the scanner at the photodetector plane in order to reduce the effects of scanner wobble; though it should be mentioned that for remote targets which are small with respect to their distance from the transmit/receive aperture, the peak-to-peak scanner deflection will typically be very small. For example, only a 0.2° peak to peak scanner deflection would be required to scan a 30 m target at a distance of 10 km.

Notice now that the contribution due to pupil U is upshifted in frequency according to the operating frequency f_c of the indicated acousto-optic modulator (AOM). After detection then, the processed target information will be contained in an intermediate frequency (IF) signal centered at frequency f_c. The photodetector is thus followed by a bandpass filter (BPF), centered at f_c, and an RF amplifier appropriate for amplifying the IF signal. Under many circumstances the IF signal will now simply be

synchronously demodulated by mixing with a pure sinusoid at frequency f_c, after which the demodulated information is available for output to some display or storage device. If, for instance, the demodulated IF signal is mapped to a display screen in synchronism with the scanner's motion, the processed image of the target Γ will be seen, as will be shown shortly. Figure 1, however, shows a more general demodulation scheme in which, prior to display or storage, the IF signal is first mixed down to a reduced intermediate frequency f_o and level shifted by the addition of the DC voltage V_{DC}. These steps are in general necessary in order to enure the preservation of IF phase information after display and/or to provide the opportunity for incorporating spatial carriers into the processed image. These more complex demodulation procedures have, for example, proven useful in the area of optical scanning holography in which a scanning processing system similar to that shown in Figure 1 has been employed [2,3].

To begin a more formal analysis of the incoherent scanning image processing technique, we first, for convenience, consider the two pupil functions U and V as they exist at the telescope's exit aperture. Specifically, we write

$$V_T(\overline{\rho}_i) = V(\overline{\rho}_o / M) \quad , \tag{1}$$

and

$$U_T(\overline{\rho}_i) = U(\overline{\rho}_o / M) \quad , \tag{2}$$

where U_T and V_T are expanded pupil functions which are subsequently transmitted toward the target, M is the telescope magnification and $\overline{\rho}_o$ and $\overline{\rho}_i$ represent the two dimensional intra-processor and transmit/receive aperture coordinate systems, respectively. As the processed target image will ultimately be resolution limited upon

detection according to the size and shape of the telescope's monostatic aperture, we also introduce the notation $W(\overline{\rho}_1)$ to represent the monostatic aperture function. After scanning and collecting the back scattered light, the time varying electrical signal produced by photodetection is then written, in voltage form, as

$$v(\overline{\rho}, z; t) = \iint_{-\infty}^{\infty} | W'(\overline{\rho}_2) * \{ [V_T'(\overline{\rho}_2 - \overline{\rho}; z) + U_T'(\overline{\rho}_2 - \overline{\rho}; z)\, e^{-j2\pi f_c t}] \Gamma(\overline{\rho}_2) \}|^2 \, d\overline{\rho}_2 \quad , \quad (3)$$

where, in the case of very distant targets, the primes indicate the far field (i.e., Fraunhoffer) diffraction patterns of the indicated pupil and aperture functions, * indicates convolution, $\overline{\rho}_2$ is the two dimensional target coordinate system, $\overline{\rho}$ is a time varying two dimensional shift variable corresponding to the motion of the scanner, and where for generality the distance z to the target has been retained as a parameter. Ideally, with an infinite receiver aperture (i.e., $W'(\overline{\rho}_2) = \delta(\overline{\rho}_2)$), equation (3) is interpreted as follows. First, the target $\Gamma(\overline{\rho}_2)$ is illuminated by the superposition of U'_T and V'_T, the far field diffraction patterns of pupil functions U_T and V_T, where U'_T is shifted in frequency by an amount f_c due the action of the AOM shown in Figure 1. As the composite scanning beam is scanned about the target then, the back scattered light is collected by the receiver aperture, focused onto a photodetector and converted to an electrical signal. Due to the finite size of the receiver aperture, though, the received optical information must first be convolved with the far field diffraction pattern of the receiver aperture function $W'(\overline{\rho}_2)$, prior to detection, in order to account for the limited spatial frequencies capable of passing through the receiver aperture. As such receiver apertures typically act as low pass spatial frequency filters, this convolution process in essence serves to limit the resolution of the scanned processed image [4,6].

Upon performing the appropriate simplifications of equation (3), the IF signal $\tilde{v}$ following the RF amplifier of Figure 1 is found to be

$$\tilde{v}(\bar{\rho}, z; t) = \Re e\Big[\int\int_{-\infty}^{\infty} \Big\{ W'(\bar{\rho}_2) * \big(V_T'(\bar{\rho}_2 - \bar{\rho}; z)\, \Gamma(\bar{\rho}_2) \big) \Big\}^{*} \times \Big\{ W'(\bar{\rho}_2) * \big(U_T'(\bar{\rho}_2 - \bar{\rho}; z)\, \Gamma(\bar{\rho}_2) \big) \Big\}\, d\bar{\rho}_2\, e^{-j2\pi f_c t} \Big] \quad , \tag{4}$$

where $\Re e$ indicates the real function and the superscripted * indicates complex conjugation. Notice that equation (4) is in the form of the inverse phasor transformation. That is, we may write

$$\tilde{v}(\bar{\rho}, z; t) = \Re e\big[\tilde{V}(\bar{\rho}, z)\, e^{-j2\pi f_c t} \big] \quad , \tag{5}$$

where the phasor $\tilde{V}(\bar{\rho}, z)$ is given by

$$\tilde{V}(\bar{\rho}, z) = \int\int_{-\infty}^{\infty} \Big\{ W'(\bar{\rho}_2) * \big(V_T'(\bar{\rho}_2 - \bar{\rho}; z)\, \Gamma(\bar{\rho}_2) \big) \Big\}^{*} \times \Big\{ W'(\bar{\rho}_2) * \big(U_T'(\bar{\rho}_2 - \bar{\rho}; z)\, \Gamma(\bar{\rho}_2) \big) \Big\}\, d\bar{\rho}_2 \quad . \tag{6}$$

Notice that equation (6) is a function only of spatial coordinates and is in general complex. In fact, equation (6), though rather cumbersome at this point, represents the complex processed image of the target $\Gamma(\bar{\rho}_2)$. We'll look at equation (6) more carefully later. For now, though, consider the demodulated signal produced by further frequency and level shifting as shown in Figure 1. In general, the signal finally available for output to a display or storage device is written as

$$v_o(\bar{\rho}, z; t) = V_{DC} + \Re e\big[\tilde{V}(\bar{\rho}; z)\, e^{-j2\pi f_c t} \big] \quad . \tag{7}$$

Note the similarity of this result and that given by equation (5). The only difference is that the temporal frequency has been changed to f_o and that a DC bias has been applied. Let's look at each of these effects separately.

First, consider that case where $f_o = 0$. That is, assume that the IF frequency is removed from equation (5) by mixing, prior to storage or display. Equation (7) then becomes

$$v_o(\bar{\rho}; z) = V_{DC} + |\tilde{V}(\bar{\rho}; z)| \cos\{\Phi_{\tilde{v}}(\bar{\rho}; z)\} \quad , \qquad (8)$$

where $\Phi_{\tilde{v}}(\bar{\rho}; z)$ is the phase of the processed image represented by equation (6). We see then that the addition of V_{DC} is required in order to ensure that $|v_o| > 0$. For instance, most display devices are incapable of responding to negative modulating signals. In order to properly preserve the phase $\Phi_{\tilde{v}}$ of the processed image then, V_{DC} is simply adjusted to ensure that $|v_o| > 0$.

Next, consider the case for which $\tilde{V}(\bar{\rho}; z) = 1$, everywhere. That is, consider $\Gamma(\bar{\rho}_2)$ to be a uniformly illuminated white background, for instance. Also consider that the signal of equation (7) is fed to the modulating input of some display device whose electron gun (in the case of a common CTR display monitor) is adjusted for raster scanning. That is, the velocity v_x of the electron gun in the x-direction, is very much greater than the velocity v_y of the electron gun in the y-direction. Equation (7) can then be written as

$$v_o(\bar{\rho}; z) = \frac{1}{2} + \cos\left(2\pi \frac{f_o}{v_x} x\right) \quad , \qquad (9)$$

where we have assumed that $V_{DC} = 1/2$. After display, then, we see that the effect of the temporal frequency f_o is to create a corresponding spatial frequency f_o/v_x. Though not always necessary, as previously mentioned the incorporation of spatial frequency carriers by this method has proven useful in the generation of off axis, spatial carrier frequency holograms by scanning techniques similar to those described here. The only requirement is that the spatial frequency f_o/v_x be resolvable by the chosen display device. Thus f_o is usually chosen much smaller than f_c, as the AOM drive frequencies are usually on the order of tens to hundreds of MHz. We therefore interpret equation (7) as representing the processed image $\tilde{V}(\overline{\rho}; z)$ of our target $\Gamma(\overline{\rho}_2)$, level shifted by V_{DC} in order to preserve the phase Φ_V, and upon which has been imposed a spatial carrier frequency f_o/v_o, as desired. Now let's look at the processed image $\tilde{V}(\overline{\rho}; z)$ given in equation (6) more carefully.

Consider for now that the monostatic transmit/receive aperture of the telescope in Figure 1 is arbitrarily large, such that $W'(\overline{\rho}_2) = \delta(\overline{\rho}_2)$. Equation (6) then becomes

$$\begin{aligned} \tilde{V}_\infty(\overline{\rho}; z) &= \int\int_{-\infty}^{\infty} V_T'^*(\overline{\rho}_2 - \overline{\rho}; z)\, U_T'(\overline{\rho}_2 - \overline{\rho}; z)\, |\Gamma(\overline{\rho}_2)|^2\, d\overline{\rho}_2 \\ &= \left(V_T'^*(\overline{\rho}; z)\, U_T'(\overline{\rho}; z)\right) \otimes |\Gamma(\overline{\rho})|^2 \end{aligned} \quad , \qquad (10)$$

where the subscript ∞ indicates an infinite receiver aperture and $\otimes$ indicates correlation. In terms of the more common convolution process we may write

$$\tilde{V}_\infty(\overline{\rho}; z) = \left(V_T'^*(-\overline{\rho}; z)\, U_T'(-\overline{\rho}; z)\right) * |\Gamma(\overline{\rho})|^2 \quad . \qquad (11)$$

Notice that equation (11) in general represents an incoherent image processing function since the complex convolution kernel operates on the target intensity distribution $|\Gamma|^2$.

Though not conveniently obvious, the more general representation of the processed image given in equation (6) is simply a spatially low pass filtered version, of sorts, of the ideal processed image given in equation (11). Furthermore, and quite interestingly, equation (10) may also be applied directly in the case where a target is not so distant from the scanner as to be in the far field. In such a case (e.g., a laboratory environment) where resolution limits due to finite receiver apertures are less severe, or possibly negligible, we simply take the primes of equation (10) to represent Fresnel diffraction and interpret $\tilde{V}_\infty(\overline{\rho}; z)$ as our complex processed image.

A greater challenge is to now determine the appropriate pupil functions U_T and V_T such that after propagation to the target the desired convolution kernel is created. In practice the two pupil synthesis process proceeds, generally, as follows. First, a decision must be made as to the processing function (e.g., edge extraction) which is to be applied to the target. A convolution kernel capable of yielding this processing function is then identified (e.g., the "Laplacian-of-the-Gaussian," or LOG kernel, for edge extraction [5]) and written as the product of two other functions $V_T'^*(-\overline{\rho}; z)$ and $U_T'(-\overline{\rho}; z)$. We then make a change of variables and back propagate these functions to the scanner plane to determine the pupil functions $V_T(\overline{\rho}_1)$ and $U_T(\overline{\rho}_1)$. As may be expected, the last step is, in general, the most difficult, and may need to be performed numerically.

COMPARISON TO SINGLE PUPIL PROCESSING:

To further investigate the advantages of the incoherent two pupil scanning image processing technique we shall now compare it to two other similar scanning image processing techniques. First consider a single pupil direct detection (i.e., no optical mixing at either the transmit or receive stages) scanning image processor. We will also, for convenience, consider an arbitrarily large receiver aperture. For such a system then,

the post detection electrical signal, represented as a voltage, will be

$$v(\bar{\rho};\, z) = \int\int_{-\infty}^{\infty} \left| U'(\bar{\rho}_2 - \bar{\rho};\, z)\, \Gamma(\bar{\rho}_2) \right|^2 d\bar{\rho}_2 = \int\int_{-\infty}^{\infty} \left| U'(\bar{\rho}_2 - \bar{\rho};\, z) \right|^2 \left| \Gamma(\bar{\rho}_2) \right|^2 d\bar{\rho}_2 \quad , \tag{12}$$

or

$$v(\bar{\rho};\, z) = \left| U'(-\bar{\rho}) \right|^2 * \left| \Gamma(\bar{\rho}) \right|^2 \quad , \tag{13}$$

where U is the single pupil function, the primes indicate either Fresnel or Fraunhoffer diffraction, as appropriate, $\bar{\rho}_2$ is the two dimensional coordinate system at the target, z is the distance to the target, $\bar{\rho}$ is a time variant two dimensional shift parameter corresponding to the motion of the scanner, and * indicates convolution. As equation (13) is seen to yield a processed version of $|\Gamma|^2$, we thus recognize equation (13) as representing the mechanism by an image is incoherently processed by single pupil scanning. Unfortunately, this technique is severely limited with respect to possible processing functions due to the strictly real/positive nature of the processing kernel $|U'|$. In fact, the optical transfer function (OTF) of a single pupil scanning system has been shown to be [1]

$$OTF = \frac{\mathcal{F}\{v(\bar{\rho};\, z)\}}{\mathcal{F}\{|\Gamma(\bar{\rho})|^2\}} = U \otimes U \quad , \tag{14}$$

where $\mathcal{F}$ indicates the spatial two dimensional Fourier transformation and where $\otimes$ represents correlation. As the OTF of an incoherent imaging system is representative of

its spatial filtering characteristics, we see that a single pupil scanning image processor is inherently limited to low pass spatial filtering operations since the OTF is simply the auto-correlation of the non-diffracted pupil function U.

Notice now that equations (11) and (13) reduce to the same form if two identical pupils are chosen in the two pupil scanning processor arrangement of Figure 1. We are thus capable of performing incoherent low pass spatial filtering operations by either the single or two pupil techniques. The two pupil technique, however, has at least one distinct benefit. That is, since two temporally offset pupils are mixed prior to transmission in the two pupil technique, upon detection the processed image information is contained in an intermediate frequency signal. This, under most circumstances, will provide increased electrical signal to noise ratios for images processed by the two pupil method, as a result of the IF signal being isolated from low frequency environmental noise. In general, the two pupil technique will not, however, provide post detection signal to noise ratios as large as those possible in scanning processing systems which provide for mixing with an optical local oscillator (LO) beam just prior to detection.

Consider now a single pupil scanning image processor which, as just mentioned, provides for local oscillator mixing immediately prior to detection. For such a system the post detection electrical signal, represented as a time varying voltage, will be

$$v(\overline{\rho}, z; t) = \int_{-\infty}^{\infty}\int \left| A + U'(\overline{\rho}_2 - \overline{\rho}; z)\, \Gamma(\overline{\rho}_2)\, e^{-j2\pi f_c t} \right|^2 d\overline{\rho}_2 \quad , \qquad (15)$$

where A is the amplitude of an assumed uniform local oscillator beam, f_c is the drive frequency of an AOM used in an arrangement similar to that shown in Figure 1 and where, for convenience, we once again assume an arbitrarily large receiver aperture. The corresponding IF signal is then given as

$$\bar{v}(\bar{\rho}, z; t) = 2A\Re e\left[\iint_{-\infty}^{\infty} U'(\bar{\rho}_2 - \bar{\rho}; z)\, \Gamma(\bar{\rho}_2)\, d\bar{\rho}_2\, e^{-j2\pi f_c t}\right] \quad , \qquad (16)$$

while the complex phasor representation of the processed image is given as

$$\tilde{V}(\bar{\rho}; z) = (2A) \iint_{-\infty}^{\infty} U'(\bar{\rho}_2 - \bar{\rho}; z)\, \Gamma(\bar{\rho}_2)\, d\bar{\rho}_2 \quad , \qquad (17)$$

or

$$\tilde{V}(\bar{\rho}; z) = (2A)\, U'(-\bar{\rho}; z) * \Gamma(\bar{\rho}) \quad . \qquad (18)$$

Note the similarities and differences between this result and both equations (11) and (13). For single pupil processing with LO mixing just prior to detection (as opposed to the two pupil technique where mixing occurs during transmission), we see that optically coherent processing of the target image $\Gamma(\bar{\rho})$ is performed, and that in general the processing kernel U′ may be complex and/or bipolar. There is thus great flexibility in the coherent single pupil technique. However, due to the coherent nature of the processing function described by equation (17), images processed by coherent single pupil methods are easily corrupted by coherent spatial noise (e.g., speckle) [4,6]. This is of especially great concern in optical remote sensing applications where the effects of atmospheric turbulence may not be ignored. Thus for remote optical sensing applications, even though the coherent single pupil technique may provide for higher post detection electrical signal to noise ratios, the higher image quality likely achievable by the complex incoherent two pupil processing techniques would tend to make scanning two pupil image processing techniques quite attractive. Recent advances in

optical amplification of weak LADAR signals [7], coupled with the fact that the two pupil incoherent scanning method will in general provide intermediate electrical signal to noise ratios, higher than those achievable by direct detection techniques, will also contribute to the effectiveness and utility of incoherent two pupil scanning image processing.

CONCLUSION:

Incoherent optical image processing has many advantages, the primary among these being the capability of producing processed images with reduced coherent speckle artifact noise. Until now, though, only very limited processing functions could be implemented in incoherent scanning image processing arrangements. We have seen, however, that scanning image processors based on the interaction of two superposed and temporally offset scanning pupil functions are capable of implementing any incoherent complex/bipolar processing function. Though the post detection signal to noise ratios possible with two pupil systems (as described within) will likely be somewhat lower than those possible with processing arrangements providing for optical local oscillator mixing immediately prior to detection, recent advances in optical amplifier technology [7] will likely serve to make two pupil scanning techniques very attractive.

Current and near term continuing research will be centered arround the implementation of various image processing functions commonly used in computer based post detection image processing. The first processing function to be investigated will be edge extraction via convolution with the LOG (i.e., Laplacian-Of-the-Gausian) kernal. As this convolution kernal can be approximated optically by the difference of two collimated Gaussian beams with slightly different variances [5] it is ideally suited for implementation in the two pupil scanning system. Other processing functions of interest may include active turbulence correction, holographic recording, and matched filter correlation.

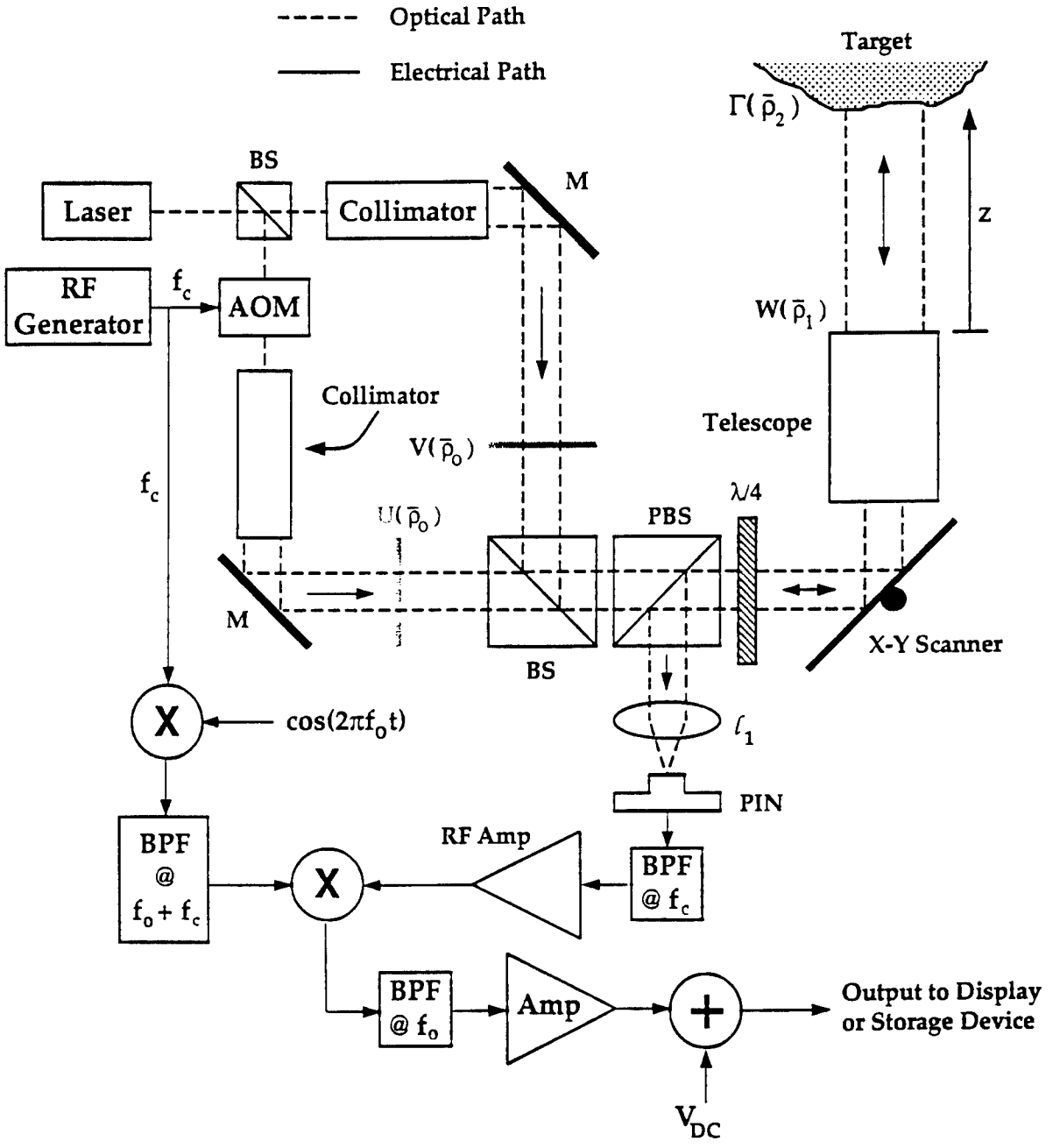


FIGURE 1: The Two Pupil Superheterodyne Incoherent Scanning Image Processor. This system is capable of producing complex processed images of remote targets via scanning with two temporally offset pupil functions. Notice that optical mixing takes place prior to transmission and that all processing effectively takes place in real time at the target.

REFERENCES

1) Ting-Chung Poon: "Scanning Holography and Two-Dimensional Image Processing by Acousto-Optic Two Pupil Synthesis," Journal of the Optical Society of America A, Vol. 2, pp. 521-527, April 1985.

2) Bradley D. Duncan and Ting-Chung Poon: "Gaussian Beam Analysis of Optical Scanning Holography," Journal of the Optical Society of America A, Vol. 9, No. 2, pp. 229-236, February 1992.

3) Bradley D. Duncan: "Real-Time Reconstruction of Scanned Optical Holograms Using an Electron Beam Addressed Spatial Light Modulator," Journal of Modern Optics, Vol. 39, No. 1, pp 63-80, 1992.

4) Francis T.S. Yu: Optical Information Processing, John Wiley & Sons, New York, 1983.

5) Robert J. Schalkoff: Digital Image Processing and Computer Vision, John Wiley and Sons, New York, 1989.

6) Joseph E. Goodman: Introduction to Fourier Optics, McGraw-Hill, New York, 1968.

7) Michael S. Salisbury, et. al.: "Signal to Noise Ratio Improvement of a One Micron Lidar System Incorporating an Optical Fiber Preamplifier," Optical Engineering, Special Issue on Pointing Tracking and Acquisition (accepted for publication) November, 1993.

APPLICATION OF GENETIC ALGORITHMS TO PATTERN THEORY

James F. Frenzel
Assistant Professor
Department of Electrical Engineering

University of Idaho
Moscow, Idaho 83844-1023
jfrenzel@groucho.mrc.uidaho.edu

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

July 1993

APPLICATION OF GENETIC ALGORITHMS TO PATTERN THEORY

James F. Frenzel
Assistant Professor
Department of Electrical Engineering
University of Idaho

Abstract

Pattern Theory is a robust technique for extracting "patterns" or "features" from a function and can be applied to many areas, such as machine learning, image processing, pattern recognition, and logic minimization. Recent efforts have focused on using pattern theory to decompose Boolean functions. However, the computational complexity of the method has currently limited its utility to functions of fewer than twenty variables. Consequently, we are interested in developing algorithms to quickly find variable partitions that lead to small decompositions.

Genetic Algorithms belong to a class of optimization methods which attempt to utilize the evolutionary mechanisms observed in Nature. Trial solutions to a particular problem are generated randomly to form an initial population. Copies of these solutions are made, proportional to the "quality" of the solution, or how well it solves the problem. Genetic operators, such as crossover and mutation, are then used to form a population of new solutions. The goal of this research was to assess the use of genetic algorithms for selecting variable partitions. A simple genetic algorithm is presented and the performance compared to other partition search techniques already in use. Finally, recommendations are made for topics of future research.

APPLICATION OF GENETIC ALGORITHMS TO PATTERN THEORY

James F. Frenzel

1 Introduction

Pattern Theory (PT) is a robust methodology for locating patterns in objects, represented as abstract mathematical functions. This representation provides a flexibility which allows the technique to be applied to many different types of items, such as images, data files, algorithms, and logic circuits. Developed at Wright Laboratory, Pattern Theory has found application in a diverse set of areas, including image recognition and enhancement, algorithm optimization, machine learning, and logic minimization.

At the heart of Pattern Theory is a method for breaking large functions into a set of smaller functions, referred to as *Ashenhurst-Curtis Decomposition* or simply *functional decomposition.* In the next section we will examine some of the important aspects of functional decomposition as they relate to this research. The interested reader is directed to the excellent tutorial developed by Ross et al. for a thorough treatment [14].

2 Problem Statement

While Pattern Theory has demonstrated its utility across many different types of problems, it continues to be limited in the size of such problems by the computational complexity of functional decomposition. For this reason, much of the logic synthesis community has long ignored functional decomposition, opting instead to pursue algebraic methods of logic minimization. However, the development of field programmable gate arrays and their popularity for rapid prototyping of digital systems has rekindled an interest. FPGAs typically have large arrays comprised of identical cells, each cell containing some type of circuitry for realizing small combinational logic functions. These cells are strictly limited in the size of the functions they can implement, typically a maximum of five variables, and techniques for

mapping digital circuits into this technology efficiently must be able to accommodate this restriction. Based upon papers presented at a recent logic synthesis workshop, functional decomposition is regarded as a promising candidate [9].

Let us now present a small example to illustrate some of the difficulties presented by functional decomposition. Assume we have an n-variable Boolean function, $f(x_1, x_2, \ldots, x_n)$, which we wish to implement. This function is said to have a *function cardinality* of 2^n. Depending upon the function, it may be possible to implement f using two smaller functions, such as

$$f(x_1, x_2, \ldots, x_n) = F[\Phi(y_1, y_2, \ldots, y_s), z_1, z_2, \ldots, z_r]$$

where $1 < s \leq n-1$, and the sets $Y = \{y_1, y_2, \ldots, y_s\}$ and $Z = \{z_1, z_2, \ldots, z_r\}$ are subsets of $X = \{x_1, x_2, \ldots, x_n\}$. Members of Y are referred to as *column* variables and Z is the set of *row* variables. If these two subsets are disjoint, such that $r = n - s$, the decomposition is said to be a *disjunctive decomposition*[1]. The total number of nontrivial, disjunctive decompositions is $2^n - n - 2$; if we allow variables to be *shared* between the two subsets, such that $Y \cap Z \neq \emptyset$, then there are on the order of 3^n different partitions. Each partition must be examined individually to determine if the decomposition is possible and cost effective. To make matters worse, it is generally desirable to continue the process, decomposing the functions F and Φ as well, commonly referred to as the *children*. In practice, this process can be continued iteratively until a predetermined stopping condition is reached or there is no further reduction in function size. The decomposed function cardinality (DFC) of a function is defined as the minimum sum of cardinalities over all subfunctions and corresponds to the implementation with the lowest complexity. A major obstacle hindering the application of PT to "real world" problems is the sheer number of partitions that need to be evaluated. Various strategies have been employed for locating "good" partitions, among them random selection and increasing row/column variable ratio [2]. The primary objective of this research was to evaluate the effectiveness of *genetic algorithms* (GAs) at searching the partition space. In the next section we will present a brief introduction to

[1]The term *disjoint* decomposition is frequently used.

genetic algorithms and discuss the particular methods which were used in this research. However, we will first identify related research being done by others.

Related Work

Although we did not find any publications describing the application of GAs to A-C functional decomposition, there are several researchers using GAs in the design of digital logic.

Rawlins Describes the use of GAs to design adders and parity checkers using primitive gates (AND, OR, etc.) [10].

Beasley A GA is presented which optimizes the design of quaternion multipliers by selecting connections which minimize the total number of individual multipliers [3].

Ghosh GAs are used to select state assignments which minimize the next state logic of finite state machines [1].

De Jong Application of GAs to the Boolean satisfiability problem. The GA produces minterms for arbitrary Boolean expressions [5].

Saab Presents results on applying GAs to VLSI test pattern generation. The results may be used to identify logic redundancy [15].

3 Methods

The mechanics of Pattern Theory and functional decomposition are currently realized in a software package called *FLASH*, implemented by the System Concepts Group of the Mission Avionics Division, Wright Laboratory [14]. Eventually, we would like to add a genetic algorithm-based search strategy to FLASH's toolbox of search strategies. However, given the time frame of this project we chose to use an existing GA package available from the University of California, San Diego [16]. This is a mature package with a large user base and supports many different options.

FLASH was run on 27 of the thirty, fully specified benchmark functions[2]. This set covers a variety of operations: parity, palindrome, addition, etc. Each is a function of 8 variables, constituting an initial cardinality of $2^8 = 256$. For each possible disjunctive decomposition, FLASH reports a figure of merit (FOM) which can be considered an estimate of the decomposition's complexity, or implementation cost. Two different FOMs produced by FLASH were used: sum of child cardinality (SOCC), and sum of grandchild cardinality (SOGCC); of these, SOGCC is the more accurate estimate of the DFC. These FOMs are used during the evaluation stage of the GA, described in the following section.

3.1 GA Basics

A genetic algorithm (GA) is an exploratory procedure that is often able to locate near-optimal solutions to complex problems. To do this, it maintains a set of trial solutions (called individuals), and forces them to "evolve" towards an acceptable solution. First, a representation for possible solutions must be developed. Then, starting with an initial random population and employing survival-of-the-fittest and exploitation of old knowledge in the gene pool, each generation should improve in its ability to solve the problem. This improvement is achieved through a four-step process involving evaluation, reproduction, recombination, and mutation. A pseudo-code implementation of a genetic algorithm is shown in Figure 1.

Representation Before applying a GA to any task, a computer compatible representation, or encoding, for possible solutions must be developed. These representations are referred to as *chromosomes*. The most common representation is a binary string, where sections of the string represent encoded parameters of the solution. For this research we represented a disjoint partition of the n variables using an n-bit binary string, where a '0' in the i-th position indicated that variable x_i was a column variable in the corresponding decomposition.

[2]Three of the functions do not decompose.

```
procedure GA;
begin
        initialize population;
        repeat
                evaluate population;
                reproduce;
                recombine;
                mutate;
        until {end condition}
end.
```

Figure 1: A Simple Genetic Algorithm

Evaluation The first step in each generation is the evaluation of the current population. This is the only step where the interpretation of the chromosome is used. Each chromosome in the population is decoded and evaluated as to how well it solves the problem; this result will be used in the next step to determine how many offspring are generated from any particular chromosome. For this research, partitions were evaluated by accessing an array where precomputed figures of merit (FOMs) were stored. The FOM was then scaled using *sigma scaling* to produce a *fitness value* [16]. Sigma scaling has been shown to maintain selection pressure, even as the population converges and becomes homogeneous.

Reproduction The next step in a generation creates a new population based upon the evaluation of the current one. For every chromosome in the current population, a number of identical copies are generated based upon the the chromosome's fitness, with the best chromosomes producing the most copies. This is the step that allows GAs to take advantage of a survival-of-the-fittest strategy.

There are several methods to calculate the number of offspring that each chromosome will be allocated. The two most popular methods are referred to as *ratioing* and *ranking*. For this research we used ratioing, under which each individual reproduces in proportion to its fitness. This has the advantage that as superior chromosomes emerge they can guide

the population quickly. The disadvantage is that if a superior individual surfaces early and dominates the population, then the population may potentially converge prematurely on a suboptimal solution.

Recombination The previous step, reproduction, creates a population whose members currently best solve the problem; however, many of the chromosomes are identical and none are different than those in the previous generation. Recombination combines chromosomes from the population and produces new chromosomes that maintain many of the features of the previous generation.

The most common method for recombination is one-point crossover. Two individuals are randomly selected from the population and, governed by a specified crossover probability, subsections of the two chromosomes are swapped about a randomly chosen crossover point. One-point crossover is said to exhibit a high *positional bias* because the probability that two bits, positioned far apart on the chromosome, will be passed together to a child is lower than that for two adjacent bits. At the other extreme is *uniform* crossover, where each bit has an equal probability of coming from a particular parent. In this work we used two-point crossover; the chromosome is treated as a ring and sections between the two crossover points are swapped to produce the children.

Mutation The last step in creating a new generation is motivated by the possibility that the initial population didn't contain all of the information necessary to solve the problem. Furthermore, it is possible that the individuals that produce no offspring may have had some information that is essential to the solution. The injection of new information into the population is called mutation. Here again, implementations vary but most simply randomly change a fixed number of bits every generation, based upon a specified mutation probability.

Summary of GA Parameters The following is a list of the particular features which characterize the genetic algorithm used in this research:

- n-bit, binary string representation;
- generational GA, as opposed to a steady-state GA[3];
- sigma scaling of the fitness function;
- proportional selection, as opposed to ranking;
- two-point crossover; and
- bit-flip mutation.

Elitism, "always evaluate children," nonrandom initial population, and super uniform initial population were also used where noted; these options are described in the GAucsd documentation [16].

3.2 Why Does it Work?

Genetic algorithms are based on two assumptions: one, that an individual's fitness is an accurate measure of its relative ability to solve the problem, and two, that combining individuals will enable the formation of improved offspring. If the first assumption is not correct, then it will be difficult for the GA to distinguish between good solutions and mediocre solutions. As a result, during reproduction both types of solutions will generate equal numbers of copies, slowing any movement towards improved solutions. Furthermore, if "good" solutions (as indicated by the fitness ranking) don't contain pieces of the best solution, it will be difficult for the GA to generate the best answer.

While the preceding discussion gives us an intuitive feel for why genetic algorithms work, John Holland developed a more formal analysis method using *schemata* (singular, schema) or, "similarity" templates. Let us assume we have a problem requiring an n-bit chromosome using a binary alphabet; thus, there are 2^n possible solutions. These solutions may be grouped according to different similarities. For example, "all solutions starting with the pattern 0110." One way of representing such a group would be using "wild card" symbols

[3] Described in Section 5.

(such as the asterisk) for positions where the particular chromosome value doesn't matter. Then the aforementioned group could be represented by the ternary string 0110***...*. This string is referred to as a *schema* and defines a set of chromosomes which match the template[4].

We can think of schemata as dividing up the solution space into regions, many of them overlapping or contained within other regions. For example, the schema 11**...* defines a region of the solution space which is contained completely within the region defined by 1***...*. In general, a schema containing m asterisks will define a region containing 2^m individuals. Looking at the problem differently, we see that a particular chromosome will match 2^n different schemata, each corresponding to a particular region. This chromosome can be thought of as sampling many regions of the solution space simultaneously. The average fitness value for individuals matching a particular schema is an indication of the quality of solutions that lie within the corresponding region. For a population of size p, there may be as many as $p2^n$ schemata represented, all being searched simultaneously. It is this implicit parallelism that makes genetic algorithms so powerful.

One question that has interested GA researchers is how the number and distribution of schemata sampled by the population changes from generation to generation. It has been shown that under ratioing, reproduction provides exponentially increasing numbers of individuals to above-average schemata and exponentially decreasing numbers to below-average schemata. But what is the effect of crossover? The probability that a given schema survives one-point crossover is proportional to its *defining length*, the distance between the first and last specified positions within the schema. For example, the schema 1*****0 with a defining length of 6 is more likely to become disrupted during one-point crossover than the schema ***01**. For the second schema to be destroyed, the crossover point must fall between the '0' and '1'; anywhere else and the schema will stay intact, independent of the other parent. Conversely, the first schema will always be disrupted, regardless of the crossover point, unless the other parent also matches the same schema. The effect of

[4]This set is often referred to as a *hyperplane*

reproduction and recombination is to create increasing numbers of short schemata with above-average fitness, often referred to as *building blocks*, which will eventually combine into superior solutions.

4 Results

Two different search methods will be used for comparison with this work: random selection and increasing row/column ratio. When possible, we will also compare our results with those obtained by Noviskey [13].

4.1 Alternative Partition Search Methods

Random Selection At the very least, we might ask the question "How does the performance compare to a random search?" Because the problems we have been working to date have a relatively small search space — 256 possible disjoint partitions for an eight-variable function — it is feasible to calculate the expected value of the best FOM if partitions are selected at random. If we know the probability that f_i was the best FOM found after t partition evaluations, then the expected value of $\mathrm{FOM}^t_{\mathrm{best}}$ is

$$\overline{\mathrm{FOM}^t_{\mathrm{best}}} = \sum_i f_i \Pr(f_i = \mathrm{FOM}^t_{\mathrm{best}})$$

Using the FOM histogram shown in Table 1 as an example, the probability that 4 is the best FOM after t trials is given as

$$\Pr(\mathrm{FOM}^t_{\mathrm{best}} = 4) = \Pr(\mathrm{FOM}^t_{\mathrm{best}} \geq 4) - \Pr(\mathrm{FOM}^t_{\mathrm{best}} > 4)$$

Assuming random selection with replacement, these probabilities are

$$\Pr(\mathrm{FOM}^t_{\mathrm{best}} \geq 4) = \left(\frac{7+5+3}{16}\right)^t$$

$$\Pr(\mathrm{FOM}^t_{\mathrm{best}} > 4) = \left(\frac{7+5}{16}\right)^t$$

yielding

$$\Pr(\mathrm{FOM}^t_{\mathrm{best}} = 4) = \left(\frac{15}{16}\right)^t - \left(\frac{12}{16}\right)^t$$

FOM	Quantity
16	7
8	5
4	3
2	1

Table 1: FOM Histogram

For random selection without replacement, the probabilities are

$$\Pr(\mathrm{FOM}^t_{\mathrm{best}} \geq 4) = \left(\frac{15}{16}\right)\left(\frac{14}{15}\right)\cdots\left(\frac{15-(t-1)}{16-(t-1)}\right) = \frac{15!(16-t)!}{16!(15-t)!}$$

$$\Pr(\mathrm{FOM}^t_{\mathrm{best}} > 4) = \left(\frac{12}{16}\right)\left(\frac{11}{15}\right)\cdots\left(\frac{12-(t-1)}{16-(t-1)}\right) = \frac{12!(16-t)!}{16!(12-t)!}$$

yielding

$$\Pr(\mathrm{FOM}^t_{\mathrm{best}} = 4) = \frac{(16-t)!}{16!}\left(\frac{15!}{(15-t)!} - \frac{12!}{(12-t)!}\right)$$

In practice, it would be computationally burdensome to perform random selection without replacement unless it was only done on a limited number of partitions, due to the overhead involved with keeping track of which partitions have been evaluated. Assuming that the number of partitions evaluated is small compared to the total number of partitions, the performance difference between the two should be very slight.

Increasing Row/Column Ratio Previous work by Axtell had demonstrated that searching partitions in the direction of an increasing row/column (IRC) variable ratio often led quickly to good partitions [2]. It is straight forward to use the FLASH output to plot the performance of an IRC search versus the number of partitions evaluated.

Noviskey's Results Preliminary work by Noviskey using SOCC data looked very promising [13]. Unfortunately, at the time of this writing only the performance after thirty evaluations is available. However, many of the techniques employed are different from those used here and warrant further investigation; these are discussed in Section 5.

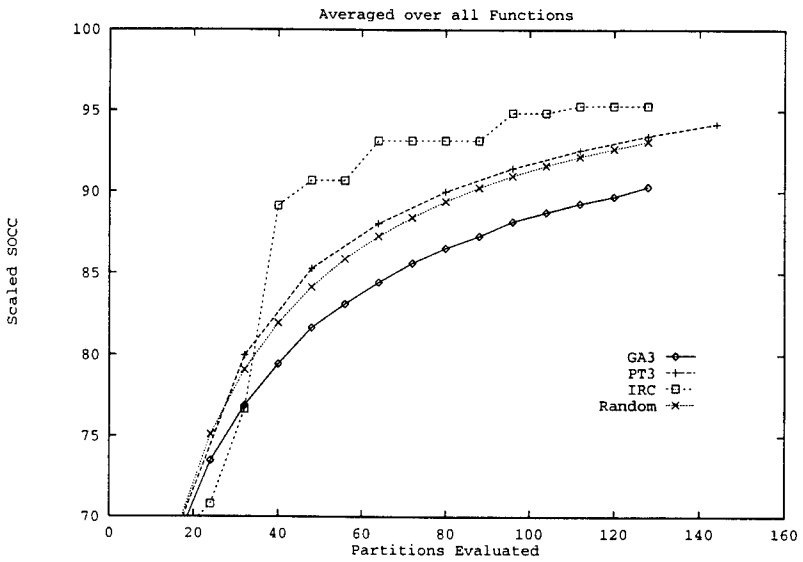


Figure 2: Performance using SOCC

4.2 Simulation Results

In this section we present the results of running the GA on the 27 FLASH benchmark functions using SOCC and SOGCC as the figure of merit. Because of the variation in decomposability between functions, we elected to present the results using a scaled, or normalized, FOM. The FOM returned by the GA was scaled by the function

$$f(\text{FOM}_{\text{GA}}) = (256 - \text{FOM}_{\text{GA}})/(256 - \text{FOM}_{\text{min}})$$

Note that FOM_{min} is not the DFC, but rather the minimum FOM found by FLASH. This presents the result as a percentage of how close the search method came to finding FOM_{min}; because this may be more important for decomposable functions, the results will be presented averaged over all functions and over functions with a DFC ≤ 64.

Figures 2 and 3 show the performance of the genetic algorithm against IRC and random search with replacement using the SOCC FOM, averaged over 200 experiments. GA3 is a genetic algorithm with the GAucsd parameters shown in Figure 4; PT3 is identical, except elitism is not used and a population size of 16 is maintained, with the initial population formed from the individuals shown. Figures 5 and 6 were produced under the same condi-

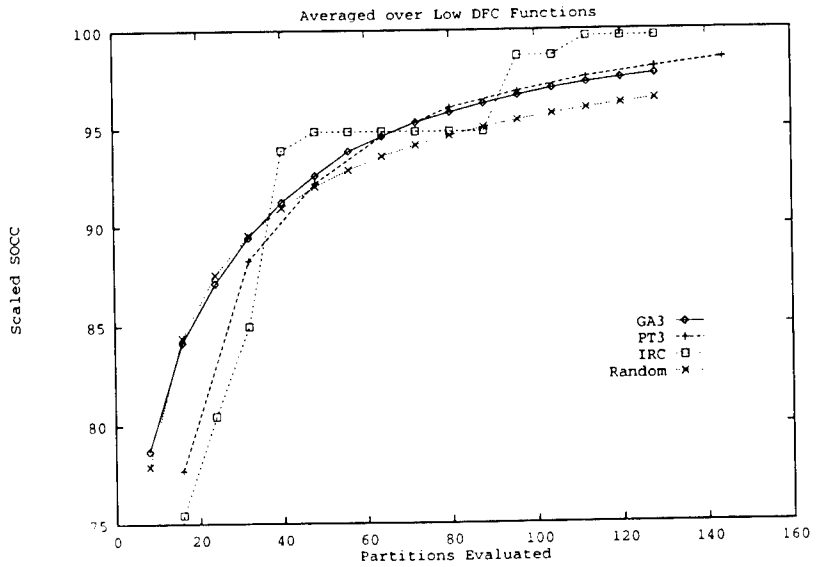


Figure 3: Performance using SOCC

Experiments =	200	10000000
Total Trials =	120	01000000
Population Size =	8	00100000
Structure Length =	8	00010000
Crossover Rate =	0.700000	00001000
Mutation Rate =	0.073000	00000100
Generation Gap =	1.000000	00000010
Scaling Window =	-1	00000001
Report Interval =	1	11000000
Structures Saved =	0	01100000
Max Gens w/o Eval =	0	00110000
Dump Interval =	0	00[illegible]000
Dumps Saved =	0	[illegible]100
Options =	aCeu	[illegible]000110
Random Seed =		00000011
Maximum Bias =	1.0	10000001
Max Convergence =	0	
Conv Threshold =	0	
DPE Time Constant =	0	
Sigma Scaling =	2.000000	

Figure 4: GA3/PT3 Parameters and PT3 Initial Population

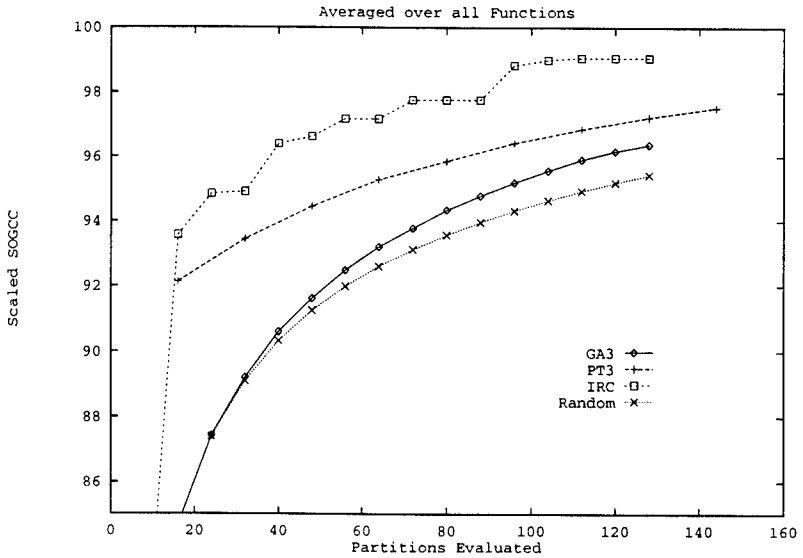


Figure 5: Performance using SOGCC

Performance after 30 Evaluations[a]								
		DFC	SOCC	Random	IRC	GA3	GA-NS	GA-NB
Unscaled FOM	All Functions	86.5	110.5	136.5	140.9	137.7	129.4	128.3
	Low DFC	32.4	48.0	70.4	78.9	71.0	58.1	63.0
Scaled FOM	All Functions		1.0	0.78	0.76	0.76	0.82	0.86
	Low DFC		1.0	0.89	0.85	0.89	0.95	0.93

[a]Results do not include function Bf11b.

Table 2: Performance using SOCC

tions, except that SOGCC was used as the FOM. Finally, Tables 2 and 3 show the results after 30 evaluations using SOCC and SOGCC as the FOM, respectively. The columns labeled "GA-NS" and "GA-NB" are results reported by Noviskey using his "structured search" (GA-NS) and his "baseline GA" (GA-NB) [13]. Again, the performance is averaged over 200 experiments and over the set of functions indicated.

We can make the following observations based upon these results: one, the IRC strategy is consistently better than random, GA3, and PT3; two, with the exception of the SOCC data averaged over all functions, GA3 and PT3 are superior to random. Furthermore, based upon the SOCC data, Noviskey's structured search looks very promising. In the final

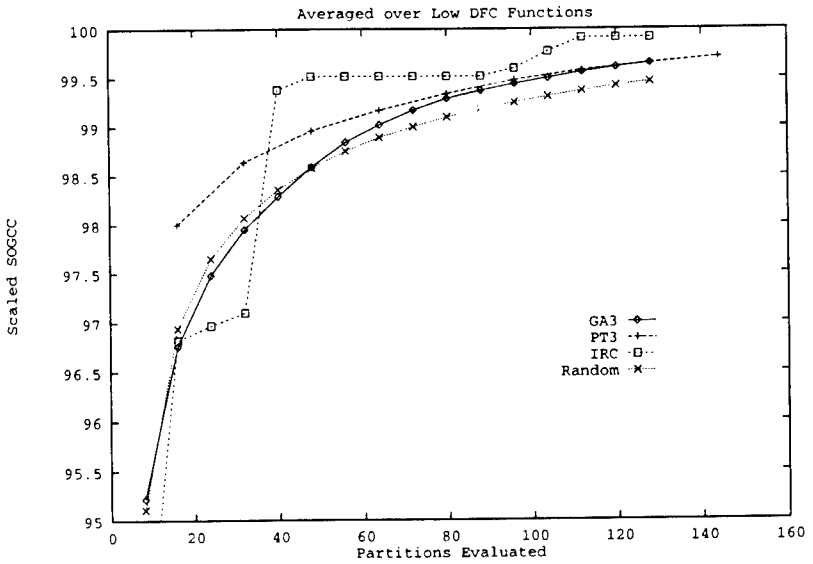


Figure 6: Performance using SOGCC

Performance after 30 Evaluations						
		DFC	SOGCC	Random	IRC	GA3
Unscaled FOM	All Functions	84.3	85.9	99.1	93.9	99.5
	Low DFC	32.0	33.3	37.6	39.5	37.8
Scaled FOM	All Functions		1.0	0.89	0.95	0.88
	Low DFC		1.0	0.98	0.97	0.98

Table 3: Performance using SOGCC

section we will attempt to offer some explanations for these results and propose topics for future research.

5 Conclusions and Future Research

There seem to be two "camps" when it comes to GAs: the generational group and the steady-state group. The generational approach is based on the model of a big pot of soup, slowly congealing, and the concerns are to keep stirring and not let it converge prematurely. The steady-state approach, on the other hand, is to maintain a stable base of good solutions and then add only one or two individuals every generation using highly disruptive or exploratory operators. The theory needed to decide which approach is better, when, and why,

is just now coming to fruition. However, some researchers believe that steady-state GAs may be more appropriate for problems with a high degree of epistasis, or gene interaction. It is worth noting that Noviskey's methods, particularly the structured search, are closer to a steady-state genetic algorithm.

There are several possible explanations for why a generational GA would not perform well:

Population Too Small: If the population is too small then there is often not enough diversity in the initial population to reach the global optimum, leading to premature convergence on suboptimal solutions. Another problem that can occur with a small population is a high sampling error of the schemata. If a particular schema is only represented by one or two individuals then it is unlikely that they will be an adequate representation of that schema's true average fitness.

Not Enough Generations: Thirty evaluations implies only fifteen crossover operations, which is generally insufficient to piece together the best solution. GAs are typically applied when there is no known direct search method and exhaustive search is infeasible.

GA Deceptive: Perhaps there is epistasis among genes, resulting in deceptive low-order schemata. This makes the performance sensitive to the representation, particularly when one-point crossover is used, because of it's high positional bias. Some researchers have opined that under this situation uniform crossover would be a better choice. Again, it is important to note that this is similar to Noviskey's uniform crossover operator[5].

5.1 Recommendations

In closing, we would like to identify several areas worthy of further exploration:

[5] Noviskey's operator was designed such that each parent contributed 50% of the alleles to each child.

- Attempt to explain the performance variations observed across the function suite. Do some functions display more epistasis? Are the low-order schemata deceptive? Determine if sampling error or premature convergence are a problem with some functions. Plot average fitness of partitions as a function of the Hamming distance from the global optimum [7, 6, 8].

- Attempt larger problems. Look at how the number of evaluations required to reach some level of performance increases with problem size for GA, IRC, and random selection. Perhaps the problem was too small to take advantage of the GA's implicit parallelism.

- Consider some of the operators and techniques used by Noviskey, such as uniform crossover, the complement operator, and rank-based selection [13]. Use a steady-state GA instead of a generational GA. Apply Noviskey's correlation work to a subset of the input space to seed an initial population or determine gene position [11].

- The improvement observed using the SOGCC data implies that the more accurate the estimate of partition quality, the better the results. Consider augmenting the FOM with additional information: column multiplicity, ratio of column types, number of minority elements in the columns. Noviskey's work on correlation and identifying row variables could also be used to differentiate between partitions with the same FOM [11, 12].

- Consider alternate representations. For epistatic problems, the performance is dependent on the ordering of genes. One possibility is to breed sets of variables and evaluate them twice: once as row variables and once as column variables[6]. A second possibility is Ross' suggestion to use a triangular matrix with real number entries from the interval $[0, 1]$. The entry (i, j) would provide a measure indicating if variables x_i and x_j should belong to the same set. This matrix would be modified using genetic operators similar to those already described.

[6]This would eliminate the need for Noviskey's complement operator.

- Speed-up partition evaluation. Consider using software profiling tools to identify the time-intensive portions of FLASH. Look at the work performed at Lund University and cited in their references for fast decomposition algorithms [9], also, Breen's work on determining column multiplicity from the compatibility matrix without forming a new partition [4].

6 Acknowledgements

I would like thank the "PT Team" and other members of the WL/AART-2 group for their encouragement and support; together they made the summer an enjoyable and rewarding one. Finally, I thank the AFOSR and the RDL for the opportunity to participate in the SFRP and their financial support.

References

[1] Jose Nelson Amaral et al. Applying genetic algorithms to the state assignment problem: A case study. In *SPIE Vol. 1706, Adaptive and Learning Systems*, pages 2–13, 1992.

[2] Mark L. Axtell. Partition selection algorithms: Row/column ratio experiment. Technical report, Veda Incorporated, c/o WL/AART-2, W/P AFB, OH 45433-7408, April 1993.

[3] David Beasley et al. Reducing epistasis in combinatorial problems by expansive coding. In *Proceedings of the Fifth International Conference on Genetic Algorithms*, 1993. *to appear.*

[4] Michael A. Breen. Some results in machine-learning. Final report, USAF AFOSR Summer Faculty Research Program, Tennessee Technological University, c/o WL/AART-2, W/P AFB, OH 45433-7408, August 1992.

[5] Kenneth De Jong. Genetic algorithms: A ten year perspective. In *Proceedings of the First International Conference on Genetic Algorithms and their Applications*, pages 169–177, 1985.

[6] Stephanie Forrest and Melanie Mitchell. What makes a problem hard for a genetic algorithm? Some anomalous results and their explanation. *to appear in* Machine Learning.

[7] David E. Goldberg. Genetic algorithms and Walsh functions: Part II, Deception and its analysis. *Complex Systems*, 3:153–171, 1989.

[8] John J. Grefenstette. Deception considered harmful. In L. Darrell Whitley, editor, *Foundations of Genetic Algorithms 2*. Morgan Kaufmann Publishers, 1992.

[9] Shousheng He and Mats Torkelson. Disjoint decomposition with partial vertex chart. In *Notes of the International Workshop on Logic Synthesis, Lake Tahoe, CA*, pages P2a-1–P2a-5, May 1993.

[10] Sushil J. Louis and Gregory J. E. Rawlins. Designer genetic algorithms: Genetic algorithms in structure design. In *Proceedings of the Fourth International Conference on Genetic Algorithms*, pages 53–60, 1991.

[11] Michael J. Noviskey. Correlation partition selection algorithm. Technical report, WL/AART-2, W/P AFB, OH 45433-7408, August 1992.

[12] Michael J. Noviskey. Row identification for function decomposition in pattern theory. Technical report, WL/AART-2, W/P AFB, OH 45433-7408, May 1993.

[13] Michael J. Noviskey et al. Application of genetic algorithms to function decomposition in pattern theory. Technical report, WL/AART-2, W/P AFB, OH 45433-7408, 1993. *in preparation.*

[14] Timothy D. Ross et al. Pattern theory: An engineering paradigm for algorithm design. Technical Report WL-TR-91-1060, WL/AART-2, Wright-Patterson Air Force Base, OH 45433-7408, July 1991.

[15] Daniel G. Saab et al. CRIS: A test cultivation program for sequential VLSI circuits. Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801, *in preparation.*

[16] Nicol N. Schraudolph and John J. Grefenstette. *A User's Guide to GAucsd 1.4*. Computer Science & Engineering Department, University of California, San Diego, La Jolla, CA 92093-0114, July 1992.

A Framework for Developing and Managing Reusable Avionics Software

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
Software Concepts Group (WL/AAF-3)
Avionics Directorate
Wright-Patterson Air Force Base
USAF Focal Point: Charles P. Satterthwaite

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

September 1993

A Framework for Developing and Managing Reusable Avionics Software

Raghava G. Gowda, Ph.D.
Assistant Professor
Department of Computer Science
University of Dayton

Abstract

Developing reusable software for avionics systems is a challenge to software engineers. The task involves foreseeing the future applications, modeling real-world events, using appropriate CASE tools throughout the development life cycle, and providing for storage and retrieval of reusable software components. This report presents a model for developing and managing reusable software components, briefly describes the software process maturity model and the integration of CASE tools with the process maturity model. It also identifies tools/ techniques and methodologies for real-time systems development, examines the critical issues in managing software projects, and offers the management a set of guidelines to introduce software engineering methodologies and CASE tools within the organization through a model project which may enforce standards for new projects.

A Framework for Developing and Managing Reusable Avionics Software

Raghava G. Gowda, Ph.D.

INTRODUCTION

It has been a great opportunity for me to work at the Wright Laboratory's Avionics Logistics Branch. During my stay, I had opportunities to do detailed literature survey on software reusability, go through some documentation on some of the projects, talk to consultants assessing software process maturity of the branch, interact with the software CASE tool vendors, talk to various project managers, and act as a consultant for one of the projects under development using object-oriented approach.

The initial objective of the summer research program was to study reusability issues in avionics software. The approach taken was to do :

A. A detailed library search in reusability and identify design criteria for reusable software
B. Gather design details about reuse efforts such as Reusable Ada Avionics Software Packages (RAASP) and Common Ada Missile Package (CAMP)
C. Summarize findings.

As it was not possible to get complete information on these projects, the research effort was directed to develop a framework for developing and controlling reusable software in the avionics domain. Based on my research, and participation in weekly review meetings of a simulation project, and my observations during my stay here, I would like to present a model for developing and managing any real-time software systems such as avionics systems. The model would be applicable for laboratories which initiate, fund, and control projects as well as to contractors who develop the systems.

Because the subject is broad, only a few aspects of it have been emphasized in this report. It is not possible to offer detailed discussions due to page constraints. After defining the reusability of software, a life cycle model is proposed which recognizes maintenance and further development as a part of the development process. This may be treated as a life cycle model of reusable software (or close to it). Then, the role of software maturity model and CASE tools in software development is discussed and suggestions are offered for integration of CASE tools acquisition and maturity model. A number of critical issues in managing software development tasks are outlined. Finally, suggestions are offered to develop a model project within an organization using appropriate methodologies and CASE tools. This will help an organization to attain higher levels in software process maturity model and offer insights for managing and integrating various projects. This experience may form a foundation for future reusable avionics software development tasks. The report consists of the following major topics:

1. Software Reusability
2. A Life Cycle Model for Reusable Software
3. Software Process Maturity Model for in-house management and external control.
4. CASE Tools
5. Software Process Maturity Model and Role of CASE Tools
6. Critical Issues in Managing Software Projects:
    - 6.1 Contents of Proposals and Deliverables
    - 6.2 Domain Analysis for Avionics and Assessment of Available Technology
    - 6.3 An Integrated View of all Projects
7. A Model For Developing and Managing Reusable Avionics Software
    - 7.1 A Model Project
    - 7.2 Software Development Processes, Methodologies, and Metrics
    - 7.3 Using CASE Tools in Projects
    - 7.4 Action Plan

1. SOFTWARE REUSABILITY

The reusability of software is the ultimate goal of any software development effort. The emphasis has shifted from project-specific systems analysis to domain analysis in order to incorporate flexibility in analysis and design so that the project-specific efforts can be reused in a broader domain. Developing reusable software for avionics systems is a challenge for software engineers. The task involves foreseeing future applications, modeling real-world events using appropriate tools, techniques, and methodologies for analysis and design of the system, and translating the specifications to software, verifying correctness of the software by testing, and providing for storage and retrieval of reusable software components. The software engineering discipline offers a number of methodologies, tools and techniques, and metrics to assist various phases of software development activities. The CASE (Computer Aided Software Engineering) technology integrates most of the tools and techniques.

Reusability has been defined differently by various authors. Kernighan [KER84] defines it as "...any way in which previously written software can be used for a new purpose or to avoid writing more software." Bott *et al.* [BOT86] give a more pragmatic definition of reusability: " a measure of the ease with which a component may be used in a variety of application contexts." In general, reusability of software can be interpreted from multiple view points. First, it can be seen as the use of existing software within a changing environment [SCH87]. Second, it can be viewed as the construction of new programs by composing such a program from software components [LYO86]. Third, it can be interpreted as the construction of programs by program transformation [CHE84].

Reuse can generally be classified as the reuse of ideas, vertical reuse, horizontal reuse, and total reuse [HAL87]. Publishing methods and techniques including algorithms leads to reuse of ideas by software professionals. Vertical reuse refers to reuse in a particular language, and horizontal reuse refers to the widely used

components in a particular environment, such as, sort utilities. Total reuse, however, is the use of complete packages after some customization.

The term reuse applies to the products developed throughout the development life cycle of software which includes requirements specifications, logical and physical design, code, and any information needed to create software. A Reuse Taxonomy by Prieto-Diaz [PRI93] shows six views of software reuse:

1. By-Substance defines the essence of the items to be reused. It consists of ideas, concepts, artifacts, components, procedures, and skills.
2. By-scope defines the form and extent of reuse. Such reuse can be classified as vertical and horizontal reuse.
3. By-mode defines how reuse is conducted. It may be planned, systematic or ad-hoc, opportunistic.
4. By-technique defines the approach used to implement reuse. It could be either compositional or generative.
5. By-intention defines how elements will be reused, as black-box (as is) or white-box (modified).
6. By-product defines what work products are reused. It includes source code, design, specifications, objects, text and architecture.

The broader concept of reuse (Basili, 1988) includes the 'use of everything associated with a software project including knowledge." The emphasis in industry has been on artifacts reuse such as Booch Ada Parts collection and the Generic Reusable Ada components for Engineers. Two of the often cited reusable projects in Avionics are Common Ada Missile Package project (CAMP) and Reusable Ada Avionics Software Packages (RAASP). The increased focus on software reusability is attributed to an overall realization of the potential benefits of not only reusing code, but also using all aspects of the development process. The documentation produced at various levels should serve as sources for maintenance and reusability. Reusability of software can not be a uniform process; it always has to be tailored for a particular domain.

1.1 ROLE OF USERS AND DEVELOPERS IN DEVELOPING REUSABLE SOFTWARE

Users of software can be classified as immediate users of the software for whom software was developed in the first place, and future users of the software. Immediate users are more concerned about the ability of the software in meeting requirements for the application, budget, and time factors. Their objectives, however, may not be the same as those of future users who tend to generalize software attributes, which in turn could lead to additional costs and delays in development efforts and enhancements. Immediate users may not favor domain analysis unless they realize the need for future developments and also its cost implications. Future users could play

a role similar to that of assembly line workers who assemble software components to produce a new product. The reuse of this same software, however, will vary depending on the ability of the user and the features of the software components.

The abilities of the users of software components are a combination of skill levels of the users, familiarity of the problem domain, and the time taken to retrieve required components from reuse repository. The tasks of the users will also be facilitated by the characteristics of the components themselves. These component characteristics are incorporated in the software by the software engineer and the development team.

It should be emphasized that the major players in software reusability are users and developers . A Reusable Components Repository is the common base through which they interact. Users are concerned with the ease of using the repository and developers are concerned with populating the repository and keeping track of engineering details.

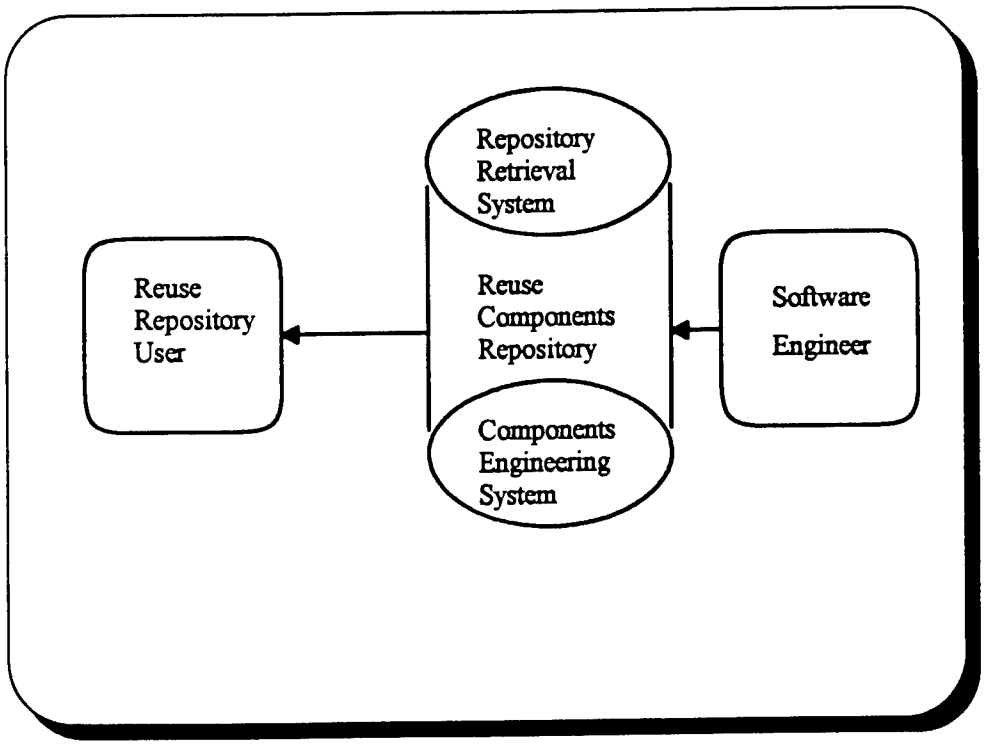


Figure 1. *Reusability Scenario*

The reusability scenario has two distinct features: a Repository Retrieval System and the Components Engineering System which consists of domain analysis, requirements analysis, design, metrics, and other configuration management details. Both the Repository Retrieval System and Components Engineering System have to be designed in collaboration with each other. A Repository Retrieval System is similar to a library retrieval system, but it has to consider graphical interfaces and capabilities to interface with other software repositories planned by DoD and commercial software vendors. Even though reusable efforts are evolving, standards for interface with repositories are not easily available. Therefore, retrieval systems need flexible designs.

Extracting reusable components from existing software is a complex activity and may not be cost effective in all cases. The basic information needed for reusable components from the users' perspective are:

a. Domain knowledge
b. Overview of reusable components
c. Details of each component which include:
    Source code adhering to a particular style
    Domain analysis
    Requirements analysis of the system for which the component was designed
    Logical Design
    System (hardware/software) constraints
    Testing / Quality Assurance reports
    Unique features etc.
    Other aspects.....

2. A LIFE CYCLE MODEL FOR REUSABLE SOFTWARE

Software Development Life Cycles offer a framework for software development. Though there is some concern about the life cycle concept itself amid its use, one has to admit their utility in identifying various tasks involved in the software development. Software professionals are well-versed with the following life cycle models:

1. Classical Software Development Life Cycle
2. Structured Software Development Life Cycle
3. Software Engineering Life Cycle
4. Information Engineering Life Cycle
5. Spiral Model of software development and enhancement

The life cycle models present different points of view for systems development. For example, the classical model assumes that the various activities are sequential. This model has been widely critiqued for its inability to

incorporate changing system needs. The *Structured Life Cycle* considers back-tracking and incorporating changes introduced at various phases. The *Software Engineering Life Cycle* closely follows the Classical and Structured life cycles and emphasizes walkthroughs, inspections, reuse, metrics, and deliverables. *Information Engineering Life Cycle* emphasizes defining data requirements of an enterprise first, and then the processes. It may be more appropriate for developing Management Information Systems projects. The *Spiral Life Cycle* deals with the prototype development environment, where neither the developer, nor the user have complete knowledge of the product to be developed. It considers risk analyses for different prototype developments. None of the above models treat software reuse or maintenance as a part of the models. Incorporating maintenance and reuse paradigm in the model forces the developer to consider a futuristic view of the system during development.

Henderson-Sellers and Edwards [HEND90] propose the **Fountain Model** (Figure 2) for the object-oriented life cycle which consists of the traditional life cycle phases and three additional phases of *Program Use, Maintenance,* and *Further Development* at the top of the Fountain. The model extends the Waterfall or Structured Life Cycle models to include reuse, maintenance, and further enhancements of software. The Fountain Model represents object-oriented software life cycle. It may also be treated as reusable software development life cycle, where *Maintenance*, and *Further Development* are some aspects of reuse efforts. We would like to add Domain Analysis prior to Requirements Analysis phase in the Figure 2 to emphasize the fact that reusability issues should be domain specific, because designing software for universal reusability may not be cost-effective and practical.

Domain analysis plays a significant role in the development of reusable software. Domain analysis refers to a detailed survey of the past efforts of an application area, current development tasks, and future needs of the problem domain. Domain analysis will allow the managers and developers to integrate past and current efforts, schedule for the optimum utilization of resources, and substantial savings in time and efforts in development, maintenance, and enhancements of systems.

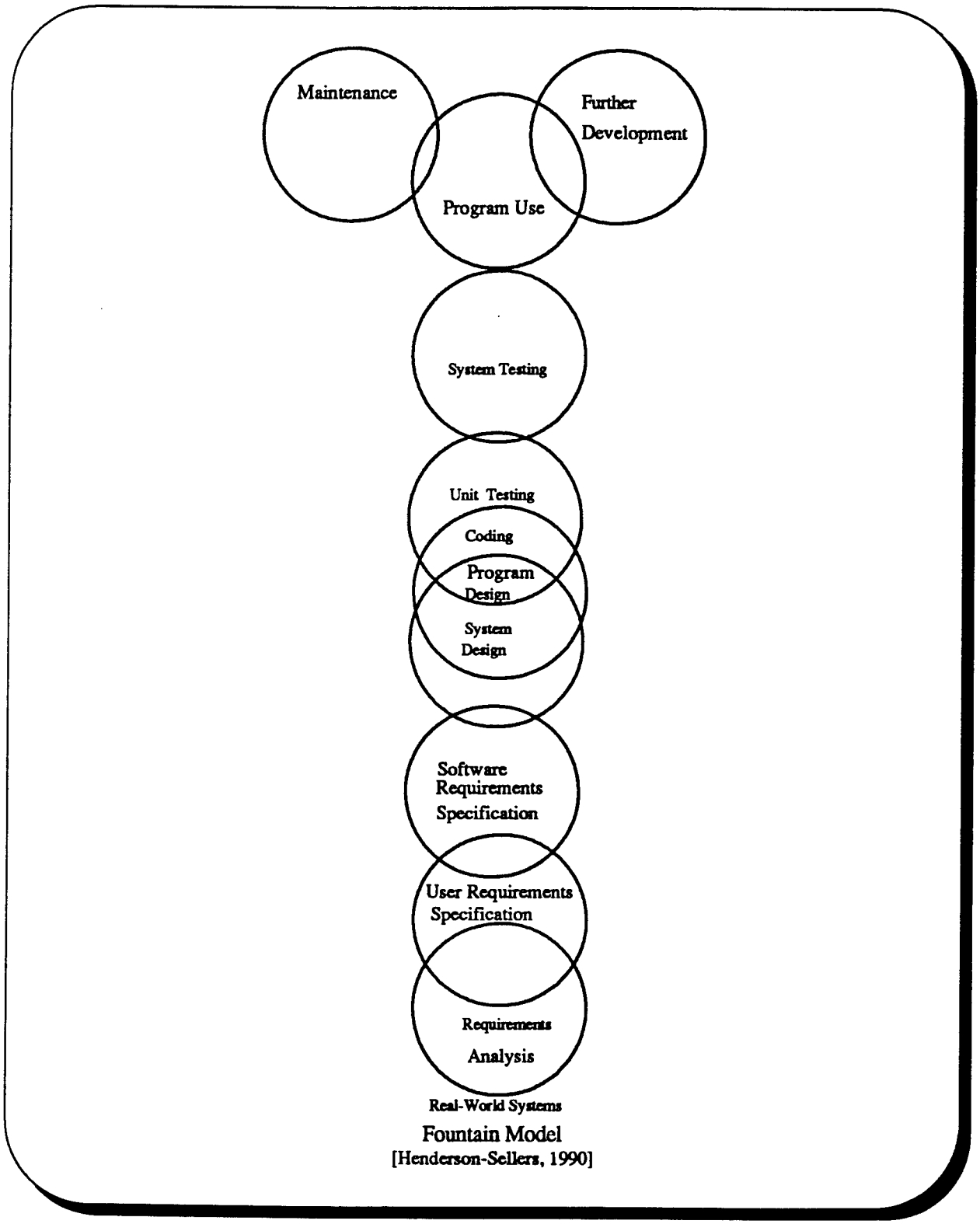


Figure 2. *Fountain Model*

3. SOFTWARE PROCESS MATURITY MODEL FOR IN-HOUSE MANAGEMENT AND EXTERNAL CONTROL

The basic elements of a management process are: planning, organizing, operating and control. For an organization such as Software Concepts Group at the Avionics Directorate of the Wright-Patterson Air Force Base, which forecasts technology for future avionics systems, translates the concepts to feasible projects, funds and monitors the projects, and diffuses the technology to its customers, and to the community at large, management issues become very complex. We can view management functions from two perspectives:

1. Internal control
2. Controlling projects developed by contractors.

The task of foreseeing the future technology is the most crucial one. As the projects are of diverse nature, it may not be possible to use the same set of tools/techniques or CASE tools in all the projects. However, a few guidelines could be offered to maintain uniformity among projects. It is necessary to:

a. Provide a common format for all projects;
b. Maintain control over all documentation and source codes;
c. Record contributions of individual projects with respect to contribution to knowledge, technology, products, etc. ;
d. Encourage use of software engineering principles and CASE tools wherever appropriate;
e. Integrate contributions of all the projects and practice them in the subsequent projects.

The Software Process Maturity Model [HUM89] would be equally applicable to Laboratories controlling projects as well as to contractors involved in software development. If the Laboratories have a high level of process maturity then they will be in a position to demand and control appropriate deliverables from their contractors more effectively. The software process is the entire set of tools, methods, and practices used to produce a software product.

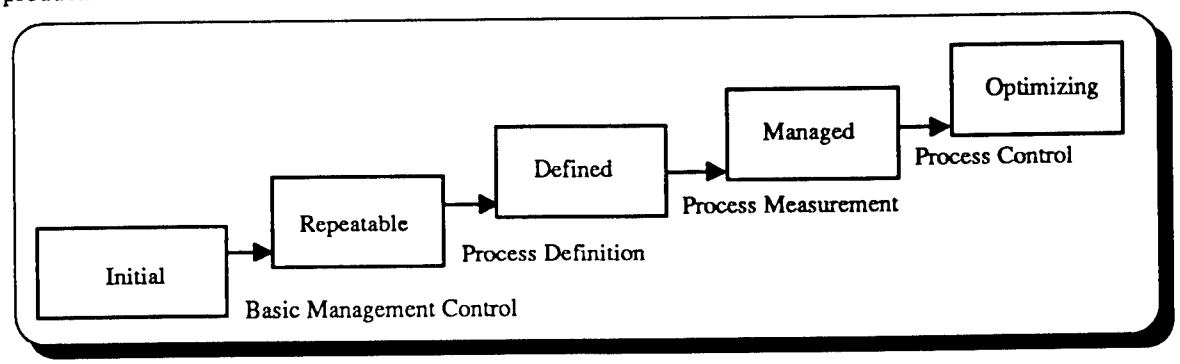


Figure 3. *Software Process Maturity Levels*

Control and improvement of tools, methods, and practices can lead the organization to higher maturity levels. Metrics or measurement of outputs can play a major role in translating the art of software development to the science of software development. CASE tools will also play a major role in the software process maturity model.

4. CASE TOOLS

This section outlines the specifics of the avionics software development process. The demands of avionics software development process are quite distinct from those of Management Information Systems or any other commercial applications. Most of the avionics applications are embedded real-time systems. The new applications are to be developed using Ada. Some of the tools/techniques and methodologies required to specify the requirements and design of the avionics systems are as follows:

A. Structured Analysis and Design Methodology

Tools/Techniques Used:

1. Data Flow Diagrams and its Real-Time extensions
2. Data Dictionary
3. Data Modeling and Entity Relationship diagrams
4. Decision Tables, Decision Trees, Action-Diagrams, Nassi-Shneiderman charts
5. State Transition Tables, State-Transition Diagrams
6. Finite-State Architecture
7. Petrinets
8. Structure Charts (in Structured Design Methodology)

B. Data Structure-Oriented Methodologies

1. Warnier/Orr Methodology
2. Jackson System Development (JSD)

C. Real-Time Design Methodologies

1. Design Method for Real-Time Systems (DARTS)
2. Structured Analysis and Design Technique (SADT)

D. Object-Oriented Analysis and Design Methodologies:

1. Bailin: Object-Oriented Requirements Specification
2. Coad and Yourdon: Object-Oriented Analysis and Object-Oriented Design
3. Shlaer and Mellor: Object-Oriented Analysis
4. Wasserman et al. Object-Oriented Structured Design (OOSD)
5. Booch: Object-Oriented Design
6. Wirfs-Brock et al. Responsibility Driven Design (RDD)

7. Rumbaugh et al. Object-Modeling Technique (OMT)
8. Embley et al. Object-Oriented Systems Analysis (OSA)

These methodologies can be implemented in most of the leading CASE tools. In general, the CASE tools can be classified from multiple view points such as upper CASE, lower CASE, and I-CASE (Integrated CASE Tools). Some of the unique components of the CASE Tool set are [BUR 89] summarized below:

a. Diagramming Tools
b. Syntax Verifiers
c. Prototyping Tools
d. Code Generators
e. Project Management and Methodology Support
f. Re-Engineering Tools
g. Central Repository

At present about 24% of the software development organizations use CASE tools. Most of the tools implement Structured Systems Analysis and Design methodologies, but they are not well integrated with code generation and testing tools. One of the major problems is the compatibility among CASE tools. Lack of industry standards makes it difficult to port systems from one set of tools to another. It also makes users dependent on a particular tool vendor. Organizations generally are hesitant to adopt this new technology because of high cost investments in these tools as well as the costs involved in training employees in the methodologies and tools. Moreover the management of an organization has high expectations about pay-off from these tools but employees are reluctant to experiment with the new technology. The opponents of CASE tools may be concerned about the following issues:

1. Training with methodologies
2. Ease of use of CASE Tools
3. Redundancies in documentation
4. Poor integration of various phases of life cycle
5. High costs of acquisition and maintenance
6. Probability and compatibility of the tools

These are some of the concerns. It should be kept in mind that CASE technology is still evolving and it might take a while to have an I-CASE in its real sense. If an organization ready to cope with the technology, it is the time to start now! We have to keep in mind that acquiring knowledge always comes incrementally. Training employees with software engineering methodologies is clearly the most crucial aspect. A careful evaluation of the tools that will be needed should be done prior to their acquisition. Consultations with users of some of the tools in similar projects seems to be the most desired method for evaluation of the tools. It would be an ideal approach,

however, to train project managers on a **Model Project** which would be of common interest. Various software engineering tools/techniques can be experimented in the Model Project. For example, an expert in the methodology may act as a moderator in the development process. In learning the development process education in software engineering disciplines would be more beneficial as compared to training with a specific CASE tool because it is not the tool which decides the success of any project but the modeling or problem solving approach which dictates how well the tool is used to develop the system.

Most CASE tools are capable of producing documents adhering to 2167A standards. However, it is the responsibility of the management to enforce uniformity in the contents of deliverables by giving specific directions regarding formats, tools/techniques, methods, and cross-referencing in the documentation. The ultimate objective is always to get an integrated product which is easy to operate, maintain, and reuse. This could be accomplished by enforcing standards with respect to the tools, techniques, and other process details.

5. SOFTWARE PROCESS MATURITY MODEL AND ROLE OF CASE TOOLS

Pfleeger [PFL91] suggests that "only when development process possesses sufficient structure and procedures does it make sense to incorporate certain kind of CASE tools in a development environment." He advocated the following CASE tools for the various process maturity levels.

	Level	Characteristics	Metric to use	CASE Tools
1.	Initial	Adhoc	Baseline	Tools that help to structure and control, estimate product size and effort
2.	Repeatable	Process dependent on individuals	Project	Tools for requirements specification, project management, and configuratiom management
3.	Defined	Process defined, institutionalized	Product	Tools to measure quality, complexity, to support design and coding, and to guide testing and integration
4.	Managed	Measured process (quantitative)	Process + Feedback	Project database, management system, simulation tools, reliability models, and impact analysis
5.	Optimizing	Imrovement fedback to process	Process + Feedback for changing process	Process programming and process simulation tools

Figure 4. *Process maturity levels related to CASE tools*

6. CRITICAL ISSUES IN MANAGING SOFTWARE PROJECTS

Issues involved in software project management are numerous and quite complex. Some of the key issues which have tremendous impact on deliverables, usability, and control of the projects are the following:

6.1 CONTENTS OF PROPOSALS AND DELIVERABLES

After going through a number of proposals, work plans, deliverables, etc. I realized the difficulties in managing the projects. Even with my adequate background in the software engineering area, I felt that I would be at the mercy of the contractor if I were a manager. The reason is that I may not have enough control over the projects. The DoD has all the controls in place for managing the projects through standards such as 2167A, but the documentation I read raised the following questions which could serve as guidelines for an organization:

1. What are the contributions of this project in terms of knowledge and technology?
2. Have similar projects been undertaken by the same contractor, by other agencies, or by other organizations in U.S. or abroad?
3. How well the contractor completed the "Related Work" section in the proposal?
4. Can we have full control over the deliverables?
5. Does the contractor make us depend on the particular contractor for related efforts in future by the nature of deliverables, or mode of information sharing?
6. How well the requirements specification is separated from design issues?
7. Has the continuity among various phases of systems development been demonstrated by the contractor?
8. How well the project is integrated with other projects in the Avionics Directorate?
9. How easy is it to follow documentation and other deliverables?
10. Do we have metrics to assess deliverables?

6.1.1 SOURCES OF CONFUSION

I strongly believe that one of the major sources of confusion can be in the proposal itself, in which the contractor fails to distinguish between the requirements of a system and the design of the system. Systems analysis or requirements analysis should emphasize only on "what" is expected of the system, and not "how" it is to be done. Unfortunately most of the proposals I read did not distinguish between analysis and design. This lack of distinction blurs the basic objectives of the project. Additionally, it is difficult to evaluate the approach as no alternatives are given. The proposal may be too detailed about a specific mode of implementation so that the reader is lost in the details or led to believe that it is the only way to do things. In such a scenario the manager has no option but to accept the tasks as outlined in the proposal. The manager will not have much control over the project as the proposal did not specify microscopic details. Only an expert in the particular area may raise

meaningful questions. The other alternative which I may be forced to adopt because of the nature of proposal, is to keep quiet and be satisfied with whatever deliverables are received from the contractor. To overcome these difficulties, I would ask the contractor to strictly follow the following guidelines:

1. Specify the system requirements without referring to any hardware or product details.
2. Discuss implementation details in the Design aspects.
3. Do not mix "What" and "How" aspects.
4. Describe contents of deliverables. Especially, identify what type of methodologies, CASE tools etc. would be used for analysis and design. Describe how documentation of one phase will lead to the successive phases.

6.1.2 TECHNOLOGY VS. APPLICATIONS

Investment in research projects related to technology development is potentially a high return area if it is conceived and managed very carefully. Otherwise investments may not produce substantial outputs. Every effort should be made to diffuse the technology to the community at large and break the monopoly of a few vendors. In avionics domain, the main thrust is both in software development and technological advancement. In both the cases, outputs of the projects need to be carefully evaluated and controlled.

6.1.3 MONOPOLY OF SOFTWARE DEVELOPERS

In the 1960's Original Equipment Manufacturers (OEM) controlled the hardware industry because of their unique coding schemes. This approach made the user dependent on a particular vendor of hardware. Introduction of ASC II solved this problem to some extent. Now a similar scenario exists today in the software industry. Though there are standards such as 2167A for documentation of software, the contents of documentation vary because of different tools and techniques used in software development, and the lack of detailed standards and process control.

In the absence of matured processes for software development, a reasonable control on the project can be achieved by defining the *contents* of deliverables. To some extent the contents would define the process and enforce uniformity in managing projects. One of the aspects which need to be emphasized, however, is to separate logical requirements from implementation details. This is the key to extend the life of the system.

6.2 DOMAIN ANALYSIS FOR AVIONICS AND ASSESSMENT OF AVAILABLE TECHNOLOGY

Domain Analysis is the formulation of the common elements and structure of a domain of applications []. It is a way of understanding and describing the past and present in order to increase productivity in the future. It

should be an ongoing process. Domain Analysis in avionics refers to surveying the avionics applications and assesses the role of the current and past projects in the avionics domain applications. One of the objectives of domain analysis is to develop a framework for integrating all tasks in avionics domain. Some of the questions which may be useful in this context are:

a. What is the knowledge base for major air frames for example, (current and future fighters, bombers, helicopters, and trainers) regarding their current functions, missions, environments and future plans?
b. What are the goals and objectives of this group with respect to software and hardware under its control?
c. Do we know the details of all current projects, and deliverables within a suitable format?
d. Do the current projects meet goals and objectives? To what extent?
e. What applications are better for reusability, in terms of :
    i) analysis
    ii) design
    iii) code
    iv) algorithms
f. Are projects aimed to develop prototypes and concepts documenting the outputs in such a manner that they could be used/implemented by any independent team?
g. Which projects undertaken by DoD, ARPA, NSF, other agencies, and private organizations are similar to the current one? Describe them.
h. Is a format for domain analysis documents developed?

6.3 AN INTEGRATED VIEW OF ALL PROJECTS

Each project should have documentation on various phases. DoD Standard 2167A deals with the details and these should be followed. For the purposes of managing projects, the Avionics Directorate may develop a document giving an integrated view of the projects under its control. A yearly report may be compiled which will consist of the following essential aspects of all the projects and how they are integrated:

1. Domain Analysis
2. Contributions of the Project
3. How these contributions are integrated with other Avionics projects or any other efforts?

The major emphasis of this yearly report should be on concepts on technology advances and on aspects which could be used elsewhere. Implementation details are available at the project level and as such should not be part of this report.

7. A MODEL FOR DEVELOPING AND MANAGING REUSABLE AVIONICS SOFTWARE

Based on the past discussions, a Model for Developing and Managing Reusable Avionics Software is offered in Figure 5.

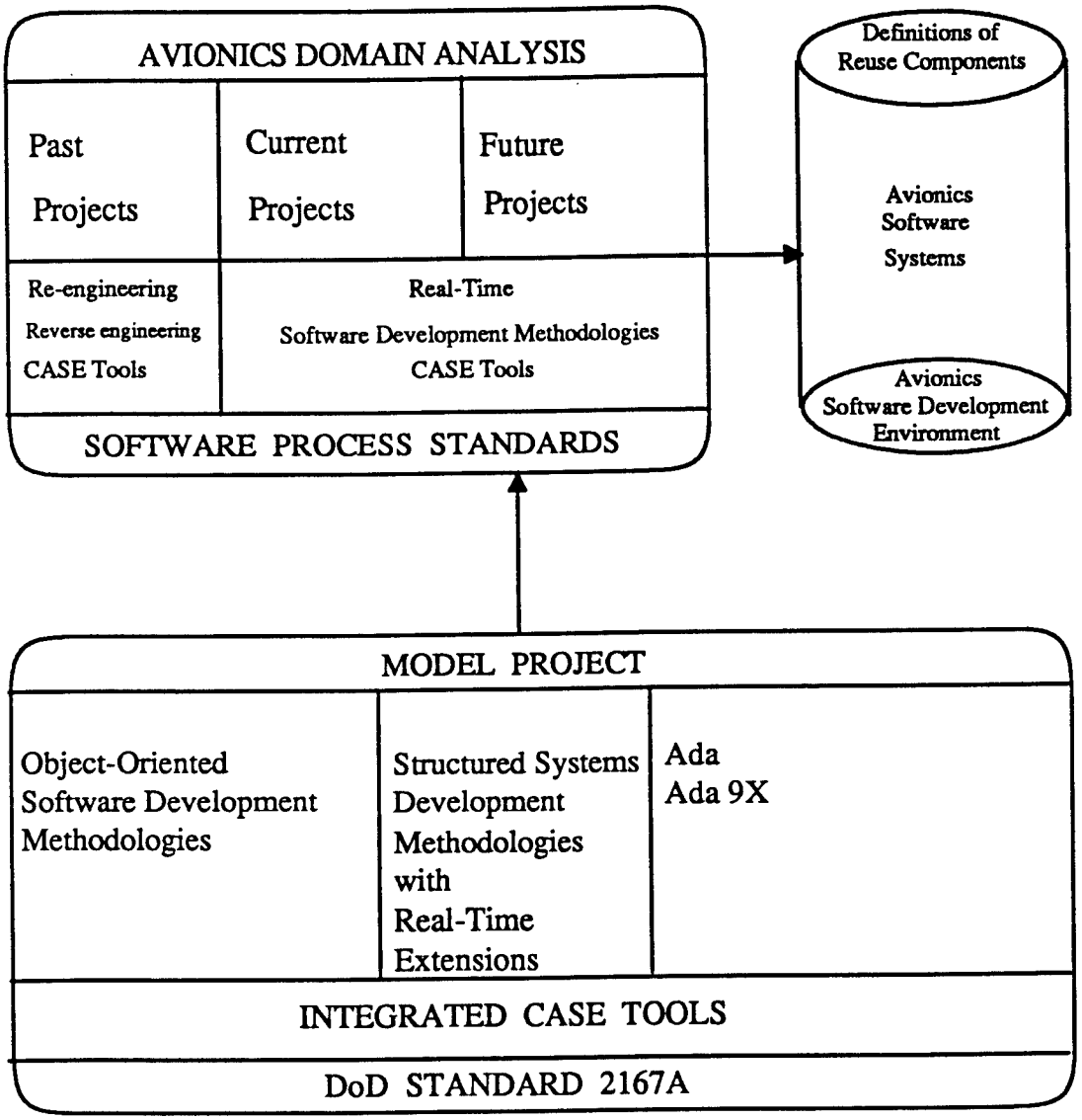


Figure 5. *A Model For Developing and Managing Reusable Avionics Software*

7.1 A MODEL PROJECT

To standardize the project formats and to appreciate technical details, it will be useful to work on a Model Project with close collaborations with a developer or consultant. It may be useful to re-engineer an existing project which will have high reusability because the existing project will have all technical details. If, however, a new project is selected the emphasis will shift to new technical details and exploration aspects. Therefore, a new project may be restrictive in exploring all tools/techniques, methodologies, and other issues. The difficulty in re-engineering an existing project would depend on the availability of documents, style of coding, and the language used.

The Model Project should be developed using appropriate methodologies, CASE tools, and documentation should adhere to standard 2167A . It should serve as a model for use of methodologies and documentation. This experience may be valuable for in-house development and in managing software development projects by contractors.

7.2 SOFTWARE DEVELOPMENT PROCESS, METHODOLOGIES, AND METRICS

Avionics software systems have a number of characteristics which make them unique compared to the traditional data processing application. In general, they are real-time systems which need to activate as a response to external events. Response time is critical for such systems. Some of them may be embedded where software is integrated with hardware components. Correctness of software, and reliability are of utmost concern to users of the software. Therefore, testing of software is a very elaborate task in its development. The system needs to be maintainable at the earliest possible time. On-board testing and maintenance efforts would minimize the machine-down-time. Most of the tools and techniques identified earlier would also be applicable to avionics software development.

Object-oriented methodologies are very well-suited for developing reusable software. The concepts of abstraction, information-hiding and polymorphism are practiced in these methodologies. Though there are a number of OOA, OOD, OOA and OOD methodologies, they are not in a matured stage such as Structured Methodologies. The Object Modeling Technique by Rumbaugh et al. [RUM91] seems to be one of the most appropriate methodologies for avionics system development as it offers Object Model, Dynamic Model, and Functional Model facilities for requirements specification of a system. It is a nice blend of structured and object-oriented concepts. The methodology may need further refinements to provide continuity among the three models and to represent collaboration between various subsystems and objects. A CASE Tool called OMTool™ implements the methodology. However, it is to be kept in mind that at present OMTool™ does not implement the Dynamic Model and Function Models of Rumbaugh's methodology. Representation of collaborations between

objects and integration of different models are the major issues regarding this methodology. The Model Project may use Object-Oriented and Structured methodologies in analysis and design phases and Ada in implementation phase. Ada 9X incorporates object-oriented programming features.

7.3 USING CASE TOOLS IN PROJECTS

At present most of the CASE tools used are in analysis and design areas. The analysis and design methodologies are not well integrated. CASE tools related to project management, code generation, and testing should also be used wherever applicable. This would integrate all the phases of life cycle. A recent report from Institute for Defense Analysis has identified that there are more than 600 testing tools available but not widely used. The CASE technology is heading towards I-CASE or Integrated CASE. Management may introduce these tools progressively in projects. As discussed earlier, acquiring different tools may be linked to different levels of process maturity.

7.4 ACTION PLAN

The action-plan may emphasize standardizing processes for a project. A participative mode of implementing a change in an organization may face less resistance. The manager may do the following for instance:

1. State the need for a Domain Analysis for Avionics to identify the future technological needs.
2. Show how the existing projects meet the needs. Emphasis should be on specifics. Global words should be avoided.
3. Identify goals for the next five years.
4. Draw an implementation plan which covers:
    a. A suggested format for summary and detailed internal reports for projects
    b. Details of deliverables from contractors
    c. Software Engineering tools/techniques to be used
    d. Identify the need for a Model Project
    e. Invite suggestions from Project Engineers
    g. Incorporate suggestions and implement the plan.

One of the difficult tasks would be to bring projects with multiple dimensions under a common thread of control. This has to be done by studying individual projects and then comparing their similarities and differences. The model project may promote uniformity in the processes i.e. in tools, methods and practices which would lead the organization to a higher level of process maturity and force the contractors to adhere to higher standards.

REFERENCES

[BAK90] Baker,T.P. Software Reuse in Real-Time Environments. Report submitted for U.S. Army HQ CECOM, Center for Software Engineering, October 9, 1990.

[BAR93] Barnes, J. Introducing Ada 9x. Ada 9x Project Report, Office of the Under Secretary of Defense for Acquisition, Washington DC, 20301, 1993.

[BUR89] Burkhard, Donald L. Implementing CASE Tools. ASM Journal of Systems Management, May 20, 1989.

[BOT86] Bott, M. F., A. Elliott and R.J. Gautier. Ada Reuse Guidelines. - Report, ECLIPSE/REUSE/DST/ADA_GUIDE/RP, Alvey ECLIPSE Project Deliverable D36, February 1986 Software Sciences Ltd.

[CAT91] Cattel, R.G.G. Object Data Management. Addison-Wesley, Reading, Mass., 1991.

[CHA92] Champeaux, Dennis de., and Penelope Faure. A Comparative Study of Object-Oriented Analysis Methods. JOOP Journal of Object-Oriented Programming, March/April 1992.

[CHE84] Cheatham, T. Reusability through Program Transformation. IEEE Transactions on Software Engineering, V 10 (5), pp. 589-594, September, 1984.

[HAL87] Hall, Patrick A. Software Components and reuse - getting more out of your code. In Will Tracz (Ed.) Tutorial: Software Reuse: Emerging Technology, The Computer Society of the IEEE, 1988.

[HEN90] Henderson-Sellers, B. and Edwards, J.M. The Object-Oriented Systems Life Cycle. Communications of the ACM, September, 1990.

[HOL90] Holmgren, Brian W. Software Reusability: A study of why software reuse has not developed into a viable practice in the Department of Defense. Thesis submitted to Air Force Institute of Technology, December 20, 1990.

[KER84] Kernighan, B.W. The UNIX System and Software Reusability. In IEEE Transactions on Software Engineering, pp. 513-518, 1984.

[HUM89] Humphrey, Watts S. Managing the Software Process. Addison-Wesley, 1989.

[INT91] Intermetrics. Draft Ada 9X Mapping Document, Volumes I and II. Mapping Specification, Ada 9X Project Report, August 1991.

[LYO86] Lyons, T. G.L. and Nissen, J.C.D. (Eds.) Selecting an Ada Environment. Cambridge University Press, 1986.

[PFL 91] Pfleeger, S.L. Process maturity as framework for CASE tool selection. Information and Software Technology, November 1991.

[PRI93] Prieto-Di'az, Ruben. Status Report: Software Reusability, May 1993, pp. 61-66.

[RUM91] Rumbaugh, James., Blaha, Michael., Premerlani, William., Eddy, Frederick., and Lorensen, Wiiliam. Object-Oriented Modeling and Design. Prentice Hall, 1991.

[SCH87] Schneidewind, N.F. Introduction to the Special Section on Software Maintenance. IEEE Transactions on Software Engineering, V13(3), pp. 303-310, March 1987.

A STUDY OF VIRTUAL REALITY AND ITS APPLICATION TO AVIONICS

Elmer A. Grubbs
Assistant Professor
Department of Engineering

New Mexico Highlands University
Las Vegas, New Mexico 87701

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

August 1993

A STUDY OF VIRTUAL REALITY AND ITS APPLICATION TO AVIONICS

Elmer A. Grubbs
Assistant Professor
Department of Engineering
New Mexico Highlands University

Abstract

A literature search on the topic of virtual reality was performed. Books, magazine articles and conference reports on the topic were read and relevant ones saved. A flicker free 3-D television system using standard television equipment with no hardware modification, which was built last summer, was updated with new circuitry developed during the school year to provide color pictures with reduced artifacts. A 486 computer system was programmed and interface circuitry built to provide both static and moving three dimensional pictures on the computer display. This system also used the flicker free methods developed last summer. New techniques for stereoscopic glasses, head mounted displays, and three dimensional displays were investigated.

A STUDY OF VIRTUAL REALITY AND ITS APPLICATION TO AVIONICS

Elmer A. Grubbs

INTRODUCTION

Virtual reality (VR) provides the capability to simulate in three dimensions nearly anything imaginary or real. VR provides the user with the ability to interact with the simulated environment and other individuals using the simulation. VR simply involves computer software and input/output devices that directly interact with our senses, producing an artificial world for us to interact in. Depending upon the particular application involved, VR is also referred to as immersion simulation, artificial reality, telepresence, virtual world, and virtual environment.

The state of the art is such that the virtual world is not completely believable. The objects represented do not have intricate structure, and are very simple and cartoon-like. Interacting with the objects may not occur in real time, i.e. there is a noticeable delay in the movements represented in the simulation space. In addition, the VR equipment is bulky and expensive. There is little in the way of tactile feedback; for example one can't feel an object as it's picked up.

The limitations to VR are being worked on by several research and development organizations. The technology for correcting most of the VR limitations already exists. However research has been conducted by a relatively small group of scientists, some who are not familiar with the breadth of the available hardware and software solutions. For example, until last year, position tracking was traditionally done with a magnetic device used since the 1970s that had a very slow response time. Only in the past year have researchers begun to use systems which overcome this limitation. Except for the tactile feedback problem, the other limitations to using VR have

solutions only a few years away.

There are many areas in the Wright Laboratory Avionics Directorate that can use VR technology. These areas will be addressed in the body of this report.

DISCUSSION

An updated literature review was conducted during the course of this assignment. Searching Computer Select (CS), and the Defense Technical Information Center (DTIC) CD roms, the WPAFB technical library index, conference reports and other sources resulted in a great deal of information. This information was organized and categorized. There are a broad range of issues, technical and otherwise, which are addressed in the literature. This report will concentrate only on the technical aspects of this emerging technology.

The hardware for a VR system consists of a computer system and input/output (I/O) devices. Input devices consist of devices such as joysticks, position sensors, television cameras, voice recognition system, data gloves, and body suits. Output devices are devices such as a head mounted display (HMD) or television/computer screen with 3-D glasses, stereo 3-D sound equipment, haptic displays, and computer speech systems. In short, anything that makes the computer easier to interface with using the human senses may be considered an I/O device to a VR system.

The software for a VR system consists of programs such as 3-D rendering, position and view calculation, I/O device interfaces, algorithms for computing constraints caused by physical laws, and motion simulation. Please refer to figure one for a block diagram of a general VR system. At this time there are VR systems available using current technology prices costing about $50,000. This price is expected to drop significantly, even as the technology improves. For example, Texas Instruments is expected to introduce a consumer product for about $400 soon. Sega is to introduce a game with a head mounted display for

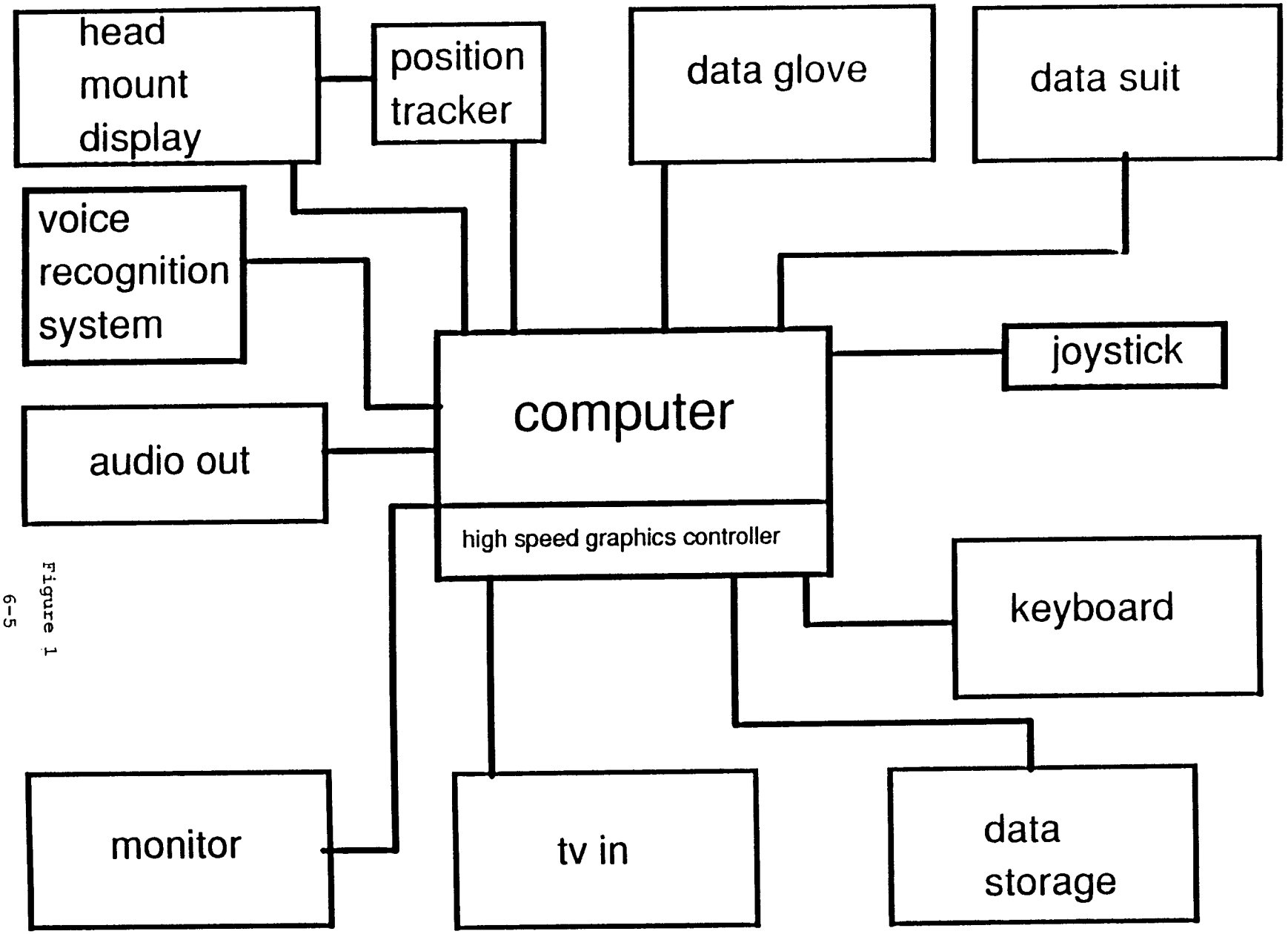

head
mount
display
position
tracker
data glove
data suit
voice
recognition
system
joystick
computer
audio out
high speed graphics controller
keyboard
monitor
tv in
data
storage

Figure 1
6-5

approximately $200 this year. Many other systems are available for $20,000 or less, but do not provide the capabilities of the more expensive systems.

There are several interesting applications of VR that are in use or being designed at this time. Several are listed below, and Appendix 1 contains a more complete list. At the University of North Carolina a VR system was prototyped for simulating the interactions between molecules. This system allows scientists to place molecules together in the correct orientation to visualize the molecular bonding. In Japan, a VR system is in use that allows customers to design a kitchen themselves and then interact and visualize their design in a three dimensional simulation. Finally, the Department of Defense has contemplated a VR system that allows a simulation of a military battle.

Another aspect of virtual reality studied this summer is the concept of three dimensional (3-D) television. Recently, with the advent of liquid crystal diode (LCD) glasses and the emergence of VR, the concept of 3-D TV has been resurrected. One of the two popular approaches is as follows. Two cameras are used, one representing left eye information, and the other representing right eye information. The left eye information is sent to the TV during the first field of a frame, and the right eye information is sent during the second, or interlaced field. During the time while the left eye information is present, the left lens of the LCD glasses is open and the right lens of the LCD glasses is closed. During the time while the right eye information is displayed, the right eye lens is open and the left eye lens is closed. Because of the fact that the field rate is 30 hz, each LCD lens is turning on and off at 30 hz. This causes an annoying flicker in the picture, which is unacceptable for general use. The second approach modifies the TV receiver so that the field rate is increased to 60 hz, thus eliminating the flicker. This approach solves one problem, but introduces another since the TV receiver needs to be modified. The high cost of this modification and the additional circuitry eliminates this alternative approach for general public use.

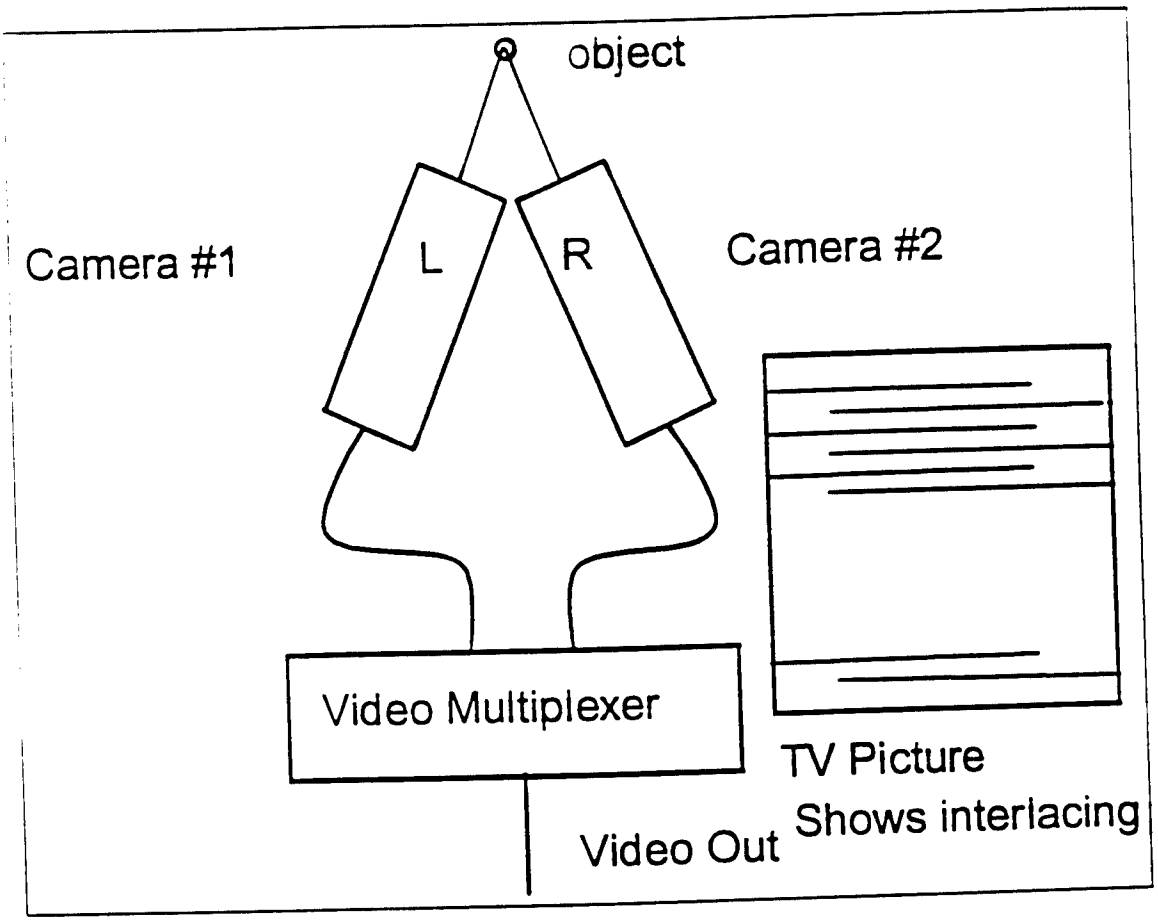
During the previous summer, circuitry was built that eliminates the 30 hz flicker. This was accomplished with no internal modifications or connections to the TV receiver! This summer the work was extended from black and white to color, and circuitry was modified to reduce the effect of artifacts or distortions, in the TV picture, which are caused by the special processing being used to reduce the flicker. Please refer to figure 2 for a block diagram of the general approach used here.

RESULTS

Modifications designed during the period 1 January through 15 May 1993 were made to the equipment designed and built during the previous summer. In addition several ideas were generated for the display of three dimensional pictures using other methods.

We first looked at possibilities for building stereoscopic glasses using something other than LCD technology. LCD technology in the short term is still expensive, and is limited by the speed with which the lens may be turned on and off. It was noted that one way mirror glass offered one potential method. If the intensity of light on the back of the glass is higher than that on the front, the mirror is transparent. If the intensity on the front is higher, the mirror is opaque, reflecting the light source back into the viewers eyes. Several test apparatus were built to demonstrate the technique. By using a light source reflecting off of a brown surface to illuminate the front of the mirror, one could either view the TV screen or a solid brown image, depending upon whether the light was on or off. This technique worked well at frequencies up to about 15 Hz. Beyond that frequency, the brain integrates the two images to form a single image. If there is a marked disparity in intensity between the two images, the brighter one will prevail. In this case, the brighter image is of the light shining on the front surface of the mirror, so the brown light is all that the viewer sees. Therefore, the method does not work at frequencies high enough to reduce flicker.

Several techniques were investigated which present the two images (left


object
Camera #1
L
R
Camera #2
Video Multiplexer
TV Picture
Shows interlacing
Video Out

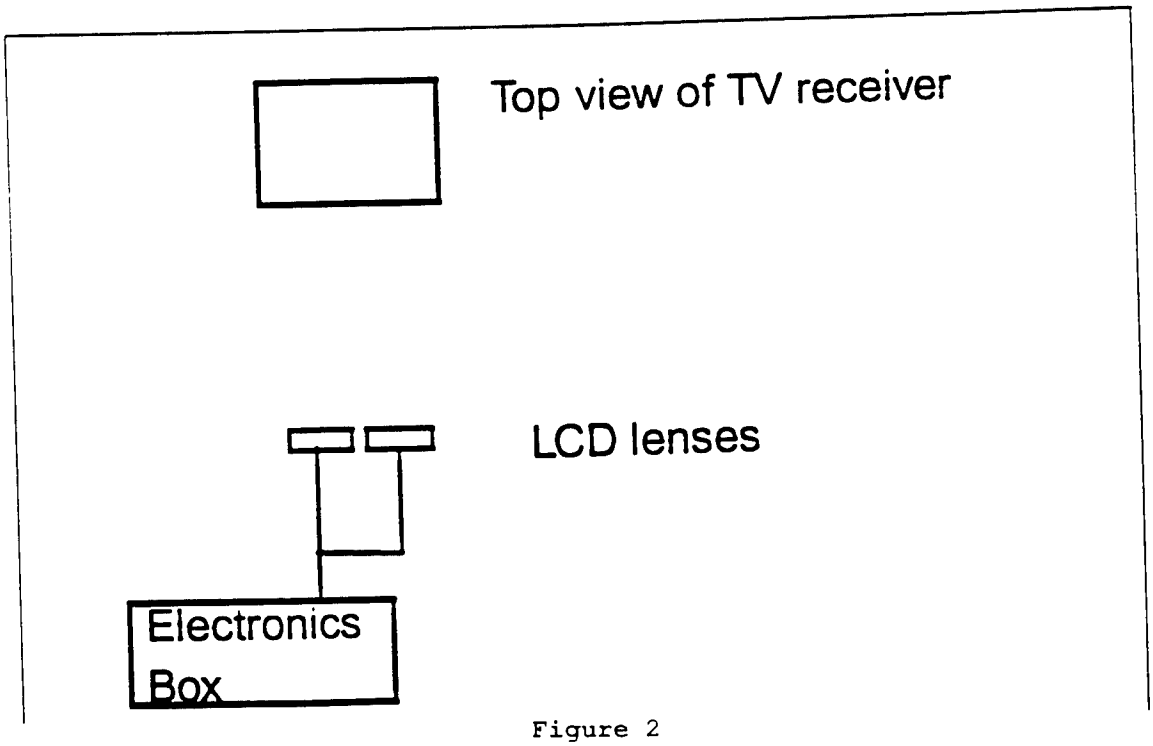

Top view of TV receiver
LCD lenses
Electronics
Box

Figure 2

and right) in a way which would allow recovery for viewing using other that LCD glasses. For example, the TV screen could be built using polarizing strips of alternate polarity covering alternate field lines. The right image would be displayed by the TV during the first field, and the left image during the interlace field. The viewer would wear glasses with a vertical polarizer on one side and a horizontal polarizer on the other, which would separate the images back out to left and right views. The brain would again integrate the two images to form a single three dimensional image, but it would still flicker at 30 Hz. This technique could be used with the flicker free method discovered last summer. It has apparently already been patented by another researcher.

Another method involves sending the two images simultaneously, one covering the majority of the TV screen, while the other is displayed on the bottom right side taking perhaps 10% or 20% of the screen. The glasses in this case would be optically designed such that the left image would be presented to the left eye of the viewer without magnification with the bottom right corner blanked out. The right image would be magnified and the entire image presented to the right eye. Although the resolution of the right image would be decreased by a factor of 5 to 10, and the left image would be missing a piece, the brain would again integrate to present an acceptable image in 3-D. The viewer would not be able to move their head appreciably using this method. An advantage is that a viewer without special glasses could still see a reasonable image in two dimensions.

The final method outlined involves a three dimensional viewing system which does not involve glasses of any kind. It would consist of four rotating LED or LCD matrices, connected to a computer system. Please refer to figure 3. Each separate pixel element would be turned on or off as it rotated depending upon whether calculations by the computer indicated that it was within the boundaries of the object to be displayed or not within the boundaries. A major problem would be turning on and off a large number of pixels in a short time,

with a limited number of connections between the computer and the rotating LCD's. This is similar to a DOD sponsored project with Texas Instruments, which uses lasers to illuminate a rotating screen.

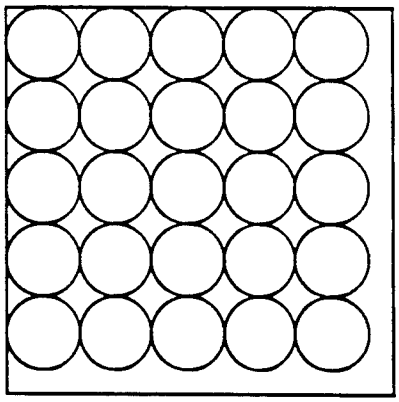
The major thrust of the last half of the summer was in applying the flicker free method developed during summer 1992 to a 486 computer system. 3-D rendering software called "Flights of Fantasy" was acquired which presented a rotating pyramid showing 3-D perspective, but displayed in two dimensions. The software was modified to use the flicker free method developed in summer 1992, as well as to interface to LCD glasses. An electronic circuit was also built which captures the vertical scan rate from the computer screen and produces the proper signal to drive the LCD glasses. Two programs are available, the first displays a stationary pyramid in three dimensions which appears to come out of the computer screen toward the viewer. The second displays a cube which rotates slowly in three dimensions as the viewer watches. Both of these programs provide flicker free images in 3-D with no internal modification to the computer, and only a small amount of external circuitry, including the LCD glasses.

Finally, an effort was begun, but not finished, to photograph an object from two separate angles at a distance of about three inches apart, digitize the images and display them on the computer in three dimensions using the flicker free technique. Unfortunately, the images were not able to be displayed on the computer due to the incorrect format of the image files. A file which presented a solid red image for the right eye and a solid blue image for the left eye was used to demonstrate the technique.

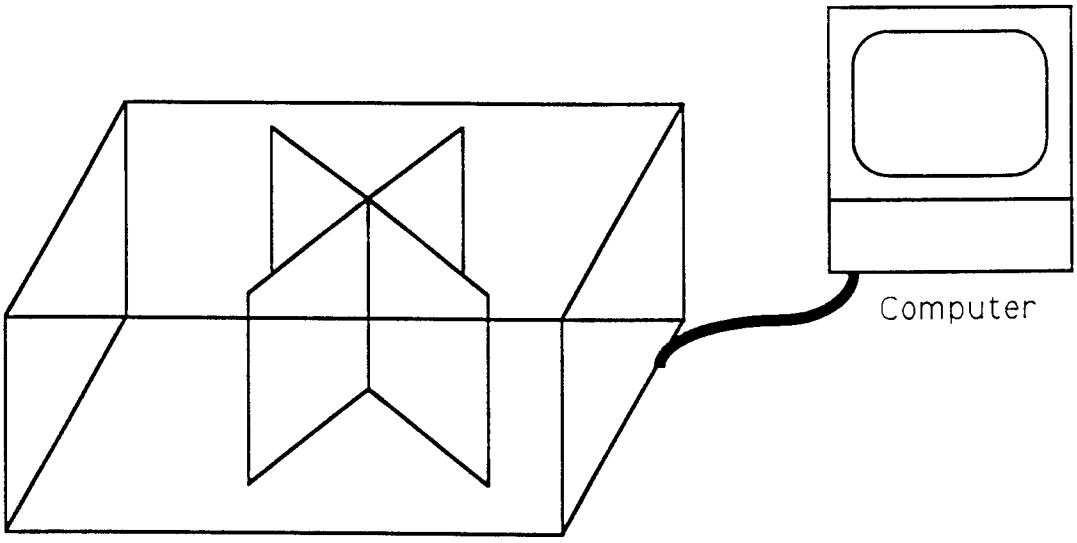
RECOMMENDATIONS

This sections outlines some the possible uses for VR in the Avionics Directorate. The section includes recommendations for using VR technology. The list presented here is not meant to be exhaustive, only representative of what is possible using this new technology.

Performance monitoring of system parameters: in software monitoring a



LED Board



Volumetric display

Figure 3

person could be placed in a virtual world of trees representing the various parameters of the avionics software. As the aircraft went through various maneuvers, the parameters could be monitored to see if any went out of bounds during any part of the flight. As various conditions occurred, the leaves or fruit on a tree could change color, alerting the user to these conditions.

Software development: Software could be developed using a visual 3-D programming language. The software engineer can arrange programming symbols in 3-D space to construct a program. In addition, the engineer can follow the flow of data through the software design, watching the data change size and shape as it is manipulated.

Navigation: A user could enter a virtual world representing the area above the earth where the aircraft was flying. The user could see "safe" corridors ahead indicating which direction to fly. Targets and threats would be represented in three dimensions augmented by other necessary data.

Packaging: Packaging of various instruments could be designed and viewed in cyberspace prior to actually assembling the device. This approach will result in errors being discovered ahead of time.

Avionics Simulation: VR can eliminate the need to have simulation domes, physical cockpits, and actual aircraft assets. The entire aircraft, its environment, and avionics can be modelled using VR. In addition, the physical test consoles and support equipment used in the simulation can be excluded by using VR technology.

Battle management: VR technology can allow missions/battles to be viewed in virtual space. For example, commanders will be able to view the battle from the perspective of the fighting units or from the enemy's point of view. Commanders could communicate with any of his assets by reaching out and "touching" the proper vehicle, unit or aircraft.

Unmanned and robotic vehicles: A pilot could fly an aircraft or cruise missile remotely using VR. With VR technology the pilot would be able to see and feel only a virtual representation of the cockpit, the instruments and the view outside the vehicle.

Facilities: Office and lab space could all be designed in cyberspace ahead of time, in fact the entire facility/lab could always be in cyberspace, where workers could perform their duties remotely. This approach would eliminate the time and money spent commuting with an automobile.

Because of the tremendous potential of this technology, I recommend WL/AAAF purchase a commercial (VR) development system. This system would serve as both an educational and prototyping medium. Individuals will be able to gain actual experience with VR technology and be able to test examples of how VR technology might be used in solving specific avionics problems.

CONCLUSIONS

A good deal of valuable research and development (R&D) work was accomplished this summer. Work begun last summer was expanded, improved, and applied directly to a 486 computer system. VR technology can be applied to a wide variety of avionics design and post deployment activities. In fact, the development and application of VR technology can not only enhance America's military superiority, but can contribute greatly to both the commercial and academic sectors of the United States. The Avionics Directorate is in an excellent position to solve existing problems with VR and refine technology for transition to Air Force users. Therefore, I believe that it is critical that VR tools be introduced into the Avionics Directorate so that work can begin. VR technology is an important technology that neither the Air Force nor the US can afford to ignore.

CENTERS OF RESEARCH

Artificial Reality Corporation

Myron Krueger

Boeing Advanced Technology Center

Human Resources Engineering Division - Armstrong Laboratory - U.S. Air Force

MIT

Marvin Minsky Thomas Sheridan

NASA - Ames Research Center

Stephen Ellis Robert Welch Elizabeth Wenzel

Telepresence Research

Scott Fisher Brenda Laurel

University of North Carolina - Chapel Hill

Frederick Brooks Henry Fuchs Warren Robinett

University of Washington - Human Interface Technology Lab

Meredith Bricken William Bricken Thomas Furness

VPL Research

VENDORS

ascension Technology Corp. P.O. Box 527 Burlington, VT 05402	Motion Trackers	(802) 655-7879
Covox, Inc. 675 Conger St. Eugene, OR 97402	Music/Voice I/O	(503) 342-1271
Crystal River Engineering 12350 Wards Ferry Rd. Groveland, CA 95321	High Speed DSP	(209) 962-6382
Digital Image Design, Inc. 170 Claremont Ave.,Suite 6 N.Y.,N.Y. 10027	3-D Computer Graphics	(212) 222-5236
Dimension International Berkshire, England	VR Computer System for PC's	(44) 734 810077
Division Ltd. Bristol, England	VR System	(44) 454 324527
Evans & Sutherland 600 Komas Dr. Salt Lake City, UT 94158	Computer Image Generator	(801) 582-5847
EXOS Inc. 8 Blanchard Rd. Burlington, MA 01803	Data Glove with Tactile Feedback	(617) 229-2075
Fake Space Labs 935 Hamilton Ave. Menlo Park, CA 94025	Alternate to HMD/Remote Camera	(415) 688-1940
Focal Point 3-D Audio 1402 Pine Ave. Suite 127 Niagra Falls, NY 14301	3-D sound	(416) 963-9188
Latent Image Development Corp. 2 Lincoln Sq. NY, NY 10023	2-D to 3-D conversion	(212) 873-5487
LEEP Systems, Inc. 241 Crescent St. Waltham, MA 02154	HMD/Optics/Cameras	(617) 647-1395
Logitech 6505 Kaiser Dr. Fremont, CA 94555	3-D Mouse/Tracker	(510) 795-8500

Polhemus Inc	Position Trackers	
P.O. Box 560 Colchester, VT 05446		(802) 655-3159
Sense8 Corp.	3-D Graphics	
1001 Bridgeway, Suite 477 Sausalito,CA 94965		(415) 331-6318
Shooting Star Technology	Position Tracker	
1921 Holdom Ave. Burnaby,BC Canada V5B 3W4		(604) 298-8574
SimGraphics Engineering Corp.	3-D Mouse/Software	
1137 Huntington Dr. South Pasadena, CA 91030		(213) 255-0900
StereoGraphics Corp.	LCD Glasses 3-D Computer/Television	
2171 E. Francisco Blvd. San Rafael, CA 94901		(415) 459-4500
StrayLight Corp.	VR System for PC's	
150 Mt. Bethel Rd. Warren, NJ 07059		(908) 580-0086
Telepresence Research	Mobile Robot	
635 High St. Palo Alto, CA 94301		(415) 325-8951
Virtual 'S' LTD	Development system/facility	
123 Mortlake High St. London, SW14 8SN Engl		(44) 81 3929000
Virtual Research	Head Mounted Display	
1313 Socorro Ave Sunnyvale, Ca 94089		(408) 739-7114
Virtual Technologies	Data Glove/Suit/Feedback	
P.O. Box 5984 Stanford, CA 94309		(415) 599-2331
The Vivid Group	VR Software + Video	
317 Adelaide St. W. Suite 302 Toronto,On Canada		(416) 340-9290
VPL Research Inc.	Data Glove/Suit/VR systems/HMD/3-D Sound	
656 Bair Island Rd. Redwood City, CA 94063		(415) 361-1710
VRG	Head Mounted Display	
800 Follin Ln Vienna, VA 22180		(703) 242-0030
VREAM Inc.	VR System for PC's	
2568 N. Clark St. Suite 250 Chicago, IL 60614		(312) 477-0425

W Industries Ltd.	VR Systems/Data Glove/Tactile Feedback/Etc	
3 Oswin Rd. Leicester LE3 1HR, England	(44) 533 542127	
Xtensory Inc.	Controller/Tactile Feedback/Software	
140 Sunridge Dr. Scotts Valley, CA 95066	(408) 439-0600	

EVENTS

Virtual Reality International	January
Virtual Reality Systems	March
Virtual Reality '94	May
International Conference on Cyberspace	May
Medicine Meets Virtual Reality	June
SIGGRAPH	August
IEEE Virtual Reality International Symposium	September
Calgary Virtual Reality Conference	October
Cyberarts	November
Virtual Reality Summit	November/December

References

1. AI EXPERT, August 1991, pp.26-39

2, AI EXPERT, August 1992, pp.22-29; 42-48

3. AI EXPERT Special Report, July 1992

4. ASEE PRISM, "The Next Best Thing to Being There", May 1992, pp.26-29

5. BYTE, April 1992, pp.135-150, 175-182

6. Communications of the ACM, "Exploring Virtual Worlds with Tom Furness", July 1991

7. Digital Media: A Seybold Report, "Stepping into Virtual Reality", Caruso, D Aug 1991

8. Edventure Holdings, "Virtual Reality:Spreadsheets for Industry",Oct 8,1990

9. EXE, "Virtual Worlds", Roth, A., Dec 1991

10. Flights of Fantasy, Christopher Lampton, Waite, 1993

11. Government Computer News, "DOD Hopes 3-D Systems Find Home in Civilian Sector",Aug 17, 1992

12. MacWeek, Bring that to the Mac, Sega", Jan 25, 1993

13. Patricia Seybold's Office Computing Report, "Cyberspace: Reality is no Longer Enough",Oct 1990

14. PIXEL, No. 6, 1991

15. Presence, MIT Press, Winter 1992

16. Science, "Looking Glass Worlds", Peterson, I.,Jan 4,1992, pp.8-15

17. Virtual Reality, Howard Rheingold, Summit, 1991

18. Virtual Reality Playhouse, Nicholas Lavroff, Waite, 1992

19. Virtual Reality, Through the New Looking Glass, Ken Pimentel and Kevin Teixeira, McGraw-Hill, 1993

CURRENT AND FUTURE APPLICATIONS OF VR TECHNOLOGY

1. A system that allows a surgeon to enter a body virtually to perform an operation using microminiature surgical tools.

2. An air traffic control system that allows the operator to see the aircraft in three dimensions and communicate with an aircraft by reaching out in cyberspace and grabbing it.

3. Systems that allow a user to remotely control a robot or combat vehicle as though he was actually present at the site.

4. Tools which allow young students to design and interact with their own virtual worlds.

5. Games that allow multiple users to interact with each other.

6. Systems that allow a user to see in three dimensions the person they are talking to via telephone

7. A maintenance system that allows a technician to take three dimensional visual parts descriptions with him to work on an aircraft.

8. A medical system that allows a doctor to see inside a patient to locate tumors or other problems

9. A virtual office where records are stored in virtual file cabinets, books and reports in virtual bookcases, and videos and records can be played.

10. A virtual school where students attend from a distance without leaving their own homes

A Framework of Multiresolutional Target Tracking

Lang Hong
Assistant Professor
Dept. of Electrical Engineering
Wright State University
Dayton, OH 45435

Final Report for
Summer Faculty Research Program
AART-1, Wright Laboratory

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

August, 1993

A Framework of Multiresolutional Target Tracking

Lang Hong
Assistant Professor
Dept. of Electrical Engineering
Wright State University
Dayton, OH 45435

Abstract

A framework of multiresolutional target tracking has been established. The wavelet transform is employed in constructing multiresolutional data and model structures. Multiresolutional tracking is performed over the multiresolutional data and model structures in a top-down fashion. The main advantages of multiresolutional target tracking include: computational efficiency, performance robustness and algorithm flexibility. Two-level joint probabilistic data association (JPDA) - nearest neighborhood (NN) and NN–JPDA target tracking algorithms are developed. Computational efficiency is achieved and advantages of both JPDA and NN are combined.

A Framework of Multiresolutional Target Tracking

Lang Hong

I. Introduction

Multiresolutional signal processing has been employed in image processing and computer vision to achieve improved performance that cannot be achieved using conventional signal processing techniques at only one resolution level. In this research, a framework of multiresolutional target tracking is established to explore the advantages of multiresolutional approaches. The foundation of this framework is the wavelet transform which has been applied successfully to many areas involving multiresolutional processing. The key ingredients in multiresolutional target tracking are multiresolutional modeling and multiresolutional measurements. Although only uniform resolutional measurements are available, multiresolutional measurements are calculated by using the wavelet transform. Multiresolutional data and model structures are then formed in a bottom-up process. Tracking algorithms are applied at each level of the multiresolutional structures in a top-down process. The advantages of multiresolutional target tracking algorithms include: computational efficiency, performance robustness and algorithm flexibility. Multiresolutional target tracking algorithms can be considered as a generalization of existing tracking algorithms.

Multiresolutional estimation with applications to multiresolutional sensor fusion has been investigated by Hong [3]. In [3], measurements are available at each resolutional level and the estimates from each level are integrated using the wavelet transform. When applied to target tracking, the greatest concern regarding the use of the wavelet transform related approaches is the capability of real-time processing. The word *real-time* is in the sense that when a single new measurement is acquired at the finest level, a new estimate can be derived based upon all measurements available at all resolutional levels at the present time index. This concern is not without reason. Until now, in the signal processing area, the wavelet transform was not applied until a batch of measurements was collected. This resulted in time delay in processing which was not allowed in many applications, such as target tracking. Hong [3] developed a semi-real-time algorithm

that divided the data into blocks, and processed the data after each block of data was acquired. In [4], a true real-time multiresolutional approach was presented. A tree-like data structure was introduced in which the bottom level of the tree corresponded to the highest resolutional level and the top level of the tree represented the coarsest resolutional level. The tree was growing and moving as new measurements were acquired and the front branches of the tree were associated with the real-time filtering process.

II. The Discrete Wavelet Transform

In this report, the discrete wavelet transform is introduced by the concept of a filter bank, Fig. 1. For a given sequence of signals $x(i,n) \in l^2(\mathbf{Z})$, $n \in \mathbf{Z}$ at resolutional level i ($x(i,n)$ is assumed a scalar sequence here), a lower resolutional signal can be derived by lowpass filtering with a halfband lowpass filter having impulse response $h(n)$. A sequence of the lower resolutional signal is obtained by subsampling the output of the lowpass filter by two,

$$x(i-1,n) = \sum_k h(2n-k)x(i,k). \tag{1}$$

Eq. (1) is a mapping from a vector space $l^2(\mathbf{Z})$ into itself $l^2(\mathbf{Z})$. An "added detail", also called wavelet coefficients, which is lost from $x(i,n)$ in lowpass filtering can be computed by first using a highpass filter with impulse response $g(n)$ and then by subsampling the output of highpass filtering by two. The added detail is given by

$$y(i-1,n) = \sum_k g(2n-k)x(i,k). \tag{2}$$

The original signal $x(i,n)$ can be recovered from two filtered and subsampled (lower resolution) signals $x(i-1,n)$ and $y(i-1,n)$. Filters $h(n)$ and $g(n)$ must meet some constraints in order to have perfect reconstruction. In addition to the regularity constraint [2], the filter impulse responses form an orthonormal set. Therefore, Eqs. (1) and (2) can be considered as a decomposition of the original signal onto an orthonormal basis and the reconstruction

$$x(i,n) = \sum_k h(2k-n)x(i-1,k) + \sum_k g(2k-n)y(i-1,k) \tag{3}$$

can be considered as summing up the orthogonal projections. In this report, filters $h(n)$ and $g(n)$ are assumed to be FIR filters. It has been shown that lowpass filter $h(n)$ must be the impulse response of a quadrature mirror filter (QMF) and $g(n)$ and $h(n)$ form a conjugate mirror filter pair [9, 10]

$$g(L-1-n) = (-1)^n h(n) \tag{4}$$

where L is the filter length (which has to be even). Eq. (4) means that once lowpass filter $h(n)$ is determined, the conjugated highpass filter can also be determined. The above decomposition can be repeated over resolutional levels. By defining

$$\begin{aligned}
&\underline{X}(i) = [x(i,1), x(i,2), ..., x(i,N)]^T, \\
&\underline{X}(i-1) = [x(i-1,1), x(i-1,2), ..., x(i-1,N/2)]^T, \\
&\underline{Y}(i-1) = [y(i-1,1), y(i-1,2), ..., y(i-1,N/2)]^T,
\end{aligned}$$

Eqs. (1) and (2) can be written in terms of matrices

$$\underline{X}(i-1) = \mathbf{H}_{i-1}\underline{X}(i) \quad \text{and} \quad \underline{Y}(i-1) = \mathbf{G}_{i-1}\underline{X}(i) \tag{5}$$

where $\mathbf{H}_{i-1}$ and $\mathbf{G}_{i-1}$ are scaling and wavelet operators which have the following properties:

$$\mathbf{H}_{i-1}^T\mathbf{H}_{i-1} + \mathbf{G}_{i-1}^T\mathbf{G}_{i-1} = \mathbf{I} \tag{6}$$

and

$$\begin{pmatrix} \mathbf{H}_{i-1}\mathbf{H}_{i-1}^T & \mathbf{H}_{i-1}\mathbf{G}_{i-1}^T \\ \mathbf{G}_{i-1}\mathbf{H}_{i-1}^T & \mathbf{G}_{i-1}\mathbf{G}_{i-1}^T \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{pmatrix}. \tag{7}$$

Similarly, Eq. (3) can also be written in the operator form

$$\underline{X}(i) = \mathbf{H}_{i-1}^T\underline{X}(i-1) + \mathbf{G}_{i-1}^T\underline{Y}(i-1). \tag{8}$$

In this report, the transform specified by Eq. (5) is called the forward wavelet transform (or simply wavelet transform) and the transform specified by Eq. (8) is called the inverse

wavelet transform.

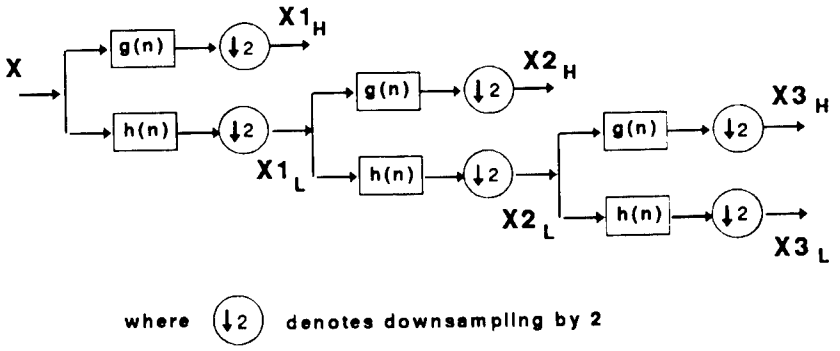


Figure 1: A filter bank implementation of the wavelet transform.

III. Multiresolutional Target Tracking

In this section, a framework of multiresolutional target tracking is reviewed. A functional block diagram of multiresolutional target tracking is shown in Fig. 2.

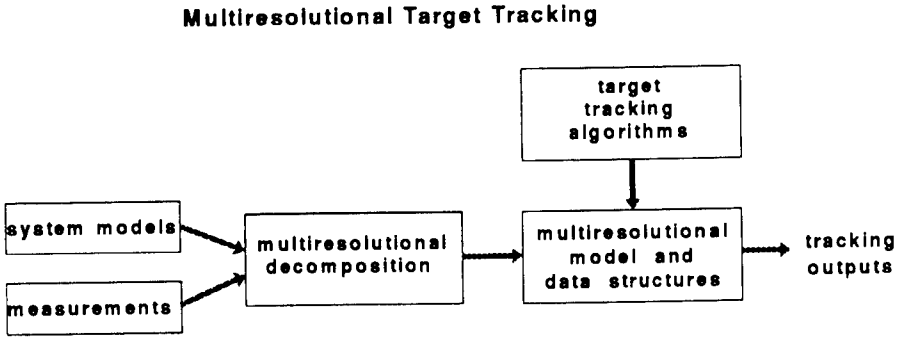


Figure 2: Functional block diagrams of multiresolutional target tracking.

The key components for multiresolutional tracking are multiresolutional decomposition and multiresolutional data and model structures which will be discussed in the sequel.

Multiresolutional Data Decomposition

Since the measurements that are available have only one resolution, a decomposition is essential to build a multiresolutional data structure. An ideal multiresolutional data structure should render itself to easy data decomposition in a bottom-up process and data composition in a top-down process. The wavelet transform provides a natural mechanism for multiresolutional decomposition and composition. A multiresolutional data decomposition process (bottom-up process) is shown in Fig. 3, where only four levels are depicted.

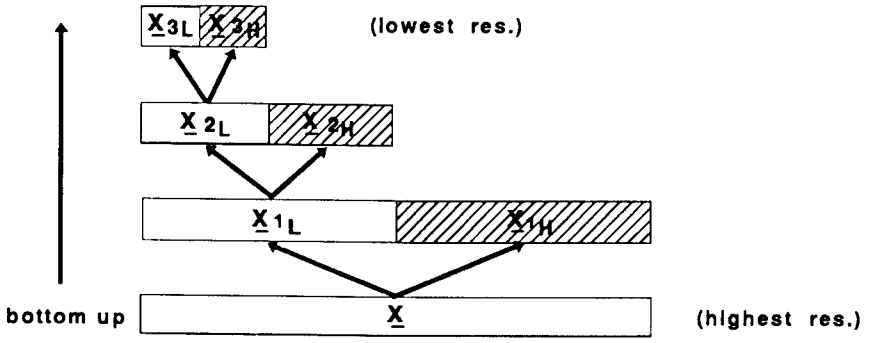


Figure 3: A multiresolutional data decomposition process.

As can be seen from Fig. 3, a given measurement sequence $\underline{X}$ is decomposed into two sequences of lower resolutional data, one in white $\{\underline{X}, \underline{X}_{1L}, \underline{X}_{2L}, \underline{X}_{3L}, ...\}$ and the other in shaded $\{\underline{X}_{1H}, \underline{X}_{2H}, \underline{X}_{3H}, ...\}$. The sequence in white is a lowpass filtered one and forms a multiresolutional data structure. The sequence in shaded is a highpass filtered one, called added details. Since the measurements have a noisy behavior, the highpass filtered data are "noise like", but not white. Having a sideproduct of added details in the decomposition process is unique and advantageous for the wavelet transform. A complete data reconstruction can be obtained by using the added details in a top-down process in the data structure.

Multiresolutional Model Decomposition

To describe target dynamics and sensor behaviors for different resolutional measure-

ments, the system model and measurement model need to be decomposed accordingly, forming a multiresolutional model structure. The model for the ith ($i = 1, .., I$) resolutional level is

$$\underline{x}^i(k+1) = \mathbf{A}^i \underline{x}^i(k) + \underline{w}^i(k), \quad \underline{w}^i(k) \sim N(0, \mathbf{Q}^i(k)), \tag{9}$$

$$\underline{y}^i(k) = \mathbf{C}^i \underline{x}^i(k) + \underline{v}^i(k), \quad \underline{v}^i(k) \sim N(0, \mathbf{R}^i(k)). \tag{10}$$

Usually, since the dimension of the state vector is smaller than that of the measurement vector, care must be given when decomposing models to avoid the problem of singularity.

Multiresolutional Tracking

A multiresolutional target tracking algorithm, in general, consists of two processes: bottom-up and top-down. The bottom-up process is a model-building process, in which given measurements and system models are decomposed into a multiresolutional data structure and a multiresolutional model structure using the wavelet transform. The top-down process is a tracking and refining process. Starting from the top level of the multiresolutional data and model structures, existing target tracking algorithms (such as the multiple hypothesis tracking (MHT) algorithm, the nearest neighborhood (NN) method and the joint probabilistic data association (JPDA) technique) can be implemented. Since the number of measurements at a coarser level, including false alarms, is much smaller than that at the conventional resolutional level, the targets can be tracked easily with less computation. For instance, when using MHT, the number of hypotheses generated at a coarser level could be significantly smaller, which results in a significant computation saving. However, due to the nature of downsampling, the tracks are less accurate at coarser levels. By utilizing the "zooming" property of the wavelet transform, the track accuracy will be refined by putting back the "added details". At the bottom level of the structures, high quality tracks are obtained. A top-down refining process is demonstrated in Fig. 4. Significant computational savings will be achieved by employing different algorithms at different resolutional levels to combine the strength of different algorithms, which is impossible for uniresolutional target tracking approaches. Algorithms of two-level NN-JPDA and JPDA-NN are presented in the next section. Also, algorithm flexibility is gained using a multiresolutional approach. For instance, when

two kinds of sensory data with different resolutions (e.g., infrared and radar images) are available, a multiresolutional tracking algorithm can be readily used to fuse information from these two sources to obtain better tracks.

Another approach introduced by Musicki and Evans [8] dealing with finite resolutional measurements is worthy of mentioning here. In their approach, measurements have physically a finite resolution and measurements falling in a resolutional cell are combined. An integrated JPDA algorithm was developed for finite resolution sensors. The multiresolutional approaches discussed in this report can be considered as dealing with measurements provided by a sensor (such as radar) which has *actively* controlled resolutional cells.

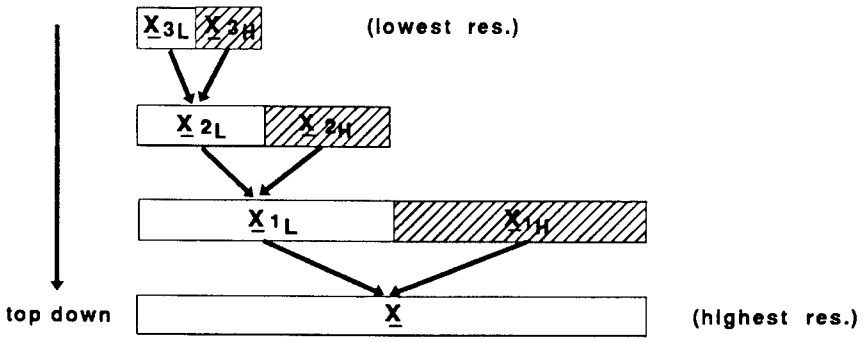


Figure 4: A top-down refining process.

IV. JPDA-NN and NN-JPDA Algorithms

Since both the NN and the JPDA algorithms are well understood by the tracking community [1], in this section, only functional descriptions of JPDA-NN and NN-JPDA are given and an emphasis is placed on combinations of the NN and the JPDA algorithms. First the NN and the JPDA are briefly reviewed. Functional diagrams of the NN and the JPDA are presented in Figs. 5 and 6.

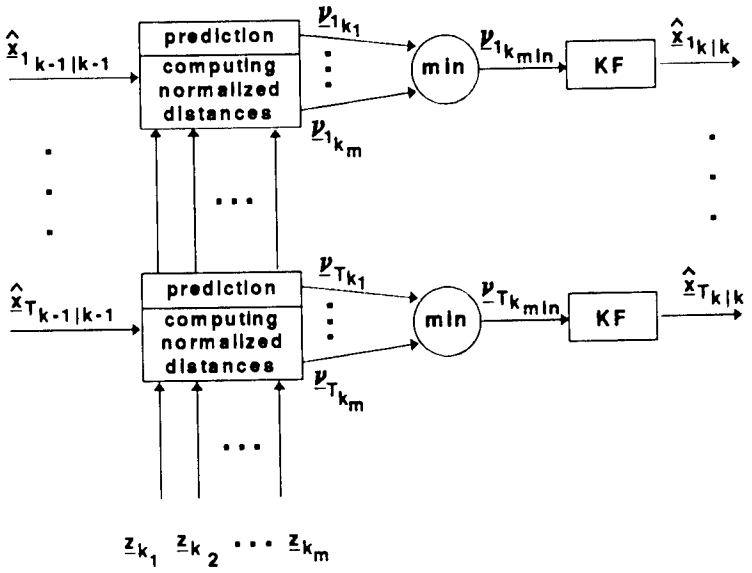


Figure 5: A functional diagram of the NN algorithm.

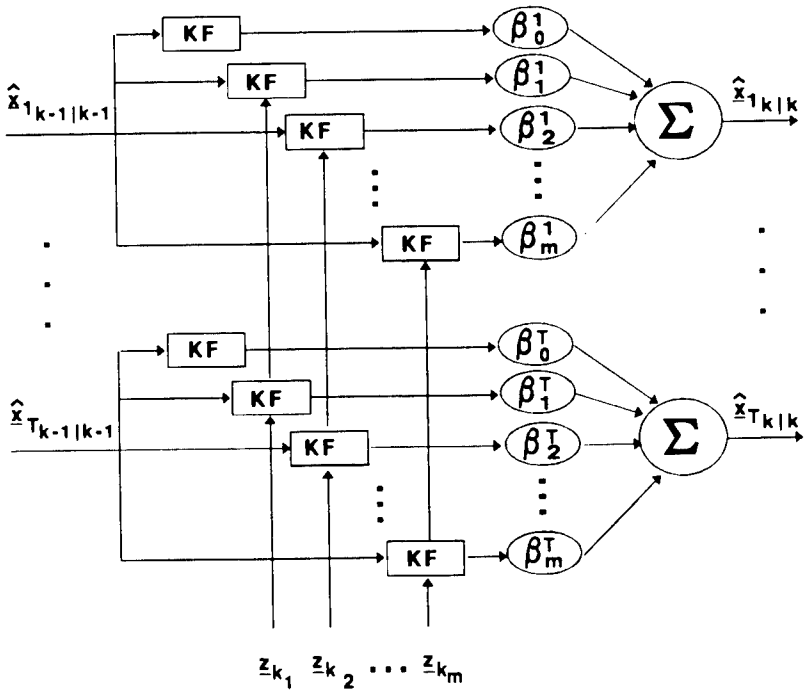


Figure 6: A functional diagram of the JPDA algorithm.

The NN algorithm at moment k takes the measurement whose normalized distance to track $\underline{\hat{x}}^t_{k|k-1}$ is minimum to update the track. The advantage of the NN algorithm is that it is simple and computationally efficient. However, tracks might be lost in the NN algorithm because only the "nearest" measurement is used in updating. On the other hand, the JPDA algorithm uses "all-neighborhood" measurements with weights according to their statistical properties. The coefficient, β^t_j, $t = 1, ..., T$ and $j = 1, ..., m$, reflects the probability that measurement j is originated from target t. Since all-neighborhood measurements are used, the JPDA has a better tracking performance than the NN algorithm. However, the computational complexity for the JPDA algorithm is much higher than that of the NN algorithm. It is naturally desired that these two algorithms be combined such that the combined algorithm is as computationally efficient as the NN algorithm, and its performance is as robust as the JPDA algorithm. This is impossible with uniresolutional approaches. With the introduction of the framework of multiresolutional target tracking, the combination of the NN and JPDA algorithms becomes not only possible but also recommended. The multiresolutional data and model structures provide an ideal platform for this combination and the wavelet transform is an ideal vehicle linking information between different resolutional levels.

There are two combinations: NN-JPDA and JPDA-NN. A functional block diagram of the JPDA-NN is shown in Fig. 7.

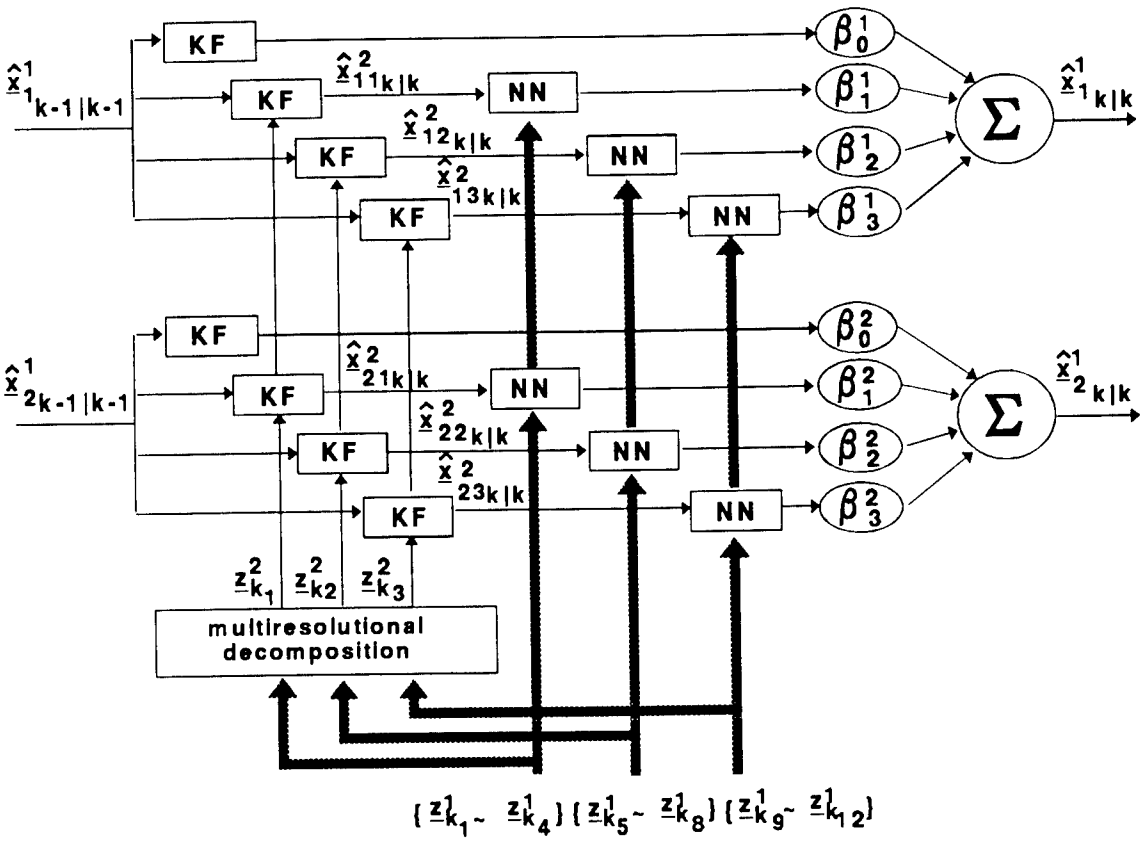


Figure 7: A functional diagram of the JPDA-NN algorithm.

Assuming that at the $(k-1)$th moment we have two tracks: $\underline{\hat{x}}_{i_{k-1|k-1}}$, $i = 1, 2$. Also assume that we have twelve validated measurements at the kth moment: $\underline{z}_{k_1}, \underline{z}_{k_2}, ..., \underline{z}_{k_{12}}$. In the following we will describe a complete track updating cycle for the JPDA-NN algorithm. The resolution of the given measurements is considered to be the highest and is assigned as resolutional level one. For notational convenience, we rewrite the tracks and measurements with a superscript denoting the resolutional level index: $\underline{\hat{x}}^1_{i_{k-1|k-1}}$, $i = 1, 2$ and $\underline{z}^1_{k_j}$, $j = 1, ..., 12$. Using a distance metric, the measurements are assumed to have been partitioned into three groups[1]: $\{\underline{z}^1_{k_1}, ..., \underline{z}^1_{k_4}\}$, $\{\underline{z}^1_{k_5}, ..., \underline{z}^1_{k_8}\}$ and $\{\underline{z}^1_{k_9}, ..., \underline{z}^1_{k_{12}}\}$

[1]The family of the wavelet transform that we used required the number of measurements be two to an integer power, i.e., 2^n. The simplest partition which meets this requirement is dividing measurements into groups with two measurements (i.e., 2^1). We are currently extending the wavelet transform to take arbitrary number of measurements.

Applying the wavelet transform to these three groups of measurements generates three coarser measurements (at level two):

$$\{\underline{z}^1_{k_1}, \underline{z}^1_{k_2}, \underline{z}^1_{k_3}, \underline{z}^1_{k_4}\} \xrightarrow{\text{wavelet transform}} \underline{z}^2_{k_1}$$

$$\{\underline{z}^1_{k_5}, \underline{z}^1_{k_6}, \underline{z}^1_{k_7}, \underline{z}^1_{k_8}\} \xrightarrow{\text{wavelet transform}} \underline{z}^2_{k_2}$$

$$\{\underline{z}^1_{k_9}, \underline{z}^1_{k_{10}}, \underline{z}^1_{k_{11}}, \underline{z}^1_{k_{12}}\} \xrightarrow{\text{wavelet transform}} \underline{z}^2_{k_3}.$$

The JPDA algorithm is applied at level two to the coarser measurements to update the tracks, resulting in two sets of updated tracks: $\underline{\hat{x}}^2_{11_{k|k}}, \underline{\hat{x}}^2_{12_{k|k}}, \underline{\hat{x}}^2_{13_{k|k}}$ and $\underline{\hat{x}}^2_{21_{k|k}}, \underline{\hat{x}}^2_{22_{k|k}}, \underline{\hat{x}}^2_{23_{k|k}}$, where subscript $\{im\}$ denotes that track $\underline{\hat{x}}^1_{i_{k-1|k-1}}$ is updated by measurement $\underline{z}^2_{k_m}$. Since only three measurements instead of twelve are involved in the JPDA tracking process at resolutional level two, the computational complexity is much lower. In other words, the data association is performed at resolutional level two among groups of measurements, instead of each individual measurement. Instead of combining tracks $\underline{\hat{x}}^2_{il_{k|k}}$, $l = 1, 2, 3$ into $\underline{\hat{x}}_{i_{k|k}}$ as regular JPDA does, each coarser track is refined by using the NN algorithm. For example, the coarser track, $\underline{\hat{x}}^2_{12_{k|k}}$ shown in Fig. 7, is refined by an NN tracker whose measurement inputs are: $\{\underline{z}^1_{k_5}, \underline{z}^1_{k_6}, \underline{z}^1_{k_7}, \underline{z}^1_{k_8}\}$. The refined tracks from the NN trackers are then combined by JPDA's weighting coefficients to generate high quality tracks at resolutional level one: $\underline{\hat{x}}_{1_{k|k}}$ and $\underline{\hat{x}}^1_{2_{k|k}}$. Notice that in the NN trackers at resolutional level one, no propagations are performed, since tracks were propagated in the JPDA trackers at the coarser level. In summary, the JPDA-NN algorithm performs data association over coarser measurements (groups of measurements) to save computation while retaining the strength of JPDA. The coarser tracks are then refined by the NN algorithm at resolutional level one to improve track accuracy. The overall behavior of JPDA-NN is similar to that of JPDA, but the computational complexity is low. The amount of computational savings depends on the measurement resolution reduction rate (the rate is 4:1 in Fig. 7). The larger the resolutional reduction rate is, the more computational savings, but the farther the behavior of the JPDA-NN algorithm from

the JPDA algorithm.

A functional block diagram of the NN-JPDA algorithm is shown in Fig. 8.

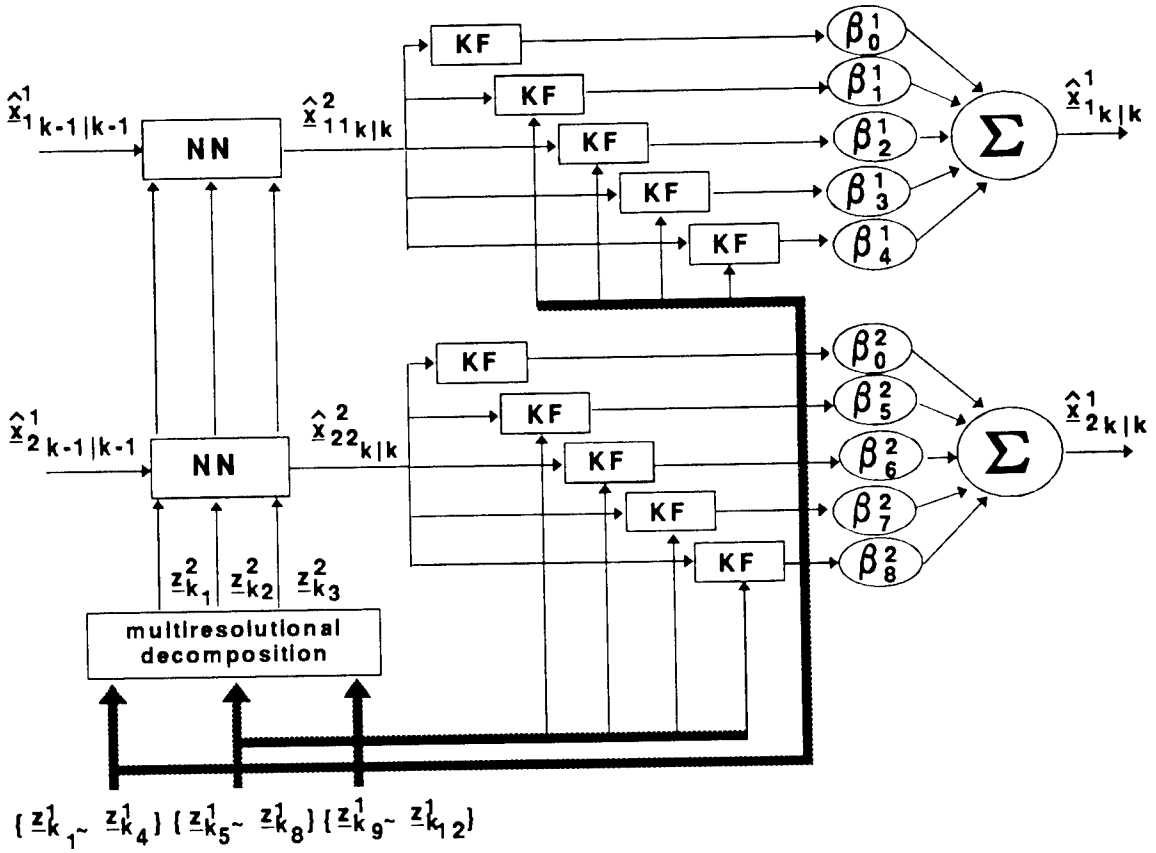


Figure 8: A functional diagram of the NN-JPDA algorithm.

We also assume that there are two tracks at the $(k-1)$th moment, $\hat{\underline{x}}_{i_{k-1|k-1}}$, $i = 1, 2$ and twelve validated measurements at the kth moment, $\underline{z}_{k_1}, \underline{z}_{k_2}, \ldots, \underline{z}_{k_{12}}$. The measurements are also assumed to have been partitioned into three groups: $\{\underline{z}^1_{k_1}, \ldots, \underline{z}^1_{k_4}\}$, $\{\underline{z}^1_{k_5}, \ldots, \underline{z}^1_{k_8}\}$ and $\{\underline{z}^1_{k_9}, \ldots, \underline{z}^1_{k_{12}}\}$. In the NN-JPDA algorithm, the NN algorithm is first applied at resolutional level two to the coarser measurements: $\underline{z}^2_{k_1}$, $\underline{z}^2_{k_2}$ and $\underline{z}^2_{k_3}$. Assume that the nearest measurement to track $\hat{\underline{x}}^1_{1_{k-1|k-1}}$ is $\underline{z}^2_{k_1}$ and the nearest measurement to track $\hat{\underline{x}}^1_{2_{k-1|k-1}}$ is $\underline{z}^2_{k_2}$. Updating by NN trackers results in coarser tracks: $\hat{\underline{x}}^2_{11_{k|k}}$ and $\hat{\underline{x}}^2_{22_{k|k}}$. Refinement of tracks is narrowed down to the groups of measurements: $\underline{z}^1_{k_1}, \ldots, \underline{z}^1_{k_4}$, and $\underline{z}^1_{k_5}, \ldots, \underline{z}^1_{k_8}$ at resolutional level one. The JPDA algorithm is then used to refine the tracks using the measurements from the groups (Fig. 8). Similarly, no propagation

operations are performed in the JPDA trackers at resolutional level one. The overall behavior of the NN-JPDA algorithm is similar to that of the NN algorithm. However, since *the nearest groups* of measurements are used by the NN trackers at resolutional level two, the NN-JPDA algorithm has better capability of maintaining tracks than the traditional NN algorithm. Therefore the NN-JPDA algorithm is nearly as efficient as the NN algorithm, but with a certain capability of the JPDA algorithm.

V. Simulations and Discussions

The JPDA-NN and NN-JPDA algorithms described in this report have been applied to various testing scenarios. Simulation results of one typical scenario are presented in this section. Fig. 9 shows a scenario of two crossing targets with constant velocities. At each scan there could be as many as ten false alarms. For the purpose of comparison, both the NN and the JPDA algorithms are applied to the scenario with an incorrect initial value and the results are shown in Figs. 10 and 11. One can see from Figs. 10 and 11 that one track was lost in the NN tracker, while JPDA maintains tracks pretty well. The results of the JPDA-NN and the NN-JPDA algorithms are shown in Figs. 12 and 13. A comparison of computational complexity of four algorithms: JPDA, NN, JPDA-NN and NN-JPDA is given in Table 1. The units in the table are numbers of floating point operations (flops). Although JPDA provides a decent result, its computational complexity is high. On the other hand, although the performance of the NN algorithm is not satisfactory in this scenario, it is extremely computationally efficient. The performance of the JPDA-NN algorithm is very compatible to that of the JPDA algorithm, but its computational complexity is much lower than that of JPDA (the resolution reduction rate is 2:1 in the implementation). The performance of NN-JPDA is much better than that of the NN algorithm and its computational complexity is compatible to the NN algorithm. By now, the beauty of the JPDA-NN and NN-JPDA algorithms is completely demonstrated: one can effectively control the trade off between the performance and computational complexity. If one wishes to achieve a performance of JPDA but cannot afford the computational burden, he/she can choose the JPDA-NN algorithm. On the other hand, if one likes to have a computational level as in the NN

algorithm, but wishes to improve the performance, he/she can choose the NN-JPDA algorithm.

V. Conclusions

A framework of multiresolutional target tracking is established and two-level JPDA-NN and NN-JPDA algorithms are developed in this report. It has been shown that the algorithms combine the strengths of JPDA and NN. By choosing JPDA-NN or NN-JPDA, one can control the trade off between performance and computational complexity.

References

[1] Bar-Shalom, Yaakov and Thomas E. Fortmann, *Tracking and Data Association*, Academic Press, Boston, MA, 1988.

[2] Daubechies, Ingrid, "Orthonormal Bases of Compact Supported Wavelets," *Communications on Pure and Applied Mathematics*, Vol. XLI, pp. 909–996, 1988.

[3] Hong, Lang, "Multiresolutional Distributed Filtering," to appear in *IEEE Trans. on Automatic Control*, Vol. 39, 1994.

[4] Hong, Lang and Todd Scaggs, "Real-Time Optimal Filtering for Stochastic Systems With Multiresolutional Measurements," *System & Control Letters*, Vol. 20, 1993.

[5] Hong, Lang, "Multiresolutional Multiple-Model Target Tracking," *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 30, No. 2, April 1994.

[6] Hong, Lang, "Multiresolutional Target Tracking Using Wavelet Transform," to appear in the *Proc. of the 32nd IEEE Int. Conf. on Decision and Control,* San Antonio, TX, Dec. 1993.

[7] Mallat, Stephane G., "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 11, No. 7, pp. 674–693, 1989.

[8] Musicki, D. and R. Evans, "Integrated Probabilistic Data Association in Clutter with Finite Resolution Sensor," to appear in the *Proc. of 32nd IEEE Int. Conf. on Decision and Control*, San Antonio, TX, Dec. 1993.

[9] Rioul, Olivier and Martin Vetterli, "Wavelet and Signal Processing," *IEEE Signal Processing Magazine,* Vol. 8, No. 4, pp. 14–38, 1991.

[10] Smith, Mark J. T. and Thomas P. Barnwell, III, "Exact Reconstruction Techniques for Tree-Structured Subband Coders," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 34, No. 3, pp. 434–441, 1986.

algorithm	number of flops
JPDA	12,703,068
NN	686,458
JPDA-NN	5,329,963
NN-JPDA	1,865,819

Table 1: A comparison of computational complexity of different algorithms.

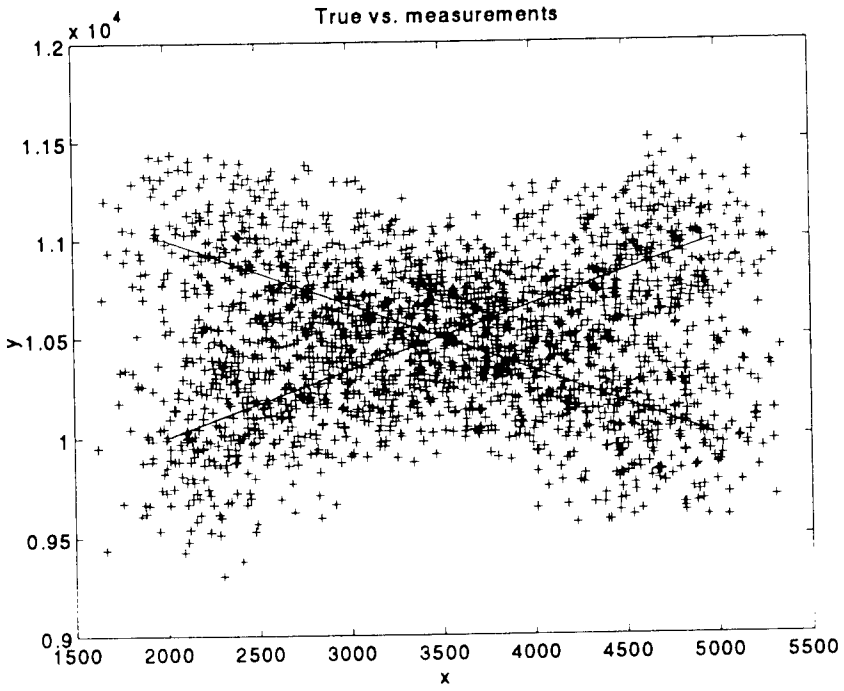


Figure 9: A scenario of two crossing targets.

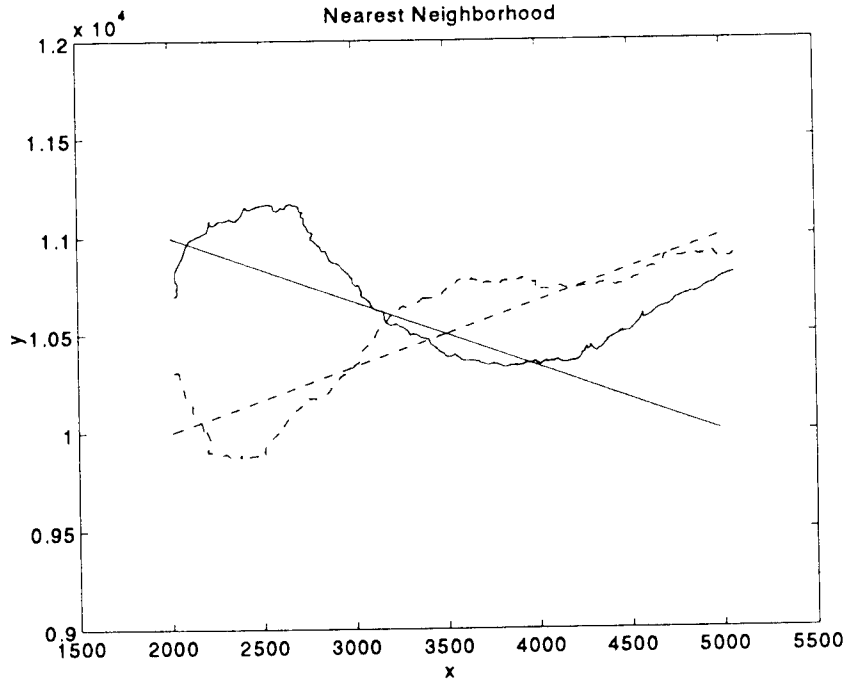


Figure 10: Estimated tracks of the target using the NN algorithm.

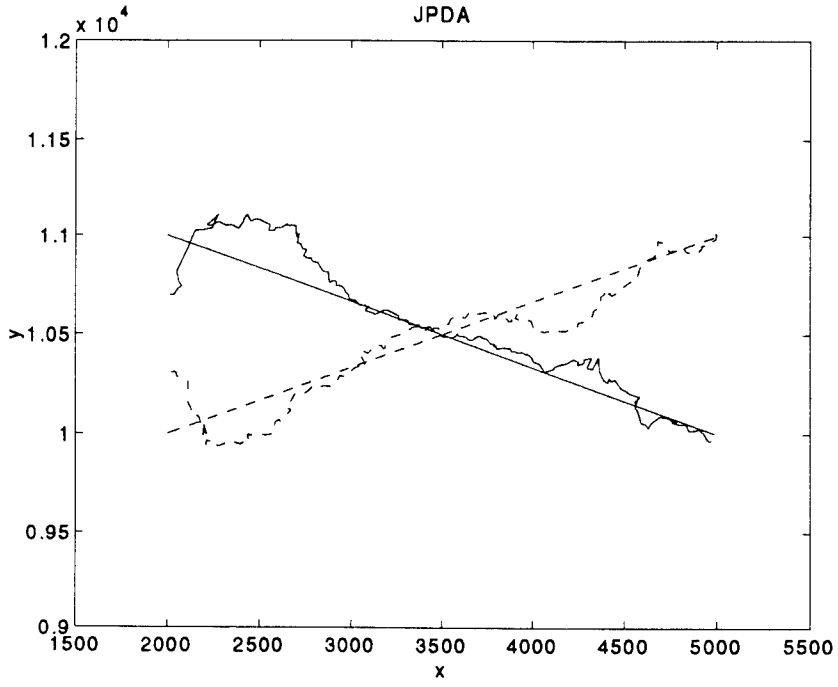


Figure 11: Estimated tracks of the targets using the JPDA algorithm.

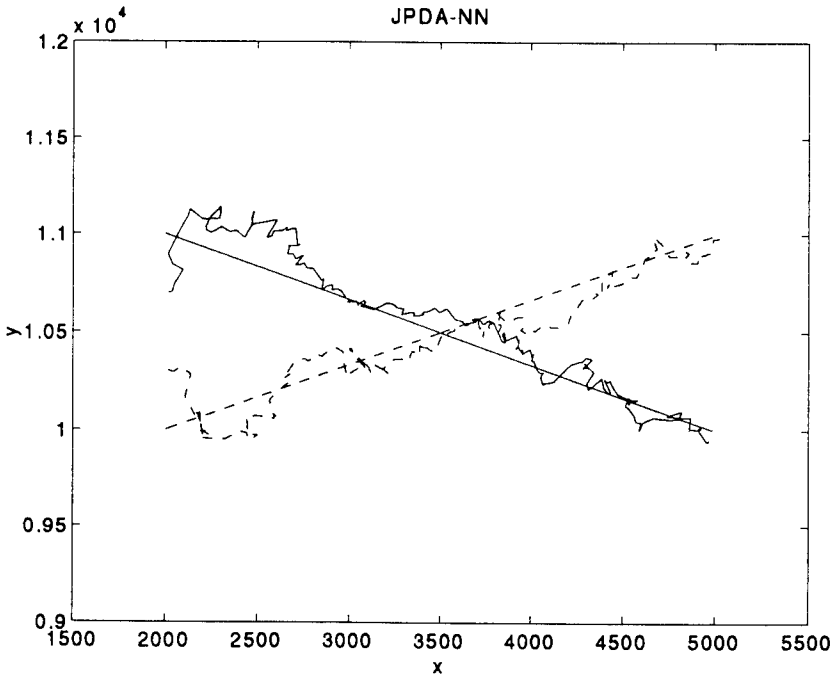


Figure 12: Estimated tracks of the targets using the JPDA–NN algorithm.

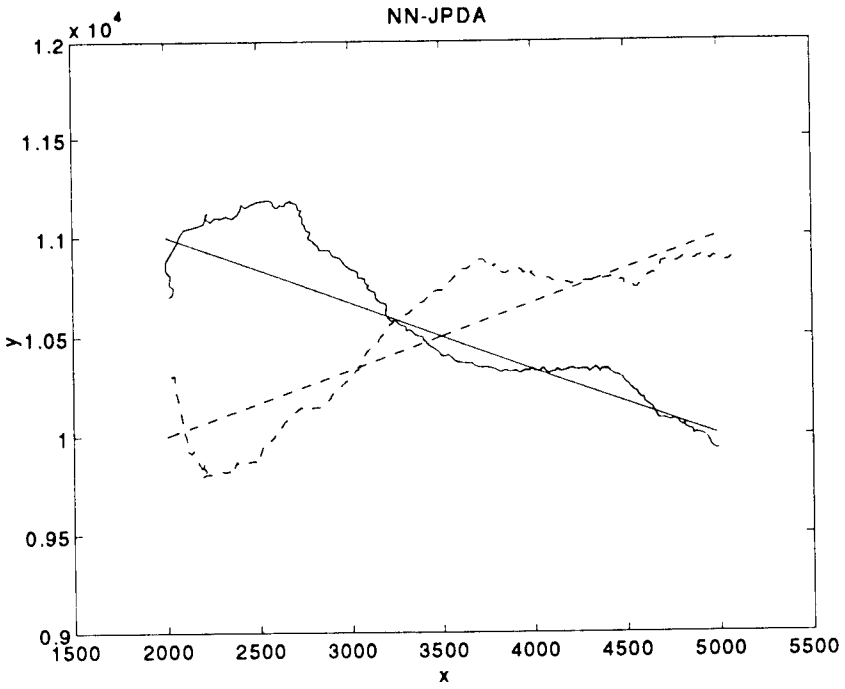


Figure 13: Estimated tracks of the targets using the NN-JPDA algorithms.

TARGET DETECTION WITH SYNTHETIC APERTURE RADAR AND COHERENT SUBTRACTION

Jian Li
Assistant Professor
Department of Electrical Engineering

216 Larsen Hall
University of Florida
Gainesville, FL 32611

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

August 1993

TARGET DETECTION WITH SYNTHETIC APERTURE RADAR AND COHERENT SUBTRACTION

Jian Li
Assistant Professor
Department of Electrical Engineering
University of Florida
Gainesville, FL 32611

Abstract

This report considers target detection with synthetic aperture radar and coherent subtraction. We shall show with experimental data that the coherent subtraction technique may be used to surpress Gaussian outliers and obtain approximate Gaussian distributions for clutter and noise. We shall also derive generalized likelihood ratio (GLR) detection algorithms that may be used with SAR images that are obtained with coherent subtraction. We shall analytically compare the performance of a) a single pixel detector, b) a detector using a complete knowledge of the target signature information and known orientation information, c) a detector using an incomplete knowledge of the target signature information and known orientation information, d) a detector using unknown target signature information and known orientation information, and e) a detector using unknown target signature information and unknown orientation information.

TARGET DETECTION WITH SYNTHETIC APERTURE RADAR AND COHERENT SUBTRACTION

Jian Li

I. Introduction

Synthetic aperture radar (SAR) technology may be used to detect radar targets of interests. High resolution SAR technology is especially useful for detecting small radar targets embedded in strong ground clutter such as in foliage. In this report, we shall consider target detection algorithms that may be used with high resolution SAR.

The first problem encountered by radar target detection researchers is how to describe and deal with the statistical properties of radar clutter and noise. In general, radar clutter and noise do not satisfy Gaussian distribution. Many statistical models, such as the well-known log-normal, Weibull, and K-distributions, have been proposed to describe the clutter and noise statistics. Although these distributions may provide better statistical models for the clutter and noise than the Gaussian distribution, the detectors that are derived based on these models may be very complicated and may involve an expensive multidimensional search over the parameter space. It is also difficult to analyze the performance of these detectors, i.e., it is difficult, if not impossible, to derive the probability of detection and probability of false alarm for such a detector, from which the threshold of the detector is determined.

Alternatively, the SAR images may be preprocessed so that the Gaussian assumption for clutter and noise is approximately valid. For example, Hunt and Cannon [1] and Reed and Yu [2] considered removing local means as such a preprocessing technique. The detectors obtained from Gaussian assumptions may avoid the multidimensional search over the parameter space. It is also much easier to analyze the performance of such detectors since many Gaussian assumption based results exist in the literature.

In this report, we consider using coherent subtraction as a preprocessing technique for SAR images. We shall show the validity of the Gaussian assumption of the clutter and noise after the coherent subtraction between two complex SAR images obtained with two identical experiments. One of the two SAR images is assumed to be target free and the other is to be sought for the presence of target. The effectiveness of the coherent subtraction technique is demonstrated with experimental data obtained by ERIM (Environmental Research Institute of Michigan).

Target detection from SAR or optical images has been considered by many authors. For example, Novak, Burl, and Irving [3] considered target detection with a polarimetric SAR. The three output images of the polarimetric SAR are first processed by a polarization whitening filter, which is derived by assuming a K-

distribution for clutter and noise. The output image of the filter is next used with a two-parameter detector for target detection. The target detection in [3] is performed one pixel at a time even though the target may occupy more than one pixel, i.e., even though the target size may be larger than the resolution of the SAR image.

Reed and Yu [2] considered generalized likelihood ratio target detection from a sequence of optical images, which are first preprocessed by removing local means so that the clutter and noise will approximately have the Gaussian distribution. In [2], each target in an image is described by a completely known template or signature with an unknown gain, which is a scalar. The algorithm, however, may not work well with SAR images. For example, for a target in foliage, the SAR target signature may change due to the interaction between target and tree trunks.

Stotts [4] considered detecting several dim targets in an image simultaneously. The image is also first preprocessed by removing local means so that the clutter and noise will approximately have the Gaussian distribution. Each dim target is described by a known template or signature with an unknown gain, which is a scalar. Stotts has shown that simultaneous detection of multiple targets may perform better than separate detection of individual target. In this report, we shall extend this idea of the simultaneous detection of multiple targets to the detection of a target with multiple pixels in a SAR image.

We shall derive generalized likelihood ratio (GLR) detection algorithms that may be used with SAR images that are obtained with coherent subtraction. The target of interest is modeled with a target template, which is large enough to cover the entire target. The size of the target template and the number of complex unknowns in it is determined by the knowledge of the target orientation information and the amount of target signature information known to the detector. Using this unifying framework, we shall analytically compare the performance of a) a single pixel detector, b) a detector using a complete knowledge of the target signature information and known orientation information, c) a detector using an incomplete knowledge of the target signature information and known orientation information, d) a detector using unknown target signature information and known orientation information, and e) a detector using unknown target signature information and unknown orientation information. We shall derive the probability of detection and the probability of false alarm of each detector. To achieve a constant false alarm rate (CFAR), each detector threshold is simply a function of the dimensional parameters of the detection problem and the desired probability of false alarm.

In Section I, we use the experimental data from ERIM to demonstrate the effectiveness of coherent subtraction as a preprocessing technique. In Section II, we formulate the target detection problem. In Section III, we present an optimal GLR detector and discuss its performance. In Section IV, we present a practical GLR detector and its performance. In Section V, we apply the practical detector to the experimental data from ERIM. Finally, Section VI contains our conclusions.

II. Coherent Subtraction

We consider below the effect of coherent subtraction as a preprocessing technique for SAR images on the statistical properties of clutter and noise. The experimental data used to demonstrate this effect are obtained by ERIM (Environmental Research Institute of Michigan) with a portable rail SAR [5]. The frequency band of the SAR is between 400 and 1300 MHz and the depression angle is 30°. The data we shall use were obtained when both the transmitter and receiver of the SAR are horizontally linearly polarized. We downsampled the original images presented in [5] by a factor of two in range and by a factor of six in azimuth since the original images are oversampled.

Figures 1(a) and (b) show the 3-dimensional (3-D) plots of the magnitudes of two complex SAR images obtained with two identical experiments. Figure 1(a) shows the 3-D plot with foliage only. The peaks in the figure correspond to the radar returns from tree trunks. Figure 1(b) shows the 3-D plot of a target in foliage. The target is a surrogate M-35 truck rotated 24° counterclockwise from end-on. Figure 1(c) shows the 3-D plot of the magnitude of the coherent subtraction between the two complex SAR images. Figure 2 is similar to Figure 1 except that the truck is broadside. Thus the target return in Figure 2 is much stronger than in Figure 1.

We note that coherent subtraction can effectively suppress the large clutter returns due to tree trunks. The large returns left in Figures 1(c) and 2(c) are due to the target and its surroundings. For example, the darkened peak to the left of the target in Figure 2(c) occurs in Figure 2(a) but not in Figure 2(b). This is because the tree that caused the darkened peak in Figure 2(a) is in the shadow of the target and thus does not show up in Figure 2(b). The presence of the darkened peak in Figure 2(c) is an additional information showing the presence of a target because its presence is due to the interaction between the target and clutter. This information is especially useful when the target return is weak.

Figure 3 shows the histograms of the real and imaginary parts of the clutter and noise in the SAR images in Figure 1 before and after coherent subtraction. The histograms are also compared with Gaussian probability density functions (pdfs). We note that before coherent subtraction, the histograms of the clutter and noise do not match the Gaussian pdfs very well due to the outliers caused by the large tree trunk returns. After coherent subtraction, however, the match is much better due to the suppression of the large tree trunk returns.

We have also found that the clutter and noise pixels in Figure 1 are independent of each other. The problem formulation below will take this result into account.

III. Formulation of the Target Detection Problem

Consider the image obtained with coherent subtraction in which a target may be present. The target may be modeled with a template consisting N pixels. The shape of the template may be arbitrary and the template may consist of areas that are not connected. Among the N pixels, K ($K \leq N$) pixels are assumed to be deterministic unknown complex scalars. The locations of the K pixels depend on the target of interest and its orientation relative to the radar. The rest $N - K$ pixels are assumed zero, i.e., they correspond to the areas of the target that does not generate radar returns.

Let $\mathbf{z}$ denote an $N \times 1$ vector consisting of the pixels of such a template in the presence of clutter and noise. Under hypothesis H_1, the target presence hypothesis, the $\mathbf{z}$ may be written

$$\mathbf{z} = \mathbf{S}\mathbf{b} + \mathbf{n}, \tag{1}$$

where $\mathbf{b}$ is the $K \times 1$ vector consisting of the K deterministic unknown complex scalars of the target. The $\mathbf{S}$ is a nonsingular $N \times K$ matrix describing the locations of the unknown scalars. Only one element in each row and each column of $\mathbf{S}$ is 1 and the rest of the elements are zero. Thus we have

$$\mathbf{S}^H\mathbf{S} = \mathbf{I}_K, \tag{2}$$

where $(\cdot)^H$ denotes the complex conjugate transpose and $\mathbf{I}_K$ denotes the identity matrix of dimension K. The $\mathbf{n}$ denotes the clutter and noise random vector and is assumed zero-mean complex Gaussian with covariance matrix $\sigma^2\mathbf{I}_N$. Under hypothesis H_0, the target absence hypothesis, the $\mathbf{z}$ may be written

$$\mathbf{z} = \mathbf{n}. \tag{3}$$

The problem of interest is to maximize the probability of detection of the target for a given probability of false alarm. We shall consider the generalized likelihood ratio target detection algorithms for the purpose. We shall consider both an optimal detector, where the clutter and noise variance is assumed known, and a practical detector, where the clutter and noise variance is unknown. The effects of the dimensional parameters such as N and K on the performance of the detectors will also be considered through performance analysis of the detectors.

We remark that the above detection problem is a generalized version of the approach of considering one pixel at a time and the approach of using the complete knowledge of the target signature. When we consider detection by using one pixel at a time, we have $N = K = 1$. When the complete knowledge of the target signature is known except for a complex gain, the $\mathbf{S}$ becomes the signature vector $\tilde{\mathbf{s}}$ whose Euclidean norm is 1, the $\mathbf{b}$ becomes the unknown complex scalar gain $\tilde{b}$, and $K = 1$. Through performance analysis of the optimal and practical detectors, we shall understand how the approaches relate to each other.

IV. Optimal Detector and Its Performance

The optimal detector is derived by assuming that the clutter and noise variance, i.e., σ^2, is known. We shall present below the generalized likelihood ratio (GLR) detector under this assumption and also present its performance.

It is shown in Appendix A of [6] that the optimal detector has the form

$$\frac{\mathbf{z}^H \mathbf{S}\mathbf{S}^H \mathbf{z}}{\sigma^2} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{4}$$

The threshold parameter γ is determined according to a given probability of false alarm. The detector (4) maximizes the probability of detection for the given probability of false alarm [7].

It is shown in Appendix B of [6] that the probability of false alarm of the above detector is

$$P_F = \sum_{k=0}^{K-1} \frac{\gamma^{K-1-k}}{(K-1-k)!} \exp(-\gamma). \tag{5}$$

We note that the P_F depends only on K, the number of unknown parameters in a target template, and the threshold parameter γ. The P_F is independent of N, the size of the template. For a given probability of false alarm, the γ in the optimal detector (4) is obtained with (5). It is also shown in Appendix B of [6] that the probability of detection of the above detector is

$$P_D = \exp(-\delta - \gamma) \sum_{i=0}^{\infty} \frac{\delta^i}{i!} \sum_{k=0}^{i+K-1} \frac{\gamma^{i+K-1-k}}{(i+K-1-k)!}, \tag{6}$$

where

$$\delta = \frac{\mathbf{b}^H \mathbf{b}}{\sigma^2}. \tag{7}$$

The δ is the signal-to-clutter-and-noise ratio (SCNR) of the template. Note that δ is the sum of the signal-to-clutter-and-noise ratios of the non-zero pixels in the template. We also note that the P_D is also independent of the template size N. The P_D depends on K, SCNR δ, and the threshold γ.

To compare the performance of the detectors described in Section I, let us consider the effect of K on the performance of the optimal detector. Figure 4 shows the probability of detection as a function of SCNR for different K. It is shown that for a given P_F and a fixed SCNR δ, the P_D of the optimal detector in (4) decreases as K increases. The four performance curves shown in Figure 4 correspond to the four scenarios shown in Figure 5. We remark that for the same target SCNR, the performance of assuming the complete knowledge of the target signature is the same as $K = 1$ in our problem formulation, which is shown in Appendix C of [6]. We note from Figure 4 that the best performance occurs when we use the complete knowledge of the target signature except for the unknown gain and when the target orientation is known. The worst performance occurs for the case of unknown target signature and unknown target orientation.

We note from Figure 4 that for a given P_F, to achieve the same P_D as for the case of $K = 1$, extra SCNR is needed. Figure 6 shows the extra SCNR needed to achieve $P_D = 0.5$ for different probabilities of false alarm P_F. We note that we have similar curves for different probabilities of false alarm P_F. The extra SCNR needed decreases slightly as P_F decreases. We also note that the extra SCNR needed increases slowly as K increases. This result may be explained as follows. As shown in Appendix B of [6], we may rewrite the optimal detector in (4) as

$$\eta' = \frac{\mathbf{z}_{2A}^H \mathbf{z}_{2A}}{K} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'. \tag{8}$$

Under hypothesis H_0, $\mathbf{z}_{2A}$ has the complex Gaussian distribution with zero-mean and covariance matrix $\mathbf{I}_K$. Under hypothesis H_1, $\mathbf{z}_{2A}$ has the complex Gaussian distribution with mean $\mathbf{b}/\sigma$ and covariance matrix $\mathbf{I}_K$. Thus under hypothesis H_0, the mean and variance of η' are 1 and $1/K$, respectively. Under hypothesis H_1, the mean and variance of η' are $1 + \delta/K$ and $1/K + \delta/K$, respectively, where δ is the SCNR given in (7). For a fixed SCNR δ, increasing K has two opposite effects on η'. On the one hand, increasing K decreases the variance of η' under both hypotheses. This effect helps increase the probability of detection for a given probability of false alarm. On the other hand, increasing K reduces the difference between the means of η' under H_0 and H_1. This effect reduces the probability of detection for a given probability of false alarm. The combined effect of K on the optimal detector is shown in Figure 6.

Although it is the best to use the complete knowledge of the target signature for target detection, the target signature may not be completely known for SAR images and even when known, the signature may change due to the interaction between the target and clutter and other factors. The change of signature may result in severe detector performance degradation. Let $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{s}}_1$ be the assumed and true target signature vectors, respectively, where both $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{s}}_1$ have Euclidean norm 1. It is shown in Appendix C of [6] that the probability of false alarm for this case is the same as (5) with $K = 1$. The probability of detection for this case has the form of (6) with $K = 1$ and

$$\delta = \frac{|\tilde{b}|^2}{\sigma^2}\rho, \tag{9}$$

where ρ is the SCNR loss factor as a result of the signature mismatch, i.e.,

$$\rho = |\tilde{\mathbf{s}}_1^H \tilde{\mathbf{s}}|^2. \tag{10}$$

It is easy to see that $0 \leq \rho \leq 1$. Figure 7 shows the SCNR loss as a function of ρ.

We now make the comparison between using the incomplete knowledge of the target signature information and the approach of using the complete knowledge of the target signature information except for the unknown gain. For example, when using an incomplete target signature described by $K = 20$ unknown parameters, the extra SCNR needed to achieve the same probability of detection as when using the complete target signature information ($K = 1$) is about 3 dB, as shown in Figure 6. If we know that the mismatch

between the assumed and true target signatures will result in a SCNR loss factor ρ of no less than 0.5, then it is better to use the complete knowledge of the target signature information in the detector. Otherwise, it is better to use the incomplete signature information and assume $K = 20$ unknowns in the target template.

The comparison between considering one target template at a time and one pixel at a time is also clear. For a template with $K = 20$ unknown parameters, for example, the extra SCNR needed to achieve the same probability of detection as for $K = 1$ is about 3 dB, as shown in Figure 6. If the template SCNR δ is at least 3 dB larger than the SCNR of the highest pixel in the template, then using the target template model is better than considering each pixel at a time. Otherwise, it is better to consider each individual pixel at a time. For the best case where all $K = 20$ pixels in the target template have equal magnitude, the template SCNR is about 13 dB more than the individual pixel SCNR. Then as compared with using one pixel at a time, the net gain of using the target template for target detection is about 10 dB because of the 3 dB loss due to the increased number of unknowns.

V. Practical Detector and Its Performance

In the previous section, we have studied the performance of the optimal detector, which assumes that the clutter and noise variance is known. In practice, however, the clutter and noise variance is unknown. We present below a practical detector for this practical situation. Our approach is similar to the one developed by Kelly [8] for target detection with a phased array airborne surveillance radar.

The practical detector we shall present utilizes both primary and secondary data of a SAR image for target detection. The data vector $\mathbf{z}$, from which the target presence is sought, is referred to as the primary data. The secondary data vectors $\mathbf{z}(1), \mathbf{z}(2), \cdots, \mathbf{z}(L)$ are assumed to be target free, i.e., they represent the target free background of a SAR image. They are assumed to have the same statistics as the primary data vector $\mathbf{z}$ under hypothesis H_0 and are statistically independent of each other and $\mathbf{z}$. The secondary data are useful for estimating the clutter and noise variance in the practical detector.

It is shown in Appendix D of [6] that the practical detector has the form

$$\eta = \frac{\mathbf{z}^H \mathbf{z} + \sum_{l=1}^{L} \mathbf{z}^H(l)\mathbf{z}(l)}{\mathbf{z}^H \left(\mathbf{I}_N - \mathbf{S}\mathbf{S}^H\right) \mathbf{z} + \sum_{l=1}^{L} \mathbf{z}^H(l)\mathbf{z}(l)} \underset{H_0}{\overset{H_1}{\gtrless}} \xi. \tag{11}$$

The threshold parameter ξ is determined according to a given probability of false alarm. The detector (11) maximizes the probability of detection [7].

It is shown in Appendix E of [6] that the probability of false alarm of the above detector is

$$P_F = \sum_{k=0}^{K-1} \frac{(\xi - 1)^{K-k-1}}{\xi^{M+K-k-1}} \begin{pmatrix} M+K-k-2 \\ K-k-1 \end{pmatrix}, \tag{12}$$

$$M = NL + N - K. \tag{13}$$

We note that for the practical detector, the P_F depends on the dimensional parameters M and K and the threshold parameter ξ. For a given probability of false alarm, the ξ in the practical detector (11) is obtained with (12).

It is also shown in Appendix E of [6] that the probability of detection of the above detector is

$$P_D = 1 - \exp(-\delta_1) \sum_{k=0}^{M-1} \sum_{i=0}^{M-k-1} \begin{pmatrix} M+K-k-2 \\ K+i-1 \end{pmatrix} \frac{\delta_1^i}{i!} \frac{(\xi-1)^{K+i}}{\xi^{M+K-k-1}}, \tag{14}$$

where

$$\delta_1 = \frac{\mathbf{b}^H \mathbf{b}}{\sigma^2 \xi}. \tag{15}$$

We note that P_D depends on the dimensional parameters M and K and δ_1, which depends on the target SCNR $\delta = \mathbf{b}^H \mathbf{b}/\sigma^2$ and the threshold ξ.

We remark that (12) and (14) also hold when L is not an integer but a rational number such that NL is an integer. The detector (11) may be changed slightly to accommodate the fact that L is not an integer.

We now examine how the different dimensional parameters affect the performance of the detector in (11). The performance of the detector is mainly determined by $M = NL + N - K$ and K, which may be seen from the performance analysis in Appendix E of [6]. Consider first $M = NL + N - K$, the total number of signal free pixels in the primary and secondary data vectors. The larger the number of signal free pixels M, the better the estimate of the clutter and noise variance, and hence the closer the performance of the practical detector to that of the optimal detector. This result may be observed from Figure 8, which shows the probability of detection as a function of SCNR for different M when $K = 2$ and $P_F = 10^{-8}$.

Consider next the effect of K, the number of unknown parameters in the target template, on the performance of the practical detector. We first explain that for large K, more number of signal free pixels M is needed by the practical detector to achieve similar performance as the optimal detector. As shown in Appendix E of [6], we may rewrite the practical detector in (11) as

$$\eta_1' = \frac{\mathbf{z}_{1A}^H \mathbf{z}_{1A}/K}{\left[\mathbf{z}_{1B}^H \mathbf{z}_{1B} + \sum_{l=1}^{L} \mathbf{z}_1^H(l)\mathbf{z}_1(l)\right]/M} \underset{H_0}{\overset{H_1}{\gtrless}} \xi'. \tag{16}$$

Under both hypotheses, the $\mathbf{z}_1(l)$ has the complex Gaussian distribution with zero-mean and covariance matrix $\mathbf{I}_N$. Under both hypotheses, $\mathbf{z}_{1B}$ has the complex Gaussian distribution with zero-mean and covariance matrix $\mathbf{I}_{N-K}$. Under hypothesis H_0, $\mathbf{z}_{1A}$ has the complex Gaussian distribution with zero-mean and covariance matrix $\mathbf{I}_K$. Under hypothesis H_1, $\mathbf{z}_{1A}$ has the complex Gaussian distribution with mean $\mathbf{b}/\sigma$ and covariance matrix $\mathbf{I}_K$. Then under hypothesis H_0, the mean and variance of the numerator of η_1' are 1 and $1/K$, respectively. Under hypothesis H_1, the mean and variance of the numerator of η_1' are $1 + \delta/K$

and $1/K + \delta/K$, respectively, where δ is the SCNR given in (7). Under both hypotheses, the mean and variance of the denominator of η_1' are 1 and $1/M$, respectively. We note that for large K and small M, the performance of (16) will be affected by the variance $1/M$ of the denominator of η_1'. Thus for large K, M must also be large in order for the practical detector in (16) to achieve similar performance as the optimal detector in (8). We found through numerical examples that for $1 \leq K \leq 1000$, the performance differences between the practical and optimal detectors are similar for different K when M is proportional to $K^{2/3}$. Figure 9 shows the probability of detection as a function of SCNR for different K when $M = 48K^{2/3}$ and $P_F = 10^{-8}$. The figure also shows the performance of the optimal detector for comparison. Note that the optimal and practical detectors have similar performances.

Figure 10 shows the probability of detection as a function of probability of false alarm, i.e., the receiver operating characteristic of the practical detector, for different K and SCNR when $M = 118$ and $P_F = 10^{-8}$. We note that to achieve the same P_D as for $K = 1$, we must significantly increase P_F for $K = 10$ and fixed SCNR. With SCNR increased by 5 dB more for $K = 10$ than for $K = 1$, however, we will have a better P_D for a given P_F.

VI. Target Detection with Experimental Data

We consider first the performance of the practical detector when used with the experimental data shown in Figure 1. Figure 11 shows the detection results before and after coherent subtraction when $M = 48$, $N = K = 1$, and $P_F = 10^{-8}$. Figure 11 shows the generalized likelihood ratios that are above the detection threshold. We note that before coherent subtraction, although the presence of the target is detected, the large tree trunk returns also result in a false alarm. With coherent subtraction, however, the false alarm is eliminated.

Consider next the experimental data shown in Figure 2. We note that the target occupies more than one pixel. Let us assume that the target orientation is known and use an incomplete target signature described by $K = 20$ unknowns. Then the template SCNR for the target is approximately 35 dB for the data shown in Figure 2(c). The largest pixel SCNR for the target is about 25 dB. Thus compared with using each single pixel for target detection, using the template with $K = 20$ unknowns results in a net gain of about 7 dB because of the 3 dB loss due to the increased number of unknowns.

VII. Conclusions

We have considered target detection with synthetic aperture radar and coherent subtraction. We have shown with experimental data that the coherent subtraction technique may be used to suppress outliers and obtain approximate Gaussian distributions for clutter and noise. We have also derived generalized likelihood ratio (GLR) detection algorithms that may be used with SAR images that are obtained with coherent subtraction. Through performance analysis, we have analytically compared the performance of a) a single pixel detector, b) a detector using a complete knowledge of the target signature information and known orientation information, c) a detector using an incomplete knowledge of the target signature information and known orientation information, d) a detector using unknown target signature information and known orientation information, and e) a detector using unknown target signature information and unknown orientation information.

References

[1] B. R. Hunt and T. M. Cannon, "Nonstationary assumptions for gaussian models of images," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 876–881, December 1976.

[2] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Transactions on Signal Processing*, vol. 38, pp. 1760–1770, October 1990.

[3] L. M. Novak, M. C. Burl, and W. W. Irving, "Optimal polarimetric processing for enhanced target detection," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 29, pp. 234–244, January 1993.

[4] L. B. Stotts, *Moving-Target Detection Techniques for Optical-Image Sequences.* Ph.D. thesis, University of California, San Diego, California, 1988.

[5] D. R. Sheen, S. C. Wei, T. B. Lewis, and S. R. D. Graaf, "Ultrawide bandwidth polarimetric SAR imagery of foliage-obscured objects," *SPIE Proceedings on OE/LASE,* Vol. 1875, Los Angeles, CA, January 1993.

[6] J. Li and E. G. Zelnio, "Target detection with synthetic aperture radar and coherent subtraction," to be submitted to *IEEE Transactions on Aerospace and Electronic Systems.*

[7] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I.* New York, NY: John Wiley & Sons Inc., 1968.

[8] E. J. Kelly, "An adaptive detection algorithm," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 22, pp. 115–127, March 1986.

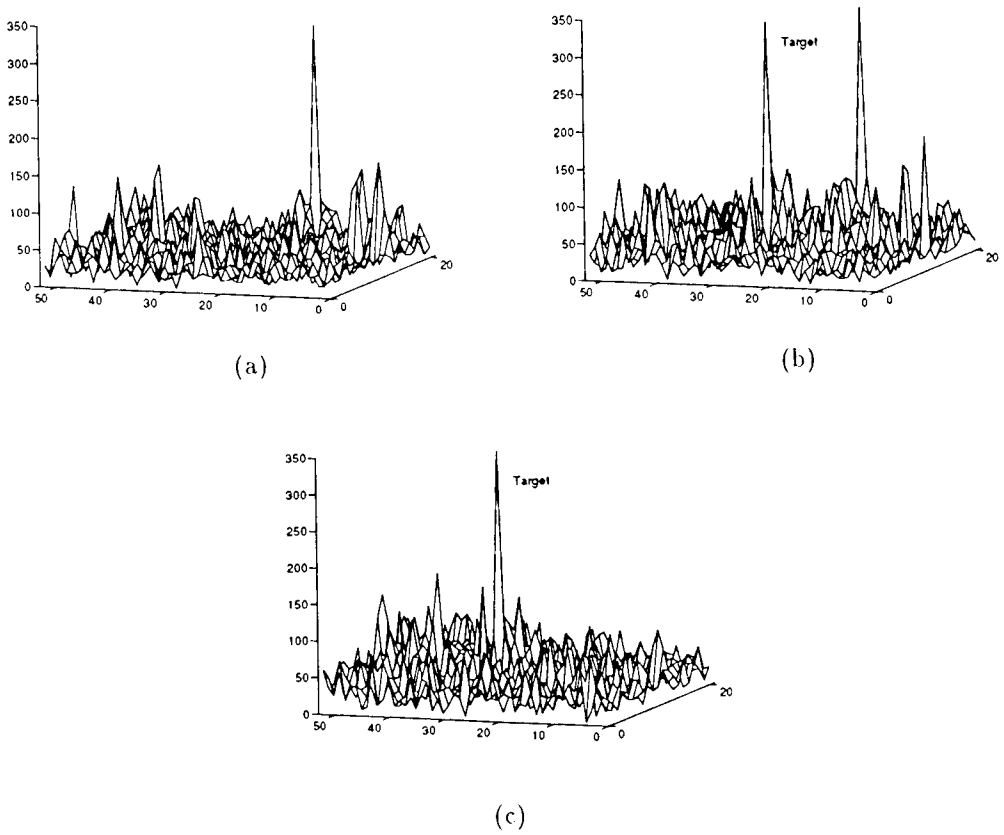


Figure 1: 3-dimensional plots of the magnitudes of complex SAR images when the target is a surrogate M-35 (end-on). (a) Foliage only. (b) Target in foliage. (c) After coherent subtraction.

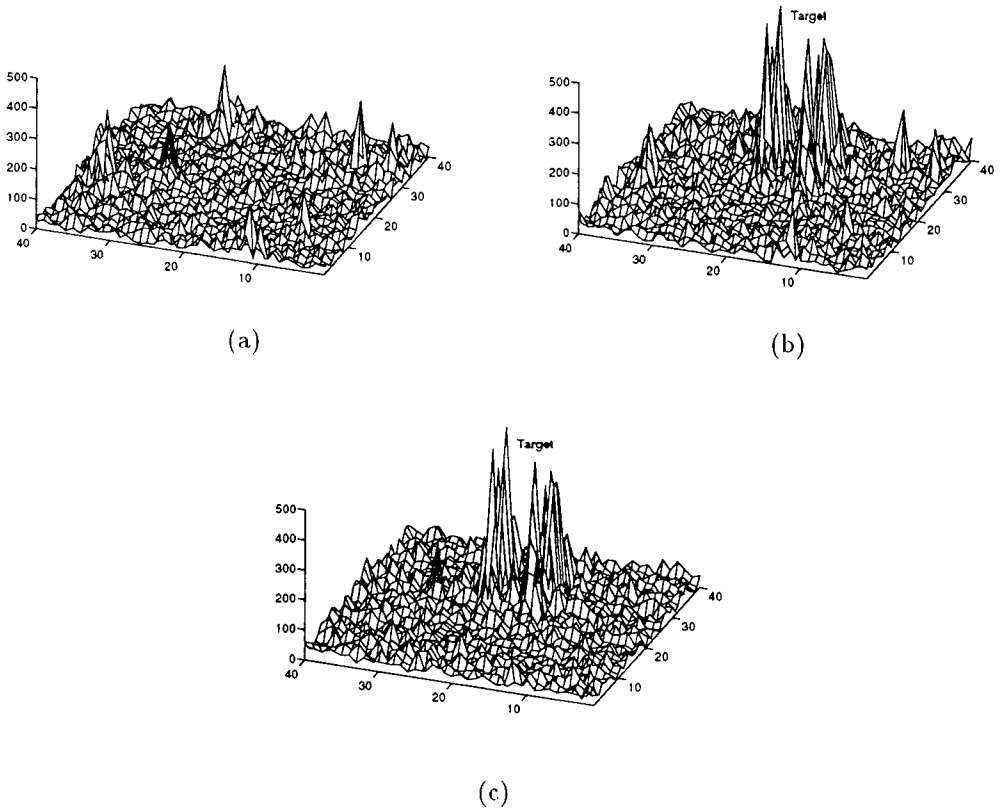


Figure 2: 3-dimensional plots of the magnitudes of complex SAR images when the target is a surrogate M-35 (broadside). (a) Foliage only. (b) Target in foliage. (c) After coherent subtraction.

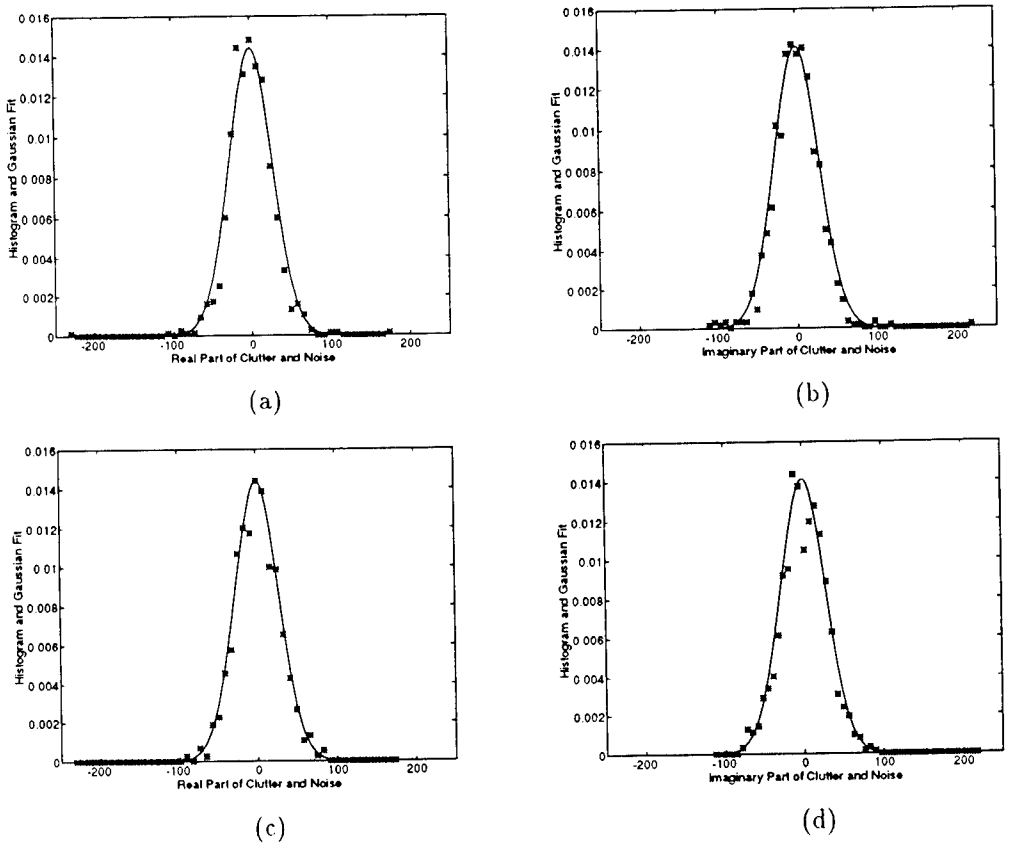


Figure 3: Histograms of clutter and noise before and after coherence subtraction and Gaussian fit. (a) Real part before coherent subtraction. (b) Imaginary part before coherent subtraction. (c) Real part after coherent subtraction. (d) Imaginary part after coherent subtraction.

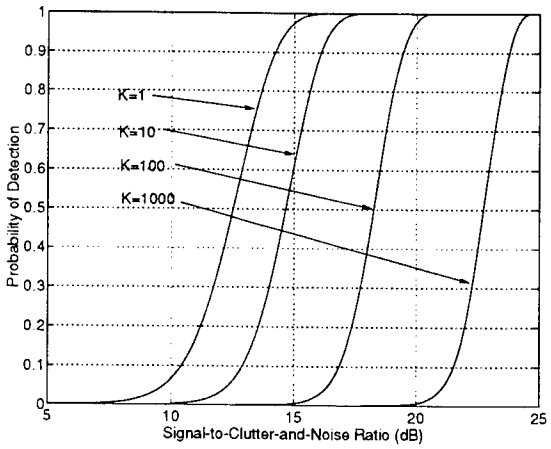


Figure 4: Probability of detection of the optimal detector vs. signal-to-clutter-and-noise ratio for different number of unknown parameters K with $P_F = 10^{-8}$.

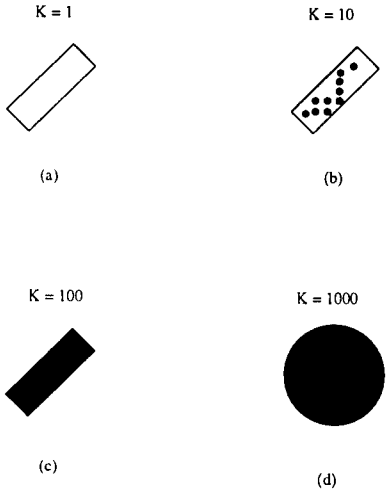


Figure 5: Target templates examples for (a) completely knowledge of target signature information except for the unknown complex gain and known orientation information, (b) incomplete knowledge of target signature information and known orientation information, (c) unknown target signature information and known orientation information, and (d) unknown target signature information and unknown orientation information.

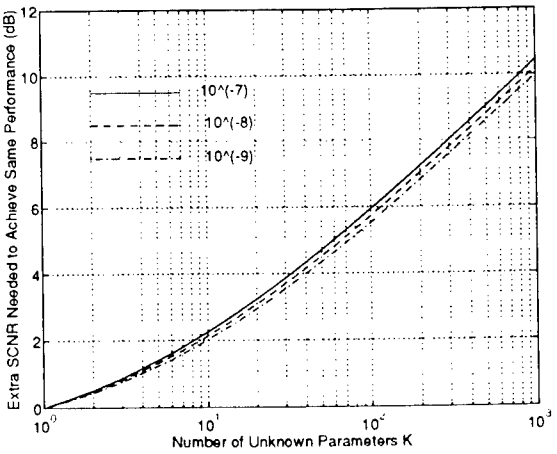


Figure 6: As compared with $K = 1$, extra SCNR needed for the optimal detector to achieve $P_D = 0.5$ vs. K for different probabilities of false alarm P_F.

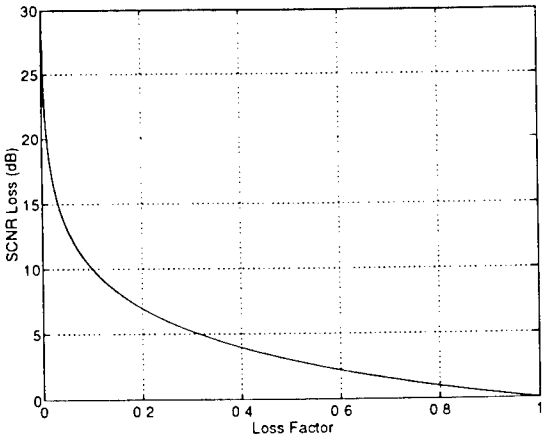


Figure 7: Signal-to-clutter-and-noise ratio (SCNR) loss for the optimal detector as a result of target signature mismatch vs. SCNR loss factor.

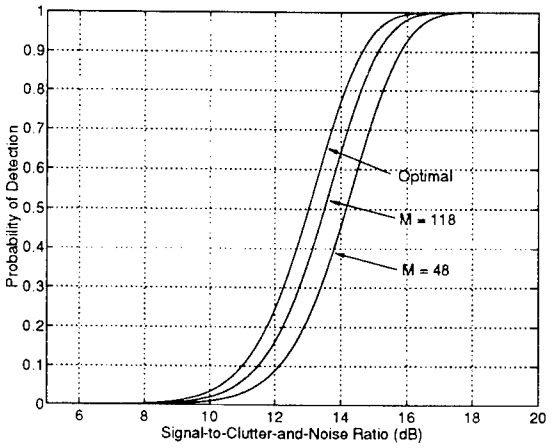


Figure 8: Probability of detection of the practical detector vs. signal-to-clutter-and-noise ratio for different L when $K = 2$ and $P_F = 10^{-8}$.

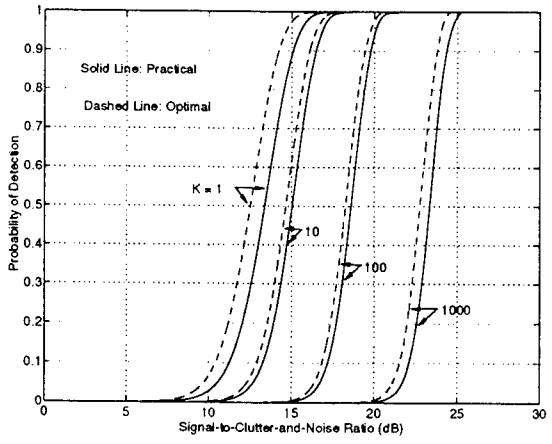


Figure 9: Probability of detection vs. signal-to-clutter-and-noise ratio for different K when $M = 48K^{2/3}$ and $P_F = 10^{-8}$.

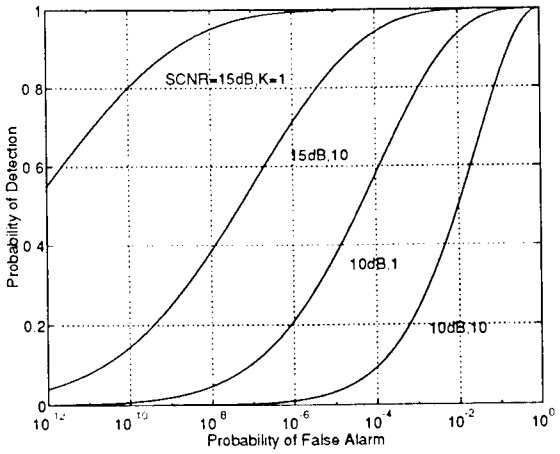
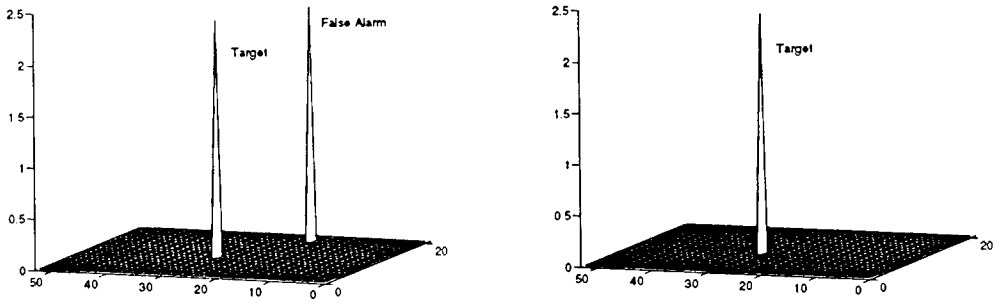


Figure 10: Receiver Operating Characteristic of the practical detector for different K and different signal-to-clutter-and-noise ratio when $M = 118$ and $P_F = 10^{-8}$.



(a) (b)

Figure 11: Target detection (end-on) with $N = K = 1$ and $M = 48$. (a) Before coherent subtraction. (b) After coherent subtraction.

BUILT-IN SELF-TESTING OF RANDOM-ACCESS MEMORIES

Carol Q. Tong

Assistant Professor

Department of Electrical Engineering

Colorado State University

Fort Collins, CO 80523

Final Report for:

Summer Faculty Research Program

Wright Laboratory

Sponsored by:

Air Force Office of Scientific Research

Bolling Air Force Base, Washington, D.C.

August 1993

BUILT-IN SELF-TESTING OF RANDOM-ACCESS MEMORIES

Carol Q. Tong

Assistant Professor

Department of Electrical Engineering

Colorado State University

Abstract

In this research, the problem of testing random-access memories (RAM) is considered. Due to the increasing density of RAM chips, built-in self-testing (BIST) has to be applied in order to save the time and cost of testing. Pseudo-random testing is evaluated as a BIST structure for RAMs. The fault coverage and test length of random testing for RAM is studied. The results show that pseudo-random testing is suitable for testing RAM using BIST.

BUILT-IN SELF-TESTING OF RANDOM ACCESS MEMORIES

Carol Q. Tong

1. Introduction

Testing semiconductor memories is becoming very important because of the increasing size and complexity of memory chips. As a result, testing with faster speed, less cost and higher reliability has become a very important issue.

This research report will discuss the problem of testing random access memories (RAM) in a computer-aided design (CAD) tool - VLSI Testability Synthesis Tool (VTST) [1] - developed at Wright State University. VTST is a synthesis tool for designing VLSI circuits incorporating built-in self-testing (BIST) structure.

A RAM chip consists of an array of memory cells. Depending on the structure of the storage element in each cell, it can be either static RAM (SRAM) or dynamic RAM (DRAM). The RAM chip also has an address decoder, address and data registers, and read/write control logic. Figure 1 shows the structure of an embedded RAM [2]. The memory array is assumed to be of the size 2^mxn, where m is the number of address lines, or the number of bits in the binary representation of an address, and n is the number of binary bits stored at each address.

Testing of RAM has been studied over many years. But only in the recent years, when the speed and cost of testing are becoming major issues, was built-in self-testing introduced to the area of testing regular structures such as RAM. BIST is expected to play a key role in the next generation of multimegabit memory devices. BIST is just the ability of a circuit or system to test itself [2]. A BIST design consists of the following parts: on-chip test generator, on-chip response analyzer and the control logic. This report will focus on the test generation part of BIST for RAM. Because of the high costs incurred by off-line testers as memories become larger, BIST has become a necessity [3].

In the next section, we will look at issues related to RAM testing, such as fault modeling. In Section 3, BIST for RAM will be discussed in detail, and the BIST design in VTST will be evaluated. Section 4 concludes the report.

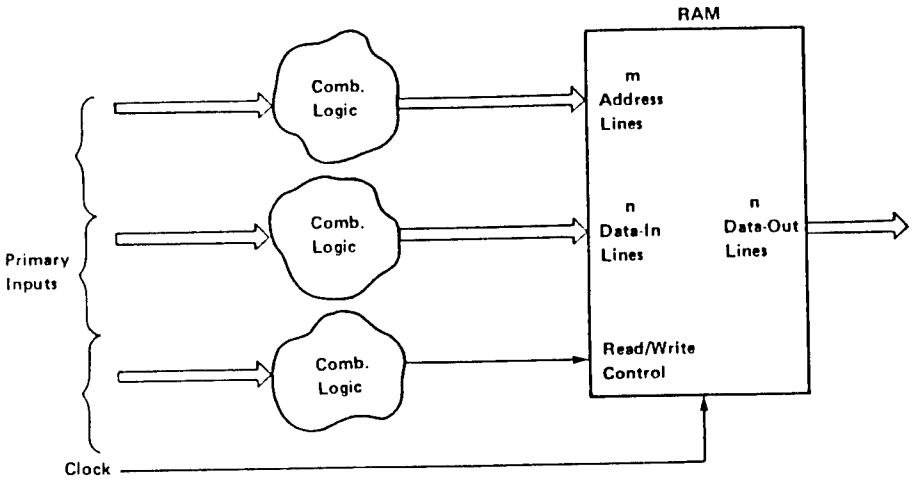


Figure 1: An embedded RAM

2. Preliminary Research on Various Issues of RAM Testing

In this section, we will introduce the fault models used in testing RAM and present some preliminary research efforts in testing RAM.

A. Fault modeling in RAM

The important fault models for testing RAM include stuck-at fault, transition fault, coupling fault, pattern-sensitive fault, decoder fault and read/write logic fault[3]. A memory cell is said to be stuck-at-1 (stuck-at-0) if its content is always logic 1 (0) no matter what is written into it. A memory cell is said to have a transition fault if the content of the cell fails to transit from 0 to 1 (1 to 0) when an 1 (0) is written into it. A pair of memory cells is said to be coupled if a transition in one of the cells changes the contents of the other cell from 0 to 1 or from 1 to 0. A memory cell is said to have a pattern-sensitive fault if its content is changed by a pattern of 0's and 1's or various transitions in a group of other memory cells. A decoder (or read/write) fault is a fault that occurs in the decoder (or read/write) circuit (outside the memory array). It has been shown that most decoder faults and read/write logic faults can be mapped to faults in the memory array [3].

B. Test algorithms and fault coverage

Deterministic testing of RAM has a long history and many algorithms exist, such as Marching Test, for generating the test vectors [3]. Each individual algorithm covers certain types of faults and the test length varies. For example, Marching Test has a test length of $O(N)$, where N is the total number of cells in the memory array. And it detects all the stuck-at faults and some coupling faults.

We studied the possibility of implementing one of these algorithms in VTST and decided that it is not feasible. The reason is that in order to implement an algorithm in BIST, a sequential circuit has to be designed to generate test vectors. If the size of the memory array changes, the design of the sequential circuit also has to be changed. This is impractical for use in a synthesis tool like VTST.

C. Parallel testing

In order to speed up the testing of memories, parallel testing has been proposed and implemented by some researchers [4]-[6]. The basic idea is that if a large memory array can be divided into many subarrays

and these subarrays are independent of each other, then these subarrays can be tested in parallel. So if the RAM chip internal structure is known, parallel testing can be applied to independent subarrays. But most of the time, the internal structure of a RAM chip is unknown to the users. Hence in VTST, we can not take advantage of parallel testing to shorten the testing time.

D. Other techniques

There are also other techniques that can be used to enhance the RAM testing. For example, using error-correcting code (ECC), self-checking can be done. If the data in a memory array can be encoded with ECC, then by checking the data read from the array, the faults in the data can be detected and corrected automaticly. This implies that we can save the time and hence the cost used for generating tests. Since this technique requires extra hardware overhead besides the overhead introduced by BIST, it is not yet adopted in VTST. But it can be an approach to reduce the RAM testing time in the future,

Another possible application of ECC in RAM testing is that it can detect transient faults. A transient fault is a fault that occurs temporarily and is harder to be detected than permanent faults, like those defined in **A.** By using ECC, whenever the transient fault occurs, it can be detected and corrected.

3. BIST of RAM

In this section, we will look at the BIST technique used in VTST.

3.1 Pseudorandom test generation

Pseudorandom testing is applying test vectors which have many random characteristics but which are generated deterministically and, therefore, repeatable. A Linear Feedback Shift Register (LFSR) is used to generate the pseudorandom test vectors. Figure 2 shows an example of an n-stage LFSR. The outputs of the LFSR are Q_i, where $i = [0, 1, ...n]$. c_i is a binary constant. When $c_i = 1$, a connection exits. When $c_i = 0$, no connection exists.

Every LFSR has a characteristic polynomial associated with it. This characteristic polynomial determines the feedback connections, and consequently the period of the vector sequences. An n-stage LFSR which produces a maximum-length sequence has a period of $2^n - 1$.

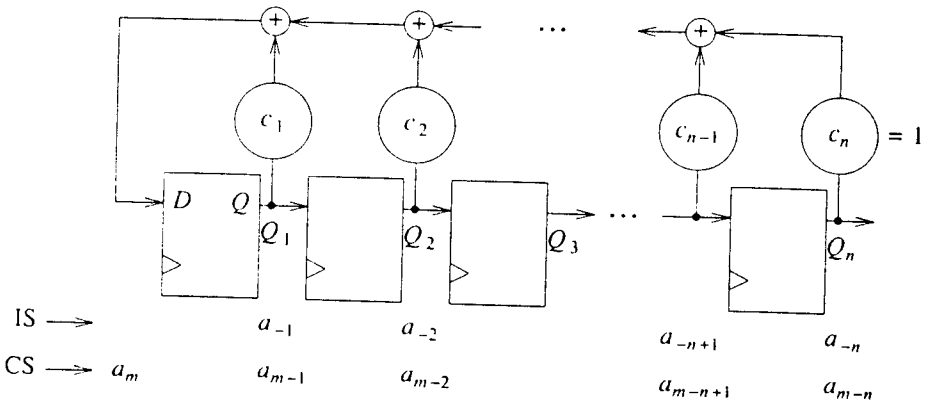


Figure 2: An n-stage LFSR

Since the test vectors are generated randomly, the actual fault coverage is not known unless fault simulation is done. Fault coverage is a quantitative measure used to determine the quality of a test set. It is defined as the ratio of faults detected by the test set to the total number of simulated faults in the circuit. The simulation and therefore the testing can be stopped whenever the desired fault coverage is reached.

3.2 Fault coverage and test length

When testing RAM using an LFSR, the address is randomly generated. Also random data is written to or read from a word at that address. First, some notations should be introduced. Assume the memory array under test has 2^m words. They define the probability of different operations:

p_d: data line has logic value 1

p_a: address line has logic value 1

p_c: control line has logic value 1 — write

$1 - p_c$: control line has logic value 0 — read

r: selecting one address with h 1's and $m - h$ 0's

$$r = {p_a}^h (1 - p_a)^{m-h}$$

$p_1 = p_d p_c r$ — write 1 to an arbitrary bit

$p_0 = (1 - p_d) p_c r$ — write 0 to an arbitrary bit

$p_r = (1 - p_c) r$ — read from the selected address

Note that $r = p_1 + p_0 + p_r$ since after an address is selected, one of the three operations has to happen at this address. Next we will look at using a Markov chain to model the testing processes and obtaining the test length for a required fault coverage.

A. Markov chain

1) stuck-at fault

The Markov chain for stuck-at-0 fault is shown in Figure 3 [2]. S_0 is the state when the cell should have logic value 0, S_1 is the state when the cell should have logic value 1, and S_2 is the state when the fault is detected. The probability of reaching state S_2 is the probability that the fault is detected. Hence the test length is the solution t of the equation $S_2(t) = 1 - e_{th}$, where e_{th} is the desired escape probability. The

initial state probabilities are $S_0(0) = 1 - i_0$, $S_1(0) = i_0$, $S_2(0) = 0$, where i_0 is the probability that the cell is initialized to 1. From the Markov chain and using the conditional probabilities, one can get the difference equations describing the transitions between states and the upper bound for test length [2]:

$$S_0(t) = (1 - p_1)S_0(t-1) + p_0 S_1(t-1)$$

$$S_1(t) = p_1 S_0(t-1) + (1 - p_0 - p_r)S_1(t-1)$$

$$S_2(t) = p_r S_1(t-1) + S_2(t-1)$$

By using Z-transform, one can obtain $S_2(t)$ and hence the test length:

$$T_0 = \left\lceil \frac{\ln(\frac{2\alpha e_{th}}{1+\alpha-2(1-p_c)i_0})}{\ln(1 - \frac{(1-\alpha)r}{2})} \right\rceil$$

where α is a constant depending on the the size of the RAM chip.

Similarly, one can derive the test length for stuck-at-1 fault, T_1 [2].

2) transition fault

The same Markov chain can be used for transition fault detection as for stuck at fault. And the required test length can be derived similarly. For example, if there is a 0 to 1 transition fault (which means that this transition can not occur), then the test length is the solution of the equation $S_2(t) \bullet S_0(t-2) = 1 - e_{th}$.

3) coupling fault

Let us look at the Markov chain for detecting one of the possible coupling faults. The others can be treated similarly. Figure 4 shows the Markov chain for detecting the following coupling fault: a 0 to 1 transition in cell j causes a change of value in cell i.

In Figure 4, $S_{0,i}$ is the state when cell j has logic 0, cell i can be 0 or 1. $S_{0,i'}$ is the state when cell j has logic 0, cell i's value has been changed because of the coupling fault. And $S_{1,i}$ is the state when cell j has logic 1, cell i can be 0 or 1. $S_{1,i'}$ is the state when cell j has logic 1, cell i's value has been changed because of the coupling fault. S_2 is the state when the fault is detected. Note that $p_i = p_c r = p_1 + p_0$, which is the probability that cell i is being written into. Similar to stuck-at fault detecting, we can obtain the difference

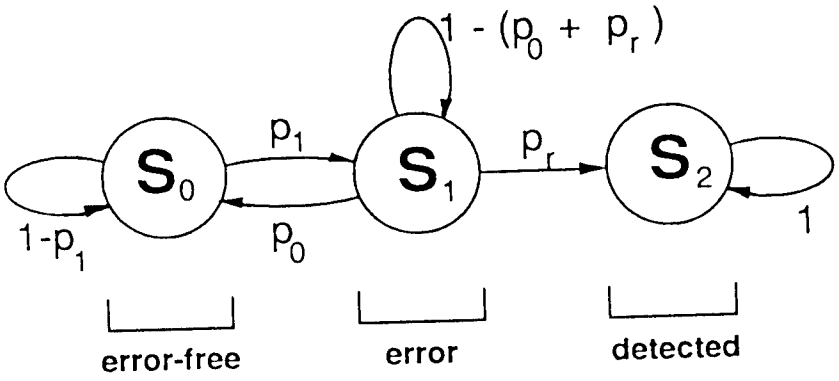


Figure 3: Markov chain of detecting a s-a-0 fault

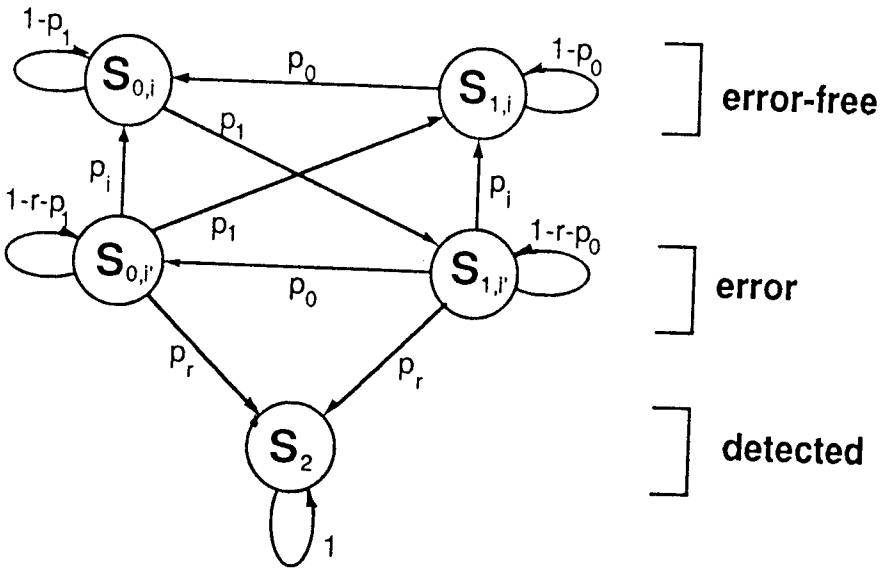


Figure 4: Markov chain of detecting a coupling fault

equations for the transitions between states:

$$S_{0,i}(t) = (1 - p_1)S_{0,i}(t-1) + p_0 S_{1,i}(t-1) + p_i S_{0,i'}(t-1)$$

$$S_{1,i}(t) = (1 - p_0)S_{1,i}(t-1) + p_1 S_{0,i'}(t-1) + p_i S_{1,i'}(t-1)$$

$$S_{0,i'}(t) = (1 - r - p_1)S_{0,i'}(t-1) + p_0 S_{1,i'}(t-1)$$

$$S_{1,i'}(t) = p_1)S_{0,i}(t-1) + (1 - r - p_0)S_{1,i'}(t-1)$$

$$S_2(t) = p_r)S_{0,i'}(t-1) + p_r S_{1,i'}(t-1) + S_2(t-1)$$

Using Z-transform we can get the expression for S_2 and then set $S_2(t) = 1 - e_{th}$ to obtain the test length.

B. Comparison with deterministic testing algorithms:

Now let us compare the test lengths of different approaches. Deterministic algorithms such as Marching Test require test length of order N, which is the number of total cells. Pseudorandom testing, on the other hand, requires a test length (e.g. T_0) nearly linear to the number of words W in the memory array. It can be shown that T_0 and T_1 are less than the test length in deterministic algorithms. Yet note that a deterministic algorithm often covers more than just stuck-at faults. Another comparison can be made with a commercial algorithm (LSI Logic). That algorithm covers stuck-at fault as well as decoder fault. Its required test length is $T = 8W \log_2(W) + 40W$. And it can be shown to be also greater than T_0 and T_1.

C. Simulation limitation

As noted before, when pseudorandom testing is used, fault simulation needs to be done in order to determine the exact fault coverage. The test lengths obtained using Markov chains are upper bounds, not the exact test lengths needed to obtain certain fault coverage. If fault simulation can be done, the actual test lengths can be much smaller than the upper bounds. But at present, there is no RAM simulator available. Hence a RAM simulation tool needs to be developed in order to reduce the test lengths in VTST. Note that the simulation needs to be implemented at the transistor level since the pass transistors in the memory cell can not be modeled at the gate level.

4. Conclusions

BIST for testing regular structures such as RAM has been shown to be feasible and necessary as the size of the memory chips keeps growing. In VTST, pseudorandom test generation is adopted for the BIST. It was shown that this approach can reduce the test length compared with other BIST structures for RAM.

Future work and improvements include evaluating test length for detecting other types of faults besides stuck-at, transition and coupling faults. Also, techniques such as error-correcting code, self-diagnosis and self-reparing may also be incorporated into VTST to further reduce the test time and cost.

References

[1] Henry Chen, "VLSI Testability Synthesis Tool (VTST)", *Wright State University Proposal,* Dec. 1992.

[2] P. Bardell, W. McAnney, and J. Savir, *Built-in Test for VLSI Pseudorandom Techniques,* John Wiley and Sons, 1987.

[3] M. Franklin and K.K. Saluja, "Built-in Self-Testing of Random-Access Memories", *IEEE Computer,* pp.45-56, Oct. 1990.

[4] T. Sridhar, "New Parallel Test Approach for Large Memories," *Proc. Int. Test Conf.,* pp.462-470, 1985.

[5] J. Inoue et al., "Parallel Testing Technology for VLSI Memories," *Proc. Int. Test Conf.,* pp.1066-1071, 1987.

[6] Y. Matsuda et al., "New Array Architecture for Parallel Testing in VLSI Memories," *Prof. Int. Test Conf.,* pp.322-326, 1989.

TEMPERATURE-DEPENDENCE OF THE INTERSUBBAND ENERGIES AND OPTICAL ABSORPTION SPECTRUM OF MODULATION-DOPED QUANTUM WELLS

Godfrey Gumbs
Professor
Department of Physics

Hunter College of the City University of New York
695 Park Avenue
New York, NY 10021

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

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

August 1993

TEMPERATURE-DEPENDENCE OF THE INTERSUBBAND ENERGIES AND OPTICAL ABSORPTION SPECTRUM OF MODULATION-DOPED QUANTUM WELLS

Godfrey Gumbs
Professor
Department of Physics
Hunter College of the City University of New York

Abstract

Infrared absorption experiments on Si-doped GaAs/AlGaAs multiple quantum wells have shown that the peak positions of the intersubband transitions are not just associated with the spacing between the conduction subbands, which only include the Hartree interaction. Numerical calculations of the quasiparticle energies show that the depolarization and exchange interactions are large and have the effect of making the transition energy between the two lowest conduction subbands decrease as the temperature increases, for a fixed electron density. This feature is quite similar to the shift in the infrared absorption peak position (a "blueshift" as the temperature is decreased). Our calculations also reveal a striking non-monatonic dependence of the blueshift of the intersubband transition energies due to the the exchange interaction as a function of the dopant density. So far, experiments have not been performed on a sufficiently large number of samples for our calculations to be compared in detail with experiment. However, the limited experimental data available seem to indicate that there is a nonlinear variation of the blueshift of the peak position of the absorption spectrum as a function of density.

Table of Contents

Topic	Page #
I. INTRODUCTION	12-4
II. QUASIPARTICLE ENERGY SPECTRUM	12-4
III. OPTICAL ABSORPTION FOR QUANTUM WELLS	12-7
IV. NUMERICAL RESULTS AND DISCUSSION	12-9
V. CONCLUDING REMARKS AND SUMMARY	12-10
REFERENCES	12-11
FIGURE CAPTIONS	12-12
Figure 1	12-13
Figure 2	12-14
Figure 3	12-15

TEMPERATURE-DEPENDENCE OF THE INTERSUBBAND ENERGIES AND OPTICAL ABSORPTION SPECTRUM OF MODULATION-DOPED QUANTUM WELLS

Godfrey Gumbs

I. INTRODUCTION

There is considerable interest in the optical absorption for electron transitions between the conduction subbands in multiple quantum wells of $GaAs$ in $AlGaAs$.[1-6] Further work on the optical properties of these systems will undoubtedly continue because of their possible use as optoelectronic devices such as photodetectors, photomodulators and lasers. Early work on intersubband transitions in Si inversion layers (for a review, see Ref. 7) showed that exchange and correlation effects play a nontrivial part on the subband structure of confined electronic systems. Bandara et al.[8] explicitly calculated the exchange correction to the Hartree term for electrons in a quantum well where the lowest subband is occupied as a result of delta doping in the well. However, the work of Ref. 8 was restricted to zero temperature. The purpose of this paper is to examine the role played by many-body effects on the temperature dependence of the intersubband transition energies for electrons in a quantum well. The many-body effects are temperature dependent since the effective Coulomb interaction between electrons depends on the thermal population of the levels. We are interested in this problem because of recent experimental work which shows an interesting blueshift in the peak position of the absorption spectrum when the temperature is reduced from 298 K to 5 K.[5] So far, all intuitive attempts to model the temperature-dependence of the absorption peak positions have shown an increase in the transition energy as the temperature is increased, contrary to experimental data.[5] The reason for this failure is an inadequate treatment of the many-body effects. We have examined the role played by the Hartree and exchange interactions to first order in the Coulomb interaction and our results show features similar to the temperature dependence of the intersubband transition energies observed experimentally for modulation-doped quantum wells.

In optical experiments such as infrared absorption, intersubband transitions, associated with the dipole absorption, will be produced. The Hartree and exchange interactions play an an important role in determining the absorption spectra, such as the variation of the peak positions with the dopant density and the temperature. In the present work, we calculate the excitation energies for intersubband transitions. For this, we determine the quasiparticle energies and, for completeness, present analytical results for the absorption coefficient since the lineshape yields additional information concerning the many-body effects.

II. QUASIPARTICLE ENERGY SPECTRUM

Our method of calculation of the self-energy in the particle-hole Green's function is based on the dielectric response function formalism of Martin and Schwinger.[9] The general formulation is now pre-

sented. We expand the Green's function in terms of a complete set of states $\phi_{\mathbf{k},n}$ which satisfy the Schrödinger and Poisson equations simultaneously. That is,

$$G(\mathbf{r},\mathbf{r}';\omega) = \sum_{\mathbf{k},n} \phi_{\mathbf{k},n}(\mathbf{r})\phi^*_{\mathbf{k},n}(\mathbf{r}')\, G_{\mathbf{k},n}(\omega) \tag{1}$$

$$\phi_{\mathbf{k},n}(\mathbf{r}) = \frac{1}{A^{1/2}} e^{i\mathbf{k}\cdot\rho}\zeta_n(z) \tag{2}$$

where $\mathbf{k}$ and ρ are two-dimensional (2D) vectors, A is the sample area and the confinement of an electron in the z direction within the quantum well leads to discrete energy levels which in the parabolic approximation are $\epsilon_{\mathbf{k},n} = E_n + \hbar^2 k^2/2m^*_n$. The envelope function $\zeta_n(z)$ and the subband edge E_n are self-consistently determined by the Schrödinger equation

$$\left(-\frac{\hbar^2}{2}\frac{d}{dz}\frac{1}{m(z)}\frac{d}{dz} + V_{\rm b}(z) + V_{\rm H}(z)\right)\zeta_n(z) = E_n\zeta_n(z) \tag{3}$$

and Poisson's equation

$$\frac{d}{dz}\varepsilon(z)\frac{d}{dz}V_{\rm H}(z) = 4\pi e^2\left[N_0^+(z) - \eta\sum_n \frac{k_BT}{E_{{\rm F},n}}|\zeta_n(z)|^2 \ln\left(1 + e^{(\mu - E_n)/k_BT}\right)\right] . \tag{4}$$

In this notation, μ is the chemical potential and $E_{{\rm F},n} = \hbar^2(2\pi n_{\rm 2D})/2m^*_n$, where m^*_n is the effective mass of an electron in the nth subband with

$$\frac{1}{m^*_n} = \int_{-\infty}^{\infty} dz \frac{|\zeta_n(z)|^2}{m(z)}, \qquad \eta = \int_{-\infty}^{\infty} dz\; N_0^+(z) . \tag{5}$$

Here, the mismatch of the electron effective mass in the z direction is given by $m(z)$ and the variable background dielectric constant is accounted for through $\varepsilon(z)$. The bare confining potential $V_{\rm b}(z)$ is taken as zero inside the well and $0.81x$ eV outside where x is the mole fraction of Al in $GaAs/Al_xGa_{1-x}As$ quantum wells. In Eq. (4), $N_0^+(z)$ is the doping concentration and the integrated total charge on the right-hand side of Eq. (4) is zero by charge neutrality. For a given temperature, the Fermi energy is determined iteratively from the charge neutrality condition. We make the standard assumption that all the donors are ionized even at low temperatures.

The poles of the Green's function expansion coefficients $G_{\mathbf{k},n}(\omega)$ in Eq. (1) are obtained from Dyson's equation. The quasiparticle energies correspond to the solutions of

$$\omega_{\mathbf{k},n} = \omega^{(0)}_{\mathbf{k},n} + \Sigma_{\mathbf{k},n}(\omega_{\mathbf{k},n}) , \tag{6}$$

where $\hbar\omega^{(0)}_{\mathbf{k},n} = \epsilon_{\mathbf{k},n} - \mu$ and the self-energy is given by

$$\Sigma_{\mathbf{k},n}(\omega) = i\int_{-\infty}^{\infty}\frac{d\omega'}{2\pi}e^{i\omega' 0^+}\sum_{n'}\int\frac{d^2\mathbf{q}}{(2\pi)^2}G_{\mathbf{k}-\mathbf{q},n'}(\omega')$$
$$\times\int_{-\infty}^{\infty}dz\int_{-\infty}^{\infty}dz'\;\zeta_n(z)\zeta_n(z')\zeta_{n'}(z)\zeta_{n'}(z')V_{\rm sc}(z,z';\mathbf{q},\omega-\omega') . \tag{7}$$

Therefore, many-body and temperature effects contribute to the self-energy through Eq. (7) in conjunction with the equation for the screened potential which is given by

$$V_{sc}(z,z';\mathbf{q},\omega) = \int_{-\infty}^{\infty} dz''\, \epsilon^{-1}(z,z'';\mathbf{q},\omega)\, v_b(z'',z';q) \ , \tag{8}$$

where the Coulomb interaction is $v_b(z,z';q) = 2\pi e^2 \exp(-q|z-z'|)/\varepsilon_s q$, with $\varepsilon_s = 4\pi\varepsilon_0\varepsilon_b$ for a background dielectric constant ε_b within the quantum well. In self-consistent-field theory, the inverse dielectric function is a solution of the integral equation

$$\epsilon^{-1}(z_1,z_2;q,\omega) = \delta(z_1-z_2) + \int_{-\infty}^{\infty} dz_3 \int_{-\infty}^{\infty} dz_4\, v_b(z_1,z_3;q)\chi^0(z_3,z_4;q,\omega)\epsilon^{-1}(z_4,z_2;q,\omega) \ . \tag{9}$$

Therefore, screening involves polarization effects which are described in terms of the density-density response function. If exciton binding within an electron-hole pair plays a role, vertex corrections to the polarization function must be included. In the ladder approximation, we have[10]

$$\chi^0(z_1,z_2;q,\omega) = \sum_{n,n'} R^{(0)}_{n,n'}(q,\omega)\zeta_n(z_1)\zeta_n(z_2)\zeta_{n'}(z_2)\zeta_{n'}(z_1) \ , \tag{10}$$

where

$$R^{(0)}_{n,n'}(q,\omega) \equiv 2\int \frac{d^2\mathbf{k}}{(2\pi)^2} \frac{f_0(\epsilon_{\mathbf{k},n}) - f_0(\epsilon_{\mathbf{k}-\mathbf{q},n'})}{\hbar\omega + \epsilon_{\mathbf{k},n} - \epsilon_{\mathbf{k}-\mathbf{q},n'} + i\gamma} \Gamma_{nn'}(\mathbf{k},\mathbf{q};\omega) \ . \tag{11}$$

Here, γ is a parameter due to impurity and phonon scattering, $f_0(\epsilon)$ is the Fermi distribution function,

$$\Gamma_{nn'}(\mathbf{k},\mathbf{q};\omega) = 1 + \int \frac{d^2\mathbf{p}}{(2\pi)^2} V^{ex}(\mathbf{k}-\mathbf{p}) \frac{f_0(\epsilon_{\mathbf{p},n}) - f_0(\epsilon_{\mathbf{p}-\mathbf{q},n'})}{\hbar\omega + \epsilon_{\mathbf{p},n} - \epsilon_{\mathbf{p}-\mathbf{q},n'} + i\gamma} \Gamma_{nn'}(\mathbf{p},\mathbf{q};\omega) \ , \tag{12}$$

and $V^{ex}(\mathbf{k})$ is the static-screened exciton interaction whose screening length depends on the density-of-states of the electron gas.

The contributions from the many-body and temperature effects to the quasiparticle energies are more clearly identified if closed-form analytic results for the self-energy are obtained. In order to simplify our calculations, we replace the Green's function in Eq. (7) by a single-particle one (i.e., the Hartree-Fock approximation). By setting $\epsilon^{-1}(z-z';q,\omega) = \delta(z-z')$, we neglect the screening effects which are given by the second term in Eq. (9). Consequently, we obtain the following approximation for the exchange part of the self-energy

$$\hbar\Sigma^{exch}_{\mathbf{k},n} = -\sum_{n'} \int \frac{d^2\mathbf{q}}{(2\pi)^2} \int_{-\infty}^{\infty} dz \int_{-\infty}^{\infty} dz' \zeta_n(z)\zeta_{n'}(z)\zeta_n(z')\zeta_{n'}(z')$$
$$\times \frac{2\pi e^2}{\varepsilon_s q} e^{-q|z-z'|} f_0\left(\epsilon_{\mathbf{k}-\mathbf{q},n'}\right) \ , \tag{13}$$

which is a generalization of the result of Bandara et al.[8] to finite temperature for the correction to the Hartree energy $\epsilon_{\mathbf{k},n}$ in Eq. (6). Since $\Sigma^{exch}_{\mathbf{k},n}$ is independent of frequency, the quasiparticle energy is given by a closed-form analytic result.

We have solved Eq. (9) for $\epsilon^{-1}(z,z';q,\omega)$ when the Coulomb potential is included and obtained the following analytic result

$$\epsilon^{-1}(z_1,z_2;q,\omega)-\delta(z_1-z_2)=\sum_{n,n'}R^{(0)}_{nn'}(q,\omega)w_{nn'}(z_1;q)\,K_{nn'}(z_2;q,\omega)\ , \tag{14}$$

where

$$w_{nn'}(z_1;q)\equiv\int_{-\infty}^{\infty}dz_3\ v_{\rm b}(z_1,z_3;q)\zeta_n(z_3)\zeta_{n'}(z_3)\ . \tag{15}$$

Our calculations show that $K_{nn'}$ in Eq. (14) is the solution of a linear matrix equation which we write as

$$\sum_{n,n'}\left[\delta_{mn}\delta_{m'n'}-R^{(0)}_{nn'}(q,\omega)u_{mm';nn'}(q)\right]K_{nn'}(z;q,\omega)=\zeta_m(z)\zeta_{m'}(z)\ , \tag{16}$$

where

$$u_{mm';nn'}(q)=\int_{-\infty}^{\infty}dz\int_{-\infty}^{\infty}dz'\ \zeta_m(z)\zeta_{m'}(z)v_{\rm b}(z,z';q)\zeta_n(z')\zeta_{n'}(z')\ . \tag{17}$$

Equations (13) through (17), along with Eqs. (7) and (8), give the quasiparticle energies with exchange and correlation effects included. However, in this paper, we restrict our numerical calculations to the Hartree and exchange contributions to the quasiparticle energies as a first step in our determination of the role played by the Coulomb interaction on the thermal variation of the quasiparticle energies. We now make use of the results we have derived so far to formulate the absorption coefficient.

III. OPTICAL ABSORPTION FOR QUANTUM WELLS

When the electron gas with a positive jellium background is perturbed by an external electric field, the density distribution of the electrons will oscillate with a normal mode frequency. The resulting density fluctuation will induce an effective dynamic dipole in the system. Since the wavelength of the incident light is much larger than the size of the sample being measured, we assume that the perpendicular electric field (along the z direction) is uniform within the sample. The coupling of the dipole to the external electric field gives rise to energy absorption which is represented by the absorption coefficient

$$\beta(\omega)=\frac{\omega}{\epsilon_{\rm b}^{1/2}c\epsilon_0 L_z}{\rm Im}\ \alpha(\omega)\ , \tag{18}$$

where L_z is the width of the quantum well and the polarizability $\alpha(\omega)$ of the electron gas is given by

$$\alpha(\omega)=-e^2\sum_{n<n'}Z_{n,n'}\chi_{n,n'}(\omega)\left[Z_{n,n'}+U_{n,n'}(\omega)\right]\ . \tag{19}$$

In Eq (19), we have introduced an irreducible polarizability function defined by

$$\chi_{n,n'}(\omega)=R_{n,n'}(0,\ \omega)+R_{n',n}(0,\ \omega)\ , \tag{20}$$

where $R_{n,n'}(0,\omega)$ is obtained from Eq. (11) by replacing the Hartree energies $\epsilon_{\mathbf{k},n}$ with the quasiparticle energies including exchange. Also, the dipole transition matrix is defined by

$$Z_{n,n'} = \int_{-\infty}^{\infty} dz\ \zeta_n(z) z \zeta_{n'}(z)\ . \tag{21}$$

Furthermore, the Coulomb interaction matrix $U_{n,n'}(\omega)$ is determined from the following set of linear self-consistent equations

$$U_{m,m'}(\omega) = \sum_{n<n'} \chi_{n,n'}(\omega) A_{n,n';m,m'} \left[Z_{n,n'} + U_{n,n'}(\omega) \right]\ , \tag{22}$$

where, in this notation, we have

$$A_{n,n';m,m'} = -\frac{2\pi e^2}{\epsilon_s} \int_{-\infty}^{\infty} dz \int_{-\infty}^{\infty} dz'\ \zeta_n(z)\zeta_{n'}(z) |z - z'| \zeta_m(z')\zeta_{m'}(z')\ . \tag{23}$$

From these general results for the absorption, we now obtain a closed form analytic result for a special case which is for a two-level system when the incident light satisfies the resonance condition. Let us consider the transition between the topmost occupied state ('0') and the lowest unoccupied state ('1') and assume that the level separation is large so that the coupling between different transitions can be ignored. In this case, we are left with a simple two-level model. A straightforward calculation shows that Eq. (19) gives

$$\alpha(\omega) = -e^2 Z_{0,1}^2 \left[\frac{\chi_{0,1}(\omega)}{1 - A_{0,1;0,1}\chi_{0,1}(\omega)} \right]\ . \tag{24}$$

In the denominator of Eq. (24), the depolarization shift from screening and the vertex correction are included in the irreducible polarizability through the ladder approximation. It is clear that there is temperature dependence in the depolarization shift as well as in the vertex correction. The peak positions in Im $\alpha(\omega)$ correspond to resonant absorption. If we neglect the vertex correction to the polarizability in Eq. (11), assume the temperature is zero and that the effective mass is the same for each of the two subbands, we obtain from Eq. (24)

$$\text{Im}\ \alpha(\omega) = \frac{F(\omega)}{\left(\hbar^2\omega^2 - E_{1,0}^2\right)^2 \left(\hbar^2\omega^2 - E_{1,0}^2 - 2n_{2D}A_{0,1;0,1}E_{1,0}\right)^2 + B(\omega)}\ . \tag{25}$$

where the folowing notation has been introduced

$$F(\omega) = 4e^2 Z_{0,1}^2 \hbar\omega E_{1,0} n_{2D} \gamma \left[\left(\hbar^2\omega^2 - E_{1,0}^2\right)^2 + 2\gamma^2 \left(\hbar^2\omega^2 + E_{1,0}^2\right) \right]\ , \tag{26}$$

and

$$B(\omega) = 4\gamma^2 \left[\left(\hbar^2\omega^2 - E_{1,0}^2\right)^2 \left(\hbar^2\omega^2 + E_{1,0}^2 + n_{2D}A_{0,1;0,1}E_{1,0}\right) \right.$$
$$\left. -2n_{2D}A_{0,1;0,1}E_{1,0} \left(\hbar^2\omega^2 + E_{1,0}^2\right) \left(\hbar^2\omega^2 - E_{1,0}^2 - n_{2D}A_{0,1;0,1}E_{1,0}\right) \right]\ . \tag{27}$$

In Eqs. (25) - (27), $E_{1,0} = E_1 - E_0$ is the difference of the subband edges. If we further assume that $\gamma^2,\ n_{2D}A_{0,1;0,1}E_{1,0} \ll \left|\hbar^2\omega^2 - E_{1,0}^2\right|$ and $\hbar\omega \approx E_{1,0}$, we obtain from Eq. (25)

$$\text{Im}\ \alpha(\omega) = f_{1,0} \frac{\delta_{1,0}}{\left(\hbar^2\omega^2 - E_{1,0}^2 - 2n_{2D}A_{1,0;1,0}E_{1,0}\right)^2 + 4\delta_{1,0}^2}\ , \tag{28}$$

where the spectrum broadening is

$$\delta_{1,0} = \sqrt{2} E_{1,0} \gamma \ , \tag{29}$$

and the spectrum amplitude is

$$f_{1,0} = \frac{4e^2 Z_{0,1}^2 E_{1,0} n_{2D}}{\sqrt{2}} \ . \tag{30}$$

Equation (28) is similar to that in Ref. 7. The peak position $\hbar\omega_{\rm p} = E_{1,0}\sqrt{1+\eta_{1,0}}$ in the absorption spectrum is displaced from the subband energy separation $E_{1,0}$ due to depolarization effects, where $\eta_{1,0} = 2n_{2D}A_{1,0;1,0}/E_{1,0}$. The peak position varies with temperature and dopant density in a similar way as the intersubband separation but they differ quantitatively due to depolarization and vertex corrections.

IV. NUMERICAL RESULTS AND DISCUSSION

We have calculated the quasiparticle energies by first solving the Schrödinger-Poisson combination in Eqs. (3) and (4) and then adding its contribution due to exchange in Eq. (7). The energy eigenvalues were calculated self-consistently when all subbands are included thereby taking into account the coupling between the ground and all the excited subbands. Figure 1 shows plots for intersubband transitions between the two lowest subbands at the Γ point as a function of the dopant density $N_{\rm d}$; for comparison, we present results at $T = 5\ K$ and $T = 298\ K$. The separation between the two curves in Fig. 1 has an interesting variation as a function of $N_{\rm d}$. This "shift" in the transition energy is plotted in Fig. 2. Some light is shed on this nonmonatonic variation by examining the Hartree and exchange contributions to the total energy separately. Our numerical calculations show that when the single-particle energy of the ground subband is plotted as a function of $N_{\rm d}$, the curve for $T = 5\ K$ is very close to the curve for 298 K in the limit of low dopant concentration ($N_{\rm d} \sim 10^{18}\ cm^{-3}$). However, as $N_{\rm d}$ increases, the separation between these curves increases with the curve for $T = 5\ K$ always above the curve for $T = 298\ K$. The same situation applies for the first excited subband. However, the curves for the exchange energies of the ground and first excited subbands show competing behaviors as functions of $N_{\rm d}$. The $T = 5\ K$ and the $T = 298\ K$ curves for the exchange energies of the ground subband are separated by about 10 meV in the limit of low density ($N_{\rm d} \approx 10^{17}\ cm^{-3}$). As $N_{\rm d}$ increases, the exchange energies for the ground subband at these two temperatures decrease at different rates so that they are almost equal at $N_{\rm d} \approx 5 \times 10^{18}\ cm^{-3}$. For this range of values of $N_{\rm d}$, the $T = 5\ K$ curve for the ground subband always lies below the $T = 298\ K$ curve. The exchange energy curve for the first excited subband at 298 K is slightly above the corresponding curve at $T = 5\ K$ at low densities. However, as $N_{\rm d}$ increases the exchange energy at $T = 5\ K$ does not decrease as rapidly as it does at 298 K so that the two curves cross and their separation increases with $N_{\rm d}$. Therefore, when the Hartree and exchange energies are combined, the shift in the transition energy would partly be determined by the dominant behavior of either subband over a range of densities. In Fig. 3, the transition energy between the ground subband and the first excited state is plotted as a function of temperature for wavevector $q = 0$ and $N_{\rm d} = 4.3 \times 10^{17}$. This reduction in the transition energy as the temperature increases is similar in nature to the experimental results obtained from infrared absorption for GaAs/AlGaAs doped quantum wells.[5] This is the first theoretical work to reproduce this temperature dependence. It is clear that exchange and depolarization effects play an important role in this thermal behavior since the Hartree energy is almost independent

of temperature.

V. CONCLUDING REMARKS AND SUMMARY

In this paper, we have presented a formulation for calculating the quasiparticle energies for modulation doped quantum wells. These quasiparticle energies are a first step in determining the peak positions in the optical absorption coefficient of modulation doped quantum wells. A first-principles many-body formalism shows that the exchange energy must be included to simulate the variation of the conduction subband energy with the dopant density and the temperature. Our calculations show that the intersubband energy is "blueshifted" with a reduction in temperature. As far as we know, this is the first calculation which shows a temperature variation that is similar to the infrared absorption experiments for the conduction bands in modulation-doped GaAs/AlGaAs quantum wells over the temperature range from 5 K to 298 K.[5] All previous attempts to model the temperature-dependence of the absorption peak positions have shown an increase in energy as the temperature is increased, contrary to the experimental data.[5] The reason for this failure is an inadequate treatment of the many-body effects. We have examined the role played by the Hartree and exchange interactions to first order in the Coulomb interaction.

The Hartree potential of the Coulomb interaction was obtained by solving Poisson's equation self-consistently with the Schrödinger equation. This procedure gives the conduction band edges as well as the envelope function describing the confinement of an electron within the quantum well. The approximation we used for the conduction band dispersion was the parabolic approximation $\epsilon_{\mathbf{k},n} = E_n + \hbar^2 k^2/2m_n^*$. Nonparabolicity in the z direction would shift down the intersubband transition energies but should not change the qualitative nature of our results.

Acknowledgments — I would like to thank Dr. John P. Loehr for useful discussions and for writing the computer programs for the research reported here. This work was supported in part by the Summer Faculty Program of the Air Force Office of Scientific Research.

References

1. L. C. West and S. J. Eglash, Appl. Phys. Lett. **46**, 1156 (1985).

2. B. F. Levine, C. G. Bethea, K. K. Choi, J. Walker, and R. J. Malik, Appl. Phys. Lett. **53**, 231 (1988).

3. A. Pinczuk, S. Schmitt-Rink, G. Danan, J. P. Valladares, L. N. Pfeiffer, and K. W. West, Phys. Rev. Lett. **63**, 1633 (1989).

4. B. C. Covington, C. C. Lee, B. H. Hu, H. F. Taylor, and D. C. Streit, Appl. Phys. Lett. **54**, 2145 (1989).

5. M. O. Manasreh, F. Szmulowicz, D. W. Fischer, K. R. Evans, and C. E. Stutz, Appl. Phys. Lett. **57**, 1790 (1990); M. O. Manasreh and J. P. Loehr in *Semiconductor Quantum Wells and Superlattices for Long Wavelength Infrared Detection*, (ed. M. O. Manasreh, Artech, MA, 1993), ch. 2.

6. M. O. Manasreh, F. Szmulowicz, T. Vaughan, K. R. Evans, C. E. Stutz, and D. W. Fischer, Phys. Rev. B**43**, 9996 (1991).

7. T. Ando, Solid State Commun. **21**, 133 (1977); T. Ando, A. B. Fowler, and F. Stern, Rev. Mod. Phys. **54**, 437 (1982).

8. K. Bandara, D. D. Coon, O. Byungsung, Y. F. Lin, and M. H. Francombe, Appl. Phys. Lett. **53**, 1931 (1988).

9. P. C. Martin and J. Schwinger, Phys. Rev. **115**, 1342 (1959).

10. G. D. Mahan, *Many-Particle Physics*, (Plenum, New York, 1990), p. 263.

Figure Captions

Fig. 1. Calculated transition energies between the two lowest conduction subbands at the Γ point as a function of the dopant density N_d; for comparison, we present results at $T = 5\ K$ and $T = 298\ K$. These calculations include the Hartree and exchange interactions.

Fig. 2. The transition energies between the ground and first excited subbands are calculated at $T = 5\ K$ and $T = 298\ K$. The "difference" between these two transition energies, shown in Fig. 1, is plotted as a function of dopant density.

Fig. 3. The transition energy between the ground subband and the first excited state is plotted as a function of temperature. Here, $q = 0$ and $N_d = 4.3 \times 10^{17}$.

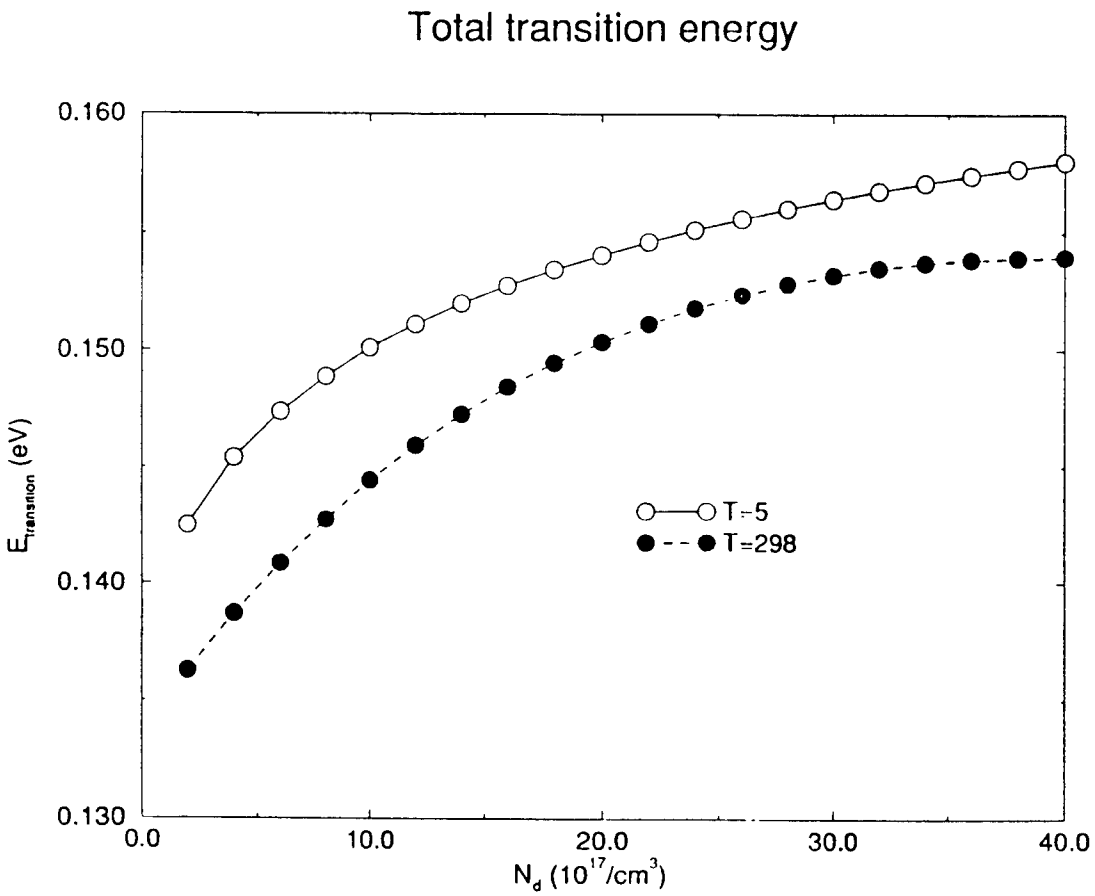

Total transition energy
0.160
0.150
0.140
0.130
$E_{transition}$ (eV)
0.0
10.0
20.0
30.0
40.0
N_d (10^{17}/cm^3)
T=5
T=298

Fig. 1

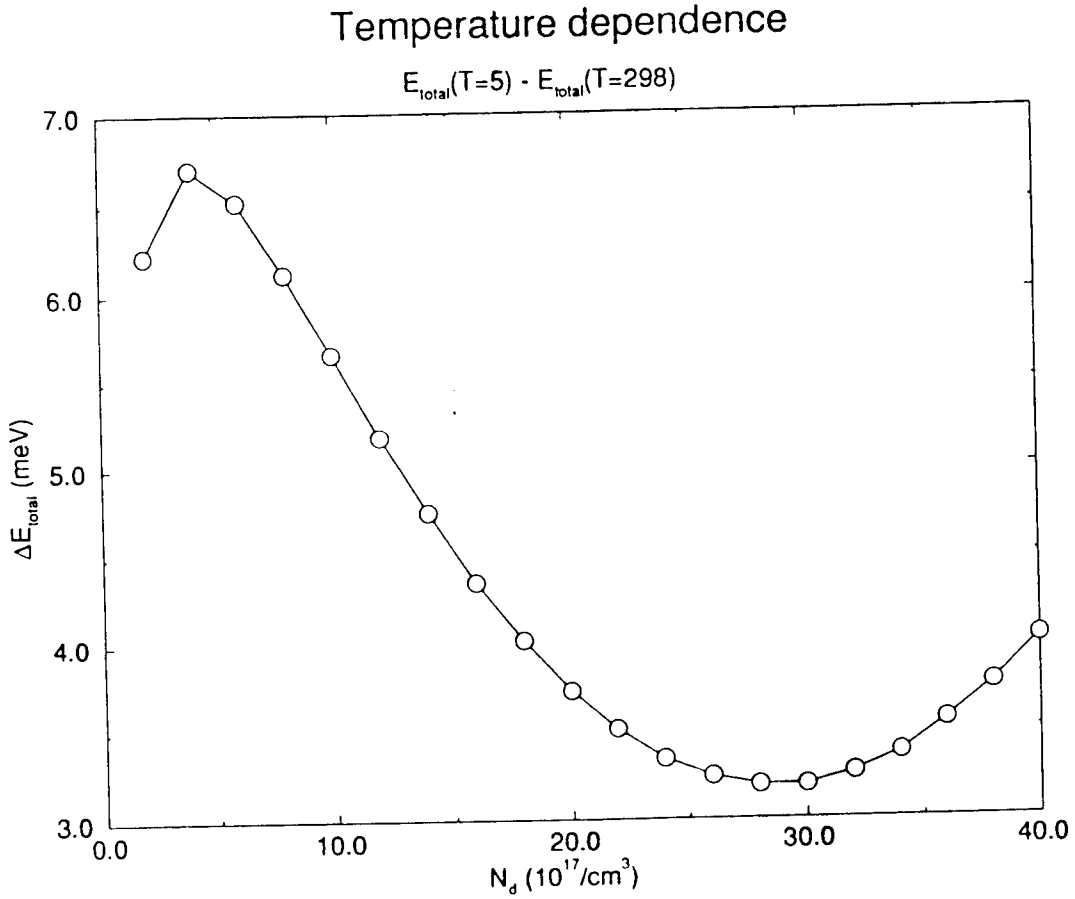

Temperature dependence
$E_{total}(T=5) - E_{total}(T=298)$
7.0
6.0
5.0
4.0
3.0
ΔE_{total} (meV)
0.0
10.0
20.0
30.0
40.0
N_d (10^{17}/cm^3)

Fig. 2

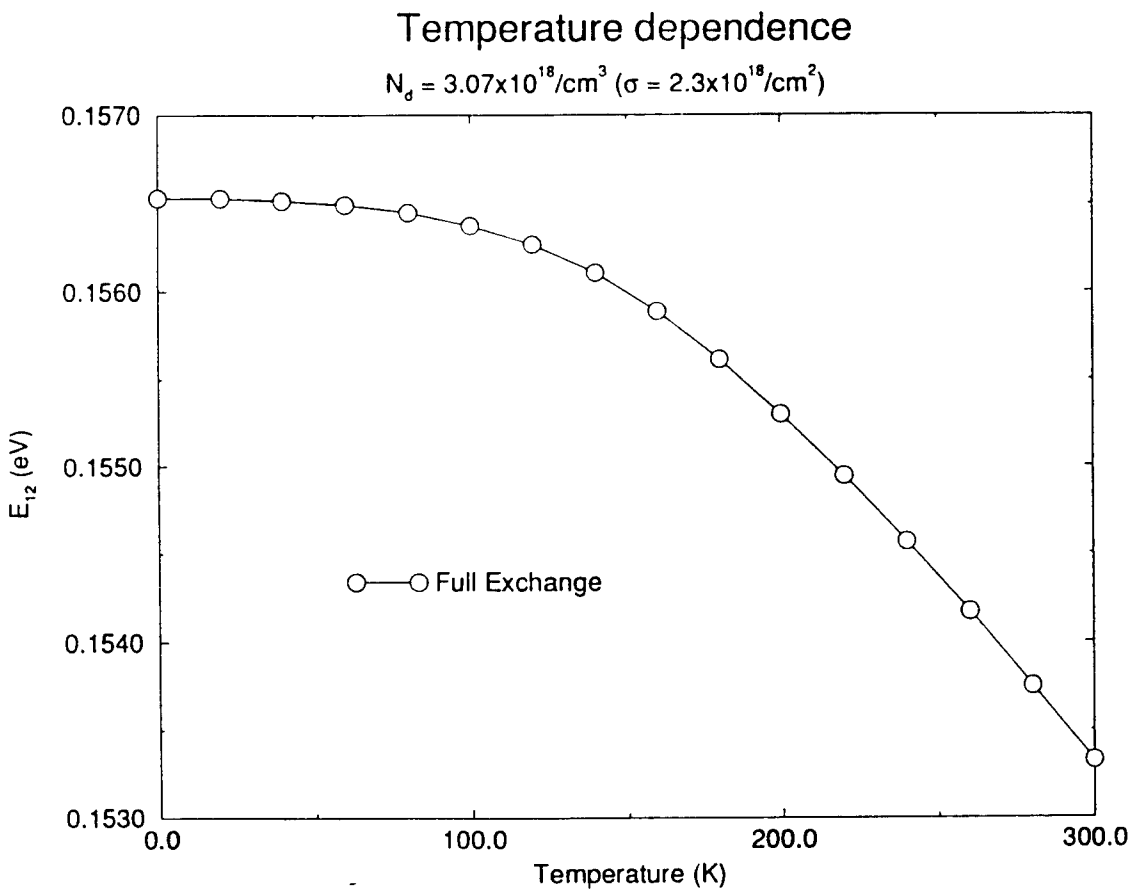

Temperature dependence
N_d = 3.07x10^{18}/cm^3 (σ = 2.3x10^{18}/cm^2)
0.1570
0.1560
0.1550
0.1540
0.1530
E_{12} (eV)
Full Exchange
0.0
100.0
200.0
300.0
Temperature (K)

Fig. 3

AN ANALYTICAL MODEL FOR AlGaAs/GaAs MULTI-EMITTER FINGER HBT INCLUDING SELF-HEATING AND THERMAL COUPLING EFFECTS

J. J. Liou

Associate Professor

Electrical and Computer Engineering Dept.

University of Central Florida, Orlando, FL 32816

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

University of Central Florida

September 1993

AN ANALYTICAL MODEL FOR AlGaAs/GaAs MULTI-EMITTER FINGER HBT INCLUDING SELF-HEATING AND THERMAL COUPLING EFFECTS

J. J. Liou
Associate Professor
Electrical and Computer Engineering Dept.
University of Central Florida, Orlando, FL 32816

Abstract

An analytical model which can be used to predict the thermal as well as electronic behavior of the multiple emitter heterojunction bipolar transistor (HBT) is presented. The model is developed from the knowledge of device make-up (doping concentrations, layer thicknesses, etc.), and relevant physics such as the effects of graded heterojunction, self-heating, thermal coupling, and ballast emitter resistance are included in a unified manner. Thermal runaway, or current crush, phenomenon observed in the multi-finger HBT at high current level has been successfully described. Experimental evidences obtained from a 6-finger and 4-finger HBTs are included in support of the model. We found that the current crush phenomenon is caused by the uneven increase of the base and collector currents at elevated temperatures due to the thermal effect.

AN ANALYTICAL MODEL FOR AlGaAs/GaAs MULTI-EMITTER FINGER HBT INCLUDING SELF-HEATING AND THERMAL COUPLING EFFECTS

J. J. Liou

1. INTRODUCTION

The advance of AlGaAs/GaAs heterojunction bipolar transistor (HBT) technology in recent years has made high output power possible and practical. The HBT's very high current handling capability and the very poor thermal conductivity of GaAs, however, often lead to significant self-heating effect which confines the HBT performance considerably below its electronic limitation [1]. For modern microwave HBTs, a multiple emitter finger structure has frequently been used, in which several HBT emitters, each with its own HBT operation, are arranged in parallel to each other with proper spacing [2]. Such a structure can reduce the HBT signal propagation delay time. In addition, it allows less current to be carried and thus less heat power to be generated in each HBT unit cell, thus making the self-heating effect less prominent compared to its single-emitter finger counterpart. Recently, a 12.5 W cw (power density of 1.74 $mW/\mu m^2$ of emitter area) monolithic amplifier constructed using 12-finger HBTs was demonstrated at 10 GHz [3].

While the multi-finger HBT offers higher output power density, it is more susceptible to a thermally limited phenomenon called thermal runaway, or current crush [4]. When the base current is fixed and is relatively large, the collector current decreases sharply (current crush) as the collector-emitter voltage is increased beyond a critical value. This phenomenon results from the combination of self-heating in the unit emitter finger and thermal coupling among the neighboring fingers. A common remedy to this problem is to use large resistors (ballast resistors) at the emitter and/or base contacts [5]. Such a resistance gives rise to a large voltage drop at the contact and thus reduce the voltage drop across the emitter-base junction. This in turn decreases the current density and the heat generated in each unit HBT. Evidently, this approach limits the output power density, not by thermal, but by electronics means. It can also

degrades the HBT high frequency performance due to the extra delay time through the ballast resistance.

Despite the fact that the multi-finger HBT has become increasingly important and popular in high-power microwave applications, efforts on analytical modeling such a device have been limited in the past. This is due in part to the complicated nature of the negative feedback of the thermal effect on the HBT current-voltage characteristics. The problem is further compounded by the thermal coupling between the neighboring emitter fingers when a multi-finger structure is considered. A few numerical models have been reported in the literature [4,6-7]. For example, solving the three-dimensional heat transfer equation, Gao et al. [6] studied the temperature in each emitter finger as a function of the emitter spacing, the number of the fingers, and the geometry of the substrate. Their results, however, are not directly suited for HBT design because the heat power on each finger, which is related to the current and applied voltage, was treated as an independent input parameter. Recently, an analytical HBT model including the self-heating effect was derived from the knowledge of HBT make-up [8], but it is only applicable for the single-finger HBT. Furthermore, the model employs the drift-diffusion theory for the charge transport in the HBT, an approximation valid only for HBTs having an ideal graded heterojunction [9].

This paper develops a physics-based and analytical model capable of predicting the multi-finger HBT current-voltage characteristics. Relevant physics such as the effects of self-heating, thermal coupling between fingers, ballast emitter resistance will be accounted for in a unified manner. In addition, the model accounts for the thermionic and tunneling mechanisms at the hetero-interface and is applicable for HBTs having a nonideal graded heterojunction.

2. ISOTHERMAL HBT MODEL

We develop an isothermal model for a graded HBT. Because the conduction band discontinuity (or spike) in an abrupt HBT can hinder the free-carrier

transport from the emitter to base, a graded layer inserted between the emitter and base is often used to improve the free-carrier injection efficiency [10-11]. Such a layer, normally having a thickness between 100 and 300 Å, can effectively remove the spike and thus make the thermionic and tunneling mechanisms at the hetero-interface less important.

Following the thermionic-field-diffusion approach developed by Grinberg et al. [12], the electron current density J_n across the hetero-interface (located at $x = -W_g$) is the difference between two opposing fluxes

$$J_n(-W_g) = qv_n\gamma[n(-W_g^-) - n(-W_g^+)] \tag{1}$$

where v_n is the electron thermal velocity, γ is the tunneling coefficient [12], and n is the electron concentration. It should be mentioned that the value of γ depends strongly on the conduction band barrier potentials V_{B1} and V_{Bgc}. In (1)

$$n(-W_g^-) = N_E\exp(-V_{B1}/V_T) \quad \text{and} \quad n(-W_g^+) = n(X_2)\exp[(V_{B2} + V_{Bgc})/V_T] \tag{2}$$

where N_E is the emitter doping density and V_T is the thermal voltage. The parameter $n(X_2)$ can be solved using the relation:

$$J_n(-W_g) = J_{SCRB} + J_{SCRG} + J_n(X_2) \tag{3}$$

where J_{SCRB} is recombination current density in the space-charge layer associated with the base layer ($x = 0$ and $x = X_2$) and J_{SCRG} is the recombination current density in the graded layer (the models for J_{SCRB} and J_{SCRG} will be developed in the later), and $J_n(X_2)$ is the diffusion-only current in the quasi-neutral base (QNB). For a very thin base,

$$J_n(X_2) = J_C = qD_n n(X_2)/(W_B + \Delta W_B + D_n/v_{sat}) \tag{4}$$

where J_C is the collector current density, D_n is the electron diffusion

coefficient in the QNB, $W_B = X_3 - X_2$ is the QNB thickness, ΔW_B is the current-induced base pushout [13], and v_{sat} (= 10^7 cm/sec) is the saturation drift velocity caused by the high field in the base-collector junction.

Combining (1)-(4) and solving for $n(X_2)$, we obtain

$$n(X_2) = [qv_n\gamma N_E \exp(-V_{B1}/V_T) - J_{SCRB} - J_{SCRG}]/\eta \tag{5}$$

where $\eta = qD_n/(W_B + \Delta W_B + D_n/v_{sat}) + qv_n\gamma\exp[(V_{B2} + V_{Bgc})/V_T]$. Thus J_C can be calculated from (4) after $n(X_2)$ is found from (5).

The components of the base current density J_B of the HBT include 1) injection of hole current density J_{RE} from the base to emitter; 2) electron-hole recombination current density J_{RB} in the quasi-neutral base; 3) electron-hole recombination current density J_{SCR} in the emitter-base space-charge layer; and 4) electron-hole recombination current density J_{RS} at the emitter and base surfaces. This injection of hole current density can be modeled using the conventional diffusion-current only approximation [14]:

$$J_{RE} = qD_pN_B\exp[-(V_{B1} + V_{B2} + V_{Bgv})/V_T]/W_E \tag{6}$$

where D_p is the hole diffusion coefficient in the emitter and $W_E = X_E - X_1$ is the thickness of the quasi-neutral emitter. The recombination current density in the quasi-neutral base is [14]

$$J_{RB} = J_n(X_2)(1 - \alpha) \tag{7}$$

where α is the base transport factor. The recombination current density J_{SCR} in the space-charge layer consists of three recombination current densities occurred in the emitter-side of the space-charge layer (J_{SCRE}), in the graded layer (J_{SCRG}), and in the base-side of the space-charge layer (J_{SCRB}). Thus

$$J_{SCR} = J_{SCRE} + J_{SCRG} + J_{SCRB}$$

$$= q \int_{-X_1}^{-W_g} U_{SRH,E} dx + q \int_{-W_g}^{0} U_{SRH,G} dx + q \int_{0}^{X_2} U_{SRH,B} dx \qquad (8)$$

where U_{SRH} is the Shockley-Read-Hall recombination rate. The surface recombination current density is influenced strongly by the fabrication process. It includes electron-hole recombination taking place at the emitter side-walls as well as at the extrinsic base surface. Empirically [15-17]

$$J_{RS} = J^* \exp(V_{BE}/nV_T) \qquad (9)$$

Here J^* and n ($1 < n < 1.33$ depending on the bias condition [17]) are the experimentally determined parameters that characterize the surface recombination current.

3. HBT MODEL INCLUDING THERMAL EFFECTS

In this section, we first review briefly an analytical single-finger HBT model including the self-heating effect developed recently [8]. The model will then be extended to the multi-finger HBT by including thermal coupling effect. The effect of the emitter and base resistances will also be included.

3.1 Single-Finger Structure

Fig. 1 shows the two-dimensional HBT structure including the sub-collector and substrate. Since the size of the intrinsic HBT (region directly underneath the emitter contact) is much smaller than that of the extrinsic HBT, we assume that the temperature in the intrinsic HBT is uniform and that the heat generated in the intrinsic HBT is dissipated primarily through the substrate. The heat power P_s (W) generated in the HBT is

$$P_s = J_C V_{CE} A_E \qquad (10)$$

where A_E is the emitter area and $V_{CE} = V_{BE} + V_{CB}$ is the applied collector-emitter

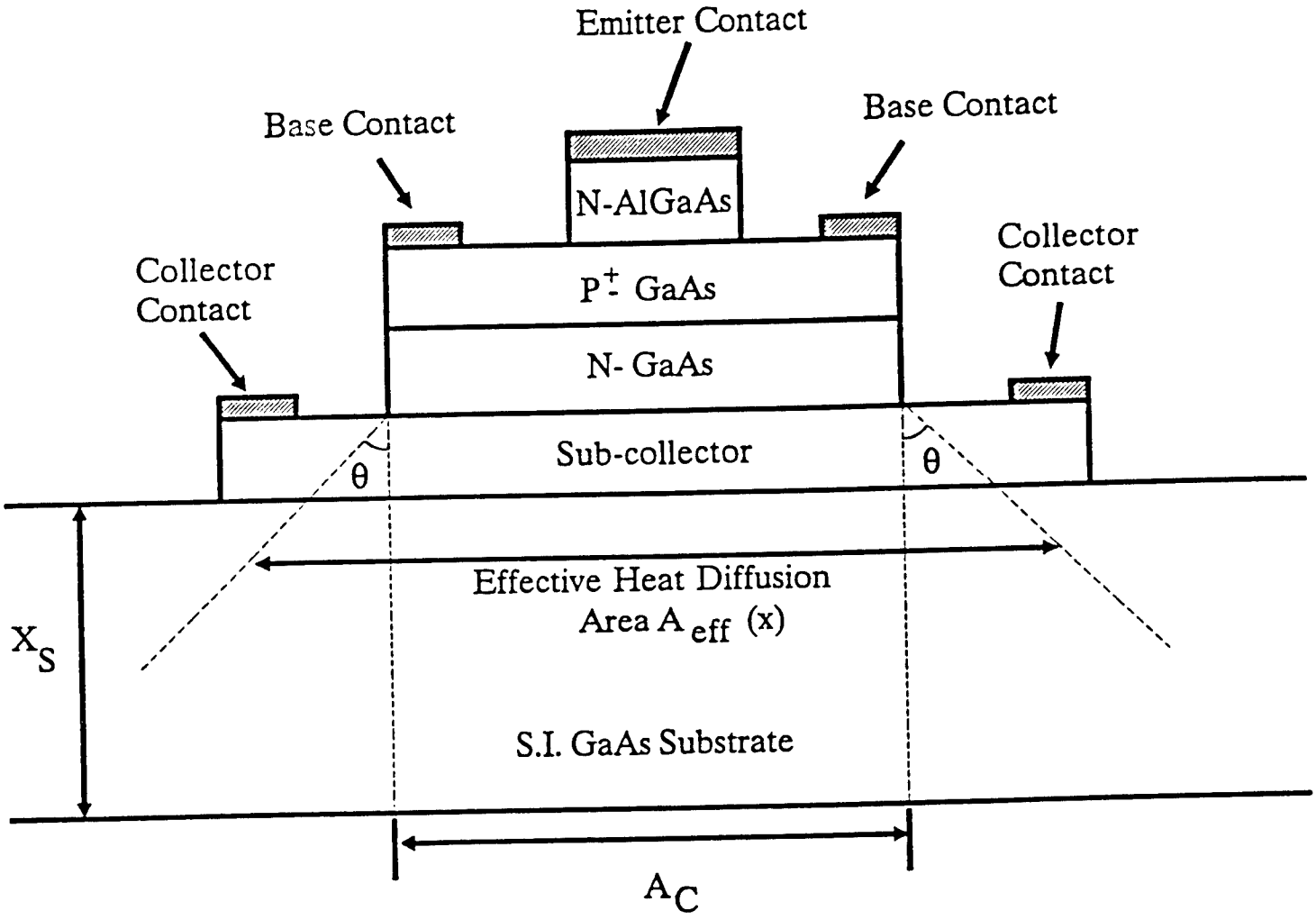


Fig. 1 Two-dimensional HBT device structure illustrating the effective area through which the heat generated in the intrinsic HBT is dissipated.

voltage. The generated heat power is related to the thermal resistance R_{th} of the substrate as

$$T - T_0 = P_s R_{th} \tag{11}$$

where T is the HBT lattice temperature and T_0 = 300 K is the ambient temperature. Assume the heat is dissipated throughout the effective area A_{eff} in the substrate with a lateral diffusion angle θ (θ = 45 is used) (Fig. 1) and the thermal conductivity K_s is proportional to $(T/T_0)^{-b}$, where b = 1.22 [18].

To account for the voltage drop caused by the emitter and base resistances, the applied base-emitter voltage V_{BE} needs to be replaced by the base-emitter junction voltage $V_{j,BE}$:

$$V_{j,BE} = V_{BE} - r_E(J_C + J_B) - r_B J_B \tag{12}$$

Here r_E and r_B are the specific emitter and base resistances (in Ω-cm^2), respectively.

3.2 Multi-Finger Structure

The foregoing approach can be extended to modeling the multi-finger HBT in which several emitter fingers are arranged in parallel with proper spacing, as shown in Fig. 2. A numerical analysis [6] solving the coupled current and heat transfer equations for the multi-finger HBT indicates that the device performance is affected strongly by both the self-induced thermal resistance R_{th} (discussed in the previous section) and the coupled thermal resistance R_c due to the heating from the neighboring emitter elements. Since the thermal coupling on the subject finger due to the nearest (primary) finger is much larger than that due to the secondary fingers, we will neglect the secondary thermal coupling effect.

To illustrate the modeling concept, let us consider a 3-finger pattern. Due to the symmetrical geometry, the two outer fingers will have identical thermal properties. The temperatures at the outer and center fingers are

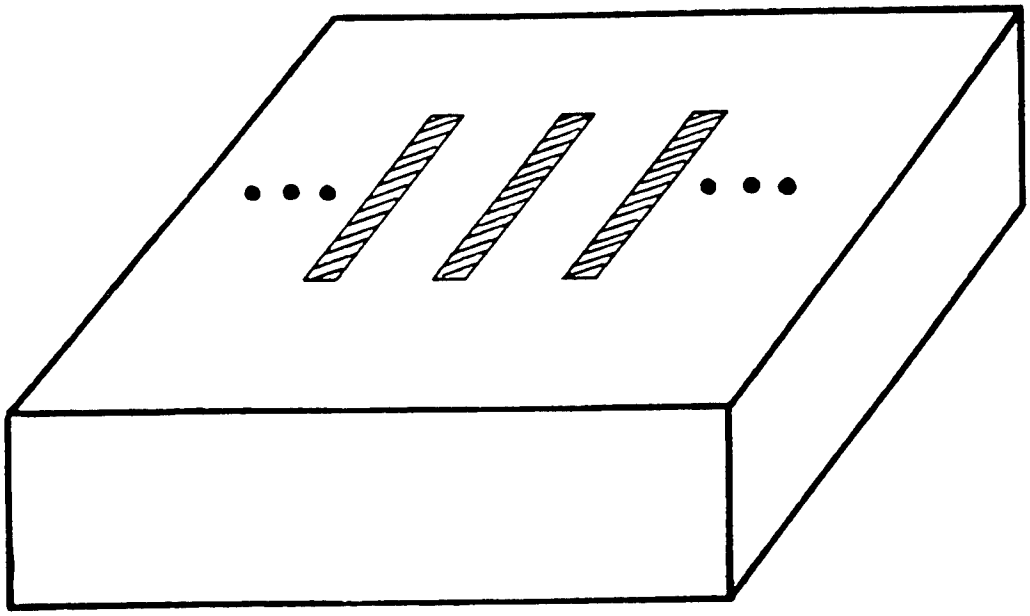


Fig. 2 Topology of the multi-emitter finger HBT.

$$T_S = T_0 + P_{s,S}R_{th,S} + P_{s,C}R_{c,C} \quad (13)$$

$$T_C = T_0 + P_{s,C}R_{th,C} + 2P_{s,S}R_{c,S} \quad (14)$$

where the subscripts S and C denote outer and center fingers, respectively. The term involving R_{th} is the temperature rise due to self-heating in the unit HBT and the term involving R_c is the temperature rise due to thermal coupling between the subject finger and the nearest (primary) neighboring finger(s). Note that the center finger is subjected to two thermal couplings whereas the outer fingers are subjected to only one thermal coupling.

The value of R_c depends on the geometry of emitter fingers and the process, including the emitter mesa etching and metalization, and is too complicated to model physically. We suggest R_c be related to R_{th} through an empirical parameter A_1 as $R_C = A_1R_{th}$. Increasing the value of A_1 will increase the temperature at the center finger and, to a lesser extent, the temperature at the two outer fingers, and subsequently increase the likelihood of current crush. For a typical HBT under study, the emitter finger spacing is 10 μm, and A_1 is empirically determined as 0.25 (A_1 = 0 for a single-finger HBT).

The initial value of T can be calculated from the above equations after the initial J_C, J_B, and P_s are calculated under room temperature. The correct T, J_C, and J_B for each emitter finger are obtained after several iterations. Summing J_C and J_B in each finger then yields the total J_C and J_B for the multi-finger HBT.

4. RESULTS AND DISCUSSIONS

Figure 3 shows the Gummel plot calculated from the model and obtained from measurement for a 6-finger HBT at V_{CB} = 0, 3 V, and 6 V. The HBT has a typical intrinsic make-up of $N_E = 5\times10^{17}$ cm^{-3}, $N_B = 8\times10^{18}$ cm^{-3}, $N_C = 5\times10^{16}$ cm^{-3}, emitter layer thickness of 1000 Å, base layer thickness of 1000 Å, collector layer thickness of 7000 Å, a finger area of 2.5x10 μm^2, and a ballast emitter contact resistance of 6×10^{-6} Ω-cm^2. The extrinsic make-up of the HBT, which is needed to calculate the thermal resistance R_{th}, is Z = 25 μm, A_C = 10x25 μm^2, and X_s =

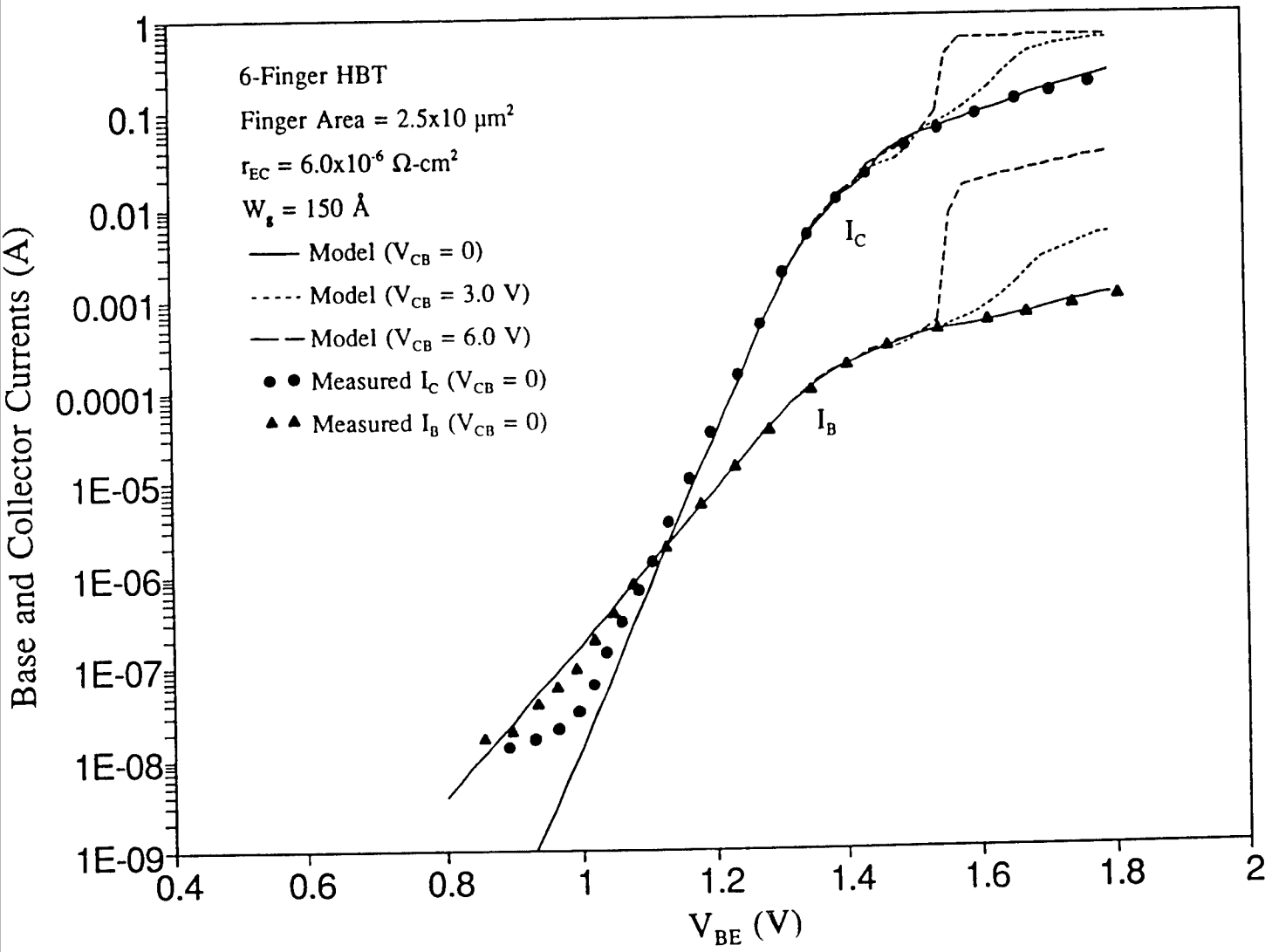


Fig. 3 Base and collector currents calculated form the present model for a 6-finger HBT at V_{CB} = 0, 3, and 6 V. Also included are the experimental data measured at V_{CB} = 0.

100 μm. This gives a thermal resistance of about 400 k/W at room temperature. As shown in the figure, both the collector and base currents (I_C and I_B) increase as V_{CB} is increased. This is due to the fact that the power generated, and therefore the temperature, in the HBT is increased as V_{CE} ($V_{CE} = V_{BE} + V_{CB}$) increases.

The model developed can also be used to calculate I_C versus V_{CE} characteristics for constant I_B. The results, together with experimental data, are given in Fig. 4. A negative slope on the I_C-V_{CE} characteristics is observed when the base current is large where the thermal effect becomes prominent. This can be attributed to the uneven increase of I_C and I_B as V_{CB}, or V_{CE}, is increased (see Fig. 3). The increased I_B due to the thermal effect reduces the base-emitter voltage V'_{BE} required to maintain that constant base current, which subsequently decreases the collector current.

Even with a ballast resistance, current crush can still prevail, only less obvious and if V_{CE} is sufficiently large. This is evidenced by the "minor crush" at $I_B = 0.35$ mA and $V_{CE} = 7$ V shown in Fig. 4. The current crush results from the even more asymmetrical increase in the base and collector currents at higher base current seen in Fig. 3 (at $V_{BE} = 1.45$ V and $V_{CB} = 6$ V). If I_B is fixed, then the voltage V'_{BE} required to maintain that I_B will also decrease sharply when V_{CE} is increased beyond a critical value, which then decreases sharply the collector current. It is said current crush occurs. As will be shown later, the current crush is much more apparent if the HBT does not have a ballast emitter resistance.

We next examine the effect of the emitter contact resistance r_{EC} on the HBT performance. Here we consider a 3-finger HBT. Fig. 5(a) shows the Gummel plot for three different r_{EC} and $V_{CB} = 2$ V. The same plot for $V_{CB} = 5$ V is given in Fig. 5(b). Clearly the increased r_{EC} suppresses both the collector and base currents at high V_{BE}. Of equal importance to note is the behavior of the collector current at large r_{EC} and when V_{CE} is high. Let us use the results in Fig. 5 to illustrate this point. Consider first the non-ballast r_{EC} case ($r_{EC} = 10^{-6}$ Ω-cm^2) and a fix $I_B = 1$ mA. When $V_{CB} = 2$ V (or $V_{CE} \sim 3$ V), $V'_{BE} \sim 1.62$ V and

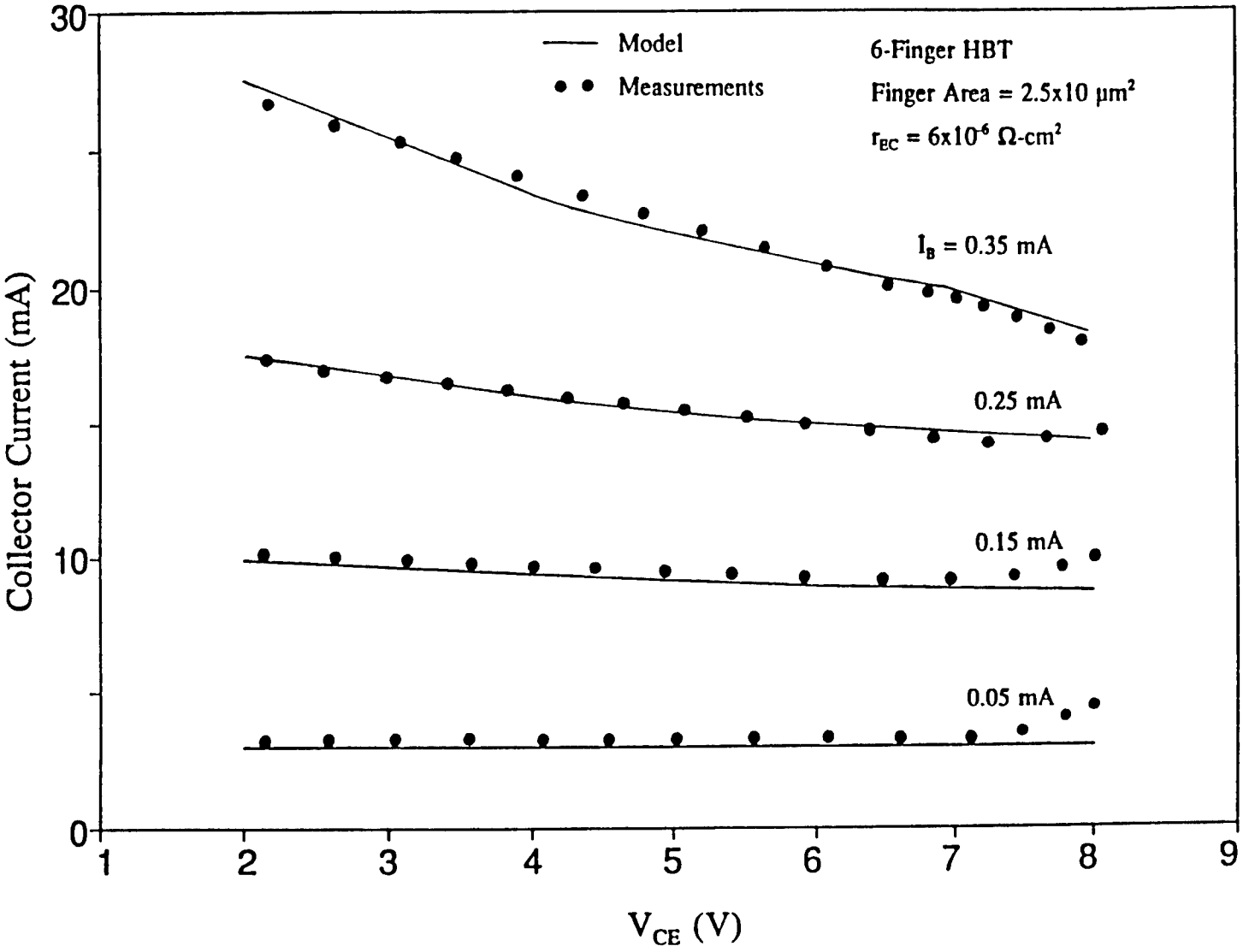
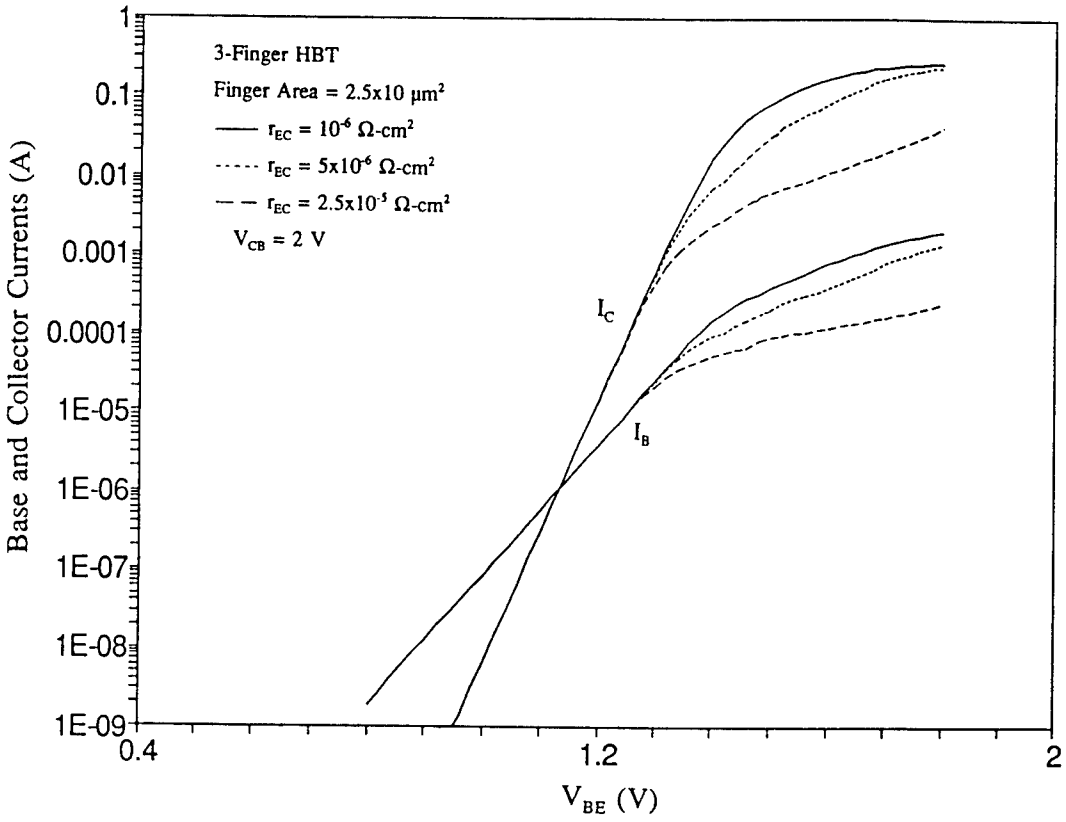
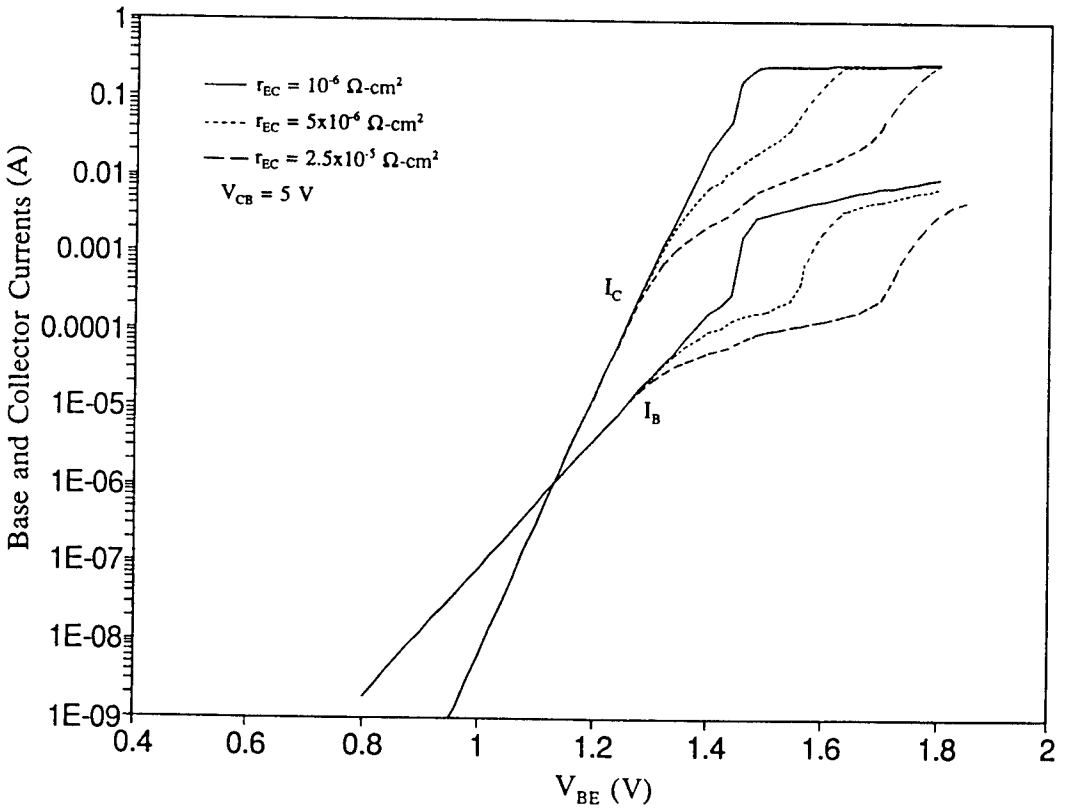


Fig. 4 Collector current vs. collector-emitter voltage characteristics as a function of constant I_B for a 6-finger HBT with ballast emitter resistance.



(a)



(b)

Fig. 5 Base and collector currents calculated from the present model for three different emitter contact resistances at (a) V_{CB} = 2 V; and (b) V_{CB} = 5 V.

$I_C \sim 0.12$ A. As V_{CB} is increased to 5 V ($V_{CE} \sim 6$ V), $V'_{BE} \sim 1.42$ V, and the corresponding I_C is about 0.06 A, which has been decreased to about 50% of its previous value. Now let us consider the HBT with a ballast resistance ($r_{EC} = 2.5\times10^{-6}$ Ω-cm^2) and fix I_B at 0.3 mA. At $V_{CE} = 3$ V, the corresponding I_C is 0.04 A, and at $V_{CE} = 6$ V, I_C is about 0.035 A. Thus the collector current in this case is decreased slightly, but not crushed, as V_{CE} is increased from 2 to 6 V.

Fig. 6 illustrates the current crush phenomenon in a multi-finger HBT without the ballast emitter resistance. The device considered has 4 emitter fingers, a finger area of 2.5×20 μm^2, and a low emitter contact resistance ($r_{EC} \sim 10^{-6}$ Ω-cm^2). The onset of current crush is observed at $V_{CE} \sim 5$ V when I_B is increased beyond 2.5 mA.

5. CONCLUSION

Temperature increase due to self-heating and thermal coupling has been known as a major factor limiting the performance of AlGaAs/GaAs multiple emitter HBTs. An analytical model has been developed to describe the d.c behavior of such devices. The effect of the graded heterojunction, a feature used frequently to improve the HBT emitter injection efficiency, is also accounted for in the model. We found that an elevated temperature in the HBT due to the thermal effect increases the base current more quickly than the collector current. This uneven current increase is the main mechanism contributing to current crush phenomenon observed in high power HBTs. Our results also suggest that while current crush can occur in all HBTs with sufficiently high current level and applied voltage, incorporating a ballast resistance in the emitter contact can reduce the extent of current crush.

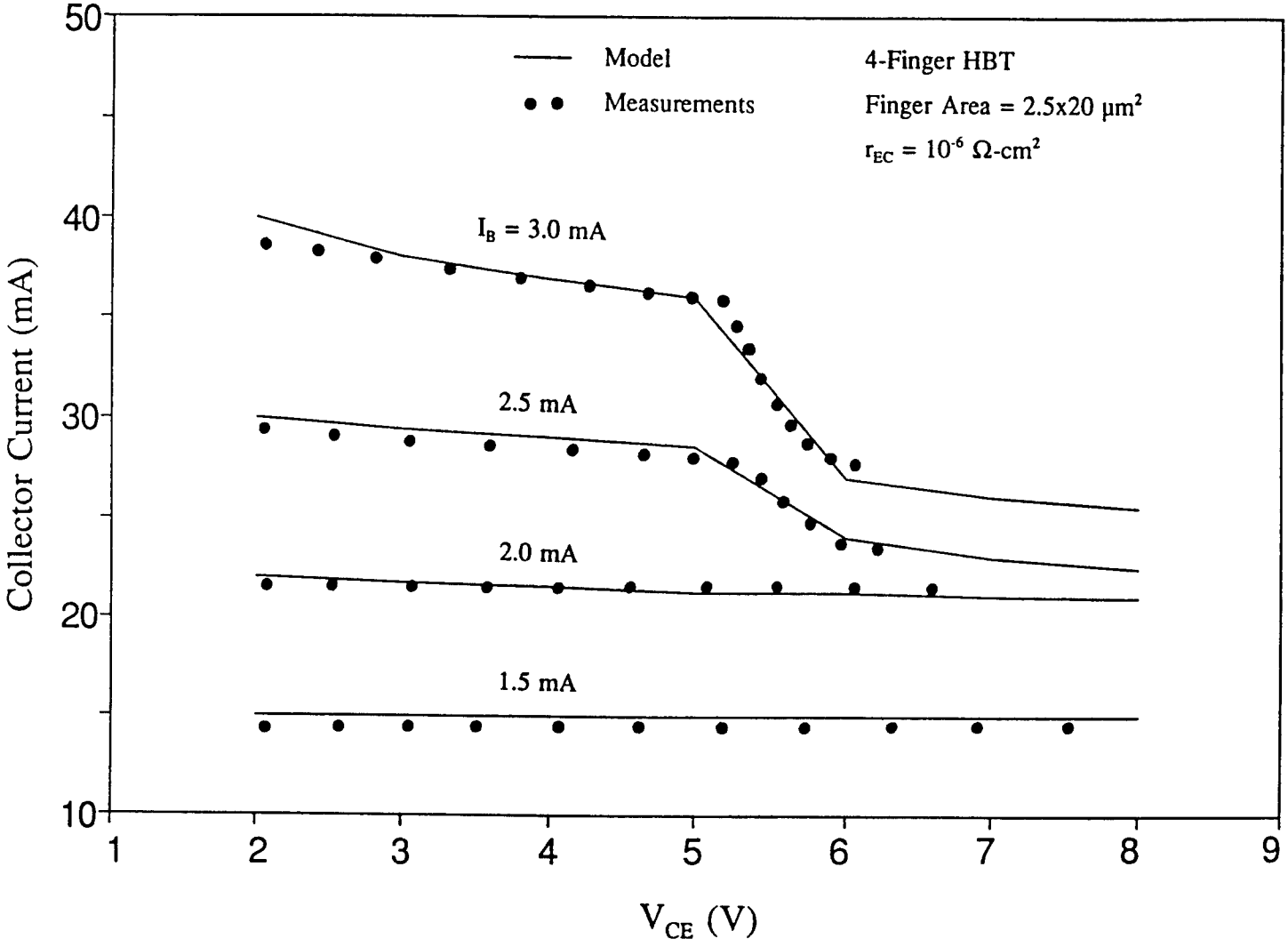


Fig. 6 Collector current vs. collector-emitter voltage characteristics as a function of constant I_B for a 4-finger HBT without ballast emitter resistance.

REFERENCES

[1] For a review, see: M. E. Kim, B. Bayraktaroglu, and A. Gupta, "HBT devices and applications," in HEMTs & HBTs: Devices, Fabrication, and Circuits (F. Ali and A. Gupta, Editors), Boston: Artech House Inc., 1991.

[2] N. L. Wang, N. H. Sheng, M. F. Chang, W. J. Ho, G. J. Sullivan, E. A. Sovero, J. A. Higgins, and P. M. Asbeck, "Ultrahigh power efficiency operation of common-emitter and common-base HBTs at 10 GHz," IEEE Trans. Microwave Theory and Tech., vol. 38, pp. 1381-1389, 1990.

[3] M. A. Khatibzadeh, B. Bayraktaroglu, and T. Kim, "12 W monolithic X-band HBT power amplifier," IEEE MTT-S International Microwave Symposium Dig., pp. 47-50, 1992.

[4] L. L. Liou, B. Bayraktaroglu, and C. I. Huang, "Thermal stability analysis of multiple finger microwave AlGaAs/GaAs heterojunction bipolar transistor," IEEE Int. Microwave Symp. Tech. Dig., 1993.

[5] G. B. Gao, M. S. Unlu, H. Morkoc, and D. L. Blackburn, "Emitter ballasting resistor design for current handling capability of AlGaAs/GaAs power heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. 38, pp. 185-196, 1991.

[6] G. B. Gao et al., "Thermal design studies of high-power heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. ED-36, pp. 854 -862, 1989.

[7] D. S. Whitefield, C. J. Wei, and J. C. M. Hwang, "Temperature-dependent large-signal model of heterojunction bipolar transistors," IEEE GaAs IC Symp. Tech. Dig., pp. 221-224, 1992.

[8] J. J. Liou, L. L. Liou, C. I. Huang, and Bayraktaroglu, "A physics-base heterojunction bipolar transistor including thermal and high-current effects," IEEE Trans. Electron Devices, to be published Sept. 1993.

[9] S. -C. Chen, Y. -K. Su, and C. -Z. Lee, "A study of current transport on p-N heterojunctions," Solid-St. Electron., vol. 35, pp. 1311, 1982.

[10] A. Das and M. S. Lundstrom, "Numerical study of emitter-base junction

design for AlGaAs/GaAs heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. 35, pp. 863, 1988.

[11] S. Tiwari and D. J. Frank, "Analysis of the operation of GaAlAs/GaAs HBT's," IEEE Trans. Electron Devices, vol. 36, pp. 2105, 1989.

[12] A. A. Grinberg, M. S. Shur, R. J. Fischer, and H. Morkoc, "An investigation of the effect of graded layers and tunneling on the performance of AlGaAs/GaAs heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. ED-31, pp. 1758, 1984.

[13] C. T. Kirk, Jr., "A theory of transistor cutoff frequency falloff at high current densities," IRE Trans. Electron Devices, vol. ED-9, pp. 164, 1962.

[14] J. J. Liou, "Calculation of the base current components and determination of their relative importance in AlGaAs/GaAs and InAlAs/InGaAs heterojunction bipolar transistors," J. Appl. Phys., vol. 69, pp. 3328, 1991.

[15] C. H. Henry, R. A. Logan, and F. R. Merritt, "The effects of surface recombination on current in $Al_xGa_{1-x}As$ heterojunctions," J. Appl. Phys., vol. 49, pp. 3530, 1978.

[16] J. J. Liou and J. S. Yuan, "Surface recombination current of AlGaAs/GaAs heterojunction bipolar transistors," Solid-St. Electron., vol. 35, pp. 805, 1992.

[17] W. Liu and J. S. Harris, Jr., "Diode ideality factor for surface recombination current in AlGaAs/GaAs heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. 39, pp. 2726, 1992.

[18] D. P. Maycock, "Thermal conductivity of silicon, germanium, III-V compound and III-V alloys," Solid-St. Electron., vol. 10, p. 161, 1967.

FIGURE CAPTIONS

Fig. 1 Two-dimensional HBT device structure illustrating the effective area through which the heat generated in the intrinsic HBT is dissipated.

Fig. 2 Topology of the multi-emitter finger HBT.

Fig. 3 Base and collector currents calculated form the present model for a 6-finger HBT at V_{CB} = 0, 3, and 6 V. Also included are the experimental data measured at V_{CB} = 0.

Fig. 4 Collector current vs. collector-emitter voltage characteristics as a function of constant I_B for a 6-finger HBT with ballast emitter resistance.

Fig. 5 Base and collector currents calculated from the present model for three different emitter contact resistances at (a) V_{CB} = 2 V; and (b) V_{CB} = 5 V.

Fig. 6 Collector current vs. collector-emitter voltage characteristics as a function of constant I_B for a 4-finger HBT without ballast emitter resistance.

GENERAL – PURPOSE ELECTROMAGNETIC MODELING OF COPLANAR WAVEGUIDE STRUCTURES IN MICROWAVE AND MILLIMETER – WAVE PACKAGES

Krishna Naishadham
Assistant Professor
Department of Electrical Engineering
Wright State University
Dayton, OH 45435

Final Report for:
AFOSR Summer Faculty Research Program
Wright Laboratory, WPAFB

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

September 1993

GENERAL – PURPOSE ELECTROMAGNETIC MODELING OF COPLANAR WAVEGUIDE STRUCTURES IN MICROWAVE AND MILLIMETER – WAVE PACKAGES

Krishna Naishadham, Assistant Professor
Department of Electrical Engineering
Wright State University
Dayton, OH 45435

ABSTRACT

Coplanar transmission lines and coplanar waveguide (CPW) discontinuities have been analyzed by the finite–difference time–domain (FDTD) method. The FDTD computational mesh is truncated by imposing absorbing boundary conditions on the walls, thus simulating outgoing waves appropriate to an open structure. The residual reflection from these boundaries introduces significant error in the frequency-domain parameters derived by Fourier transformation of the time–domain voltages and currents calculated by FDTD at appropriate reference planes. In this research, we have developed a new computationally-efficient method called the geometry rearrangement technique (GRT) to cancel the dominant contribution to the residual reflection from absorbing boundaries. We have applied the GRT to compute the effective dielectric constant of coplanar lines as a function of frequency, and the computed results have been found to be in good agreement with published data, thus indicating the effectiveness of the GRT in canceling residual reflection from absorbing boundaries. We have developed a computer program to calculate the S–parameters of CPW discontinuities. As a test case, we have computed the S–parameters of a coplanar line with an air–bridge, and the results are in excellent agreement with measurements reported elsewhere. We are continuing to validate our program by investigation of other CPW discontinuities such as L–bend with air–bridges and/or dielectric overlay, open-circuited stub, etc. This research is applicable to efficient characterization of MMIC elements and discontinuities, and high–density microwave and millimeter–wave packages, which are currently being investigated in aerospace research. We conclude the report with a summary of potential aerospace–related problems which can be solved with the tools developed in this research.

GENERAL – PURPOSE ELECTROMAGNETIC MODELING OF COPLANAR WAVEGUIDE STRUCTURES IN MICROWAVE AND MILLIMETER – WAVE PACKAGES

Krishna Naishadham

I. INTRODUCTION

A wide variety of transmission lines, discontinuities and components are employed in both microwave and high–speed digital circuits to accomplish various design objectives. These structures employ microstrip, coplanar waveguide (CPW), coplanar slotline, or a similar environment to support signal propagation, and are not amenable to characterization over a wide range of frequency, geometrical and physical parameters by simple quasi–static methods [1]–[3] or canonical measurements [4]. Therefore, the general–purpose analysis of these transmission line structures tends to be difficult and computer–intensive. With the high density of integrated circuits available today in both high–speed digital circuits and microwave and millimeter–wave integrated circuits (MMICs), the electronic packaging that encompasses these circuits has become very sophisticated and must be accurately characterized for reliable circuit design. For example, multilayer packaging incorporates vias which connect signal lines from one layer to another, or air bridges and wire bonds to connect the chip to a circuit component. Because these discontinuities can introduce complicated frequency–dependent capacitive and inductive effects, radiate energy, and excite unwanted package modes in the circuits, reliable methods of analysis are needed to develop equivalent circuit models and general–purpose CAD tools for these discontinuities. This report summarizes our effort to develop such tools for the analysis of CPW transmission lines and discontinuities. The CPW configuration has received considerable atten–tion in the recent past because of several factors which include the ease of circuit interconnection, low radiation loss, and the promising flip–chip packaging technique [5], [6].

A few accurate methods have been applied to analyze coplanar waveguides and CPW discontinuities. Knorr and Kuchler [7] determined the dispersion characteristics and characteristic impedance of CPWs using the spectral–domain method of moments (MoM). Similar analysis by Hasnian et al. [8] yielded the dispersion characteristics of picosecond pulses in CPWs. A full–wave analysis of CPW and slotline using the finite difference time domain (FDTD) method has been performed by Liang et al. [9] to extend the dispersion characteristics into the terahertz regime. The analyses of CPW discontinuities, however, have been quite sparse in comparison with CPWs. Simons and Ponchak [10] presented for the first time equivalent circuit models derived from measured S–parameters for open circuit, series gap, symmetric step and right angle bend CPW discontinuities. Dib et al. [11] analyzed shielded CPW discontinuities with air bridges using the spectral domain moment method. The air bridge, however, was not considered as an integral part of the MoM computation in [11]. Rittweger et al. applied the FDTD method for a full three–dimensional analysis of a CPW band–reject filter with air bridges [12]. Omar and Chow [13] obtained the S–parameters of a CPW with air bridges using the moment method and a simplified representation of the Green's functions in terms of complex images.

In this report, we employ the FDTD method to analyze CPWs and CPW discontinuities. A novel technique based on transmission line analysis is used to cancel the spurious reflection from the far–end absorbing boundary used to terminate the FDTD computational mesh. The method is illustrated by application to CPW transmission lines with and without air bridges, and CPW right angle bends. The FDTD method is chosen over other techniques, such as the MoM, because (a) a single time–domain simulation provides characterization of the discontinuities from DC well into the sub–millimeter wave frequency regime, (b) it can be readily extended to analyze multilayered package structures, and complicated, yet realistic configurations such as a right angle bend compensated by a dielectric overlay in the inner slot [10], packages lined with walls made of absorbing material, etc. Unlike the moment method, the FDTD method does not involve cumbersome analytical preprocessing pertinent to the derivation of Green's functions, and the solution of a system of linear equations by the inversion of a large dense matrix.

II. FINITE DIFFERENCE TIME DOMAIN METHOD

In this section, we summarize the salient features of the FDTD method, adopting the notation in [14]. The propagation of the fields in the structure of interest is governed by the Maxwell's equations

$$\mu \frac{\partial \mathbf{H}}{\partial t} = -\nabla \times \mathbf{E} \tag{1}$$

$$\epsilon \frac{\partial \mathbf{E}}{\partial t} = -\nabla \times \mathbf{H} \tag{2}$$

The media are assumed to be isotropic, lossless and piecewise homogeneous. The problem is discretized over a finite three–dimensional computational volume consisting of rectangular cubes, known as Yee unit cells. The six field components in (1) and (2) are interleaved in space on the six faces of the unit cell [14, Fig. 1]. Such an arrangement automatically satisfies the continuity of the tangential field components. The spatial and time derivatives in Maxwell's equations are replaced by their central difference approximations defined over the surface of the unit cell, resulting in a set of nodal E–fields, and a corresponding orthogonal set of nodal H–fields. The reader is referred to [14, eqs. (3)–(8)] for the finite difference approximations of the nodal fields derived from (1) and (2). With this scheme, the H–field node is displaced in space from the E–field node by half space step, and is updated (in time) one–half time step ahead of the E–field. This half–step staggering of the fields in time and space allow for the solution of the difference equations by using an explicit, iterative, leap–frog scheme, whereby the field at a given position and time instant is updated in an explicit manner utilizing the nearest–neighbor fields. For modeling dielectric interfaces, the interface is assumed to pass through the center of a unit cell, and a dielectric constant equal to the average of those of the two adjacent media is enforced at the interface. Since the material constants can vary arbitrarily from cell to cell, the FDTD method can conveniently analyze structures filled with arbitrarily inhomogeneous media. Conducting surfaces are treated by setting the tangential electric field components to zero.

The computational domain must be truncated into a region which is of finite and manageable size.

This is accomplished by introducing fictitious boundaries, where the so–called absorbing boundary conditions (ABCs) [15], [16] are specified locally such that the reflection from these boundaries is minimal. Thus, these boundaries simulate the condition of outgoing waves appropriate to open–region problems. The use of ABCs makes it feasible to solve a wide variety of microwave circuit and packaging problems using the FDTD method. For the structures considered in this report, the pulses on the transmission lines will be assumed to be normally incident on the mesh walls (or absorbing boundaries), so that Mur's first–order ABC can be used. We implement this ABC in difference–form as

$$E_0^{n+1} = E_1^n + \frac{v\,\Delta t - \Delta \ell}{v\,\Delta t + \Delta \ell}\left[E_1^{n+1} - E_0^n \right] \tag{3}$$

where E_0 represents the tangential electric field component on the mesh wall, E_1 represents the same component one node inside of the mesh wall, v is the maximum velocity of light in the computational volume, the superscript n is the time index, $\Delta \ell$ and Δt are the spatial and time steps, respectively. For a stable solution, the maximum time step is restricted by

the Courant condition [14, eq. (9)]. The side walls should be chosen far enough from the circuit in consideration so that the fringing fields which propagate tangential to the walls have neglible amplitude on the walls. This is especially critical for CPWs, whose transverse fields are quite spread out in phase, and whose absorbing boundaries on the sides have a metal sheet sandwiched between media with different propagation velocities. In Sec. III, we will address a novel method, termed as the Geometry Rearrangement Technique (GRT), which we have developed to minimize influence of the reflection from absorbing boundaries.

The excitation pulse used in this research has been chosen to be Gaussian in shape since its Fourier transform is also Gaussian centered at zero frequency. This property makes it useful in the analysis of frequency–dependent characteristics of planar transmission lines and discontinuities. The parameters of the Gaussian pulse are chosen according to the criteria detailed in [17]. The frequency-dependent circuit parameters are computed as described in [14].

III. GEOMETRY REARRANGEMENT TECHNIQUE

The FDTD computational domain is artificially terminated by absorbing boundaries in order to make the problem size manageable. Since these boundaries are not completely transparent, the results obtained by the FDTD method are subject to error caused by the reflection from the boundaries. Although various methods such as superabsorption [16] have been proposed to minimize this reflection in the time domain, the error is still large in the frequency domain, since the Fourier transform is very sensitive to even relatively small error caused by imperfect boundaries. For example, when the FDTD method is used to calculate the effective dielectric constant of a microstrip transmission line with the first-order Mur boundary condition [15], we have found that the frequency-domain result shows periodic oscillation with a peak value around 10–20% of the result predicted by an empirical dispersion formula [18]. This large error is caused by imperfect absorbing boundaries. In this research, we have developed a new simple method called the geometry rearrangement technique which can minimize the reflection from the far-end absorbing boundary to a much smaller level than possible by using conventional ABCs. An attractive feature of the GRT is that an accurate numerical result can be obtained directly from the Fourier-transformed frequency-domain data. In this section, we will summarize the main feature of this method. The reader is referred to [18] for more details. The improvement of the computed FDTD results on CPWs, accomplished by using the GRT, is demonstrated in the next section. Although we describe the method below for calculating the effective dielectric constant, we have successfully applied GRT to compute accurately the S-parameters of planar circuit discontinuities [19].

The conventional method of calculating the effective dielectric constant with the FDTD method [17] is by Fourier transformation of the voltage (or the electric field) at two different points on the transmission line. With V_{p1} and V_{p2} denoting the transforms of voltages at the points P_1 and P_2 (see Fig. 1), we have

$$e^{-\gamma(\omega)\Delta\ell} = \frac{V_{p1}}{V_{p2}} \tag{4}$$

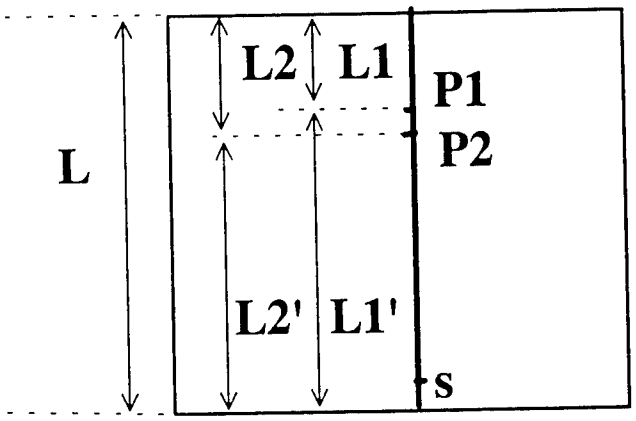


Fig. 1. Transmission line

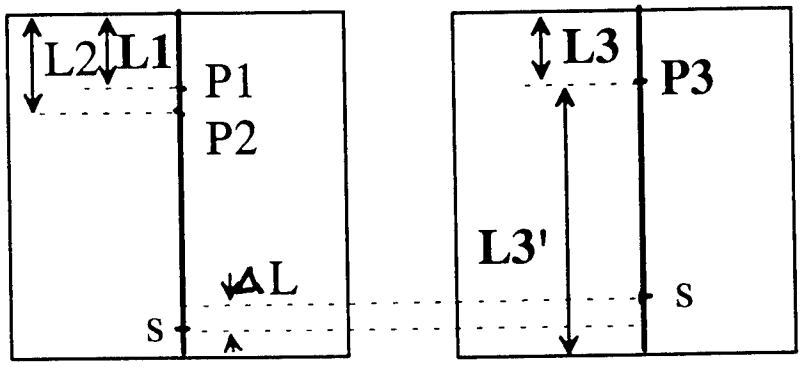


Fig. 2. Two identical transmission lines with source relocation in the second

where

$$\gamma(\omega) = \alpha(\omega) + j\beta(\omega) \tag{5}$$

$$\Delta\ell = \ell_2 - \ell_1 \tag{6a}$$

$$\beta(\omega) = \frac{1}{\Delta\ell}\,\mathrm{Ang}\left[\frac{V_{p1}}{V_{p2}}\right] \tag{6b}$$

The effective dielectric constant is given by

$$\epsilon_{reff}(\omega) = \frac{\beta^2(\omega)}{\omega^2\epsilon_0\mu_0} \tag{7}$$

where ω is the angular frequency and μ_0, ϵ_0 are the constitutive parameters of free space. We now examine how errors caused by reflection from the absorbing boundaries influence the computed dielectric constant. Let the reflection coefficient at the far–end boundary be Γ_f and that at the source–end boundary be Γ_s. Assume that the reflection caused by side, top and bottom walls is very small compared to that caused by the two end walls indicated above. The voltages V_{p1} and V_{p2} would then be the superposition of an incident wave and multiply reflected waves:

$$V_{p1} = V_{p1in}\left[1 + \Gamma_f e^{-2\gamma(\omega)\ell_1}\left[1 + \Gamma_s e^{-2\gamma(\omega)\ell_1'}\left[1 + \Gamma_f e^{-2\gamma(\omega)\ell_1}(1 + \ldots)\right]\right]\right] \tag{8}$$

$$V_{p2} = V_{p2in}\left[1 + \Gamma_f e^{-2\gamma(\omega)\ell_2}\left[1 + \Gamma_s e^{-2\gamma(\omega)\ell_2'}\left[1 + \Gamma_f e^{-2\gamma(\omega)\ell_2}(1 + \ldots)\right]\right]\right] \tag{9}$$

where V_{p1in}, V_{p2in} are incident voltages at points P_1, P_2. Eqs. (8) and (9) can be summed up in closed form:

$$V_{p1} = V_{p1in}\,\frac{1 + \Gamma_f\, e^{-2\gamma(\omega)\ell_1}}{1 - \Gamma_f\Gamma_s\, e^{-2\gamma(\omega)\ell}} \tag{10}$$

$$V_{p2} = V_{p2in}\,\frac{1 + \Gamma_f\, e^{-2\gamma(\omega)\ell_2}}{1 - \Gamma_f\Gamma_s\, e^{-2\gamma(\omega)\ell}} \tag{11}$$

where $\ell = \ell_1 + \ell_1' = \ell_2 + \ell_2'$. From (4), (10) and (11), we obtain:

$$e^{-\gamma(\omega)\Delta\ell} = \frac{V_{p1in}}{V_{p2in}} \frac{1 + \Gamma_f e^{-2\gamma(\omega)\ell_1}}{1 + \Gamma_f e^{-2\gamma(\omega)\ell_2}} \tag{12}$$

$\gamma(\omega)$ calculated from the above equation has an error caused by Γ_f. The true value of $\gamma(\omega) \equiv \tilde{\gamma}(\omega)$ should be calculated from

$$e^{-\tilde{\gamma}(\omega)\Delta\ell} = \frac{V_{p1in}}{V_{p2in}} \tag{13}$$

for the transmission line without boundaries (without reflections).

2.1. Source Relocation

If $\ell_1 = \ell_2$ in eq. (12), eq. (12) would be the same as eq. (13), which means $\gamma(\omega) = \tilde{\gamma}(\omega)$. But this cannot be realized on the single transmission line, since the two points P_1 and P_2 coalesce into one and $\Delta\ell$ becomes zero. However, if two transmission lines with the same characteristics are used instead of one, as in Fig. 2, we can use voltage V_{p3} at point P_3 on the second transmission line to *simulate* the voltage V_{p2} at point P_2 on the first transmission line. This is accomplished by moving the *source* in the second line (identical tothat in the first line) by a distance $\Delta\ell$ closer to the far–end boundary, so that the distance between P_3 and the source location S_2 on the second line is the same as the distance between P_2 and the source location S_1 on the first line. Eq. (12) then becomes:

$$e^{-\gamma(\omega)\Delta\ell} = \frac{V_{p1in}\,[1 + \Gamma_f e^{-2\gamma(\omega)\ell_1}]}{V_{p3in}\,[1 + \Gamma_f e^{-2\gamma(\omega)\ell_3}]} = \frac{V_{p1in}}{V_{p3in}} \tag{14}$$

since $\ell_3 = \ell_1$ and $V_{p3} = V_{p2}$. Therefore, the result calculated from voltages at the two points P_1, P_3 on the two transmission lines with the source simply relocated in the second, is an accurate numerical result. It does not contain the error introduced by Γ_f or Γ_s. A similar result can be obtained by moving the far–end boundary in the second line $\Delta\ell$ closer to the source, keeping the source position unchanged [18]. The latter rearrangement, called *boundary relocation*, would be convenient in those instances when

the reflection from the side–walls is significantly affected by the source location, such as CPW with low dielectric constant substrates [9].

IV. TEST RESULTS AND DISCUSSION

We first present computed results for the effective dielectric constant and characteristic impedance of a coplanar transmission line and compare our results with [9] to validate the FDTD algorithm.

The parameters used in the comptuation are (see Fig. 3):

Dielectric constant of the substrate	$\epsilon_r = 13.0$
Width of the center strip	W = 0.135 mm
Slot width	S = 0.065 mm
Width of the lateral strips	$W_\ell = 1$ mm
Thickness of the substrate	h = 0.5 mm
Cell size	$\Delta x = \Delta y = \Delta z = 0.0135$ mm
Time step	$\Delta t = 0.0176$ ps
Number of cells	Nx × Ny × Nz = 85 × 120 × 60
Number of time steps	N = 4096

The same geometry (Fig. 3) has been analyzed by using FDTD in [9]. However, instead of using the time–consuming superabsorption boundary condition [9], we use the simple first-order Mur condition. We use the magnetic wall symmetry around the x = 0 plane to reduce size of the computational domain.

Unlike the microstrip transmission line, the electric field distribution in the CPW is loosely bound to the substrate over the slots. It is difficult to simulate accurate absorbing boundary conditions on the side, top and bottom walls. Although the field of the wave traveling longitudinally along the CPW is mainly concentrated in the slot, the fringing field spreads out laterally away from the slot [20], and does not decay to small values near the absorbing boundary walls on the sides. To better simulate the distribution of the field in the slot, the amplitude of the Gaussian pulse for the excitation field is chosen as [20]:

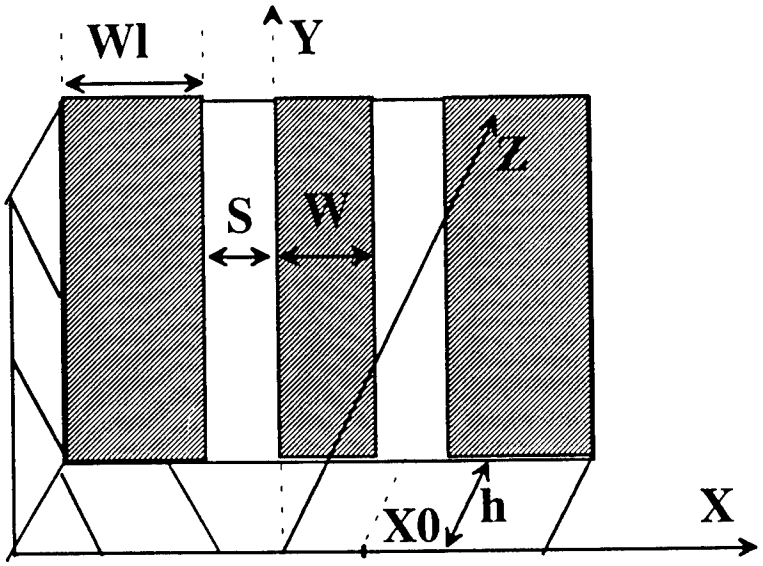


Fig. 3. Coplanar waveguide geometry

$$E_x = C\left[1 - \left[\frac{2(x-x_0)}{S}\right]^2\right]^{-\frac{1}{2}}, \quad |x-x_0| < \frac{S}{2} \tag{15}$$

at the air–dielectric interface, and as

$$E_x = C\left[1 + \left[\frac{2(x-x_0)}{S}\right]^2\right]^{-\frac{1}{2}}, \quad |x-x_0| < \frac{S}{2} \tag{16}$$

for a few cells above and below the air–dielectric interface over the excitation plane. In (15) and (16), C is an arbitrary constant and x_0 denotes the center of the slot (Fig. 3).

The numerical results for the CPW are shown in Figure 4. Four curves are shown — the conventional Fourier transform result, the GRT result, least squares fit to the GRT result, and an empirical result from [8]. It is observed that the maximum difference between the geometry rearrangement and the least squares results is less than 1.5 percent. The conventional Fourier transform result oscillates around the least squares or the geometry rearrangement result, but the difference between the conventional Fourier transform and the least squares is more than 12 percent. The least squares fit to the GRT result agrees reasonably well with the empirical formula. Therefore, considerable improvement in the accuracy of the effective dielectric constant can be obtained by using the geometry rearrangement technique with a simple absorbing boundary condition instead of the conventional FDTD implementation.

The characteristic impedance of the coplanar line depicted in Fig. 3 has been computed using FDTD, and is displayed in Fig. 5 as a function of frequency. Excellent agreement is observed between our result and that in [9, Fig. 6a].

Next, we consider a coplanar line with an air–bridge. The geometry is the same as that in Fig. 3, but an air–bridge with a height of 0.0135 mm, a width of 0.027 mm, and a length of 0.189 mm, has been positioned 0.729 mm from the $z = 0$ plane. The two S–parameters, S_{11} and S_{21}, where port 1 is located 0.594 mm from the $y = 0$ plane and port 2 is located 0.675 mm from the far–end wall, have been computed as described in [14], and compared with the measurements in [21]. The results, shown in Figs. 6 and 7, are in good agreement. No measured results are available beyond 30 GHz.

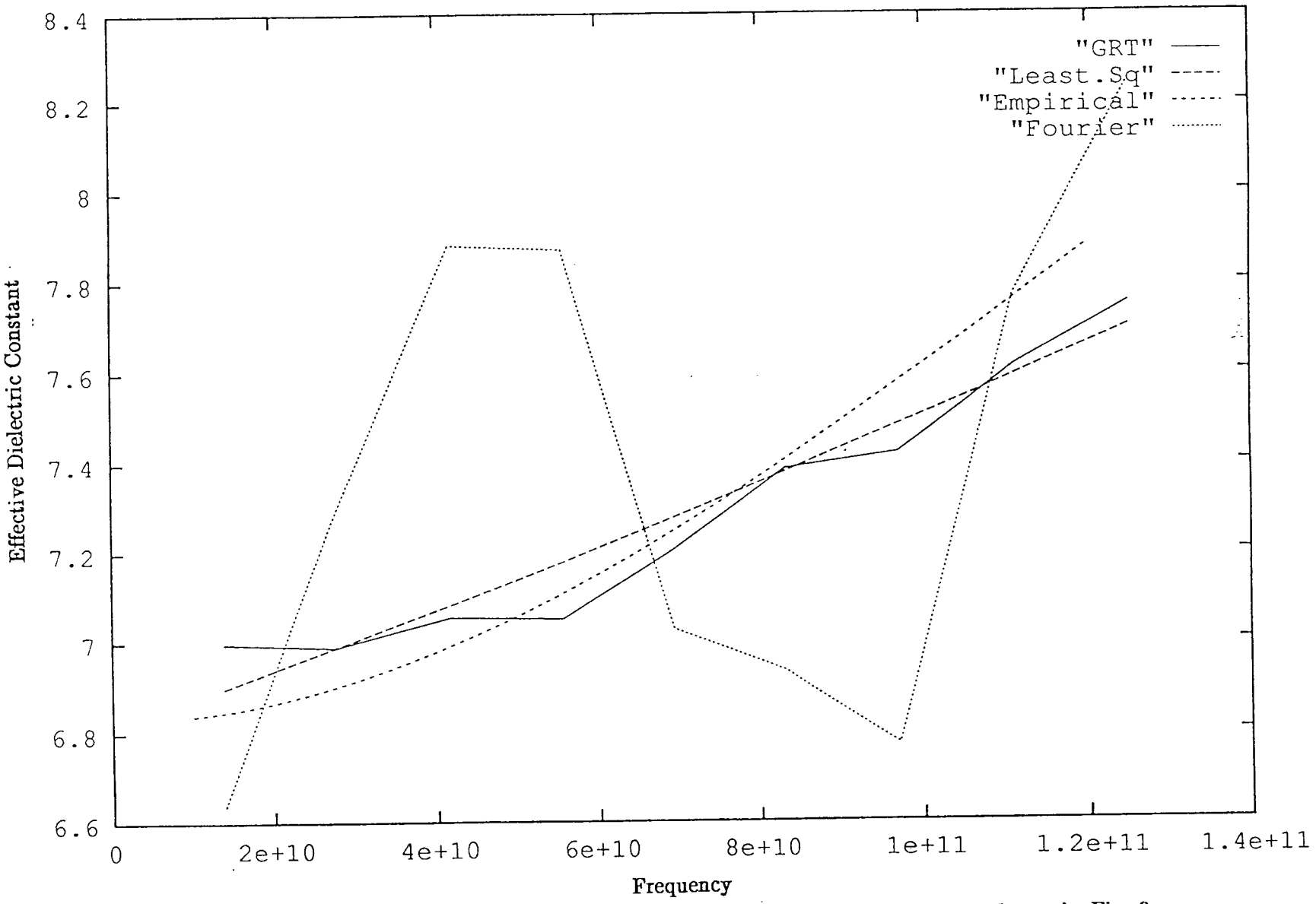


Fig. 4 Effective dielectric constant for the coplanar transmission line shown in Fig. 3.

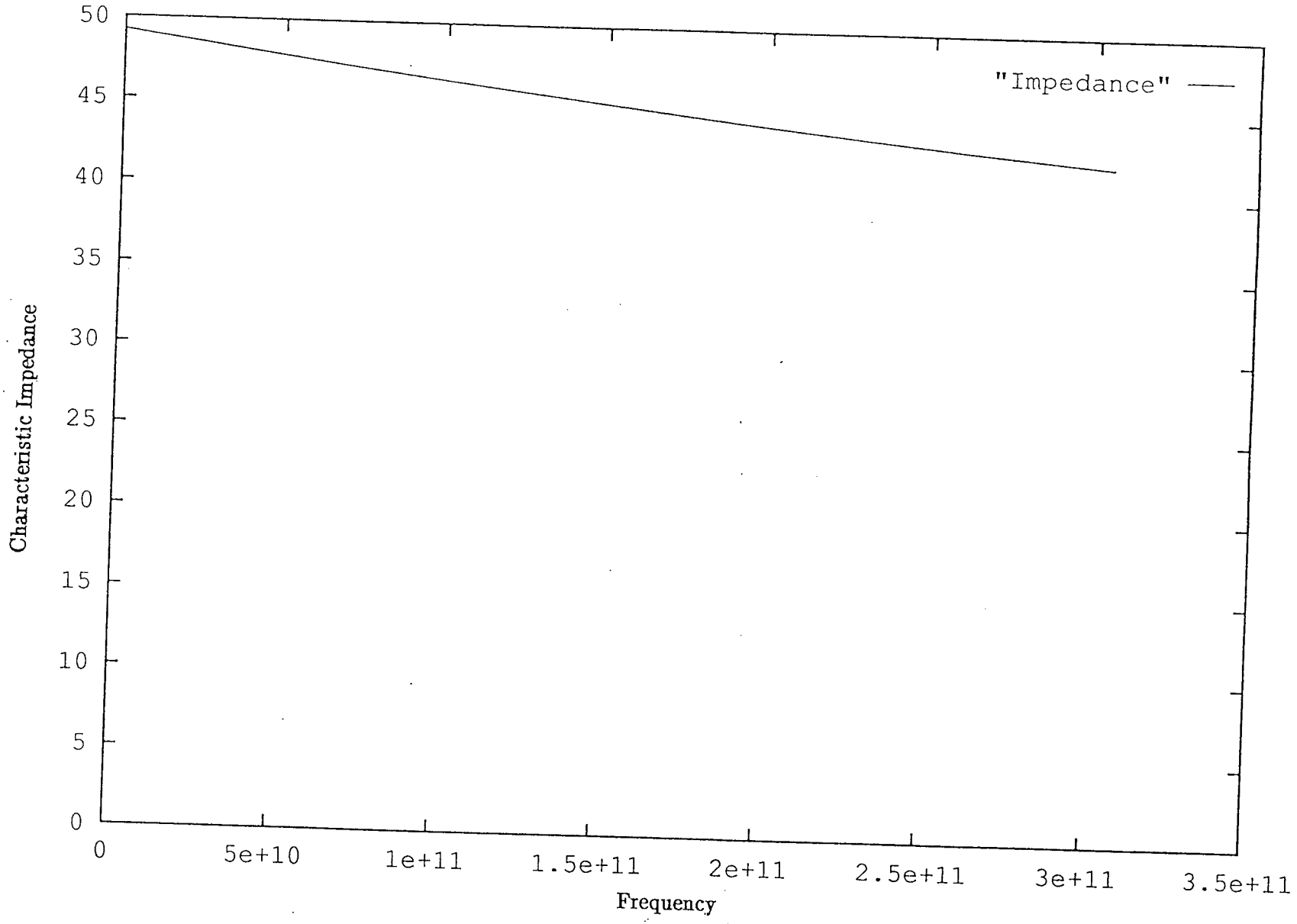


Fig. 5 Characteristic impedance of the coplanar transmission line shown in Fig. 3.

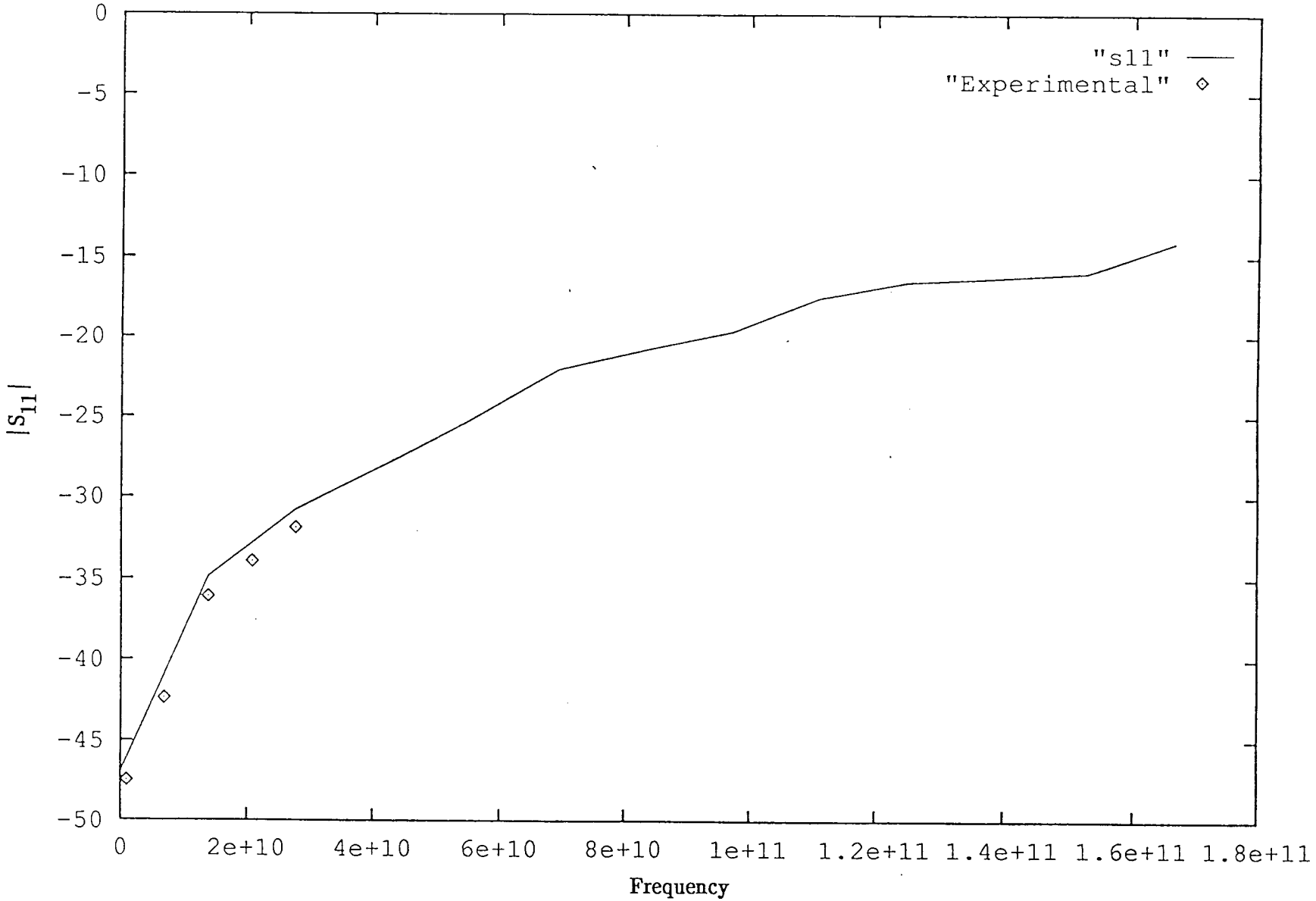


Fig. 6 $|S_{11}|$ **for the CPW in Fig. 3 with an air–bridge added.**

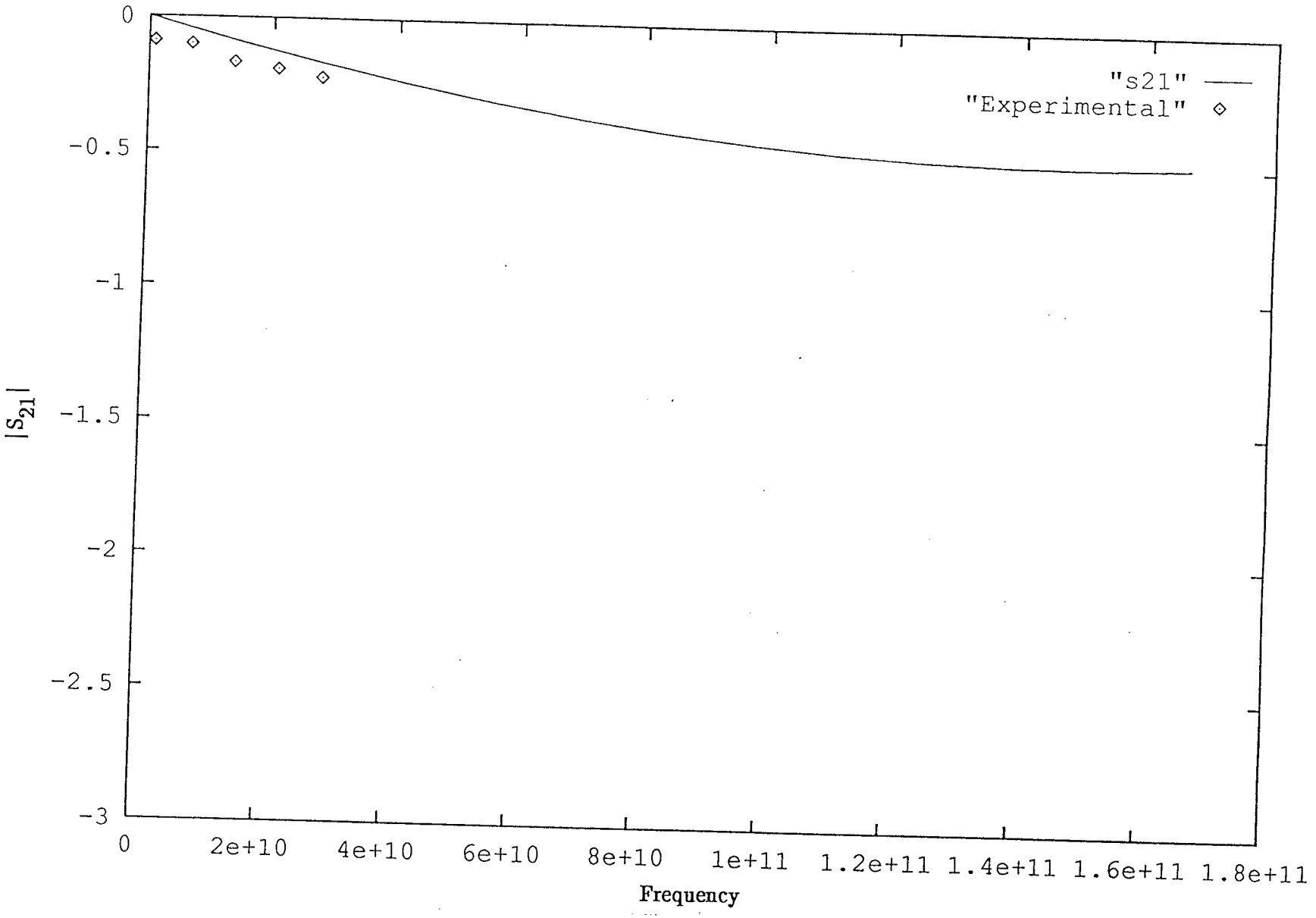


Fig. 7 $|S_{21}|$ for the CPW in Fig. 3 with an air–bridge added.

V. SUMMARY AND CONCLUSIONS

We have applied the finite–difference time–domain technique to analyze coplanar transmission lines and discontinuities. Unlike the microstrip geometry, the fields in CPW are spread out in space, and the absorbing boundaries used to terminate the computational mesh need special care. Although advanced boundary conditions such as superabsorption may be used, we have developed a simple method known as the geometry rearrangement technique for the FDTD analysis of planar transmission lines and discontinuities. We have determined the frequency–dependent effective dielectric constant of microstrip lines [18] and CPW, and computed results are found to be in good agreement with those obtained from empirical formulas and published literature. We are currently applying this method to compute the S–parameters of passive planar components in both microstrip and CPW configurations, and are examining how one may use th reflection coefficient estimated in the GRT to improve the FDTD result further. In comparison with advanced boundary conditions such as superabsorption, the GRT can save computer memory as well as computational time, and is considerably easier to implement on the computer. Also, the GRT *directly* minimizes the error introduced by Fourier transform operation without any need for curve–fitting. Our computed characteristic impedance for coplanar lines also agrees well with published data.

As a test case for CPW discontinuities, we have used our FDTD program to compute the S–parameters of a CPW with air–bridge. The computed results show excellent agreement with experimental data in [21]. We are currently applying the FDTD method to analyze other CPW structures such as bends and stubs in a multilayered structure with air–bridges and dielectric overlays on discontinuities.

VI. FUTURE WORK

The time span available on the AFOSR Summer Research Program is quite short for a comprehensive investigation. Given adequate resources, we wish to apply our FDTD algorithm to investigate the following problems relevant to MMICs, high–speed digital circuits, high–density microwave and millimeter–wave packaging:

a) study the influence of walls lined with surface–impedance, lossy or absorbing boundaries on the CPW characteristics and on coupling between discontinuities,

b) study the resonant behavior of packages lined with metallic walls and/or partitions — in particular the resonant coupling, Q–factors and losses,

c) extension of our algorithm to CPW discontinuities with inclined edges, such as mitered CPW bends,

d) develop and validate a design–base of equivalent circuits and S–parameters for CPW discontinuities in a general multilayered environment,

e) a thorough investigation of the loss mechanism in CPWs well into the terahertz regime.

REFERENCES

[1] C. Wei, R. F. Harrington, J. R. Mautz, and T. K. Sarkar, "Multiconductor transmission lines in multilayered dielectric media, "*IEEE Trans. Microwave Theory Tech.*, vol. MTT–32, no. 4, pp. 439–450, 1984.

[2] M. Kirschning, R. H. Jansen, and N. H. L. Koster, "Accurate model for open end effect of microstrip transmission lines," *Electron. Lett.*, vol. 17, pp. 123–125, 1981.

[3] P. Silvester and P. Benedek, "Microstrip discontinuity capacitance of right–angle bends, T–junctions, and crossings," *IEEE Trans. Microwave Theory Tech.*, vol. MTT–21, pp. 341–346, 1971.

[4] M. Caulton, J. J. Hughes, and H. Sobol, "Measurements on the properties of microstrip transmission lines for microwave integrated circuits," *RCA Review*, vol. 27, pp. 377–391, 1966.

[5] E. E. Davidson and G. A. Katopis, "Package electrical design," in *Microelectronics Packaging Handbook*, (R. R. Tummala and E. J. Rymaszewski, eds.), New York, NY: Van Nostrand Reinhold, ch. 3, pp. 111–165, 1989.

[6] Department of Defense (DoD) Advisory Group on Electron Devices (AGED) Special Technology Area Review (STAR) Report on *Microwave Packaging Technology*, February 1993.

[7] J. B. Knorr and K. Kuchler, "Analysis of coupled slots and coplanar strips on dielectric substrates," *IEEE Trans. Microwave Theory Tech.*, vol. MTT–23, no. 7, pp. 541–548, 1975.

[8] G. Hasnain, A. Dienes, and J. R. Whinnery, "Dispersion of picosecond pulses in coplanar transmission lines," *IEEE Trans. Microwave Theory Tech.*, vol. MTT–34, no. 6, pp. 738–741, 1986.

[9] G. Liang, Y. Liu, and K. K. Mei, "Full–wave analysis of coplanar waveguide and slotline using the time–domain finite–difference method," *IEEE Trans. Microwave Theory Tech.*, vol. MTT–37, no. 12, pp. 1949–1957, 1989.

[10] R. N. Simons and G. E. Ponchak, "Modeling of some coplanar waveguide discontinuities," *IEEE Trans. Microwave Theory Tech.*, vol. MTT–36, no. 12, pp. 1796–1803, 1988.

[11] N. I. Dib, P. B. Katehi, and G. E. Ponchak, "Analysis of shielded CPW discontinuities with air bridges," in *Proc. IEEE Microwave Theory and Techniques Symposium*, pp. 469–472, 1991.

[12] M. Rittweger, M. Abdo, and I. Wolff, "Full–wave analysis of coplanar discontinuities considering three–dimensional bond wires," in *Proc. IEEE Microwave Theory and Techniques Symposium*, pp. 465–468, 1991.

[13] A. A. Omar and Y. L. Chow, "A solution of coplanar waveguide with air–bridges using complex images," *IEEE Trans. Microwave Theory Tech.*, vol. 40, no. 11, pp. 2070–2077, 1992.

[14] D. M. Sheen, S. M. Ali, M. D. Abouzahra, and J. A. Kong, "Application of the three–dimensional finite–difference time–domain method to the analysis of planar microstrip circuits," *IEEE Trans. Microwave Theory Tech.*, vol. 38, no. 7, pp. 849–857, 1990.

[15] G. Mur, "Absorbing boundary conditions for the finite difference approximation of the time domain electromagnetic field equations," *IEEE Trans. Electromagn. Compat.*, vol. EMC–23, no. 4, pp. 377–382, 1981.

[16] K. K. Mei and J. Fang, "Superabsorption — A method to improve absorbing boundary conditions," *IEEE Trans. Antennas Propagat.*, vol. AP–40, no. 9, pp. 1001–1010, 1992.

[17] X. Zhang, J. Fang, and K. K. Mei, "Calculations of the dispersive characteristics of microstrips by the time–domain finite–difference method," *IEEE Trans. Microwave Theory Tech.*, vol. MTT–36, no. 2, pp. 263–267, 1988.

[18] X. P. Lin and K. Naishadham, "Geometry Rearrangement Technique — A new method for the FDTD analysis of dispersion in planar transmission lines," *IEEE Trans. Microwave Theory Tech.*, submitted.

[19] X. P. Lin and K. Naishadham, "Application of the geometry rearrangement technique to the computation of S–parameters of microstrip discontinuities by the FDTD method," in preparation.

[20] S. B. Cohn, "Slot–line field components," *IEEE Trans. Microwave Theory Tech.*, vol. MTT–20, no. 2, pp. 172–174, 1972.

[21] A. A. Omar, Y. L. Chow, L. Roy, and M. G. Stubbs, "Effects of air–bridges and mitering on coplanar waveguide 90° bends: theory and experiment," in *Proc. IEEE Microwave Theory and Techniques Symposium*, pp. 823–826, 1993.

WAVE PROPAGATION DURING HIGH VELOCITY IMPACT ON COMPOSITE MATERIALS

Serge ABRATE
Department of Mechanical and Aerospace Engineering
and Engineering Mechanics
University of Missouri-Rolla

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, D.C.

September 1993

WAVE PROPAGATION DURING HIGH VELOCITY IMPACT ON COMPOSITE MATERIALS

Serge ABRATE
Department of Mechanical and Aerospace Engineering
and Engineering Mechanics
University of Missouri-Rolla

Abstract

Understanding the response of laminated composite materials to ballistic impacts is of interest to those responsible for designing and maintaining airplanes that must be able to withstand some level of damage. As a first step towards understanding the process of damage development under high velocity impacts, an extensive series of tests was conducted in order to measure transverse normal stresses, at several locations through the thickness of the laminate, during high velocity impact. Significant damage is expected to be introduced from the initial passage of the impact generated stress wave and the tensile reflected wave. High compressive stresses induce shattering or shear failure near the impacted face while tensile stresses induce delaminations near the back face. These two zones are thought to be defined in the early stages of the impact.

The objective of the present invistigation is to develop a mathematical model capable of analyzing wave propagation through the thickness of the laminate during the early stages of the impact event in order to fully explain what is recorded during the experiments. A one dimensional analysis is developed accounting for nonlinear material behavior, the presence of adhesive layers where stress gages are located, and viscoelastic effects. A finite element model is developed and the governing equations are solved used Newmark's step by step integration. Results are presented showing that the numerical model is able to reproduce most of the significant features of the measured signals.

WAVE PROPAGATION DURING HIGH VELOCITY IMPACT ON COMPOSITE MATERIALS

Serge ABRATE

Introduction

Low velocity impacts can be expected to occur during manufacturing, normal operation, or maintenance. Tool drops or flying debris during take-off and landing of aircrafts are common examples. Ballistic impacts resulting in complete penetration are also to be considered for military aircraft in particular. Finally, hyper-velocity impact in which melting and vaporization of the target occurs is considered in relation to impacts of spacecrafts with meteorites or space debris at velocities measured in km/s. An understanding of impact damage is needed in order to design composite structures since impacts can reasonably be expected to occur during the life of the structure and the induced damage will significantly affect its behavior. Extensive efforts have been directed towards understanding the low velocity impact problem and resulted in a large number of publications which were reviewed comprehensively [1-3]. In comparison, ballistic impacts have received much less attention.

This investigation is concerned with the analysis of laminated composite materials subjected to ballistic impact. An extensive experimental program was conducted at Wright Patterson Air Force Base in order to measure transverse normal stresses under the impactor at several locations through the thickness. The initial phase of impact, in which a compressive wave travels through the thickness, is reflected at the back surface and travels back towards the impactor, is of particular interest. During that phase, high compressive stresses are generated near the impacted face. Subsequent tensile stresses are then generated near the back face which induce delaminations. Understanding the stress distribution through the thickness of the laminate during this first phase requires the development of a mathematical model capable of duplicating the signals recorded during the experiments. Such a model will further explain damage patterns observed during experiments.

In this report, the experimental results obtained prior to the present investigation will be reviewed first. Next, a discussion of the material properties of composite materials in the through-the-thickness direction is presented in order to justify the choice of constitutive relation to be employed in a model. The development of a finite element model for analyzing this problem is described next. Results are presented to show the ability of the model to explain the main features of the transverse normal stress history measured during the experiments. Suggestions for future work are also provided.

Experimental results

In this section, results of an extensive experimental study conducted prior to the currect investigation are reviewed. In this study, impacts by a 1/2" steel ball on a 128 ply $[(0/90/45/-45)_{16}]_s$ graphite-epoxy laminate will be considered, but similar results are also available for impacts on laminates with different number of plies. In order to place stress gages at different locations through the thickness (Fig. 1), the 128-ply symmetric balanced laminate is made-up of eight 16 ply $[0,90,\pm 45]_{2s}$ sublaminates that were cured separatedly. Stress gages were placed on selected interfaces and the sublaminates were then bonded together.

Considering a particular case (Panel D7.1), in which the initial velocity of the ball is 1280 ft/s, Fig. 2-a indicates that, after contact is made with the front face (line A in Fig. 2-a), the stress at the first gage location remains zero until the compressive wave arrives (line B). The transverse normal stress increases until it reaches a plateau (point C) and is followed by a second increase. Notice that compressive stresses are taken as positive here whereas, in solid mechanics, compressive stresses are negative. This convention is often used in impact

studies. The signal is then cut off due to saturation of the gage. At gage 3 (Fig. 2-b), the same pattern is observed: a first pulse arrives, the amplitude reaches a plateau and then a second pulse increases the amplitude significantly until the gage is saturated. At gage 5 (Fig. 2-c), the first pulse can clearly be seen and the second compressive pulse is also present. However, by the time the secondary pulse arrives at the gage location, the initial compressive pulse has reached the free surface (back face), was reflected as a tensile wave and has travelled back to the gage location. Therefore, for large times, the total transverse stress at a given depth is the sum of the primary and secondary stress wave amplitudes. This situation, with two waves travelling with different velocities, is similar to what is observed during impact on elasto-plastic metals in which an elastic wave is followed by a plastic wave [4].

Experiments indicated that the first compressive wave travels with a wave velocity of 3108 m/s which is approximately three times that of the second pulse. The amplitude of the first pulse is also attenuated as the wave propagates through the thickness and reaches the various gages successively. As an incident wave reaches an interface between two different materials, it is split into an reflected and a transmitted wave (Fig. 3). The ratio between the amplitudes of the transmitted and the incicident waves is dependent on the material properties at the interface [5]. It can be shown that, for differences in material properties that can be expected, the ratios obtained are of the same order of magnitude as those measured during experiments. Here, an adhesive layer between two 16-ply sublaminates introduces two such interfaces. If R is the attenuation introduced by one adhesive layer, and the amplitude of the first pulse at gages 1, 3 and 5 are called σ_I, σ_{III}, σ_V respectively, we have $\sigma_{III} / \sigma_I = R$, $\sigma_V / \sigma_I = R^3$.

Material properties

While no testing was performed to characterize the material used in this particular study, several publications report on the elastic properties of AS4/3501-6 graphite-epoxy. According to Iarve et al [6],

$E_1 = 140$ GPa, $E_2 = E_3 = 15.1$ GPa, $G_{12} = G_{13} = 5.51$ GPa, $\nu_{12} = .47$, $\rho = 1490$ kg/m^3

According to Lee and Sun [7, 8], for the same material,

$E_1 = 138$ GPa, $E_2 = E_3 = 9.65$ GPa, $G_{12} = G_{13} = 5.24$ GPa, $G_{23} = 3.24$ GPa, $\nu_{12} = \nu_{13} = .3$, $\nu_{23} = .49$

while AdTech Systems Research Inc. [9] used

$E_1 = 138$ GPa, $E_2 = E_3 = 10.3$ GPa, $G_{12} = G_{13} = 5.5$ GPa, $G_{23} = 3.1$ GPa.

Even though the modulus in the fiber direction is consistent, large variations in the other elastic properties are noticed. In particular, the through-the-thickness modulus E_3 which is of primary importance here, varies between 9.65 and 15.1 GPa.

In order to simplify their analyses, most investigators [6-9] model composite laminates as homogeneous anisotropic solids when studying impact problems. That is, a laminate with many layers is replaced by a single layer of anisotropic material that has equivalent elastic properties. The problem of determining the equivalent properties for a particular lay-up given the elastic properties of a single lamina has been addressed in several studies. Here, the approach taken by Sun and Li [10] is adopted and a simple program was written to perform that task for any lay-up. As an example, the variation of the three independent Poisson's ratios for angle-ply laminates as a function of fiber orientation were determined. Material properties given by Lee and Sun [8] were used and the results (Fig. 4) indicate that, as reported by previous authors [10, 11], Poisson's ratios can become negative. The equivalent modulus in the through-the thickness direction was shown to depend on the lamination scheme. The calculation of equivalent elastic properties will be discussed in more details elsewhere.

Taking E_3 = 15 MPa, we find that, with a compressive stress of 2.275 MPa for the the amplitude of the first pulse at gage 1 (Fig. 2-a), the strain is 15 %. One must then question the use of a linear elastic model for such loading levels. The nonlinear behavior of composite materials was reported by many authors. Of particular interest here is the work reported by Zhu et al. [12] who reported on quasi-static and dynamic compression tests for polyester resin and for Kevlar/polyester laminates in the through-the-thickness direction. Quasi-static tests were performed at a constant loading rate using an MTS machine and for the dynamic tests, a Hopkinson bar apparatus was used. For pure resin, the quasi-static tests showed a strong nonlinear behavior as strains become larger than 7 or 8 % (Fig. 5-a). With the much higher strain rates achieved in the dynamic tests, the response is substantially different, the initial modulus has increased significantly, suggesting strong viscoelastic effects. Results for through the thickness compression tests (Fig. 5-b) show the same trend. Ishai and Cohen [13] presented a detailled study of the non-linear behavior of filled epoxy loaded in compression and also discussed strain rate effects in those materials. Harding [14] presented a review of experimental work directed towards characterizing the material behavior of composite materials under impact using the split Hopkinson bar technique. Even though specimens were only subjected to inplane loadings, nonlinear behavior at high strains and strain rate effects are shown. El-Habak [15] also used the split Hopkinon bar technique to characterize composites loaded in compression and showed significant nonlinear behavior at high strain levels. Weirick [16] conducted a series of tests designed to characterize the mechanical behavior of an epoxy resin filled with glass microballons. The shock wave observed consisted of an elastic and a plastic wave due to the change in material behavior. The intial modulus of the microballon filled epoxy is lowered as the microballons are crushed and then increases as the material becomes more dense after the crushing is completed. In a study of contact laws for composites, Cairns [17] considered that AS4/3501-6 graphite-epoxy composites can be modelled as elasto-plastic materials. This brief literature survey indicates that, for the type of loading considered here, nonlinear behavior must be considered and that strain rate effects might be important.

In this investigation, a bilinear model (Fig. 6) is used for the stress-strain behavior of the material in the through the thickness direction. Knowing the density of the material, the modulus E_1 is selected so that the wave velocity matches that observed during the experiments. The modulus E_2 is taken as E_1 / 9 so that the velocity of the second pulse matches that observed during the experiments. The critical strain ϵ_{cr} is unknown and several values will be used in a parametric study to determine its influence. Since this is the first model of wave propagation through the thickness of the laminate during high velocity impact, strain rate effects are not modelled. Viscoelastic effects can easily be included in the model if proven necessary by a first analysis.

Formulation

For low velocity impacts, wave propagation through the thickness is not important and only the overall deformation of the structure must be accounted for in a mathematical model [1]. In the present case, on the other hand, wave propagation through the thickness is of primary interest since significant damage can introduced before overall plate deflections take place. That initial damage determines in a large measure the final damage pattern. Only a few studies [9, 18] were concerned with predicting stresses under the impactor during high velocity impact of composite laminates. In all cases, a linear elastic model was used in a two-dimensional elasticity approach. Even though the overall stress distribution is complicated, there is a small region under the impactor in which unidirectional behavior is observed. Previous studies [18-20] and detailled finite element analyses performed by this author, using the ADINA code, indicate that, considering the material to be linear elastic and homogeneous through the thickness, several important features of the gage response cannot be predicted. With such a model, a single pulse is obtained and therefore, several complicating factors must be accounted for in order to adequately model the problem. Since the laminate is made-up of 8 sublaminates bonded together after curing, an incident wave will split into a transmitted wave and a reflected wave as it reaches each interface between a sublaminate and an adhesive layer (Fig. 3). The presence of finite adhesive layers with different material properties must be accounted for. Interfaces between plies inside a sublaminate are not considered because it is assumed that there will be no resin rich region inside a sublaminate. Therfore, the modulus of elasticity and the density will be uniform through the thickness of each sublaminate. The nonlinear material behavior at high

strain values must also be accounted for in order to explain the presence of two pulses travelling at different velocities.

Therefore, as a first approach to this problem, a one dimensional analysis is proposed. Once proven successful, a two dimensional model can be developed. The equation of motions were developed using standard finite element procedures and Newmark's integration procedure was selected for time integration. For nonlinear cases, the initial stiffness approach was used and several iterations were performed for each time steps in order to eliminate unbalanced forces introduced by nonlinearities. The computer program runs on a personal computer and can easily be modified to include additional complicating effects.

Results

Example 1: Uniform rod impacted by a mass

This first example is selected in order to verify the present formulation on a simple case for which an exact closed-form solution is available. Graf [5] showed that the compressive strain in a uniform, semi-infinite, rod subjected to the impact of a rigid mass is given by:

$$\epsilon(x,t) = -\frac{V}{C} . H\langle ct-x\rangle . e^{-\frac{EA}{mc^2}(ct-x)} \qquad (1)$$

where ϵ is the strain, V the initial velocity of the projectile, c the wave velocity, EA the axial rigidity of the rod and m the mass of the projectile. In Eq. 1, $H\langle x\rangle$ is the Heavyside function where $H\langle x\rangle = 1$ when $x>0$ and $H\langle x\rangle = 0$ when $x<0$. At a given location, the magnitude of the stress decays exponentially with time and the rate depends on the mass of the projectile relative to that of the rod. With small projectiles, significant decays are observed as the projectile slows down significantly. With heavier projectiles, the decay is much smaller and it becomes negligible when the mass of the projectile becomes 100 times that of the rod. Notice that, in this example, the wave propagates in the x-direction without attenuation. That is, if ϵ is plotted versus x at time t_1, at time t_2, the graph is simply shifted in the x direction by $c(t_2-t_1)$. It is known [5], that a compressive wave is reflected as a tensile wave as it impinges on a free boundary. Then, Eq. 1 can be used to construct the solution for a finite rod impacted on the left by a rigid mass and free at the other end.

Fig. 7 shows that the finite element solution is in good agreement with the exact solution for the case where EA = c = V = 1. The stress history at the impacted end and at the midpoint on the rod are predicted accurately except where sharp discontinuities are present. The numerical solution can be improved by increasing the number of elements and reducing the time step. Eq. 1 implies that the initial strain under the projectile is given by the ratio of the projectile velocity to the velocity of compressive waves in the material.

$$\epsilon = \frac{V}{C} \qquad (2)$$

This relationship was used by Robinson and Davies [21] to differentiate between low velocity impacts, for which wave propagation through the thickness is not important, from high velocity impacts for which damage is indroduced during the first travel of the compressive wave through the thickness. It can be used to estimate the strain induced during impact and was found to provide good estimate when compared to either experimental data or published results ([9]).

This example indicates that the basic formulation employed here and its computer implementation are correct so that the program can now be used to investigate more complex cases.

Example 2: Rod with a bond impacted by a mass

The same problem discussed previously was modelled using 85 elements. In order to model the presence of an adhesive layer, the modulus for the element at the center of the rod was taken to be equal to .5, while the modulus for all other elements was equal to 1. The impact response is identical to that shown in Fig. 7, until the compressive wave reaches the adhesive layer. Afterwards, the behavior is substantially different due to multiple reflections from that layer (Fig. 8). Therefore, the presence of adhesive layers can substantially affect the transient response to impact loading.

Example 3: Impact on a 128-ply laminate

The case of a 128-ply laminate impacted by a rigid mass is considered next. The analysis does not account for the presence of adhesive layers at the gage locations. The evolution of the stress under the impactor and at the first two interfaces is shown in Fig. 9. An elastic pulse is then followed by a slower moving "plastic" pulse. Since the appropriate stress-strain behavior of the material is not known, it was decided to vary the value of the critical strain (4, 8 and 12 %). A total of 80 identical elements where used and the stress in elements 1, 10 and 20 represents the stress under the impactor, and at gages 1 and 2 respectively. Fig. 9-a shows the two pulses propagating at different velocities. The amplitude of the first pulse is equal to the product of E_1 by the critical strain which is 4 % in this case. This value is much lower than what is observed during experiments. Fig. 9-b shows the stress in elements 1 and 10 again for a critical strain of 8 %. The amplitude of the first pulse becomes closer to the experimental value. The dashed lines show the effect of reducing the time step and it indicates that the elastic pulse is a square pulse while the second pulse is largely unaffected by the change. The results in Fig. 9-c indicate that, as ϵ_{cr} increases, stresses increase but that the difference between the "yield" stress and the maximum stress becomes smaller. Fig. 7 indicates that, as the time step is increased, the elastic pulse becomes sharper. Therefore, with this model, as the elastic pulse arrives at a given location, the material experiences an infinite strain rate as the stress increases instantaneously from zero to the yield stress. This is then followed by a period in which the stress remains constant before the second pulse arrives. However, as shown by the quasi-static and dynamic tests conducted by Zhu et al [12], strain rate effects have a very significant influence on the stress-strain behavior of the material. These strain rate effects are not included in the present study and, therefore, a sharp pulse is predicted instead of the more and more rounded pulse observed in the experiments (Fig. 2).

Conclusions

The present study is a first attempt at modeling the transient response of a composite laminate to high velocity impact. A finite element model was developed in order to perform a one-dimensional analysis of the transient response of a composite laminate subjected to high velocity impact. It was found that, in order to capture the main features of the response determined experimentally, the nonlinear behavior of the material which is to be expected at strains exceeding 7-8 % must be included in the model. It is then shown that the transverse normal stress at a given point consists of an elastic pulse which is followed by a second pulse propagating at a much slower velocity. The presence of adhesive layers, indroduced in order to imbed stress gages through the thickness of the laminate, accounts for significant attenuation of the amplitude as the wave propagates through the thickness. Numerical results indicate that the initial loading rate is very high and that very rapidly the stress reaches a constant level until the second pulse arrives. The experiments indicate that the loading occurs more progressively and that the wave front is not sharp as predicted by the model. In addition, the wave front becomes more and more rounded as the wave travels through the thickness. This difference is attributed to strain rate effects which are not accounted for in the present model but which are known to be significant in the through the thickness direction.

The present investigation should be extended in order to strengthen the conclusions reached here. From the analysis of the experimental results and the numerical model developed here, nonlinear and viscoelastic effects are expected to aplay an important role in the high velocity impact of graphite-epoxy laminates. Experiments should be conducted to verify this conclusion and obtain data to be used for better correlation between experiments and analysis. Quasi-static through the thickness tests should be conducted for strains up to at least 15 % in order to validate the bi-linear model used and to obtain precise material data for the material used. More complex material models can easily be incorporated in the model if necessary. Hopkinson bar tests can be conducted in order to quantify the strain rate effects. Strain rate effects can easily be included in the model and in combination with the use of accurate material properties from experiments good agreement between experimental and numerical results will be obtained. In a second phase, the model can be extended to three dimensions in order to determine the extent of impact induced damage throughout the thickness. Detailled suggestions for future work will be submitted in an upcoming proposal.

References

1- Abrate S., "Impact on Laminated Composite Materials, "Applied Mechanics Reviews, Vol 44, No 4, pp 155-190, 1991

2- Abrate S., "Impact on Laminated Composite Materials: Recent Advances," Proc of CSME FORUM 1992, Montreal, Canada, June 1-5, 1992

3- Cantwell W.J., Morton J., "The Impact Resistance of Composite Materials - A Review," Composites, Vol 22, No 5, pp 347-362, 1991

4- Zukas J.A., Nicholas T., Swift H.F., Greszczuk L.B., Curran D.R., Impact Dynamics, J. Wiley & Sons, 1981

5- Graf K.F., Wave Motion in Elastic Solids, Ohio State University Press, 1975

6- Iarve E.V., Haq I.U., Soni S.R., "The Effect of Structural Inhomogeneity of Composite Laminates in Modeling the Impact Induced Deformation," J. Reinforced Plastics and Composites, Vol 12, pp 404-413, 1993

7- Lee S.W.R., Sun C.T., "Modeling Penetration Process of Composite Laminates Subjected to a Blunt-Ended Punch," Int. SAMPE Tech. Conf., Vol. 23, Kiamesha Lake, NY, pp 624-638, Oct 21-24, 1991

8- Lee S.W.R., Sun C.T., "A Quasi-Static Penetration Model for Composite Laminates," J. Composite Materials, Vol 27, No 3, pp 251-271, 1993

9- AdTech Systems Research, Inc., "Dynamic Response of Composite Plates to Impact Load," Final Technical Report submitted to Wright Laboraties, Flight Dynamics Directorate, Vehicle Subsystems Division

10- Sun C.T., Li S., "Three-Dimensional Effective Elastic Constants for Thick Laminates," J. Composite Materials, Vol 22, pp 629-639, 1988

11- Herakovich C.T., "Composite Laminates with Negative Through the Thickness Poisson's Ratios," J. Composite Materials, Vol 18, pp 447, 1984

12- Zhu G., Goldsmith W., Dharan C.K.H., "Penetration of Laminated Kevlar by Projectiles - Experimental Investigation," Int. Solids Structures, Vol 29, No 4, pp 399-420, 1992

13- Ishai O., Cohen L.J., "Effect of Fillers and Voids on Compressive Yield of Epoxy Composites," J. Composite Materials, Vol 2, No 3, p 302, 1968

14- Harding J., "Mechanical Behavior of Composite Materials Under Impact Loading," in Shock-Wave and High-Strain-Rate Phenomena in Materials, M.A. Meyers, L.E. Murr, K.P. Staudhammer, Eds, M. Dekker Inc, N.Y., 1992,pp 21-34

15- El-Habak A.M.A., "Mechanical Behaviour of Woven Glass Fibre-Reinforced Composites under Impact Compression Load," Composites, Vol 22, No 2, pp 129-134, 1991

16- Weirick L.J., Shock Characterization of Epoxy-42 Volume Percent Glass Microballons," in Shock-Wave and High-Strain-Rate Phenomena in Materials, M.A. Meyers, L.E. Murr, K.P. Staudhammer, Eds, M. Dekker Inc, N.Y., 1992, pp 935-946

17- Cairns D.S., "Simple Elasto-Plastic Contact Laws for Composites," J. Reinforced Plastics and Composites, Vol 10, No 4, pp 423-433, 1991

18- Bogdanovich A.E., Yarve E.V., "Numerical Analysis of the Impact Deformation of Plates Made of Composites," Mechanics of Composite Materials, Vol 25, No 5, pp 586-599, 1990

19- Liu G.R., Tani J., Ohyoshi T., Watanabe K., "Transient Waves in Anisotropic Laminated Plates, Part 1: Theory," J. Vibration and Acoustics, Vol 113, pp 230-234, 1991

20- Liu G.R., Tani J., Ohyoshi T., Watanabe K., "Transient Waves in Anisotropic Laminated Plates, Part 2: Application," J. Vibration and Acoustics, Vol 113, pp 235-239, 1991

21- Robinson P., Davies G.A.O., "Impactor Mass and Specimen Geometry Effects in Low Velocity Impacts of Laminated Composites," Int. J. Impact Eng., Vol 12, No. 2, pp 189-207, 1992

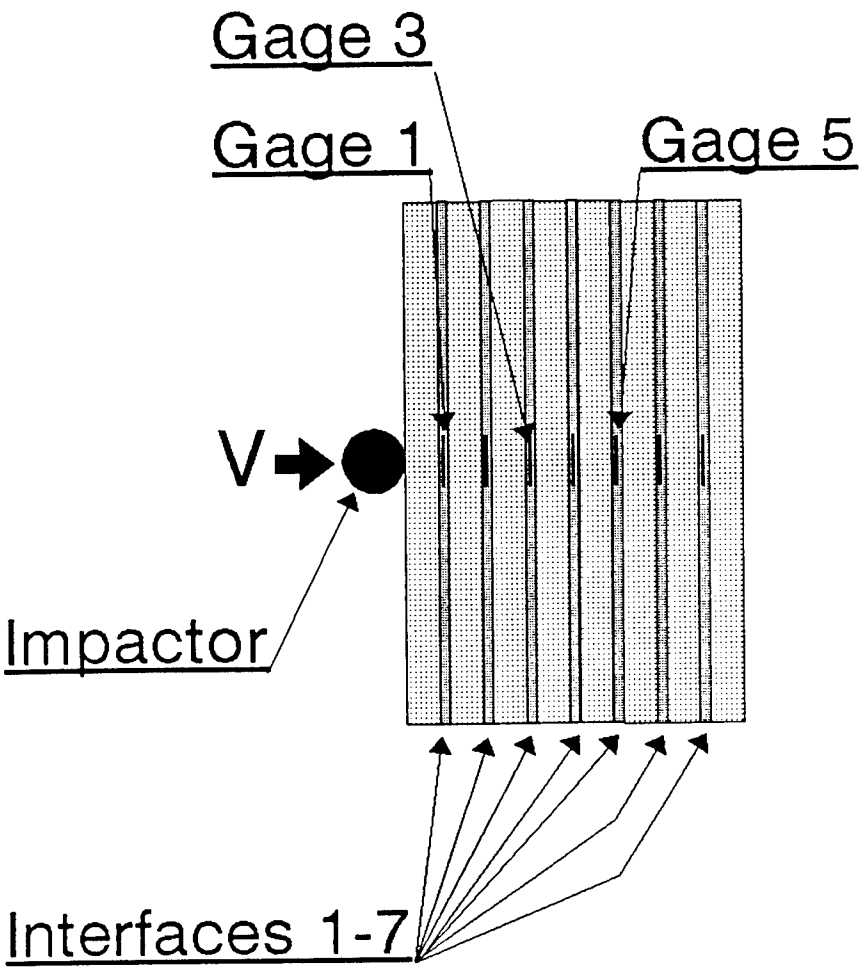


Figure 1: Location of gages in 128-ply lay-up

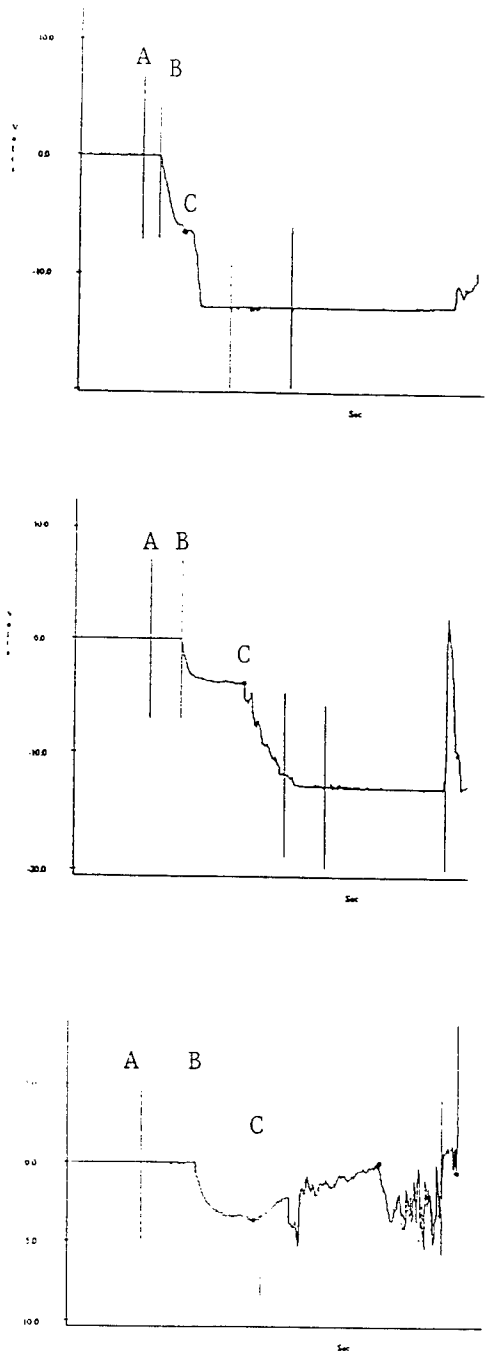


Figure 2: Transverse normal stresses under the impactor, a- Gage 1, b- Gage 3, c- Gage 5

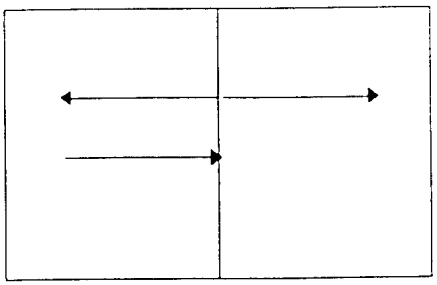


Figure 3: Reflection and transmission of an incident wave at the interface between dissimilar materials

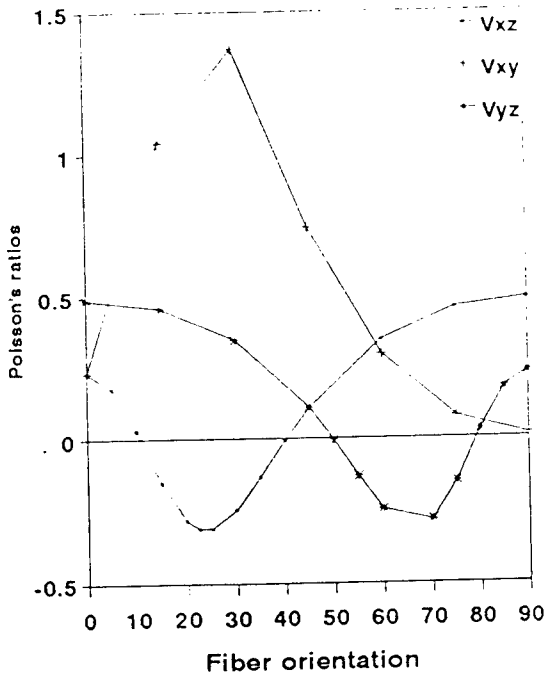
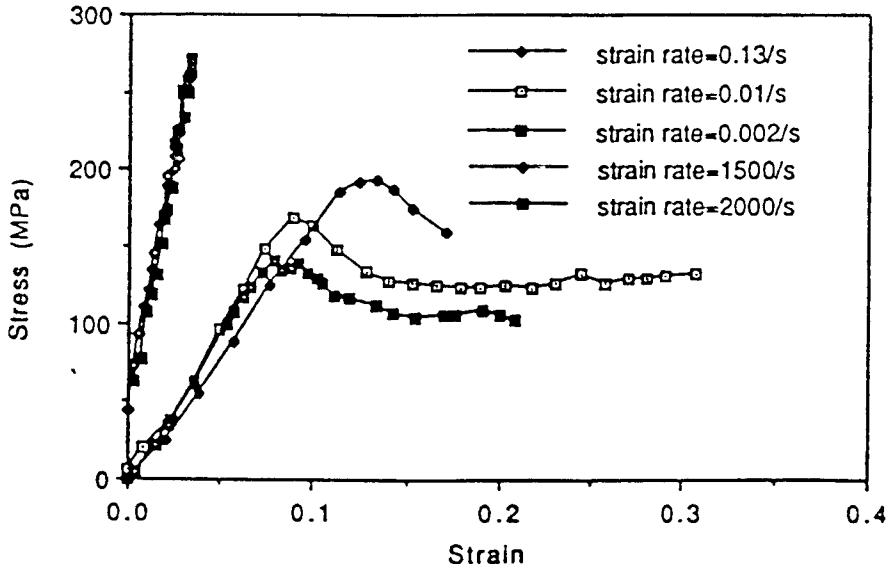
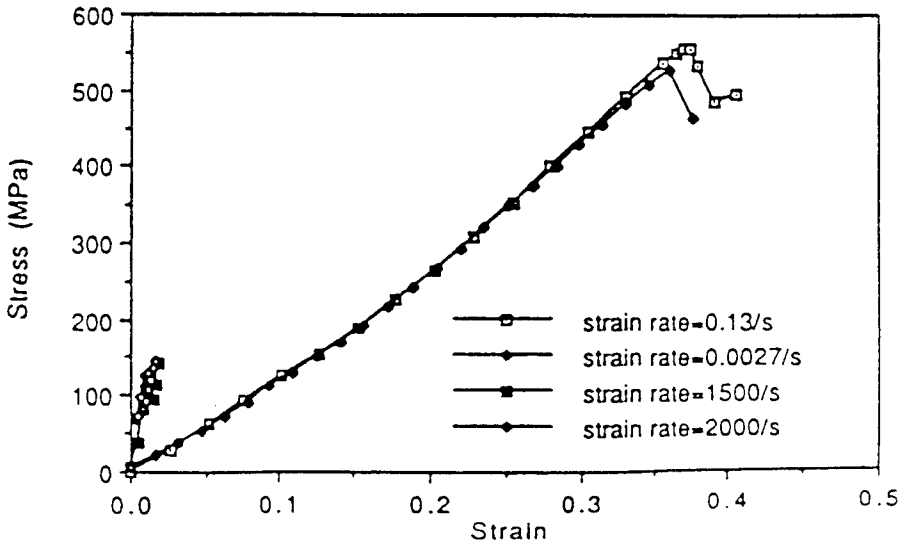


Figure 4: Poisson's ratio for angle-ply laminates as a function of orientation angle



Quasi-static and dynamic compression test results for polyester resin.



Comparison of quasi-static and dynamic through-thickness compression test results for Kevlar/polyester laminates.

Figure 5: Stress-strain curves in compression (Zhu et al. [12]). a- pure resin, b- laminate

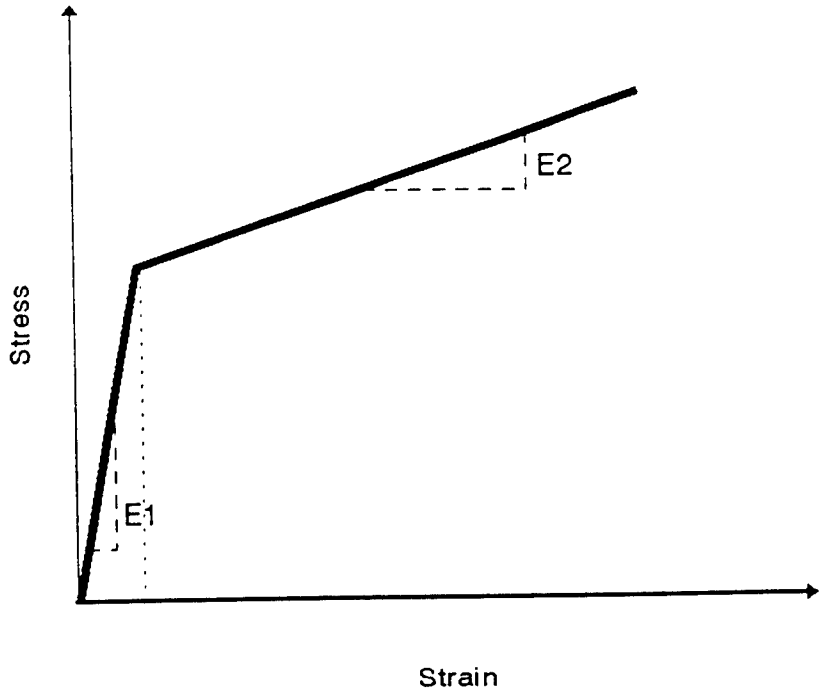


Figure 6: Idealized Stress-strain curve

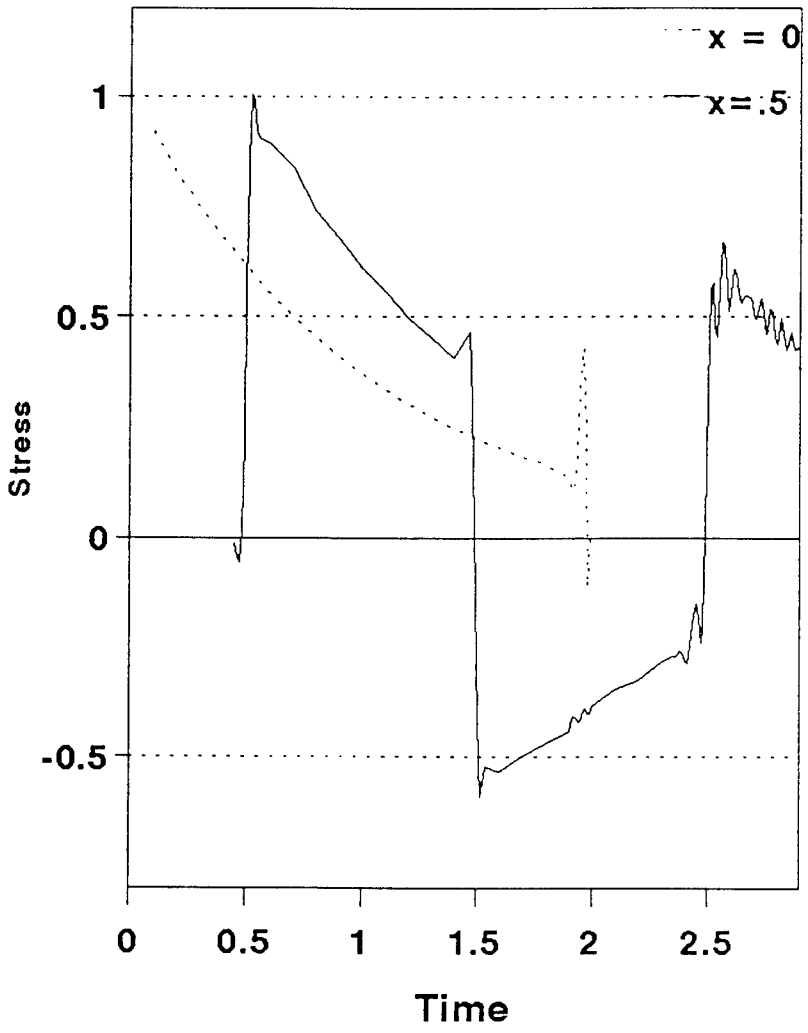


Figure 7: Impact on a finite uniform bar

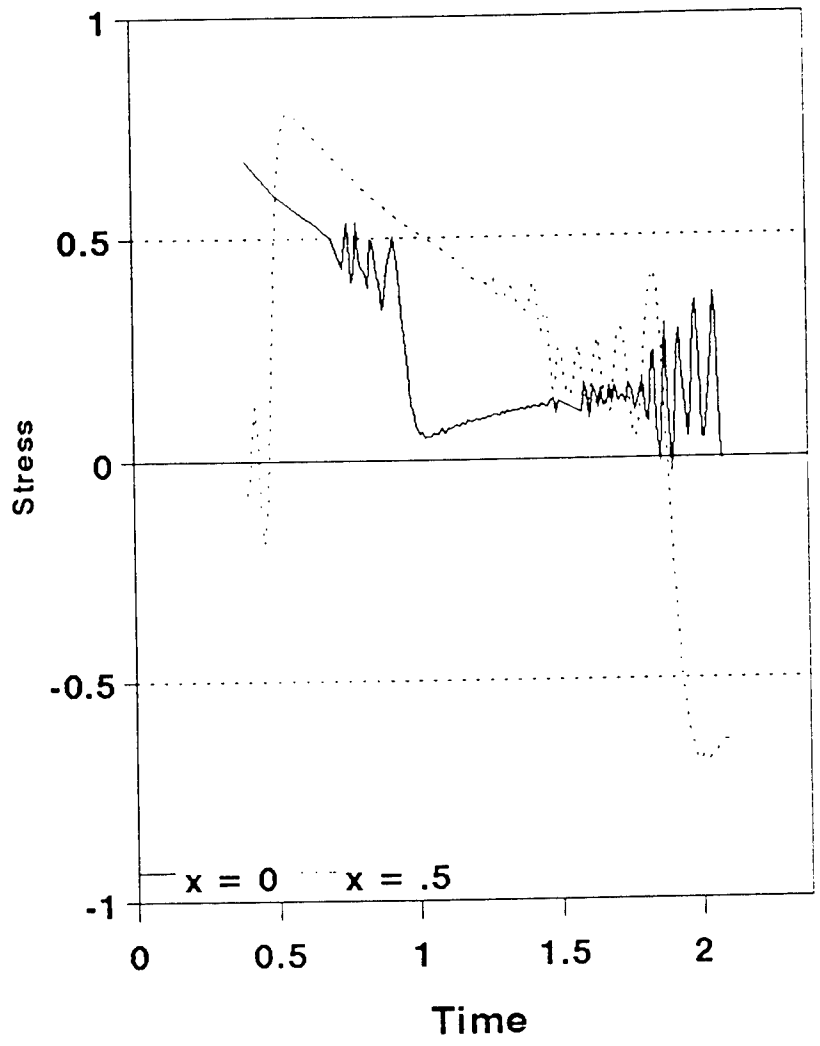


Figure 8: Impact on a rod with a bond

(a)

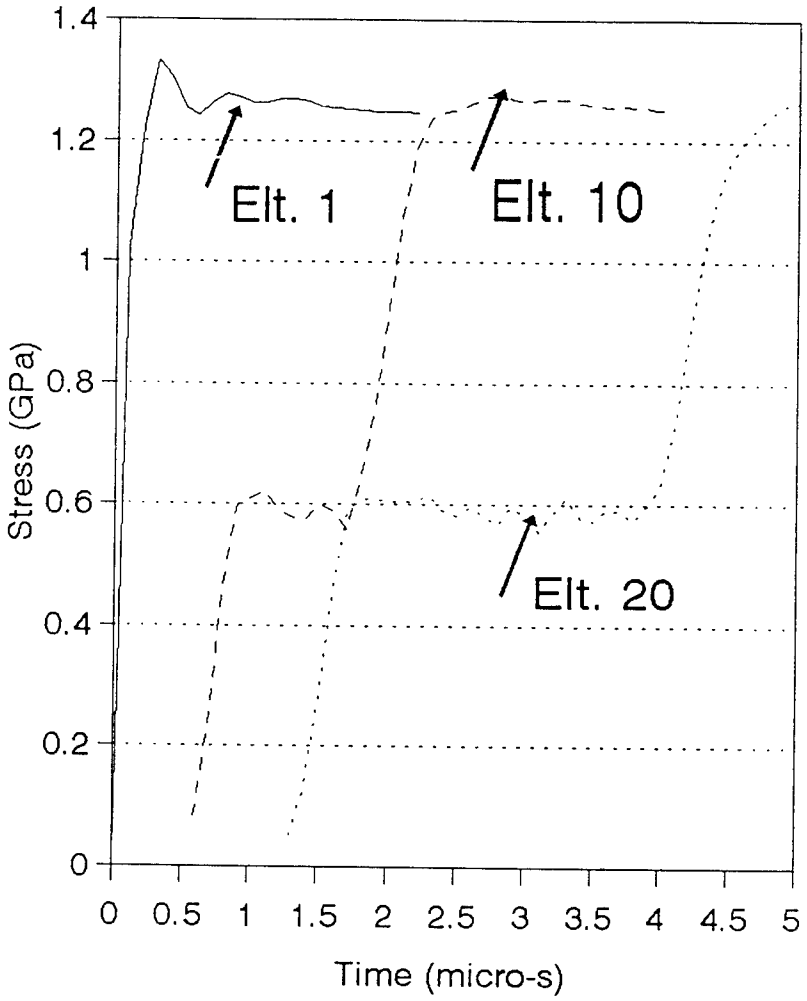
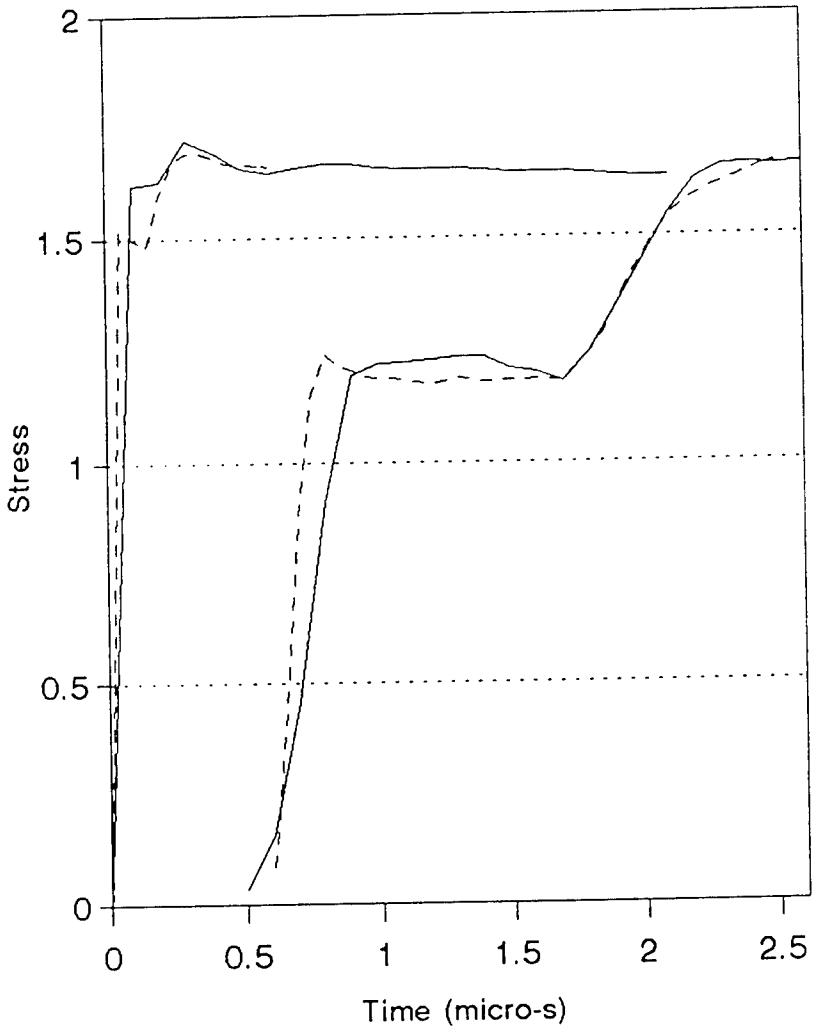
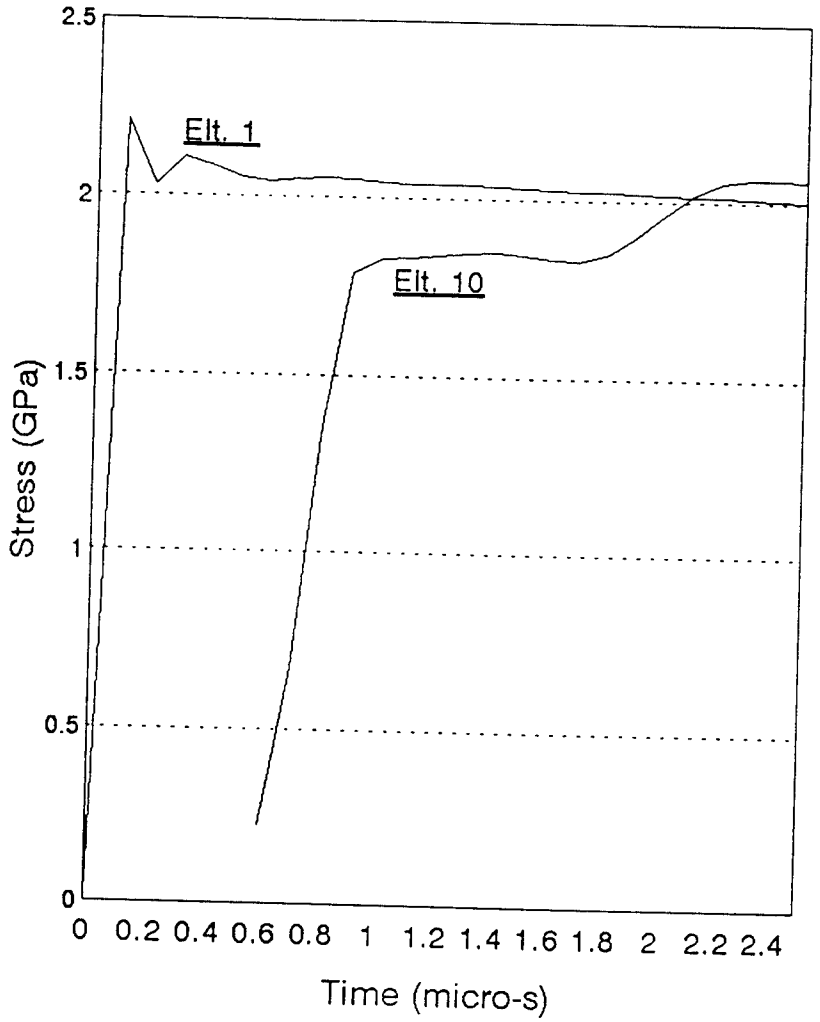


Figure 9: Impact on a 128-ply laminate a- ϵ_{cr} = 4 % , b- ϵ_{cr} = 8 % , c- ϵ_{cr} = 12 %

(b)



(c)


2.5
2
1.5
1
0.5
0
Stress (GPa)
Elt. 1
Elt. 10
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2 2.2 2.4
Time (micro-s)

INVESTIGATION OF FUEL NEUTRALIZATION AGENTS

William W. Bannister

Professor, Department of Chemistery

University of Massachusetts at Lowell

1 University Avenue

Lowell, Massachusetts 01854

Final Report for:

Summer Faculty Research Program

WL/FIVCF, Tyndall AFB, Florida

Sponsored by:

Air Force Office of Scientific Research

Bolling Air Force Base, Washington, DC

INVESTIGATION OF FUEL NEUTRALIZATION AGENTS

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

ABSTRACT

This study involved an investigation of fuel neutralization (FN), i.e., the rendering of spilling or spilled fuel non-burnable, extending the time to reignite, and/or facilitating the washing away of spilled fuel. FN agents which were studied were subjected to ignition and burn-back testing, chemical oxygen demand (COD) testing, and biological oxygen demand (BOD testing. The combination of these tests provided information about the FN capabilities as well as the environmental impact of the agents. This report reviews the experimental procedures used, the results obtained, and evaluations still pending, and gives conclusions based on work to date and recommendations for future work.

INVESTIGATION OF FUEL NEUTRALIZATION AGENTS

William W. Bannister

INTRODUCTION

WL/FIVCF has an ongoing program to evaluate commercially available firefighting agents to further improve airfield rescue, fire suppression, and hazardous material mitigation. Such agents must be environmentally acceptable and compatible with existing delivery systems. One area of concern is fuel neutralization (FN), the rendering of spilling or spilled fuel non-burnable, extending time to reignition (measured by "burn-back" time). FN approaches include the following technologies:

1. Reducing fuel vaporization by adding gelling agents, resulting in a significantly reduced rate of evaporation. Moreover, the gelled fuel can be disposed of much more readily.

2. Mixing extinguishing water into the non-burning fuel by use of self-mixing emulsifiers.

3. Application of absorbent polymers which entrain the fuel into the polymeric matrix.

A preliminary investigation was conducted by Beltran, Inc. in 1989, whose final report will be cited extensively in this discussion. Most of this work involved formulations of agents which could emulsify Jet Propulsion (JP) fuels in water. The most efficient agents (in terms of forming fine microemulsions) were also the best in self-mixing, but these also proved unsatisfactory in achieving non-combustible fuel-water blends. Several Beltran agents were good emulsifiers and also made the emulsified fuel non-combustible. These emulsions were coarser and were harder to mix. The most efficient of these were the sodium salt of dioctylsulfosuccinate (MO70; Alcolak, Ltd.; and DV1875; Mona Corp.). A very similar agent was

dioctylsulfosuccinate as the free acid (DOSS), which was very efficient in small burn tests when formulated with Film-Forming Fluoro-Protein (FFFP), as agent "FM". Small-scale lab tests resulted in two optimum formulations: FN-1 (fluorosurfactant "Fluorads" [3M] non-ionic FC100, 0.2%; anionic FC135, 0.1%; "Sticky Water" [SW], 0.5%; and AFFF, 3%); and FN-2 (1.25% DOSS substituted for the 3% AFFF in FN-1). In small-scale lab tests FN-1 and FN-2 successfully emulsified and inerting JP-4/water. In field tests, however, both FN-1 and FN-2 were less effective in burn-back than AFFF alone. This was attributed to the difficulty of mixing of the surfactants into the fuel. It was observed that water structuring and thickening agents such as SW significantly improved sealing of fuels by AFFF and FFFP, and it was suggested that addition of small amounts of SW into the make-up water streams of AFFF could improve burnback efficacy.

METHODOLOGY

A. INSTRUMENT AND TEST APPARATUS

Bunsen burner

Petri dishes, mixing bowls, stirring rods, 250 ml beakers

Stop watch, pH meter

Foam Maker Apparatus

Hach COD Reactor with COD vials

Hach 3000 Spectrometric colorimeter

YSI Model 58 Dissolved Oxygen meter for BOD tests

B. TEST METHODS

1. Ignition and Burn-Back Testing

The flame from the Bunsen burner was passed over various mixtures in the glass bowl at predetermined intervals, and to ignite surface fires in petri dishes for extinguishment tests.

2. Laboratory Extinguishment Apparatus

This apparatus consisted of a flow meter, a glass tube with a glass frit 0.5 inch from the bottom, and a petri dish.

The agent was injected by a syringe into the glass tube above the frit. Air was pumped at a controlled flow rate through the frit to aereate the agent while being delivered to a flame ignited on the surface of 20 mL of JP-4 in a 3.5-inch diameter petri dish.

3. Chemical Oxygen Demand (COD; Standard Methods, 17th ed.)

COD testing was performed by placing 2 mL of oxidizing solution in a COD reactor vial,heating this to 150°C, with 2 mL of test solution then added to the vial. Contents were incubated at 150°C for two hours. Results were compared colorimetrically with standard solutions to obtain COD test values.

4. Biological Oxygen Demand (BOD; Standard Methods, 17th ed.)

BOD testing was performed by placing the test sample, diluted in accordance with its predetermined COD value, in a BOD reactor vial. After addition of the BOD innoculating agent, the BOD reactor was incubated for five days at 20°C.

RESULTS AND DISCUSSION

A. AGENT COMPONENTS

Considerable test protocol development was accomplished since the Beltran project. The current work confirms the efficiency of many Beltran agents; see Table 1. Our work also indicated that self-mixing was attainable in spray applications by incorporating several new, commercially available proprietary formulations such as PETROL SAFE and Sticky Water (SW) into the FN agent formulations. These will be discussed later in this report.

B. OPTIMIZED WL/FIVCF AGENT SYSTEMS

Based on the measurements obtained in this phase of the research and development work on FN agent systems (summarized in Table 2) the following initial conclusions were reached:

1. The best FN agents appear to be those which contain the Fluorad agents FC-99 and FC-100. Compositions containing these agents showed most resistance to ignition.

Table 1. TESTED CHEMICALS FROM THE BELTRAN REPORT

NAME	CHEMICAL	CATEGORY
DOSS	Dioctyl sulfosuccinate	Anionic surfactant
MO70R	Sodium dioctylsulfosuccinate	Anionic surfactant
DV-1875	Sodium dioctylsulfosuccinate	Anionic surfactant
TX-45	Octylphenoxypolyethylene oxide	Nonionic surfactant
FC-99	Fluorocarbon surfactant	Anionic surfactant
FC-100	Fluorocarbon surfactant	Nonionic surfactant
FC-129	Fluorocarbon surfactant	Cationic surfactant
FC-135	Fluorocarbon surfactant	Anionic surfactant
PVP	Polyvinyl pyrolidine	Water absorbent gel
PG	Propylene glycol	Alcohol
PAM	Polyacrylamide	Water absorbent gel
AFFF	Extinguishment foam	Commercial product

2. Unfortunately, FC-99 and FC-100 had the greatest resistance to biodegradability. Thus, environmentally these appear to be unacceptable.

3. Composition #4 in Table 3, containing no fluorosurfactant components, provided th best FN characteristics of the remaining choices. "Haloing" effects (formation of a circular band of flickering but not sustained flame) would probably be acceptable, since in such cases no sustained ignitions were observed for more than an hour after any initial burn-back had been undertaken.

Table 2. BEST FORMULATIONS PREPARED BY WL/FIVCF

NAME	1	2	3	4	5	6
DOSS	1.0%			1.0%		
MO70R	1.0%	1.0%	1.0%	1.0%	0.5%	
DV-1875						
TX-45		0.5%	0.2%	0.5%	1.0%	
FC-99	1.0%	0.5%			1.0%	
FC-100	1.0%	0.5%	1.0%			
PVP		0.2%	0.5%	0.4%	1.0%	
RESULTS						
SELF MIXING						
TPS (min)						
TIME TO HALO (sec)						
TIME TO BURN (min)						
INVERTED MIXING						
TPS (min)						
TIME TO HALO (sec)						
TIME TO BURN (min)						
STIRRED MIXING						
TPS (min)						
TIME TO HALO (sec)						
TIME TO BURN (min)						
SHAKING MIXING						
TPS (min)	NT	NT	NT	60min+	NT	
TIME TO HALO (sec)	50min	60min	68min+	0	69min	
TIME TO BURN (min)	90min+	60min+	68min+	60min+	69min	

Key

TPS Time to Phase Separation
(+) Mixing very good; almost no fuel left on top
(O) Less than 5 mL left on top
(X) Phase separation immediate; 25 mL fuel left on top
NA Not Applicable
NT Not Tested

C. EVALUATION OF COMMERCIALLY AVAILABLE FN AGENTS

Since termination of the Beltran project, there have been a number of proprietary products submitted by private contractors for evaluation of FN agents:

* FYREZYME (Ecology Technologies Int'l, Inc., Mesa, AZ)

* BIOSOLVE (Southeast BioSolve, Inc., Jacksonville, FL)

* FIRE-TECH (Int'l Envioronmental Technologies, Inc., Washington, DC)

SAF (Stable Aqueous Firefighting Foam; Adherent Technologies, Inc. Albuquerque, NM)

** UNI-SNUFF (Inferno Snuffers, Inc., College Station, TX)

** SP-911 (Safety Products Group, Inc., Peachtree City, GA)

** FIREFREEZE

---

* Claims include bioremdiation; these claims not under investigation by WL/FIVCF.

** Recent submissions; as a result, these agents have not yet been fully investigated by WL/FIVCF.

A summary of FN tests on these proprietary formulations is presented in Table 3, showing how easily agents can create fuel/water emulsions to prevent ignitability and burn-back. Results of the tests performed on each of the proprietary formulations are provided in the following sections.

1. FIREZYME

FIREZYME has a COD of 856,000 mg/L, a high BOD at 78,000 mg/mL, and a percent biodegradability (BOD/COD times 100) of 9%. Being comprised of bioemulsifiers, enzymes, amino acids, and sugars it should be easily biodegraded. This is evidenced by the high BOD and BOD/COD percentage values. With a low pH of 3.9, FIREZYME is more acidic than any other agents tested. It can be anticipated that FIREZYME will be corrosive to metals over time. There is concern about possible relationships between the high BOD value and the low pH: in the environment, dilution should quickly elevate the pH to a point where, if biodegradability is a function of pH, the agent might not biodegrade.

Tests to determine FIREZYME's effectiveness as a fuel neutralizing agent are summarized in column 1 of Table 1. After shaking a sample of FIREZYME/fuel/water, fuel vapor odors were still present, indicating poor emulsification; separation

Table 3. FUEL NEUTRALIZATION MATRIX

NAME	1	2	3	4	5	6
JP-4	20 ml	20 ml	20 ml	20 ml	20 ml	20 ml
Water	20 ml	20 ml	20 ml	20 ml	20 ml	20 ml
Firezyme	20 ml					
Biosolve		20 ml				
FUEL BUSTER			20 ml			
FireFreeze				20 ml		
Micro Blazeout					20 ml	
Firezyme						20 ml
RESULTS						
SELF MIXING	X	X	X	X		
TPS (min)	NA	NA	NA	NA		
TIME TO HALO (sec)	NA	NA	NA	NA		
TIME TO BURN (min)	NA	NA	NA	0		
INVERTED MIXING	X	O	+	+		
TPS (min)	NT	NA	NA	NA		
TIME TO HALO (sec)	NT	NA	NA	NA		
TIME TO BURN (min)	NT	NA	O	2 secs		
STIRRED MIXING	+	+	+	+	+	+
TPS (min)	NA	NA	NA	NA	NA	NA
TIME TO HALO (sec)	NA	5 secs	0	NA	4 secs	0
TIME TO BURN (min)	NA	10min+	10min+	10min+	10min+	0
SHAKING MIXING				+		
TPS (min)				NA		
TIME TO HALO (sec)	NT			10 sec		
TIME TO BURN (min)	10 sec			10min+		

Key

TPS Time to Phase Separation
(+) Mixing very good; almost no fuel left on top
(O) Less than 5 mL left on top
(X) Phase separation immediate; 25 mL fuel left on top
NA Not Applicable
NT Not Tested

occurred immediately, and in burn-back testing reignition of the fuel occurred ten seconds after mixing stopped. In testing for extinguishing abilities, using a flow rate of 0.38 liters/minute,

6% FIREZYME extinguished the fire in 25 seconds; the FIREZYME sank to the bottom of the fuel in a matter of minutes. Table 4 shows the results of a number of trial tests using FIREZYME.

Table 4. FIREZYME

#	Flow Rate	Extinguishment Time
1	380 ml/min	25 sec
2	380 ml/min	19 sec
3	380 ml/min	20 sec
4	380 ml/min	25 sec
5	190 ml/min	None

2. BIOSOLVE

BIOSOLVE has a COD of 685,000 mg/L, a BOD of 219,000 mg/L, and a BOD/COD biodegradation ratio of 32% for a five-day BOD period; over a longer time degradation would be even more complete. The BOD/COD ratio indicates an acceleration in hydrocarbon biodegradation when the agent is distributed over a fuel spill, since BIOSOLVE emulsifies and encapsulates the hydrocarbon as an emulsion of tiny droplets. The pH was between 7.9 and 8.4, indicating a fairly neutral formulation with little concern for metal corrosion. With stirring, BIOSOLVE/JP-4/water emulsions did not separate until after several hours. However, enough fuel vapors were escaping from the emulsion to cause a fire halo when flame tested. After 24 hours the mixture had formed an even stronger emulsion which did not burn when ignited.

5. FUEL BUSTER

FUEL BUSTER has a COD of 345,000 mg/L. a BOD of 24,000 mg/L, and a BOD/COD ratio of 6%; AFFF's BOD/COD is only 2%. Thus. FUEL BUSTER should be more biodegradable than AFFF. The pH of FUEL BUSTER is 5.6, which is only slightly acidic. Therefore, corrosion should not be a big problem.

FN testing was performed in accordance the manufacturer's suggestions. Water, JP-4 and FUEL BUSTER were placed in a one square foot metal pan, and mixed using a high pressure water

stream. The fuel/water/FUEL BUSTER mixture mixed well and did not reignite when flame was applied in the burn-back testing.

FUEL BUSTER was also tested in standard FN matrix testing, with results presented in column 3 of Table 3, and in Table 5. FUEL BUSTER usually worked well when applied with vigorous agitation. The amount of water did not make much difference in the amount of time to prevent burn-back: no ignition occurred within 30 minutes. Using the recommended 1:2:4 FUEL BUSTER/JP-4/water formulation, no reignition occurred in the burn-back testing even 12 hours after mixing.

Foam expansion of FUEL BUSTER was measured using the laboratory extinguishing apparatus. In one minute, 75 mL of foam formed using 10 ml of FUEL BUSTER concentrate. The foam burned off as quickly as it was applied to the fuel fire. Although a good FN, FUEL BUSTER is not an effective extinguishing agent.

4. STABLE AQUEOUS FOAM (SAF)

SAF has a COD of 12,300 mg/mL, a BOD of 5,000 mg/mL in its 3% concentration form, and a BOD/COD of 41%. Comprised of a natural polymer and an alpha olefin sulfonate surfactant and with no fluorocarbons, it is therefore more biodegradable than AFFF.

SAF did not perform as an FN agent in its 3% or 6% concentrations, nor is it formulated as a FN agent. It required 31 seconds to extinguish a fire and had poor burn-back resistance.

5. FIREFREEZE

FIREFREEZE has a COD of 292,000 mg/L, a BOD of 12,000 mg/L, and a BOD/COD of 4%. (more biodegradable than AFFF). With a pH of 7.4, it should not be corrosive over extended time.

FIREFREEZE was an ineffective firefighting agent, burning up almost as quickly as it was sprayed onto the fire. As an FN agent it made a good emulsion that would not ignite in burn-back testing. The following day the sample had separated into three layers (water, emulsion and 2 ml of overlying JP-4).

Table 5. FUEL BUSTER

NAME	1	2	3	4	5	6
FUEL BUSTER	25 ml	25 ml	25 ml	50 ml	25 ml	
JP-4 FUEL	50 ml	50 ml	50 ml	50 ml	50 ml	
WATER	50 ml	25 ml	125 ml	50 ml	125 ml	
RESULTS						
SELF MIXING TPS (min) TIME TO HALO (sec) TIME TO BURN (min)	X NA NA 0			X NA NA 0		
STIRRED MIXING TPS (min) TIME TO HALO (sec) TIME TO BURN (min)	+ NA 0 0			+ NA 0 10min+		
SHAKING MIXING TPS (min) TIME TO HALO (sec) TIME TO BURN (min)	+ NA 0 0					
SPRAY MIXING TPS (min) TIME TO HALO (sec) TIME TO BURN (min)	+ NA 30min+ 30min+	+ NA 30min+ 30min+	+ NA 30min+ 30min+		+ NA 30min+ 12hrs+	

Key

TPS Time to Phase Separation
(+) Mixing very good; almost no fuel left on top
(0) Less than 5 mL left on top
(X) Phase separation immediate; 25 mL fuel left on top
NA Not Applicable
NT Not Tested

Table 6. STABLE AQUEOUS FOAM (SAF)

#	Flow Rate	Extinguishment Time
1	250 ml/min	31 seconds
2	190 ml/min	45 seconds

6. MICRO BLAZEOUT

MICRO BLAZEOUT is not claimed by the manufacturer as being a FN agent. In a test using 20 mL of agent, 40 mL JP-4 and 40 mL of water, MICRO BLAZEOUT proved very ineffective in stopping reignition during burnback tests. MICRO BLAZEOUT has a pH of 8.7, which could be alkaline enough to provide corrosion problems for aluminum or similar active metals. The COD is 535,000 mg/L, the BOD is 65,000 mg/L, and the BOD/COD ratio is 12% over a five-day BOD period. MICRO BLAZEOUT should be more biodegradable than AFFF.

7. UNI-SNUFF

UNI-SNUFF forms a good emulsion when 12 ml of this agent are mixed with 24 mL each of JP-4 and water. During the mixing process the agent absorbs the fuel and water, forming a very thick mixture. Even after 24 hours the sample would not reignite in burnback testing. As a spill neutralization agent the clean-up capabilities may be limited by virtue of the thickness of the emulsion. In its pure state the pH is 10.0. This could be alkaline enough to provide corrosion problems for aluminum or similar active metals. The COD value is quite high, 2,510,000 mg/L, the BOD is 750,000 mg/L, and the BOD/COD ratio is 30% biodegradability over a five-day period.

8. HAZCLEAN

HAZCLEAN is not effective as a FN agent. It did not mix well with JP-4, and had a highly obnoxious stench. The pH was 8.4. The COD is 458,000 mg/L, the BOD is 34,000 mg/L, and the BOD/COD ratio is 7% over a five-day BOD period.

9. PETROSEAL (FFFP)

PETROSEAL (FFFP) provides a good foam blanket on top of fuel when stir mixed. A pH of 7.5 indicates this agent is not anticipated to be corrosive to metals. When first stir mixed and subjected to ignition in burnback testing, there is some flame halo effect but not significant burn-back. After standing

overnight the sample separated into two layers, with free JP-4 floating on top. When flame was applied the sample readily ingited. The COD is 760,000 mg/L, the BOD is 61,500 mg/L, and the BOD/COD ratio is 8% over a five-day period. Further BOD tests will have to be performed over a longer incubation period.

10. AQUEOUS FILM FORMING FOAM (AFFF)

AFFF is a fire extinguishing agent only and is not not a FN agent. The pH of AFFF is 7.6 which is close to neutral. Thus, AFFF is not highly corrosive to metals. The COD is 845,000 mg/mL; the five-day BOD is 15,000 mg/L; and the BOD/COD ration is 2% over a five-day BOD period. This indicates a biodegradability problem. Various AFFF foam tests were done using the standard foam generator with 20 mL of JP-4 in a 3.5 inch diameter petri dish fire. With 3% AFFF, an air flow rate of 0.19 liters/minute was required to obtain a good extinguishment time; see Table 7.

Table 7. Time to Extinguish with Various AFFF Concentrations

#	Flow Rate	ML used	EXT Time
1	.19 lit/min	15 ml	25 sec
2	.19 lit/min	10 ml	24 sec
3	.19 lit/min	15 ml	21 sec
4	.19 lit/min	20 ml	22 sec
5	.19 lit/min	10 ml	14 sec
6	.19 lit/min	15 ml	17 sec
7	.19 lit/min	20 ml	24 sec

CONCLUSIONS

The results of testing of commercially available proprietary FN candidates are presented in Table 8 and Figure 3.

Table 8. Summary of Results

NAME	COD mg/l	BOD mg/l	Ratio	FN	EX	BB
FIREZYME	856,000	78,000	9%	Poor	Fair	Poor
BIOSOLVE	685,000	219,000	32%	Fair	Poor	V.Good
FUEL BUSTER	395,000	24,000	6%	Good	Poor	Poor
SAF	12,300	5,000	41%	Poor	Fair	Poor
FIREFREEZE	292,000	12,000	4%	Good	Poor	Excell
MICRO-BLAZEOUT	535,000	65,000	12%	Poor		Poor
UNI-SNUFF	2,510,000	750,000	30%	Good	Poor	Excell
EFFF/HAZCLEAN	458,000	34,000	7%	Poor		
PETROSEAL	760,000	61,500	8%	Poor		Poor
AFFF	845,000	15,000	2%	Poor	Excell	Fair

FN - Fuel neutralization testing
EX - Extinguishment testing
BB - Burn-back testing

The data provided in Table 8 and Figure 3 are summarized as follows:

1. No one agent has superior characteristics for all three categories of fuel neutralization, extinguishment and burnback resistance.

2. AFFF remains the agent of choice for extinguishment and burnback resistance characteristics.

3. UNI-SNUFF, FIREFREEZE and BIOSOLVE provided good FN and burn-back control. PETROSEAL was shown to be a fairly good FN agent, but was not as effective as the other three agents in burn-back control.

4. All of the agents require thorough mixing to achieve FN and burn-back control. Unrealistic requirements exist for stir mixing or for spray mixing for all the agents, and this has proven difficult to achieve.

5. UNI-SNUFF, BIOLSOLVE, and SAF demonstrated the best biodegradability, in five-day testing. FIREZYME and HAZCLEAN also showed good biodegradability, but lack good FN properties in comparison to BIOSOLVE, UNI-SNUFF, and FIREFREEZE.

6. FIREFREEZE was a good FN agent, but lacked good biodegradability in five-day BOD testing. Further tests will be conducted on this agent, as well as some of the other promising agents, to check biodegradability over longer periods of time.

RECOMMENDATIONS FOR FURTHER WORK

1. Continue lab evaluation of FN agents, with emphasis on investigation of ways and means of reducing volume requirements of FUEL BUSTER as a FN agent.

2. Investigate use of EXXON "CORREXIT" fluorosurfactant agents as sealing agents. In work done by EXXON it was shown that one gallon of this agent was capable of sealing eight million square feet of fuel surface within minutes of application, with 90% reduction of volatility and fire hazards.

3. Investigate the possible use of amine gelling agents for FN by gelation.

4. Investigate use of halon-impregnated silica agents for FN by gelation.

5. Investigate observations made in the Beltran report regarding utilization of water-structuring and thickening agents (e.g., "SW") for improvement of AFFF, FFFP and similar sealing agent formulations. In the case of AFFF, the sealing characteristics of the film over hydrocarbon fuel were greatly improved by addition of very small amounts of SW. It was suggested that the addition of small amounts of SW into the AFFF make-up water stream

could greatly improve the burn-back efficiency of AFFF.

6. Investigate relationships which may exist between the very high BOD values and very low pH of some of the agents studied in this project. When distributed in the environment it would be anticipated that dilution effects would quickly elevate the pH to a point where, if biodegradability is a function of pH, the agent may thereby suffer a loss of biodegradability.

7. Investigate a simple, easy to construct and easy to operate "Dynamic Volatility Apparatus" (essentially a small wind tunnel) which was designed at the University of Massachusetts/Lowell. The device holds two thermostated sample cups, one to contain a standard fuel sample and the other to hold fuel treated with a FN agent. Weighing the two samples before and after blowing air across the cup surfaces allows the efficiency of the agent's FN sealing effects to be accurately gauged.

REFERENCES

1. Report from Environmental Laboratory, Water Spigot, Inc. (5806 E. Hwy 22, Panama City, FL, 32401), Nov. 1992.

2. "Development of Fuel Neutralization Agents to Prevent Flashback on Aircraft Fires", Beltran Report, August 1987 - September 1989.

3. "Standard Guide for Room Fire Experiments", 1991 Annual Book of AST Standards.

4. Standard Methods Edition, "For the Examination of Water and Waste Water", 17th ed., 1989; BOD, Section 5.1 - 5.10; COD, 5.10 - 5.16.

ACKNOWLEDGMENTS

Research was supported by a Summer Faculty Research Program award from the Air Force Office of Scientific Research, and administered by the Research & Development Laboratories of Culver, California. I am also grateful for valuable assistance provided by my mentor, Mr. Douglas Nelson, and by my fellow laboratory workers, Mr. Barry, Mitchell, Mr. Richard Hunter, Ms. Lisa Lopez, and Mr. Richard Hunter; Mr. Richard Vickers, Chief of the Fire Protection Research Section of the Wright Laboratories, and Mr. Andrew Poulos, Head Librarian at the Air Force Civil Engineering Services Agency (both at Tyndall AFB);

The Preliminary Numerical Results and Mathematical Analysis for Least–Squares Finite Element Method for Incompressible Flow

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

Cleveland State University

September 1993

The Preliminary Numerical Results and Mathematical Analysis for Least-Squares Finite Element Methods for incompressible Flow

Ching Lung Chang
Associate Professor
Department of Mathematics
Cleveland State University

Abstract

Since the beginning of the 1970's, people have developed mixed finite element methods for incompressible flow. Engineerers and mathematicians have achieved great results in this field. The Galegin Method solves this elliptic boundary problem successfully. But the velocity and pressure interpolations are required to satisfy a LBB condition which precludes many natural elements. Over the past 20 years, most mathematicians and engineers believed this to be necessary. In this research, a least–squares method for these problems is proposed. This method leads to a minimization problem, the divergence free condition is no longer forced to be zero but is minimized with another equations. And thus it is not subject to the restriction of the inf-sup condition. Piecewise linear elements or piece quadratic element with equally order interpolation can be applied for both the approximation functions and the test functions. Thus simplest and natural elements are easy to program. By the analysis and numerical tests this kind methods achieve optimal rates of convergence in L-2 and in H-1 norms.

There are two parts in this report, analysis in least-squares finite element of stress–velocity–pressure version is in part one and the numerical experience for the vorticity-velocity-pressure version is in part two.

PART ONE

A Least–Squares Finite Element Method for Incompressible Flow in Stress-Velocity-Presure Version

Ching Lung Chang
Cleveland State University

1. Introduction.

Over the passed few years, a series of research papers have been published concerning the least-squares finite element methods for incompressible flow [6,7,14,15]. Such least-squares methods relax the divergence free restriction, div$\underline{u}$ is no long forced to be zero but is minimized with other equations. Therefore the regular finite element spaces using piecewise polynomial with equal order interpolation can be applied for both the test and trial functions. Another advantage of least-squares methods is that the matrix is positive definite and symmetric for a given regular finite element space. This allows the use of efficient schemes, such as the general SOR or conjugate gradient methods, to solve large systems.

During the past two decades many engineers and mathematicians have done research in the above problem. The mixed Galerkin method solves this problem successfully. In most cases the elements are required to satisfy a saddle point condition [2,13], which may not be necessary for the method which we will introduce in the next section.

The least–squares method relaxes the exact divergence free condition with a small non–vanishing div $\underline{u}$, therefore the elements may require less restriction. For example, all of the velocities $\underline{u}$, pressure p and stress $\phi's$ are allowed to be approximated by piecewise linear functions in $H^1(\Omega)$, we will show that the method achieves an optimal rate of convergence.

Weighted least–squares methods were used by Bramble, Schatz, Glowinski, Fix, Gunzburger, Nicolaides, Oden, Carey, Zienkiewicz and many others [4,11,17,20]. In this paper we are going to apply the theory of [Wend] type first order linear systems in the plane to the weighted least-squares methods. The work of Aziz, Kellogg and Wendland [1,19] gave a general theory for this method. Jiang, Povinelli, and Chang have successfully transfered the Stokes problem into a first order system in a two dimensional region and then treat it by least–squares method.

So far people have concentrated their attention in these methods to the velocity-vorticity-pressure formulation [3,9,14,15]. Numerical examples and mathematical analysis have achieved very good results in this field. The difficulty is how can we set a correct boundary condition satisfing the analysis and error estimates, which may lead the numerical approximation efficiently? Of course we can set $\underline{u} = 0$ on a solid body such as wall, cylinder or airfoil. But there is argument in the mathamatical analysis in spite of the numerical examples showed that the convergent rates seem optimal. In the work[3], they use $p = P$ on the boundary Γ, where p is the total head as $p = \hat{p} + \frac{1}{2}|\underline{u}|^2$ and $\hat{p}$ denotes the real pressure. For Stokes problem or Navier-Stokes equations we have $\underline{u} = 0$ on Γ. It may not be always easy to set p on the boundary.

This paper developes a stress-pressure-velocity version least-squares finite element method. Using the stresses as auxiliary variables, the Stokes problem can be written into a linear system with six equations and corresponding boundary conditions. The proof of convergence of this method is provided and the numerical experiences in a cavity driven support this analysis.

2. The Stress-Velocity-Pressure Version:

We introduce the auxiliary variables

$$\begin{cases} \phi_1 = \frac{\partial u_1}{\partial x} \\ \phi_2 = \frac{\partial u_1}{\partial y} \\ \phi_3 = \frac{\partial u_2}{\partial x} \\ \phi_4 = p. \end{cases} \tag{2.1}$$

By the incompressible condition, we have $\phi_1 = -\frac{\partial u_2}{\partial y}$. Then the dimensionless dynamic equations of the Stokes problem may be written in the form

$$\begin{cases} -\frac{\partial \phi_1}{\partial x} - \frac{\partial \phi_2}{\partial y} + \frac{\partial \phi_4}{\partial x} = f_1 & \text{in } \Omega \\ \frac{\partial \phi_1}{\partial y} - \frac{\partial \phi_3}{\partial x} + \frac{\partial \phi_4}{\partial y} = f_2 & \text{in } \Omega. \end{cases} \tag{2.2}$$

The compatibility condition gives

$$\begin{cases} \frac{\partial \phi_1}{\partial x} + \frac{\partial \phi_3}{\partial y} = 0 & \text{in } \Omega \\ \frac{\partial \phi_1}{\partial y} - \frac{\partial \phi_2}{\partial x} = 0 & \text{in } \Omega. \end{cases} \tag{2.3}$$

To recover the velocity, we have

$$\begin{cases} \text{div}\underline{u} = 0 & \text{in } \Omega \\ \text{curl}\underline{u} = \phi_3 - \phi_2 & \text{in } \Omega. \end{cases} \tag{2.4}$$

Where the Ω is a bounded and connected subset of $\Re^2$ with a piecewise smooth boundary Γ and $\underline{f} \in [L^2(\Omega)]^2$ is a given function of body force.

If we write the above equations together, we have

$$L\underline{U} = \begin{bmatrix} -\nu\frac{\partial\phi_1}{\partial x} - \nu\frac{\partial\phi_2}{\partial y} + \frac{\partial\phi_4}{\partial x} \\ \nu\frac{\partial\phi_1}{\partial y} - -\nu\frac{\partial\phi_3}{\partial x} + \frac{\partial\phi_4}{\partial y} \\ \frac{\partial\phi_1}{\partial x} + \frac{\partial\phi_3}{\partial y} \\ \frac{\partial\phi_1}{\partial y} - \frac{\partial\phi_2}{\partial x} \\ \mathrm{div}\underline{u} \\ \mathrm{curl}\underline{u} - \phi_3 + \phi_2 \end{bmatrix} \underline{U} = \begin{bmatrix} f_1 \\ f_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \text{in } \Omega \tag{2.5}$$

And the boundary condition $\underline{u} = 0$ on Γ, implies that the tangential derivatives of u_i are vanished or

$$\begin{cases} n_1\phi_2 - n_2\phi_1 = 0 & \text{on } \Gamma \\ n_1\phi_1 + n_2\phi_3 = 0 & \text{on } \Gamma. \end{cases} \tag{2.6}$$

The zero boundary condition of $\underline{u} = 0$ is equavelent to $n_1u_1 + n_2u_2 = 0$ as well. So we have

$$R\underline{U} = \begin{bmatrix} -n_2 & n_1 & 0 & 0 & 0 & 0 \\ n_1 & 0 & n_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & n_1 & n_2 \end{bmatrix} \underline{U} = \underline{0} \qquad \text{on } \Gamma. \tag{2.7}$$

where $\underline{U} = [\phi_1, \phi_2, \phi_3, p, \underline{u}]^t$. The above system should be weighted which is required by the analysis in the following sections.

3. Numerical Formulation of Stress-velocity-pressure Version

Throughout this paper, we will employ standard notations of the Sobolev spaces and their associated norms [10,18]. We let $H^m(\Omega)$ denote the Sobolev space of functions having square integrable derivatives of order up to m over Ω, as

$$H^m(\Omega) = \{v \in L^2(\Omega);\ \partial^\alpha v \in L^2(\Omega) \ \text{ for } \ |\alpha| \le m\}. \tag{3.1}$$

We define the norms by $\|u\|_m^2 = (u,u)_m$. and the inner product in $H^m(\Omega)$ is defined as

$$(u,v)_m = \sum_{|\alpha|\le m} \int_\Omega \partial^\alpha u \cdot \partial^\alpha v\ . \tag{3.2}$$

We also define the space for our problem,

$$S = \{\underline{V} \in [H^1(\Omega)]^6;\quad R\underline{V} = 0 \ \text{ on } \Gamma\}. \tag{3.3}$$

From the work of [4], we will use finite dimensional subspace $S_r^h \in S$ of functions to approximate our solutions. The parameter h,which represents a mesh spacing, is used to indicate the approximation property

of S_r^h. In this paper we say S_r^h approximates optimally with respect to r if for every $\underline{V} \in S \cap [H^{r+1}(\Omega)]^6$, there exists $\underline{V}^h \in S_r^h$ such that

$$h\|\underline{V} - \underline{V}^h\|_1 + \|\underline{V} - \underline{V}^h\|_0 \leq Ch^{r+1}\|\underline{V}\|_{r+1} \tag{3.4}$$

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

Then we can define the least–squares quadratic functional

$$J(\underline{V}) = \int_\Omega (L\underline{V} - \underline{F}) \cdot (L\underline{V} - \underline{F}) \qquad \text{for } \underline{V} \in S. \tag{3.5}$$

Consider the problem that if $\underline{U}$ minimizes $J(\underline{V})$ over $\underline{V} \in S$, it is easy to have

$$\int_\Omega L\underline{U} \cdot L\underline{V} = \int_\Omega \underline{F} \cdot L\underline{V} \qquad \text{for any } \underline{V} \in S \tag{3.6}$$

and that a solution of (2.5) and (2.7) is also a solution of (3.6) and that the sufficiently smooth solution of (3.6) also solves (2.5) and (2.7).

A finite element approximation to the solution of (2.5) and (2.7) or (3.6) is defined as a solution of the problem

$$\text{Min } J(\underline{V}^h) \qquad \text{over } \underline{V}^h \in S_r^h. \tag{3.7}$$

Similar to (3.6), the solution of $\underline{U}^h$ of (3.7) satisfies the corresponding finite algebraic equations

$$\int_\Omega L\underline{U}^h \cdot L\underline{V}^h = \int_\Omega \underline{F} \cdot L\underline{V}^h \qquad \text{for any } \underline{V}^h \in S_r^h. \tag{3.8}$$

Once a basis for S_r^h is chosen, evidently, we can see that (3.8) is equivalent to a symmetric linear algebraic system. Moreover, in the next sections we will show that this algebraic system is also positive definite.

4. The *a priori* Estimates

In this section we will apply the theory given by W.L.Wendland [19]. We try to show that L is an elliptic operator, and that the boundary operator R satisfies the complementing condition. If $\int_\Omega p = 0$, the boundary value problem gives rise to our desired inequality for the Stokes problem. Eq.(2.5) can be rewritten as a matrix form

$$L\underline{U} = A\underline{U}_x + B\underline{U}_y + C\underline{U} = \underline{F} \tag{4.1}$$

where A, B and C are 6×6 constant matrices. Furthermore, we multiply A^{-1} to each term in (4.1), and let $\tilde{B}$ and $\tilde{C}$ be $A^{-1}B$ and $A^{-1}C$. Then note the formulas:

$$\tilde{B} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1/\nu & 0 & 0 \\ 0 & -\nu & -\nu & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix},$$

and

$$\tilde{C} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}.$$

Following the procedure in [19], we check the polynomials

$$\det(\xi I + \eta\tilde{B}) = (\xi^2 + \eta^2)^2$$

$$\neq 0 \qquad \text{for real} \quad [\xi, \eta]^t \neq \underline{0}.$$

Therefore the operator L defined in (4.1) is an uniformly elliptic system. In this paper we discuss the problem with constant coefficients. So that the position variable (x, y) is dropped. We review and check the Lopatinski condition which is fulfilled for our problem.

Without loss of generality, we let $\nu = 1$ in following steps. After elementary operations, we can find the eigenvalues of matrix $\tilde{B}^T$ are i and $-i$, both having multiplicities 3.

Consider the eigenvalue in the upper plane $\lambda_+ = i$, to which belongs a chain of linearly independent generalized eigenvectors $\underline{p}_1$ and $\underline{p}_2$ of $\tilde{B}^T$ defined by

$$\begin{cases} \tilde{B}^T p_1 - \lambda_+ p_1 = 0 \\ \tilde{B}^T p_2 - \lambda_+ p_2 = p_1 \end{cases}$$

and a third $\underline{p}_3$ satisfing

$$\tilde{B}^T \underline{p}_3 - \lambda_+ \underline{p}_3 = 0,$$

where

$$\begin{cases} \underline{p}_1 = [0, 1, -1, -i, 0, 0]^T \\ \underline{p}_2 = [-2, 0, 2i, -1, 0, 0]^T \\ \underline{p}_3 = [0, 0, 0, 0, 1, -i]^T. \end{cases}$$

The ellipticity of (4.1) implies that the complex 6×6 matrix

$$(\underline{p}_1, \bar{\underline{p}}_1, \underline{p}_2, \bar{\underline{p}}_2, \underline{p}_3, \bar{\underline{p}}_3)^T$$

is nonsingular. The inverse matrix can be written as

$$Q = (\underline{q}_1, \bar{\underline{q}}_1, \underline{q}_2, \bar{\underline{q}}_2, \underline{q}_3, \bar{\underline{q}}_3)$$

or

$$Q = \begin{bmatrix} -\frac{i}{4} & \frac{i}{4} & -\frac{1}{4} & -\frac{1}{4} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & -\frac{i}{4} & \frac{i}{4} & 0 & 0 \\ 0 & 0 & -\frac{1}{4} & -\frac{1}{4} & 0 & 0 \\ \frac{i}{2} & -\frac{i}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{i}{2} & -\frac{i}{2} \end{bmatrix}.$$

And the vector $\underline{p}_j$ and $\underline{q}_k$ satisfies the equations

$$\begin{cases} \underline{p}_j^T \underline{q}_l = \delta_{jl} \\ \underline{p}_j^T \bar{\underline{q}}_l = 0 \end{cases}$$

for all $j, l = 1, \cdots .6$. The solvability thory for the boundary value problems depends on the following *Lopatinski* condition,

$$\begin{aligned} \det(2\underline{r}_j \underline{q}_k) &= 8 \det \begin{bmatrix} \frac{2n_1+n_2 i}{4} & \frac{n_2-n_1 i}{4} & 0 \\ \frac{-n_1 i}{4} & \frac{-n_1-n_2 i}{4} & 0 \\ 0 & 0 & \frac{n_1+n_2 i}{2} \end{bmatrix} \\ &= \frac{-1}{4}(n_1 + n_2 i)^3 \neq 0, \end{aligned}$$

since $(n_1, n_2) \neq (0, 0)$.

By the work of [19], we can now state

Theorem 1. *For $\ell \geq 0$ there is a constant $C > 0$ such that*

$$\|\underline{U}\|_{\ell+1} \leq C \left(\|L\underline{U}\|_\ell + \|R\underline{U}\|_{\ell+\frac{1}{2}} + \|\underline{U}\|_0 \right). \tag{4.2}$$

It can be shown that the boundary value problem associated with (3.1),(3,2) has a unique solution. Therefore the term $\|\underline{U}\|_0$ can be dropped from (4.2). Since we discuss our problem in the space S, we have the inequality,

$$\|\underline{U}\|_{\ell+1} \leq C\|L\underline{U}\|_\ell. \tag{4.3}$$

The inequality (4.3) is crucial for our least-squares error analysis. It is interesting to note that (4.3) contains, in particular, the usual shift inequality for the system (1.3). Let $[\underline{u}, p]$ solve (1.3) and define the variables in (2.1). it yields the usual *a priori* inequality,

$$\|\underline{u}\|_{\ell+1} + \sum_{i=1}^{4} \|\phi_i\|_{\ell+1} \leq C\|\underline{f}\|_\ell,$$

or

$$\|\underline{u}\|_{\ell+2} + \|p\|_{\ell+1} \leq C\|\underline{f}\|_\ell \tag{4.4}$$

for solutions of (1.3).

5. Error Estimates

In this section we will discuss the numerical scheme defined by (3.11). Denote the bilinear form as

$$a(\underline{U}, \underline{V}) = \int_\Omega L\underline{U} \cdot L\underline{V}. \tag{5.1}$$

Thus, (3.9) and (3.11) can be rewritten as: find $\underline{U} \in S$ (defined by (3.6)), such that

$$a(\underline{U}, \underline{V}) = \int_\Omega \underline{F} \cdot L\underline{V} \qquad \text{for any } \underline{V} \in S \tag{5.2}$$

and find $\underline{U}^h \in V_r^h$ (defined by 3.10)), such that

$$a(\underline{U}^h, \underline{V}^h) = \int_\Omega \underline{F} \cdot L\underline{V}^h \qquad \text{for any } \underline{V}^h \in S_r^h. \tag{5.3}$$

By inspection, a is symmetric and $a(\underline{U}, \underline{U}) \geq 0$. Furthermore, if $a(\underline{U}, \underline{U}) = 0$, from (3.7) we get $\underline{U} = 0$. Hence the matrix associated with the linear system (5.3) is positive definite.

Combining (5.2), (5.3), we have

$$a(\underline{U} - \underline{U}^h, \underline{V}^h) = 0 \qquad \text{for any } \underline{V}^h \in V_r^h. \tag{5.4}$$

Our error estimate is contained in the following theorem.

Theorem 2. *Suppose S_r^h approximates optimally with respect to r. Let $[\underline{u}, p]^t$ be the solution of (1.3), defining the auxiliary variables in (2.1) , and let $\underline{U} \in S$,$\underline{U}^h \in S_r^h$ be the solution of (3.6),(3.8) respectivly. Then*

$$\|\underline{U} - \underline{U}^h\|_1 \leq Ch^r \|\underline{U}\|_{r+1}. \tag{5.5}$$

[proof] Using (3.10) with $\ell = 1$, (5.5) and (5.1), we have for any $\underline{V}^h \in S_r^h$,

$$\|\underline{V}^h\|_1^2 \leq C \cdot a(\underline{V}^h, \underline{V}^h).$$

Applying this inequality to $\underline{U}^h - \underline{V}^h \in S_r^h$ and using (5.4),

$$\begin{aligned}
\|\underline{U}^h - \underline{V}^h\|_1^2 &\leq Ca(\underline{U}^h - \underline{V}^h, \underline{U}^h - \underline{V}^h) \\
&= C(a(\underline{U}^h - \underline{U}, \underline{U}^h - \underline{V}^h) + a(\underline{U} - \underline{V}^h, \underline{U}^h - \underline{V}^h)) \\
&= Ca(\underline{U} - \underline{V}^h, \underline{U}^h - \underline{V}^h) \\
&\leq C_1 \|\underline{U} - \underline{V}^h\|_1 \cdot \|\underline{U}^h - \underline{V}^h\|_1.
\end{aligned}$$

Hence $\|\underline{U}^h - \underline{V}^h\|_1 \leq C\|\underline{U} - \underline{V}^h\|_1$. Using the optimal approximation property of $\underline{S}_r^h$, choose $\underline{V}^h$ so that $\|\underline{U} - \underline{V}^h\|_1 \leq Ch^r \|\underline{U}\|_{r+1}$. Then $\|\underline{U}^h - \underline{V}^h\|_1 \leq Ch^r \|\underline{U}\|_{r+1}$, so

$$\begin{aligned}
\|\underline{U} - \underline{U}^h\|_1 &\leq \|\underline{U} - \underline{V}^h\|_1 + \|\underline{U}^h - \underline{V}^h\|_1 \\
&\leq Ch^r \|\underline{U}\|_{r+1},
\end{aligned} \tag{5.6}$$

which is the desired result. Above analysis concerning the error estimates is given by standard techniques. The velocity boundary conditions can be applied for. The experimence for solving the Navier-stokes equations shows us, it is not so easy to set the correct boundary conditions. And the least-squares of all equations in all of the part of our region may neglect some inportant information. Therefore the technique of computation is also a very difficulty one.

The analysis of least-squares for non-linear Navier-Stokes equations is still an open problem, even the preliminary numerical experience is satisfied.

6. Numerical Experiences.

We take for our domain the unit square $\Omega = (0,1) \times (0,1)$. The example we present has the smooth exact solution which is very similar to the one given by [17]. Let $\psi = x^2y^2(1-x)^2(1-y)^2$ then u, v and ϕ_i will be the derivatives or double derivatives of ψ, as

$$\begin{cases} u = x^2(1-x)^2 2y(1-3y+2y^2) \\ v = -2x(1-3x+2x^2)y^2(1-y)^2 \\ \phi_1 = 4xy(1-3x+2x^2)(1-3y+2y^2) \\ \phi_2 = 2x^2(1-x)^2(1-6y+6y^2) \\ \phi_3 = -2(1-6x+6x^2)y^2(1-y)^2 \\ \phi_4 = p = x^2+y^2. \end{cases} \tag{6.1}$$

Let $\nu = 1$, then from (2.5) we can easily have,

$$\begin{cases} f_1 = -4(1-6x+6x^2)y(1-3y+2y^2) + 12x^2(1-x)^2(1-2y) + 2x \\ f_2 = 4x(1-3x+2x^2)(1-6y+6y^2) - 12(1-2x)y^2(1-y)^2. \end{cases} \tag{6.2}$$

Substitute (6.2) into (2.5) then solve (5.3). We compare the numerical solution and the exact solution. The error with repect to $u, v, \phi_1, \phi_2, \phi_3, p$ are listed in the following tables.

Table 1 and Table 2 exhibits the numerical results of the errors in $L-2$ and $H-1$ norms for the piecewise quadratic elements in the bilinear quadrilaterals, respectively. It indicates the size of the error $e_i = \phi_i - \phi_i^h$ for $i = 1,2,3$,$e_u = u - u^h, e_v = v - v^h$ and $e_p = p - p^h$ with respect to both the H^1 and L^2 norms.

Table 1.

h^{-1}	$h^{-3} \cdot \|e_u\|_0$	$h^{-3} \cdot \|e_v\|_0$	$h^{-3} \cdot \|e_p\|_0$	$h^{-3} \cdot \|e_1\|_0$	$h^{-3} \cdot \|e_2\|_0$	$h^{-3} \cdot \|e_3\|_0$
10	7.964e-3	7.964e-3	1.574e-2	3.090e-2	3.581e-2	3.581e-2
20	7.803e-3	7.803e-3	5.964e-3	2.762e-2	3.563e-2	3.563e-2
30	7.787e-3	7.787e-3	3.586e-3	2.720e-2	3.563e-2	3.563e-2

Table 2.

h^{-1}	$h^{-2} \cdot \|e_u\|_1$	$h^{-2} \cdot \|e_v\|_1$	$h^{-2} \cdot \|e_p\|_1$	$h^{-2} \cdot \|e_1\|_1$	$h^{-2} \cdot \|e_2\|_1$	$h^{-2} \cdot \|e_3\|_1$
10	2.875e-2	2.875e-2	7.085e-2	1.201E-1	1.342e-1	1.342e-1
20	2.747e-2	2.747e-2	1.502E-2	1.189e-1	1.337e-1	1.337e-1
30	2.749e-2	2.749e-2	8.886e-3	1.185e-1	1.338e-1	1.338e-1

The results in Table 2 support the error analysis given in (5.6). The results in Table 1 are errors in L-2 norms. Although we have not given the L-2 error analsis yet, they seem to achieve the optimal convergent rates.

PART TWO

A Preliminary Report for Numerical Results for Incompressible Flow by Finite Element Method in Velocity-Vorticity -Presure Version

Ching Lung Chang
Cleveland State University

1. Introduction.

Recently there has been substantial interest in least-squares finite element methods for velosity-vorticity-pressure formulations of the incompressible Navier-Stokes equations. Such least-squares methods relax the divergence free restriction, div$\underline{u}$ is no long forced to be zero but is minimized with other equations. Therefore we can apply the piecewise linear continuous functions to be our trial and test functions. The advantage of least-squares methods is that the matrix is positive definite and symmetric. This allows the use of preconditioned conjugate gradient methods, to solve large systems, using upper-storage-by-rows

Let Ω be a bounded and connected subset of $\Re^2$ with a piecewise smooth 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:

$$\begin{cases} -\nu\Delta\underline{u} + \underline{u}\cdot\text{grad}\underline{u} + \text{ grad } p = \underline{f} & \text{in } \Omega \\ \text{div } \underline{u} = 0 & \text{in } \Omega \\ \underline{u} = 0 & \text{on } \Gamma \end{cases} \tag{1.1}$$

where $\underline{u}$, p with $(p, 1)=0$, ν are velocity, pressure and kinematic viscosity (constant), ν is the inverse of the Reynolds number Re, all of which are assumed to be nondimensionalized. The velocity-vorticity-pressure version has the following form if we introduce vorticity $\omega = \text{curl } \underline{u}$,

$$\begin{cases} -\nu\text{curl}\underline{\omega} + \underline{u}\cdot\text{grad}\underline{u} + \text{ grad } p = \underline{f} & \text{in } \Omega \\ \text{curl}\underline{u} - \omega = 0 & \text{in } \Omega \\ \text{div } \underline{u} = 0 & \text{in } \Omega \\ \underline{u} = 0 & \text{on } \Gamma \end{cases} \tag{1.2}$$

At first, we pay our to the limit of above equation, as Re $\rightarrow$ 0, the Stokes problem. Eliminating the non-linear terms from above system,

$$\begin{cases} -\nu\text{curl}\underline{\omega} + \text{ grad } p = \underline{f} & \text{in } \Omega \\ \text{curl}\underline{u} - \omega = 0 & \text{in } \Omega \\ \text{div } \underline{u} = 0 & \text{in } \Omega \\ \underline{u} = 0 & \text{on } \Gamma \end{cases} \tag{1.3}$$

This work was performed with First Lieutenent John Nelson in WP/FI AFB,1993

The linear system is elliptic in any sence. The difficulty is how can we set a correct boundary condition satisfing the analysis and error estimates, which may lead the numerical approximation efficiently.

2. Numerical Formulation

Throughout this paper, we will employ standard notations of the Sobolev spaces and their associated norms [10,18]. the Sobolev space of functions having square integrable derivatives of order up to m over Ω, as

$$H^m(\Omega) = \{v \in L^2(\Omega);\ \partial^\alpha v \in L^2(\Omega) \ \text{ for } \ |\alpha| \le m\}. \tag{2.1}$$

We rewrite the equations (1.3) into a matrix form as

$$L\underline{U} = A\underline{U}_x + B\underline{U}_y + C\underline{U} = \underline{F} \tag{2.2}$$

where A, B and C are 4×4 constant matrices.

$$A = \begin{bmatrix} \tilde{u}_1 & 0 & 0 & 1 \\ 0 & \tilde{u}_1 & -\nu & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} \tilde{u}_2 & 0 & \nu & 0 \\ 0 & \tilde{u}_2 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

and

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The $\tilde{u}_1$ and $\tilde{u}_2$ in matrices A and B represent the previous u_1 and u_2 when we perform the non-linear iteration; they are zeros when we solve the Stokes problem. The zero boundary condition can be presented as

$$R\underline{U} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \underline{U} = \underline{0} \qquad \text{on } \Gamma. \tag{2.3}$$

where $\underline{U} = [\underline{u}, \omega, p]^t$. We will employ the definitions of the spaces S and S_r^h in (3.3) and (3.8) in Part one, then we can define the least–squares quadratic functional

$$J(\underline{V}) = \int_\Omega (L\underline{V} - \underline{F}) \cdot (L\underline{V} - \underline{F}) \qquad \text{for } \ \underline{V} \in S. \tag{2.4}$$

Consider the problem that if $\underline{U}$ minimizes $J(\underline{V})$ over $\underline{V} \in S$, it is easy to have

$$\int_\Omega L\underline{U} \cdot L\underline{V} = \int_\Omega \underline{F} \cdot L\underline{V} \qquad \text{for any } \underline{V} \in S. \tag{2.5}$$

Similar to (2.5), the solution of $\underline{U}^h$ in finite element space satisfies the corresponding finite algebraic equations

$$\int_\Omega L\underline{U}^h \cdot L\underline{V}^h = \int_\Omega \underline{F} \cdot L\underline{V}^h \qquad \text{for any } \underline{V}^h \in S_\tau^h. \tag{2.6}$$

Once a basis for S_τ^h is chosen, (2.6) is equivalent to a symmetric linear algebraic system. It is also positive definite.

3. Numerical Experience

I felt it was important to test the applicability of the least-squares method by attempting to numerically simulate incompressible flows whose flow characteristics are well known or have an analytical solution. To begin with I chose two of the simplest model incompressible flows: plane Couette flow and plane Poiseuille flow. Both flows concern the flow of a single fluid contained between two flat-infinite plates, as shown in figure 1.

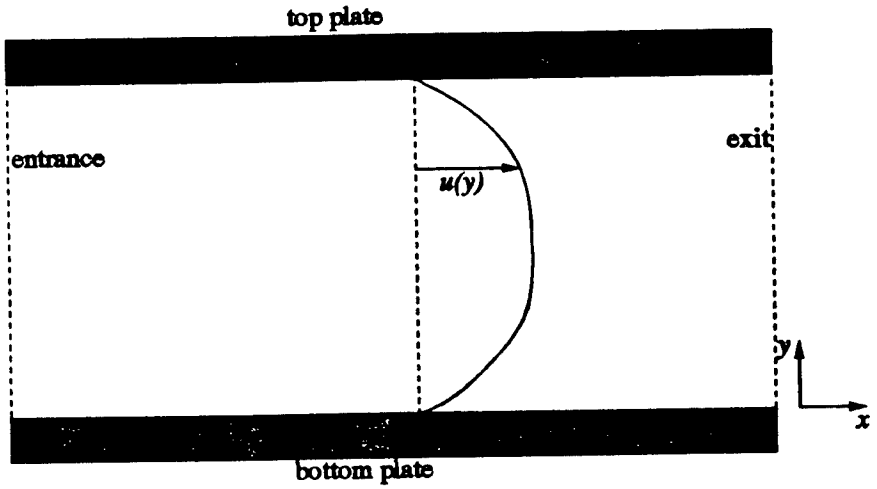


Figure 1. Setup of channel flow.

In Couette flow, the pressure gradient $\frac{\partial p}{\partial x} = 0$. The top plate is moves in the positive x direction at a constant speed and the bottom plate is fixed motionless. Distances are normalized by the distance between the plates, and velocities are normalized by the speed of the top plate. After this normalization, the plates

are located in the plane $y = 0$ and $y = 1$. This flow then has an analytical solution for all Reynolds numbers given by

$$u(x, y) = y, v(x, y) = 0, p(x, y) = \text{constant}.$$

In Poiseuille flow the top and bottom plates are held fixed while a constant pressure gradient is applied. In the simulations, the pressure was normalized so that it took the value 1 at the entrance and 0 at the exit. In the simulations, the grid was made to be 5 units long. This flow then has an analytical solution for all Reynolds numbers given by

$$u(x, y) = 0.025y(1 - y), v(x, y) = 0, p(x, y) = -G(x - x_o) + p(x_o),$$

where G is a constant and represents the applied pressure gradient.

In order to numerically simulate these two flows, we created a unstructured, triangular element mesh using the Flight Dynamics Interdisciplinary and Applied CFD section's unstructured grid generation and post-processing package TOPDUUG. A simulation of both Couette and Poiseuille flow was completed on the grid shown in figure 2.

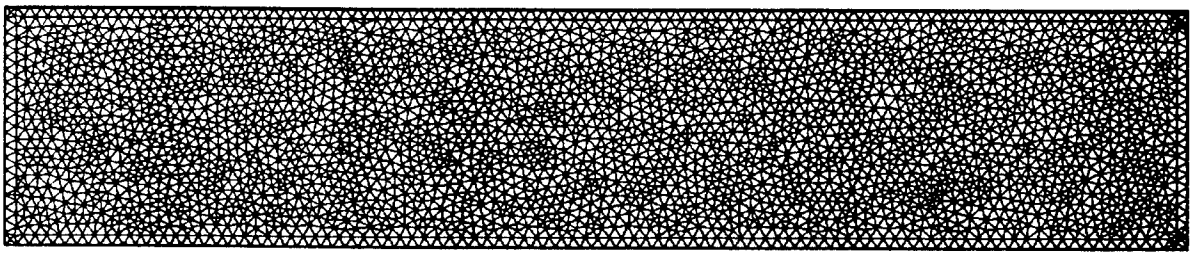


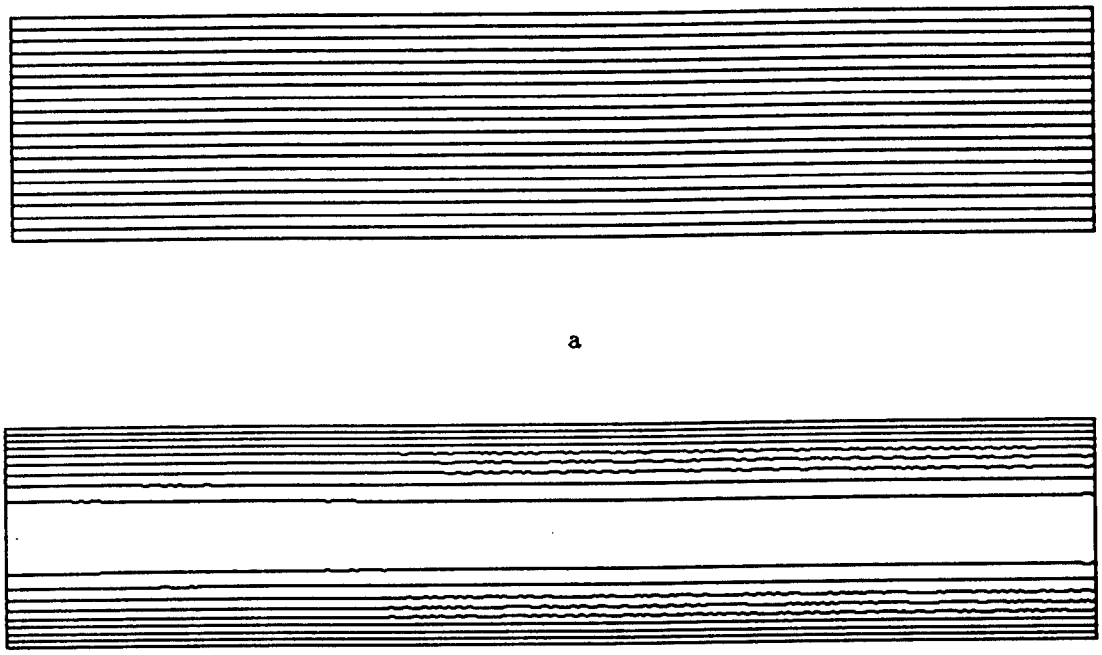
Figure 2. Grid used to simulate plane Couette and plane Poiseuille flows.

In order to complete a simulation of each flow, two boundary conditions were required on each boundary for each flow. The boundary conditions applied for each flow are shown in table 1.

flow	boundary	boundary condition
Couette	bottom plate	$u = 0, v = 0$
	exit (right side of grid)	$v = 0, p = 0$
	top plate	$u = 1, v = 0$
	entrance (left side of grid)	$v = 0, p = 0$
Poiseuille	bottom plate	$u = 0, v = 0$
	exit (right side of grid)	$v = 0, p = 0$
	top plate	$u = 0, v = 0$
	entrance (left side of grid)	$v = 0, p = 1$

Table 1. Boundary conditions used for similations of Couette and Poiseuille flows

The least squares method was successful in simulating both flows. The flow was found to be independent of x, as shown in figure 3.

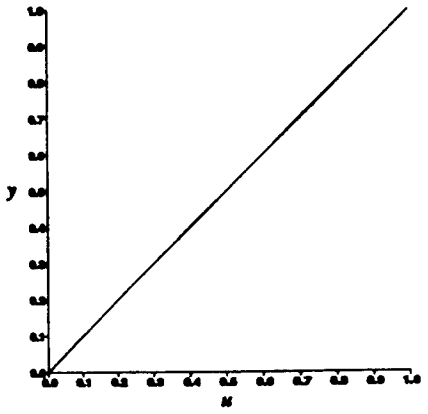


a

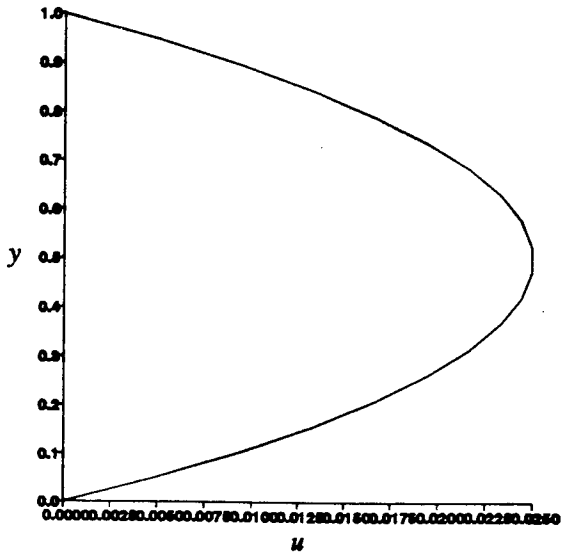
b

Figure 3. Level contours of the horizontal velocity component u for (a) Couette flow, and (b) Poiseuille flow.

The u versus y profiles calculated by the simulations are shown in figure 4.



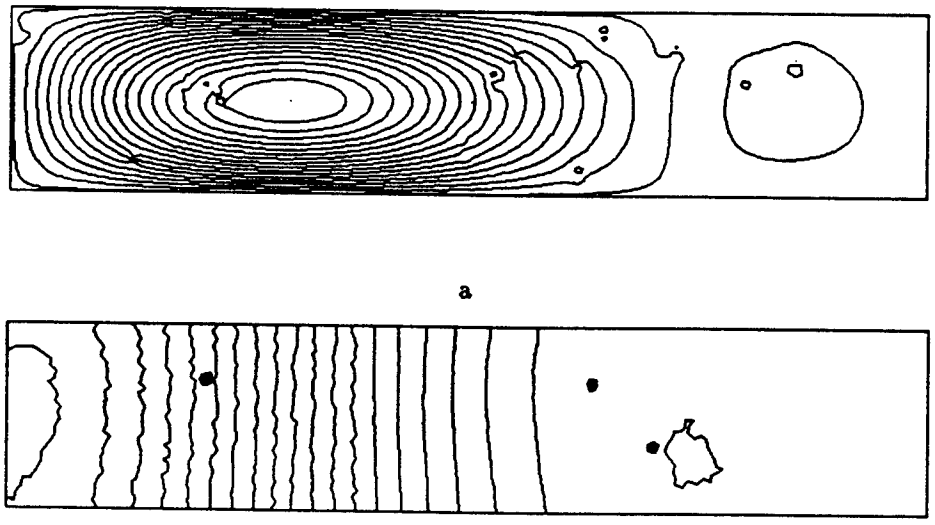
a



b

Figure 4. Plots of u versus y for calculated by the simulation for (a) Couette flow, (b) Poiseuille flow.

For Couette flow, the analytical solution requires v and p to be 0, and ω to be 1. In our simulation p was found to be zero everywhere in the domain to within machine accuracy. The maximum magnitude of v was 0.0005, and ω never differed from 1 by more than 0.001. Plots of v and ω are shown in figure 4.



a

b

Figure 4. Level contours for Couette flow of (a) the vertical velocity component v, and (b) the vorticity ω.

For Poiseuille flow, the analytical solution requires that v be 0, p decrease from 1 at the entrance to 0 at the exit in a linear fashion, and w to be linear in y with $\omega = -1$ at the bottom plate and $\omega = 1$ at the

top plate. Below are plots of level contours for v,ω, and p from the simulation of Poiseuille flow.

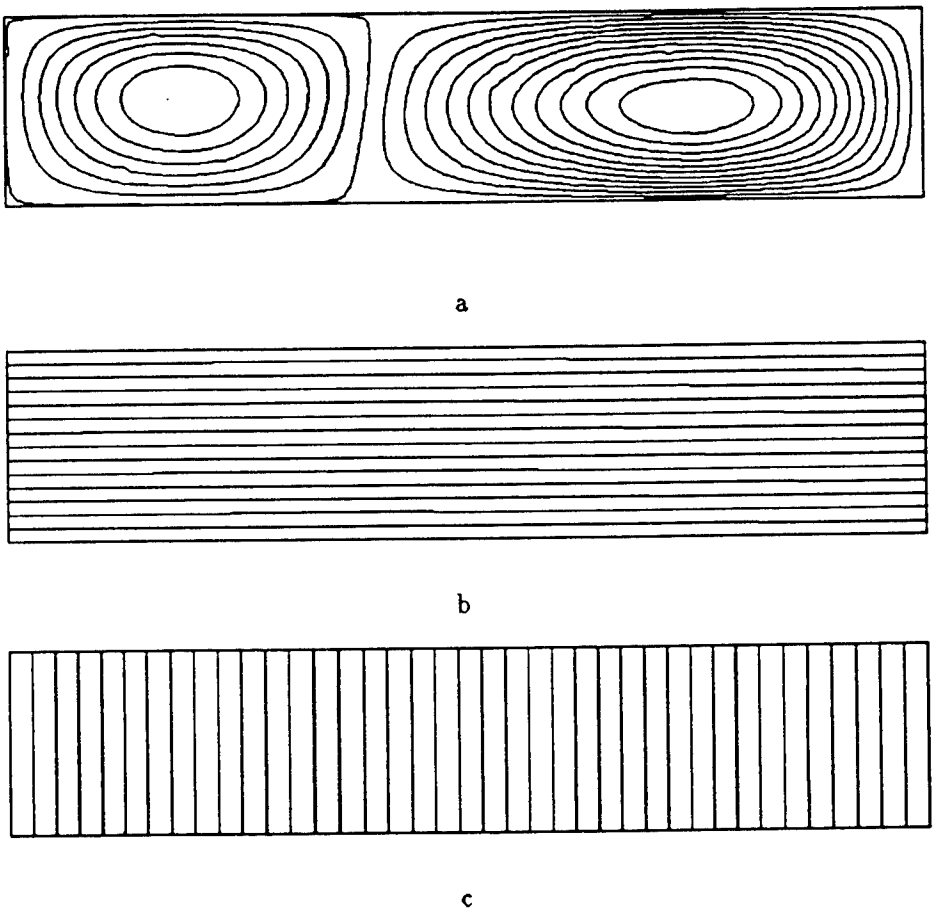


Figure 5. Level contours for Poiseuille flow of (a) the vertical velocity component v, (b) the vorticity ω, and (c) the pressure p.

References

[1] 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.

[2] I. Babuška, J. T. Oden and K. Lee, "Mixed–hybrid finite element approximations of second–order boundary value problems", Comput. Methods Appl. Mech. Engrg., Vol. 11, 175–206, 1977.

[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, June 1993.

[4] J. H. Bramble and R. Scott, "Simultaneous approximation in scales of Banach spaces", Math. of Computation, Vol. 32, 947–954, 1978.

[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", Mathematics of Computation, to appear, 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] L. Franca, T. Hughes and R. Stenberg, "Stabilized finite elements," Incompressible Computational Fluid Dynamics: Trends and Advances, (Ed. by M.Gunzburger and R.Nicolaides), Cambridge, to appear.

[13] V. Girault and P. A. Raviart, "Finite element methods for Navier–Stokes equations" Springer–Verlag, Berlin, 1986.

[14] B. N. Jiang and C. L.Chang, "Least–squares finite elements for Stokes problem", Comput. Methods Appl. Mech. Engrg, Vol. 78, 297–311, 1990.

[15] B. N. Jiang and L. A. Povinelli, "Least–squares finite element method for fluid dynamics", Comput. Meth. Appl. Mech. Engrg., Vol. 81, 1990.

[16] J. Oden and J. Reddy, "An introduction to the mathematical theory of finite element", Wiley, New York, 1976.

[17] J. T. Oden and G. F. Carey, "Finite Elements", Prentice–Hall,Inc., Englewood Cliffs, NJ, 1984.

[18] R. Temam, "Navier–Stokes equations and nonlinear functional analysis", SIAM, Philadelphia, 1983.

[19] W. L. Wendland, "Elliptic system in the plane", Prentice–Hall,Inc., Englewood Cliffs, NJ, 1984. London, 1979.

[20] O. C. Zienkiewicz, "The finite element method", Vol. 1, 4th ed., McGraw–Hill, 1989.

Anti-Penetration Laboratory Data Acquisition and Control Systems

Bobby L. Green

Assistant Professor

Department of Engineering Technology

Texas Tech University

Box 43107

Lubbock, TX 79409-3107

Final Report for:

Summer Faculty Research Program

Wright Laboratory/Air Base Survivability Section

Tyndall AFB, FL

Sponsored by:

Air Force Office of Scientific Research

Bolling Air Force Base, Washington, D. C.

Summer 1993

Anti-Penetration Laboratory Data Acquisition and Control Systems

Bobby L. Green
Assistant Professor
Department of Engineering Technology
Texas Tech University
Lubbock, TX

Abstract

Over the past several years there have been a growing number of manufactures developing PC data acquisition and control systems. The Anti-Penetration Laboratory at Tyndall AFB, Florida acquired several examples of the PC data acquisition and control systems for experimental use and evaluation. It was discovered that ease of operation of the various data acquisition systems varied, some were straight forward, and some take a great deal of familiarization to operate effectively. However in all the cases the users of any of the several data acquisition systems must extensively familiarize themselves with each particular data acquisition system to be used then frequently refresh their memories to remain familiar with the data acquisition system. Any break in the use of a particular data acquisition system usually requires retraining. The retraining is time consuming, but lack of continued familiarity with the system will lead to data error or lost data. The PC data acquisition systems are designed to fill several niches. The faster PC data acquisition systems currently operate around 40 MHz, will measure voltages or currents and store data for later evaluation, maximum data storage before a break is usually limited to the size of the RAM on the data acquisition board. Available RAM on a data acqusition system is up to 8-megabytes and growing.

Anti-Penetration Laboratory Data Acquisition and Control Systems

Bobby L. Green
Assistant Professor
Department of Engineering Technology
Texas Tech University
Lubbock, TX

Introduction

The survivability section purchased a single IBM compatible PC computer, an expansion chassis, several different types of PC data acquisition systems and some peripheral equipment. Some of the data acquisition systems were purchased with control systems to actuate output devices with the data acquisition systems. Most of the data acquisition system are expandable, that is, the software used to control the data acquisition system boards will control several data acquisition system boards of the same family allowing multiple channel data acquisition.

Following is a list of several of the data acquisition systems and controllers available to the anti-penetration laboratory. The list is representative of most of the PC data acquisition systems and control system types available in today's market. All of these data acquisition and control systems were purchased from CyberResearch Inc., Bradford, CT [1] and have been listed with the CyberResearch part number first, the manufacturers part name or number in parentheses, and the manufacturers name and phone number.

a) Inst-140 (CompuScope 220), Gage Applied Science, Montreal, Canada

b) CYDAS-16F (CIO-AD16) and CYCTM-05 (CIO-CTR), ComputerBoards Inc., Mansfield, MA phone 508-261-1123

c) ACPC-1616 and ACAO-128, Strawberry Tree Inc., Sunnyvale, CA phone 408-763-8800

d) Inst-292 (Programmable Stimulator and Control System) and INST-601 (Computerscope), RC Electronics, Goleta, CA phone 805-685-7770

e) DIO-32F (32-bit Digital I/O Board), National Instruments, Austin, TX phone 512-794-0100

f) COHM-16 (Serial Expansion Ports), Small Port Expansion Boards, Nashville, TN phone 615-834-8000

g) DAP-2405 (Data Acquisition System), Microstar Laboratories, Redmond, WA phone 206-881-4286

All the various data acqusition and control systems were installed in a single PC expansion chassis. The bus hardware addresses were changed in order to address all the data acqusition systems pieces in the same expansion chassis. For field use the different pieces of equipment should be placed in separate PC chassis and their hexadecimal base addresses should be set to their original factory specifications or to a non-conflicting bus address in their respective PC's. The PC bus addresses are well standardized for peripheral equipment so when placed into individual PC's the cards should experience no address conflicts when using the factory settings. If an address conflict is experienced it is usually a straight forward matter to change the hardware address and solve the problem.

installation and set up of hardware and software for various data acquisition systems

The PC Data acquisition systems are usually laid out on a single PC card or a half card module that will plug directly into an expansion slot in a PC chassis. In our case, with an expansion, chassis we were able to install all the systems into a separate expansion chassis and operate them form a single PC. Operating all the data acquisition systems from a single PC is an excellent way to familiarize one with several different types of systems in a short period of time, but it does not provide a useful platform for a single data acquisition system.

Each of the data acquisition systems are usually supplied with software written for the piece purchased. Some were supplied with software written specifically for the piece, additional freeware, and instructions to be patient while learning to operate the software and equipment. Each piece of software is usually system specific and will operate only its single family of equipment. To operate a conglomeration of data acquisition systems simultaneously it is necessary to write new software to control the conglomeration. All the data acquisition systems can then be operated from a single piece of software, but they will still operate serially. Operating a series of different data acquisition systems from a single PC slows down all the individual data acquisition systems, due to excess computational overhead, and is not an efficient use of the individual data acqusition systems.

There are several disadvantages in installing more than one type of data acquisition systems in a single PC, among the disadvantages are bus addressing and software compatibility. The software supplied with a system is nearly always system specific for only one piece of equipment so only one manufactures family of equipment may be

operated by the software at a time. Similar pieces of data acquisition equipment performing the same tasks are usually placed into the same hardware address. If more than one piece of hardware is set to the same hardware address the software will address incompatible pieces of equipment, the conflict usually causes an operating system failure. Manufactures provide for the addressing problems resulting in mixing manufacturers equipment, but the simple solution is choose one hardware manufacturer for any single task for assured hardware and software compatibility.

general operations of PC based data acqusition systems

The two computerscopes, Gage Applied Sciences' CompuScope 220 [2], R.C. Electronics' Computerscope [3], and R.C. Electronics' RC-200 Programmable Stimulator and Control System [4] operate in very similar manners. They are two PC based, menu driven, storage, oscilloscopes with on board RAM and a programmable signal generator with on board RAM. The memory for the computerscopes is a First In Last Out (FILO) type memory, as new data in input into the RAM memory on the computer card the oldest data is pushed out of memory and lost. Storing the data in FILO format means the computer scopes can remember past history if properly triggered. Manufactures usually call past history a pre trigger mode because a current trigger retains data previous to the trigger event and stores new data until any remaining memory is exhausted. The data taken may be down loaded onto a more permanent storage media such as magnetic tape, floppy disks, hard disks, etc. During the down loading period the data acquisition systems cannot take data, if an important event occurs during the down load period it will be lost. For slow events having periods of hours, minutes, or seconds down loading data poses no data interruption problems, for faster events milliseconds, microseconds, or nanoseconds, the devices are reduced to one shot devices and will miss rapidly occurring

events if they occur during a data transfer cycle. To operate these devices one must have their instruction manual constantly at hand until their menu items are committed to memory. Then the manual must be kept at hand to refresh ones memory.

The CompuScope 220 is a two channel, 40 megasample-per-second device so the CompuScope will use up to 40 megabytes of memory per second. The Computerscope is a 1 megasample-per-second 16 channel device, so the Computerscope uses up to 1 megabyte of memory per second. It is obvious the on board RAM for both these devices will be filled very quickly when using their maximum sampling rates. The stored data may be downloaded and evaluated with software supplied by the manufacturer or the data may be evaluated with a third parties software.

The Strawberry Tree Inc., ACPC-1616 [5] and ACAO-128 [6] can be menu driven with the software included with the devices or they can be driven with Icon based software Strawberry Tree's QuickLog PC software [7]. The ACPC-1616 will take data at up to 10,000 samples-per-second using a variety of transducers, the software is capable of controlling multiple cards with up to 240 analog inputs. The some of the transducers available for the ACPC include temperature compensated thermocouples, RTD's, strain gages, pressure transducers, flow meters, switches, photocells, DC voltage or DC current input, etc. At 10,000 samples-per-second it is clear the ACPC is designed for data acquisition of slowly varying signals and the controllers control processes with long time constants.

The menu driven software is as tedious to learn as most menu driven software, however the QuickLog PC icon driven software makes the control system more user friendly and gives a schematic type environment for developing data acquisition and control schemes.

ComputerBoards' CIO-AD16 [8] and CIO-CTR [9], both assume moderate to extensive computer programming skills and moderate to extensive skills in reading computer timing diagrams. The manuals supplied with the acquisition card and the control card indicate that the cards are very versatile if the manuals are intensely studied, the supplied instructions are carefully followed and the programming is correctly written.

National Instruments' DIO-32F [10], assumes moderate to extensive computer programming skills and moderate to extensive skills in reading computer timing diagrams. The DIO-32F is a 32-bit, Parallel, Digital, I/O (Input/Output) interface for communicating with other digital parallel interface equipment. It is not a data acquisition or control system but it is a digital I/O communications link between a microprocessor and peripheral digital equipment.

Microstar Laboratories' DAP-2405 [11], is supplied with the DAP-2504 card, DAP software, a 120 page "Hardware manual", a 140 page "Applications Manual", a 200 page "Systems Manual", a 270 page "DAPL Manual" and a one of an array of terminal panels for analog or digital I/O data. The DAP-2504 assumes moderate to extensive computer programming skills, moderate to extensive skills in reading computer timing diagrams and requires an intense period of study to turn on and operate properly. The DAP-2504 is a very power piece of equipment requiring a certain amount of time for start up. New windows software makes the device more user friendly.

Small Port Expansion Boards' COHM-16 [12], is a straight forward serial expansion kit with software. The Smart Port allows expansion of a single serial port into an array of 4, 8, 16, or 32 separate serial ports for interfacing multiple serial devices from the same microprocessor.

disadvantages of PC data acqusition systems

Most of the data acquisition systems, whether they are digital I/O or analog A/D conversions are not user friendly devices. Nearly all of them take several hours of dedicated familiarization time and many of them may take a little reprogramming or a lot of reprogramming to operate properly. They are limited to 40 or 50 megasamples-per-second because the computer clocks operate around 25 to 66 MHz. Events that occur in the nanosecond range are currently outside the range of PC data acquisition systems. Events in the microsecond range, however fall well within the PC data acquisition systems range making PC data acqusition systems an excellent tool for millisecond and microsecond events.

advantages of PC data acqusition systems

Even with the programming trouble the PC data acquisition systems are still mighty fine pieces of equipment. They are simple to use, readily available, and accurate in the microsecond range, and have on board capability to perform real time data reduction and control at a very reasonable cost. If the goal is to have a dedicated data acquisition and control system or multiple identical dedicated data acquisition and control system it becomes very economical to invest in dedicated programming time for task specific software.

conclusion

PC data acquisition systems purchased by the anti-penetration laboratory should be separated into several PC's and operated as distinct data acquisition systems. They will then be capable of performing tasks in parallel and will provide the laboratory with an array of very effective data logging and data reduction tools.

The PC data acquisition systems are very powerful and simple to use after an initial period of familiarization. To utilize a PC data acquisition system the user must stay familiar and current in operations of the pieces of equipment he will be working with. Intermittent use means the user must spend time retraining in the proper use of a specific PC data acquisition system before the system is brought on line and often errors are generated due to lack of familiarity. It is not necessary to have a specialist for a single piece of equipment but it is necessary to have a specialist can be the field to operate the several pieces of equipment. A specialist should dedicate time to staying current and familiar with several pieces of similar equipment and be capable of operation any single piece of equipment immediately and efficiently.

There is a need for more user friendly software for the PC data acquisition systems. As the market has grown the PC data acquisition system have become more user friendly but they have a bit to go before they have an intuitive feel about their use and operation.

references

[1] CyberResearch Inc., "PC systems Handbook for Scientists & Engineers", CyberResearch, New Haven, CT, Vol. 7, No 2., fall 1990, 10 th anniversary edition, 1993/1994.

[2] Gage Applied Science, "CompuScope 220 Digital sampling Oscilloscope Installation & Users Guide', Gage Applied Science, Montreal Canada, February 1988.

[3] R.C. Electronics, "Computerscope ISC-16 Reference Manual", R.C. Electronics, Goleta, GA, December 1990.

[4] R.C. Electronics, "RC-200 Programmable Stimulator and Control System User's Manual Version 1.3", R.C. Electronics, Goleta, GA, December 1990.

[5] Strawberry Tree Inc., 'Analog Connection ACPC Data Acquisition & Control System for the IBM PC and Compatible Computers", Strawberry Tree Inc. Computer Instrumentation and Controls, Sunnyvale, CA, June 1988.

[6] Strawberry Tree Inc., 'Analog Connection ACAO Data Acquisition & Control System for the IBM PC and Compatible Computers", Strawberry Tree Inc. Computer Instrumentation and Controls, Sunnyvale, CA, June 1988.

[7] Strawberry Tree Inc., 'QuickLog PC user's Guide, Applications Software for the IBM PC and Compatible Computers", Strawberry Tree Inc. Computer Instrumentation and Controls, Sunnyvale, CA, June 1988.

[8] ComputerBoards Inc., "CIO-AD16 User' Manual Revision 4 May 1990 & CIO-AD16 Utility Disk Revision 2.3", ComputerBoards Inc. Mansfield, MA, May 1990.

[9] ComputerBoards Inc., "CIO-CTR User' Manual Revision 1.1 March 1990 & CIO-CTR Utility Disk Revision 1.1", ComputerBoards Inc. Mansfield, MA, March 1990.

[10] National Instruments Corporation, "AT-DIO-23F Users Manual, January 1991 Edition Part Number 320147-01", National Instruments Corporation, Austin, TX, January 1991.

[11] Microstar Laboratories Inc., "DAPL Manual Data Acquisition Processor L Analog Accelerator Series Version 3.3A", Microstar Laboratories Inc., Redmond, WA, 1991

[12] Arnet Corporation, "Smartport Expansion Board User's Manual, MAN-0177-010 Rev. 1.0 July 1990", Arnet Corporation, Nashville, TN, 1990.

SKIN-FRICTION AND FLOW DIRECTION MEASUREMENT BY SURFACE-OBSTACLE INSTRUMENTS

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 Research Program

Wright Laboratory

Sponsored by:

Air Force Office of Scientific Research

Bolling Air Force Base, Washington, DC

August 1993

SKIN-FRICTION AND FLOW DIRECTION MEASUREMENTS BY SURFACE-OBSTACLE INSTRUMENTS

Raimo J. Hakkinen
Professor and Director, Fluid Mechanics Laboratory
Department of Mechanical Engineering
Washington University

Abstract

Calibration data of surface-obstacle skin friction meters, including blocks, fences, Preston tubes and Stanton tubes, were examined throughout the available ranges of dimensionless wall-shear-stress and pressure-differential parameters. The calibration relationships, including the effect of compressibility, were reformulated in terms of variables containing physical quantities at the wall and including the probe size in only one parameter. In view of the current trend toward miniaturization, special attention was given to the range where the flow disturbance introduced by the probe remains essentially within the linear part of the velocity profile. Design criteria were derived for differential for given flow properties and shear stress at the wall. Detail design of a specific adjustable/retractable surface-obstacle device for the Wright Laboratory M3 and M6 supersonic wind tunnels was initiated, and a test program was proposed.

SKIN FRICTION AND FLOW DIRECTION MEASUREMENT BY SURFACE-OBSTACLE INSTRUMENTS

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 [1]. 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 there exists a great variety of experimental techniques 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 Rechenberg [2], Winter [3], Settles [4], and Hakkinen [5].

The 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 study is intended to provide the groundwork for the development of 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 proposed instrument 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 is 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 results of the project are presented in this final report in two parts: (a) evaluation and reformulation of the calibration relations for surface-obstacle devices and derivation of design criteria for choosing the smallest possible probe size for given flow conditions; and (b) description of a specific design for use in the Wright Laboratory supersonic/hypersonic wind tunnels.

Technical Background

Relative to the well-established techniques for direct local measurement of other fluid-dynamic parameters, 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 been limited to simple configurations, generally flat plates or axisymmetric surfaces, where no significant streamwise pressure gradients are present.

The fundamental problem of direct force measurement under general conditions stems from two physical considerations: (a) the force-sensing element is isolated from the surrounding surface by gaps, which allow a pressure gradient across the element to exert an error force on the upstream and downstream edges of the element; and (b) a curved element not pivoting around its center of curvature also experiences a pressure-gradient-induced additional force in the direction of the measured shear force. As discussed in e.g. [5], in the limiting case of a flat element the error caused by consideration (b) can be eliminated by the use of a parallel-shifting support linkage instead of a single pivot.

The use of a parallel-moving element does not eliminate the gap force of consideration (a). Suppose that shear stress τ acts on a flat sensing element of streamwise length l; thus, the shear force to be measured is $\tau\, l$ per unit width of element. If a pressure gradient dp/dx acts in streamwise direction over the element, the force exerted over an effective edge depth t is $(dp/dx)\, t\, l$, giving a relative error

$$\varepsilon = \frac{(dp/dx)\, t\, l}{\tau l} = \frac{[dc_p / d(x/c)]}{c_f\ (t/c)}$$

It is essential to note that the relative error is independent of the streamwise extent of the element, l, and

also of the reference length, c, which is introduced here solely for convenient estimation of ε in practical cases. As an example, consider an airfoil with chord c = 1 m. Excluding the much stronger favorable gradients in the nose region, in the front part of the upper airfoil surface one may have adverse pressure gradients of $dc_p/d(x/c)$ = 10 and skin friction coefficient c_f = 0.002, yielding for a t = 1 mm effective gap height an error ε = 5, or 500%. In most cases, it is practical to bevel both moving and stationary edges to reduce its effective height, but a reduction to t = 0.1 mm would still yield an error ε = 0.5, or 50%. The situation is somewhat better on the rear half of a typical airfoil, where $dc_p/d(x/c)$ = 1 would be more likely; for c_f = 0.002 the errors would still be ε = 0.5 (50%) and 0.05 (5%) for t = 1 mm and 0.1 mm, respectively.

Although quantitative edge effects have been investigated using known pressure gradient/shear stress relationships in pipes, and an ingenious compensation technique using surface tension phenomena in liquid-filled gaps has been developed in Switzerland [6], there is no accurate correction procedure applicable under general flow conditions. Therefore, the floating element has been relegated primarily to direct shear force measurement on flat or cylindrical surfaces with negligible pressure gradients. It does, however, have an important role as an absolute calibration reference for indirect techniques.

It is re-emphasized that, in a given flow condition, the relative magnitude of the pressure-gradient-induced error is independent of the streamwise size of the floating element; hence, miniaturization does not improve its accuracy unless the edge depth facing the surrounding isolation gap is significantly reduced below the already small dimensions found in traditional designs. A development in this direction is the microdeposited device of Schmidt, Howe, Senturia and Haritonidis [7], where the exposed element has an edge depth of only t = 0.03 mm; however, flow underneath the element may introduce additional errors. The development of a moving-belt-type direct-force sensor by Vakili and Wu [8] should also be mentioned.

Because of the error potential and operational limitations inherent in traditional floating element devices, various indirect instruments have been developed for measurement of skin friction under practical flow situations. The most prominent of these instruments is the Preston-tube [9] which consists of a round total head probe placed on the surface. The large data base of Preston-tube calibrations [9,10,11,12,13,14,15,16,17] covers both low-speed flows and compressibility effects. The Preston-tube belongs to the class of surface-obstacle instruments where the sensed quantity is the pressure differential between the pressure on the upstream face of the instrument and the local static pressure in undisturbed flow; this pressure differential balances the shear force driving the flow region immediately upstream of the obstacle, and a calibration relationship can be established in terms of dimensionless parameters that

involve flow properties on the surface. These relationships will be reviewed and reformulated in this report; while the bulk of the data derives from experiments on Preston-tubes, available data on related devices, such as Stanton-tubes and sublayer fences are included.

In view of new opportunities and needs for miniaturization of instruments, it is interesting to note that the minimum size (in terms of lower limit of operating range) often specified for standard Preston-tubes does not result from manufacturing limitations, but from the cessation of validity of the accepted calibration correlations as the values of the dimensionless shear-stress and pressure-differential parameters decrease. Corresponding to a change in the dominating physical environment, a different calibration law assumes validity when the disturbance created by very small probes no longer extends beyond the near-wall, near-linear range of the velocity profile. Stanton-tubes and sublayer fences have been frequently operated successfully in this range, and some data exist on small Preston-tubes. These data allow formulation of a general understanding of the underlying calibration framework, and will be used as a basis for design criteria of small surface probes.

At first, surface-obstacle measurements in pressure gradients would appear to suffer from the same, large errors found in the case of floating elements. Fortunately, available data on Preston-tube errors in pressure gradients, especially the meticulous work of Hirt and Thomann [13], indicate that for a probe protrusion height equal to an effective edge depth the errors are reduced by an order of magnitude.

Most direct force sensors and surface-obstacle instruments are inherently unidirectional. However, directionally sensitive floating element designs have been manufactured [18,19], but the additional complexity and still limited applicability have precluded their adoption. Preston- and Stanton-tubes are unidirectional; directionally-sensitive variations of blocks have been explored by Dexter (as reported by Winter in [3]) and Iuso, Onorato and Spazzini [20]. Fiore and Scaggs [21] used a rotatable fence to determine flow direction. Although heat-transfer instruments calibrated as skin-friction meters are not within the scope of this study, the successful experiments conducted with directionally sensitive hot-film arrays should be mentioned, e.g. McCroskey and Durbin[22]. Directionally sensitive oil film instruments have also been designed, such as that of Seto and Hornung [23].

In view of this background, it appears 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. The remainder of this report addresses design criteria for such an instrument.

General Calibration of Surface Obstacle

As will be seen in the following, all practical calibration relationships can be expressed in terms

of dimensionless variables containing the surface shear stress τ, the pressure differential Δp on the face of the obstacle, the protrusion height of the obstacle h, and fluid properties at the wall: density ρ, dynamic viscosity μ, kinematic viscosity ν, and pressure p_e, assumed constant across the boundary layer and equal to the external static pressure. The general calibration relationship can then be expressed as

$$\tilde{p} = f(\tilde{\tau}, \tilde{p}_e) \tag{1}$$

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 Preston-tube and some other calibration relationships to this form is reviewed in the following and summarized in Fig. 1. It is expected that the surface-obstacle design developed in this project would be characterized by a similar calibration pattern.

For the purpose of selecting the size of a surface obstacle for a given flow situation, the general calibration relationships can be expressed as

$$\frac{\Delta p}{p_e} = \frac{\tilde{p}}{\tilde{p}_e} = \frac{1}{\tilde{p}_e} f\left(\frac{\tau}{p_e}\tilde{p}_e, \tilde{p}_e\right) = \bar{f}\left(\frac{\tau}{p_e}, \tilde{p}_e\right) \tag{2}$$

where the probe size, h, appears in only one parameter, $\tilde{p}_e$. A plot in these variables is presented in Fig. 2, primarily on the basis of the extensive data on Preston-tubes, and it is anticipated that a similar guide chart can be prepared for selecting the proper protrusion of the proposed adjustable instrument.

The dimensionless parameter $\tilde{p}_e$ is introduced to the calibration relationship (1) by compressibility effects, and disappears from most equations in their absence. However, in form (2) p_e is always present because of the normalization of τ and Δp by p_e.

It should be noted that if compressibility effects are determined solely by the flow around the probe, free-stream conditions other than $\tilde{p}_e$ should then be necessary only if it is desired to relate e.g. surface temperature to free-stream temperature and external Mach number. However, some calibration correlations are based on the concept of reference temperature, as defined by Sommer and Short [24], and thereby contain an implicit dependence on T_e. In [12], Keener and Hopkins state that use of the reference temperature instead of actual wall temperature does not markedly improve accuracy of the calibration laws, and suggest the use of wall temperature in the non-dimensional parameters. Except as specifically noted, the actual wall temperature will be used throught the rest of this report.

Preston-tube Calibrations

A great variety of Preston-tube calibrations is found in the literature, starting with the initial work of Preston [9]. These empirical relationships are curve fits of analytical expressions of various complexities to experimental data, and in most cases agree within a few per cent, except at the lower

values of the dimensionless shear parameter where the influence of the outer velocity profile is gradually diminishing. In the following, some of the generally used relationships are given first in their original form and then recast in the form of equation (1):

Compressibility effects are 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 relation

$$\frac{\Delta p}{p_e} = \frac{\tilde{p}}{\tilde{p}_e} = \left(1 + \frac{\gamma - 1}{2} M_p^2\right)^{\frac{\gamma}{\gamma - 1}} - 1 \tag{3}$$

or, at supersonic values of M_p by the normal shock loss combined with isentropic expansion

$$\frac{\Delta p}{p_e} = \frac{\tilde{p}}{\tilde{p}_e} = \frac{\left(\frac{\gamma + 1}{2} M_p^2\right)^{\frac{\gamma}{\gamma - 1}}}{\left[\left(\frac{\gamma - 1}{\gamma + 1}\right)\left(\frac{2\gamma}{\gamma - 1} M_p^2 - 1\right)\right]^{\frac{\gamma}{\gamma - 1}}} - 1 \tag{4}$$

It is convenient to define a "probe dynamic pressure"

$$\tilde{q}_p = \frac{\gamma}{2} \tilde{p}_e M_p^2 \tag{5}$$

If the compressible calibration laws are then replotted with replacement of $\tilde{p}$ by $\tilde{q}_p$, a common calibration chart can be prepared covering both compressible and incompressible regimes (Fig. 1). It is easily seen that as $M_p \to 0$, $\tilde{q}_p \to \tilde{p}$, and in most cases the parameter $\tilde{p}_e$ disappears from the equation..

Preston [9]

$$\log_{10}\left(\frac{\tau d^2}{4\rho\nu^2}\right) = -1.396 + \frac{7}{8}\log_{10}\left(\frac{\Delta p d^2}{4\rho\nu^2}\right) \tag{6a}$$

$$\tilde{p} = 48.02\,\tilde{\tau}^{\frac{8}{7}} \tag{6b}$$

Allen (11):

$$\log_{10}\left(\sqrt{\frac{\rho}{\rho_e}}\frac{\mu_e}{\mu} R_d \sqrt{c_f}\right) = 0.02139\left[\log_{10}\left(\sqrt{\frac{\rho}{\rho_e}}\frac{\mu_e}{\mu} R_d \frac{u_{pt}}{u}\right)\right]^2 + 0.7814\log_{10}\left(\sqrt{\frac{\rho}{\rho_e}}\frac{\mu_e}{\mu} R_d \frac{u_{pt}}{u}\right) - 0.4723 \tag{7a}$$

$$\log_{10}\tilde{\tau} = 0.0062(\log_{10}\tilde{q}_p)^2 + 0.785\log_{10}\tilde{q}_p - 1.0098 \qquad (7b)$$

Allen also gives a simple power-law which,when recast in the same parameters as in (7a), reduces to

$$\tilde{q}_p = 37.50\,\tilde{\tau}^{1.132} \qquad (7c)$$

Allen's wall parameters in [11] were evaluated at the reference temperature of Sommer and Short [24].

Sigalla [14]:

$$\frac{\rho}{\rho_e}\frac{\mu_e}{\mu}R_d\frac{u_{pt}}{u} = 5.13\left(\sqrt{\frac{\rho}{\rho_e}}\frac{\mu_e}{\mu}R_d\sqrt{c_f}\right)^{1.146} \qquad (8a)$$

$$\tilde{q}_p = 29.12\,\tilde{\tau}^{1.146} \qquad (8b)$$

Fenter and Stalmach [15]:

$$\frac{\rho}{\rho_e}\frac{\mu_e}{\mu}\frac{\sqrt{5+M_e^2}}{M_e}R_d\sin^{-1}\left(\sqrt{\frac{\rho}{\rho_e}}\frac{\mu_e}{\mu}R_d\frac{u_{pt}}{u}\right) = \sqrt{\frac{\rho}{\rho_e}}\frac{\mu_e}{\mu}R_d\sqrt{c_f}\left[4.06\log_{10}\left(\sqrt{\frac{\rho}{\rho_e}}\frac{\mu_e}{\mu}R_d\sqrt{c_f}\right)\right] + 1.77 \qquad (9a)$$

$$\tilde{p}_{FS} = \frac{2\gamma}{\gamma-1}\left(\frac{T_0}{T}\right)\tilde{p}_e\left\{\sin^{-1}\left[\sqrt{\left(\frac{\gamma-1}{2}\right)\left(\frac{T}{T_0}\right)}\,M_p\right]\right\}^2 = \tilde{\tau}(0.8816\ln\tilde{\tau} + 2.38)^2 \qquad (9b)$$

It is easily seen that as

$$M_p \to 0, \tilde{p}_{FS} \to \tilde{q}_p \to \tilde{p}.$$

Bradshaw and Unsworth [16]:

$$\frac{\Delta p}{\tau} = 96 + 60\log_{10}\left(\frac{d\sqrt{\tau\rho}}{50\mu}\right) + 2.37\left[\log_{10}\left(\frac{d\sqrt{\tau\rho}}{50\mu}\right)\right]^2 + 10^4\frac{\tau}{a^2\rho}\left[\left(\frac{d\sqrt{\tau\rho}}{50\mu}\right)^{0.30} - 2.38\right] \qquad (10a)$$

$$\tilde{p} = \tilde{\tau}\left[0.903 + 25.973(\log_{10}\tilde{\tau}) + 0.5933(\log_{10}\tilde{\tau})^2 + 10^4\frac{\tilde{\tau}}{\gamma\tilde{p}_e}\left(\tilde{\tau}^{0.15} - 2.38\right)\right] \qquad (10b)$$

Bertelrud [17]:

$$\frac{\Delta p}{\tau} = 87.77\log_{10}\left(\frac{u_\tau d}{\upsilon}\right) - 51.93 \qquad (11a)$$

$$\tilde{p} = \tilde{\tau}(43.885\log_{10}\tilde{\tau} - 51.93) \qquad (11b)$$

Keener and Hopkins [12]:

$$\sqrt{\frac{\rho}{\rho_e}}\frac{\mu_e}{\mu}R_d\frac{M_{pt}}{M_e}=5.74\left(\sqrt{\frac{\rho}{\rho_e}}\frac{\mu_e}{\mu}R_d\sqrt{c_f}\right)^{1.132} \tag{12a}$$

$$\tilde{q}_p=36.10\,\tilde{\tau}^{1.132} \tag{12b}$$

Patel [10], as modified by Frei [6]:

$180<\tilde{\tau}<12600$:

$$\log_{10}\left(\frac{\tau d^2}{4\rho\nu^2}\right)=1.6167-0.7405\log_{10}\left(\frac{\Delta p d^2}{4\rho\nu^2}\right)+0.2914\left[\log_{10}\left(\frac{\Delta p d^2}{4\rho\nu^2}\right)\right]^2-0.0177\left[\log_{10}\left(\frac{\Delta p d^2}{4\rho\nu^2}\right)\right]^3 \tag{13a}$$

$$\log_{10}\tilde{\tau}=2.7741-1.1106\log_{10}\tilde{p}+0.3234\left[\log_{10}\tilde{p}\right]^2-0.0177\left[\log_{10}\tilde{p}\right]^3 \tag{13b}$$

$12600<\tilde{\tau}<10^5$:

$$\tilde{p}=\tilde{\tau}\left[1.77\log_{10}(1.27\tilde{\tau})\right]^2 \tag{13c}$$

If the actual wall temperature is used to evaluate flow properties at the surface, no free-stream flow parameters are included in the above relationships besides the static pressure which is assumed to be constant across the boundary layer (and a wall-to-stagnation-temperature-ratio in the the Fenter and Stalmach relation (15)). However, it is often convenient to express surface properties in terms of free-stream Reynolds number R_x, Mach number M_e or skin friction coefficient c_f, for instance

$$\tilde{\tau}=\left(\frac{c_f}{2}\right)\left(\frac{\mu_e}{\mu}\frac{T_e}{T}\right)^2 R_c^2\left(\frac{h}{c}\right)^2, \tag{14}$$

where c is an arbitrary reference length, such as the airfoil chord. The viscosity and temperature ratios may be expressed in terms of M_e for an adiabatic wall with a known recovery factor.

Because of the expressions used for description of compressibility effects in the relationships of Fenter and Stalmach and of Bradshaw and Unsworth, the parameter $\tilde{q}_p$ does not appear conveniently; however, both relationships can be shown to reduce to the basic dependence of $\tilde{p}$ on $\tilde{\tau}$ as M_p approaches zero. When the calibration relations are examined in terms of τ/p_e and $\Delta p/p_e$ ratios, the replacement of $\tilde{p}$ by $\tilde{q}_p$ is seen to be equivalent to a non-linear distortion of the $\Delta p/p_e$-scale.

Some of the calibration relations discussed above are given in the form of a simple power-law

$$\tilde{q}_p=a_n\,\tilde{\tau}^n \tag{15}$$

while others are more complex non-linear expressions in terms of the same variables. In fact, the power-law relations may be considered tangents drawn on the logarithmic plot to the more general calibration

relationship at different values of τ. Such tangents to Patel's lower-range equation, (13b), have $n = 1.298$ at $\tilde{\tau} = 867$ and $n = 1.165$ at $\tilde{\tau} = 1.27 \times 10^5$, in good agreement with the simpler relationships. In the higher range, the expression

$$\tilde{q}_p = 35.55 \tilde{\tau}^{1.13} \tag{16}$$

agrees with Allen's square-logarithmic equation (7b) within +/- 0.6%.

In the middle of the lower range of Preston's calibration, at $\tilde{\tau} = 866$, expression (16) is too high by 17 %, which is not surprising because Preston's lower range is influenced by a few data points obtained in the Stanton-tube range characterized by the 5/3-power relationship. In fact, the tangent to Preston's expression at its lower limit, $\tilde{\tau} = 165$, has the exponent $n = 1.72$.

Stanton-tube (5/3-power) range

At the lowest values of $\tilde{\tau}$ in Preston's tube data the exponent of the correlation curve given in [7] increases to 1.72, and both Preston [9] and Patel [10] refer to a square relationship between $\tilde{\tau}$ and $\tilde{p}$ for lower values of $\tilde{\tau}$. However, in this regime, data on Stanton-tubes seem to follow a lower exponent, such as the 5/3-power law identified experimentally as 5/3 by Hakkinen [5]. In [24] Trilling and Hakkinen showed that the 5/3-exponent indeed results theoretically from the assumption that the disturbance created by the probe is confined to the near-linear part of the boundary layer next to the surface, and fitted then available data to this form of calibration. Recently, data following the 5/3-law have been obtained by Weiser, Nitsche and Renken on sublayer fences [26].

The few points measured by Preston [9] near the upper limit of the range under consideration appear to be consistent with a 5/3-power law as well as with the square relationship suggested by Preston and Patel; the 5/3-power law is certainly consistent with Patel's equation, (13b).

Patel [7] proposed for $\tilde{\tau} < 125$ the relation

$$\tilde{p} = 0.211 \tilde{\tau}^{\,2} \tag{17}$$

which matches the value, but not the local exponent, of the more complicated higher range relationship. In view of the above discussion, it appears reasonable to match both the values and the exponents of the 5/3-power and the higher-range correlations. For the original Preston-correlation, this condition is satisfied at $\tilde{\tau} = 182.8$, $\tilde{p} = 6{,}577$, yielding

$$\tilde{p} = 1.117 \, \tilde{\tau}^{\,5/3} \tag{18}$$

Frei [6] has adopted an improved correlation for the Preston-tube range, starting at $\tilde{\tau} = 179.1$; the value and exponent can be matched at $\tilde{\tau} = 199.7$, $\tilde{p} = 7622$, to yield the same coefficient.

While the coefficient, $n = 1.117$, can be expected to be valid only for round surface-tubes; the

general calibration relationship has been observed for other types of obstacles. The coefficient depends on the particular probe geometry but probably remains within an order of magnitude [20,26,27]. It is likely that even the coefficient determined above from the match with Preston-tube correlation will be modified when more low-range data on round surface-tubes become available.

Very few data are 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, $\tilde{q}_{p_c}$, in place of $\tilde{p}$ would appear reasonable in compressible boundary layers. If the entire disturbance is contained within the near-linear layer next to the wall, the calibration in the Stanton-range should also be independent of the state of the boundary layer being laminar or turbulent; as suggested by the limited data in [2]; however, differences in calibration will probably occur in the transition from the lower to the higher range. Therefore, for $\tilde{\tau} > 200$ the above relationships must be used with caution if there is a possibility of the boundary layer being laminar.

The purpose of the present study is not to advocate extension of Preston-tube methodology to the lower range, but to obtain an overall characterization of the calibration laws governing surface obstacles in general. For this purpose, the most comprehensive data set is available for Preston-tubes.

Characterization of Surface Obstacle Calibration

Using the Preston-tube data base as the working example, the following expressions are proposed for general characterization of surface-obstacle calibration:

$12{,}600 < \tilde{\tau} < 100{,}000$:

$$\tilde{q}_p = 35.55\ \tilde{\tau}^{1.13} \tag{19a}$$

$200 < \tilde{\tau} < 12{,}600$:

$$\log_{10}\tilde{\tau} = 2.7741 - 1.1106\log_{10}\tilde{q}_p + 0.3234\left[\log_{10}\tilde{q}_p\right]^2 - 0.0177\left[\log_{10}\tilde{q}_p\right]^3 \tag{19b}$$

$10 < \tilde{\tau} < 200$:

$$\tilde{q}_p = 1.117\ \tilde{\tau}^{\ 5/3} \tag{19c}$$

$\tilde{\tau} < 10$:

$$\tilde{p} \approx 1.2\tilde{\tau} \tag{19d}$$

for the specific configuration of Taylor [28].

The calibration correlations in the two upper ranges have been established for turbulent boundary layers; in the two lower ranges the boundary layers are most likely laminar. At values of $\tilde{\tau}$ in the upper hundreds, 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.

To estimate the performance of a given probe size in a particular experiment, the calibration laws written in the form of Equation (2) are most useful. For a subsonic probe Mach number M_p,, substitution of the probe dynamic pressure (5) to correct for compressibility yields, from (3),

$$\frac{\Delta p}{p_e} = \mathrm{F}(M_p{}^2) = \mathrm{F}\left[\frac{2}{\gamma \tilde{p}_e} f\left(\frac{\tau}{p_e}, \tilde{p}_e\right)\right] = \left[1 + \frac{\gamma - 1}{\gamma \tilde{p}_e} f\left(\frac{\tau}{p_e}, \tilde{p}_e\right)\right]^{\frac{\gamma}{\gamma - 1}} - 1 \tag{20}$$

If M_p exceeds one, i.e. $\Delta p/p_e > 0.8929$, Equation (4) must be used for $\mathrm{F}(M_p{}^2)$ in (20). It is now convenient to express the parameter $\tilde{p}_e$, which contains neither the shear stress nor the face pressure differential, as a dimensionless probe size parameter:

$$\tilde{p}_e = p_e h^2/\rho \nu^2 = (h/h_O)^2, \tag{21}$$

where

$$(h_O)^2 = \rho \nu^2/p_e = \mu^2/p_e \rho = \mu^2 RT/p_e \tag{22}$$

For perfect gases, the surface protrusion scale

$$h_0 = \frac{\mu\sqrt{RT}}{p} \tag{23}$$

can be expressed in terms of readily available parameters: surface temperature T, which also determines surface viscosity μ, and the boundary layer static pressure p_e. Another interpretation is

$$h_0 = \frac{\sqrt{\gamma}\ \nu}{a} \tag{24}$$

which converts h/h_O essentially into a Reynolds number based on the speed of sound at the wall and probe protrusion. It is interesting to note also that h_O has the same physical dependence and order of magnitude as the mean free path; for practical probe sizes, h/h_O and $\tilde{p}_e$ are therefore large numbers.

In experiment planning, the essential consideration is to select the minimum probe protrusion to give a satisfactory Δp-reading for a given set of flow parameters. Thus, p_e and minimum Δp_e will be known, and if reasonable estimates can be made for τ, the required minimum probe protrusion h can then be determined from a chart where $\Delta p/p_e$ is plotted against τ/p_e, with h/h_O as a parameter. Such a chart can be prepared for Preston-tubes using the calibration laws identified above, and the objective of the present study is to obtain a similar guide for an adjustable, directionally-sensitive surface-obstacle probe.

For simplicity, the round surface-tube calibration is approximated on the logarithmic plot by two straight lines:

$$\tilde{q}_p = 1.117\tilde{\tau}^{5/3} \tag{25}$$

and

$$\tilde{q}_p = 35.55\ \tilde{\tau}^{1.13} \quad . \tag{26}$$

These tangent lines intersect at $\tilde{\tau} = 630$ and $\tilde{q}_p = 51630$, where the deviation from the Patel-curve peaks at about 22%. Both relationships can be expressed in the form

$$\frac{\Delta p}{p_e} = \left[1 + \frac{\gamma - 1}{\gamma} a_n \left(\frac{\tau}{p_e} \right)^n \left(\frac{h}{h_0} \right)^{2(n-1)} \right]^{\frac{\gamma}{\gamma - 1}} - 1 \tag{27}$$

and are plotted in Fig. 2. This chart can be used as a rough guide for selecting the proper Preston-tube size when it is desired to minimize the disturbance introduced into the boundary layer. A more refined plot can be prepared using the actual Patel-calibration curve in the changeover region. It is expected that a characterization and sizing guide similar to Fig. 2 can be prepared for any type of surface-obstacle instrument once adequate calibration data are available.

Remaining Concerns

The above integration of the Preston-tube calibration laws with the lower range is 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.

A possible source of deviation from the general calibration is the manner in which the face pressure is obtained; in [26], the orifice slit in front of a round surface wire is specified to have a width equal to the wire diameter. From the observed small effects of pressure gradient on Preston tubes, one might infer that the face pressure build-up occurs close to the instrument face, and the smallest possible orifice size is therefore advisable. In some instruments [21,26], the pressure differential is measured between similar orifices on the front and back faces of the obstacle, and the dependence of the base pressure on the shear stress enters into the calibration.

Design of instrument for Wright Laboratory:

The instrument proposed for use in the M3 and M6 tunnels has the following specifications: (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 at 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; and (f) use of commercially available pressure transducers, such as Baratron 225D series starting at $2mm_{water}$ (20 Pa) full scale. An accurate internal or external means is required for measurement of the protrusion height of the obstacle. A sketch of this instrument is presented in Fig. 3.

Proposed calibration test program

A considerable data base exists 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 [29,30]. A test matrix or the surface-obstacle instrument will be planned to utilize an adequate number of these known test conditions for maximum coverage of the operating parameter ranges of the instrument. At each test point, a sequence of protrusion heights will be tested, either by advance manual positioning or by remote control. The protrusion sequence, starting with the flush position, will be initially estimated from the Preston-tube chart in Fig. 2, and is expected to produce a similar chart specific to the obstacle instrument in the range of parameters covered in the tests. Schedule of the test program will be contingent on completion of the instrument and the availability of wind tunnel facilities. A parallel low-speed calibration program is being proposed for the 2-by-2-foot wind tunnel at Washington University.

Future work

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 similar to those used in recent high-temperature floating element designs [31].

References:

1. Szodruch, J., Viscous Drag Reduction on Transport Aircraft, AIAA 91-0685, 29th Aerospace Sciences Meeting, Reno, NV, Jan. 1991.

2. Rechenberg, I., Messung der turbulenten Wandschubspannung, Z. Flugwiss. Vol. 1, No. 11, Nov. 1963, pp. 429-438.

3. 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.

4. Settles, G. S., Recent Skin Friction Techniques for Compressible Flows, AIAA 86-1099, May 1986.

5. Hakkinen, R. J., Measurement of Skin Friction in Transonic Flow, Ph. D. thesis, California Institute of Technology, 1954.

6. Frei, Dieter, Direkte Wandschubspannungsmessung in der turbulenten Grenzschicht mit positivem Druckgradient, Dr. techn. Wiss. thesis, Federal Institute of Technology, Zurich, 1979.

7. Schmidt, Martin A., Howe, Roger T., Senturia, Stephen D., and Haritonidis, Joseph H., Design and Calibration of a Microfabricated Floating-Element Shear-Stress Sensor, IEEE Transactions on Electron devices, Vol. 35, No. 6, June 1988.

8. Vakili, A. D. ,and Wu, J.M., Wall Shear Stress Measurement Using a New Transducer, AIAA 86-1092, 1982.

9. Preston, J. H., The Determination of Turbulent Skin Friction by Means of Pitot Tubes, J. Roy. Aeron. Soc., Vol. 58, Feb. 1954.

10. Patel, V. C., Calibration of the Preston tube and limitation on its use in pressure gradients, J. Fluid Mech., vol. 23, part 1, pp. 185-208, 1965.

11. Allen, J., Reevaluation of compressible-flow Preston tube calibrations, NASA TM-X-3488, Feb. 1977.

12. Keener, E. R., and Hopkins, E. J., Use of Preston tubes for measuring hypersonic skin friction, NASA TN D-5544, Nov., 1969.

13. Hirt, F., Anzeige des Prestonrohres im Druckgradient, Dr. techn. Wiss. thesis, Federal Institute of Technology, Zurich, 1984.

14. Sigalla, A., Calibration of Preston Tubes in Supersonic Flow, AIAA J,vol. 3, no. 8, Aug.1965,p.1531.

15. Fenter, F. W., and Stalmach, C. J., Jr., The Measurement of Local Turbulent Skin Friction by Means of Surface Impact-Pressure Probes, DRL-393, CM-878, Univ. of Texas, Oct. 1957 (also AIAA J.).

16. Bradshaw, P., and Unsworth, K., Comment on "Evaluation of Preston Tube Calibration Equations in Supersonic Flow", AIAA J.,vol. 12, No. 9, Sept. 1974.

17. Bertelrud, A., Preston Tube Calibration Accuracy, AIAA J., Vol. 14, No. 1, 1976, pp. 98-100.

18. Tennant, M. H., Pierce, J. F., and McAllister, J. E., An Omnidirectional Wall Shear Meter, J. Fluids Engg., Vol. 102, March 1980, pp. 21-25.

19. Tcheng, P., Development of a Servo Transducer at LARC for Direct Skin Friction Measurement, 61st Annual Meeting of the Supersonic Tunnel Association, March 1984.

20. Iuso, G., Onorato, M., and Spazzini, P. G., Skin Friction Measurements in 3-D Boundary Layers, IEEE International Congress on Instrumentation in Aerospace Simulation Facilities, Silver Spring, MD, Oct. 27-31, 1991, pp.442-448.

21. Fiore, A., and Scaggs, N. E., Boundary Layer fence Method for Measuring Surface Shear Stress in a Supersonic, High Reynolds Number Flow Field, AFFDL Tech. Rept. AFFDL-TR-78-89, 1978.

22. McCroskey, W. J., and Durbin, E. J., Flow Angle and Shear Stress Measurements Using Heated Films and Wires, J. Basic Engg., Vol. 94, 1, 1975, pp. 46-52.,

23. Seto, J., and Hornung, H..,Two-Directional Skin Friction Measurement Utilizing a Compact internally-Mounted Thin-Liquid-Film Skin Friction Meter, AIAA Paper 93-0180, 31st Aerospace Sciences Meeting & Exhibit, Reno, NV, Jan. 1993,

24. Sommer, S. C., and Short, B. J.,Free-Flight Measurements of Turbulent-Boundary-Layer Skin Friction in the Presence of Severe Aerodynamic Heating at Mach Numbers From 2.8 to 7.0, NACA TN 3391, 1955. .

25. Trilling, L., and Hakkinen, R. J., The Calibration of the Stanton Tube as a Skin Friction Meter, in 50 Jahre Grenzschichtforschung, F. Vieweg & Sohn, 1955.

26. Weiser, N., Nitsche, W., and Renken, F., Wall Shear Stress Determination by Means of Obstacle-Wires, Eighth Symposium On Turbulent Shear Flows, Technical University Of Munich, Sept. 9-11, 1991.

27. Fage, A., and Falkner, V. M., An Experimental Determination of the Intensity of Friction on the Surface of an Airfoil, Proc. Roy. Soc., Series A, Vol. 129, 1930.

28. Taylor, G. I., Measurements with a Half-Pitot Tube, Proc.Roy.Soc., Ser.A.,Vol.166, 1938, pp.476-481.

29. 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.

30. Wagner, M. J., Skin Friction and Heat Transfer Measurements in Mach 6 High Reymolds Number Flows, 15th Internatiomnal Congress on Instrumentation in Aerospace Simulation Facilities, Saint-Louis, France, 20-23 September, 1993.

31. DeTurris, D. J., and Schetz, J. A., Direct Measurements of Skin Friction in a Scramjet Combustor, AIAA 90-2342, AIAA/SAE/ASME/ASEE Joint Propulsion Conference, Orlando, FL, July 1990.

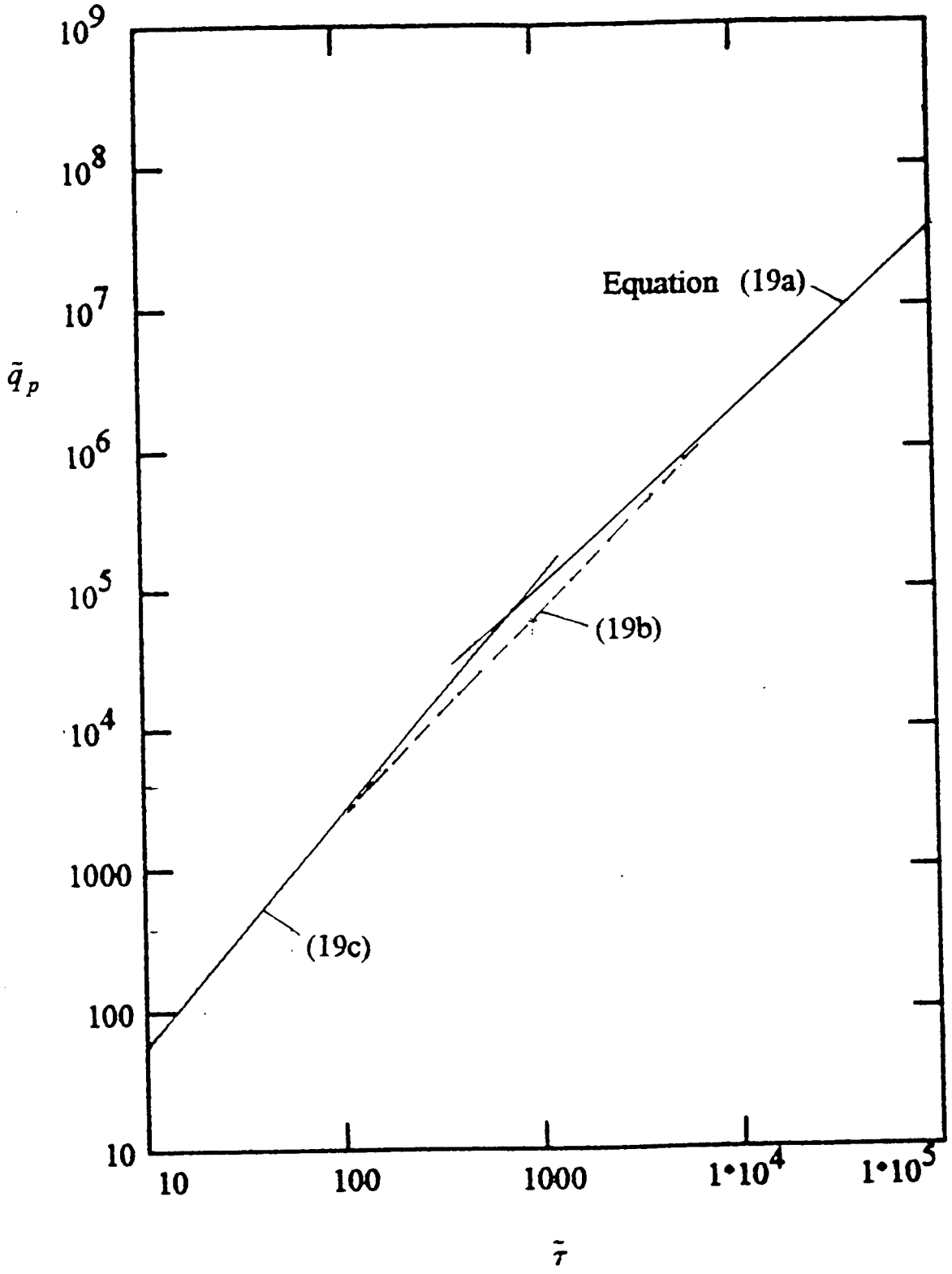


Fig. 1 Simplified calibration law of round surface tubes

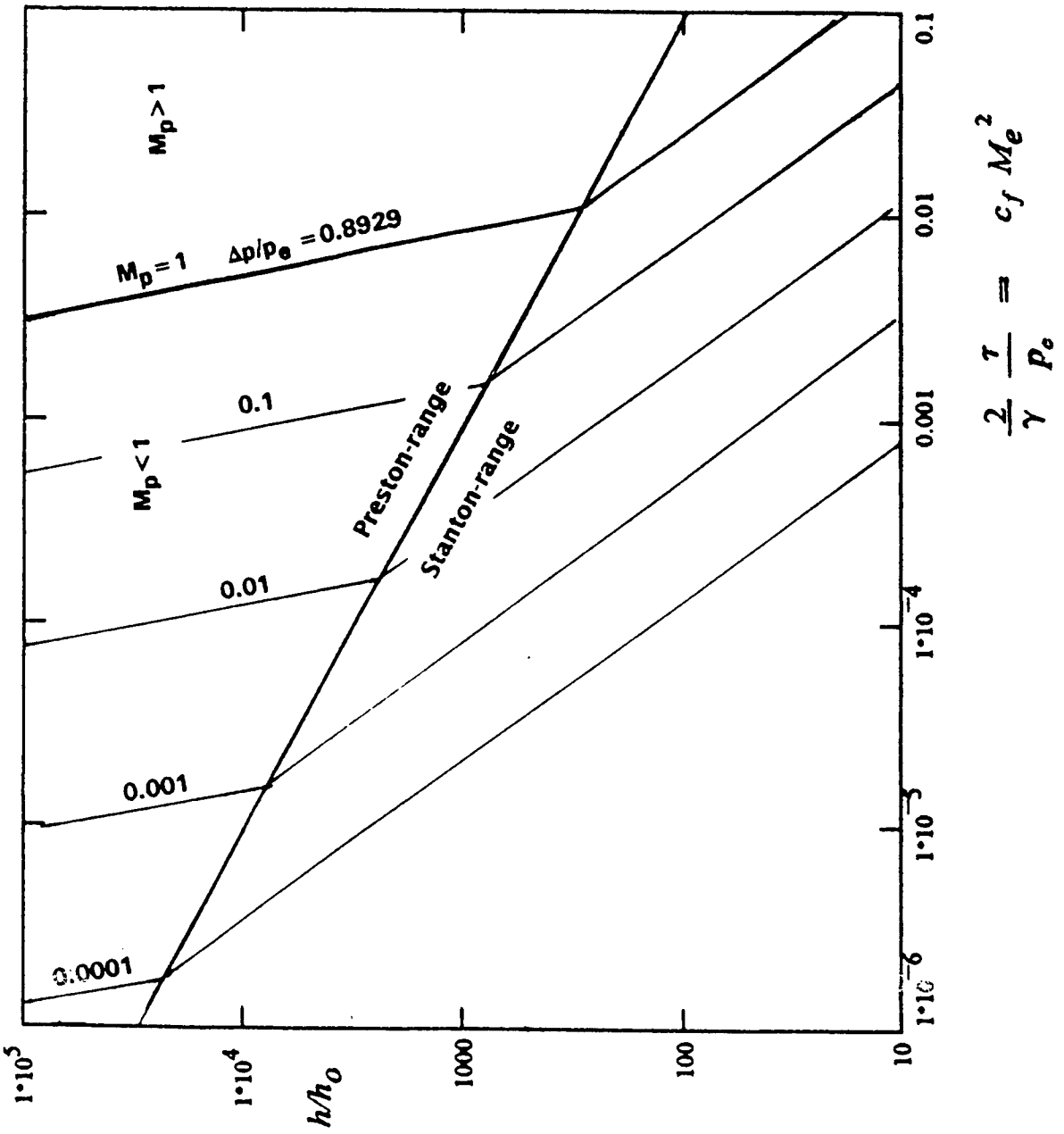


Fig. 2 Sizing chart of round tubes as function of shear stress and face pressure differential

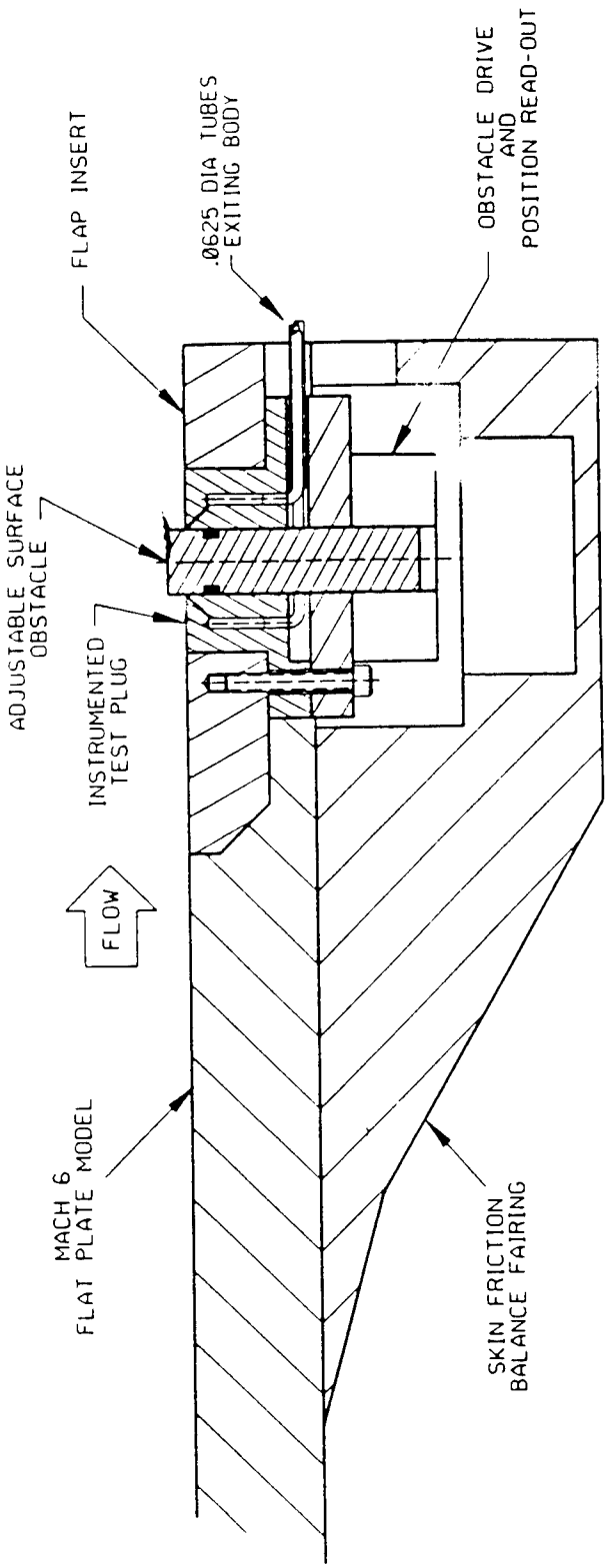


Fig. 3 Proposed design for Wright Laboratory M6 wind tunnel

NONLINEAR FEEDBACK CONTROL OF LINEAR DYNAMIC SYSTEMS

Charles E. Hall, Jr.
Assistant Professor
Department of Mechanical and Aerospace Engineering

North Carolina State University
Raleigh, NC 27695-7910

Final Report for:
Summer Research Program
Wright Laboratory

Sponsored by:
Air Force Office Of Scientific Research
Bolling Air Force Base, Washingron, D.C.

August, 1993

NONLINEAR FEEDBACK CONTROL OF LINEAR DYNAMIC SYSTEMS

Charles E. Hall, Jr.
Assistant Professor
Department of Mechanical and Aerospace Engineering
North Carolina State University

ABSTRACT

The control of various systems it is desirable to have two types of feedback control for a system depending on the distance that the system has to travel through the state space. Using a nonlinear feedback law allows for the changing of the feedback gain as a function of the system's error. The technique is primarily directed at a third order system. This limitation was primarily for the reduction of the complexity of the calculations. The feedback gain is calculated by standard linear control techniques for large and small error values. These two gain matrices are combined in a nonlinear function. For zero error, the nonlinear feedback function is equal to the small error feedback matix. For large, approaching infinity, error values the nonlinear feedback function approaches the gain matrix for large errors. One possible nonlinear function is formulated. The requirements for the stability of the system as a function of these two gain endpoints are formulated using Lyapunov's Second Method. These techniques are applied to an F-15 aircraft. This example system is a longitudinal controller based on the Short Period Approximation and a first order servo model. The nonlinear feedback function blended two LQR feedback control law designs. The results are not as dramatic as anticipated, but indicated that further study is warranted.

NONLINEAR FEEDBACK CONTROL OF LINEAR DYNAMIC SYSTEMS

Charles E. Hall, Jr.

INTRODUCTION

It is sometimes desirable to use different feedback gains in various regions of the state space to enhance the response of a system to large deviations from the desired steady state values. This has classically been accomplished through the use of gain scheduling. More recently the use of fuzzy logic controllers have been proposed to apply nonlinear feedback to control systems. Fuzzy logic controllers, despite the statements of their enthusiasts, are not the cure all and involve a more complicated design precedure. Fuzzy logic controllers are deterministic systems with a more complicated design process due to the shape of the member sets[1]. Another way of implementing the different feedback gains is by designing a nonlinear feedback controller. The remainder of this paper discusses the design of one particular nonlinear feedback technique for a system with linear plant dynamics.

THEORETICAL DEVELOPMENT

The objective is to use nonlinear state feedback on a system with linear dynamics with a 3-dimensional state space with a single input. State feedback was chosen so that an LQR feedback design could be implemented. A state estimator could be formulated for those systems in which the system's output is not simply the system's state. The following analysis works can also be used for systems employing output feedback. The limit of the dimension of the state space was chosen to reduce the complexity of the calculations. In addition this method was designed with a aircraft longitudinal control system in mind, which is 3-dimensional when the Short-Period Mode and a first order servo are used. This system is described by

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a_1 & a_2 & a_3 \end{pmatrix}; \qquad B = \begin{pmatrix} 0 \\ 0 \\ b_1 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$\dot{\vec{x}} = A\vec{x} + B\vec{u}$$

$$\vec{y} = C\vec{x}$$

with

$$\vec{x} \in R^3; \qquad \vec{u} \in R^1; \qquad \vec{y} \in R^3.$$

For this linear system two sets of desired feebdback gains are determined. The first set of feedback gains are chosen for the case where the deviation of the state from the desired values are small. These gains are denoted by $K^s = (\, k_1^s \quad k_2^s \quad k_3^s \,)$. A second set of feedback gains are designed for the system's response to a large error in the system's state from the desired values and are denoted by $K^l = (\, k_1^l \quad k_2^l \quad k_3^l \,)$. Both sets of feedback gains must stabilize the system, as the case of an unstable feedback system was not studied. Any method of determination of the feedback gain sets can be used. Other requirements on the eigenvalues of the systems with feedback with both large and small deviations are at the discression of the designer.

These two gain sets form the endpoints for the system's feedback design. These feedback gains are then combined in a nonlinear function, such that for small state errors the effective gain is equal or close to the set K^s. While for large perturbations of the system the feedback gains equal or approach K^l. One method is to use a function $r(\vec{x}_d - x)$, where $0 \leq r(\vec{x}) \leq 1$ with $r(0) = 0$ and $r(\infty) = 1$. There are many, if not an infinite number, of other possible candidate functions. The requirements on the range of the function $r(\vec{x})$ that were stated can be modified. This functional description was chosen to simplify the calculations. The primary requirement is that as the error $\vec{x}_d - \vec{x}$ approaches zero the feedback gain approaches K^s, and as the error gets large the feedback gains approach K^l. A candidate nonlinear function is

$$r(\vec{x}_d - x) = 1 - e^{-\left(\frac{(x_{1d}-x_1)^2}{\sigma_1^2} + \frac{(x_{2d}-x_2)^2}{\sigma_2^2} + \frac{(x_{3d}-x_3)^2}{\sigma_3^2}\right)}.$$

This function is combined with the designed gain matrices to yield

$$K(\vec{x}, \vec{x}_d) = K^s + (K^l - K^s) r(\vec{x}_d - x)$$

for the nonlinear gain function. It can be seen that for errors that are small with respect to the respective σ_i the value of K approaches K^s, with $K = K^s$ for $\vec{x} = \vec{x}_d$. For errors signals that are large, the feedback gain K approaches K^l.

For this particular function, $r(\cdot)$, the parameters σ_i can be tailored to emphasize or deemphasize a certain error signal. For instance, if $\sigma_i = \infty$ the i'th error signal is completely removed from affecting the function $r(\cdot)$.

This function results in the feedback gain of the system being constant on a ellipsiod centered on the origin of the error space. If a particular σ_i is set to infinity, then the ellipsiod becomes an ellipsiodal column centered on the i'th axis.

The linear system with this feedback becomes a nonlinear system. Described by

$$\dot{\vec{e}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a_1 - b_1(k_1^s + (k_1^l - k_1^s)r(\vec{e}) & a_2 - b_1(k_2^s + (k_2^l - k_2^s)r(\vec{e}) & a_3 - b_1(k_3^s + (k_3^l - k_3^s)r(\vec{e}) \end{pmatrix} \vec{e}$$

with

$$\vec{e} = \vec{x}_d - \vec{x}.$$

This formulation allows for the the desired state value, $\vec{x}_d$, to be considered as an input to the system. Since this system is not a nonlinear system that is input affine, due to the function $r(\cdot)$, normal nonlinear controllablility checks can not be performed. But it is easy to see that the system can be driven to any state.

This system is trivially observable, since each state is a separate output. This may not always be the case. In the case that the output vector is not equal to the state vector, the Implicit Function Theorem can be used to determine if the system is locally observable. While no formal proof is presented it is expected that if the original linear system is observable then the linear system with nonlinear feedback is also observable. In order to avoid any observability problem the state vector and the output vector are set equal.

Stability of the feedback system is a concern that must be addressed. Simply because the endpoints of the feedback law result in stable systems does not guarantee an overall stable system. Further examination demonstrates that the system may be unstable for intermediate gain values.

Criteria for stability is shown throught the use of the Second Method of Lyapunov[2]. The Lyapunov function is set to be the energy of the system, both potential and kenetic. This results in a positive definite matrix, M, such that the system's energy is given by $\vec{x}^T M \vec{x}$. The energy is equal to zero only at the state, $\vec{x}_d$. It is easily shown that $\vec{x}_d$ is the only equilibrium point. The matrix A_f is the dynamics matrix of the system with the nonlinear feedback law. The matrix MA_f is negative definite if all of the eigenvalues of the

matrix, A_f, have negative real parts. This requires that all the eigenvalues of the A_f matrix are in the left half plane. Thus, the normal methods for determining stability can be used.

Routh-Horowitz Criteria is used to examine the location of the eigenvalues. The system dynamics matrix with feedback A_f is given by $A - BK(\cdot)$. For this the characteristic equation of the matrix is formed. This is a third order equation in λ, with all but the λ^3 coefficient dependent on the funtion $r(\cdot)$. The first requirement is that all of the coefficients are of the same sign. Thus, all of the coefficients must be greater than zero. This yields for the λ^3 to λ^0 coefficients

$$1 > 0$$

$$-a_3 + b_1(k_3^s + (k_3^l - k_3^s)r(\cdot)) > 0$$

$$-a_2 + b_1(k_2^s + (k_2^l - k_2^s)r(\cdot)) > 0$$

$$-a_1 + b_1(k_1^s + (k_1^l - k_1^s)r(\cdot)) > 0.$$

For the given system, we need to examine the first column of the Routhian array. The λ^3, λ^2 and λ^0 first column elements are the coefficients the respective terms of the characteristic equation. Thus it can be seen that they satisfy the stability requirements. Not only at the endpoints, but for all values of the function $r(\cdot)$ in the range of 0 to 1.

The λ^1 element is given by, and must satisfy

$$-a_2 + b_1(k_2^s + (k_2^l - k_2^s)r(\cdot)) - \frac{-a_1 + b_1(k_1^s + (k_1^l - k_1^s)r(\cdot))}{-a_3 + b_1(k_3^s + (k_3^l - k_3^s)r(\cdot))} > 0.$$

It would seem that the easiest way of checking for stability would be to insert numerical values into this equation. But this must be performed enough times to insure stability in the entire range of $r(\cdot)$. Analysis of the above relation, provides insight into the requirements for stability. Since the demoninator of the second term is greater than zero, both sides of the relation is multiplied by $-a_3 + b_1(k_3^s + (k_3^l - k_3^s)r(\cdot))$. This results only in a change of scale of the polynomial, while the sign remains unchanged. This yields a polynomial relation in the function $r(\cdot)$.

$$b_1^2(k_2^l - k_2^s)(k_3^l - k_3^s)r^2(\cdot) + (-a_2b_1(k_3^l - k_3^s) + b_1^2k_2^s(k_3^l - k_3^s) - a_3b_1(k_2^l - k_2^s) + b_1^2k_3^s(k_2^l - k_2^s) - b_1(k_1^l - k_1^s))r(\cdot) +$$

$$(a_2a_3 - a_2b_1k_3^s - a_3b_1k_2^s + b_1^2k_2^sk_3^s + a_1 - b_1k_1^s) > 0$$

This can be rewritten as

$$\alpha r^2(\cdot) + \beta r(\cdot) + \gamma > 0.$$

Since then endpoints of the gain function by design yield stable systems, and the function $r(\cdot)$ are equal to 0 and 1at the endpoints, it is known that

$$\gamma > 0$$

and

$$\alpha + \beta + \gamma > 0.$$

The question as to if this system is stable for all values of $r(\cdot)$ in the range of 0 to 1 is now addressed by examination of the polynomial of $r(\cdot)$. Let

$$g(r) = \alpha r^2(\cdot) + \beta r(\cdot) + \gamma.$$

Figure 1 shows the four possible plots of $g(r)$ as a function of $r(\cdot)$. Only Case 3 is unstable. All other cases are stable. One method of determination if the system is stable can be performed by locating the roots of $g(r)$, and if the roots are not in the interval of 0 to 1 then the system is stable. There is one case that is not shown, and that is when $\alpha = 0$. This case is must be stable since the locus of $g(r)$ is a straight line, $g(r) = \beta r(\cdot) + \gamma$, since the endpoints are both stable.

There is an additional way of determining stability. This is accomplished by using basic analysis of the function $g(r)$ with an increasing amount of calculations. It can be seen that since $g(r)$ is a quadratic function of $r(\cdot)$, the plot must be a parabola. This parabola can have either a positive or negative second derivative. If the second derivative is negative, which corresponds to Case 4, then the system must be stable. This is equivalent to

$$(k_2^l - k_2^s)(k_3^l - k_3^s) < 0.$$

On the otherhand if the second derivative is positive, then the location of the minimum of $g(r)$ can be found be equating the first derivative to zero. This is equivalent to

$$r_{min} = \frac{\beta}{-2\alpha}.$$

If $r_{min} < 0$ or $r_{min} > 1$ then the system is stable. While if r_{min} is in the range of 0 to 1 then $g(r_{min})$ must be evaluated. The system is stable if $g(r_{min}) > 0$.

It should be noted that if α and β are both positive which corresponds to Case 1, and this is a stable system.

In the case of a longitudinal controller for an aircraft, using the short period approximation, the gains are such that $\alpha < 0$. Thus, always yielding a stable system for this type of feedback scheme. This is due to the sign convention used on aircraft dynamics.

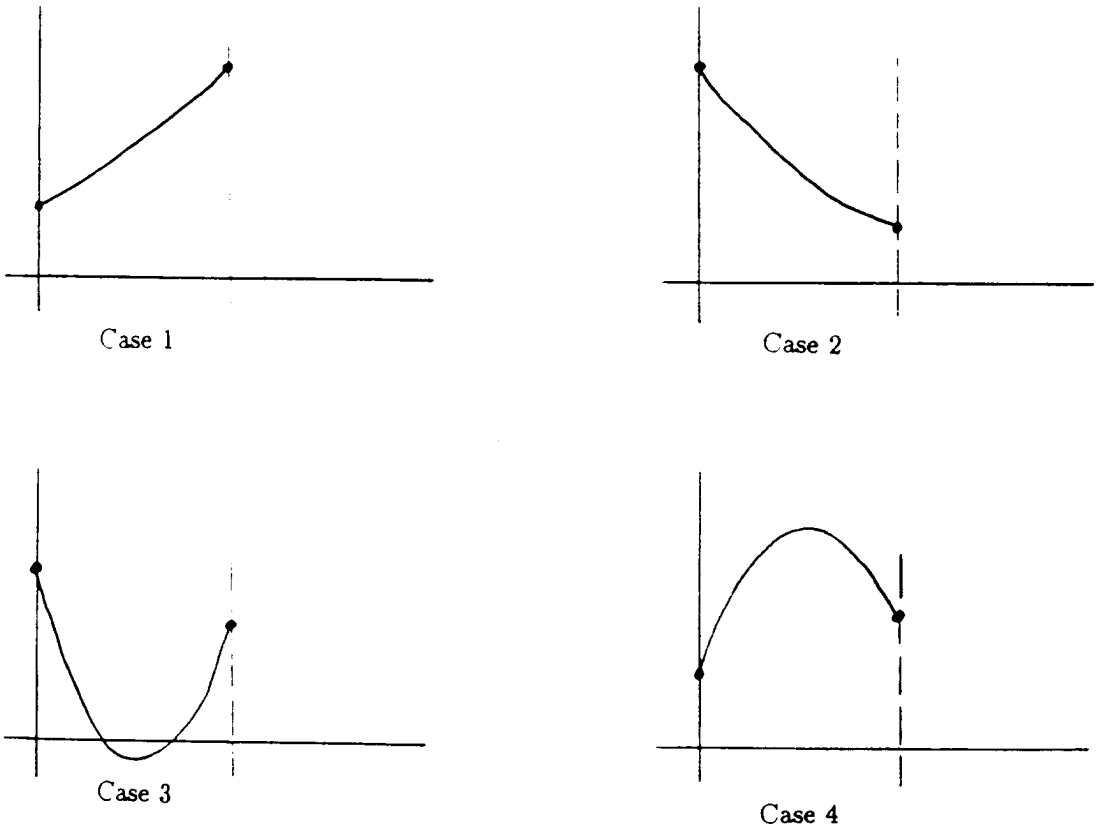


Figure 1. Possible plots of $g(r)$.

This has shown that a nonlinear feedback methodology is feasible. The nonlinear feedback offers a way of changing gains in a linear system to achieve advantages in the time response of the system. The nonliear feedback has been shown to be stable under certain circumstances. The system remains both controllable and observable.

EXAMPLE: F-15 LONGITUDINAL CONTROL SYSTEM.

A linear dynamic model of an F-15 flying as 5000 ft at a speed of 877.68 fps[3] was generated using a Short Period Approximation to the longitudinal motion of the aircraft[4]. The only input to the system was the elevator. The elevator was driven by a servo. The servo was considered to be of first order with a corner frequency of 6.3662 Hz. This yielded a 3-dimensional system with the states being: Angle of attack, α. Pitch rate, q and Elevator angle, δ_e. The input was the commanded elevator angle, δ_e^c. The output will be the state of the system.

$$A = \begin{pmatrix} -0.9074 & 1.0000 & 0.0000 \\ -4.1420 & -0.7313 & -7.1414 \\ 0.0000 & 0.0000 & -20.000 \end{pmatrix} \quad B = \begin{pmatrix} 0.000 \\ 0.000 \\ 20.000 \end{pmatrix}$$

$$C = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}$$

$$\vec{x}^T = (\alpha \quad q \quad \delta_e)$$

$$\vec{u} = \delta_e^c$$

The eigenvalues for this system are at -20.0000 and $-0.8193 \pm 2.0333i$.

The goal is to design two LQR feedback systems. One for small errors and the other for large errors. Both sytems assigned a cost to deviations in the outputs α and q. While there was no cost assigned to δ_e. The costs on the output deviations were the same in both designs. The output cost matrix, Q is given by the following.

$$Q = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{pmatrix}$$

The input cost was varied. For the small deviation feedback the input cost, R, was set equal to 1.0. The feedback gain matrix corresponding to this cost arrangement was call $K_{1.0}$. While for the large deviation system the input cost was set at 0.2, with the resulting gain matrix denoted by $K_{0.2}$.

$$K_{1.0} = (0.1411 \quad -0.8733 \quad 0.2742)$$

$$K_{0.2} = (\ -0.1086 \quad -2.1146 \quad 0.5843\ )$$

The logic for the feedback design was that it was desired that the system's input was freer in the large disturbance case. This would give a system that would quickly slew to a region close to the desired state, but as the system approached that state the slewing would slow down. This could be avantageous for the aircraft.

For most of the simulations run, both σ_1 and σ_2 were set equal to 0.1. The state δ_e effect on the nonlinear feedback was eliminated by setting σ_3 to ∞. The was done so that errors in α and q would be the driving force in increasing the system's gain to speed up the response. There were simulations that were run with $\sigma_2 = 0.05$.

This results of the simulations are presented in Figures 2-8. Figures 2-5 are from simulations with the inital state $\vec{x}^T(0) = (\ 0.0 \quad 0.2 \quad 0.0\ )$. Figure 2 shows the unaugmented dynamics of the F-15 aircraft with the stick-fixed. Figure 3 is the time response of the system using the feedback matrix, $K_{1.0}$. Figure 4 is similar to Figure 3 but using $K_{0.2}$. Figure 5 demonstrates the blending of the two systems with nonlinear feedback. Figures 6-8 show the response to the LQR system using $K_{1.0}$, the nonlinear feedback system with $sigma_2 = 0.1$ and $\sigma_2 = 0.05$ to an input doublet in the commanded pitch rate q^{com}. This was done by setting $\vec{x}_d = (\ 0.0 \quad 0.2 \quad 0.0\ )$ from time equal 0 to 1 seconds, then giving the negative of $\vec{x}_d$ for the time between 1 and 2 seconds.

It can be seen from the plots that the nonlinear feedback exhibits the dynamics of both linear feedback gain matrices. At large errors the fast slew rate of the $K_{0.2}$ system can be seen. While for small errors, the smoother response of the $K_{1.0}$ system can be seen. It must be noted that in the simulations of the nonlinear feedback system, the change in gain is a smooth function. Figure 9 shows the function $1 - r(\cdot)$ and δ_e^c corresponding to Figure 5. This type of control law exhibits a smooth overall response and could advantageous to apply for certain applications.

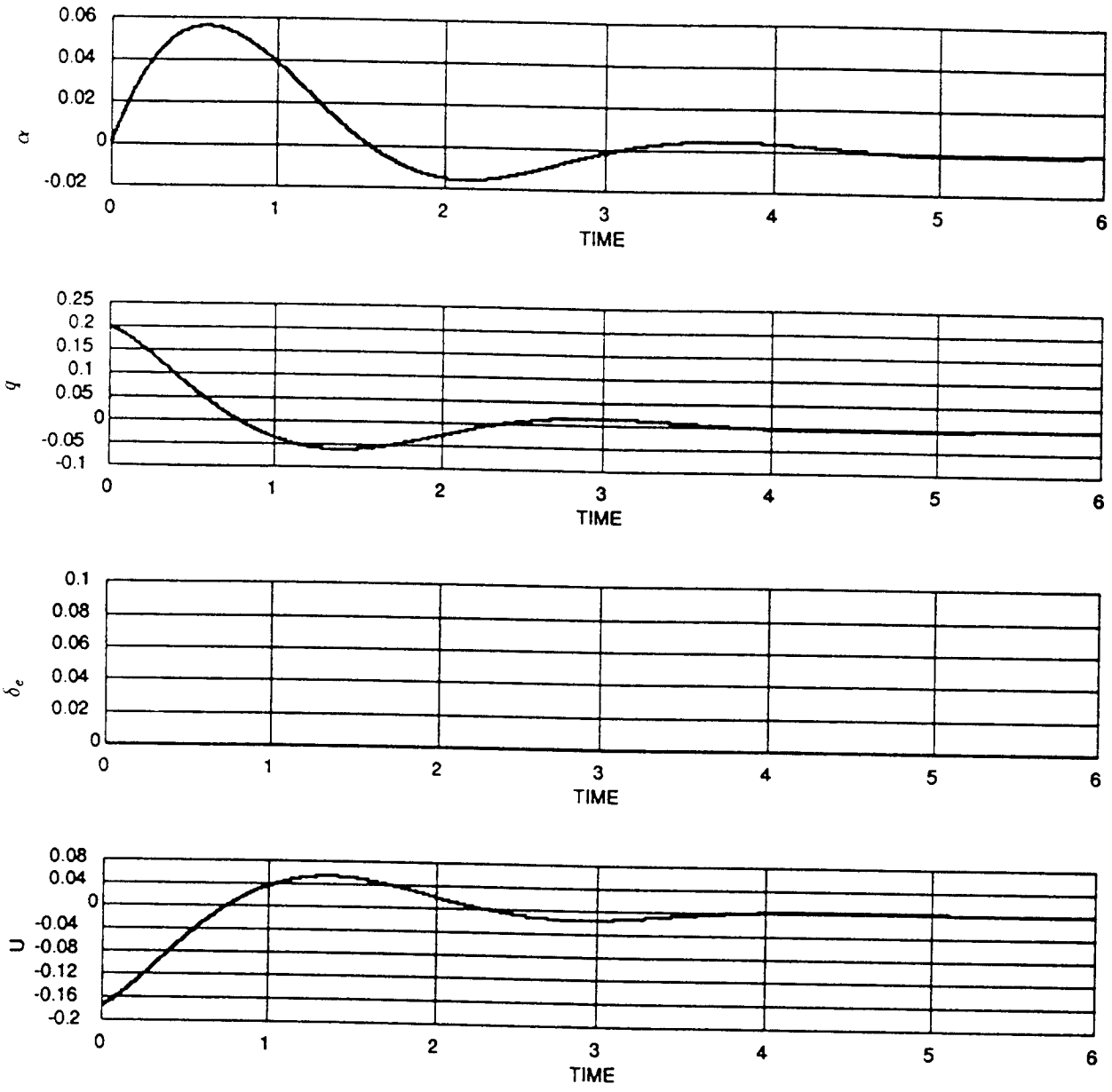


Figure 2. Unaugmented dynamics of the F-15 aircraft with the stick-fixed.

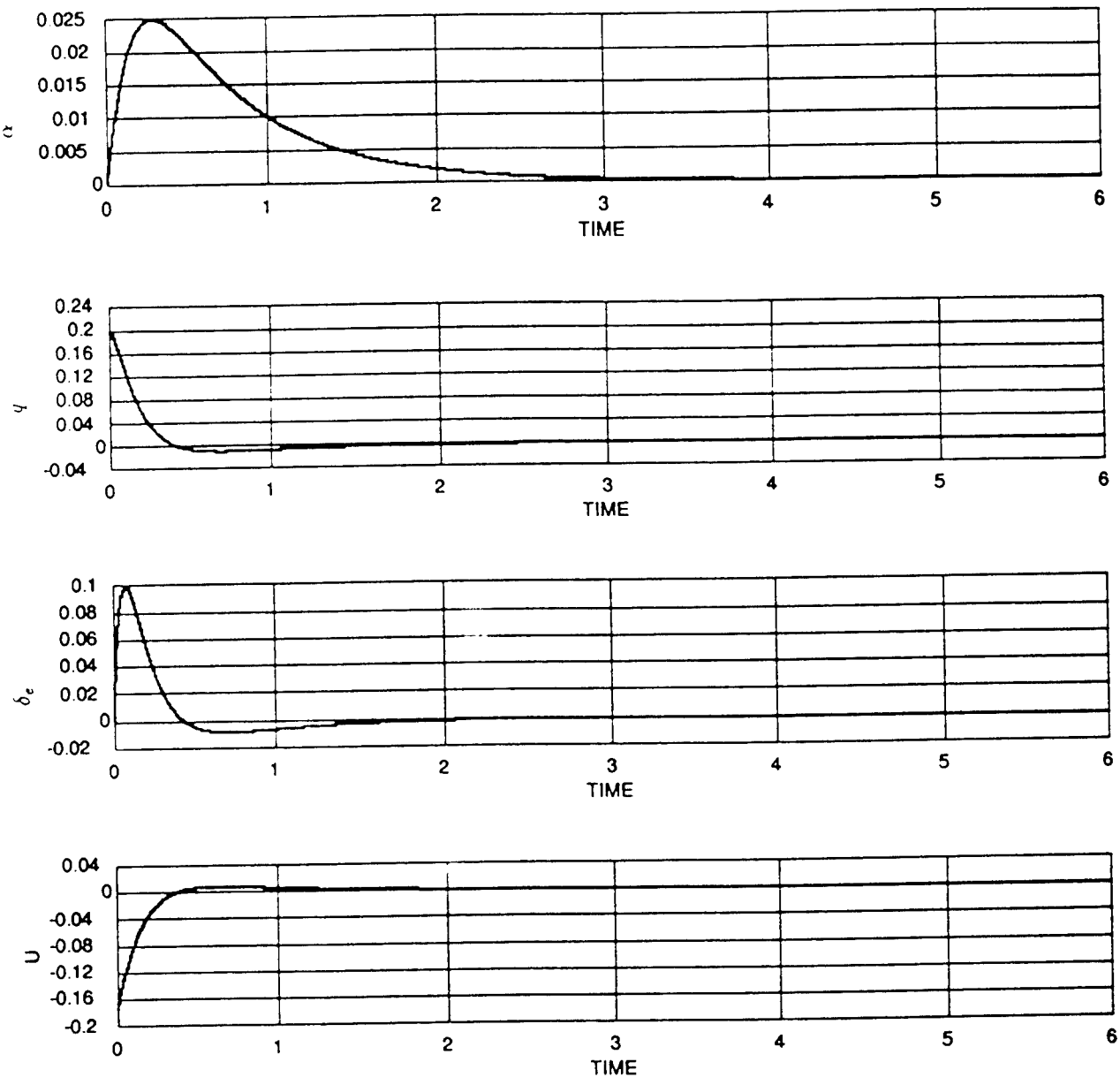


Figure 3. F-15 aircraft with the feedback matrix K_{10}.

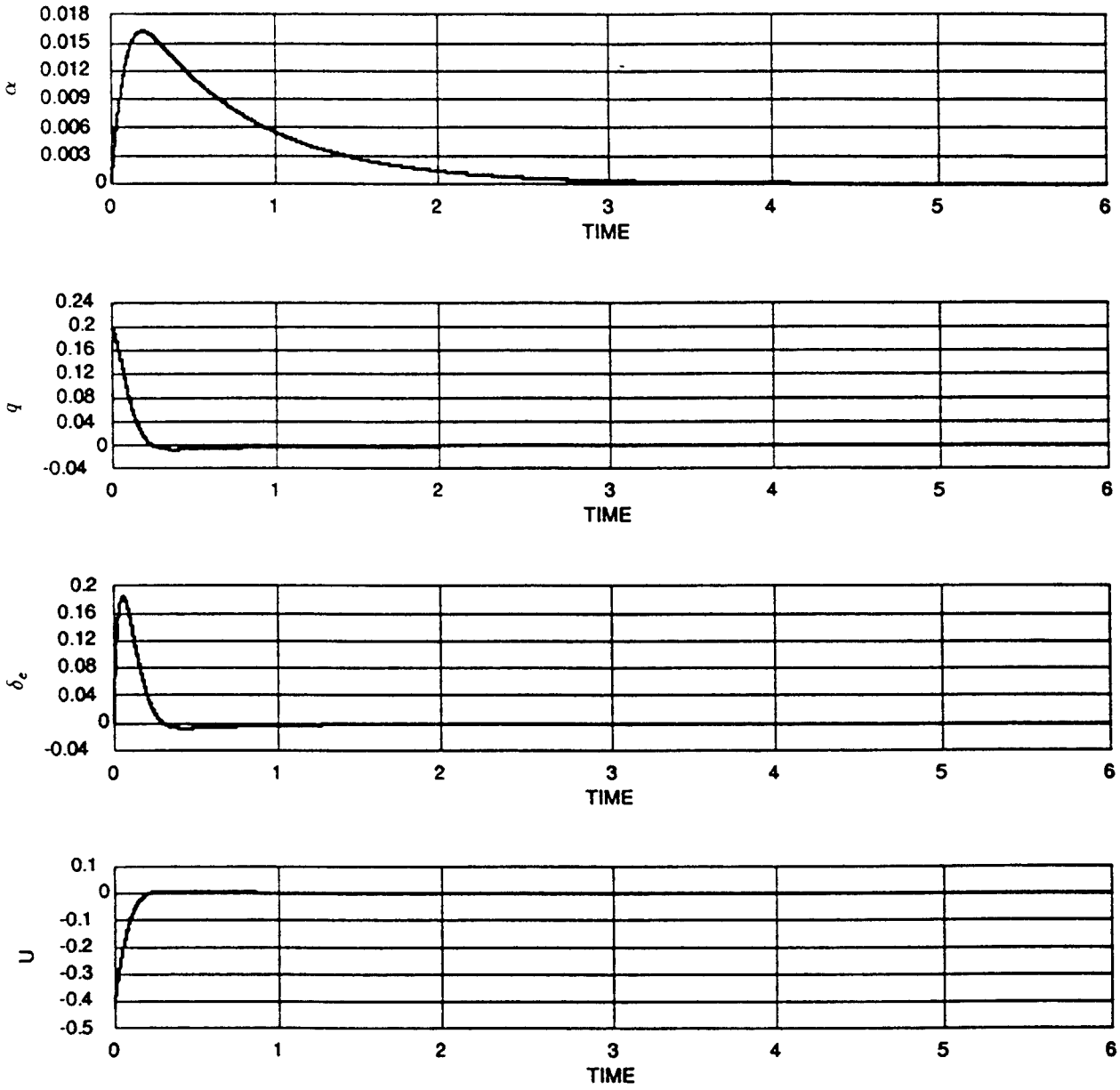


Figure 4. F-15 aircraft with the feedback matrix $K_{0.2}$.

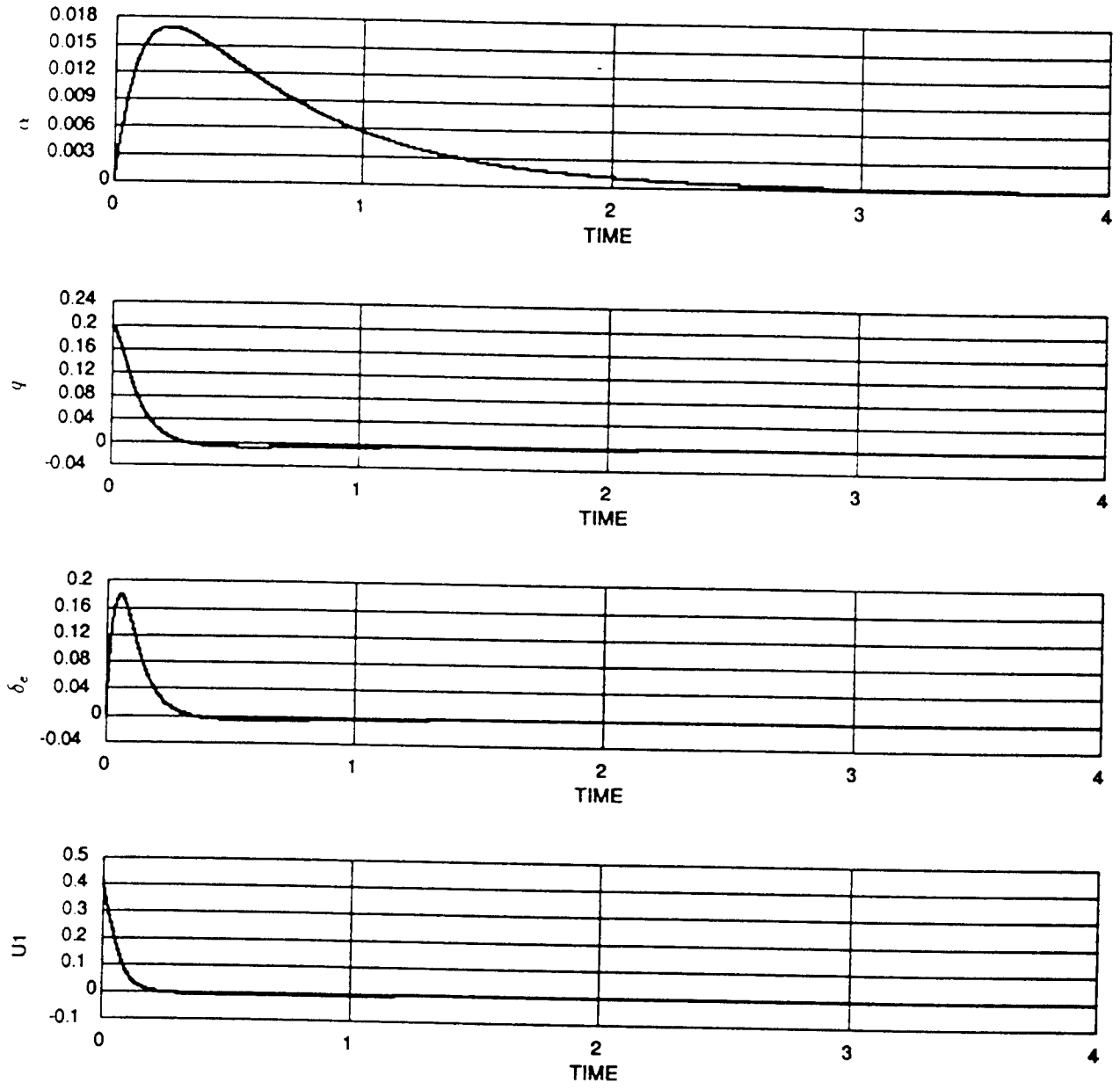


Figure 5. F-15 aircraft with nonlinear feedback.

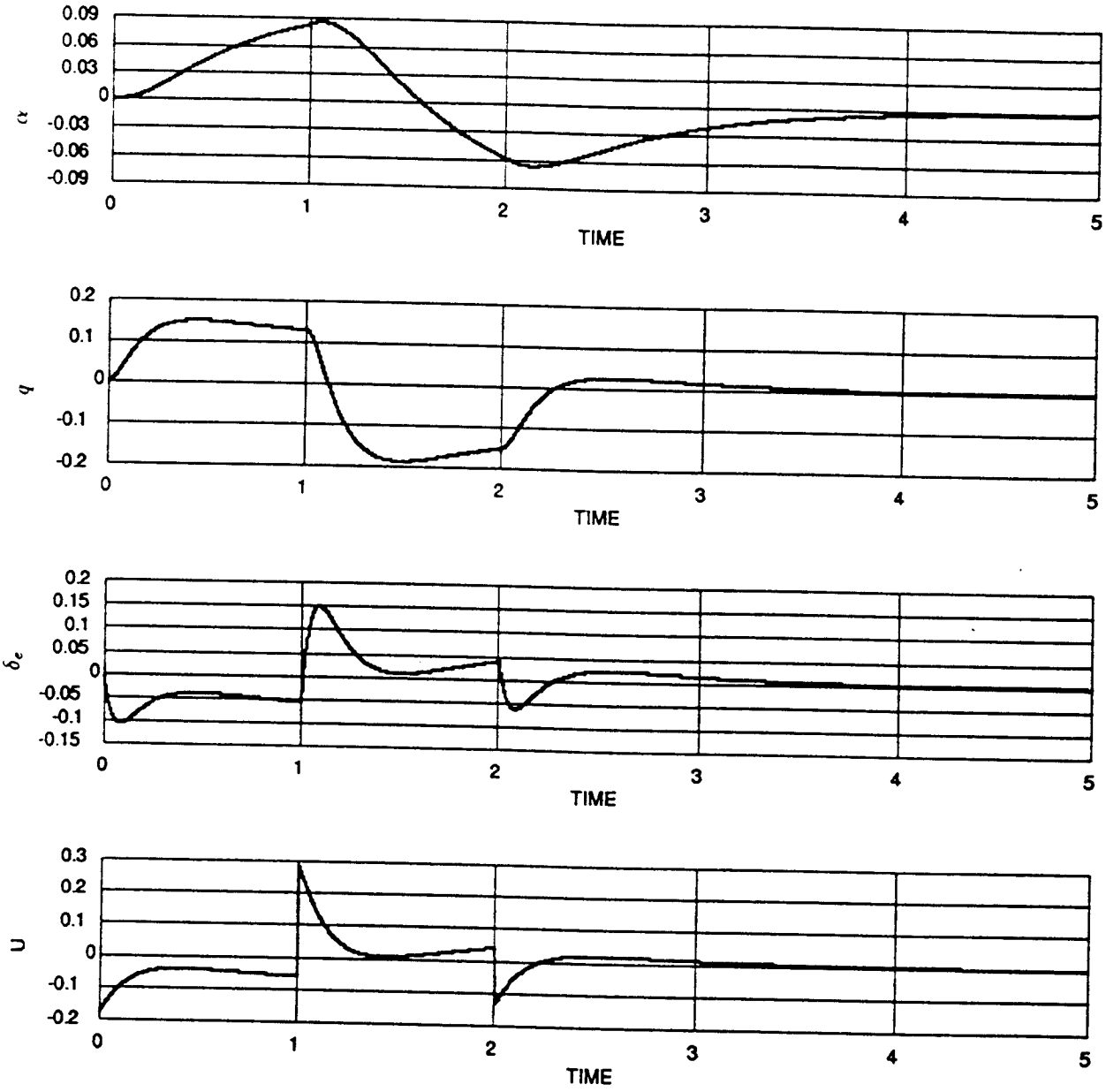


Figure 6. Doublet response with the linear feedback matrix $K_{1.0}$.

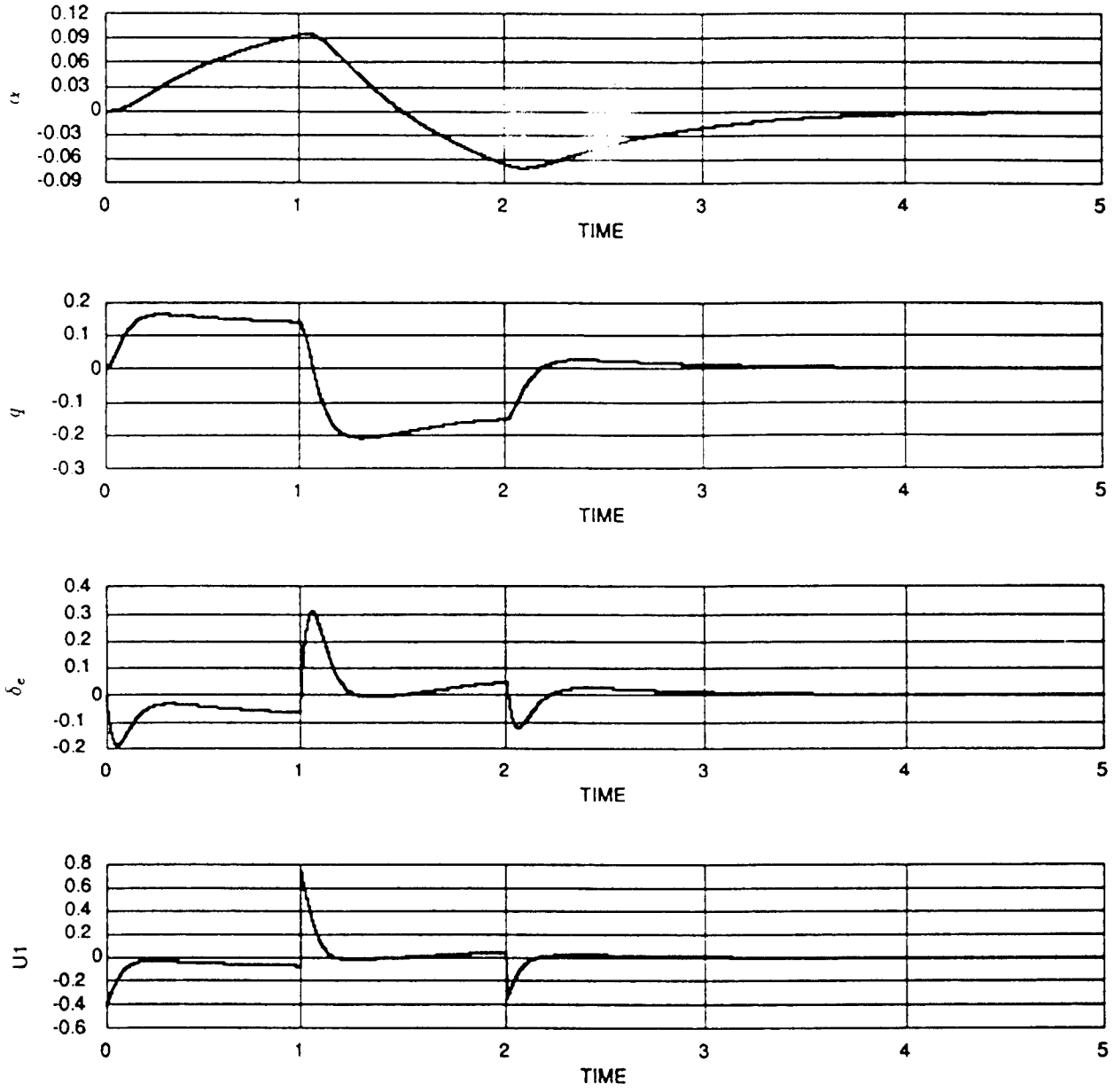


Figure 7. Doublet response with nonlinear feedback, $\sigma_2 = 0.1$.

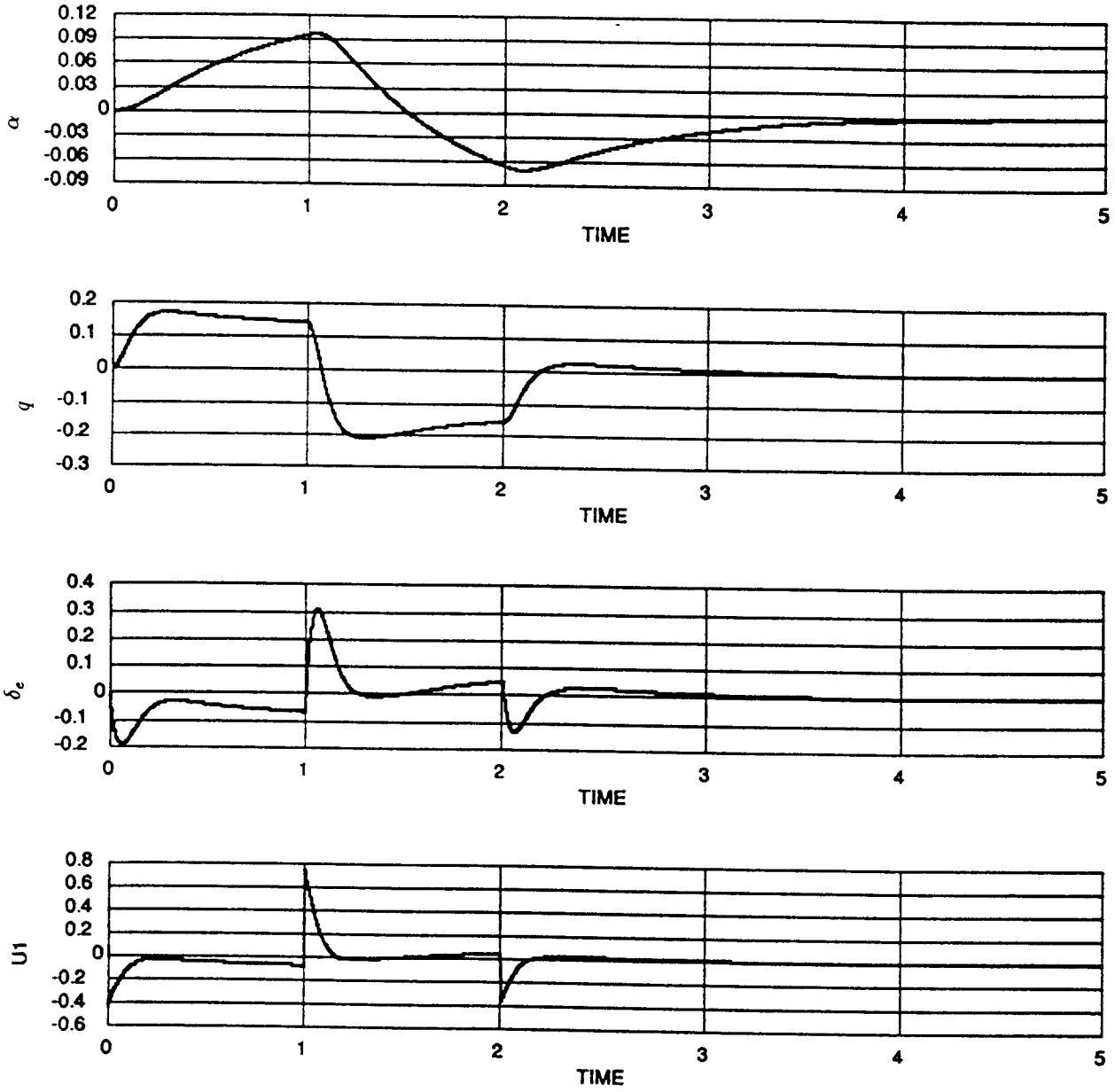


Figure 8. Doublet response with nonlinear feedback, $\sigma_2 = 0.05$.

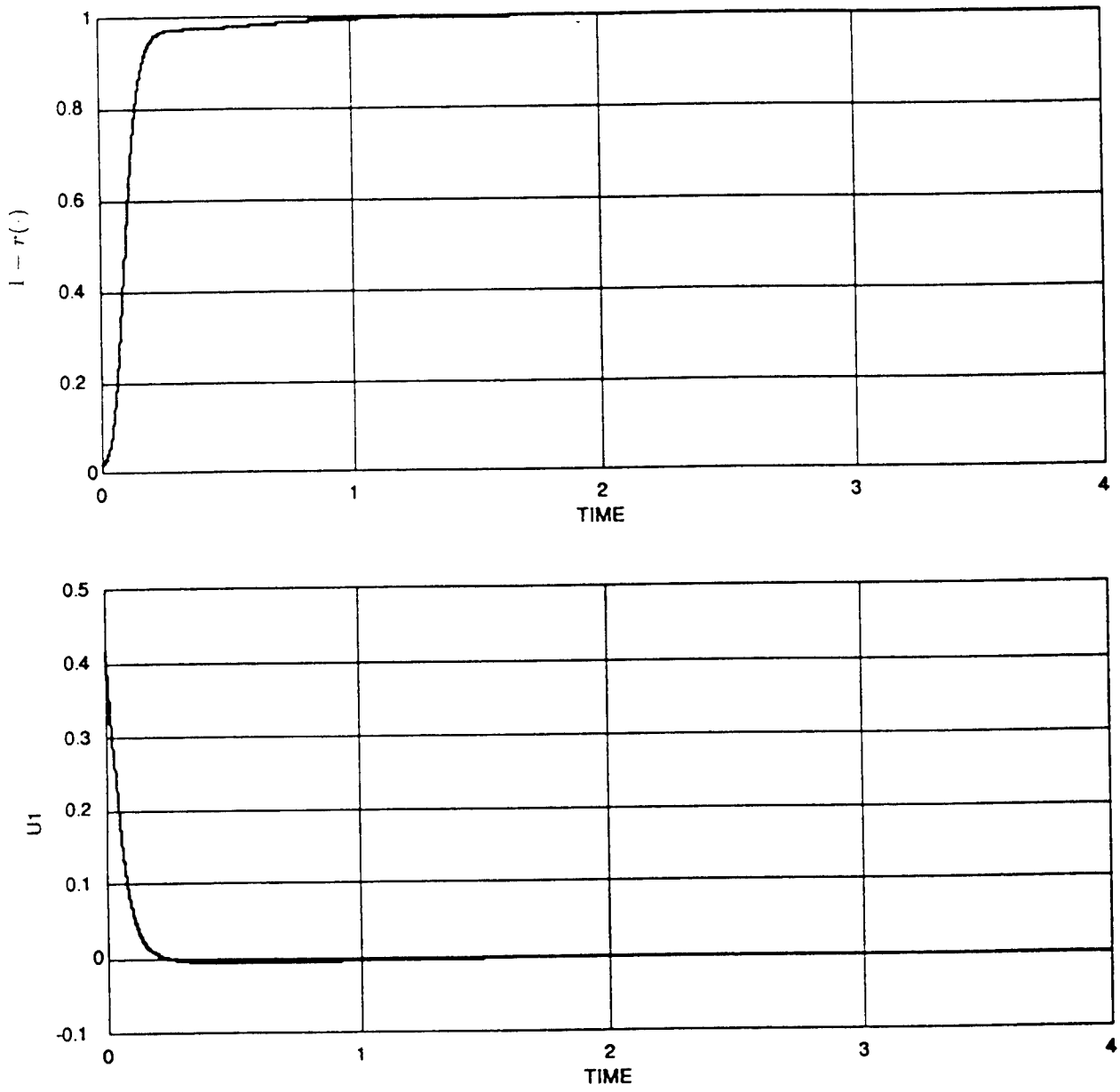


Figure 9. The function $1 - r(\cdot)$ and δ_ϵ^c corresponding to Figure 5.

FAILED APPROACHES

The method and example detailed above was not the most desired method of implementing a nonlinear feedback design on a linear system. The main attempt was to aimed at transforming the linear system with nonlinear feedback to a linear system. In this technique the nonlinear control system in the $\vec{x}$ state space would be transformed into a $\vec{z}$ state space. In the $\vec{z}$-space the system dynamics, to include feedback, would be described by a linear system. This does necessitate that the $\vec{x}$-space to $\vec{z}$-space transformation would be a nonlinear mapping. This mapping could be generated from either the Lie Controllability Algebra or the Observability Algebra. This would have allowed the system with nonlinear feedback to be designed, which would be compatible with the transformation technique, with stardard linear design methods in the transformed space. The primary complication of this method was the need to know the $\vec{z}$-space to $\vec{x}$-space transform, the inverse mapping. Even for relatively simple systems this inverse map could become quite involved. MACSYMA was required to calculate these inverse maps. These maps were generated for a couple of nonlinear feedback functions, but the resulting $\vec{z}$-space systems were not linear. This problem lead to the discovery of a misprint in the theorem pertaining to the transformation[2]. While the controllability or observability algebra had to be of full rank, the system was controllable or observable, higher orders of the algebras components were required to be not only combinations of the first n components but the coefficients for this must also be constants. No nonlinear feedback laws that were applied to linear systems could be found to satisfy this requirement. While at the same time it could not be proved that there were no nonlinear feedback laws that could satisfy these requirements. Considerable time was spent on the method, as even simple systems became extremely involved and only a few tries were made. Additional time was unavailible to find if there were any nonliear feedback functions that would satisfy the requirements or if none would. This is to be left to later work.

CONCLUSIONS

The use of nonlinear feedback in the control of systems with linear dynamics does provide advantages to the designer. The control gains can be varied with the as a function of the error signal. This allows for enhanced system response to large disturbances. The system can be shown to be stable. The transistion

between outer and inner gain matrices is smooth. The added complexity in the design and implementation is minimal.

It is desirable to determine a more general approach using modern nonlinear control methods. As mentioned time was unavailible to accomplish this objective.

ACKNOWLEDGEMENTS

I thank Capt. Stuart Sheldon and Mr. Thomas Molnar of the Control System Development Branch at Wright Laboratory for their help and assistance on this research. This research was sponsored by the Air Force Office of Scientific Research.

REFERENCES

1.Hall, C. E., "Enhancement of the Time Response on Linear Control Systems via Fuzzy Logic and Nonliear Control", Air Force Office of Scientific Research, 1992.

2. Nijemeier, H. and van der Schaft, A., **Nonlinear Dynamical Control Systems**. Springer-Verlag, New York ©1990.

3. Blakelock, J. H., **Automatic control of Aircraft and Missiles**. Second EditionsJohn Wiley and Sons, New York, ©1991.

4. McLean, **Automatic Flight Control systems**. Prentice Hall International, Ltd. Englewood Cliffs, NJ ©1990.

OPTICAL TECHNIQUE FOR MEASURING TIRE DEFORMATION AND STRAINS

Paul P. Lin
Associate Professor
Mechanical Engineering Department
Cleveland State University
Euclid Avenue at E. 24th St.
Cleveland, OH 44115

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

December, 1993

OPTICAL TECHNIQUE FOR MEASURING TIRE DEFORMATION AND STRAINS

Paul P. Lin
Associate Professor
Mechanical Engineering Department
Cleveland State University
Euclid Avenue at E. 24th St.
Cleveland, OH 44115

Abstract

The main objective of this research was to investigate the feasibility of applying an optical technique called fringe projection to quantifying aircraft tire deformation and strains. Three different types of tires in static and dynamic conditions, subjected to different amount of tire deflections, were tested. The experimental results indicate that the proposed measuring system and the optical technique were capable of measuring tire deformation and strains. Some results of data analyses are included in this report. It was found that the key to more accurate three dimensional geometry determination is to keep track of the geometric change of the reference point on a tire when subjected to loading. Finally, the conclusions for this work are made, and the future work is recommended.

OPTICAL TECHNIQUE FOR MEASURING TIRE DEFORMATION AND STRAINS

Paul P. Lin

Introduction

In non-contact measurement, several optical techniques are available. laser ranging can yield a dense set of depth values with which surface structure can be obtained through surface fitting or approximation (Vemuri and Aggarwal, 1984). This technique, however, is usually slow and expensive. Stereo vision utilizes the disparity between the projected positions of a point in two images to infer the depth of this point (Marr and Poggio, 1976). But, the correspondence between points in the stereo images is difficult to establish, and the computation is sensitive to errors introduced in digitization and camera calibration. The well known Projection Moire technique uses a white light or laser light source and two gratings of the same pitch (one in front of light and the other one in front of camera) to generate Moire interference patterns. The image is recorded in a single CCD camera, in lieu of two cameras used in stereo vision. Another very similar technique, Shadow Moire, uses only one grating near the object to generate the Moire patterns. The Moire contours thus obtained, however, do not make a difference between peaks and pits unless prior information or additional algorithms are applied. Furthermore, this technique is very sensitive to vibration due to the necessity of superimposing two images. The most accurate optical technique available today is phase-shifting interferometry. It takes time to make three or four consecutive phase shifts to form interferograms, and therefore cannot be used for measurement of dynamic motion. This technique, by nature is also very sensitive to vibration. The proposed fringe projection technique (Lin and Parvin, 1990; Lin, et. al.,1991) uses a single light source and only one grating in front of light projector to generate optical fringes. No image superposition is required. In comparison with the Moire or phase-shifting technique, the fringe projection technique is less sensitive to vibration and much more computationally efficient.

In this report, the measuring system and methodology are described, and the results of preliminary data analyses are discussed. The limitation of the

measuring system is addressed, and the recommendations for improving the measurement accuracy of this system are included.

Methodology

The measuring system consists of

(1) Optical equipment: White light projector, grating, optical rails, laser diode and CCD camera, etc.

(2) Imaging equipment: Frame grabber and image data acquisition software.

(3) Recording equipment: High resolution super VHS video recorder (VCR).

(4) Computing equipment: 486-based 66 MHZ micro-computer and High resolution RGB monitor.

The light produced by the white light projector passes through the grating of two directional grid and illuminate the tire surface (see Figure 1). The image captured by the CCD camera is recorded in a super VHS video recorder, and then sent to the frame grabber where the image data are digitized and processed. The digital image is then displayed in a high resolution RGB monitor. The frame grabber and CCD camera both have the same resolution of 512 by 480 pixels. Although the highest shutter speed available from the CCD camera is 100 micro seconds, a lower speed at 250 micro-seconds was used in order for the camera to receive sufficient light. The camera frame rate is 1/30 seconds, so are the VCR and frame grabber.

The principle of this optical measurement is based on the curvature changes of fringe centers and the spacing between them, which in turn, translates into the change of surface height. It is necessary to specify the location of the reference point on the tire, and a reference plane (xy plane) passing through this point and perpendicular to the viewing direction. It is worthwhile to note that when a tire is loaded, not only the location (x and y components) of the reference point changes, the height (z component) of the point changes as well.

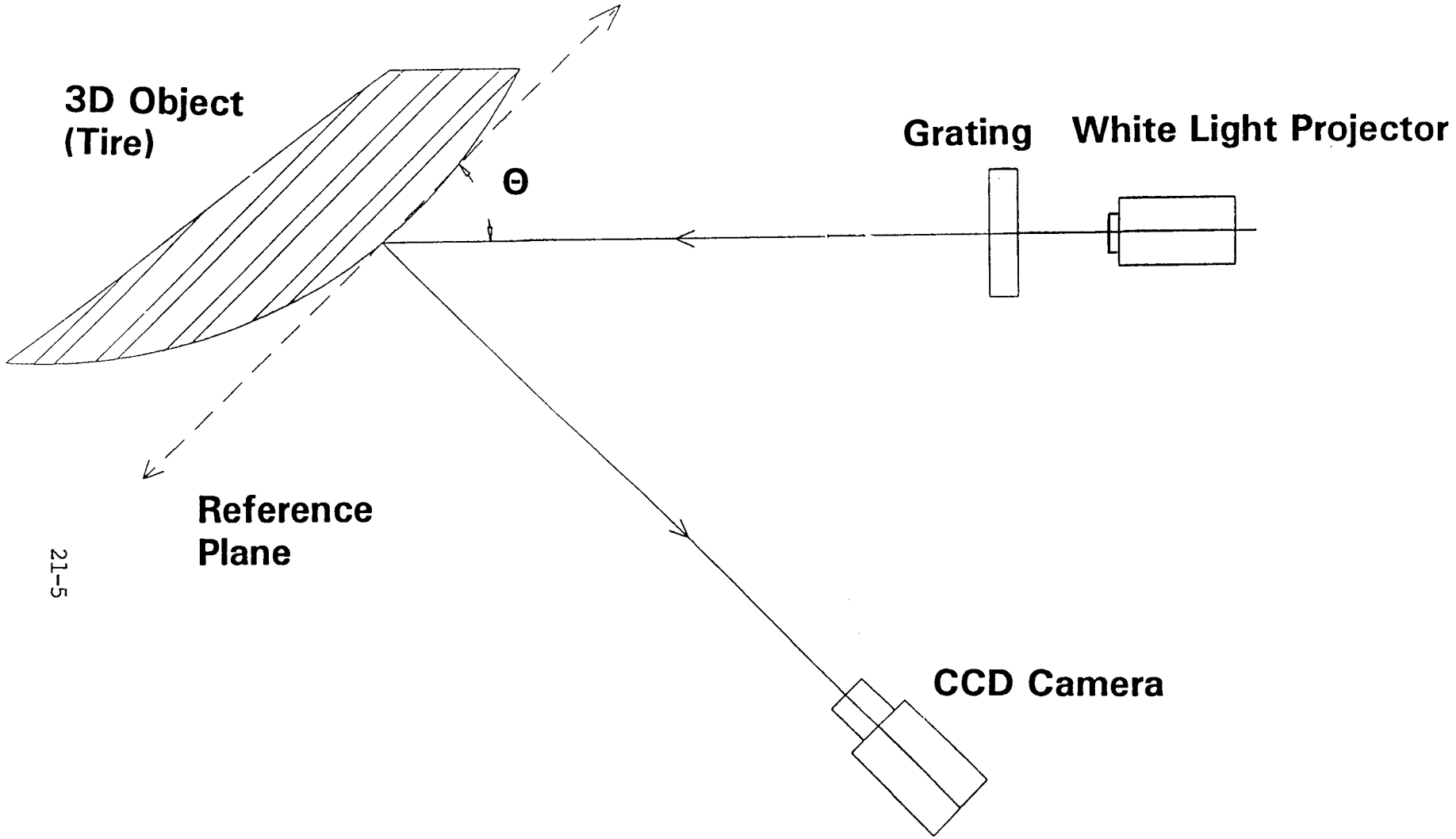


Fig.1 Experimental Arrangement for 3D Geometry Measurement

The acquired images were filtered and analyzed with and without the use of grating. Without grating, some points of interest in the tire were marked with white dots. The displacements of these points were traced when the tire was subjected to different loads. These displacements, however, are limited to what the camera actually sees in two dimensions. With grating, two directional optical fringes were generated, and the fringe centers were accurately detected. In the follow-up project, the image containing many fringe centers will be scanned from top to bottom and left to right (see Figure 2). Finally, the in-house developed computing algorithm will be applied to determining the three dimensional geometry of tire deformation. In addition, strains can be calculated by detecting the length changes of certain line segments.

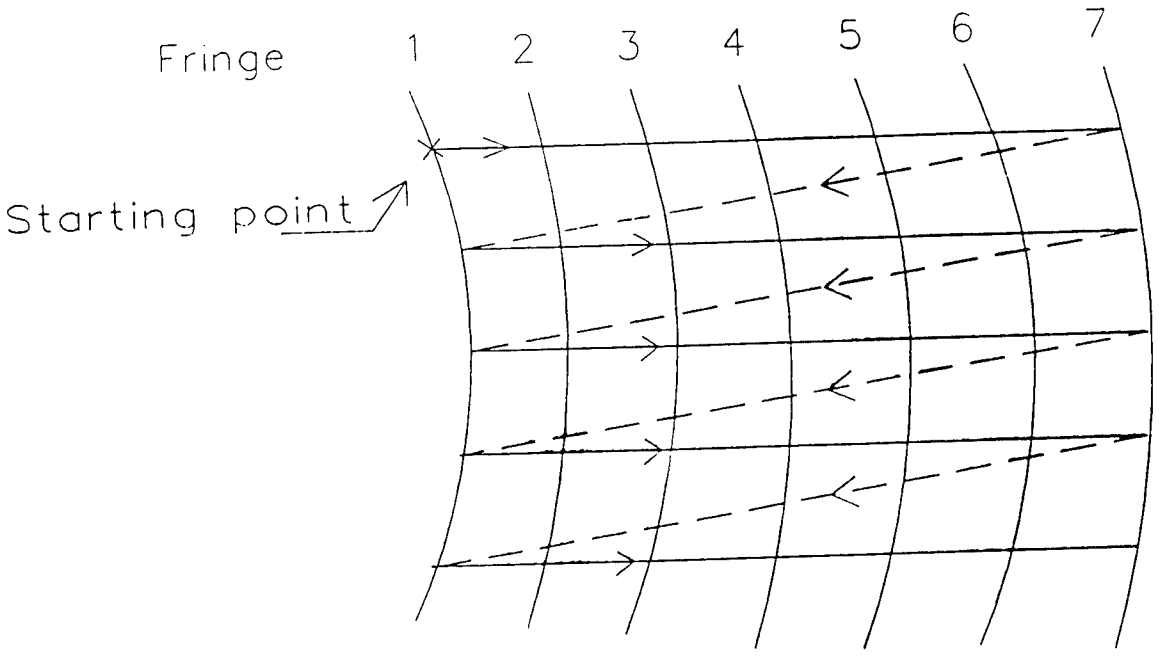


Fig. 2 Image Data Scanning Process

Results and Discussion

Several weeks were spent in conducting tests on three different aircraft tires: F-111, F-16 and KC-135. Due to time constraint, only the results of two dimensional tire deformation analyses are reported. The detailed three dimensional deformation and strains will be analyzed later. The results for the three different tires are summarized as follows:

(1) F-111 Tire:

The displacements subjected to different loading conditions are summarized in TABLE I. With flat plate, the displacements are generally at least 30% more than those with flywheel. It is interesting to note that there is no substantial difference in terms of displacement between "rated pressure, corrected load" and "corrected pressure, rated load".

(2) F-16 Tire:

The displacements subjected to different loading conditions are summarized in TABLE II and TABLE III. Although this tire is much smaller than F-111 tire (about half of F-111), the directions of displacements when subjected to an applied load exhibit about the same pattern as the F-111 tire. However, the displacement magnitudes are considerably smaller (about from 50 to 100% smaller than those of F-111 tire). This can be understood that a small tire is very much confined by the tire's wheel, and thus there is not much room for the tire to deform in the tire surface plane. There is a possibility that a small tire might exhibit a relatively large out-of-plane deformation. It is true that the magnitude of deformation has to do with the tire structure as well.

(3) KC-135 Tire:

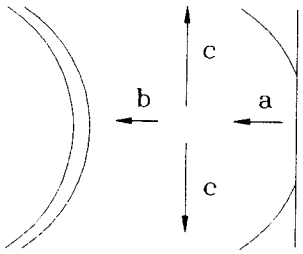
Figure 3 shows the digitized image of KC-135 tire with white dots marked for the purpose of tracing the displacements of the points of interest. Figures 4-6 show the same tire with the projected gird pattern subjected to loading. In

addition to tracing the displacements of some points as done for F-111 and F-16, the displacements of points along 0°, 30°, 60°, 90°, 120°, 150° and 180° lines were also under study (see Figure 7). It was to investigate how much the displacements of points at different angular positions change as a tire rotates. Some interesting results are shown in TABLE IV and TABLE V. It was found that the highest magnitude of displacement always occurs along the 30° line. Furthermore, with flat plate the displacement magnitudes are generally 30 to 50 % more in the region of ±90° when compared to those with flywheel. Beyond the region the difference of displacement magnitude between flat plate and flywheel becomes much less noticeable.

TABLE I

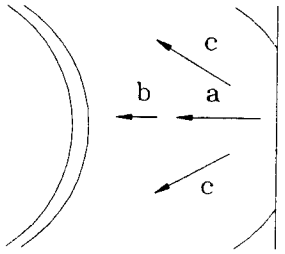
F - 111 TIRE with FLAT PLATE
(static load)

(A) 10 % deflection



c >> a ≃ b

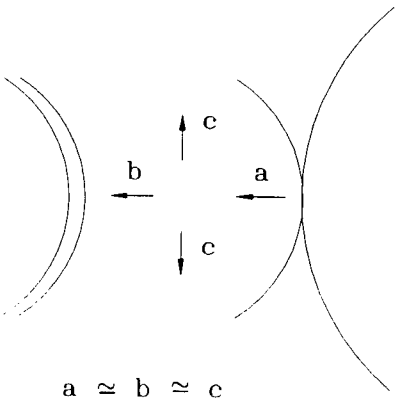
(B) 30 % deflection



a ≃ 1.1c & a >> b

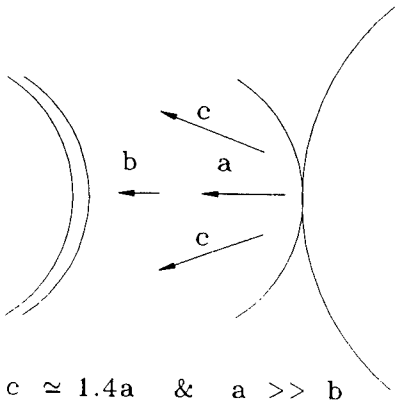
F - 111 TIRE with FLYWHEEL
(static load)

(A) 10 % deflection
: Rated pressure, corrected load



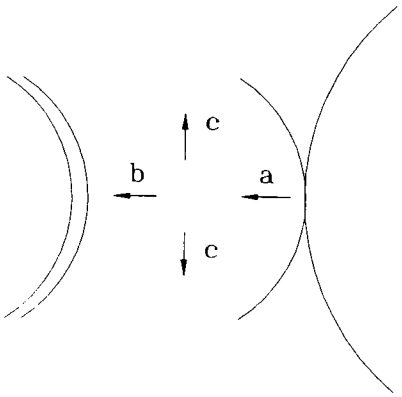
a ≃ b ≃ c

(B) 30 % deflection
: Rated pressure, corrected load



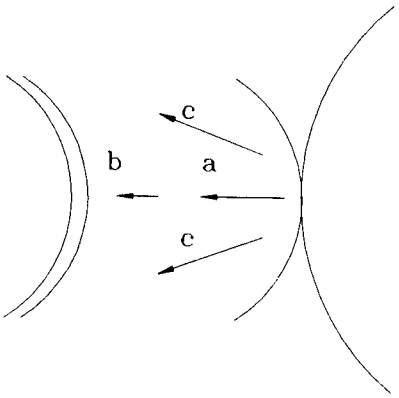
c ≃ 1.4a & a >> b

: Corrected pressure, rated load



a ≃ b ≃ c

: Corrected pressure, rated load

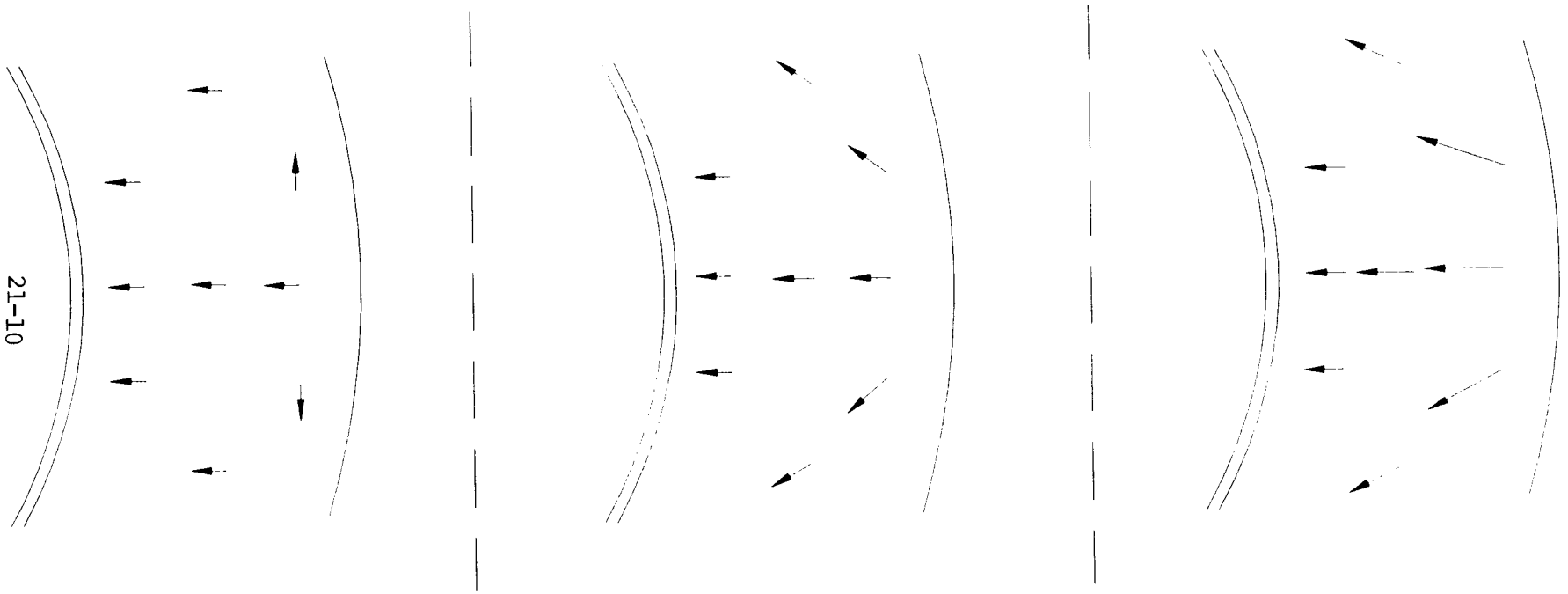


c ≃ 1.5a & a >> b

T E II

F - 16 TIRE :th FLAT PLATE (static load)

(A) 10% deflection (B) 30% deflection (C) 40% deflection



NOTE : The displacement magnitude of any point shown here is proportional to the lenfth of each vector.

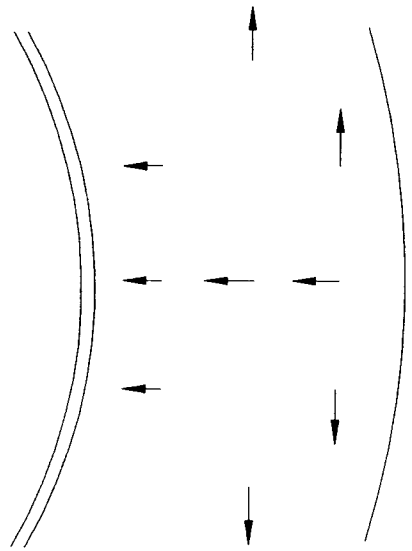
TABLE III

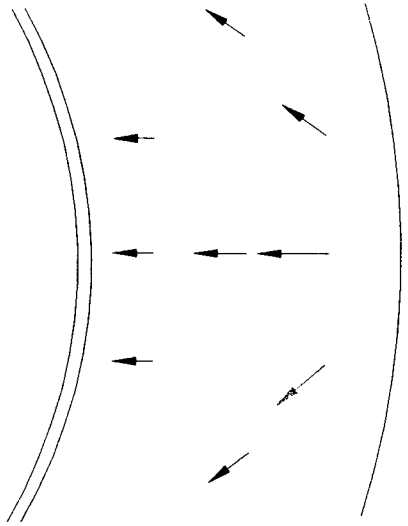
F - 16 TIRE with FLYWHEEL (static load)

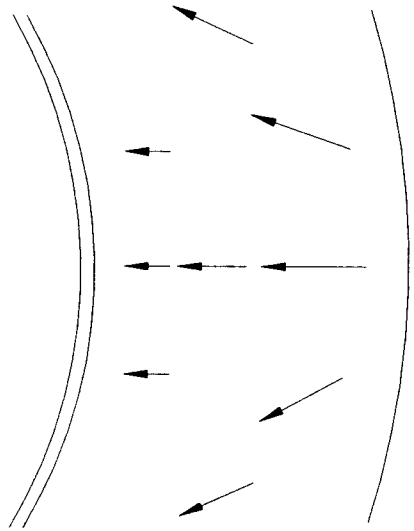
(A) 10% deflection
: Rated pressure,
Corrected load

(B) 30% deflection

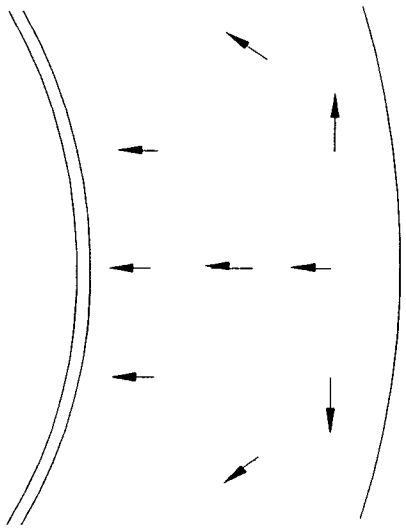
(C) 40% deflection

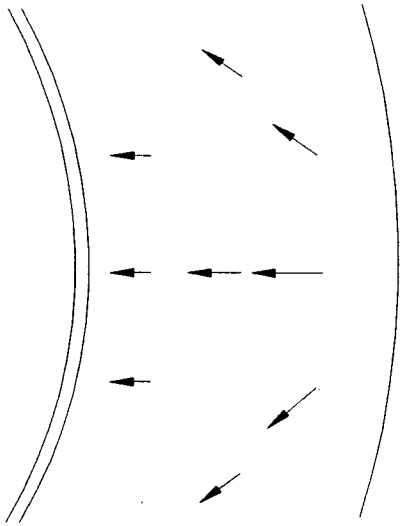


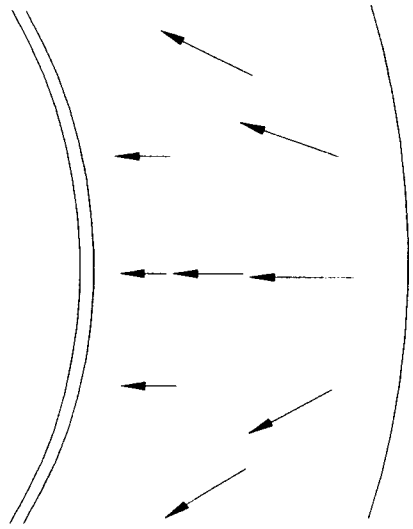




: Corrected pressure,
Rated load







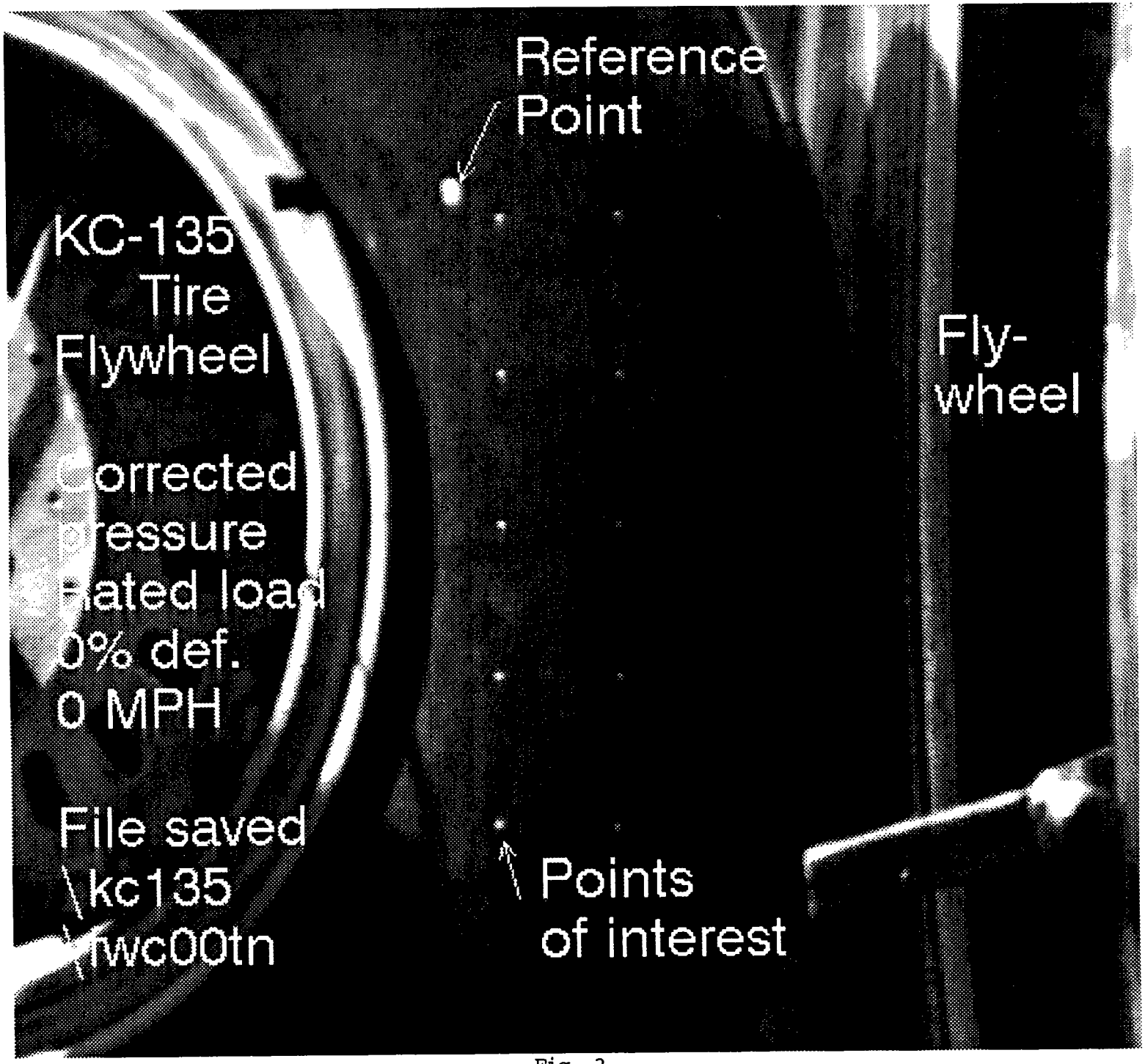

Reference
Point
KC-135
Tire
Flywheel
Corrected
Pressure
Rated load
0% def.
0 MPH
File saved
\kc135
\fwc00tn
Points
of interest
Fly-
wheel

Fig. 3


KC-135 Tire
Flat plate
Deflection:
0 %
Speed:
0 MPH
File saved:
\kc135
\fpr00n

Fig.4

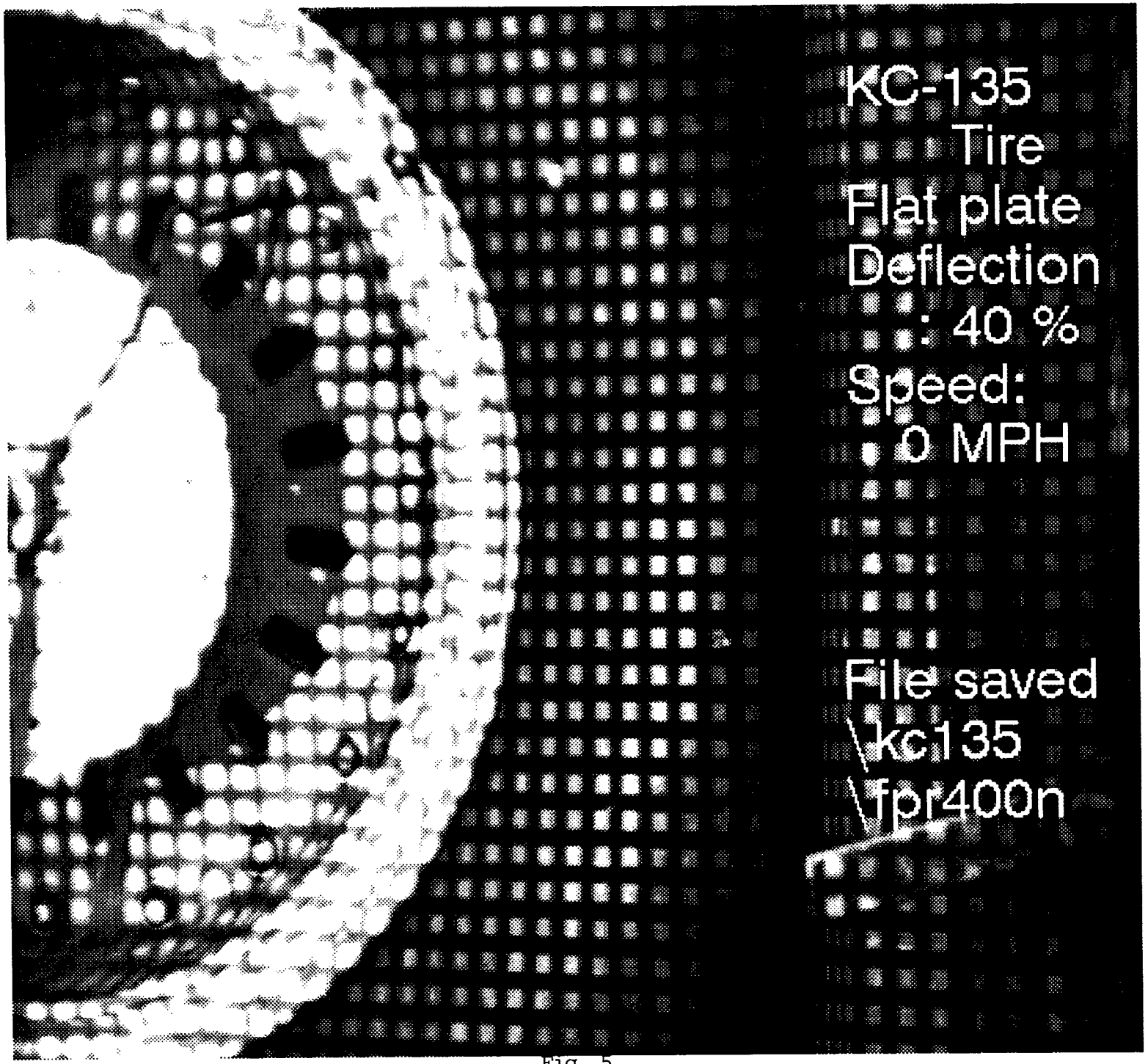

KC-135
Tire
Flat plate
Deflection
: 40 %
Speed:
0 MPH
File saved
kc135
fpr400n

Fig. 5

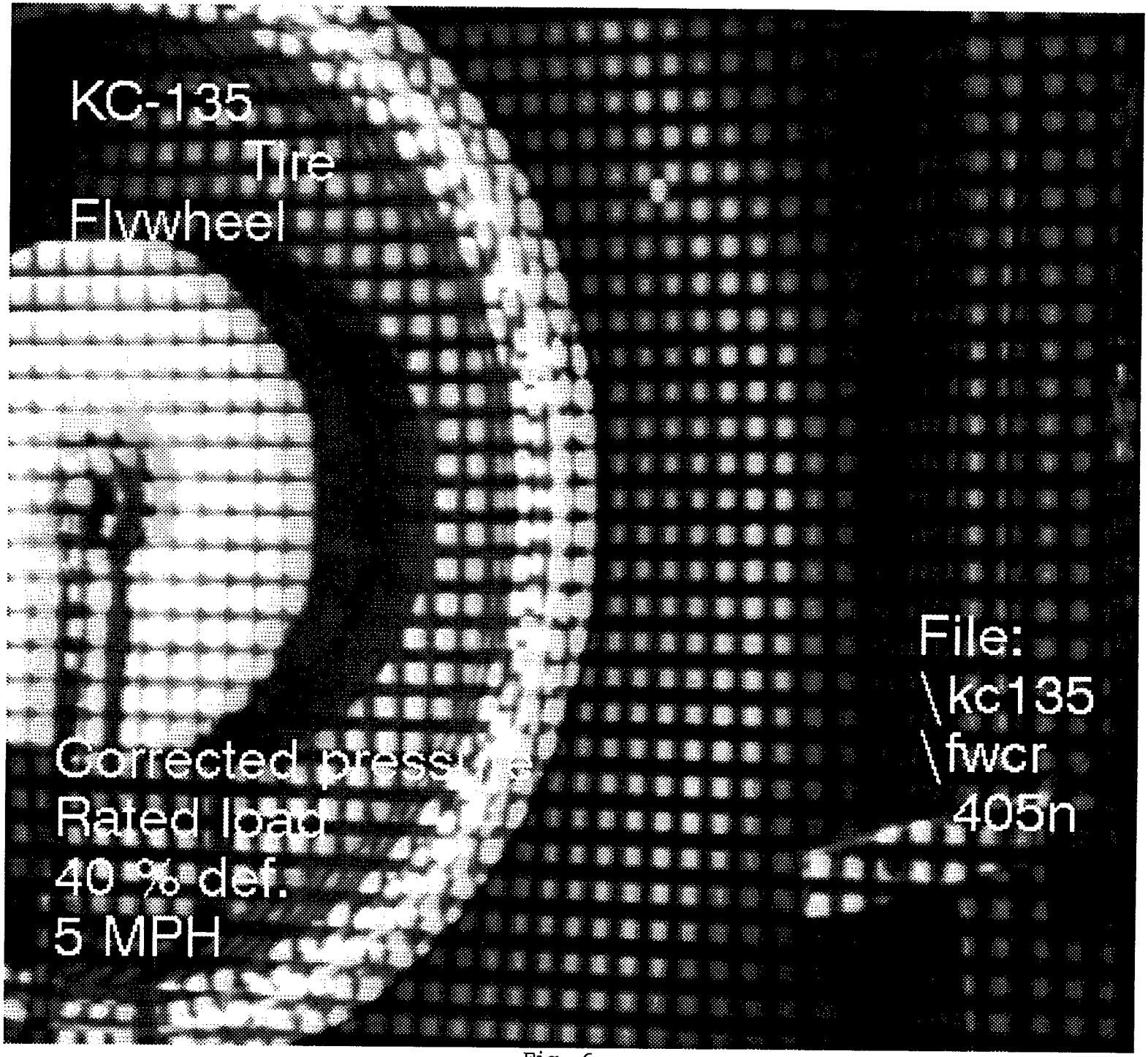

KC-135
Tire
Flywheel
Corrected pressure
Rated load
40 % def.
5 MPH
File:
\kc135
\fwcr
405n

Fig. 6

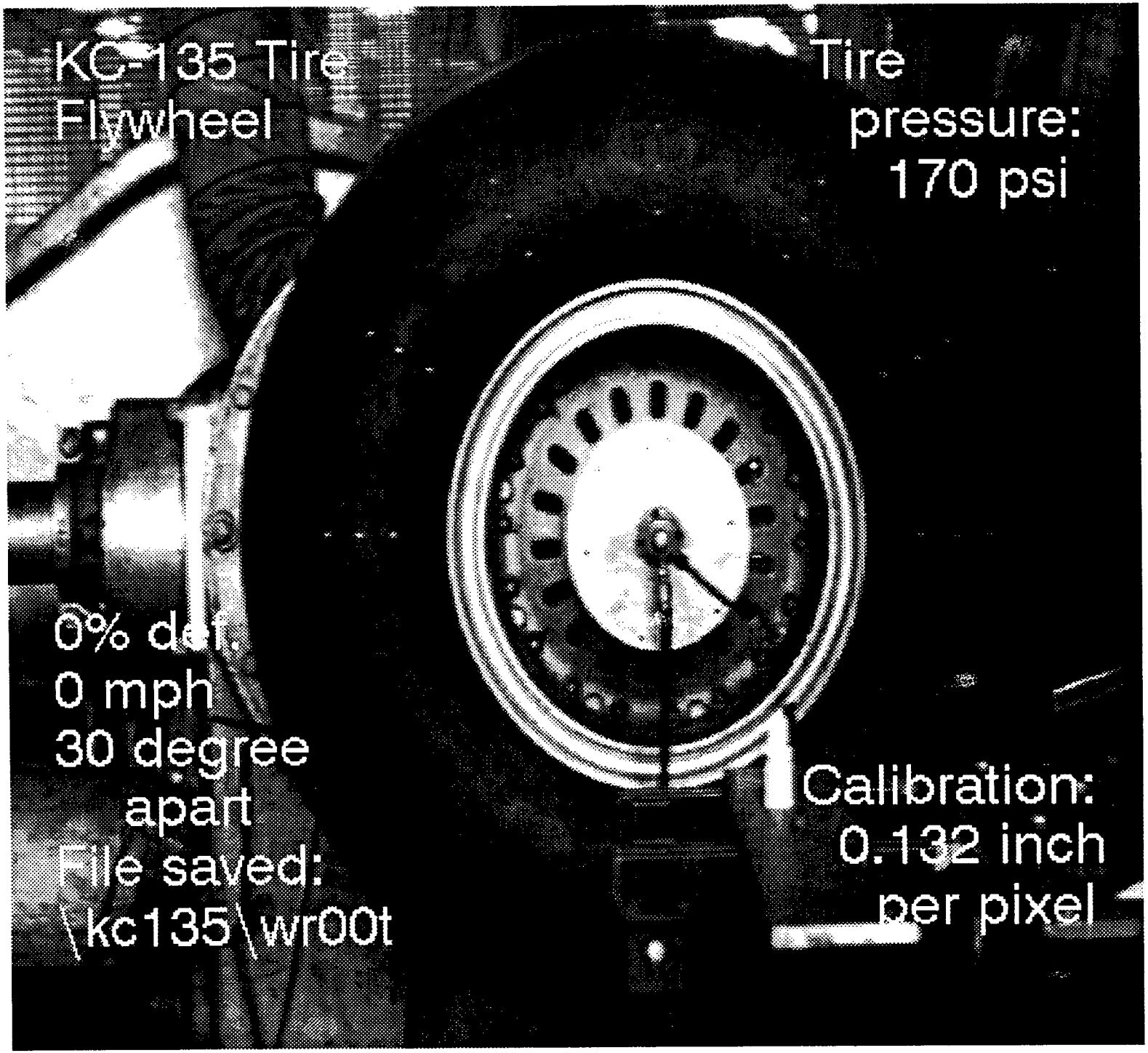
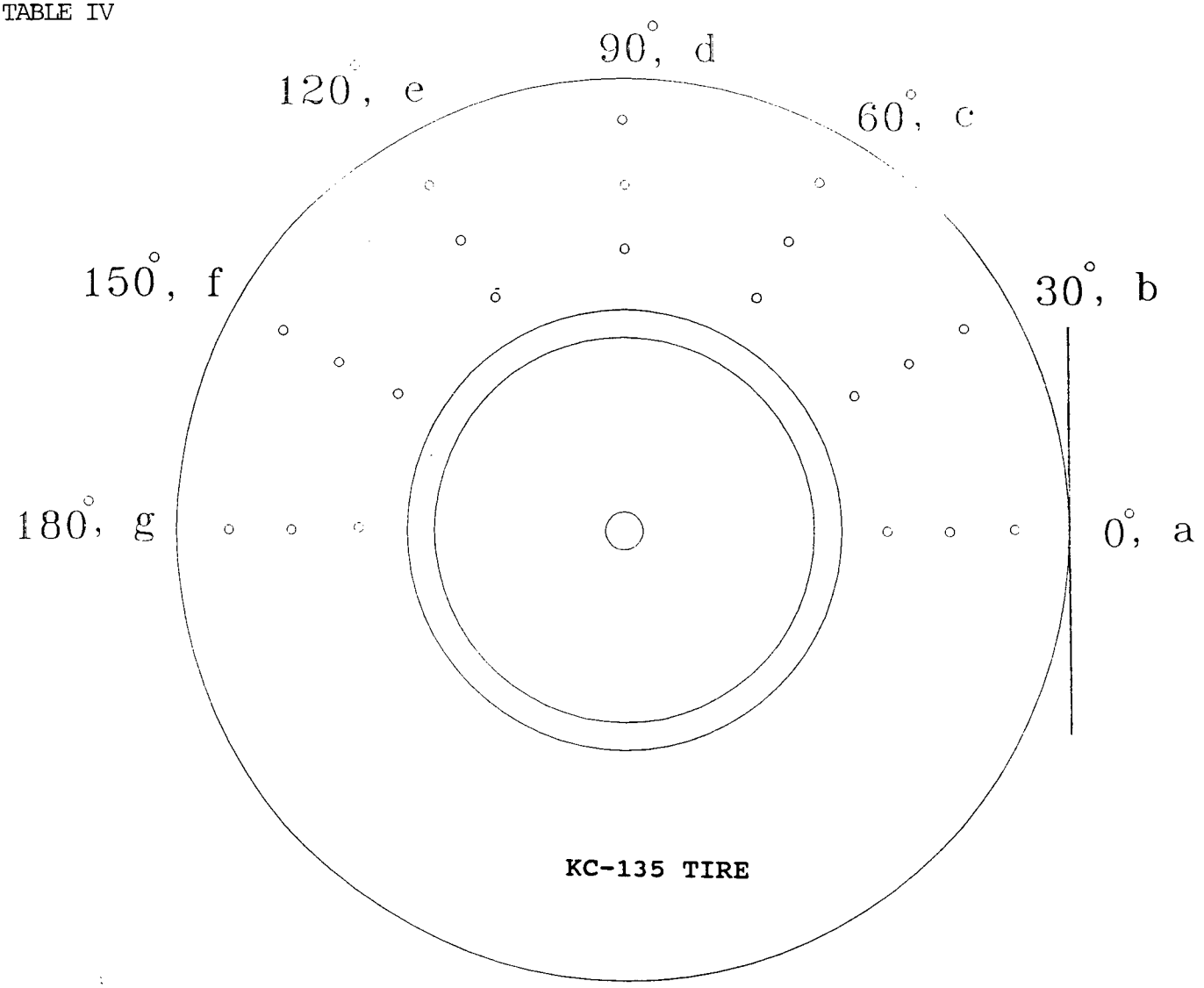

KC-135 Tire
Flywheel
Tire
pressure:
170 psi
0% def.
0 mph
30 degree
apart
File saved:
\kc135\wr00t
1
Calibration:
0.132 inch
per pixel

TABLE IV



With Flat Plate

Let a, b, c, d, e, f, and g represent the magnitudes of two dimensional displacements along 0, 30, 60, 90, 120, 150 and 180 degrees, respectively. Note that the displacements should be symmetric about the horizontal line (i.e. 0 degree line).

(1) 10% Deflection: b > c > a > d ≈ e ≈ f ≈ g
(about the same displacement between 90° and 180°).

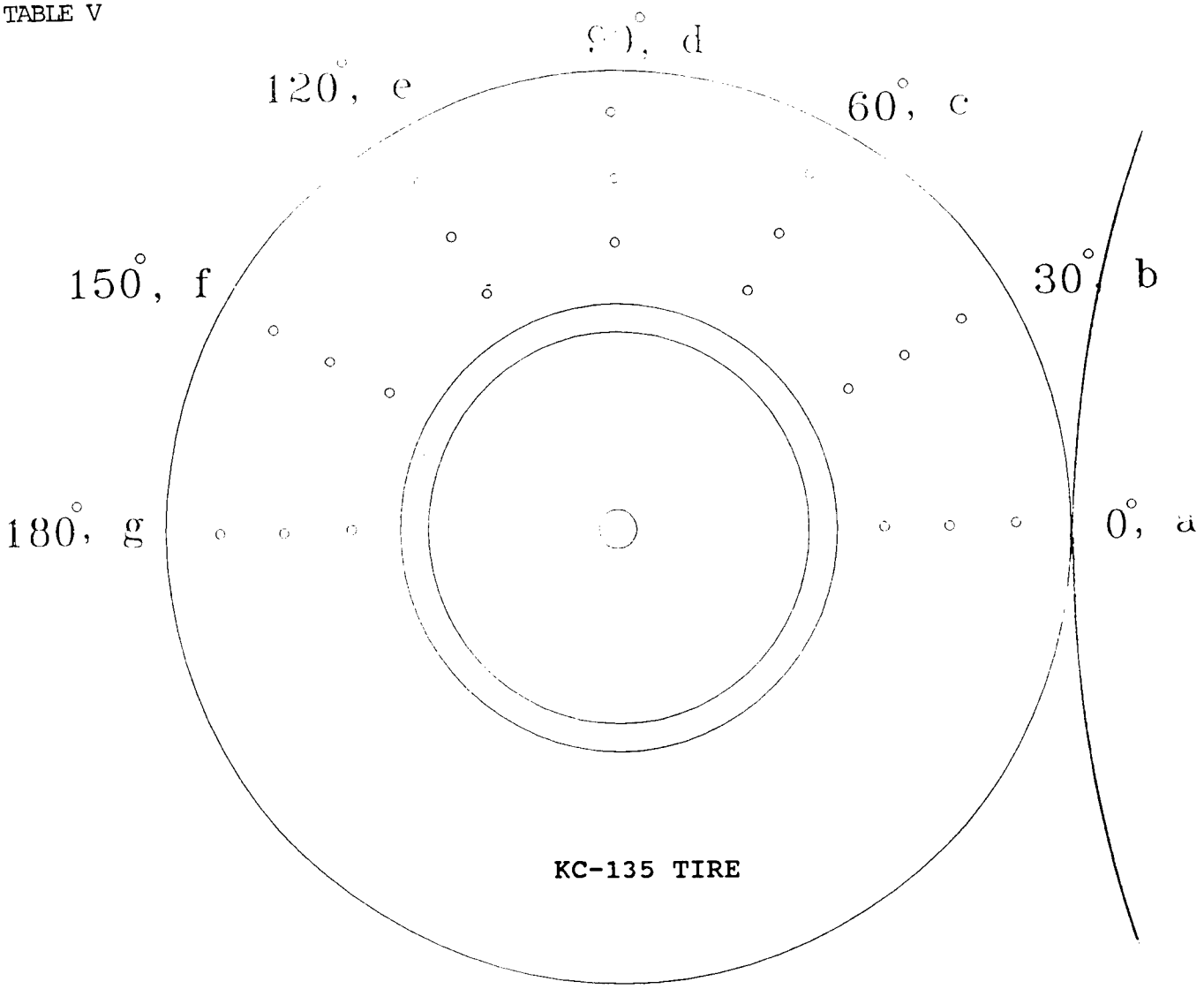
(1) 20% Deflection: b > c > a > d > e ≈ f ≈ g
(about the same displacement between 120° and 180°).

(1) 30% Deflection: b > c > a > d > e > f ≈ g
(about the same displacement between 150° and 180°).

(1) 40% Deflection: b > c > a > d > e > f > g
(about the same displacement along 180° only).

NOTE: Substantial vertical component of displacement exists along 30° and 60° lines, and there is no noticeable vertical component of displacement elsewhere.

TABLE V



With Flywheel

Let a, b, c, d, e, f, and g represent the magnitudes of two dimensional displacement along 0, 30, 60, 90, 120, 150 and 180 degrees, respectively. Note that the displacements should be symmetric about the horizontal line (i.e. 0° line).

(1) 10% Deflection: b > a > c ≈ d ≈ e ≈ f ≈ g
(about the same displacement between 60° and 180°).

(1) 20% Deflection: b > a > c > d ≈ e ≈ f ≈ g
(about the same displacement between 90° and 180°).

(1) 30% Deflection: b > a > c > d > e ≈ f ≈ g
(about the same displacement between 120° and 180°).

(1) 40% Deflection: b > a > c > d > e > f ≈ g
(about the same displacement between 150° and 180°).

NOTE: Substantial vertical component of displacement exists alone 30° line only; and there is no noticeable vertical component of displacement elsewhere.

Conclusions

This study shows that it is feasible to use the proposed measuring system to measure the three dimensional tire deformation and strains. The fringe projection optical technique worked very well for quantifying tire deformation. Some conclusions can be drawn as follows:

(1) The magnitude of deformation subjected to the same loading varies from one tire to another.

(2) Depending upon the location, the magnitude of deformation varies within the same tire.

(3) For the same tire, there is almost no difference between "rated pressure, corrected load" and "corrected pressure, rated load". However, there is a substantial difference between using a flat plate and a flywheel. In general, the displacement magnitudes with a flat plate are much larger than those with a flywheel. Thus, when we use a flywheel to spin a tire and to test the tire deformation, the magnitude of measured displacements should be magnified.

(4) Although the measuring system is not fast enough to trace the movement of some points at certain angular positions in a rotating tire, it is capable of quantifying the dynamic change of three dimensional tire deformation at some arbitrary angular position. To monitor the geometric change at the same angular position per revolution, a motion synchronization device is needed.

Recommendations for Future Work

In order for the measuring system to perform better, it is recommended that two additional pieces of equipment be added in the future tests. They are

(1) Fiber-optic displacement sensor: This sensor will provide an accurate detection of the reference point's height change.

(2) Synchronization device: With this device, the wobbling effect can be eliminated, since an image is taken only when points of interest in a rotating tire is at the same angular position.

Future work includes the determination of three dimensional tire deformation and strains, and tests for measuring the tire deformation subjected to braking and various yaw directions.

REFERENCES

1. Lin, P. P. and Parvin, F., "Edge Detection with Subpixel Resolution and its Application to Radius Measurement via Fringe Projection Technique," SME Technical Paper, MS90-576, 1990, pp. 4-13 - 4-27.

2. Lin, P. P., Parvin, F., and Schoenig, Jr., F. C., "Optical Gaging of Very Short-term Surface Waviness," Transactions of NAMAR/SME, 1991, pp. 327-322.

3. Marr, D. and Poggio, T., "Cooperative Computation of Stereo Disparity, " Science, V. 194, 1976, pp. 283-287.

4. Vemuri, B. C. and Aggarwal, J. K., "3-Dimensional Reconstruction of Objects from Range Data," Proc. of 7th Int. Conf. on Pattern Recognition, V1, 1984, pp. 752-755.

EXPERT SYSTEM RULE-BASE EVALUATION USING REAL-TIME PARALLEL PROCESSING

James L. Noyes
Professor of Computer Science
Department of Mathematics and Computer Science

Wittenberg University
Springfield, OH 45501-0720

Final Report for:
Summer Faculty Research Program

Wright Laboratory - Flight Dynamics Directorate
WL/FIPA Bldg 146
2210 Eighth Street Ste 1
Wright-Patterson Air Force Base, OH 45433-7511

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

August 1993

EXPERT SYSTEM RULE-BASE EVALUATION USING REAL-TIME PARALLEL PROCESSING

James L. Noyes
Professor of Computer Science
Department of Mathematics and Computer Science
Wittenberg University

Abstract

A large rule-based expert system having $O(10^3)$ rules, each involving $O(10^1)$ out of $O(10^5)$ possible Boolean conditions, can require a significant amount of processing time to evaluate. This time can be reduced if all rules have a single consequent and have antecedents which contain only conjunctions of the Boolean conditions or their complements. If the consequents do not insert new facts into the rule-base, then parallel processing can be used with even greater efficiency. Fast processing is necessary if real-time execution constraints are imposed, such as those associated with civilian and military aircraft cockpits during flight operations. This paper presents efficient data structures and algorithms to process that type of rule-base.

EXPERT SYSTEM RULE-BASE EVALUATION USING REAL-TIME PARALLEL PROCESSING

James L. Noyes

INTRODUCTION

The value of a rule-based expert system (ES) to help solve a variety of diagnostic and advisory needs has been well-demonstrated over the last two decades [2]. Sometimes, a large number, say $O(10^3)$, of the ES rules must be continuously checked in real-time (e.g., every $O(10^{-1})$ seconds) due to stringent requirements imposed by the problem. In addition, while each rule may use only $O(10^1)$ conditions, there may be a very large number of possible conditions, say $O(10^5)$, for the entire rule-base that must be checked during each time-step. Because of these timing demands, parallel processing may be deemed necessary. Parallel processing has become increasingly important in order to accelerate a variety of computations [2,8]. This paper discusses research connected to the development of a data structure and an algorithm to evaluate this type of rule-base and the estimation of the processor speeds necessary to evaluate these rules within the required time. The particular application for this real-time ES is a rule-base to aid the pilot of modern fighter or transport aircraft and the remainder of the paper will address this application. However, the results of the research presented here could be used in other applications as well.

EXPERT SYSTEM FORMULATION

This ES rule-base formulation depends upon a state vector, a criteria vector, a response (action) vector, and a set of rules.

The aircraft state vector **s** consists of continuous and discrete components (state variables) completely describing the state of the aircraft at a given time-step t_k, of magnitude $O(10^{-1})$ seconds. These values are determined by a collection of on-board sensors and there may be $O(10^2)$ of these. For example, state variable s_{12} might represent the number of gallons of fuel currently in the

fuel tank. The aircraft criteria vector **c** is a vector of m Boolean (True or False) variables. Each of these variables is based upon a value of one or more of the variables in the state vector. For example, criterion c_{33} might represent the relationship between the current amount of fuel and a minimum fuel reserve (e.g., $c_{33} = [s_{12} \geq f_R]$) and c_{33} is True as long as there is enough fuel in reserve.

A set of n rules, of order $O(10^3)$, defines the on-board expert system that will advise the pilot and, with the pilot's consent, act on his or her behalf. Each ES rule can be formulated in terms of a conjunction of simple Boolean conditions which lead to a single action. If all of a given rule's conditions are true (based upon the elements of the corresponding criteria vector), an action will result. This action could either be an activity that is automatically performed for the pilot or it could be a recommendation to the pilot. All of these actions define an action vector **a** of size n. Each rule is expected to involve only a relatively small number of m possible conditions. For this paper, m is of order $O(10^5)$. For example, each rule may have up to 10 conditions. The rule-base is built off-line, and not modified during the search process. For example, a typical rule might look like this ("~" means "NOT"):

$$\text{Rule } R_{123}: \quad \text{action}_{12} \Longleftarrow c_1 \;\&\; \sim c_5 \;\&\; c_6 \;\&\; c_{18} \;\&\; \sim c_{47} \;\&\; \sim c_{99}$$

This rule is interpreted as stating that action_{12} will be taken if c_1, c_6, and c_{18} are all true while c_5, c_{47}, and c_{99} are all false.

In a typical ES, the inference engine performs three standard operations: the match operation matches the criteria against the rules to see which could fire, the resolve operation chooses which of these rules will fire, and the execute operation actually fires these rules and updates working memory. For the given problem, these operations can be simplified into a simple match-fire operation with no resolution operation nor updates.

While it is assumed that no action alters the criteria vector **c** in any way at any time-step t_k, it is possible that different rules can have the same action. Hence, by expressing each rule only in terms of simple AND and NOT logic, its evaluation can be done very efficiently and independently. (Note that OR constructs are equivalent to multiple rules which specify the same action.)

DATA STRUCTURE AND ALGORITHM DEVELOPMENT

The data structure and algorithms developed to evaluate this ES, are designed to be used by a single fast processor or by parallel processors which can have a correspondingly slower clock-speed. This data structure utilizes the notion of a blackboard which contains the state and criteria vectors described above. In addition, three other vectors, the action, query, and index vectors completely define the rule-base. Unlike the **s** and **c** vectors, these three vectors are not updated, and can reside either in the blackboard or some other data storage area.

In general, a blackboard is a global and dynamic data base for the communication of independent asynchronous knowledge sources for related aspects of a given problem. The aircraft system blackboard will contain the state vector **s** and criteria vector **c**. These vectors will be updated by an independent on-board computer (not involved in the rule search) at each time-step. Each update of the vector **c** will immediately initiate a new rule search, so the rule search must be complete for the criteria vector at time-step t_k before the criteria vector is updated at time-step t_{k+1}. Hence, this blackboard must also be accessible by the computer that executes the rule processing algorithm. Criteria vector updates are discussed in [5,6].

While it is possible that a fast single processor computer could be used on the aircraft, the most likely hardware configuration for the rule processing algorithm will involve eight (8) parallel processors. Transputers are expected to be used. One of these eight processors will serve as the combined master and I/O processor, and will have one of its four serial I/O ports connected to the

common data bus on the aircraft. This processor will accept the criteria vector and possible pilot input and provide the ultimate rule search output. The remaining seven (7) processors may use each of their four I/O ports to connect to any other processor. A pre-set architecture will be employed.

Both algorithms presented in this paper should be considered as prototypes and have been implemented in Pascal. If a particular algorithm is sufficiently successful, it will be eventually be implemented in Ada. This will permit a ready transition to operational aircraft. Mil-std versions of transputers exist. Together, transputers programmed in Ada represent a mature and installable parallel processing capability that takes advantage of modern processor architectures.

Method-1

The simplest method for this ES evaluation is based upon assuming prioritized if-then action rules. This is equivalent to a priority-oriented backward chaining method. This is the obvious choice when n << m and no other assumptions are made about available data. (Note that if these rules were not prioritized, then this first algorithm could be viewed as a forward chaining algorithm.) Because no OR-logic is present in a given rule, the current rule-processor should stop with the first False for c_i (or first True for $\sim c_i$). If these rules were ranked and evaluated from highest to lowest priority, then the first action produced (if any) would be the most important from the pilot's point of view. If required, different levels of parallelism could be employed during this evaluation process. If the processing time is not fast enough, then rules having the same priority could be grouped according to their number of criteria in order to equalize the work among the parallel processors [7]. A simple example of a rule-base with four rules is:

Rule R_1: $action_1$ <== c_1 & c_3 & $\sim c_4$ & c_{40} & $\sim c_{98}$ & c_{99} (Highest Priority)
Rule R_2: $action_1$ <== c_2 & c_4 & c_{22} & $\sim c_{85}$
Rule R_3: $action_2$ <== c_5 & c_{99}
Rule R_4: $action_3$ <== c_1 & c_{50} (Lowest Priority)

These could be represented efficiently by using three vectors: the previously discussed action vector **a**, a query vector **q**, identifying which criteria have to be checked, and an index vector **End**, that delimits the criteria that appear in each of the rules. For the above rule-base, consider:

Rule 1: $a_1 = 1$; $q_1 = 1$, $q_2 = 3$, $q_3 = -4$, $q_4 = 40$, $q_5 = -98$, $q_6 = 99$, so
$Start_1 = 1$; $End_1 = 6$
Rule 2: $a_2 = 1$; $q_7 = 2$, $q_8 = 4$, $q_9 = 22$, $q_{10} = -85$, so
$Start_2 = 7$; $End_2 = 10$
Rule 3: $a_3 = 2$; $q_{11} = 5$, $q_{12} = 99$, so
$Start_3 = 11$; $End_3 = 12$
Rule 4: $a_4 = 3$; $q_{13} = 1$, $q_{14} = 50$, so
$Start_4 = 13$; $End_4 = 14$

Here **q** employs positive integers to indicate the indices of the criteria used in the rules and negative integers for the indices of the criteria complements (NOT-conditions). From the previous example, one has the 14-element query vector:

q:

1	3	-4	40	-98	99	2	4	22	-85	5	99	1	50

This allows for direct and very fast access to the **c** vector stored on the blackboard (only one internal integer multiplication and addition are needed to compute any cell address). If multiprocessors are used, this Boolean criteria vector **c** can be accessed from the blackboard by all of these parallel processors. If multicomputers are used, **c** would be communicated to the local memory of each processor and this communication time will need to be considered [1]. Each processor also must use components from the query vector **q**. Note the relationship $Start_{j+1} = End_j + 1$ with $Start_1 = 1$, so only the **End** unsigned integer index vector is actually needed by the algorithm. In this example, one has:

End:

6	10	12	14

which implies **Start:**

1	7	11	13

This method yields Algorithm-1, presented next, which is a relatively simple and straightforward algorithm that can utilize these data structures.

```
Forall i := 1 to n do in parallel
   begin
      if i = 1
         then j := 1
         else j := End_{i-1} + 1;
      Fired := TRUE;
      while j ≤ End_i and Fired do
         begin
            k := q_j;
            if k ≥ 0 and not c_k then
               Fired := FALSE
            else if k < 0 and c_-k then
               Fired := FALSE;
            j := j + 1
         end;
      if Fired then perform action a_j
   end
```

In Algorithm-1, the Forall statement creates up to n parallel processes. If p is the number of parallel processors and $p \geq n$, then this loop completes as soon as the slowest of these processes has finished execution. Here the total parallel processing time at a given time-step is the maximum of these times. If $p < n$, then the next available processor would evaluate the next unprocessed rule, hence the total parallel processing time at a given time-step is then the maximum of all of the sums of the individual processor times. Notice that this reduces to a normal sequential processing algorithm when $p = 1$.

For example, if $p \geq 4$ and it takes an estimated average of 50 microseconds to check each condition in the previous 4-rule situation, then 4 copies of the loop body will be created on 4 different processors, each with its own value of the loop control i-variable. These will execute in parallel with respective times of 300, 200, 100, 100 microseconds, at most (as soon as a FALSE is determined, the process stops for the current rule). This would then take at most 300 microseconds in parallel versus at most 700 microseconds if done sequentially, giving a speedup of 7/3 or approximately 2.3. Here the action performance time (e.g., displaying an information screen) was not considered, nor was processor-assignment overhead or communication time. Of course, any of these three times can have a significant effect on this ES evaluation process.

Method-2

The previous method does not take advantage of searching in any informed way whenever a state variable (and hence a criterion) changes, because the indexing is in the opposite direction from rule to criterion. A second, combined forward-backward chaining method, could be used to check only the rules whose criteria values have changed since the last evaluation of the rule-base. To do this, one could also index in the opposite direction, checking only the rules having newly changed (currently "active") criteria. The forward phase identifies the changed criteria and rules that use these criteria. The backward phase is the same as before with presumably fewer rules to process. For example, using the same four rules as before, one could have something like:

Criterion c_1: $NeedToCheck_1$ = False; $First_1$ = 1; $Last_1$ = 2; $r_1 = 1$, $r_2 = 4$
Criterion c_2: $NeedToCheck_2$ = True; $First_2$ = 3; $Last_2$ = 3; $r_3 = 2$
Criterion c_3: $NeedToCheck_3$ = False; $First_3$ = 4; $Last_3$ = 4; $r_4 = 1$
Criterion c_4: $NeedToCheck_4$ = True; $First_4$ = 5; $Last_4$ = 6; $r_5 = 1$, $r_6 = 2$
Criterion c_5: $NeedToCheck_5$ = False; $First_5$ = 7; $Last_5$ = 7; $r_7 = 4$
Criterion c_6: $NeedToCheck_6$ = True; $First_6$ = 0; $Last_6$ = 0; not in any rule
.....
Criterion c_{99}: $NeedToCheck_{99}$ = False; $First_{99}$ = 13; $Last_{99}$ = 14; $r_{13} = 1$, $r_{14} = 3$

Assuming criteria c_2, c_4, and c_6 were the only ones that changed (their **NeedToCheck** components would be set to True in the blackboard), the above would cause $Rule_2$, $Rule_1$, and $Rule_2$ to be consolidated into the set { $Rule_1$, $Rule_2$ } with the components $NeedToCheck_2$, $NeedToCheck_4$, and $NeedToCheck_6$ being reset back to False. The efficiency of this method is related to the number of active (recently changed) criteria at any time-step. The number of criteria that change at any time-step is highly dependent upon the application. The fewer the criteria that change, the faster this method will be, but this method is more complex and requires both more data and storage than the previous method.

Each change in the state vector **s** at time-step t_k can cause the status of the Boolean criteria vector **c** (and its corresponding **NeedToCheck** vector) to change. Each criteria vector change, in turn, causes a set (or prioritized list) of rule numbers to be defined. Each rule in the set would contain at least one of the

changed criteria and only the rules in this set need to be checked to see if all conditions hold. Once these rules have been identified, the actual criteria checking itself is done in the same manner as in Algorithm-1.

DEALING WITH UNCERTAINTY

In practice, one or more sensor failures may lead to undetermined (uncertain) components of the state vector **s**, which may lead to one or more unknown components of the criteria vector **c**. For any given rule, one of three situations must hold at time-step t_k: (1) all its criteria are known, (2) there are unknown criteria, but at least one of the known criteria fails to be satisfied, (3) all of the known criteria are satisfied, but there are still unknown criteria. The first two situations are easily addressed, since it can be exactly determined if the rule will fire or not (in the second case it will not fire). In the third situation, the values of the unknown criteria determine whether the rule will fire or not. Because of the possible interdependence of criteria, it is very difficult to determine any type of formal probability measure associated with the firing of this rule since multi-variable conditional probabilities are involved. However, it is possible to report a possible action by simply keeping count of the number of criteria that are unknown for the given rule. This requires that each component of the criteria vector **c** have one of three conditions (True, False, Unknown), instead of just True or False as used in Algorithm-1.

The algorithm to do this is a variation of Algorithm-1, but is slightly more complex and takes more processing time. This is because an additional IF-test is needed and two additional counting operations are necessary for the reporting when one or more of the necessary **c** values are unknown. The reporting of the Ucount/Ncrit ratio is intended to give the pilot some measure of exactly how many unknown criteria (Ucount) exist relative to the total number of criteria (Ncrit) that are used in the given rule. For example, if there are ten criteria in the rule and a possible action is reported with a ratio of 1/10, then the pilot might place more confidence in it than if a ratio of 7/10 was presented.

The algorithm designed to deal with this uncertainty is presented below as Algorithm-1u:

```
Forall i := 1 to n do in parallel
   begin
      if i = 1
         then j := 1
         else j := End_{i-1} + 1;
      Fired := TRUE;
      Ncrit := 0;
      Ucount := 0;
      while j ≤ End_i and Fired do
         begin
            k := q_j;
            if c_|k| is Unknown then
               Ucount := Ucount + 1
            else if k ≥ 0 and c_k is False then
               Fired := FALSE
            else if k < 0 and c_-k is True then
               Fired := FALSE;
            j := j + 1;
            Ncrit := Ncrit + 1
         end;
      if Fired then
         if Ucount = 0
            then perform action a_j
            else report possible a_j with Ucount/Ncrit ratio
   end
```

This algorithm could also be modified to report exactly which unknown criteria caused the problem. When considered in the total application context, it may also be useful to report the failed sensors that caused the unknown criteria.

SIMULATION GUIDELINES AND RESULTS

The software and hardware realization associated with the rule processing algorithm will depend upon the amount and frequency of the available data and the real-time constraints for the solution. To see if an algorithm is acceptable, it could be implemented within a specially written Turbo Pascal simulation program such as PASIM (Pilot's Associate Simulator). This simulator can be used to test Algorithm-1 and estimate both the sequential and parallel processing speeds. Because of the interest in handling uncertainty, the PARSIM program was developed to test Algorithm-1u. PARSIM can be thought of as an extension of

PASIM that also allows the user to incorporate an uncertainty percentage that will also simulate sensor failures throughout the flight. In this section, sample simulation results are given for both of these algorithms. However, most of the emphasis is placed upon Algorithm-1u as implemented by the PARSIM program.

The current PARSIM program parameters include: (1) a maximum of 10,000 rules that are assumed to be in priority order, (2) a maximum of 10,000 different actions (during a given time-step, actions can be listed for all the rules that are fired), (3) a maximum of 32,760 different criteria can be used altogether (this is the largest single block of data allowed in Turbo Pascal and 32,767 is the largest positive integer), (4) a maximum of 8 transputer processors can be used (one of these used strictly for I/O). The use of dynamic (array) variables in the program was ultimately necessary to allow the sizes achieved above.

In order to perform an effective simulation, one needs to know the processor speed in: (1) evaluating a single Boolean condition c_i, and (2) performing any recommended action produced by a rule. The Inmos transputers to be simulated are T800-20 32-bit models with math coprocessors. These transputers have a clock-speed of 20 MHz and up to 4 megabytes of memory. The transputer rule processing speeds are unknown, but the more critical condition evaluation speed can be estimated or bounded by empirically timing this evaluation on the processors below (all including math coprocessors). Here are the approximate averages of the measured processing speeds required to evaluate a single criterion based upon Algorithm-1:

Intel 80386/16MHz: 81 microseconds = 8.1×10^{-5} seconds
Intel 80486/33MHz: 16 microseconds = 1.6×10^{-5} seconds
Intel 80486/50MHz: 12 microseconds = 1.2×10^{-5} seconds

The 386/16 processor is the slowest of these and presumably the closest in processing speed to the T800-20 transputer. For Algorithm-1u with the 386/16 processor, the approximate average criterion evaluation speed was found to be 115 microseconds (i.e., it takes almost 42% longer to evaluate a single criterion).

PARSIM randomly generates up to a given number of conditions for each of n rules. It also generates random Boolean values for m conditions, and updates them at random for each time-step. Many random numbers are required during a typical simulation. The built-in Turbo Pascal pseudo-random number (PRN) generator called RANDOM, did not produce a sufficient amount of PRNs, so the uniform (0,1) real full-period PRN generator (implemented with 32-bit integers) is employed [3,4]. This is done with the seed update s := 16807s mod 2147483647 and producing the uniform PRN by using u := s/2147483647. As usual, at the start of a typical simulation, the "randomized" initial seed s is obtained from the system clock.

A nominal mission length for a fighter aircraft (such as an F-16) might range from 1-2 hours up to as many as 5 hours of flying time. If one assumes that sensor updates all occur at 0.1 second intervals, this dictates the simulation time-step. For example, a 90 minute flight would take 54,000 time-steps (90x60x10), and if there were 10,000 rules with up to 10 criteria per rule (an average of 5 criteria per rule), it would take about eighteen hours to run the PASIM code on a 486/33 machine. The PARSIM code takes longer, for the reasons already indicated, hence a much shorter number of time-steps was used.

In the simulation, all rules in the rule-base are evaluated for each time-step in the flight, this is to produce the "worst-case" situation so that any necessary processor speed-up can be identified. Obviously the search could be significantly faster if the algorithm could terminate after the first action was performed.

Figure 1 indicates essentially what is shown on the screen when PARSIM executes (the screen size has more columns than this page of text so the wording has been slightly modified). The underlined quantities represent the simulation input and output values. As with any simulation, these values contain a measure of uncertainty.

Specifically, Figure 1 shows the input and output of a short PARSIM simulation with 4,000 rules with up to 10 criteria per rule generated. The number of unique actions chosen does not affect the simulation and was arbitrarily chosen to be 4,000 also. The input of 1.15e-4 indicates the estimated average time, in seconds, that it will take the on-board rule-processor to process a single criterion (typical of a 386/16). The action time is input as zero, since it is assumed that the ES triggers another computer to perform this action. No intermediate output is requested (it is only feasible to do this when the number of rules is very small). Here 12,000 is the number of time-steps. Depending upon the sensor update time, this could correspond to different flight times. For example, if 0.1 seconds is the update time, then 0.1x12000 seconds corresponds to 20 minutes of flight time, but if 0.5 seconds is the sensor update time, the flight is 1 hour and 40 minutes in length. The number 16027 is the total number of unique criteria generated. When the user enters 10, it indicates that 10% of these are expected to fail before the end of the flight.

```
                    Rule Simulation Program
This simulates the real-time processing of a set of expert system rules.
Logical contradictions within the generated rules are not guaranteed, nor are
the absence of duplicate rules.  Neither should significantly effect the
simulation.  It is guaranteed that there will be no duplicate criteria in a rule.
INPUT:
Enter the number of rules to generate (1,..,10000): 4000
Enter the number of different actions (1,..,10000): 4000
Enter the criteria limit for each rule out of 32760 possible (1,..,28761): 10
Enter the simulated time needed to evaluate a single criterion: 1.15e-4
Enter the simulated time needed to perform a single action: 0
Enter the number of parallel processors (2,..,8): 8
Enter the amount of intermediate output desired (0 is nominal)
- None(0), First Action(1), Rules & Actions(2), Rules, Actions & Criteria(3): 0
Enter the (non-negative) number of simulation time-steps: 12000
Enter the uncertainty percent for the 16027 criteria generated [0.0,100.0]: 10
OUTPUT:
The actual lapsed system clock a-time was 1.499850E+0003 seconds, with 21966
criteria processed and 1723 unique rule(s) out of 4725280 fired.
On average there were 5 criteria per rule with 3.125E-0005 seconds
needed to process each rule and 1.250E-0001 seconds for the entire rule-base.
The simulated sequential s-time (one CPU) was 1.034389797E+0004 time units.
The simulated parallel p-time (8 processors) was 1.529650535E+0003 time units.
Max{a-time}= 1.7000E-0001, Max{s-time}= 9.0988E-0001, Max{p-time}= 1.3720E-0001
```

Figure 1

In the output, the a-time is the total time that it takes the computer executing PARSIM to process all the criteria in all the rules for the entire flight. (The example in Figure 1 was run on a 486/33 and took 1,499.85 seconds to do this.) Its main purpose is to help determine the overall average single criteria evaluation time for the particular computer being used. If the user enters a 1 for the single criterion evaluation time, then the s-time (simulated sequential time) is simply a count of the number of criteria processed. By dividing this count into the a-time, one obtains the average time that it takes to process a single criterion for the processor on the computer currently being used. By making several runs of this type (e.g., 15 runs), one can obtain a reasonable estimate of this overall average.

The key output values are the last two given in this figure and represent the maximum simulated sequential and parallel times over the entire flight that it will take to evaluate the rule-base. From these two values, it can be determined if the entire rule-base can be processed in less time than the sensor update time. For example, if the sensor updates are done every 0.1 seconds, then the rule processing time at any time-step must not be slower that this. Here a single processor takes slightly over 0.9 seconds to process the entire rule-base with the sequential form of Algorithm-1, while the 8-processor parallel version of the same algorithm takes slightly over 0.1 seconds. For this single simulation, neither the sequential nor the 8 parallel processors (only 7 actually processing the rules) will process the rule-base within an acceptable amount of time. However, if the sensor update time is 0.5 seconds, then the parallel processing is fast enough since $0.1372 < 0.5$, but the sequential processing is still not fast enough. If this algorithm is implemented on multicomputers that have no common memory, then data, such as the c vector, will have to be communicated to each local rule-processor for each sensor update cycle. This takes an additional amount of time. For example, if this time were 0.2 seconds, the parallel processing is still fast enough since 0.3372 is still less than 0.5.

Of course, conclusions such as the above should not be based upon just one simulation run. One should run several simulations, at least 10, with the same number of rules for long time periods (e.g., equivalent to 5 hours of flight time) in order to draw conclusions with a sufficient amount of confidence. Since it will take approximately 1 hour and 24 minutes to perform this simulation on a 486/33 microcomputer, a lot more running time is needed.

If uncertainty is not a concern, then Algorithm-1 can be used as implemented by PASIM. Using the same inputs as above, except that 8.1e-5 is used in place of 1.15e-4 (no uncertainty percent is needed), one finds that the maximum sequential time is almost 0.6 seconds while the maximum parallel time is under 0.09 seconds. Hence, based upon just one run, on would conclude that parallel processing of 4,000 rules is fast enough even when 0.1 second sensor updates are used. This does not take any additional communication time into account.

There are some PARSIM (and PASIM) system limitations that should be mentioned: (1) Due to the problem requirements as well as the clock precision on the simulation computer (1/100th of a second), one should not expect reliable timing estimates with fewer than 1,000 rules. (2) The upper limit is 10,000 rules, but if one simulates a rule-base of around 10,000 rules and chooses a maximum of, for example 10 criteria per rule, then on average 50,000 criteria would have to be kept in the **q** array. But, this array is limited to 32,760 locations, so the code will automatically reduce the number of criteria per rule near the end of the generated rule-base to ensure that each rule has at least one criterion. Hence, the actual average number of criteria per rule may be closer to 3 than to 5 because of all of the rules that must have only one criterion. It is up to the PARSIM user to determine if this average is realistic. (3) PARSIM times do not take into account any processor assignment or data communication times. These depend upon both the architecture and hardware being used.

Most of the simulations that were run during this investigation used rule-bases of sizes from 1,000 up to 6,000. The number of distinct actions was arbitrarily input to be the same as the number of rules since this has no effect on the simulation timing at all. (However, if one wishes exact simulation reproducability, this can be input as a negative number, and the absolute value of this number is used as the initial PRN seed instead of using the system clock.) Two different rule limits were investigated, up to 10 criteria per rule and up to 20 criteria per rule. For example, if a rule is needed to identify a radar according to six pairs of parameter ranges, then up to 12 criteria may be needed in this rule (e.g., $710 \leq f_1 \leq 855$ yields $c_1 = f_1 \geq 710$ and $c_2 = f_1 \leq 855$).

Two different time-steps were also studied, 0.1 and 0.5 seconds. These were the thresholds used to determine when the simulated maximum sequential or parallel times were good enough, meaning smaller than 0.1 or 0.5 seconds, respectively. The criterion evaluation times depended upon the algorithm being used. As stated earlier, on average, it was found that Algorithm-1 used 81 microseconds and Algorithm-1u used 115 microseconds to evaluate a single criterion for the 16MHz Intel 386 processor which is the available processor closest in clock speed to the 20MHz transputer.

CONCLUSIONS

Due to the short project time available and the large amounts of running time required, only one or two simulations were made for each rule number and criteria limit combination. All of the conclusions below are based upon these limited cases and should be viewed accordingly. In particular, for maximum rule-base evaluation times close to any desired threshold (e.g., 0.1 or 0.5), more simulations will be necessary.

Based upon the simulations using Algorithm-1 with the PASIM program (no uncertainty addressed), eight parallel processors with a clock speed of 16MHz or faster were always able to process a rule-base of 4,500 or fewer rules having a

maximum of 10 criteria per rule. Single processors at this same speed were unable to do this in under a tenth of a second. The average maximum speed-up was 6.6, using 1 master and 7 rule-processors. A faster single processor, such as the 50MHz Intel 486, would be able to process the same 4,500 rules within this same time.

Once uncertainty is introduced into the criteria vector, the processing time increases. This was investigated by using Algorithm-1u in the PARSIM code. Here the estimated average time to process a single criterion goes from 81 microseconds to 115 microseconds. In order to process all of the rules within a tenth of a second, with up to 10 criteria per rule, the simulation showed that number of rules in the rule-base had to be reduced to 2,500. It appears that up to 2,000 rules, each having up to 20 criteria can be executed in parallel under a tenth of a second (with no extra communication time taken into account). Without parallel processing, not even 1,000 rules could be evaluated. If the sensor update time is increased to a half-second, then up to 6,000 rules with up to 20 criteria each can be processed in parallel (even with a 0.2 second communication time added).

ACKNOWLEDGEMENTS

First, I would like to thank Major Peter G. Raeth, Chief of the Pilot/Vehicle Interface Branch (WL/FIPA) and my research advisor during this ten-week period. His guidance was very clear and he went out of his way to see that the problem was well-defined and helped me obtain the related literature and computing resources. Next I would like to thank Lt Col. Tim Kinney, Deputy of the WL/FIP Division, and Dr. Jim Olsen, Chief Scientist of WL/FI for their invitation to do this type research at Wright Laboratory. Finally, I would like to thank Andy Probert, WL/FIPC, for his insights into what pilots really expect from an on-board expert system.

REFERENCES

1. Bruce P. Lester, The Art of Parallel Programming, Prentice Hall, Englewood Cliffs, NJ, 1993.

2. James L. Noyes, Artificial Intelligence with Common Lisp: Fundamentals of Symbolic and Numeric Processing, D. C. Heath, Lexington, MA, 1991.

3. Stephen K. Park and Keith W. Miller, "Random Number Generators: Good Ones are Hard to Find," Communications of the ACM, Vol. 31, No. 10, Oct. 1988, pp. 1192-1201.

4. William H. Press, et al., Numerical Recipes in Pascal: The Art of Scientific Computing, Cambridge University Press, New York, NY, 1989.

5. Peter G. Raeth, "An Expert Systems Approach to Decision Support in a Time-Dependent, Data Sampling Environment," in Expert Systems: A Software Methodology for Modern Applications, Edited by Peter G. Raeth, IEEE Computer Society Press, Los Alamitos, CA, 1990, pp. 170-177.

6. Peter G. Raeth, "Expert Systems in Process Observation and Control," AI Expert, Vol. 5, No. 12, Sep. 1990, pp. 40-45.

7. K. R. Tout and D. J. Evans, "Parallel Forward Chaining Technique with Dynamic Scheduling, for Rule-Based Expert Systems," Parallel Computing, Vol. 18, No. 8, Aug. 1992, pp. 913-930.

8. Robert R. Trippi and Efraim Turban, "Parallel Processing and OR/MS," Computers & Operations Research, Vol. 18, No. 2, 1991, pp. 199-210.

A NEW MODELING TECHNIQUE FOR PIEZOELECTRICALLY ACTUATED BEAMS

Mo-How Herman Shen

Assistant Professor

Department of Aeronautical and Astronautical Engineering

The Ohio State University

2036 Neil Ave.

Columbus. Ohio 43210

Final Report for:

AFOSR Summer Research Program

Wright Laboratory

Wright Patterson AFB

September 27, 1993

A NEW MODELING TECHNIQUE FOR PIEZOELECTRICALLY ACTUATED BEAMS

Mo-How Herman Shen
Assistant Professor
Department of Aeronautical and Astronautical Engineering
The Ohio State University

ABSTRACT

A one dimensional theory is developed for modeling the analysis of beams containing piezoelectric sensors and actuators. The equation of motion and associated boundary conditions are derived for the vibrations of piezoelectrically actuated beams. A generalized variational principle is used to formulate the equation of motion, taking into account the interfacial shear stress concentration near the ends of the actuators. This is accomplished by introducing a "stress function" into the beam's compatibility relations. This function has its maximum value at the ends of a piezoelectric actuator and decays exponentially in the longitudinal direction. The effect of coupling between longitudinal deflection and bending deflection is investigated in the present study. For the practical applications, in according with the proposed beam theory, a one-dimensional finite element formulation is presented. The proposed beam theory as well as the finite element approach can be easily used in developing a formal two-dimensional theory for piezoelectrically actuated composite plates and shells or other physical systems.

A NEW MODELING TECHNIQUE FOR PIEZOELECTRICALLY ACTUATED BEAMS

Mo-How Herman Shen

INTRODUCTION

One of the major areas of DoD, NASA, and the Air Force research in engineering involves structural mechanics which is concerned with developing lighter, stronger, and more durable structures that can be applied to a variety of flight vehicles ranging from helicopters to interplanetary spacecraft. The future direction in this area is to expand the limits of the vehicles' structural performance which are able to sense, to respond, and to control their own characteristics and states, so as to achieve much higher levels of operational performance to meet mission requirements. In general, these types of structures are called **Adaptive Structures** or **Smart Structures**. Realization of these new structures is possible only by extensive research in the areas of materials, monitoring, diagnostics and identification, and adaptive control integration. Toward that goal, in the beginning, a considerable amount of research effort has been expended in attempting to understand and construct new materials that would be able to sense the environmental uncertainties and actively adapt themselves to the environment.

Piezoelectric materials have been widely used as sensory and active materials in many applications in the area of Adaptive Structures. One of the engineering challenges is the modeling of the actuation mechanisms in the actuators and substructure. Extensive research[1-8] have been done analytically as well as experimentally on the implementation of mechanics of structure and the electric energy of the piezoelectric actuators. Almost all the studies so far were conducted based on the conventional beam or plate theories where a local stress/strain field is not considered. However, it has been found out in many applications that the models based on those conventional beam or plate theories cannot achieve

an accurate desired structural form in according with the provided electric energy. In turn, a more detailed modeling technique is needed in order to accurately and effectively predict the actuation mechanisms of integrated active structural system.

Recently, several studies[1-4] have been expended in attempting to understand and model the actuation mechanism in beams with piezoelectric actuators. Such a task is complicated by the fact that the transverse shear stress (possible the transverse normal stress) is concentrated at the ends of the actuators. It has been clearly indicated in the studies by Robbins and Reddy[1] and Lin and Rogers[2,3] that the stress concentration is indeed significantly affecting the actuation mechanisms. Consequently, in order to design an adequate piezoelectric actuators and to provide an accurate actuation mechanism, the stress concentration must included in the mathematical models, particularly for those cases requiring relatively thick actuators.

The objective of the study is to further develop an effective one-dimensional mathematical model for determining the dynamic responses of beams with piezoelectric actuators (see Figure 1). This model is based on Timoshenko beam theory and general kinematic assumptions which account for the coupling effects between longitudinal vibration and bending vibration as well as the effects of an interfacial shear stress concentration near the ends of the piezoelectric sensors. A corresponding experimental study is also carried out to verify the analytical predictions. The use of this new theoretical model to structural damage detection and identification and active controls of structural systems is also discussed.

THEORETICAL DEVELOPMENT

Main structure

The geometry of a beam is shown in Figure 1. In the main structure, indicated by beam 1 in Figure 1, only transverse vibration is considered. The assumptions using the

Timoshenko beam theory to include shear deformation effects are summaried as following:

$$\begin{cases} u_y = 0, \;\; u_z = w(x,t), \;\; u_x = -z(w' + \beta) \\ \epsilon_{xx} = u' \;\; \epsilon_{xz} = -\beta \\ \epsilon_{yy} = \epsilon_{zz} = -\nu\epsilon_{xx} \\ \epsilon_{xy} = \epsilon_{yz} = 0 \\ \sigma_{yy} = \sigma_{zz} = \sigma_{xy} = \sigma_{yz} = 0 \\ X_x = X_y = X_z = 0 \end{cases} \tag{1}$$

where u_i are the displacements referring to cartesian axes x, y, z; σ and ϵ represent stress and strain, β is the angle due to shear force, and X_i and p_i are the body forces and velocity components, respectively.

The equations of motion can be derived in terms of displacement w and shear angle β through classical variational principles such as Hamilton's principle. This gives

$$E^*I(w'''' + \beta''') + \rho A\ddot{w} = 0, \tag{2}$$

$$E^*I(w''' + \beta'') - A\kappa G^*\beta = 0, \tag{3}$$

along with the associated boundary conditions. Here, κ is the shear correction factor, and E^*, G^* are the effective beam bending and shear moduli, respectively.

Piezoelectric devices

In Figure 1, the piezoelectric devices are divided into the upper (beam 2) and lower devices (beam 3). The governing equation of motion of each device is derived according to the same procedure as for the main structure (beam 1).

Since the distribution of the normal stress (σ_{xx}) is related to the electric filed, the constitutive relations of the piezoelectric devices are undefined. In the absence of the subsidiary compatibility and constitutive conditions, classical variational principles cannot be applied to the present problem. However, these principles can be generalized by the introduction of Lagrange multipliers to yield a family of variational principles that includes the Hellinger-Reissner principle in elastodynamical problems and the Hu-Washizu principle in elastic static problems.

In this work, the **Hu – Washizu** principle will be modified to include the virtual work done by the inertial and electrical forces. This yields the following functional:

$$J = \int_{t_1}^{t_2} \{ \int_v [\rho p_i \dot{u}_i - \frac{1}{2}\rho p_i p_i - \underline{A}(\epsilon_{ij}) + D_i \bar{E}_i + (\epsilon_{ij} - \frac{1}{2}(u_{i,j} + u_{j,i}))\sigma_{ij} + X_i u_i]dv + \int_{s_1} \bar{g}_i u_i ds + \int_{s_2} g_i(u_i - \bar{u}_i)ds + \int_{s_a} QV ds\}dt \quad (4)$$

where ρ is the density, $\underline{A}(\epsilon_{ij})$ is the strain energy density function, the g_i are the respective surface tractions, Q is the electrical charge applied on the surface s_a, V is the electric potential, D_i are the electric displacements, $\bar{E}_i$ are the electric field, v is the total volume of the system, s is its external surface, and s_a is the external surface that covered by the piezoelectric actuators and sensors.

The functional J of Eq. (4) has stationary values for the actual solution for the independent quantities u_i, p_i, ϵ_{ij}, and σ_{ij}. Therefore, from variational principle, for arbitrary independent variations of δu_i (within conditions $\delta u(t_1) = \delta u(t_2) = 0$), δp_i, $\delta \epsilon_i$, and $\delta \sigma_i$, the first variation of the functional J equals zero, i.e., $\delta J = 0$, and is listed as follows:

$$\delta J = \int_{t_1}^{t_2} \{ \int_v \{(\sigma_{ij,j} + X_i - \rho \dot{p}_i)\delta u_i + (\sigma_{ij} - \underline{A}_{,\epsilon_{ij}})\delta \epsilon_{ij} + D_i \delta \bar{E}_i + [\epsilon_{ij} - \frac{1}{2}(u_{i,j} + u_{j,i})]\delta \sigma_{ij} + [\rho \dot{u}_i - (\frac{1}{2}\rho p_i p_i)_{,p_i}]\delta p_i\} dv + \int_{s_1} (\bar{g}_i - g_i)\delta u_i dS + \int_{s_2} (u_i - \bar{u}_i)\delta g_i ds + \int_{s_a} Q\delta V ds\} dt = 0 \quad (5)$$

The overbarred quantities $\bar{g}_i$ and $\bar{u}_i$ denote the prescribed values of surface tractions and surface displacements, respectively.

The kinematic assumptions for beam 1 and beam 3 is defined as follows:

$$\begin{cases} u_y = 0, \quad u_z = w(x,t), \quad u_x = -z(w' + \beta) \\ p_x = 0, \quad p_y = 0, \quad p_z = P(x,t) \\ \epsilon_{xx} = u' \quad \epsilon_{xz} = -\beta \\ \epsilon_{yy} = \epsilon_{zz} = -\nu\epsilon_{xx} \\ \epsilon_{xy} = \epsilon_{yz} = 0 \\ \sigma_{xx} = \sigma_{xx}(x,z,t) + d_{xz}E(-\bar{E}_z) \quad \sigma_{xz} = \sigma_{xz}(x,z,t) \\ \sigma_{yy} = \sigma_{zz} = \sigma_{xy} = \sigma_{yz} = 0 \\ X_x = X_y = X_z = 0 \end{cases} \quad (6)$$

where $u_i (i = x, y, z)$ and β are the displacements and the shear angle referring to the beams 2 and 3's cartesian axes x, y, z: σ and ϵ represent stress and strain, and X_i and p_i are the body forces and velocity components, respectively.

This kinematic assumptions (6) are substituted into the above formulation (5), whereby the problem is reduced to a form corresponding to the beam model. The equations of motion can be derived in terms of displacement w_i and shear angle β_i by subsequently integrating by parts each term of Eq. (5) and then integrating over the section. This gives

$$E_p I_p (w_j'''' + \beta_j''') + \rho_p A_p \ddot{w}_j = 0. \tag{7}$$

$$E^p I_p (w_j''' + \beta_j'') - A_p \kappa G_p \beta_j = 0. \tag{8}$$

and

$$- d_{xz} Q_p E_p (w_j'' + \beta_j'') + (\xi_{zz}^{\sigma} - E_j d_{xz}^2) \frac{d^2 V}{dz^2} = 0; \quad Q_p = \int z dA_p, \quad j = 2, 3, \tag{9}$$

where A_p, E_p, and G_p are the cross sectional area, the piezoelectric beam bending and shear moduli of the piezoelectric devices, respectively. ξ_{zz}^{σ} and d_{xz} are piezoelectric strain constant and electrical permittivity, respectively.

In Figure 2, the piezoelectric devices are bonded to the upper surface (beam 2) and the lower surface (beam 3). The governing equations of motion (Eqs. 7-9) of each device have been derived. However, the effect of in-plane displacement u_2 and u_3 on the location of the neutral planes of beams 2 and 3, respectively, should be considered in order to achieve the matching conditions.

$$u_2(x) = e_2 w_2'(x); \quad u_3 = e_3 w_3'(x),; \quad w_1(x) = w_2(x) = w_3(x) \tag{10}$$

along the each piezoelectric devices. The distance between neutral planes of the upper device (beam 2) and the main structure (beam 1) is designated e_2. Similarly, e_3 is the distance between the neutral plane of the lower device (beam 3) and the main structure (beam 1). In the piezoelectrically covered region longitudinal motion is thereby coupled with bending

motion, and the longitudinal vibration in the piezoelectric devices changes the bending vibration of the main structure. In other words, in order to describe the bending motion of the beam with bonded piezoelectric devices, in addition to analyze Eqs. (7-9), the governing equation for the longitudinal displacement u,

$$E_p A_p u_j'' - \rho_p A_p \ddot{u}_j = 0, \quad j = 2,3, \tag{11}$$

should be also included.

Local Rayleigh-Ritz Method

In this study, a local Rayleigh-Ritz Method is used, to calculate a piecewise continuous fit to the deflection shape. The displacements, $\hat{w}(x)$, $\hat{\beta}$, and $\hat{u}$ are approximated by using cubic and linear polynomials defined over specific segments of the structure, here it is called subbeam. The coefficients of the polynomials are determined uniquely in terms of the displacements at the end points. The displacements at a point within the ith subbeam are approximated as

$$\hat{w}_i(\eta) = \underline{F}^T(\eta)\underline{w}_i, \quad 0 \le \eta \le l_s, \tag{12}$$

$$\hat{\beta}_i(\eta) = \underline{H}^T(\eta)\underline{\beta}_i, \quad 0 \le \eta \le l_s, \tag{13}$$

and

$$\hat{u}_i(\eta) = \underline{H}^T(\eta)\underline{u}_i, \quad 0 \le \eta \le l_s \tag{14}$$

where $\underline{F} = [F_1, F_2, F_3, F_4]^T$ and $\underline{H} = [H_1, H_2]^T$ are vectors of prescribed (shape) functions of position and $\underline{w}_i$, $\underline{\beta}_i$ and $\underline{u}_i$ are vectors of end transverse deflection and its slope, shear angle, and longitudinal displacement for the i-th subbeam. The shape functions $(F_j)_{j=1,\ldots,4}$ and $(H_j)_{j=1,2}$ are listed in the Appendix.

The static problem is expressed as

$$[K_\epsilon]\,\underline{d} = \underline{f} \tag{15}$$

where $\underline{d}$) is the vector of nodal displacements for the beams 1, 2, and 3 ($\underline{d} = (\underline{w}, \underline{\beta}$ of beam 1: $\underline{d} = (\underline{w_j}, \underline{\beta_j}, \underline{u_j}, V_u, V_b)^T$ for the beam 2, if $j = 2$ and beam 3, if j=3), and $[K_e]$, $\underline{f}$ are the global stiffness matrix and load vector, respectively. The assembly process used to obtain these matrices is symbolically described by

$$(\underline{d}, \underline{f}, [K_e]) = \sum_{i=1}^{N} (\underline{d_i}, \underline{f_i}, [k_i]) \tag{16}$$

where $\underline{d_i}$, $[k_i]$ and $\underline{f_i}$ are the nodal displacement, stiffness matrix and nodal load, respectively, for the i-th element, and the summation extends over all N elements.

The stiffness and mass matrices have dimension 6×6 for the elements located in the main structure (beam 1) and 10×10 in the piezoelectric devices.

RESULTS AND DISCUSSION

The following examples of continuous or distributed sensing/actuation problems were constructed. The first example, as shown in Figure 2, is a cantilevered bimorph piezoelectric beam, which consists of two identical PVDF beams. The second case is a cantilevered aluminum beam with six pairs of piezoelectric PZT sensors bonded to the top and bottom surfaces of the beam.

Piezoelectric PVDF bimorph beam

The first example corresponds to the study deflection of a piezoelectric bimorph beam which is made of two layers of PVDF (piezoelectric polymeric polyvinylidene fluoride) with opposite polarity. material properties of the PVDF is listed in Table 1.

The finite element mesh with ten identical elements to model the bimorph beam is shown in Figure 2. Five elements are used upper and lower layers to approximate the deflection in bending and stretching produced from an applied voltage. The effect of coupling between bending and stretching deflections is included by enforcing the matching conditions of Eq. (10). The static deflections of the 5 nodes are calculated when a voltage, 1 V,

applied across the thickness of the beam. The results are presented in Table 2, in which the nodal deflections calculated by the proposed modeling technique are within 4.4% in comparing with the theoretical predictions obtained by Tsou and Tseng[8]. The predications from their's three dimensional finite element approach using five eleven-noded hexahedron element are 4.4% to 10.0% off of the solution. This probably because the 3-D finite element used in their study was relatively rigid than the present beam element.

The tip deflection of the beam is also calculated and compared with the results from Ref. [8] for various applied voltages rang from 1 to 200 V as shown in Figure 3. The comparison between the calculated deflections based on the present modeling technique and the theoretical predications is in relatively good agreement.

In order to exame the sensing capability of the modeling technique, the output voltages of each sensor (element) are calculated when the tip deflection reaches to 1 cm under an applied tip load. The results are compared with the 3-D finite element predications[8] in Figure 4. It is important to note that the resulted voltage of each sensor calculated in the present approach is a constant value which is quite different than the continuous voltage distribution along the beam presented in Ref. [8].

Piezoelectric PZT sensered/actuated beam

An example of an active beam that was tested is illustrated in Figure 4. The beam is cantilevered and is made of aluminum having dimensions: $\frac{1}{16}$ in. thick × 1.5 in. wide × 15 in. long. Six pairs of piezoelectric actuating and sensing pads (1.0 in. long × 0.01 in. thick PZT having top & bottom electrodes) is epoxied to the beam, symmetric about the beam axis centered and 1 in. from the support and the free ends. The material properties of the aluminum and PZT are provided in Table 1.

The output voltages of the each sensor pair are calculated when the tip deflection reaches to $\frac{3}{16}$ in. (=0.047674 cm) under an applied tip load. Calculated voltage of sensor pairs 1, 2, and 3 presented in Table 3 show in close agreement with the experimental done

by Dr. Joseph J. Hollkamp at the Structural Dynamic Branch, Structures Division, Wright Laboratory.

CONCLUSIONS

A modeling technique for the flexural motion of a beam containing distributed piezoelectric devices is presented. It is based on a key kinematic assumption made to satisfy the compatibility requirements in the vicinity of the interfaces between the piezoelectric devices and the main structure. The idea is to include the coupling between longitudinal deflection and bending deflection in the analysis process.

The governing equation and associated boundary conditions are derived under the generalized variational principle. The derivation procedure is used for the cases of short beams or composite beams which the effect of shear stress concentration near the crack-tips becomes important. The validity of the theory is established by examining static deflections a cantilevered piezoelectric PVDF bimorph beam as well as a cantilevered aluminum beam containing 6 pairs of distributed PZT sensors and actuators. The analytical solutions show excellent agreement with experimental results and theoretical solutions.

Th the proposed modeling technique does not required to integrate the piezoelectric devices into the governing equation derivation process for the main structure. In other words, the proposed modeling processes don't need a special procedure (or a special finite element) to compute the stiffness and mass matrices for the beam sections which are bonded by the piezoelectric devices. Therefore, the proposed modeling technique can be easily implemented to any general-purpose finite element code or other models for structures. Furthermore, the proposed modeling technique could be further developed as a design procedure for designing 2-D and 3-D complicated structures containing distributed piezoelectric sensors and actuators.

Appendix

The shape function for the sub-beams Eqs. (12)-(14) are given by:

$$F_1 = 1 - 3(\frac{\eta}{l})^2 + 2(\frac{\eta}{l})^3 \quad F_2 = \eta - 2\frac{\eta^2}{l} + \frac{\eta^3}{l^2}$$

$$F_3 = 3(\frac{\eta}{l})^2 - 2(\frac{\eta}{l})^3 \quad F_4 = -\frac{\eta^2}{l} + \frac{\eta^3}{l^2}$$

and.

$$H_1 = 1 - \frac{\eta}{l_s} \quad H_2 = \frac{\eta}{l_s}$$

References

[1] Robbins, D. H. and Reddy, J. N., "Finite Element Analysis of Piezoelectrically Activated Beams," *Computers and Structures*, Vol. 41, No. 2, 1991, pp. 265-279.

[2] Lin, M. W. and Rogers, C. A., "Analysis of A Beam Structure With Induced Strain Actuators Based On An Approximated Linear Shear Stress Field," *Proceeding of the Conference on Recent Advances In Adaptive and Sensory Materials and Their Applications*, 1992, pp. 363-376.

[3] Lin, M. W. and Rogers, C. A., "Modeling of The Actuation Mechanism In A Beam Structure With Induced Strain Actuators," *Proceeding of the 34th Structures, Structural Dynamics and Materials Conference*, AIAA/ASME/ASCE/AHS, 1993, pp. 3608-3617.

[4] Kulkarni, G. and Hanagud, S. V., "Modeling Issues in The Vibration Control With Piezoceramic Actuators," *Proceeding of 1991 ASME Winter Annual Meeting, Smart Structures and Materials*, AD-Vol. 24/AMD-Vol. 123, pp. 7-17.

[5] Crawley, E. F. and de Luis, J.,"Use of Piezoelectric Actuators as Elements of Intelligent Structures," *AIAA Journal*, Vol. 25, No. 10, 1987, pp. 1373-1385.

[6] Crawley, E. F. and Lazarus, K. B.,"Induced Strain Actuation of Isotropic and Anisotropic Plates," *AIAA Journal*, Vol. 29, No. 6, 1991, pp. 944-951.

[7] Ha, S. K., Keilers, C. and Chang, F. K.,"Finite Element Analysis of Composite Structures Containing Distributed Piezoceramic Sensors and Actuators," *AIAA Journal*, Vol. 30, No. 3, 1992, pp. 772-780.

[8] Tzou, H. S. and Tseng, C. I., "Distributed Vibration Control and Identification of Coupled Elastic/Piezoelectric Systems: Finite Element Formulation and Application," *Mechanical Systems and Signal Processing*, Vol. 5, No. 3, 1991, pp. 215-231.

PVDF		PZT		Aluminum	
E_p	2.0×10^9	E_p	63.0×10^9	E	73.0×10^9
G_p	7.75×10^8	G_p	23.3×10^9	G	45.6×10^9
ρ	1800 kg/m^3	ρ	7600 kg/m^3	ρ	26300 kg/m^3
ν	0.29	ν	0.35	ν	0.3
d_{31}	2.2×10^{-11} C/N	d_{31}	37.0×10^{-11} C/N		

Table 1: Material properties of aluminum beam and piezoelectric materials.

Node	1	2	3	4	5
Theory[8]	0.138	0.552	1.24	2.21	3.45
3-D EFM[8]	0.124	0.508	1.16	2.10	3.30
Erroe (%)	(10.0)	(8.0)	(6.2)	(5.1)	(4.4)
1-D FEM	0.132	0.528	1.188	2.112	3.30
Error (%)	(4.3)	(4.3)	(4.2)	(4.4)	(4.4)

Table 2: Static deflection of the piezoelectric bimorph beam (10^{-7}m).

Sensor	1	2	3	4	5	6
Specimen 1	11.0	8.8	7.26	N.A.	N.A.	N.A.
Specimen 2	11.5	8.4	7.23	N.A.	N.A.	N.A.
Specimen 3	11.25	8.9	7.25	N.A.	N.A.	N.A.
1-D FEM	10.98	9.04	7.1	5.15	3.2	1.26
Error (%)	(2.4)	(3.8)	(1.9)	N.A.	N.A.	N.A.

Table 3: Sensor voltage distribution of the bending deflection.

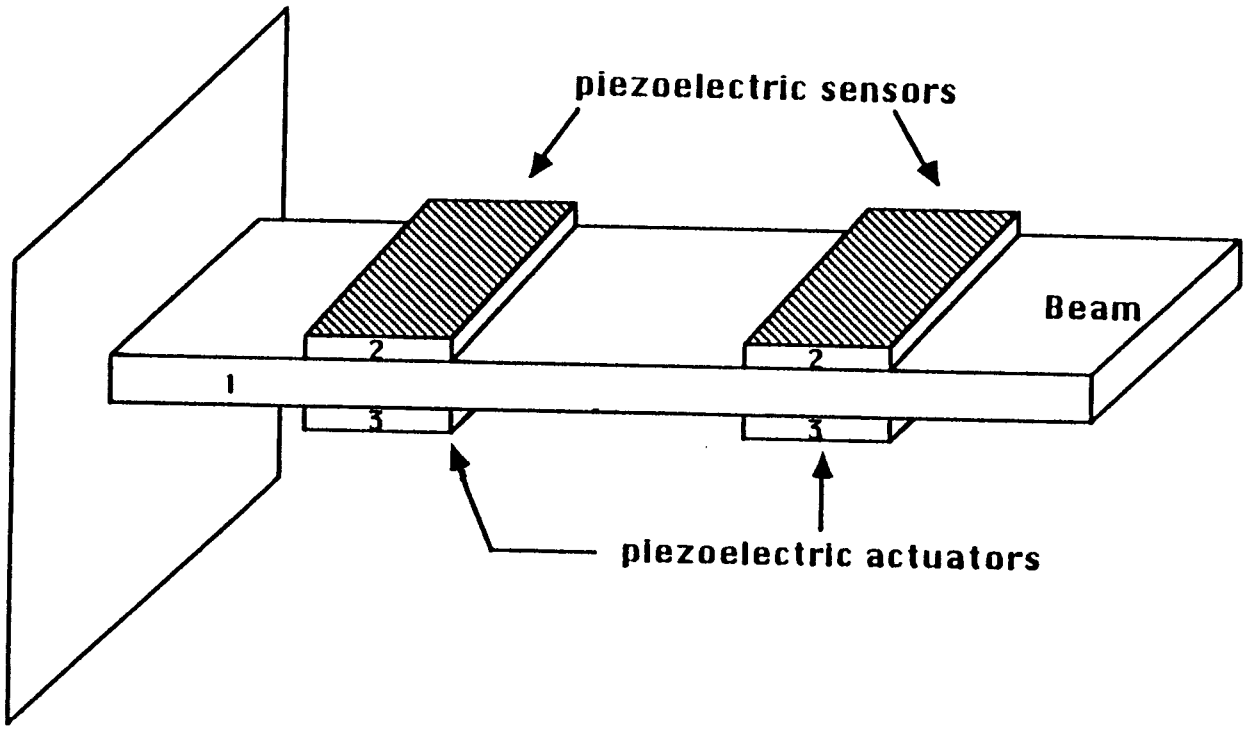


Figure 1. Geometry of cantilevered beam with piezoelectric devices.

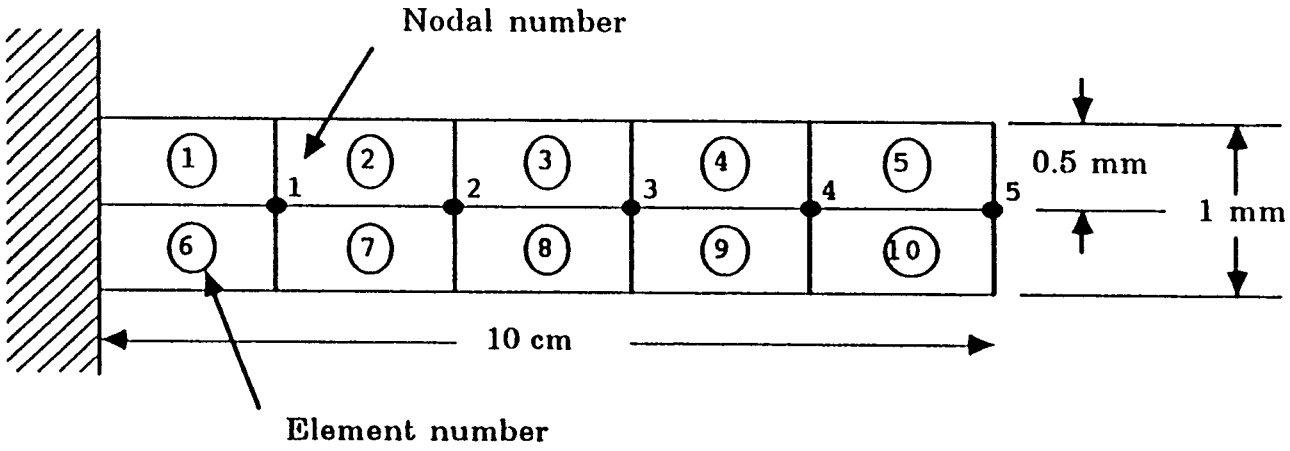


Figure 2. Finite element mesh for the piezoelectric bimorph beam.

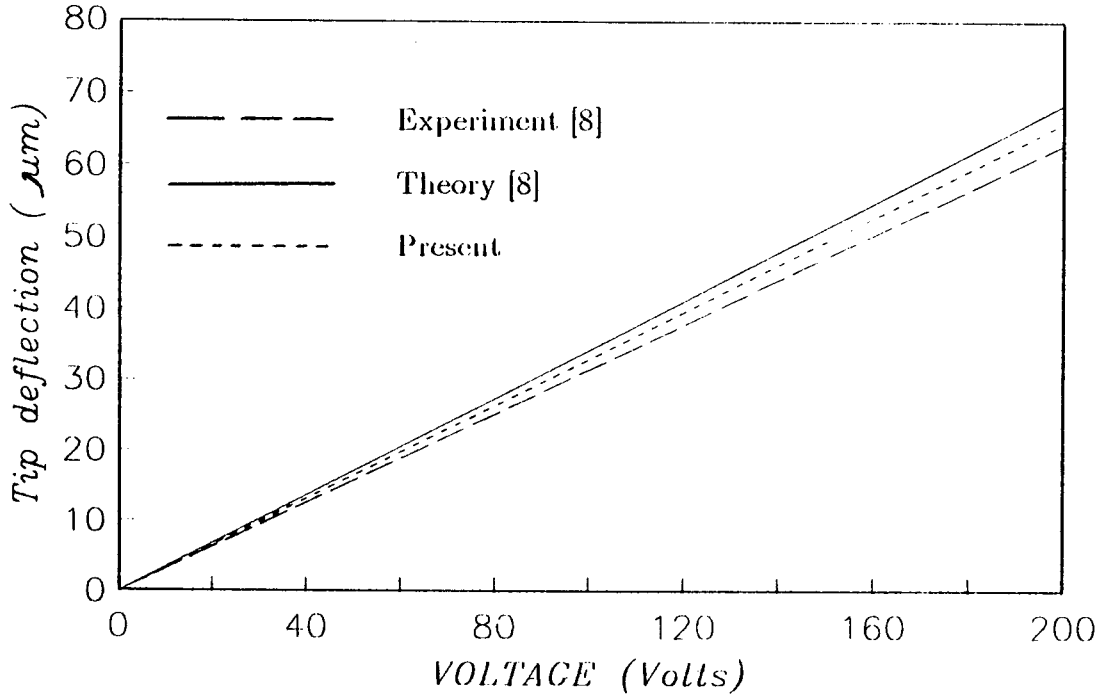


Figure 3. Tip deflection of the bimorph beam in terms of input voltage.

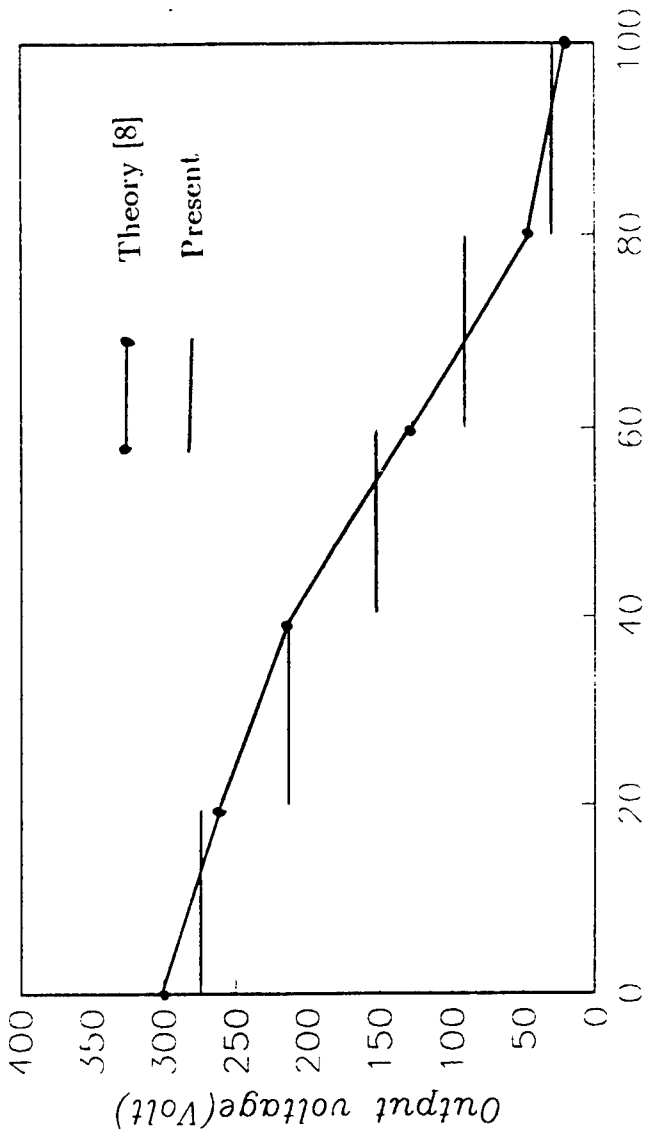


Figure 4. Sensor voltage distribution for 1 cm tip bending defection.

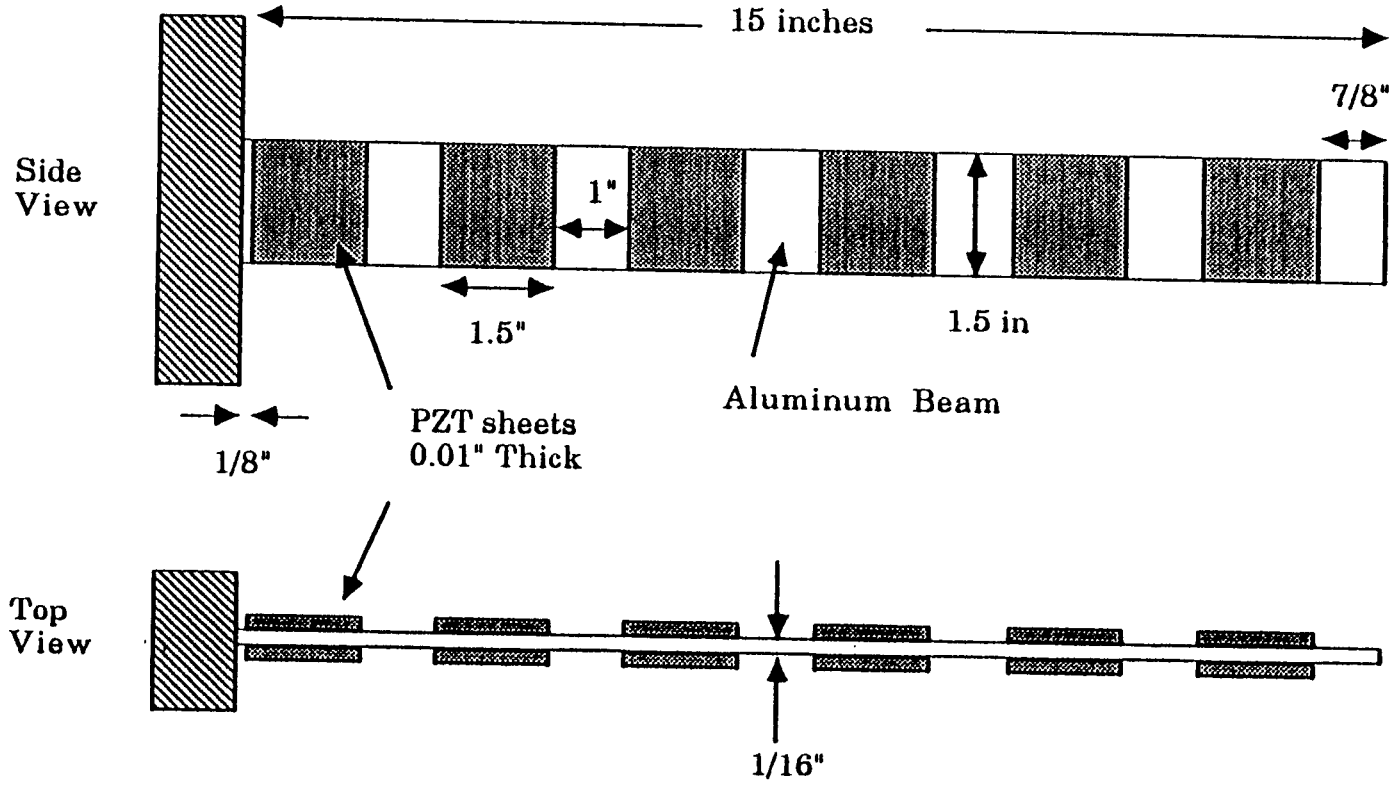


Figure 5. Geometry of cantilevered beam with 6 pairs of piezoelectric devices.

COLLOCATED INDEPENDENT MODAL CONTROL WITH SELF-SENSING ORTHOGONAL PIEZOELECTRIC ACTUATORS (Theory and Experiment)

H. S. Tzou[1] and J. J. Hollkamp[2]

[1] Department of Mechanical Engineering
University of Kentucky
Lexington, KY 40506–0046

[2] Wright Laboratory
Flight Dynamics Directorate
WL/FIBG, WPAFB Ohio 45433

Final Report for:

Summer Research Extension Program
Wright Laboratory

Sponsored by:

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

University of Kentucky

and

Wright Laboratory
Flight Dynamics Directorate

August 20, 1993

COLLOCATED INDEPENDENT MODAL CONTROL WITH SELF–SENSING ORTHOGONAL PIEZOELECTRIC ACTUATORS

(Theory and Experiment)

H. S. Tzou[1] and J. J. Hollkamp[2]

[1] Department of Mechanical Engineering
University of Kentucky
Lexington, KY 40506–0046

[2] Wright Laboratory
Flight Dynamics Directorate
WL/FIBG, WPAFB Ohio 45433

ABSTRACT

Distributed self–sensing piezoelectric actuators provide a perfect collocation of sensors and actuators in closed–loop structural controls. To achieve independent control of various natural modes, spatially distributed self–sensing orthogonal piezoelectric actuators are proposed in this study. A generic spatially shaped orthogonal sensor/actuator theory is derived first, followed by an application to a Bernoulli–Euler beam. Spatially distributed orthogonal sensors/actuators are designed based on the modal strain functions and they are fabricated using a 40μm piezoelectric polymer. A cantilever beam laminated with these self–sensing orthogonal piezoelectric actuators is tested. Collocated independent modal control of the cantilever beam with spatially distributed self–sensing orthogonal actuators is demonstrated and control effectiveness studied.

COLLOCATED INDEPENDENT MODAL CONTROL WITH SELF–SENSING ORTHOGONAL PIEZOELECTRIC ACTUATORS (Theory and Experiment)

H. S. Tzou and J. J. Hollkamp

INTRODUCTION

A perfect sensor/actuator collocation usually provides a stable performance in closed–loop feedback controls. A self–sensing piezoelectric actuator is a single piece of piezoelectric device simultaneously used for both sensing and control. (The sensor signal is separated from the control signal by using a differential amplifier; this signal is then amplified and fed back to induce control actions.) Self–sensing piezoelectric actuators have been proposed in recent years. Dosch, Inman, and Garcia (1992) proposed a self–sensing piezoelectric actuator for collocated control of a cantilever beam. Anderson, Hagood, and Goodliffe (1992) presented an analytical modeling of the self–sensing actuator system, and studied its applications to beam and truss structures. Rectangular–shape piezoelectric devices attached near the fixed end were used in both studies.

It is known that the spatially distributed orthogonal sensors and actuators are sensitive to a mode or a group of natural modes (Tzou, 1993; Lee, 1992). Spatially distributed piezoelectric sensors and actuators were investigated in a number of recent studies, such as beams, plates, rings, shells, etc. (Lee and Moon, 1990; Lee, 1992; Anderson and Crawley, 1991; Collins, Miller, and von Flotow, 1991; Hubbard and Burke, 1992; Tzou and Fu, 1993a&b; Tzou and Tseng, 1990; Tzou, Zhong, and Natori, 1993; Tzou, 1993; Tzou, Zhong, and Hollkamp, 1994). Based on the modal orthogonality, a spatially shaped self–sensing orthogonal modal actuator is effective to only a single mode; consequently, each vibration mode can be independently controlled, *independent modal control*, while the feedback control system is kept simple. This paper is to investigate the sensing and control characteristics of self–sensing orthogonal modal actuators. A generic theory for a spatially distributed self–sensing orthogonal actuator is proposed first, followed by an experimental study of self–sensing orthogonal piezoelectric actuators. Independent modal control with the self–sensing orthogonal actuators are demonstrated. (Note that the emphasis is placed on the experimental aspect.)

THEORY

It is assumed that a spatially distributed piezoelectric layer is laminated on a one–dimensional (1–D) structure, such as arches, rings, beams, rods, etc., Figure 1. Both the piezoelectric layer and the elastic continuum have a constant thickness. It is assumed that the piezoelectric material is hexagonal symmetrical such that the piezoelectric constants $e_{31} = e_{32}$.

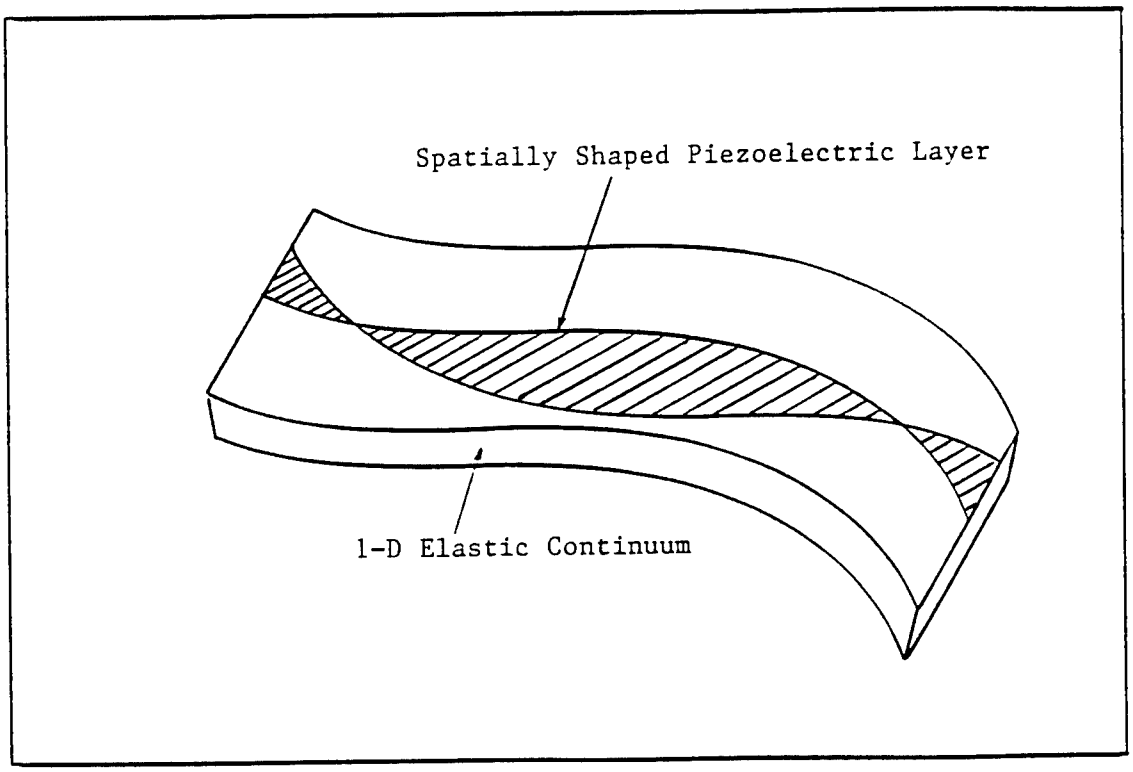


Fig.1 A 1–D spatially distributed orthogonal sensor/actuator.

An open–circuit sensor signal ϕ_3^s from a 1–D spatially distributed orthogonal sensor can be estimated from its strains:

$$\phi_3^s = -\frac{e_{31}}{\epsilon_{33}S^e}\int_{\alpha_1} \mathrm{sgn}[U_3(\alpha_1)]\left[h^s(S_{11}^o+S_{22}^o) + 0.5h^s(h+h^s)(k_{11}+k_{22})\right] A_1A_2\int_0^{W_s(\alpha_1)} d\alpha_2\, d\alpha_1$$

$$= -\frac{e_{31}}{\epsilon_{33}S^e}\int_{\alpha_1} W_s(\alpha_1)\ \mathrm{sgn}[U_3(\alpha_1)]\ \Big[h^s(S^o_{11}+S^o_{22}) + 0.5h^s(h+h^s)(k_{11}+k_{22})\Big]\ A_1A_2\ d\alpha_1\,, \tag{1}$$

where e_{31} is the piezoelectric constant; ϵ_{33} is the dielectric constant; S^e is the effective electrode area; $W_s(\alpha_1)$ is a 1–D shape function; $U_3(\alpha_1)$ is an orthogonal function; h^s is the thickness of the sensor layer; h is the continuum thickness; S^o_{ii} is the membrane strain; k_{11} denotes the bending strain; A_1 and A_2 are Lamé parameters; and sgn[·] is a signum function which defines the polarity changes of the orthogonal sensor. Note that S^o_{22} and k_{22} are usually neglected since $\partial(\cdot)/\partial\alpha_2 = 0$. In addition, the first term (leading by h^s) denotes the membrane strain contribution to the sensor output, and the second term (leading by $0.5h^s$) the bending strain contribution. The total output signal is contributed by the sum of membrane and bending strain components. In a 1–D elastic continuum with finite radius of curvatures, e.g., arches and rings, both membrane and bending components contribute to the output signal. However, for flat 1–D continua with infinite radius of curvature, e.g., beams, the output signal is contributed either by the membrane component, e.g., rods, or the bending component, e.g., beams (Tzou, 1993).

The distributed velocity (strain–rate) feedback can be derived using the modal expansion method and a spatially distributed modal feedback force (Tzou, Zhong, Hollkamp, 1994). The k–th modal equation can be written as

$$\ddot{\eta}_k + \frac{c}{\rho h}\dot{\eta}_k + \omega_k^2\eta_k = \frac{1}{\rho h N_k}\sum_{j=1}^{3}\sum_{m=1}^{\infty}\int_{\alpha_1}\int_{\alpha_2}\Big(\mathcal{G}^{vf}_{jm}(\alpha_1,\alpha_2)\dot{\eta}_m(t)U_{jk}(\alpha_1,\alpha_2)\Big)A_1A_2d\alpha_1d\alpha_2\,. \tag{2}$$

where η is a modal coordinate; c is the damping constant; $\mathcal{G}^{vf}_{jm}(\alpha_1,\alpha_2)$ is the distributed *velocity feedback function*; and $N_k = \int_{\alpha_1}\int_{\alpha_2}[\sum_{j=1}^{\infty}(U_{jk})^2]A_1A_2d\alpha_1d\alpha_2$. Using the modal orthogonality, one can write the distributed ***velocity feedback function*** as

$$\mathcal{G}^{vf}_{jm}(\alpha_1,\alpha_2) = \mathcal{G}^{vf}_{i}U_{3n}(\alpha_1,\alpha_2)\,, \tag{3}$$

where $\mathcal{G}_i^{vf}$ is a velocity weighting factor (gain constant). All modes other than the n = k mode are filtered out due to their orthogonalities. Considering the transverse oscillation only, one can derive the n–th modal equation (*independent modal control equation*) with the ***velocity modal feedback control force*** as

$$\ddot{\eta}_n + \frac{1}{\rho h}\left(c - \frac{\mathcal{G}_3^{vf}}{N_n}\int_{\alpha_1}\int_{\alpha_2} U_{3n}^2\, A_1 A_2 d\alpha_1 d\alpha_2\right)\dot{\eta}_n + \omega_n^2\eta_n = 0\,. \tag{4}$$

For 1–D continua, the transverse mode shape U_{3n} is only a function of one coordinate α_1, e.g., the circumferential direction in rings and arches, the longitudinal direction in beams and rods, etc. If electrode areas of the actuators are designed as a 1–D shape function of $W(\alpha_1)$, the modal control force for a 1–D spatially shaped actuator can be rewritten as

$$\ddot{\eta}_n + \frac{1}{\rho h}\left(c - \frac{\mathcal{G}_3^{vf}}{N_n}\int_{\alpha_1} W(\alpha_1) U_{3n}^2\, A_1 A_2 d\alpha_1\right)\dot{\eta}_n + \omega_n^2\eta_n = 0\,. \tag{5}$$

Note that the modal coupling and the spillover from all other natural modes are eliminated. This modal filtering characteristics will be demonstrated in an experimental study on a cantilever beam laminated with orthogonal sensors/actuators presented later.

Orthogonal Sensor/Actuator for a Cantilever Beam

A 1–D cantilever Bernoulli–Euler beam usually exhibits transverse oscillations only. (The in–plane longitudinal oscillation is neglected.) The Lamé parameters for a flat uniform beam are $A_1 = 1$, $A_2 = 1$; the radii are $R_1 = \infty$ and $R_2 = \infty$. In addition, $\partial(\cdot)/\partial\alpha_2 = 0$. Accordingly, the closed–loop equation of motion of a cantilever beam can be derived.

$$\rho h\ddot{u}_3 + YI\frac{\partial^4 u_3}{\partial x^4} - b\,\frac{\partial^2 (M_{11}^a)}{\partial x^2} = bF_3\,, \tag{6}$$

where ρ is the mass density; Y is Young's modulus; I is the area–moment of inertia; b is the beam width; M_{11}^{a} is the induced control moment; and F_3 is the external mechanical force. As discussed previously, the orthogonal modal sensors/actuators are designed based on the *modal function* $U_{3m}(x)$:

$$U_{3m}(x) = \frac{1}{\lambda_m^2} \frac{d^2 U_{3m}(0)}{dx^2} \left[C(\lambda_m x) - \frac{A(\lambda_m L)}{B(\lambda_m L)} D(\lambda_m x) \right] , \tag{7}$$

where

$$A(\lambda_m x) = 0.5[\cosh(\lambda x) + \cos(\lambda x)] , \tag{8a}$$

$$B(\lambda_m x) = 0.5[\sinh(\lambda x) + \sin(\lambda x)] , \tag{8b}$$

$$C(\lambda_m x) = 0.5[\cosh(\lambda x) - \cos(\lambda x)] , \tag{8c}$$

$$D(\lambda_m x) = 0.5[\sinh(\lambda x) - \sin(\lambda x)] , \tag{8d}$$

where x defines the distance measured from the fixed end. The eigenvalue λ_m is determined by its characteristic equation:

$$\cos(\lambda L) \cosh(\lambda L) + 1 = 0 , \tag{9}$$

where $\lambda_1 L = 1.875$; $\lambda_2 L = 4.694$; $\lambda_3 L = 7.855$; $\lambda_4 L = 10.996$; $\lambda_5 L = 14.137$; etc. L is the beam length. The first derivative $\frac{d}{dx}\left[U_{3m}(x)\right]$ is the *modal slope function* and the second derivative $\frac{d^2}{dx^2}\left[U_{3m}(x)\right]$ is the *modal strain function.* The *modal strain function* is used to define the shapes of orthogonal modal sensors/actuators:

$$U''_{3m}(x) = \left[\left[e^{(\lambda L - \lambda x)} [e^{\lambda L} + \cos(\lambda L) + \sin(\lambda L)]/2[e^{2\lambda L} + 2e^{\lambda L} \sin(\lambda L) - 1] \right] \right.$$
$$\left. + \left[- e^{\lambda x} \{ e^{\lambda L} [0.5\cos(\lambda L) - 0.5\sin(\lambda L)] + 0.5 \} \right. \right.$$

$$+ e^{2\lambda L}[0.5\cos(\lambda x) - 0.5\sin(\lambda x)] + e^{\lambda L}[\cos(\lambda x)\sin(\lambda L) - \sin(\lambda x)\cos(\lambda L)]$$
$$- 0.5\cos(\lambda x) - 0.5\sin(\lambda x)\Big] / [e^{2\lambda L} + 2e^{\lambda L}\sin(\lambda L) - 1]\Big] / (\lambda L)^2 . \qquad (10)$$

Note that each orthogonal modal sensor/actuator has a distinct shape based on its modal strain function and eigenvalue. Detailed layouts of the spatially shaped orthogonal sensors/actuators are presented next.

MODEL FABRICATION AND EXPERIMENTAL SETUP

The shapes of distributed orthogonal sensors/actuators follow the definitions of modal strain functions defined by their eigenvalues. The first four modal function are plotted in Figure 2, and their modal strain functions are plotted in Figure 3. Note that the effective regions are from zero to one, since they are normalized in the length direction.

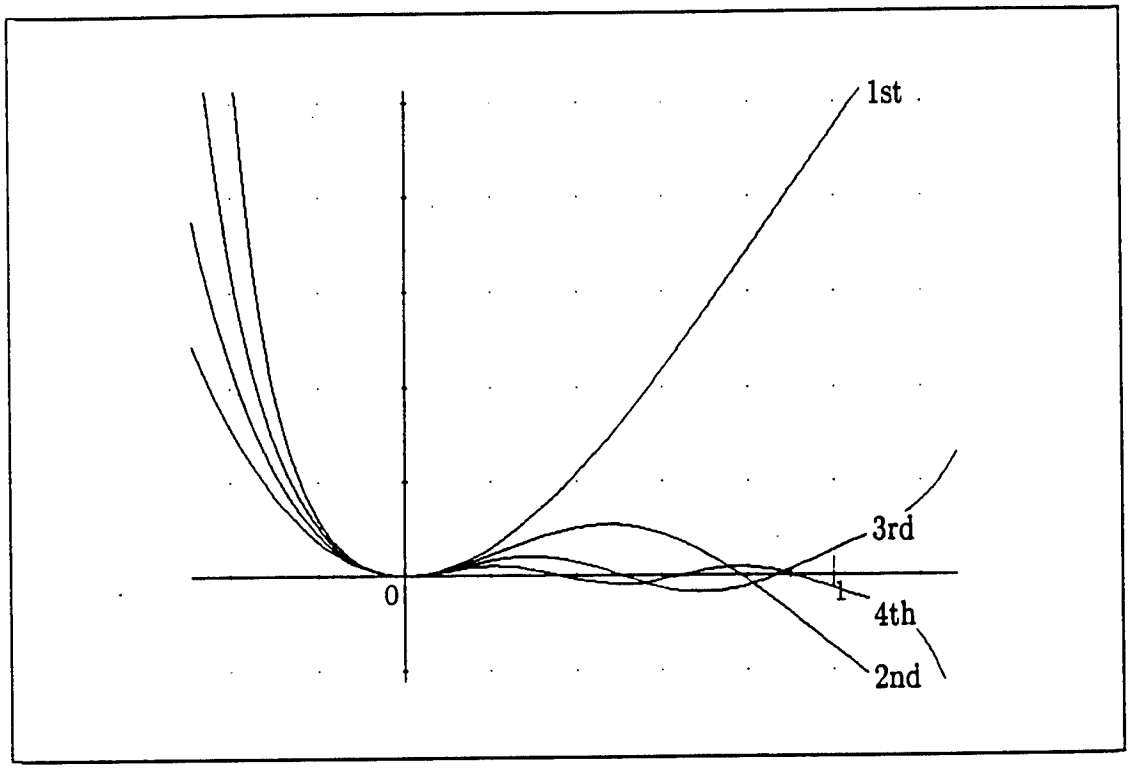


Fig.2 Mode shape functions of the cantilever beam.

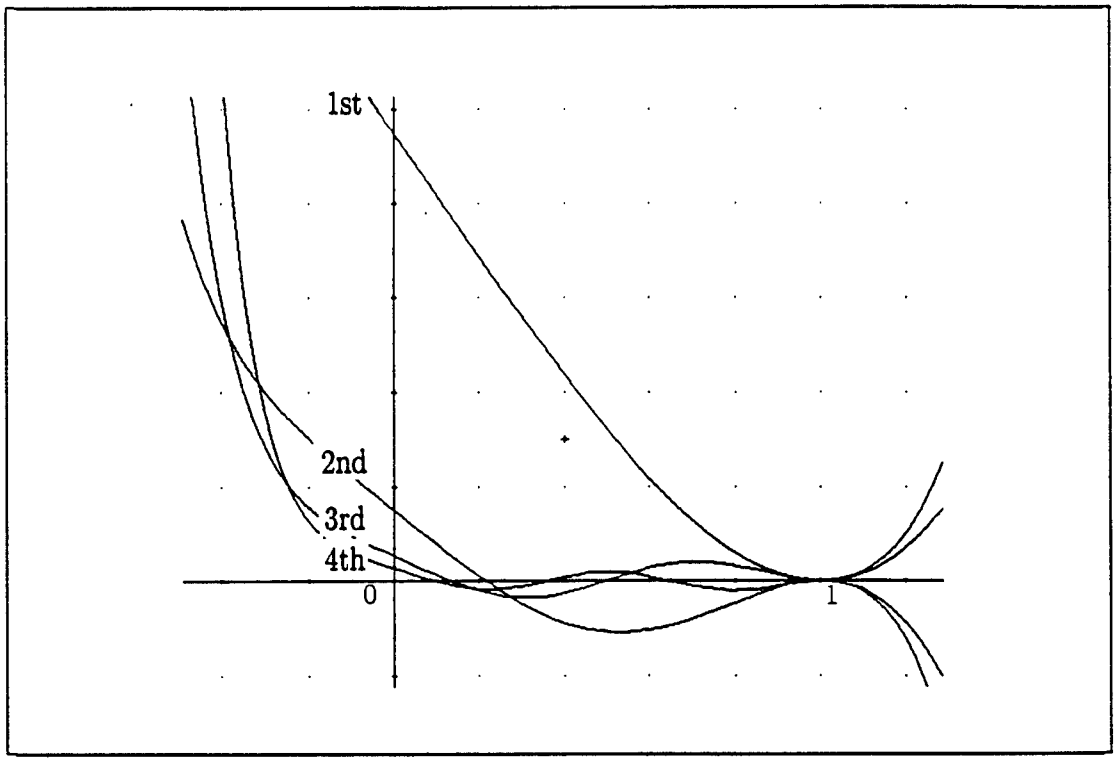


Fig.3 Modal strain functions of the cantilever beam.

Polymeric piezoelectric polyvinylidene fluoride (PVDF) is flexible and easy to cut into various shapes in a laboratory environment. In this study, a 40μm biaxially oriented PVDF sheet is used for the orthogonal sensors/actuators. These sensor/actuator layers are cut according to their strain functions and then glued on a plexiglas beam (15×1×1/8–in). Patterns of surface electrodes are first laid out on the plexiglas beam using a thin–film silver paste to ensure a good electrical conductivity. Individual silver electrodes are connected by either thin silver–paste lines (internal connections) or .5mm Teflon coated surgical wires (external connections). Polarity changes are achieved by reversing the cut PVDF sheets. The finished PVDF/plexiglas beam is shown in Figure 4.

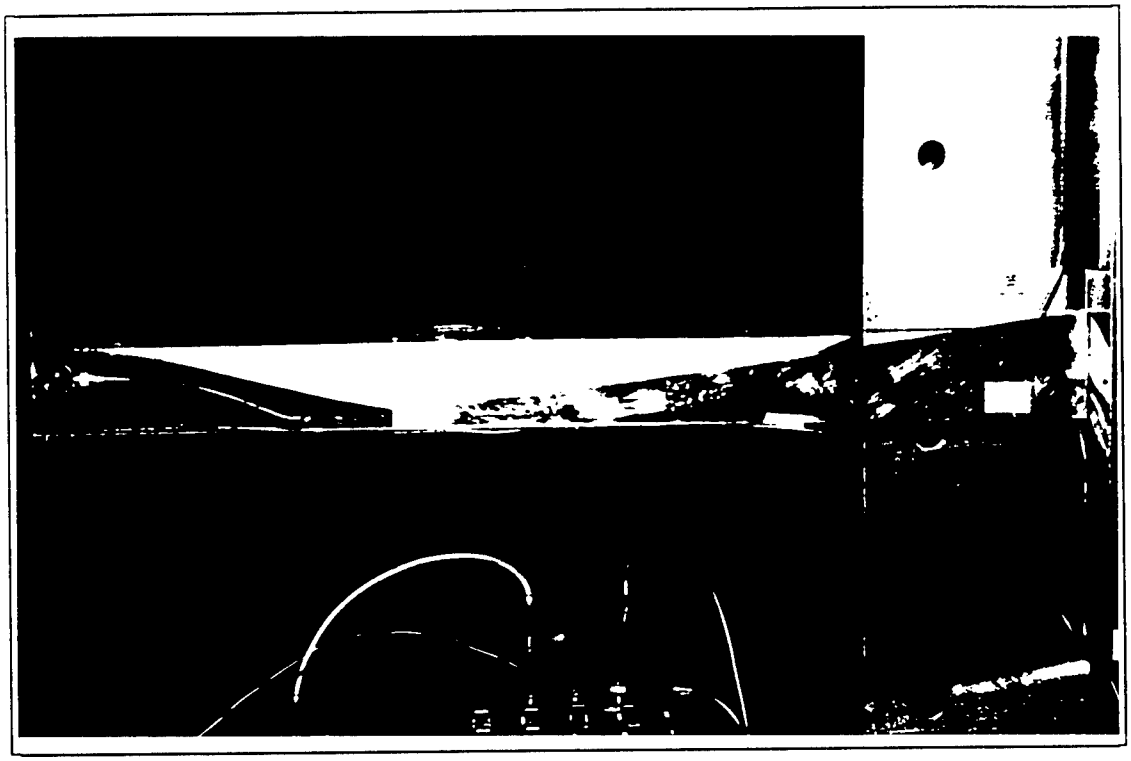


Fig.4 Experimental beam model with orthogonal PVDF sensors/actuators.

Apparatus

A self–sensing feedback control circuit is setup with two current amplifiers and a differential circuit (Anderson, et al., 1992; Dosch, et al., 1992). Three operational amplifiers (AD–711JN), six resistors (24.9kΩ, 8MΩ, and 16MΩ) and a capacitor (14nF) are used to build the circuits for the first and second orthogonal modal sensors/actuators. Figure 5 shows the circuit. The capacitor is used to match the capacitance of the orthogonal piezoelectric sensor/actuator. A power amplifier (BK–1651) supplies a 30V to the operational amplifiers, and an signal amplifier is used to amplify the sensor signal to induce control actions in the piezoelectric layers. A reference accelerometer (Kistler 5205) is mounted at the free end to provide a reference signal. All signals are input into an HP data acquisition system (HP3566A) for signal processing and recording.

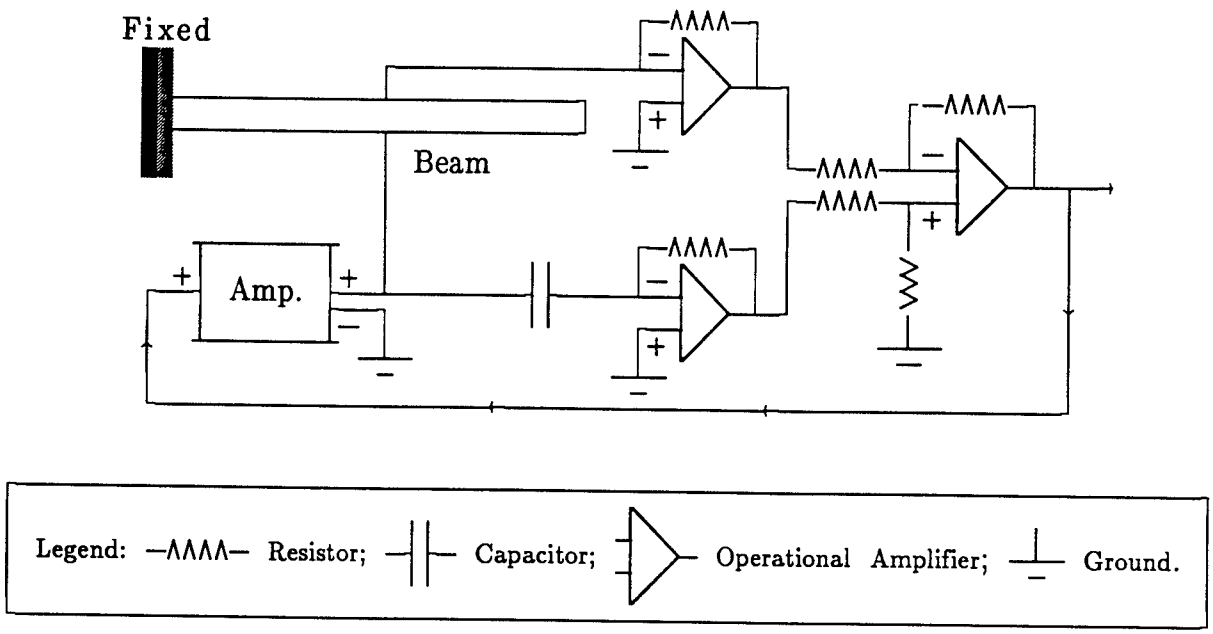


Fig.5 A self–sensing feedback control circuit.

Experimental Procedures

There are two sets of experiments carried out in this study. The first set is to test the modal orthogonality of orthogonal modal sensors; the second set is to evaluate the control effectiveness of the self–sensing orthogonal actuators.

The first set involves two tests: 1) strain signals and 2) strain–rate signals. The strain signal is contributed by elastic bending strains of the cantilever beam; it is ultimately related to the transverse deflection u_3. Thus, the strain signal is often referred as "displacement" signal; the strain–rate can be regarded as the "velocity" signal. The signs of these signals are individually checked to ensure correct feedback signals in the self–sensing feedback control.

A self–sensing feedback control circuit, discussed previously, is used in the second set experiments. Controlled time histories from the accelerometer are acquired; modal damping ratios are calculated using the eigensystem realization algorithm (ERA) method (Juang and Pappa, 1985).

RESULTS AND DISCUSSION

There are two sets of experiments carried out in this study. The first set is to test the modal orthogonality of orthogonal modal sensors; the second set is to evaluate the modal control effectiveness of the self–sensing orthogonal piezoelectric actuators.

Modal Sensing

In order to evaluate the sensing effectiveness of orthogonal modal sensors, the spatially shaped piezoelectric layers were subjected to external excitations and their dynamic responses recorded. Strain and strain–rate responses were also tested using a strain–rate circuit (Lee, 1992). Figure 6 shows the spectra of the first modal sensor, the second modal sensor, and the accelerometer. It is observed that the accelerometer senses multiple modes of the cantilever beam, and the modal sensors only respond to their respective modes.

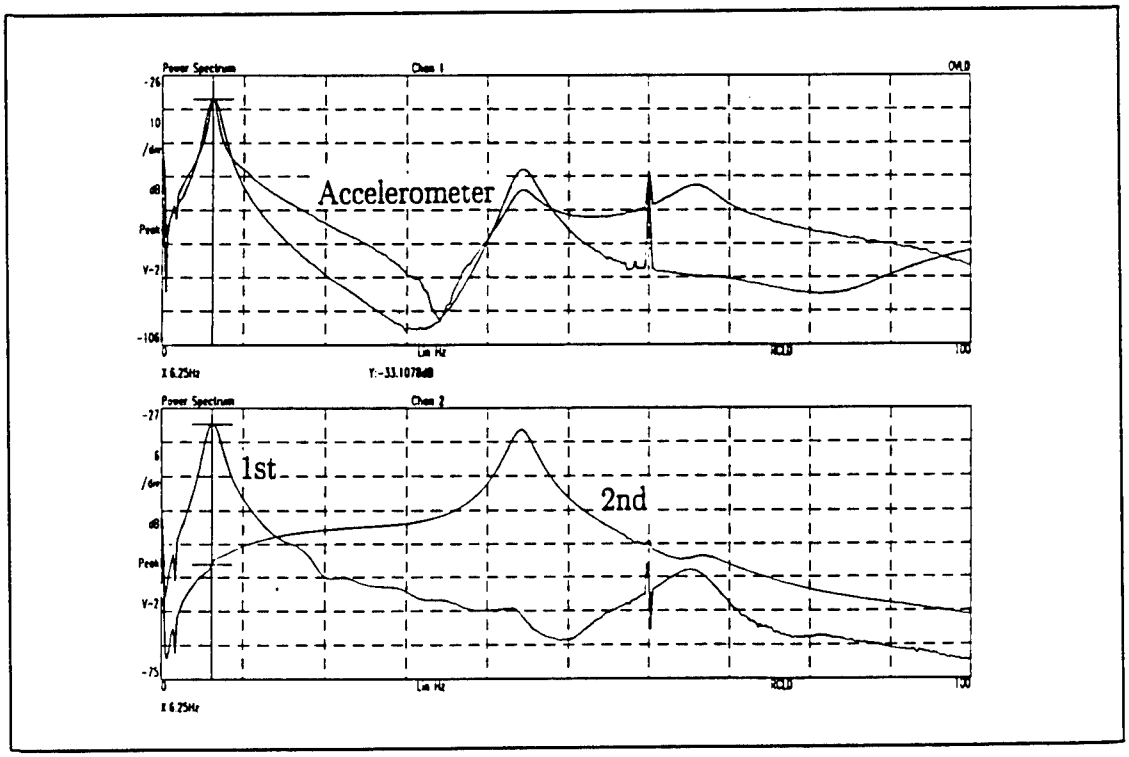


Fig.6 Spectra of the orthogonal modal sensors and the accelerometer.

Self–Sensing and Feedback Control

As discussed previously, a self–sensing piezoelectric actuator provides a perfect collocation of sensor and actuator. In this section, free oscillation and controlled time histories are presented and their respective damping ratios are calculated.

1) Free Oscillations

For the first mode, an initial displacement was applied to the free end and the snap–back response recorded. The free oscillation time histories of strain and strain–rate signals of the first modal sensor/actuator are plotted in Figures 7 and 8, and those of the second sensor/actuator are plotted in Figures 9 and 10, respectively. Note that those time histories of the second sensor/actuator were obtained via impulse excitations.

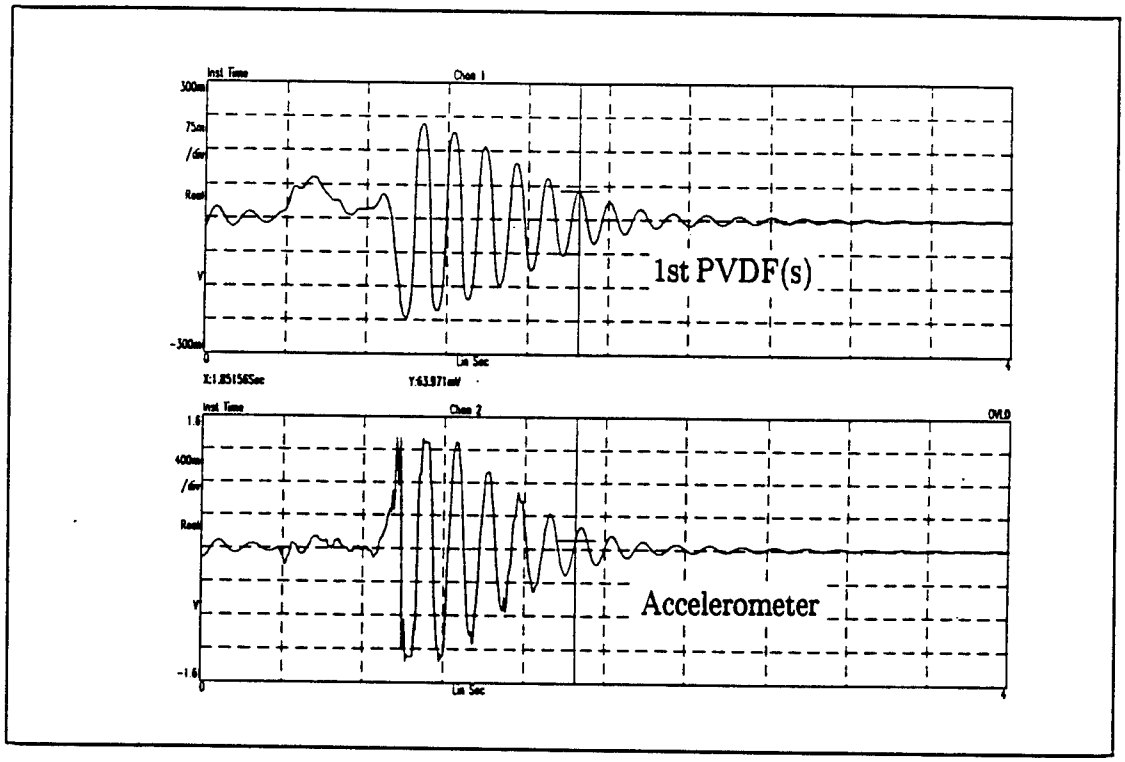


Fig.7 Free oscillation of the plexiglas beam (1st strain).

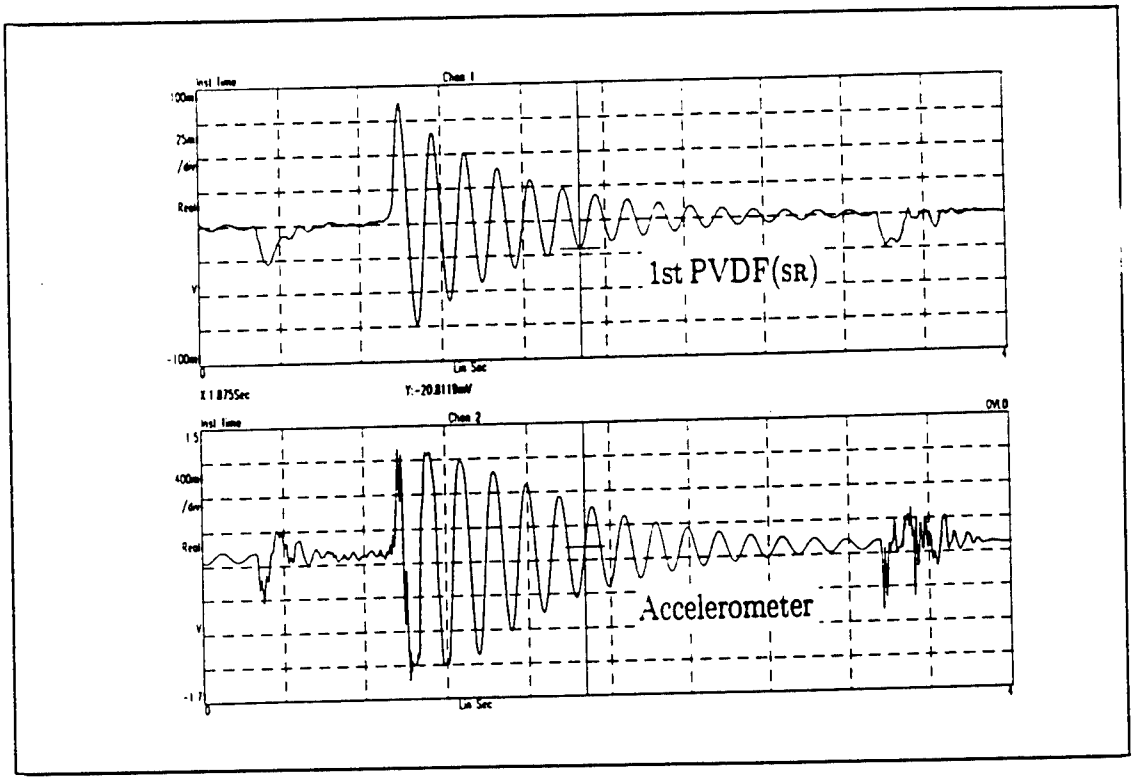


Fig.8 Free oscillation of the plexiglas beam (1st strain–rate).

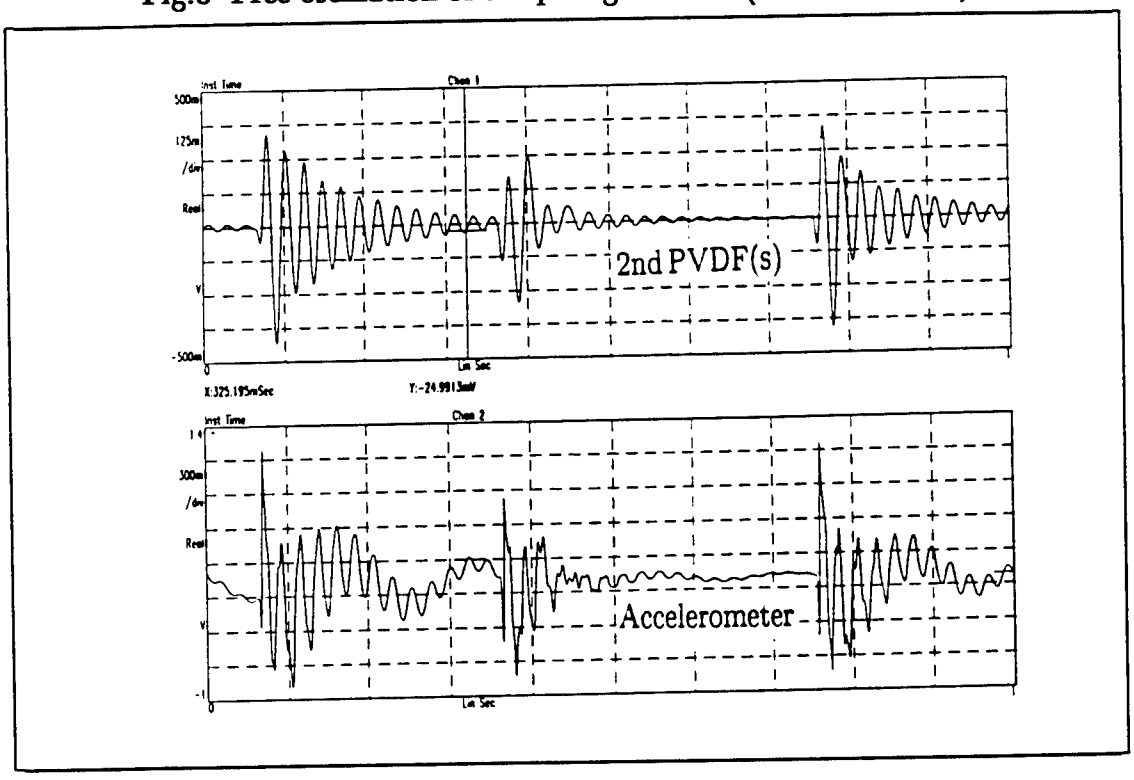


Fig.9 Free oscillation of the plexiglas beam (2nd strain).

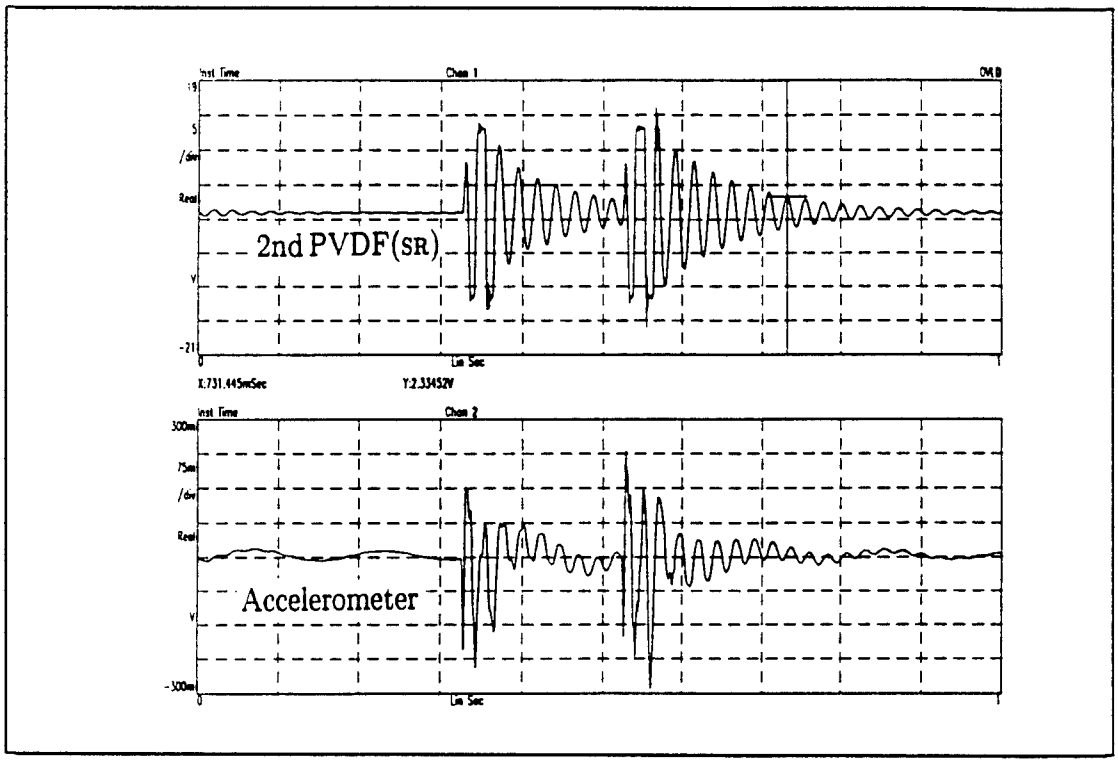


Fig.10 Free oscillation of the plexiglas beam (2nd strain–rate).

Damping ratios of the first and second modes were calculated using the free vibration time histories. The first modal damping ratio is 4.0% and the second modal damping ratio is 3.4%. (Note that these data were calculated using more than five sample time histories.) It should be pointed that there was an accelerometer cable taped on the plexiglas beam, which caused a higher damping for the first natural mode.

2) Self–Sensing Control – Independent Modal Control

Control effectivenesses of the self–sensing orthogonal actuators were evaluated when the self–sensing control circuit was powered on. The sensing (strain–rate) signal was separated from the actuating signal via the circuit shown in Figure 5. The controlled responses (via the accelerometer signals) of the plexiglas beam were plotted in Figures 11 and 12.

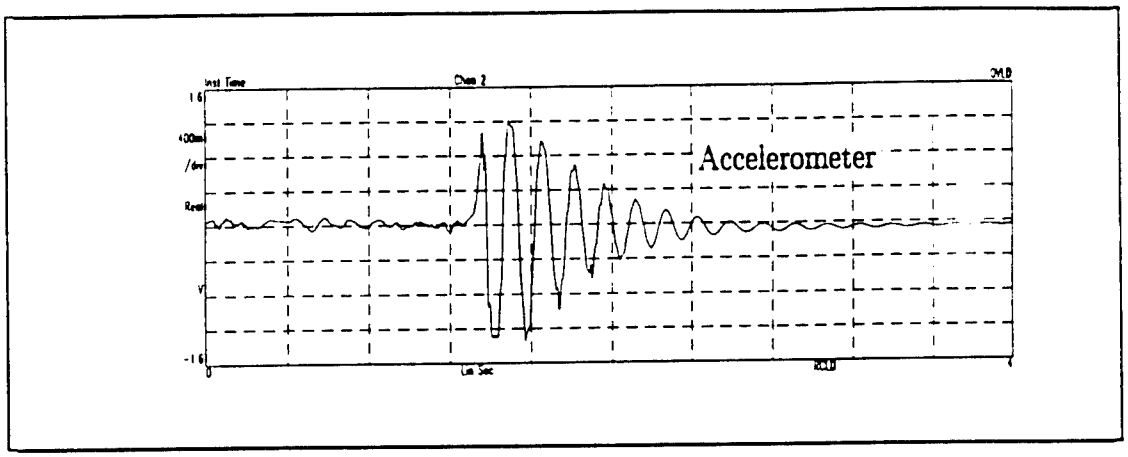


Fig.11 Controlled time history of the plexiglas beam (1st).

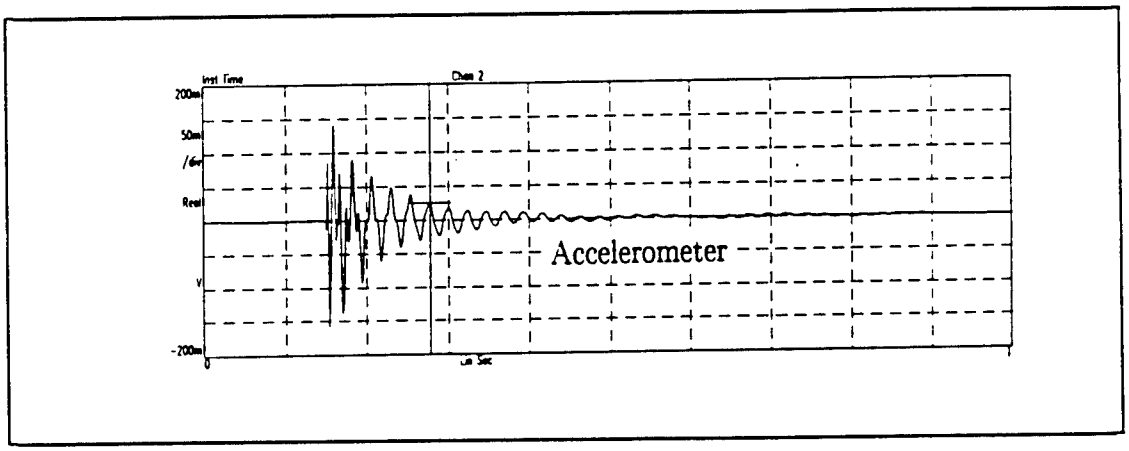


Fig.12 Controlled time history of the plexiglas beam (2nd).

It was noted that the strain–rate signals are rather noisy, which is probably introduced by the electrical line noises, the circuit, the amplifier, etc. Although the strain–rate signals are very noisy, the plexiglas beam is controlled well via the self–sensing orthogonal piezoelectric actuator. The averaged damping ratios of the controlled responses are 7.1% for the first mode and 4.2% for the second mode. (Note that since the second mode was relatively difficult to excite by an impulse excitation, twelve samples were used to obtained the averaged data. Control of the second mode could be improved by changing the resistors in the circuit.) It should be pointed out that the convergence of a modal response is determined by the product of the damping ratio and the modal frequency, i.e.,

$e^{-\zeta_n \omega_n t}$. Consequently, responses of the higher modes usually converge much faster than those of the lower modes.

SUMMARY AND CONCLUSION

Distributed self–sensing piezoelectric actuators provide a perfect collocation of sensors and actuators in closed–loop structural controls. To separate control actions for different natural modes (independent modal control), spatially distributed self–sensing orthogonal piezoelectric actuators were proposed in this study.

A generic orthogonal sensor/actuator theory was presented first, followed by an application to a Bernoulli–Euler beam. Spatially distributed orthogonal sensors/actuators were designed based on the modal strain functions. A physical model was fabricated and its self–sensing control effectiveness tested. A 40μm polymeric piezoelectric PVDF sheets were cut and laminated on a plexiglas beam. Surface electrodes were connected by either silver pastes or surgical wires. A self–sensing feedback control circuit was setup and tested.

Experimental results showed that the orthogonal modal sensors are sensitive to their respective modes. Free and controlled (via the self–sensing feedback control circuit) time histories were recorded and their modal damping ratios calculated. The calculated results suggested that the modal damping ratios were enhanced by 77.5% for the first mode and by 23.5% for the second mode. The convergence of modal responses is determined by the product of the modal damping and the modal frequency. Thus, the independent modal control of continua can be effectively achieved by using the spatially distributed self–sensing orthogonal piezoelectric actuators.

ACKNOWLEDGEMENT

A fellowship supported by the Wright Laboratory (Flight Dynamics Directorate) under the AFOSR RDL Program is gratefully acknowledged. Contents of the information do not necessarily reflect the position or the policy of the government, nor should official endorsement be inferred.

REFERENCES

Anderson, M.S. and Crawley, E.F., 1991, "Discrete Distributed Strain Sensing of Intelligent Structures," *Proceedings of the Second Joint Japan/USA Conference on Adaptive Structures*, pp.737–754.

Anderson, E.H., Hagood, N.W., and Goodliffe, J.M., 1992, "Self–Sensing Piezoelectric Actuation: Analysis and Application to Controlled Structures," AIAA Paper: AIAA–92–2465–CP, 33rd SDM Conference.

Colins, S.A., Miller, D.W., von Flotow, A.H., 1991, "Piezoelectric Spatial Filters for Active Vibration Control," *Recent Advances in Active Control of Sound and Vibration*, Technomic, pp.219–234.

Dosch, J.J., Inman, D., and Garcia, E., 1992, "A Self–Sensing Piezoelectric Actuator for Collocated Control," *J. of Intelligent Material Systems and Structures*, Vol.(3), pp.166–185, Jan. 1992.

Hubbard, J.E. and Burke, S.E., 1992, "Distributed Transducer Design for Intelligent Structural Components," *Intelligent Structural Systems*, Tzou and Anderson (Ed.), Kluwer Academic Publishers, Dordrecht/Boston/London, August 1992, pp.305–324.

Juang, J.N. and Pappa, R.S., 1985, "An Eigensystem Realization Algorithm for Modal Parameter Identification and Model Reduction," *J. of Guidance and Control*, Vol.8, No.5, pp.620–627.

Lee, C.K., 1992, "Piezoelectric Laminates: Theory and Experimentation for Distributed Sensors and Actuators," *Intelligent Structural Systems*, Tzou and Anderson (Ed.), Kluwer Academic Publishers, Dordrecht/Boston/London, August 1992, pp.75–167.

Lee, C.K. and Moon, F.C., 1990, "Modal Sensors/Actuators," ASME *Journal of Applied Mechanics*, Vol.(57), pp.434–441.

Tzou, H.S., 1993, *Piezoelectric Shells (Distributed Sensing and Control of Continua)*, Kluwer Academic Publishers, February 1993.

Tzou, H.S. and Fu, H., 1993a, "A Study on Segmentation of Distributed Sensors and Actuators, Part–1, Theoretical Analysis," *Journal of Sound & Vibration*, Vol.(168), No.19, November 1993. (To appear)

Tzou, H.S. and Fu, H., 1993b, "A Study on Segmentation of Distributed Sensors and Actuators, Part–2, Parametric Study and Vibration Controls," *Journal of Sound & Vibration*, Vol.(168), No.19, November 1993. (To appear)

Tzou, H.S. and Tseng, C.I., 1990, "Distributed Piezoelectric Sensor/Actuator Design for Dynamic Measurement/Control of Distributed Parameter Systems: A Finite Element Approach," *Journal of Sound and Vibration*, Vol.(138), No.(1), pp.17–34.

Tzou, H.S., Zhong, J.P., and Natori, M.C., 1993, "Sensor Mechanics of Distributed Shell Convolving Sensors Applied to Flexible Rings," *ASME Journal of Vibration & Acoustics*, Vol.(115), No.1, pp.40–46, January 1993.

Tzou, H.S., Zhong, J.P., and Hollkamp, J.J., 1994, "Spatially Distributed Orthogonal Piezoelectric Shell Actuators (Theory and Applications)," *Journal of Sound & Vibration*, 1994. (To appear)

(RDL–Report93.Wp/Fibg1)

Development of Air Force Superconductivity Power Technology

Xingwu WANG
Associate Professor
Department of Electrical Engineering
Alfred University
26 North Main Street
Alfred, NY 14802

Final Report for:
Summer Research Extension Program
Wright Laboratory

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

and

Alfred University

September 1993

Development of Air Force Superconductivity Power Technology

Xingwu WANG
Associate Professor
Department of Electrical Engineering
Alf ed University

Abstract

This study is an extension of previous research on Superconductive Magnetic Energy Storage (SMES) technique, carried out last summer. As superconductor market gradually matures, it is necessary to examine other superconductive power techniques such as generators, motors, transformers, chokes, and power lines. Following criteria are applied in the evaluation: efficiency, weight, size, cost, durability, maintainability, and operability. Based on existing status of air base power system, it is suggested that superconductive rotational machinery (>300 KW) is beneficial, and other power devices may be beneficial. Besides further feasibility study, in-house experimentation at an air base is needed to make realistic assessment. In particular, tests on high Tc superconductors, wires and coils are required. Based on estimated savings, it is suggested that R&D funding on Air Force superconductive generators should be approximately one million dollars in 1994-5, which is approximately 0.28% of federal R&D funding on superconductors. Research on superconductive power systems is a suitable topic for air base engineers/scientists due to following reasons: 1. emphasis on mobility and critical needs on new power systems; 2. federal government's requirement on high tech development and military/civilian dual usage.

Development of Air Force Superconductivity Power Technology

Xingwu WANG

I. Introduction

Last summer, we carried out a feasibility study on Superconductive Magnetic Energy Storage (SMES) technique. Following two items were examined: 1. existing low Tc superconductive SMES, 2. critical needs for power conditioning and energy storage. We concluded that SMES will improve existing Air Force power/energy systems; and recommended that a small SMES should be tested in an Air Force facility.[1]

In the past one year, there has been great progress towards fabrications of practical high Tc superconductors.[2] First, practical BiSrCaCuO coils can be manufactured by powder-in-tube techniques.[3] Second, large critical currents, Ic, can be achieved in long rods.[4] Third, operational ranges of these superconductors are wider than before.[5a] Fourth, HgBaCaCuO superconductor has been discovered with Tc of 133 K.[5b] Since the superconductor market has gradually matured, we should examine different superconductive power techniques such as generators, motors, transformers, chokes, and power lines. Following criteria will be applied in the evaluation: efficiency, weight, size, durability, maintainability, and operability.

II. Methodology

Since the discoveries of high Tc oxide superconductors in 1987 and 1988, several DOD agencies have initiated and/or expanded their research efforts on superconductivity and applications.[6a-7b] However, existing investigation on air base power systems is very

limited.[8a]

In this study, we have utilized following resources: 1. publications in scientific journals; 2. documents in DOE laboratories; 3. literatures in Defense Technical Information Centers; 4. discussions with experts in universities, DOE and DOD agencies; 5. consultations with managers and engineers in Headquarters Air Force Civil Engineering Support Agency (AFCESA) and associated agencies familiar with air base power requirements.[8b-d]

III. Superconductive Generators

A. Existing air base generators

Power level of air base generators varies from several KW to several hundred MW, as illustrated in Fig. 1.[8e] While voltage varies from several KV to several hundred MV, Fig. 2. Specific weight of a generator is approximately 10-100 lb/KW;[8f] and specific volume is approximately 0.02-0.05 m^3/KW.[8g] To purchase and install a stationary generator, cost per KW is $300 - $350 in class A generator (750 KW range, full time usage, continuous operation). The cost will be approximately $400 per KW for a power plant generator (larger than 1 MW). To transport a generator via a cargo plane, cost is approximately $0.0003 per pound per mile.[9a] For a 1 MW generator, operating cost is $0.072-$0.079 per kilowatt-hour; and maintenance cost is $0.0002-$0.0030 per kilowatt-hour. These values will be the basis for the evaluation and development of superconductive generators.

B. Conventional generators and superconductive generators

Conventional generators are described in many textbooks and

handbooks.[9b-c] Essential materials of a generator are: winding(s) and electromagnet(s). The coil winding(s) may be a magnetic field winding, and/or an armature winding. These windings are made of a normal conductor with certain resistivity. To conduct large current, the wire cross sectional area of a winding has to be large. The electromagnets may be associated with the field windings and/or the armature windings to create magnetic field distributions with required flux densities. These electromagnets are usually massive. The field winding and associated electromagnet are referred as field winding assembly; and the armature winding and associated electromagnet are referred as armature assembly. If the field winding assembly is rotational, the armature winding will be stationary; and vice versa. The rotational part is called rotor, and the stationary part is called stator.

Conventional generators are heavy and large. There exist many techniques to reduce weight/size, and to increase power generation efficiency. As an example, the field winding assembly can be replaced by a set of permanent magnets. As another example, electrical coil windings can be replaced in cryogenic temperature environment, and resistive dissipation can be reduced.

Since the discovery of superconductive materials in 1911, especially since the commercialization of high current density superconductors in 1960's, superconductive generators have been invented, designed, and tested.

C. Development of superconductive generators

Most of the pioneer work was carried out in the United States.

First superconductive generator was constructed in mid 1960's. Since then, different rotor/stator arrangements have been designed and tested. In earlier designs, the stator has a superconductive field winding; and the rotor has a conventional armature.[10] In recent designs,[11a-b] the rotor has a superconductive field winding and the stator has a conventional armature as illustrated in Fig. 3.[11c] There exist several different kinds of superconductive rotors; i.e., slow response type A, slow response type B, and quick response type.[12a] In some other designs, both windings are all superconductive.[12b-c]

Power levels of US experimental generators have been varied: 45 KVA in 1968, 5 MVA in 1972, 20 MVA in 1978, 50 MVA in 1990, and 300 MVA design goal in mid 1980's.[13a-d] After developed a 30 KVA generator in 1972, Japanese research effort has become intensive, with current design goal of 200 MVA[13e]. Started at 1.5 MVA in 1974, the current design in Russia (USSR) is 1200 MVA. Germany's generators are 0.3 MVA in 1983, 120 MVA in 1987, and 850 MVA (design) in 1993. France's generators are 1 KVA in 1965, 0.5 MVA in 1977, and 20 KVA in 1987. In addition, China had a design of 0.4 MVA in 1975. Load tests on superconductive generators have demonstrated technical availability of superconductive generators. In Appendix I, a summary of worldwide activities is provided.[14]

As far as DOD is concerned, Navy has the largest and longest programs, with the focus on the ship propulsion systems.[15a-16] Air Force has expertise in airborne generators.[17a-b] In appendix II, DOD support for superconductor research is tabulated, along with other federal agencies.[17c]

D. Operation benefits

Operation benefits include technical and economical factors. On the technical side, superconductive generators have following advantages: high efficiency (>99.5%), light weight, small size, and long lifetime. In Table 1, specific weight and volume of several generators are listed. On the economical side, overall cost includes capital, operational, and maintaining costs. Capital cost of superconductive generator may be higher than that of conventional generator. However, operational cost of superconductive generator is lower than that of conventional generator. In a Westinghouse study, the operational cost is lowered by a factor of three or four. This is advantages because operation cost is the largest portion in overall cost of air base generator, see section A. In addition, the lifetime of superconductive generator is longer than that of conventional generator, and the maintaining cost of superconductive generator will be lower. Thus, overall cost of superconductive generator system is estimated to be lower than that of conventional generator.[18a] As an example, let us consider a small operation 3,000 miles away for one year, supported by a new generator of 900 KW (40,000 lb). If a conventional generator is utilized, the overall cost will be $997,488, in which operation cost shares 62.4%. However, if a superconductive generator is utilized, the total cost will be $881,992.80; or 11.6% savings. (Itemized costs are listed in Table 2 for the comparison between a conventional generator and a superconductive generator.) If the operation is longer than one year, there will be more savings. When 500-1,000

units are superconductive, the total savings will be $57,728,000-115,456,000, or more.

E. Recommended program development approach

As far as air bases are concerned, it is feasible to utilize superconductive generators with powers larger than 300 KW.[18b] Development on high Tc superconductive wires and coils should continue. (Appendices III-V illustrate recent development on Jc of wires, and performance of coils.[18c]) Experimentation on existing low Tc superconductive generators should also continue. For example, there exists a 10 MW generator at MIT.[19a] Its superconductive rotor field winding is workable, while its normal conductive stator armature winding is not.[19b] An estimated cost to repair/remodel the unit is $ 200K - $ 300K. (Estimated costs of superconductive generator and other power devices are provided in Appendix VI.) A joint sponsorship between DOD and DOE may allow further experimentation on this generator with low cost and low risk.[20] Based upon the estimated savings in Section D, it is reasonable to invest one million dollars in Air Force R&D on superconductive generators.

IV. Superconductive Motors

A. Air base motors

Since motor inventory data were not available during this study, a telephone survey on large motors was conducted. At Arnold AFB, Tullahoma, TN, there exist approximately 150 units in 300 horse power (h.p.) range, 70 units in 300-500 h.p. range, 40 units in 500-700 h.p. range, and several units with power larger than 700 h.p. In a test lab of Wright-Patterson AFB, Dayton, OH, there

exist 4 units in 450-600 h.p. range, 9 units in 1,000-2,000 h.p. range, 11 units in 2,100-5,500 h.p. range, and 3 units in 12,000-44,000 h.p. range.

B. Superconductive motors

When a rotational machine is operating in generator/motor mode, it is called alternator. There have been numerous studies on superconductive alternators.[21] Operational principles of a superconductive motor is similar to that of a superconductive generator, with a different energy conversion direction.

Since 1960's, various superconductive motors have been developed.[22-23] There exist several designs of superconductive motors: DC, synchronous, induction, induction/synchronous hybrid, reluctance, and homopolar inductor motors.[24a-c] In addition to rotational motors, linear motors have been constructed and tested.[25a-b] Since 1987, high Tc superconductor motors have been designed and demonstrated.[26a-c]

C. Recommended development approach

A bench top demonstration unit based on high Tc superconductor should be constructed and tested. The construction may consist of two phases: 1. motors based on bulk superconductors; 2. motors based on superconductive coil windings. Phase 1 project can be an improvement on existing motors.[27-28] Phase 2 project can be divided into two steps: A. evaluation of high Tc superconductors and wires/coils; B. construction of motors. Step A may require 1 year to finish. Since initial part of Step A belongs to 6.1 category, funding from internal sources and AFOSR will be needed. Superconductors can be supplied by several sources,[29] and evaluated

in different labs.[30] Step B may require 2 years, and experimentation should be carried out in one of the Air Force Labs.

V. Superconductive Transformers

The basis of transformer technique is coils which have been utilized in SMES, generators, motors, and magnets. So far, there has been no published result on the construction/test of a full scale power transformer. A 1 GVA transformer has been considered by DOE labs.[31]

From available information, it is estimated that transformer power in air bases varies from several hundred KW to 50 MW.[32] To further study the feasibility and reliability of superconductive transformers, we suggest that a small scale transformer (200 - 700 KW) should be designed, built, and tested.

VI. Superconductive Chokes

A choke is a coil which can limit current flow in a circuit of a power station or substation. Other names are: current limiter and reactor. It has been demonstrated that high Tc superconductors can be utilized as chokes.[33] An experiment should be conducted in an Air Force lab to study the reliability of the superconductive chokes.

VII. Power Lines

Power lines are divided into transmission lines or distribution lines. In Air Force installations, the number of transmission lines are very limited, but the number of distribution lines are countless. A typical distribution line has a length of 25 - 30 miles, with the maximum current of approximately 1,000 amperes. All these characteristics are compatible with

capabilities of high Tc superconductors, based on existing studies.[34] If all the distribution lines are buried underground, the reliability and security of Air Force power systems will be enhanced.

VIII. Overall Recommendation

Currently, Air Force is going through a critical period. Its future operation will be highly mobile. Air base power systems should meet following criteria: high efficiency, high reliability, small size, light weight, long lifetime, and easy operation/maintenance. Superconductive techniques offer great opportunity to Air Force civil engineering community, i.e., to reconsider/redesign existing power systems during this peaceful period. Furthermore, superconductor research is consistent with current federal government policy: high tech development and military/civilian dual usage.

To make realistic assessments on superconductive power technology, air base engineers/scientists should carry out in-house experiments. To share resources, they should collaborate with other researchers in DOD, DOE, and university labs.

Acknowledgement

The author would like to thank Mr. Tom Hardy and Mr. Ed. Alexander for their support; and Mr. Tom Kent, Mr. Dave Conkling, Mr. Larry Strother, and Mr. Reza Salavani for their help. Part of data was complied by Mr. Rick O'Neil, a graduate student at Alfred, partially supported by 1992 Summer Research Extension Program of AFOSR.

References

1. As informed by Mr. Tom Hardy of WL/FIVCO, a SMES device will be tested at Tyndall AFB, FL, in 1994.
2. Between May and June, 1993, the author was a summer faculty at Argonne National Lab, and participated research on the practical high Tc superconductors.

3. A coil fabricated by American Superconductor Co. has been delivered to Dr. Chuck Oberly's group at Wright-Patterson AFB, OH. The length of the coil is approximately 300 meters. Another company, Intermagnetics General Co., is also able to produce coils with similar properties.

4. Critical current, Ic, is the maximum current flowing through a superconductor specimen without destroying superconductivity. Ic is a term suitable for power applications, given in Amperes. While critical current density, Jc, is the maximum current per unit cross sectional area without destroying superconductivity, normally expressed in Amperes per centimeter squared. Jc is a term suitable for electronics applications.

5a. At 77 K, Jc of BiSrCaCuO superconductor decreases drastically as applied magnetic field increases. However, at temperatures below 20 K, Jc does not decrease substantially as field increases. Thus, BiSrCaCuO superconductor may be useful at low temperatures. In contrast, Jc of $Tl_1Ba_2Ca_2Cu_3O_x$ superconductor does not decrease substantially as field increases at 77 K.

5b. A. Schilling, M. Cantoni, J. D. Guo, and H. R. Ott, Nature, Vol. 363, 6 May, 1993, pp. 56-58.

6a. "Report of the Defense Science Board Task Force on Military System Applications of Superconductors", Office of the Under Secretary of Defense for Acquisition, Washington, D.C., October 1988.

6b. "Military Applications of Superconductors", Military Technology (MILTECH), 5/89, pp. 59 - 65.

6c. "The DOD Workshop Report on Superconductivity", ed. by L. Cohen, and E. A. Edelsack, IDA Memorandum Report M-482, 1988.

7a. "Navy Superconductivity Program", Volume 1-3; Naval Consortium for Superconductivity, Office of Naval Research, and Office of Naval Technology; May 1989.

7b. Air Force Office for Scientific Research (AFOSR) has been sponsoring programs on superconductivity and electronic applications such as thin films/devices. Dr. Chuck Oberly's group at Wright-Patterson AFB, OH, has done research on cryogenic and superconductive generators for airborne applications. In addition, WL/FIVCO (formerly RACO/AFCESA) has completed two SMES projects in an SBIR program (phase I).

8a. Currently, Alfred University is conducting feasibility studies for Air Force: 1. applications of superconductive devices (sponsored by AFOSR via RDL); 2. hybrid systems of SMES and Fuel Cells (sponsored by WL/FIVCO).

8b. "A study of air base facility/utility energy R&D requirements", ESL-TR-91-44, April 1992, Air Force Civil Engineering Support Agency, Tyndall AFB, FL 32403.

8c. Bare base power supply systems", ESL-TR-87-25, June 1988, Air Force Civil Engineering Support Agency, Tyndall AFB, FL 32403.

8d. "Large-area emergency power conceptual design report", ESL-TR-91-07, June 1991, Air Force Civil Engineering Support Agency, Tyndall AFB, FL 32403.

8e. Inventory data provided by AFCESA, with 647 units having power larger than 300 KW. The exact number of bare base generators (750 KW and/or larger than 300 KW) is not known. Based on a survey in two air bases, there exist 199 units of 750 generators.

8f. Data based on 750 KW bare base generator, 500-1,000 KW generators in several air bases, and 1-10 MW commercial products.
8g. Data based on 750 KW bare base generator and 750-1,100 KW diesel fueled generator set manufactured by Onan Co., Minneapolis, Minnesota 55432.
9a. The cost is $0.0003272620 per pound per mile when the total weigh is between 2,200 and 3,599 pounds; and $0.0002881265 per pound per mile when the weight is over 3,600 pounds.
9b. I. L. Kosow, "Electric Machinery and Control", (Prentice Hall, Inc., Englewood Cliffs, N.J., 1964).
9c. "Standard Handbook for Electrical Engineers", 7th ed., ed. by D. G. Fink, and H. W. Beaty, (McGraw-Hill Book Company, New York, 1978).
10. H. H. Woodson, Z. J. J. Stekly, and E. Halas, IEEE Transactions on Power Apparatus and Systems, Volume PAS-85, No. 3, March 1966, pp. 274 - 280.
11a. C. J. Oberhauser, and H. R. Kinner, Advances in Cryogenic Engineering, Volume 13, Ed. by K. D. Timmerhaus, 1968.
11b. P. Thullen, J. C. Dudley, D. L. Greene, J. L. Smith, Jr., and H. H. Woodson, IEEE Transactions on Power Apparatus and Systems, Volume PAS-90, No. 2, March/April 1971, pp. 611 - 619.
11c. J. L. Kirtley, and F. J. Edeskuty, Proceedings of the IEEE, Volume 77, No. 8, August 1989, pp. 1143 - 1154.
12a. N. Higuchi, H. Fukuda, T. Ogawa, Y. Nakabayashi, Y. Kobayashi, M. Ogihara, H. Sawazaki, Y. Yagi, A. Ueda, and T. Kitajima, IEEE Transactions on Applied Superconductivity, Volume 3, No. 1, March 1993.
12b. O. Tsukamoto, N. Amemiya, T. Takao, S. Akita, K. Ohishi, H. Shimizu, Y. Tanaka, and Y. Uchikawa, IEEE Transactions on Magnetics, Volume 28, No. 1, January 1992.
12c. Y. Brunet, P. Tixador, T. Lecomte, and J. L. Sabrie, Electric Machines and Power Systems, No. 11, 1986, pp. 511 - 521.
13a. S. K. Singh, and C. J. Mole, IEEE Transactions on Magnetics, Volume 25, No. 2, March 1989, pp. 1783 - 1786.
13b. As a summer faculty in a DOE lab, the author obtained some research notes and publications donated by Dr. M. Atoji.
13c. "Superconducting Generator Design", EPRI EL-577, November 1977.
13d. "Superconducting Generator Design", EPRI EL-663, March 1978.
13e. A conceptual design of 600 MW generator has been described by A. Ueda, T. Hirao, and T. Takeshita, Electrical Engineering in Japan, Volume 111, No. 2, 1991, pp. 102 - 114.
14. Similar data can be found in an article by N. Maki, and K. Yamaguchi, Hitachi Review, Volume 39, No. 1, 1990, pp. 25 - 30; and in Reference 13a.
15a. J. L. Smith, Jr., J. L. Kirtley, Jr., and F. C. Rumore, Proceedings of Spring Meeting/STAR Symposium, New London, Conn., April 26-29, 1978, pp. 21-1 - 21-11, Society of Naval Architects and Marine Engineers, One World Trade Center, Suite 1369, New York, NY 10048.
15b. T. L. Doyle, J. H. Harrison, D. W. Taylor, and A. Chaikin, Proceedings of Spring Meeting/STAR Symposium, New London, Conn., April 26-29, 1978, pp. 20-1 - 20-7, Society of Naval Architects and Marine Engineers, One World Trade Center, Suite 1369, New York, NY

10048.
16. H. O. Stevens, Jr., R. W. Meyer, L. T. Dunnington, M. D. Miller, and W. B. Mangis, "Development of superconducting machinery", Naval Ship Research and Development Center, Washington, D.C. 20034, Report 3556, October 1971.
17a. Dr. Chuck Oberly's group. Cryogenic generators do not contain superconductive components.
17b. "High power study: superconducting generators", AFAPL-TR-76-37, Wright - Patterson AFB, Ohio 45435, March 1976.
17c. Figures were provided by White House, and Tables were published in Superconductor Week, January 18, 1993, p 3.
18a. There exist many analyses of the economical factor. See, for example, E. L. Daniels, and J. L. Kirtley, Jr., "Generators", in "Advances in Applied Superconductivity: A Preliminary Evaluation of Goals and Impacts", ed. by A. M. Wolsky, E. J. Daniels, R. F. Giese, J. B. L. Harkness, L. R. Johnson, D. M. Rote, and S. A. Zwick, Argonne National Laboratory, ANL/CNSV - 64, January 1988, pp. 39 - 56.
18b. For the power less than 300 KW, we should develop high efficiency generators by using permanent magnets and modern controllers.
18c. Developments in Superconductivity, Interdivisional EPRI Newsletter, Summer 1993.
19a. J. L. Kirtley, Jr., J. L. Smith, Jr., and S. D. Umans, IEEE Transactions on Energy Conversion, Volume 6, No. 2, June 1991, pp. 274 - 281.
19b. Private communication with Dr. Jim Kirtley at MIT. Terminal voltage was 13.8 KV in earlier design, and will be 4 KV in new design.
20. Both DOD and DOE have funded MIT superconductive generator programs since 1960's.
21. P. Thullen, and J. L. Smith, Jr., Advances in Cryogenic Engineering, Volume 15, Proceedings of 1969 Cryogenic Engineering Conference, Los Angeles, California, June, 1969, (Plenum Press, New York, 1970).
22. T. A. Buchhold, in "Applied Superconductivity", ed. by V. L. Newhouse, (Academic Press, New York, 1975).
23. K. F. Schoch, Advances in Cryogenic Engineering, Volume 6, Proceedings of 1960 Cryogenic Engineering Conference, Boulder, Colorado, August 1960, ed. by K. D. Timmerhaus, (Plenum Press, New York, 1961).
24a. E. J. Daniels, B. W. McConnell, and T. A. Lipo, in "Advances in Applied Superconductivity: A Preliminary Evaluation of Goals and Impacts", ed. by A. M. Wolsky, E. J. Daniels, R. F. Giese, J. B. L. Harkness, L. R. Johnson, D. M. Rote, and S. A. Zwick, ANL/CNSV-64, Argonne National Lab, January 1988.
24b. P. Tixador, C. Berriaud, and Y. Brunet, IEEE Transactions on Applied Superconductivity, Volume 3, No. 1, March 1993, pp. 381 - 384.
24c. C. H. Joshi, and R. F. Schiferl, IEEE Transactions on Applied Superconductivity, Volume 3, No. 1, March 1993, pp. 373 - 376.
25a. O. Tsukamoto, Y. Tanaka, K. Oishi, T. Kataoka, Y. Yoneyama, T. Takao, and S. Torii, in "Advances in Superconductivity II", Proceedings of 2nd International Symposium on Superconductivity,

November, 1989, Tsukuba, Japan, ed. by T. Ishiguro, and K. Kajimura, (Springer-Verlag, Tokyo, 1990), 1047 - 1050.
25b. H. Nagano, M. Kinugasa, T. Tokizawa, K. Hayakawa, and K. Sasaki, in "Advances in Superconductivity II", Proceedings of 2nd International Symposium on Superconductivity, November, 1989, Tsukuba, Japan, ed. by T. Ishiguro, and K. Kajimura, (Springer-Verlag, Tokyo, 1990), pp.1055 - 1058.
26a. J. D. Edick, R. F. Schiferl, and H. E. Jordan, IEEE Transactions on Applied Superconductivity, Volume 2, No. 4, December 1992, pp. 189 - 194.
26b. J. S. Edmonds, D. K. Sharma, H. E. Jordan, J. D. Edick, R. F. Schiferl, IEEE Transactions on Energy Conversion, Volume 7, No. 2, June 1992, pp. 322 - 329.
26c. "Electric Motors Using Superconducting Materials Applied to Power Generating Station Equipment", EPRI TR-101127, September 1992.
27. X. W. Wang, J. Koprevich, R. C. Ward, III, N. R. Mannur, H. Petersen, W. B. Carlson, and W. A. Schulze, in "Superconductivity and Ceramic Superconductors II", ed. by K. Nair, et al., Ceramic Transactions, Volume 18, (American Ceramic Society, Inc., Westerville, Ohio, 1991).
28. A. Takeoka, A. Ishikawa, M. Suzuki, K. Niki, and Y. Kuwano, IEEE Transactions on Magnetics, Volume 25, No. 2, March 1989, pp. 2511 - 2514.
29. For example, American Superconductor Co., Intermagnetics General Co., Argonne National Lab, and Alfred University.
30. For example, WL/POOX-2 lab in Wright-Patterson Air Force Base, Ohio, National High Magnetic Field Lab in Tallahassee, Florida, New York State Institute on Superconductivity in Buffalo, New York, and Alfred University, Alfred, New York.
31. R. F. Giese, and B. W. McConnell, in "Advances in Applied Superconductivity: A Preliminary Evaluation of Goals and Impacts", ANL/CNSV-64, Argonne National Lab, January 1988.
32. This estimation is based on an energy survey conducted by RACO/AFCESA (WL/FIVCO) several years ago. We have assumed that each base has one main transformer.
33. American superconductor Co., Watertown, MA.
34. EPRI, Pirelli Cable Co., Underground Systems, Inc. Techniques will be: tapes, coatings of high Tc superconductor materials on copper tubes, and others.

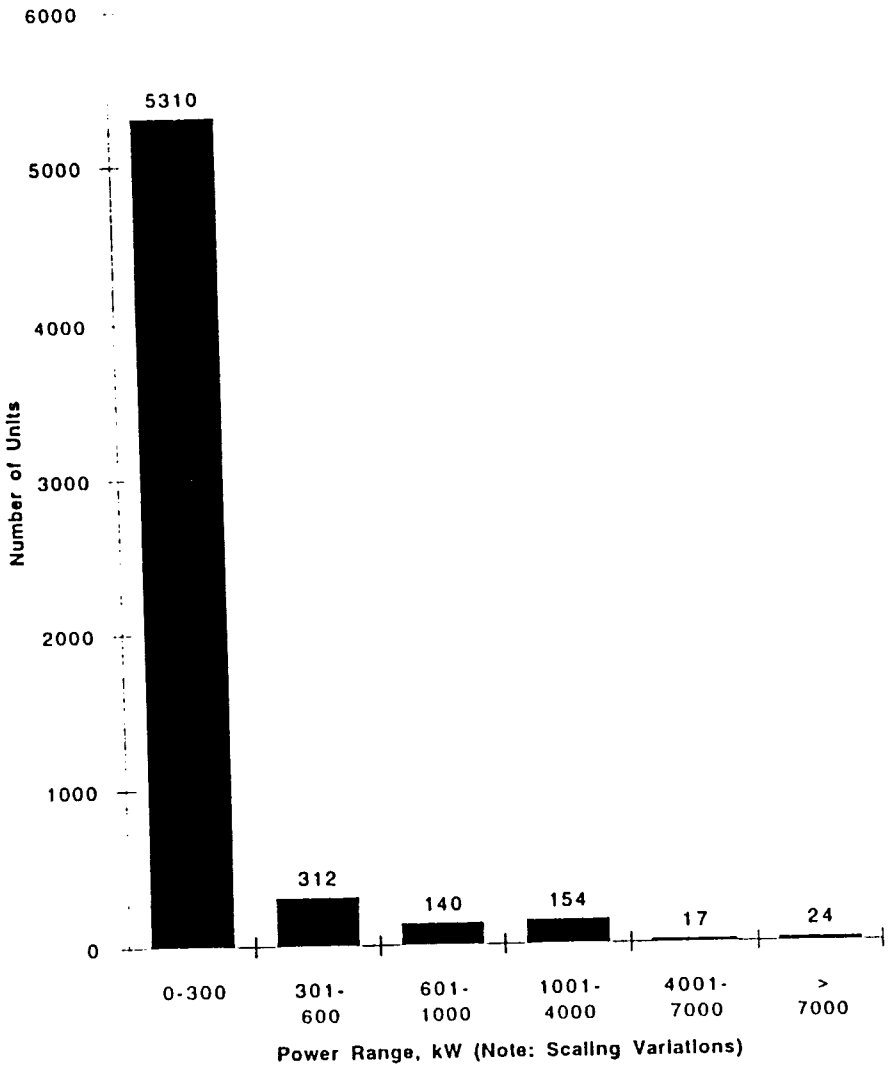


Figure 1. Number of Units vs. Generator Power.

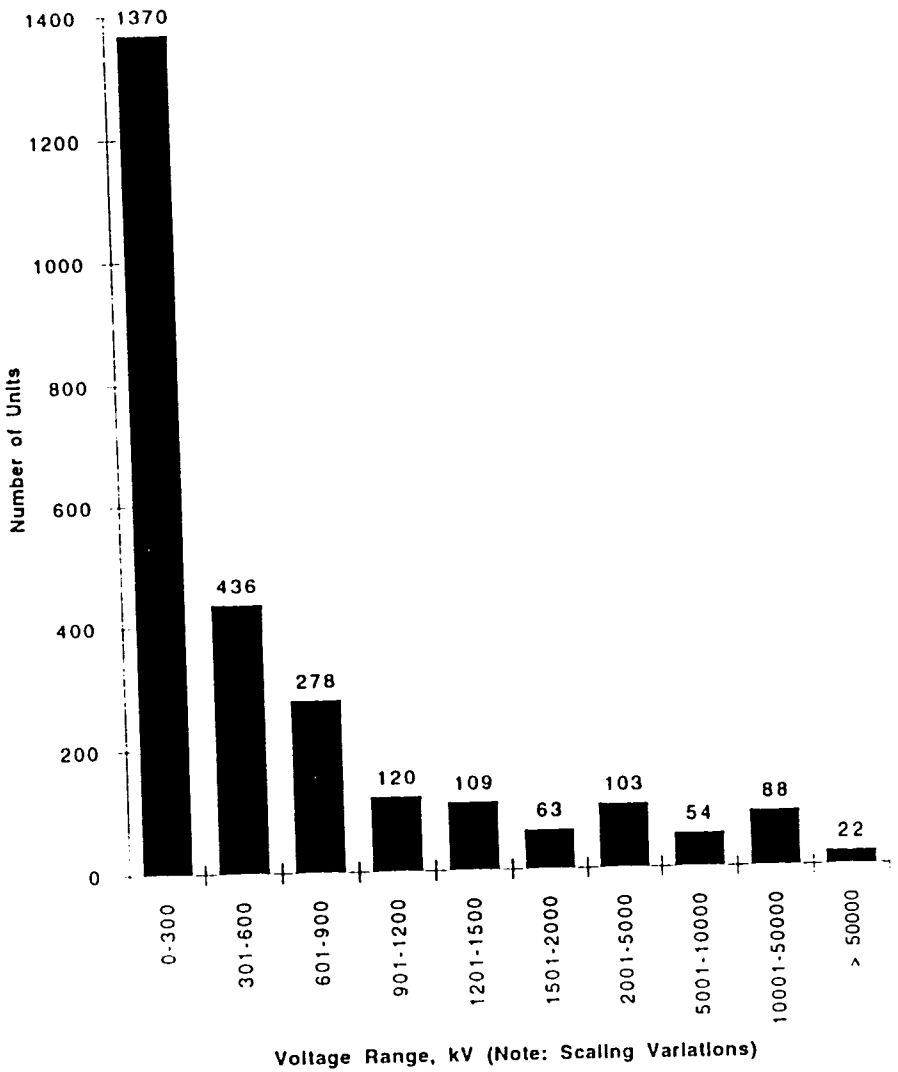


Figure 2. Number of Units vs. Generator Voltage.

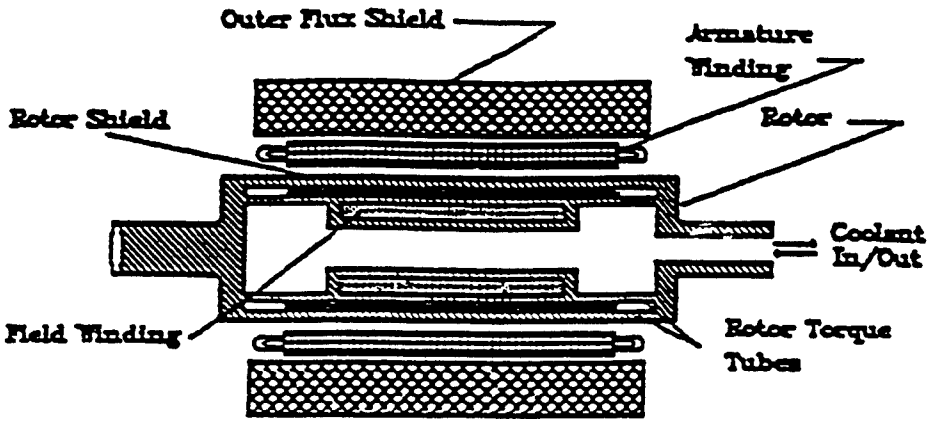


Fig. 3. Cross-sectional view of superconducting generator

Table 1. Specific weight and volume of superconductive generators

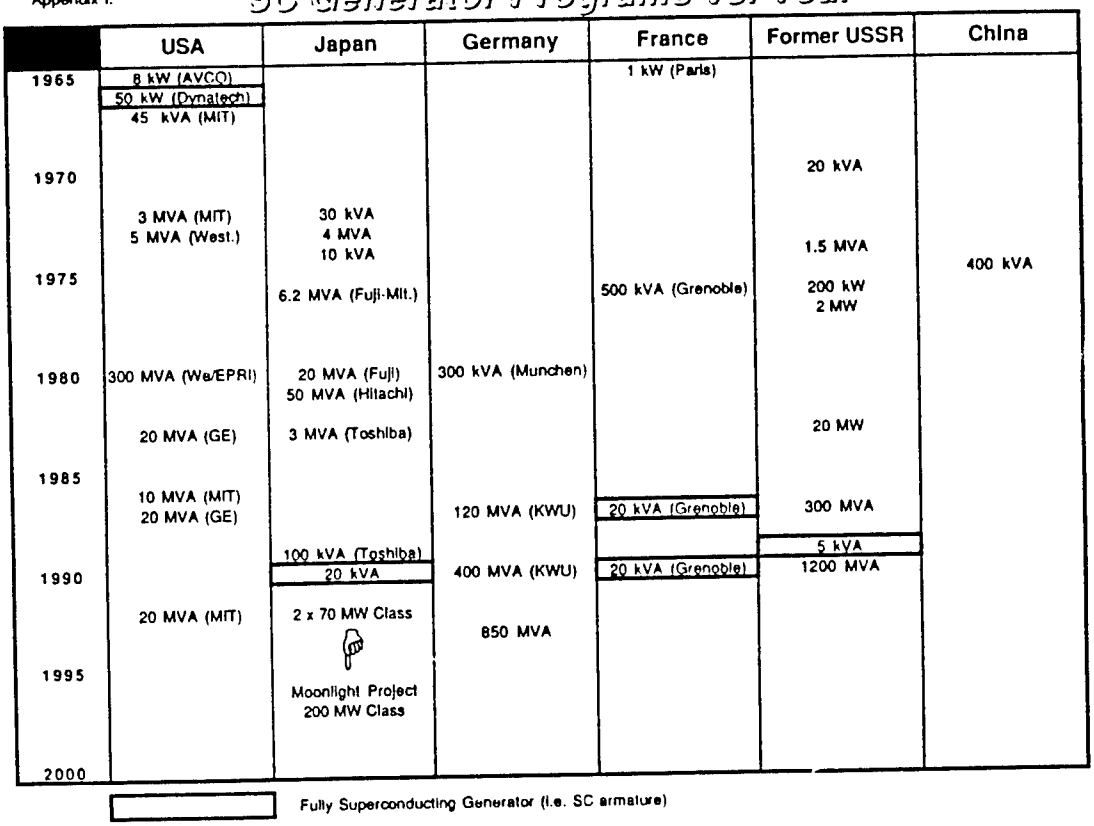
Power	Specific Weight (lb/KW)	Specific Volume (m^3/KW)
18.5 KW		0.00520
45 KW		0.00234
5 MW		0.000408
10 MW		0.0000229
6 MW	0.1852	
1 GW	0.992	
1 GW	0.988	
1 GW	1.103	

Table 2. Comparison between a conventional generator and a superconductive generator

Cost	Purchasing & installing	Transport	Operation	Maintenance
Convention unit price	$350/KW	$0.0003/lb-mile	$0.079/KW-hr	$0.003/KW-hr
Convention total price	$315,000	$36,000	$622,836	$23,652
Ratio super/conv	1.5	0.6	0.6	0.6
Super-conductor	$472,500	$21,600	$373,701.60	$14,191.20

Appendix I.

SC Generator Programs vs. Year



Appendix. II

Table 1. Federal Support for High Temperature Superconductivity R&D

	FY1987	FY1988	FY1989	FY1990	FY1991	FY1992(est)
DOE	12.3	26.7	38.0	42.1	46.1	49.5
DOD	19.0	43.7	58.0	58.7	54.2	63.2
NSF	11.7	17.1	23.3	20.2	21.8	22.0
DOC(NIST)	1.1	2.8	4.8	2.8	3.4	4.9
NASA	0.5	3.3	4.8	4.8	4.7	4.6
DOI	0.1	0.1	0.1	0.1	0.1	0.0
HHS(NIH)	0.0	0.0	0.0	0.0	3.0	0.0
DOT	0.0	0.1	0.1	0.1	0.3	1.2
Total High T_c R&D	44.9	93.8	129.1	128.7	130.7	145.4

Table 2. Federal Support for Low Temperature Superconductivity R&D

	FY1987	FY1988	FY1989	FY1990	FY1991	FY1992(est)
DOE	28.5	30.7	64.5	85.7	104.6	74.4
DOD	7.0	16.1	15.0	14.3	14.3	17.1
NSF	2.0	3.8	3.0	3.0	3.0	5
DOC(NIST)	0.0	0.6	0.5	0.5	0.5	1
NASA	2.1	3.7	2.4	1.9	1.9	3
DOI	0.0	0.0	0.2	0.2	0.2	0
HHS(NIH)	0.0	3.3	2.8	4.0	3.4	3.5
DOT	0.1	0.0	0.0	0.0	0.0	0.0
Total Low T_c R&D	39.7	57.2	88.3	109.5	127.8	100.6

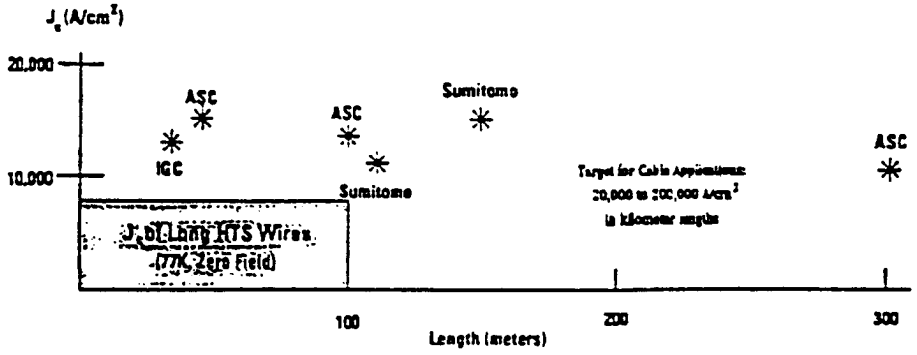
Table 3. Federal Support for Procurement of Superconducting Devices

	FY1987	FY1988	FY1989	FY1990	FY1991	FY1992(est)
DOE	1.0	1.5	5.8	16.5	39.7	115.8
DOD	0.2	14.4	17.4	18.0	20.0	7
Total Procurement	1.2	15.9	23.2	36.5	69.7	115.8

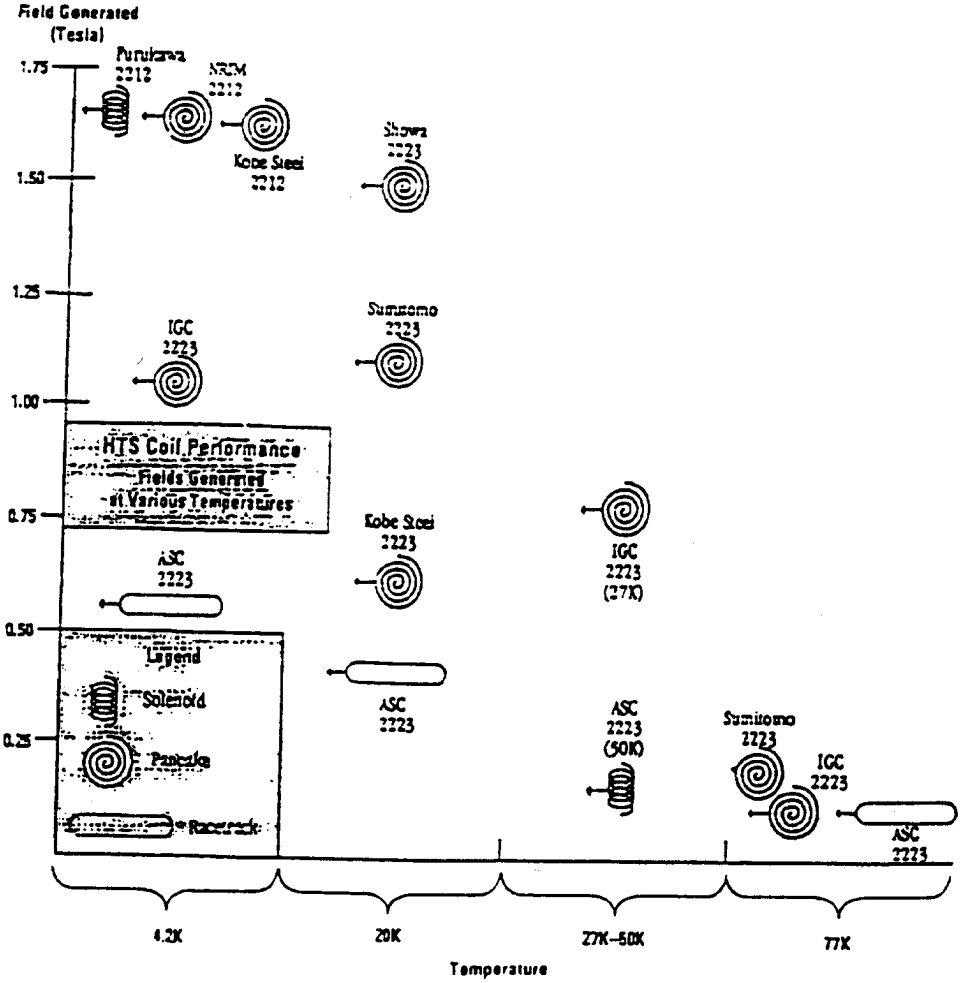
Source: White House

(in Millions)

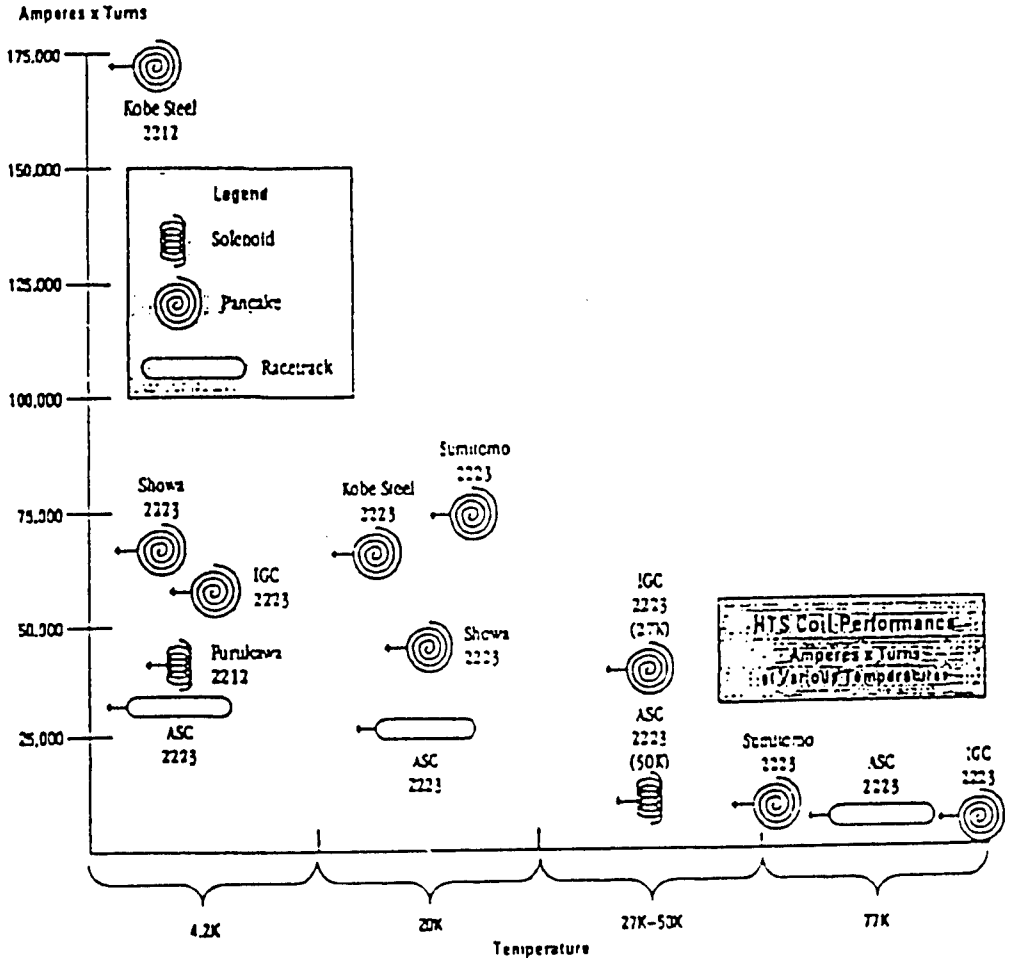
Appendix III.
Jc of long wires (77K, 0T)



Appendix IV.
Magnetic field generated in coil



Appendix V.
Amperes X turns



Appendix VI. Preliminary Cost Breakdown

Following statement is provided in an article by L. S. Greenberger, published in Public Utilities Fortnightly, August 1, 1991, pp. 38-40.

For each facet of a superconductive system: $60 million to produce a 20-megawatt prototype generator; $50 million for a 100 meter long, 500-megavolt transmission line; $55 million for a 1-megawatt SMES; $30 million for a 4-kilovolt/138-kilovolt transformer; and $55 million for a 5,000 horsepower electric motor.

A STUDY FOR PRODUCT DESIGN, MATERIALS SELECTION, PROCESSES, AND INTEGRATION USING LEARNING APPROACHES

C. L. Philip Chen

Assistant Professor
Department of Computer Science and Engineering
Wright State University
Dayton, OH 45435
E-mail: pchen@valhalla.cs.wright.edu

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

September 1993

A STUDY FOR PRODUCT DESIGN, MATERIALS, SELECTION, PROCESSES, AND INTEGRATION USING LEARNING APPROACHES

C. L. Philip Chen
Assistant Professor
Department of Computer Science and Engineering
Wright State University

Abstract

The relation on product design, materials selection, and processes has been studied. The procedures for design, materials selection and processing is summarized in this study. To design a product, material properties and materials processing has to be considered simultaneously. The study s ows that association among design, materials, and processes must be built and identified. Several learning approaches that can be used as tools for building an inductive and deductive coupling system for materials research are discussed. Future research opportunities in this area are also presented.

A STUDY FOR PRODUCT DESIGN, MATERIALS, SELECTION, PROCESSES, AND INTEGRATION USING LEARNING APPROACHES

C. L. Philip Chen

1. Introduction

In recent years, the research in product design, materials, and processes has been studied significantly. However, an integrated method for associating design, materials, and processes has not been reported. Why product design, materials, and processes association is so important? As we know, a particular design is eventually made into a product, design, materials, and process must be intimately interrelated. Each component of a product must be designed so that it not only meets design requirements, but also can be manufactured efficiently and economically. To be able to manufacture economically and efficiently, the materials of the component and the manufacturing processes must be competitive with the design specification. The *concurrent engineering* approach is an integration approach to achieve this goal. However, currently research in concurrent engineering falls into integration of database and indispensable computer programs [1] or expert systems [2]. The development of learning-based approach still under infancy. In the next section, the overview of product design, materials selection, and selection of manufacturing processes will be discussed followed by discussion of learning approaches and future research opportunities.

1.1. The Design Process

The first step in design is to understand the function and performance of the product. Then the design concept follows. The concept probably is an innovative, a creative, or probably modification of a previous experience. To avoid design turn-around time, the designer must also have the knowledge of materials and processes or production cost. However, in most cases, materials selection and manufacturing processes are aided by other organizations (This is why a learning-based, induction and deduction integration are needed). Table 1 roughly shows traditional shapes and some common methods of production [3]. Figure 1 also shows the traditional design process [4]. Design For Manufacture (DFM) and Design For Assembly (DFA) are other factors that a designer must consider. Analytical model of product cost is another consideration in design. Finally, a system that transfer a product design, to design concept, materials and processes selection will be implemented.

1.2. Selection of Materials

Table 1. Shapes and some common methods of production [3]

SHAPES AND SOME COMMON METHODS OF PRODUCTION

SHAPE	PRODUCTION METHOD
Flat surfaces	Rolling, planing, broaching, milling, shaping, grinding
Parts with cavities	End milling, electrical-discharge machining, electrochemical machining, ultrasonic machining, cast-in cavity
Parts with sharp features	Permanent-mold casting, machining, grindi- fabricating
Thin hollow shapes	Slush casting, electroforming, fabricating
Tubular shapes	Extrusion, drawing, roll forming, spinnir trifugal casting
Shaping of tubular parts	Rubber forming, expanding with hydr ssure, explosive forming, spinning
Curvature on thin sheets	Stretch forming, peen forming, fabricating
Openings in thin sheets	Blanking, chemical blanking, photochemical blanking
Reducing cross-sections	Drawing, extruding, shaving, turning, centerless grinding
Producing square edges	Fine blanking, machining, shaving, belt grinding
Producing small holes	Laser, electrical-discharge machining, electrochemical machining
Producing surface textures	Knurling, wire brushing, grinding, belt grinding, shot blasting, etching
Detailed surface features	Coining, investment casting, permanent-mold casting
Threaded parts	Thread cutting, thread rolling, thread grinding, chasing
Very large parts	Casting, forging, fabricating
Very small parts	Investment casting, machining

The first criterion of materials selection is to consider mechanical properties such as strength, toughness, ductility, hardness, elasticity, fatigue, creep, and/or correlation of above. The mechanical properties specify the limitation of the formation. Based on the functi of designed product, the mechanical properties determine what materials the component ne product should be. These properties also determine the selection of the manufacturing processes.

The cost of raw materials and manufacturing processes are also the major concern for the selection. For the commercial and marketing point of view, design engineer must consider appearance, product cost, and product life as well. Materials properties and manufacturing processes also relate to the product appearance. For example, manufacturing processes determines surface texture of the product. For the environmental point of view, design for an environmentally friendly production, disposal and recycling are considerations that relate to materials and processes.

1.3. Manufacturing Processes selection

Selection of manufacturing processes depends on shape and geometry of the product, the type of material, nical properties of materials, and the cost of processes. Since not all processes produc final products, additional finishing operations such as grinding, polishing, machining n e necessary. These processes will also add additional cost to the

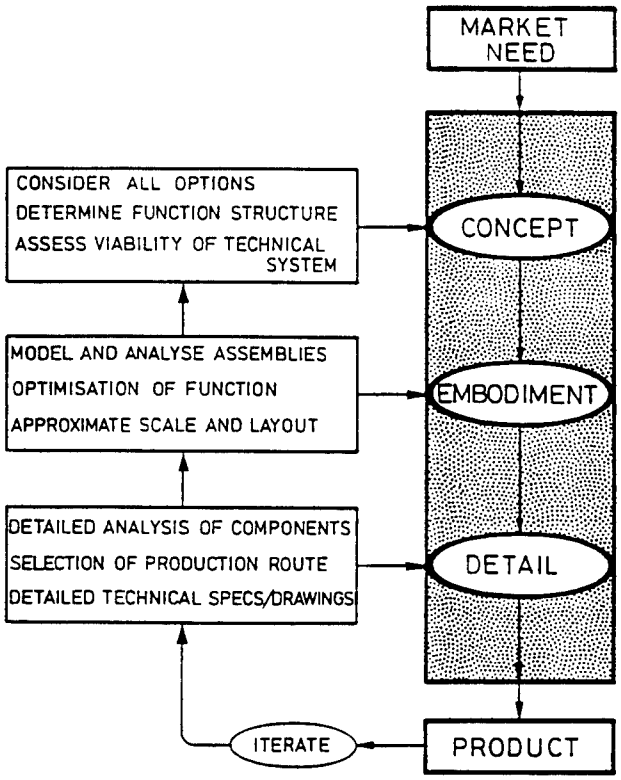


Figure 1. The design flow chart

product.

1.4. Other Considerations

Assembly operation is to assemble individual parts that have been manufactured to a product. To save the cost of the assembly, automatic assembly and design for assembly can contribute significantly reduction of the product.

2. Materials, Design, and Processes

This section briefly reviews fundamental aspects of materials, design, and processes and their relationship.

2.1. Material Properties

Material properties successfully explore their potential and characteristics for a product design. There are six important classes of materials for product design: metals, polymers, elastomers, ceramics, glasses and composites. Several standard properties such as general cost, mechanical, thermal, wear, corrosion/oxidation are listed in Table 2.

Table 2. Material properties

Class	Property	Symbol and Units	
General	Relative Cost	C_R	(--)
	Density	ϱ	(Mg/m^3)
Mechanical	Elastic moduli	E, G, K	(GPa)
	Strength (yield/ultimate/fracture)	σ_f	(MPa)
	Toughness	G_c	(kJ/m^2)
	Fracture toughness	K_{Ic}	(MPa $m^{1/2}$)
	Damping capacity	η	(--)
	Fatigue ratio	f	(--)
Thermal	Thermal conductivity	λ	(W/m K)
	Thermal diffusivity	a	(m^2/s)
	Specific heat	C_p	(J/kg K)
	Melting point	T_m	(K)
	Glass temperature	T_g	(K)
	Thermal expansion coefficient	α	($°K^{-1}$)
	Thermal shock resistance	ΔT	(°K)
	Creep resistance	–	(--)
Wear	Archard wear constant	K_A	(MPa^{-1})
Corrosion/ Oxidation	Corrosion rate	--	(--)
	(Parobolic rate constant)	K_p	(m^2/s)

Each class of materials has its only certain common characteristics: metals are ductile and conduct heat well; polymers are light and are able to expand, and so on. Based on these properties, classification can be very useful for the selection of materials for a product design. However, the classification only relates the function of product, in the design stage, to material properties with a general idea. The detailed relations between the design, properties, name of materials, furthermore, manufacturing processes are needed to be investigated.

2.2. Materials vs. Design

The materials properties can be displayed as material selection charts [4]. The charts summarize the engineering information in a compact way. Each chart displays the material classes with respect to the given properties. Eighteen material property charts are given:

Young's Modulus/Density	Strength/Density
Fracture/Toughness/Density	Young's Modulus/Strength
Specific Modulus/Specific Strength	Fracture Toughness/Modulus
Fracture Toughness/Strength	Loss Coefficient/Young's Modulus
Thermal Conductivity/Thermal Diffusivity	Thermal Expansion/Thermal Conductivity
Thermal Expansion/Young's Modulus	Normalized Strength/Thermal Expansion
Strength/Temperature	Young's Modulus/Relative Cost
Strength/Relative Cost	Normalized Wear Rate/Bearing Pressure
Young's Modulus/Energy Content	Strength/Energy Content

According to Ashby, the most striking feature of the charts is the way in which members of a material class cluster together. To relate design with materials, the first step starts with full menu of materials, by applying primary constraints (defined by the designer), defining performance index, maximizing performance index, a list of materials can be narrowed down. The material is selected by further narrowing, imposing secondary constraints, identifying material clusters, selection of best material in the cluster, and analysis. Figure 2 shows this process [4]. Depending on shape, loading, and design, the performance indices are minimizing material weight while maximizing strength ($\frac{\sigma_f{}^x}{\rho}$, or $\frac{failure\ strength}{density}$), stiffness ($\frac{E}{\rho}$), and crack ($\frac{K^x{}_{IC}}{\rho}$). Elastic design, damage-tolerant, and thermal design can be one of thei performance indicies.

The narrowing process can be done by transferring the constraints into performance index as a function of the material property charts mentioned above. A block of area can be identified and selected. The same selection procedure applies to the selection of shape of materials. The design with shape involves the section shape of material as a variable. Four shape factors are given: elastic bending, $\phi^e{}_B$; elastic twisting, $\phi^e{}_T$; failure in bending, $\phi^f{}_B$; and failure in twisting, $\phi^f{}_B$; The shaped material can be selected from the material property charts. This is done by rewriting the performance index as a function of shape factors. Embedding the shape factor in the material property functions implicitly moves the selection line, or the selection area.

However, the selection process requires matching the performance index across the property charts if several design goals are desired. Justification of the selection process requires an experienced design engineer. Further development of automatically selection of materials by a given design or a performance index is necessary.

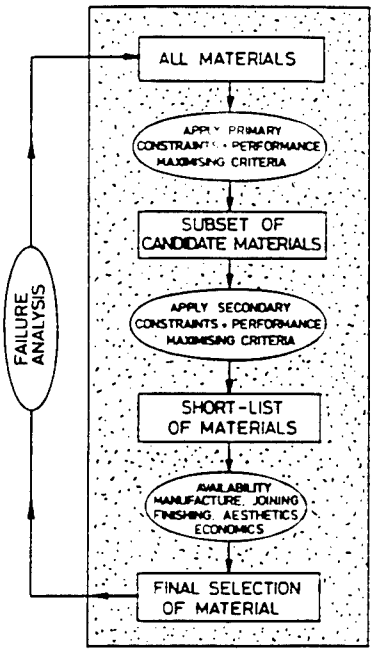


Figure 2. A flow chart for selection procedure

2.3. Materials vs. Processes

Material processes depend on material , shape, and design. The selection of processes is very difficult because of tremendous shaped materials and finishing methods. Like the aspect of design, process selection is an iterative procedure. Figure 3 shows the flow chart of process selection and the relation to the design [4].

In general, manufacturing processes can be classified to nine processes: casting, pressure molding, deformation processing, powder, special, matching, heat treatment, joining, and finishing. Similar to the design and material selection, process selection charts are defined. Based on the defined attributes, each process occupies a certain area of charts. The attributes are size, shape, complexity, precision, surface roughness, etc. Five charts are given:

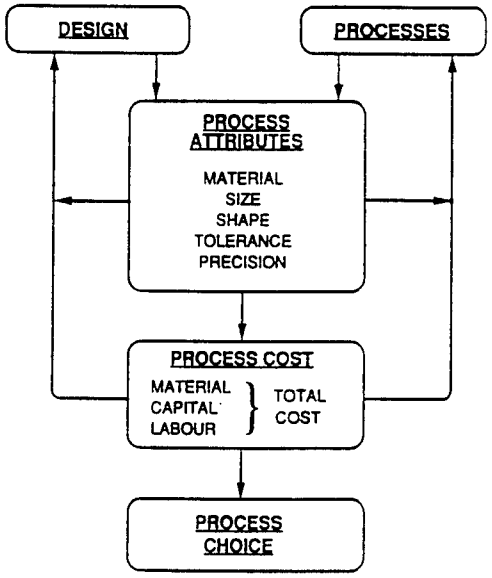


Figure 3. Process selection and design

Size/Shape(or Surface Area/Section)
Size/Melting Point
Tolerance/Surface Finish
Information Content/Size
Hardness/Melting Point

Process selection is achieved by setting the upper and lower limits of in the chart specified by the design. The processes which lie within or bounded by the search area are candidates. The procedure is repeated through all the charts, narrowing the selection and finally a subset of process capable of achieving the design goal. Cost issues may determine the final choice. However, very few article discusses the cost issue. Currently, concurrent engineering and agile manufacturing have brought this important issue [5, 6]. Based on the study above, there is a need for developing an intelligent system that integrates design, materials, and processes. In the following section, several learning approaches and opportunities for materials research are discussed.

3. Inductive Learning Techniques for Concept Formation

Inductive learning, which is learning a generation from a set of example, is one of the most fundamental learning tasks. Concept learning, a typical inductive learning problem, is to infer a definition that will allow the learner to correctly recognize future generalized instances of that concept for given some concept examples. Given a set of objects exhibiting various

properties, conc formation is a process to divide the objects into useful categories. Concept formation is a fundamental problem in unsupervised learning. The features (or attributes) extraction and defining boundary of the categorization is the first stage of the concept formation.

There are many different approaches to Inductive learning. The techniques such as machine learning paradigm, fuzzy systems, neural networks, and genetic algorithms are of particular interested recently. The goal of materials, design, and processes research will be to develop an intelligent induction and deduction coupling techniques. These new learning disciplines can be applied to build a general learning platform. The platform is a domain-independent so that it can be used in any material domains. Following is the brief introduction for the above learning disciplines and some issues related to the materials research.

3.1. Machine Learning: The ID3 algorithm

ID3 is a decision tree induction algorithm. It represents concepts as decision trees. The decision trees allow us to determine classification of an object by testing its values for certain properties. ID3 constructs decision trees in a top-down fashion. It selects a property to test at the current node of the tree and uses this test to partition the set of examples. The algorithm then recursively constructs a subtree for each partition [7]. Because the order of test is critical to constructing a decision tree, ID3 heavily relies on its selection criterion for the root of each subtree. ID3 has impressive results on several applications. The work on modification of ID3 algorithm to improve the complexity has been done.

According to the study in Ashby's book, the basic selection charts have been built. These charts can be treated as categorization of materials selection, design, and processes. There is no need for us to using ID3 algorithm for this kind of categorization. However, the categorizations only represent two-dimensional relationship. Cross categorization for materials, design, and processes are needed to done. ID3 algorithm can be used in cross-categorization application.

3.2. Learning using Fuzzy Rules and Concept

Fuzzy logic and systems have rapidly become one of the most successful learning tools. The fuzzy approximate reasoning provides decision-support and expert systems with powerful reasoning capabilities. The components of conventional fuzzy systems include: fuzzifiers, defuzzifiers, and a set of inference rules. Fuzzifiers convert inputs into their fuzzy representations. Defuzzifiers convert the output of the fuzzy process logic intro crisp solution. The inference rules represent a collection of linguistic rules. These linguistic rules can be represented in a matrix form, or call a fuzzy knowledge base, with actions (or output

variables) in the entities and input control variables as the causes. In fuzzy systems, the values of fuzzified input execute all the rules in the fuzzy knowledge base that have the input variables. This process generates a new fuzzy set representing each output or solution. Defuzzification process creates a value for the output variable.

Based on the introduction mentioned above, fuzzy inference rules are created by the designer. In other words, he first needs to know what information is available, then inference rules are defined. The problem arises if there is an unknown complicated environment that no mathematical model exists. The task is how to discover rules. The first step is concept formation or clustering from problem description. Using the formatted concept a set of fuzzy rules, hopefully, can be created. For materials, design, and processes, the concept formation is to discover clusters and associative relation among them, and to develop algorithms that convert formatted concept to fuzzy rules. Further work on development of fuzzifiers and defuzzifiers are also needed to be considered. In materials, design, or processes there are some attributes that deals with numerical data such as production cost and materials cost. Several issues such as how to combine linguistic concept together with numerical data to generate fuzzy rules may arise.

3.3. Genetic Algorithms, Classifier, and Fuzzy Systems

A Genetic Algorithm is a search algorithm modeled on the mechanics of natural selection [8]. Classifier systems are massively parallel, message-passing, rule-based systems that learn through credit assignment (the bucket brigade algorithm) and rule discover (the genetic algorithms) [9]. The paradigm are well known and can be found easily from several literatures. Application of genetic algorithms to concept learning problems is an approach for induction concept formation. Basically, the learning is supervised. Issues such as internal representation of the search space, define an evaluation function are needed to be concerned.

There are several hybrid systems that combine the paradigm of Fuzzy logic, GAs, and Classifier. For the Fuzz-GA systems, the obvious applications are to find optimal membership functions of fuzzifiers for given control problems such as control of water level in the tank, space rendezvous [10]. However, the achievement basically neglect the power of the integrated systems. One opportunity for materials research is to integrate fuzzy systems with GA for searching new combination of materials (structure such as Nye's diagram [13]). The basic concept block diagram is shown in Figure 4.

In this system, GAs search for membership function or combination of the materials in the fuzzy systems. In other words, the GAs search for the best rules set for the fuzzy systems. The materials, design, and processes are inter-associate internally. For Fuzzy-Classifier systems, the basic application is to modify the fuzzy message matching and fuzzy message

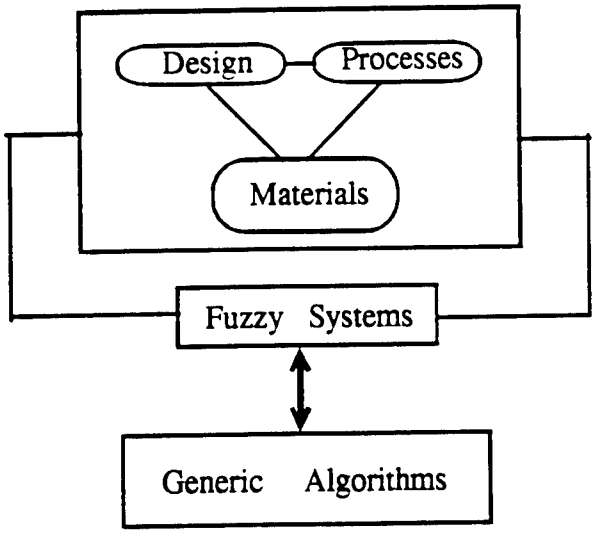


Figure 4. Basic Digram for intergrating Fuzzy system and GAs

generation of the original classifier.

3.4. Neural Networks and Association

Neural networks help solve various engineering problems with natural mechanisms of generalization recently [11]. Neural networks deals with uncertainty, association by parallel-distributed structure. Two basic learning algorithms exist in the neural networks: supervised and unsupervised learnings [12]. Applications of individual learning can be found in literature. However, integration of supervised and unsupervised has not been applied widely. The unsupervised learning portion discovers clusters based on similarity, and the supervised learning portion obtains the generalized mapping from the known input-output relationship. For materials application, the association can be expanded to more than 3 dimensions, e.g., association among materials, design, processes, and/or shapes. The opportunity for developing a hybrid system that integrates neural networks and fuzzy systems is worth to pursue.

4. Conclusion and Future Research Opportunities

From the review of materials, design, processes selection and relationship, and existing machine learning approaches, there is a need to design a self-improving system that is capable of performing concept formation for materials and process design using inductive and deductive coupling techniques. A study of using supervised and unsupervised techniques, genetic algorithms and classifier systems to develop such a system will be investigated. The future work in this area is to design a bi-directional learning system that (1) automates materials and

process design based on product design information, (2) advises product and geometric design based on materials processing and process, (3) analyzes cost information in the early stage of product design, and (4) optimizes material properties for product and process design.

5. References

[1] J. Yu, S. Krizan and K. Ishii, "Computer-aided design for manufacturing process selection," *Journal of Intelligent Manufacturing,* Vol. 4, 1993, pp. 199-208.

[2] M. Y. Demerc, *Expert systems applications in materials processing and manufacture,* TMS Publications, Warrendale, PA, 1990 .

[3] S. Kalpakjian, *Manufacturing Processes for Engineering Materials,* Addison-Wesley Publishing Co., NY., 1991.

[4] M. F. Ashby, *Materials Selection in Mechanical Design,* Pergamon Press, Oxford, UK., 1992.

[5] A. Kusiak, ed., *Concurrent Engineering: Automation, Tools, and Techniques,* John Wiley & Son, Inc., NY, 1993.

[6] R. Nagel and R. Dove, "21 Century Manufacturing Enterprise Strategy., Vols. 1 and 2, *Iacocca Institute,* Lehigh University. Bethlehem, PA., 1991.

[7] J. Quilan, "Induction of Decision Trees," *Machine Learning,* Vol. 1, No. 1, pp. 81-106, 1986.

[8] J. Holland, *Adaptation in natural and artificial systems,* U. of Michigan Press, Ann Arbor, MI, 1975.

[9] L. B. Booker, D. E. Goldberg, and J. H. Holland, "Classifier Systems and Genetic Algorithms," *Artificial Intelligence,* Vol. 40, 1989, 235-282, 1989.

[10] C. L. Karr, "Genetic algorithms for fuzzy logic controllers," *AI Expert,* Vol. 6, No. 2, pp. 26-33, 1991.

[11] C. Dagli et al. ed., *Artificial Neural Networks in Engineering,* ASME press, 1991, 1992.

[12] Y. H. Pao, *Adaptive Pattern Recognition and Neural Networks,* Addison Wesley, NY., 1989.

[13] J. F. Nye, *Physical Properties of Crystals,* Oxford University Press, London, 1957.

STRUCTURE AND COMPOSITION CHARACTERIZATION OF GaAs, AlGaAs LAYERS AND SUPERLATTICES GROWN ON GaAs BY MOLECULAR BEAM EPITAXY

Alfred T. D'Agostino

Assistant Professor

Department of Chemistry

University of South Florida

4202 E. Fowler Avenue

Tampa, Florida 33620-5250

Final Report for:

Summer Faculty Research Program

Wright Laboratory

Sponsored by:

Air Force Office of Scientific Research

Bolling Air Force Base, Washington, D.C.

September 1993

STRUCTURE AND COMPOSITION CHARACTERIZATION OF GaAs, AlGaAs LAYERS AND SUPERLATTICES GROWN ON GaAs BY MOLECULAR BEAM EPITAXY

Alfred T. D'Agostino
Assistant Professor
Department of Chemistry
University of South Florida

Abstract

High resolution x-ray double crystal diffractometry was used to characterize thin film layers grown on III-V monocrystalline material by molecular beam epitaxy. In particular, the thickness, composition and structure of epilayers and superlattices grown on (001) GaAs were determined. GaAs and $Al_xGa_{1-x}As$ epilayers, high electron mobility transistor structures and quantum well superlattices were characterized by evaluating rocking curves and performing simulations based on dynamical x-ray theory. Data was also obtained to support projects which focus on the low temperature growth process of GaAs on GaAs and the relationship between growth conditions and structure of III-V films.

STRUCTURE AND COMPOSITION CHARACTERIZATION OF GaAs, AlGaAs LAYERS AND SUPERLATTICES GROWN ON GaAs BY MOLECULAR BEAM EPITAXY

Alfred T. D'Agostino

Introduction

Gallium arsenide (GaAs) has been the subject of intense research for some time now because of its interesting chemical properties and potential for use in electronic device applications. In particular, its high mobility and high saturated drift velocity (as compared to silicon) and its ability to produce semi-insulating substrates has made it possible to produce true monolithic circuits that can operate in the microwave regime [1]. Application areas in which GaAs based devices may be used include: communications, radar, detectors, electronics, and high speed computing.

One important type of device utilizing GaAs is the field effect transistor (FET). A number of different types may be identified and includes the metal semiconductor FET and microwave devices like the IMPATT and Gunn diodes. Devices may also be made by incorporating other elements into the lattice as is done in high electron mobility transistors (HEMT), where for example AlGaAs is a component and doping is required. For fundamental research projects, quantum well structures have been constructed for investigation. As growth capability and fabrication technology improve, the potential for use of GaAs based devices will expand greatly.

Two general methods exist for forming conductive surface layers on GaAs substrates: epitaxy and ion implantation. The method of ion implantation will not be discussed here, however. Within the category of epitaxial growth techniques are liquid phase epitaxy and chemical vapor deposition procedures. These methods will not receive attention here so that the emphasis can be placed on molecular beam epitaxy. Molecular beam epitaxy (MBE) is the most recent major method developed for epitaxial growth [2]. Typically, under ultrahigh vacuum, a monocrystalline substrate (rotating and heated) is exposed to effusion cell sources of various elements (*e.g.* Ga, As, Al, Sb, *etc.*) to produce crystalline epilayers of desired composition and thickness. In the case of epitaxial growth on GaAs, temperature, flux density and other factors are controlled to produce high quality crystalline films of specific composition. The formation of complex superlattices of various designs are also possible using molecular beam epitaxy. MBE is versatile in that layers of any composition (including graded layers) and thickness can be produced with high uniformity and accuracy. Superlattice structures with thin layers (in the nanometer regime) of semiconducting material, which are separated by insulating layers, to form quantum well structures are readily produced by MBE. To aid in development, growth and testing of such structures, various techniques are used for film analysis.

With good analytical data, attempts can be made to understand epitaxial growth processes and thereby improve processing methods and achieve better results. Two important facets of GaAs process technology then become: 1) the relationship between growth conditions and epilayer parameters and 2) the non-destructive evaluation of

resultant structures. In practical terms, two of the more important goals in process technology have been to reduce the temperature at which quality monocrystalline epilayers are produced and to develop analytical techniques that furnish detailed structure and composition information about MBE grown structures. The work described in this report reflects attempts to characterize III-V material grown on GaAs by MBE in terms of appropriate models, processing conditions, structure, and composition.

Characterizing Epitaxial Layers and Superlattices

The use of a high resolution x-ray double crystal diffractometer was made in determining the composition, thickness and structure of molecular beam grown films and multilayers on GaAs (001) substrates. It has been shown that diffraction techniques may be used as an integral part of III-V semiconductor film characterization [3]. Its use in the analysis of AlGaAs epilayers on GaAs, for example, has been of particular interest recently [4,5]. The present discussion will highlight the way in which high resolution diffractometry was used in the characterization of GaAs and AlGaAs layers on the (001) GaAs substrate; it will include a description of the analysis of superlattices composed of these materials. However, research objectives have also included the study of other III-V substrates, epilayers and structures (*e.g.* utilizing GaSb, InAs, GaSbAs, *etc.*) by diffraction pattern analysis.

Strain is an important characteristic of an epitaxial film and may be used as a parameter to assess film properties [3]. The cause of strain in epitaxial films is primarily due to the difference of the bulk lattice spacing of substrate and film parallel to the

interface, the so called lattice mismatch. By analyzing high resolution diffraction "rocking curves", structure and composition data may be derived for the III-V systems under study.

The film strain parallel to the interface between two perfect cubic crystals is given by $e_{epilayer} = (a_s - a_f / a_f)$, where a_s and a_f are the lattice parameters of the unstrained substrate and epilayer film respectively. A number of factors be identified as contributing to the production of strain in a system may be identified and described by a strain tensor (the components of which and whose form will not be described in detail). A discussion of the mechanism by which strain is relaxed will not be considered in this discussion. It will suffice to say however that deduction in strain in thick films is caused by misfit dislocations. A more thorough description may be found elsewhere [3,6].

If the strain tensor for a system is determined, the film strain ϵ_{hkl} parallel to the unit vector normal to the (hkl) planes can be computed. In the case of the systems described herein (e.g. $Al_xGa_{1-x}As$ on GaAs), the rocking curves of the (hkl) planes typically show two peaks, from the substrate and film respectively. They are separated by $\Delta w_{hkl} = \Delta\theta_{hkl} + \Delta w_{\Delta\phi}$ where the first term represents the difference of Bragg angles due to the difference in lattice spacings between substrate and film and where the last term represents the change w due to the tilt in lattice planes by $\Delta\phi_{hkl}$. If the sample is rotated around the diffraction vector, $\Delta\theta_{hkl}$ can be determined as the mean value of two measurements of Δw taken at two azimuths 180° apart.

The mismatch between epilayer and substrate may be described by perpendicular and parallel components. By using reflections from the (004) planes, the perpendicular

component can be evaluated. A parameter known as the relaxed mismatch may be defined as the mismatch the layer would have if it were totally relaxed. If the mismatch and layer thickness are not too large, the layer will tetragonally distort and maintain a coherent interface with the parallel component equal to zero. If relaxation occurs, this component will be non zero and asymmetric reflections (*e.g.* (115)) need to be considered to evaluate it.

This procedure forms the basis on which III-V epilayers are analyzed by x-ray diffraction and thereby provides important structural and composition information.

Experimental

Films and superlattices were grown on (001) oriented GaAs wafers in a Varian Gen II Molecular Beam Epitaxy system using heated solid source effusion cells. The system was equipped with a reflection high energy diffraction (RHEED) facility.

Samples were characterized using a Rigaku double crystal high resolution x-ray diffractometer. Cu K_α radiation (40 kV, 30 mA) was used with GaAs reference crystal to give non-dispersed beam from (004) reflections through a 1 mm aperture.

Rocking curves were obtained for samples and include the (004) symmetric reflections at 0 and 180 degree rotation, and asymmetric (115) reflections at glancing exit and glancing incidence angles. Rocking curves were obtained by scanning at 2 to 4 arsecond steps with counting times of about 10 seconds. Lattice mismatch factor, epilayer composition, and superlattice period were obtained for samples as necessary. Rocking curve simulations were performed.

Epitaxial Growth on GaAs

GaAs was grown on (001) GaAs by MBE under various conditions to assess effects on physical parameters. Deposition temperature, flux and sticking coefficients are important factors influencing growth. In ultrahigh vacuum, film growth is governed by kinetics of interaction of the molecular beams and not thermodynamic equilibrium.

For III-V materials it is observed that when the lattice mismatch is small and the thickness of the epilayer is not large, the elastic strain in the layer is not relieved by the formation of misfit dislocations. It is hoped that by studying the growth of GaAs on GaAs, where no lattice mismatch is present, as a function of temperature, the interfacial strain may be observed and assessed by high resolution x-ray diffraction.

Representative rocking curves for 2μ MBE grown GaAs films at 250, 325 and 350 °C are shown in Figure 1. Table I summarizes the data obtained for this series of samples. At a growth temperature of 350 °C, a single sharp (004) reflection is observed. At lower temperatures, the epilayer peak begins to be resolved. As strain is reduced, the splitting between substrate and epilayer peak is lessened. By comparing this data with that of results from rocking curves for other reflections, lattice parameters for the films may be calculated and reconciled with the mechanism of growth described below for a low temperature process. These analyses are still underway.

TABLE I Effect of Temperature on the MBE Growth of GaAs
High Resolution X-Ray Diffraction Rocking Curve Data

GROWTH TEMPERATURE (°C)	PEAK WIDTH SUBSTRATE (arcsecs)	PEAK WIDTH OVERLAYER (arcsecs)	PEAK SPLITTING (arcsecs)
200	21	40	120
250	21	22	47
300	26	-	-
325	48	-	-
350	19	-	0
400	21	-	0

To understand the growth process of GaAs epilayers at various temperatures, it is convenient to begin a discussion with a description of the dynamics of adsorption and desoprtion of the substrate material. The (001) oriented GaAs surface is polar and may be terminated by either Ga or As atoms or a combination of both. Above 300 °C a surface which is arsenic rich, will lose up to about 0.5 monolayer of arsenic as As_2, leaving a Ga rich surface. At temperatures above 550 °C, the dissociative Langmuir evaporation of GaAs becomes significant [7]. Desorption is congruent below 630 °C (that is, the fluxes, J_i, leaving the surface are related by $J_{Ga} = 2J_{As2}$). In this temperature range the evaporation rate of the compound is determined by the desorption rate of Ga with arsenic evaporating as As_2; between 550 °C and 630 °C, evaporation rates may be between 0.01 and 1.0 monolayers per second. Above 630 °C, As_2 is lost preferentially and the free Ga on the surface aggregates to form liquid droplets.

Arsenic fluxes may be comprised of either As_2 or As_4 molecules while the incident gallium specie is monatomic. The interaction of As_2 with a Ga monolayer terminated

surface of GaAs occurs with a sticking coefficient of unity. Stoichiometric GaAs will be produced provided $J_{Ga} < 2J_{As2}$; any excess As_2 lost by desorption. Above about 300 °C other processes become significant. Between 300 °C and 600 °C the sticking coefficient increases. Thus the total flux of As_2 leaving the surfaces is made up of two parts; one from the dissociation of GaAs, which in turn creates a Ga surface population, and from incident molecules which are not absorbed. The sum is constant, but the ratio is dependent on temperature. The surface concentration of gallium and arsenic present during growth will therefore depend on substrate temperature and relative flux intensities.

Growth processes involving As_4 flux are more complex than those involving the dimer [8]. When Ga and As_4 beams interact on a GaAs surface the relative flux ratios strongly influence the As_4 sticking coefficient For $J_{Ga} << J_{As4}$ the sticking coefficient is proportional to J_{Ga} and stoichiometric GaAs is produced. When $J_{Ga} \leq J_{As4}$, however, S_{As4} becomes independent of J_{Ga} but never exceeds 0.5. With As_4 in this temperature range, excess Ga is incorporated into the growing GaAs film despite the fact that not more than half the As_4 supplied to the substrate surface is consumed.

The growth process involving Ga-As_2 interactions on GaAs is a simple first order dissociative chemisorption of As_2 dimer on a Ga surface atom (with the possibility of an association reaction to form As_4 at lower temperatures and of some GaAs dissociation at higher temperatures. The important feature in the model for the growth of GaAs from Ga and As_4 is the pairwise dissociation of As_4 molecules adsorbed on adjacent Ga atoms. From two As_4 molecules four As atoms are incorporated in the GaAs lattice and the

other four desorb as an As_4 molecule.

Single Layer $Al_xGa_{1-x}As$ Analysis

If the strain tensor is obtained for a film/substrate system, the strain free lattice parameter of the film can be determined. In substitutional systems, *e.g.* (Ga,Al)As, the composition of the film can be determined from the strain free parameters by Vegard's Law [9]. Thus the composition of AlGaAs films grown on GaAs were determined in this study.

Since the AlGaAs layer is tilted with respect to the substrate, that is, misoriented from (001) GaAs, its effect on $\Delta\theta$ (the measured angular spacing between the substrate and layer peak) must be considered prior to calculation of Al mole fraction. The tilt is therefore determined by measuring the peak splitting before and after rotating the sample. The epitaxial layer can also influence $\Delta\theta$. This effect is dependent on both the thickness and strain in the layer. Therefore, both symmetric and asymmetric reflections were used to derive Al concentration.

$Al_xGa_{1-x}As$ layers were grown on (001) GaAs using As, Al and Ga sources as described above. Rocking curves taken for the (004) symmetric reflection at 0° and 180° rotation and (115) reflections at glancing incidence and glancing exit angle were recorded and analyzed to obtain the Al composition. A representative (004) rocking curve for an MBE grown $Al_xGa_{1-x}As$ epilayer on (001) GaAs is shown in Figure 2. The epilayer and substrate peaks were split by 79 arcseconds and have full widths at half maximum intensities of 21 and 15 arcseconds respectively. The composition of the

epilayer was determined to be 0.237 Al from the double crystal diffractometer rocking curve data as illustrated by the method below.

$\Delta\theta$ for the (004) and (115) reflections were determined and used to calculate $\Delta d/d$, the strain perpendicular to the interface, using the expression $\Delta d/d = -\cot(\theta^S)\Delta\theta$ [4] (where θ_S is the substrate Bragg angle). Since $\Delta d/d_{(004)}$ is related to $(c - a_s)/a_s$ and $\Delta d/d_{(115)}$ is related to $(a - a_s/a_s)$ (where a, c, and a_s are lattice constants), the following expression could be constructed and solved for $\Delta a/a$: $\Delta d/d = (h^2 + l^2 + k^2)^{-1} \cdot [(h^2 + k^2) \cdot \{\Delta a/a\} + l^2 \cdot \{(c - a_s/a_s\}]$.

Using the average Poission ratio, ν, for GaAs and AlAs, and the expression $c - a_{equiv}/a - a_{equiv} = -2\nu/1-\nu$, values for a and c were obtained ($c = \Delta d/d_{(004)} \cdot a_s + a_s$; $\Delta a/a + a_s$) and used to calculate a_{equiv}. Thus $a_{equiv} = ((2\nu/1-\nu)\cdot a + c)\cdot((1+(2\nu/1-\nu))^{-1}$. To provide x, the composition, $(a_{equiv} - a_{GaAs}) / (a_{GaAl} - a_{GaAs})$ was calculated.

GaAs/AlGaAs Superlattices

The modulation of strain and composition in III-V superlattices may also be determined from rocking curves as a special case of the laminar structures described above [10]. The existence of the superlattice period has been shown by small angle x-ray diffraction and high angle diffraction experiments [11]. The structure profile, representative of the systems under study, is given in Table II.

TABLE II Profile of 6 nm n-Type GaAs Quantum Well Structure

	PROFILE	*MOLE %*	*THICKNESS*	*DOPANT*	*[Nd-Na]*
	n GaAs		0.5 um	Si	1×10^{18}
	i AlGaAs	15 %	50 nm	None	-----
	i GaAs		0.5 nm	None	-----
	n GaAs		6 nm	Si	1×10^{18}
50X	i GaAs		0.5 nm	None	-----
	i AlGaAs	15 %	50 nm	None	-----
	n GaAs		0.8 um	Si	1×10^{18}
	i GaAs buffer		0.2 um	None	-----
	si GaAs substrate				

To a first approximation, the superlattice structure can be regarded as having a unit cell that has the same lattice parameters parallel to the substrate and a near multiple of the lattice parameter perpendicular to the substrate. The spacings in the system will depend on the lattice constants of the individual layer components and also on the elastic strain components introduced by the lattice mismatch between them. By using simulations, and appropriate models it is possible to determine structure and composition information about superlattice structures.

Figure 3 shows the experimental and computed symmetric (004) rocking curves for a 6.0 nm n-type GaAs quantum well superlattice (with period of 57 nm) deposited on (001) GaAs by MBE. The experimental curve, shown in Figure 3, yielded a periodicity of 57.0 nm with Al mole fraction of 0.171. To confirm the physical parameters, simulations were performed. The model used for computation is given by

a laminar periodic structure with a superlattice period consisting of two layers, GaAs and $Al_xGa_{1-x}As$, each with its own structure factor, thickness and strain and based on dynamical theory of x-ray diffraction [12]. The fitting of the experimental data to the theoretical curve was achieved by trial and error adjustment of the structural parameters (using the RADS simulation program published by Bede Scientific, Ltd.). It was determined that the supperlattice period was 56.3 nm with unit layer thicknesses of 48.5 nm and 7.7 nm for the GaAs and AlGaAs components respectively. The Al mole fraction was determined to be 0.2. This data being in good agreement with experimental and projected profile. However, background and other corrections (for instrument function and crystal curvature) are still under study. Other well structures were studied; including those fabricated with p-GaAs and whose thickness dimensions were as small as 2 nm.

HEMT's and Other Heterostructures

In general, the profile of high electron mobility transistor structures studied were: n-GaAs | $Al_xGa_{1-x}As$ | n-$Al_xGa_{1-x}As$ | $Al_xGa_{1-x}As$ | SI GaAs (where $x \approx 0.27$ and $n = 2x10^{18}$ cm^{-3}). n-GaAs and AlGaAs layer thicknesses varied between 5 to 10 nm and 5 to 50 nm respectively. Systems had been grown on 3" SI Gas wafers from Sumitomo. Source temperatures were typically: Ga, 970 °C; As, 300 °C; Al, 1000 °C; and Si, 1125 °C. The V/III BEP ratio was approximately 11.5. The substrate temperature was typically 900 °C. Hall mobility and density measurements were available for the samples.

Symmetric (004) rocking curves were used to provide the basic data necessary to evaluate the structure and composition of HEMT structures. In general the diffraction

peaks for the samples were sharp with broad pronounced tail structure. Attempts were made to resolve the layer and substrate peaks with curve resolution and rocking curve simulation programs. To date no definitive information on the results is available. Studies of a number of other systems, *e.g.* GaSb and InAs on GaAs, were initiated; however, due to the preliminary nature of the data, results will not be discussed here.

Summary

High resolution x-ray double crystal diffractometry has been used to determine the structure and composition of MBE grown III - V layers on GaAS with the aid of rocking curve simulation. GaAs growth on GaAs, AlGaAs epitaxy, superlatices, HEMT structures, and quantum well structures were studied. Detailed results of this effort will be presented in at least two papers; to be submitted to Physical Review Letters and Materials Research Bulletin. Collaborative efforts between the author and the Surface Interactions Group in MLBM at Wright Laboratory are sought and will be considered in AFOSR extension funding.

Acknowledgement

Support for the author from the Air Force Office of Scientific Research is gratefully acknowledged. Special thanks are extended to Drs. Michael Capano and Walt Haas for sponsoring the author's visit to the Materials Directorate.

References

1. R. E. Williams, Gallium Arsenide Processing Techniques, Artech House, Inc., Dedham, MA, 1984.

2. L. L. Chang and K. Ploog (eds.), Molecular Beam Epitaxy and Heterostructures, Martinus Nijhoff Publishers, Dordrecht, 1985.

3. A. Segmüller, *Thin Solid Films* 154 (1987) 33.

4. M. S. Goorsky, K. F. Keuch, M. A. Tischler, R. M. Potemski, *Appl. Phys. Lett.* 59 (1991) 2269.

5. B. K. Tanner, A. G. Turnbull, C. R. Stanley, A. H. Kean, M. McElhinney, *Appl. Phys. Lett.* 59 (1991) 2272.

6. R. Bennett, J. del Alamo, *J. Appl. Phys.* 73 (1993) 195.

7. C.T. Foxon, J. A. Harveyand B. A. Joyce, J. *Phys. Chem. Sol.* 34 (1973) 1693.

8. C. T. Foxon and B. A. Joyce, *Surf. Sci.* 50 (1975) 434.

9. V. Swaminathan, A. T. Macrande, Materials Aspects of GaAs and InP, Prentice Hall, Englewood Cliffs, Nj, 1991.

10. A. Segmüller, P. Krishna, L. Esaki, *J. Appl. Cryst.* 10 (1971) 1.

11. L. L. Chang, L. Esaki, A. Segmüller, R.Tsu, *Proc. 12th Intl. Conf. Phys. Semicond.* Stutgart, Germany, July 15-19, 1974, pp. 688-692.

12. D. K. Bowen N. Loxley, B. K. Tanner, L. M. Cooke M. A. Capano, *Mater. Res. Soc. Symp. Proc.* 208 (1991) 113.

Fig. 1. - HRDCD Rocking Curves for (004) Reflection From GaAs Epilayers on GaAs

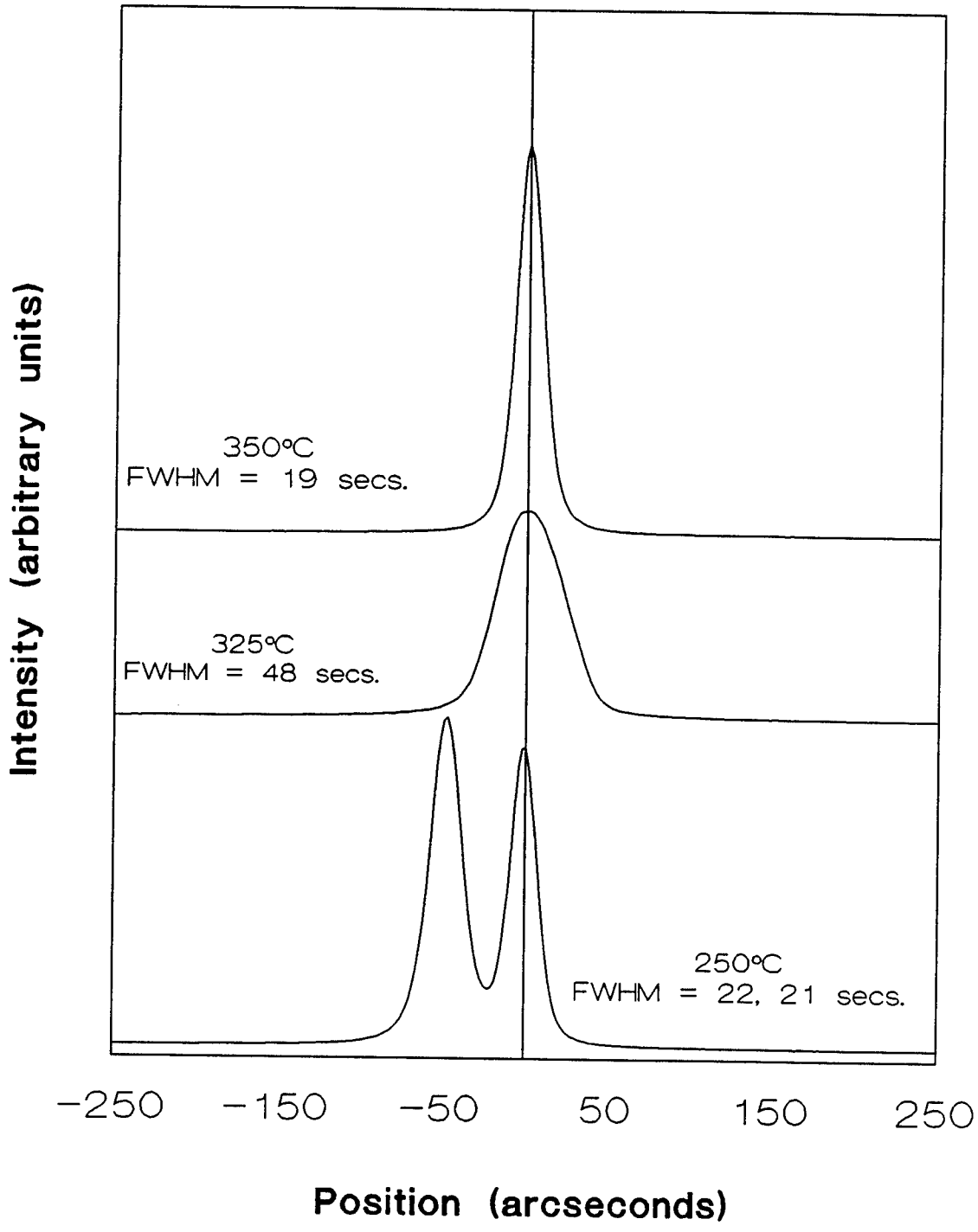


Figure 2. HRDCD Rocking Curve - (004) Reflection of AlxGa1-xAs on (001) GaAs

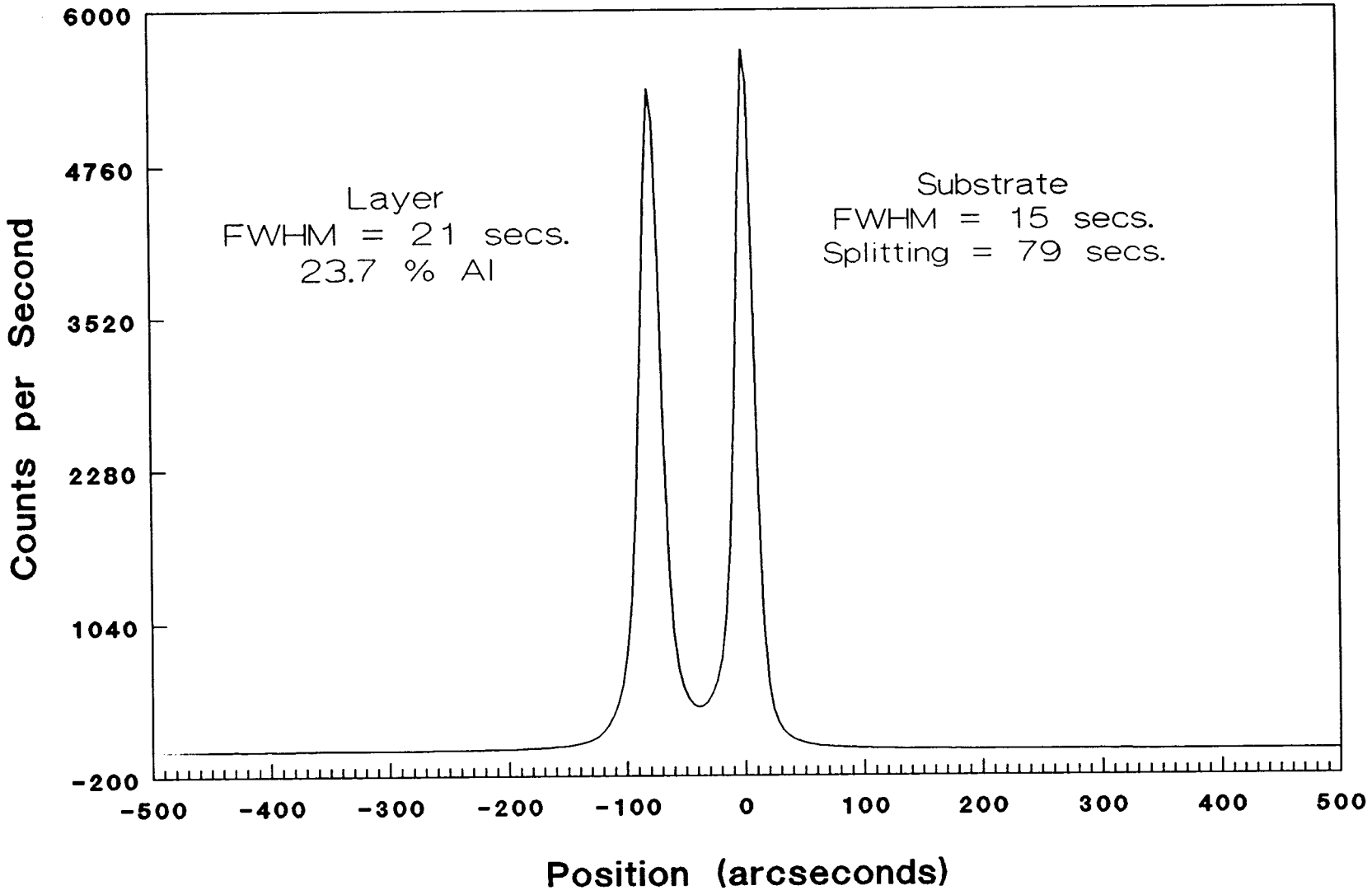
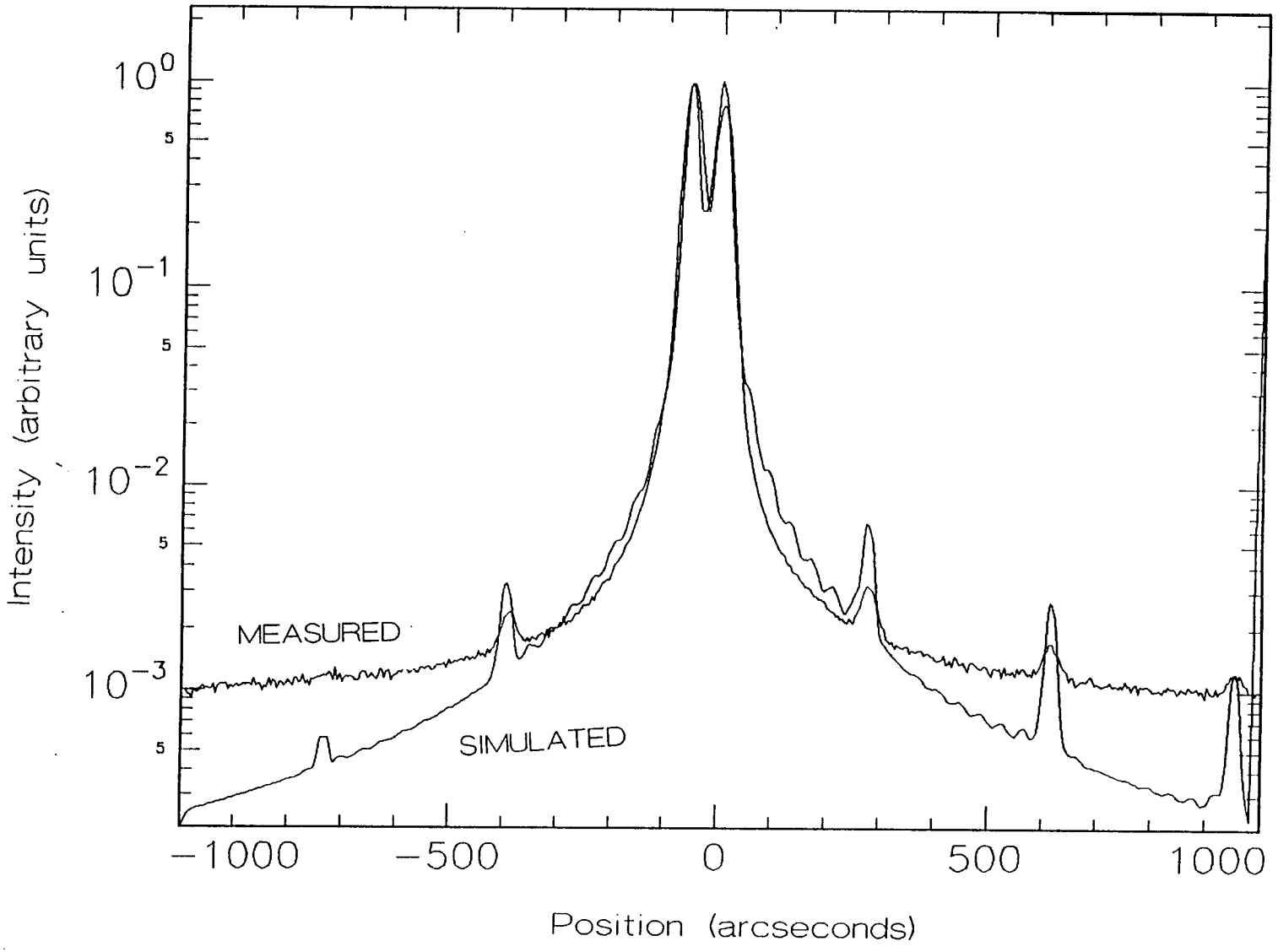


Fig. 3.- Computed and Experimental (004) Rocking Curves of 6 nm Well Superlattice



EXPERIMENTAL STUDIES OF SECOND-HARMONIC GENERATION IN GLASS

Vincent G. Dominic
Assistant Professor
Electro-Optics Program

University of Dayton
300 College Park
Dayton, Ohio 45469-0227

Final Report for:
Summer Faculty Research Program
Wright Laboratories - Materials Laboratory
WL/MLPO

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

August 1993

EXPERIMENTAL STUDIES OF SECOND-HARMONIC GENERATION IN GLASS

Vincent G. Dominic
Assistant Professor
Electro-Optics Program
University of Dayton

Abstract

Focusing intense laser light along with some of its second harmonic into a glass sample transforms the glass into a frequency doubler. We present a new method to measure the optical phase shift between the second-harmonic beam used to seed the glass and the second-harmonic beam subsequently produced by the glass sample. Determination of this phase shift is essential for understanding the growth dynamics of the effect, and its value can discriminate between proposed theoretical models. We also investigated and characterized a second, non-optical technique to transform an ordinary piece of glass into a frequency-doubling element. This second method relies on electric field poling in which a thin sample is heated to ~300 °C and immersed in a strong dc electric field (3 kV/mm). After cooling the sample in the presence of the field a permanent second-order optical nonlinearity is induced in the glass. We studied the polarization properties of the induced nonlinearity to determine if the effect arises from a dc field locked inside the material. The polarization studies show that the optical nonlinearity cannot be explained by simply invoking an internal dc electric field.

EXPERIMENTAL STUDIES OF SECOND-HARMONIC GENERATION IN GLASS

Vincent G. Dominic

Part I - Introduction: Phase-shift measurement

In 1986 Österberg and Margulis[1] observed that illuminating a glass optical fiber with intense infrared light for several hours eventually caused green light to emerge from the far end of the fiber. Because glass has inversion symmetry, such frequency doubling in glass should be forbidden. Although a virgin glass fiber exposed to intense infrared light initially produced no second-harmonic signal, during ~10 hours of infrared illumination green light appeared and slowly increased in strength. A year later Stolen and Tom[2] showed that launching some green light into the fiber along with the infrared light dramatically increased the speed of the process, so that the fiber could now perform as a frequency doubler after only minutes (instead of hours) of illumination. These effects remained unexplained for 6 years but are now understood; the incident green and infrared beams cause a dc electric field to spring up in the glass, and this semi-permanent dc electric field both ruins the inversion symmetry of the glass and permits periodic phase-matching of the frequency-doubling process.

Previously, we showed that the dc electric field is created by charges that migrate in the glass after multiphoton ionization from the intense incident optical fields.[3] Using an optical probe, we experimentally mapped out the detailed shape of the dc electric field in the transverse plane (call it the x-y plane), i.e., in the plane perpendicular to the path of the two original incident light beams. In this paper we measure the relative spatial phase shift $\Delta\theta$ in the propagation direction between the green beam used to seed the glass and the second-harmonic beam generated inside the glass.

Why is this spatial phase shift worth measuring? Because it provides a check on the validity of current theories of second-harmonic generation in glass. The injected light beams at ω and 2ω create a dc electric field in the glass, which then allows second-harmonic generation. Is the green light produced by second-harmonic generation in phase with the green light that originally seeded the process? Recent experiments[4-6] have measured these two green beams to be out of phase by 90°, which is exactly the wrong value for this process to be able to bootstrap up and grow in strength. (Adding a small vector at 90° to an existing vector only rotates the vector's direction but does not increase its magnitude. If the phase is initially different from 90° this paradox is avoided.) We show that the phase shift is not necessarily 90° in all glasses, and is ~44° in our Schott SK5 glass samples.

There have been four previous measurements[4-7] of the phase shift between the seeding and fiber-generated green beams. All these measurements use the interference between light beams that were doubled in separate nonlinear materials. These types of experiments have a long history[8-16] and generally make use of only two nonlinear elements. In the present case we want to compare the relative phase of the seeding green beam and the glass-generated green beam. However, because the generated green light is much weaker than the seeding green beam, in order to directly interfere these two beams the seeding green beam must first be greatly attenuated. The only place to attenuate the seeding green beam and not the glass-generated green beam is before the glass sample, but inserting an attenuator there inevitably shifts the phase of the seeding green beam. Margulis *et al.*[4] studied a germanium-doped glass fiber. They greatly attenuated their seeding green light and measured a phase shift of $\Delta\theta = +99^\circ \pm 9.2^\circ$ (The + sign means that they had to *increase* the optical path length of the seeding green in order to reach the interference fringe peak.) A second measurement[5] also used germanium-doped fibers and a complicated scheme in which the second-harmonic generating crystal that provided the green seeding beam was intentionally phase-mismatched. Determination of the phase shift in this case requires careful consideration of the beam walkoff in the doubling crystal. Their measured phase-shift was $\Delta\theta = -88^\circ \pm 4^\circ$. In ref. [6] the phase shift is measured in a bulk glass sample (Soviet glass ZhS-4) by utilizing non-colinearly polarized green and infrared seeding beams (see also ref. [14]). In this case the generated green and the seeding green are polarized differently and an analyzer may be used to equalize their magnitudes for good fringe visibility. For three different relative orientations of the seeding polarizations the phase shift was found each time to be nearly 90°. In this experimental arrangement, as well as the previous two discussed above, the act of measuring the phase shift tends to perturb it as well. Dianov *et al.*[7] utilized temperature dephasing to deduce that the phase shift was 135° (no error reported), but their conclusion hinges on the charge transport model that they used, which we believe is flawed because it doesn't predict the observed signal growth.

Experimental technique

We developed a new technique to measure the phase shift between the seeding and the glass-generated green light beams. Our technique is similar to the methods discussed above,[4-6] except that we use three frequency-doubling elements and so avoid perturbing the original seeding setup, as shown in Fig. 1. Three doubling elements produce three different green beams: (i) The doubling crystal (LBO) located before the glass sample makes the original green seeding beam. (ii) The glass sample creates the green signal beam. (iii) Another doubling crystal (KTP) located after the glass sample provides a green reference beam. In order to determine the relative phase of

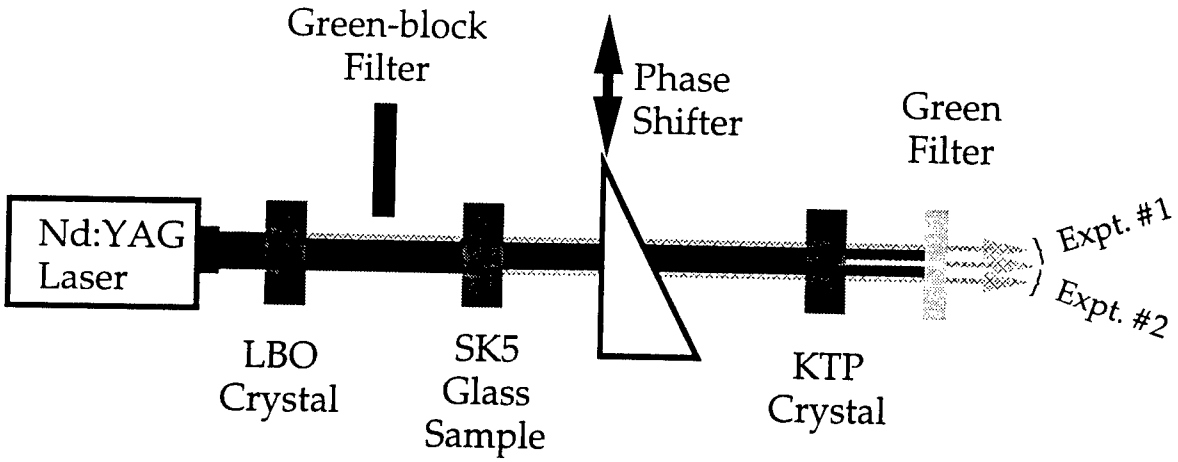


Figure 1. Heuristic experimental schematic showing that we interfere the second harmonic generated in either the LBO crystal (green-block filter out, experiment #1) or the "seeded" SK5 glass sample (green-block filter in, experiment #2) with the second harmonic produced in the KTP crystal. Translation of the glass prism shifts the relative phase of the green (either LBO-generated or SK5-generated) compared to the infrared which is later doubled in the KTP crystal.

beams (i) and (ii), we measure two separate interference patterns, namely that between beams (i) and (iii) and also between beams (ii) and (iii). We then compare the phases of these two interference patterns. If they are the same, it implies that beams (i) and (ii) were in phase, since both interference patterns share the common reference beam (iii). However, if the phases of the two interference patterns differ by an amount $\Delta\theta$, then beams (i) and (ii) must differ by the same relative phase $\Delta\theta$. It is precisely this phase shift between the green seed and the green signal that we want to measure. Notice that instead of attenuating the seeding green beam, we simply block it and replace it with the infrared seeding beam, which we then double in a KTP crystal *after* the infrared light has passed through the glass sample. Any phase shift that the green-blocking filter imparts to the infrared beam appears on *both* the SK5-generated second harmonic and the KTP-generated beam, and so is automatically canceled. This cancellation occurs because second-harmonic generation is a coherent process in which the phase of the generated beam bears a definite phase relationship to the source beam.

The infrared beam incident on the KTP crystal acts as a phase reference in our measurements. Consequently, all phase shifts introduced by optical dispersion between the green and infrared beams are irrelevant unless they occur between the glass sample and the KTP crystal. For example, any phase shifts caused by the multitude of polarizers and waveplates located between the laser and the glass sample will not affect our measurements; they cancel out when we take the difference of the two interference patterns. The relative phase between the green beam incident on the KTP crystal and the green beam generated by that crystal will shift as the two beams propagate through the crystal, because the crystal is birefringent and the polarizations of the

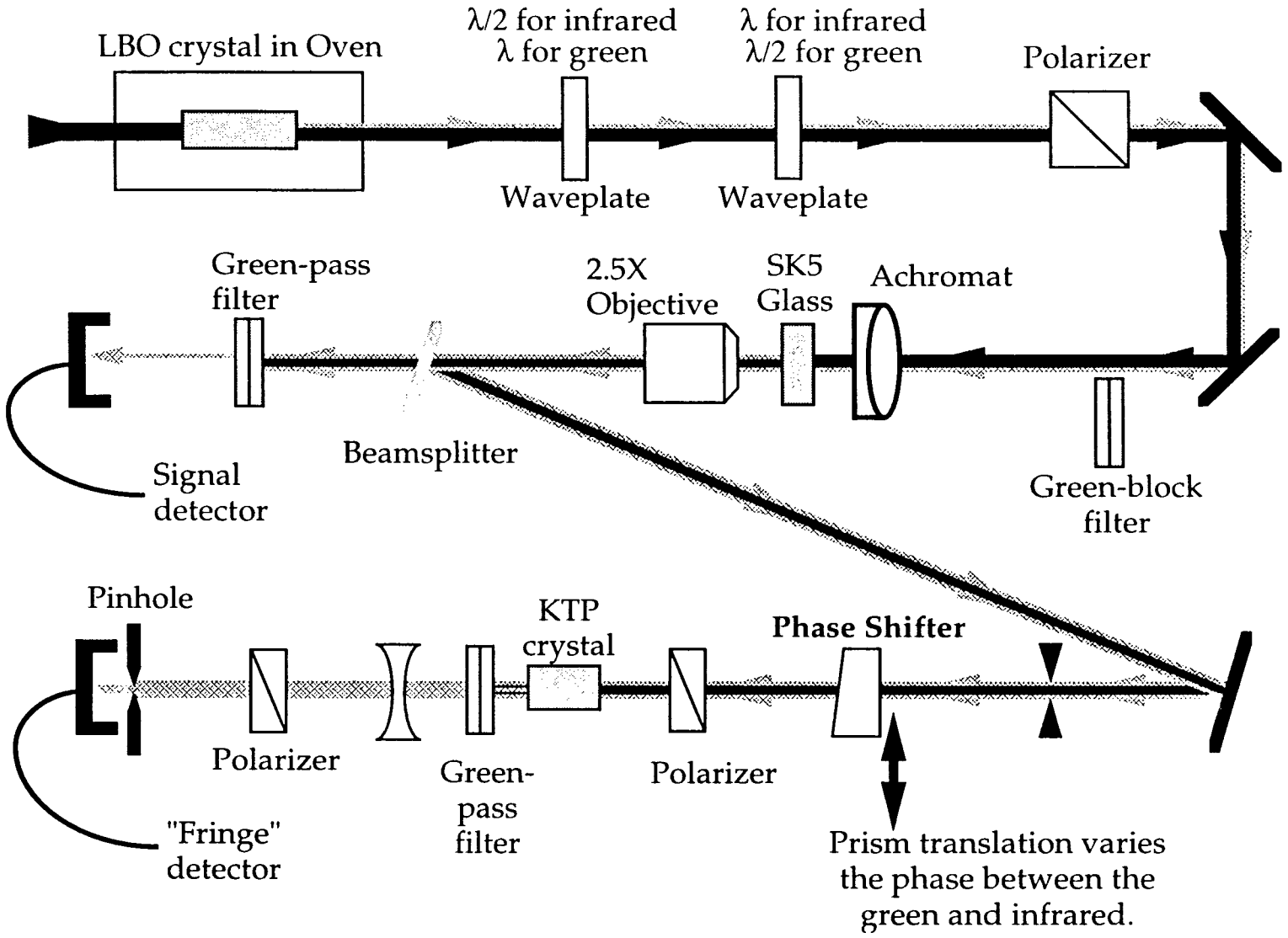


Figure 2. Full experimental layout for measuring the phase shift between the green produced by the LBO doubling crystal (which is used for seeding) and the green produced by the SK5 glass sample. The third frequency doubler, the KTP crystal, provides a reference green beam that we interfere with either the LBO- or glass-generated second harmonic.

two green beams are different. However, this phase shift also cancels out because it is present in both interference patterns.

Our full experimental setup is shown in Fig. 2. We use a mode-locked (76 MHz repetition rate) and Q-switched (1 kHz pulse rate) Nd:YAG laser (Coherent Antares). Some of the infrared (1.064 μm) from the laser is frequency doubled in a lithium triborate (LBO) crystal that is non-critically phase matched, so that the green and infrared beams follow the same optical path. We adjust the powers of the infrared and green seeding beams with two harmonic waveplates and a polarizer. We use 2.5 W of IR power and 1.5 mW of green power to seed a virgin location in our Schott SK5 glass sample. At the focus these powers correspond to 4×10^{11} W/cm^2 and 3.2×10^8 W/cm^2 of peak power for the infrared and green beams, respectively. We monitor the growth of the second-harmonic light generated in the glass sample by momentarily and periodically blocking the green seeding light with a spinning wheel, and any green light generated by the glass sample is detected by a photomultiplier tube. After the green signal becomes sufficiently strong (~1 hour), we monitor, with an apertured detector, the interference between the green beam generated by the original LBO crystal and the green beam generated by doubling of the infrared beam in the KTP crystal. The dispersion of our prism conveniently separates the centers of these two beams in the horizontal direction, so that a high-contrast interference pattern can be obtained by placing a aperture closer to the weaker of the two beams and translating the prism to sweep the interference pattern past the aperture. (The green generated by the SK5 sample is much too weak to affect this interference pattern, since the infrared-to-green conversion efficiency is less than 10^{-5}). After accumulating several fringes, we block the green seeding light from the LBO crystal with a green-block/infrared-pass filter, and return the prism to its original position, being careful to eliminate backlash. We then monitor the interference fringes between the green generated by the glass sample and by the KTP crystal, as shown in Fig. 3. Notice that both of the measured interference patterns share the green light generated from doubling the infrared in the KTP crystal. Because this green beam is a common reference for the two interference patterns, a comparison of the two patterns directly reveals the phase shift between the LBO-generated green (seeding beam) and the SK5-generated green (glass signal beam). We measured the phase shift fifteen times, and we found that the average phase shift (and its standard deviation) was $\Delta\theta = -43.7° \pm 2.6°$. In Fig. 3 an increase in the prism position corresponds to <u>increasing</u> the glass thickness in the optical path. We found that moving the prism by 305.3 μm gave one full fringe. From the dispersion of the BK7 prism we expected a value of 309.2 μm.

We found an interesting complication in our experiment: the measured phase shift $\Delta\theta$ had different values if we probed slightly different locations in the glass sample, and the phase shift also depended on the location of the aperture in front of the "interference fringe" detector. Suppose

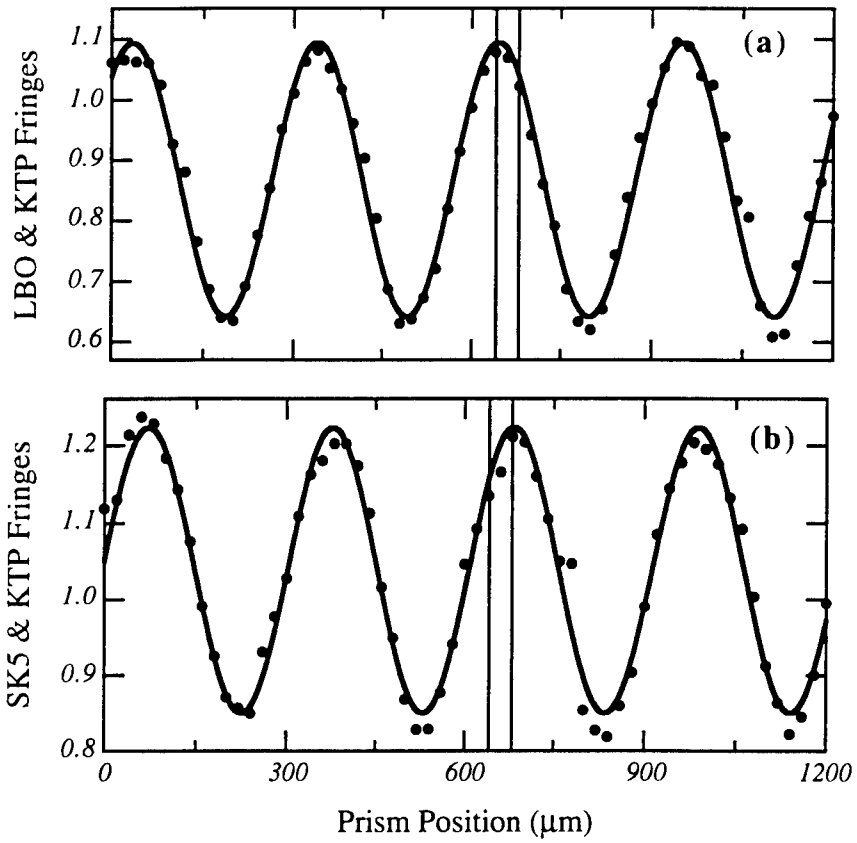


Figure 3. (a) First we observe the interference pattern between the LBO- and KTP-generated second-harmonic beams by translating a phase-shifting glass prism. (b) Next, we block the green seeding beam (LBO-generated) and measure the interference between the SK5- and KTP-generated second-harmonic beams. The phase shift between the LBO-generated seeding beam and the SK5-generated signal beam is shown directly as the shift between these two patterns.

the incident seeding beams propagated in the z direction and both beams were polarized in the y direction. Figure 4 shows the variation in the measured phase shift as the glass sample was translated in the y direction, that is, transverse to the direction of the light beams. In Fig. 4a we show that if the aperture is above the centerline of the two interfering beams, then the measured phase shift changes rapidly for small vertical displacements of the sample. Fig. 4b shows the converse case when the aperture is below the centerline. To obtain Fig. 4c we carefully centered the aperture and observed that the measured phase shift is uniform in the central region, but undergoes an abrupt step of 180° towards the wings. Horizontal displacement of the sample did not affect the measured phase shift. This dependence of the phase shift on the position of the probing beam is expected from the spatial shape of the dc electric field locked inside the glass sample. For the case in which the seeding beams are both vertically polarized, Fig. 5 shows the

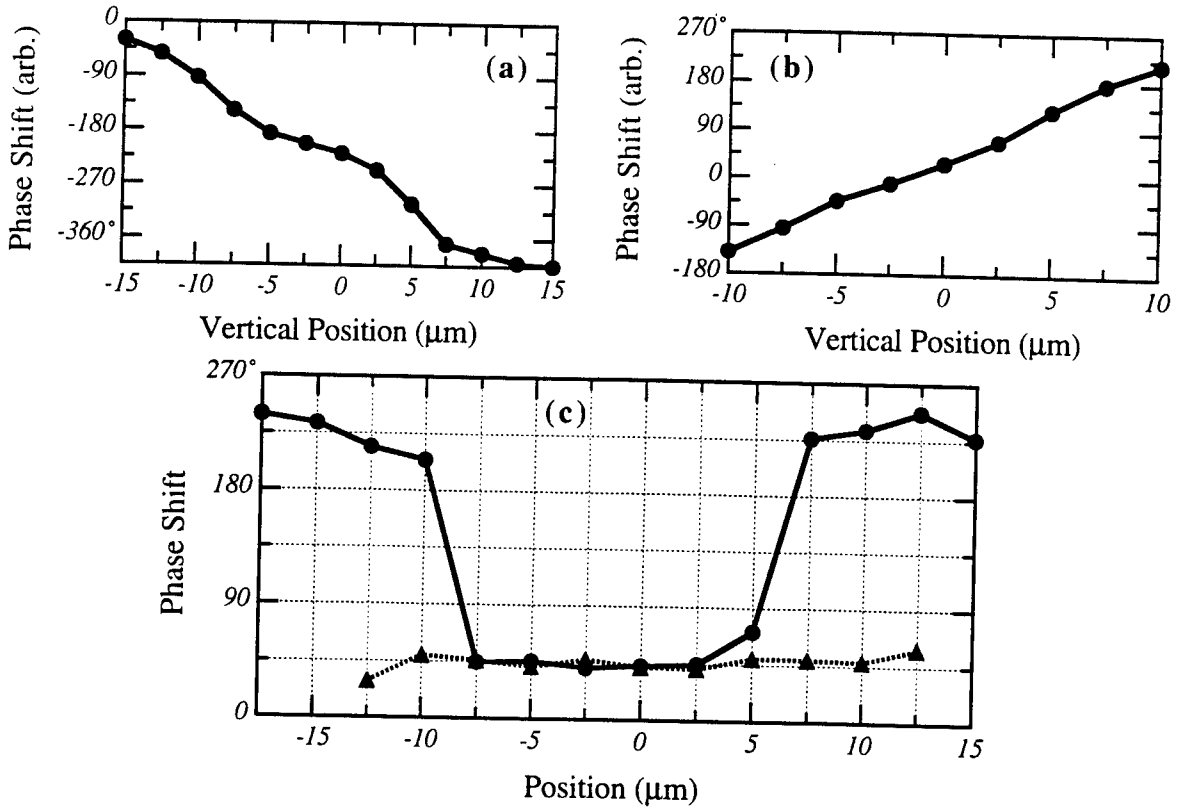


Figure 4. We demonstrate here that the measured phase shift is sensitive to both the position of the aperture before the "interference fringe" detector and the location that the reading beam probes the dc electric field. (a) The aperture is too high and the phase shift sharply decreases as the probe beam is scanned vertically. (b) Same as (a) but here the aperture is too low. (c) When the aperture is on the centerline, the measured phase shift is flat in the central region but jumps by π as the probe is moved vertically. The dashed line in (c) shows that the phase shift does not change as the horizontal position of the probe is varied.

measured shape of the dc-electric field in the glass.[17] Imagine probing this transverse dc field pattern in its center, where the dc electric field is directed up, in the +y direction. In this central region the phase of the dc electric field is not changing, but if the probing beam is moved either up or down it eventually reaches a region where the dc electric field switches sign, and now points down in the −y direction. Switching the sign of the dc electric field imparts an extra 180° phase shift on the green beam generated in the glass sample (as we will show in Eq. 2 below), and explains the abrupt phase change seen in the data of Fig. 4c. Notice that translating the probe horizontally along the centerline scans a dc field that always points in the same direction, and so we expect and observe no dependence of the measured phase shift as we scan the horizontal position of the probing beam, as shown by the dashed line in Fig. 4c. Only by carefully adjusting both the aperture position and the sample position is the flat-top data of Fig. 4c obtained, and the actual phase shift is revealed. If these precautions are not taken, then an arbitrary phase shift value can be obtained. This complication was not recognized in previous phase-shift measurements.

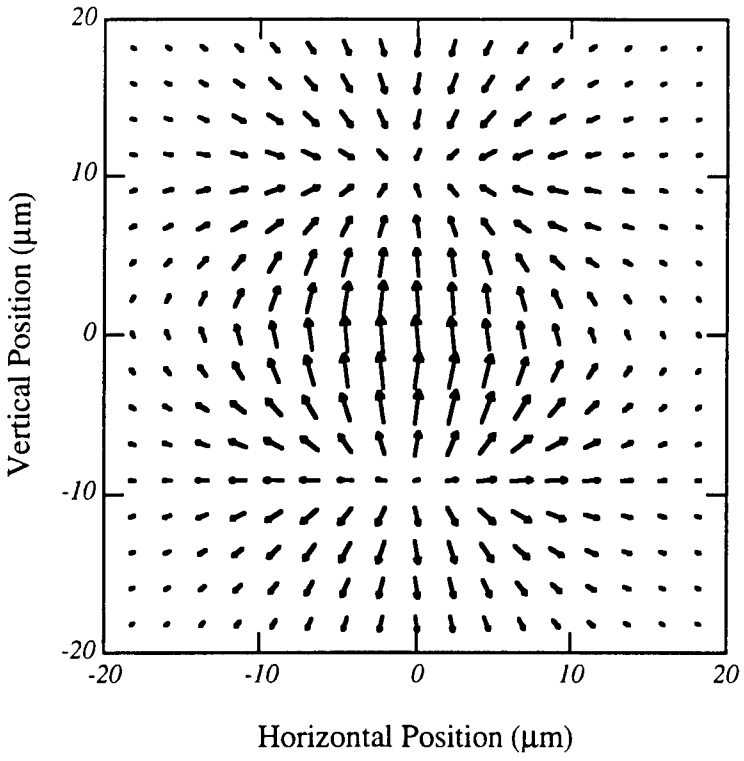


Figure 5. Measured shape of the light-induced dc electric field for the case where both of the seeding beams are vertically polarized. We map the field direction in the plane transverse to the light propagation direction. As you move along the vertical axis of symmetry the field switches direction at two points. As you move along the horizontal axis of symmetry, the field direction is "up" along the entire line. The shape of this dc field explains the phase-shift behavior demonstrated in Fig. 4.

Phase shift of the dc electric field

In our experiment, the intensity of the glass-generated green is always much less than that of the LBO-generated green that seeds the glass. In this case the LBO-generated green beam seeds the dc field and sets the field's phase, and the weak, glass-generated green beam does not affect its own generation. The simplest model[2] for producing the light-induced dc electric field leads to the following relationship between the dc electric field and the seeding optical fields:

$$\begin{aligned} E_{dc} &= \Gamma\, E_\omega\, E_\omega\, (E^*_{2\omega})_{LBO}\, e^{i(2k_\omega - k_{2\omega})z} \\ &= |\Gamma| E_\omega\, E_\omega\, (E^*_{2\omega})_{LBO}\, e^{i(2k_\omega - k_{2\omega})z - i\phi_\Gamma} \end{aligned} \qquad (1)$$

where ϕ_Γ is the phase-shift between the dc electric field and the interference pattern, and $(E_{2\omega})_{LBO}$ is the green seeding beam generated in the LBO crystal. This simple equation suffices for heuristics, as we discuss below. We wish to determine the value of ϕ_Γ. The coupled-mode equation for dc field-induced second-harmonic generation is:

$$\frac{dE_{2\omega}}{dz} = i \frac{3\omega}{2n_{2\omega}c} \chi^{(3)} E_{dc} E_{\omega} E_{\omega} e^{i(2k_{\omega} - k_{2\omega})z} \quad (2)$$

Inserting Eq. 1 into Eq. 2 and keeping only the phase-matched terms in the interaction, we find:

$$\frac{dE_{2\omega}}{dz} = i \frac{3\omega}{2n_{2\omega}c} \chi^{(3)} |\Gamma| |E_{\omega}|^4 (E_{2\omega})_{LBO} e^{i\phi_{\Gamma}} \quad (3)$$

Note that the phase-shift $\Delta\theta$ between the glass-generated green and the seeding green is $\phi_{\Gamma} = \Delta\theta - 90°$; that is, the generated green is 90° degrees shifted from ϕ_{Γ} because of the factor of i in the wave equation.

Our measurements show that the dc field induced inside of SK5 is not in phase with the three-field interference pattern $E_{\omega}^2 E_{2\omega}^*$ that served to create the dc field. This photo-induced effect is nonlocal and the phase shift is $\phi_{\Gamma} = -133.7°$. We measured a similar value for the phase shift in Schott glass SK4 under the same seeding conditions. These measurements contrast sharply with the three reported[4-6] phase shift values of 90° in Ge-doped fused silica[4,5] and in Soviet glass ZhS-4, where a local response was found.

Complications

We discuss potential sources of error in our measurement. First, we observed that even after one hour of seeding the signal generated by the SK5 sample had not completely saturated. Our phase-shift measurements were all performed after one hour of seeding, and the phase shift for a saturated signal might be different. As evidence for this fact we observed that increasing the power of the green or the infrared seeding beams and seeding for one hour led to larger values for the measured phase shift. (Higher seeding power caused the second-harmonic signal to saturate sooner.) Also, when we monitored the phase shift as a function of time we found that it slowly increased. Therefore, the measured phase shift depends on the exposure time and the seeding powers.

Our light beams are focused rather tightly into the SK5 glass sample, and so we must consider nonlinear optical phase shifts such as self-phase modulation and cross-phase modulation. If the infrared power were the same for both fringe-shift measurements, then self-phase modulation of the infrared and cross-phase modulation of the infrared on the green would be common to both fringe measurements and so would cancel. Unfortunately, the green-blocking filter that we insert to reflect the LBO-generated green unavoidably decreases the infrared power by 20%. The change in the self-phase modulation caused by this 20% change is common to both the glass- and the KTP-green, since both are generated from the infrared. However, the effect of

cross-phase modulation will not cancel, since the glass-generated green will suffer from this phase modulation but the KTP-generated green will not. The effect of importance here is the difference between the cross-phase modulation suffered by the LBO- and the glass-generated green. This produces an additional phase of:

$$\Delta\theta_{XPM} = 4\pi\, n_2\, \Delta I_\omega \frac{L}{0.532\ \mu m} \tag{4}$$

where ΔI_ω is the change in the infrared intensity and n_2 is the intensity coefficient of the refractive index ($3.2\text{x}10^{-16}$ cm^2/W in fused silica and similar in SK5). If the interaction length is the depth of focus of the infrared beam (measured to be ~340 μm in air and so inferred to be ~510 μm inside the glass sample), then the phase change from cross-phase modulation will be $\Delta\theta_{XPM} = 17.8°$. However, the flexibility of our experimental technique allows us to measure the dependence of the phase shift on the infrared reading power without disturbing the induced nonlinearity (because we block the seeding green to avoid perturbing the induced effect while still measuring the relative phase of the sample-generated green signal). We found that cross-phase modulation altered the measured phase shift with a slope (in terms of the average infrared power) of ~6°/W. The 20% change in the infrared reading power caused by inserting the green-block/infrared-pass filter will therefore alter the measured phase shift by only 3°. The extreme intensity dependence of the second-harmonic generation growth rate[18] implies that the actual interaction length in Eq. 4 is probably much less than the gaussian focal depth of the writing beams. This partially explains the lower-than-expected dependence of the phase shift on the reading intensity.

Equation 1 states that the optical field product $E_\omega^2\, E_{2\omega}^*$ is the source term for the dc field (in the usual complex representation of the optical fields). Although the actual source term may contain more complicated dependencies on the optical fields, these extra terms always contain the product of $E_\omega^2\, E_{2\omega}^*$ *and* pairs of fields appearing in both conjugated and unconjugated form, so that their phase information cancels. Thus, while $E_\omega^2\, E_{2\omega}^*$ may not describe the entire intensity dependence of the production of the dc electric field, it is precisely the spatially periodic driving term responsible for the spatially periodic dc field created in the glass.

What is the physical origin of the phase shift between the dc electric field and the driving term $E_\omega^2\, E_{2\omega}^*$? This phase shift arises from the quantum mechanical scattering phase shift of an ionized electron's wavefunction compared to a free wave expanding from the same location. The electron is moving away from an attractive potential, and so it is delayed compared to a freely traveling electron. The value of this scattering phase shift depends on the form of the binding potential that tugs on the electron as it flees, and also on the momentum of the electron. For the

case of a Coloumbic potential the phase shift can be given in closed form. In more complicated cases the phase shift is not known.

Conclusions - Phase-shift measurement

The multiphoton ionization interference[19,20] model for the light-induced optical nonlinearity in glass leads to a natural phase shift between the driving term for the interference ($E_\omega^2\ E_{2\omega}^*$) and the dc electric field. This phase shift arises from the quantum mechanical scattering shifts inherent in multiphoton ionization. This spatial phase shift affects the growth and saturation of the dc electric field in much the same manner that spatial and temporal phase-shifts affect other light-induced optical nonlinearities such as the photorefractive effect.[21] We show that the phase shift is not 90° as previously reported, but is –44° in our glass samples. This value of the phase shift was very reproducible under our seeding conditions, but we also observed that the phase shift depends on the seeding duration and the optical seeding powers. We will extend our work by measuring the phase shift in Ge-doped fused silica, and in other bulk glasses to determine whether the phase shift depends on the glass composition.

We believe that our experimental technique offers several benefits for studying the phase shift. Because we can determine the phase shift without disturbing the seeding setup, we can easily monitor the phase shift as a function of the seeding time and as a function of the reading intensity. This flexibility results from introducing the controllable path length variation *after* the glass sample. Also, the horizontal displacement of the infrared and the green from the phase-shifting prism means that we can prudently choose a location (closer to the weaker green beam) where the fringe visibility is high. We have also showed that the probe beam and the aperture must be placed properly for accurate determination of the phase shift. This complication especially impacts experiments in bulk glasses, but may also play a role in fiber experiments if the fiber is not single mode at the second-harmonic wavelength.

Part II - Introduction: Poling of fused quartz

Application of a dc electric field to an isotropic material turns the sample into an electro-optic device. The electric field breaks the original inversion symmetry and creates a second-order optical nonlinearity that generally disappears upon removal of the applied field. Recently, however, Myers *et al.*[22] discovered a way to electrically pole ordinary fused quartz by heating the glass to ~300 °C and applying a dc electric field. This caused the sample to retain its second-order optical nonlinearity indefinitely after cooling. The resulting strength of the nonlinearity is ~1 pm/V which compares favorably with the best nonlinear crystals (lithium niobate: $LiNbO_3$ ~ 30 pm/V).

Since glass is cheap, sturdy, and easily fabricated into optical waveguides, this induced nonlinear optical effect will be extremely useful in making electro-optic waveguide devices.

The originally proposed microscopic mechanism for this poling effect in amorphous quartz invoked migration of mobile ions (probably Na^+) and the subsequent formation of a space-charge electric field near the anode surface of the glass sample.[22] The optical nonlinearity arises from electric-field induced second-harmonic generation which obeys the following relationship between the nonresonant third-order susceptibility $\chi^{(3)}$ (symmetry-allowed in all materials) and the second-order nonlinear susceptibility $\chi^{(2)}$:

$$\vec{\vec{\chi}}^{(2)} = 3\vec{\vec{\chi}}^{(3)} \cdot \vec{E}_{dc}. \tag{5}$$

The direction of the dc electric field $\vec{E}_{dc}$ produced by poling the sample is along the direction of the applied voltage. The symmetry properties of $\vec{\vec{\chi}}^{(3)}$ are well known and it is easy to show that the electric polarization at the second-harmonic frequency $\vec{P}_{2\omega}$ produced by the $\vec{\vec{\chi}}^{(2)}$ nonlinearity in Eq. 5 is related to the probing fundamental field $\vec{E}_{\omega}$ by the following vector relationship:

$$\vec{P}_{2\omega} \propto \chi^{(3)}_{xyyx}\, \vec{E}_{dc}\left(\vec{E}_{\omega} \cdot \vec{E}_{\omega}\right) + 2\chi^{(3)}_{xxyy}\, \vec{E}_{\omega}\left(\vec{E}_{\omega} \cdot \vec{E}_{dc}\right). \tag{6}$$

Only the component of the polarization density that is perpendicular to the propagation direction of a wave radiates an optical beam into the far field. If the direction of propagation of the generated second-harmonic beam is labeled $\hat{k}_{2\omega}$ (parallel to $\hat{k}_{\omega}$ in isotropic samples) then the component of the polarization that radiates the second harmonic is given by:

$$\vec{P}^{\perp}_{2\omega} = \vec{P}_{2\omega} - \hat{k}_{2\omega}\left(\hat{k}_{2\omega} \cdot \vec{P}_{2\omega}\right). \tag{7}$$

We choose $\hat{z}$ as the direction of propagation, $\hat{x}$ is vertical and $\hat{y}$ is horizontal. If we express all vectors in this coordinate system then we find that only the component of the dc field $\vec{E}^{\perp}_{dc} = \left(\vec{\bar{1}} - \hat{z}\hat{z}\right) \cdot \vec{E}_{dc}$ contributes to the signal. If the second-harmonic signal is detected after the beam passes through an analyzer characterized by a polarization vector $\hat{e}_a$, then the signal strength is:

$$I_{2\omega} \propto \left|\chi^{(3)}_{xyyx}\left(\hat{e}^{*}_{a} \cdot \vec{E}^{\perp}_{dc}\right)\left(\vec{E}_{\omega} \cdot \vec{E}_{\omega}\right) + 2\chi^{(3)}_{xxyy}\left(\hat{e}^{*}_{a} \cdot \vec{E}_{\omega}\right)\left(\vec{E}_{\omega} \cdot \vec{E}^{\perp}_{dc}\right)\right|^2. \tag{8}$$

Since the direction of the applied poling electric field is known, these polarization properties are easily checked by controlling the state of polarization of the probing fundamental beam and

viewing the produced second-harmonic beam behind a polarization analyzer. We performed these experiments and found discrepancies with the theoretical predictions.

Description of the experimental apparatus

We used a poling rig designed for contact-poling of organic polymer waveguides. The electrodes contact both sides of a microscope slide sample (2" x 1" x 1 mm) composed of commercial-grade fused quartz. We heated the electrodes (2 cm diameter, brass) to ~350 °C and applied 3 kV across the thin dimension of the sample. The sample remained at high temperature for ~20 minutes with the poling field applied before we turned the electrical heater off. After cooling for ~40 minutes the glass slide returned to room temperature and the poling field was removed. After poling we noticed a change to the surface of the glass slide that made it frost when damp, moist air was directed at the surface, however, this change occurred even without application of the poling field. The visible appearance of the sample was otherwise unchanged by the poling procedure and the measured absorption spectrum showed no distinct alterations.

The optical setup used to probe the induced second-harmonic generation effect is shown in Fig. 6. We used a Q-switched (1 kHz) and mode-locked (82 MHz) Nd:YAG laser operating at 1.064 μm with approximately 300 mW of average power focused with a 50 mm lens. The peak power is ~100 kW. We set the input infrared polarization with the λ/2 quartz half-wave waveplate immediately preceding the input lens and we monitor the second-harmonic signal behind a dichroic sheet polarizer that follows the re-collimation lens. A harmonic beamsplitter (reflects 532 nm, transmits 1.064 μm), an infrared absorbing glass filter, a 532 nm interference filter, and a photomultiplier tube serve to isolate and detect the signal. The output of the photomultiplier tube is connected to a 1 MΩ oscilloscope and a Stanford Research SR510 Lockin amplifier. The λ/2 waveplate and the analyzing polarizer are mounted on motorized rotation stages controlled by a Newport 855 motion controller. The data acquisition process is computer controlled.

The sample is attached to a vertical post so that the probe beam angle-of-incidence may be altered to perform the Maker-fringe experiment.[23,24] We rotate the sample about a vertical axis so that the proposed dc electric field lies entirely in the horizontal plane. If the thickness of the nonlinear region (spatial extent of the internal dc field) is much less than a coherence length ($l_c = \lambda/2/[n_{2\omega} - n_{\omega}]$ = 48 μm in fused silica), then we won't observe any Maker fringes as the sample is rotated. This is exactly what we observe as shown in Fig. 7. At normal incidence we observe no second-harmonic signal since the dc electric field has no component perpendicular to the propagation direction. As the incidence angle is increased, the projection of the dc field increases, resulting in increased second-harmonic generation. At very high angles of incidence,

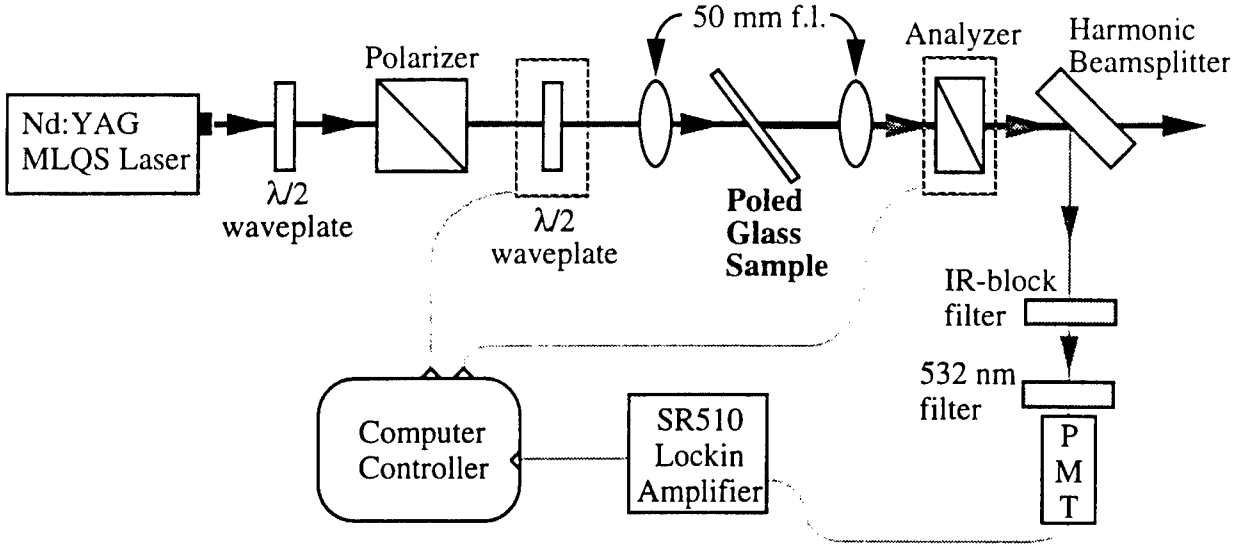


Figure 6. Schematic showing the experimental arrangement used to measure the polarization properties of the poled fused quartz samples. The half-wave waveplate (immediately before the input lens) and the analyzer are mounted on motorized rotation stages. We can rotate the probing polarization and the analyzer transmission axis to any orientation. We measure the complete behavior of the second-harmonic signal as either the input polarization, the analyzer orientation, or both are rotated.

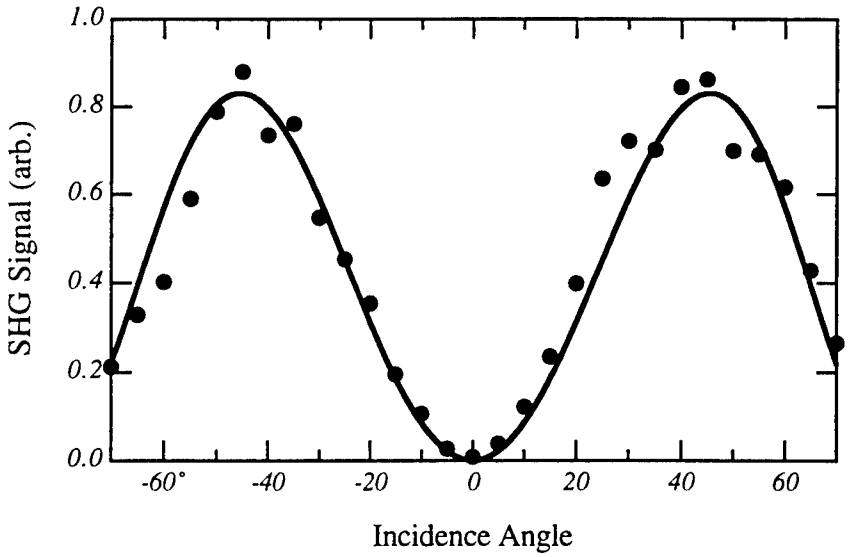


Figure 7. Results of the Maker fringe experiment in which we measured the second-harmonic generation (SHG) signal as the angle of incidence of the probe beam on the poled glass sample is varied. Also displayed is the single-parameter theoretical fit to the data which assumes that the active nonlinear region of the sample is much less than the coherence length.

the Fresnel reflectivity of the air-glass interface begins to reflect a significant fraction of the incident infrared probe beam and the generated signal and the detected signal falls accordingly. Figure 7 shows the data as well as a theoretical fit utilizing only an overall magnitude parameter. The

behavior displayed in Fig. 7 demonstrates that the nonlinear region is much shorter than one coherence length since we don't observe an oscillating signal versus incidence angle.[23,24]

We determined the polarization dependence of the generated second-harmonic signal by performing the following experiments: 1) Set the analyzer horizontal (p-polarized) and rotate the input polarization starting from horizontal. 2) Set the analyzer vertical (s-polarized) and rotate the probing polarization. 3) Set the input polarization horizontal (p-polarized) and rotate the analyzer. 4) Set the input polarization vertical (s-polarized) and rotate the analyzer. 5) Rotate the input polarization and the analyzer together so that they remain parallel ($\hat{e}_\omega \parallel \hat{e}_a$). 6) Rotate the input polarization and the analyzer together so that they remain crossed ($\hat{e}_\omega \perp \hat{e}_a$). Equation 8 shows that experiments of type #5 and #6 both determine the projection of the internal dc electric field onto the direction of the analyzer. We show in Figs. 8-10 the data from these six experiments and we have grouped them as rotating-input-polarization experiments (Fig. 8), rotating-analyzer experiments (Fig. 9), and rotate-both experiments (Fig. 10). In Fig. 10 we plot the data in cylindrical coordinates so that for a given orientation angle θ of the analyzer, the distance from the origin in this direction gives the strength of the measured signal. In all these figures we also show the best theoretical fit to the data obtained using Eq. 8 along with the condition $\chi^{(3)}_{xyyx} = \chi^{(3)}_{xxyy}$, which is true for non-dispersive materials (Kleinman's conjecture). We see that there is stark disagreement between the measured data and the theoretical fits utilizing the internal dc electric field model.

We performed many checks to determine whether our experimental setup contained an artifact responsible for the disagreement between the theory and the data. We tried probing different locations in the poled region, including the very center, and found no difference in the observed signal. We changed the focusing/re-collimation lenses to 5X objectives to determine if the disagreement was caused by the focusing conditions and found no difference. We found no evidence for birefringence in the poled region of the sample or infrared leakage into the photomultiplier detector. We checked for and ruled out any polarization sensitivity of the detection system. We changed the polarization analyzer to a Glan-Thomson prism and observed no change in the measured dependencies. We insured that the 1.064 μm $\lambda/2$ waveplate gave very pure polarization rotation. We inspected the optical probing beam and observed no spatial asymmetry in its shape. We performed all six experiments for different amounts of reading power and at various angles of incidence and in numerous different poled quartz samples. In no case did we observe significant departure from the behavior displayed in the following figures.

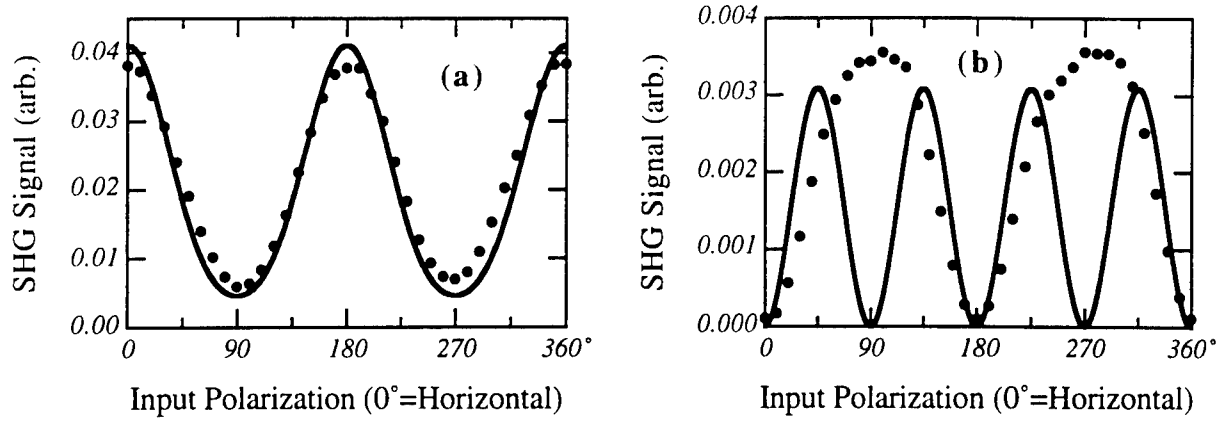


Figure 8. We rotated the input polarization and set the analyzer either horizontal (a) or vertical (b). The experimental data is shown by the circles and the solid lines indicate the best fit using Eq. 8.

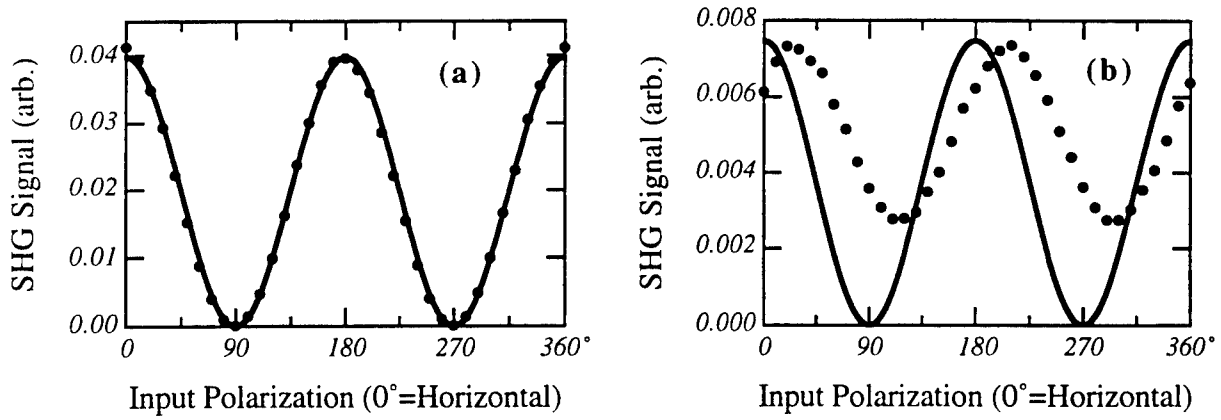


Figure 9. We fix the input polarization either (a) horizontal or (b) vertical and rotate the analyzer.

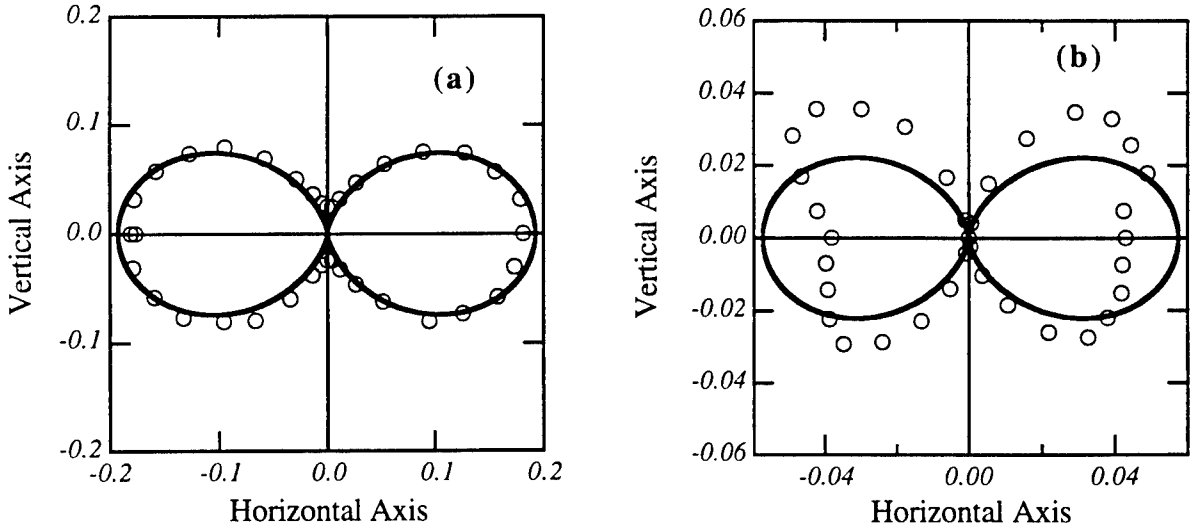


Figure 10. Here we rotate the input polarization **and** the analyzer. In (a) they are parallel, in (b) they are crossed.

Conclusions - Poled fused quartz

We have found strong disagreements between the proposed model for second-harmonic generation in poled fused quartz and our experimental observations of the polarization dependence of the effect. We have checked for and eliminated many possible experimental errors in our setup. We conclude that the induced second-order nonlinearity cannot be explained by simply invoking an internal dc electric field. Several other mechanisms are possible such as microcrystallite formation or alignment of dipolar moieties inside the glass host. We are currently conducting investigations to probe for these possibilities.

Even without a complete understanding of the microscopic mechanism responsible for this unusual nonlinear optical effect, we believe that poled fused quartz holds great promise as a waveguide material for electro-optic and second-harmonic generation devices. This is also demonstrated by the rapidly increasing interest in this phenomenon.[25-28] Glass is cheap, mechanically rugged, easily formed into waveguides, and highly resistant to optical damage. We have begun investigating the electro-optic properties of our poled glass samples and we will extend this research to include a complete characterization of the poling process and and to investigate alternate glasses.

Acknowledgements

The author thanks Rama Vuppuladhadium for assisting in the experimental investigations of poled fused quartz. I would also like to thank Pat Hemenger for making this research possible, and I thank Uma Ramabadran and Dave Zelmon for helpful discussions. Thanks also to Paul VonRichter who designed and built the poling rig.

References

1. U. Österberg and W. Margulis, "Dye laser pumped by Nd:YAG laser pulses frequency doubled in a glass optical fiber," Opt. Lett. **11**, 516-518 (1986); "Experimental studies on frequency doubling in glass optical fibers," Opt. Lett. **12**, 57-59 (1987).
2. R. H. Stolen and H. W. K Tom, "Self-organized phase-matched harmonic generation in optical fibers," Opt. Lett. **12**, 585-587 (1987).
3. V. Dominic and J. Feinberg, "Multiphoton ionization interference explains second-harmonic generation in glass," submitted to Phys. Rev. Lett.
4. W. Margulis, I. C. S. Carvalho, and J. P. von der Weid, "Phase measurement in frequency-doubling fibers," Opt. Lett. **14**, 1346-1348 (1989).
5. K. Koch and G. T. Moore, "Two-color interferometry using a detuned frequency-doubling crystal," Opt. Lett. **16**, 1436-1438 (1991).
6. M. A. Bolshtyansky, V. M. Churikov, Yu. E. Kapitzky, A. Yu. Savchenko, and B. Ya. Zel'dovich, "Phase properties of $\chi^{(2)}$ gratings in glass," Opt. Lett. **18**, 1217-1219 (1993).
7. E. M. Dianov, P. G. Kazansky, D. S. Starodubov, D. Yu. Stepanov, "Observation of phase mismatching during the preparation of second-order susceptibility gratings in glass optical fibers," Sov. Lightwave Commun. **1**, 395-398 (1991).

8. R. K. Chang, J. Ducuing, and N. Bloembergen, "Relative phase measurement between fundamental and second-harmonic light," Phys. Rev. Lett. **15**, 6-8 (1965).
9. J. J. Wynne and N. Bloembergen, "Measurement of the lowest-order nonlinear susceptibility in III-V semiconductors by second-harmonic generation with a CO_2 laser," Phys. Rev. **188**, 1211-1216 (1969).
10. R. C. Miller and W. A. Nordland, "Absolute signs of second-harmonic generation coefficients of piezoelectric crystals," Phys. Rev. **B2**, 4896-4902 (1970).
11. F. A. Hopf, A. Tomita, and G. Al-Jumaily, "Second-harmonic interferometers," Opt. Lett. **5**, 386-388 (1980).
12. H. W. K. Tom, T. F. Heinz, and Y. R. Shen, "Second-harmonic reflection from silicon surfaces and its relation to structural symmetry," Phys. Rev. Lett. **51**, 1983-1986 (1983).
13. G. Berkovic, Y. R. Shen, G. Marowsky, and R. Steinhoff, "Interferences between second-harmonic generation from a substrate and from an adsorbate layer," J. Opt. Soc. Amer. B **6**, 205-208 (1989).
14. N. B. Baranova, A. N. Chudinov, A. A. Shulginov, and B. Ya. Zel'dovich, "Polarization dependence of interference between single- and two-photon ionization," Opt. Lett. **16**, 1346-1348 (1991).
15. B. Ya. Zel'dovich, Yu. E. Kapitzky, and A. N. Chudinov, "Interference between second harmonics generated into different KTP crystals," Sov. J. Quantum Electron. **20**, 1120-1121 (1990).
16. A. N. Chudinov, Yu. E. Kapitzky, A. A. Shulginov, and B. Ya. Zel'dovich, "Interferometric phase measurements of average field cube $E_\omega{}^2E_{2\omega}{}^*$," Opt. and Quant. Electron. **23**, 1055-1060 (1991).
17. V Dominic and J. Feinberg, "Spatial shape of the dc electric field produced by intense light in glass," Opt. Lett. **18**, 784-786 (1993).
18. V. Dominic and J. Feinberg, "Growth rate of second-harmonic generation in glass," Opt. Lett. **17**, 1761-1763 (1992).
19. N. B. Baranova and B. Ya. Zel'dovich, "Physical effects in optical fields with nonzero average cube, $\langle E^3\rangle \neq 0$," J. Opt. Soc. Amer. B **8**, 27-32 (1990).
20. D. Z. Anderson, V. Mizrahi, and J. E. Sipe, "Model for second-harmonic generation in glass optical fibers based on asymmetric photoelectron emission from defect sites," Opt. Lett. **16**, 796-798 (1991).
21. V. L. Vinetskii, N. V. Kukhtarev, S. G. Odulov, and M. S. Soskin, "Dynamic self-diffraction of coherent light beams," Sov. Phys. Usp. **22**, 742-756 (1979).
22. R. A. Myers, N. Mukherjee, and S. R. J. Brueck, "Large second-order nonlinearity in poled fused silica," Opt. Lett. **16**, 1732-1734 (1991).
23. P. D. Maker, R. W. Terhune, M. Nisenoff, and C. M. Savage, "Effects of dispersion and focusing on the production of optical harmonics," Phys. Rev. Lett. **8**, 21-22 (1962).
24. J. Jerphagnon and S. K. Kurtz, "Maker Fringes: A detailed comparison of theory and experiment for isotropic and uniaxial crystals," J. of Appl. Phys. **41**, 1667-1681 (1970).
25. A. C. Liu, M. J. F. Digonnet, and G. S. Kino, "Electro-optic phase modulation in a silica channel waveguide," *Integrated Photonics Research Conference*, paper PDP6, Optical Society of America, Washington, D.C., 1993.
26. P. G. Kazansky, A. Kamal, and P. St. J. Russel, "Erasure of thermally poled second-order nonlinearity in fused silica by electron implantation," Opt. Lett. **18**, 1141-1143, (1993).
27. R. A. Myers, N. Mukherjee, and S. R. J. Brueck, "Large second-order nonlinearity bulk and thin film SiO_2," *Digest of Conference on Laser and Electro-Optics*, **12**, paper CThD6,Optical Society of America, Washington, D.C., 1993.
28. A. Okada, K. Ishii, K. Mito, and K. Saski, "Phase-matched second-harmonic generation in novel corona poled glass waveguides," Appl. Phys. Lett. **60**, 2853 (1992).

FIBER MATRIX INTERFACE - INFORMATION FROM EXPERIMENTS VIA SIMULATION

George N. Frantziskonis
Department of Civil Engineering
and Engineering Mechanics
University of Arizona
Tucson, Arizona 85721

Report for:
Summer Research Program
Metals and Ceramics Laboratory
Nondestructive Evaluation Branch
Wright-Patterson Air Force Base

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

July 1993

AN AMENDMENT CONTAINS FURTHER RESULTS AS WELL AS THE MATERIAL PRESENTED HEREIN AND IS WITH:

Dr. Tom Moran

WL/MLLP

Bldg 655

Wright Patterson AFB

OH 45433-7817

FIBER MATRIX INTERFACE - INFORMATION FROM EXPERIMENTS VIA SIMULATION

George N. Frantziskonis
Department of Civil Engineering
and Engineering Mechanics
University of Arizona

Abstract

This study explores a novel procedure for obtaining quantitative information on the mechanical properties of the fiber-matrix interface in composite materials. The method simulates actual experiments in detail, including fiber breakage, matrix yiel ' and/or cracking, and interface failure. The paper concentrates on two commonly performed experiments, the so-called fragmentation test for metal matrix, and the pushout/pullout test for metal as well as ceramic matrix composites. Based on the documented capability of the technique to simulate actual experimental data, reliable values of interface (homogenized) properties can be obtained. In addition, the simulations provide further understanding of the mechanisms involved during the relevant testing. Although this study presents results from basic problems, the method is general enough to include effects of residual stress, of high temperature environment, of dynamic crack propagation, as well as three-dimensional details of the interface failure process. The potential exists for simulating non destructive wave based techniques aimed at evaluating interface properties.

INTERFACES IN COMPOSITES - INFORMATION FROM EXPERIMENTS VIA SIMULATION

George N. Frantziskonis

Introduction

Physical reasoning suggests that the mechanical properties of composite materials rely significantly on the nature of the interface between fiber reinforcement and matrix. It is the interface that delivers information (kinematic and dynamic quantities) from the matrix to the fiber and vise versa. Failure of composites involves not only failure of fibers and matrix, but also involves the propagation of cracks along (and across as explained subsequently) interfaces. The properties of such cracks, i.e. dissipated energy during propagation, their interplay with matrix/fiber, etc., are decisive for the macroscopic properties of a composite. It is therefore important to understand the interface properties and its role in the overall mechanical performance of a composite. Consequently, interfacial characterization has received intensive attention, from the experimental as well as the analytical point of view.

Various experimental procedures addressing interfacial properties have been designed. Mechanical destructive tests have been and are being used. Recently, attempts to characterize interfacial properties non-destructively, i.e. Karpur et al (1993), have also been examined. It is not intended herein to provide a thorough review of the literature on interfacial properties and testing. However, reference is given to those works directly relevant to the present study. For reviews and trace of the literature we refer to Metcalfe (1974), Evans et al (1991), and the works cited therein. For analysis of micro-mechanical stresses involved we refer to Pagano (1991) and McCartney (1990).

In general, a "universal" experimental procedure designed to identify interface properties for various material combinations has not been identified. This is due to the fact that it is very difficult, if not impossible, to examine interface properties directly - to isolate the interface response. The relevant test measurements are sensitive to the properties of the matrix, the fiber(s),

the interface(s) present, and the geometry and load conditions of the test setup. For example, while the so-called single fiber fragmentation test has proven effective for metal and polymer matrix composites, it is rather inappropriate for ceramic matrix ones. This is discussed further in the following.

The present study focuses on: (a) the single fiber fragmentation test, often performed on metal matrix composites, (b) the pushout and pullout tests on metal and ceramic matrix composites. The following section concentrates on the information available from such tests, followed by the description of the numerical simulation procedure used herein, and presentation of relevant results. Throughout, the paper discusses the method and results critically, and evaluates the potential of the approach. It is stressed that more needs to be understood in this important area of interface properties identification. It seems that the success in doing so depends heavily on close cooperation between experimental work, destructive and non-destructive, and analytical, simulation work.

Experimental Information

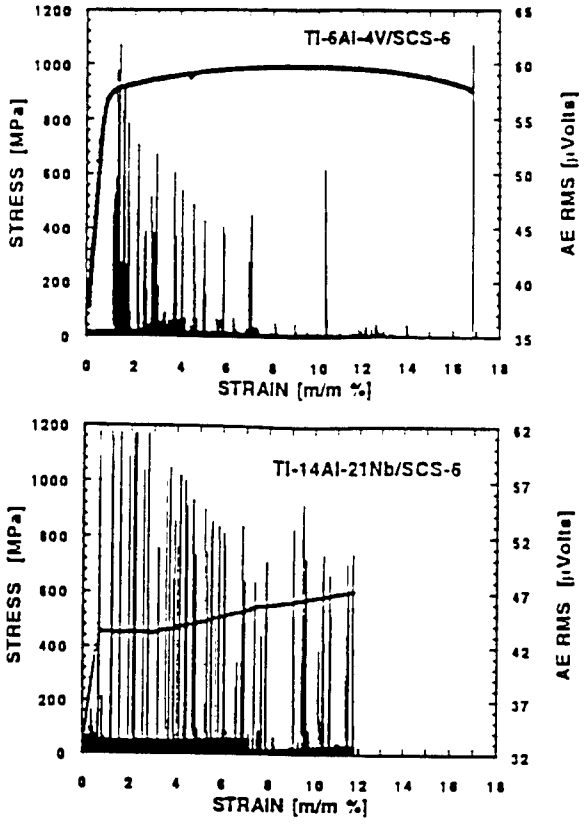
In the single fiber fragmentation test a fiber is embedded in a ductile matrix. The sample is subjected to tensile loading along the fiber axis. Through transfer of load from the matrix to the fiber, at some point the fiber breaks. Further loading results into that the fiber breaks successively into smaller fragments until the fragments become too short to enable further increase in the fiber stress level. Figure 1 (from Roman et al 1993, where also an overview on the single fiber fragmentation test is given) contains typical results obtained from fragmentation test on SCS-6 SiC fiber with Ti-6Al-4V and Ti-14Al-21Nb (wt. %) matrix. According to Roman et al (1993) the Ti-6Al-4V matrix possesses enhanced ductility and shows continuous yielding without yield drop or shear band or localized deformation zone formation. The Ti-14Al-21Nb shows much more complicated response at post yield strains. This study concentrates on the Ti-6Al-4V matrix. As shown in Figure 1 the specimen shows significant amount of plastic yielding. Since the acoustic emission bursts correspond mainly to fiber fracture, Roman et al (1993), it is seen that fiber fragmentation initiates after the matrix (Ti-6Al-4V) has reached its yield stress.

This information is important for identifying interface properties as shown subsequently.

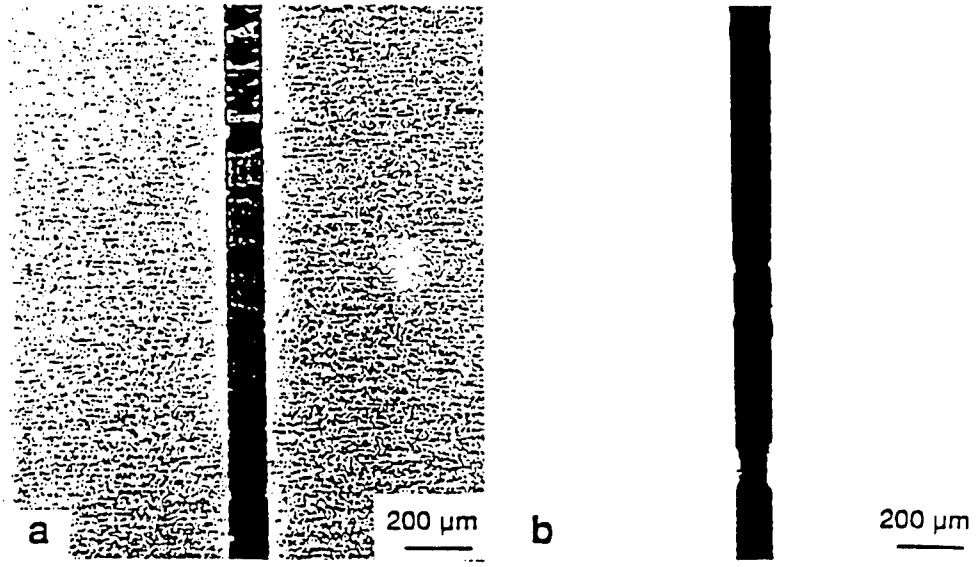
Fragmentation tests are often performed on metal matrix as well as polymer matrix composites. The multiple fracture behavior has been studied mostly through the so-called shear lag analysis which provides a relation among the critical aspect ratio of the fiber, tensile strength of the fiber, and the interfacial shear stress. Several limitations of the method have been identified, since the method neglects the dependence of the interfacial shear stress on the volume and strain hardening characteristics of the matrix, the modulus of the matrix, and the strengths of fiber and matrix. Also, as mentioned by Roman et al (1993) the interfacial characteristics predicted by that method are often very unrealistic. Ochiai and Osamura (1986a,b) have attempted to overcome some of the limitations of the shear lag analysis by considering the details of stress transfer (from matrix to fiber) and plastic stress-deformation response for the matrix. They have also reported numerical results by assigning a Weibull distribution to the fiber spatial strength.

The so-called pushout and pullout tests are commonly performed on ceramic and metal matrix composites. Figure 2 shows a typical configuration for a pushout test. The pullout configuration is similar, where tensile load is applied on the fiber. For the numerical simulations described in the next section we consider pushout and pullout of a SiC (SiC-6) fiber embedded in a Ti-6Al-4V matrix, and in a glass matrix. The length of the fiber pulled/pushed out in metal matrix is much shorter than the length in ceramic matrix composites. This is mostly due to experimental difficulties in testing long fiber lengths in a metal matrix. As shown herein, these geometrical differences have important consequences in the information obtained from the tests.

The literature on the pushout and pullout test is rich. For a review of the reported experimental, analytical work in this area we refer to Kerans and Parthasarathy (1991) for ceramic matrix and to Watson and Clyne (1992) for metal matrix composites. A large number of parameters influence the results from such tests, i.e. non-uniformities due to end effects, residual (radial and axial) stresses, the stability of interface crack propagation, the elastic and fracture parameters of the fiber and matrix. Relevant analytical works examine some of the underlying mechanisms, the result being a better understanding of the problem, i.e. Atkinson et al (1982), Kerans and Parthasarathy (1991), Watson and Clyne (1992).



Typical tensile stress-strain curves and acoustic emission RMS-strain plots for the two single fiber composite systems at room temperature.



Optical micrographs showing a portion of the fragmented fiber in the two composites after tensile testing: (a) Ti-6Al-4V/SCS-6 and (b) Ti-14Al-21Nb/SCS-6.

Figure 1: Typical Load-Displacement Curves and Fiber Fragmentation Pattern Observed in Testing SCS-6 Fiber in Ti-6Al-4V, Ti-14-Al-21Nb Matrices, After Roman et al (1993).

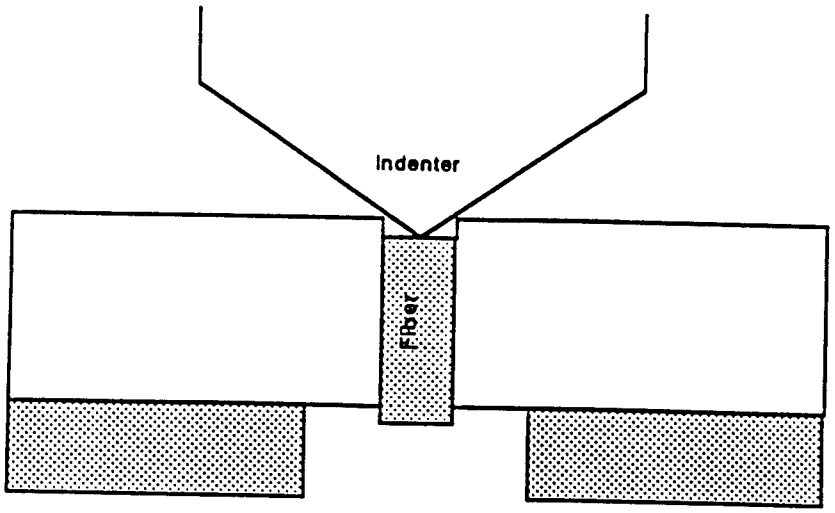


Figure 2: Schematic of Pushout Test

Simulations

Materials, mechanics research has been traditionally carried out through experiments and theoretical analysis. Recently, however, a new trend of computer assisted research has evolved. This branch has been triggered by the rapid progress in computer performance, and from the increasing need for the understanding of systems far more complex than traditional techniques have ever handled. For example, in a typical single fiber fragmentation test, one can identify many complicated processes that take place concurrently - matrix yielding, fiber breaking, interface failure, transfer of stress through the interface and matrix and subsequent fiber fragmentation. It is important that such processes are understood so that the role of the interface can be identified and quantified.

In order to simulate such problems numerically, one may automatically think of the finite element method (FEM). However, within the FEM framework it is difficult to simulate the fracture processes occurring at the microlevel. Such a process would require use of elements much smaller than the crack sizes, significant mesh refinement, in addition to the requirement for a continuum based fracture criterion (at the micro-level) that may be difficult to specify.

In this work, a microscopic representation of the fiber, matrix and interface is achieved

through a so-called lattice. Lattices are being used extensively in different scientific fields, i.e. fluid mechanics, physics etc particularly as a tool to solve differential equations. It seems that such a method for solution of problems within linear elasticity was first investigated by Hrennikoff (1941). There it is shown that a lattice provides a consistent approach to the solution of elasticity problems - the solution converges to the exact elasticity one with lattice spacing reduction. The advantages of using a lattice (over FEM) become evident (as further explained in the sequence) when micro-fracturing is important. Such advantages have been realized by a branch of statistical physics where micro-fracturing in statistically heterogeneous solids are examined extensively, Herrmann and Roux (1990), Charmet et al (1990). A number of works simulating the process of micro-fracturing in composite materials using a lattice discretization have been reported in the literature recently. It seems that this approach is receiving increasing attention. Here we mention the work of Schlangen and vanMier (1992) in modeling microcracking in cement base composites, the works of Murat et al (1992) and Monette et al (1992) in modeling the behavior of short fiber reinforced composites, and the work of Dai and Frantziskonis (1993) in modeling the statistical fracturing of cementitious composites and correlating it with ultrasonic non-destructive measurements.

In this study we utilize a triangular lattice. The properties of the unit cell are, for the case of linear, isotropic elasticity: Young's modulus equal to the modulus assigned to the bonds of the unit cell and a Poisson ratio that depends explicitly on the (constant) angular stiffness between bonds. In the absence of angular stiffness the Poisson ratio of the unit cell is equal to 1/3. For a thorough presentation of the lattice properties and different lattice types we refer to Hrennikoff (1941), Herrmann and Roux (1990) and Murat et al (1992). It is also possible, without extensive effort, to consider anisotropy within a unit lattice cell, non linear effects etc. Also, by assigning beam (bending stiffness) to the bonds micro-rotational (Cosserat) effects are recovered. However, in this study we consider the simplest possible case which calls for a triangular lattice without angular stiffness. Besides simplicity, the following advantages can be identified. Since interest is on micro-fracture at the unit cell level, using such a lattice there is only one choice to serve as bond-fracture criterion, namely the level of stress or the level of the corresponding strain at a bond. This is important since it is very difficult to identify (experimentally) the local conditions under which failure at the microlevel occurs. Since in a composite material properties

vary spatially (i.e. transition from matrix to interface to fiber) ambiguities related to the angular stiffness at the transition zones - that may render the problem non unique - are not present when using such a "central force" lattice. On the other hand, the Poisson ratio of 1/3 may not be precise. For the material combinations considered herein such a value is not unreasonable for the matrices considered. For the fiber material the problem of Poisson ratio determination is a difficult one and rigorous methods for its determination have not been established. Furthermore, it is very difficult to determine (experimentally) the local characteristics (i.e. Poisson effects) of interfaces/interface reaction zones.

In short, there are several issues to be resolved before an accurate representation of Poisson, of local anisotropy, of length scales present, and perhaps of local rotational (Cosserat) effects come into the picture. Thus we proceed in this study by considering the simplest possible case, the central force triangular lattice throughout the domain of interest.

The Interface

Figure 3 shows a 40x120 triangular lattice which is one of the lattices used for simulating the single fiber fragmentation test. The fiber is placed in the center of the lattice in figure 3, parallel to the y-direction and together with the interface is considered to be four lattice spacings wide, ninety four spacings long (4x94). The rest of the lattice is assigned matrix properties, and a single lattice spacing is assigned interface properties. From a first evaluation it may seem that the interface region is considered too "thick". This brings up the problem of interface thickness, and, as will be explained, there is a simple way to account for this in the analysis.

As far as terminology is concerned the term "transition region" (cf following discussion) may be more appropriate than interface. A transition region allows elastic deformation within the "interface" before fracture. However, both terms are used in the following, hoping that confusion is not possible.

During processing of a composite material, a "reaction zone" is formed, i.e. an interface that may impart bonding between the matrix and the fiber, Metcalfe (1974). Several works have examined the material in the vicinity of the fiber identifying significantly different properties than

the matrix. That region is often called mesophase, Theocaris (1987). Despite extensive work in this area, the behavior and properties of such a "mesophase", or "transition region" or "interface" or "interface reaction zone" have not been understood well.

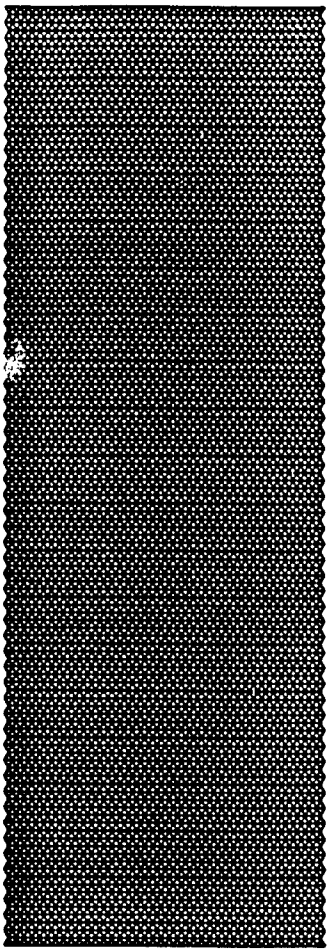


Figure 3: A 40x120 triangular lattice.

The SCS-6 (SiC) fiber is approximately 140 μm in diameter and contains 2 μm thick alternating outer layers of C and nonstoichiometric SiC that protect it from cracking during handling. In a composite made of titanium alloy matrix and SiC fiber with carbon coating a reaction zone consisting of several layers of Ti_xC_y and Ti_xSi_y is formed and is modulated by the alloying composition of the matrix and the processing procedure. Similar considerations hold for ceramic matrix composites. A thorough presentation on this subject can be found in Karpur et al (1993) and the references cited there, where also the concept of the "equivalent elastic interface" (EEI) is introduced. The concept is relevant to, and complements the, present study, thus, the ideas behind EEI, Karpur et al (1993), are briefly described here. Since the thickness of the transition region (interface) and the spatial variation of properties along it are difficult to quantify, it is advisable to consider an equivalent homogeneous transition region. By doing so we are allowed to consider apparently different material compositions, i.e. metal matrix, ceramic matrix, within the same framework.

It is difficult to specify the exact thickness of the (homogeneous) transition region, and the analysis would be sensitive to changes in thickness. In order to overcome this difficulty, the properties of the interface can be defined in a way that its response is independent of thickness. Thus interface properties should be defined in a way that delivery of information, the jump in

displacement and the transmitted stress across the interface, is consistently independent of thickness. This is accomplished precisely by dividing the relevant quantities, modulus of the homogenized region and failure stress, by the transition region thickness. By defining quantities such as the "stiffness coefficient" (modulus over thickness) and "failure stress coefficient" (failure stress over thickness) the need for precise specification of the thickness is overcome. Indeed, it is experimentally possible to measure such coefficient nondestructively, i.e. the "shear stiffness coefficient," Matikas and Karpur (1993).

In our lattice discretization, the smallest discretization scale (the lattice spacing) is assigned to the interface thickness. Thus the interface is considered homogeneous by definition. A stiffness coefficient S (elastic modulus over thickness) and a failure stress coefficient F (failure stress over modulus) are assigned to it, and its failure is considered brittle. As mentioned previously the (shear) stiffness coefficient can be evaluated non-destructively, Karpur et al (1993), Matikas and Karpur (1993). At this time such relevant experiments are being conducted (private communication with Karpur & Matikas) and correlations with the present study will be examined later. Also, such correlations are not straight-forward as discussed in the following, cf section titled "discussion". At this stage, the relevant interface properties will be deducted from the destructive experiments via back analysis and physical reasoning.

Simulation Results

The following have been assumed in the numerical simulation procedure. Inertia effects and body forces are neglected, and, load is applied slowly enough so that there is enough time for redistribution of stress before failure/yield proceeds further. The fiber, interface and ceramic matrix are considered brittle - when a bond fails its load is reduced to zero and the released load is redistributed by solving the problem again with the broken bond absent. The difference for ductile matrix bonds (metal matrix) is that after the yield stress has been reached the modulus is changed to the (linear) hardening modulus E_{mh}. Thus the simulation procedure involves the following steps: (a) discretize the structure into a lattice; (b) assign failure/yield stress to each bond, depending on whether it is spatially within matrix (brittle or ductile), fiber or interface; (c)

similarly to step (b) assign stiffness to each bond; (d) apply an increment of external displacement or load until the failure or yield criterion is satisfied by the bond carrying the maximum load - the problem being linear makes identification of that load easy; (e) if that bond is brittle release the load carried by it and solve the problem for the released load applied at the joints adjacent to it. If another bond fails during this process repeat step (e). If that bond is ductile apply the new modulus (E_{mf}) to it; (f) increment the externally imposed boundary condition until the next bond fails or yields and repeat the previous step. Allowing failure or yield of one bond at a time, together with the linearity of the problem during each step, assures a unique solution.

Simulation Results - Fragmentation Test

The following properties for the fiber and matrix are considered and assigned to the corresponding lattice bonds. For the SiC (SCS-6) fiber a Young's modulus E_f=393 GPa, and a failure stress σ_f=3.5 GPa. Fiber failure is perfectly brittle - when a bond fails its load carrying capacity is reduced to zero. The Ti-6Al-4V matrix is ductile, with a Young's modulus E_m=110 GPa and yield stress σ_m=0.83 GPa. For the linear hardening post-yield response the modulus is a fraction of E_m. Typically the strain hardening modulus E_{mh} = 1/100 E_m, however, as shown in the following the fiber fragmentation pattern depends on E_{mf}. The matrix is not allowed to fail - due to its ductility, failure occurs at large strains and the simulation is not carried out to such levels.

For identification of interface properties the following can be considered. We employ the experimental evidence that no fiber failure occurs in the linear regime of the specimen's stress-strain response, figure 1, but fiber failure initiates in the vicinity of the deviation from linearity. Having the matrix and fiber properties fixed, the interface properties have to be such that the predicted load displacement response and fragmentation pattern match the experimental results as close as possible. These impose important restrictions on the properties of the interface. Thus, together with the 9step-wise) linearity of the problem a few simulations can identify the range of interface properties. Before we identify such ranges, we present the conclusions from the

simulations.

(a) the fiber fragmentation pattern depends strongly on the volume of the matrix present. This is better understood if we consider the following. In the limit case of very small or negligible matrix volume, a single fiber failure will occur. As the amount of matrix surrounding the fiber increases there should be a threshold where multiple fiber fragmentation occurs. By increasing the amount of matrix further, the result is reduction in the average fragmentation length. The saturation limit, if it exists, depends on the properties of the interface. This important fact is rarely mentioned or addressed in the literature although it has been observed experimentally, Krishnamurthy (1993). Our analysis showed that the threshold for the present material combination is approximately at a ratio matrix over fiber approximately 10-12. The approximation is due to the following.

(b) the matrix hardening modulus E_{mh} influences the fragmentation pattern. The influence is not sensitive - only large changes in E_{mf} influence the fragmentation pattern, i.e. average length between fragments. For $E_{mf} = 0$ a single fiber fracture occurs, or even no fiber fracture at all, depending on the interface properties.

(c) both the interface modulus and strength, and their relative values influence the fragmentation pattern. A weak interface will fail even before the matrix yields, the result being that the fiber will simply act as an inclusion in the matrix. This also depends on the modulus. For example, if the interface and fiber are subjected to the same strain, the fiber will fail first only if the interface relative values (strength, modulus) allow so. Interface modulus has not a strong effect within 100% or so.

(d) the simulations showed evidence that it is practically impossible to achieve fragmentation lengths (average) of the order of or lower than the fiber diameter. Perhaps this provides goals as far as tailoring the interface properties is concerned. In passing, it is noted that the optimum interface properties depend on the geometric and loading conditions of a specific test and/or configuration. For example if fibers are close together without "enough" matrix material in between, the optimum interface properties are likely to be different than the ones implied by a single fiber fragmentation test. This important area seems to be totally unexplored. It is currently being addressed.

(e) immediately after the fiber breaks at some location, tension cracks along the interface propagate. In order that fragmentation occurs these cracks have to be arrested and the arresting

length (and thus the fragmentation pattern) are influenced strongly by the interface strength. This provides significant limitations to the range of interface strength.

(f) the load-displacement or stress-strain response of the specimen is practically insensitive to the interface properties (within orders of magnitude), thus the observed fragmentation pattern and interface failure are the ones providing information on interface properties.

Figure 4 shows the load-displacement (stress-strain) response obtained from 40x120

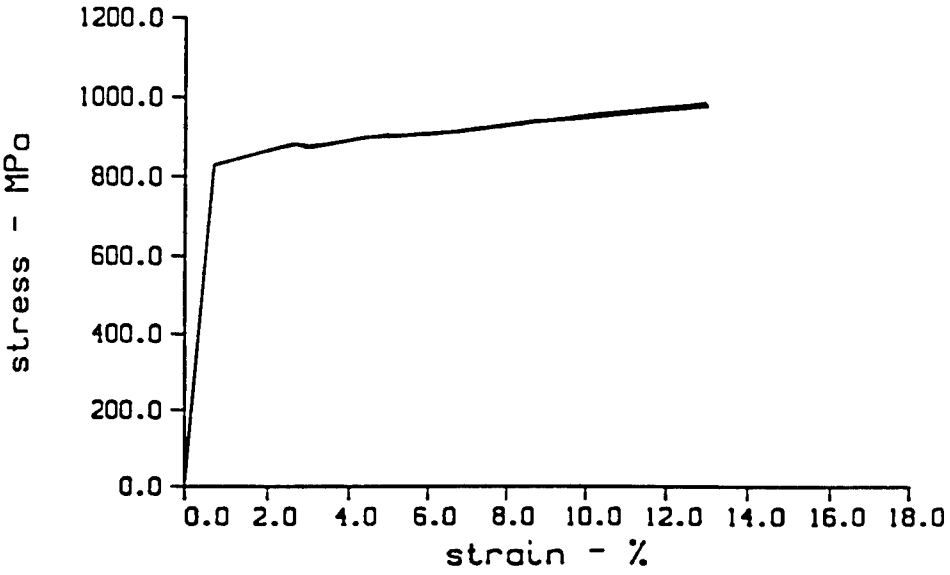


Figure 4: Two Stress Strain Responses Obtained with Identical Material Properties Except the Interface Strength Coefficient, S = 11 MPa/μm and S = 6 MPa/μm.

lattices. The material properties used for the matrix and fiber are the ones given above. For the interface the modulus coefficient used is 857 MPa/μm and the failure strength coefficient for one of them is 11 MPa/μm and for the second one 6 MPa/μm. The meaning of such interface properties is discussed later (section titled discussion). It is clear that the interface strength has practically no influence on the load-deformation response. Similar results hold for the interface modulus. The (small) stress drops in the curves correspond to fiber breaks. It is noted that experimentally, often, no load drop is observed. Since it is known that the fiber breaks at those levels of stress, this may be attributed to small snap-through, to rate of loading effects, or to the sensitivity of the equipment used. The predicted stress-strain response (figure 4) correlates well with the experimental one (figure 1b) where yield initiates at about 850 MPa, at a strain close

to the 2% level. Further it is noted that with these values for the interface properties most of the matrix bonds yield before fiber breaks are initiated.

The simulation is two-dimensional while the actual fragmentation test, figure 1, is three-dimensional! The diameter of the SiC fibers is approximately 0.14 mm. The samples, Roman et al 1993, were 1.50 mm thick with 19.05 mm x 6.35 mm gage sections. Thus the problem is not axisymmetric. It was mentioned above that the fragmentation pattern depends on the amount of matrix surrounding the fiber. In the three dimensional case it is not clear if the minimum matrix dimension (1.50 mm for the tests), or the area of the matrix (1.50 x 6.35 mm^2) or both are decisive with respect to the fragmentation pattern. For the test configuration, the ratio (matrix/fiber) with respect to the minimum dimension is approximately 1.5/0.14. Experimentally this ratio seems to govern the fragmentation pattern rather than the cross-sectional area ratio. Three-dimensional simulations are not attempted at this time, due to excessive computer time requirements. Further, several issues need to be understood before 3-D effects come into picture.

Although the two different values for interface strength coefficient yielded similar load deformation response, the fragmentation patterns that evolved were different. In the following figures, for effective presentation, a broken bar is represented by a thin, short line transverse to its original direction. Figure 5 shows some of the evolution stages of fiber fragmentation and interface failure for S = 11 MPa/μm, and figure 6 shows some of the stages for S = 6 MPa/μm. The following can be identified:

- Interface cracks at the fiber ends initiate and quickly become arrested. This is consistent with the analysis and experiments by Atkinson et al (1982) on the stability of interface cracks near the fiber end, for an embedded fiber.
- For both cases the fiber breaks first in the middle of its length. The reason why this happens is explained subsequently.
- After, or concurrently with, the first full fiber crack development (over its width) cracks propagate along the interface and at some point are arrested, figures 5b, 6b.
- The crack length along the interfaces is important with respect to subsequent fiber breaks, and thus with respect to the final fragmentation pattern. If figures 5 and 6 are compared, it is seen that the interface crack length is smaller in figure 5 than figure 6. This is the decisive reason for the final fragmentation pattern. In figure 5 the average fragmentation

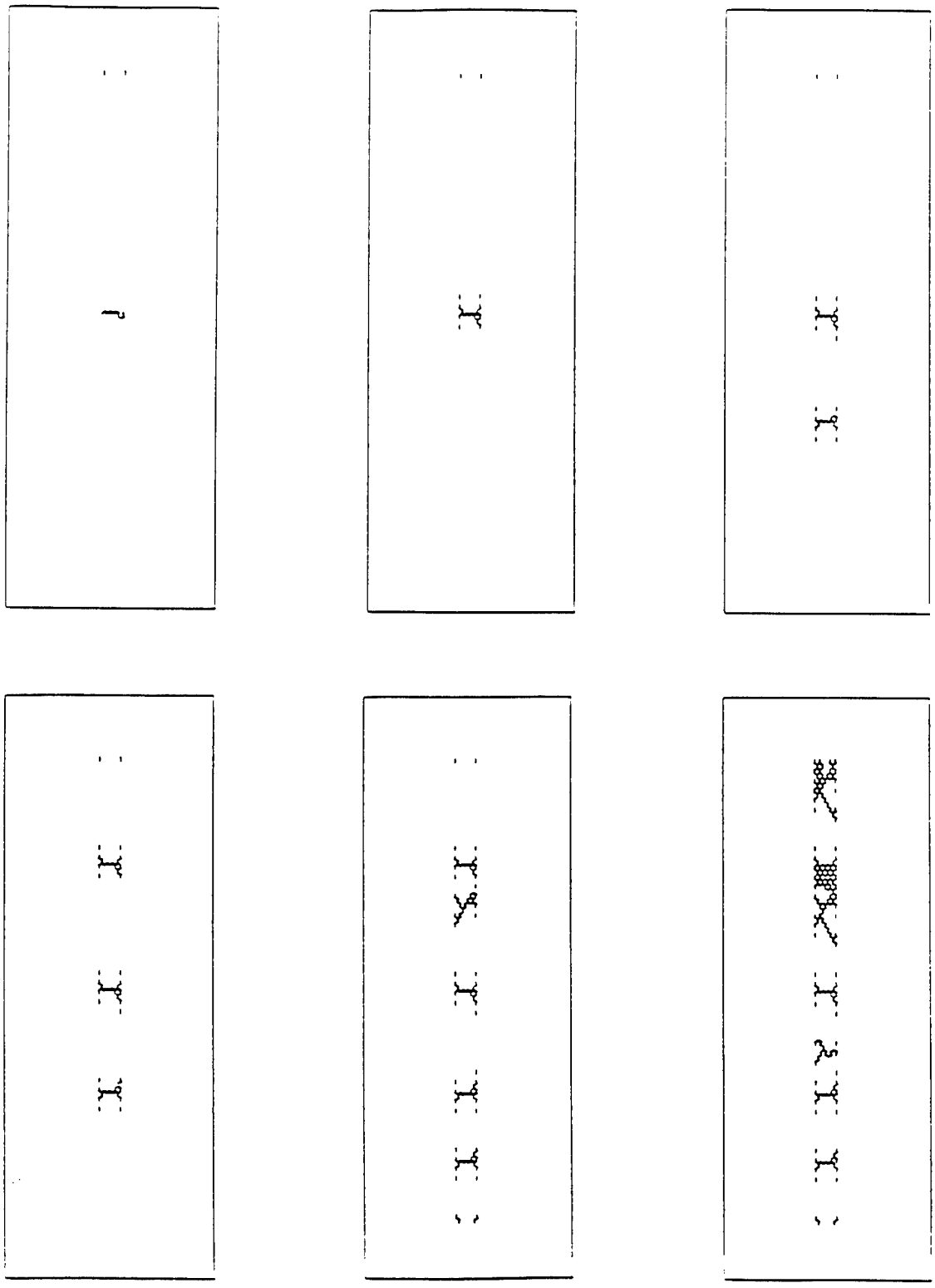


Figure 5: Fiber Fragmentation and Interface Failure Stages for Interface Strength Coefficient $S = 11\ MPa/\mu m$.

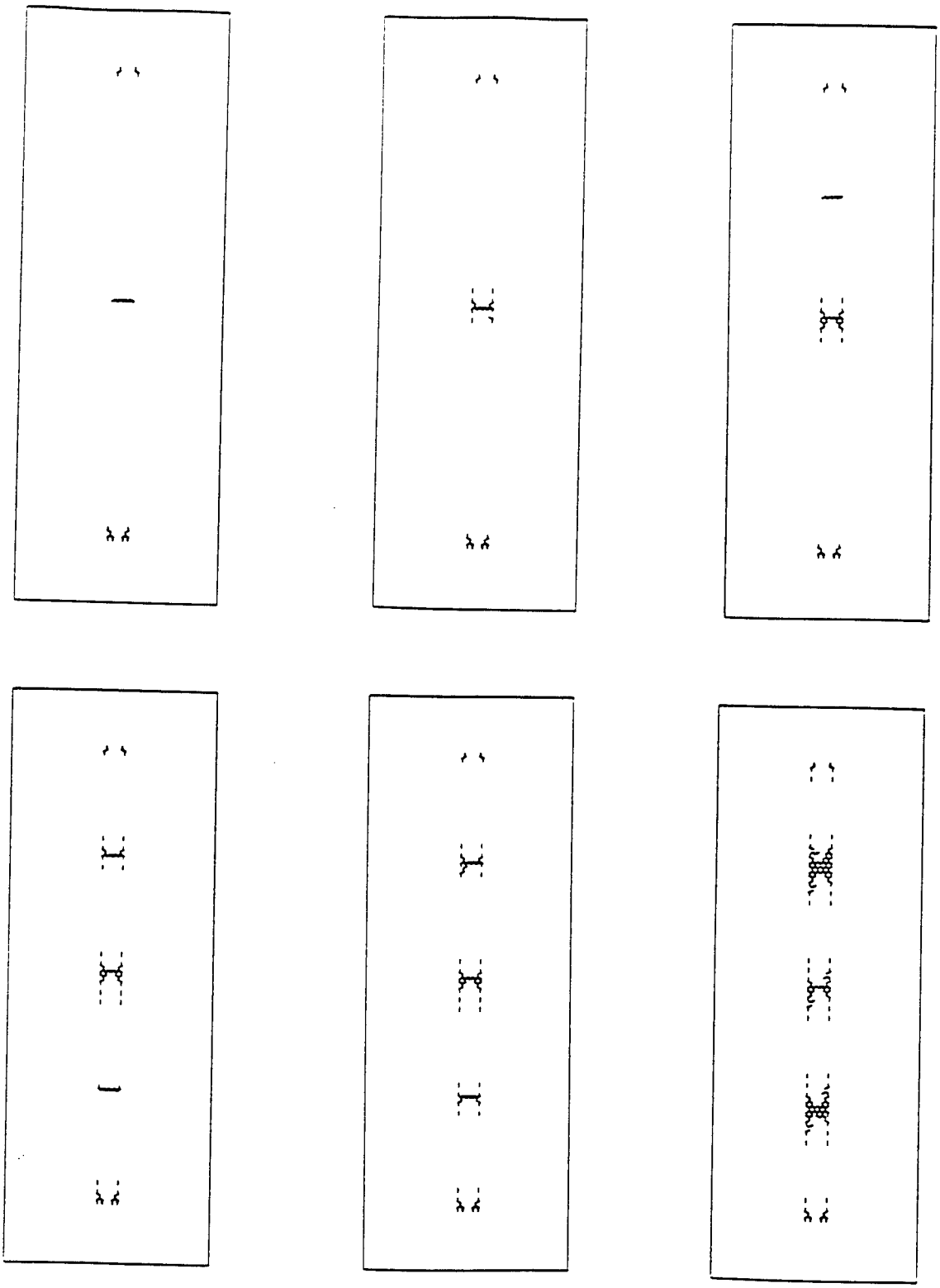


Figure 6: Fiber Fragmentation and Interface Failure Stages for Interface Strength Coefficient $S = 6\ MPa/\mu m$.

length is about half the one in figure 6.

- The differences between figures 5 and 6 can be explained by the length of interface cracks. In figure 6, for example a greater total length of interface is required to transmit enough load to the fiber - capable to break it. The interface shear strength is important here. Figure 7 shows the crack evolution when an interface shear strength coefficient S = 1 MPa/μm is considered. For strengths even lower than that, no fiber breakage is observed.

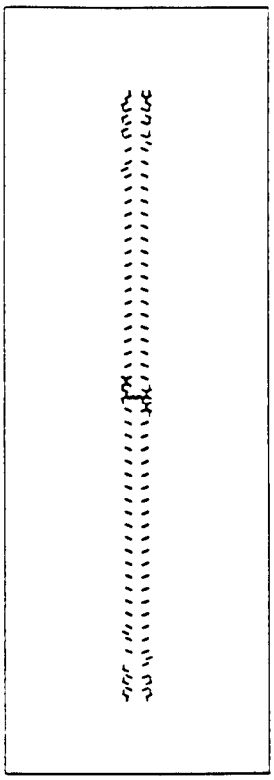


Figure 7: Crack Pattern for Interface Strength Coefficient S = 1 MPa/μm

So now the question why the first fiber break develops in its middle comes into picture. We find this opportunity to discuss further, at the same time, the problem of the ratio width of sample (or matrix) over width of fiber. Let us reduce the amount of matrix surrounding the fiber, say a lattice of 30x120 with the same dimensions for the fiber as before. For interface properties same as those that produced figure 5 a single fiber break will develop (no fragmentation) for this

shows such distribution. That figure is plotted as follows: Only bonds exceeding a certain level of strain are drawn. Clearly the stress concentration is in the middle. In order that the fiber breaks there enough load capable of doing so must be built up in the matrix and the interface has to be capable of transmitting the force to the fiber. For the present fiber, matrix and interface properties, the horizontal dimension of 30 units is approximately the limit that this will happen.

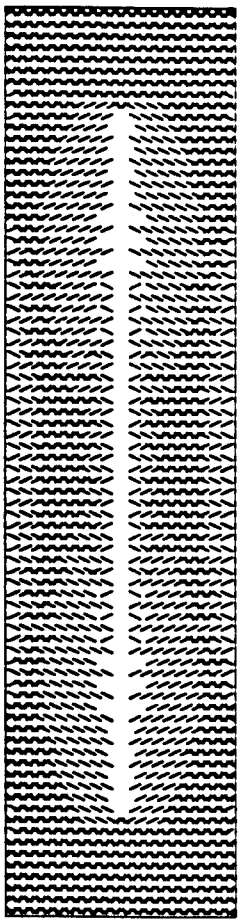


Figure 8: Strain Distribution Before Fiber Fracture.

The interface region is considered homogeneous. This, of course is an approximation of the reality. In general at the micro-level (at the length scale of a material's micro-structure) it may be argued that failure is predominantly in tension. Then, compressive failure is the integrated, phenomenological process of several tensile micro-failures. By considering the interface as being homogeneous, we indirectly imply/assume that the actual response is homogenizable. Then failure in compression is possible. For the single fiber fragmentation test, compressive failure proved not important.

REFERENCES

Atkinson C, Avila J, Betz E & Smelser RE, (1982), "The Rod Pull Out Problem, Theory and Experiment," *J. Mech. Phys. Solids*, 30, 97-120.

Charmet JC, Roux S & Guyon E, editors (1990), *Disorder and Fracture*, Plenum press.

Coker D, Ashbough NE & Nicholas T, (1991) in Proc. ASTM STP Proc. Symp. on Thermomechanical Fatigue Behavior of materials, San Diego, California.

Dai H & Frantziskonis G, (1993), "Heterogeneity, Spatial correlations, Size Effects and Dissipated Energy in Brittle Materials," *Mechs of Matls.*, preprint.

Eldridge JI, (1992), "Fiber Pushout Testing of Intermetallic Matrix Composites at Elevated Temperatures," Proc. Mat. Res. Soc. Symp., vol 273.

Evans AG, Zok FW & Davis J, (1991), "The role of Interfaces in Fiber-Reinforced Brittle Matrix Composites," *Comp. Sci. Tech.*, 42, 3-24.

Herrmann HJ & Roux S, editors, (1990), *Statistical Models for the Fracture of Disordered Media*, North-Holland.

Hrennikoff A, (1941), "Solution of Problems of Elasticity by the Framework Method," *J. Appl. Mechs. ASME*, A169-A175.

Karpur P, Matikas T & Krishnamurthy S, (1993), "A Novel Parameter to Characterize the Fiber-matrix Interphase/Interface for Mechanics of Continuous Fiber Reinforced Metal Matrix and Ceramic Matrix Composites," *Comp. Sci. & Tech.*, under review.

Kerans RJ & Parthasarathy TA, (1991), "Theoretical Analysis of the Fiber Pullout and Pushout Tests," *J. Amer. Cer. Soc.*, 74, 1585-1596.

Krishnamurthy S (1993) private communication.

Majumdar BS, Newaz GM & Ellis JR, (1993), "Evolution of Damage and Plasticity in Titanium-Based, Fiber-Reinforced Composites," *Mettal. Trans A*, 24A, 1993-1597.

McCartney LN, (1990) "New Theoretical Model of Stress Transfer Between Fibre and Matrix in a Uniaxially Fibre-Reinforced Composite," *Proc. Roy. Soc. London*, A425-442.

Matikas T & Karpur P, (1993) "Ultrasonic Reflectivity Technique for the Characterization of Fiber-Matrix Interface in Metal Matrix Composites," *J. Appl. Phys.*, to appear.

Metcalfe AG, (1974) *Interfaces in Metal Matrix Composites*, Academic Press, New York.

Monette L, Anderson MP, Ling S & Grest GS, (1992), "Effect of Modulus and Cohesive Energy on Critical Fibre Length in Fibre-reinforced Composites," *J. Mater. Sci.*, 27, 4393-4405.

Murat M, Anholt M & Wagner HD, (1992) "Fracture Behavior of Short-Fiber Reinforced Materials," *J. Mater. Res.*, 7, 3120-3131.

Ochiai S & Osamura K, (1986a), "Stress Distribution of a Segmented Fibre in Loaded Single Fibre - Metal Matrix Composites, *Z. Metallkde*, 77, 249-254.

Ochiai S & Osamura K, (1986b), "Multiple Fracture of a Fibre in a Single Tungsten Fibre - Copper Matrix Composite, *Z. Metallkde*, 77, 249-254.

Pagano, NJ, (1991), "Axisymmetric Micromechanical Stress Fields in Composites," in Local Mechanics Concepts for Comp. Matl. Systems, Reddy JN & Reifsnider KL edts.

Roman I, Krishnamurthy S, & Miracle DB, (1993), "Fiber-matrix Interfacial Behavior in SiC-Titanium Alloy Composites,"

Schlangen E, & Van Mier JGM, (1992) "Simple Lattice Model for Numerical Simulation of Fracture of Concrete Materials and Structures," *Mater. & Struct.*, 25, 534-542.

Theocaris PS, (1987), *The Mesophase Concept in Composites*, Springer-Verlag, New York.

Watson MC & Clyne TW (1992), "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. Mat.*, 40, 131-139, II. Application of the Test to Composite Materials, 40, 141-148.

THE AMENDMENT, SEE PAGE ONE, CONTAINS THE RESULTS FOR FIBER PUSHOUT, PULLOUT, AS WELL AS THE MATERIAL PRESENTED HEREIN

THERMOMECHANICAL FATIGUE OF TITANIUM MATRIX COMPOSITES

Zhanjun Gao
Department of Mechanical & Aeronautical Engineering
Clarkson University
Potsdam, NY 13699

and

Brian P. Sanders
Wright Laboratory Materials Directorate
WL/MLLN Bldg. 655
2230 Tenth Street, Suite 1
Wright-Patterson AFB, OH 45433-7817

Final Report for:
Summer Research Program
Wright Laboratory

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

August 1993

THERMOMECHANICAL FATIGUE OF TITANIUM MATRIX COMPOSITES

Zhanjun Gao
Clarkson University

Brian P. Sanders
Wright Laboratory Materials Directorate

ABSTRACT

The objective of the research is to develop a mechanistic model to study the response and damage development of metal matrix composite laminates under a general thermomechanical fatigue loading. The research consists of two parts. The first part deals with fatigue life prediction of SCS_6/Ti-15-3 metal matrix composite laminates. Damage accumulation and property degradation of each lamina in the laminate are studied. Internal load re-distribution among all the laminae is evaluated from the model of stiffness reduction. It is shown that within the first few cycles, stiffness reduction saturates to a characteristic level which is independent of cyclic and static loading history. As far as the fatigue life prediction is concerned, it is not important to know how the damage is accumulated, or what the stress-strain response is during the first few cycles. The fatigue life is determined by the saturated stiffness and the applied stress amplitude. The predictions of fatigue life agree well with experimental data of $[0/90]_{2s}$, $[0_2/\pm 45]_s$ and $[0/\pm 45/90]_s$ (SCS_6/Ti-15-3) laminates. The data needed to predict the life of these laminates under variable amplitude fatigue are those of its laminae. Therefore, it is possible to establish a data base for different unidirectional laminae and predict the fatigue life of a laminate based on the data of its laminae and the laminate stacking sequence.

The second part of the research deals with micromechanical analysis of local damage and viscoplastic behavior of titanium matrix composites under a thermomechanical fatigue loading. The present micromechanical analysis is based on Eshelby's solution of an ellipsoidal inclusion [2.1] and Mori-Tanaka's concept of mean field theory [2.2]. The model is applicable to study deformation and stress of a composite under general nonisothermal in-phase and out-phase fatigue loading. Interfacial damage and debonding are represented in the model. Such an analysis make it possible to determine the possible failure modes (matrix failure or fiber failure or combined) under different thermomechanical fatigue loading. The predicted creep behavior of unidirectional composite under a load in fiber direction agrees well with experimental data.

THERMOMECHANICAL FATIGUE OF TITANIUM MATRIX COMPOSITES

Zhanjun Gao and Brian P. Sanders

PART 1: FATIGUE LIFE PREDICTION

Local damage and its evolution are very critical in determining the fatigue life of metal matrix composites. A great deal of effort has been made in studying the local damage such as interfacial failure, matrix yielding and fiber fracture and their implications on fatigue life [1.1, 1.2]. However, such an approach is fairly complicated, and determination of damage is possible only within the first few number of loading cycles. This research takes a novel approach to fatigue life prediction of a laminate. Instead of determining damage events and their evolutions, we focus on the consequence of the damage on stress redistribution among 0 degree plies and off-axis plies. The 0 degree plies are critical plies whose failure will lead to final rupture of the laminate. Therefore, the fatigue life of the laminate can be determined if the load share of the 0 degree plies is evaluated.

Figure 1.1 shows a typical stiffness reduction for a $[0/90]_{2s}$ lay-ups of SCS_6/Ti-15-3 laminate at room temperature [1.1]. The elastic modulus dropped about 22% during the first few cycles, then remained almost constant until just before the final failure. This indicates that the load re-distribution for such a laminate occurs only during the first few loading cycles. After the first few cycles, the damage in the laminate reaches a stable level, called the characteristic damage state (CDS) (Reifsnider, etc. [1.2]). Our experience with different composite materials systems indicates that the CDS is completely defined by the properties of the fiber, matrix, interface and fiber orientations, etc. and is independent of loading history. Figure 1.2 shows the spacing between cracks in the -45 degree plies of a $[0/90/\pm45]_s$ AS-3501-5 graphite epoxy laminate as a function of quasi-static load level and cycles of loading [1.2]. Cracks develop quite early in the life and quickly stabilize to a very nearly constant level with a fixed spacing for both quasi-static and cyclic loading. The crack patterns for two types of loading are essentially identical, regardless of load history. A schematic representation of damage in terms of stiffness reduction is shown in Fig. 1.3. If the stress amplitude is high enough, the reduced axial stiffness at CDS level can be reached by one cycle. For a lower stress amplitude, it takes more loading cycles to reach CDS. The important fact is that the saturated axial stiffness of the laminate is independent of loading history. As far as the fatigue life prediction is concerned, it is not important to know how the damage is accumulated, and what the stress-strain response is during the first few cycles before the CDS.

The fatigue life is determined by the stiffness at CDS and the applied stress amplitude as to be seen later.

Prediction of stiffness at CDS

Since the stiffness at the CDS level, E_{xc} is a characteristic material property of the laminate, and independent of loading history, it is to our advantage to determine E_{xc} under a static loading condition. Our attention is focused on cross-ply laminate first. Under a static loading, as soon as the loading increases up to a level, σ_A, at which the transverse stress in the 90 degree plies equals to the transverse strength Y of the unidirectional laminate, damage begins to initiate in 90 degree plies. The transverse modulus, E_2, begins to degrade. The shear modulus G_{12} has no effect on axial stiffness change of the laminate.

After damage initiates, the stress σ_{22} in 90 degree plies, denoted as $\sigma_{22}(90^o)$, is not completely relaxed. Instead, it is necessary that this stress remains at a level equal to transverse strength Y so that additional damage can be accumulated up to the CDS level. Thus the reduced modulus E_2 in 90 degree plies can be approximately obtained by solving the equation

$$\sigma_{22}(90^o) = Y. \tag{1.1}$$

The algorithm is as follows:

(1) For an applied load σ_x larger than σ_A, let all the elastic parameters in 0 degree plies be the initial values. E_2 in 90 degree plies is reduced by one step length E_2. Stress distributions in each ply are obtained through laminate analysis.

(2) If the stress $\sigma_{22}(90^o)$ is larger than Y, repeat the first step until the Eqn. (1) is satisfied.

(3) Use the final reduced value of E_2 for 90 degree plies to obtain the axial stiffness of the laminate.

Figure 1.4 shows the stiffness reduction analysis using the above procedure for a SCS_6/Ti-15-3 laminate with $[0/90]_{2s}$ lay-ups. The reduction of transverse stiffness E_2 of the 90 degree plies results in a loss of laminate axial stiffness E_x, as shown in the figure. When the load approaches to the ultimate tensile strength (UTS) level, stress-strength ratio σ_{11}/X of 0 degree plies (H_0 in the figure) approaches to 1, indicating the 0 degree plies are about to fracture. The axial stiffness E_x also approaches a stable value. Therefore, the characteristic stiffness E_{xc} can be approximately obtained as the stiffness value corresponding to applied load equal to the ultimate tensile strength, which is 944 MPa for the laminate [1.1]. The predicted value of E_{xc} is equal to

78% of the initial axial stiffness, which agrees with experimental data shown in Fig. 1.1.

For $[0_2/\pm45]_s$ laminates, the transverse modulus E_2 of 45 and -45 degree plies has little influence on the shear stress of these plies and axial stiffness of the laminate. This implies the stress transfer from the off-axis plies to the 0 degree plies is caused by a reduction of shear modulus G_{12} of the off-axis plies. The equation corresponds to Eqn. (1.1) to determine the reduction of G_{12} is changed to

$$\tau_{12}(45^o) = S \qquad (1.1')$$

where $\tau_{12}(45^o)$ is the shear stress of 45 or -45 degree plies, S is the shear strength of the lamina. The procedure to determine the characteristic stiffness E_{xc} due to reduction of G_{12} is similar to that used for $[0/90]_{2s}$.

For $[0/90/\pm45]_s$ laminate, since there are two kinds of off-axis plies, more stress components need to be considered, which makes the problem more complicated. Usually, the damage in the 90 degree plies occurs earlier than that in 45 and -45 degree plies. In fact, 90 degree plies often delaminate from the rest of plies before the final failure of laminate, which implies there is no significant sharing of load by 90 degree plies after the CDS. Therefore, if only the characteristic damage is concern, all the elastic parameters in 90 degree plies can be set zero. The problem then is reduced to one in a similar case as in $[0_2/\pm45]_s$ laminate except there is a nonzero value of thickness of 90 degree plies for the laminate analysis here.

Fatigue life of laminates containing 0 degree plies

As discussed earlier, damage in a laminate can be characterized by the axial stiffness reduction. The damage accumulates with the increase of number of cycles resulting stress re-distribution up to a saturated state called the characteristic damage state (CDS). When the CDS level in off-axis plies is reached, a stabilized axial stiffness is reached, and no stress redistribution takes place between 0 degree plies and off-axis plies. Since the 0 degree plies are critical plies whose failure results in final rupture of the laminate, the fatigue life of the laminate is determined from the life of the 0 degree plies. Of course, the stress redistribution among 0 degree plies and other off axis plies has to be represented correctly through stiffness reduction analysis, as previously explained.

The fatigue life for a laminate containing 0 degree plies contains two part, CDS life(the number of cycles necessary to reach the CDS) and the fatigue life to failure of the 0 degree plies. For as-fabricated SCS_6/Ti-15-3 laminate, the

axial stiffness stabilized after a few cycles reported by Johnson etc. [1.1]. This indicates that the CDS life is quite small and can be neglected.

The S-N curve of 0 degree laminate can be represented as

$$\text{Log } \sigma_{11}(0^{o}) = a\ (\text{Log N})^2 + b\ \ (\text{Log N}) + \text{Log X} \qquad (1.2)$$

where $\sigma_{11}(0^{o})$ is the cyclic load amplitude of 0 degree laminate, N the number of cycles to failure, X its ultimate tensile strength in fiber direction, and a and b are material constants determined from experimental tests. The values of these constants for SCS_6/Ti-15-3 0 degree laminate are: X=1518 MPa, a=-0.0496, b=0.1548.

When the CDS level is reached, the axial stiffness degrades to its characteristic value, E_{xc}, which can be obtained by the procedure described in the previous section. The cyclic stress amplitude of the 0 degree plies is determined as

$$\sigma_{11}(0^{o}) = \sigma_x \frac{E_1}{E_{xc}} \qquad (1.3)$$

where E_1 is the initial axial stiffness of the 0 degree plies, and σ_x is the stress amplitude of the global applied load on the laminate. Combining Eqns. (1.2) and (1.3), the fatigue life, N, of the laminate is determined from the following equation

$$\text{Log } (\sigma_x \frac{E_1}{E_{xc}}) = a\ (\text{Log N})^2 + b\ \ (\text{Log N}) + \text{Log X}. \qquad (1.4)$$

The proposed model was used to predict the stiffness reduction and fatigue life of as-fabricated SCS_6/Ti-15-3 laminates. The characteristic stiffness E_{xc} for each laminate was obtained, and then Eqn. (1.4) was used to evaluate the fatigue life, N, of the laminate.

The predicted S-N curves for $[0/90]_{2s}$, $[0_2/\pm45]_s$ and $[0/\pm45/90]_s$ laminates agree very well with the experimental data by Johnson etc. [1.1] as shown in Figs. 1.5, 1.6 and 1.7. Overall, the results of this paper are consistent with the experimental observation of Johnson etc. [1.1] that the stress in the 0 degree fiber could be used to correlate the fatigue life of different laminates containing 0 degree plies.

The predicted values of E_{xc} for $[0/90]_{2s}$, $[0_2/\pm45]_s$ and $[0/\pm45/90]_s$ laminates are 141.1 GPa, 98.4 GPa and 98.4 GPa, respectively. The predicted life depends on the characteristic stiffness E_{xc}. A better prediction of the S-N curve is obtained for the cross-ply laminate. This may be attributed to the

accurate prediction of E_{xc} compared with the experimental result [1.1]. As for the other laminates, $[0_2/\pm 45]_s$ and $[0/\pm 45/90]_s$, no experimental data about E_{xc} are available. In general, the predicted life for these two kinds of laminate by the proposed model tends to be a little conservative. That may result from the smaller predicted value of E_{xc}. Since the reduced axial stiffness is obtained under ultimate tensile strength not the CDS load, it is over-reduced to a smaller value. The smaller is E_{xc} the larger is the equivalent cyclic stress, which causes a smaller fatigue life.

PART 2: LOCAL DAMAGE GROWTH AND GLOBAL VISCOPLASTIC BEHAVIOR

This part of the research deals with micromechanical analysis of local damage and viscoplastic behavior of titanium matrix composites under a thermomechanical fatigue loading. The model is applicable to study deformation and stress of a composite under general nonisothermal in phase and out phase fatigue loading. Interfacial damage and debonding are represented in the model.

1. Significance of the approach

(1). A three-dimensional nondilute stress solution which takes into consideration the contribution of randomly distributed fibers is obtained.

(2). The stress-strain relation includes creep and plastic deformation.

(3). Other type of constitutive models, such as the unified constitutive model of Bodner [2.10] can be implemented without difficulties.

(4). The stress analysis can be extended to include an interphase (a layer with finite thickness between fiber and matrix materials). The effects of the interphase and interphasial failure on thermomechanical behavior of metal matrix composites can be investigated.

(5). Deformation and stress of a composite under general nonisothermal in phase and out phase fatigue loading can be studied. Such an analysis will make it possible to determine the possible failure modes (matrix failure or fiber failure or combined) under different thermomechanical fatigue loading.

(6). The proposed approach utilizes Eshelby's elasticity solution of single inclusion [2.1] and Mori-Tanaka's concept of mean field theory [2.2], which greatly reduces the complexity of the analysis involved. While maintaining a relative high accuracy, the three-dimensional stress is determined in a closed form at any given time step.

(7). It is conceivable to use a similar approach to calculate the local stress (point-wise stress instead of mean stress) in the ductile matrix and establish a constitutive relation for each point of the matrix material.

However, it is expected that under a loading in the fiber direction, the mean value approach will provide a result with a good accuracy, and

(8). The proposed approach is applicable to particle reinforced composites, short fiber composites, and continuous fiber composites.

2. Dilute solution: single fiber in an infinite matrix

A model of composites is shown in Fig. 2.1, where identical ellipsoidal inhomogeneities(or fibers, these two terms are used alternatively in this paper) are aligned along the same direction. A traction force is prescribed on the boundary of the material. The elastic stiffness of the matrix and inhomogeneities are C^m and C^f, respectively. Here C^m and C^f are 6 by 6 stiffness matrix.

We will first consider the situation of a single inclusion in an infinite matrix as shown in Fig. 2.2. The matrix material D, contains an ellipsoidal inhomogeneous inclusion Ω. The traction force prescribed on the boundary of the material produces a uniform stress σ^0 and strain ε^0 when the material does not contain any inhomogeneities, i.e.

$$\sigma^0 = C^m \varepsilon^0,$$

where $\sigma^0 = [\sigma^0_{11}, \sigma^0_{22}, \sigma^0_{33}, \sigma^0_{12}, \sigma^0_{13}, \sigma^0_{23}]^T$, $\varepsilon^0 = [\varepsilon^0_{11}, \varepsilon^0_{22}, \varepsilon^0_{33}, \varepsilon^0_{12}, \varepsilon^0_{13}, \varepsilon^0_{23}]^T$, "T" denotes transport of a vector. Throughout this paper all other stress and strain vectors are defined in a similar fashion.

Due to the existence of the inhomogeneities and nonelastic strain ε^{ne} in Ω, the stress σ^0 and strain ε^0 have perturbations, $\tilde{\sigma}$ and $\tilde{\varepsilon}$, respectively, from their original values. In the following, superscripts f and m denote quantities of fibers (or inhomogeneities) and matrix, respectively, The averages of these quantities over their defined region are expressed as <.>. For instance, $<\sigma^f>$ is the average fiber stress, while $<\sigma^m>$ denotes the average matrix stress.

From Hooke's law

$$\begin{aligned} \sigma^0 + \tilde{\sigma} &= C^f(\varepsilon^0 + \tilde{\varepsilon} - \varepsilon^{ne}) && \text{in } \Omega, \\ \sigma^0 + \tilde{\sigma} &= C^m(\varepsilon^0 + \tilde{\varepsilon}) && \text{in D-}\Omega, \end{aligned} \tag{2.1}$$

where

$$\sigma^0 = C^m \varepsilon^0 \qquad \text{in D.} \tag{2.2}$$

The inhomogeneous inclusion is simulated by an inclusion in the homogeneous material with nonelastic strain ε^{ne} plus an equivalent eigenstrain ε^*,

$$\sigma^0 + \tilde{\sigma} = C^m(\varepsilon^0 + \tilde{\varepsilon} - \varepsilon^{ne} - \varepsilon'') \qquad \text{in } \Omega,$$
$$\sigma^0 + \tilde{\sigma} = C^m(\varepsilon^0 + \tilde{\varepsilon}) \qquad \text{in D-}\Omega, \tag{2.3}$$

The eigenstrain, ε'' is a fictitious nonelastic strain introduced to simulate the stress disturbance due to the existence of fiber and ε^{ne}. The equivalency between (2.1) and (2.3) holds when ε'' is chosen in a way such that the following equation is satisfied

$$C^f(\varepsilon^0 + \tilde{\varepsilon} - \varepsilon^{ne}) = C^m(\varepsilon^0 + \tilde{\varepsilon} - \varepsilon^{ne} - \varepsilon'') \qquad \text{in } \Omega. \tag{2.4}$$

Using the Eshelby's solution of homogeneous inclusion problem with a uniform nonelastic strain $\varepsilon^{ne} + \varepsilon''$, the strain $\tilde{\varepsilon}$ is found as

$$\tilde{\varepsilon} = S(\varepsilon^{ne} + \varepsilon'') \tag{2.5}$$

where **S** is the Eshelby's tensor given in reference[2.1].

When Eqn. (2.5) is substitute into Eqn. (2.4), the following relation is obtained to determine the eigenstrain ε''

$$C^f[\varepsilon^0 + S(\varepsilon^{ne} + \varepsilon'') - \varepsilon^{ne}] = C^m[\varepsilon^0 + S(\varepsilon^{ne} + \varepsilon'') - \varepsilon^{ne} - \varepsilon''] \qquad \text{in } \Omega. \tag{2.6}$$

The eigenstrain ε'' is solved from Eqn. (2.6) as

$$\varepsilon'' = [C^m(I - S) + C^f S]^{-1} (C^m - C^f)[\varepsilon^0 + (S - I)\varepsilon^{ne}], \tag{2.7}$$

where "-1" denotes the inverse of a matrix, and I is a 6 by 6 unit matrix. The stress in fiber is then given as

$$\begin{aligned} \langle\sigma^f\rangle &= \sigma^0 + \tilde{\sigma} = C^f[\varepsilon^0 + S(\varepsilon^{ne} + \varepsilon'') - \varepsilon^{ne}] = C^f[\varepsilon^0 + (S - I)\varepsilon^{ne} + S\varepsilon''] \\ &= C^f\{\varepsilon^0 + (S - I)\varepsilon^{ne} + S[C^m(I - S) + C^f S]^{-1} (C^m - C^f)[\varepsilon^0 + (S - I)\varepsilon^{ne}]\} \\ &= C^f[(C^m)^{-1}\sigma^0 + (S - I)\varepsilon^{ne}] + H[(C^m)^{-1}\sigma^0 + (S - I)\varepsilon^{ne}] \end{aligned} \tag{2.8}$$

where $H = C^f S[C^m(I - S) + C^f S]^{-1} (C^m - C^f)$

3. Nondilute solution: Mori-Tanaka method

When the volume fraction of fibers is small, Eshelby's solution [2.1] for the single inclusion (inhomogeneity) in an infinite medium can be used to represent the stress and strain fields in the inhomogeneity and matrix region. However, the interactions between the inhomogeneities have to be taken into account when inhomogeneities are close to each other such as the case of composite materials.

The concept of an average field [2.2] in inhomogeneities and their surrounding matrix has been introduced in an attempt to include the interaction between the inhomogeneities. The method involves rather complex manipulations of the field variables along with special concepts of equivalent inclusions. Slightly different derivations of the method have appeared in the literature [2.3-2.7]. Benveniste [2.8] has provided a valuable contribution to the Mori-Tanaka's method by giving a much more direct and simplified derivation of the method. After some mathematical description and careful scrutiny of the existing works dealing with equivalent inclusion and the average stress method, he revealed that Mori-Tanaka's method leads to the nondilute stress expression by solving the single inclusion problem with a single fiber in an infinite matrix, but by replacing the applied load by the average stress of the matrix in the composite, i.e., $\langle\sigma^m\rangle$. A schematic demonstration is shown in Fig. 2.2. Therefore, the non-dilute solution is obtained as long as the single inclusion problem is solved, and the constant applied strain term ε^0 is replaced by the unknown quantity $\langle\varepsilon^m\rangle$.

The nondilute fiber stress is obtained as

$$\langle\sigma^f\rangle = \mathbf{C}^f[(\mathbf{C}^m)^{-1}\langle\sigma^m\rangle + (\mathbf{S} - \mathbf{I})\varepsilon^{ne}] + \mathbf{H}\,[(\mathbf{C}^m)^{-1}\langle\sigma^m\rangle + (\mathbf{S} - \mathbf{I})\varepsilon^{ne}] \tag{2.9}$$

Note that

$$v_f\langle\sigma^f\rangle + (1-v_f)\langle\sigma^m\rangle = \sigma^0 \tag{2.10}$$

where v_f is the composite fiber volume fraction. Equations (2.9) and (2.10) are used to determine the average stress in matrix, $\langle\sigma^m\rangle$.

$$\langle\sigma^m\rangle = [(1-v_f)\,\mathbf{I} + v_f(\mathbf{C}^f + \mathbf{H})\,(\mathbf{C}^m)^{-1}]^{-1}\,[\sigma^0 + v_f(\mathbf{H} + \mathbf{C}^f)(\mathbf{I} - \mathbf{S})\varepsilon^{ne}] \tag{2.11}$$

For a general thermomechanical loading, we have

$$\begin{aligned}\sigma^0(t) &= \sigma^1 \sin \omega t + \sigma^2 \\ \varepsilon^{ne}(t) &= -\varepsilon^c - \varepsilon^p - (\alpha_m - \alpha_f)\,T(t)\mathbf{I}_1 \\ T(t) &= T^1 \sin(\omega t + \varphi) + T^2\end{aligned} \tag{2.12}$$

where $\mathbf{I}_1 = [1, 1, 1, 0, 0, 0]^T$, ε^c and ε^p are creep strain and plastic strain, respectively, α_m and α_f are the coefficients of thermal expansion for the matrix and fibers, T(t) is the difference between current temperature and the stress free temperature. φ is the phase difference between thermal and mechanical loading.

4. Constitutive equations

This section deals with the constitutive relationships of the matrix material. Stress and strain to be used should be understood as the average stress and strain of the matrix material, although the superscript "m" is omitted.

A one dimensional creep response is given by Dorn's law [2.9]

$$\dot{\varepsilon}^c = A\{\frac{|\sigma|}{G}\}^n \frac{Gb}{kT} D_0 \exp\{-Q/(RT)\} \tag{2.13}$$

where $\dot{\varepsilon}^c$ and σ are the creep rate and stress of the material, A and n are constants that can be determined experimentally, G is the shear modulus, b is the Burger's vector, k is Boltzmann's constant, and D_0 and Q are pre-exponential constant and activation energy for self-diffusion.

For two-dimensional and three-dimensional problems, the Prandtl-Reuss relations can be used for computing the creep increments. Thus we assume an equivalent stress defined the same way as in plasticity theory and an equivalent creep strain increment and write

$$\dot{\varepsilon}^{c*} = A\{\frac{\sigma^*}{G}\}^n \frac{Gb}{kT} D_0 \exp\{-Q/(RT)\} \tag{2.14}$$

where $\dot{\varepsilon}^{c*}$ and σ^* are the von Mises' effective strain rate and effective stress (of matrix) defined as

$$\sigma^* = \sqrt{(\frac{3}{2}\sigma'_{ij}\sigma'_{ij})} \tag{2.15}$$

and

$$\dot{\varepsilon}^{c*} = \sqrt{(\frac{2}{3}\dot{\varepsilon}^c_{ij}\dot{\varepsilon}^c_{ij})} \tag{2.16}$$

respectively. σ'_{ij} is the deviatoric stress (of matrix). The creep rate components are taken to follow the Prandtl-Reuss relation

$$\dot{\varepsilon}^c_{ij} = \frac{3}{2}\frac{\dot{\varepsilon}^{c*}}{\sigma^*}\sigma_{ij} \tag{2.17}$$

Plastic deformation is treated in a similar manner. Plastic flow rules are given by

$$d\varepsilon_{ij}^{p} = \frac{3}{2}\frac{d\varepsilon^{p*}}{\sigma^{*}}\sigma_{ij}^{'} \tag{2.18}$$

and the experimentally determined instantaneous stress-strain curve

$$\varepsilon^{p*} = f(\sigma^{*}) \tag{2.19}$$

where $\dot{\varepsilon}^{p*}$ is the von Mises' effective plastic strain defined as

$$d\varepsilon^{p*} = \sqrt{(\frac{2}{3}d\varepsilon_{ij}^{p}\, d\varepsilon_{ij}^{p})}\cdot \tag{2.20}$$

Equation (2.18) is valid only if the body is being loaded and has yielded, for example, if

$$\sigma^{*} > \sigma_{y} \tag{2.21}$$

where σ_{y} is the yield stress in tension. This is the so called von Mises yield condition for multiaxial stress states. If the body is unloading or if $\sigma^{*} < \sigma_{y}$, then no plastic strain occurs.

5. Procedure for solution

Equations (2.11), (2.14) - (2.17) are written in an incremental form

$$<\sigma^{m}> = [(1-v_f)\, I + v_f(C^{f} + H)\,(C^{m})^{-1}]^{-1}\,[\sigma^{0}(t) + v_f(H + C^{f})(I - S)$$
$$(-\varepsilon^{c} - \varepsilon^{p} - \Delta\varepsilon^{c} - \Delta\varepsilon^{p} - (\alpha_{m} - \alpha_{f})\, T(t)I_{1})] \tag{2.11'}$$

$$\Delta\varepsilon^{c*} = A\{\frac{\sigma^{*}}{G}\}^{n}\frac{Gb}{kT}D_{0}\,\mathbf{exp}\{-Q/(RT)\}\Delta t \tag{2.14'}$$

$$\sigma^{*} = \sqrt{(\frac{3}{2}\sigma_{ij}^{'}\sigma_{ij}^{'})} \tag{2.15'}$$

$$\Delta\varepsilon^{c*} = \sqrt{(\frac{2}{3}\Delta\varepsilon_{ij}^{c}\Delta\varepsilon_{ij}^{c})} \tag{2.16'}$$

$$\Delta\varepsilon_{ij}^{c} = \frac{3}{2}\frac{\Delta\varepsilon^{c*}}{\sigma^{*}}\sigma_{ij}^{*} \qquad (2.17')$$

Equations (2.18) and (2.19) are combined to yield

$$\Delta\varepsilon_{ij}^{p} = \frac{3}{2}\frac{f'(\sigma^{*})\Delta\sigma^{*}}{\sigma^{*}}\sigma_{ij}^{'} \qquad (2.18')$$

where $f'(\sigma^{*})$ is the derivative of the experimentally determined function, $f(\sigma^{*})$, $\Delta\sigma^{*}$ is the increment of effective stress corresponding to the time increment Δt, i.e., $\Delta\sigma^{*} = \sigma^{*}(t+\Delta t) - \sigma^{*}(t)$.

The procedure for solving total nonelastic strain as a function of time is as follows:

(1). At the start of the first time interval $t=\Delta t$, ε^{C} and ε^{P} are zero. Assuming that $\Delta\varepsilon^{C}$ and $\Delta\varepsilon^{P}$ are zero and substituting these values into Eqn. (2.11') gives a first approximation of stress in the matrix at $t=\Delta t$.

(2). Substitute the approximation of stress in the matrix into Eqn. (2.15') and then Eqn. (2.14') to obtain the first approximation of incremental equivalent creep strain.

(3). Substitute the above equivalent creep strain into Eqn. (2.17') to obtain first approximation of the components of incremental equivalent creep strain.

(4). If Eqn. (2.21) is not satisfied or the material is under unloading, let the incremental plastic strain components be zero and go to next step, else compute the incremental plastic strain components from Eqns. (2.18').

(5). The first approximations of stress, incremental creep strain and plastic strain are substituted into Eqn. (2.11') again and the iteration proceeds from Eqns. (2.15'), (2.14'), (2.17') and (2.18'), etc. until the procedure converges to the correct set of incremental creep strain and plastic strain.

(6). At the beginning of the second time increment, the creep strain and plastic strain (ε^{C} and ε^{P}) are known and are equal to incremental creep strain and plastic strain developed during the first time interval. In fact, the creep strain and plastic strain at the beginning of any time interval will always be known and will be equal to the accumulated incremental strains up to that time interval. The procedure for calculating the average stresses and strains for the other time interval is the same as in steps 1 to 5.

Predictions of creep strain of SCS_6/Ti-β-21 unidirectional laminate at a load of 586.5 MPa and temperature of 650^{o} C are compared with experimental

data in Fig. 2.4. In this example, the creep equation for the matrix material, Eqns. (2.14) and (2.14') are replaced by

$$\varepsilon^{c*} = a_0 e^{a_1\sigma^*}[1 - e^{-a_2(\sigma/a_3)^{a_4}t}] + a_5 e^{a_6\sigma^*}t \qquad (2.22)$$

where constants a_0 to a_6 are determined from creep tests of the matrix material [2.11] as $a_0 = 0.00625$, $a_1 = -0.00254$, $a_2 = 0.00087$, $a_3 = 103$, $a_4 = 4.867$, $a_5 = 1.608*10^{-9}$, $a_6 = 0.06008$. The above equation includes both primary and secondary creep of the matrix material. While the predictions agrees reasonable well with experimental data, the model overestimates the creep strain for t less than 2000 seconds, and underestimates the creep strain when t larger than 2000 seconds. Such errors are believed to be caused by fiber fractures in the materials which is experimentally observed but not included in the calculation of Fig. 2.4. Furthermore, the creep strain equation for the matrix, Eqn. (2.22) is obtained based on tests of load range 70 MPa to 103 MPa, but the stress in matrix of the composite under the applied load 586.5 MPa varies in a broader range which cannot be represented by Eqn. (2.22) with values of a_0 to a_6 used.

6. Fiber debonding

Two different fibers are embedded in the matrix, perfectly bonded fibers (Ω_b) and debonded fibers (Ω_d) with volume fraction v_b and v_d, respectively, Fig. 2.5. Under the applied stress σ^0, the average total stress in the matrix is $\sigma^0 + C^m\bar{\varepsilon}$, where $\bar{\varepsilon}$ is the total strain perturbation in the matrix. If a single fiber is introduced into the composite, the Eshelby's equivalent inclusion method [2.12, 2.13] yields the fiber stress as

$$\sigma^0 + \sigma^b = C^f(\varepsilon^0 + \bar{\varepsilon} + S\,\varepsilon^b - \varepsilon^{ne}) = C^m(\varepsilon^0 + \bar{\varepsilon} + S\,\varepsilon^b - \varepsilon^b - \varepsilon^{ne}) \quad \text{in } \Omega_b \qquad (2.23)$$

where ε^b are fictitious nonelastic strains in bonded fibers (Ω_b). The stress and strain in the debonded fibers is expressed as

$$\sigma^0 + \sigma^d = C^f(\varepsilon^0 + \bar{\varepsilon} + S\,\varepsilon^d - \varepsilon^{ne}) \qquad (2.24)$$

and

$$\varepsilon^0 + e^d = (\varepsilon^0 + \bar{\varepsilon} + S\,\varepsilon^d - \varepsilon^{ne}) \qquad (2.25)$$

It is noted that for a debonded void enveloping a high modulus fibers, the actual stress in X_1 direction and strain in X_2 and X_3 direction are zero, i.e.,

$$(\sigma^0 + \sigma^d)_{11} = 0, \qquad (\varepsilon^0 + e^d)_{22} = 0, \qquad (\varepsilon^0 + e^d)_{33} = 0 \tag{2.26}$$

The summation of the stress perturbation is equal to zero, i.e.,

$$v_b \langle \sigma^b \rangle + v_d \langle \sigma^d \rangle + (1 - v_b - v_d)\, C^m \bar{\varepsilon} = 0 \tag{2.27}$$

Eqns. (2.23)-(2.27) can be used to solve the unknowns , $\bar{\varepsilon}$, e^d and e^b.

The transverse behavior of SCS_6/Ti-β-21 unidirectional laminates at a temperature of 650^o C is considered using the proposed model. Figure 2.6 shows the growth of void as a function of applied load, where a is the radius of the fibers, c is the larger half axis of the elliptic voids, Fig 2.5. In Fig. 2.7, creep strain of the composite is shown as a function of applied load in MPa.

REFERENCES

[1.1]. Johnson, W. S., Lubowinski, S. J. and Highsmith, A. L., 'Mechanical Characterization of Unnotched SCS_6/Ti-15-3 Metal Matrix Composites at Room Temperature' ASTM 1080, J. M. Kennedy, H. H. Moeller, and W. S. Johnson, Eds., American Society for Testing and Materials, Philadelphia (1990) pp 193-218.

[1.2]. Reifsnider, Fatigue of Composite Materials, Elsevier Science Publishers B. V. (1990).

[2.1]. Eshelby, J. D. 'The Determination of the Elastic Field of an Ellipsoidal Inclusion, and Related Problems' *Proceedings of Royal Society, London, Series A* Vol 241 (1957) pp 376-396.

[2.2]. Mori, T. and Tanaka, K. 'Average Stress in Matrix and Average Elastic Energy of Materials with Misfitting Inclusions' *Acta Metall* Vol 231 (1973) pp 571-574.

[2.3]. Taya, M. and Mura, T. 'On the Stiffness and Strength of an Aligned Short-Fiber Reinforced Composite Containing Fiber-End Cracks Under Uniaxial Applied Stress' *ASME Journal of Applied Mechanics* Vol 48 (1981) pp 361-367.

[2.4]. Taya, M. and Chou, T. W. 'On Two Kinds of Ellipsoidal Inhomogeneities in an Infinite Elastic Body: An Application to a Hybrid Composite' *International Journal of Engineering Science* Vol 17 (1981) pp 553-563.

[2.5]. Weng, G. J. 'Some Elastic Properties of Reinforce Solids with Special Reference to Isotropic Ones Containing Spherical Inclusions' *International Journal of Engineering Science* Vol 22 (1984) pp 845-856.

[2.6]. Norris, A. N. 'An Examination of the Mori-Tanaka Effective Medium Approximation for Multiphase Composites' *Journal of Applied Mechanics* Vol 56 (1989) pp 83-88.

[2.7]. Tandon, G. P. and Weng, G. J. 'Stress Distribution in and Around Spheroid Inclusions and Voids at Finite Concentration' *Journal of Applied Mechanics* Vol 53 (1986) pp. 511-518.

[2.8]. Benveniste, Y. 'A New Approach to the Application of Mori-Tanaka's Theory in Composite Materials,' *Mechanics of Materials* Vol 6 (1987) pp 147-157.

[2.9]. Mukherjee, A. K., Bird, J. E. and Dorn, J. E., *Trans. ASM*, Vol 62 (1969) pp 155-179.

[2.10]. Bodner, S. R., 'Review of a Unified Elastic-Viscoplastic Theory,' Unified Constitution Equations for Creep and Plasticity, *Elsevier Applied Science Pub.*, England (1987) pp 273-301.

[2.11]. M. Boyle, unpublished work.

[2.12]. Taya, M. and Patterson, W. 'Growth of a Debonded Void at a Rigid Secondary Particle in a Viscous Metal,' *J. Materials Science*, Vol 17 (1982) pp 115-120.

[2.13]. Kyono, T., Hall, I. and Taya, M. 'The Effect of Isothermal Exposure on the Transverse Properties of a Continuous Fibre Metal-Matrix Composite,' *J. Materials Science*, Vol 21 (1986) pp 4269-4280.

ACKNOWLEDGMENT

We would like to especially thank Drs. J. Larsen, T. Nicholas and S. Mall for helpful discussions and guidance. Our thanks are also extended to Dr. R. Neu, J. Kroupa, M. Boyle, Dr. M. Khobaib, D. Coker and H Zhao for their assistance. Support of AFOSR to the first author through the Summer Research Program is gratefully acknowledged.

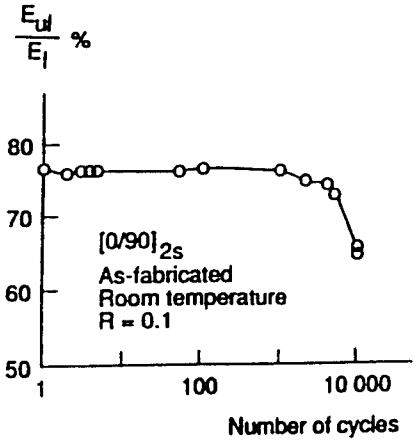


Fig. 1.1. Change in elastic unloading modulus for [0/90]2s layups of SCS_6/Ti-15-3 [1.1].

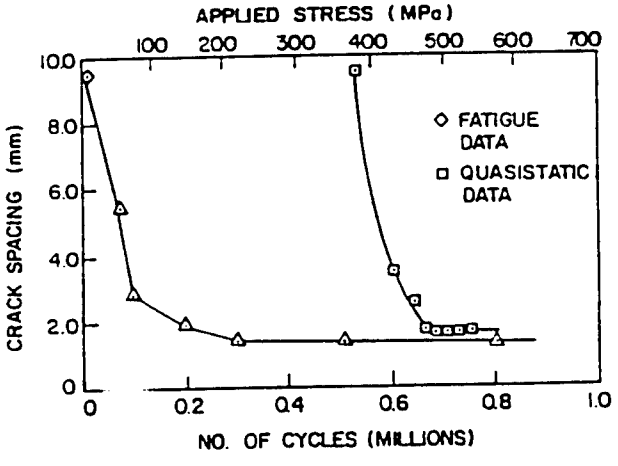


Fig. 1.2. Crack spacing in 45^0 plies of a [0/90/45/-45]s graphite epoxy laminates as a function of increasing quasi-static load or cycles [1.2].

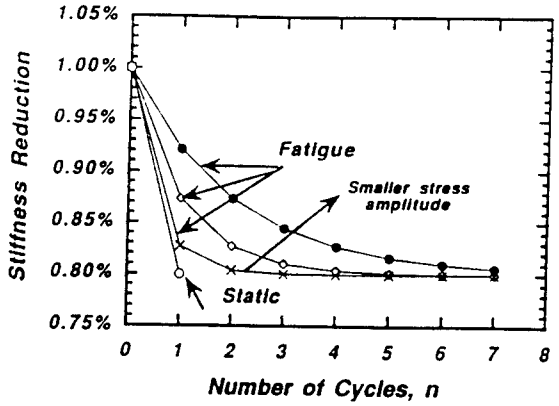


Fig. 1.3. A schematic representation of stiffness reduction as a function of loading cycle.

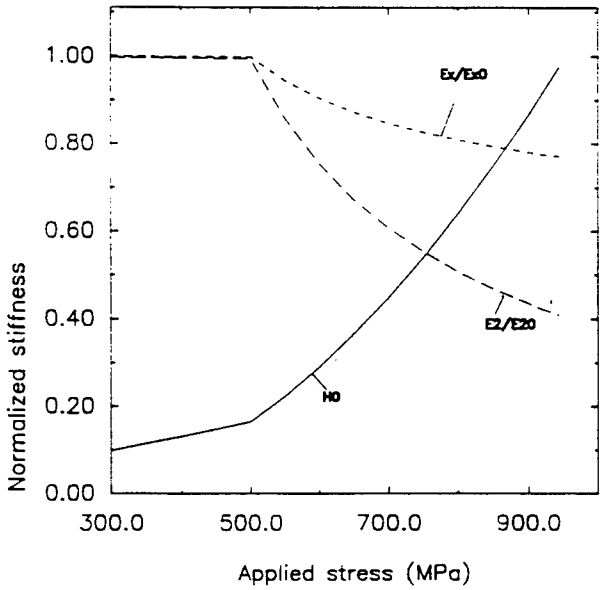


Fig. 1.4. Stiffness reduction and stress redistribution of a SCS_6/Ti-15-3, [0/90]2s laminate. $Ho=\sigma_{11}(0^0)/X$, the stress to strength ratio of the 0^0 plies.

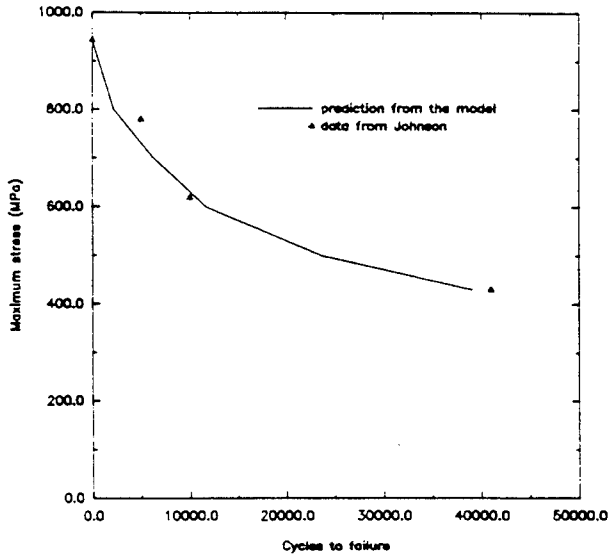


Fig. 1.5. Fatigue life prediction and test data for $[0/90]_{2s}$ laminates.

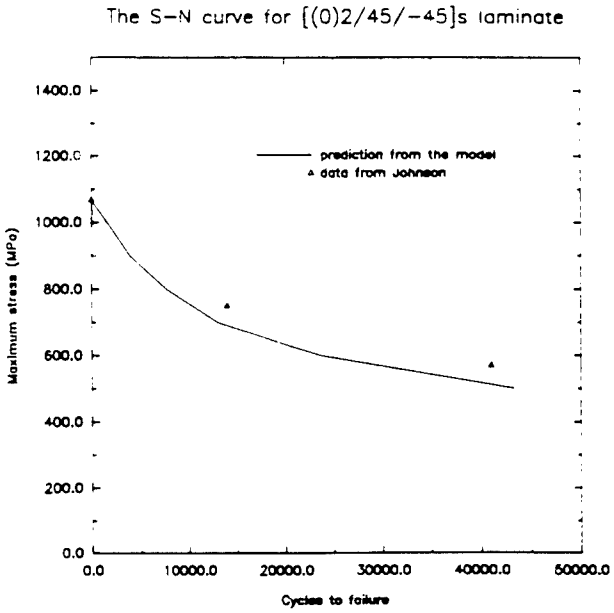


Fig. 1.6. Fatigue life prediction and test data for $[0_2/45/-45]s$ laminates.

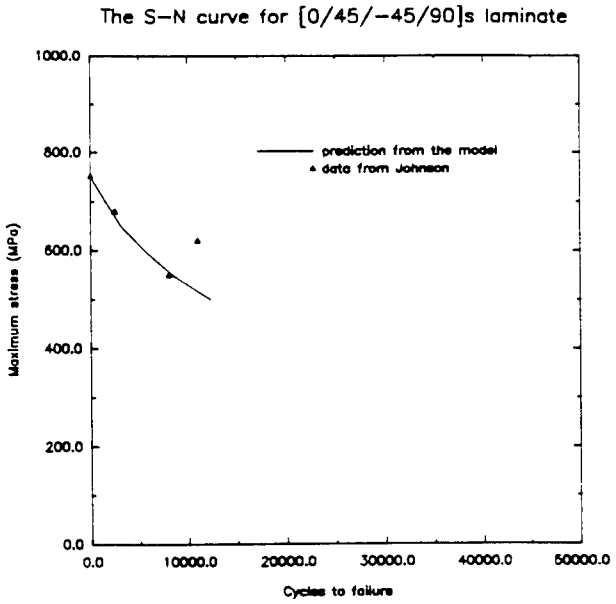


Fig. 1.7. Fatigue life prediction and test data for [0/45/-45/90]s laminates.

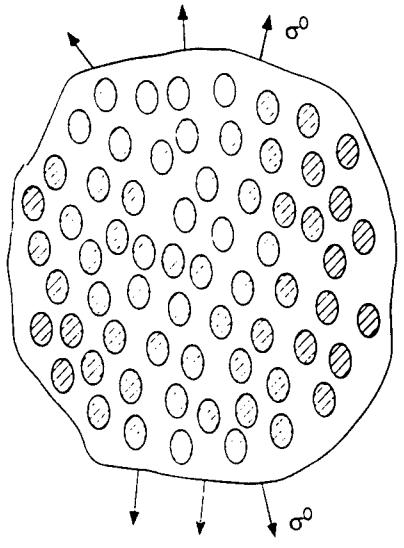


Fig. 2.1. A schematic diagram of composites showing the matrix, with stiffness C^m, and fibers with stiffness C^f.

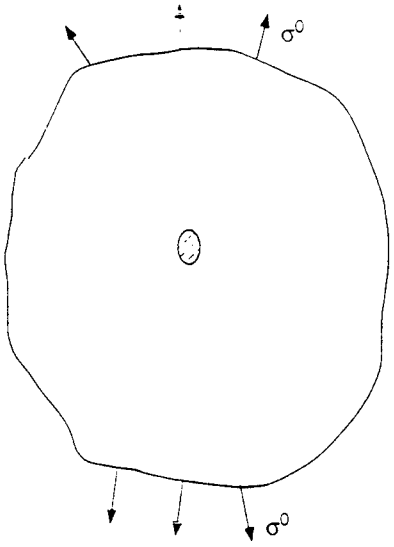


Fig. 2.2. A single inclusion in an infinite solid under an applied load, $\sigma^0 n$.

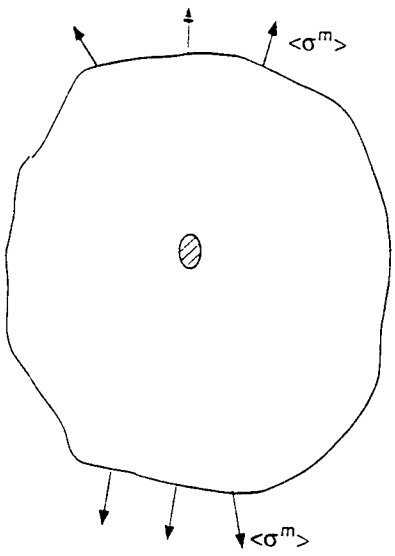


Fig. 2.3. The non-dilute solution of stress in the inhomogeneities is obtained by solving the single inclusion problem with applied load replaced by the average stress of the matrix of the original problem, $<\sigma^m>$.

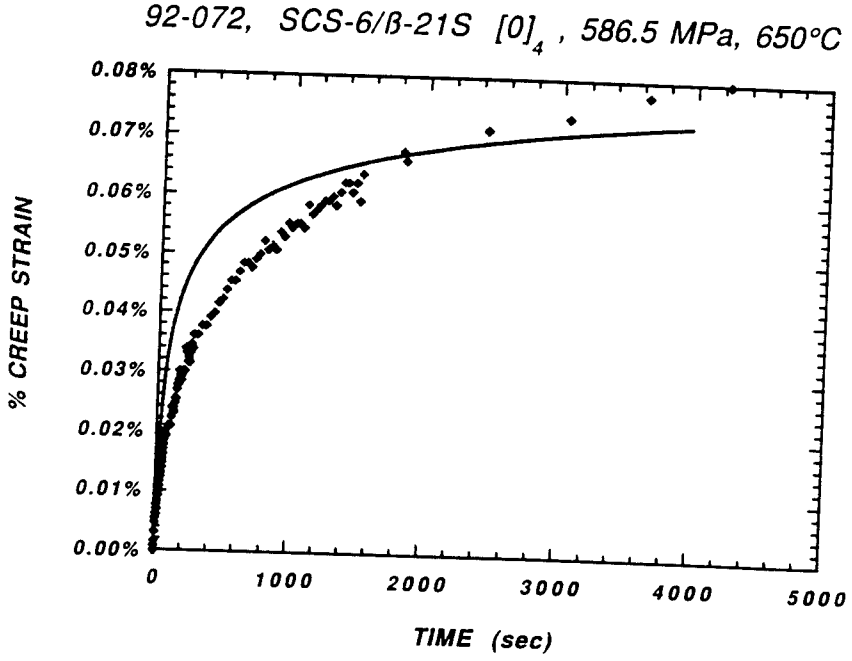


Fig. 2.4. Prediction (solid line) and test data of creep strain of SCS_6/β-21S unidirectional laminate under a constant load of 586.5 MPa at 650^0C.

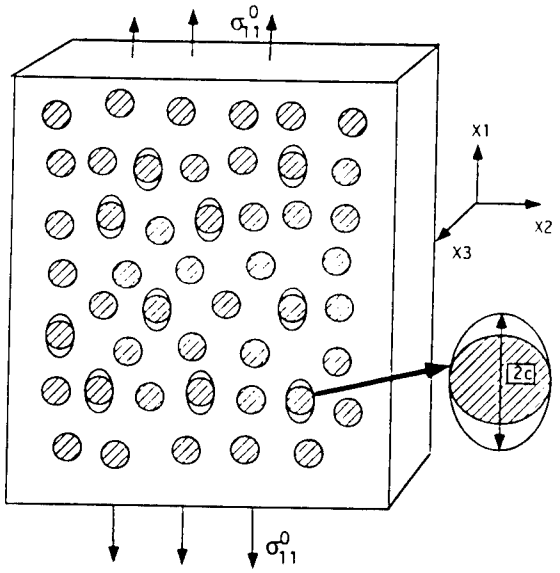


Fig. 2.5. A schematic view of a laminate under a transverse loading. Debonding occurs at a result of interfacial failure.

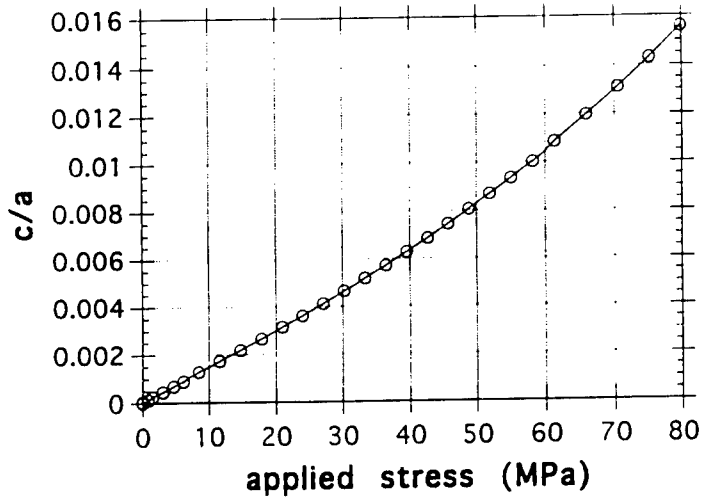


Fig. 2.6. Normalized debonding length as a function of applied stress at 650^0C. c: the larger half axis of the elliptic voids, a: radius of fibers.

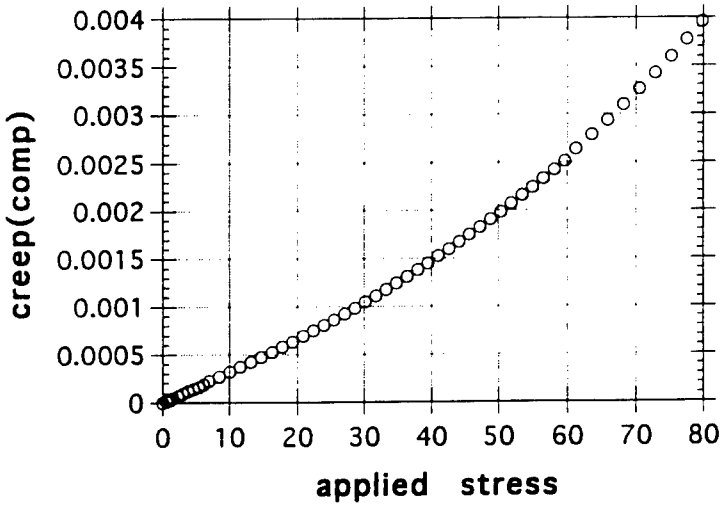


Fig. 2.7. Composite creep strain as a function of applied stress at 650^0C.

REAL-TIME IN SITU ELLIPSOMETRY OF POLYMER FILMS PRODUCED BY FLOWING AFTERGLOW SYNTHESIS

Joyce A. Guest
Associate Professor
Department of Chemistry

University of Cincinnati
Cincinnati, OH 45221-0172

Final Report for:
Summer Faculty Research Program
Wright Laboratory/MLPJ

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

September, 1993

Real-Time In Situ Ellipsometry of Polymer Films Produced by Flowing Afterglow Synthesis

Joyce A. Guest
Associate Professor
Department of Chemistry
University of Cincinnati

Abstract

A helium-neon laser-based ellipsometer has been constructed and coupled to a flowing afterglow plasma reactor to monitor in real time the growth of polymeric thin films formed from benzene and thiophene precursors. This phase modulated ellipsometer is shown to be robust and to provide time-dependent data that vary sensitively with film thickness. Progress has also been made in evaluating the utility of benzene/thiophene precursor injection using a pulsed injection valve downstream of the argon metastable excitation source. This system shows promise for development into an apparatus for controlled deposition of multilayer polymer films of known morphology.

Real-Time In Situ Ellipsometry of Polymer Films Produced by Flowing Afterglow Synthesis

Joyce A. Guest

Introduction

Multilayer polymer thin films have many possible applications as nonlinear optoelectronic devices[1] and as rugate filters,[2] optical coatings, and organic electroluminescence cells. [3] Flowing afterglow plasma synthesis is a unique and versatile method to generate dense polymeric structures with promising morphology and optical properties.[4] At Wright Laboratory, Peter Haaland (of Lawrence Associates) and coworkers are exploring these syntheses in reactors which separate the active plasma from the injection region for the organic precursors. The organic species are separated effectively from ultraviolet light, electrons, ions, and fields generated in the plasma, and instead collide predominately with highly excited argon metastable atoms. Plasma conditions, substrates, and precursors have been varied, and a variety of single and multilayer polymer films have been produced and then evaluated *ex situ.*

The plasma polymer deposition apparatus at Wright Laboratory includes a quadrupole mass spectrometer, which has been used to evaluate homogeneous kinetics in the gas phase during polymer deposition. We decided that incorporating a non-invasive real-time *in situ* probe of the polymer film *during deposition* would be a powerful tool for understanding and controlling the characteristics of these materials. Such a tool would enhance the capabilities of the plasma reactor significantly for producing useful multilayer films.

We chose to construct an ellipsometer for this apparatus for a variety of reasons. Reflectance ellipsometry is a powerful and well-understood non-destructive technique in which the polarization transformation in a light beam upon reflection from a surface is

measured and interpreted.[5,6] The optical properties of the material on which the light impinges determine the phase and amplitude change in the light upon reflection. The method is extremely sensitive to the refractive index, absorption coefficient, and thickness of the film. There are a wide variety of commercial and custom-designed ellipsometers that are appropriate for different situations, and vary over a range of complexity. For example, the films produced in this laboratory are typically evaluated after deposition by a commercial spectroscopic ellipsometer[7], and the wavelength-dependent results are used in modeling the sample refractive index and thickness. Interest has accelerated recently in developing ellipsometry as a real time monitor of material growth processes.[8]

The present studies were performed with a home-built polarization modulated ellipsometer. This technique was introduced by Jasperson and Schnatterly in 1969,[9] and has been explored thoroughly by a small number of research groups over the ensuing years.[10] The method is particularly well-suited to real-time measurement and control of multilayer structure processing such as etching,[11] and has just recently begun to be utilized for such studies. The heart of the polarization modulated ellipsometer is a photoelastic modulator (PEM), which is a 50 kHz oscillating wave plate of controllable amplitude based on the piezo-optic effect.[12] Use of the PEM permits the construction of an ellipsometer with no mechanical moving parts and contributes greatly to ease of use. Sampling could potentially be done in a time period as short as 20 μs, or as long a duty cycle as desired.

At the same time as the PME was being constructed and evaluated in this laboratory, we set up and tested a new plasma reactor[13] in collaboration with Hao Jiang (Lawrence Associates). This new reactor incorporates an inductively coupled rf discharge source and two pulsed injector valves downstream for precursor introduction. The preliminary studies of this reactor will be discussed in this report in conjunction with the ellipsometer development.

Experimental Method

The experimental description will focus primarily on the ellipsometer and its integration into the plasma reactor. Secondarily, the changes to the plasma reactor will be discussed briefly as they relate to the overall project.

Polarization modulated ellipsometer. Several groups have exploited and advanced the understanding of this powerful PEM-based non-mechanical ellipsometry method, but a PEM-based ellipsometer has not been produced commercially, to our knowledge. There has been a revival of interest in polarization modulated ellipsometry (PME) recently, because of its potential for sub-millisecond temporal resolution, and improvements in computational capabilities that permit rapid data analysis for real-time process control.[11]

The PME constructed for this project is of standard reflection design. The light source is a 1 mW HeNe laser operating at 632.8 nm. The laser light passes through an alignment iris, a Glan-Taylor polarizer (polarization axis 45° from vertical), PEM (Hinds International PEM-FS-4 at 0° orientation), and then through a window into the reactor, striking the sample at ~70° from normal incidence. The reflected light passes through an exit window, a Glan-Taylor analyzer (±45° orientation), and an alignment iris, and finally strikes a photodiode detector (EG&G FND-100). The detector output is processed (1 kΩ input impedance) by a lock-in amplifier (SRS model SR530) that is interfaced to a 486DX-based computer. The interface program permits lock-in signal acquisition at the first and second harmonics of the reference frequency, which is the PEM oscillation frequency. The SR530/computer combination can also acquire and process two additional photodiode outputs if required for normalization of laser power variations.

The measured signals, S_ω, $S_{2\omega}$, and S_0 at 50 kHz, 100 kHz, and DC, can be calibrated and converted to intensities I_ω, $I_{2\omega}$, and I_0 that are directly related to the

ellipsometric angles Ψ and Δ by the following relations:

$$I_\omega = I_o \sin 2\Psi \sin \Delta$$
$$I_{2\omega} = I_o \sin 2\Psi \cos \Delta$$

The calibration consists of determining the system transfer characteristics by prescribed methods,[9,10] and by setting the angular orientation of the sample with respect to the ellipsometer beam path. The ellipsometric angles Ψ and Δ are related to the ratio of the complex Fresnel reflection coefficients r_p and r_s for p and s polarizations,

$$\frac{r_p}{r_s} = \tan \Psi e^{i\Delta}$$

Hence $\tan \Psi = |r_p|/|r_s|$, and $\Delta = \delta_p - \delta_s$, where the δ_i are the phase changes upon scattering for each polarization.

Even without a complete system calibration, one can evaluate thickness and refractive index changes during deposition by monitoring the 50 kHz and 100 kHz lock-in signals. This should be true for our apparatus, since multilayer depositions are performed using films of known refractive index n and extinction coefficient k.

Plasma reactor. The new flowing afterglow reactor is a modification of the apparatus described in Haaland and Targove. Argon (99.999%) carrier gas is flowed at 200 sccm at a pressure of 0.5 Torr, and excited with ~20 W of inductively coupled radiofrequency power at 13.56 MHz. Following recombination of electrons and ions downstream, the precursor monomer species are injected into a flow region where argon metastable species dominate tens of milliseconds downstream. A right-angle bend following the plasma source minimizes the amount of ultraviolet light that will strike the precursor and potentially lead to undesired electronic excitation/photochemistry.

The modified reactor includes fittings onto which two Lasertechnics LPV pulsed valves (0.5 mm orifice) are attached. These valves permit controlled injection of precursor gases, and facilitate examination of the effect of gas pulse length and frequency on the polymer film deposition rate. The precursor gas flow rate was typically 0.2-0.5 sccm, produced in 2 ms pulses at 5 or 10 Hz. The film deposition rate is gauged by quartz crystal deposition monitor measurements under a variety of argon flow, pressure, rf power, and precursor conditions, as well as a function of substrate position along the reactor tube.

For the ellipsometry studies, the sample substrate was mounted 20° from the long axis centerline of the reactor, as shown in Figure 1. In this configuration, the input laser passes through the tube or an attached window 40° from the reactor centerline and strikes the sample. The reflected beam exits nearly along the centerline of the reactor, through a flat window. This arrangement was chosen for ease of construction and to keep the ellipsometer windows in a region where deposition is minimized (at the entrance window), and nonexistent (at the exit window upstream from the precursor injection line). We found that the sample could be translated about 3 mm on its manipulator without affecting the laser beam path significantly.

Results and Discussion

Figure 2 shows the unnormalized $S_{2\omega}$ signal variation over a three-hour period during film deposition on a silicon wafer substrate with benzene as a precursor. The film thickness change over this period is estimated to be less than 200 nm, based on optical constants and film thickness derived from *ex situ* measurements using a Rudolph Instruments S2000 spectroscopic ellipsometer after the real-time studies on the sample were complete. Deposition rates in the new system are presently low: <0.02 nm/s. Nonetheless, the real-time method is sensitive to thickness changes on the order of 0.1-0.2 nm for this sample,

which has a derived complex refractive index N of about 1.63 - 0.04i in the 600 nm region. It is remarkable that the first test of this device shows such sensitivity and stability over a several hour period.

Following improvements to the data acquisition software, we examined film deposition using thiophene as a precursor. Figure 3a shows both the 50 kHz signal (S_ω) and 100 kHz signal ($S_{2\omega}$) during plasma deposition over nearly 4.75 hours. Again, the deposition rate is believed to be very low, about 0.02 nm/s, but both signals vary rather smoothly over time, considering that they are not normalized for incident light intensity. Figures 3b and 3c show 500 second timescale expansions of Figure 3a, corresponding to about 10 nm polymer deposition thickness. These figures show that the present system is sensitive to angstrom-scale changes in the film. Note also that the rf power was increased after 6700 s, leading to an increase in the deposition rate and hence signal variation. The joint trends in the 50 kHz and 100 kHz signals during deposition are sensible, since the former varies as sin 2Ψ sin Δ and the latter varies as sin 2Ψ cos Δ.

We can explore the temporal variations in the PME signal further, and extend the results to predictions for multilayer films, using the methods to compute reflection amplitudes and comp..te reflection matrices for multilayer structures from MacLeod.[14] The ellipsometric angles Ψ and Δ can be extracted from the complex reflection amplitude ratio for p- and s-polarized light, and then converted to I_ω and $I_{2\omega}$, which are proportional to the signals measured in the PME experiment. Figure 4 shows the relative I_ω and $I_{2\omega}$ calculated for a two-layer, slightly absorbing film, the first with refractive index N=1.6 -0.05i (1000 nm thick), and the second overlayer ..h N=1.8-0.05i (1000 nm thick). The intensities vary in amplitude and period as the refractive index is changed, as expected. Figure 5 shows the Ψ - Δ trajectory for the same model system. Because the films are absorbing, the Ψ - Δ trajectory does not

repeat cyclically.

During these studies, we noticed that the scattered laser beam alignment remained constant when the sample was translated up to 3 mm along the reactor centerline. Since the HeNe beam is <1.5 mm, this means that we can assess the film thickness profile over the translation region. Improvements in the translator and sample mounting could make possible the *in situ* study of the polymer film over a variety of positions with minimal realignment.

The ease of use and rapid alignment of the *in situ* ellipsometer is due in part to the fact that its light source is a collimated laser, compared to the broadband sources used with spectroscopic ellipsometers. A limitation is that only one wavelength is available from the HeNe laser used for the present studies, although laser sources that operate at two or more wavelengths could be used if necessary. The important requirement for monitoring multilayer film deposition is that the refractive indices of the films from the different precursor materials are known and differ at the PME monitor wavelength.

For deposition of multilayer films of known composition, one can evaluate layer thickness changes during deposition by monitoring the 50 kHz and 100 kHz lock-in signals as a function of time. Each film deposition process has a history that is preserved throughout the data acquisition. This history can be used on-line with model calculations to compute the thickness and growth rate of each polymer layer, utilizing the cyclical nature of the ellipsometric signal during deposition. The temporal information is thus much more powerful than single-point analysis for process control.

Conclusions and Future Directions

We have demonstrated that polarization modulated ellipsometry is a rugged and practical technique for real-time monitoring of polymer film growth in a flowing afterglow reactor. The preliminary working design successfully overcomes several potential pitfalls,

among them coating of the reactor windows by polymeric and other debris, and physical instabilities leading to signal instabilities. The planned construction of a permanent *in situ* ellipsometer for the flowing afterglow reactor is expected to make a major contribution to the evolution of the plasma technique into a useful method for custom fabrication of multilayer heterostructures such as organic rugate filters and NLO devices. This real-time ellipsometer is a cost-effective instrument that will efficiently guide the technical development of these polymer multilayer growth processes.

The next generation ellipsometer will be based closely on the one constructed for this project. The sample holder and manipulator will be improved to facilitate alignment reproducibility, which is critical for quantitative measurements. The system needs to include light intensity normalization and optical system calibration, both of which can be patterned after previously demonstrated procedures. This new version of the flowing afterglow system can be optimized, utilizing the results of fluid dynamics calculations and rf power coupling enhancement to increase and control uniform polymer deposition. A variety of *ex situ* spectroscopic and microscopic techniques can peg the identity and overall morphology of the polymer films, while the real-time *in situ* ellipsometer evaluates the growth dynamics.

Acknowledgments

I am indebted to Peter Haaland and Hao Jiang for their generous collaboration and efforts contributing to the realization of this project. I also thank Robert Crane for his support and valuable advice throughout the summer.

References

1. P. Prasad and D. Williams, *Introduction to Nonlinear Optical Effects in Molecules and Polymers* (Wiley, New York, 1991).

2. W. E. Johnson and R. L. Crane, Proc. SPIE **2046** (August, 1993); B. A. Tirri, J. E. Lazo-Wasem, and T. D. Rahmlow, Jr., Proc. SPIE **2046** (August, 1993).

3. C. Hosokawa, H. Higashi, and T. Kusumoto, Appl. Phys. Lett. **62**, 3238 (1993).

4. P. Haaland and J. Targove, Appl. Phys. Lett. **61**, 34 (1992); P. D. Haaland and S. J. Clarson, Trends in Polymer Sci. (February, 1993); P. Haaland and H. Jiang, Polymer Preprints **34**, 675 (1993).

5. R. M. A. Azzam and N. M. Bashara, *Ellipsometry and Polarized Light* (North-Holland, New York, 1987).

6. H. G. Tompkins, *A User's Guide to Ellipsometry* (Academic, San Diego, 1993).

7. Rudolph Instruments S2000 Spectroscopic Ellipsometer.

8. N. J. Ianno, S. Nafis, P. G. Snyder, B. Johs, and J. A. Woolam, Appl. Surf. Sci. **63**, 17 (1993); H. Schwiecker, D. B. Dang, H. P. T. Thanh, J. Zilian, U. Schneider, and J. Heland, Proceedings SPIE **1746**, 222 (1992); R. W. Collins, Rev. Sci. Instrum. **61**, 2029 (1990); W. H. Southwell and W. J. Gunning, Proc. SPIE **1019**, 84 (1988).

9. S. N. Jasperson and S. E. Schnatterly, Rev. Sci. Instrum **40**, 761 (1969).

10. G. E. Jellison, Jr., and F. A. Modine, Appl. Opt. **29**, 959 (1990); O. Acher, E. Bigan, and B. Drevillon, Rev. Sci. Instrum. **60**, 65 (1989); V. M. Bermudez and H. Ritz, Appl. Opt. **17**, 542 (1978).

11. W. M. Duncan and S. A. Henck, Appl. Surf. Sci. **63**, 9 (1993).

12. J. C. Kemp, J. Opt. Soc. Am. **59**, 950 (1969).

13. Designed by Peter Haaland.

14. H. A. MacLeod, *Thin-Film Optical Filters*, (McGraw-Hill, New York, 1989).

Figure Captions

Figure 1: Schematic of laser beam path in flowing afterglow reactor.

Figure 2: 100 kHz lock-in signal from *in situ* polarization modulated ellipsometer during film deposition (benzene precursor gas).

Figure 3: (a) 50 kHz and 100 kHz lock-in signal from *in situ* polarization modulated ellipsometer during film deposition (thiophene precursor gas). (b) and (c) are 500 second regions of the deposition process measured in (a).

Figure 4: Predicted polarization modulated ellipsometer intensities over 2000 nm film thickness for two-layer film: 1000 nm (N=1.6-0.05i), then 1000 nm (N=1.8-0.05i) over substrate with n=1.5. Solid line is I_{ω}, dashed line is $I_{2\omega}$.

Figure 5: Predicted Ψ-Δ trajectory for evolution of two-layer film of 1000 nm each as described for Figure 4.

Figure 1

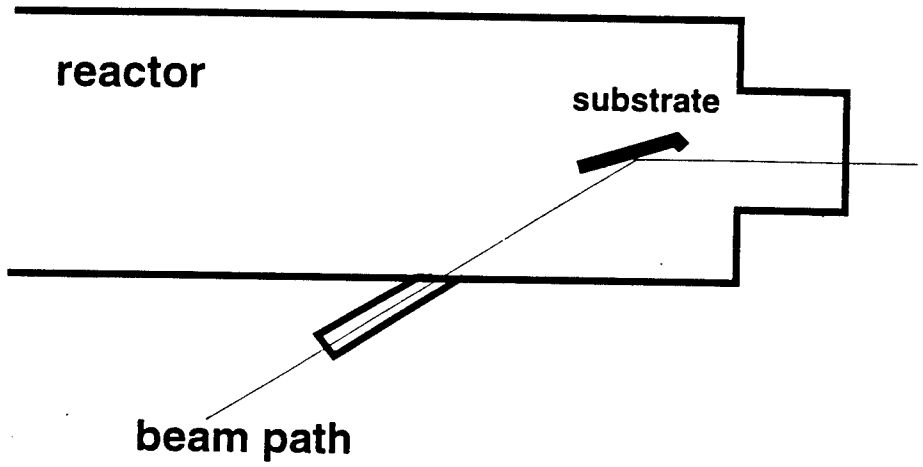


Figure 2

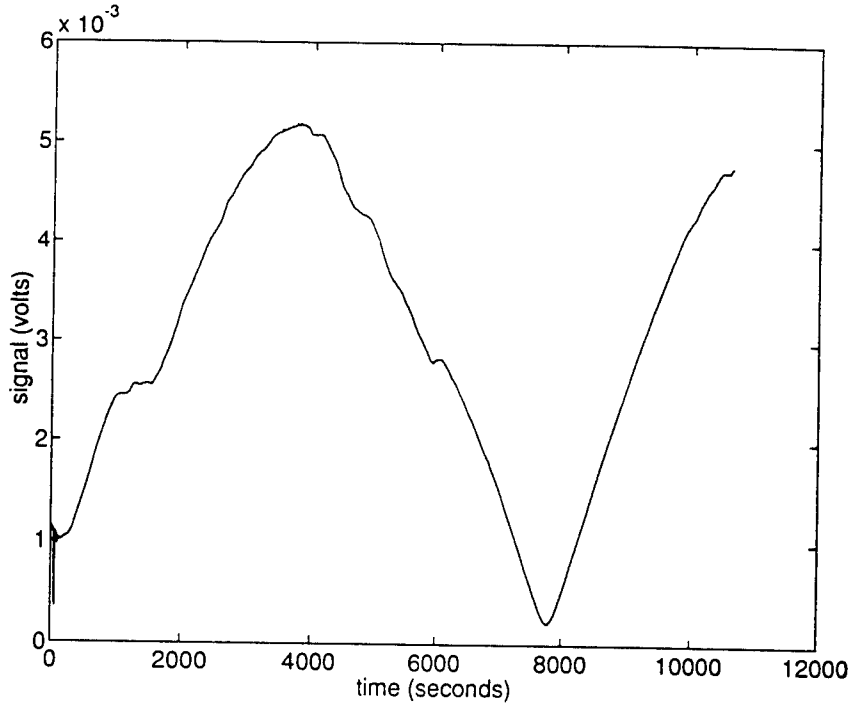


Figure 3a

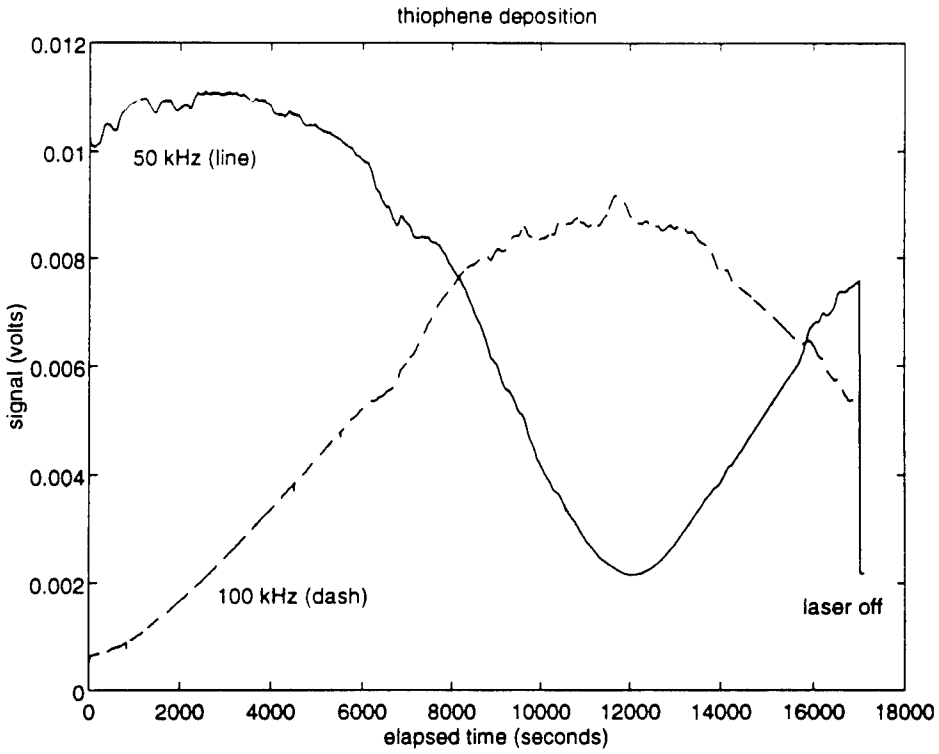


Figure 3b

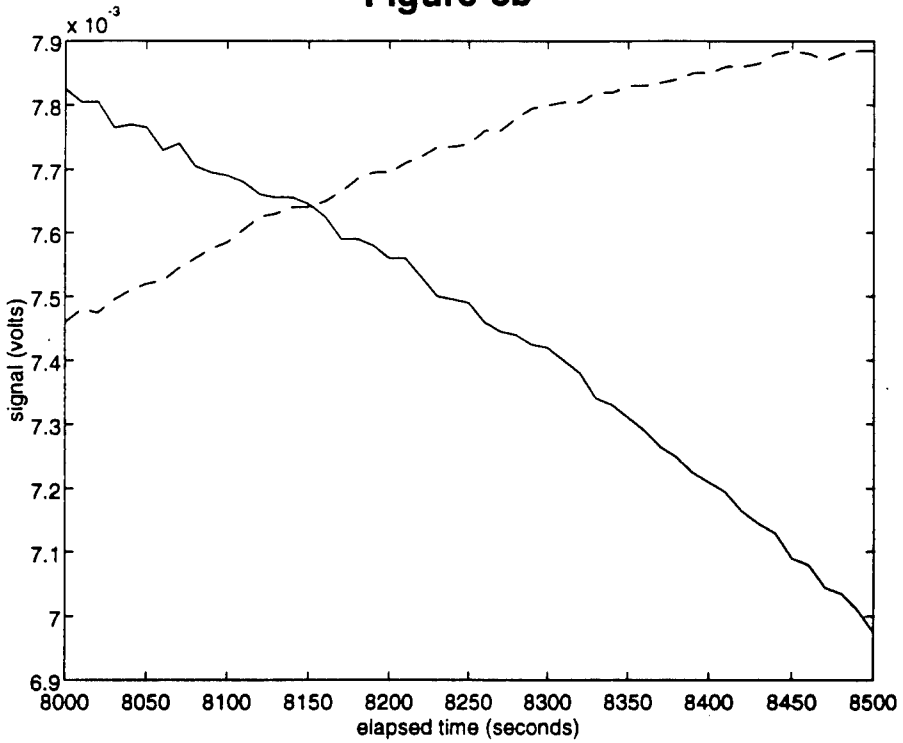


Figure 3c

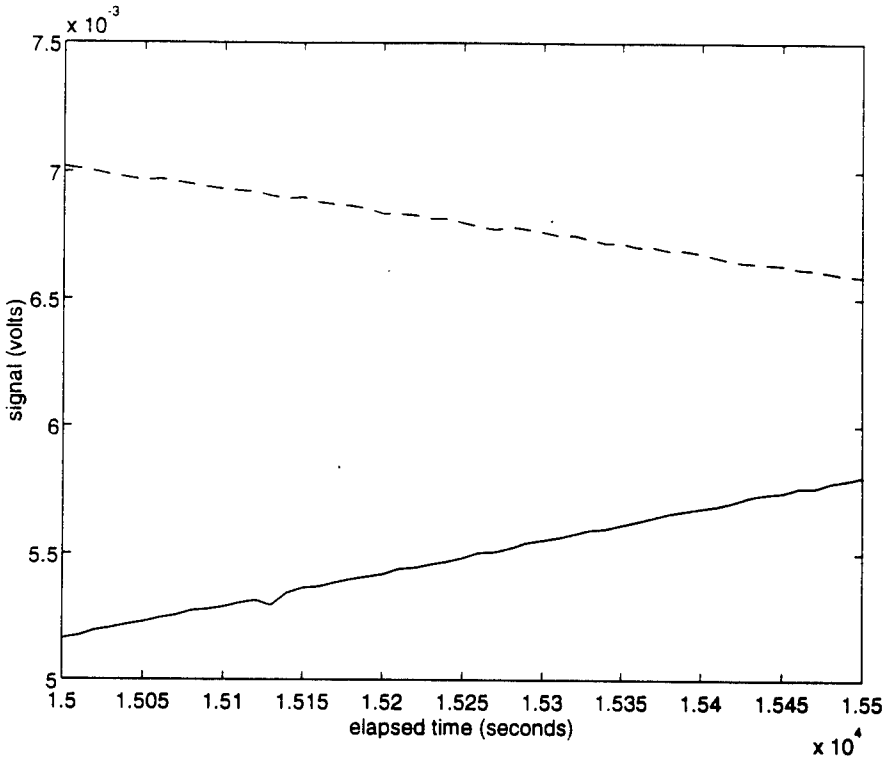

x 10^-3
7.5
7
6.5
6
5.5
5
signal (volts)
1.5
1.505
1.51
1.515
1.52
1.525
1.53
1.535
1.54
1.545
1.55
elapsed time (seconds)
x 10^4

Figure 4

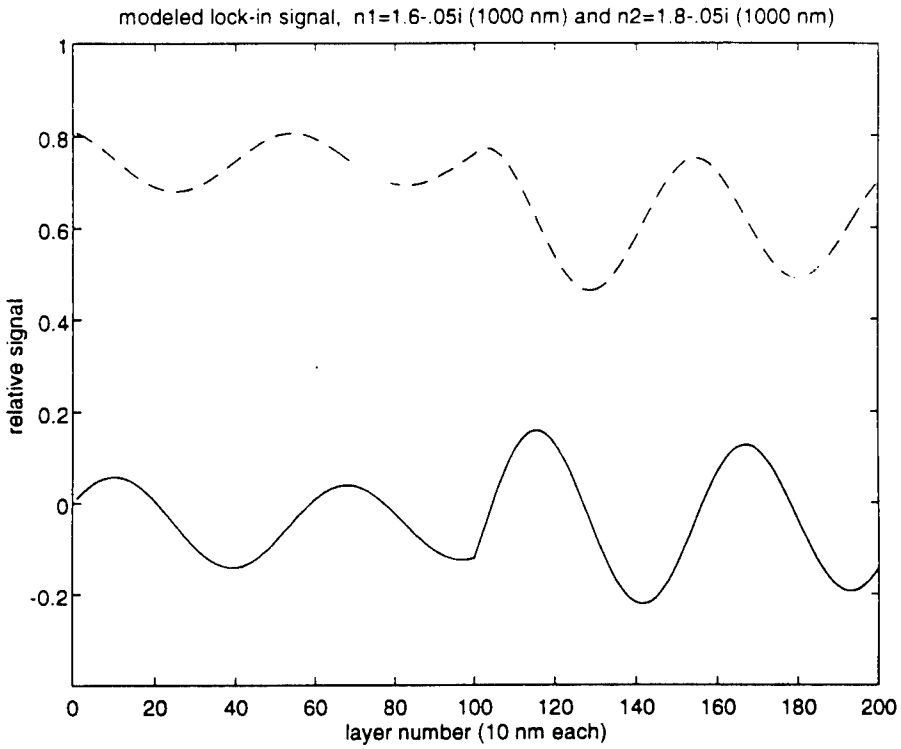
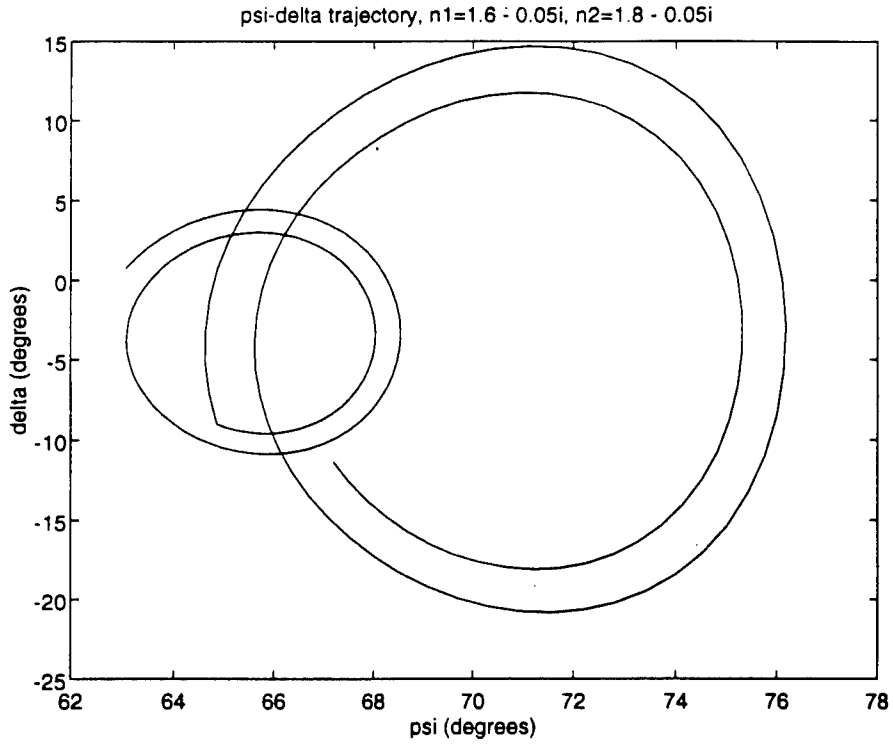


Figure 5



MOLECULAR DYNAMICS SIMULATION OF B2 NiAl

John A. Jaszczak
Assistant Professor
Department of Physics

Michigan Technological University
1400 Townsend Dr.
Houghton, MI 49931-1295

Final Report for:
Summer Faculty Research Program
Wright Laboratory

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

August, 1993

MOLECULAR DYNAMICS SIMULATION OF B2 Ni-Al

John A. Jaszczak
Assistant Professor
Department of Physics
Michigan Technological University

Abstract

The development of alloys based on the ordered intermetallic alloy NiAl in the are of interest for high temperature aerospace components. NiAl in the B2 structure is also a popular material for basic scientific research in ordered alloys. In this regard, there has been recent interest in understanding dislocation mobility in NiAl at the atomic level through atomistic computer simulation. Atomistic simulations employing realistic elastic boundary conditions of straight dislocations have been used to estimate relative dislocation mobilities by determining the stress required to initiate dislocation motion. However, these studies have been confined to zero temperature. Many properties, such as the thermal activation of kink pairs and the activation pathway for dislocation cross slip, are best studied using finite-temperature molecular dynamics simulations.

In order to test and develop molecular dynamics simulation capabilities for dislocation-mobility studies, as well as to begin maintaining a database of thermal properties of the NiAl potential, preliminary molecular dynamics studies have been carried out to study stoichiometric NiAl in the B2 structure at 400°K using an empirical embedded-atom-method potential developed explicitly for NiAl. Computer codes have been developed and tested to statistically process data generated in the course of the simulations. Such thermodynamic quantities as the energy, heat capacity, temperature, stress tensor, volume and pair distribution functions can be calculated in various thermodynamic ensembles. In preparation for the simulation of dislocations and dislocation mobility, a capability has been developed to track the mis-coordination of atoms and their average spatial coordinate through the course of the simulation.

MOLECULAR DYNAMICS SIMULATION OF B2 NiAl

John A. Jaszczak

Introduction

The development of alloys based on the ordered intermetallic alloy NiAl are of interest for high temperature aerospace components and other engineering applications.[1] Because of its simple crystal structure, highly ordered lattice, range of compositional stability, and shape memory effect, NiAl is also a popular compound for basic scientific research in ordered alloys. In order to understand and, hopefully, improve the ductility of such alloys, there has been recent interest in understanding dislocation mobility in NiAl at the atomic level through atomistic computer simulation. Atomistic simulations employing realistic elastic boundary conditions[2-4] of straight dislocations have been used to estimate relative dislocation mobilities by determining the stress required to initiate dislocation motion.[5] However, these studies have largely been confined to zero temperature. Many properties important to dislocation mobility, such as the thermal activation of kink pairs and the activation pathway for dislocation cross slip, are best studied using finite-temperature molecular dynamics (MD) simulations.[6] MD simulations that have been performed to study the mobility of straight dislocations in NiAl have been deficient in that they have used potentials fitted to Ni_3Al and have not used elastic boundary conditions.[7] Simulations of kinks in BCC metals[8] and in NiAl[9] have been restricted to zero temperature.

During the AFOSR Summer Faculty Research Program, we have tested and begun extending the molecular dynamics simulation capabilities of the code DYNAMO. With the goal of investigating dislocation-mobility in NiAl at non-zero temperatures, we have begun using MD to compile a database of thermal properties of stoichiometric NiAl in the B2 structure using an empirical embedded-atom-method potential developed explicitly for NiAl.[5] Such thermodynamic quantities as the energy, heat capacity, stresses, lattice parameters, elastic constants and pair distribution functions have been calculated at 400°K. The lattice parameters and elastic constants

are of particular importance in developing the elastic boundary conditions for simulations of dislocation mobility.[4,5] In preparation for the simulation of dislocations and dislocation mobility, a capability has been developed to track the mis-coordination of atoms and their average spatial coordinates through the course of the simulation. Initial results, along with a description of the codes developed for producing them, are presented below.

Simulation Methods

The atomistic simulation tools used are based on the codes known as CREATOR (or LATTICE) and DYNAMO, both originally issued by Sandia National Laboratories. LATTICE is used to generate ideal crystal lattices, strained lattices, and single dislocations with strain fields consistent with anisotropic continuum elastic theory. The molecular dynamics simulations were carried out using the enhanced capabilities of DYNAMO.

Perhaps one of the most obvious practical differences between conducting zero-temperature lattice statics simulations (which are also performed within DYNAMO) and non-zero temperature molecular dynamics simulations is that the information gained from MD is statistical in nature. Data describing the instantaneous states of the system are saved throughout the course of the simulation in files that are then post processed to compute average thermodynamic quantities. An important part of this process is the determination of transients and correlation times and the computation of standard deviations. For an excellent overview of molecular dynamics methods as well as other simulation methods, see Ref. 10. The heart of the molecular dynamics simulation is the integration of Newton's equations of motion for the atoms comprising the system of interest. In DYNAMO, this is accomplished using a fifth-order Nordsieck-Gear predictor-corrector method.[10] Various ensembles can be simulated by controlling different independent variables. The number of particles, N, remains fixed in all of the ensembles considered here.

Of particular interest in this work is the microcanonical, canonical, and isothermal-isostress ensembles. In the microcanonical ensemble the volume is fixed and energy of the system is

conserved while the stresses and temperature fluctuate during the simulation. This ensemble is quite useful for checking the correctness of the simulation code and for optimizing the time step by monitoring energy conservation. Most of the simulations of interest will be carried out in the canonical ensemble, in which the temperature and the volume are held fixed. Although it does not formally generate states in the canonical ensemble, temperature control in DYNAMO is accomplished via the "gentle" velocity rescaling approach of Berendsen et al.,[10,11] in which the system is driven to a set temperature at a rate determined by a "thermal equilibration time", set at 0.01 psec in the following isothermal simulations. As one typically desires to simulate the system under zero average stress for a given temperature, some simulations are carried out in the isothermal-isostress ensemble, in which the temperature and the stresses are equilibrated to constant, predetermined values. The method employed for obtaining constant stresses is based on the method of Andersen,[10] generalized for simulation cells of orthorhombic symmetry, and including a damping term to control undesired oscillations in the computational-cell parameters. In principle, the heat capacity can be determined from the energy fluctuations in the isothermal ensembles, and such a computation has been included in the post-processing codes. However, the effect of varying the thermal equilibration time constant on the heat capacity has not been determined here so this quantity is not reported in the following.

While previous many-body potentials used to study NiAl have been the embedded-atom-method (EAM) potential developed by Voter and Chen[12] for Ni_3Al, we have employed an EAM potential developed specifically for NiAl by Rao, Woodward and Parthasarathy.[5,13] The cut-off range of the potential is 5.5626 Å. The determined zero-temperature elastic constants and lattice parameter for this potential are given in Table 1.

For most simulations of perfect crystal NiAl, LATTICE was used to generate a structure with 128 atoms in a symmetrical, cubic simulation cell oriented with the axes x,y,z parallel to [100], [010] and [001], respectively. The B2 structure can be considered as two simple-cubic sublattices, one of Ni and the other of Al, displaced with respect to each other by $\sqrt{3}/2a$ along the

cubic unit cell diagonal, where a is the lattice constant. Thus the computational cell consists of for B2 unit cells on each edge, with one Ni and one Al atom in each cubic unit cell. Given the zero-temperature lattice parameter and the cut-off of the potential, this geometry satisfies the minimum image convention[10] and ensures that no atom interacts with any of its periodic images. Three-dimensional periodic boundary conditions have been used throughout.

Simulation Results

1. Energy Conservation

Initial MD simulations were carried out with fixed volume and without temperature control (microcanonical ensemble). Such simulations are useful for checking the simulation code, and for determining a reasonable simulation time step. One desires to use as large a time step as possible to allow for simulations over longer times, however, too large a time step can cause integration errors which manifest themselves in non-conservation of the energy. After slightly modifying the EAM potentials for NiAl to be consistent with the way DYNAMO handles different force cut-offs, a time step of 0.001 psec resulted in energy conservation to one part in 10^6. A smaller time step was even better, while a time step of 0.002 psec resulted in energy conserved to only one part in 10^5. As a result, an integration time step of 0.001 psec was chosen for all simulations.

2. Autocorrelation

Since the data generated in the course of an MD simulation is highly correlated from one time step to the next, it is important to estimate the correlation time of quantities of interest in order to calculate accurate standard deviations of the mean.[10] For example, if n measurements are made of the pressure in a simulation and the standard deviation is found to be σ, then the standard deviation of the mean pressure is computed as $\frac{\sigma}{\sqrt{n}}$, provided that the n measurements are statistically independent. If only every m-th measurement can actually be considered statistically

independent, then the standard deviation of the mean should be computed as $\frac{\sigma}{\sqrt{n}}\sqrt{m}$. A convenient method of estimating the correlation time for a quantity Ω is through its autocorrelation function, which is defined as

$$A_\Omega(\tau) = \frac{<\Omega(t+\tau)\Omega(t)> - <\Omega>^2}{<\Omega^2> - <\Omega>^2}, \tag{32-1}$$

where the thermodynamic averages $<\Omega>$ are given by

$$<\Omega> = \frac{1}{n}\sum_{t=1}^{n}\Omega(t), \tag{32-2}$$

t is the time step, and is τ the time shift between measurements being compared. As $\tau \rightarrow 0$, $\Omega(t+\tau)$ and $\Omega(t)$ become perfectly correlated. In this case, $<\Omega(t+\tau)\Omega(t)> \rightarrow <\Omega>^2$, and $A_\Omega(\tau)$ approaches 1. On the other hand, as τ increases, $\Omega(t+\tau)$ and $\Omega(t)$ become uncorrelated, $<\Omega(t+\tau)\Omega(t)> \rightarrow <\Omega>^2$, and $A_\Omega(\tau)$ approaches 0.

For B2 NiAl equilibrated at 400°K, the autocorrelation function of the pressure, $A_\Omega(\tau)$, is shown in Fig. 1. Autocorrelations of the stress behave similarly. As each data point on the curve represents 0.02 psec, the correlation time for NiAl is estimated to be 0.05 psec.

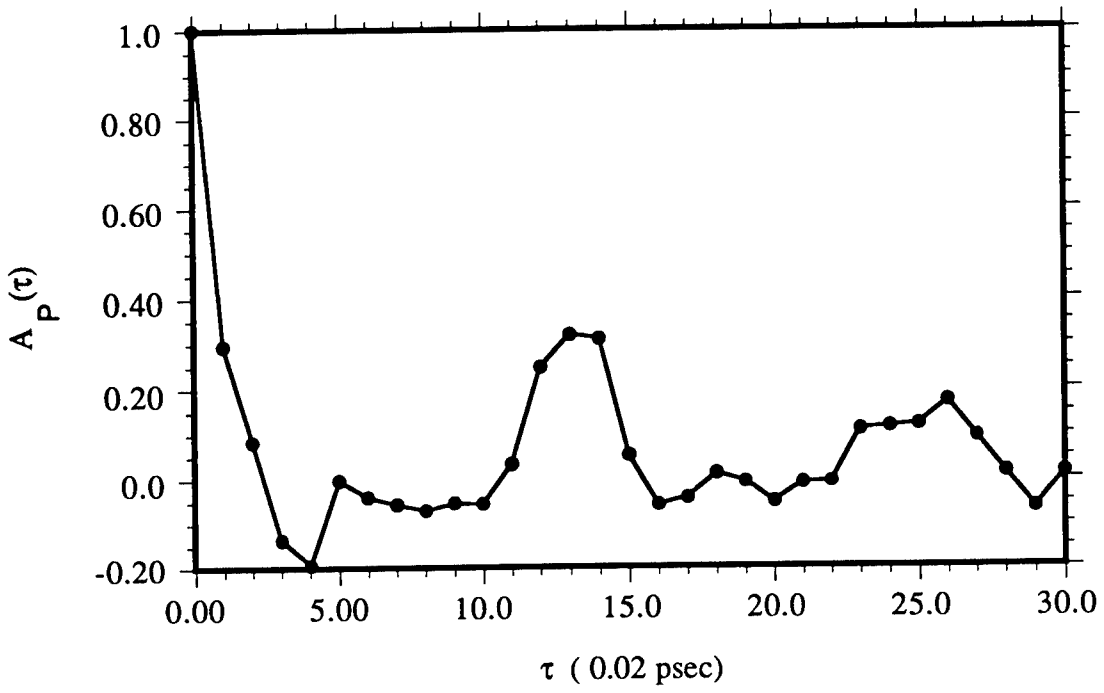


Fig. 1. Autocorrelation function of the pressure $A_P(\tau)$ as a function of time separation τ for NiAl at 400°K, averaged over a 20 psec simulation with 128 atoms.

3. Thermal Expansion

To determine the lattice parameter corresponding to zero stress for NiAl at 400°K, simulations were performed in the isothermal-isostress ensemble. Simulations on the order of 6 psec determined an average lattice parameter a=2.90943 Å. This lattice parameter, in a 20 psec fixed-volume simulation, determined the average residual stresses shown in Table 1. These residual stresses and their fluctuations set the scale for determining elastic constants by stress-strain simulations. The resulting linear coefficient of thermal expansion is approximately 1.6 times larger than the experimental value[1] of 15.1×10^{-5} K^{-1}, and reflects a "thermal softness" of the potential.

Table 1. Elastic Constants, C_{ij}, Bulk Modulus, B, Lattice Parameter, a, and average residual stresses, σ^o, for the NiAl potential using 128 atoms in all cases except for the determination of C_{44}, which used 500 atoms.

	T = 0°K	T = 400°K
C_{11} (GPa)	190	144 ± 2
C_{12} (GPa)	126	100 ± 2
C_{44} (GPa)	129	105 ± 10
B (GPa)	147	118.7 ± 0.5
a (Å)	2.8815	2.90943
σ^o_{ii} (GPa)	9.89×10^{-5}	$(-2.6 \pm 2.1)\times10^{-4}$
σ^o_{ij} (GPa)	0	$(0.8 \pm 1.3)\times10^{-4}$

4. Elastic Constants

Several means are available to calculate elastic constants in MD and Monte Carlo Simulations.[14] While it is not uncommon to use the second-order strain dependence of the energy for zero-temperature elastic constants, the free energy is more difficult to determine and is therefore seldom used to determine non-zero temperature elastic constants. Fluctuation methods are efficient for the calculation of non-zero temperature elastic constants of crystals with a monatomic basis, but formalisms have generally not been worked out to correctly account for sublattice displacements and inhomogeneous strains in more complicated systems.[14,15] A simple and relatively straightforward means of determining elastic constants at non-zero temperatures is to investigate the first-order strain dependence of the virial stresses, which are already computed in DYNAMO. Relative to the cubic coordinate system, C_{11} and C_{12} are easily determined (see Table 1) from the generalized Hooke's law relation (in Voigt notation)

$$\sigma_i = C_{ij}\ \varepsilon_j \qquad (32\text{-}3)$$

by computing the stresses resulting from various magnitudes of an applied strain ε_1 (with all other strain components equal to zero). For example, Fig. 2 shows the dependence of the stress σ_1 on the strain ε_1, with a slope whose magnitude is C_{11}. Stresses were typically determined in 15 psec simulations after an initial 7-psec equilibration.

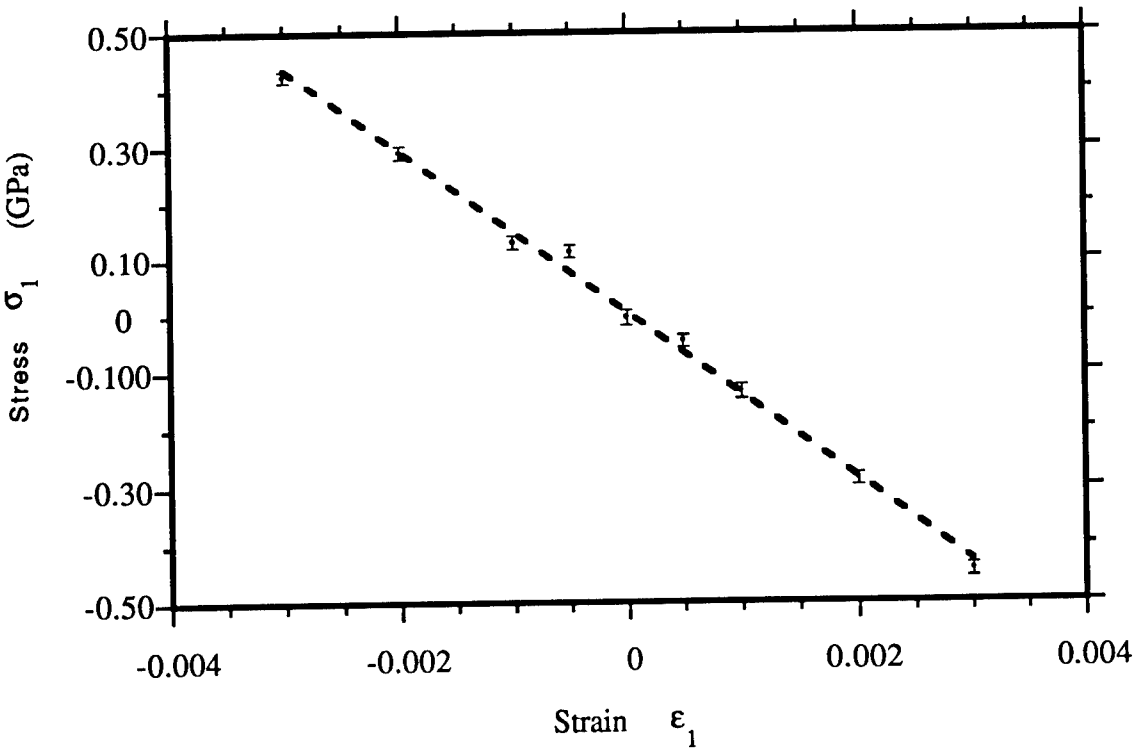


Fig. 2. Stress-strain curve to determine C_{11}. The magnitude of the slope from a least squares linear fit (dashed line) is $C_{11} = 144 \pm 2$ GPa. Error bars represent statistical standard deviations of the mean.

The determination of C_{44} is slightly more complicated, as DYNAMO can only perform simulations using a computational cell with orthogonal axes. By performing the following rotation on the NiAl

$$x \parallel \langle 100 \rangle \rightarrow x' \parallel \langle 110 \rangle \tag{32-4a}$$

$$y \parallel \langle 010 \rangle \rightarrow y' \parallel \langle 1\bar{1}0 \rangle \tag{32-4b}$$

$$z \parallel \langle 001 \rangle \rightarrow z' \parallel \langle 001 \rangle \ , \tag{32-4c}$$

the elastic constants in the cubic coordinate system (unprimed) can be related[16] to the elastic constants in the rotated system (primed) as

$$2C_{44} = C'_{11} - C'_{12} \, . \tag{32-5}$$

In the the rotated coordinate system, a small applied strain $\varepsilon'_1 = \varepsilon'_2 = \varepsilon$, $\varepsilon'_3 = 0$, gives two independent determinations of C_{44} via the relations

$$\sigma'_1 = (C'_{11} - C'_{12})\varepsilon = 2C_{44}\,\varepsilon, \tag{32-6a}$$

and

$$\sigma'_2 = (C'_{12} - C'_{11})\varepsilon = -2C_{44}\,\varepsilon, \tag{32-6b}$$

while $\sigma'_3 = 0$. In the rotated coordinate system, in order to satisfy minimum image conditions, a minimum computational cell size of 3×3×4 unit cells, or 144 atoms could be used. However, due to the breaking of the symmetry of the z' direction compared to the x' or y' directions in the rotated computational cell, finite-size effects were found to be significant. A computational cell of 5×5×5 unit cells, or 500 atoms, was found to yield satisfactorily small stresses for the previously determined lattice parameter, and was thus the cell size of choice in the determination of C_{44}. It was found that strains less than 0.0005 and simulations on the order of 100 psec for each strain, were necessary to obtain a satisfactory determination of C_{44} (Table 1). As can also be seen from

Table 1, it is not possible to tell the sign of C_{12}-C_{44} due to the rather large uncertainty in C_{44}. Additional simulations are underway to determine C_{44} more accurately.

The thermal softening of the elastic constants of metals is known to primarily be a result of the thermal expansion of the crystal.[17] The more rapid softening with temperature of the elastic constants for the NiAl potential, shown in Table 1, as compared with experimental determinations,[18] is related to and is consistent with the potential's larger thermal expansion. While this fact could be used to aid development a new potential for NiAl, we would propose using the same potential in initial MD studies of dislocations to facilitate a direct comparison with work already done at zero temperature.[5,8]

5. Pair Distribution Function

A useful characterization of structure and structural disorder in condensed systems is obtained from the pair distribution function, G(r), which gives the probability of finding a pair of atoms a distance r apart, relative to a completely random distribution at the same density.[10,17] With increasing temperature, the centers of the peaks can be seen to shift to larger separations due to the thermal expansion, and the widths of the peaks can be seen to broaden due to the increasing thermal disorder (related to the Debye-Waller factor). Shown in Fig. 3 are the partial pair distribution functions for B2 NiAl equilibrated at 400°K. The first peak, centered at $r=\sqrt{3}/2a$ represents the eight nearest Ni-Al neighbors in the B2 structure, where a is the lattice parameter at 400°K. The second peaks at r=a represent the six next-nearest neighbors, which are of type Ni-Ni and Al-Al, etc.

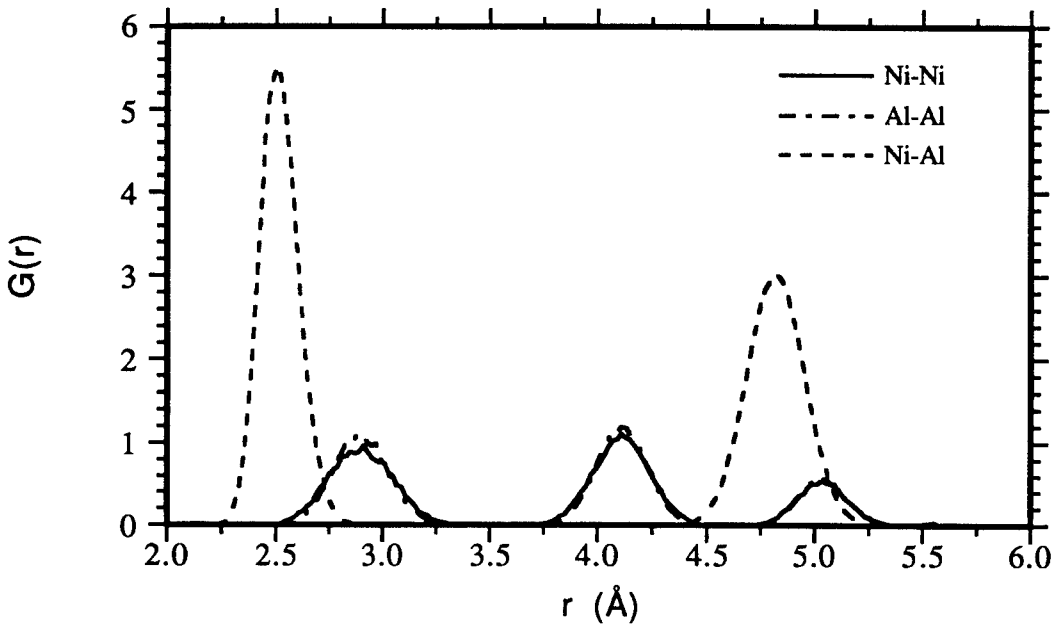


Fig. 3. Partial pair distribution functions for NiAl equilibrated at 400°K, and averaged over a 2-psec simulation.

Conclusions and Outlook

Molecular dynamics simulations have been carried successfully out for stoichiometric NiAl at 400°K in order to test and develop such capabilities. At this temperature, pair distribution functions, lattice parameters, residual stresses and elastic constants have been determined for a many-body EAM potential previously developed explicitly for NiAl. The EAM potential for NiAl was found to be rather thermally soft as compared to actual NiAl.

We propose to extend this work, and investigate the stress required to move a single straight dislocation at 400°K with realistic elastic boundary conditions. These molecular dynamics simulations can be compared with similar zero temperature simulations using the same potential.[5] At the same time, the database of NiAl thermal properties can be refined and expanded. In order to efficiently track the motion of the dislocation in the course of the simulation, work has begun to

compute and monitor the average position of near-neighbor mis-coordinated atoms, which are located only at the dislocation cores.[19] Stress and mis-coordination calculations will have to be modified for the dislocation simulations which are carried out in a cylindrical simulation cell with a boundary of "frozen" atoms at the curved surface and periodic boundaries at the flat surface. Initial simulations of an <001>{110} screw dislocation will be carried out on a small system with approximately 2000 atoms. More realistic simulations satisfying the minimum image condition in the periodic direction (parallel to the dislocation line) will be conducted next for the same dislocation using up to 8000 atoms. Successful simulations of this dislocation will lead to simulations of other dislocations and other temperatures, and will set the stage for the non-zero temperature simulation of kinks.

Acknowledgements

Collaboration and discussions with C. Woodward, S. I. Rao and D. M Dimiduk are gratefully acknowledged. This work was supported by the AFOSR Summer Faculty Research Program.

References

[1]For a recent review, see D. B. Miracle, *Acta Metall. Mater.* **41** (1993) 649.

[2]M. H. Yoo and B. T. M. Loh, (Oak Ridge National Laboratory unpublished report) 1971.

[3]A. N. Stroh, *Phil. Mag.* **3** (1958) 625.

[4]J. P. Hirth and J. Lothe, *Theory of Dislocations,* 2nd ed. (Wiley, New York, 1982).

[5]T. A. Parthasarathy, S. I. Rao and D. M. Dimiduk, *Phil. Mag. A.* **67** (1993) 643.

[6]M. S. Duesbery and G. Y. Richardson, *Solid State Mater. Sci.* **17**(1), (1993) 1.

[7]A. Moncevicz, P. C. Clapp and J. A. Rifkin, *MRS Symp. Proc.* **209** (1991) 213.

[8]M. S. Duesbery, *Acta Metall.* **31** (1983) 1747; M. S. Duesbery, *Acta Metall.* **31** (1983) 1759.

[9]T. A. Parthasarathy, D. M. Dimiduk and G. Saada, *MRS Symp. Proc.* **288** (1992) 311.

[10]M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford, New York, 1987).

[11]H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.* **81** (1984) 3684.

[12]A. F. Voter and S. P. Chen, *MRS Symp. Proc.* **82** (1987) 175.

[13]S. I. Rao, C. Woodward and T. A. Parthasarathy, *MRS Symp. Proc.* **213** (1991) 125.

[14]J. M. Rickman and J. A. Jaszczak, *Phys. Rev. B.* (1991)

[15]J. Ray, *Comput. Phys. Rep.* **8** (1988) 111.

[16]See Ref. 4. pp. 430-435.

[17]J. A. Jaszczak and D. Wolf, *Phys. Rev. B.* **46** (1992) 2473.

[18]N. Rusović and H. Warlimont, *Phys. Stat. Sol.* **44** (1977) 609.

[19]D. Wolf and J. A. Jaszczak, In *Materials Interfaces: Atomic-Level Structure and Properties.* (Chapman and Hall, London, 1992) p. 662.

APPENDIX: Summary of Data Files and Post-Processing Codes

Several Fortran codes along with Unix command files have been developed to simplify the analysis of data generated in the course of the MD simulations. The following output-data files are generated in DYNAMO, and typically have a number appended to the end, here taken as 1 for example, for the purpose of processing and bookkeeping:

gr.out1	Output data file of partial pair distribution functions and integrated numbers of pairs, versus pair separation.
len.stl1	Lists the instantaneous computational cell parameters versus time for isostress simulations.
md.stl1	Lists the instantaneous kinetic energy, potential energy, total energy, temperature and pressure with time.
mold.out1	Standard output containing general simulation parameters.
stress.stl1	Lists the instantaneous components of the stress tensor with time.

The following command files look for the appropriate data files, appended with a specified number as above, to process with the like-named Fortran codes:

auto.com 1	Computes the autocorrelation function (see pp. 192-195 of Ref. 10) for any column of data in md.stl1. This gives a measure of simulation time scale for sampling independent data, and is used to correct the standard deviations of the mean computed in other post-processing codes.
post.com 1	Computes averages and standard deviations of data in md.stl1. The heat capacity is also computed, as appropriate for simulations with fixed temperature. The effect of the thermal equilibration time parameter on the heat capacity has not been investigated. Code can generate smaller files convenient for plotting of optionally binned data.

stress.com 1 Computes the average and standard deviations of the six stress components in stress.stl1. Code can generate smaller files convenient for plotting of optionally binned data.

Other utilities:

polfit.com f.dat Searches for the file f.dat and performs least-squares polynomial fits to the data. The order of the polynomial and the type of statistical weighting of the data is determined interactively by the user.

Files under development for mis-coordination computation and dislocation tracking:

B2NNCUT Input data file giving structural information such as lattice parameter, nearest-neighbor coordination cut-off distances and ideal nearest-neighbor coordinations for each pair type. One means of obtaining reasonable cut-offs is from the pair distribution functions.

nncoord.out1 Output data file showing average position of mis-coordinated atoms and actual atom coordinations and mis-coordinations for each type of nearest-neighbor pairs (based on B2NNCUT).