AD_____

Award Number: DAMD17-98-2-8003

TITLE: Massachusetts Institute of Technology Consortium Agreement

PRINCIPAL INVESTIGATOR: Haruhiko H. Asada, Ph.D.

CONTRACTING ORGANIZATION: Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

REPORT DATE: March 1999

TYPE OF REPORT: Final I of Phase 2

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1999	Final (31 Dec 97 - 31 Dec 98)

4. TITLE AND SUBTITLE
Massachusetts Institute of Technology Consortium Agreement

5. FUNDING NUMBERS
DAMD17-98-2-8003

6. AUTHOR(S)
Haruhiko H. Asada, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

E-Mail:

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

14. SUBJECT TERMS

15. NUMBER OF PAGES
333

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	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

H. Asada

PI - Signature Date

Introduction

We at the Brit and Alex d'Arbeloff Laboratory for Information Systems and Technology at the Massachusetts Institute of Technology are pleased to deliver to our sponsors the First Progress Report of Phase 2 of the Home Automation and Healthcare Consortium. This report describes all major research accomplishments within the last six months since we launched the second phase of the consortium. It contains new findings, concepts, implementation, and experiments in diverse fields of home automation and healthcare research, ranging from human physiological modeling, patient monitoring, and diagnosis to new sensors and actuators, physical aids, and human-machine interface.

Several of these accomplishments have led to patentable ideas that have been filed as provisional patent applications. The MIT Technology Licensing Office will notify the sponsors of these provisional patent applications. Should you be interested in pursuing the possibility of using the technologies, please reply to the Technology Licensing Office within six months after signing a non-disclosure agreement.

This summer our faculty and staff members participating in the Consortium will visit the sponsors to report their work on the consortium project and related topics. We will notify you as soon as our schedule is determined. We welcome the sponsors to visit MIT to see the latest results of consortium research. Laboratory tours and demonstrations will be arranged as well as appointments to meet our faculty and staff members.

For your convenience, the contents of this report will be posted at the web site of the MIT d'Arbeloff Laboratory. The Consortium sponsors alone can access the technical part of the web site by typing a keyword, which will be notified shortly. Please direct your colleagues and technical staff to the Consortium web site, should they be interested in our project.

We appreciate your sponsorship that enables us to conduct these exciting projects in the diverse fields of home automation and healthcare. We look forward to meeting you soon.

H. Harry Asada
Principal Investigator
Ford Professor and Director
d'Arbeloff Laboratory

Table of Contents

10. Nickel-Titanium Shape Memory Alloy Actuators for Home Automation
S. Lafontaine, I. Hunter

11. Noninvasive Blood Glucose Quantitation using Spectroscopic-based Optical Technique
K. Youcef-Toumi, V. Saptari

12. Signal to Noise Enhancement for an Invisible Marking System Using an Infrared Activation System
H. Asada, R. Doubleday

Physical Aids

13. The Superchair: A Holonomic, Omnidirectional Wheelchair with a Variable Footprint Mechanism
H. Asada, M. Wada

14. Smart Mobility and Monitoring Aid: A Helping Hand for the Elderly
S. Dubowsky

15. Large Scale, Mechanical Surface Waves for Elastic Body Transport-The Hyperbed
H. Asada, J. Spano

16. Design and Prototyping of a Surface Wave Actuator Using Shape Memory Alloy Fibers
H. Asada, W. Finger

Human-Machine Interface

17. Human Machine-Interface and Interactive Control
Part 1: Instrumented Nails and Virtual Switch Panels
H. Asada, S. Mascaro, K-W. Chang

18. Human Machine-Interface and Interactive Control
Part 2: Human Machine Interactive Control Using Dual Petri Nets
H. Asada, S. Mascaro

Home Networking

19. A Modular, Minimum Complexity, High-Resolution and Low Cost Field Device Implementation for Home Automation and Healthcare
S. Martel, S. Lafontaine, I. Hunter

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Human Physiological Modeling

CHAPTER 1

Virtual Human Project

P. Hunter, I. Hunter

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Virtual Human Project

Peter Hunter[1] and Ian Hunter

§1 Introduction

Currently medical measurements such as electro-cardiograms (ECG), blood pressure, respiration rate, etc are stored in medical databases. Databases are not models and as such are unable to make quantitative predictions. One of the objectives of the virtual human project is to facilitate the trend away from medical databases to sophisticated dynamic medical/physiological models of humans. We anticipate that occupants within a home will have their personal virtual human model stored in the home. Eventually each person should carry around their own virtual human counterpart in their wearable (or implanted) computer. The virtual human model should be the place where all biophysical and healthcare related measured are stored or referenced. Differences between the real-time output of the virtual human model and the actual person should be used for signaling potential health problems.

The development of good virtual human models is dependent on three main areas:

1. **Instrumentation** to measure relevant physiological subsystem parameters: (the initial virtual human model needs to be parameterized for a specific individual via detailed measurements).
2. **System identification** techniques to represent the various physiological subsystem dynamics: (many of the subsystems are highly nonlinear: nonlinear system identification techniques are required).
3. **Continuum modeling** techniques to represent the three dimensional distribution of tissue properties (optical, thermal, electrical, mechanical, etc).

Areas 2 and 3 above are mathematical techniques that have been developed by very different engineering and mathematical disciplines. In a collaboration between our groups: largely system identification expertise (Ian Hunter) in Mechanical Engineering at MIT and largely continuum modeling expertise (Peter Hunter) in Engineering Science in the University of Auckland, we are bringing these two powerful mathematical techniques together to create powerful Virtual Human healthcare related models. In this report we focus on the continuum modeling framework which is largely centered on and utilizes the CMISS computer modeling package (mostly written by Peter Hunter). In the next report we will focus on the system identification techniques which are largely centered on modern derivatives of the NEXUS [1]computer language (written by Ian Hunter). A subsequent report will be devoted to the unique combination of these two mathematical modeling approaches.

Virtual Human Modeling

The virtual human model is designed to operate at multiple scales. At one end of the scale the model will provide a 'lumped parameter' description of various physiological subsystems for use in prediction and diagnostics. For example, a lumped parameter model of the

[1] Professor, Engineering Science, University of Auckland, New Zealand and visiting Professor, MIT

cardiovascular system would be developed for use in predicting heart rate and blood pressure changes in response to exercise. At another level, appropriate to the Physiome Project (see below), the virtual human model provides an anatomically accurate continuum mechanics description of the body in which subcellular properties are encapsulated in empirical 'constitutive laws'. At a still finer scale these constitutive laws are interpreted with anatomically detailed subcellular models which contain empirical descriptions of membrane ion channels and pumps and subcellular signaling pathways. Below this the ion channels and receptor binding sites will be modeled with molecular models and eventually link into the Genome Project.

An important benefit of this hierarchical approach is to allow functionally accurate modeling at a coarse scale using system identification and parameter estimation based on the next finer scale down (illustrated in Fig.1). For example, a simple lumped parameter model of the heart for use in heart rate and blood pressure predictions can best be established by modeling the coupled electromechanical-biochemical activity of the heart with continuum mechanics models based on finite element techniques (see later). The simple lumped parameter model is obtained by performing system identification and parameter estimation on the continuum model. The hierarchical approach also makes it possible to predict the changes in system parameters resulting from specific diseases since the parameters are interpreted in terms of underlying anatomically based models.

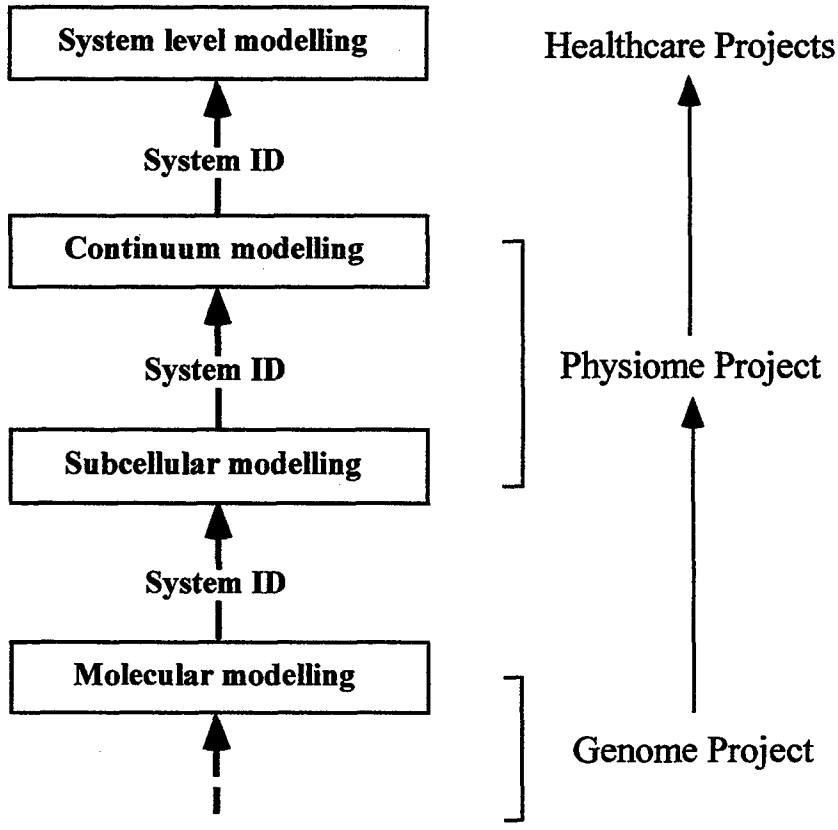


Figure 1. The relationship between the different modeling levels. Each level uses an approximation of the properties determined from the level below.

Another advantage of this hierarchical approach is that it brings into one rational framework a number of research disciplines operating at grossly different scales: e.g. human system level

modeling, the Physiome Project and the Genome Project. The Genome Project (http://www.genome.com) to map out the human genome (i.e. identify the entire sequence of base pairs in human DNA) is nearing completion - about one third of the 100,000 genes have now been sequenced and the remainder will be identified within about three years. This information has already had a major impact on medical science by identifying the genes that code for critical proteins in certain genetic diseases. The next and much greater challenge is to understand how the code for these building blocks determines the function of various organs. The Physiome Project (http://www.physiome.com), an attempt to coordinate various models of organ function, is now being taken up by several international bodies such as the International Union of Physiological Sciences (IUPS). The University of Auckland heart/lung model developed by one of us (Peter Hunter) has been identified as one of a few key 'proof of concept' projects which will establish the utility of this modeling based approach to understanding human physiology and treating human diseases.

The next three sections describe the continuum level and sub-cellular level modeling software (CMISS) being developed in Auckland. An example of how this software has been used to derive the form and parameters for a lumped parameter model is given in Section 5. A brief outline of future developments is given in Section 6.

§2 CMISS overview

CMISS began life in the early 1970s as a finite element program for large deformation biomechanics problems, principally for stress analysis in the heart. It subsequently evolved into a general purpose biological systems modeling tool in the areas of (and hence the name): Continuum Mechanics, Image analysis, Signal processing and System identification. Over the last few years it has also been developed as a tool for creating virtual environments for medical and other applications.

CMISS consists of two main parts: the computational engine and the graphical user interface (GUI). A common mode of operation is to have the computational engine running on a fast parallel-processing server while the GUI runs on a local workstation. The relationship between the computational engine and the graphical interface is shown in Fig.2. The database containing anatomical and material property data for various organ models is linked into the main computational code and various application-specific GUIs are available from the general purpose CMISS graphical interface.

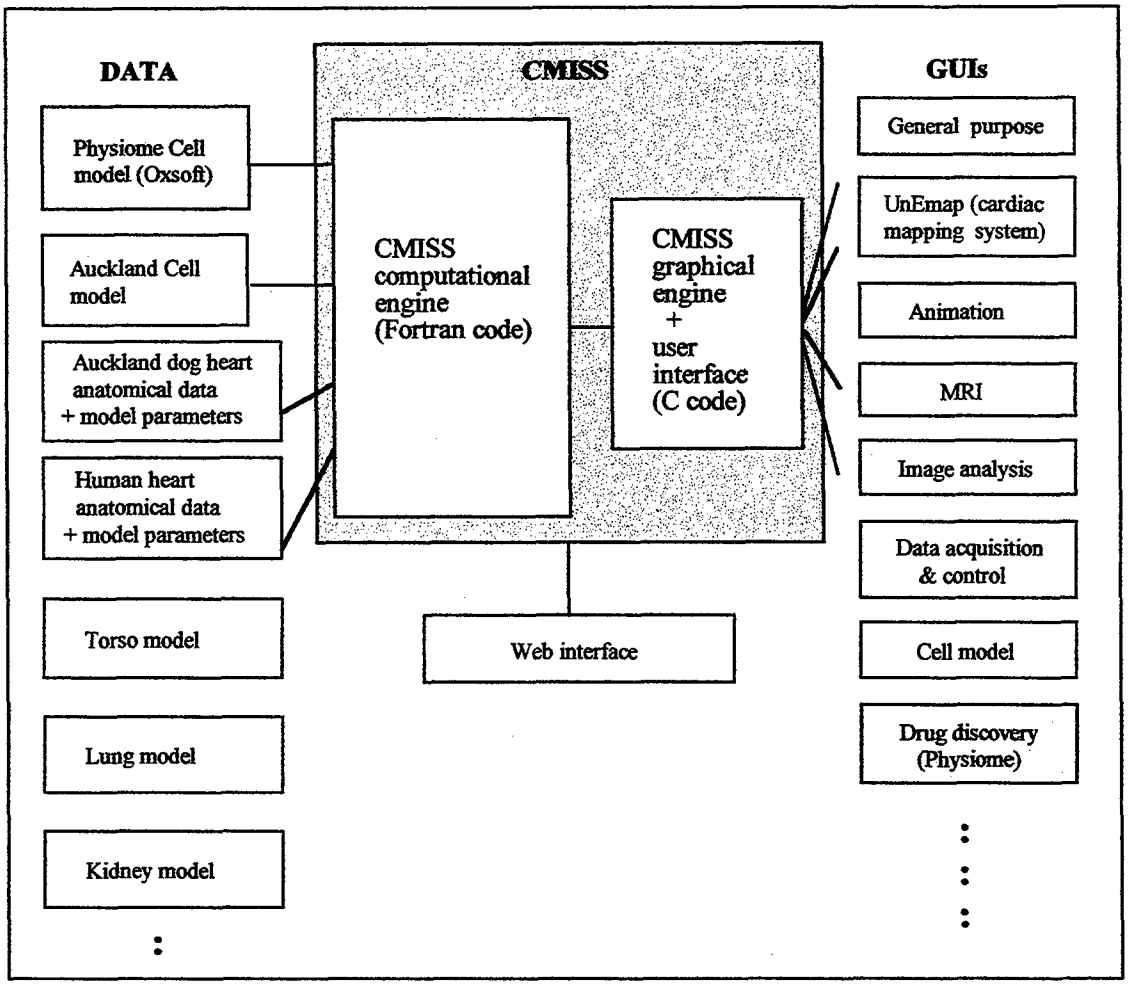


Figure 2. The relationship between CMISS, the cell models, the anatomical data bases and the application-specific graphical user interfaces.

§3 CMISS computational engine

This program encompasses a number of computational algorithms designed to handle the particular problems of modeling biological structures and systems. Major features are:

3.1 Multiple variables and equations

Biological problems involve multiple field variables governed by multiple systems of equations that interact with each other. For example, computing mechanical stress and deformation (three-dimensional displacement fields) during the heartbeat involves solving equations derived from physical conservation laws (conservation of mass and conservation of momentum). Closely coupled to this are several other field variables, such as the cell transmembrane potential (governed by the transmembrane ionic currents and pumps encapsulated within the diFrancesco-Noble equations and the associated diffusion processes); the extracellular potential (governed by extracellular diffusion); the oxygen partial pressure (governed by metabolic demand and coronary flow equations); flow amd pressure variables in the ventricles (governed by the Navier-Stokes equations). Examples of the interactions between these processes are: the

interaction between ventricular fluid pressure and wall stress; the direct effect that stress and deformation have on the ionic currents and hence the waves of activation (see Fig.2); the interaction between coronary flow, oxygen delivery and metabolic demand (both via mechanical work and activation processes).

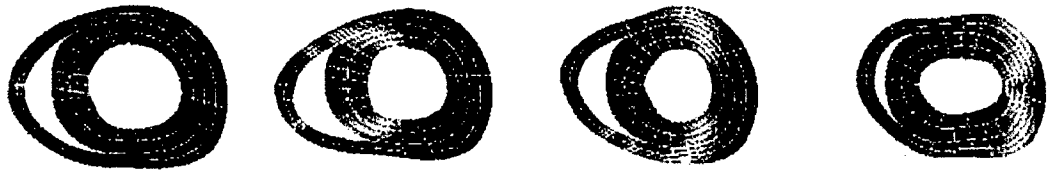


Figure 3. An example of a coupled problem being solved by CMISS. A cross-section of the electro-mechanical heart model is shown in several states of contraction as a wave of excitation propagates from initial stimulus sites on the right ventricle and septum.

CMISS is designed to handle any number of these interacting systems of equations representing different physical processes. The equations can be nonlinear, time-dependent and defined in one, two or three spatial dimensions in a number of alternative coordinate systems (see below).

3.2 Complex geometries

An important difference between biological structures and engineering structures is that biology, unlike engineering, relies on complex three-dimensional shapes which have no axes of symmetry to simplify the analysis. An ability to mathematically represent these complex structures efficiently is crucial to effective biological modeling. For example, the three-dimensional geometry of the heart could be modeled by linear or quadratic finite elements (the tools of traditional engineering analysis) but it turns out to be far more efficient to use fewer higher order elements.

CMISS is, in one respect, a biological CAD package. It contains a wide range of basis functions for dealing with complex biological geometries and sophisticated fitting techniques have been developed to generate the mathematical surface descriptions from geometric data (see below and Fig.3). The data can some from many alternative sources. For example, there are extensive facilities for handling geometric data from X-ray and MRI images. Another special feature of CMISS is the ability to use various coordinate systems (rectangular Cartesian, cylindrical polar, spherical polar, prolate spheroidal, oblate spheroidal). This can often greatly simplifying the task of modeling a particular organ (e.g. prolate spheriodal for the heart, spherical polar for the skull). A major feature of the program and the main reason why it can be used in conjunction with virtual environment graphics is the availability of C^1-continuous elements. These elements employ cubic Hermite basis functions that give slope continuity for both the geometric and solution variables. Thus geometrically complex shapes can be modeled efficiently in a manner which preserves their visual appearance and also allows them to be used in mathematical models based on, for example, the equations of motion. To the best of our knowledge, this feature is not available on any other commercial finite element program.

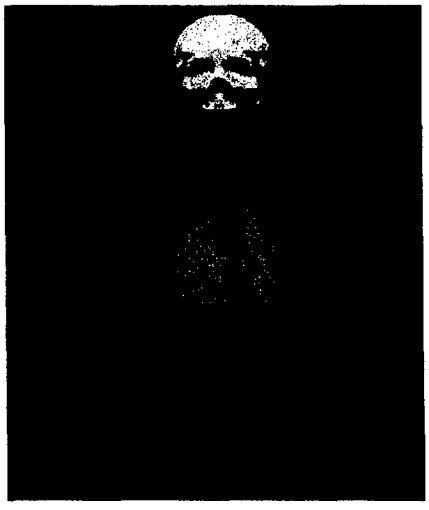
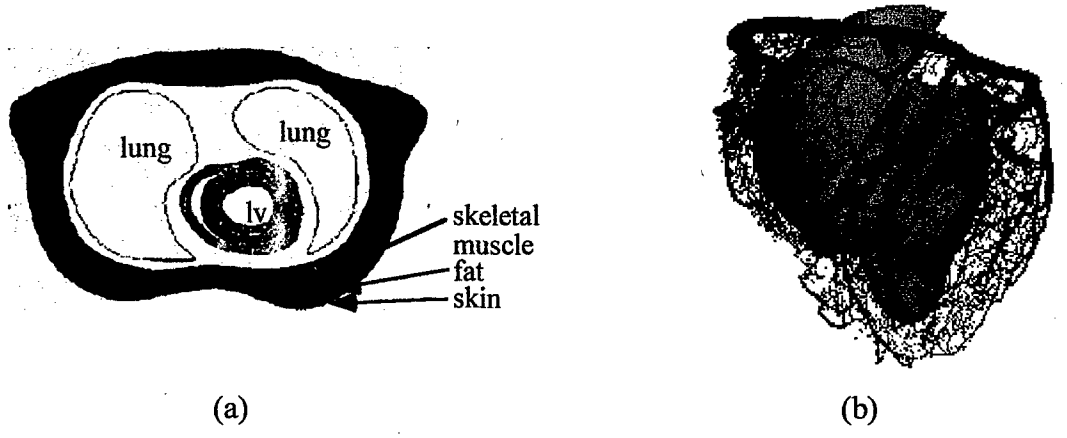


Figure 4. An example of the efficient handling of anatomical structures: The torso/head model shown here contains the skull, the heart, the lungs, and the layers of skeletal muscle, fat and skin, all modeled to sub-millimeter accuracy from MRI and X-ray data.

2.3 Multiple domains

There are many areas in biological modeling where multiple coupled equations need to be solved on multiple spatial regions (domains). For example, current flows from the interstitial myocardial domain of the heart to the surrounding torso and thus gives rise to the potentials picked up by ECG electrodes. The forward problem of electro-cardiology is to predict the distribution of potential on the body surface generated by current sources in the heart arising from the electrical activation of cardiac muscle. Solution of the inverse problem - estimating the electrical events in the heart from measurements of body surface potentials - is used clinically to diagnose conduction abnormalities. To solve the forward and inverse problems of electro-cardiography requires that all components of the torso be modeled - both their anatomy (geometry and anisotropic structure - see Fig.4) and the appropriate governing equations (which are different for the heart from the rest of the torso - see Fig.5 (a)). Another example is solving the large deformation elasticity equations on the heart myocardial domain and the Navier-Stokes equations of fluid flow in the ventricular cavity domain. Adding the solution of flow and oxygen transport in the coronary vessels constitutes another separate physical domain with its own set of governing equations (see Fig.5 (b)).

CMISS is designed to handle any number of these general coupled equations acting on separate spatial domains. The solution of such coupled problems requires the use of a variety of computational techniques because each type of equation has its own most appropriate technique - the Galerkin finite element method is ideal for solving the equilibrium equations of myocardial mechanics; finite difference collocation is more efficient in dealing with the fine spatial scale of current flow in the myocardium; and the boundary element method is best suited to modeling the complex anatomy of the torso.



(a) (b)

Figure 5. Examples of problems with multiple domains: (a) The electro-mechanical heart model is coupled into a model of the torso (shown here in cross-section) to predict the distribution of electrical potentials on the skin; (b) Coronary flow and pressure and oxygen transport are computed in a model of the coronary circulation, coupled to the electro-mechanical heart model.

3.4 Anisotropic structures

All biological structures are anisotropic (i.e. have material properties which are different in different material directions). This anisotropy is often closely coupled to the underlying geometry. CMISS has the ability to represent material anisotropy in relation to the description of the underlying geometry by means of fiber direction fields with appropriately chosen basis functions. For example, the fibrous-sheet structure of the heart (see Fig.6) is represented by spatially varying geometric angles defined with respect to the geometric material coordinates so that as the heart deforms, the correct fiber angles are preserved. For the muscle fiber angle the basis functions are bilinear in the plane of the heart wall but cubic through the wall - to reflect the dominant direction of anisotropy.

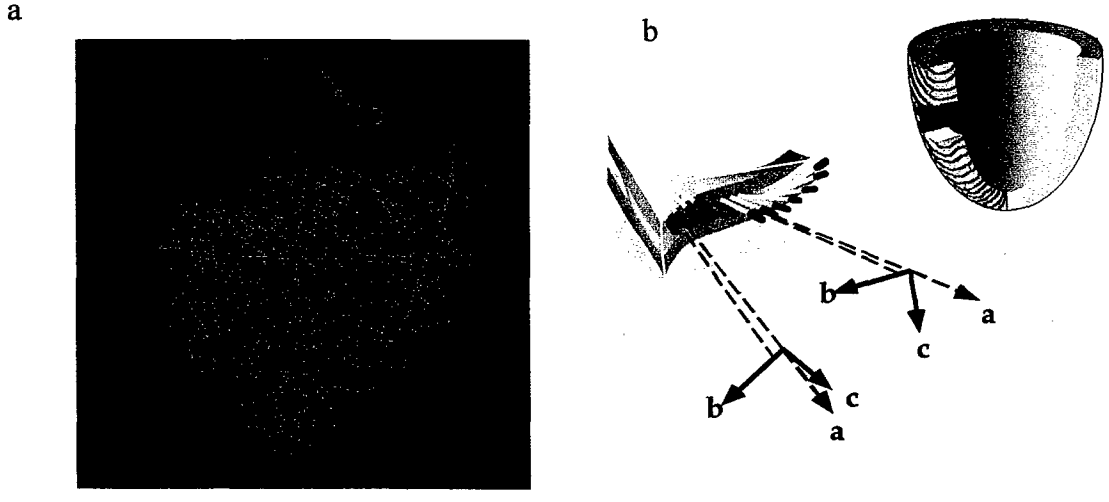


Figure 6. The fibrous-sheet structure of the heart. The coordinate systems for the solution of both the mechanics and the electrical activation of the heart are based on

these spatially varying coordinates. (a) The fiber field shown on the epicardium. (b) The orthotropic material axes defined in relation to the fibrous-sheet structure of the heart.

3.5 Nonlinear material properties

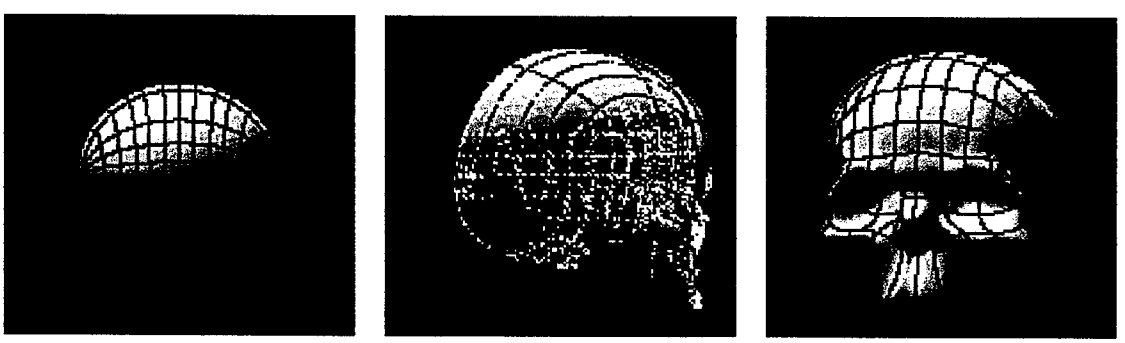
Biological models, unlike most engineering models, almost always require the use of nonlinear material property laws. The passive elastic properties of the heart, for example, are highly nonlinear as well as being anisotropic. These constitutive laws are also very different from any engineering material laws. It is very important to formulate the material properties in a way that leads to efficient numerical computation since the material law is evaluated many times during the continuum model computations.

CMISS has both very general-purpose material property descriptions and ones specialized for particular materials such as the myocardium.

3.6 Least-squares data fitting and optimization

A common problem in biological modeling is fitting geometric or other anatomical data with a mathematical model. A considerable amount of anatomical detail is needed to model the complex three-dimensional geometry and fibrous structure of many biological systems.

Algorithms have been developed in CMISS for fitting various types of model including bicubic Hermite nodal parameters which preserve arclength derivatives. This can be either linear or nonlinear fitting and include various degrees of smoothing. An extensive anatomical database is being developed using finite element descriptions of geometry and structural anisotropy.



(a) (b) (c)

Figure 7. Least-squares fitting of skull surface data with a bicubic Hermite model of the skull. (a) The data points projected onto the surface of the initial spherical model; (b) the model and data projections part way through the fitting process; (c) the final fitted skull surface.

Another important aspect of biological analysis addressed by the program is the optimization of material, shape or other parameters of a model to minimize some specified objective

function, and the system identification and parameter estimation of distributed parameter systems.

§4 CMISS graphical user interface

This program is based on X/Motif and C code and includes command parsing and graphical display. The 3D graphics window has general rotation and zoom controls as well as stereo viewing, animation and video output. The following facilities are used for creating virtual environments:

4.1 CMISS command window and command files

Nearly all actions in CMISS can be controlled by commands entered in the center pane of the command window shown in Fig.8. Many actions can also be controlled by using 'point and click' in the graphical window (which then initiate the commands). Command files are accessed through the 'File' menu on the command window and these appear as illustrated in Fig.9.

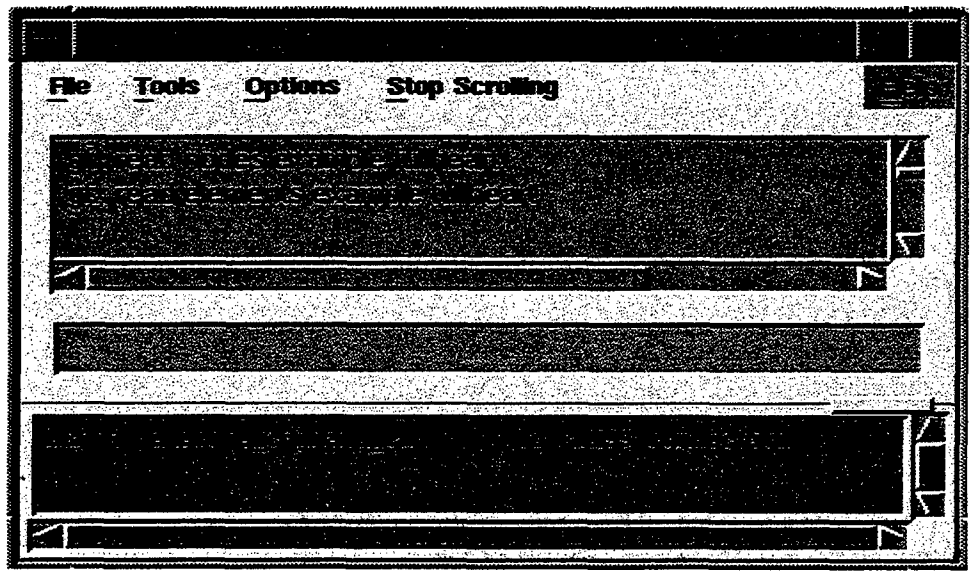


Figure 8. CMISS command window.

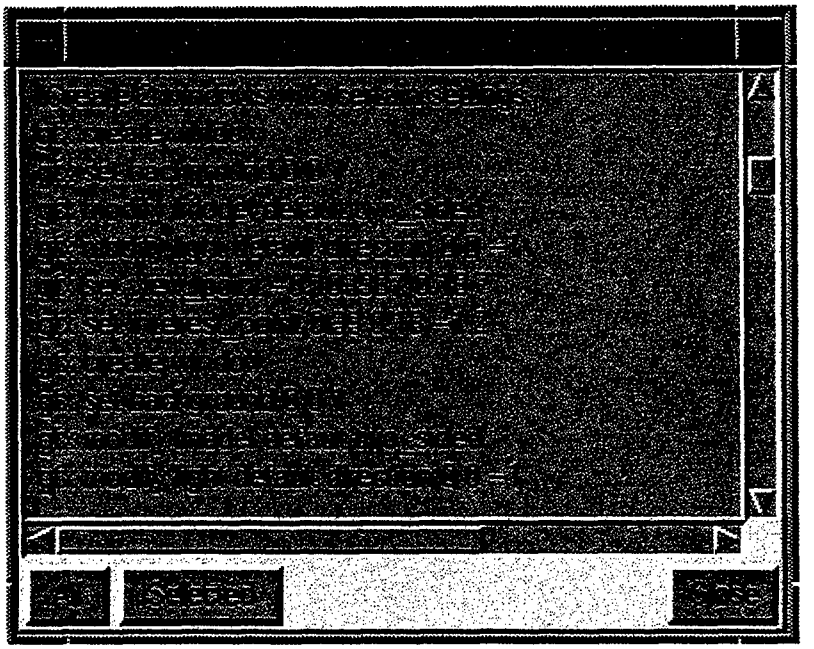


Figure 9. A CMISS command file.

The following graphical tools are currently available from the 'Tools' menu:

- 3D window
- Image processing
- Volume editor
- Graphical element attribute editor
- Graphical material editor
- Node editor
- Point editor
- 3D digitiser
- Input controller
- 2D projections
- UnEmap
- Cell Model

The first of these, a 3D graphical output window shown in Fig.10 and discussed in the next section, is used to display the finite element models and solution data. The second is a general purpose image processing environment and the third is a volume editor. The graphical element attribute editor and the material editor are discussed below. UnEmap is a signal processing environment for electrophysiological mapping studies. The Cell Model is a graphical environment for interacting with the parameters of the Auckland cardiac cell model.

4.2 3D graphics window

Graphical display of 3D models and the display of output from these models is a very important part of CMISS. Fig.10 shows a 3D heart model with solid shading of the endocardial surfaces and translucent shading of the epicaridal surface. The lines show the finite element boundaries.

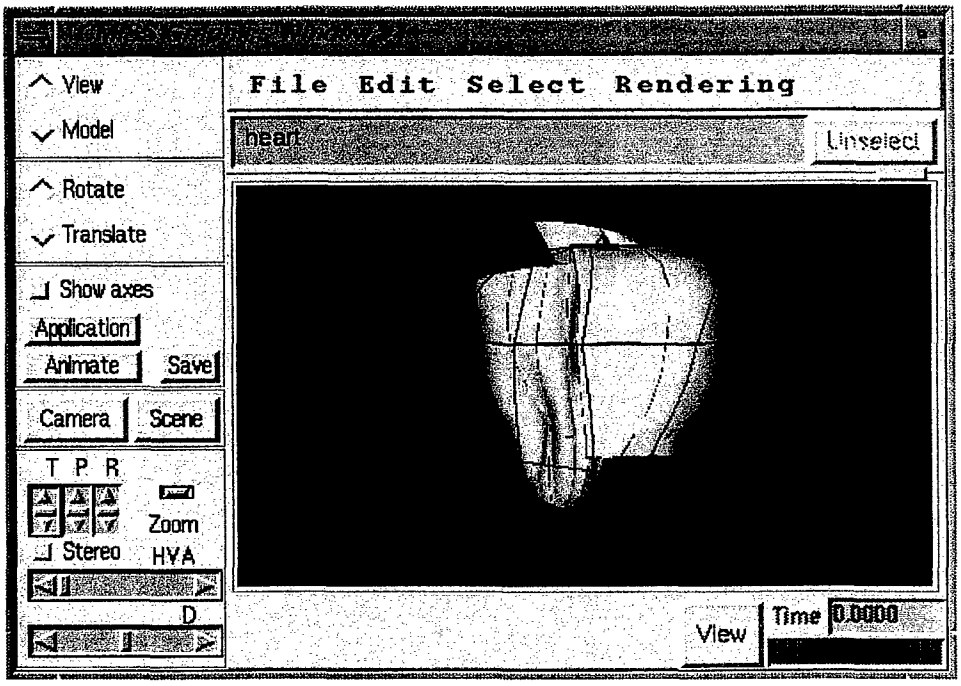


Figure 10. 3D graphics window showing heart model.

Some of the tools which have been developed to interact with the models are:

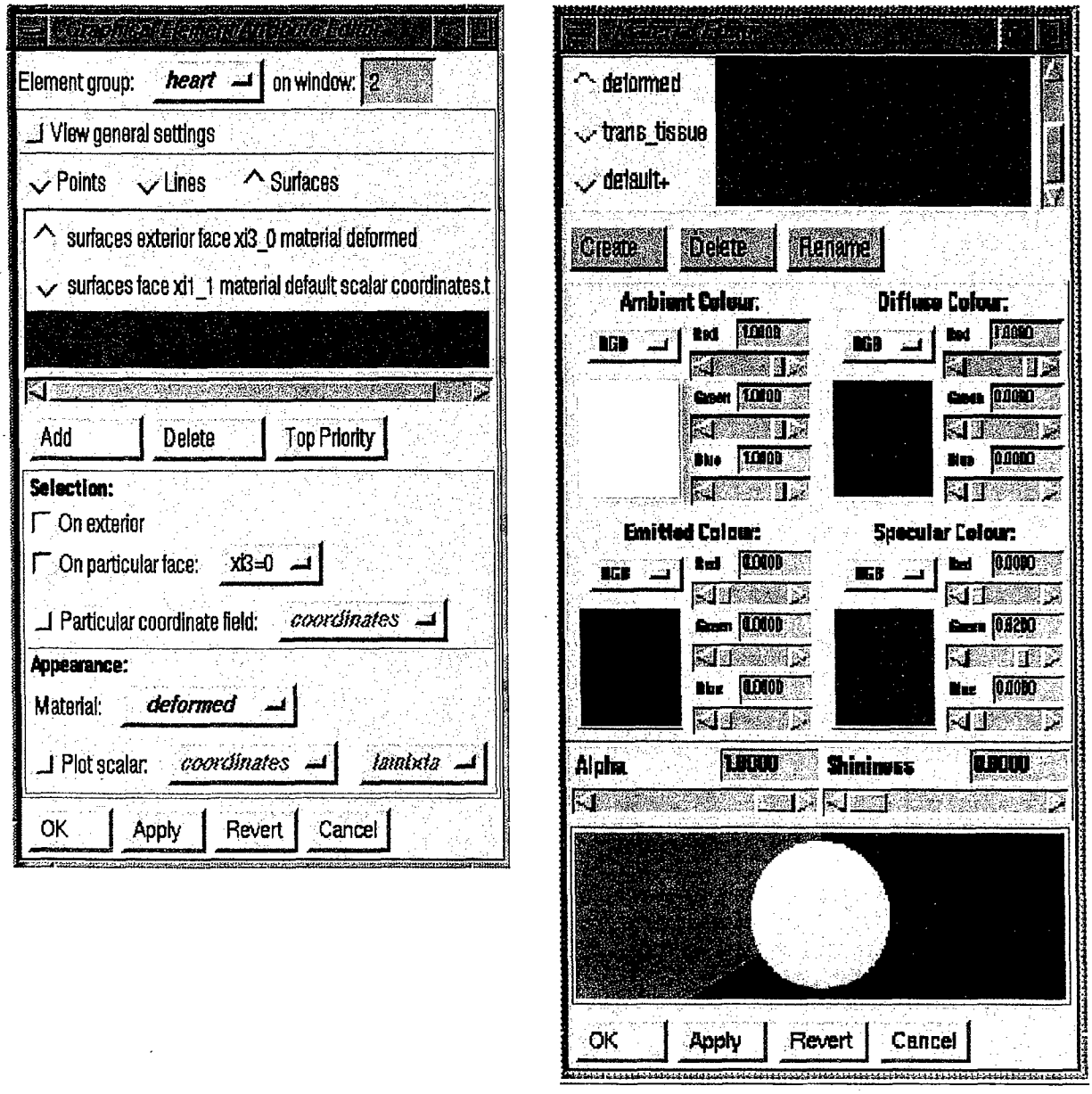
- **Texture editor**

permits complex 2D and 3D textures to be created by various methods including inserting textiles/voxels of specified color, reflectance etc into a normalized texture cell which maps onto the surface or volume of specified objects. Examples are shown in Fig.11.

Figure 11. An example of adding 3D textures to: (a) the endocardial surface of the heart model, and (b) the epicardial surface (the atria and blood vessels are also added this way).

4.3 Graphical attributes editor and material editor

When the files defining the finite element models are read into CMISS, the models are displayed by default as wireframes. The graphical attributes of the models can then be specified with a graphical attributes editor and a material editor, as shown in Fig.13. For example, the surfaces may be rendered with specified colors, textures and light reflectance properties, or colored with a field variable such as electrical potential.



(a) (b)

Figure 13. (a) Graphical finite element attributes editor, (b) Material editor.

4.4 Auckland cell model

Another graphical interface accessible from the 'Tools' menu on the Command window is the Auckland cell model shown in Fig.14. This model is an implementation of various ionic current of cardiac cell electrophysiology models (Beeler-Reuter, diFrancesco-Noble, Luo-Rudy) coupled in with the models of troponin/tropomyosin kinetics and cross-bridge mechanics (Hunter-McCulloch-ter Keurs) and cardiac metabolism (Loiselle-Rouhard-Zahalak).

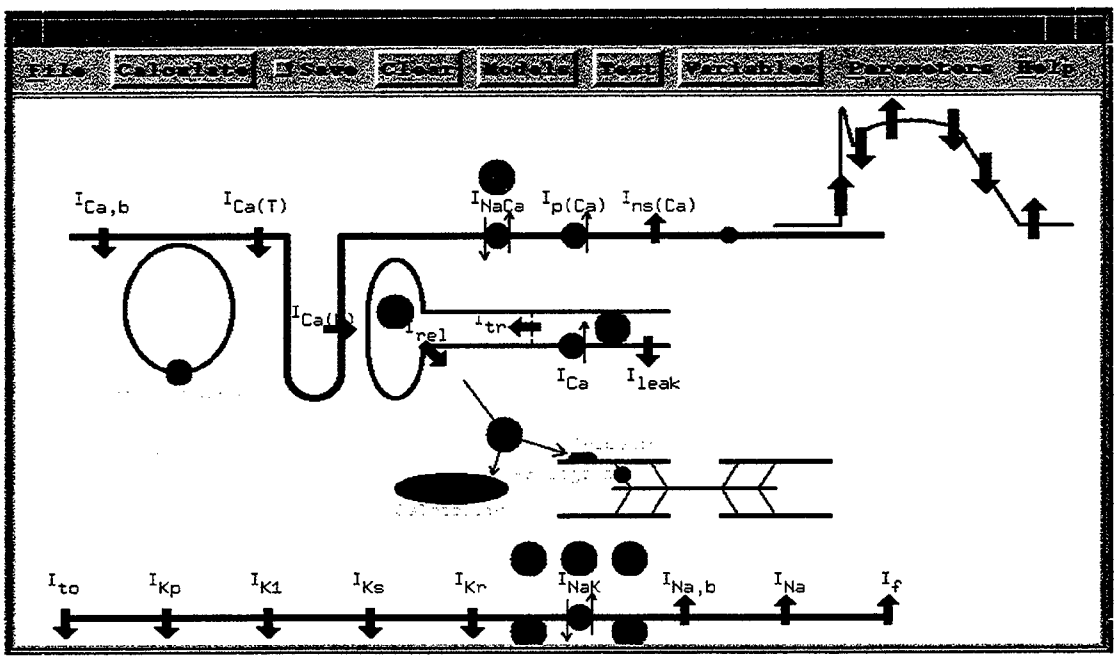


Figure 14. Auckland cell model.

§5 From continuum model to lumped parameter system

An example of the use of continuum level modeling to create a lumped parameter or black-box model is the acinus of the lung. Respiration in the human lung occurs at the level of the acinus, each of which contains 12 generations of branching airways. There are about 26,000 acinii at the ends of the conducting airways. To define suitable empirical relations for a black-box model of the acinus, and to identify its parameters, an anatomically accurate model of the 12-generation acinus is established and the advection-diffusion equations of gas transport in this continuum model are solved. By running the anatomically accurate acinus model through a range of boundary conditions a simple regression model is established which accurately mimics its behavior. Respiration in the whole lung is then modeled with an anatomically accurate model of the conducting airways (see Fig.17) in which the black-box acinar model is attached to each of the 26,000 endings of this model.

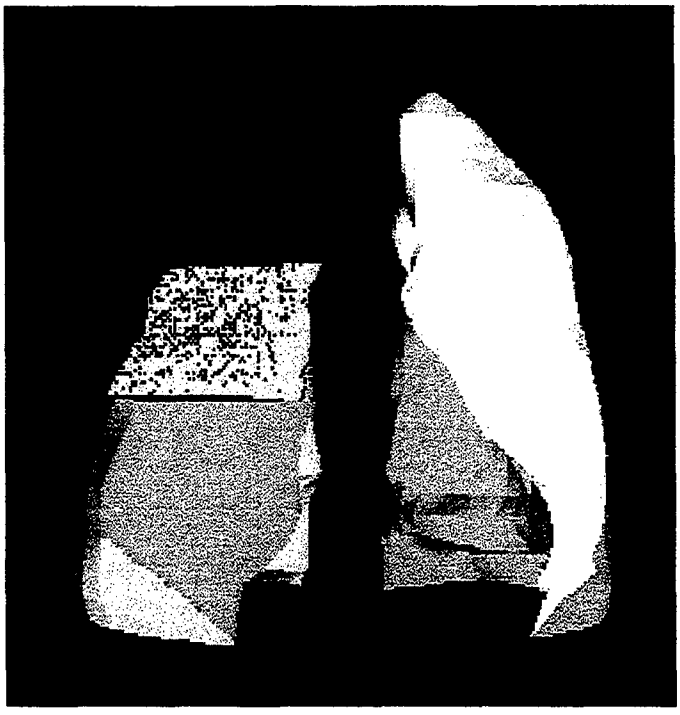


Figure 17. Finite element model of the lungs. The conducting airways are shown in one lobe.

§6 Future work

The current virtual human model includes anatomically accurate continuum models of:

1. the heart and lungs
2. the layers of skeletal muscle and fat around the torso
3. the skull and facial muscles
4. the hip bones and femur

An anatomically accurate model of the blood vessels of the thorax and the ribs and spine is currently under development. When these latter are complete we will have a reasonably comprehensive model of the human thorax for cardiovascular and respiratory studies. At that stage we will begin system identification studies with this model to derive suitable lumped parameter models for healthcare projects.

§7 References

Some recent publications which present the mathematical methods underlying the continuum analysis of soft tissue function are:

1. Hunter, P.J., Smaill, B.H.., Nielsen, P.M.F. and LeGrice, I.J. A mathematical model of cardiac anatomy. Chapter 6 in "Computational Biology of the Heart", Eds. A. Panfilov and A. Holden. John Wiley Series on Nonlinear Science. pp173-217. 1996.

2. Hunter, P.J., Nash, M.P. and Sands G.B. Computational electro-mechanics of the heart. Chapter 12 in "Computational Biology of the Heart", Eds. A. Panfilov and A. Holden. John Wiley Series on Nonlinear Science. pp347-409. 1996.

3. Costa, K.D., Hunter, P.J., Rogers, J.M., Guccione, J.M., Waldman, L.K. and McCulloch, A.D. A three-dimensional finite element method for large elastic deformations of ventricular myocardium: Part I - Cylindrical and spherical polar coordinates. *ASME J. Biomech. Eng.* 118:452-463, 1996.

5. Costa, K.D., Hunter, P.J., Wayne, J.S., Waldman, L.K., Guccione, J.M. and McCulloch, A.D. A three-dimensional finite element method for large elastic deformations of ventricular myocardium: Part II - Prolate spherical coordinates. *ASME J. Biomech. Eng.* 118:464-472, 1996.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Human Physiological Modeling

CHAPTER 2

Hemodynamic Modeling and State Estimation for Clinical Assessment of Cardiovascular Disorders

R. Kamm, Y. Huang

d'Arbeloff Laboratory for Information Systems and Technology
MIT

The Home Automation and Health Care Consortium

Phase 2

Progress Report - Hemodynamic Modeling and State Estimation for Clinical Assessment of Cardiovascular Disorders

Roger D. Kamm and Yaqi Huang

Introduction

Changes in the arterial pressure and flow pulses are a reflection of variations in the hemodynamic state of an individual. These changes are based on such factors as arterial geometry, peripheral resistance, and cardiac contractility. Thus, the potential exists to extract clinically useful data from analysis of the shape, magnitude, and timing of the pulse. The premise behind this study is that changes in the pressure pulse are indicative of changes in the hemodynamic state of the patient and that quantitative analysis of the pressure and/or flow traces can provide estimates of important hemodynamic parameters. A numerical model can provide the means for understanding the connection between the measured arterial pulse and the hemodynamic parameters. Such an algorithm may be easily integrated into standard hospital monitoring techniques, thus potentially reducing the need for traditional, more invasive, hemodynamic monitoring methods.

The work in this project will lead to a method and apparatus for continuous monitoring of the hemodynamic state of a subject, either at home or in the hospital. It makes use of recently developed computational methods for simulating the cardiovascular system with a realistic model, capable of predicting with considerable accuracy the time-varying pressure and flow traces throughout the arterial network. It also utilizes methods of parameter estimation based on comparison between measured pressure or flow traces and the corresponding traces produced by the computational model and stored in a database or "library". The method requires only a single measurement of the pressure or flow trace at one easily accessible location in the arterial system and has the potential for making detailed predictions of many patient parameters of clinical importance. Among these are peripheral resistance, arterial elastance, cardiac contractility, diastolic filling volume and cardiac output. The measurement can be made by using a sensing device worn by the subject as a wrist watch or ring with a signal sent by telemetry to a remote receiver for analysis. The result of the analysis could be a warning sent either to the subject or directly to the physician or hospital in the event that one or more parameter values deviate from normal range. Alternatively, the results could simply provide a means of observing normal variations in the hemodynamic state of the subject over time.

Background

There is increasing pressure on the medical profession to provide quality health care at low cost. There is also a compelling need to minimize the length of hospitalization, increasing the need for home health care monitoring equipment. Finally, as the population ages, the incidence of diseases of the elderly, heart disease being among the most prevalent, will continue to increase.

At the same time, computational methods and computational hardware continue to advance at rapid rates. Methods now exist to simulate the entire cardiovascular system with great detail. These methods have the capability of simulating subject-specific

behavior simply by unique specification of the many parameters of the model. Finally, noninvasive methods and devices for monitoring the time-varying pressure or flow traces at different locations in the arterial network are currently available and others are under development.

This sets the stage for the present approach which combines current computational methods and solutions obtained by high-speed workstations with readily obtained measurements on a particular subject. Recognizing that the detailed pattern of the pressure or flow trace represents a unique combination of the parameters characterizing the patient (suitably reduced to become tractable), it is possible to infer clinically-useful information from a comparison of a measurement to the solutions obtained from the computer simulation. This invention represents a method for doing so.

It is a primary purpose of this study to provide a computational model of sufficient realism and complexity to capture the behavior of the real cardiovascular system. Solutions obtained from this model using a high-speed computer in which a carefully selected subset of the entire parameter set that governs system behavior have been systematically varied are stored in a database or library for later use. Another important object is to represent the solutions obtained from this model in terms of a collection of "features" that capture the essential character of the computed pressure or flow traces. These features are expressed in mathematical form as functions (surrogates) of the several dominant parameters and used to assist in the parameter estimation procedure.

The estimation of cardiovascular parameters for a particular subject is based on a comparison of a measured flow or pressure trace from the individual to the many solutions contained in the library to identify the closest match. This can be accomplished in a variety of ways, but in the preferred method, is done by solution of the surrogate equations based on the features extracted from the measured trace.

The other critical aspect of this project is the development of a method for and device used in measuring the pressure or flow trace of an individual and either storing the trace for later analysis, analyzing the trace immediately and storing its features, or sending by telemetry the trace or features to a remote site for further analysis. Analysis consists of comparing the measured trace with the traces contained in the library, predicted by the computer model, to determine the one that best fits the measurement. This may be accomplished by extracting the features from the curve and using these features, along with the mathematical functions described above, to solve for the corresponding parameter values, or by direct comparison of the measured trace with the computed ones, using any of a number of pattern recognition methods to identify the best fit.

Work Completed

The Computational Model. The crux of this study is the development of a computational model, based on the following assumptions:

- Blood is approximated as an incompressible, Newtonian fluid.
- Blood flow in the aortic tree is approximately one-dimensional, justified by the unidirectional, primarily axial nature of blood flow in arteries (Pedley, 1980).
- The artery walls may be treated using a viscoelastic model.
- Viscous friction is approximated by considering the periodic behavior of wall shear, when appropriate.
- Curvature is ignored. The segments are assumed to be linearly tapered with respect to the cross-sectional area between bifurcation regions, and the angle of departure of a daughter branch from the main branch and the additional losses associated with the branched flow are taken into account.

•Flow into minor branches may be treated as a distributed leakage. This leakage is a function of the arterial-venous pressure gradient.

The analysis begins with consideration one dimensional flow in an elastic artery using the basic equations for momentum and continuity. allowing for frictional loss (including time-dependent effects), as well as distributed wall leakage along each arterial segment where the driving force for flow is the pressure drop between the local arterial pressure and the uniform venous pressure. A pressure-area relation or "tube law" is formulated to provide a third independent equation that includes viscoelastic wall damping. The set of hyperbolic, partial differential equations for the collection of arterial segments are solved using an adaptation of the MacCormack two step predictor-corrector method.

This model allows for linear segments that represent the larger vessels in the arterial tree, but to model the finer terminal branching structure in this manner is impractical. Rather, the terminal vessels are modeled as a lumped parameter Windkessel. The model is advantageous in that it allows the behavior of an entire arteriolar and small vessel bed to be captured using few parameters. It is also advantageous in that the phenomenon of peripheral wave reflections is well approximated.

Since our primary interest is in the behavior of the arterial system, the venous and pulmonary circulations are modeled as lumped parameters, rather than as a distributed branching model as employed on the arterial side. Note that the model contains unidirectional valves for entry and exit of blood into the right ventricle, which is also driven by a specified time-varying compliance. Venous inflow from the capillary beds is obtained by summing the total outflow from the arterial system; the flowrate to the left atrium is calculated from the left ventricular model outlined above.

The arterial tree is represented by a network consisting of 28 elements shown schematically in Figure 1. Each element is assigned a number, as is each bifurcation. The numbered elements correspond to specific major arteries. Values for the various parameters of the model were taken from the literature or estimated based on available knowledge as described in Ozawa (1996). Using these values, the validity of the model can be assessed by comparing simulation results to measurements from the literature. One such sample comparison is shown in Figure 2 between calculations using the standard case parameter values and measurements of Mills (1970). For other validation results, see Ozawa (1996).

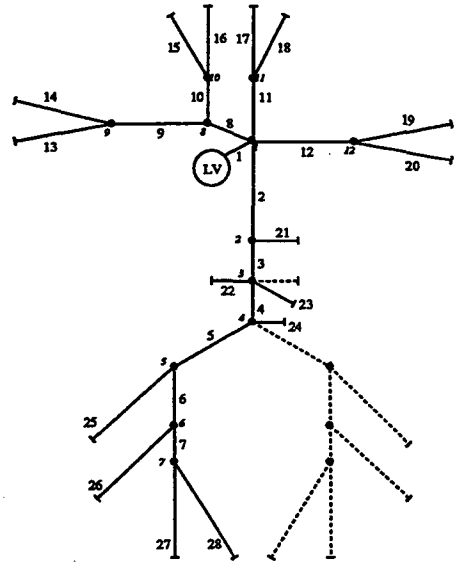


Fig. 1. The arterial network used in the simulatins.

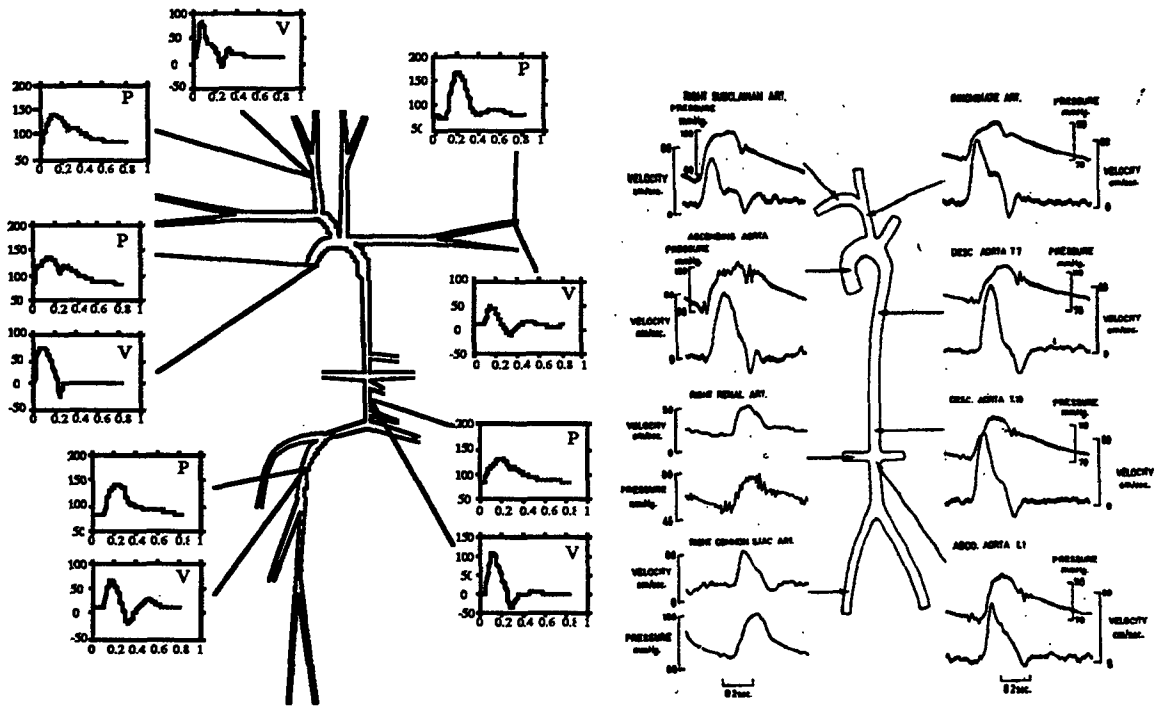


Fig. 2. Computed traces (left) of pressure (p) and velocity (v) compared to measurements of Mills (1970) (right).

Parameter estimation.. In the approach currently being pursued, the hemodynamic parameters are estimated by solution of the matrix equations expressing the functional relationship between model parameters and extracted features. These functional relationships are obtained by analysis of the solution library using locally linear or quadratic fits to the computational data. This matrix equation, if linear, can simply be inverted to obtained the parameter values corresponding to a given set of feature values. If it is deemed that a nonlinear representation is required, then other, more computationally intensive approaches will need to be invoked. Current studies have focused on the functional dependence of the features upon the parameters is well represented by low-order polynomials, suggesting that either a linear or quadratic representation will be adequate for parameter estimation. Techniques similar to those of Yesilyurt and Patera (1995) will be implemented and tested against computer generated "patient" data.

The parameters that have been determined to be most influential are: heart rate, systolic fraction, end diastolic filling volume, maximum left ventricular contractility, and systemic vascular resistance. Of these, end diastolic filling volume, ventricular contractility and systemic vascular resistance have clear clinical an diagnostic value. If a parameter space can be defined for N influential model parameters which affect the behavior of the arterial system, then an objective function which gives an indication of the error between the output of the model for a given parameter set and the actual patient data can be defined. The terms "model output" and "patient data" refer to the pressure and velocity tracings versus time as measured at various anatomical locations throughout the body for the numerical model and the *in-vivo* case, respectively. If a random sampling of the N-parameter space is performed, surrogates which help to describe the "error" as a function of the parameters can be constructed using the N-dimensional interpolation scheme. From the polynomial fit, the "best fit" can be located and the

corresponding parameters associated with the point of minimum "error" extracted. This procedure depends on the ability to generate enough points to obtain a representative sampling of the entire parameter space from which physiological behavior is expected to result. Additionally, it depends on the fact that a unique solution can be obtained, when several solutions may in fact exist. Another implication of sampling the parameter space is that as the number of parameters N increases, the number of required sample points increases exponentially (as does the time needed to accomplish all of the necessary computational runs).

It is technically possible to identify thousands of parameters associated with the numerical model. The prospect of attempting to determine exact behavior for a system with thousands of degrees of freedom, however, is an impossible task. Thus, in order to accomplish the present objectives, certain simplifying assumptions are made, based on self-similarity, parameter screening to identify the most critical parameters, dimensional analysis to express all the parameters in dimensionless form, a simplification of the model by elimination of all venous and pulmonary circulation elements, and other approximations, the number of critical parameters that exert the strongest influence on the solution is reduced significantly.

Using a grid discretized into four points along each axis, this results in a total of 4^6 or 4096 runs. Among these cases, many combinations produce results that are unrealistic in that the predicted cardiac output or mean arterial pressure is too low. To identify and eliminate these cases, a simple lumped parameter model of the circulation is used to make a rough prediction of the solution, from which a determination is made as to whether or not to run the full, non-linear simulation for that particular set of parameter values. This procedure reduced the number of simulations to be run from 4096 to 337.

Thus, unique sets of six parameters on a 4x4x4x4x4x4 grid were identified, where each axis was bounded by the maximum and minimum values selected for each individual parameter. Each axis was discretized into four nodes and each parameter assigned a dimensionless scale value of 0.0, 0.33, 0.66, and 1.0. The 337 gridpoints that fell within the physiological space were run and the data for the state variables as a function of time were stored for one steady-state cycle. The specified parameter set for each of the 337 runs was also stored for later processing.

The data from the 337 runs were used to construct the *parameter space library* from which surrogates were subsequently generated. Thus, the model is no longer required following construction of the library, unless in the future additional modifications are required to increase the accuracy of the estimation method, which is in turn a function of the accuracy of the simulations.

As mentioned above, many possible methods exist to characterize the shape of the curve so as to allow comparison to other similar curves, and any of these procedures would be appropriate in the present case. Such methods have been developed for a variety of applications.

In the preferred embodiment, the solution for the unknown parameters is treated as a series of six equations with six unknowns (six hemodynamic parameters). If one considers the complexity of the pressure pulse, the six independent characteristics or *features* of the pulse can conceivably be extracted and quantified. Such a process is termed *feature extraction* to distinguish it from parameter estimation. A feature may take the form of any value that can be used to quantify the pulse. One must therefore make the assumption that each pulse is specific for each given set of parameter values, and therefore that six features are adequate to *uniquely* specify one pulse. Thus, the six equations developed using the Shepard routine (described below) become the dependance of each of the six features on the set of six parameters.

The features employed in the present embodiment are those indices most often used by physicians. These include mean pressure, peak systolic dP/dt, the slope of the pressure upstroke during early systole, and the systolic ejection period, or the time during the cardiac cycle that the left ventricle is actively contracting. One can also envision

several more features that would be expected to change as a function of the hemodynamic parameters: the peak pressure, the pressure at end diastole, *dP/dt* during diastole, the time to peak pressure, the time to the dicrotic notch, and so forth. Additional features such as the amplitude and timing of secondary and reflected peaks may also be included, although a systematic measurement these features is difficult and unreliable, as not all pressure pulses measured at different times or different locations may contain such information. The velocity waves at various locations may also be analyzed in similar manner, although it is less clear what features may follow the alterations in parameter values. These features and the concept of "feature extraction" are summarized graphically in Figure 3.

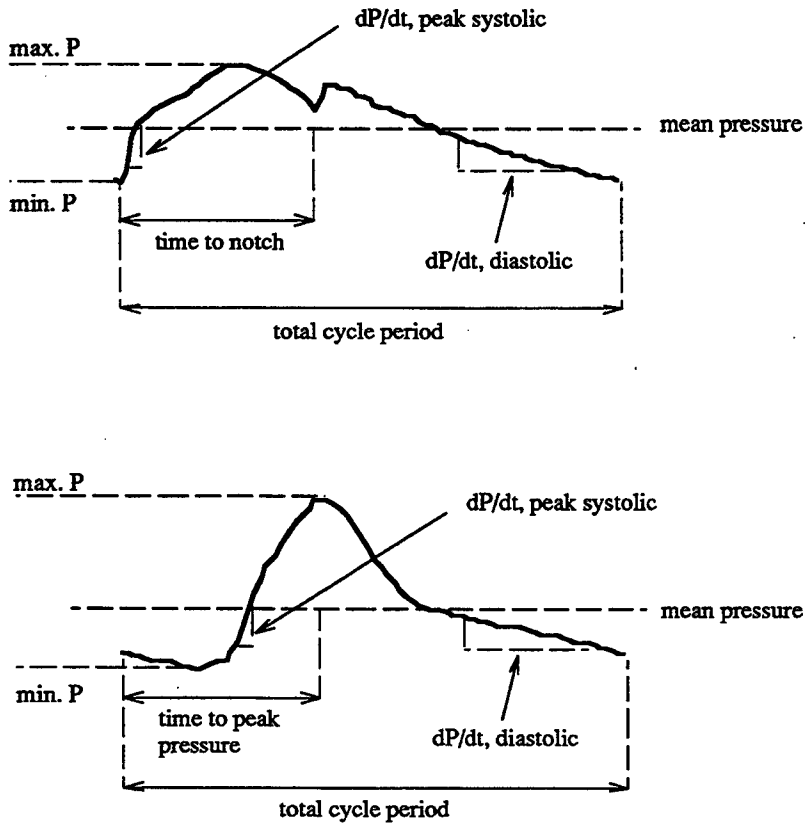


Figure 3. Features associated with a computed aortic pressure pulse (top) and a radial pressure pulse (bottom).

The features selected should be functions of the six parameters, and should be as independent from each other as possible. They should also be representative of the system as a whole, rather than depend primarily upon local hemodynamic conditions. One is not restricted, however, to using physical features that are visually obtained. Any six indices used to define a specific curve can be used. For instance, it may be possible to uniquely define a specific curve through the first several coefficients from the Fourier decomposition of a curve. Other mathematical decompositions may be applied to extract similar information. For present purposes, the set of features used are the following those outlined in Figure 3, applied to the radial artery and carotid artery pressures.

While it is clearly possible to calculate sample "points" throughout the parameter space values of the objective function, sparse points in space are inadequate for describing the complete behavior of the objective function. One can, however, make an

estimation of a continuous function that describes the relationship between the objective function and the set of parameter values. This continuous function is called a "surrogate", and acts as a model for the functional relationship that we wish to estimate. The surrogate is a multi-dimensional quadratic interpolation of a field of sample points. An existing quadratic interpolation routine known as the Shepard quadratic interpolation can be implemented with little modification to the problem at hand. The original formulations were first described in detail by Franke and Nielson and Renka for the two and three dimensional cases (Franke and Nielson, 1990; Renka, 1988). Yesilyurt (1995) expanded the routine to arbitrary dimensions, following the formulation of Renka.

The Shepard routine can be implemented to first determine the coefficients for the surrogate using the sample runs generated by the cardiovascular model. Thus, a set of parameters may be sent to the code, which then returns the corresponding set of feature values based upon the interpolation between points.

Plans for the Next Six Months

At this stage, several possible approaches can be employed to select the appropriate set of parameters given a specific feature set obtained from the measurements. One possibility is to invert the relationship between parameters and features so that all one needs to do is insert the values of the features and the individual parameters are directly computed. This may not be possible, however, given the likely nonlinearity of the parameter-feature relationship. One alternative is to solve the system iteratively using an optimization routine that minimizes the error between the features extracted from the measured trace and the features corresponding to a set of parameters from the computational model.

The selection and evaluation of various approaches to parameter estimation will be the primary aim of the next period. These will initially be tested against numerically-determined "test data", and subsequently against actual measurements.

Bibliography

Franke, R., Nielson, G.M. Scattered data interpolation and applications: A tutorial and study, in: H. Hagen and D. Roller, eds., Geometric Modelling: Methods and Applications. (131-160). Springer-Verlag, Berlin, 1990.

Mills, C.J., et al. Pressure-flow relationships and vascular impedance in man. Cardiovascular Research, 4, 405-417, 1970.

Pedley, T.J. The Fluid Mechanics of Large Blood Vessels. Cambridge University Press, 1980.

Renka, R.L. Algorithm 660: QSHEP2D: Quadratic Shepard method for bivariate interpolation of scattered data. *ACM TOMS*, 14, (149-150), 1988.

Yesilyurt, S. Construction and validation of computer-simulation surrogates for engineering design and optimization. Ph.D. thesis, Dept. of Mechanical Engineering, Massachusetts Institute of Technology: April, 1995.

Yesilyurt, S., Patera, A.T. Surrogates for numerical simulations: optimization of eddy-promoter heat exchangers. *Computer Methods in Applied Mechanics and Engineering* 121, 231-257, 1995.

Patient Monitoring and Diagnosis

CHAPTER 3

Hyper Ring Project

B-H Yang, H. Asada, K-W. Chang,
S. Rhee, Y. Zhang

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Progress Report, Total Home Automation and Healthcare/Elder Care Consortium
April, 1998

Hyper Ring Project

Boo-Ho Yang, H. Harry Asada
Co-Investigators

Kuowei Chang
Senior Lecturer

Sokwoo Rhee, Yi Zhang
Graduate Research Assistant

Abstract

The Hyper Ring is a new wearable monitoring system in a double-ring configuration, where not only temporal measurement but also spatial assessment on an arterial blood flow are allowed. This report presents preliminary research efforts for developing new instrumentation methodologies in utilizing the Hyper Ring configuration. In the Hyper Ring, each ring is equipped with LEDs and photodetectors for photoplethysmography. With dual, concurrent finger photoplethysmograms, we can monitor the pulse wave velocity (PWV) and assess the elastic property of the digital arteries. Four electrodes are also installed in the two-ring configuration for electrical impedance plethysmography (EIP). A new mathematical model for each of the instrumentation techniques is presented.

d'Arbeloff Laboratory for Information Systems and Technology
MIT

1. Introduction

As the population of aged people increases, close and continuous monitoring becomes more important. Real-time, continuous monitoring allows not only for emergency detection but also for long-term assessment for establishing the right dose and timing of medication. Especially, an ambulatory system that allows long-term monitoring of otherwise extremely noncompliant patients such as demented elderly people is highly demanded. To answer these demands, we developed a compact, wearable monitoring system in a ring configuration that can be comfortably worn by the patient twenty-four hours a day and that transmits data to a computer through a wireless communication. As a continuing effort to expand the functionality of the ring sensor, we have also developed a concept of a "Hyper Ring," where not only temporal measurement but also spatial assessment on an arterial blood flow are allowed with a two-ring configuration, as shown in Figure 1.

In the new sensor system, each ring is equipped with LEDs and photodetectors for photoplethysmography. With dual, concurrent finger photoplethysmograms, we can monitor the pulse wave velocity (PWV) and assess the elastic property of the digital arteries. Four electrodes are also installed in the two-ring configuration for electrical impedance plethysmography (EIP). The EIP is known to provide absolute measurement of volumetric change of an arterial segment.

The objective of this report is to present preliminary research efforts to fully utilize the double-ring configuration of the Hyper Ring for blood flow monitoring. A new pulse-wave model is presented taking into account the viscosity of blood for digital arteries. The model provides better fidelity with published experimental data than other well-known models of the pulse wave velocity. A preliminary model of the electrical behavior of finger tissues is also presented for better EIP instrumentation.

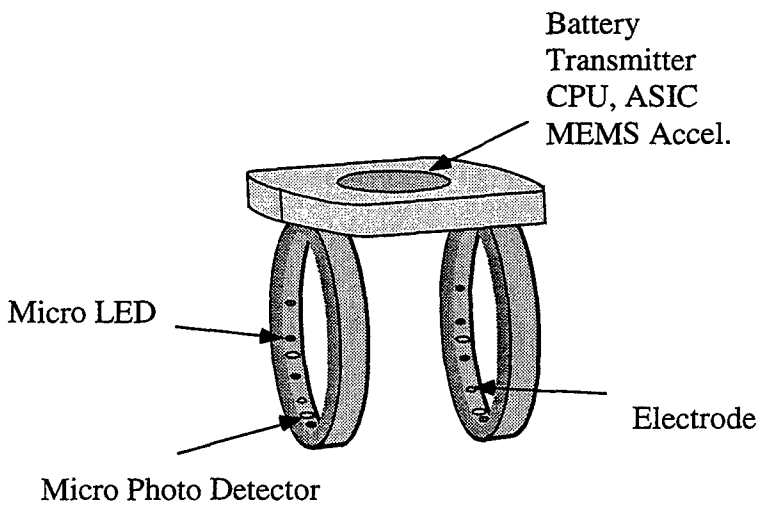


Figure 1: Conceptual diagram of the Hyper Ring

2. Hemodynamic Model for Pulse Wave Velocity

In the Hyper Ring, pulse wave velocity and the radius of the digital arteries on the finger can be measured by non-invasive optical. In addition to photo-plethysmography, we can get the blood pressure information continuously by combining these two measurable parameters and a proper model for digital arteries in the finger.

Many researchers have developed a variety of models for blood vessel, some of which are very simple and the others very complicated. To be adopted for our ring sensor-based approach, there are a few points that have to be observed by the model. First, as the diameter of the digital artery in the finger is relatively small (less than 1 mm), the effect of viscosity of the blood that flows in the artery must be considered significantly. Second, the model must be represented in a handy-closed form so that the real time computation is possible. This excludes the possibility of using many available models for pulse wave propagation which are composed of a set of nonlinear equations. In this section, we develop a new model which fits in our needs.

2.1 Blood Vessel Modeling for Pressure Change Monitoring

The blood vessel can be thought of as an elastic tube carrying a viscous fluid. The vessel expands and contracts due to the pressure of the blood in it. Therefore, if we can measure the elasticity and the volume of the blood vessel at a given point, we can derive the pressure. By using the Navier-Stokes equation and the mass conservation equation, we will derive this equation.

If the fluid is inviscid, the pressure can be easily obtained by applying basic wave equations. But if the viscosity of the fluid cannot be neglected, the formula would be quite different.

We start with the following basic equations. Figure 2 shows an idealized model of blood vessel. We measure the pressure at a short section by the ring sensor. Thus we can neglect the influence of tapering and branching of the blood vessel.

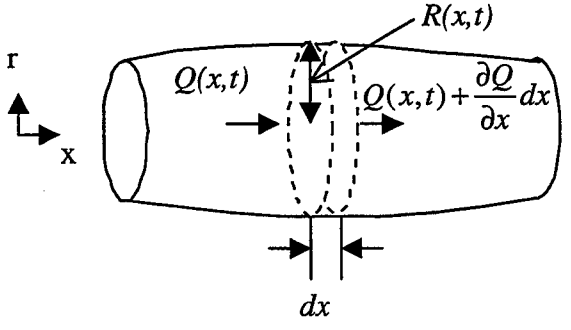


Figure 2: Idealized model of blood vessel

We first apply the standard Navier-Stokes equation :

$$\rho\left[\frac{\partial v_x}{\partial t}+v_r\frac{\partial v_x}{\partial r}+\frac{v_\theta}{r}\frac{\partial v_x}{\partial \theta}+v_x\frac{\partial v_x}{\partial x}\right] = -\frac{\partial P}{\partial x}+\mu\left[\frac{\partial^2 v_x}{\partial r^2}+\frac{1}{r}\frac{\partial v_x}{\partial r}+\frac{1}{r^2}\frac{\partial^2 v_x}{\partial \theta^2}+\frac{\partial^2 v_x}{\partial x^2}\right]+\rho G_x \quad (1)$$

where μ is the viscosity, and ρ is the density of the blood.

In this equation, x, r, θ represent longitudinal, radial, and rotational axes along tube respectively. This partial differential equation cannot be solved explicitly. Therefore we have apply several assumptions so that this equation can be reduced to a solvable form.

We make the following assumptions.

(1) No change along $\theta \Rightarrow \frac{\partial}{\partial \theta}=0$

(2) Neglect $v_r \Rightarrow v_r \approx 0$ (Wall velocity of r direction is very small compared with v_x.)

(3) As $r << x$, $\frac{\partial}{\partial r} >> \frac{\partial}{\partial x} \Rightarrow \frac{\partial}{\partial x} \approx 0$ (by order of magnitude analysis)

(4) No body force $\Rightarrow G_x = 0$

All of these are the standard assumptions that are made whenever the longitudinal length of the tube is much larger than its radius. With these assumptions and setting $u \equiv v_x$, the Navier-Stokes equation reduces to,

$$\frac{\partial u}{\partial t}=-\frac{1}{\rho}\frac{\partial P}{\partial x}+\nu\left[\frac{\partial^2 u}{\partial r^2}+\frac{1}{r}\frac{\partial u}{\partial r}\right] \qquad \nu = \text{kinetic viscosity } (\nu \equiv \mu/\rho) \quad (2)$$

where ν is the kinetic viscosity ($\nu \equiv \mu/\rho$)

As there are still three independent variables, we have to make another assumption. The heart beat is actually more like an impulse function. But if we approximate it as a sine wave, we can reduce the above equation into a modified Bessel equation, which can be solved easily. Approximating the heart beat as a sine wave means that the pressure wave form is approximate as a sine wave. In addition, we can also approximate the velocity of viscous blood in the tube as a sine wave form since this velocity is essentially generated by the pressure. As we consider the blood to be viscous, we also assume that the blood is locally fully developed flow. The Reynolds number in the digital arteries in a finger is around 100 to 250, and this verifies our reasoning that the blood cannot be considered as inviscid.

$$P = P_0 + P_1 e^{i\omega(t-\frac{x}{c})} \quad (3)$$

$$u = u_0 + u_1 e^{i\omega(t-\frac{x}{c})}, \quad u_1 = u_1(r) \quad (4)$$

ω = Angular pulse wave frequency

c = Pulse wave velocity

P = Pressure

u = Velocity of fluid (blood)

where P is the pressure as a function of x and t, and the P_0 is the constant. P_1 is a real number representing pressure amplitude. u is the velocity of the blood also represented as a function of x and t. u_1 is a function of r and is a complex number that includes a phase lag. The pulse wave velocity c is also a complex number including attenuation effect. The real part of c is mainly related to the actual pulse propagation speed that we measure, and the imaginary part is related to the attenuation effect as the wave travels along x-direction.

Inserting equations (3) and (4) into equation (2), we can get the following PDE.

$$\frac{\partial^2 u_1}{\partial r^2}+\frac{1}{r}\frac{\partial u_1}{\partial r}-\frac{i\omega}{\nu}u_1=-\frac{i\omega}{\nu\rho c}P_1 \tag{5}$$

Boundary Conditions are,

$$\begin{aligned}&\frac{\partial u_1}{\partial r}=0 \ \ at \ r=0\\ &u_1=0 \ \ at \ \ r=R(x,t)\end{aligned} \tag{6}$$

This is the form of modified Bessel equation. After we solve this and apply the boundary condition, we get the following value of u_1.

$$u_1=\frac{P_1}{\rho c}\left[1-\frac{J_0\left(i^{\frac{3}{2}}\sqrt{\frac{\omega}{\nu}}r\right)}{J_0\left(i^{\frac{3}{2}}\sqrt{\frac{\omega}{\nu}}R\right)}\right] \qquad J_0: \text{Bessel function of 0th kind} \qquad i^{\frac{3}{2}}=\frac{1}{\sqrt{2}}-\frac{1}{\sqrt{2}}i \tag{7}$$

So the velocity profile in the blood vessel is,

$$u(r,x,t)=u_0+\frac{P_1}{\rho c}\left[1-\frac{J_0\left(i^{\frac{3}{2}}\sqrt{\frac{\omega}{\nu}}r\right)}{J_0\left(i^{\frac{3}{2}}\sqrt{\frac{\omega}{\nu}}R\right)}\right]e^{i\omega(t-\frac{x}{c})} \tag{8}$$

Now we will put this result into the mass conservation. First, we can apply the continuity equation to the model on Figure 2.

$$\frac{\partial}{\partial t}\iiint_{CV}\rho dV+\iint_{CS}\rho v_x dS=0 \Rightarrow \frac{\partial}{\partial t}(\pi R^2\rho dx)+\rho(Q+\frac{\partial Q}{\partial x}dx-Q)=0 \tag{9}$$

Then this equation reduces to the following result.

$$2\pi R\frac{\partial R}{\partial t}+\frac{\partial Q}{\partial x}=0, \tag{10}$$

where Q = flow rate (m^3/s) and $R=R_0+R_1 e^{iw(t-x/c)}$ ($R0 >> R1$).

R is the radius of blood vessel at given x and t. R_1 is also a complex number that includes a phase difference. Magnitude of R_1 represents the amplitude of the change of radius, and the phase of R_1 represents the phase lag. As the amplitude of elastic deformation of blood vessel is around 5% ~ 10% of its original radius, we can consider that R_0 is much larger than R_1.

The flow rate (Q) is represented as follows,

$$\begin{aligned} Q &= \int_0^R 2\pi r u(r,x,t)dr \\ &= \pi R^2 u_0+\frac{\pi R^2 P_1}{\rho c}\left[1-\frac{2J_1(i^{\frac{3}{2}}\alpha)}{J_0(i^{\frac{3}{2}}\alpha)i^{\frac{3}{2}}\alpha}\right]e^{i\omega(t-\frac{x}{c})}, \quad \alpha\equiv R\sqrt{\frac{\omega}{\nu}}: \text{Womersley Number} \end{aligned} \tag{11}$$

If we put (11) into (10), we get the following.

$$2\pi RR_1-\frac{\pi R^2 P_1}{\rho c^2}\left[1-\frac{2J_1(i^{\frac{3}{2}}\alpha)}{J_0(i^{\frac{3}{2}}\alpha)i^{\frac{3}{2}}\alpha}\right]=0 \Rightarrow c=\sqrt{\frac{\pi R^2 P_1}{2\pi RR_1\rho}\left[1-\frac{2J_1(i^{\frac{3}{2}}\alpha)}{J_0(i^{\frac{3}{2}}\alpha)i^{\frac{3}{2}}\alpha}\right]} \tag{12}$$

From the definitions of P_1 and R_1, we can say $P_1 \approx \Delta P$, $R_1 \approx \Delta R$.

$$c=\sqrt{\frac{R\Delta P}{2\rho\Delta R}\left[1-\frac{2J_1(i^{\frac{3}{2}}\alpha)}{J_0(i^{\frac{3}{2}}\alpha)i^{\frac{3}{2}}\alpha}\right]}\equiv\sqrt{\frac{R\Delta P}{2\rho\Delta R}[1-f(\alpha)]} \tag{13}$$

This is the pulse wave velocity (c) represented with the pressure change (ΔP), the radius of blood vessel (R), the change of radius (ΔR), and the viscosity (ν). The pulse wave velocity and the radius of blood vessel are the quantities that can be measured with optical sensors or using EIP (Electro-Impedence Plethysmography), and the heart rate (ω) can be also measured by either of the two methods. Also, the viscosity of blood is a generally known constant. With these data we can derive pressure change information by using the following equation.

$$\Delta P = \frac{2\rho c^2 \Delta R}{R[1 - f(\alpha)]} \tag{14}$$

Let us assume the following stress-strain relationship for a thin-walled tube,

$$\Delta R = \frac{R^2}{eY} \Delta P$$

where Y is Young's Modulus and e is the thickness of the wall. Also, if we assume that the blood is inviscid, we get:

$$\nu \to 0 \Rightarrow \alpha \to \infty \Rightarrow f(\alpha) \to 0$$

Inserting the following definitions,

$$S = \pi R^2, \Delta S = 2\pi R \Delta R,$$

the equation (13) reduces to,

$$c_{inv} = \sqrt{\frac{S}{\rho}\frac{\Delta P}{\Delta S}} : \text{Bramwell and Hill equation}$$

$$= \sqrt{\frac{eY}{2\rho R}} : \text{Moens - Korteweg equation}$$

This verifies that our new equation of the pulse wave velocity matches with classical equations at the extreme ($\nu = 0$). This also verifies that our approach is valid.

2.2 Comparison with Other Models and Experimental Results

In 1980, Pedley derived a set of nonlinear equations that described the behavior of a blood vessel. The model included viscosity as well as the considerations of the dynamics of the wall of the blood vessel. He linearized it around R=0, and the result came out as follows.

$$c = \frac{R}{\sqrt{3}} \sqrt{\frac{\omega}{\nu}} \sqrt{\frac{eY}{2\rho R}} \qquad \text{around R=0}$$

A graphical comparison of Moens-Korteweg equation, Pedley's analysis, and the new equation is on Figure 3. This figure shows that the Moens-Korteweg equation goes wrong as R becomes smaller. It also shows that the new equation shows a fair match with the linearized model by Pedley around R=0. As the digital arteries in the finger has diameter

of around 0.5mm, both of the Moens-Korteweg equation and Pedley's analysis show considerable errors in the region we need, which makes it impossible for us to use them for our device to compute blood pressure information. But the new equation can be used even in this region.

Also a comparison with experimental results (obtained by Caro, Pedley, and Seed in 1974 from a dog) is on Table 1. This experimental results show that the error of Moens-Korteweg equation increases as R becomes smaller, but the new equation keeps its error less than 5 percent even in the area that Moens-Korteweg equation shows more than 20 percent of error. Figure 4 graphically shows the comparison of experimental result, this new model and the Moens-Koreteweg equation.

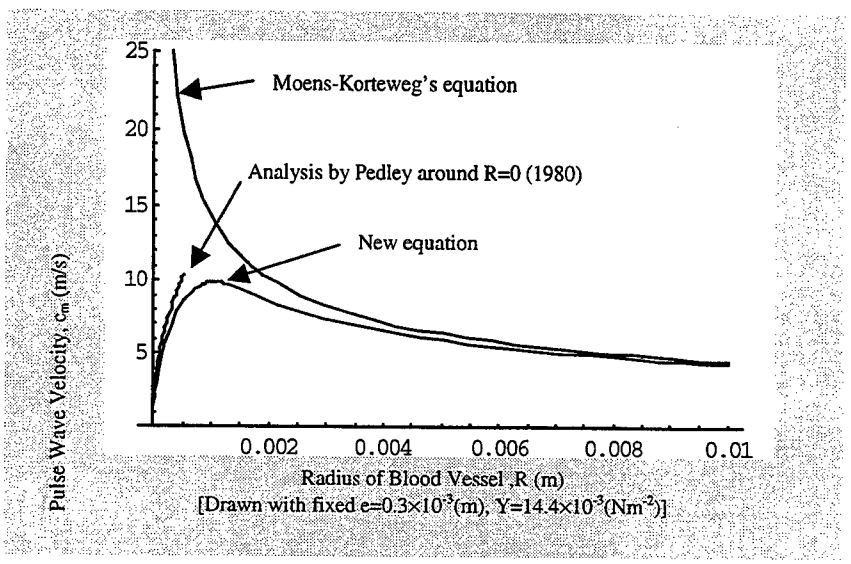


Figure 3: Comparison of the new model and other models of PWV

Table 1: Empirical data of blood nessls and the calcuation results of PWV

Cardiovascular Parameters of a Dog (Caro, Pedley & Seed (1974))

Site	Abdominal Aorta	Carotid Artery	Femoral Artery
Internal Radius, R (mm)	4.5	2.5	2.0
Wall Thickness, e (mm)	0.5	0.3	0.4
Static Young's Modulus (10^5 Nm^{-2})	10	9	10
Dynamic Young's Modulus, Y (10^5 Nm^{-2}) (Bergel, 1961)	12	14.4	13
Measured Wave Speed (m/s)	7.0	8.0	9.0

Calculation Result (Heart rate = 2 Hz, ρ=1.055 $\times 10^3$, ν=4 $\times$ 10^{-6})

Calculated Wave Speed from Moens - Korteweg equation (m/s) (Error %)	8.0 (14.3%)	9.1 (13.8%)	11.1 (23.3%)
Calculated Wave Speed from new equation (m/s) (Error %)	7.3 (4.3%)	7.9 (1.3%)	9.4 (4.4%)

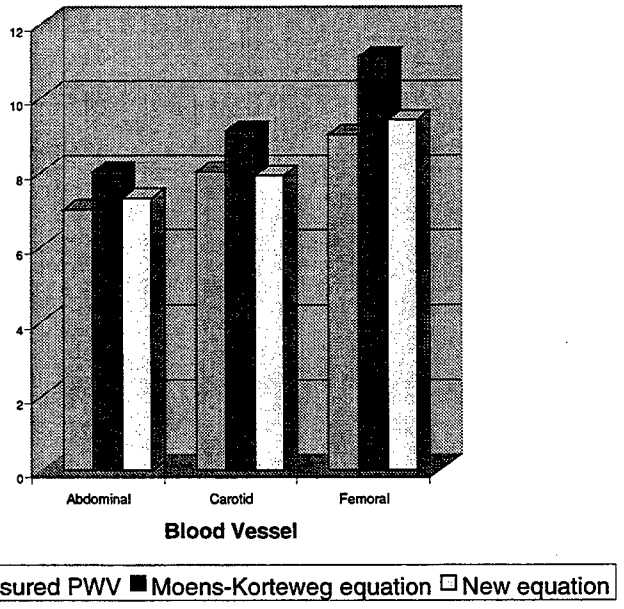


Figure 4: Graphical comparison of calculation results of PWV

These results show that the blood pressure change can be effectively obtained as we measure the pulse wave velocity and the radius of a digital artery.

2.3 Conclusions

In this section, a new model for a blood vessel for pressure change monitoring was presented. Although the existing model for pulse wave velocity has more and more error as the diameter of the blood vessel becomes smaller, the new model showed better fidelity with the experimental results and other models built for the region that the radius is very small. With this model and the double ring configuration, we can continuously monitor some information about the blood pressure, which will be very useful for the people who have a cardiovascular disorder.

3. Electrical Impedance Plethysmograph for the Hyper Ring

Digital blood flow has been proven to be one of many useful indicators for physiological and pathological changes in the peripheral circulatory system. It can provide certain valuable diagnostic indices for general circulatory insufficiency, especially in the peripheral vascular disease area including vasoparesis and atherosclerosis. Accurate measurement of the digital blood flow is useful in assessing the progress of medical treatment and surgical intervention.

The human finger is highly sensitive to physical and environmental changes, such as temperature and posture. The variation of blood flow in the finger is one of the most important physiological factors in the human thermal-regulatory system. For this reason, there is also a demand for non-invasive measurement of digital blood flow in management of the thermal environment.

There are several ways to measure blood flow and the Electrical Impedance Plethysmography (EIP) has been proven to be most promising because EIP has a unique advantage of flexibility not shared by other methods. The underlying principle of EIP is as follows: the human body consists of a variety of ionicly conducting tissues and body fluids, each of which has a different conductivity. As a result, the electrical impedance contribution of an anatomical unit depends not only on its conductivity but also on the volumetric proportionality of these components. Under the assumption that the impedance of all the tissues remains constant, the impedance change of a finger segment is caused mainly by the blood volume change within the segment brought on by a pulsatile blood flow in the section.

In this section, the conceptual design of the blood flow measurement using a ring sensor is described in Section 3.1. The various key issues are presented in Section 3.2. A detailed parallel conductor model of the human finger is described in Section 3.3. In the last section, the potential applications of electrical impedance plethysmography are discussed and future work plans are also presented.

3.1 Conceptual design of Electrical Impedance Plethysmograph for Ring Sensors

3.1.1 Principle of Electrical Impedance Plethysmograph

The fundamental law governing the electrical resistance is

$$R = \rho \frac{l}{A} \tag{15}$$

where,

A : Cross section of the conductor

l : Length of the conductor

R : electrical resistance of the conductor

ρ : resistivity of the conductor

By applying the above equation to blood and multiplying the right hand side of equation (15) by 'l/l', we obtain the basic formula for electrical impedance plethysmography. Here we assume that the resistivity of blood is constant, and

$$Z = \rho \frac{l^2}{V} \tag{16}$$

where,

Z : Electrical impedance of the blood (time varying because of prapagation of blood along the vessel)

V : Volume of the blood in the measured segment (time varying because of prapagation of blood along the vessel)

l : Length of the blood vessel being measured

ρ : electrical resistivity of blood, $150\Omega \cdot$ cm for normal haematocit

We replace 'R' with 'Z' because the electrical impedance of human tissue is not purely resistive, and it also has a capacitive component as well (to be discussed in a later section).

Re-arranging equation (16), we obtain

$$V = \frac{\rho l^2}{Z} \tag{17}$$

In the traditional EIP technique, the following assumption is made that the volume and impedance changes in the finger is entirely due to the blood flow, which means that the impedance of other tissues in the finger remains constant. Besides, by applying certain frequency current, the capacitance components of the tissue are supposed to be minimized to zero. Then applying impedance plethysmograph, we first measure the impedance change dZ/dt, and then calculate the volumetric change of the blood due to the heart beat in the measured segment, by the following equation:

$$\frac{dV}{dt} = -\frac{\rho l^2}{Z^2}\frac{dZ}{dt} \tag{18}$$

where: dV/dt is the change rate of blood flow(ml/sec);

ρ is the resistivity of blood(Ω·cm);

L is the length of the measured section(cm);

Z is the basal impedance of the measure section(Ω);

dZ/dt is the change rate of the electrical impedance(Ω/sec).

By integrating both sides of equation (18), we can obtain the pulsatile blood flow in certain time interval.

This relationship is valid only when:

(1) the electrical properties of tissue and blood are isotropic

(2) the tissue with the exception of the blood vessel has a constant cross section

(3) the electric field is uniform throughout the cross-sectional area

3.1.2 Conceptual Design of the dual-ring sensor

Based on these restrictions, we make the following design for the hyper ring sensor as shown in Figure 5. It has following features:

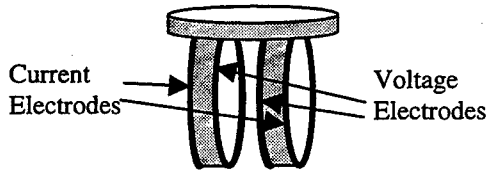


Figure 5: Conceptual design of the hyper ring

(1) The hyper ring consists of two rings, which are 1cm apart from each other.

(2) Two annular electrodes consist of thin silver wires are embedded into the outside of the ring, which introduce a constant, high-frequency and low-amperage current. In a similar arrangement, two additional wires are embedded into the inner side of the ring, which detect the voltage change of the section.

3.2 Issues

- Electrical safety

A major point of concern for any medical instruments is its electrical safety. This is particularly important in the ring sensor since it supplies current directly to the patient. As a rule of design, the sensor must be electrically isolated from the AC power lines.

During the development, we consider two aspects:

First of all, to choose the minimum current based on signal and noise considerations so as to minimize the effect of current on the human subjects.

Second, to use a medical grade isolation transformer to isolate the sensor, which is attached to the human subject directly, from various electrical equipment.

- Electric field artifact

One of the conditions for successful EIP measurement is that the electric field distribution in the section must be uniform. This means applying a current source rather than a voltage source on the segment.

It has been demonstrated that the use of annular electrodes can provide a uniform electric field distribution in the central portion of the section provided that some distance exists between the current and potential electrodes in order for a uniform axial field to develop.

Based on this requirement, we use four electrodes in the ring sensor design, rather than four small electrodes which may save space but is totally ineffective.

- Real and Imaginary parts of electrical impedance

In the traditional EIP technique, the frequency is chosen to minimize the effect of the reactance component in the finger. It assumes that at this frequency, the capacitance of the human tissue is negligible so that the measured impedance is resistive, and the measured resistance is directly related to the volumetric change of the finger.

3.3 Parallel Conductor Model of the Finger

Because of the presence of non-conducting cell membrane, the body tissue is not purely resistive, which means that there is a phase shift between the current and voltage. In present EIP application, a constant current of a fixed frequency (20 to 100KHz) is chosen to minimize the effect of reactive component. Such simplification results in a neat equation but it is also responsible for creation of inaccuracy as well. This dilemma can be resolved by formulating an accurate and practical model of electrical impedance of the human finger section.

The study of digital circulation is of significant interest not only in the monitoring of peripheral circulation but also because it is an important index of the regulations of the circulation system to the changing environment. The anatomy of a human finger is shown in Figure 6.

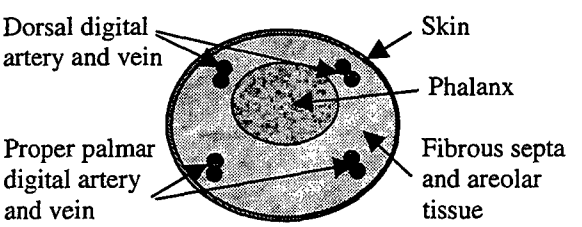


Figure 6: Cross sectional view of the human finger

Considering the contributions of all the relevant components, including external electrodes, blood, bone, muscle, skin, and other tissues in the finger, we obtain a parallel conductor model for the finger section shown in Figure 7.

- all of the parallel components of the finger, such as blood, skin, muscle, bone, have both electrical resistance and capacitance.
- There exists a contact resistance and parallel capacitance between the skin and conducting gel applied on the skin surface to minimize interfacial impedance.
- R_3 represents the transverse impedance of the finger, which is much smaller in comparison with the longitudinal impedance, and it is usually ignored in EIP measurement.

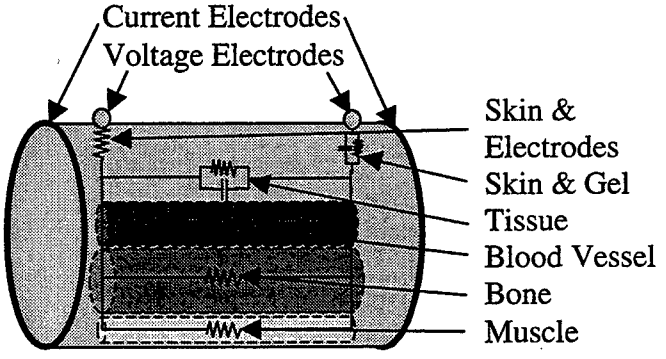


Figure 7: Electrical model of the finger section

The transfer function of the model is given by

$$\frac{U(S)}{I(S)} = \frac{R_1}{R_1 C_1 S + 1} + \frac{R_2}{R_2 C_2 S + 1} + R_3 \tag{19}$$

where: U(S) is the Laplace Transform of the voltage;
I(S) is the Laplace Transform of the current;
R_1 is the equivalent parallel resistance of blood, muscle, bone and skin in the finger
C_1 is the equivalent parallel capacitance of blood and other tissues in the finger (caused by the dielectric property of the cell membrane);
R_2 is the contact resistance between the skin and conducting gel;
C_2 is the contactcapacitance between the skin and conducting gel;
R_3 is the resistance of the skin and electrodes.

By using the resistivity values provided by Geddes and Baker (1967), we obtain the simulation result of the frequency response of the finger section.

The geometry parameters of the finger section that we used in the model are as followings:

Length of the section $L = 1cm$

Perimeter of the section $P = 6cm$

Width of the electrodes $w = 0.1cm$

So : Radius of the section $r = \frac{P}{2\pi} = \frac{6}{2\pi} = 0.955cm$

Area of the section $A = \pi r^2 = 2.865cm^2$

Volume of the section $V = AL = 2.865 \times 1 = 2.865cm^3$

The electrical properties and anatomical parameters of the finger section we used in the model are:

Resistivity of the blood $\rho_{blood} = 150\Omega \cdot cm$ Percent of the blood in the finger : 5%

Resistivity of the bone $\rho_{bone} = 1800\Omega \cdot cm$ Percent of the bone in the finger : 40%

Resistivity of the skin $\rho_{skin} = 289\Omega \cdot cm$ Percent of the skin in the finger : 50%

Resistivity of the muscle $\rho_{muscle} = 245\Omega \cdot cm$ Percent of the muscle in the finger : 5%

Contacting capacitance of the skin and gel $C_{skin} = 1\mu F / 100cm^2$

Capacitance of the cell membrane $C_{membrane} = 1\mu F / cm^2$

Here are the calculations of the resistance and capacitance in the parallel conductor model:

$$R_{blood} = \rho_{blood} \frac{L^2}{V_{blood}} = 150 \times \frac{1^2}{2.865 \times 7\%} = 748\Omega$$

$$R_{bone} = \rho_{bone} \frac{L^2}{V_{bone}} = 1800 \times \frac{1^2}{2.865 \times 30\%} = 2094\Omega$$

$$R_{skin} = \rho_{skin} \frac{L^2}{V_{skin}} = 289 \times \frac{1^2}{2.865 \times 60\%} = 168\Omega$$

$$R_{muscle} = \rho_{muscle} \frac{L^2}{V_{muscle}} = 245 \times \frac{1^2}{2.865 \times 3\%} = 2850\Omega$$

Thus, the equivalent resistance and capacitance are:

$$R_1 = \frac{1}{\frac{1}{R_{blood}} + \frac{1}{R_{bone}} + \frac{1}{R_{skin}} + \frac{1}{R_{muscle}}} = \frac{1}{\frac{1}{748} + \frac{1}{2094} + \frac{1}{168} + \frac{1}{2850}} = 123\Omega$$

$$C_1 = C_{membrane} \times A = 1 \times 2.865 = 2.865\mu F$$

$$R_2 = R_{skin} // R_{gel} \approx R_{gel} = 100\Omega$$

$$C_2 = C_{skin} \times A_{electrode} = 1/100cm^2 \times (2 \times 6 \times 0.1) = 0.012\mu F$$

$$R_3 = 2\rho_{skin} \frac{L}{A_{electrode}} = 289 \times \frac{0.1}{6 \times 0.1} \times 2 = 96\Omega$$

With this estimation, we can obtain the electrical model of the finger section in term of transfer function:

$$\frac{U(S)}{I(S)} = \frac{R_1}{R_1 C_1 S + 1} + \frac{R_2}{R_2 C_2 S + 1} + R_3$$

$$= \frac{123}{123 \times 2.865 \times 10^{-6} S + 1} + \frac{100}{100 \times 0.012 \times 10^{-6} S + 1} + 96$$

$$= \frac{4.06 \times 10^{-8} S^2 + 8.76 \times 10^{-2} S + 319}{4.23 \times 10^{-10} S^2 + 3.56 \times 10^{-4} S + 1}$$

Figure 8 is the simulation result of the frequency response of the model. It shows the phase shift between the voltage and current which was found to be decreasing between the frequency range from 10KHz to 100KHz. The upper limit of 100 KHz was the frequency used in traditional EIP.

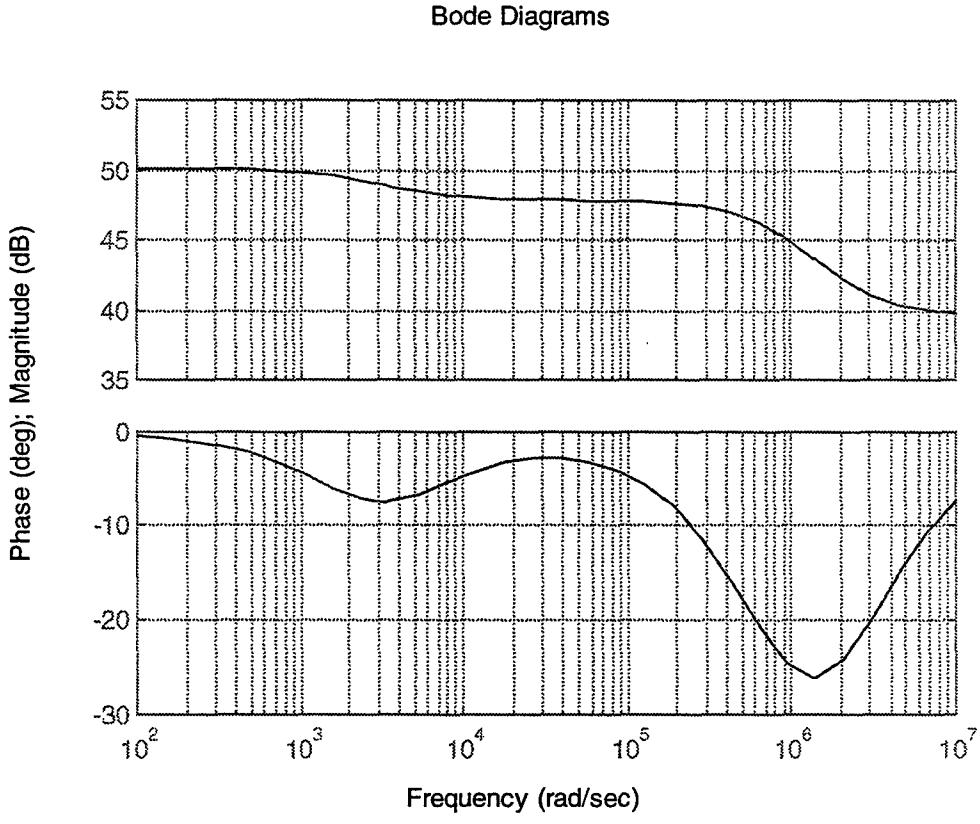


Figure 8: Simulation of the electrical model of the finger

3.4 Prospect of EIP application

- Identification of composition of the human tissue

Since the electrical property of the human tissue may vary with time, we can hardly describe it by a constant-coefficient model. What we can do, after successful formulation of the initial model, is to use the adaptive observer technique to obtain the dynamic electrical model of the human body.

3.5 Future work

- Prototype development of the dual-ring sensor with EIP implementation
- Miniaturization of the dual-ring sensor
- Explore further utility of EIP in body composition estimate, cardiac output measurement, etc.

References

Baker, L. E., 1989, "Principles of the Impedance Technique," IEEE Eng Med Biol Mag, March, pp. 11-15

Geddes, L. A., Baker, L. E., 1967, "The Specific Resistance of Biological Material – A Conpendium of Data for the Biomedical Engineer and Physiologist," Med & Biol Eng, Volume 5, pp. 271-293

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Patient Monitoring and Diagnosis

CHAPTER 4

Miniaturization of the Ring Sensor

B-H Yang, H. Asada, K-W. Chang,
S. Rhee, Y. Zhang

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Progress Report, Total Home Automation and Healthcare/Elder Care Consortium
April, 1998

Miniaturization of the Ring Sensor

Boo-Ho Yang, H. Harry Asada
Co-Investigators

Kuowei Chang
Senior Lecturer

Sokwoo Rhee, Yi Zhang
Graduate Research Assistant

Abstract

In this report, the issues relating to miniaturization of a ring sensor for photo-plethysmography are discussed. The problems that arise from microscopic-level implementation are presented, and the solutions to these problems are suggested. The importance and methodology of power saving, largely due to the limited capacity of a small-sized battery, is also discussed. The process of fabricating the miniaturized version of finger ring sensor is described in detail.

d'Arbeloff Laboratory for Information Systems and Technology
MIT

1. Introduction

A finger ring sensor for 24-hour patient monitoring was already developed in Phase I of the consortium. However, the size of the current prototype, as shown in Figure 1-(a), was too large to be of use in real life. For the ring sensor to be practical, it must be reduced to a size no larger than a college ring. If various electronic components including the battery can be packed within this size constraint, people will just wear it, be incognizant of its presence, and may even take a shower while wearing it. Furthermore, the current prototype is too heavy and unwieldy, a factor responsible for the majority of the motion artifact observed. Therefore, the ring must be compact and light enough such that the person who wears it would not feel its presence. In our initial effort, we plan to consolidate all the circuit elements on to a small, single-layer printed circuit board as shown in Figure 1-(b).

The ultimate and most ideal way to reduce size is to design a mixed signal ASIC chip so that all the circuitry can be condensed and packed up in one masked-chip. Unfortunately, this approach is too costly, it requires a long lead-time, and was found to be impractical at the present stage of development. Since we cannot go for an ASIC at this time, we have no alternative but to use commercially available discrete electronic components for ring sensor construction.

As is well known, there are certain limits in the size of electrical components. The conventional integrated circuit, for example, is too large to be considered for ring construction if it is used in their regular plastic package. However, a substantial size reduction can be achieved by putting IC chips, in die form, directly on a circuit board without the plastic package. It would also be advantageous to use discrete resistors and capacitors of the smallest size available. As soldering is obviously difficult with these components, a thermal-sonic wire-bonding machine using 0.001″ gold wire are used to connect the components to the circuit board.

Besides size reduction, power consumption minimization is another obstacle. As the size of the ring is reduced, it is no longer practical to put relatively large batteries on the ring sensor. As a result, the operating life of the ring is substantially reduced with small batteries, and the power saving strategy becomes far more critical. Nevertheless, we have devised a number of important power reduction schemes which successfully extended the useful life of these small batteries by adopting faster LEDs, higher CPU clock speed, and faster transistor for the RF transmission circuit.

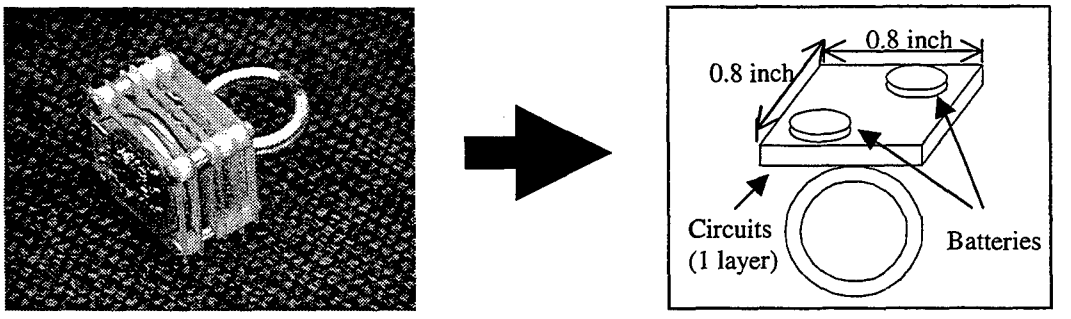


Figure 1: The prototype ring sensor (a) and a new design of miniaturized ring sensor (b)

2. Issues on Miniaturization

2.1. How do we reduce size?

The components we usually use have their own limits. For example, the size of an IC chip is characterized by its package. If we stick to these kind of normal components it is impossible to reduce the size of the ring to the extent desired.

Despite its packaged size, the functional core of an IC chip is actually very small. This core part is called a "die" and its size is far smaller than the package itself. The reason that a die is embedded in a large plastic package is to make soldering and handling easier. Although difficult, we can use the IC chips in their "die" form rather than in their plastic package, thereby resulting in a ring construction of minimal size. Since these chips cannot be connected with normal soldering iron, we need to use the equipment called "wire-bonding machine". This machine connects components with very thin gold wire. Basically it works like an extremely tiny welding machine which melts gold with ultrasonic energy at an elevated temperature of 150 °C. The inter-connections between various active and passive components are manually made under a microscope with the aid of a mechanical mouse for positioning of the circuit substrate relative to the welding head. A schematic diagram depicting the operation of this machine is shown in Figure 2.

For passive components such as resistors and capacitors, we can purchase them in very small thin or thick film style which are designed for wire bonding. These components are very small compared to conventional ones we normally use (Figure 3). The pads (called "termination") which replace the leads of conventional components are made of gold or aluminum when they are designed for wire bonding. There are certain kinds of components for which their pads made of nickel or silver, they are not designed for wire bonding, but rather, for surface mount assembly. If we have to use this type of components for practical considerations dictated by commercial availability, we cannot resort to wire bonding. In this particular case, we would use conducting epoxy to make connections on the circuit board.

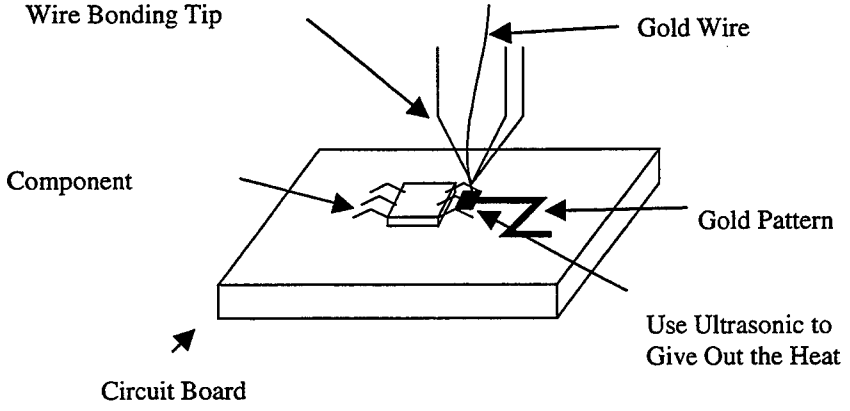


Figure 2: Wire Bonding Machine

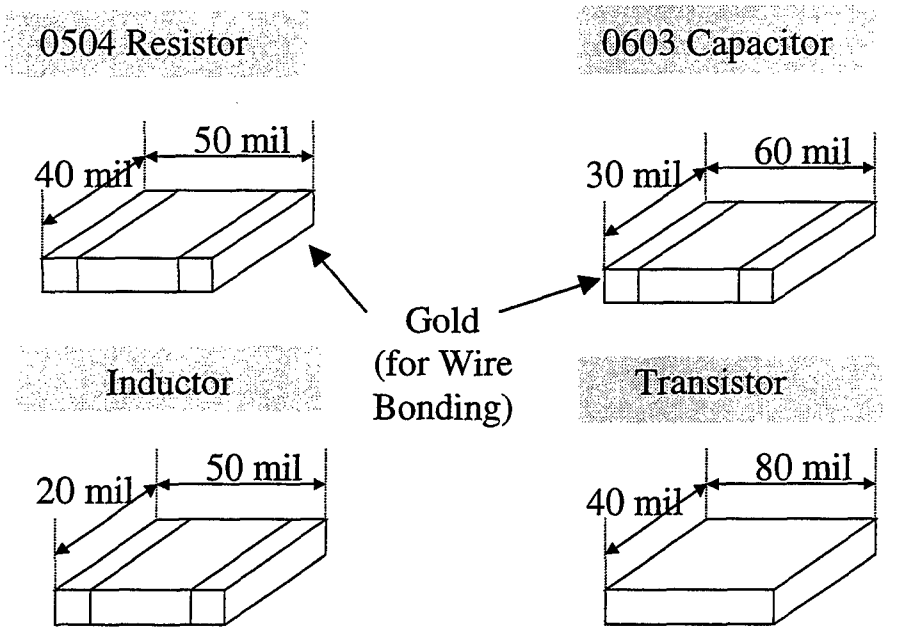


Figure 3: Dimensions of components

2.2. What kind of circuit boards will we use?

There are two choices for the circuit board material and ceramic substrates are more commonly used. Ceramic substrate is rigid, strong, and it can withstand the high temperature encountered in the thermal sonic wire bonding operation. More importantly, all the resistors used in our design can be deposited as film resistors on the substrate without actually taking up any physical space because these thin film-type resistors can be over-coated with a thin layer of insulating glazing material such that other components such as capacitors, transistors and IC dies can be situated on top of the underlying resistors. Since the film resistors are deposited at the factory by screen printing, our work load is substantially reduced during fabrication. Despite the advantage in size and work load reduction, ceramic substrate was not used in our current development due to high cost and a prohibitively long lead time.

One practical solution lies in the use of conventional G-10 epoxy printed circuit board material. This approach offers some unique advantages. One can make multiple circuit patterns on multiple layers in the same board. The benefit is that the space taken up by the inter-connecting circuit tracings is greatly reduced, and we can make the circuit board and therefore, the ring smaller. In addition, the fabrication cost for printed circuit boards is less than that for ceramic substrates, and the lead time is much shorter (usually only a couple of days). Of course, in this case we cannot enjoy the benefit of ceramic circuit board which comes with film resistors already put on by the factory.

2.3. How do we reduce the power consumption?

Power saving is of critical concern in a miniaturized ring sensor application, since, inevitably, tiny batteries must be chosen in place of those used in the first prototype. In the first place, power consumption can be reduced by shortening the duty rate of the LED

blinking cycle. To implement this, we need a photodiode which can respond to the short duration light pulse much faster. For this reason, we need to use a PIN photodiode rather than an ordinary PN photodiode. PIN type photodiode usually has 10 times faster response time than PN diode. To reduce the on time of LEDs, it is necessary to turn LEDs on and off faster. For this, we also need a higher CPU clock speed.

For telemetry transmission, we can save power by increasing the baud rate of RS-232 protocol. Since the number of bits to be transmitted per unit time is fixed, we can reduce the time duration when the transmitter is turned on by making each bit shorter through the use of a higher baud rate. For this to be successful, the rise and fall time of the transmitter need to be reduced, and this can be achieved by adopting a faster transistor in the Rf transmitter circuit.

In conclusion, in order to save power, we need to adopt a faster photodiode, a higher CPU clock, and a faster transistor for the transmitter.

3. Process of Fabrication

3.1. Finalize the circuit and collect the necessary components.

Since we have already built a prototype, we have a finalized circuit diagram. The circuit we will use for building a miniaturized ring sensor is almost the same as that of the previous prototype ring except for a few changes. (The circuit diagram is given in Appendix 1 and 2.) The changes were made largely due to power saving considerations. First, the number of red LED has been reduced to one from two in the previous prototype. This will save about 500 microamperes. Although we reduced the number of LED's, we have already verified that it is still sufficient for the operation as we expected. The crystal for CPU clock was also changed to 120 MHz to reduce the duty rate of LED's.

Ambient light elimination was implemented from the software side. The circuit is equipped with two channels of signal conditioners. To activate ambient light elimination, we just turn off one of the two LED channels permanently. In other words, we turn on the detector while no light source is turned on. The signal coming to the detector at this moment will be purely from the ambient light sources. We subtract this from the value measured when the one of the LED 's is on, so that we get a value that the influence of ambient light is removed.

About the components, all the IC chips used in this circuit were in die form. The sizes of these die form chips range from about 50 by 50 mil to more than 100 by 100 mil, which is still much smaller than the packaged type of surface mount type.

3.2. Design a conducting pattern to be put on the ceramic substrate or the printed circuit board.

The next step is to design the circuit pattern of the board. This is the most time consuming and takes a lot of brain power. First, we measure the dimensions of all components, and determine the positions of the components on the board. To make the circuit board as small as possible, the spaces between components must be minimized as long as it does not break the design guideline. A typical example of the design guideline is shown on Appendix 3. Basically, after we decide the positions of the components, the

VCC and GND supply lines are drawn first, and the other lines must be drawn later. The complete pattern is shown on Appendix 4 and 5.

3.3. Make the circuit board using gold as conducting material

As mentioned earlier, we can make the circuit board out of ceramic material or the conventional printed circuit board. At present, we adopted the ordinary printed circuit board to reduce the lead time. As a result, we must install all of the resistors manually ourselves, which not only takes more time but also may increase the possibility of mistakes.

3.4. Put the components on the board and connect using wire bonding technique or conducting epoxy.

The ceramic substrate is designed basically for wire bonding technique. In case that the components are not for wire bonding (that is, the terminations are not made of gold or aluminum), we use conducting epoxy to connect the terminations to the gold pattern on the board. The wire bonding must be done very carefully because the gold wire is extremely easy to break, and it is hard to get rid of the broken wire and to do new bonding. Even after finishing the bonding, the wire bonded components must be gently taken care of. To prevent the gold wire from being broken, it is recommended to put non-conducting epoxy on the bond and harden it with heat, so that the wire is shielded by the epoxy. The picture of the complete circuit board of miniaturized ring is shown on Figure 4.

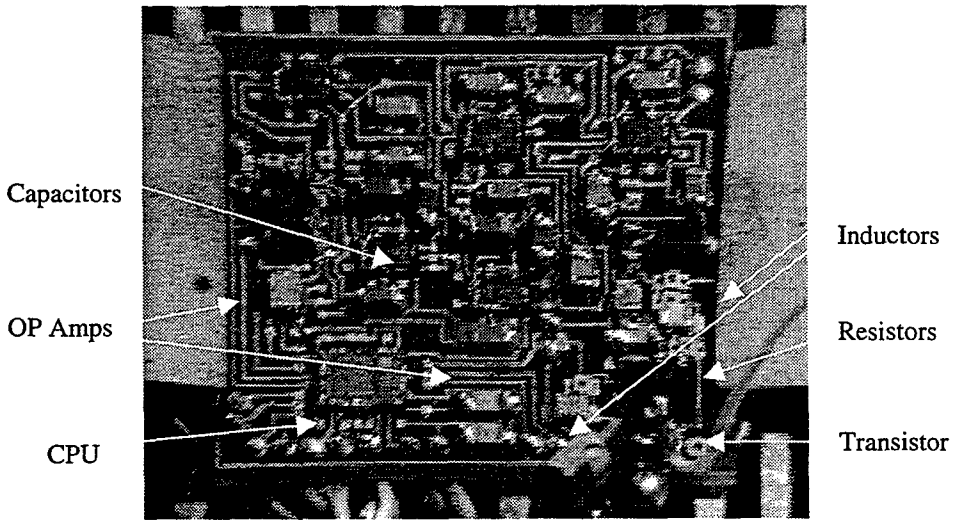


Figure 4: Picture of the miniaturized circuit board

3.4. Do external wiring and Debugging

After finishing wiring the board, we have to do some external wiring, since the large components such as batteries, oscillators and switches are not bonded on the board. They have to be connected to circuit by external wires. In addition, there is another small circuit board for the LED's and the photodiode. This small board will be glued on the opposite side of ring from the main circuit board. This small board also must be connected to the main board by thin ordinary wire. If we want to put external antenna, this must be wired, too.

The last step is debugging. It would be almost a miracle if all the circuit works just after finishing the fabrication, especially for this kind of small device. Most of all, the gold wire is so liable to break, as was mentioned already. Check all the gold wire bonding if they are not broken. If any kind of soldering was done, check if it is done completely. Another important checkpoint is the components themselves. The die form chips are very weak to electrostatic and must be checked carefully in case that the ring doesn't function. The picture of a complete miniaturized ring sensor is shown in Figure 5.

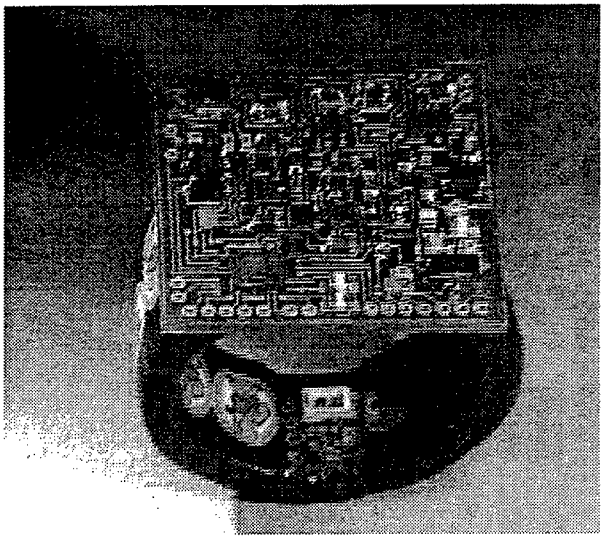


Figure 5: A miniaturized ring sensor

3.6. Software – In circuit Programming

The CPU used in the present miniaturized ring sensor was designed to allow in-circuit programming. We can erase the assembly codes in the CPU by exposing the EPROM portion of the CPU to an ultraviolet light source from a standard EPROM eraser. After the CPU memory becomes blank, it can then be reprogrammed by connecting five wires from the board to the corresponding pins of an EPROM writer. This useful feature allows

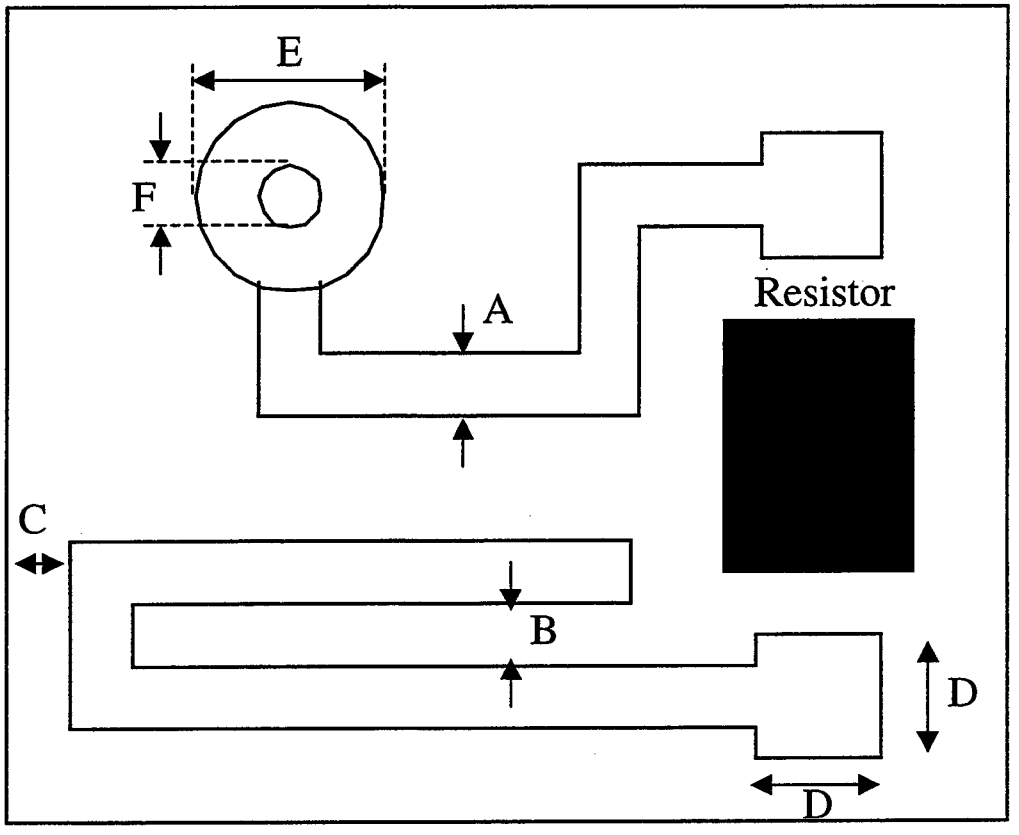
the ring sensor to become totally programmable. The operating software for the ring sensor can be upgraded whenever an improved version of software becomes available.

4. Conclusion and Future Work

In this report, the issues concerning miniaturization and fabrication of the ring sensor were discussed in detail. It was shown that the previous bulky prototype ring sensor can indeed, be effectively reduced to the size of an ordinary ring without sacrificing any of the functionality. As a matter of fact, the hardware design was actually improved so that more useful functions such as power saving algorithm and ambient light cancellation can be implemented by software.

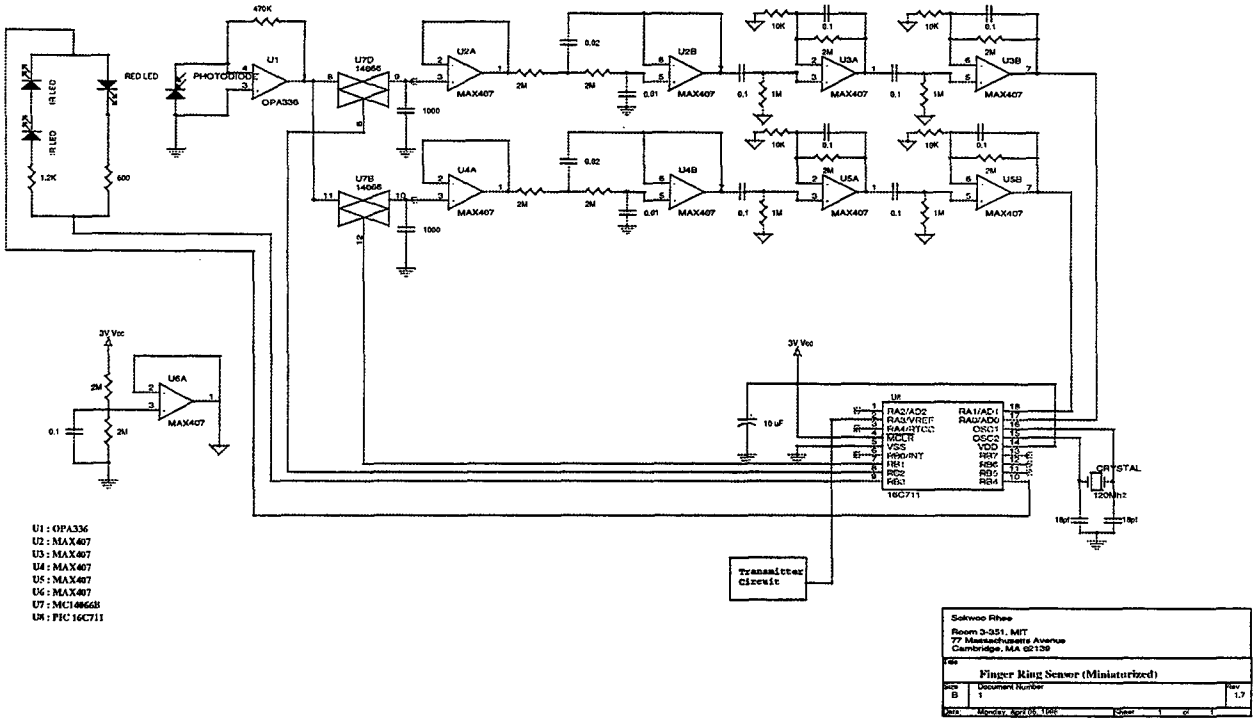
In the future, the dual ring configuration for continuous blood pressure monitoring will also be implemented in a miniaturized format after the efficacy, usefulness and functionality are fully established.

Appendix 1: Typical Values of Dimensions on the Substrate Design

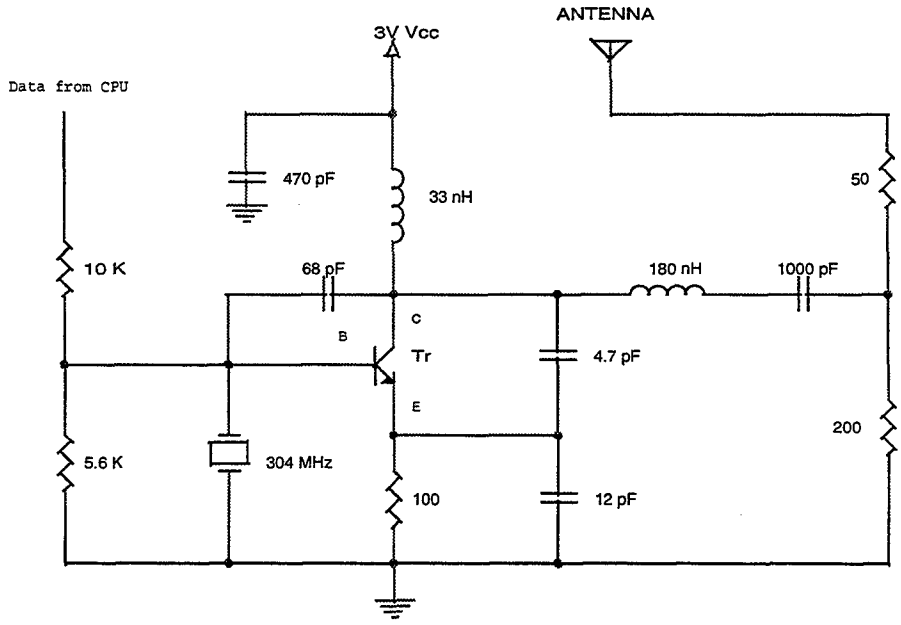


	Patterns	Typical Values (inch)
A	Conductor Width	0.06
B	Conductor to Conductor Space	0.06
C	Conductor to Substrate Edge	0.06
D	Wire Bond Pad	0.0125
E	Ring	0.03
F	Hole Diameter	0.01

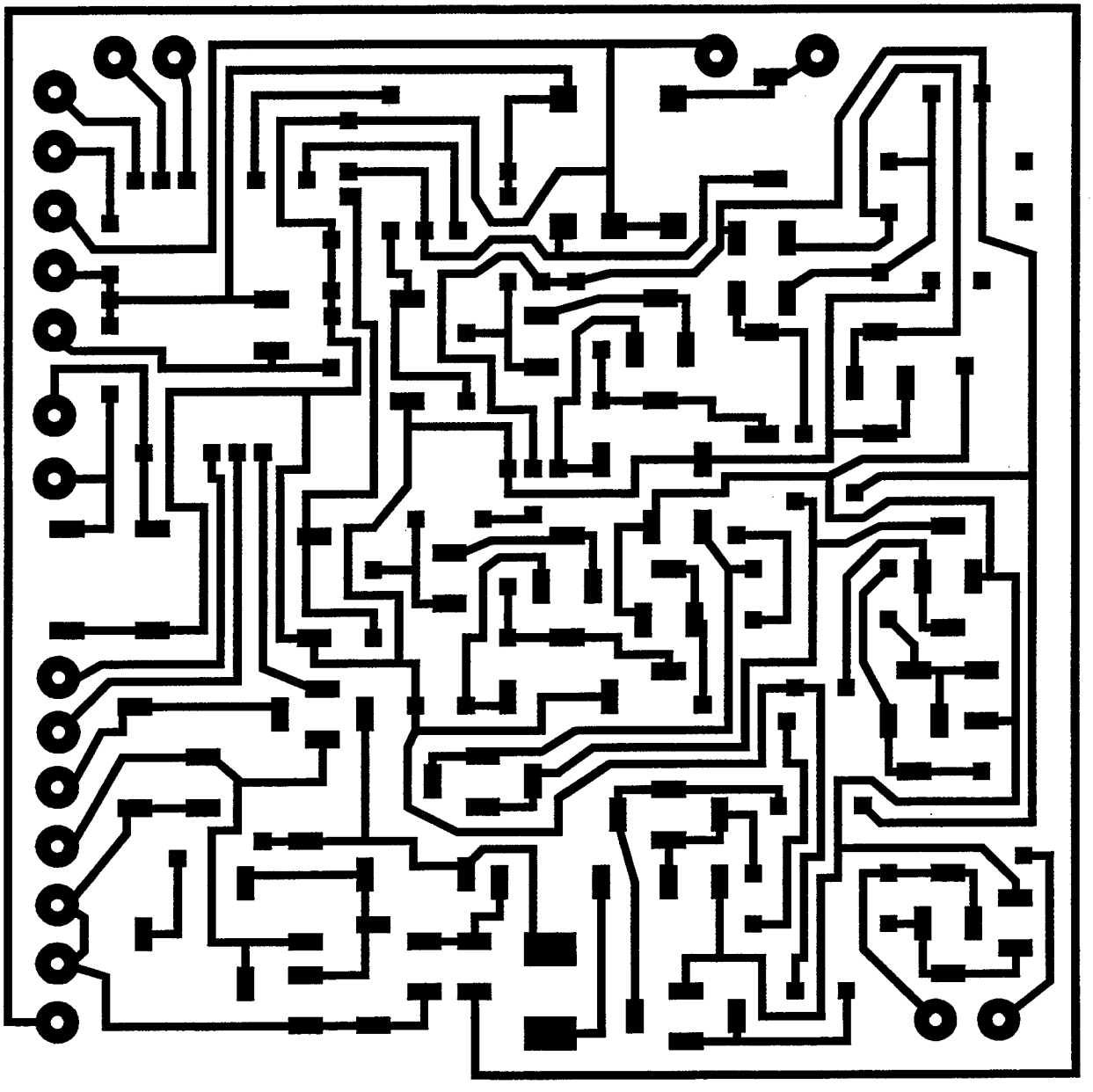
Appendix 2: Circuit Diagram of the Ring Sensor (Detection and Signal Processing Part)

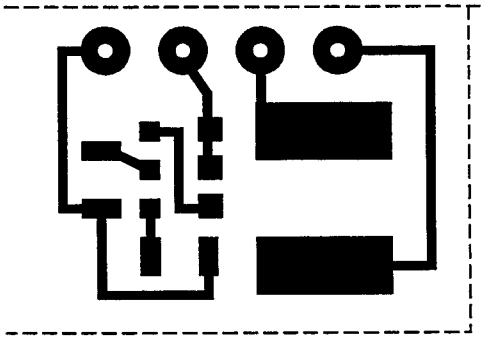


Appendix 3: Circuit Diagram of the Ring Sensor (Transmitter Part)

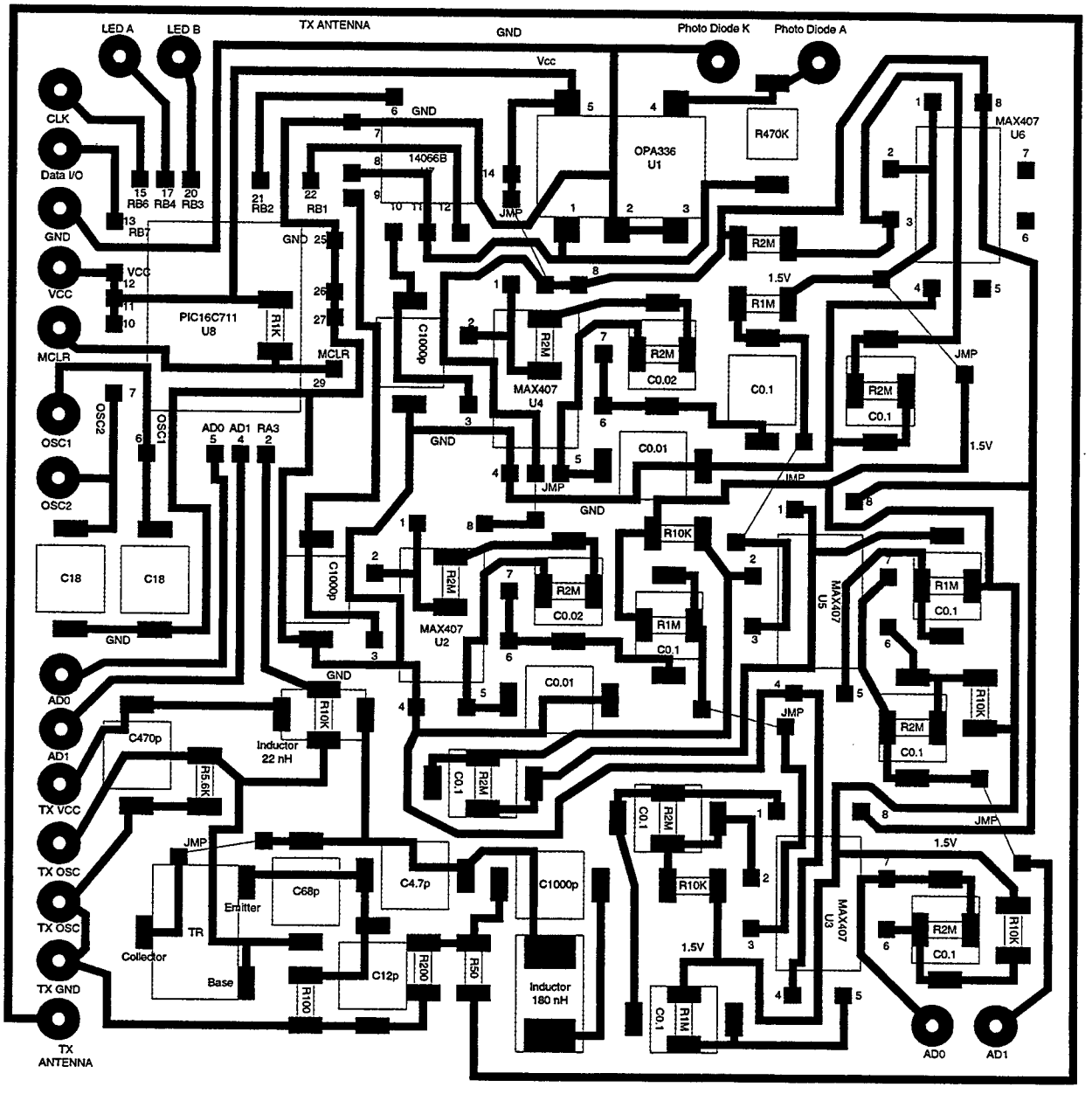


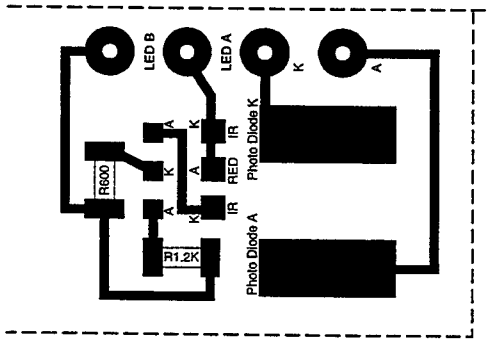
Appendix 4: Pattern for the Miniaturized Ring Sensor





Appendix 5: Labeled Pattern for the Miniaturized Ring Sensor





Patient Monitoring and Diagnosis

CHAPTER 5

SIMSUIT and Biochair Projects

L. Jones, J. Tangorra, E. Liu

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Total Home Automation and Health Care Consortium

March 31, 1998

SIMSUIT and Biochair Projects

Lynette Jones

Principal Research Scientist

James Tangorra

Graduate Research Assistant

Eric Liu

Undergraduate Research Assistant

Abstract: The SIMSUIT project has as its objective the development of a wearable, modular, health monitoring system that is can make measurements of a number of physiological variables including heart rate, blood pressure, respiration rate, and core body temperature. The monitoring systems must be lightweight, wireless, non-invasive and non-intrusive. The SIMSUIT is not only designed to measure these variables but also to evaluate the status of different systems by perturbing them and measuring the responses to these perturbations. It is often under these conditions of active stimulation that problems in the functioning of a system may first emerge. One aspect of the SIMSUIT project therefore, is to develop appropriate testing protocols and analysis procedures for evaluating different physiological systems. In this report the development of the vestibular-ocular testing apparatus will be described, together with the research being conducted on the development of an ambulatory blood pressure cuff based on shape memory alloy fibers and initial work on determining the characteristics of a wearable thermometer.

Vestibular-ocular Testing Device

A portable vestibular-ocular testing apparatus has been designed and is undergoing further development and testing. The apparatus is being developed with two objectives in mind. First, it is to be used as a clinical evaluation tool to examine the functioning of the human vestibular-ocular system, and second it is to serve as a device that can be used to measure the level of alertness in human operators controlling vehicles or machines.

A working prototype of a portable vestibular ocular reflex testing device has been completed (see Figure 1). The helmet-based apparatus uses stochastic system identification techniques to evaluate the performance of a test subject's vestibular ocular reflex (VOR) and combined head and eye gaze response. Small torque perturbations, less than 1 N-m in amplitude and at frequencies of up to 10 Hz, are delivered to the subject's head while he or she tracks a computer-controlled laser target. The protocol evaluates the VOR and gaze response under normal system operating conditions, where natural head and eye movements are used to follow the visual target.

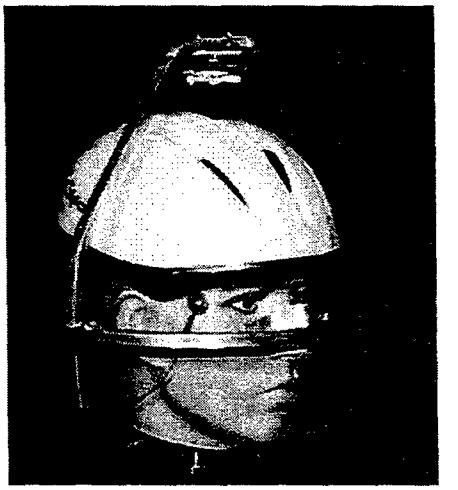


Fig 1: Portable VOR Testing Apparatus

Head movements are monitored with magneto-hydrodynamic rotational velocity sensors, and eye motions are recorded with an electro-oculograph (EOG) built using a signal conditioning board that was designed in the lab. The helmet's torque perturbations, and the visual target's trajectory are controlled with a Visual Basic 5.0 program. Data on head and eye movements are collected with a National Instruments data acquisition board, and stored in text files for later analysis.

System identification algorithms that analyze the head and eye movements and conduct a system level evaluation of the VOR are being developed with MathCad 7.0 software. A

multi-input, non-linear system identification algorithm is needed to assess accurately the VOR and gaze response systems. The VOR analysis algorithm is being developed by building upon algorithms developed for simpler single input, linear, and non-linear systems. A single input, single output, linear system identification algorithm was created that enables the helmet to be used for stochastic characterization of the rotational dynamics of the head and neck. A pseudo-random binary input torque is delivered to the head, and the impulse response function of the head and neck is estimated from the auto-covariance of the helmet's input torque, and the cross covariance of the perturbation torque and the resulting rotational velocity of the head (see Figures 2, 3, 4).

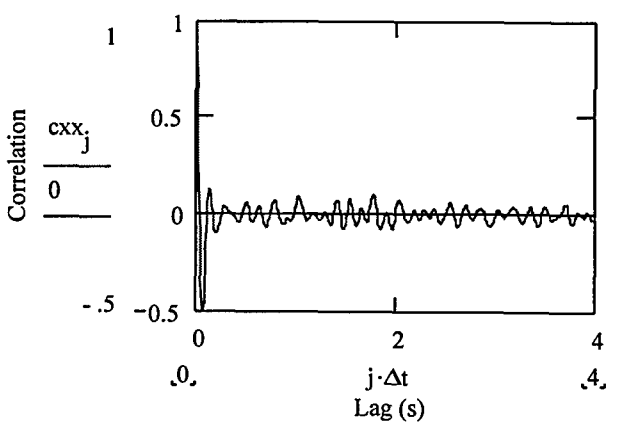


Fig 2: Auto Covariance of perturbation to head.

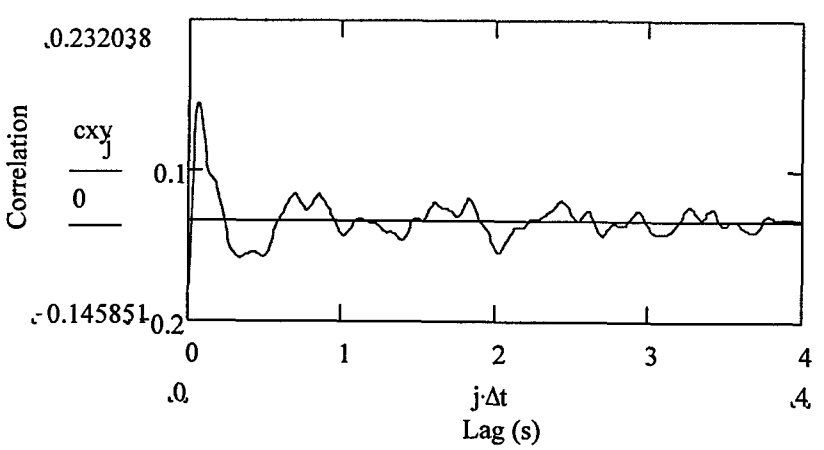


Fig 3: Cross Covariance of head perturbation and head movement

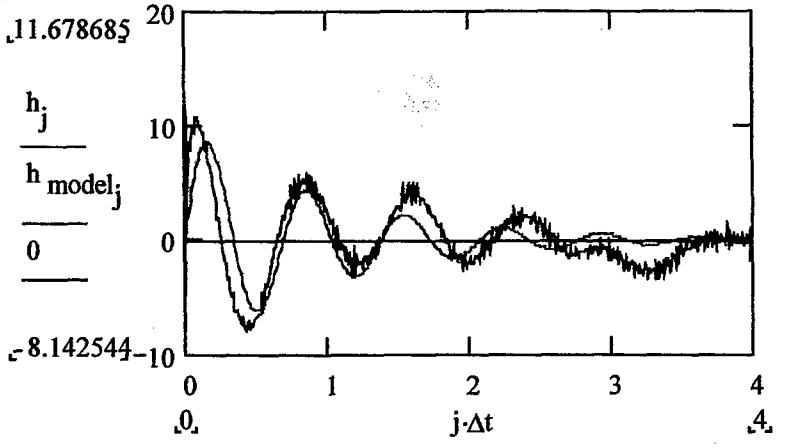


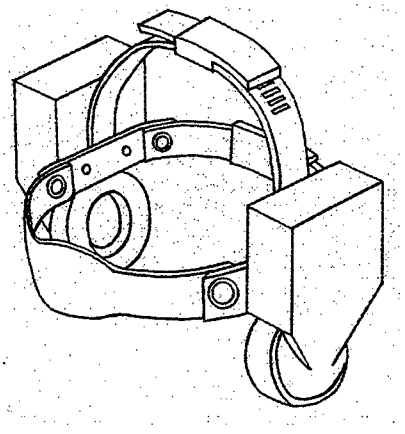
Fig 4: Impulse response of the head and neck (red) and second order linear model fitted to impulse response (blue).

A parametric model of the head and neck is made by fitting a second-order model to the impulse response with a Levenberg-Marquardt sum of squares minimization technique. The linear stochastic analysis and second order model does not result in a very accurate estimate of the head and neck dynamics, but it does produce a rough estimate of the system parameters. As with many physiological systems, in which non-linearities enable the system to perform as required, it may be simplistic to assume that the head and neck can be approximated by a simple arrangement of idealized linear components. The neck's damping and stiffness probably change with head position and the rotational velocity of the head, and may be better evaluated using non-linear techniques. A non-linear analysis technique is being explored at present to see whether it provides a better estimate of the head and neck's response function. Once a non-linear algorithm can be satisfactorily applied to the single-input, single-output head and neck system, it will be extended to analyze the multi input VOR system.

The equipment and protocols being developed are intended to be tools for clinical VOR evaluations as well as for routine, unobtrusive, testing in industrial environments. It is suspected that a correlation exists between the level of alertness of a human operator and the performance of the vestibular system and in particular the VOR. Periodic system level testing of the vestibular system may give information about the changes in a person's attention throughout the workday, and may provide insight as to when a person is at their optimum in terms of alertness. The helmet-based perturber and external visual target work well in a laboratory environment where dedicated space is available for the apparatus, but may not be ideal for an industrial environment. To get an uncorrupted view of the vestibular system's performance

performance during the day, it is necessary to test a person without changing their mental set to a "test-mode". Ideally, the test equipment should be part of their work uniform and the person should not have to stop work to put on a piece of equipment, such as the helmet perturber, or have to leave the work environment to go to an evaluation area with a visual target. A second prototype of the vestibular ocular testing device is therefore being developed and built to replace the first prototype shown in Figure 1. The new system is a headphone-based apparatus that combines a head perturber with a 3-D acoustic target presented by earphones as shown schematically in Figure 5.

Fig 5: Headphone Perturber. Perturbation modules located above each earphone



The headphones and the perturbation modules attach to an adjustable headband, similar to those worn by surgeons or found inside safety helmets. The headphones are comparable to those worn by military pilots and aircrew, in that they provide hearing protection as well as being capable of producing sound. With this prototype, the perturbation torque is created by accelerating masses back and forth inside each perturbation module. An electric coil is permanently fixed inside each module and permanent magnets, attached to slider bars, are accelerated by driving current through the coil. Different configurations of the Alnico and ceramic permanent magnets are being explored to limit the overall weight of the device and to maximize the reaction force. Control of the perturbation torque will be through a Visual Basic 5.0 program.

A considerable amount of development is required to create a realistic 3-D acoustic target that results in a test subject consistently directing his or her eyes to specific locations. At

present, a significant amount of research is being done in this area by companies involved with the development of virtual reality systems. None of the commercially available systems that we have evaluated are adequate for our testing protocol, but it is believed that it will be more efficient to wait for a commercial product, than to develop a 3-D auditory target system in-house.

Ambulatory Blood Pressure Monitoring Device

Ambulatory blood pressure monitoring devices are used at present in a number of situations, and are most frequently used to evaluate the efficacy of medications prescribed for hypertension, or for monitoring episodic hypertension. Many of the existing devices are cumbersome to wear and the act of inflating and deflating the cuff is noisy and obtrusive. There are many other situations in which regular and non-intrusive monitoring of blood pressure would be extremely desirable given the risks associated with elevated blood pressure, a known risk factor for cardiovascular disease, and the difficulties associated with acquiring sufficient data to make a reliable diagnosis.

A lithium-ion battery-powered ambulatory blood pressure monitoring device is being developed that will be incorporated into the sleeve of the SIMSUIT above the elbow (Provisional patent application: 60/072230; filed 1/23/98). Blood pressure will be measured by occluding the brachial artery just above the elbow joint which is the conventional position for making blood pressure measurements. This site has been selected in preference to the wrist or finger in order to minimize the interference of the cuff with movements of the arm and hand. The goal of non-intrusiveness is an extremely important criterion for systems incorporated into SIMSUIT as it is envisaged that it will be worn in some situations for prolonged periods of time. The cuff is made up of shape memory alloy fibers that are embedded in an insulating sleeve and contract and relax in response to constant current pulses. Research to date has been devoted to determining the characteristics of the occlusive part of the blood pressure monitoring system, which is based on nickel titanium (NiTi) fibers.

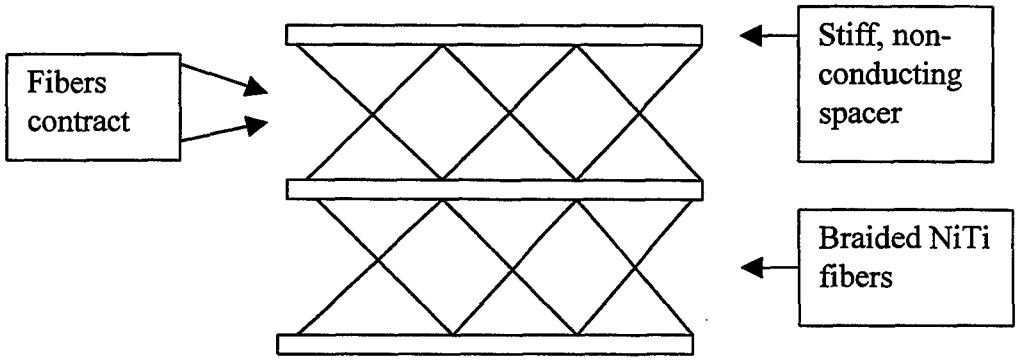
Shape memory alloy fibers such as NiTi undergo a structural transition when they are heated from a martensite phase to an austenitic phase, which is characterized by a higher crystalline symmetry. With cooling they return to the martensite phase. An external stress is applied to the fibers in order to achieve the shape memory effect and they are heated by passing a current through them (Joule heating). The efficiency of these fibers is usually found to be less than 3%, when calculated as the ratio of the work done by the fiber to the input electrical energy. There are several features of shape memory alloy fibers that make them extremely attractive in this medical application, namely that they are compact, lightweight, and have excellent power-to-mass ratios. NiTi fibers generate much larger forces per cross-sectional area than any other actuator technology, with peak stresses of approximately 200 MPa.

NiTi fibers are biocompatible, but because the transition temperature is 70 deg C the fibers cannot be in direct contact with the skin (which burns at temperatures above 45 deg C) and so there must be suitable electrical insulation between the skin and the fibers. Another disadvantage of these fibers is that their contractions result from changes in their bulk material properties, and so contractions are limited to approximately 8% of their total length (strain). A further limitation is that their total lifetime, defined in terms of the number of contractions that can be obtained before the response amplitude diminishes or fails altogether, is shortest for large contractions, which restricts their use in many applications. This is not a major impediment to their use in a blood pressure cuff as measurements of systolic and diastolic blood pressure are not taken continuously, and so high efficiency is not essential, and in many situations measurements will be taken at relatively long time intervals. It is also possible to overcome this limitation by building redundancy into the design, which can be achieved by cycling through a number of fibers. The extremely small diameter and weight of these fibers means that it is possible to mount a number of them in a small workspace.

Shape memory alloy fibers such as NiTi can contract very rapidly but the relaxation speed is usually limited by heat dissipation from the fiber. It can take over 300 ms to return to 50% of the maximum strain and full recovery can take over 1 second. This slow relaxation time and hence low bandwidth has limited their use in many potential applications (e.g.

robotics). A process has been developed by Hunter and Lafontaine (U.S. Patent Number: 5,092,901) that uses very brief current pulses (1-10 ms) while the fiber is stretched which modifies the NiTi fiber so that it can contract and elongate more rapidly. After this conditioning, NiTi fibers recover 50% of the strain in less than 40 ms and the fall time (10-90%) is less than 130 ms. This process is being used in the NiTi blood pressure cuff.

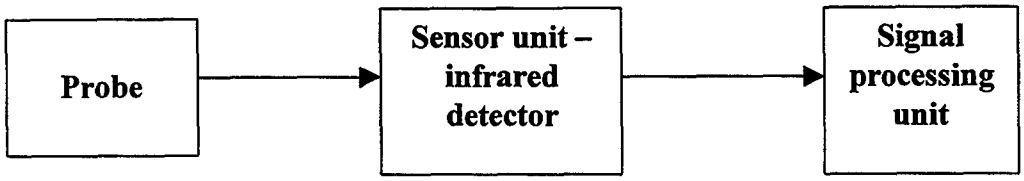
A number of experiments have been conducted in which measurements have been made of the contraction and relaxation times of NiTi fibers to establish that they can contract and relax within the time intervals mandated for ambulatory blood pressure monitoring systems. Such systems are required to allow pressure to be controlled and maintained at a rate between 1.0 mm Hg per second and 20 mm Hg per second from initial differential pressures of 250 mm Hg, 150 mm Hg and 50 mm Hg. The release rate of the cuff is supposed to allow a pressure of 250 mm Hg to be reduced to a pressure of 20 mm Hg in 4 seconds. The forces produced during contraction have been measured in order to determine which configuration of the NiTi fibers is optimal for occluding the brachial artery. Given cuff dimensions of 0.08 m and 0.2 m, it is estimated that the maximum pressure required is 38,700 kPa (290 mm Hg) which would occur with approximately a 10% contraction in the fibers. With single fibers in a range of configurations the maximum contractions achieved have been 4-5%, however, with NiTi fibers braided (as shown below) the contractions are now in the order of 7-8%. Once the design of this aspect of the device is finalized, blood pressure itself will be determined on the basis of Korotkoff's sounds (auscultatory method) detected using piezoelectric microphones and from oscillations in flow in the brachial artery which will be transduced using infra-red detectors (oscillometric method).



Wearable Infrared Tympanic Thermometer

A number of infrared tympanic thermometers have been developed and marketed over the past 8 years which have been widely accepted in pediatric care and emergency room medical treatment due to their ease of use, lack of distress to the patient and the brief period of time required to take a thermal measurement (Betta et al., 1997). These thermometers use an infrared sensor such as a pyroelectric element (Thermoscan, E-Z Therm) or a thermopile (FirstTemp Genius) to detect the infrared radiation from the tympanic membrane in the auditory canal. As the tympanic membrane shares a vascular supply with the hypothalamus it is considered an excellent site for measuring a temperature that is thought to be a reliable indicator of core body temperature. Aural temperatures are, however, different from temperatures recorded at other body sites such as the armpit, rectum and mouth and so some of the existing infrared thermometers have the facility of converting the aural temperature into an equivalent axillary, rectal or oral temperature. The algorithm implemented by the signal-processing unit usually just adds a fixed offset which ranges from 0.4 °(oral) to 0.8° C (rectal). The error of measurement that is accepted from existing devices is $\pm$ 0.1° C over a temperature range of 37-39° C and $\pm$ 0.2° C when temperatures are between 36-37 ° C or 39-41° C.

There are three basic components to an infrared tympanic thermometer: a probe that directs the infrared radiation from the thermal target to the infrared sensor; a sensor unit consisting of an infrared detector that converts the thermal radiation into an electrical signal and a signal-processing unit (as shown below). The selection of an infrared sensor for a wearable infrared thermometer is based on a number of considerations, including coast, reliability, accuracy and susceptibility to fluctuations in ambient temperature. In general, infrared tympanic thermometers are susceptible to errors if the ambient temperature is outside a specified range which is usually between 15-40° C. This means that a wearable device based on thermal infrared sensors will not be able to be used in extremely hot or cold environments. Some existing infrared thermometers are unable to measure body temperatures lower than 34° C or higher than 41.1° C.



We are studying a number of infrared detectors to establish which would be optimal in this application. The signal processing unit and power supply will be mounted in the headset that the person wears and the probe will protrude from one of the ear phones. Although such a system would impair hearing due to its position in the ear canal, it is not envisaged that core temperature would need to be recorded continuously except under unusual conditions during which loss of hearing in one ear may not be detrimental to the person's functioning.

References:

Betta, V., Cascetta, F., Sepe, D. (1997). An assessment of infrared tympanic thermometers for body temperature measurement. Physiological Measurement, 18, 215-225.

Hunter, I.W., & Lafontaine, S.R. (1992). Shape memory alloy fibers having rapid twitch response. U.S. patent 5,092,901.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Patient Monitoring and Diagnosis

CHAPTER 6

An Intelligent Cardiopulmonary System for use in the Care of the End Stage Cardiac Patient: A Concept Paper

T. Sheridan, J. Thompson

d'Arbeloff Laboratory for Information Systems and Technology
MIT

An Intelligent Cardiopulmonary System for use in the Care of the End Stage Cardiac Patient: A Concept Paper

Thomas B. Sheridan
Principal Investigator

James M. Thompson
Co-Principal Investigator

1. INTRODUCTION

The goal is this project is the development of a system designed to acquire, process and analyzed blood pressure, heart rate, oxygen saturation, and thoracic signals to make a decision on the relative health of the patients cardiovascular system. Unfortunately most patients do not have the tools necessary to become an active and informed participant in their own health care. Many of those who end up with serious health problems enter the health care system too late, and thus require more extensive and costly care. The patients selected for monitoring by the intelligent system are those patients who are at a higher risk for decompensation as compared to the general population. These "high risk" patients frequently enter the health care system too late and thus require more extensive and costly care in addition to the emotional and physical strain to themselves and their families. The goals of this program are to decrease the initial acuity, length of hospital stay and readmission rates for patients with congestive heart failure. This will result in substantial savings in health care costs with a decreased burden on the acute health care system.

Why focus on the cardiovascular and pulmonary system? It is estimated that 65 of 239 million Americans have cardiovascular disease. One million die annually, and this is one

of every two deaths in the United States. The mortality from cardiovascular disease exceeds that of all other diseases combined. Congestive heart failure (CHF) is estimated to affect three million people in the United States. It is the final pathway of a variety of primary cardiovascular disease entities, such as coronary artery disease, hypertension, valvular heart disease, genetic disorders, diabetes and the sequelae of infection or toxin exposure, among others. Hospitalizations and mortality from CHF have increased steadily since 1968, despite the overall improvement in mortality from cardiovascular disease. Heart failure is now the underlying cause of death in over 39,000 persons annually. In 1992, it was the first listed diagnosis in 822,000 persons and is the most common hospital discharge diagnosis in persons over 65 years of age. The incidence of death from CHF is 1.5 times as high in black Americans as in whites. The estimated direct economic cost of CHF in the United States be reported to be $10.2 billion annually. The problem will only get worse, as the elderly segment of the population is increasing at a rate 5.6 times that of the other age groups. There are currently 25 million Americans greater than 65 years of age and 2.7 million Americans greater than 85. Over the next 50 years the >65 age group will see a 140% increase versus 25% in the other age groups. At present, the only cure for end-stage CHF is cardiac transplantation.

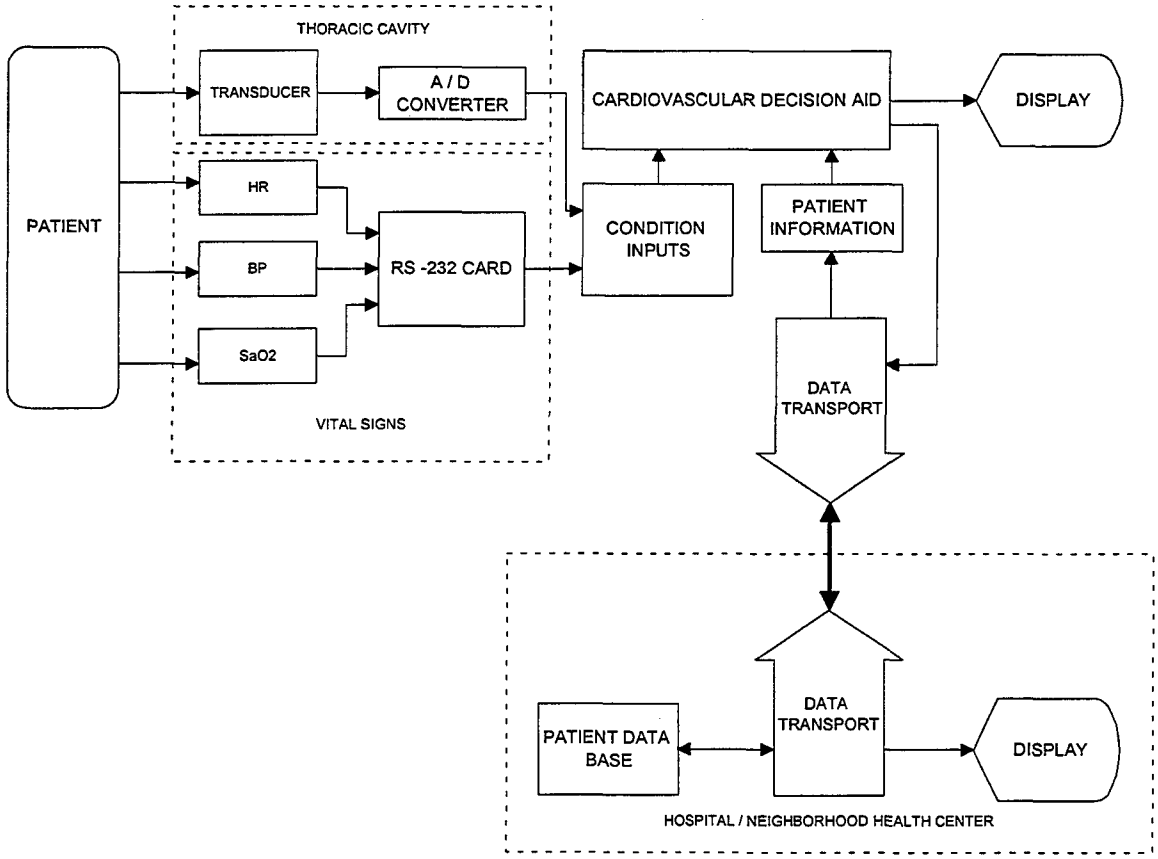
Studies have shown that intervention can improve care and decrease costs by decreasing hospital admissions, which account for a large portion of their health costs. Investigators in Los Angeles found that interventions (invasive tests, medication adjustment, patient education and follow-up at a heart failure center) decreased the number of hospital admissions from 429 in the six months before referral to 63 in the six months after referral.[1]

2. PROPOSED SYSTEM

As stated earlier, the goal is this project is the development of a system designed to acquire, process and analyzed blood pressure, heart rate, oxygen saturation, and thoracic signals to make a decision on the relative health of the patients cardiovascular system. The patients selected for monitoring by the intelligent system are those patients who are at a higher risk for decompensation as compared to the general population. These "high

risk" patients frequently enter the health care system too late and thus require more extensive and costly care in addition to the emotional and physical strain to themselves and their families.

We propose to produce two types of intelligent systems. The emphasis will initially be placed on a portable system that will be carried by the home health care professional on a visit to the patients' home. A second system will be a permanent home based system for use by the primary caregiver and / or fragile (CHF) patient. In either scenario, the operator of the Intelligent Cardiopulmonary Decision System (ICDS) will be directed by the ICDS on where to place various sensors and what measurements to take. The ICDS



will then process the data and make a recommendation to the patient concerning further care. There will be human factor issues on the user interface, as well as some type of patient feedback so that they can actively participate in their care.

3. SYSTEM DESIGN

The structure of the ICDS is shown in figure 1. The patient's vital signs and oxygen saturation will be acquired first and will be evaluated by the ICDS. The ICDS will then direct the acquisition of other information as needed by the system. This additional information will be conditioned, digitized, and processed before being put in a form that could be inputted into the ICDS. The ICDS will then make an initial assessment of the patients current state, compare it with a predetermined "optimal state" and make a decision on where to proceed from that point. Options include requesting additional information from the patient, patient education, instructions to hold or to take an additional dose of a medication, decision to re-evaluate after a waiting period, contact the primary care physicians office, connect to the central system, or call an ambulance for transportation to a medical facility.

Figure 1. Structure of the Intelligent Cardiopulmonary Decision System (ICDS)

An outline of the system follows.

A. CLINICAL INPUTS

The inputs into the system will be thoracic acoustic signals, heart rate, blood pressure, and oxygen saturation. This is in addition to information about the patient present in the patient information system.

1). Thoracic signals

We will be using various transducers, placed under the direction of the ICDS, to acquire the signal of interest for that specific patient. The signal will undergo conditioning and be put into a form useful for the expert system.

2). Heart rate

The heart rate will be acquired from a standard portable monitor (via RS-232 input) and used by the ICDS as an additional piece of information about the patient's current state. Heart rate is important in patients' who currently have marginal coronary flow and are sensitive to the physiological consequences of tachycardia. Tachycardia decreases coronary diastolic filling time, which decreases the supply of oxygen to myocardial tissue, especially endocardial. In addition, tachycardia increases oxygen demand, which further contributes to

negative myocardial oxygen balance. This initially results in regional wall motion abnormalities, which causes a rise in ventricular end diastolic and end systolic pressures, which further decrease diastolic blood flow, starting the cycle to heart failure. Bradycardia can also have deleterious effects on certain pathologic states. Patients with mitral or aortic regurgitation can go into congestive heart failure, depending on the magnitude of the regurgitant fraction and the degree of bradycardia. Changes in the other inputs would affect the magnitude of the changes in heart rate that would start the cycle toward CHF.

3). Blood pressure

Both an increase and a decrease in blood pressure can have an effect on cardiovascular dynamics that would have a deleterious effect on cardiac patients. Certain types of congestive heart failure are sensitive to changes in afterload, and the presence of blood pressure changes in these patients could start the process toward congestive heart failure.

4). Oxygen saturation.

B. DATA ACQUISITION

1). Inputs

There will be two types of information obtained from the patient. The first will be vital signs, which can be obtained from standard portable monitors. Oxygen saturation will also be acquired. Additional thoracic signals will be acquired through transducers as requested by the ICDS.

2). Signal Conditioning

3). Data Acquisition Hardware

C. SIGNAL PROCESSING

D. OUTPUT CONDITIONING

E. MEDICAL MODEL

1). Past Medical History

2). Past Surgical History

3). Weight

4). Cardiovascular Evaluations

a). Noninvasive Methods

i). Echocardiography

ii). Ultrafast Computed Tomography

iii). Radionuclide Angiography

iv). Gated Magnetic Resonance Imaging

b). EKG

c). Arterial Blood Gas

d). Cardiac Hemodynamic Data

i). CVP, PCWP, RVEDP, RVEDV, PAP, LVEDP, LVEDV, EF, CO, mVO_2.

e). Thallium Scan

f). Hemoglobin

g). Left Ventricular Pressure-Volume Loop

h). Quantitative Angiocardiography

5). Medications

a). Types of Agents

b). Drug Effect Transducer

6). Other Therapeutic Agents

a). Oxygen (FiO_2)

7). Pathophysiological Issues

a). Increased Afterload

b). Parasympathetic inhibition

c). Sympathetic activation

d). Frank-Starling Mechanism

e). Neurohumoral stimulation

f). Changes in Myocardial beta-Adrenoceptor Density

g). Myocardial energy requirement

h). Salt and Waster Retention

i). Arterial Vasodilation

j). Venodilation

k). Endothelial dysfunction

l). 2,3-Diphosphoglyceride activity

m). Vascular Wall Changes

n). Ventricular Dilation

o). Ventricular Hypertrophy

p). Diastolic dysfunction

F. CARDIOPULMONARY DECISION AID

1). New Clinical Inputs

The most recent clinical inputs (thoracic signals, heart rate, blood pressure, and oxygen saturation) are analyzed and considered by the ICDS.

2). Other Inputs

a). Temperature

3). Integration with the Medical Model

4). Determine the Current State of the Patients Cardiopulmonary System

5). Compare the Patients Current State with a Predetermined "Optimal State".

6). Ascertain What Maneuvers Exacerbate Patients Condition

7). Determine if Intervention is Necessary

8). Determine the Type and Scope of Intervention

9). Evaluate Patients Response to Therapeutic Interventions

G. HUMAN FACTORS ENGINEERING

1). Clinical Interface for Patient and Primary Caregiver

2). Clinical Interface for Health Care Worker

---

[1] Fonarow GC et al. *Impact of a comprehensive heart failure management program on hospital readmission and functional status of patients with advanced heart failure.* J Am Coll Cardiol 1997 Sept; 30:725-32.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Patient Monitoring and Diagnosis

CHAPTER 7

Using HVAC Systems for Cardiovascular Stress Tests in the Home: Initial Modeling and Experiment of Coupled Cardiovascular/Thermoregulatory Dynamics

B. Gu, H. Asada, S. Liu

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Using HVAC systems for Cardiovascular Stress Tests in the Home: Initial Modeling and Experiment of Coupled Cardiovascular/Thermoregulatory Dynamics

Bei Gu
Prof. Harry Asada
Dr. Sheng Liu

Abstract

This research is aimed to develop new methods of identifying key parameters in human physiological models based on measurement of cardiovascular activities in response to thermal environment variation. The main idea is to thermally excite human cardiovascular responses in an amiable and non-intrusive manner so as to identify human model parameters that are otherwise not identifiable in a stationary condition. In this report, a model describing the dynamic behavior of human cardiovascular and thermoregulatory systems is presented. Preliminary experimental data are also included to verify the effectiveness of this physiological model.

1. Introduction

Common cardiovascular diseases include heart disease, cerebrovascular disease, hypertension and atherosclerosis. Cardiovascular diseases account for 43 per sent of annual mortality in the United States [1]. More than one in four Americans suffer from cardiovascular diseases at an estimated cost in 1994 of $128 billion in medical expenses and lost productivity. When the average age of population gets older and older, these diseases become more dominant in the health related problems in our society. It is well known early detection and diagnosis can substantially increase the treatment success rate of these diseases.

1.1 Current Methods of Cardiovascular System Evaluation

Blood pressure is a crucial piece of information to assess the cardiovascular system. A test of cholesterol concentration can show the risk of arteriosclerosis. If a patient shows the sign of cardiovascular disorder, a treadmill stress test will be suggested to detect coronary problems. Pharmacological stress test is a substitute for the exercise stress test. The basic idea of the stress tests is to increase the load on the heart and to determine its maximum pumping capacity.

1.2 Models of Circulatory and Thermoregulatory System

There have been a number of mathematical models, including lumped parameter and distributed parameter models, for characterizing cardiovascular and thermoregulatory systems. Hale began the circulatory system modeling with a windkessel (air chamber) model in 1769. He used compartments to represent ventricles, blood vessels, and viscera. Between the elastic chambers, blood is flowing through resistive tubes. Therefore the entire cardiovascular system resembles an electrical circuit of resistors and capacitors. Lumped parameter models for the entire cardiovascular system range from simple 5 or 6 compartments to hundreds of compartments depending on the purpose and accuracy of simulation. The RC circuit type lumped parameter model is quite successful in revealing the fundamental dynamic characteristics of human circulatory system. In recently years, some distributed parameter models have been developed. These models can provide more detailed information that lumped parameter models overlook, such as pulse wave transmission, etc. Ozawa [6] developed a comprehensive distributive cardiovascular model at the MIT Fluid Mechanics Lab.

For physiological models of thermal regulation, the simplest form is a two-node, lumped parameter thermal capacitance model that divides the human body into a core region and a shell region [2]. More complex models including finite element and finite difference models are also available [7]. These models can be used to simulate the human dynamical response to environmental temperature change. Some can also provide cardiac output prediction based on the corresponding metabolic rate [2].

Despite numerous models reported in the literature so far, most of these models consider the circulatory and thermoregulatory systems separately. Although some models do

consider the interaction between two systems, none can effectively capture the dynamic characteristics of the interaction.

1.3 Research Objective

The main objective in this research is to develop an effective, non-invasive cardiovascular test system based on thermal excitation generated by the home heating, ventilation, and air conditioning (HVAC) system. As opposed to most cardiovascular stress tests that require elaborated exercise and invasive measurement and must be performed in clinical facilities, this new method can be implemented in the home with little human assistance.

At the current stage of the research project, an analytical model that explicitly describes the cardiovascular responses to the thermal excitation has been developed. In this report, derivation of this model that characterizes both human cardiovascular and thermoregulatory systems is presented. Preliminary experimental results are also included. It is proposed that this new cardiovascular/thermoregulatory model will be utilized for processing physical signals of human for early diagnosis of cardiovascular problems.

2. Main Idea

Since cardiovascular system and thermoregulatory system are strongly coupled, changes in thermal environment that affect the human thermoregulatory system can also excite dynamic responses of the cardiovascular system. Therefore, it is proposed that the cardiovascular system can be tested for any ailment based on its dynamic response to a well controlled thermal excitation. The main advantage of this method is that elderly people being tested do not need to actively perform a certain exercise to undergo a stress test. The cardiovascular evaluation is done in a non-invasive and non-intrusive manner. This method can be implemented in the home so that the monitor and test are continuous as shown in Figure 1.

Compared to the traditional methods that stress the circulatory system to its full extreme, this method does not apply excess stress on human body. The main challenge of this new method is that it does not capture the information of highly stressed

cardiovascular responses critical to the diagnosis. To compensate for the lack of this information, patients are tested and monitored continuously. More importantly, this method essentially utilizes multiple new sensors and models of cardiovascular and thermoregulatory systems for interpreting sensor signals. This new technology will make circulatory assessment possible without stressing the patients excessively.

2.1 Coupled Dynamics of Cardiovascular and Thermoregulatory Systems

Human body is a complex system consists of many subsystems, such as: cardiovascular system, thermoregulatory system, neural system, muscles, etc. Among these subsystems, some of them are highly coupled. For example, the thermoregulatory system regulates the body temperature by means of circulatory system, and therefore circulatory system abnormality may results in failure in thermoregulation. On the other hand, any failure in the thermoregulatory system certainly can impact on circulatory system. Higher mortality rate of cardiovascular diseases under extreme weather is an example of this coupled complexity.

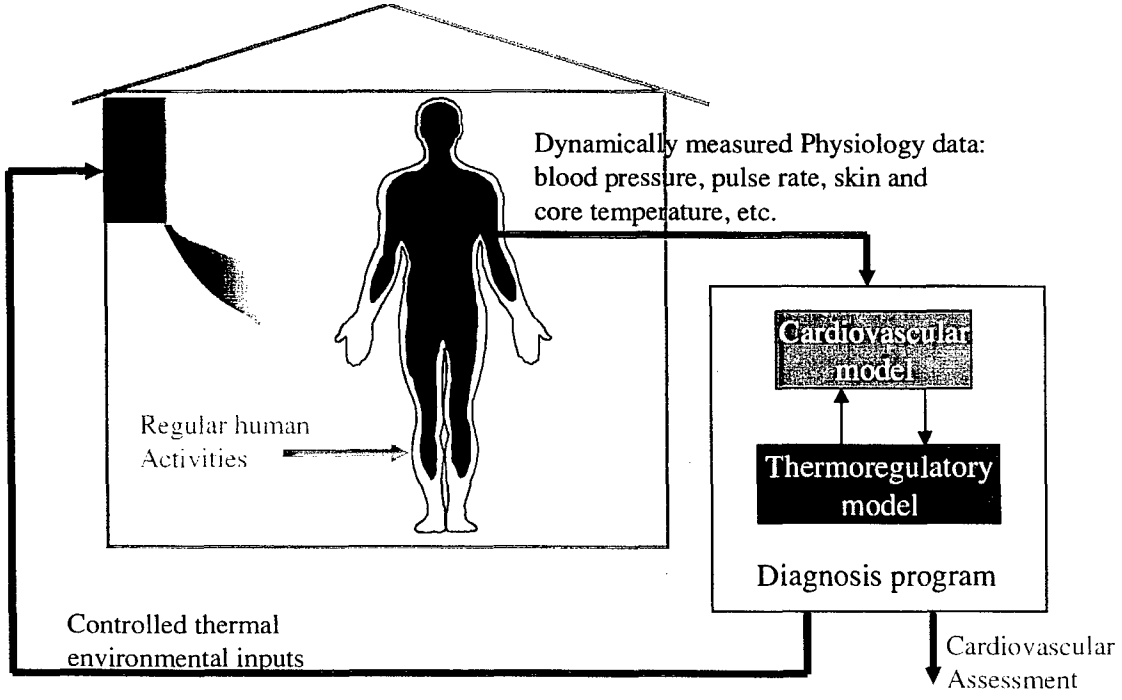


Figure 1: Thermal Excited Model Based Cardiovascular System Evaluation System.

From the energy point of view, the inputs to the thermoregulatory system are the metabolic heat generation and temperature of the environment. The body can be treated as a core region with internal heat generation and a shell region as shown in Figure 2. Between the core and shell, there are two means of heat transfer. The mode with large heat transfer capacity is the heat convection by the blood flow circulation from the core to the shell. The convection mode is important not only because of its higher rate of heat transfer but also its controllability by human autonomic control system. Human body controls the blood flow from the core to shell efficiently so that it constantly balances the heat generation and dissipation rates. Clearly, blood flow circulation plays a key role in thermal regulation of a human body. The other heat transfer mode is conduction through tissues from the core to the shell. The heat transfer rate in this mode is almost invariant.

The cardiovascular system consists of heart, blood vessels and other organs. This system can be well represented by a lumped parameter system if detailed convective processes can be ignored. The inputs to this system are the fluid flow pumped by the heart and ambient pressure. The blood flows through a series of resistive and elastic elements. Among all blood vessels, the subcutaneous vessels carry blood from the core to the shell of human body. Human body controls the shrinkage and expansion of these vessels to modulate the convective heat transfer rate from the core to shell. Therefore the circulatory system changes its properties when the human is subjected to different thermal conditions.

2.2 Cardiovascular Response to Thermal Excitation

Dynamic characteristics of physical systems can be best studied by exciting the systems and observing their transient responses. A lot more information about a system can be drawn from its dynamic behavior than from its steady-state condition. Applying external excitation is a common practice in engineering field for system identification. The treadmill stress test for evaluation of the cardiovascular system is a concept of excitation as well. When a person is engaging in an intense physical work, muscles consume far more nutrition than under the resting state. The body core temperature raises due to the heat dissipated from these working muscles and organs. Both of the nutrition requirement and raised core temperature request the circulatory system to supply more

blood to accommodate their needs. Thus the heart is pumping more and more blood. The heart itself needs nutrition as well. As it pumps more to the circulatory system, it consumes more oxygen. When coronary arteries fail to supply enough blood, the heart may soon fail. The stress test pushes the heart close to the edge of failure so test providers can observe substantial changes in the electrocardiograph signal.

The method to apply stress to the heart is not unique. The pharmacological stress test is using drugs to excite the heart and detect problems. Pure thermal excitation is also possible to generate the same effect as the treadmill test, since thermal excitation can effectively raise the core body temperature. To generate the same intensity of cardiovascular stress as in the treadmill test, extreme thermal condition must be applied, which will cause severe discomfort, and with its risk on the heart, it should be performed only in clinical facilities by professionals.

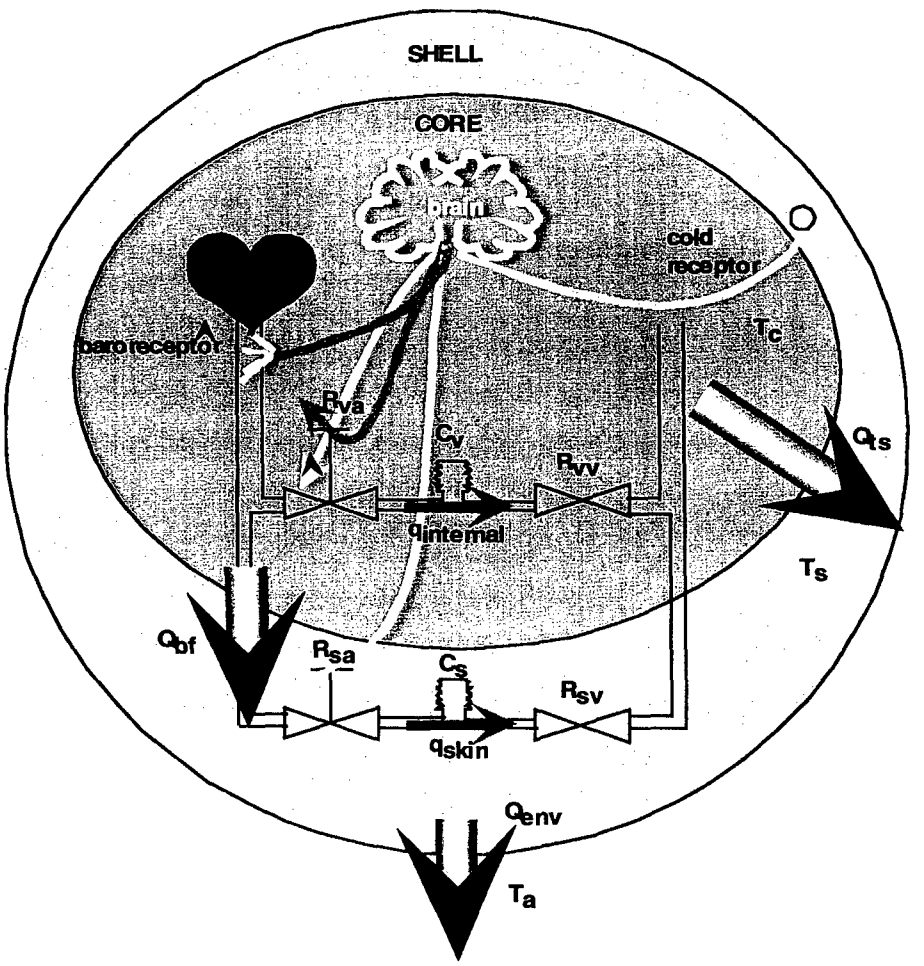


Figure 2: Coupled Cardiovascular and Thermoregulatory System.

Since the objective of the consortium is to develop technology that can be implemented at home and with least management, we are interested in cardiovascular evaluation without the patient's active participation and can be continuously performed at home. Moderate thermal excitation will be suitable for this purpose.

The most important question we have to answer is whether a moderate thermal excitation can generate enough information for cardiovascular assessment. To seek the answer, we establish a coupled dynamic model of circulatory and thermal regulation as well as a thermal excitation test procedure. The test is performed dynamically and continuously. Both circulatory and thermoregulatory data are logged and analyzed. We can control a HVAC system that regulates the thermal environment to various conditions so as to excite cardiovascular responses. For example, consider a patient has a higher than normal cholesterol value. It can not be concluded the patient is healthy only depending on his or her cholesterol level. If the arteries of this patient are not stenosis, the patient is still healthy. Our new cardiovascular assessment method may provide information of the patient's arterial system. We change the properties of the patient's circulatory system by changing the ambient thermal environment while the responses of the patient are recorded. The dynamic responses to different thermal excitations will depend on patient's cardiovascular properties, such as arterial resistance and compliance. The responses can be examined and interpreted based on the use of the model. Responses to more severe conditions may be extrapolated from the test data. Suggestions for health care can be made based on these predictions.

3.1 Lumped Parameter Models of Cardiovascular and Thermoregulatory Systems

Although there exist many circulatory and thermoregulatory models that are effective to simulate complex human circulation and thermal responses, we start the modeling from the simplest form that can describe the fundamental phenomenon of interest. The most critical characteristic we want to capture is the blood pressure fluctuation during a thermal environment transient. For this purpose, we use lumped-parameter dynamic model to describe each system.

3.1 Energy Based Modeling

As mentioned before, one of the main goals of this research is to establish a model that can be utilized to process the physiological response data for diagnosis of cardiovascular system. In particular, well developed techniques of system identification and parameter estimation will be applied. Therefore, a key objective of modeling is to develop a model in standard form such that these advanced techniques are readily applicable. To this end, and energy based modeling approach is employed here.

All of the existing cardiovascular models do not explicitly specify what are the energy sources of the system. For example: the six-compartment model developed by Sah [9] treated the two ventricles as two modulated capacitors; blood vessels and viscera were modeled as capacitors and resistors. In that model, energy source was not explicitly stated, but hidden in the modulated capacitors. As a result, the differential equations derived based on this modeling method are not in the desired standard form.

One of the most common representation forms of system dynamics is the state space model that consists of a number of first order differential equations. In this research, the energy-based bond graph method is utilized that can readily lead to state-space models [8]. The bond graph model clearly conveys the characteristics of the coupled circulatory and thermoregulatory system. The order of the system, the energy sources and sinks, and the modulated dissipative elements can be shown systematically.

3.2 Circulatory System

Only systemic part of circulation is modeled in the current analysis. The pulmonary circulation is relatively independent of the thermal regulation. As shown in Figure 3, the blood flow starts from the left ventricle. Part of blood circulates through peripheral skin blood vessels to control heat dissipation, and the other part of blood goes through internal viscera. The systemic blood then flows to the right atrium that is the end of systemic circulation. Figure 3 presents the schematic of the cardiovascular system.

The left ventricle is treated as a flow source that regulates the blood flow rate. This is based on the function of the heart. The blood transports all nutrition that body needs. If the concentrations of nutrition are assumed constant, the blood flow rate determines the nutrition supply. The circulatory system has its resistance to the fluid

flow so that a pressure differential is need to push the blood at a given flow rate. For a required blood flow rate, blood pressure closely depends on the resistance of the circulatory systems. When the resistance increases, ventricles have to do more work to maintain a demanded blood flow rate, and the blood pressure become higher. This is consistent with the physiology nature of a common type of hypertension.

Since the blood is viscous, when blood flows through blood vessels, especially arterioles and capillary vessels, viscous effect causes flow resistance. Therefore the viscous effect is accounted for as resistors in the circulatory system model, as shown in Figure 3.

Arterial wall is made up of tissues with surrounding muscles. Blood vessels change their volume depending on blood pressure. This behavior is known as blood vessel compliance. The compliance effect is modeled as capacitive elements in the model.

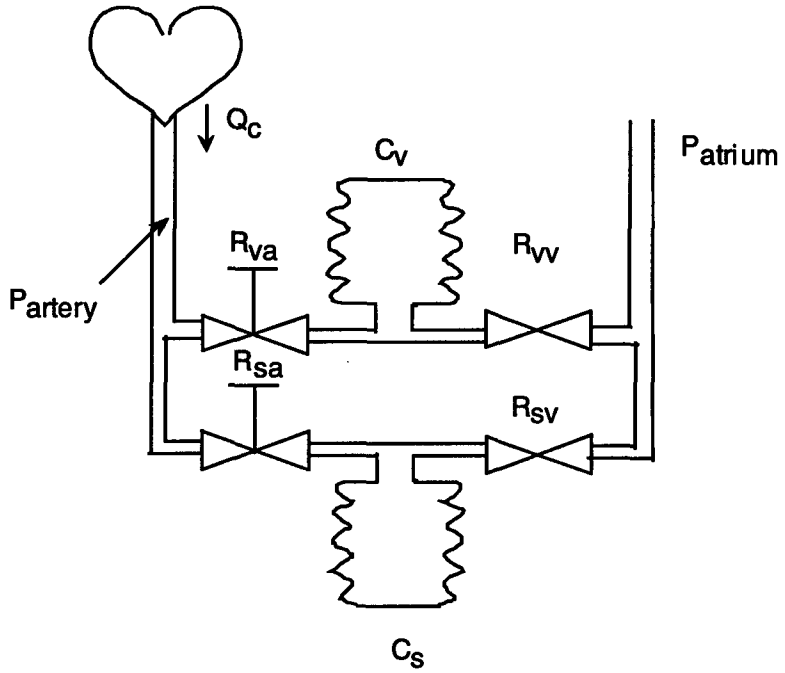


Figure 3: Schematics of Circulatory System.

At the right atrium, the pressure is low, and is approximately fixed at the extra cardiac pressure. Therefore, the right atrium is modeled as a constant pressure source. The bond graph of circulatory system is shown in Figure 4.

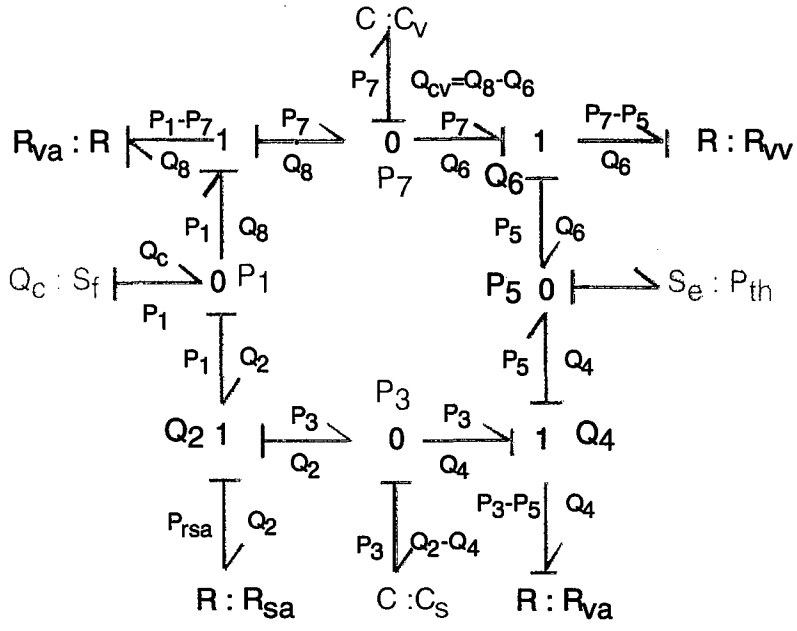


Figure 4. Bond graph representation of circulatory system.

3.3 Thermal Regulatory System

Among all lumped parameter and distributed parameter models, a two-node model developed by Gagge et al [2] is adopted for its ability to describe the system and its simplicity. The human body is lumped into two thermal capacitors: the body core and the body shell as shown in Figure 5. Some literatures referred to the capacitor as thermal mass. However, from energy-based modeling point of view, thermal capacitors are more appropriate. The capacitor integrates the input flow and provides an effort output. A thermal capacitor accumulates the entropy inflow and uniquely results in a certain temperature. In bond graph terminology, entropy flow rate is a flow variable, while temperature is an effort variable.

Figure 6 shows the bond graph representation of the dynamic system of thermoregulation. The energy source for the thermal regulatory system is the metabolic heat. The rate of metabolic entropy supply is modeled as a flow source. The effort sink of the thermal system is the ambient air temperature. The heat is dissipated through

convection from the skin surface to the ambient air. The ambient air temperature is independent of the human body.

Between the body core heat capacitor and the body shell heat capacitor, there are two means of heat transfer: heat conduction through tissues and heat convection by blood flow. The thermal resistance is assumed relatively constant because the properties of human tissue do not change fast enough compared to the time constant of the model system. The heat convection by blood flow is the major mechanism that a human body regulates its temperature. When the core temperature is higher than a certain value, the blood vessels in the subcutaneous region expand. The vasodilatation causes more blood flow from the core to the shell of body. The increased blood flow carries more heat from the core to the shell and the heat dissipation rate increases. When the human body is subjected to cold environments, the subcutaneous blood vessels contract. The vasoconstriction can cause the peripheral blood flow rate decrease to almost zero to maintain the core temperature. Therefore, the convection heat transfer is modeled as a resistance modulated by the peripheral blood flow which is a state dependent variable in the circulatory model.

The heat transfer between the skin and ambient air consists of convection, evaporation and radiation. The proposed test procedure will impose moderate thermal condition, heat lost due to evaporation can be ignored. The radiative heat transfer on subjects should not cause sweating and therefore is negligibly small. The convection heat transfer is modeled as constant resistance for simplicity.

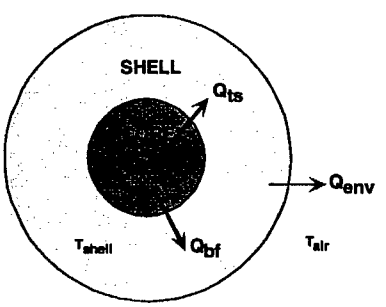


Figure 5: Two-node Model of Thermoregulatory System.

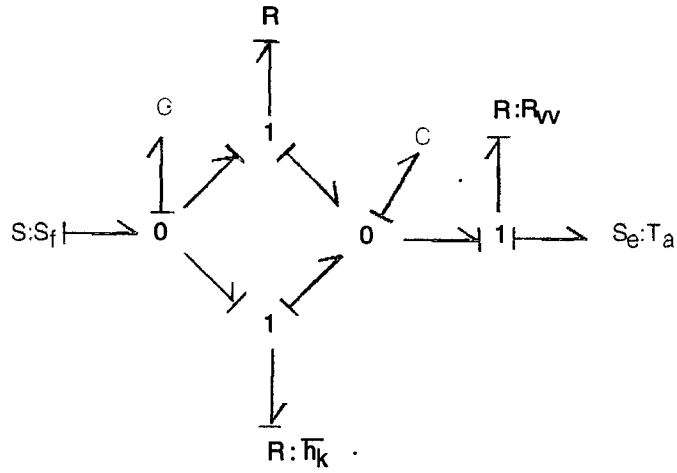


Figure 6: Bond Graph Representation of Thermoregulatory System.

3.4 Integrated Model and System Equations

In a human body, the central nerve system governs the operation of most of the body functions. In principle, there is the autonomic control system that coordinates all physiological control and the subsystems. Two autonomic controls are considered in current analysis: the body temperature control and the blood pressure control. Thermal receptors all over the body send the temperature information to the central nerve system. The nerve system adjusts the status of peripheral blood vessels to change the thermal resistance between the core and shell. This temperature control mechanism is represented in the model by the skin blood vessel resistance modulated by the error of core and shell temperatures. For regulating blood pressure, the baroreflex system senses the blood pressure and sends the signal to the nerve system and autonomic control system then controls the blood vessel resistance and ventricle contraction to maintain the blood pressure within the normal range. For example, when a person feels cold, vasoconstriction happens in his subcutaneous tissue, and the overall blood vessel resistance increases. This results in a higher blood pressure if the cardiac output remains unchanged. However, the autonomic system will expand the large blood vessels to decrease the flow resistance. The blood pressure then decreases back to the normal level. The particular baroreflex control considered here should not be confused with the cardiac

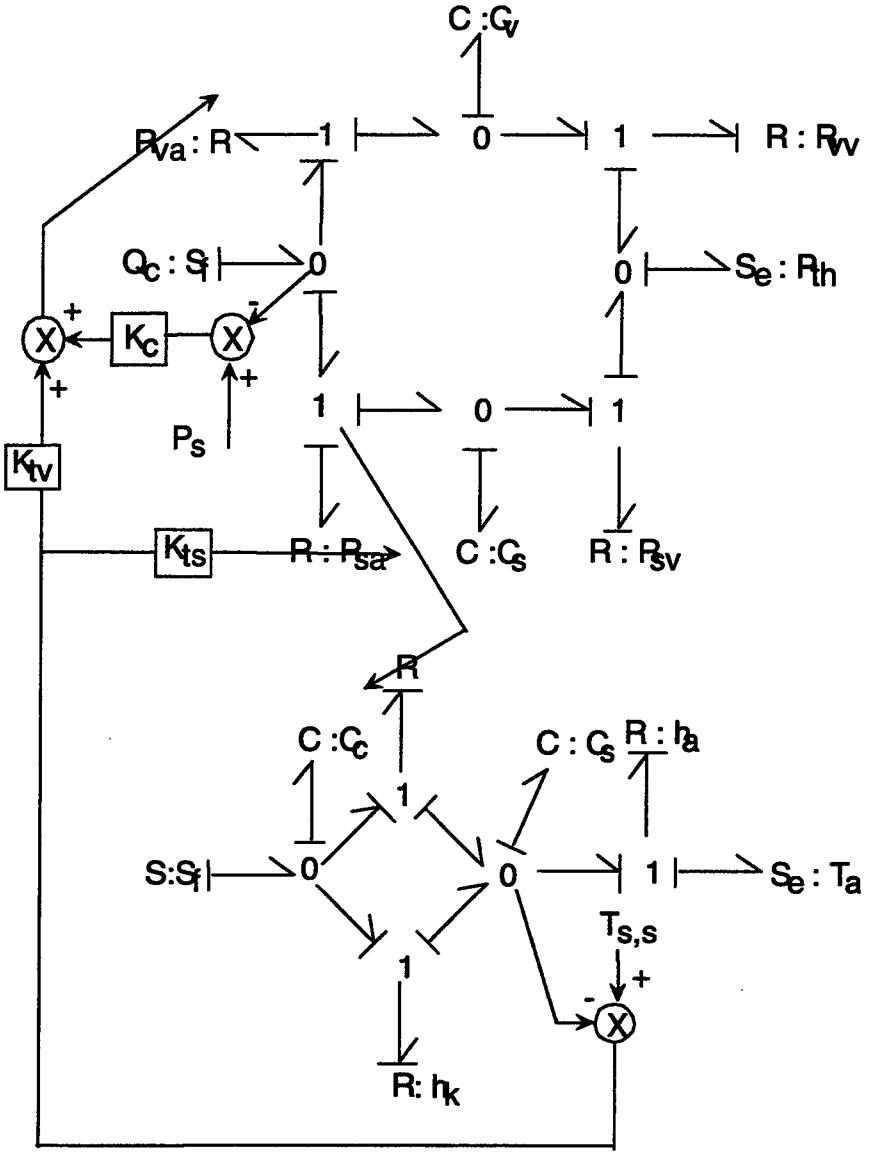


Figure 7: Bond Graph of Overall System with Autonomic Controls.

output and blood pressure interaction during exercises, which also causes the increase of both cardiac output and blood pressure.

The models of circulatory and thermoregulatory systems with the two autonomic controls can be integrated into a single model to explicitly characterize cardiovascular response to a thermal excitation. Combining bond graph models shown in Figure 5 and 6, the bond graph representation of this single model is shown in Figure 7. When a

human body is close to a thermally neutral condition, all parametric correlations can be approximated to be linear. The resultant model in a standard state space is then given as:

$$\begin{pmatrix}\dot{P}_s\\ \dot{P}_v\end{pmatrix}=\begin{bmatrix}-\frac{1}{C_s}\left(\frac{1}{R_{va}+R_{sa}}+\frac{1}{R_{sv}}\right) & \frac{1}{C_s}\frac{1}{R_{va}+R_{sa}}\\ \frac{1}{C_v}\frac{1}{R_{va}+R_{sa}} & -\frac{1}{C_v}\left(\frac{1}{R_{va}+R_{sa}}+\frac{1}{R_{vv}}\right)\end{bmatrix}\begin{pmatrix}P_s\\ p_v\end{pmatrix}+\begin{bmatrix}\frac{1}{C_s}\frac{R_{va}}{R_{va}+R_{sa}} & \frac{1}{C_sR_{sv}}\\ \frac{1}{C_v}\frac{R_{sa}}{R_{va}+R_{sa}} & \frac{1}{C_vR_{vv}}\end{bmatrix}\begin{pmatrix}Q_c\\ P_{th}\end{pmatrix}$$

$$\begin{pmatrix}P_1\\ Q_s\\ Q_v\end{pmatrix}=\begin{bmatrix}\frac{R_{va}}{R_{sa}+R_{va}} & \frac{R_{sa}}{R_{sa}+R_{va}}\\ \frac{1}{R_{sv}} & 0\\ 0 & \frac{1}{R_{vv}}\end{bmatrix}\begin{pmatrix}P_s\\ P_v\end{pmatrix}+\begin{bmatrix}\frac{R_{sa}R_{va}}{R_{sa}+R_{va}} & 0\\ 0 & -\frac{1}{R_{sv}}\\ 0 & -\frac{1}{R_{vv}}\end{bmatrix}\begin{pmatrix}Q_c\\ P_{th}\end{pmatrix}$$

$$\begin{pmatrix}\dot{T}_c\\ \dot{T}_s\end{pmatrix}=\begin{bmatrix}-\frac{\overline{h_k}+(\rho C_p)_{be}Q_s}{\overline{C_c}} & \frac{\overline{h_k}+(\rho C_p)_{be}Q_s}{\overline{C_c}}\\ \frac{\overline{h_k}+(\rho C_p)_{be}Q_s}{\overline{C_s}} & -\frac{\overline{h_k}+\overline{h_a}+(\rho C_p)_{be}Q_s}{\overline{C_s}}\end{bmatrix}\begin{pmatrix}T_c\\ T_s\end{pmatrix}+\begin{bmatrix}\frac{1}{A_{Du}\overline{C_c}} & 0\\ 0 & \frac{\overline{h_a}}{\overline{C_s}}\end{bmatrix}\begin{pmatrix}H\\ T_a\end{pmatrix}$$

$$R_{sa}=R_{sa,s}+K_{ts}(T_{s,s}-T_s)$$
$$R_{va}=R_{vas,s}+K_{tv}(T_{s,s}-T_s)$$
$$R_{va}=R_{va,s}+K_c(P_s-P)$$

$$\mathbf{X}=\begin{bmatrix}P_s & P_v & T_s & T_c\end{bmatrix}^T$$
$$u=T_a$$
$$\dot{\mathbf{X}}=f(\mathbf{X},u)$$
$$P=g(\mathbf{X})$$

4. Experimental Setup and Procedure

The experiments are being conducted in the d'Arbeloff lab at MIT. Two rooms connected by a door are used in the experiment. The two rooms are maintained at different temperatures. The hot room is about 35 °C and the cold room is 8 °C. Room temperatures and skin temperatures are measured by thermocouples and signals are logged into a computer. A skin blood flow sensor consists of a photo plethysmograph sensor is attached to the right index finger. Infrared thermometers measuring ear drum temperature provide the representation of core temperature. Two of these thermometers

are used, one is in the hot room and the other is in the cold room. The blood pressure is measured by an automatic blood pressure meter which is applied on the left arm. Both the skin blood flow sensor and the blood pressure meter provide heart rate signal.

One of the researchers of this project is the first subject of the experiment. The subject is a health male student in his late twenties. The subject wears only shorts during the test. The experimental setup is shown in Figure 8.

The test starts in the hot room. As soon as the subject steps in the hot room, all

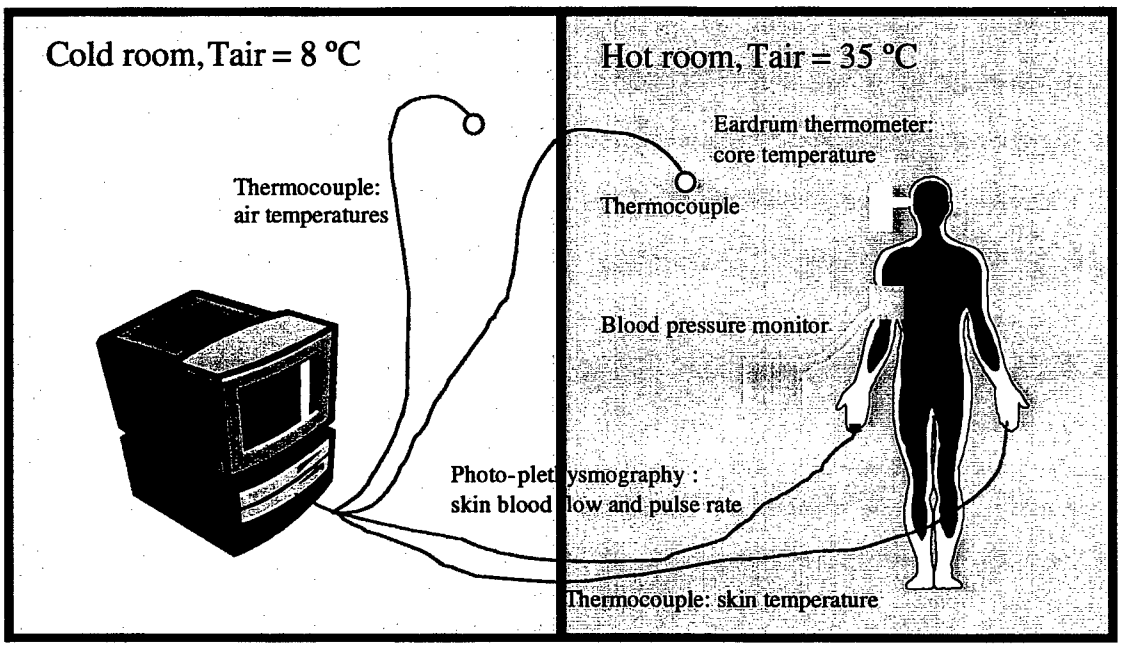


Figure 8: Experimental Setup.

the sensors are connected and data are taken. After about one hour when the physiological condition of the subject reaches steady state, he walks into the cold room. The subject stays in the cold room for more than ten minutes and goes back to the hot room when he feels too cold. He then remains in the hot room until all physiology signals are steady. The subject walks into the cold room and repeats the experiment. No apparent shivering and sweating occurs during the experiment.

Temperatures are measured continuously and logged automatically by the computer. The skin blood flow signals are collected continuously by the sensor. A transceiver receives the raw signal from the blood flow sensor and forwards the skin blood flow and pulse rate signals to the computer. The blood pressure signal, however, is taken about once every two minuets due to the limitation of the oscillometric method.

The blood pressure meter provides systolic and diastolic pressures as well as heart rate data on a LCD panel. The core temperature is taken at about one minute interval due to the limitation of the measuring device. The core temperature is also displayed on a LCD panel. The blood pressure and core temperature are measured and recorded by an assistant.

5. Results and Discussions

A typical set of data obtained in the experiment is plotted in Figure 9. The experiment shows that the ambient temperature drop has elevated the blood pressure. Although there are some fluctuations in the blood pressure, both systolic and diastolic pressure remain higher in the cold room than in the warm room. The pulse rate drops during the period of

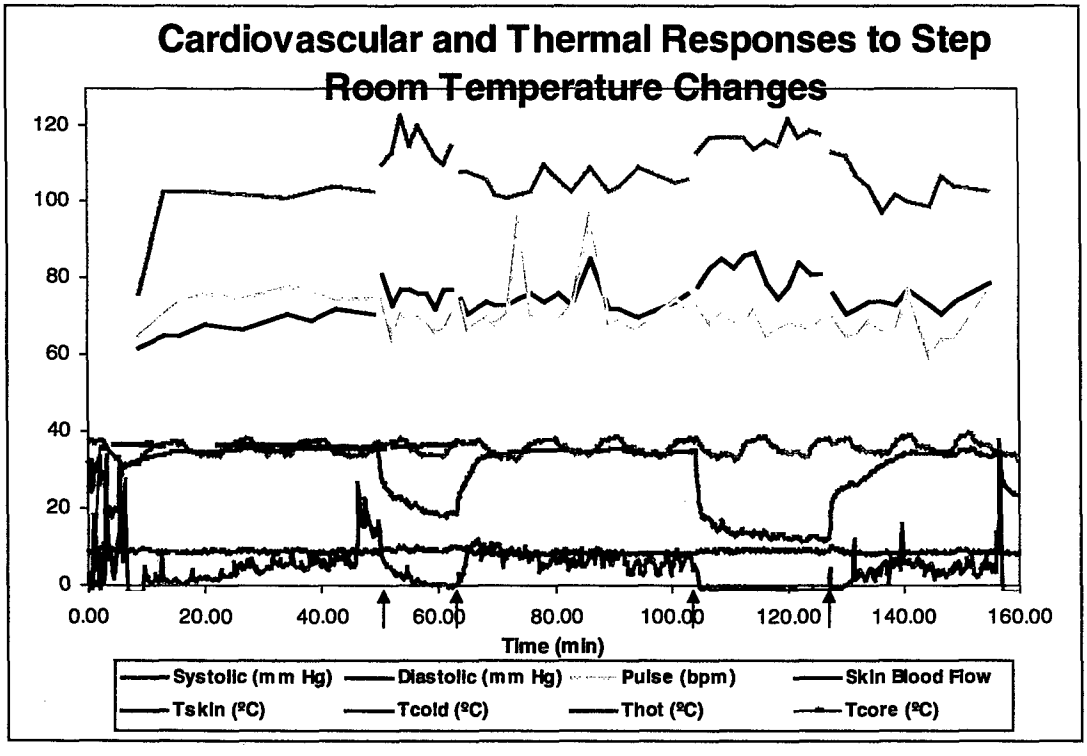


Figure 9: A Set of Experiment Results. (The Arrows Indicate the Ambient Temperature Change).

blood pressure elevation. This qualitatively justifies the assumption that the total cardiac output does not change during the exposure to cold air.

5.1 Comparison of the Experimental and Analytical Results

We have done numerical simulation based on the dynamical system model of the coupled cardiovascular and thermoregulatory system. The circulatory system parameters are adopted from the work of White et al [3]. The thermoregulatory parameters are taken from previous work in this laboratory [4]. All the control gains are obtained based on the measured data. Since all parameters are derived based on average human subject, the simulation results are normalized to be compared with experimental results.

The comparison is shown in Figure 10. The smooth curve is the simulation result plotted against all the experimental data points. The simulation is able to predict the general trend of blood pressure. The model shows a steady state blood pressure elevation which is reported by Ihenacho [5]. However, the rate of blood pressure increase at the beginning of cold exposure is somewhat lower than that in the actual human response. This can be the result of sensor error in the rapid transient situation or the negligence of the change of blood vessel compliance due to vessel shrinkage.

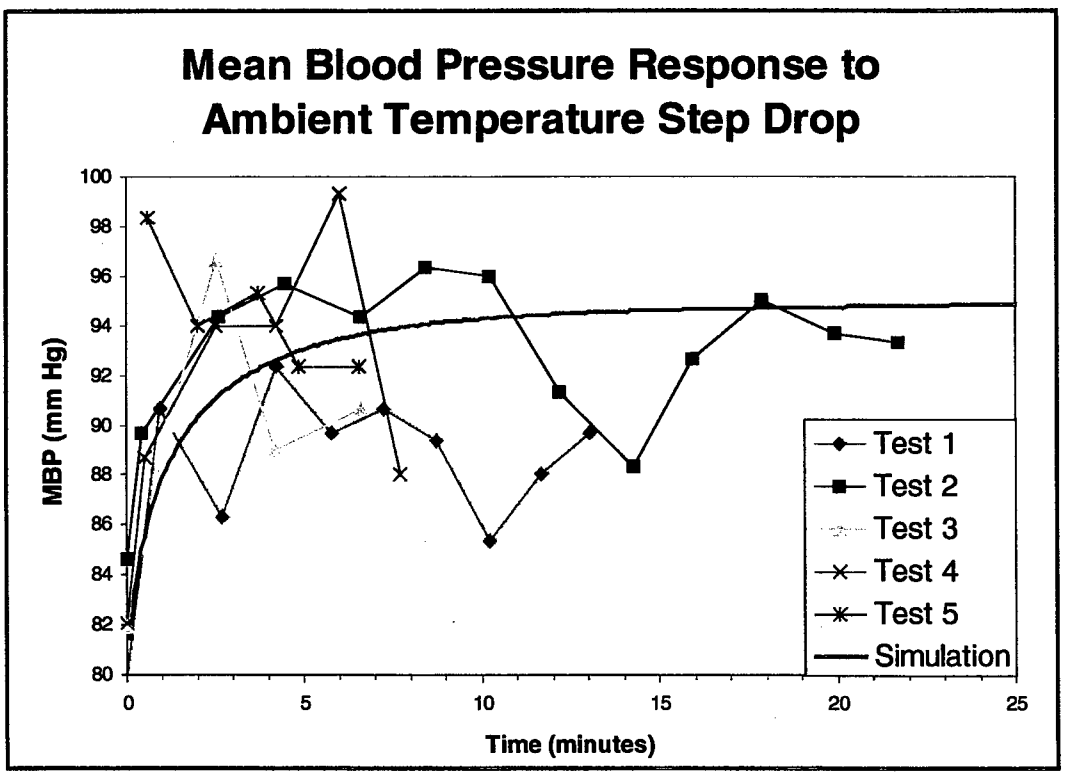


Figure 10: Comparison Between Experiment and Simulation.

5.2 Future Work

Based on the preliminary results, the following tasks are proposed.

(1) The coupled system model needs to be refined. For example, the coronary blood flow of normal people will increase during the cold exposure as the result of additional heart work needed to pump blood with higher pressure. Patients with coronary artery disease fail to increase the coronary blood flow during the cold stress. With a model including coronary vessels, we should be able to predict the risk of ischemia due to the cold stress.

(2) New sensors are needed for continuous and distributed measurement of blood flow. The blood pressure transient may play an important role in a patient's health. In addition, continuous measurement of blood pressure signal can help verify and improve the mathematical model. Sensors that measure blood pressure or flow at different positions are also critical and are currently being developed.

The HVAC stress test technique will not be confidently implemented without testing patients with known diseases. These tests will help us to verify our model, interpret simulation results, and design test inputs.

Reference

1. http://www.social.com/health/nhic/data/hr0100/hr0142.html, American Heart Association, 1998
2. Gagge, A.P., Stolwijk, J.A.J., and Nishi, Y., "An Effective Temperature Scale Based on A Simple Model of Human Physiological Regulatory Response," *ASHRAE Transactions*, Vol. 77, Part 1, pp. 247-262, 1971
3. White, R.J., Fitzjerrell, D.G., and Croston, R.C., "Fundamentals of Lumped Compartmental Modelling of the Cardiovascular System," *Adv. Cardiovascular Phys.*, Vol. 5, Part 1, pp. 162-184, 1983
4. Zhou, M., "Human-Centered Control of the Indoor Thermal Environment," *Ph.D. Thesis, M.I.T.*, 1998
5. Ihenacho, H.N.C., "Air-Conditioning and Health: Effect on Pulse and Blood Pressure of Young Healthy Nigerians," *Central African Journal of Medicine*, Vol. 36, No. 6, pp. 147-150, 1990
6. Ozawa, E.T., "A Numerical Model of the Cardiovascular System for Clinical Assessment of the Hemodynamic State," *Ph.D. Thesis, M.I.T.*, 1996
7. Werner, J., "Thermoregulatory models," *Scandinavia Journal of Work Environmental Health*, Vol. 15, suppl. 1, pp. 34-46, 1989.
8. Karnopp, D.C., and Margolis, D.L., *System Dynamics: A Unified Approach*, 2nd Edt. John Wiley & Sons, New York
9. Sah, R. and Moody, G., *User Manual for the Cardiovascular Simulator*, 1985

Nomenclature

Ps	Blood pressure in skin vessels
Pv	Blood pressure in viscera vessels
Cs	Compliance of skin blood vessel
Cv	Compliance of viscera blood vessel
Rsa	Artery resistance of skin blood vessel
Rsv	Venous resistance of skin blood vessel
Rva	Artery resistance of viscera blood vessel
Rvv	Venous resistance of viscera blood vessel
P	Mean artery pressure
Qs	Skin blood flow
Qv	Viscera blood flow
Qc	Cardiac output
Pth	Right atrium pressure
Tc	Core temperature
Ts	Shell temperature
ρ	Blood density
Cp	Specific heat of blood
$\overline{hk}$	Tissue conductance
$\overline{ha}$	Skin heat transfer coefficient
$\overline{c}_s$	Shell thermal capacitance
$\overline{c}_c$	Core thermal capacitance
Adu	Skin area
Ta	Air temperature
H	Metablolic heat generation
Ts,s	Skin temperature set point
Ps	Mean artery pressure set point
Kts	Skin thermal reflex gain
Ktv	Viscera thermal reflex gain
Kc	Baroreflex gain
Rsa,s	Skin artery resistance set point
Rva,s	Viscera artery resistance set point
$\boldsymbol{X}$	Vector of state variable
u	System input

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Sensors and Actuators

CHAPTER 8

Conducting Polymer Sensors for the Home

P. Madden, J. Madden, T. Kanigan, I. Hunter

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Conducting Polymer Sensors for the Home

Peter G. Madden, John D. Madden

Dr. Tanya Kanigan and Professor Ian W. Hunter

Introduction

Conducting polymers exhibit a wide range of tunable properties that make them ideal for use as sensors. Examples of applications are chemical, mechanical, optical, thermal, acoustic, and electrical sensors. Most of these sensor modalities involve changes in the electronic structure of the polymer resulting from chemical, mechanical, optical, thermal and acoustic stimuli.

Among the important characteristics of polymers that must be understood in order to develop and optimize sensors are the conductivity, composition, and the thermal and mechanical behaviors. In this section we discuss methods for and results from determination of composition and thermal behavior, as well as providing a theoretical framework for understanding the unique conduction mechanisms observed in these materials.

For example, to construct conducting polymer mechanical sensors, the variation in conductivity with mechanical strain, with temperature, and with the ambient environment must be understood. The principle of operation of these mechanical stress and strain sensors is quite simple. Applied strains reversibly affect polymer morphology and dimensions, leading to changes in resistance. Polymers provide tremendous advantages over traditional strain gauges (silicon or metal based) because they can undergo recoverable deformations that are one to three orders of magnitude larger. However, polymeric materials can be more sensitive to temperature variations. The ratio of polymer to deliberately introduced dopants, the solvent, and the ionic content will affect the operation of the sensor performance. Changes in resistance are dependent on conductivity, which may be highly anisotropic and process dependent in conducting

polymers. Therefore an understanding of the fundamental mechanisms of electron transport is critical.

In the following sections we begin by reviewing two important thermal characterization techniques coupled with molecular identification. These techniques allow 1) the determination of the ratios of polymer, solvent and ionic contents; 2) the determination of the glass transition temperatures which are critical for an understanding of mechanical behavior; and 3) the investigation of temperatures at which thermal degradation occurs and of the mechanisms associated with this degradation. The latter is critical for the specification of the operating temperature range. Results from polypyrrole, which is employed as a mechanical sensor, a chemical sensor, in transistors, tunable optics and in high energy density batteries, are shown.

After discussing the analytical techniques and results, background material on conduction mechanisms in polymers is presented. A proper understanding of these is important in applications of conducting polymers because electron transport is often involved. These mechanisms are very different from those in metals, leading to important ramifications for material processing and application.

Polymer Characterization

Polymeric materials often coexist with liquid, solid and ionic phases of other organic and inorganic materials. In order to study the chemical, electrical, thermal, optical and mechanical properties of polymers, and compare and model results, it is important to determine relative composition of the phases. Also, in polymers the key parameter that can affect all of these properties is the glass transition temperature. The investigation that follows seeks to discover if a glass transition temperature exists in the electrochemically synthesized polypyrrole investigated, as well as get a quantitative description of thermal degradation. This section begins with a description of two important tools used in polymer characterization. These are thermal gravimetric analysis (TGA), and differential thermal

analysis (DTA). Mass spectroscopy (MS) is briefly described as it is employed for molecular analysis of material released from TGA/DSC samples as temperature is increased. Some of the principles and capabilities of these techniques are first described, followed by some data obtained from polypyrrole samples.

Thermal Gravimetric Analysis

Thermal gravimetric analysis, abbreviated TGA, enables the investigation of thermal decomposition of an analyte. It can provide information both on the temperatures over which the material being investigated decomposes, and also the relative mass fractions of the various constituents within a sample. Furthermore, the rate of mass loss as a function of temperature can be employed to study reaction kinetics and reaction mechanisms. Thus TGA is an important tool used to determine thermal stability, reaction kinetics and stoichiometry.

The technique is of particular interest in the investigation of polymers because they often contain several distinct phases in an unknown proportion. Polypyrrole as grown in our laboratory, for example, contains solvent (propylene carbonate), ions (tetraethylammonium hexafluorophosphate) and polymer. The relative concentrations of the various constituents affect optical, electrical and mechanical properties. Furthermore, in creating products from polymers it is essential to determine their working temperature ranges. Determining decomposition temperature puts an upper bound on the working temperature range.

Procedure

In TGA the temperature of a sample is controlled. The change in sample mass is measured as a function of temperature. Typically temperature is ramped linearly.

Apparatus

TGAs consist of a balance, a furnace, a temperature control loop and a balance feedback loop. The balance usually consists of a galvanometer to which a beam is attached. The sample sits on the beam, generating a torque. The torque is balanced by applying current

to the galvanometer coil in order to maintain constant position. Displacement of the beam is measured using a slit on the beam, which is irradiated by an LED. Displacement is recorded as the change in output from a photocell. The mass resolution for such a balance is typically $< 0.1\mu g$.

Temperature is typically ramped linearly. However, other temperature cycles can be useful, such as making the rate of temperature rise proportional to the rate of mass loss. Gas flows over the sample to remove volatiles emitted by the analyte. Either inert gases (e.g. Argon) or reactive gases (e.g. air or Oxygen) may be employed depending on the conditions of interest. After passing over the sample the gas composition may be analyzed. TGA is often used in conjunction with mass spectrometry or Fourier transform infrared adsorption spectroscopy (FTIR). These methods allow the identification of volatile materials being released.

Design and Experimental Considerations

Ideally, mass changes would be perfect steps. However, mass changes due to a given reaction or vaporization often occur over temperature ranges of as much as 100 K or more. As a result, mass changes due to several reactions are often superimposed, making analysis difficult.

How can the mass changes due to various reactions be distiguished? If mass losses occur over very narrow temperature ranges then they can more easily be distinguished. Slowing the rate of temperature rise, providing more time for reactions to occur, temperatures to stabilitize and mass transport through the sample to take place, improves resolution. If reaction kinetics are simply slow compared to the temperature ramp rate, then either when temperature change is slowed, or even stopped, to allow a component to be expelled, loss steps are sharpened. By increasing gas flow, or by employing a vacuum, rates can be increased by reducing the partial pressure of the reaction product. Sometimes mass transport of the reactant to the sample surface can be an issue and products may be trapped within a sample. Sample geometry can increase resolution in such cases; namely surface area exposed to gas flow should be maximized.

However, some width to the change is inevitable. Clearly, reaction rates change exponentially with temperature, as given by the Arrhenius equation. Depending on the order of the reaction, the step width will be at least 10 K. Solvents will simply vaporize at their boiling point, but will evaporate at lower temperatures, as dictated by partial pressures. Elemental, molecular and or spectroscopic analysis of the mass effluent is then useful to determine if overlapping mass losses due to several reactions or phase changes are occuring. Data from combined TGA/ Mass spectrometry are shown in the results section.

Sources of error in the mass measurement include changes in buoyancy, and turbulence from the flowing gas. Buoyancy of gas is a function of temperature. Between room temperature and 1000° C, buoyancy of a gas changes by a factor of 4.4. The change is readily calculated from the ideal gas law, in which, under isobaric conditions, the ratio of initial to final density is equal to the ratio of final to initial temperatures. Employing a differential measurement between a sample balance and a nearly identical reference balance is often used to compensate for change in buoyancy. The second balance ideally remains at a fixed mass and is exposed to the same temperature and environment as the balance holding the sample. Alternatively, when only a single balance is present, a blank run may first be performed; the results of which are then used to correct for buoyancy changes in the sample run.

High flow rates are desirable because partial pressures of the volatile components are reduced and hence mass transfer rates are increased. As mentioned, sharper steps in mass loss are thus obtained and therefore higher resolution. However, high flow rates and hence Reynolds number may lead to turbulence. The appropriate shaping of the furnace minimizes turbulence due to gas flow. Alternatively, a vacuum may be applied to the sample.

Condensations of products onto cold portions of the balance, and the build-up of electrostatic forces within the furnace are further sources of error. Temperature resolution and control are functions of heating rate, sample and furnace geometry's, thermal conductivity, enthalpy of the process. Rapid exothermic reactions, for example, may change sample temperature by hundreds of degrees.

In summary TGA involves the recording of mass change as a function of temperature. Studying mass change as a function of temperature in turn provides information on thermal decomposition, degradation, reaction kinetics and composition. Attention is now turned to methods that provide more quantitative analyses of thermodynamics and kinetics.

Differential Scanning Calorimetry (DSC) and Differential Thermal Analysis (DTA)

Differential Scanning Calorimetry (DSC) and Differential Thermal Analysis (DTA) are employed to identify the kinetics, enthalpies and temperatures of onset of reactions and phase transitions. In polymer samples DSC and DTA are useful in identifying glass transition temperatures, melt temperatures and solvent evaporation. When used in conjunction with thermal gravimetric analysis (TGA), DSC and DTA provide additional information in identifying mass loss steps. The use of differential thermal techniques dates back to Le Chatelier in the 1880s.

Procedure and Apparatus

In differential thermal analysis both a sample and a reference are heated (or cooled). The reference consists of inert material having a heat capacity close to that of the sample. The temperature difference between sample and reference as a function of sample temperature, reference temperature or time is then recorded. Given matched heat capacities, and a well-designed furnace, temperature difference will be zero or constant as long as no reactions or phase transitions occur. A difference in heat capacity will lead to a

temperature difference that increases linearly with temperature. An exothermic reaction will lead to a rise in the sample temperature, and hence a peak in the sample minus reference temperature plot. Conversely, an endothermic reaction will lead to a valley. The vaporization of a liquid, for example, will lead to a dip in sample temperature relative to the reference. The areas of the peaks and valleys are proportional to the enthalpy of reaction. Generally, however, DTA is used for qualitative rather than quantitative analysis.

Differential scanning calorimetry was first developed in the early 1960s at Perkin Elmer. Two matched furnaces are employed. One contains the sample and the other the reference. The temperature difference between the sample and the reference are once again determined. The temperature difference is driven to zero via a feedback loop that proportions power between the two furnaces. A second feedback loop is used to drive the average temperature of the two furnaces along a temperature profile. In DSC, the difference in power provided to each furnace is plotted as a function of temperature. Because sample and reference are both maintained at the same temperature, DSC is appropriate for quantitative analysis. This is because both sample and reference as maintained at the same temperature.

Mass Spectrometry

Mass spectrometry (MS) is employed to perform elemental and molecular identification based on charge to mass ratio on molecules of up to approximately 100,000 Daltons. Polymers generally have higher molecular weights, so MS used identify decomposition products, solvent and ionic content. There are two principal methods of mass spectroscopy, namely magnetic MS and time of flight-based MS. There are also many methods of sample preparation, as samples must form plasma.

Methods

Many means of ionization are available including inductively coupled plasma (ICP) formation, electron and ion bombardment, and laser ablation. In time of flight MS, ions are accelerated by kilovolt electric fields. Their time of flight down an evacuated column to a detector is then measured. The time, t, is a function of the column length, L and the velocity, v, which in turn is related to mass, m, charge, q and the accelerating voltage, V, by

$$v = (2Vq/m)^{1/2}$$

$$t = v/L.$$

Alternatively, the charged particles can be accelerated using an electric field, and then deflected in a magnetic field applied perpendicular to the ion velocity, v. Given a magnetic field strength, B, a distance of travel through the magnetic field, L, and calculating the centripetal acceleration due to the magnetic field on the particle travelling at velocity, v, the angle of deflection, a, is related to the mass to charge ration by:

$$\sin(a)=LqB/mv.$$

Velocity, v, is given as above, leading to a result that the sine of the angle, a, is proportional to the square root of the charge to mass ratio. This dependence of deflection on the charge to mass ratio is the basis of detection by mass spectroscopy.

Data and Results

Polypyrrole is a very versatile polymer that is employed in optical, electrical, biological, electrochemical, and mechanical systems. It is composed of polypyrrole chains, as well as solvent (propylene carbonate abbreviated PC) and ions (tetraethylammonium hexafluorophosphate, $Et_4N^+PF_6^-$). Three key questions arise, namely (1) what is the relative content of solvent, ions and polymer, (2) what is the thermal decomposition temperature and (3) is there a glass transition observed in this polymer? The combination of TGA, DTA and MS were employed to help find the answers.

Polymer Growth

A mixture of 0.06*M* freshly distilled pyrrole monomer and 0.05*M* tetra ethyl ammonium hexafluoro phosphate is prepared in propylene carbonate. Polypyrrole is then electrodeposited onto a glassy carbon substrate. A copper counter electrode is employed and in galvanostatic (constant current) mode, the deposition current density on the glassy carbon is 1.25 $A\ m^{-2}$. Deposition takes place at -30°*C* in a nitrogen atmosphere. The resulting films have densities of 1440 kg/m^3 and their counter ion content is thought to be 0.27 PF_6^- per monomer.

The low temperatures used appear to limit side reactions, which can degrade conductivity. Glassy carbon, which is polished to a 0.05 micrometers finish, allows ready removal of the film either by direct peeling or thermal shock with liquid nitrogen.

With this fabrication procedure, conductivity's exceeding $3x10^4$ *S/m* and tensile strengths of greater than 25 *MPa* are routinely achieved. Typical films are 30 micrometers thick, 10 *mm* wide and 30 *mm* long.

Thermal Analysis

First TGA and DTA were performed on the solvent and ionic components of the polymer. Figures 1 and 2 depict results of combined TGA/DTA performed on propylene carbonate,

and tetraethyl ammonium hexafluorophosphate, respectively.

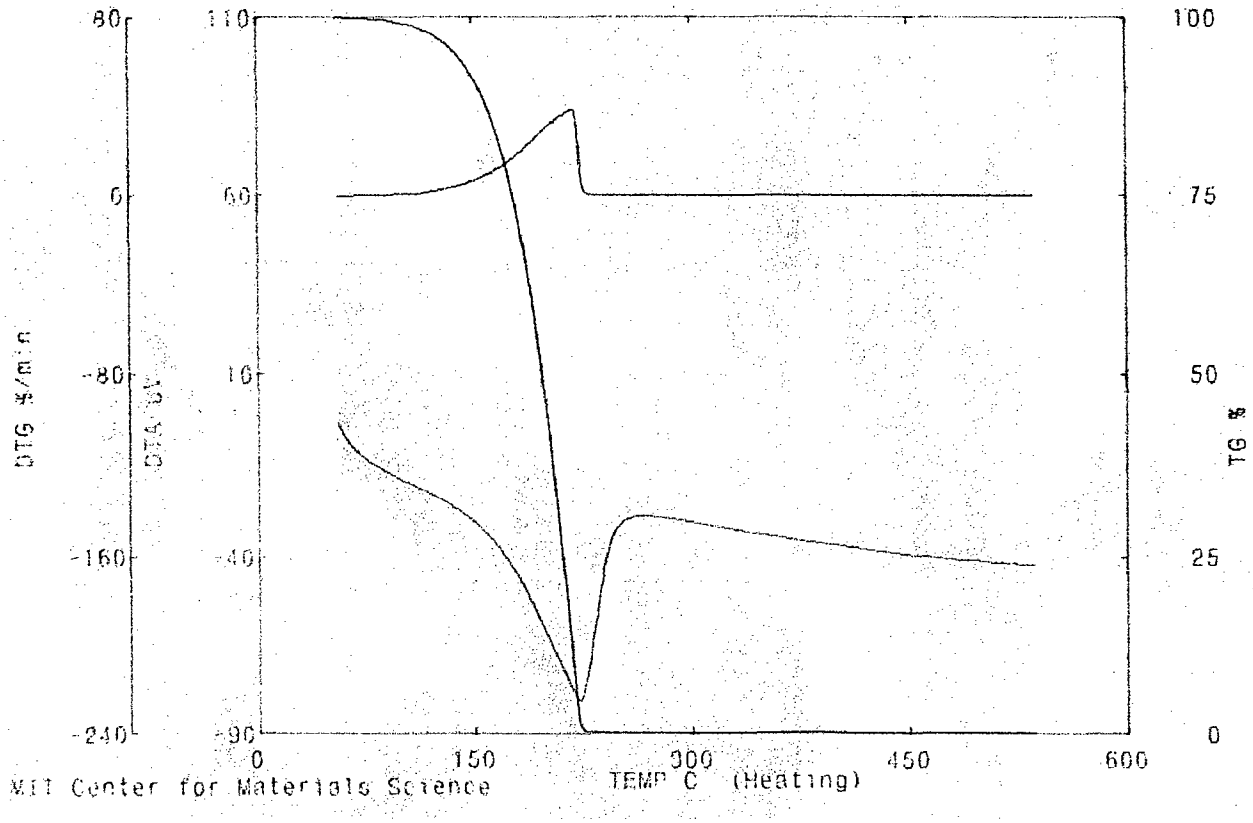


Figure 1: Sample: propylene carbonate, mw = 102.09
Initial weight: 26.778 mg
Final weight: 0.034 mg (no visible liquid remains)
Temperature range: 50 C - 550 C
Heating rate: 20 C / min
Sampling rate: 1 / s
Gas: N2, 150 mL / min

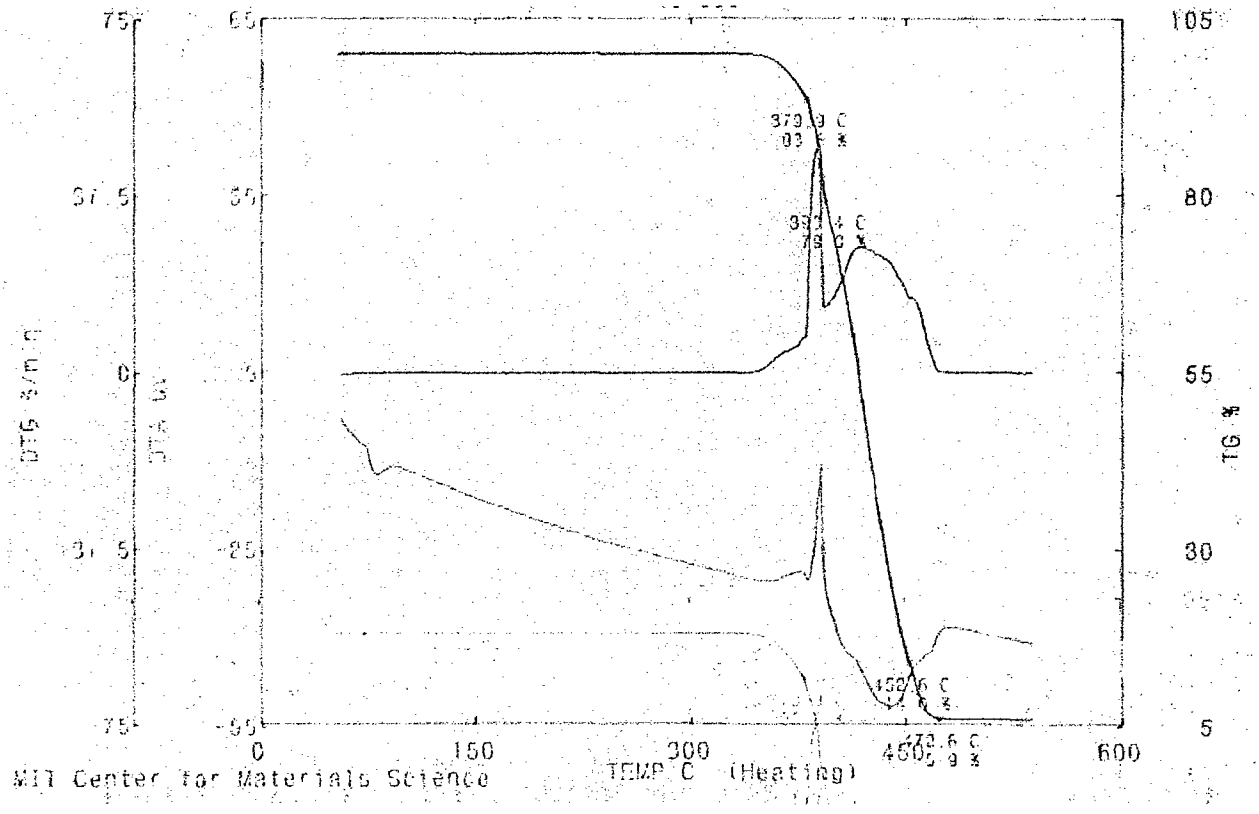


Figure 2: Sample: tetraethylammonium hexafluorophosphate, mw = 275.22

Initial weight: 7.894 mg
Final weight: 0.6446 mg
Temperature range: 50 C - 550 C
Heating rate: 20 C / min
Sampling rate: 1 / s
Gas: N2, 150 mL / min

The red curve represents the mass as a function of temperature, the blue the rate of mass loss and the green the temperature difference between sample and reference. The boiling point of propylene carbonate is 242 degrees Celcius. Note that in Figure 1 there is a large temperature lag between the sample relative to the reference as this temperature is approached, indicating an endothermic reaction. An endothermic reaction is expected as solvent evaporates.

Note the breadth of the rate of loss peak. This is characteristic of an evaporating liquid, which in liquid state still maintains a significant vapor pressure. Propylene carbonate is also known to thermally degrade at temperatures below its boiling point, which contributes to broadening. The degradation is known to begin above 150 degrees C. The sudden drop in rate at 242 degrees C is due to the complete loss of sample, as would be expected at the boiling point for pure solvent.

Note that even after the removal of all the sample a temperature difference is still maintained between the sample and the reference, and in fact the temperatures diverge beyond 300 degrees Celsius. The slope is constant, indicating that this is due to a difference in heat capacity between the sample and reference. While ideally the heat capacities are balanced, it is difficult to eliminatthe differences completely, resulting in a sloping baseline for the DTA.

The decomposition of tetraethylammonium hexafluorophosphate shows several peaks in both the DTA and the rate of mass loss. The first peak endothermic valley occurs at 100 degrees C. This is almost certainly due to the loss of water that is trapped within the salt. Very little mass loss occurs indicating that the salt has a low hydration level.

Two peaks occur further in rate of loss, one at 380 degrees C and one at 452 degrees C. The first is the result of an exothermic reaction and the second and endothermic process. DTA indicates that another exothermic process may be occurring at about 360 degrees C. The technique does not indicate what these peaks correspond to. However, if such peaks occur in the full polymer data, they are an indication that tetraethylammonium hexafluorophosphate is likely present. Note also that 5.9% of the mass still remains at the end of the temperature scan. The technique does not indicate what the composition of the residue is. However, the ratio of cation to anion masses (0.9) reveals that both ions must decompose to a large extent.

Note that it is quite unlikely that both the tetraethylammonium cation and the hexafluorophosphate anion are present in equal concentrations within the polymer. The as grown polymer is formed with a positively charged backbone, and it is believed that the anion balances charge. The cation in fact may not be present at all. As a result, the shapes and number of peaks in the rate of mass loss may differ from what is observed in the pure salt.

Figure 3 presents the results for the total polymer systems, namely polypyrrole with some component of propylene carbonate, tetraethylammonium cation, hexafluorophosphate anion and perhaps some water. The temperature was generally ramped uniformly, but held at 242 degrees C (boiling point of propylene carbonate) to allow boil off. Temperature was also held constant at 394 and 430 degrees C. Temperatures were held until the rate of mass loss dropped to near zero. Figure 4 is the same experiment repeated, but with double the temperature range and no temperature holds.

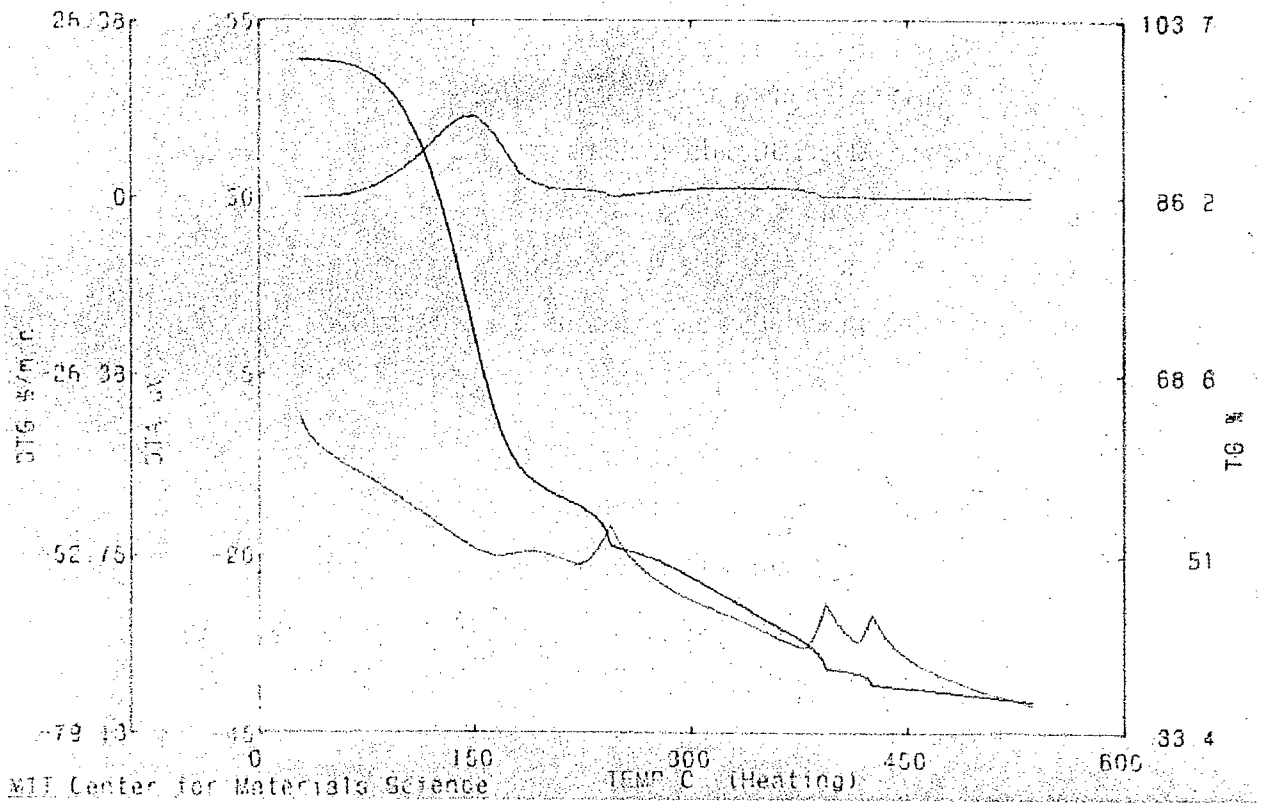


Figure 3: polypyrrole, reference batch 3/7/98 #2
Initial weight: 5.687 mg
Final weight: 2.066 mg
Temperature range: 30 C - 500 C
Heating rate: 20 C / min
Held temperature constant at: 242 C, 388 C, 420 C.
Sampling rate: 1 / s
Gas: N2, 150 mL / min

The data becomes somewhat more difficult to interpret for this multi-component system. First, there does not appear to be a clear glass transition. A glass transition is essentially a phase change in the polymer with no associated mass loss. While not strictly a second order phase transition, there is a latent heat associated with the glass transition. No large temperature lag is visible that cannot be associated with mass loss.

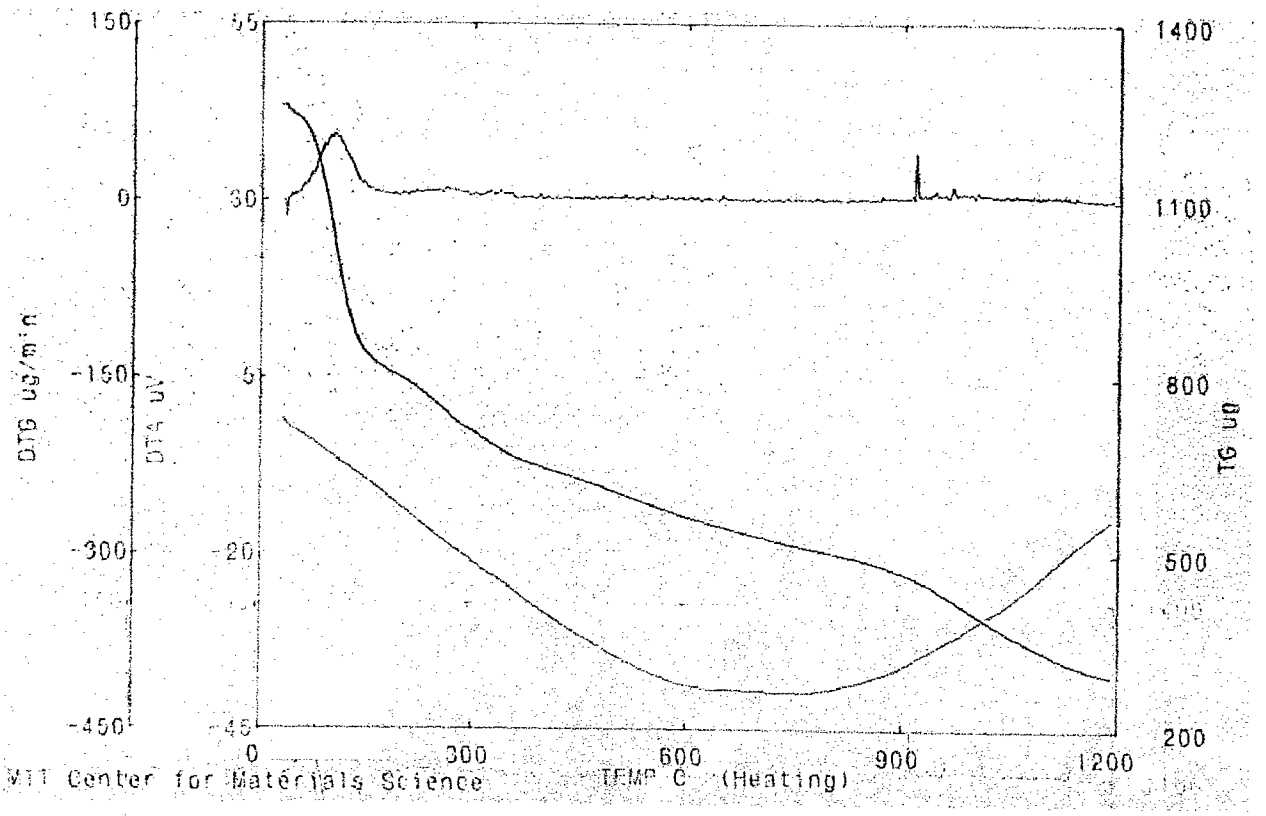


Figure 4: Sample: polypyrrole, 3/7/98 #2
Initial weight: 1.261 mg
Temperature range: 30 C - 1200 C
Heating rate: 8 C / min
Sampling rate: 1 / s
Gas: N2, 150 mL / min

The absence of a clear glass transition temperature in polypyrrole is not a surprise. The rigid nature of the polymer backbone tends to favor a semi-crystalline morphology. Hence any glass transition would only take place in the amorphous regions between crystalline domains. Other processes could easily mask such a transition.

There appears to be solvent loss, which is a maximum at 148 degrees Celsius and finishes near 179 degrees Celsius. The mass loss to that point represents 40% of the sample. A further mass loss step (8%) occurs at 244 degrees C. This step is likely due to some final quantities of solvent still trapped within the sample. The relative rise in sample temperature seems to indicate that the reaction occurring is exothermic and therefore not

solvent boiling. The DTA peak, however, is due only to the fact that temperature ramping was halted to complete boil off. The results indicate that likely 48% of the sample mass consists of solvent.

The origin of further losses is unclear. There is likely some polymer degradation occurring between 240 degrees C and 430 degrees C, given the broad peak in the rate of mass loss and the fact that no salt degradation was observed until 380 degrees C. The higher temperature data adds little insight, except to show that rapid deterioration of the sample occurs above 900 degrees Celsius.

How much of the loss is polymer degradation and how much is salt? Is solvent still being evolved? The data cannot give us a clear answer. In order to check, we must add elemental or molecular analysis tools to analyze the materials evolved. The next figures provide results from a TGA system whose output flows directly into a mass spectrometer.

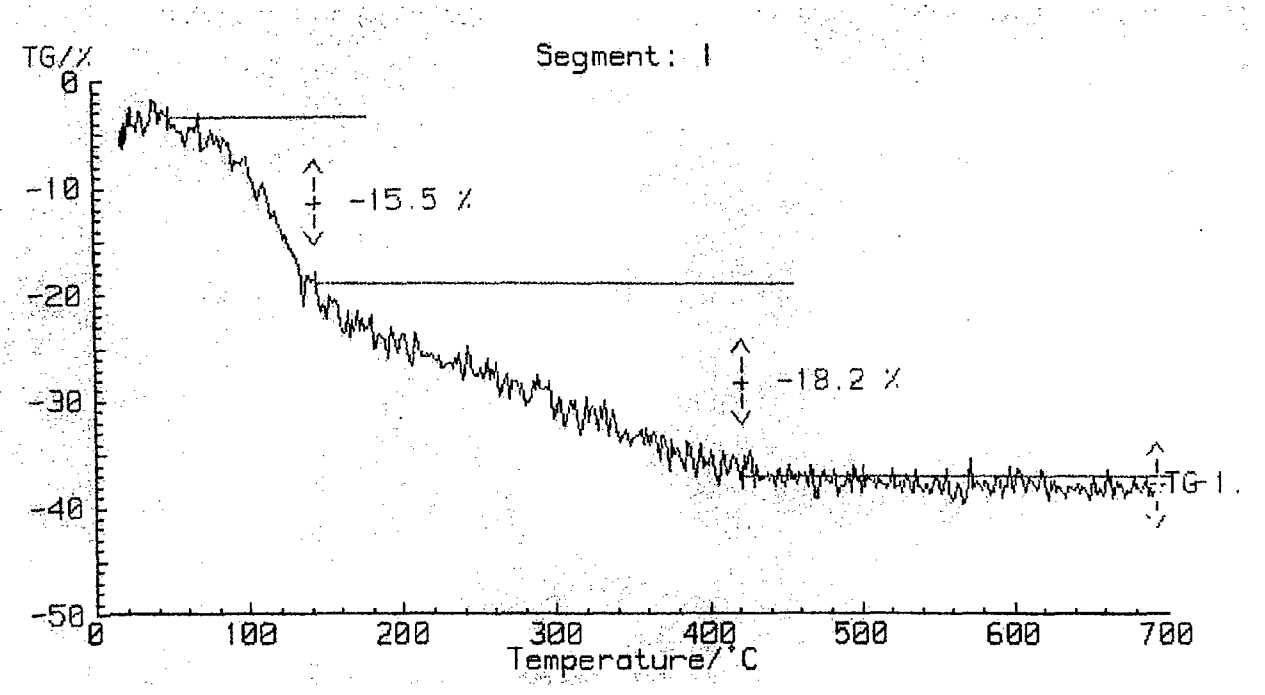


Figure 5: Sample: polypyrrole, 3/7/98 #2
Initial weight: 7.61 mg, Range: 30-700 C
Heating rate: 10 C / min Sampling rate: 3 / s
Gas: Ar, 50 mL / min

Figure 5 shows the results of TGA using the combined Quadrupole MS-TGA system (Netzsch STA 409-QMS). Results are qualitatively similar to those in Figures 4 and 5, with the first mass loss step possibly occurring at a lower temperature. Further information is gained by examining the combined MS and TGA data, Figures 6 and 7, and plotting the temperature dependence of the various MS charge to mass ratio peaks observed, Figures 8 and 9.

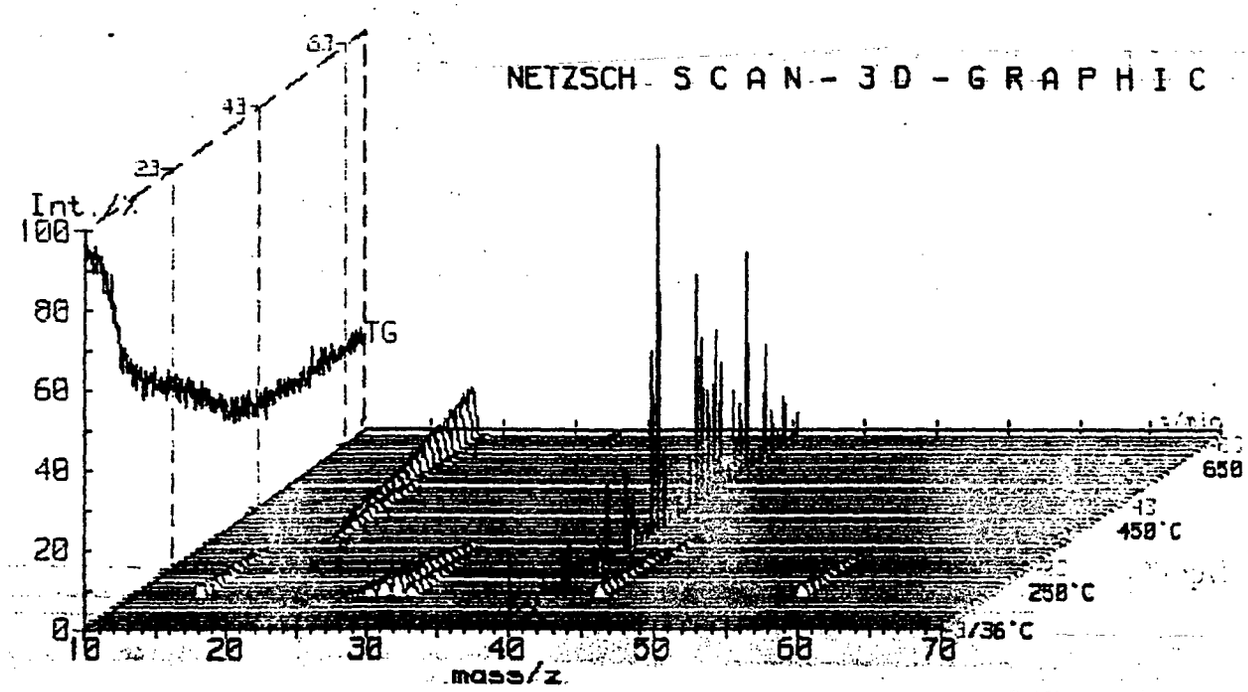


Figure 6: Intensity of mass peaks vs. temperature.
Mass to charge ratios 10:70.

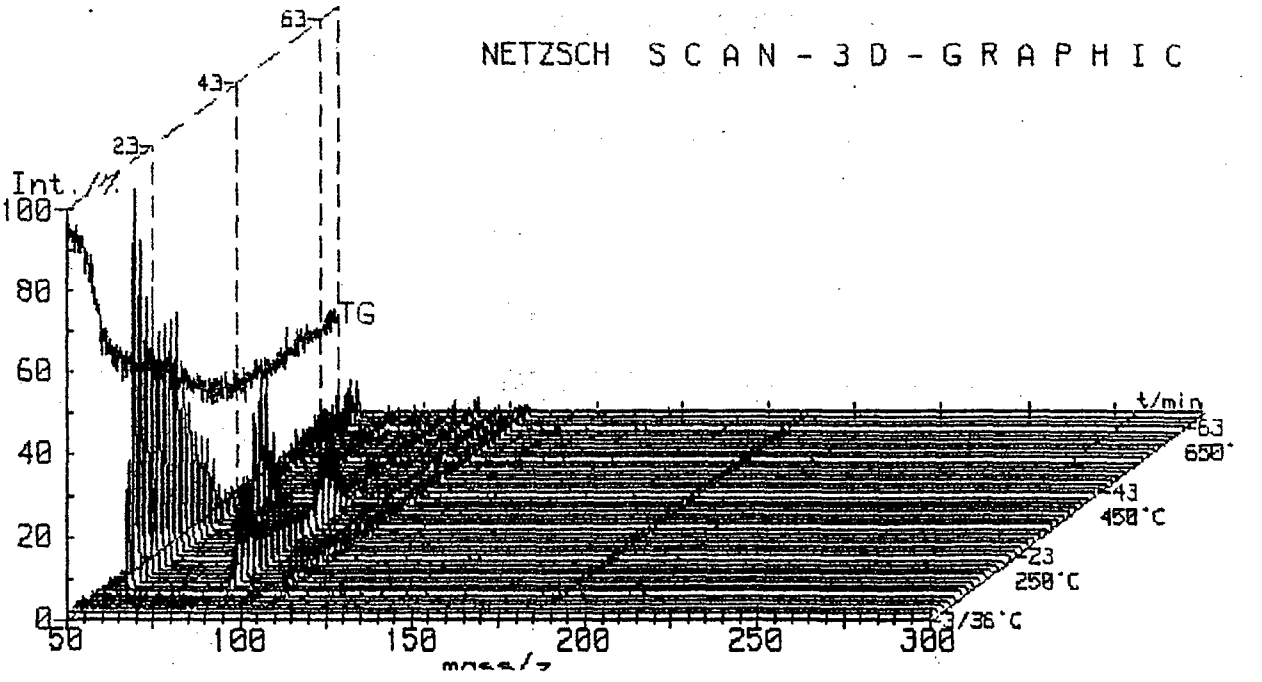


Figure 7: Intensity of mass peaks vs. temperature.
Mass to charge ratios 50:300.

Several peaks clearly emerge. These are listed in table 1, along with their probable identification. Note that mass spectroscopy of molecules does not always provide clear molecular identification. The probable decomposition products and their most likely ionization states must be examined. Thus some a priori knowledge of the sample constituents is almost always required, particularly for organic materials.

The peak at 40 mass/charge units and the shadow at 20 are clearly due to the Argon carrier gas that flows through the TGA cell, transporting evolved materials to the MS. Other peaks have been identified using knowledge of the probable decomposition products of propylene carbonate, hexafluorophosphate and polypyrrole.

Table 1: Mass Peaks	Temperatures (Celsius)	Identification
15	100-310	CH_3
18	350-700	NH_4
19	300-500	F
20	All	Argon, Ar^{++}
28	120-300; 580-700	CO; N_2, CHNH
29	120-370	CH=O
40	All	Argon, Ar^{+}
43	100-330	CH_3CO
44	130-370	CO_2
57	120-370	C_3H_6O, propylene oxide $CH_3CH_2CH=0$, propion-aldehyde $CH_2=CHCH_2OH$, allyl alcohol
87	120-370	Ethylene carbonate (PC – CH_3)
102	120-320?	Propylene carbonate

Figures 8 and 9 show the amplitudes of mass peaks as functions of temperature. No peaks were found to correspond to mass losses below 100 degrees C. The 15.5% mass loss below 100 degrees C could be associated with water loss, or undetected PC. Further investigation is required. It should be noted that some molecules do not all ionize equally easily, and therefore many molecules, such as water on occasion, may pass undetected.

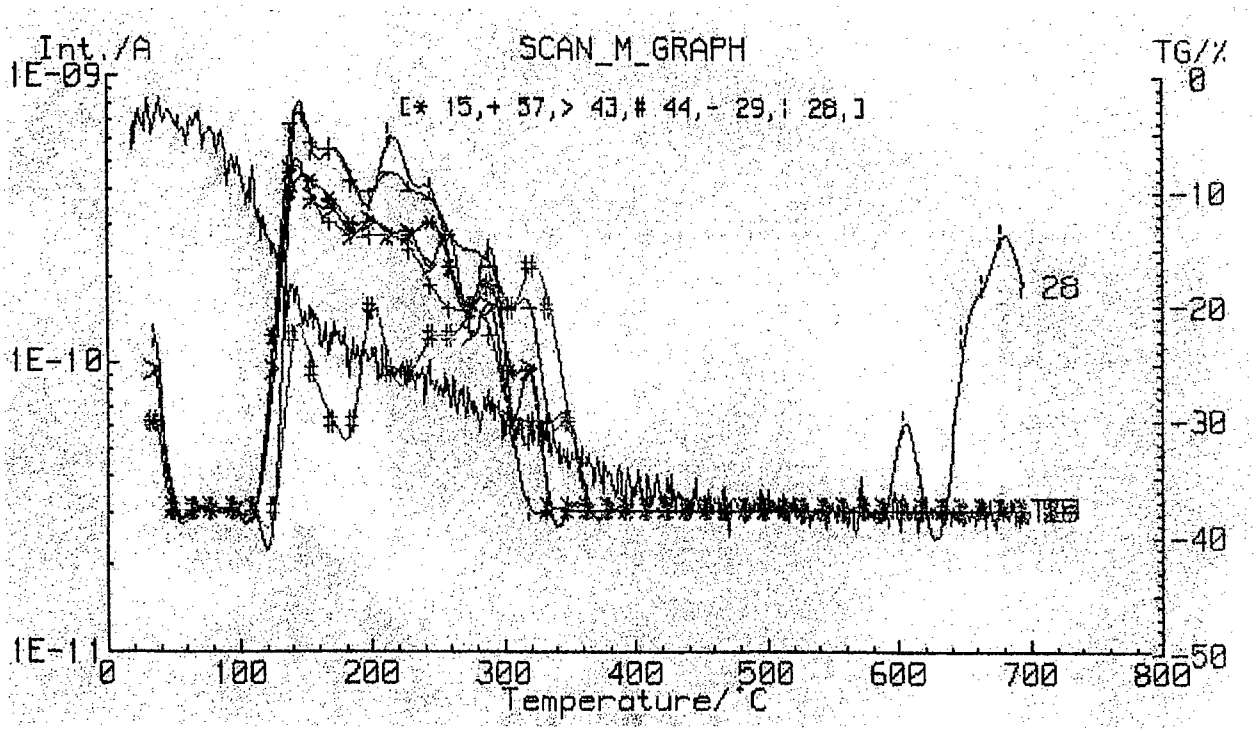


Figure 8: Selected mass loss peaks as a function of temperature, also showing TGA results.

Mass loss between 100 and 370 degrees C appear to be predominantly decomposition and boiling of propylene carbonate. Every peak visible in that temperature range is associated with the breakup of PC.

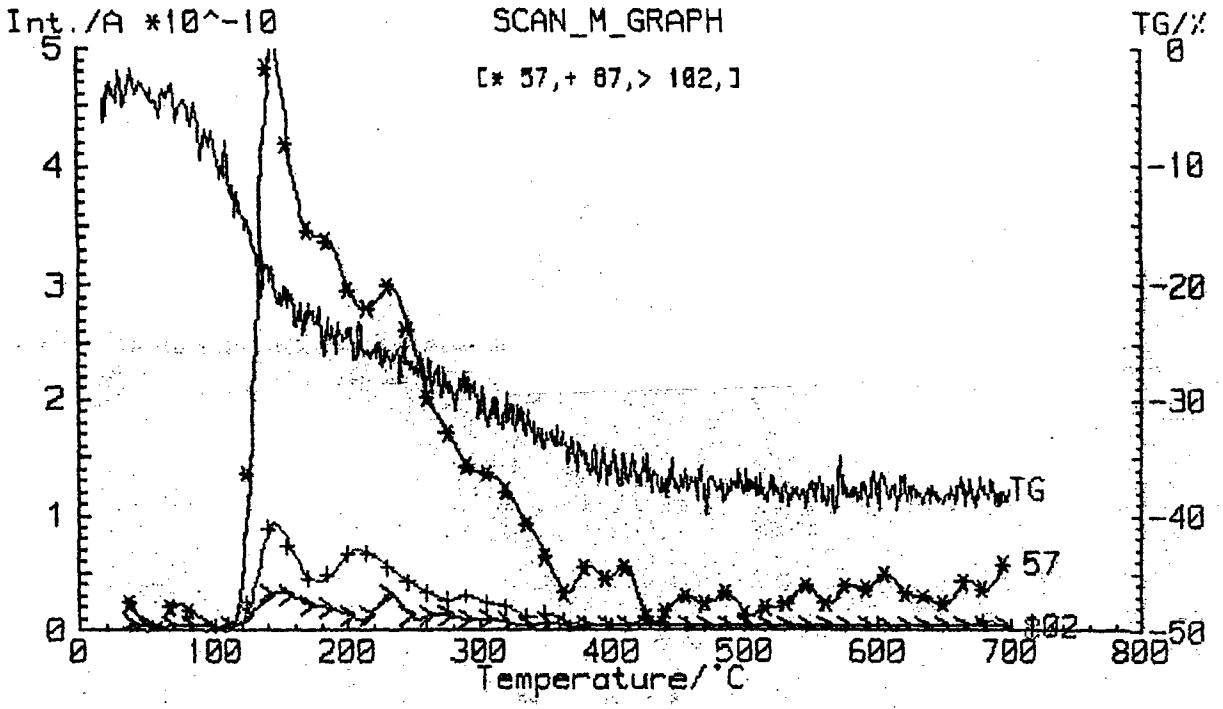


Figure 9: Selected mass loss peaks as a function of temperature also showing TGA results.

A fluorine peak (mass to charge 19) appears between 300 and 500 degrees Celsius with a peak loss around 400 degrees C. This loss pattern corresponds well with results in Figure 2.

Evidence of polymer decomposition is visible in mass losses at 18 and 28, corresponding to the loss of Nitrogen containing molecules. It appears that the onset of polymer chain degradation occurs near 350 degrees C.

Discussion and Conclusion

Three fundamental tools for polymer materials characterization have been discussed, and results presented. The results indicate that polypyrrole, synthesized as described, does not have a pronounced glass transition temperature. Solvent is evolved between 100 and 350 degrees C, and the hexafluorophosphate salt appears to decompose at about 400 degrees

C. The polymer chains are thermally stable (in an Argon atmosphere) until 350 degrees C. As much as 50% of the mass appears to be solvent, largely if not exclusively propylene carbonate. Higher resolution or other techniques are required to determine the percent ionic concentration present in the polymer.

A Review of Conduction Mechanisms in Conducting Polymers

Valence Bands, Conduction Bands, and the Bandgap

Many of the basic principles of polymer conduction arise from an understanding of conduction in crystalline solids.

For a single atom, the energy levels that can be occupied are discrete rather than continuous. The number of electrons that can fill a particular energy level is determined through Schrodinger's equation and is the number of unique quantum states for that level. This is known as the Pauli Exclusion Principle which can also be stated as 'No two electrons in a system can have the same quantum numbers.' When an electron changes levels, it will absorb or emit an energy quanta equal to the difference in energy between its initial and final level.

When two atoms are brought together to form a molecule, for instance H_2, the original energy levels of the hydrogen atoms split into two distinct energy levels. If the energy levels did not split but instead overlapped, it would be possible to have two electrons, one from each of the hydrogen atoms, with the same quantum state. By splitting the levels, the two electrons are both accommodated in the molecule with distinct quantum numbers and therefore do not violate the Pauli Exclusion Principle.

As more and more atoms bond together to form a crystalline solid, the energy levels split into finer and finer increments, eventually (in the limit of a very large number of molecules) becoming continuous energy bands of allowable energy instead of discrete energy levels.

If the temperature approaches absolute zero, the electrons in a solid will go to the minimum possible energy and will drop to the lowest energy bands, always obeying the

Pauli Exclusion Principle. Bands which are filled (in the outermost orbitals the bands might be partly filled) at absolute zero are called valence bands. Bands of energy above the valence band are called the conduction bands.

In the absence of an applied electric field, the net flow of electrons is zero; the number of electrons with momentum in one direction is equal to the number of electrons with momentum in the opposite direction. When an electric field is applied, electrons must change their energy level if they are to accelerate parallel to the electric field. Acceleration will only take place if there are vacant energy levels for the electron to move into. Thus current will only flow if there are vacant energy levels.

When empty energy levels in the valence band are available or if the conduction band overlaps the valence band, the material will conduct even at very small electric fields. If the valence band is full and there is a gap between the valence band and the conduction band, the electron must be able to gain enough energy from the applied field to jump across the gap into the conduction band.

Insulating materials have a very large band gap between valence and conducting energy bands (polyethylene for instance has a band gap of 8 eV). In semiconducting materials, the band gap is small. Silicon for instance has a band gap of about 1eV. At room temperature, the average electron thermal energy kT is about 1/40 eV. Because of the distribution of energies, many electrons will have enough energy to cross the semiconductor bandgap giving the material a small conductivity. The very high conductivities of metals result when the conduction band and valence band overlap. Electrons in the valence band of metals can be excited very easily into the conduction band.

Electrons and Holes

When an electron is excited out of the valence band into the conduction band (for example by an electric field, absorption of a photon, or via thermal excitation) there will be a resulting absence of charge in the valence band. The absence of charge is called a hole and has a positive charge. In a crystalline lattice, the hole is typically associated with a particular lattice site or particular atom. However, an electron from an adjacent atom can

jump over to the empty site, filling one hole and simultaneously creating a new one. The migration of the negatively charged electron from a full lattice site to an empty one is equivalent to the migration of a positively charged hole in the opposite direction.

In silicon and other inorganic semiconductors, holes or conduction band electrons can be deliberately introduced by adding a very small percentage of acceptor atoms, which accept atoms from the surrounding Si lattice, creating a hole, or donor atoms, which donate an electron into the conduction band. The addition of acceptor or donor atoms is called doping and can have significant effects on the conductivity even in concentrations of one part per million.

Conduction in Polymers

Conduction in polymers differs from conduction in inorganic materials because of the anisotropic nature of chemical bonding. Along the polymer chain, covalent bonds hold the carbon atoms tightly together. Perpendicular to the polymer chains, there are no strong covalent or ionic bonds. Attraction is due to the much weaker van der Waal's forces or to hydrogen bonding. The resulting conductivities are highly anisotropic on a molecular scale.

The weak interactions with neighboring polymer chains make it easy for the polymer molecules to distort. The changes in morphology in turn affect the electronic band structure and produce new states that lie in the bandgap. Polymer folds, chain ends, impurities, and differences between amorphous regions and crystalline regions will also affect the band structure of the polymer.

The uncertainties in structure make it very difficult to predict precise values for the conductivities. Preparation methods for the same polymer can yield very different properties. Solvents, chemical concentration, temperature, potential for electrodeposition, and other variables during the polymerization reaction can all affect structure.

Solitons, Polarons, and Bipolarons

One of the simplest conducting polymers to model is polyacetylene (see Figure 1). Polyacetylene has two structures known as cis and trans. The simplest model of polyacetylene is one where the carbon atoms on the backbone are equally spaced and the resulting bonds are one and a half strength. However, it is found that this structure is not energetically favorable[i,ii, iii]

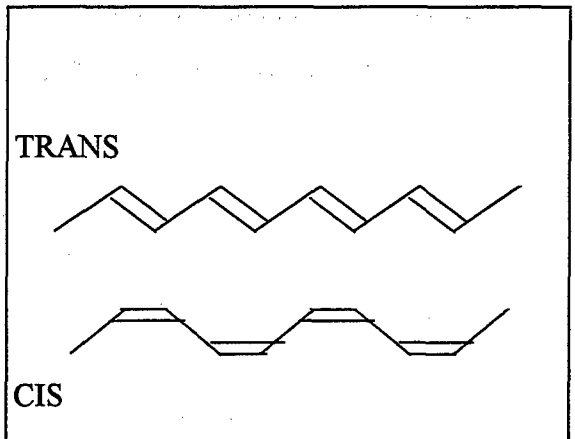


Figure 1: Structures of polyacetylene (PA). a) cis-PA. b) trans-PA.

The most favorable electronic configuration is one where single and double bonds alternate. As the polymer chain is very flexible, the change in electronic structure is accompanied by a change in mechanical structure. Part of the energy gained from the more stable electronic state is lost to the mechanical reshaping but the net energy is lower. The localization of the bonds on alternating sites also creates an energy gap between the valence and conduction bands. The more stable configuration with alternating single and double bonds is known as the Peierls transition[iv]. Other conducting polymers also have alternating single double bond structures.

A very important feature of the alternating bonds is that they have both relatively low ionization energy and relatively high electron affinity. The polymer chain can quite easily lose or gain an electron, or simply displace an electron along the polymer backbone. Such changes in the electronic structure are very important to the conduction mechanism. Again it is important to emphasize that the changes in electronic structure are accompanied by a rearrangement of the molecule and that the energy of the polymer molecules is affected by both electronic and physical structure.

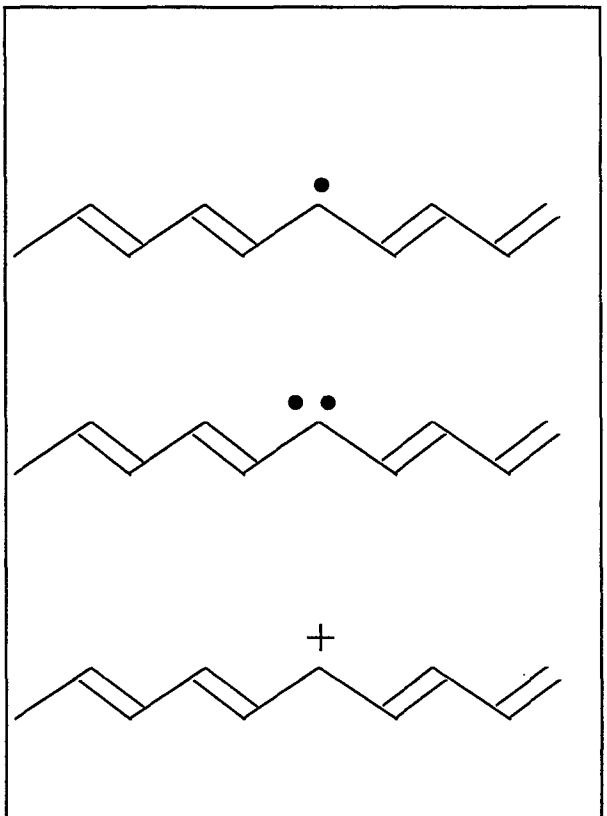


Figure 2: A neutral, a negative, and a positive soliton in trans-polyacetylene. (from p. 694, Bloor).

One of the simplest changes in electronic structure for polyacetylene is shown in Figure 2. There is a change in the bond alternation which creates a radical (a site with an unpaired electron), which is called a soliton. If the unpaired electron is removed, the soliton is known as a positive soliton. Likewise, if the electron is paired, the soliton is a negative soliton. The energy needed to create a soliton is small and for polyacetylene, the soliton is stable because the two different bond alternations have the same energy. The energy of soliton states are at energies in the middle of the bandgap, and so increase the conductivity of the polymer chain by providing lower energy states for electrons to jump to. Positive or negative solitons can propagate along the polymer backbone to carry current (Pople and Wallmsley and later Su, Schrieffer, and Heeger used a series of modeling assumptions based on the Huckel method to solve for the energies of formation and activation of solitons in polyacetylene[i,ii] and demonstrated that solitons can act as current carriers).

A neutral and a charged soliton can also become coupled to form what are known as polarons. As for solitons, polarons correspond to a new energy state in the bandgap. In polymers that do not have the same energy for different bond alternations, a single soliton is unstable. In poly(para-phenylene) (PPP) a single bond is more stable than a double bond between adjacent phenyl groups. A single soliton in a PPP molecule would lead to a large increase in the structural energy of the molecule. A pairing of two solitons into a polaron returns the single-double alternating bond to the more stable configuration after a short segment and therefore requires much less energy to form (Figure 3). Just like

solitons in polyacetylene, polarons can propagate along a polymer backbone to conduct current.

Finally, a bipolaron can coalesce from two polarons. Bipolarons will form if the increase in energy due to the proximity of the electric charge is more than offset by the decrease in energy brought about by the resulting morphological change. Bipolarons provide a conduction mechanism by propagating along the polymer backbone.

The morphological changes that arise from the formation of solitons, polarons, and bipolarons have a significant effect on their energies of formation. However, the changes do not result in significant macroscopic dimensional changes. Conducting polymers being investigated for application as actuators (contractile polymers) are not thought to change shape because of the changes in molecular shape induced by solitons, polarons, or bipolarons. Rather, the dimensional changes are thought to be due to the expansion of contraction of the polymer matrix due to ion flow into and out of the matrix.

Doping of Conducting Polymer Chains

The formation of solitons, polarons, and bipolarons is influenced by the local chemical environment of the polymer molecule. In particular, acceptor (oxidizing) or donor (reducing) entities in the vicinity of the polymer chain make the formation of solitons, polarons, and bipolarons more energetically favorable. Dopants therefore result in an increase in the number of conducting segments on the polymer chain and can increase the conductivity. It is also believed that dopants contribute to interchain conductivity by transferring charge.

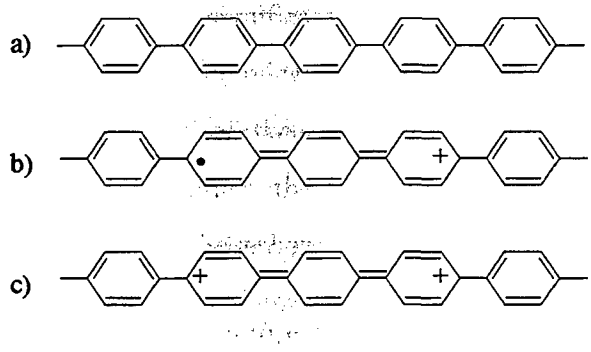


Figure 3: a) The standard bond structure or poly(para-phenylene) b) polaron bond structure and c) bipolaron bond structure.

Dopant ions dissolved in an electrochemical solution can be transported into or out of a polymer matrix by applying an electric field. Conductivity

changes can be controlled by applying such an electric field to control the number of solitons, polarons, or bipolarons on the polymer molecules[v].

Dopant ions or molecules can also diffuse into the polymer structure from the ambient atmosphere. Chemical sensors are based on the diffusion of specific chemicals into the polymer and the measurement of the resulting change in polymer electronic structure.

Polymer photodetectors rely on the energy of the photon to excite electrons from the valence band into the conduction band or into midband states of solitons, polarons, or bipolarons.

In a real polymer, there are several factors that complicate the mechanisms of conductivity presented above. Impurities, folds in the chain structure, chain ends, and differences between amorphous and crystalline regions can all contribute to local electronic states that act as traps for conduction band electrons. Conduction between the localized states occurs by hopping or quantum mechanical tunneling. Tunneling effects are also believed to be important in interchain conductivity.

References

[i] J.A. Pople and S.H. Walmsley, 'Bond alternation defects in long polyene molecules', *Molecular Physics*, 1962, **5**, p 15.

[ii] W.P. Su, J.R. Schrieffer, and AJ. Heeger, 'Solitons in Polyacetylene', *Physical Review Letters*, 1979, **42**, p. 1698.

[iii] D. Bloor, 'Electrical Conductivity', in **Comprehensive Polymer Science, Pergammon Press, New York, 1989, p. 687.**

[iv] R.E. Peierls, 'Quantum Theory of Solids', Clarendon Press, Osford, 1964.

[v] M.S. Wrighton, *Science*, 1986, **231**, p. 32.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Sensors and Actuators

CHAPTER 9

Diffractive Chemical Sensor Plastic Wrap

T. Kanigan, C. Brenan

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Diffractive Chemical Sensor Plastic Wrap

Dr. Tanya Kanigan and Dr. Colin Brenan

This section presents the concept of a plastic sheet imprinted with diffractive chemical sensing elements which change spatial period, and thus appearance, when chemical analytes are absorbed on the sensing elements. A method for fabricating diffractive chemical sensing elements in cellulosic films, and the envisioned utility of such a product in conjunction with the FOS spectrometer are also described.

Chemical Diffractive Sensors

A material having a periodic pattern of grooves in its surface may diffract electromagnetic radiation of wavelength λ according to the Bragg's law

$$m\lambda = n\Lambda\sin\theta$$

where m is the diffraction order, n is the refractive index, Λ is the grating spatial period and θ is the angle at which a diffraction maxima is formed.

For a grating with a spatial period of Λ, thickness of t and a refractive index modulation of Δn, the diffraction efficiency for light of wavelength λ incident at angle ψ is given by the following equation (Kogelnik, 1969):

$$\eta = \frac{[\sin(\nu^2 + \xi^2)^{1/2}]^2}{1 + \nu^2/\xi^2}$$

where

$$\nu = \frac{\pi\Delta n d}{\lambda\cos\psi}$$

$$\xi = \frac{\pi\Delta\psi t}{\Lambda}$$

and where $\Delta\psi$ is the deviation from Bragg angle, θ.

The wavelength of the light most efficiently diffracted by the grating depends on the grating spatial period. Polymer materials, such as polymer hydrogels, can be made to swell and contract when exposed to specific molecules (Holtz and Asher, 1997; Ley *et al.*, 1997). This is achieved by adding chemical recognition sites to the material. These chemical recognition sites may be functional groups covalently bonded to the polymer backbone, or separate molecules dispersed in the polymer film. When a chemical binds to such a recognition site, the material swells due to increased osmotic pressure. As the material expands, the spatial period of a diffraction grating changes producing a change in the grating spectral properties.

By choosing an appropriate spatial period and depth, the grating can be designed to diffract light efficiently enough to be seen with the naked eye (see Figure 1). Grating parameters may also be selected such that the binding of a chemical species of interest to chemical recognition sites within the grating material produces a visible change in the color of the diffracted light, or in the appearance or disappearance of the grating itself. This latter condition will occur when swelling or contraction causes the grating spatial period to shift into or out of the visible range.

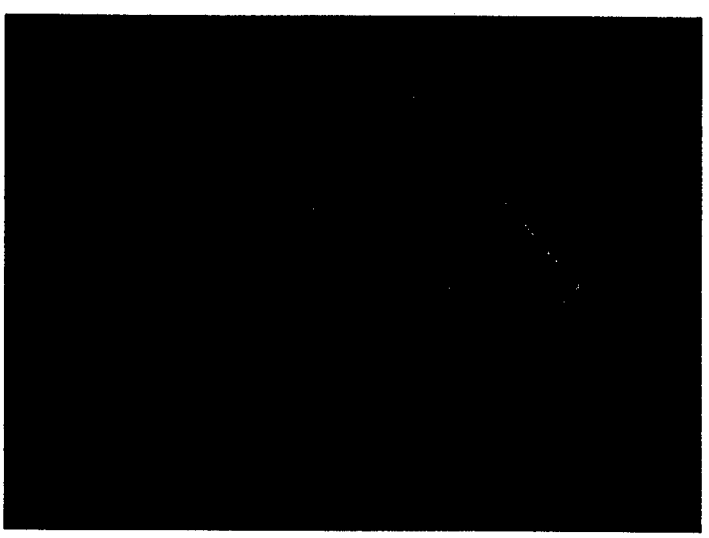


Figure 1 Surface relief grating etched in fused silica substrate (left) and its cellulose acetate replicate (right).

Fabrication of Diffractive Sensors in Cellulosic Thin Films

In order for diffractive chemical sensing elements to find widespread use in the home and healthcare industries, a versatile practical method for producing these structures in biocompatible materials. Our fabrication method is based upon the replication of a surface relief grating in a cellulose acetate film and its subsequent chemical conversion to a hydrophilic, insoluble, transparent, amorphous cellulose hydrogel layer.

Cellulosic polymers such as cellulose acetate, cellulose nitrate and regenerated cellulose are commonly as membranes to remove small molecules from solution. One of the most important uses of cellulose acetate is in reverse-osmosis membranes. Cellulose acetate reverse-osmosis membranes are integrally skinned asymmetric membranes comprised of a thin dense skin layer (0.1 to 0.25 μm) on top of a porous sublayer (~ 100 μm). They are prepared by a phase inversion process (Loeb and Sourirajan,) in which a film is cast from dilute solution (usually in a solvent mixture), partially dried, then plunged into a non-solvent. The skin layer controls the transport properties of the membrane and the porous sublayer provides mechanical support. This combination results in a membrane with high selectivity and throughput, suitable for water desalination.

Cellulose itself is an important membrane material used primarily for hemodialysis. These membranes are prepared from regenerated cellulose. *Regenerated cellulose* refers to both cellulose that has been dissolved in a solvent then reprecipitated, and to cellulose that has been recovered from a cellulose derivative by removing the functional groups. Most cellulose membranes are prepared by casting from cuprammonium hydroxide solution. This results in a membrane with a substantial degree of crystallinity. Although cellulose is very hydrophilic, it is insoluble in water and most other solvents due to the high density of hydrogen bonds which effectively crosslink the sample. Regeneration from a derivative with more favorable solution properties is the most convenient route to forming glassy cellulose films.

The fabrication of cellulose acetate/regenerated cellulose chemical diffractive sensors consists of three steps: replication of a surface relief grating etched in glass in a cellulose

acetate film, conversion of the grating from cellulose acetate to regenerated cellulose and functionalization of the cellulose with side chains that selectively immobilize specific chemicals of interest (analytes).

1. *Grating Replication: Transferring a Diffraction Grating from a Glass Master to Cellulose Acetate*

We have transferred a series of holographic gratings from a master pattern etched in fused silica onto a cellulose acetate film in the following manner. Cellulose acetate film were cast on the etched glass surface from a 15 % weight per volume solution in acetone, by pouring a small amount of the solution on to the glass, then tipping it on its side to drain and dry. Once the films had dried at room temperature, they were soaked in distilled water until delamination of the cellulose acetate film occurred. The gratings are visible in white light as colored reflections.

2. *Conversion to Regenerated Cellulose*

Cellulose acetate films are deacetylated by soaking in aqueous ammonia for several hours. Since this process occurs without imparting the polymer chains with enough mobility to form crystalline regions, the preferred thermodynamic state, the material remains optically homogeneous (i.e., transparent). The replacement of acetate groups with hydroxyl groups results in an increase in porosity and in the hydrophilicty of the material. The resulting cellulose film is more highly swellable in aqueous environments.

Figure 2 shows an image of a replicated grating collected with an atomic force microscope. The grating spatial period is approximately 410 nm. Figure 3 shows an atomic force microscope image of the same grating after deacetylation in a humidified environment. The spatial period has clearly increased by more than a factor of two.

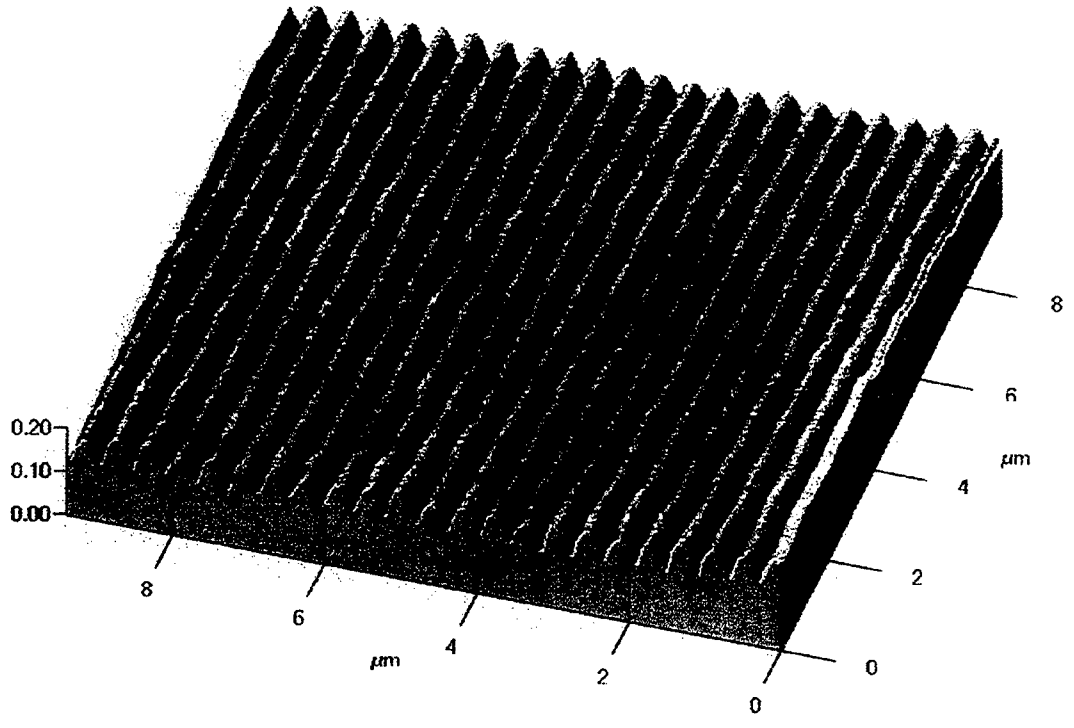


Figure 2 Atomic force microscope image of a cellulose acetate surface relief grating.

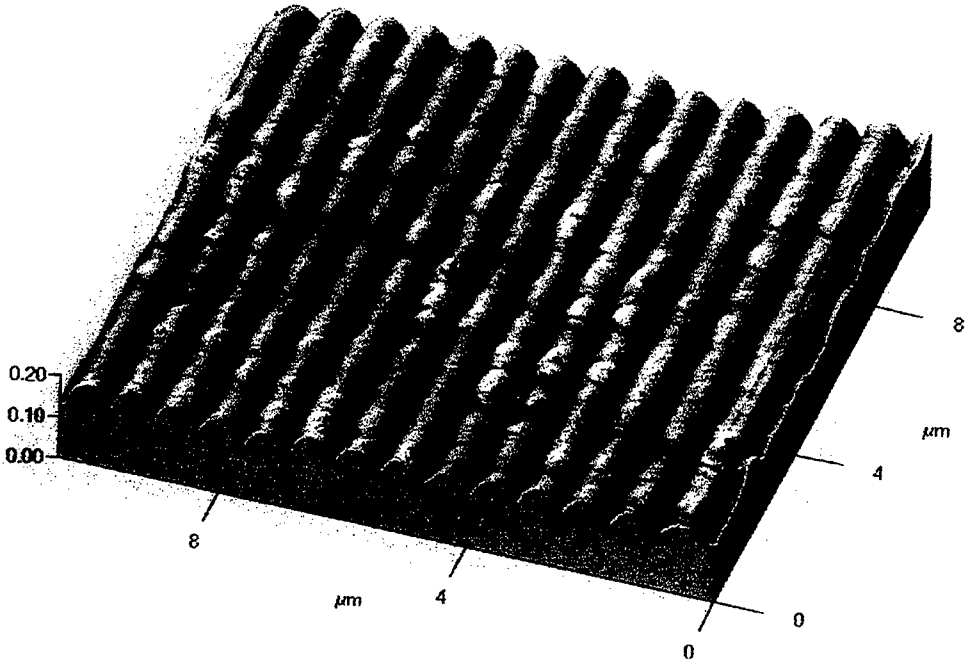


Figure 3 Atomic force microscope image of a surface relief grating in a regenerated cellulose film.

3. Functionalization: Sensitizing the Grating to Specific Chemicals

We are currently investigating methods of sensitizing our gratings to specific chemicals. One promising route is via chemical conversion of the hydroxyl groups on the cellulose backbone to carboxymethyl groups. The resulting carboxyl groups can be used to covalently bind antibodies to the hydrogel via their amino groups. A similar method has been used to immobilize antibodies on dextran, a polysaccharide very similar in structure to cellulose (Cush *et al.*, 1993)..

Cellulose repeat unit
R = OH or OAc.

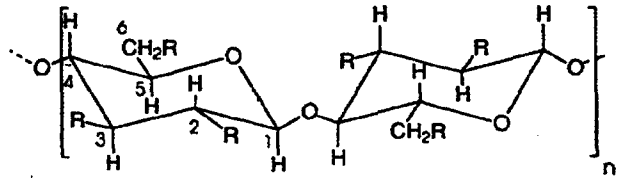


Figure 4 Chemical structure of the cellulosic repeat unit. For cellulose acetate R represents either a hydroxyl group (-OH) or an acetate group ($-O_2CCH_3$). For cellulose, R represents a hydroxyl group.

Collection, Detection, Inspection

The hydrolysis process which converts cellulose acetate to regenerated cellulose, proceeds at a faster rate on grating side of the film than it does on the opposite side of the film due to the densified skin layer that forms during the film casting process. Thus films can be prepared with cellulose hydrogel gratings on one surface while the opposite surface is comprised of denser cellulose acetate. This renders the film more mechanically robust and less penetrable to small molecules. If such diffractive chemical sensors are incorporated into a polymer film large enough to enclose some airspace or liquid space of interest, than this plastic wrap may function both as a chemical sensor and as a sample collector.

Although the presence or absence of absorbed chemicals can be determined by observing the color of the diffracted light when the grating is illuminated with a white light at some given angle, a simple optical system for quantifying the change in grating spatial period is desirable for more sensitive detection and quantification of the amount of analyte absorbed. We are in the process of designing such a system comprised of an inexpensive diode laser, a grating mount and a linear CCD array to record the spatial diffraction pattern.

Additional information about the chemical structure of trapped analytes can be obtained using non-destructive spectroscopic techniques such as Raman spectroscopy. The sensor wrap could be inserted directly into the compact fast orthogonal search (FOS) spectrometer being designed in our laboratory. If the presence of the grating interferes with the spectroscopic measurement, a proximal section of the film which lacks a grating, but does have chemical recognition sites can be sampled. The FOS spectrometer provides a means to confirm the identity of analytes and to discriminate between chemical species that respond to the same recognition sites.

Potential applications of this technology are monitoring of

- Water desalination and analysis: chemical diffractive sensors could be incorporated into cellulosic reverse osmosis and other membrane structures used for water purification.

- Chemical off-gassing: By covering an object with the sensing wrap, a headspace is established so that toxins released slowly and at low concentration can be collected over an extended period of time.

- Food Spoilage: The sensor wrap could be used by itself or as a layer in a plastic laminate in food enclosing products such as sandwich to detect chemicals released from rotting food.

REFERENCES

Cush, R., Cronin, J. M., and Stewart, J. (1993). The resonant mirror: a novel optical biosensor for direct sensing of biomolecular interactions. II. Applications., *Biosensors & Bioelectronics*, **8**, 355.

Holtz, J. H. and Asher, S. A. (1997). Polymerized colloidal crystal hydrogel films as intelligent chemical sensing materials. *Nature*, **389**, 829.

Kogelnik, H. (1969). Coupled wave theory for thick hologram gratings. *Bell. Syst. Tech. J.*, **48**, 2909.

Ley, C., Calderara, and Lougnat, D. J. (1997). Holographic gratings recorded in polymer hydrogels-an original application as a sensor in aqueous environment. *Measurement Science & Technology*, **8**, 997.

Loeb, S. and Sourirajan, S. (1964). U.S. Pat. 3,133,132.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Sensors and Actuators

CHAPTER 10

Nickel-Titanium Shape Memory Alloy Actuators for Home Automation

S. Lafontaine, I. Hunter

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Nickel-Titanium Shape Memory Alloy Actuators For Home Automation.

Dr. Serge Lafontaine and Prof. Ian Hunter

Introduction

In this report we review some of the experimental data acquired to study the characteristics of Nickel-Titanium (NiTi) actuators in home devices such as the SIMSUIT. The NiTi muscle-like actuators could be integrated in the suit to remove or simulate tremor. Other applications would include their use in mini or micro-scanning devices such as ultrasound, thermal or spectroscopic imaging systems.

It became clear after an extensive survey of current motor technologies that the lack of high-performance motors is one problem that limits the development of robotic devices (Hunter and Lafontaine, 1992a; Hunter *et al.*, 1991; Hollerbach *et al.*, 1991). This is also one of the main reasons why paraplegics and lower-extremity amputees still have to move in wheelchairs instead of walking with powered artificial limbs. Current motors cannot generate enough power per unit mass and enough force per cross-sectional area to be concealed in an artificial limb that could then actively assist in climbing stairs. This is also the reason why there are no adequate force reflecting gloves or power assisted force suits.

Nature has developed an actuator that is used throughout the animal kingdom. Muscles, on the other hand, have been in widespread use for over 300 million years throughout the animal kingdom, even though their overall performance is rather modest in many respects. Muscles are exceptional for their large range of motion (approximately 15% of their body length in-situ) and very long life time (over 2.5 billion cycles). However efficiency of conversion of chemical to mechanical energy is moderately low (35%), as are the force per unit area (350 kPa), power to mass ratio (50 W/kg, 200 W/kg peak) and bandwidth (10 to 20 Hz) (Hunter and Lafontaine, 1992). They are scaleable in design such that more force or displacement can be obtained by adding muscle fibers in parallel or series. Their stiffness changes over a range of 100:1 and can be modulated by co-contraction. Muscles have local energy storage for about 35 full contractions so an

external energy supply does not need to be provided for a high-speed low delay muscle response. On the other hand they have no "catch state" and require continuous energy expenditure to maintain a fixed position even though no mechanical work is done. Finally muscles cannot be used as a generator to recover energy from mechanical work.

In a review of new actuator technologies Hunter and Lafontaine (1992) evaluated a number of materials that could be used in artificial muscle-like actuators as fibers, films or rods and that could be assembled in a series or in a parallel configuration and controlled like muscle fibers. Out of that study two materials became of great interest: nickel-titanium (NiTi) shape memory alloys and electrically conducting polymers. NiTi alloys have reached the stage of being an engineering material that can be used in actual designs (Funakubo 1987; Duerig *et al*., 1990). They generate huge forces of more than 180 MN/m^2, about 700 greater than muscle, large displacements (>7%), and have very large power to mass ratios (>100 kW/kg). The main limitations are their very low efficiency (<2%) and a limited lifetime that can be compensated by redundancy given high power-to-mass ratio and force per unit area.

NiTi fibers and springs have frequently been proposed for use as artificial muscles because of their favorable material properties, low toxicity level, bio-compatibility, and reasonable cost (Bergamasco *et al.*, 1989; Duerig *et al.*, 1990a,b; Funakubo, 1987; Hirose *et al.*, 1990, 1989a,b, 1984; Hunter *et al.*, 1991, 1990; Homma *et al.* 1989). In 1989, Oaktree Automation Inc., developed the Finger-Spelling Hand, an anthropomorphic robotic device based on 36 NiTi actuators located in the forearm and hand, which duplicated the motions of a human hand accurately enough to form the characters of the sign language (Boggs, 1993). Hitachi (Nakano *et al.*, 1984) has developed a robot hand with three fingers. Reynaerts and Van Brussel (1994) developed a two-fingered hand with five degrees of freedom. Most of these devices were too slow to be used commercially.

Speed of NiTi Actuators

NiTi actuators are normally assumed to be slow actuators. The process by which they contract is one of phase transformation. They change from a body cubic centered (BCC) lattice structure above a certain temperature when they are in their Austenite

phase, to an face cubic centered (FCC) lattice structure below a lower transformation temperature where they are in a Martensite phase. The lattice structure is well defined in the Austenite phase. Cooling down the alloy results in a twinned martensite, which can be "de-twinned" and deformed passively when an external force is applied. Heating the alloy to bring it to the Austenite phase can be achieved extremely rapidly. A very brief, very high-energy electrical pulse is applied to the fiber and Joule heating can bring it to the Austenite phase in milliseconds. This is show in Figure 1, where the sudden increase in strain in the first 10 milliseconds corresponds to the change to the Austenite phase.

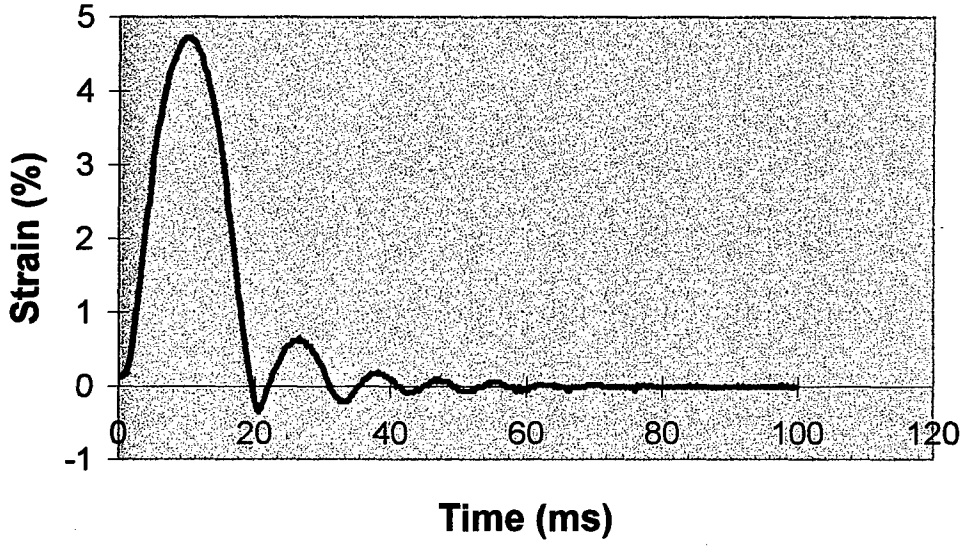


Figure 1 NiTi Pulse Response

Cooling the fiber however is inherently a slow, passive process. This is the main reason why NiTi fibers are believed to be slow actuators. However when short current pulses are used a series of phenomena permit give rise to a fast contraction. One of them seems to be inherent to a thermo-mechanical modification of the fiber. After the fiber has been conditioned with a number of such pulses an active expansion is also observed. Brief pulses may also give rise to a skin heating process that creates a layer of vapor, which insulates the NiTi fiber and constrained heating to the fiber itself. When the fiber contracts and the vapor is shed a fast cooling of the fiber occurs. Pulses however as shown in Figure 1 are faster than muscle fiber twitches and could be used to generate a quick opposing forces to cancel tremor in a human muscle.

Using two fibers in series and activating them sequentially produced even better temporal resolution. The speed of shortening and lengthening that is possible is seen in

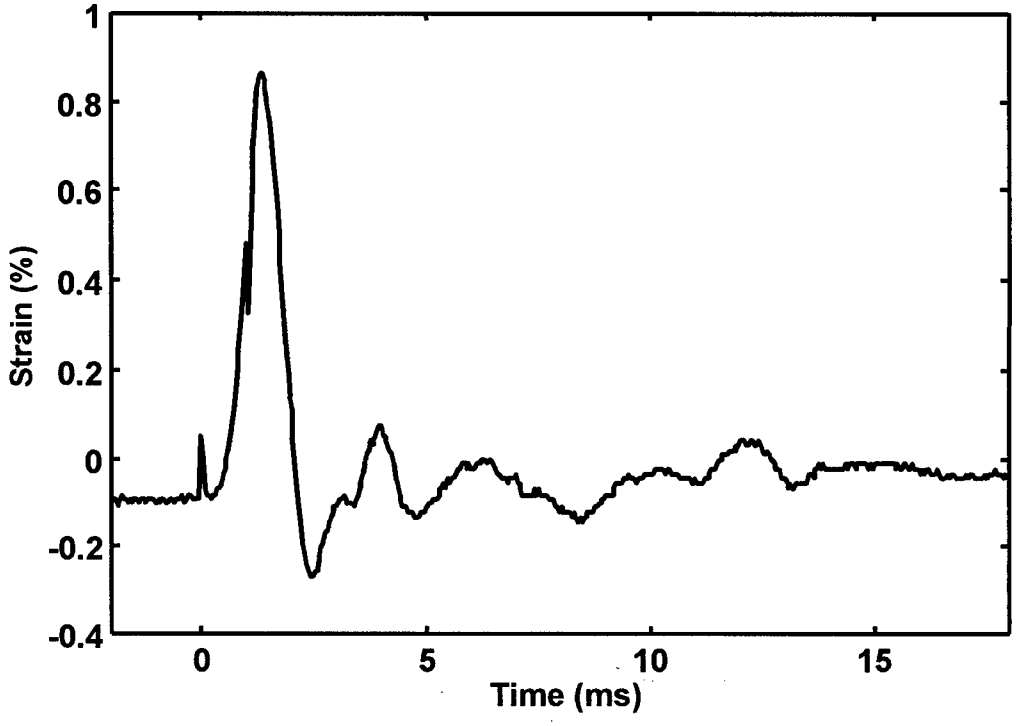


Figure 1 Response of NiTi fiber to double electrical pulses.

Figure 2 in which the interval between the two pulses was shortened to 1 ms. The width of the response is only 1.6 ms, which is more than an order of magnitude better than fast twitch muscle fibers. The velocity of shortening was nearly 1.7 m/s or 17 fiber lengths/s, which again is an extraordinarily high value, compared to muscle.

The use of very short pulses has other important consequences. Figure 3 shows the effect of different duration pulses in terms of the energy requirements in VA ms.. There is little difference in the length change produced for a given energy input with inputs briefer than about 2 ms, but the efficiency is less with longer pulses (5 or 10 ms).

Efficiency was studied more carefully in a second series of experiments. The integral of force (stress) times change in length (strain) represents the work done (work/m^3), which can be compared to the electrical energy input. The efficiency is low (little work is done) for energy inputs less than 1 Joule, but increases rapidly to between 2

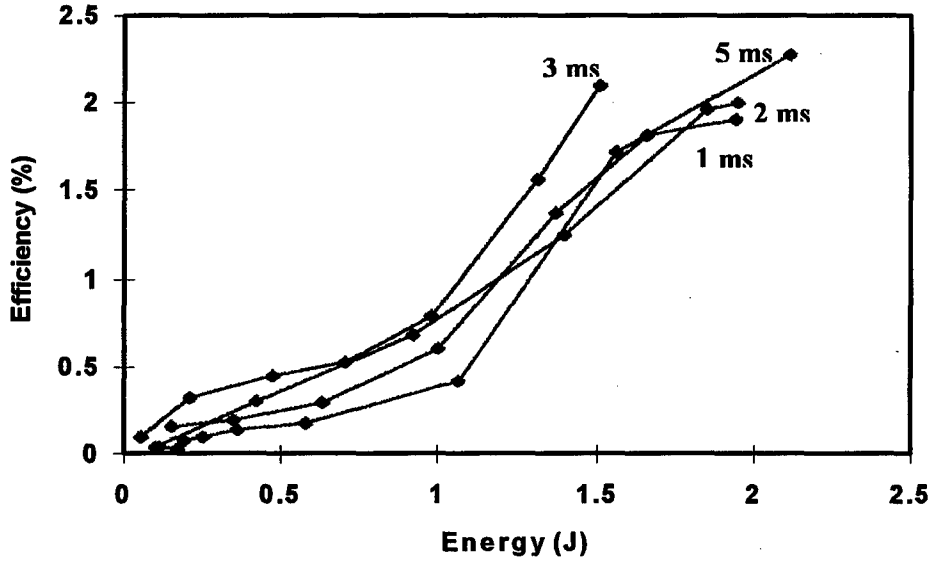


Figure 3 Efficiency of conversion of electrical energy to mechanical energy as a function of pulse duration.

and 3% for larger energy inputs. A steady stress of 40 MPa was applied in this series of experiments, which were conducted with the fibers in water at room temperature. The

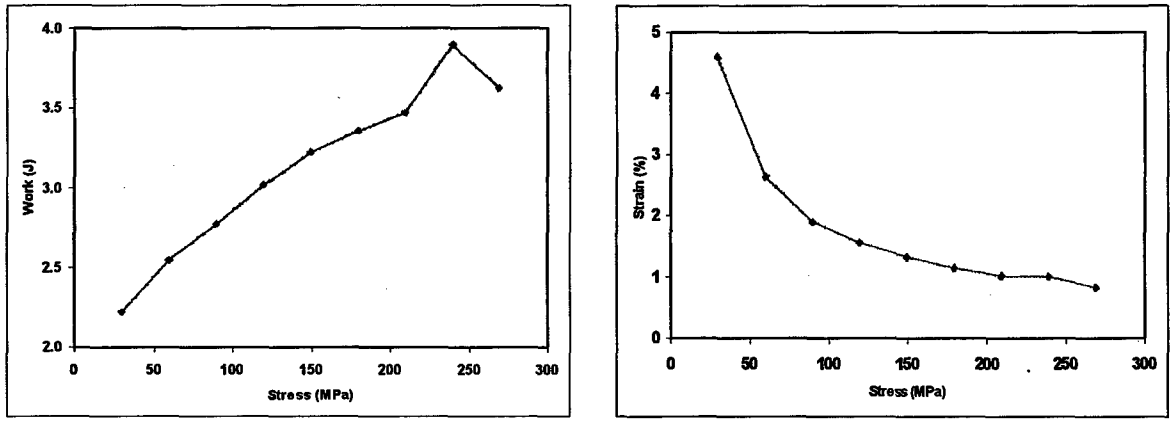


Figure 4 The amount of mechanical work done by the fiber depends on the stress maintained in the fiber and the strain in the fiber. Figure 9A at the left displays the mechanical work as a function of the stress. Figure 9B at the right displays the corresponding change in length.

energy efficiency was relatively constant for pulses below 5 ms, but declined abruptly for longer duration pulses (not shown).

The energy efficiency is relatively low, but was not optimized in these experiments. Two factors could be varied to improve efficiency: first the stress could be altered since the strain change depends on stress.

Figure 4A shows the effect of varying the stress on the strain and the product of stress and strain. The optimal efficiency was obtained at a stress level of between 200 MPa and 250 MPa in this experiment, but was only 1.3%. The speed of shortening is also reduced at higher stress levels as shown in Figure 4B.

Response to multiple pulses

One other point of interest is the response of the system to series of pulses. In Figure 3 responses are superimposed for 1, 2, 5, 10, 15 and 25 pulses at 5 ms intervals. Each pulse was 55 V and 1 ms in duration. With the fiber in water, the total response to one pulse was quite brief (20 ms), but subsequent pulses could add to the response and maintain a partially fused contraction. At higher rates the response would be completely

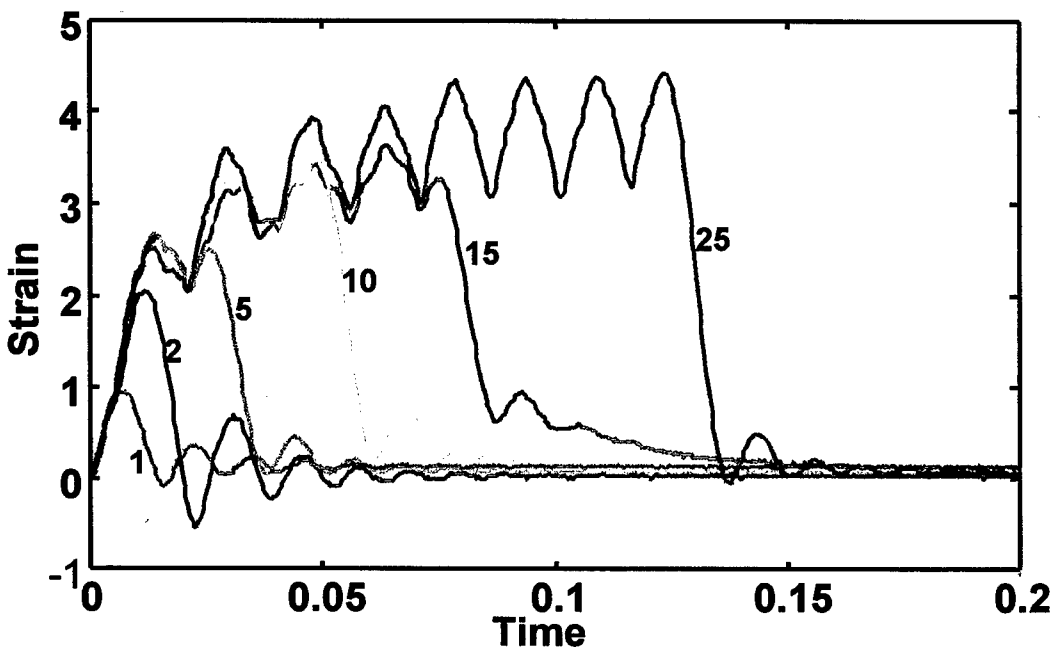


Figure 5 Response of NiTi fiber to multiple pulses.

fused. This behavior is reminiscent of the twitch and tetanic behavior of muscle fibers when stimulated with a brief pulse (the action potential) and a series of such pulses at

various frequencies. This pulse code modulation may be a more efficient way of activating NiTi fibers, as well as muscle fibers.

Variation in NiTi fiber stiffness

It is now well established that in the phase transitions of NiTi alloys, there are one or two intermediate phases as described by Beyer (1995). It is often assumed that there is a large change in stiffness associated with R-phase transformations as described by Jordan *et al.* (1994) and Wu *et al.* (1995). Figure 6 shows the dynamic variation in stiffness as the temperature is slowly cycled from slowly from -10 to +110 °C. The stiffness was measured by varying the stress applied to the fiber in a triangular manner and measuring the corresponding changes in strain. The stress-strain curve (not shown here) displayed large nonlinearities and the term stiffness is used here in a loose sense since this large change in stiffness is not observed under stationnary conditions. This apparent dynamic "pseudo-stiffness" changes over a large range in a way similar to

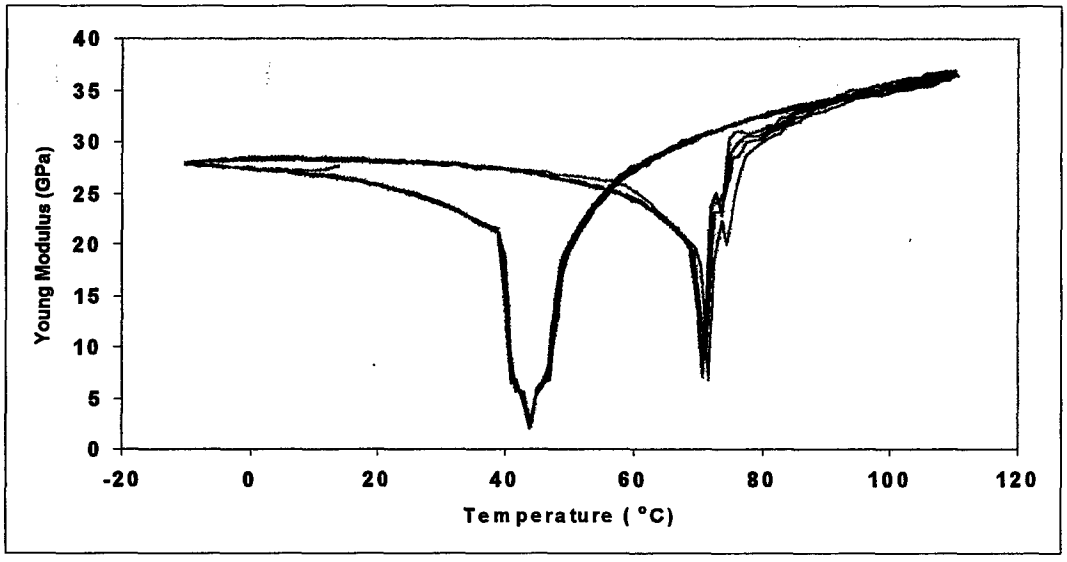


Figure 6 Dynamic stiffness of NiTi fiber.

muscle. This raises however the possibility of controlling tremor by stiffness control of NiTi fibers in the SIMSUIT.

Reproducibility of NiTi cycles.

As can be seen in Figure 6 the change in properties of NiTi fibers is highly reproducible as they undergo complete cycling in temperature changes. This is also the case for the change in length of NiTi fibers subjected to a constant stress as show in figure 7. A number of cycles are plotted and they all superimpose accurately on each

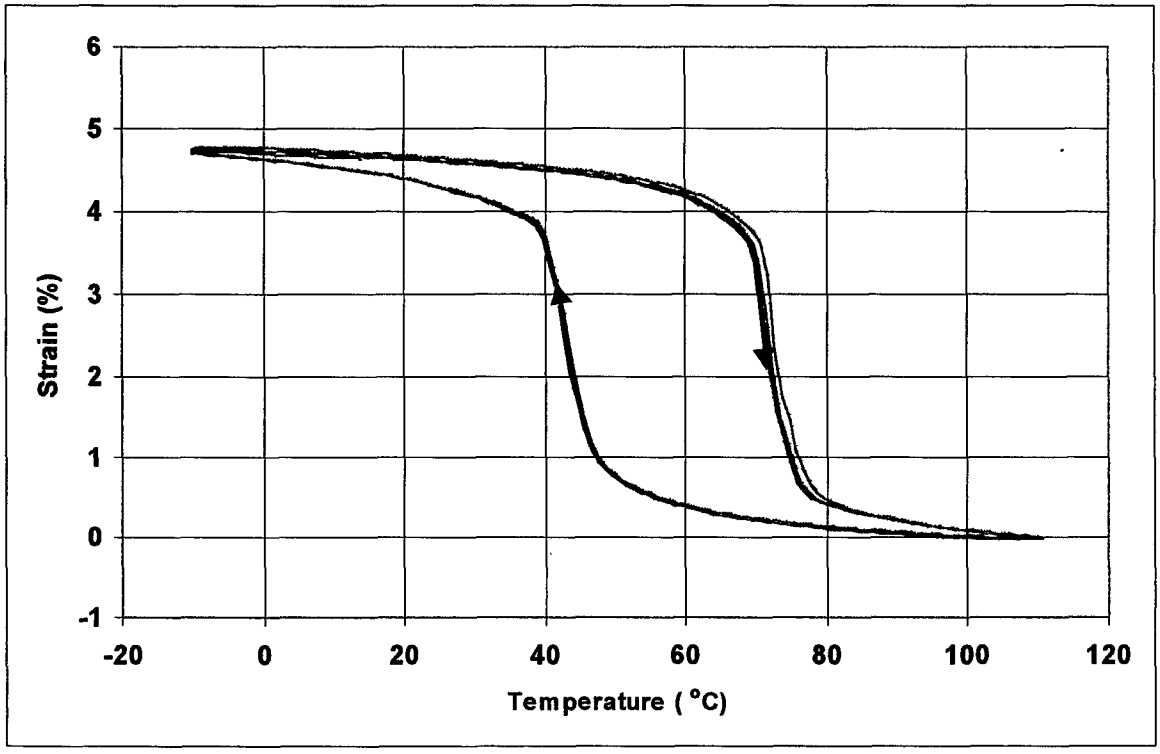


Figure 7 Repeatability of strain changes in temperature cycling of NiTi fibers.

other. This property could be used to implement an inexpensive mini-scanning system for laser beams or ultrasound. This is particularly the case for the cooling part of the cycle. The repeatability is such that open-loop control could be sufficient for scanning. Given that a good controller would have to take into account the large nonlinearities of these fibers, in a scanning device a higher resolution could be obtained by measuring the position of a mirror or sensor. It is important to mention also that this repeatability is observed only after a number of cycles when complete phase transformations are achieved. If for a cycle a partial transformations only was obtained, the next cycles would deviate markedly from the steady-state one.

Figure 7 also shows the large hysteresis in temperature observed in the change in

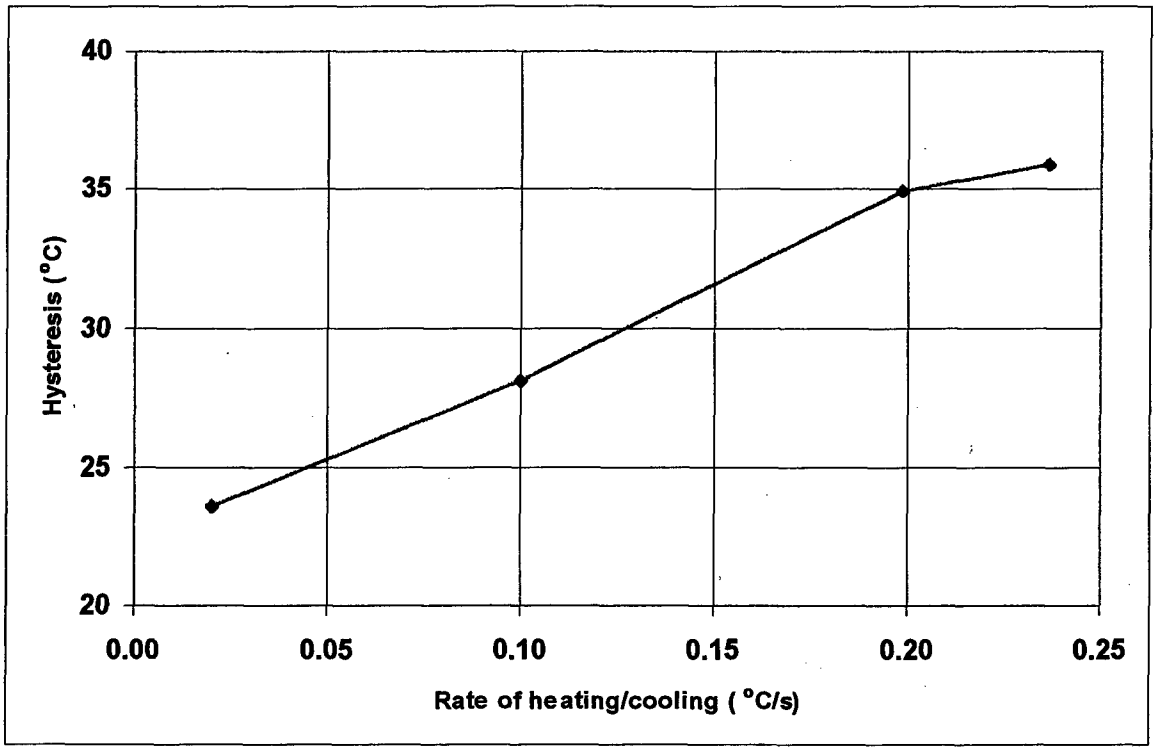


Figure 8 Variation in hysteresis of NiTi strain changes with rate of heating.

strain. The fiber shortens in this case at a temperature of approximately 70°C while in re-lengthen only at approximately 45°C. This large hysteresis of 25°C is normally observed in NiTi alloys and displays the highly nonlinear behavior of these fibers.

This change in hysteresis seems to somewhat decrease with rate of temperature change but only to a small extent. Figure 8 shows a plot of how the hysteresis changes with the rate of temperature changes. It can be seen that even with very slow cycling the hysteresis never tends to zero. When extrapolated the hysteresis would tend to 22°C for very slow changes in rate. Even though this effect is small, precise scanning would require precise control in temperature rate changes also.

However there might be situations where a much more linear response of the NiTi fibers might be obtained. This could be obtained by limiting the phase transformation to from the Austenite phase to the R-phase and vice-versa. This is shown in Figure 9 which displays the change in resistivity of a NiTi fiber with temperature cycling. The resistivity

is a function of phase transformation and temperature. The bulk resistivity can be described in a first approximation by the usual rule of mixed fractions of solutions. In Figure 9, the low resistivity at high temperature is characteristic of a pure Austenite phase. The higher resistivity is characteristic of an intermediate rhomboedric (R) phase, between the Austenite and Martensite phases. The hysteresis observed in this case is less than 1.5% and is characteristic of a secondary phase transformation. This transformation results in much smaller strains, limited to approximately 1.5% as opposed to 5-8% for the full martensitic phase transition, and much smaller forces, approximately 40MPa. The R-phase transformation however is not observed in all NiTi alloys and special alloys or

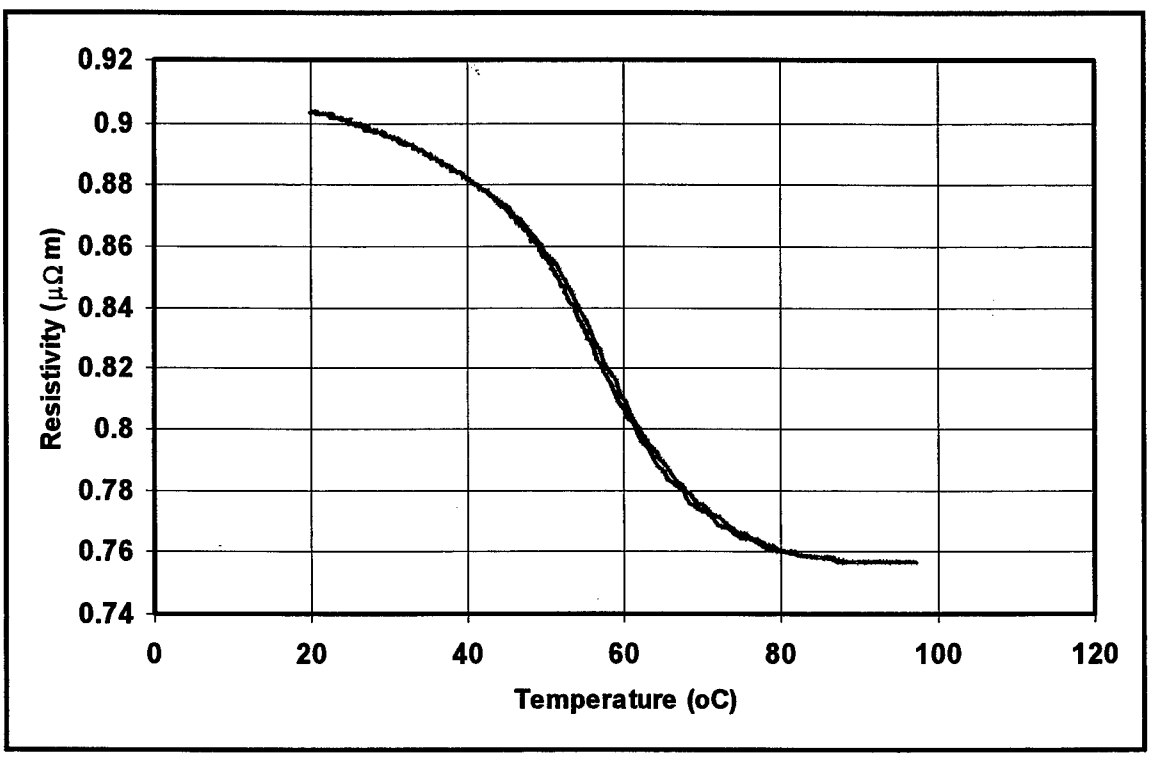


Figure 9 Hysteresis of the R-phase transition in some NiTi alloys.

thermo-mechanical processing of the fibers is required.

The electrical impedance of a NiTi fiber was measured in a frequency range extending from 100 Hz to 40 MHz. The results are shown in Figure 10. The data

indicates that the NiTi fiber was purely resistive up to several hundred kHz. The implication of this result is that if the resistance of the fiber can be measured at the same time as the fiber is electrically heated for control purposes, a high frequency electrical signal can be used for impedance measurements in a frequency band outside the frequency band used to control the fiber.

However the resistivity is not a simple measure of change in length or tension. Another resistivity plot is given in Figure 11 where the temperature cycling goes from. At temperatures below 20°C the resistivity is not anymore a simple one-to-one function of temperature, while there is no appreciable change in length. When a stretch is applied the diameter and length of the fiber change, as well as a small change in volume occurs with phase transformations. All of these would have to be taken into consideration in converting a resistance measurement to a resistivity measurement. The resistivity thus obtained would indicate the ratio of phase transformation. In this alloy however there are 3 phase transformations in the cooling process and mainly 2 phase transformations in the heating phase. All of these would need to be integrated in a proper model in order to use meaningfully resistance measurement for control purposes.

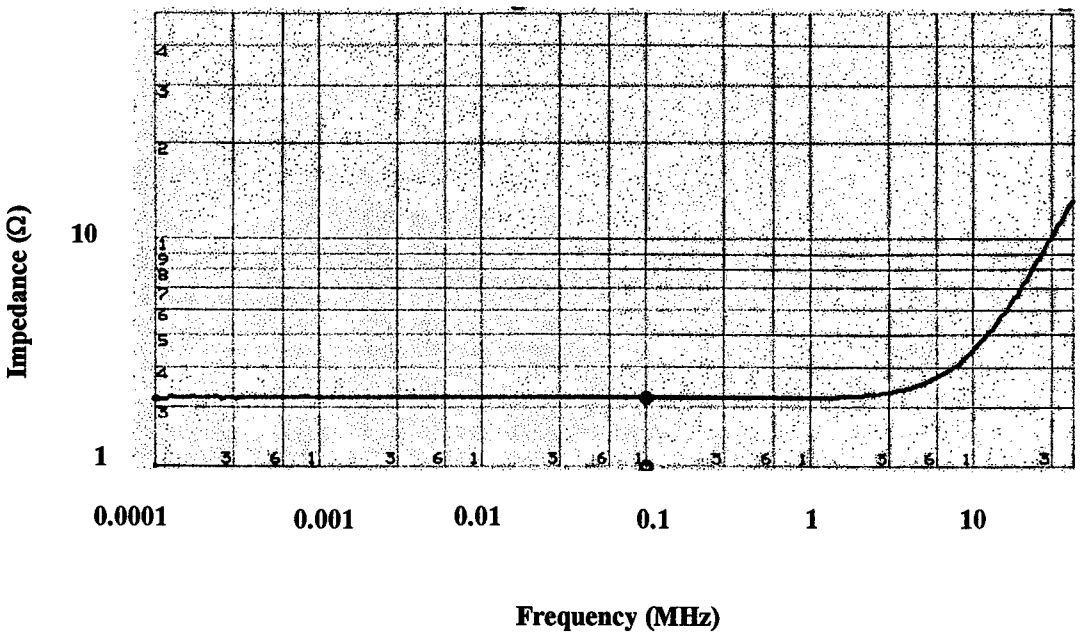


Figure 10 Impedance measurement of NiTi fiber at 20 °C.

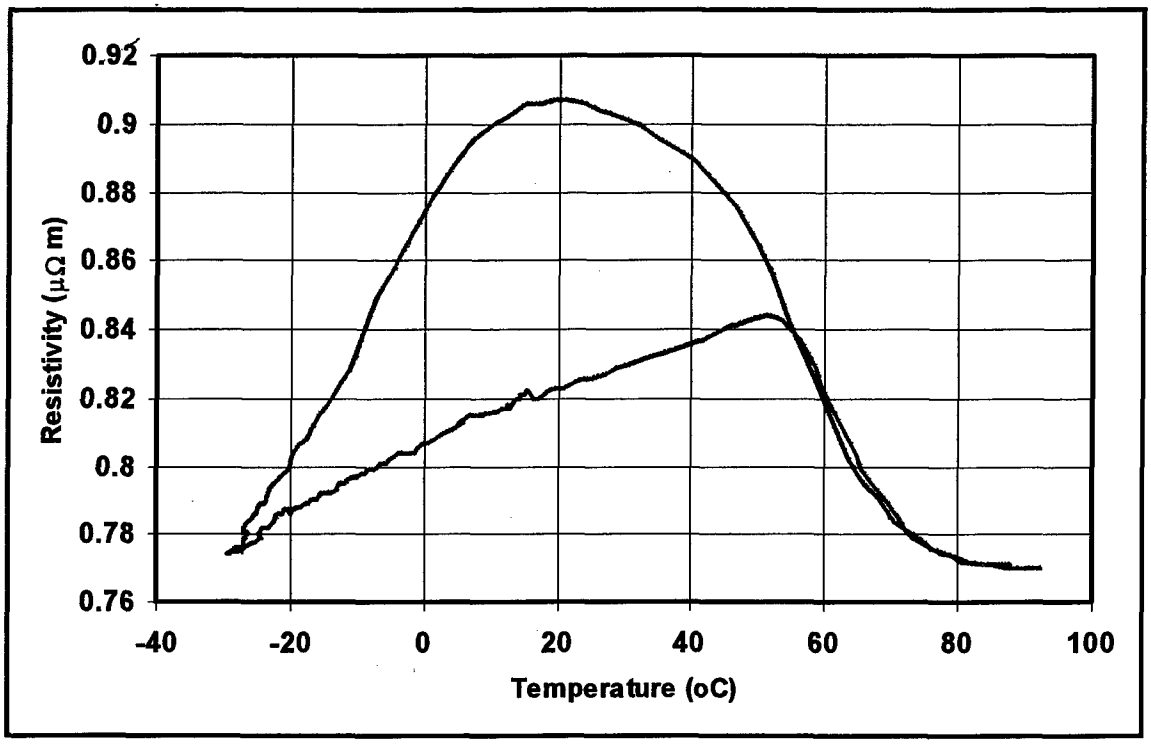


Figure 11 Resistivity curves of a commercially available NiTi alloy for a complete phase transformation.

Conclusion

We have briefly discussed in this report some results made on NiTi fibers which indicate how they can used for applications such as physiological tremor removal or simulation or in scanners where they would move a mirror or sensor. Besides being highly nonlinear, NiTi alloys are known also to have a limited life time and low efficiency. However, there are numerous applications where efficiency is not a critical issue and where power is readily available. Furthermore, material research may lead to more efficient shape memory alloy actuators. Efficiencies over 5% have been measured in a new type of shape memory alloy which undergoes martensitic transformations through the application of a magnetic field (Ullakko *et al.*, 1997).

The number of contractions, active or passive, which can be achieved before the amplitude of the response decreases or failure occurs in the material. The number of cycles depends on several factors such as the alloy used, the amount of strain and stress obtained in each contraction and the type of thermo-mechanical process used to achieve

the final properties of the SMA. In modern alloys several millions of cycles can be obtained given limits are not exceeded on stress and strain. This limitation can be further taken into account in system designs and circumvented in several ways. Redundancy or cycling through a number of fibers can be used in many situations given that the fibers can generate large forces per unit area and a number of them can be mounted in a small workspace.

Bibliography

Bergamasco, M., Salsedo, F. and Dario, P. A linear SMA motor as direct-drive robotic actuator. *IEEE International Conference on Robotics and Automation*, 1989a, 1, 618-623.

Bergamasco, M., Salsedo, F. and Dario, P. Shape memory alloy micromotors for direct-drive actuation of dexterous artificial hands. *Sensors and Actuators*, 1989b, *17*, 115-119.

Beyer, J. Recent advances in the martensitic transformations of Ti-Ni alloys. *Journal de Physique IV*, 1995, *5(Colloque C2)*, C2-433-443.

Duerig, T.W., Melton, K.N., Stöckel, D. and Wayman, C.M. *Engineering aspects of shape memory alloys.* Boston: Butterworth-Heinemann, 1990.

Funakubo, H. *Shape memory alloys.* New York, NY: Gordon and Breach, 1987.

Hunter, I.W. and Lafontaine, S. Shape memory alloy fibres having rapid twitch response. US patent 5,092,901, 1992.

Hirose, S., Ikua, K. and Umetani, Y. Development of shape-memory alloy actuators. *RoManSy '84, the 5th CISM-IFTOMM Symposium*, 1984.

Ikuta, K., Tsukamoto, M. and Hirose, S. Mathematical model and experimental verification of shape memory alloy for designing micro actuator. *Proceedings of IEEE Micro Electro Mechanical Systems*, 1991.

Jordan, L., Chandrasekaran, M., Masse, M. and Bouquet, G. Study of Phase Transformations in Ni-Ti based shape memory alloys. *Journal de Physique IV*, 1995, *5(Colloque C2)*, C2-489-494.

Lu, X. A systems approach to modelling and design of high strain shape memory alloy actuators. Master of Engineering Thesis, Department of Electrical Engineering, McGill University, Montreal, 1997.

Shaw, J.A., Kyriakides, S. On the nucleation and propagation of phase transformation fronts in a NiTi alloy. *Acta Materialia*, 1997, 45(2), 683-700.

Silling, S.A. Dynamic growth of martensitic plates in an elastic material. *Journal of Elasticity*, 1992, 28(2), 143-64.

Ullakko, K., Hunag, J.K., Kokovin, V.V. and O'Handley, R.C. Magnetically controlled SMA in Ni2MnGa intermetallics. *Scripta Materiala*, 1997, *36(10)*, 1133-1138.

Wu, K., Dalip, S.K., Liu, Y. and Pu, Z. Damping characteristics of R-phase NiTi shape memory alloys. *Smart Structures and Materials SP IE-Int. Soc. Opt. Eng.*, 1995.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Sensors and Actuators

CHAPTER 11

Noninvasive Blood Glucose Quantitation Using Spectroscopic-based Optical Technique

K. Youcef-Toumi, V. Saptari

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Noninvasive Blood Glucose Quantitation using Spectroscopic-based Optical Technique

Prof. Kamal Youcef-Toumi
Principal Investigator

Vidi A. Saptari
Graduate Research Assistant

March 31, 1998

Abstract

Quantitation of blood components is a routine and significant task in medical diagnostics. The established methodology uses different enzymatic assays implementing photometric and electrochemical detection, which are time and labor consuming. The main research emphasis is on the development of a noninvasive, spectroscopic-based blood glucose monitoring device. Upon accomplishment of this objective, noninvasive detection of several blood analytes may be considered. At this stage, the feasibility of such device is being studied, and proper instrumentation is being investigated. Literature review of several techniques researched by other investigators has been completed, with the purpose of identifying an appropriate technique for this task.

1. Introduction

Great interest in research has been directed toward development of novel, non-invasive technologies as diagnostic tools. One such desired technology is a spectroscopic-based optical system for blood compositional analysis. This technology will provide fast and simple method, replacing older chemical techniques that are labor-intensive and time-consuming.

Few of the important blood components for chemical pathology are glucose, protein, cholesterol, alkaline phosphatase and calcium. Among them, glucose stands out as the most pursued substance in this area of research. This is due to the importance of its monitoring in diabetic patients, which will eliminate the need for painful withdrawal of blood four to six times a day (further explained on the following section).

Several groups around the world have sought to realize this goal without considerable success. It is in this area that our research will be mainly emphasized, although our ultimate goal is to measure several blood analytes simultaneously for medical diagnostics purposes. The intended plan for the first phase of this project was to examine the different methods being investigated by these groups, and to identify the limiting problems of each technique. The methods were compared and the most promising one was decided on. In this paper we report our findings.

1.1. Diabetes Facts: Motivation

Diabetes mellitus is a chronic metabolic disease in which there is a deficiency of, or a resistance to an effective use of insulin. This disorder, along with its associated complications, is ranked as the seventh leading cause of death in the United States [1]. Diabetes is grouped into two types: type I or insulin-dependent, and type II or non-insulin-dependent.

Type I diabetes affects approximately 700,000 people in the United States alone. People who have this illness are dependent on insulin injection to live (at least one shot a day). It is very important for them to check their blood sugar levels as frequent as possible to ensure that they are within acceptable limits at all time.

Type II diabetes affects approximately 15 million people in the United States, which corresponds to 9 out of 10 of all cases of diabetes. Although they are not dependent on insulin injection, they also need to practice good blood-sugar control, to avoid or reduce the risk of complications due to high glucose level. Diet, exercise and pills usually suffice in managing this type of diabetes.

Although the evidence is now overwhelming that frequent glucose monitoring and tighter control of blood glucose can significantly reduce the risk of complications associated with diabetes [2], patients are often reluctant to perform the procedure, because the common "finger-prick" technique is invasive, painful and inconvenient. Therefore, an elegant, non-invasive measurement/monitoring technique is desirable.

1.2. General Aspects for Noninvasive Spectroscopic Measurement

For a non-invasive spectrometric assay, usually the skin or some other part of the body tissue is integrally probed. The metabolic information of interest is usually gained from the blood, which constitutes only a relatively small fraction of the tissue volume under investigation (figure 1). Several investigators have considered measuring glucose in the interstitial fluid of the subcutaneous tissue minimal-invasively, by extracting the fluid using microdialysis, suction effusion or reverse iontophoresis technique[3-5]. However, there is still some controversy regarding the relation of glucose concentration contained in the interstitial fluid and in the blood. Some groups claim that under steady state conditions, glucose concentration in the tissue is practically identical to that in the blood. However, delay of 5 to 15 minutes is observed when glucose concentration increases sharply, for instance, after meal or sugar intake [3]. Furthermore, not all blood constituents have identical concentration in the subcutaneous tissue [4]. Due to these reasons, blood might be the best specimen for measuring metabolite concentrations.

As mentioned before, the signals from the blood constitutes only a small fraction of the spectrum. Furthermore, the metabolites of interest usually have low concentrations (glucose concentration in the blood is about 0.1% for normal individuals [13]). As a result, the necessary selectivity for reliable glucose or other analytes prediction may only be gained by multivariate data analysis exploiting information from broad spectral ranges.

1.3 Emerging Technologies for Noninvasive Glucose Measurement

Several optical techniques for noninvasive measurement of blood and tissue glucose have been proposed. This includes near infrared absorption/diffuse reflectance, Raman scattering, polarimetry and near-infrared scattering technique. This section gives an overview and describes the significant features of each technique.

1.3.1. Near-Infrared Absorption/Diffuse Reflectance

This technique is the oldest and the most widely researched method. Near-infrared refers to light radiation in the wavelength region between 700 to 2,500 nm, which has the advantage of being able to probe more deeply into tissue compared to the visible, ultraviolet and mid-IR. When a tissue is irradiated with near-infrared radiation, it either absorbs, transmits or scatters the light. The absorption in this wavelength region is due to the overtone and combination vibrations of molecules, where the scattering is due to discontinuities in refractive index of tissue on the microscopic level. The absorption and scattering coefficients denoted by μ_a and μ_s respectively, in units of mm^{-1} are the rates of radiant energy loss $-d\varphi/dz$ due to absorption and scatter per incremental unit photon length dz in a tissue [6]. In diffuse reflectance the light that is scattered back to the surface after "sampling" the intended tissue volume is analyzed (figure 1).

Several investigators have shown that this method, coupled with statistical multivariate data analysis can predict clinically useful glucose concentration [7,8]. The biggest problem, however, is the irreproducibility of the result due to *in vivo*, physiological variations such as the skin temperature, as well as procedural variations,

such as measurement probe pressure [7]. In addition, each individual has a unique skin and tissue property, which necessitates calibration to be performed for each user.

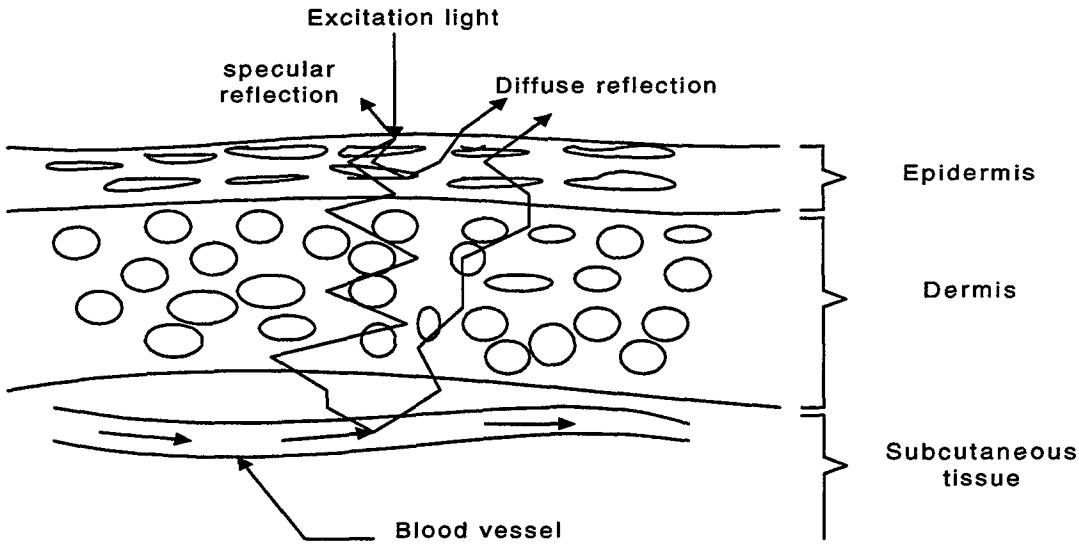


Figure 1. Schematic presentation of probable photon paths in the skin and tissue.

1.3.2. Near-Infrared Raman Spectroscopy

Raman spectroscopy has been utilized over the past thirty years mainly by physicists and chemists. The inherent difficulty of this method is that its signals are very weak, having the intensity of about 10^{-3} of the Rayleigh scattered light. Only recently, with the replacement of slow photomultiplier tubes with faster CCD arrays and the manufacture of higher power near infrared laser diodes, has the technology become available to allow researchers to consider the possibility of tissue diagnosis and blood chemicals analysis *in vivo* and in real time.

The phenomenon of Raman scattering is observed when a monochromatic light is incident upon an optically transparent (negligible absorption) media. A small portion of the light is scattered inelastically, exhibiting frequency shifts, which are associated with transitions between rotational, vibrational and electronic levels [9]. For *in vivo* tissue studies, a laser in the near-infrared region is usually used as an excitation source to minimize fluorescence background signal.

A significant advantage of Raman technique over near-infrared absorption technique is that its spectrum has distinct and pronounced peaks, easing the task of separating signals that generate from the metabolites of interest. The primary problem of this method, however, is the inherent weakness of Raman signal, creating the need for prohibitively high excitation power and relatively long signal collection time in order to get an acceptably useful spectra. Photothermal damage of the tissue is of great concern for *in vivo* measurements using this technique.

1.3.3. Polarimetry Technique

The rotation of linearly polarized light by optically active substances has been used for many years to quantitate the amount of substance in solutions[10]. The concept behind this technique is that the amount of rotation of polarized light by an optically active substance depends on the thickness of the layer traversed by the light, the wavelength of the radiation, the temperature and the pH of the solvent, and the concentration of the optically active material [11]. This technology is used industrially to measure the concentration of sugar in foods.

For polarimetry to be used as a noninvasive method, the signal must be able to pass from the source, through the body, and to a detector, without total depolarization of the beam. Since the skin possesses high scattering coefficients, maintaining polarization information in a beam passing through a thick piece of tissue including skin (for eg. a finger) would not be possible. Several investigators have suggested to measure glucose in the aqueous humor of the eye [11]. The path length would be on the order of 1 cm, which is the average width of the anterior chamber of a human eye.

There are several problems with this approach for glucose measurement. First, the signal size is small. The angle of rotation for a 1 cm thick tissue compartment would be $< 0.00004^0$ per 1 mg/dl increment in glucose concentration. Furthermore, other optically active substances contribute to the signals, increasing or decreasing the polarization angle, thus questioning the specificity. Second, there is a time lag between blood and aqueous humor glucose concentrations during periods of rapidly shifting blood glucose concentrations. Other problems include corneal rotation and eye motion artifact.

1.3.4. NIR Scattering Technique

Where NIR absorption technique considers both the absorption and the scattering of the light, this technique only measures the latter. It has been found that the presence of glucose in an aqueous solution increases its refractive index and therefore has an influence upon the scattering properties of particles suspended in solution [12]. As the glucose induced scattering changes are small, any possible application for noninvasive glucose monitoring has to rely on an accurate separation of scattering and absorption coefficients. Kohl et al.[12], used an intensity-modulated frequency domain NIR spectrometer to separate the scattering coefficient from the absorption coefficient.

The biggest concern with this method is its specificity, since other physiologic effects unrelated to glucose concentration could produce variations on the scattering coefficient. For example, changes in skin temperature would affect the scattering coefficient even greater than the changes due to glucose concentration variation alone. The measurement precision of the scattering coefficient and separation of scatter and absorption changes is another concern with this technique.

2. Discussion

	NIR absorption	NIR Raman	NIR scattering	Polarimetry
Specificity	moderate	high	low	low
Signal strength	high	low	high	high
Stability	low	high	low	na

Table 1. Method Comparison

Although we will focus on glucose measurement exclusively as the short-term plan of this project, we keep in mind that the ultimate goal is noninvasive multicomponent blood analysis. Therefore, it is essential that our method of choice has the capability for such task.

For the plausibility of a multicomponent blood analysis, specificity is the most important factor. Signals from different blood constituents should be able to be distinguished from each other, so that each can be quantitated independently. In this regard, Raman technique is superior. NIR absorption, NIR scattering and polarimetry technique should be thought of as purely empirical methods, relying heavily on statistical multivariate analysis. Another significant factor to consider is stability; that is the sensitivity of the technique to disturbances or noises. The most significant noise that has been a problem for NIR absorption and scattering methods is the skin temperature variability. It has been found that Raman spectra are not as sensitive to temperature changes as the NIR absorption and scattering spectra. However, this still needs to be confirmed.

The biggest problem that faces Raman technique, however, is its intrinsic weak signals. Furthermore, significant part of the spectra collected will be generated from the skin and tissue above the blood vessels. As a result, in order to get an acceptable signal-to-noise ratio, this method requires an unsuitably high excitation power and long collection time for *in vivo* and in real time measurement.

In conclusion, among the four potential optical methods, Raman spectroscopy is the most promising for accurate and reliable multicomponent blood constituents quantitation. The spectra has sharp and distinct peaks, allowing greater discrimination among closely-spaced signals and consequently for more accurate extraction of concentrations from the spectral data. However, this method is still faced with several critical problems. It is our plan in this research to address them.

3. Present Status and Research Objective

In this first phase of our research, feasibility studies and identification of the most appropriate instrumentation are currently underway. Overview of other emerging technologies has been completed. Raman-based optical technique has been identified as an appropriate method for the objective of this research.

In this project, we plan to create an enabling technology to for noninvasive blood glucose measurement with clinical accuracy. The system that we develop would have the capability for further improvements; that is for making multicomponent blood analysis possible. First, the two limiting problems mentioned before have to be addressed:

1. acquiring acceptable signals, given an allowable excitation power and collection time
2. dealing with "unwanted" signals generated from the skin and tissue

4. Conclusion

It is anticipated that Raman spectroscopy will play an important role in clinical chemistry in general, and in blood assay specifically. Development of new and improved optical instruments/components will be extremely beneficial. However, procedural and instrumental design for a specific application is also essential for realization of such goal. It is in this area that our research will to focus on.

References

1. Barnwell M. et. al., Diabetes, Skidmore-Roth Pub., El Paso, Tex., 1995
2. The Diabetes Control and Complications Trial Research Group, New Engl. J. Med. 329, 977-986 (1993)
3. Pfeiffer E.F., "The *Ulm Zucker Uhr system* and its consequences", Horm. Metab. Res. 26: 510-514, 1994
4. Kimura J., "Noninvasive blood glucose concentration monitoring method with suction effusion fluid by IFSET biosensor", Applied Biochemistry and Biotechnology 41: 55-58, 1993
5. Rao G. et al., "Reverse iontophoresis: noninvasive glucose monitoring *in vivo* in humans", Pharm. Res. 12 (12): 1869-1873, 1995
6. Wilson B.C., Jacques S.L., "Optical reflectance and transmittance of tissues: principles and applications", IEEE Journal of Quantum Electronics 26(12): 2186-2199, 1990
7. Marbach R., Heise H.M. et al., "Noninvasive blood glucose assay by near-infrared diffuse reflectance spectroscopy of the human inner lip", Appl. Spect. 47 (7): 875-881, 1993
8. Muller U.A. et al., "Non-invasive blood glucose monitoring by means of near infrared spectroscopy: methods for improving the reliability of the calibration models", The International Journal of Artificial Organs 20 (5): 285-290, 1997
9. Colthup N.B., "Introduction to Infrared and Raman Spectroscopy, 1990 (pg. 62-64)
10. Browne, C.A., and Zerban, E.W. (1941), Physical and Chemical Methods of Sugar Analysis, 3rd ed., John Wiley & Sons, New York, pp. 263-265
11. Cote G.L. et al., "Noninvasive optical polarimetric glucose sensing using a true phase measurement technique", IEEE Trans. on Biomed. Eng. 39 (7): 752-756, 1992
12. Kohl M. et al., "Glucose induced changes in scattering and light transport in tissue simulating phantoms", SPIE Vol. 2389: 780-788, 1995
13. Marbach R., Heise H.M., "Optical Diffuse reflectance accessory for measurements of skin tissue by near-infrared spectroscopy", Applied Optics 34(4): 610-621, 1995

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Sensors and Actuators

CHAPTER 12

Signal to Noise Enhancement for an Invisible Marking System Using an Infrared Activation System

H. Asada, R. Doubleday

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Signal to Noise Enhancement for an Invisible Marking System Using an Infrared Activation System

Haruhiko H. Asada
Principal Investigator

Rolland L. Doubleday, Jr.
Graduate Research Assistant

Abstract

A system was developed that uses an IR laser diode to fluoress an ink that absorbs at 785 nm and has emission at 830 nm, and a CCD camera is used to capture this low strength ink emission. There are many things that have emission in this band, and for this reason, the signal-to-noise ratio is very small. Extremely robust methods were developed to deal with the noise in the viewable frame. The Perturbation/Correlation method is the most noteworthy method, and uses a sine wave superimposed on a DC value for laser intensity, and then monitors the change in intensity of the returned signal over this perturbation in laser intensity. These changes in viewable frame intensity are then correlated to the known perturbation in laser intensity, giving a large enhancement in the signal to noise ratio.

Most of the work that has been performed so far has been in the development of methods that would result in a greater S/N ratio at any given laser power. Now, the actual characteristics of the noise are being investigated to determine when and how these signal processing techniques should be applied, as well a general classification of sources of noise for later intelligent scan planning.

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Table of Contents

List Of Figures

1. Introduction

One of the main obstacles to home automation, such as home robotics, is the inability to locate and identify thousands of objects in a highly unstructured environment. In a well-defined environment, such as a laboratory, pattern recognition techniques have been used to identify and locate a few items within a small area of search. However, in the home environment, this stringent definition of environment does not exist. For this reason, a system was developed to locate and identify objects without any requirements placed on item orientation or search area. Since it was to be used in the home, the system had to be non-intrusive as well as meeting ANSI standards on eye and skin safety.

The system developed is composed of a laser diode that emits at 785 nm and fluoresces a laser dye that has peak absorption at this wavelength. The ink emission is at 833 nm, and a CCD camera that is sensitive in this near IR region is used to capture this emission signal. Since both the laser and ink emission are at wavelengths above what the human eye can see, both the activation and acquisition are invisible to the human, meeting the non-intrusive requirement.

The laser dyes that have absorption and emission in the near IR also have low quantum efficiencies associated with them, which is just the ratio of photons returned in emission to the total photons that are used to fluoress the ink. Due to this low returned signal strength, very robust signal processing methods had to be developed to enhance the signal-to-noise ratio (S/N). With every increase in S/N ratio, the laser power could be reduced, making the system even more eye and skin safe. However, with too great a reduction in laser power, the S/N ratio would drop and errors in the read would occur.

This research will address such issues as the system architecture, the signal processing algorithms used to deal with the low signal strength, and some work in the description of the noise in the viewable frame. Some alternate designs will be presented with regard to light diffusion and modulation, and some ideas for future applications of this work will be discussed.

Some possible application of this technology is in object location and identification of parts in parts feeders, counterfeit checking, spotlight tracking of performers, surgical tool preparation, stamp and money counting, and any inventory or logistics system. A provisional patent application was made on this system.

2. A System Description

The system consists of a laser diode that emits at 785 nm with a nominal power of 20mW. This emission wavelength is the peak absorption frequency of a special formulation of IR 125 laser dye and solvent. This solvent/dye combination has a peak fluorescent emission at 833 nm. This laser light is taken through a spatial filter to reduce the effects of spatial aberrations in the laser source optics and provide a nice gaussian distribution in power. This light is then taken though a double convex and concave lens combination that is used to expand the 6.96 mm elliptical beam to 42 mm (1:6).

On the side of image acquisition, the image is taken through a notch filter with a pass range of 825-835 nm, with the center wavelength roughly the wavelength of the peak ink fluorescent emission. This is passed through a 10:1 motorized zoom lens. Three channels of D/A are used to control the iris, zoom, and focus of the lens, and one channel of A/D is used to bring in the zoom position information.

The ink emission signal is fairly weak due to an associated low ink quantum efficiency, and for this reason, an image intensifier was used to amplify the light signal by a factor of 30,000. A monochrome CCD camera is then used to measure this conditioned image intensity signal, with a PCI bus frame grabber used to bring this information into the computer.

2.1 The System Design

A requirement was set early on that the system had to be non-intrusive. To meet this requirement, it was desired to provide a system that was totally invisible to the humans within the home environment, as well as meet eye and skin safety standards. In addition to this, a practical requirement was to provide a signal that could be easily differentiated from noise. To accomplish this, a detection method had to be found that would detect at wavelengths above or below what a human could detect. A silicon CCD chip was found to have a spectral response that started at around the same wavelength as the human eye, but extends roughly 200 nm above the top end of the human eye. Figure 2.10 gives a graphical illustration of this feature. The red area denotes the desirable design range for picking an ink and its corresponding detection system.

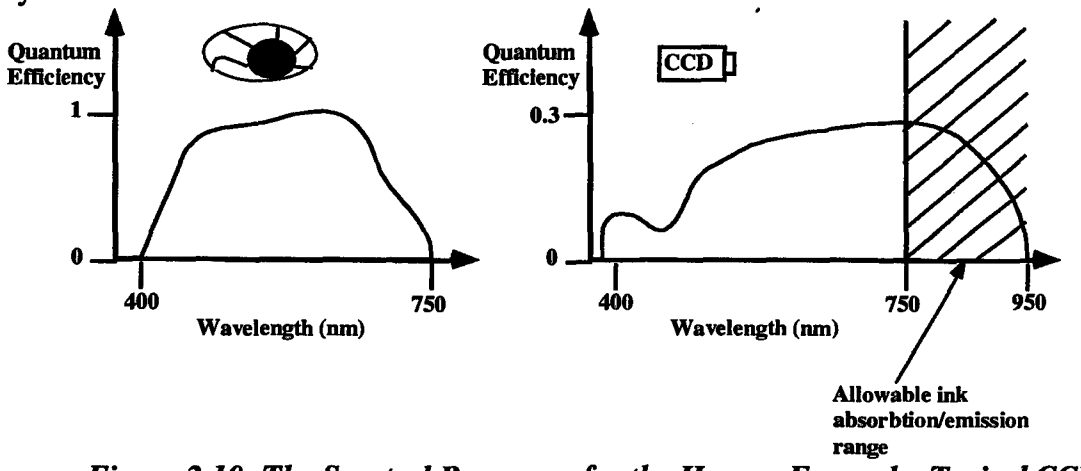


Figure 2.10: The Spectral Responses for the Human Eye and a Typical CCD Chip

Some basic design rules were developed at this point to constrain the design direction for the laser dye activation system. First, the bottom end of the ink absorption curve had to be higher than the top end of the human eye spectral response. Specifying this rule leads to a search of all inks/dyes that have absorption at just above 750 nm, which corresponds to the top end of human spectral response. Second, the peak of the ink emission curve had to be below the top end of the CCD spectral response. This rule is used to keep the cost of detection relatively cheap, since cameras designed specifically for IR applications are upwards of $10,000, and a monochrome CCD detector can be obtained for a few hundred dollars. Third, the top of the laser emission curve had to be below the bottom end of the notch filter response curve. If this rule were not obeyed the laser emission would overlap the region of ink emission, making detection of the low strength ink emission very difficult. This also gives motivation for finding a laser source with a fairly narrow wavelength operating range. The fourth rule that was used to select an ink\activation system was that the product of the quantum efficiencies for the camera, camera optics, ink, and filter must be such that at a given laser power flux, the ink is detectable. This is the most limiting rule, and does not take into consideration such issues as light diffusion and search time. Finally, the "nominal" power flux provided by the laser source and diffuser combination had to be eye and skin safe. A strong coupling between laser light diffusion and signal detectability should be noted. Figure 2.11 shows, for the chosen system, the ink absorption and emission, the laser emission and the notch filter pass range.

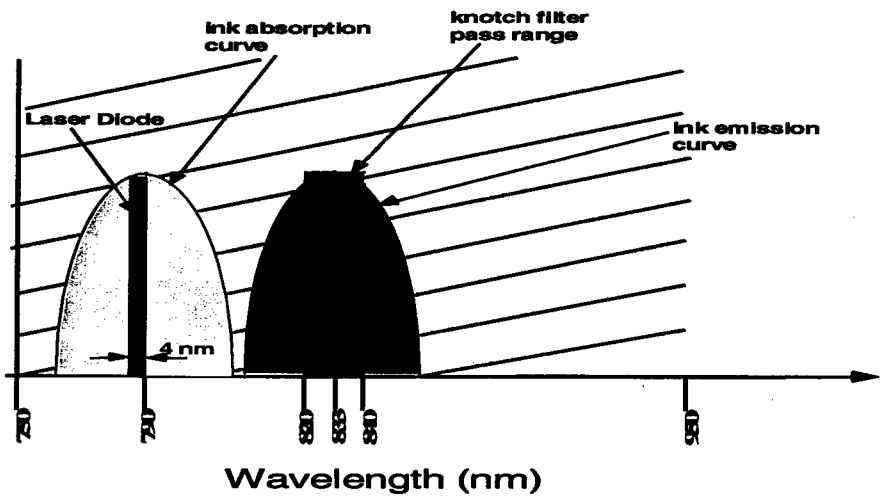


Figure 2.11: Relative Spectral Locations of Ink Activation System

3. Noise Reduction Methods

3.1 The Perturbation Correlation Method

The Perturbation/Correlation method works very well in reducing noise in the viewable frame when the noise is of the constant, repeatable type. Examples of such noise sources are tungsten lamps, sun reflections, and other types of light or heat sources that have a wide band of emission.

The Perturbation\Correlation Method is designed specifically to increase the signal to noise ratio by actively changing the parameters that are known to bring the most change in the signal, and then performing a correlation to the known perturbation input. The ink was known to fluoress at a given wavelength, and the fluorescent emission strength was found to be a function of the power of the light source at this wavelength.

Due to a lack of sensitivity to this parameter change, the magnitude of repeatable noise remains fairly constant over this perturbation. When a correlation is made between the noise+signal measurement and the perturbation, the noise falls away due to a weak correlation and only the ink signal remains. It should be mentioned that this method need not be restricted to a power variance of the light source, but could be applied to any measurable parameter that is sensitive to perturbation.

For the case of this application, a sinusoid perturbation in light power is applied, and a certain number of frame array (intensity) are sampled over the sine period, and for each frame, a spatial gradient is performed. This gives an idea of the areas of the frame that are most sensitive to the given perturbation. A sequence is then formed that is composed of pixel intensity values for the same pixel over the many frames, and a discrete time correlation is made of this sequence to the perturbation input. The following is the sequence that would be constructed:

$$Y[n,i,j] = [A_0(i,j), A_1(i,j), A_2(i,j), A_3(i,j), .., A_N(i,j)] \tag{3.10}$$

This sequence can then be correlated with the known power perturbation sequence:

$$X[n,i,j] = [B_0 + B_1 \sin(2\pi n / T_k)] \tag{3.11}$$

The correlation function, in discrete time, is then written as:

$$\phi[n] = \sum_{k=-\infty}^{+\infty} x[n+k]\,y[k] \tag{3.12}$$

Again, $x[n]$ is the lamp or laser power, and $y[n]$ is the sensed intensity function for each pixel over the discrete sample period.

3.2 Noise Modeling

A lot of work has been done with regard to signal to noise enhancement through the above mentioned Perturbation/Correlation method and signal time averaging, and very remarkable gains in S/N ratio have been realized. However, up until now, these methods have been applied

blindly to both the large amplitude noise case as well as the low amplitude noise case, without any real means of capturing the characteristics of the noise in the viewable frame at any instant in time. With some knowledge of the noise makeup, it is possible to apply these image processing algorithms in a smart manner, as well as to find an optimal search path based on the probability of signal detection at any spatial location (if the signal were present).

3.3 Hardware Setup For Noise Measurement

Figure 3.30 shows a simple spectrum sampler that was constructed by using a bank of photo-diodes that were placed behind notch filters at different center pass wavelengths (refer to Figure 3.31 for specific filter information). The photo-diodes were chosen on the basis of the high gain requirement as well as the linearity of the spectral response for the photo-diode for wavelengths between 430 nm and 880 nm (the visible to the near IR range). Some signal conditioning was performed to change the small amplitude current, produced from incident radiation on the detector's surface, to a 0-5 V signal that could be read by the DAS 1800 A/D. Op-amps were used to provide isolation to the converted current to voltage signal, as well as to provide a "zero and span" for the calibration of the individual detection system. A requirement was found for this calibration since each diode had a different base voltage signal (voltage when there is no radiation on the detector surface). This problem was taken care of by using the zero portion of the circuit, and the different filter pass gains could be nulled by appropriate settings of the span circuit.

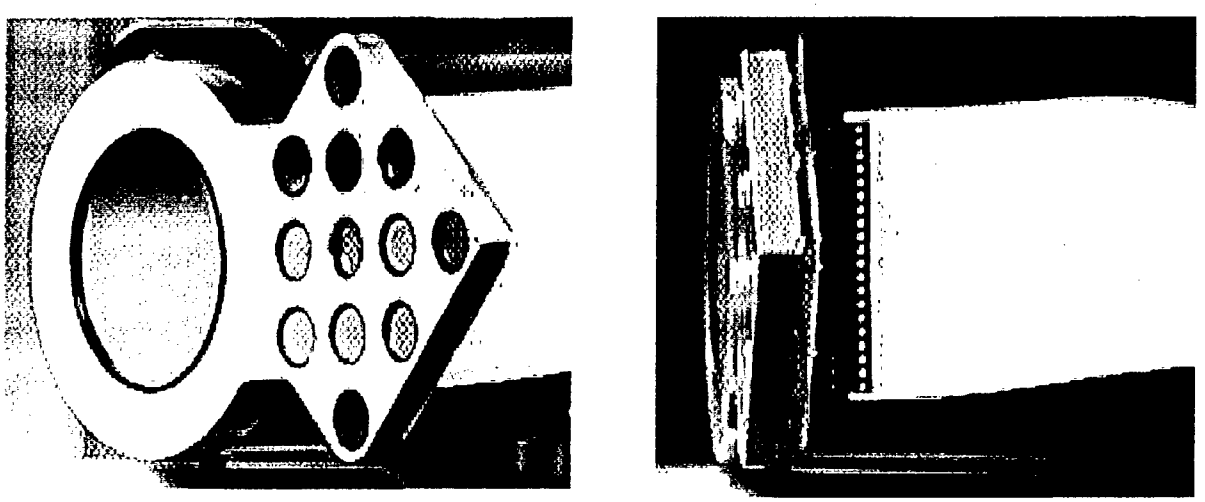


Figure 3.30:The Noise Measurement System

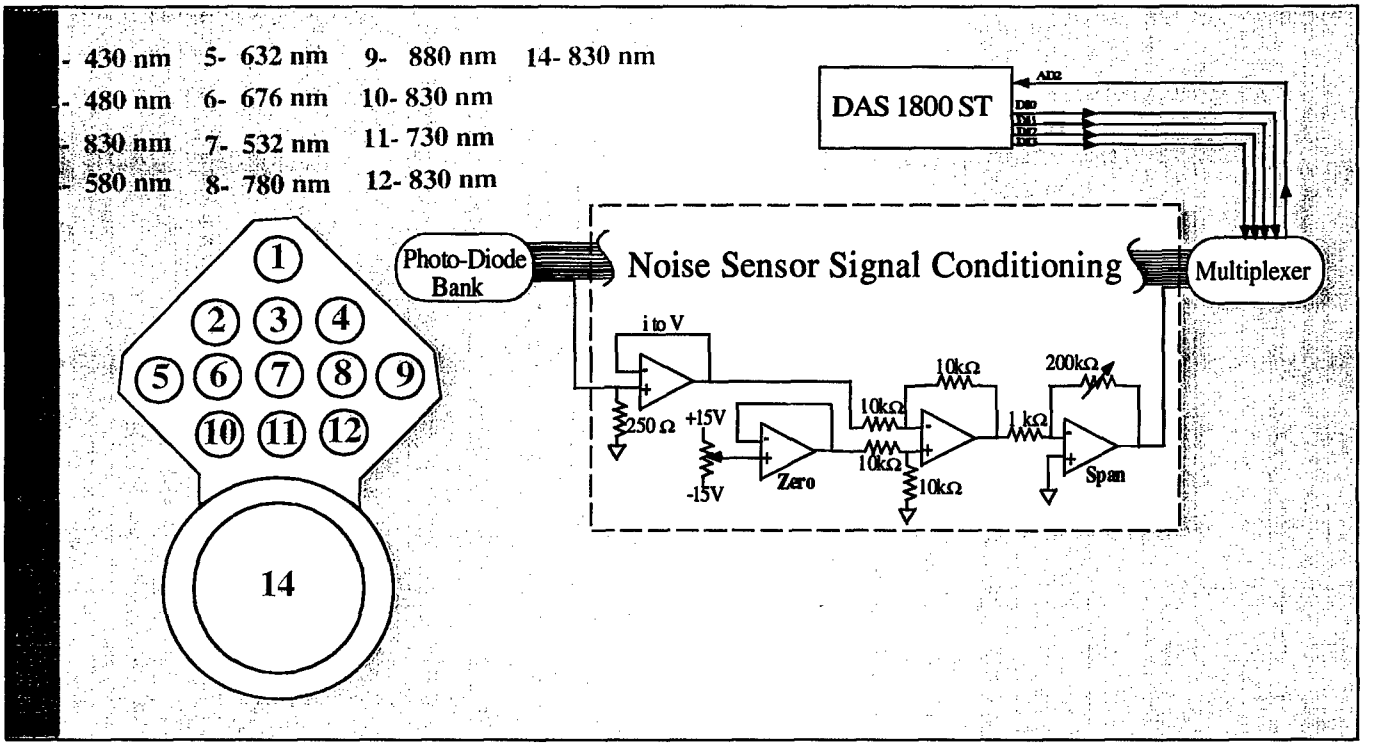


Figure 3.31: Signal Conditioning and Pass Ranges for the Spectrum Sampler

3.4 Calibration Procedure for the Noise Measurement System

The circuitry developed for the collection of the incident radiation at a particular wavelength was calibrated by using a narrow band light source of known intensity. Before the filter was placed in front of the photo-diodes, each diode was covered with black electricians tape and the zero circuit potentiometer was adjusted to give zero volts out for each photo-diode, measured at the output of the signal conditioning circuitry. Then, with the tape removed, a hollow tube was placed over the diode and the light source was shined upon the detector surface. The span potentiometer was then adjusted to give 5V at the output of the signal conditioning circuitry. This was done for each of the twelve diodes.

At this point, the calibration became rather intuitive. The feedback resistance of each span element was measured at the given calibration. Then, after checking the pass gain found from the spectral response curve for each filter, this gain was divided into 1 and multiplied by the span resistance measured. This gave a new resistance to which the span potentiometer was then set. This, in effect, compensated for the attenuation of the signal as it was passed through the notch filter. By doing this, it was possible to equate a 5V signal to the wattage of the light source used in the calibration, but now at a given wavelength. From this calibration procedure, it was possible to develop curves that represented the noise in the viewable frame and find the power of this noise signal over a given wavelength range.

3.5 One Motivation for this Noise Power Estimation

Often, a simple, black body emission model is used to describe noise sources such as tungsten lamps and other broad band emitters. For instance, a good approximation for the tungsten

element in an incandescent light bulb is a 1 mm black body emitter at 3000K. For these particular black body approximations, a relationship for the monochromatic emissive power for a black surface is used:

$$E_{b\lambda} = \frac{C_1 \cdot \lambda^{-5}}{e^{\left[C2(\lambda T)^{-1}\right]} - 1} \tag{3.50}$$

Where $C_1=3.742x10^8$ $W\mu m^4/m^2$ and $C_2=1.4389x10^4$ μm K (Reference Mills), and $E_{b\lambda}$ is in $[W/(m^2\mu m)]$. Figure 3.50 gives a plot of the black body emissive power as a function of wavelength and at five discrete temperatures.

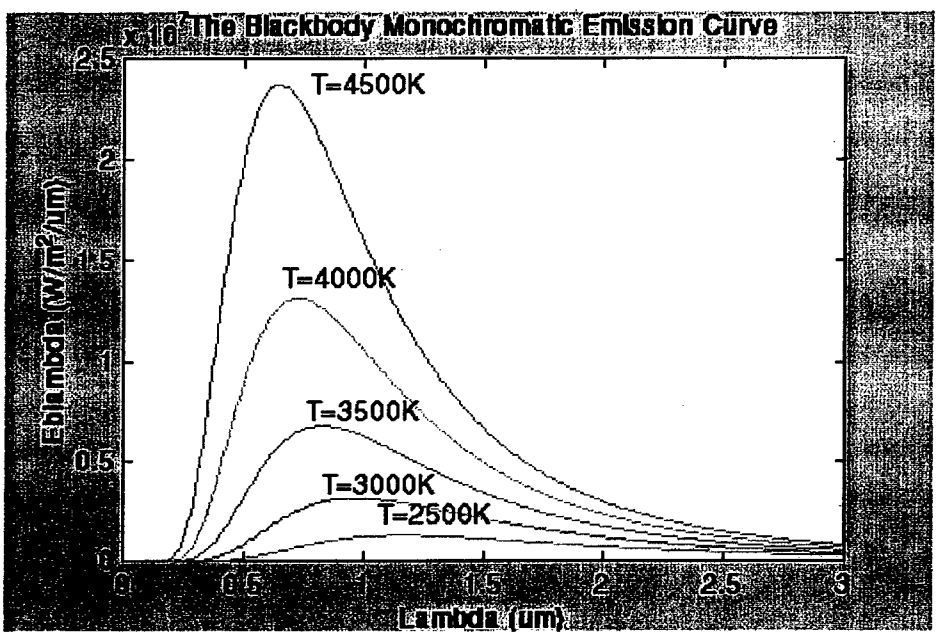


Figure 3.50: The Black Body Emission Curve for Varying Body Temperatures

The total emission of a black body over all wavelengths is:

$$E_{bb} = \sigma T^4 \tag{3.51}$$

The amount of power contained between zero and some wavelength, λ, can be found by integrating equation 3.50 with respect to lamda, and the percentage of total power found in this range can be found by dividing by the total black body emissive power. This new quantity is referred to as the external fractional function (Mills):

$$f_e(\lambda,T) = \frac{\int_0^{\lambda} E_{b\lambda}(T,\lambda)\,d\lambda}{\sigma T^4} \tag{3.52}$$

This fractional function is shown in Figure 3.51

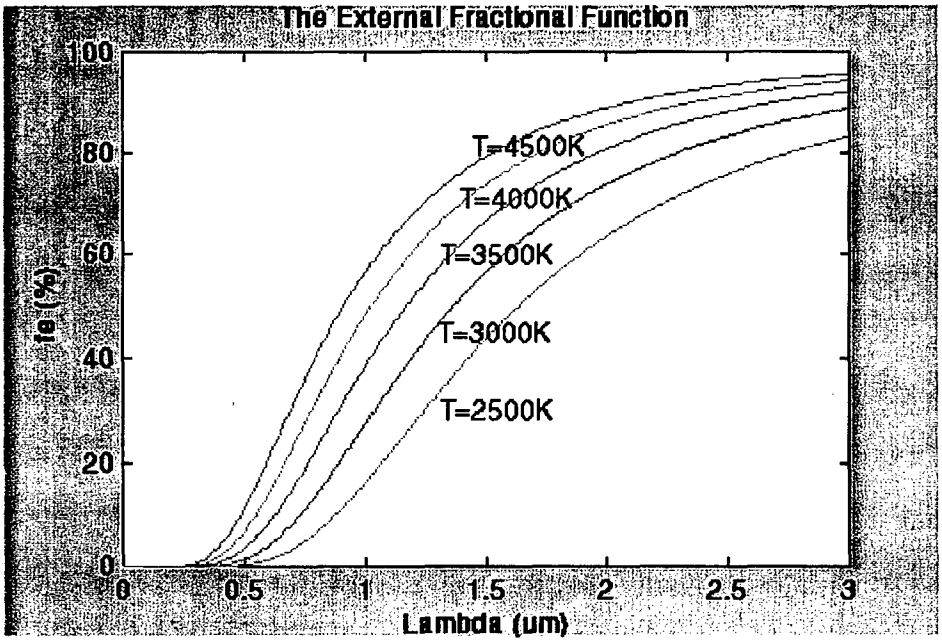


Figure 3.51: A Plot of the External Fractional Function For Varying Body Temperatures

In the case of noise measurement in a specific band, what is required is the emissive power between two wavelengths, which is obtained by first subtracting the fractional function at the bottom wavelength from the fractional function of the top wavelength:

$$\Delta f_e = f_e(\lambda,T)_{top} - f_e(\lambda,T)_{bottom} \tag{3.53}$$

This quantity is then multiplied by the total blackbody emissive power:

$$\Delta f_e \sigma T^4 = \int_0^{\lambda_{top}} E_{b\lambda}(T,\lambda)d\lambda - \int_0^{\lambda_{bottom}} E_{b\lambda}(T,\lambda)d\lambda \tag{3.54}$$

which can also be written as:

$$\Delta f_e \sigma T^4 = \int_{\lambda_{bottom}}^{\lambda_{top}} E_{b\lambda}(T,\lambda)d\lambda \tag{3.55}$$

This quantity now represents the power contained between two wavelengths for a noise source modeled as a black surface and at a given temperature. The two wavelengths, in this case, are the bottom and top cutoff wavelength for the notch filters. However, this does not give the

amount of radiation incident at the detector, but rather the total power flux emitted into the environment between two wavelengths.

To find the actual radiation that is incident upon the surface of the photo-diode detector, the total power flux from the noise source must be multiplied by a view factor. By modeling the noise source to detector combination as two parallel coaxial disks, a view factor may be found:

$$F_{nd} = (1/2)\left[\left[1+\left[\frac{1+\left(\frac{r_d}{d^2}\right)^2}{\left(\frac{r_n}{d}\right)^2}\right]\right]-\left[\left[1+\left[\frac{1+\left(\frac{r_d}{d^2}\right)^2}{\left(\frac{r_n}{d}\right)^2}\right]\right]^2-4\cdot\left(\frac{r_d}{r_n}\right)^2\right]^{\frac{1}{2}}\right] \tag{3.56}$$

It should be noted that $\mathbf{r_n}$ is the radius of the noise source, $\mathbf{r_d}$ is the radius of the photo-diode detector surface, and **d** is the distance from the noise source to the detector. Although the tungsten lamp filament is spherical in nature, only the frontal area of the filament is seen by the detector, so the coaxial disk approximation is justified.

Assuming a general form for the transfer function for the photo-diode as $G_d = G_d(\lambda)$ and for the filter as $G_f = G_f(\lambda)$, then the total power flux incident upon the photo-diode detector is (3.57):

$$P_{measured} = (1/2)\, G_d(\lambda)G_f(\lambda)\left[\left[1+\left[\frac{1+\left(\frac{r_d}{d^2}\right)^2}{\left(\frac{r_n}{d}\right)^2}\right]\right]-\left[\left[1+\left[\frac{1+\left(\frac{r_d}{d^2}\right)^2}{\left(\frac{r_n}{d}\right)^2}\right]\right]^2-4\cdot\left(\frac{r_d}{r_n}\right)^2\right]^{\frac{1}{2}}\right]\int_{\lambda_{bottom}}^{\lambda_{top}} E_{b\lambda}(T,\lambda)\,d\lambda$$

This quantity is actually a power flux. To get the actual power received by the detector, then the area of the detector must be multiplied by the flux. It should be noted that the measured power is very sensitive to the distance **d** from the noise source to the detector. However, since each detector is mounted very close to one another on the mount, the radiation must travel approximately the same distance to each detector. This approximation breaks down at extremely low distances in which the distances are not of the same order for each detector, as well as large angles of rotation of the detector mount. Also, at very long distances, the view factor is very tiny and the measured radiation is very small in magnitude, which gives a breakdown in this noise measurement system.

The motivation of this exacting measurement of noise is that it makes it possible to get an idea of the overall wavelength makeup of the noise in the viewable frame. If the noise characteristics follow that of a blackbody surface at a given temperature, then the noise is of the repeatable type, and the location of this noise source can be determined as well as a basic classification for the noise. Also, if the noise is repeatable in nature, the Perturbation/Correlation method works very well to remove this noise. If it doesn't follow the black body surface characteristics, then it is random, dark current noise, and can be effectively removed by just time averaging the frames at this given spatial location.

3.6 Explanation of Measurements Made by the Noise Detection System

Figures 1 and 2 in Appendix A are actual plots of the noise measured by this detection system. The measurements are of power flux incident on the detector, with the 0-255 representing a 0 to P_{MAX} of the calibration light source. Figure 3.60 then plots the noise, as a function of wavelength, for two different spatial locations; one with a noise source present and one without. This basically gives the spectral response curve for two different spatial locations; one that contains large amplitude, repeatable noise; and the other, low amplitude dark current or random noise. Figure 3.61 is a weighted spectral response for the viewable frame. More weight is given to the noise that falls in or near the band of the notch filter (825-835nm), and less weight is given to the noise well outside the notch filter range. A gaussian function is used in picking the weights, with the center of this corresponding to the center of the notch pass range. This is in effect is a "penalty schedule" for the viewable frame, and can be used in the design of search patterns for a particular bar code. The probability of detection, given the presence of a bar code, goes down with an increase in elevation of the penalty schedule, so the intuitive search would be the shortest path to connect all areas of equal and least elevation.

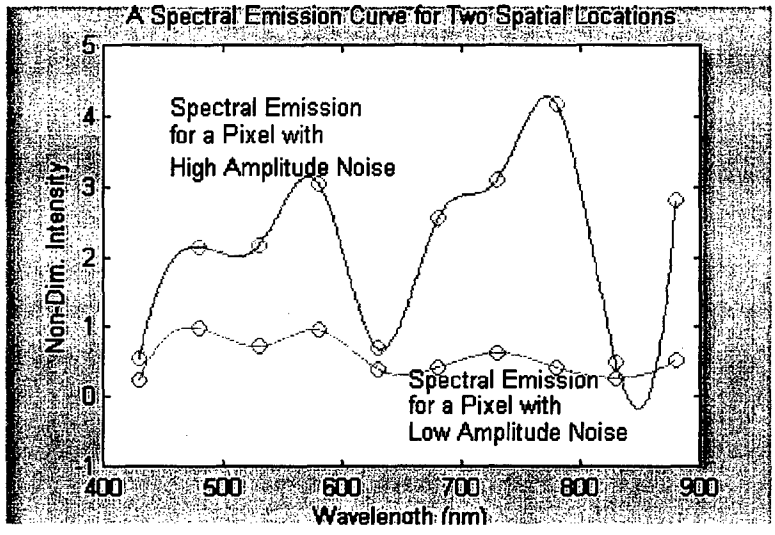


Figure 3.60: A Spectral Emission Curve for Two Spatial Locations

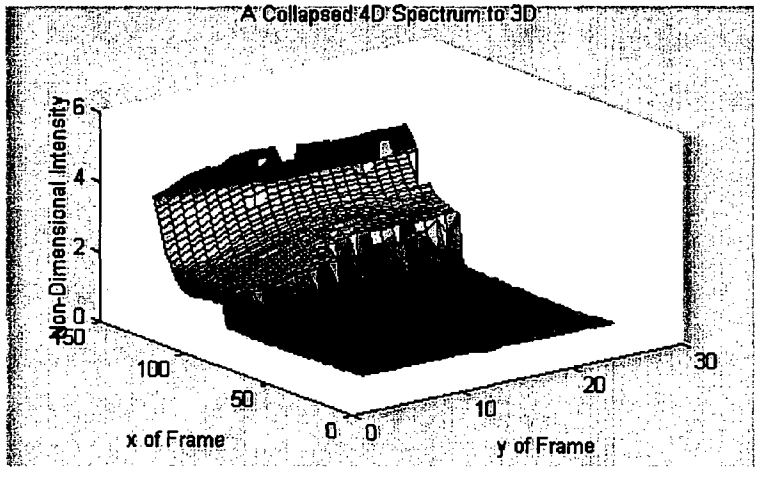


Figure 3.61: A "Penalty Schedule" for Scan Trajectory Planning

3.7 An Alternative Approach to Light Modulation

Up until now, the light modulation has been through motion of a pan and tilt, which, because of it's weight, is very slow. The light diffusion is accomplished through a galilean beam expander, but other methods of diffusion were tried. With the current setup, the more the laser is diffused, the less signal that will be received at the CCD. To measure the emission of the ink at varying radial distances, a motorized zoom lens is used, but there is definitely time associated with this process when changing fields of view. The frame rate of the frame grabber is the limiting factor to search velocity, and for these reasons, an alternate design was made for future implementation of this technology.

Figure 3 in Appendix A depicts a two axis galvo scan system. The laser diode beam is taken through a collimator and is reflected off the notch filter (this works since the filter blocks this wavelength). This light is then taken through the two axis's of the galvos, and when it returns, the image returns on the same axis of the outgoing beam. The part of the image that corresponds to ink emission is then passes through the notch filter. This is taken to a mirror-type beam splitter that breaks into two perpendicular beams of 50% power each. These beams are then taken to another set of beam splitters which splits them again. The overall beam (image) intensity is now 25% of the original beam intensity. At the output of each beam splitter a double convex and double concave lens are used to give a fixed magnification zoom, and another double convex lens is used to make a radiometer to focus this image energy on to an avalanche photo-diode. Four distinctly different magnification settings are used to a system that can, almost instantly, look at four different radial ranges. This removes the slow zoom and focus of the motorized zoom lens. The avalanche photo-diode was selected because it produces a current proportional to intensity but with a very large gain. The sample time for these modules can be up to 10^{12} Hertz, which is compared to the 30 Hz frame rate of the frame grabber.

4. Conclusion and Future Work

A system was presented that can be used to measure the fluorescent emission of a bar code printed in laser dye. The dye absorption was at 785 nm, and the peak ink emission was at 833 nm. The system architecture that was used in the prototype in the "read" of this bar code was also presented.

Some of the methods that were used to enhance the signal to noise ratio (S/N) were presented, with emphasis on the Perturbation/Correlation method. This method used a perturbation in laser intensity to give a corresponding change in ink emission magnitude, and these changes were correlated to separate the noise in the viewable frame from the signal.

With noise reduction methods outlined, research in the description of the noise was also presented. A system that could be used to determine the wavelength content of the noise in the viewable frame was presented. The calibration of this system was also discussed, with the basic architecture of the system described. Some motivation for this type of analysis was shown through black body model approximations to measure the relative intensity of the power flux at a given wavelength for a particular noise source. Examples of this noise measurement were presented, and some uses of this information were discussed.

For future work, a system was presented that would be one alternative to the prototype of the bar code reader that exists in the lab. It utilizes galvos to modulate the outgoing laser beam, as well as to bring in the incoming image. It also reduces the time required to sample the emission at different radial distances from the detector, which was found to be quite large with the motorized zoom lens. Finally, it uses avalanche photo-diodes in the emission detection since they have quite a large gain, which is useful for low-signal strength radiation measurement. In addition, they have a very high sample rate, much higher than the frame rate of the frame grabber used in the current system.

Appendix A

Figures Included with the Noise Measurement of Section 3.6 and the Alternate Light Modulation Design of Section 3.7

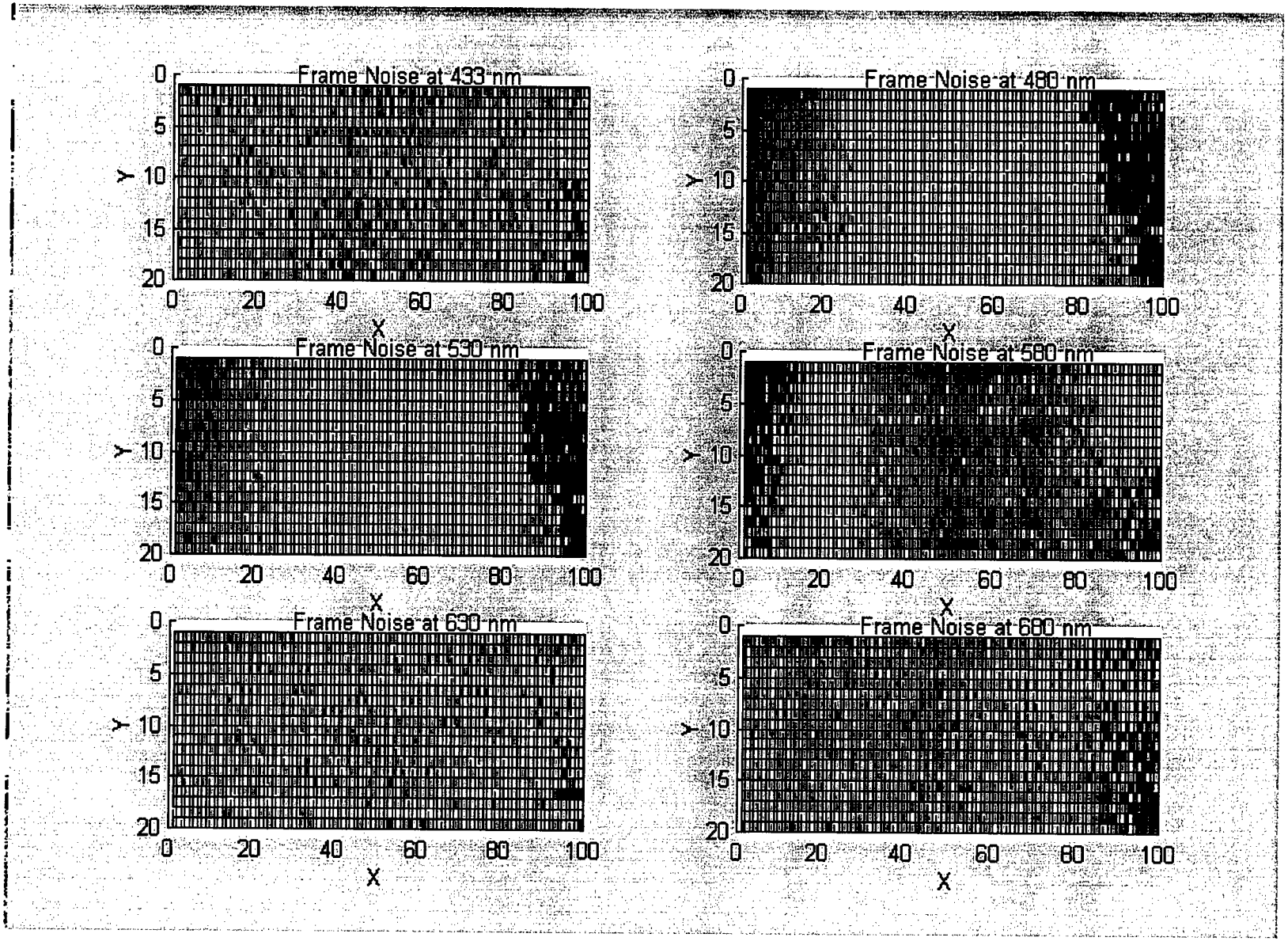


Figure 1: Frame Noise at 433, 480, 530, 580, 630, and 680 nm

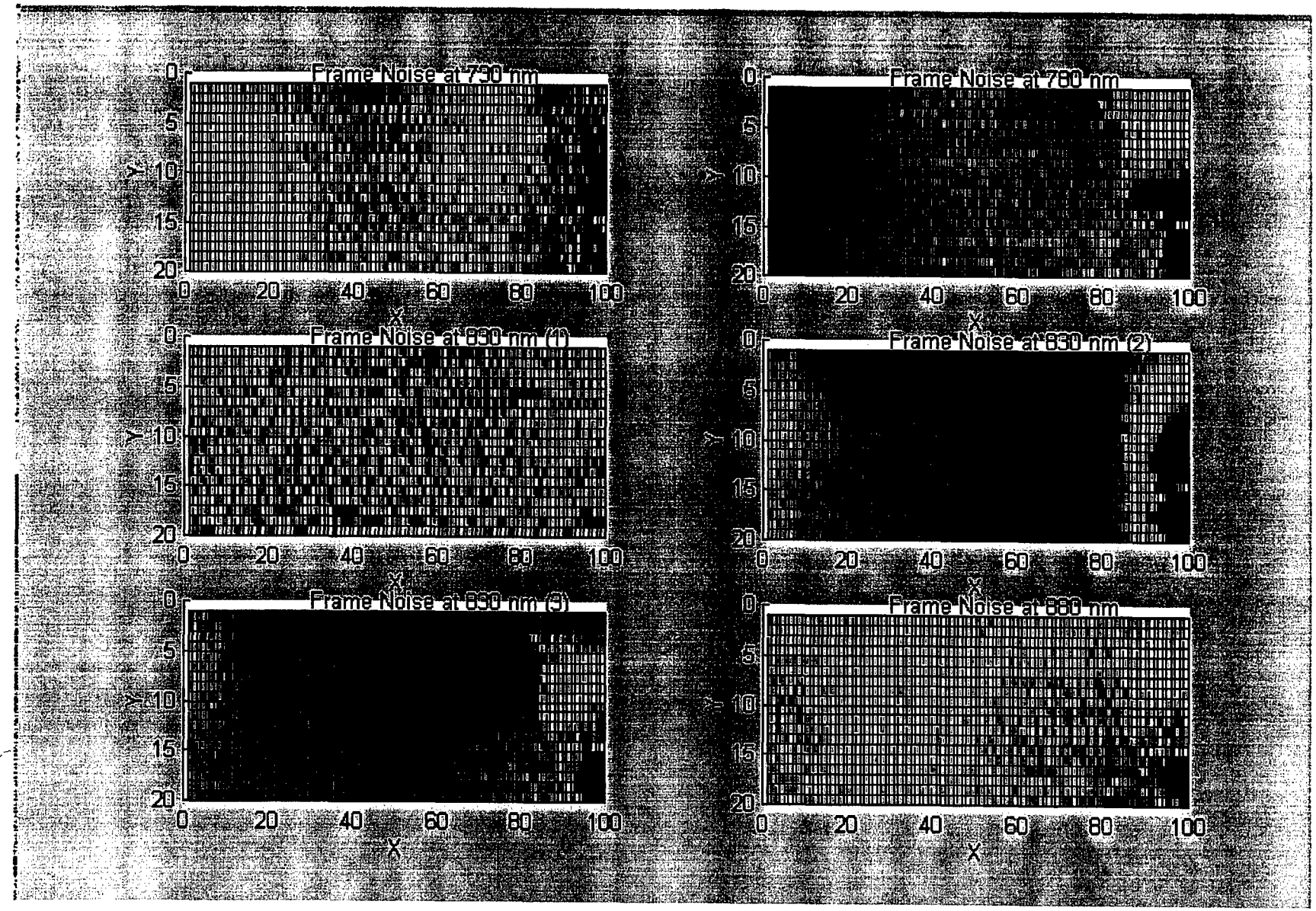


Figure 2: Frame Noise at 730, 780, 830, 830, 830, and 880 nm

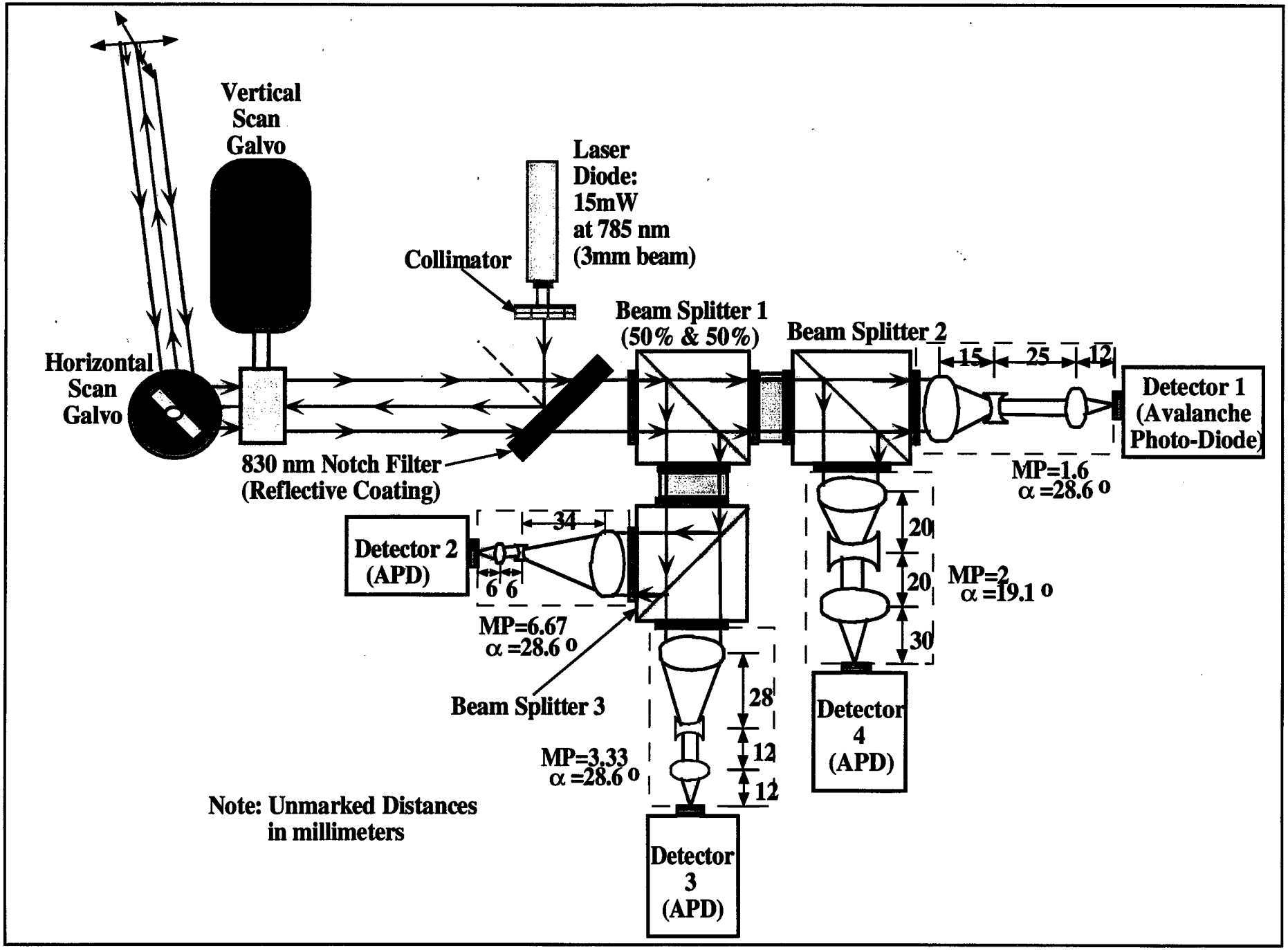


Figure 3: An Alternative Design for Light Modulation

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Physical Aids

CHAPTER 13

The Superchair: A Holonomic, Omnidirectional Wheelchair with a Variable Footprint Mechanism

H. Asada, M. Wada

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Progress Report, Total Home Automation and Health/Elderly Care Consortium
March 31, 1998

The Superchair : A Holonomic Omnidirectional Wheelchair with a Variable Footprint Mechanism

Haruhiko H, Asada
Principal Investigator

Masayoshi Wada
Visiting Scientist

Abstract

A reconfigurable mechanism for varying the footprint of a four-wheeled omnidirectional vehicle is developed and applied to wheelchairs. The variable footprint mechanism consists of a pair of beams intersecting at a pivotal point in the middle. Two pairs of ball wheels at the diagonal positions of the vehicle chassis are mounted, respectively, on the two beams intersecting in the middle. The angle between the two beams varies actively so that the ratio of the wheel base to the tread may change. Four independent servomotors driving the four ball wheels allow the vehicle to move in an arbitrary direction from an arbitrary configuration as well as to change the angle between the two beams and thereby change the footprint. The objective of controlling the beam angle is threefold. One is to augment static stability by varying the footprint so that the mass centroid of the vehicle may be kept within the footprint at all times. The second is to reduce the width of the vehicle when going through a narrow doorway. The third is to apparently change the gear ratio relating the vehicle speed to individual actuator speeds. First the concept of the varying footprint mechanism is described, and its kinematic behavior is analyzed, followed by the three control algorithms for varying the footprint. A prototype vehicle for an application of wheelchair platform is designed, built, and tested.

d'Arbeloff Laboratory for Information Systems and Technology
MIT

1. Introduction

A holonomic omnidirectional vehicle is a highly maneuverable vehicle that can move in an arbitrary direction from an arbitrary configuration. Unlike traditional nonholonomic vehicles, the holonomic vehicle can move in an arbitrary direction continuously without changing the direction of the wheels. It can move back and forth, slides sideways, and rotates at the same position. Therefore the holonomic vehicle would be useful for wheelchairs, which need to maneuver in crowded locations such as residential homes, hospitals and long-term care units as well as factories.

In the past decades, a variety of holonomic omnidirectional vehicles have been developed. The Swedish Wheel[1] is the first to accomplish omnidirectional motion without changing the direction of the wheels. The Swedish Wheel has been applied to a wheelchair[2] and other applications[3]. Pin and Killough developed a unique omnidirectional vehicle with powered wheel units consisting of a pair of round wheels that alternately touch the floor[4]. The Omni-Track with ball wheels arranged in a crawler mechanism allows for sideway motion with large traction forces[5]. The VUTON omnidirectional vehicle consisting of arrays of cylindrical tires combined with an unique crawler mechanism is capable of carrying a large payload[6]. The Ball Wheel omnidirectional vehicle developed by the authors' group uses spherical tires held by a novel ring roller mechanism that transmits an actuator torque to the ball wheel[7]. This Ball Wheel Vehicle exhibits smooth motion with no shimmy and jerk along with highly maneuverable and precise movements, all of which are desirable features for wheelchair applications.

To apply the Ball Wheel to a wheelchair, however, the vehicle must meet several requirements for complex indoor applications. First, the vehicle body must be compact enough to go through narrow doorways. Standard doors are limited in width; the vehicle's tread and chassis width must confirm to the dimensional constraints. A narrow tread, however, may incur instability of the vehicle. As the patient moves, the mass centroid of the vehicle may shift in a wide range. Moreover, infirm patients cannot sit up in the middle of the chair, but tend to lean towards the arm rests. The footprint of the wheelchair[1] must be wide enough to prevent the patient from falling on the floor. A large footprint is therefore desirable for stability and safety, while wheelchairs must conform to dimensional constraints. Also the footprint must be compact since a large footprint does not allow the vehicle to maneuver in a closely confined place. Stability and maneuverability are therefore conflicting requirements.

[1] In this paper, footprint refers to the area enclosed by the contact points of the vehicle wheels.

Traditional vehicle designs with fixed footprint configurations would not provide an efficient solutions to this stability-maneuverability trade-off problem.

In addition, the original Ball Wheel Vehicle has three wheels to achieve 3 DOF motion. Its footprint is a triangular area, which is inadequate for maintaining stability. A four-wheeled vehicle is desirable, but incurs an over constraint problem between the active wheels and the ground since the vehicle has only three DOF while the four motors drive the four wheels independently. The over constraint problem may result in slip at the wheels or generate unwanted internal forces within the vehicle chassis.

In this paper, a novel reconfigurable footprint mechanism will be developed to augment stability and enhance maneuverability as well as to resolve the over-constraint problem. This new mechanism would allow to vary the ratio of wheel base to tread so that the vehicle could go through a narrow doorway and that the mass centroid could be kept within the footprint at all times. Furthermore, this varying footprint mechanism would function as a kind of continuously varying transmission (CVT) that changes the gear ratio between the actuator speed and the resultant vehicle speed. Therefore, the vehicle would be able to meet diverse requirements for speed and torque, exhibiting enhanced maneuverability and efficiency. In the following sections, the new mechanism will be described together with the original ball wheel mechanism. Its kinematic and static behavior will be analyzed, and algorithms for stability augmentation and transmission control will be developed. Experiments by using a prototype vehicle will be presented at the end to demonstrate the feasibility and validity of the proposed method.

2. Mechanical Design

2.1 The Ball Wheel Mechanism[7]

The Ball wheel mechanism with a special roller ring, shown in Figure 1-(a), is used for the vehicle. The ball is held by roller ring *A* at a great circle together with a set of bearings *B* arranged on another great circle. The roller ring is rotated by a servo motor to drive the ball wheel. Since the ring is inclined, a traction force is created between the ball wheel and the floor, as shown in the plane view, Figure1-(b). The stationary bearings *B*, arranged on the second great circle, allow passive rotation of the ball about an arbitrary axis within that great circle. As a result, the ball is free to move in the

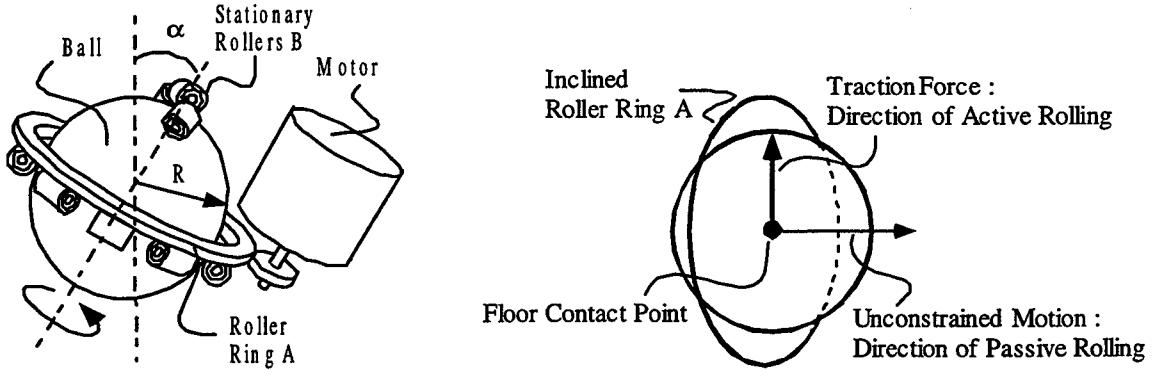
direction perpendicular to the traction force, as shown in Figure 1-(b). The vehicle must have at least three ball wheels, each generating a traction force in a different direction. The resultant force acting on the vehicle is given by the vectorial sum of the traction forces. Varying the combination of the traction forces creates an arbitrary force and moment driving the vehicle.

The vehicle consisting of these ball wheels can move in an arbitrary direction with an arbitrary linear velocity and rotational velocity at an arbitrary position and orientation. There is no singular point in this mechanism, hence it is omnidirectional and holonomic. Moreover, this ball wheel vehicle allows for smooth motion with no shimmy and jerk, all of which are desirable for wheelchairs transporting patients.

2.2 A Reconfigurable Footprint Mechanism

Figure 2 shows the schematic of a new reconfigurable footprint mechanism for a four-wheeled holonomic omnidirectional vehicle. All the wheels are the ball wheels described above with independent suspensions. Two pairs of the ball wheels at diagonal positions are fixed to the tips of two beams intersecting at a pivotal joint in the middle. The two beams rotate about this pivotal joint so that the ratio of wheelbase to tread can vary. To go through a narrow doorway, the tread becomes narrow while the wheelbase becomes long, as shown in Figure 3-(a). To increase sideway stability, the tread is expanded, as shown in Figure 3-(c). To be isotropic, the two beams intersect at the right angle, as shown in Figure 3-(b).

One design issue with this reconfigurable footprint mechanism is that the chair mounted on the vehicle must be kept aligned with the bisector of the two beams intersecting at the pivotal joint, although both beams rotate about the joint. To this end, a differential gear mechanism is used for the pivotal joint. As shown in Figure 4, the three bevel gears form a differential gear mechanism. The middle bevel gear, Gear 3, rotates freely about the horizontal shaft β that is fixed to the vertical shaft α. Bevel gear 1 is fixed to beam A, while bevel gear 2 to beam B. When beam A rotates about the vertical shaft α together with bevel gear 1, bevel gear 3 rotates. As a result, bevel gear 2 rotates the same amount but in the opposite direction to bevel gear 1. In consequence, the chair mounted on the vertical shaft α is kept at the bisector position of the intersecting beams, A and B. The angle between the two beams is measured by a potentiometer, as shown in the figure.



(a) Ball Wheel Unit (b) Plane View seen from the top of the Ball Wheel

Figure 1: Ball Wheel Mechanism

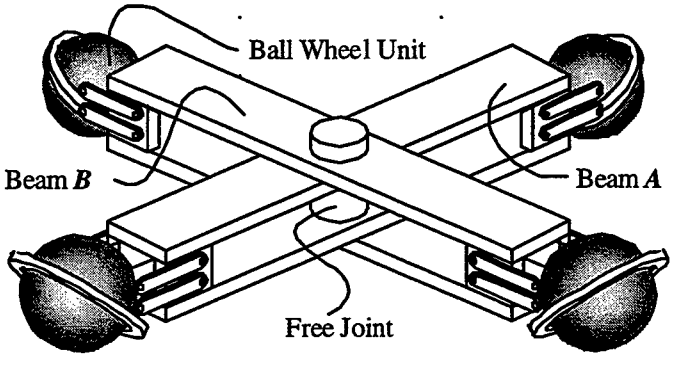
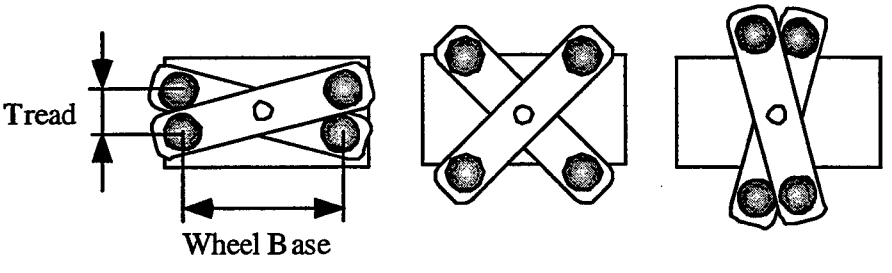


Figure 2: Omnidirectional Reconfigurable Vehicle



(a)Narrow Configuration (b)Isotropic Configuration (c)Wide Configuration

Figure 3: Reconfiguration of the Footprint of the Vehicle

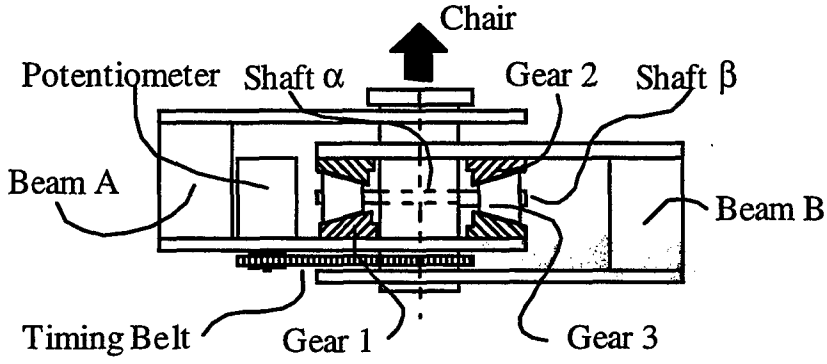


Figure 4: Pivotal Joint of the Vehicle

3. Kinematic Analysis

3.1 Ball Wheel

Consider the *i*-th ball wheel and half the beam holding the ball wheel, as shown in Figure 5. As ball rolls on the floor, i.e. the X-Y plane, the contact point with the X-Y plane moves together with the beam. The time rate of change of the contact point is called ball velocities, v_{xi} and v_{yi}, with reference to the fixed frame O-XY. The pivotal joint of the vehicle, denoted O_v in the figure, moves at v_{xv} and v_{yv} and the angular velocity of the i-th half beam is denoted ϕ_i. The ball velocities, v_{xi} and v_{yi}, are given by

$$\begin{bmatrix} v_{xi} \\ v_{yi} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -L\sin\phi_i \\ 0 & 1 & L\cos\phi_i \end{bmatrix} \begin{bmatrix} v_{xv} \\ v_{yv} \\ \dot{\phi}_i \end{bmatrix} \qquad (1)$$

where L is the distance between the pivotal joint O_v and the contact point of the ball wheel. The ball wheel rolls in one direction, and is free to roll in the direction perpendicular to the active direction, as mentioned in section 2.1. Let ψ be the angle pointing in the direction of active rolling on the O-XY plane, as shown in Figure 5. Note that ψ is measured relative to the beam to which the ball wheel unit is fixed. The ball velocities, v_{xi} and v_{yi}, can be decomposed into the velocity in the active direction, v_{ai}, and the one in the passive direction. Using eq.(1), the active velocity component v_{ai} is given by

$$v_{ai} = v_{xi}\cos(\phi_i + \psi) + v_{yi}\sin(\phi_i + \psi)$$
$$= v_{xv}\cos(\phi_i + \psi) + v_{yv}\sin(\phi_i + \psi) + \dot{\phi}_i L\sin\psi \qquad (2)$$

On the other hand, the i-th active velocity component v_{ai} is a function of the angular velocity of the i-th actuator, ω_i, since the ball is driven by the actuator in that active direction. As shown in Figure 1, let R be the radius of the spherical tire and α the angle between the vertical line and the direction of the inclined roller ring. The active velocity v_{ai} is given by

$$v_{ai} = \rho R\sin\alpha \cdot \omega_i \qquad (3)$$

where ρ is the gear reduction ratio associated with the roller ring and the gear of the motor.

Figure 6 shows the whole vehicle with four ball wheels. Frame O_v-X_vY_v is attached to the pivotal joint, where the X_v axis is the bisector of the angle between the two beams, 2ϕ. Let ϕ_v be the angle of the X_v axis measured from the X axis. The direction of each half beam is given by

$$\begin{aligned} \phi_1 &= \phi_v + \phi, & \phi_2 &= \phi_v - \phi + \pi \\ \phi_3 &= \phi_v + \phi + \pi, & \phi_4 &= \phi_v - \phi \end{aligned} \qquad (4)$$

Our objective is to obtain the relationship between the active ball wheel movements driven by individual actuators and the resultant vehicle motion. To describe the entire vehicle motion including the variable footprint mechanism, four generalized velocities are needed; two translational velocities of the pivotal joint, v_{xv}, and v_{yv}, angular velocity of the vehicle chassis, $\dot{\phi}_v$, and the time rate of change of the angle between the two beams, $\dot{\phi}$. Substituting eq.(4) into (2) and rotating the coordinate system to the one parallel to the vehicle coordinate system,

$$\begin{bmatrix} v_{a1} \\ v_{a2} \\ v_{a3} \\ v_{a4} \end{bmatrix} = \begin{bmatrix} \cos(\phi+\psi) & \sin(\phi+\psi) & L\sin\psi & L\sin\psi \\ \cos(\phi+\psi) & -\sin(\phi+\psi) & L\sin\psi & -L\sin\psi \\ -\cos(\phi+\psi) & -\sin(\phi+\psi) & L\sin\psi & L\sin\psi \\ -\cos(\phi+\psi) & \sin(\phi+\psi) & L\sin\psi & -L\sin\psi \end{bmatrix} \begin{bmatrix} v_{xv} \\ v_{yv} \\ \dot{\phi}_v \\ \dot{\phi} \end{bmatrix} \qquad (5)$$

For the prototype vehicle to be described later in section 7, the angle of active rolling direction, ψ, is 90 degrees, and the above relationship reduces to:

$$\begin{bmatrix} v_{a1} \\ v_{a2} \\ v_{a3} \\ v_{a4} \end{bmatrix} = \begin{bmatrix} -\sin\phi & \cos\phi & L & L \\ -\sin\phi & -\cos\phi & L & -L \\ \sin\phi & -\cos\phi & L & L \\ \sin\phi & \cos\phi & L & -L \end{bmatrix} \begin{bmatrix} v_{xv} \\ v_{yv} \\ \dot{\phi}_v \\ \dot{\phi} \end{bmatrix} \tag{6}$$

The above 4 by 4 matrix in eq.(5) is invertible for all the vehicle configuration, as long as $\cos\phi \sin\phi \neq 0$.

$$\mathbf{V}_v = \mathbf{J}\mathbf{V}_a \tag{7}$$

where

$$\mathbf{V}_v = [v_{xv} \quad v_{yv} \quad \dot{\phi}_v \quad \dot{\phi}]^T$$

$$\mathbf{V}_a = [v_{a1} \quad v_{a2} \quad v_{a3} \quad v_{a4}]^T$$

$$\mathbf{J} = \begin{bmatrix} \frac{1}{4\cos(\phi+\psi)} & \frac{1}{4\cos(\phi+\psi)} & \frac{-1}{4\cos(\phi+\psi)} & \frac{-1}{4\cos(\phi+\psi)} \\ \frac{1}{4\sin(\phi+\psi)} & \frac{-1}{4\sin(\phi+\psi)} & \frac{-1}{4\sin(\phi+\psi)} & \frac{1}{4\sin(\phi+\psi)} \\ \frac{1}{4L\sin\psi} & \frac{1}{4L\sin\psi} & \frac{1}{4L\sin\psi} & \frac{1}{4L\sin\psi} \\ \frac{1}{4L\sin\psi} & \frac{-1}{4L\sin\psi} & \frac{1}{4L\sin\psi} & \frac{-1}{4L\sin\psi} \end{bmatrix} \tag{8}$$

Note that matrix **J** is the Jacobian relating the vehicle velocity vector to the ball velocities in the active directions. The above analysis shows that the four independent actuators driving the four ball wheels completely determine the vehicle velocity as well as the angular velocity of the footprint reconfiguration mechanism. Note that there is no singular point in the Jacobian and that no over constraint situation occurs in this mechanism.

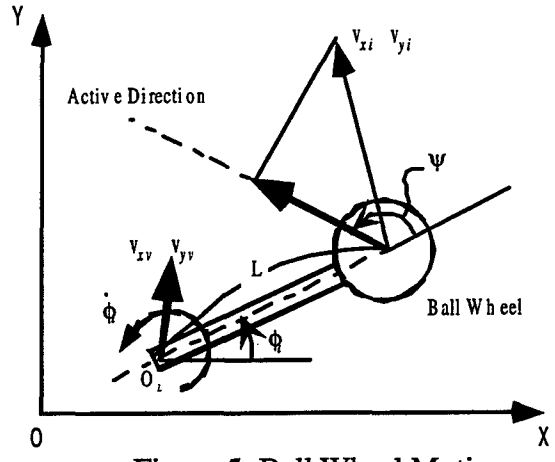


Figure 5: Ball Wheel Motion

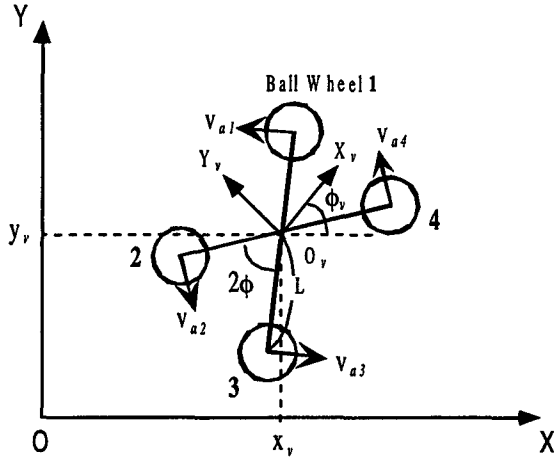


Figure 6: Coordinate System of the Vehicle

4. Static Stability Augmentation

Static stability is among the most critical requirements for wheelchairs. In this paper, the varying footprint mechanism is used for augmenting the vehicle stability. The objective of static stability augmentation is to keep the position of the mass centroid within the footprint of the platform by varying the joint angle between the two beams. A method for estimating the centroid position and obtaining an optimal joint angle will be presented in this section. Let m be the total inertial load, i.e. the mass of the chair, patient, and vehicle excluding the ball wheels. Let (x_c, y_c) be the coordinates of the mass centroid with respect to the vehicle coordinate frame, as shown in Figure 7. Each ball wheel is equipped with a load cell to monitor the load distribution. Let F_i be the vertical force acting on the i-th ball, then the mass centroid position is given by

$$x_c = \frac{L}{mg} F_x \cos\phi \ , \quad y_c = \frac{L}{mg} F_y \sin\phi \tag{9}$$

where

$$\begin{aligned} F_x &= F_1 - F_2 - F_3 + F_4 \\ F_y &= F_1 + F_2 - F_3 - F_4 \end{aligned} \tag{10}$$

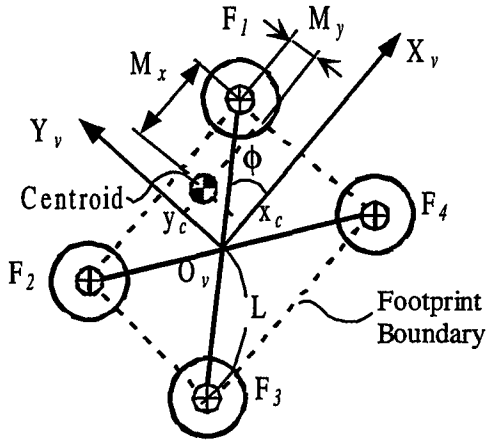


Figure 7: Static Stability Margin

When the mass centroid is on the boundary of the vehicle footprint shown by the broken lines in the figure, the vehicle is critically stable. Static stability margin is therefore defined to be the minimum distance from the mass centroid position to the footprint boundary. Since the footprint is a rectangular area parallel to the boundary of which is the X_v and Y_v axes, the stability margin can be determined by evaluating the distances to the four sides of the rectangle. Let M_x and M_y be the distances to the boundary in the x and y directions, respectively. As shown in the figure, static stability margin M is given by

$$M = Min(M_x, M_y) \tag{11}$$

where

$$M_x = L\cos\phi - |x_c| \ , \ M_y = L\sin\phi - |y_c| \tag{12}$$

The optimal footprint configuration is then given by the pivotal joint angle that maximizes the static stability margin given above:

$$\phi^0 = \arg \underset{0<\phi<\frac{\pi}{2}}{Max}(Min(M_x, M_y)) \tag{13}$$

This is a type of max-min strategy, which best augments the stability in the worst direction. Figure 8 shows the plot of M_x and M_y against ϕ. The optimal joint angle ϕ^0 is provided at the intersection of the two curves, M_x and M_y. Equating M_x and M_y yields

$$\phi^0 = \tan^{-1} \frac{mg - |F_x|}{mg - |F_y|} \tag{14}$$

The centroid is located in an area where $x_c > 0$, $y_c > 0$, that is the first quadrant of the vehicle coordinate frame, then $F_x > 0$, $F_y > 0$. Therefore the optimal joint angle in the first quadrant is given by

$$\phi^0 = \tan^{-1} \frac{F_2 + F_3}{F_3 + F_4} \tag{15}$$

Optimal angles in other quadrants can be obtained in the same manner. In summary, the optimal angle ϕ^o is given by the following form,

$$\phi^0 = \tan^{-1} \frac{Min\{(F_1 + F_4),(F_2 + F_3)\}}{Min\{(F_1 + F_2),(F_3 + F_4)\}} \tag{16}$$

Note that this method does not need the vehicle weight, patient weight and the absolute value of each ball wheel load, but simply needs the ratio of the wheel load distribution.

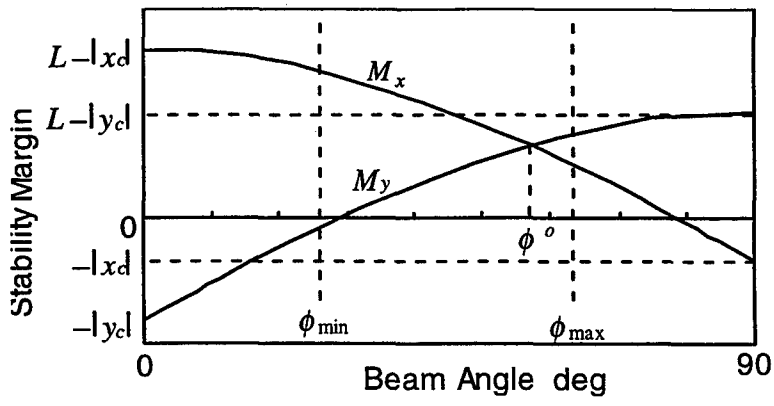


Figure 8: Plot of the Static Stability Margin against the Pivotal Joint Angle

5. Transmission Control

Since the Jacobian given by eq.(7) is a function of pivot angle ϕ, the vehicle velocity varies depending on the footprint configuration, although the individual actuator speeds remain the same. This implies that the varying footprint mechanism would change the kind of transmission ratio between the actuators and the vehicle. In other words, the transmission of the vehicle drive train can be changed from a low gear to a top gear by changing the footprint configuration. Depending on diverse requirements for vehicle speed and traction force, one can change the transmission ratio simply by changing the pivot angle ϕ. In this section, we will analyze this varying transmission, and discuss its utility.

Suppose that the vehicle is commanded to move forward, i.e. the direction of the X axis. Substituting $v_{xv}=V$, $v_{yv}=0$, $\dot{\phi}_v=0$ and $\dot{\phi}=0$ into eq.(6) yields the velocities of the individual ball wheels in the active rolling direction; $v_{ai}=-V\sin\phi$, $v_{a2}=-V\sin\phi$, $v_{a3}=V\sin\phi$ and $v_{a4}=V\sin\phi$. Figure 9 shows these ball wheel velocities and the relationship with the pivot angle ϕ. Note that the actual ball wheel motion is the vectorial sum of the active rolling driven by the actuator and the passive rolling in the perpendicular direction. As pivot angle ϕ decreases, the passive rolling vector becomes longer whereas the active rolling decreases. Therefore the ratio of the vehicle velocity to the active rolling part of the ball wheel velocity increases. This is why the reconfigurable footprint mechanism serves as a variable transmission.

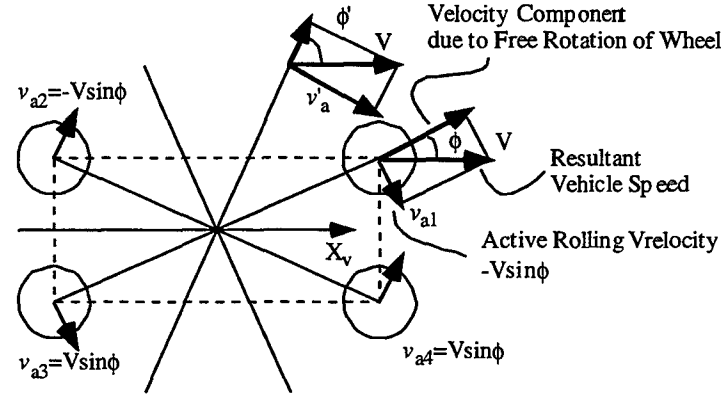


Figure 9 : Kinematics of Variable Transmission

The above argument on the vehicle transmission ratio in one direction can be extended to that of two-dimensional motion.

Consider the translational part of vehicle motion alone. Translational vehicle velocities $\mathbf{v}_t=[v_{xv}, v_{yv}]^T$ are related to the ball wheel velocity vector $\mathbf{v}_a$ by

$$\mathbf{V}_t = \mathbf{J}_t \mathbf{V}_a \tag{17}$$

where $\mathbf{J}_t$ consists of the first two rows of the Jacobian $\mathbf{J}$.

$$\mathbf{J}_t = \frac{1}{4}\begin{bmatrix} \frac{-1}{\sin\phi} & \frac{-1}{\sin\phi} & \frac{1}{\sin\phi} & \frac{1}{\sin\phi} \\ \frac{1}{\cos\phi} & \frac{-1}{\cos\phi} & \frac{-1}{\cos\phi} & \frac{1}{\cos\phi} \end{bmatrix} \tag{18}$$

The transmission ratio of the vehicle drive system is defined as

$$\lambda = \frac{|\mathbf{V}_t|}{|\mathbf{V}_a|} \tag{19}$$

where $|\mathbf{x}|$ represents the norm of vector $\mathbf{x}$. Note that, since the vehicle is a multi degree-of-freedom system, the standard scalar quotient, i.e. v_t/v_a, cannot be used. Therefore, the quotient of the vector norms is used in eq.(19). The rotational transmission ratio can be defined in a form similar to eq.(19). Note, however, that the rotational transmission does not vary depending on the footprint configuration, since the third and fourth rows of Jacobian $\mathbf{J}$ are not functions of pivot angle ϕ.

The transmission ratio varies depending on the direction of the vehicle motion. The maximum and minimum of λ and their directions of motion are obtained from the singular value decomposition of Jacobian $\mathbf{J}_t$.

$$\mathbf{J}_t = [\mathbf{u}_1 \quad \mathbf{u}_2]^T \begin{bmatrix} \frac{1}{2\sin\phi} & 0 & 0 & 0 \\ 0 & \frac{1}{2\cos\phi} & 0 & 0 \end{bmatrix} [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \mathbf{v}_3 \quad \mathbf{v}_4] \tag{20}$$

where $1/2\sin\phi$ and $1/2\cos\phi$ are singular values of matrix $\mathbf{J}_t$, and $\mathbf{u}_i$ and $\mathbf{v}_i$ are left and right eigenvectors, respectively. The two singular values provide the maximum and minimum transmission ratios. Namely, for $0< \phi \leq \pi/4$, the transmission ratio takes the maximum, $\lambda_{max}=1/2\sin\phi$, when the vehicle moves in the direction along the corresponding left eigenvector $\mathbf{u}_1^T=[1, 0]$, i.e. the X axis, with the distribution of actuator speeds given by the right eigenvector $v_1^T=[-0.5,-0.5,0.5,0.5]$. The minimum transmission ratio, $\lambda_{min}=1/2\cos\phi$, takes place when the vehicle moves in $\mathbf{u}_2^T=[0, 1]$, i.e. the Y-axis, with the actuator speed distribution of $v_2^T=[0.5,-0.5,-0.5,0.5]$. When the actuator speed distribution is v_3 or v_4, no translational velocity is generated.

Figure 10 shows the directions of the maximum and minimum transmission ratios, and Figure 11 shows the plot of the max/min transmission ratios against the pivot angle, ϕ. Note that the transmission ratio varies continuously as pivot angle ϕ varies. Therefore, the variable footprint mechanism can be treated as a continuously variable transmission (CVT).

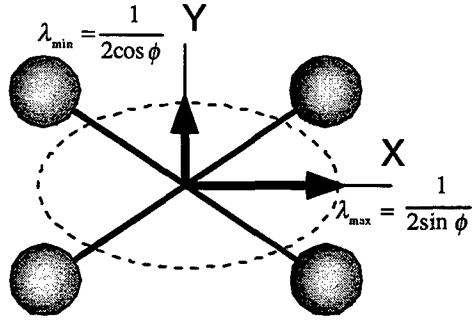


Figure 10: Maximum and Minimum Transmission Ratios

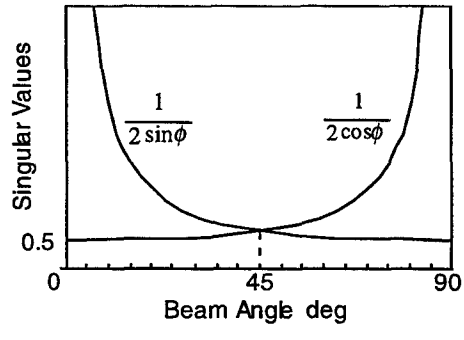


Figure 11: Plots of Singular Values against Beam Angle

Note that at $\phi = \pi/4$, the vehicle has an isotropic transmission ratio of $\sqrt{2}/4$ in all directions. Note also that, as the transmission ratio becomes larger, the traction force generated becomes smaller, hence the acceleration of the vehicle becomes smaller. The mechanical advantage is given by the reciprocal of the transmission ratio:

$$MA = \frac{1}{\lambda} = \frac{|\mathbf{F}_t|}{|\mathbf{F}_a|} \tag{21}$$

where $\mathbf{F}_a$ is a 4×1 vector comprising the traction forces at the four ball wheels, and $\mathbf{F}_v$ is the resultant force acting on the vehicle. From the above analysis it follows that

- Pivot angle ϕ must be small in order to move at a high speed in the X-direction. For traveling a long distance at a high speed, the footprint should be long in the longitudinal direction.
- For rapid acceleration, the footprint should be shortened in that direction
- Isotropic speed and traction characteristics can be achieved when $\phi = \pi/4$

Three strategies can be used for determining the footprint configuration along with the stability augmentation scheme.

6. Subsumption Control Architecture

As mentioned previously, several kinds of functionality are needed for a wheelchair. Different control modes and control objectives must be selected depending not only on an operator's commands but also on the situations the vehicle is involved. In order to coordinate diverse control modes and objectives, a subsumption control architecture[7] is used for the system.

To implement the subsumption architecture, the vehicle behavior must be decomposed into four tasks (A-D).

Task *A* is the most fundamental task in which the vehicle moves in any direction and/or rotates about any point as the operator requests while the vehicle stability is maintained. When the operator requests

the vehicle to move by using a joystick, the vehicle would move in a given direction at a given velocity. At the same time the footprint configuration would be automatically controlled to maximize the stability margin based on the load distribution among the wheels. The vehicle motion control and the footprint configuration control can be achieved simultaneously and independently by driving four wheels.

Task *B* achieves the efficient power driving for long distance traveling. When the operator commands the vehicle to move at high speed, the footprint configuration would be varied to change the transmission ratio of the wheel and the vehicle. The transmission ratio could be decided to maintain the maximum margin of the traction force. During the task execution, maximum velocity of the translational sideways and rotational motion are restricted to avoid the vehicle to be unstable in sideways.

In task *C*, the vehicle achieves the control mode for going through narrow doorways or maneuvering in crowded areas. The vehicle has to minimize its width for going through doorways and minimize the diameter of the vehicle footprint for maneuvering in crowded areas. All vehicle motions should be restricted in slow speed. The task *C* would be triggered and reset by the operator's command sent via buttons on the joystick.

In Task *D*, it is assumed that the vehicle may not fall down at all times. The static stability margin would be monitored and if the margin hits the minimum margin, the all vehicle motions would be stopped and the footprint configuration would be varied to maintain the limit stability margin. At the same time, a warning signal would be sent to the operator in order to let the patient aware of the risk.

These tasks are assigned to four layers(zero-th to 3rd) of the subsumption architecture, respectively. The layer of large number has higher priority than that of smaller number, i.e., task A occupying the zero-th layer has the lowest priority, and task D occupying the 3rd layer has highest priority. Details of each task are described as follows.

- Task *A* (Zero-th Layer) : Vehicle Traveling with Static Stability Augmentation
  - *A* -0 : Detect the changing of position of mass centroid.
  - *A* -1 : Change the footprint shape so as to optimize the stability margin.
  - *A* -2 : Move vehicle if the operator request the vehicle to move.

- Task *B* (1st Layer) : Getting Efficient Power in a Higher Speed of the Vehicle

B -0 : Detect the vehicle velocity exceeding the certain value.

B -1 : Change the footprint shape to change the transmission ratio depending on the velocity of the vehicle.

B -2 : Restrict the lateral velocity and rotational speed of the vehicle.

B -3 : Monitor the velocity of the vehicle to be reduced below certain speed, then reset.

- Task *C* (2nd Layer) : Improvement of Maneuverability [Going through doorways or moving around crowded areas]

C -0 : Detect the command from a switch.

C -1 : Change the footprint shape to minimize the width or rotational radius of the vehicle.

C -2 : Slowdown the vehicle speed.

C -3 : Detect the reset command from the switch, then reset.

- Task *D* (3rd Layer) : Prevention of Falling Down

D-0 : Detect the stability margin hitting the minimum limit.

D-1 : Change the footprint shape to prevent the falling.

D-2 : Stop the vehicle.

D-3 : Give a caution to the operator by beeps.

D-4 : Check the stability margin to be recovered, then reset.

The system has hierarchical configuration and would be able to adapt to multiple situations or requirements by changing the layer taking the control. Figure 11 illustrates a schematic of the vehicle control system.

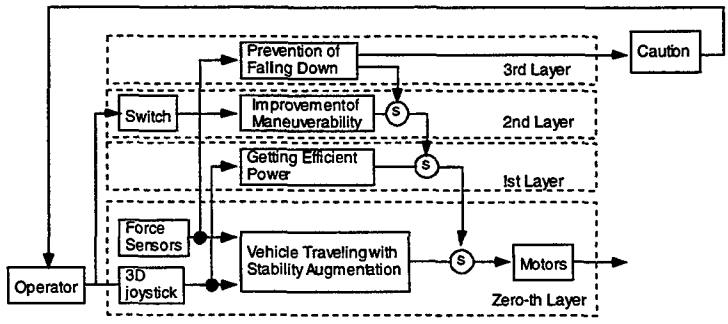


Figure12: Vehicle Control System

7. Prototyping

A prototype wheelchair with the reconfigurable footprint mechanism has been designed and built. Figure 13 shows the overview of the prototype. All the vehicle drive components, including actuators, ball wheels, and cross beams, are placed beneath the rectangular platform of 510mm wide and 610mm long. The platform is only 190mm above the floor. A commercially available chair with an aluminium and grass-fiber structure is mounted on the platform. A three degree-of-freedom joystick is attached to one of the arm rests.

Figure 14 shows the bottom view of the prototype wheelchair. The diagonal distance of the footprint, that is, the distance between the floor contact points of the two balls at diagonal positions is 700mm. The joint angle between the two beams varies from 55deg.(ϕ_{min} = 55/2deg.) to 125deg. (ϕ_{max} = 125/2deg.). This means that the wheel base and tread of the vehicle varies between 323mm and 620mm. The absolute angle of the pivotal joint is measured by a potentiometer.

Figure 15 shows a ball wheel unit with an independent suspension mechanism. The ball, 108mm in diameter, is a stainless steel sphere with 3mm thick outer coating of rubber. The oblique roller ring driven by a DC servo motor is arranged in such a way that a traction force is generated in the direction perpendicular to the longitudinal direction of the beam, namely, ψ = 90deg. The suspension mechanism allows the ball wheel to move about 25mm in the vertical direction. A parallelogram mechanism is used for the suspension so that the ball wheel unit may keep the same orientation relative to the vehicle chassis. To minimize the height of the wheel mechanism, coil springs of the suspension are placed horizontally within the parallelogram mechanism. The spring force is transmitted to the top of the ball through a transmission linkage. The vertical load acting on each wheel can be detected by measuring the displacement of the coil spring with a linear potentiometer attached to the side of the spring. Furthermore, an incremental encoder is mounted on each servo motor to measure the ball rotation in its active direction.

Figure 15 shows a differential gear mechanism used for the pivotal joint in order to keep the chair orientation aligned with the bisector of the two beams.

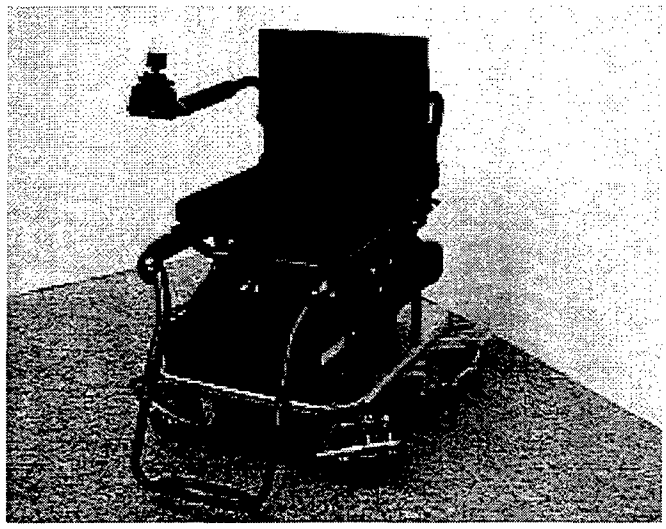


Figure 13: Wheelchair Prototype

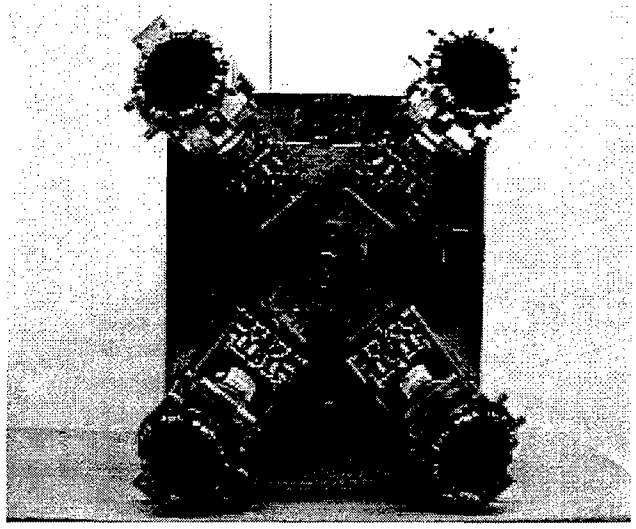


Figure 14: Reconfigurable Vehicle (Bottom View)

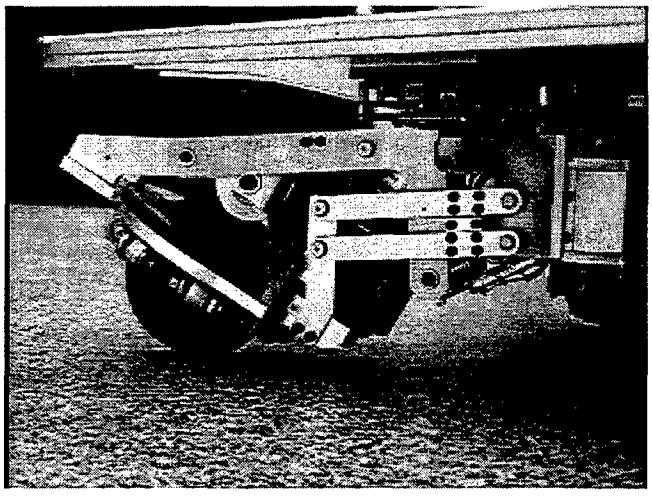


Figure 15: Ball Wheel Unit

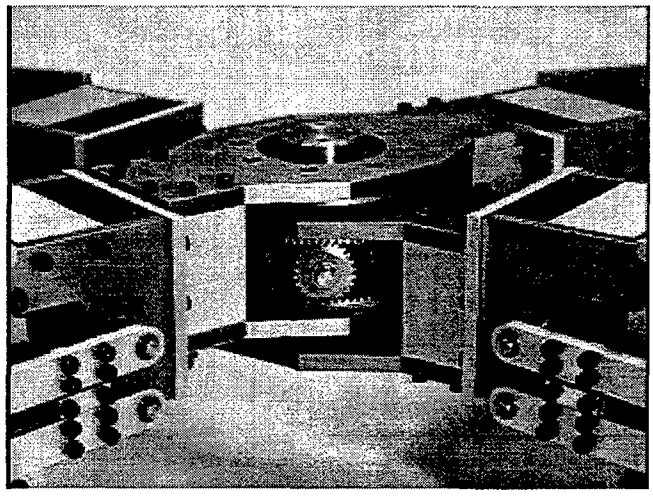


Figure 16: Pivotal Joint with Differential Gears

8. Experiments

8.1 Kinematics

First the kinematic relationship described by the Jacobian was verified through experiments. The four motors were commanded to move at constant speeds, and the resultant vehicle speed was measured. The experiment was repeated for different footprint configurations. Figure 17 shows one of the experimental results, where each ball wheel speed is kept at 7.5rpm, 15rpm or 22.5 rpm. The

resultant vehicle speed in the x direction varied depending on pivot angle ϕ. Overall the experimental results agree with the theoretical curves derived from the Jacobian given by eq.(6). Other experiments of vehicle motion in oblique directions and rotational ones showed good agreements with the theoretical Jacobian as well.

8.2 Static Stability Augmentation

The stability augmentation algorithm described in section 4 provides the optimal pivot angle that maximizes the stability margin based on the measured distribution of load over the four wheels. As the load shifts to one side, the estimated position of the centroid would move accordingly, and the pivot angle would be changed so as to keep the maximum stability margin. Figure 18 shows the experiments that demonstrates this stability augmentation behavior. A mass of 65kg was applied to the point on the Y axis at distance L from the pivotal joint, and the distance L increased to increase a moment about the X axis. The actual centroid location, which depends on this load and the mass of the chair itself, shifts a shown by a solid curve in the Figure. The estimated centroid position showed a good agreement with the theoretical curve in most of the load range. Errors in the higher moment range are due to the nonlinearity of the coil springs and the suspension mechanism. The optimal pivot angle started at 45 degree when no moment was applied. As the moment increased, the angle increased to make the footprint wider. When the moment reached 100Nm, the optimal pivot angle hit the upper limit of the angle, ϕ_{max}=62.5deg, and beyond this point the optimal angle was kept at the upper limit although the centroid position shifted further. As a result, the stability margin decreased more quickly than that in the range where the optimal angle was lower than ϕ_{max}. Nevertheless, the stability margin did not vanish until the moment reached approximately 300Nm, which is an extraordinary case. In contrast, stability margin vanishes soon when the proposed stability augmentation was not used. As shown in the figure, stability cannot be maintained in a broad range when the footprint configuration is fixed at ϕ=45deg. This shows a significant advantage of the stability augmentation control implemented on the prototype wheelchair. The vehicle remained stable even when a patient of 100kg in weight fully extended his upper body towards one side of the chair.

8.3 Continuously Variable Transmission

The transmission ratio varies about 1.5 times (45 to 27.5deg.) or about 2times (62.5 to 27.5deg.) depend on the footprint configuration shown in Figure16. This characteristics would be used for the continuously variable transmission (CTV) of the vehicle.

Figure19 shows experimental results of variable transmission control. Figure19(a) shows plots of the pivot angle ϕ, the vehicle velocity reference V^*_v, the actual vehicle velocity V_v and the motor angler velocity ω when the vehicle changed the velocity with the beam angle fixed at 45degrees; the isotropic configuration. Since the transmission ratio between wheel angler velocity and vehicle velocity is the same all the time, motor torque hit the limit at a certain velocity and the vehicle velocity is saturated at the value. Figure19(b) is the experimental result when the pivot angle ϕ was varied by the variable transmission algorithm. Since the direction of vehicle motion which might be given by the operator can not be predicted, the pivot angle should be kept at around the isotropic configuration at zero or slow speed. For this purpose, the reference of the pivot angle larger than 45 degrees given by the variable transmission algorithm would be ignored, i.e., the layer of this algorithm does not take control.

By means of the variable transmission control, both the wheel angler velocity ω and the pivot angle ϕ are varied continuously and smoothly. As a result, the maximum vehicle velocity has been increased about 15% higher than that achieved by the isotropic footprint configuration.

The other hand, varying the footprint configuration can also change the traction force between ball wheels and the ground. Large traction force is needed not only for getting high accelerations or decelerations but also for climbing up ramps. The small traction force would restrict the vehicle to go through steep ramps in crowded area, especially in the residential homes. The wider footprint configuration in lateral direction allows the vehicle to provide the larger traction force. Figure 20 is the photo of the wheelchair climbing up a 10degrees' ramp. Table 1 shows the experimental result of going up ramps with various footprint configurations. "O" indicates the wheelchair successfully climbing up the ramp and "X" indicates the failure. The prototype wheelchair could climb up 12.5 degrees' ramp with the wider configuration, ϕ=62deg. This allows the wheelchair to ride on a low-floor van without any powered lifting mechanism.

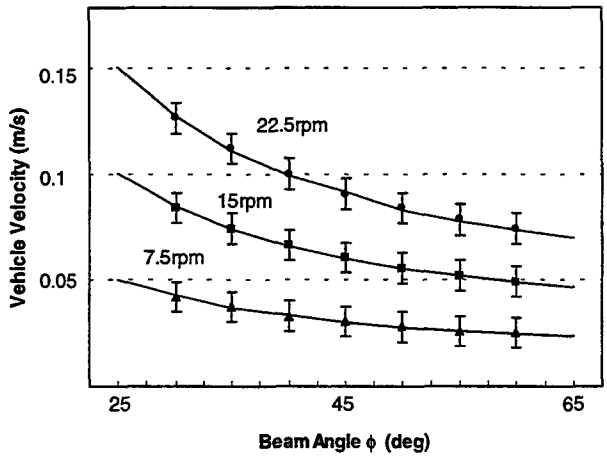


Figure 17: Vehicle Velocities against Beam Angle

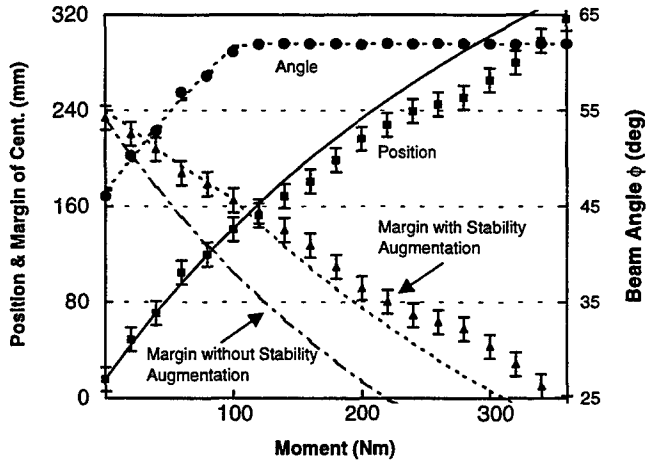
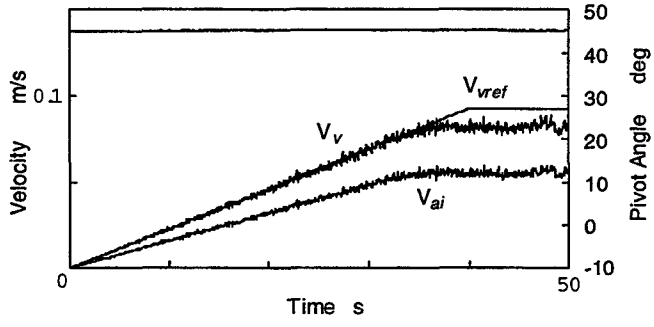
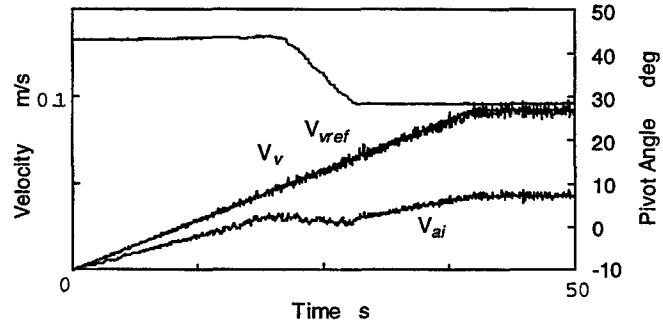


Figure 18: Stability Augmentation Control



(a) Fixed at Isotropic Beam Angle($\phi = 45$deg)



(b) With Continuously Variable Transmission

Figure 19 : Variable Transmission Control

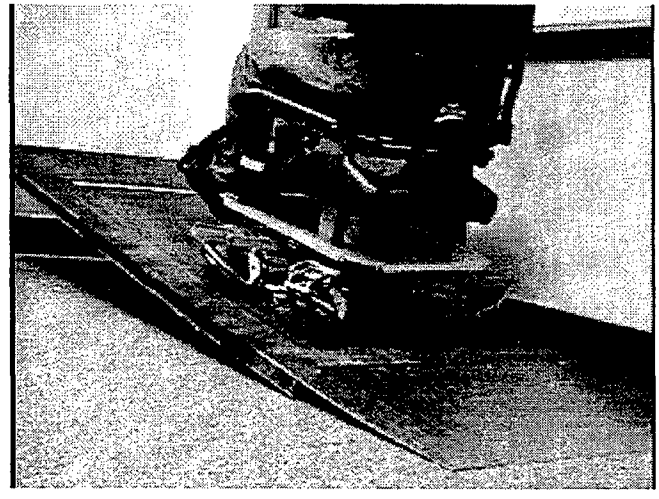


Figure 20 : Wheelchair Climbing up a 10degrees Ramp

Table 1 : Limit Angle of the Ramp along the Vehicle Footprint Configuration

Inclination of Ramp (deg.) \ Footprint Configuration ϕ (deg.)	28	365	45	53.5	62
2.5	O	O	O	O	O
5	X	O	O	O	O
7.5	X	X	O	O	O
10	X	X	O	O	O
12.5	X	X	X	X	O
15	X	X	X	X	X

9. Conclusion

A new mechanism for varying footprint for a four wheeled holonomic omnidirectional vehicle has been proposed and applied to a mobile platform of a wheelchair. A reconfigurable mechanism consists of two beams has been developed and the kinematic model has been obtained. An extra 1DOF for varying footprint not only provides an additional functionality to the vehicle but also solves an over constraint problem of four wheeled vehicles. The vehicle's 4DOF including a freedom of the reconfiguration of the footprint can be controlled independently by four motors driving ball wheels, i.e., driving the four ball wheels allow the vehicle to move in an arbitrary direction with an arbitrary angular velocity and change the footprint configuration at the same time. Then we have established a stability augment control algorithm, variable transmission algorithm and footprint shape control based on the proposed reconfigurable mechanism. Tasks required to the wheelchair with reconfigurable mechanism have been decomposed, analyzed and coordinated by the subsumption control architecture. These control methods has been implemented to a wheelchair prototype. Experimental results using the prototype have shown augmentation of the stability, smooth changing the transmission ratio and providing the large traction force of the wheelchair. These different tasks are coordinated by the subsumption architecture and share the vehicle control properly depending on situations of the wheelchair and requirements of an operator.

References

[1] H.P.Moravec, Ed., : "Autonomous Mobile Robots Annual Report - 1985," Robotics Institute Technical Report, Carnegie Mellon University, Pittsburgh, PA. , 1986.

[2] U. Borgolte, et al. : "Intelligent Control of a Semi-Autonomous Omnidirectional Wheelchair," Proc. of the 3rd International Symposium on Intelligent Robotic Systems `95 (SIRS `95), 1995.

[3] M. Gerke and H. Hoyer : "Planning of Optimal Paths for Autonomous Agents Moving in Inhomogeneous Environments," 8th Int. Conference on Advanced Robotics (ICAR97), July 1997

[4] F.G.Pin and S.M.Killough : "A New Family of Omni-directional and Holonomic Wheeled Platforms for Mobile Robots," IEEE Transactions on Robotics and Automation, Vol.10, No4, pp480-489, 1994

[5] M.West and H.Asada : "Design of a Holonomic Omnidirectional Vehicle," 1992 IEEE Int. Conf. on Robotics and Automation, pp97-103, May.1992.

[6] S.Hirose and S.Amano : "The VUTON : High Payload High Efficiency Holonomic Omni-Directional Vehicle," 6th Int. Symp. on Robotics Research, October.1993.

[7] M.West and H.Asada : "Design and Control of Ball Wheel Omnidirectional Vehicles," 1995 IEEE Int. Conf. on Robotics and Automation, pp1931-1938, May.1995.

[8] S.Mascaro, J.Spano and H.Asada : "A reconfigurable Holonomic Omnidirectional Mobile Bed with Unified Seating (RHOMBUS) for Bedridden Patients," 1997 IEEE Int. Conf. on Robotics and Automation, pp1277-1282, April.1997.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Physical Aids

CHAPTER 14

Smart Mobility and Monitoring Aid: A Helping Hand for the Elderly

S. Dubowsky

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Home Automation and Health Care Consortium Progress Report

Smart Mobility and Monitoring Aid: A Helping Hand for the Elderly

Professor Steven Dubowsky
Department of Mechanical Engineering

Reporting Period October 1, 1997 - March 31, 1998

1. Background

The objective of this three-year research program -- which started October 1, 1997 -- is to develop the fundamental technology for a Smart mobility Aid and Monitoring (SAM)system to meet the needs of elderly living independently or in senior assisted-living facilities.

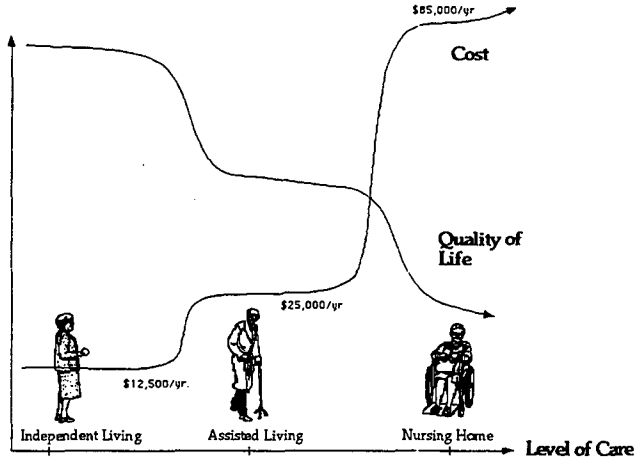


Figure 1: Progression of Elderly into Nursing Homes.

1.1 Motivation

The motivation behind this project is captured by the graph in Figure 1. As the elderly individual moves toward higher levels of care (i.e., from independent living to assisted living facilities to nursing homes), costs increase and quality of life decreases rapidly. The largest change occurs during the transition into a nursing

Table 1

Need	Physical Deficiency	Cause
Guidance	Failing memory, disorientation	Senile dementia, including Alzheimer's.
Physical support	Muscular- skeletal frailty, instability	Osteoporosis, Diabetes, Parkinson's, Arthritis, lack of exercise, failure of vestibular organs.
Health Monitoring	Poor cardiovascular function, susceptibility to strokes and heart attacks.	Poor diet, old age, lack of exercise, illness (e.g., flu or pneumonia)
Medicine Scheduling	Need to take a variety of medicines coupled with failing memory and disorientation,	Senile dementia, general failure health.

home, so delaying the onset of this transition will be extremely beneficial for the individuals and society. The transitions into higher levels of care are often mandated by specific needs which are summarized in Table 1. To prevent the transition into a nursing home, assisted-living facility aides can usually support these needs but it would still expensive. One shift of an aide costs $20-30k per year. The objective of this proposed work is to develop a less expensive and more dignified alternative for keeping the elderly as independent as possible.

1.2 Schedule

The first year of this work is focused on defining the technical challenges of the development of such a system, identifying potential solutions to these problems and performing a preliminary analysis, simulations and experiments to gauge the effectiveness of these solutions.

The key milestones of this year will be

(1) System design concept of a Smart Monitoring and Mobility Aid.

(2) The identification of key enabling technology that must be developed to meet the system-design objectives.

(3) The development of a program of research to generate the required technology and resources.

2. Reporting Period Technical Progress

Progress has been made on all three of the key milestones, as described below.

2.1 System Level Design

2.1.1 Specifications and Features

A preliminary list of specifications for a prototype SAM (dubbed a "smart-walker") has been generated. The primary user would be an elderly person who needs some physical and mental assistance to walk. Members of our team visited an assisted-living facility of the elderly, interviewed the residents and observed their living patterns. These visits will continue in order to maintain interaction with the residents. We expect that some future experimental evaluation of the SAM will involve these residents. Based on our studies and observations, we defined the technical specifications of our system which are summarized in Table 2.

Table 2: Specifications for Smart-walker.

User and operation environment characteristics	
Potential Users	Elderly with mobility difficulty due to physical frailty and/or disorientation due to aging and sickness.
Environment	Assisted-living facilities. Known structured indoor environment with random obstacles such as furniture and people. Flat and semi-hard floor or ramps less than 5 degrees.
System function and features	
Physical stability	Provide equal or better of a standard walker.
Guidance and obstacle avoidance	Provide guidance to destinations via global sensing, planning and obstacle avoidance strategies.
Health monitoring	Provide continuous health monitoring (details TBD).
Communication	Able to communicate with patients and caretakers.
Mechanical specifications	
Mobility device	Compact and robust wheel-based mobility platform with design reconfigurability .
Speed	Able to assist the elderly walking up to 0.5m/s.
Loading capacity	Able to support average body weight of an elderly person and provide 2 to 4 kg pulling force for stability and guidance.
Weight	Approximately 15 kg.
Physical size	Approximately equal to a conventional walker
Battery life	About 5 hours between charges.
Sensing and computing	
Computing power	On-board computers sufficient for planning, control, health monitoring and communication.
Sensors and aides to Navigation.	Vision based global sensing for high level planning. Ultrasonic based sensors for obstacle avoidance. Optical encoders for dead reckoning and motion control. Map based localization.

2.1.2 Test-bed: Smart-Cane

While the longer-term goal is to design a smart walker, in the short-term a smart-cane (Figure 2) will be developed. The smart-cane will serve as a test-bed for various technologies that the smart-walker will require. The smart-cane would have less functionality, but be easier and cheaper to design and build than the smart-walker. With the smart-cane, we plan to test core technologies such as human-machine interface, low-level control, obstacle avoidance, high-level control and health monitoring.

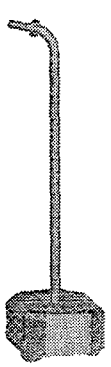


Figure 2: Smart-cane

Figure 3 shows the general design of the smart-cane. It consists of a mobility platform, ultrasonic sensors for obstacle detection, health sensors, a six-axis force/torque sensor for the human-machine interface, an on-board computer and interface electronics. The main considerations in the selection of an appropriate drive and steering configuration for the mobility platform were maneuverability, controllability, traction and stability, navigation, environment impact and simplicity. Various mobility forms and wheel configurations have been investigated. Holonomic configurations built from omni-directional wheels exhibit good maneuverability in tight quarters, but they are in general complex and may cause serious reliability and cost problems for this application. Conventional two degree-of-freedom (DOF) wheeled (non-holonomic) mobility devices can be difficult to control in tight spaces due to their over-constrained nature. The control problems for using the non-holonomic drives, however, appear to be more tractable for the smart-cane than those of using omni-directional drives. This question will be revisited for the walker.

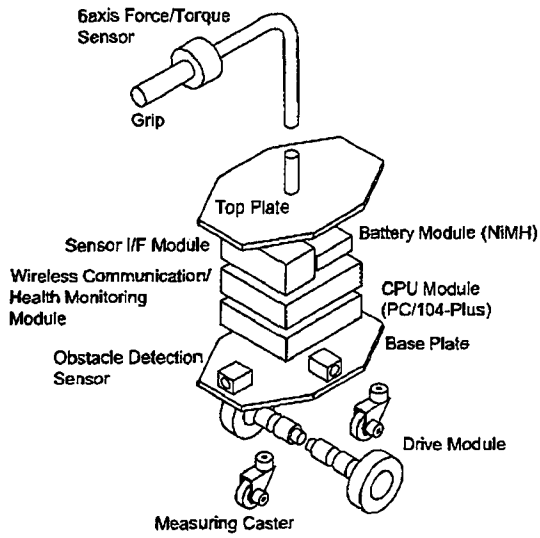


Figure 3: Smart-cane Design.

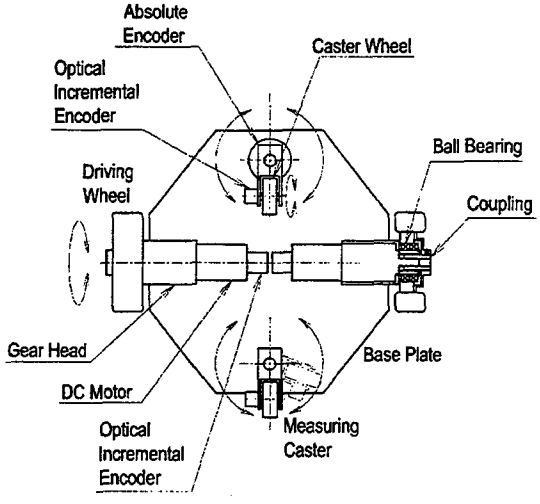


Figure 4: Layout of Motors, Casters and Sensors.

The drive module uses a common two-wheel skid-steering drive because of its simplicity in construction. It has two individually controlled driving wheels and two passive casters, as shown in Figure 4. This configuration has good maneuverability in tight environments as it allows an on-spot spin. Each drive motor will have an incremental optical encoder for motion control and dead-reckoning. The inevitable slippage in the driving wheels together with other sources of error limit accuracy of the dead-reckoning based solely on the encoders on the driving wheels. The design places additional encoders on the casters, for wheel rotation and headings, to improve the navigation accuracy of the system.

The two DC motors have a gearhead of 29:1. The gearhead will be connected to the drive with a specially designed flexible coupling. that makes wheel/motor alignment easy and reduces the effects of wheel impact on the gearhead. We will mount the driving wheel to the chassis by two ball bearings. The motors and custom-designed casters will use bearing caps to prevent dirt contamination. Most of the materials for the smart-cane will be aluminum or plastic.

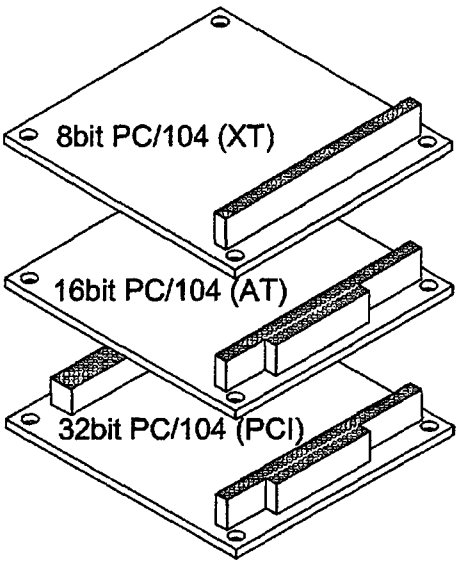


Figure 5: PC/104 Card Stack

2.1.3 Computer Architecture

The heart of the SAM will be a computer capable of performing the planning, control, health monitoring and communication tasks simultaneously. Additional constraints in the selection of the computer are small package size, high reliability, rugged design, and low electrical power consumption. To satisfy these requirements, the PC/104-plus architecture was selected as the CPU of our smart-cane. It uses small cards (3.55 x 3.755 inches) that stack together (see Figure 5) without the need for a motherboard or a back plane. This feature lets one improve the performance of the computer without having to replace the entire system. The PC/104 architecture is relatively fast with the theoretical bus transfer rate at 133MB/sec. The architecture is also based on mass-produced standard chip sets for desktop PC's so the system costs are relatively low. The consumption of the computer system would be around 10 watts, so two or three high capacity NiMH

batteries will be adequate for the smart-cane. The suitability of this architecture for the smart-walker will be reevaluated during the testing of smart-cane.

2.2 Required Technology and Research

Various technologies will need to be developed for use with the smart-cane and walker. In the recent work period, we identified several of these and list them below.

2.2.1 Mechanical Design

The design of such a system poses interesting mechanical design challenges. The optimal design of the SAM will be based on data we collect with prototypical devices (i.e., the smart-cane). It will therefore be useful to be able test various configurations of the device. The device must also be reconfigurable to meet the needs of different users. The SAM must adapt its configuration to an optimal form for support and stability of the elderly user. To design this feature into the SAM, one must understand the effects of different configurations on the elderly person's gait and stability.

The driving and steering configuration of the mobility aid will continue to be studied in order to achieve the an efficient and reliable mechanism with enough maneuverability to navigate the congested indoor environment of a typical assisted-living facility. A modular driving and support structure will enable the SAM to function in various environments using different configurations.

The SAM must also alter its geometry to maintain stability, safety and maneuverability. Certain situations, such as an encounter with a narrow portal, or a change in the user's biomechanics (i.e., gait) may require the SAM to make an on-line mechanical configuration change. To implement capability to alter configuration requires a kinematic study of both the expected users and the mobility device.

The arrangement and packaging of health-monitoring equipment, navigation sensors, electronics and computers will be based on compactness of design, optimal sensor configuration, communication issues and noise rejection.

2.2.2 Dynamics, Control and Planning

There are a variety of challenges related to the dynamics and control of the SAM. The smart-cane will use skid-steering, and it quite possible that the walker may use it too. The dynamics of this type of driving strategy are non-holonomic, which presents certain challenges in the area of control. While a non-holonomic system may have no configuration constraints, it does have motion constraints that will affect both obstacle avoidance and global path planning.

The SAM's fundamental tasks for planning are to determine where it is, where it is going and how to get there. The estimate of the location and heading can be obtained from localized sensors (such as encoders) and a dead-reckoning algorithm. The dead-reckoning error, however, is accumulative and can go unbounded. In order to achieve accurate positioning, the system must make periodical corrections using a globalized sensor system, such as beacons and triangulation.

A map-based positioning and planning strategy is proposed for the SAM. As the SAM is targeted for indoor environment, it can start with a global map based on the known structure of the environment. The map would need to be constantly updated using the vision and/or ultrasonic sensors. Accurate positioning is obtained by pattern matching and then fed to the path planing module. The path planning module will create the destinations and trajectories for the system that will be used to drive the SAM. The trajectory might be modified after encountering previously undetected local obstacles. The building of an accurate local map based on limited sensor input is a demanding task. The system must learn from its experience and adapt to new situations, including adpating to the users behavior patterns. Planning in the presence of random obstacle and with the interaction with the human dynamic system is another challenge.

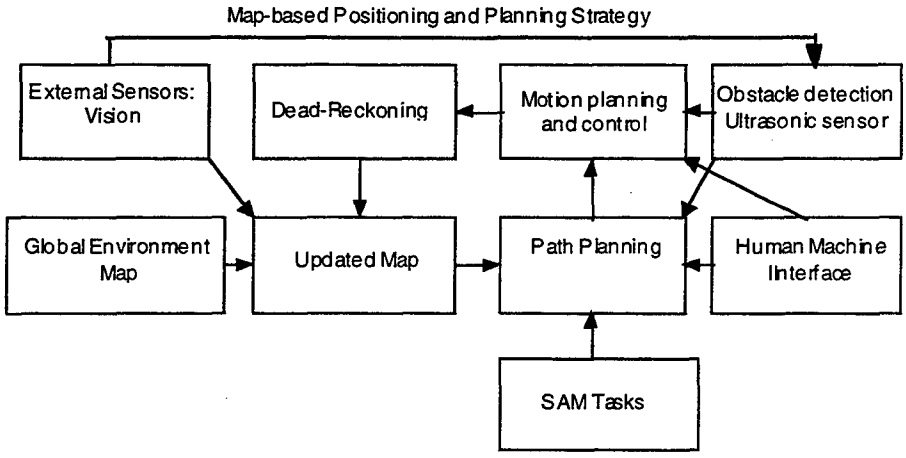


Figure 6: Map-Based Planning and Control

2.2.3 Human Factors

There are a variety of human-factor issues that must be tackled during the course of this project. This goes well beyond the simple requirement for the SAM to be ergonomic. The device must provide both physical and mental support simultaneously. One significant question that arises from the SAM's requirements is "how do we guide an elderly person without causing her to lose her balance and fall?" While there have been volumes of research done on balance problems of the elderly, very little research has been done on mechanics of guiding an elderly person without making her fall. A mockup of smart-cane was built for estimating the force required to guide a cooperative elderly person. While this data was useful for selecting the motors for smart-cane, it cannot substitute for a deeper understanding of the mechanics of stable guiding of the elderly.

It is clear that the smart-cane must consider not only its own dynamics, but also that of the user. Smart-cane must regulate both the user's walking (e.g., path, speed and gait) and vital signs (e.g., heart rate, blood pressure) simultaneously. Most users will probably have similar biomechanical models relating the walking states to the "vital" states. It is expected, however, that each individual will have different biomechanical parameters and thus the SAM must be able to adapt to these.

The SAM-user might be in varying modes of supervision depending on their clarity of mind. During periods of user-lucidity, for example, the SAM must take a passive role in aiding the user, but during periods of user-confusion the SAM must become more active. It is not yet clear how to measure the users confusion. This is an issue that must be resolved to prevent the SAM from becoming an "active-aid" at the wrong time and possible causing injury to or exhaustion of the user.

One of the strategies we are considering for controlling the "feel" of smart-cane is an admittance based control structure. Admittance control (illustrated in Figure 7) regulates the dynamic feel of the device (i.e., mass and damping). With this strategy, the SAM will be able to create zones of varying dynamic properties, or "virtual terrain" in the assisted living facility. To prevent elderly people from accidentally walking into a off-limits room, one could create a virtual "ramp" or a virtual one-way dashpot to discourage motion in the direction of the danger. This type of passive system could also be used to keep the user along a planned path. One could create a virtual canyon that would be difficult for the users to "escape" from.

2.3 Program of Research

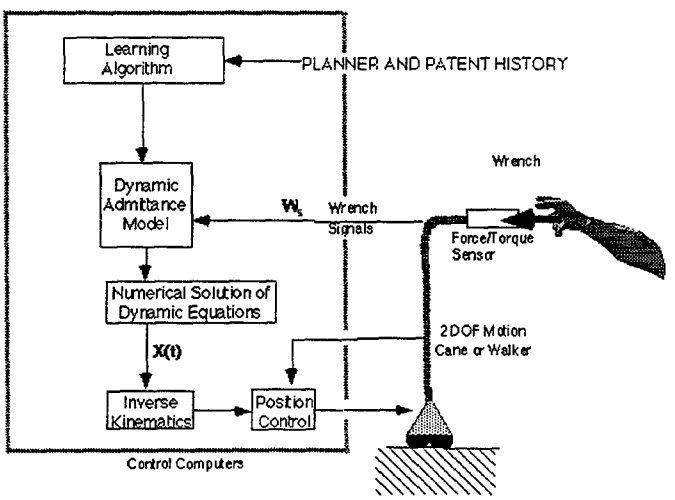


Figure 7: Admittance-Based Control

Two Ph.D. graduate students and two visiting scientists are working on the project. One student is focusing on low-level control and design of the smart-cane, while another is focusing on the high-level control and design. Work on the human-machine interface is split appropriately among the two students. One of the visiting

scientists is concentrating on the electronics and computer architecture, while the other is assisting with the mechanical design. Our current program consists of developing the smart-cane and performing field-tests of it with elderly people. We hope to learn more about the dynamics, kinematics and medical aspects of guiding an elderly person with a mobility device. This will help us to design a better smart walker.

3. Work to be Completed During Reporting Period

April 1, 1997 - June 31, 1998

We will finish the fabrication of the smart-cane in this quarter and begin testing various control strategies. We will incorporate several of the medical sensors being developed by our consortium colleagues in other labs. A hierarchical control system will be developed that implements low-level obstacle avoidance, health monitoring and regulation, and simple path planning. Several models of human balance and walking will be developed for use in the control-system. Lastly, data on walking forces will be collected using the smart-cane to determine the accuracy of the models.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Physical Aids

CHAPTER 15

Large Scale, Mechanical Surface Waves for Elastic Body Transport – The Hyperbed

H. Asada, J. Spano

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Large Scale, Mechanical Surface Waves for Elastic Body Transport - The Hyperbed

Joseph Spano and Haruhiko H. Asada

d'Arbeloff Laboratory for Information Systems and Technology
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139
spano@mit.edu, asada@mit.edu

Abstract

Surface waves are shown to be a viable transport mechanism for large-scale, elastic bodies given certain conditions and constraints on the design. Elasticity theory is employed to elucidate the interaction between the elastic body and the mechanically generated surface wave. Assuming a general friction model, conditions are derived for transport performance. Specific case studies are presented to illustrate solution results, consequences, and impact on system design. In addition a prototype design is presented. Applications of this system include a myriad of manipulation possibilities for human subjects and in a more general sense, elastic bodies.

1 Introduction

The task goal of this research is the manipulation of the human body. In particular this manipulation includes reorientation and reconfiguration of the overall body posture and the translation of the human body across the surface which holds the human body in space. Large scale, mechanically generated surface waves are explored as the tangential transport mechanism necessary to realize the task objectives. An array of active nodes are coordinated to generate a psuedo-continuum surface wave behavior that can be used as a tangential transport mechanism, allowing flexible routing of tangential forces. With appropriately designed periodicity, amplitude, and wave velocity the resulting wave behavior can optimally propagate both rigid or elastic, large-scale bodies.

This work is significant because it offers an alternative manipulation scheme for the flexible transportation of both elastic and rigid bodies, and most importantly humans. Traditional transport devices consist of rigidly fixed belts or chains or units that operate in a rigidly coordinated fashion. The disadvantage of this is that the when the manipulation task is altered the system must be redesigned and rebuilt. By utilizing a highly distributed, flexible design framework, these changes can be made by control software alterations rather than hardware adjustments.

A host of applications for the manipulation of humans can be suggested. The first obvious application is the physical positioning of debilitated individuals. This can take a variety of forms such as transporting a human from bed to chair and moving in and out of a car. Other possibilities include the fine positioning of the human body for medical imaging. Use of the distributed actuator system for skin therapy applications is clearly possible. Wave shapes and trajectories can be

designed to alter normal and tangential forces on the human skin surface to provide healthy maintenance of skin condition by massage.

Work related to this project includes efforts by researchers involved in the exploitation of surface wave behavior for micro-manipulation. [Tadokoro, 1997] [Suzumori, 1996] shows the use of coordinated pneumatic actuators for small-scale manipulation. Other related work includes that which uses a variety of mechanisms such as omni-directional wheels [Luntz and Messner 1996] for rigid body manipulation. However there appears to be no work seeking to exploit surface wave behavior for the manipulation of large-scale, elastic bodies such as humans.

2 Kinematic Analysis

2.1 Basic Properties of Surface Waves

Surface waves are a particular type of wave created by a periodic elliptical motion of each particle on the surface with a phase lag between adjacent particles. Note that surface waves are not merely a longitudinal wave nor a transverse wave, but are a mixture of the two. As shown in Figure 1, each particle moves in both longitudinal and transverse directions. For the sake of simplicity, it is assumed in this paper that the particle trajectory is a complete circle.

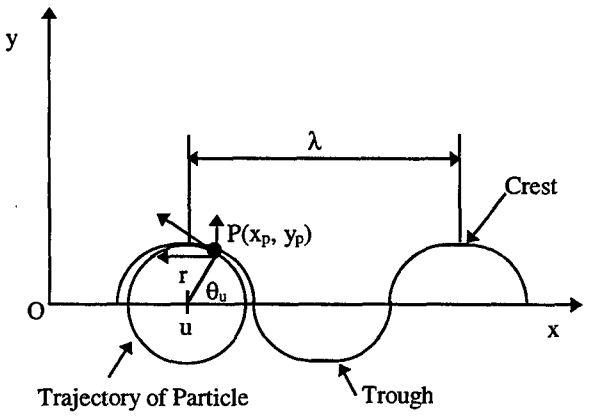


Figure 1

Let P be a particle at distance u from origin O along the surface when the wave is not present. The coordinates of particle P when waves are created are given by

$$x_p(u,t) = r\cos(\theta_u) + u$$

$$y_p(u,t) = r\sin(\theta_u)$$

where r is the radius of the circular trajectory and θ_u is the angle given by

$$\theta_u = \omega t + 2\pi\frac{u}{\lambda}$$

where ω is the angular velocity of the circular motion, and the second term, $2\pi u/\lambda$, is phase lag. The phase lag varies continually along the surface in proportion to distance u. Parameter, λ, is the wave length, that is, the distance between two successive crests or troughs, as shown in Figure 1. The frequency ω provides the number of waves passing a fixed point per unit time. Therefore, the wave velocity, v_{wave} that is the velocity at which wave crests appear to move, is given by:

$$v_{wave} = -\frac{\omega}{2\pi}\lambda$$

The velocity of the particle at each crest is given by

$$v_{crest} = \left.\frac{\partial x_p}{\partial t}\right|_{\theta_u=\pi/2} = -r\omega$$

while the velocity at the troughs is given by

$$v_{trough} = \left.\frac{\partial x_p}{\partial t}\right|_{\theta_u=-\pi/2} = r\omega$$

Note that at the crest the wave velocity and the particle velocity are aligned, but at the trough they are opposed.

2.2 Surface Waves for Transporting Humans and Flexible Objects

The goal of this paper is to transport a human lying on a bed by creating surface waves on the bed surface. As shown in Figure 2, a human body can be supported at the crests of the surface waves and be moved horizontally along the bed surface. As each particle moves along a circular trajectory, first it moves upwards, (a), contacts the human body, supports the body weight, moves in a horizontal direction, (b), moves downwards, detaches from the human body by, (c), and moves horizontally back to the original point, (d). When the particle is in contact with the human body, it moves together with the body in the horizontal direction and thereby transports the body.

There are several issues which need to be overcome in order to transport a human by surface waves. Since a human body is a multi-d.o.f. system consisting of many flexible bodies connected by articulated joints, it may conform to the wave surface as shown in Figure 2. As it slacks, the body may contact the trough side of the wave surface, which moves in the direction opposite the crests. Therefore the human body may be dragged backwards. This results in low efficiency and , more importantly, leaves the human in an uncomfortable situation.

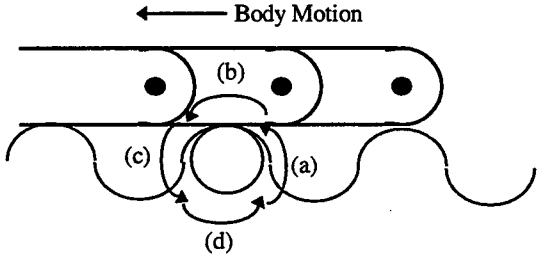


Figure 2 - Flexible object transported by surface waves

To avoid slack and uncomfortable situations,

- The wavelength λ must be shortened,
- The waves must be deep, and
- The crest must be gentle and wide (long).

The shorter the wave length, the less the body slacks, and the more crests the bed surface generates to support the body. The deeper the troughs become, the less likely the body will be to contact the troughs. The design parameters, λ and r,

must be chosen to meet these requirements. The third design guideline addresses the shape of the crests, providing an effective solution to the slack and comfort problem.

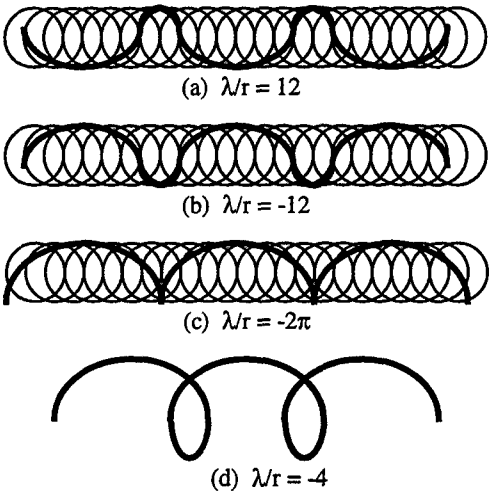


Figure 3 - Surface Waveforms for different λ/r

Figure 3b was created by turning the waveform in Figure 3a upside-down. Note that one side of the wave becomes gentle and long, hence the body can be supported by a broader area of the bed surface. The sharp crests in Figure 3a cause a concentration of stress, which provides an uncomfortable and even dangerous situation for patients with fragile skin conditions. The gradual contours in Figure 3b would significantly reduce the stress concentration and gently support the body. This gentle curvature can be generated by replacing the wave length, λ, by its negative value, -λ. Negative wavelength does not exist in natural systems, such as deep water waves, but can be created by artificial means. As demonstrated later in Section 5, the generation of surface waves with a negative wave length is feasible and quite useful for resolving the stress concentration and body slack problems.

In Figure 3, various waves are plotted for different wave lengths, λ. As the absolute value of the wave length becomes smaller, the proportion of the gradual side to the sharp side becomes larger. However, at a certain point the trough becomes a sharp edge and the waves collapse beyond this point. Namely, as shown in Figure 3d, the contour of the wave surface crosses over. This waveform, although very gradual on one side, is not physically realizable because no continuous surface can be manufactured that can continually wrap over itself. Therefore the wave form with the sharp edge, i.e. Figure 3c, provides the lower limit of the wavelength |λ|. The sharp edge is a stagnation point, where the contour of the wave surface has a zero gradient in the u direction. Namely,

$$\frac{dx_p}{du} = 0$$

Evaluating the gradient at θ_u=-π/2 where the sharp edge exists,

$$\left.\frac{dx_p}{du}\right|_{\theta_u=-\pi/2} = \frac{\partial x_p}{\partial u} + \frac{\partial x_p}{\partial \theta_u}\frac{\partial \theta_u}{\partial u} = 1 + \frac{2\pi}{\lambda}r$$

Therefore, the stagnant condition is given by

$$\lambda = -2\pi r$$

Namely, the lower bound for the absolute value of the wave length is given by

$$|\lambda| > 2\pi r$$

This means that the absolute value of the wavelength must be longer than the circumference of the circular trajectory of radius r.

By choosing a small wavelength $|\lambda|$ and a relatively large radius r that satisfy condition (9), one can obtain a broad area of contact surface supporting the human body and thereby reduce the stress concentration. In the following sections, we will analyze quantitatively the stress distribution over the contact surface as well as the speed at which the human body is moved by the surface waves.

3 Contact Problem Formulation and Assumptions

Peak interaction pressure and efficiency of transport are the two primary factors of concern in the design of the surface wave actuator for transporting humans. Our analysis should provide insight into how wave parameters can be tuned to enhance the efficiency of the transmission of motion and reduce peak pressure. We seek quantitative relationships between our two performance criteria and wave parameters.

To begin we will ignore the spatial discretization of the prototype and move forward with the analysis assuming that we can generate arbitrary continuous, sinusoidal surface waves. We must model the physical scenario of the human lying on the bed surface. To reveal the interaction forces between the two bodies we will use elasticity theory to predict the local deformations and pressure profiles resulting from the interaction. By following the assumptions utilized in the theory we can derive the quantitative results regarding interaction pressure in the normal direction.

Figure 4 illustrates the body lying on a wavy surface. This situation has been idealized as a perfectly elastic plate lying on a series of cylinders. To make this idealization we have made a series of approximations.

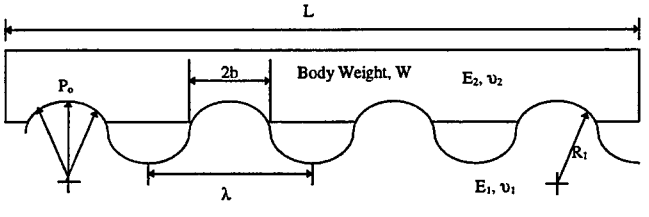


Figure 4

These include the following:

- Linearly elastic, isotropic materials
- Contact area is small compared to radius of curvature
- Sinusoidal surface can be approximated by surface of second degree
- Surfaces are perfectly smooth i.e. Only normal pressures which arise during contact are considered.

The human is not linear, isotropic, but wave parameters can be designed such that the remaining assumptions are reasonable.

From the results of the solution to the contact problem of two elastic bodies due to Hertz [Flugge, 1962], we find that the normal pressure profile at the interface is

$$P_n = P_o\sqrt{1-\frac{x^2}{b^2}}$$

where peak pressure is

$$P_o = \left(\frac{2N'}{b\Pi}\right)$$

where b is half the width of the contact patch

$$b = 2\left(\frac{N'R_o}{E_o\Pi}\right)^{1/2}$$

where E_o is given by

$$\frac{1}{E_o} = \left[\frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}\right]$$

and R_o is given by

$$\frac{1}{R_o} = \frac{1}{R_1} + \frac{1}{R_2}$$

Note that in the case of a cylinder in contact with a flat plate $R_2 = \infty$. Given a human body length, L, and weight per unit depth, W', the body weight per unit depth per wave crest is given by

$$N' = W'\frac{\lambda}{L}$$

4 Sensitivity of Peak Contact Pressure on Parameters

If we invoke the results of elasticity theory of two elastic bodies in contact we can determine the effect of surface wave parameters on peak pressure. This can be shown by starting with the expression for peak pressure

$$P_o = \left(\frac{2N'}{\pi b}\right)$$

By substitution of b and reduction, peak pressure is

$$P_o = \left(\frac{N'}{\pi}\frac{E_o}{R_o}\right)^{\frac{1}{2}}$$

We can express

$$R_o = -\frac{1}{K}$$

where K is the curvature of an arbitrary function. For our sinusoidal surface wave at the wave peak K becomes

$$K = -\left(\frac{2\pi}{\lambda}\right)^2 A$$

Substituting back into the expression for peak pressure we obtain

$$P_o = \left(\frac{4\Pi E_o W'}{L}\left(\frac{A}{\lambda}\right)\right)^{\frac{1}{2}}$$

Clearly to reduce peak pressure we must reduce the tunable ratio A/λ.

5 Derivation of Transmission Ratio

From our study of wave behavior earlier in the paper we know that the x-component of velocity over the surface of the wave is not constant. However, the body propagation velocity is constant. For this reason slip must occur between the interface of the body and the wave surface assuming the skin surface is not compliant. This slip will generate traction forces governed by some unknown nonlinear friction law that is a function of the normal pressure.

$$F_{tang} = f(P_n)$$

For a constant body velocity we must conclude that the sum of the forces applied to the body by the wave in the x-direction must be equal to zero to obey Newton's Law. Forces due to normal pressure cancel in the x-direction. Traction forces which are generated by relative motion between the body and wave do not cancel unless traction forces are generated in opposite directions along the wave surface in contact with the body. For this to occur the relative velocity which defines the direction of slip and the direction of the tangential forces generated must undergo a sign change at some point along the wave surface. At this point no slip will occur and the body velocity will be equal to the x-component velocity derived in section 2 of the wave surface at that point. This velocity is

$$V_{body} = A\omega\cos\left(\frac{2\pi}{\lambda}\right)x_{ns}$$

If we take the ratio of the body velocity with the wave velocity we obtain the effective transmission ratio of the transport mechanism.

$$T.R. = \cos\left(\frac{2\pi}{\lambda}\right)x_{ns}$$

By exploiting symmetry we can analyze just the RHP to determine the point of no-slip. By applying Newton's law, the equilibrium requirement that must be satisfied to solve for the x-position of the point of no slip is given by

$$\int_0^{x_{ns}} F_{tangx}\,dx + \int_{x_{ns}}^{b} F_{tangx}\,dx = 0$$

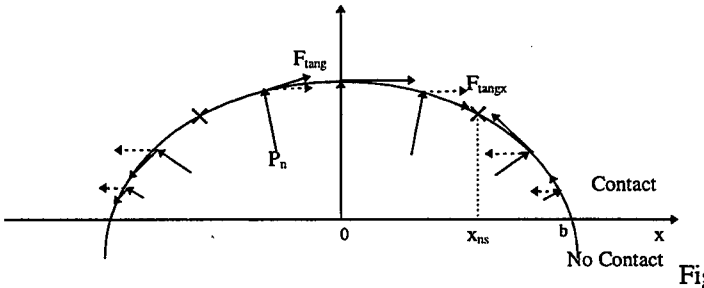


Figure 7

Let us explicitly determine the x-component of tangential force in terms of the tangential force and wave parameters for a sinusoidal wave as shown in Figure 7. By simple trigonometric arguments it can be shown that

$$F_{tangx} = F_{tang} \frac{1}{\sqrt{1+\left(-A\frac{2\pi}{\lambda}\sin\left(\frac{2\pi}{\lambda}\right)x\right)^2}}$$

With the assumption of small angles which is completely consistent with the assumptions made previously to justify the use of elasticity theory we can further simplify the expression to

$$F_{tangx} = F_{tang} \frac{1}{\sqrt{1+A^2\left(\frac{2\pi}{\lambda}\right)^4 x^2}}$$

One interesting result of this analysis is that for the body to have tangential velocity on top of the surface wave actuator requires the defining property of the surface wave that the surface particles have an x-component of velocity which requires movement in an elliptical (or circular) pattern. For instance if the wave was a transverse wave and the particles (or nodes for our prototype) moved only vertically there would be no body propagation. So we have shown that only a surface wave can display the behavior that we desire. For this reason the velocity of body propagation is independent of the direction of wave propagation, instead depending solely on the direction in which the elliptical displacement node paths are traced. In other words body propagation velocity is independent of the sign of the internodal phase difference, allowing us to utilize negative wavelength as discussed earlier to improve our design.

6 Design Trade-Offs Associated with Simple, Special Case Friction Law

By assuming some simple, special case friction law we can illustrate the connection between the tunable surface wave parameters and system performance with respect to peak pressure and transmission ratio.

Beginning with the expression for peak pressure that we derived earlier we can introduce the non-dimensional variable

$$\gamma = 2\pi\frac{A}{\lambda}$$

Peak pressure can also be converted to a non-dimensional form

$$P_o^* \equiv \frac{P_o L}{W'} = \left[\frac{2E_o L}{W'}\gamma\right]^{\frac{1}{2}}$$

Ideally we would like to derive an explicit expression for transmission ratio as a function of the parameter γ. To begin the derivation we must introduce an idealized friction model and calculate the position of no-slip. We assume a friction law of the form shown in Figure 9.

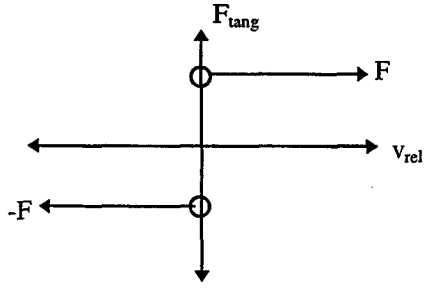


Figure 9

Applying our force equilibrium criteria to solve for the x position of the point of no slip we obtain

$$\int_0^{x_{ns}} F\frac{1}{\sqrt{1+A^2\left(\frac{2\pi}{\lambda}\right)^4 x^2}}dx + \int_{x_{ns}}^{b} -F\frac{1}{\sqrt{1+A^2\left(\frac{2\pi}{\lambda}\right)^4 x^2}}dx = 0$$

After integration and significant algebraic manipulation, indeed one can obtain an explicit expression for the transmission ratio as a function of the non-dimensional parameter γ. The expression is

$$TR = \cos\left[\frac{K\gamma^{1/2} + \sqrt{1+K^2\gamma} - 1}{2}\sqrt{\frac{1}{K\gamma^{1/2} + \sqrt{1+K^2\gamma}}}\frac{1}{\gamma}\right]$$

where

$$K = 2\left[\frac{W'}{L}\frac{2}{E_o}\right]^{\frac{1}{2}}$$

By numerical calculation we can make a parametric plot of both of our performance parameters, transmission ratio and non-dimensional peak pressure and make rational design decisions based on examining the contour. Figure 10 illustrates the results for the following range of γ.

$$.7 \le \lambda \le 1.0$$

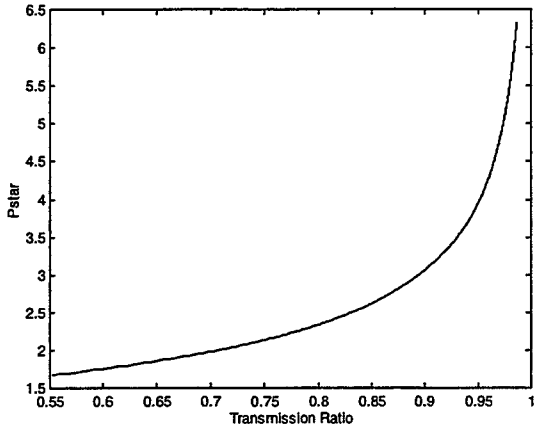


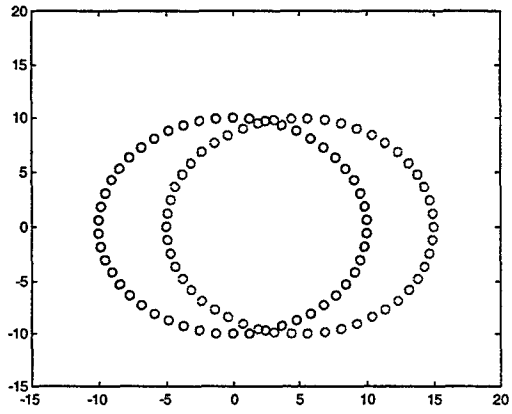
Figure 10

From the parametric plot the designer can with good accuracy, pick a design point that will specify the appropriate amplitude to wavelength ratio that will give the desired performance characteristics. Clearly from a design perspective of handling humans, it is desirable to choose a point in a region with both high transmission ratio and low pressure which corresponds with the lower, right hand region of Figure 10.

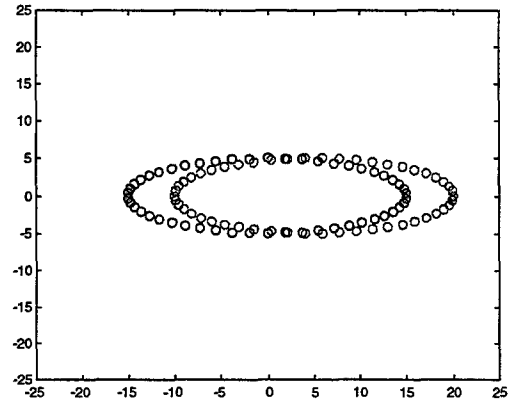
7 Expanding Our Design Space

Our analysis to this point has been focused on circular nodal trajectories as the building blocks of surface waves. We have seen that a kinematic limit exists with circular trajectories leaving us with the optimal overall surface shape corresponding

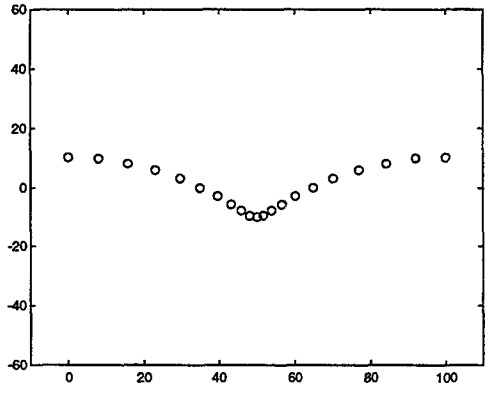
to a cycloid. Analysis of peak interaction pressure and tangential transport efficiency indicates that it is desirable to flatten the overall waveshape and it would be desirable to do so beyond the cycloid limit. To reach this end, elliptical trajectories have been explored as a way to extend our range of options and produce more efficient and comfortable surface waves. The analytical description of elliptical trajectories becomes quite complicated and the equations are not easily solved, so simulation was utilized as an exploration tool. The simulation was set up to provide a trace of two adjacent nodal paths and also to plot the overall surface wave shape. This gives a visual image of the node shape being simulated and the overall waveshape plot is used to evaluate whether looping is occurring.



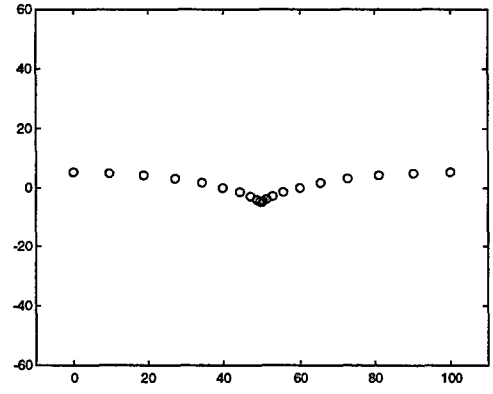
Trace of Adjacent Nodes-Circular



Trace of Adjacent Nodes-Elliptical



Surface Wave Shape-Circular



Surface Wave Shape-Elliptical

It is evident that the introduction of elliptical trajectories does, indeed flatten out the overall surface wave shape and does offer an improvement in overall performance. However, two issues should be noted. The first is that the elliptical paths tend to loop and any proposed set of parameters should be simulated first to ensure that this does not occur. A second issue is that adjacent nodes do move relative to one another. In the contact patch where the human is supported by several nodes this relative motion can cause a stretching of the skin and clearly this must be evaluated to ensure that the stresses on the human induced by this relative motion is not excessive.

8 Prototype Development

A prototype surface wave actuator has been developed in the d'Arbeloff Laboratory. (Figure 11) The prototype is made up of a series of mechanical nodes that are coordinated to generate an overall surface wave behavior. Successive mechanisms trace an elliptical trajectory and cross bars are set out of phase with each adjacent node, creating a discrete surface wave.

Figure 11 - Surface Wave Actuator

A matching pair of mechanical nodes are driven in tandem and connected by a long bar. Twenty eight of these pairs are driven in a coordinated manner to generate a wave motion that travels across the discretized surface of bars. Overall wave shapes are made by setting adjacent nodes out of phase with one another by the desired amount.

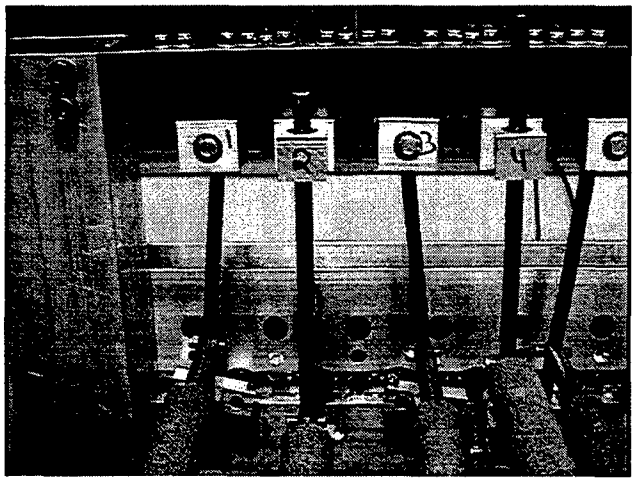


Figure 12 - Detail of Slider-Crank Node Mechanisms

A slider-crank mechanism was chosen as the fundamental mechanism for each node of the surface wave actuator. Using this mechanism one can adjust the shape of the nodal trajectory. By adjusting the length of the radius of the crank one can adjust the vertical displacement of the node and consequently the amplitude of the waveform generated by a set of nodes. By adjusting the position of the cross-bar on the connecting rod one can adjust the horizontal displacement of the node. The drive motor is an Aerotech model 1960 motor with tachometer and encoder used for experimentation. In addition a 25:1 gearhead is utilized to increase torque and reduce speed to levels reasonable for this system. Speed control of this motor sets the angular velocity of the nodes.

The bed dimensions were specified to be 84 inches by 78 inches. These dimensions were chosen to impose waves on both the major and minor axis of the human body. This allows the use of the one degree of freedom surface wave actuator to test the effects of wave propagation on two orthogonal axes of the human body.

The surface wave prototype has been utilized to successfully propagate humans across its surface. This success has illustrated that the surface wave actuator concept indeed shows promise as a unique actuator system capable of exerting tangential forces on elastic bodies, opening a host of application possibilities.

9 Conclusion and Future Work

The use of elasticity theory requires small deformations which puts limitations on the values for body weight and elasticity. Clearly we cannot control these parameters, and recognize that more complicated modeling techniques are necessary to gain a more detailed analysis of the transport phenomena of human bodies. However, with this simplified analysis we are still able to explain the qualitative behavior observed in prototype testing. Further effort must be directed toward improving friction models and relaxing the reliance on elasticity theory.

References

[Tadokoro, 1997] Satoshi Tadokoro, et.al., An Elliptic Friction Drive Element Using an ICPF Actuator, IEEE Control Systems Magazine, June 1997

[Luntz and Messner, 1997] Jonathan Luntz and William Messner, A Distributed Control System for Flexible Materials Handling, IEEE Control Systems Magazine, February 1997

[Flugge 1962] W. Flugge, Handbook of Engineering Mechanics, McGraw-Hill, 1962

[Suzumori, 1996] K. Suzumori, A Linear Pneumatic Rubber Actuator Driven by Mechanical Waves, Japan Society of Mechanical Engineering, Workshop on Robotics and Mechatronics, 1996

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Physical Aids

CHAPTER 16

Design and Prototyping of a Surface Wave Actuator Using Shape Memory Alloy Fibers

H. Asada, W. Finger

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Design and Prototyping of a Surface Wave Actuator Using Shape Memory Alloy Fibers

Haruhiko H. Asada
Principal Investigator

Bill Finger
Graduate Research Assistant

ABSTRACT

Surface Wave Actuators are a new technology for moving bed-ridden patients. The surface of the bed upon which the patient lies is itself the actuator. Discrete points on the bed are actuated in a wave pattern, and the patient is transported by the crests of the waves. The points are moved in a three dimensional trajectory in order to generate two dimensional motion. Each point is slightly out of phase with those adjacent to it. In this way, some points are always in contact to support the patient, while others detach and return.

The design of the trajectory used by the surface wave actuators is of critical importance, and is discussed at length. Also described are possible technologies that could be implemented as actuators, specifically shape memory alloy fibers. This paper also includes a detailed description of the design of the current surface wave prototype using those fibers.

d'Arbeloff Laboratory for Information Systems and Technology
MIT

1. Introduction

This report summarizes the research conducted at MIT on the implementation of surface wave actuators, using shape memory alloy fibers. A surface wave actuator is an active surface which is capable of horizontally moving an object placed upon it. These surface waves would be useful in the health care industry for moving bed-ridden patients, who are unable to help themselves. We have designed such a surface, using shape memory alloy fibers as the main actuator.

2. Surface Waves

2.1 Definition of Concept

Surface wave motion is accomplished by nodes, which are distributed across the surface. See Figure 1. A node is an element of the surface which contacts the body and supports it, yet is free to move in the horizontal directions. Each node can also move in the z direction to detach from the patient. The coordination of motion in the horizontal and vertical directions allows the node to form a rectangular or circular trajectory. At the top of the path, the node contacts the body, and accelerates it in the desired direction. At the end of its stroke, the node retracts away from the body so that it can return to its starting position.

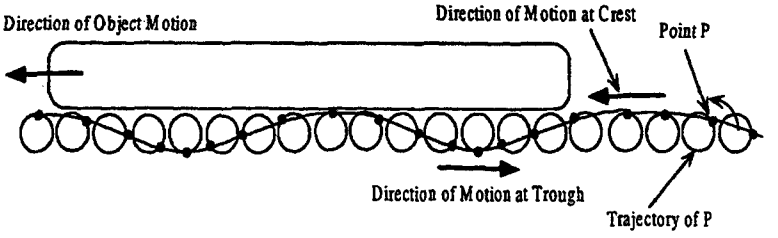


Figure 1: Surface Wave Actuation of an Object

2.2 Analysis of Requirements

There are several features which will be general to any surface wave actuator design. One, there must be more than one set of nodes, because at least one set must be in contact with the body at all times, while other sets return to begin the stroke anew. Each set of

nodes consists of nodes in phase, meaning they are at the same position in their trajectories at all times. The members of the set will be distributed evenly across the bed surface. Adjacent nodes will be slightly out of phase with their neighbors, so that some nodes are always in contact with the body.

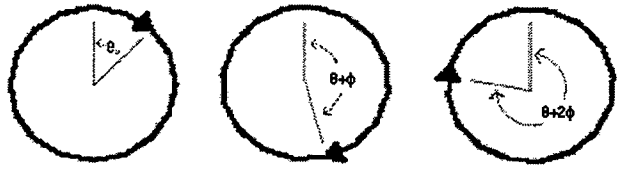


Figure 2: Phase Shift Between Nodes

All designs will have certain key parameters. x_s is the maximum range of travel of the node in the horizontal direction, z_s is the maximum displacement of the node in the z direction, and T is the period of the wave form. The maximum velocity of the body will be:

$$v = \frac{x_S}{T}$$

It therefore behooves us to maximize the stroke, and minimize the period of the wave shape. The actual realized velocity will depend on sundry other factors, such as amount of time the body is in contact with the node, the number of nodes in contact during motion, the compliance of the body, and slip. The details of trajectory design will be covered later in this report.

3. Actuators

3.1 Actuator Arrangement

A critical issue regarding actuator design is the orientation of the actuators with respect to the axes of motion. Two feasible possibilities have been explored in this research.

The first possible arrangement are decoupled axes, as shown in Figure 3.

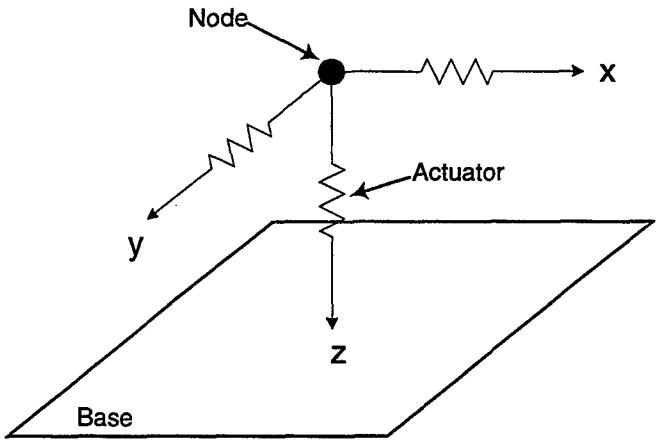


Figure 3: Decoupled Axes

Each direction of motion is controlled by one independent set of actuators. The advantage of this system is that motion in the principle directions is simple to control, and sensors mounted to the actuators will provide direct feedback of the appropriate direction. This arrangement is well suited to square node trajectories. There are several disadvantages to this configuration, however. The actuators will tend to interfere with each other in the horizontal directions. Because they must be located at the edges of the bed in order to have a stationary reference from which to pull, the actuators must be connected to the nodes with cables. These cables will overlap nodes lying in the same row or column of the matrix. Therefore, the cable attachment locations must be slightly offset between nodes of the same row or column. The cables must be placed below the plane of the surface, so they do not interfere with the z axis motion of other nodes. Another disadvantage is that if a passive force is used to support the body, which is a likely situation in cases of many dense actuators, each vertical actuator will be required to overcome that force to cause z axis motion.

Another possible arrangement of nodes is a tripod arrangement, shown in Figure 4.

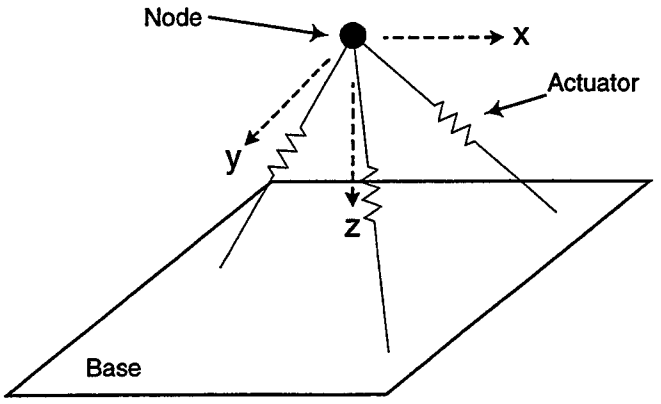


Figure 4: Tripod Arrangement

In this arrangement, the actuators must cooperate in order to generate motion in the horizontal or vertical directions. There are several important advantages to this design. Actuator interference is reduced, allowing higher node density. All three actuators cooperate to produce z-axis motions, reducing the size requirements of the actuators. There are also several disadvantages. The control of the system is more difficult, as the actuators must be well coordinated for planar motion to occur. Accurate sensing of the stroke of the actuators is required. Also, the range of output motion will be reduced, since the actuators are at angles with the directions of motion. See Figure 5.

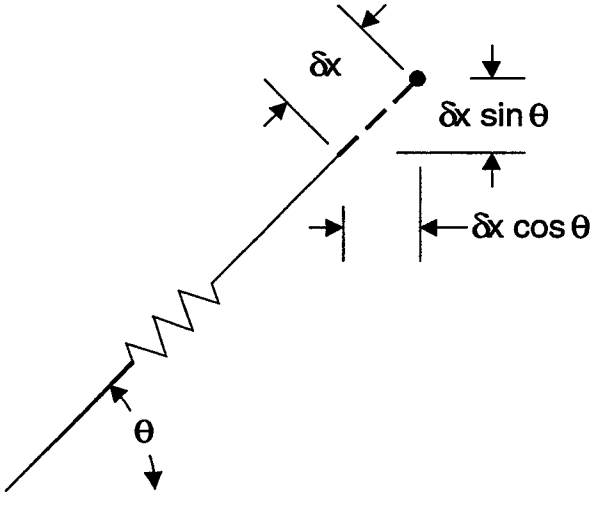


Figure 5: Actuator Stroke Usage

This arrangement is well suited for circular trajectories.

3.2 Shape Memory Alloys

3.2.1 General Description

Shape memory alloys are materials that can change their shape through the application of heat. The material undergoes a phase change between martensite and austenite when it reaches the transition temperature. This phase change results in a shift in the shape of the crystal structure of the metal, but does not rearrange its structure. If the material is then deformed in the martensite phase, it will regain its original shape when heated to the austenite phase. If this process is repeated, the alloy will remember both its original and deformed shape, and can cycle between them. Note that significant hysteresis (about 20 degrees) exists between the transition curves travelling in the two directions.

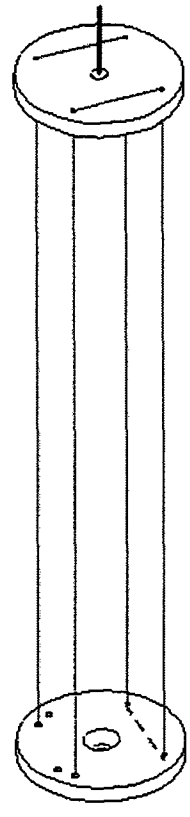


Figure 6: SMA Arrangement

To actuate the surface of the bed, one possible actuator choice is shape memory alloy fibers. The phase transition causes these wires to contract. These wires are very thin, about 6-8 thousands of an inch, but are capable of producing large forces, on the order of 5-10 Newtons. Because of their small size, these wires can be woven tightly together throughout the bed. To produce large forces, the fibers are woven in a parallel configuration, but are connected electrically in series. This causes the forces in the wires to add constructively, and the electrical series connection allows the resistances to add, reducing current requirements. See the appendix for a summary of the model used to describe shape memory behavior.

3.2.3 Issues

Although many challenges are associated with them, shape memory alloys fibers are representative of the kinds of actuators required for this application. They have a very high force density, which is required to get the density of nodes for smooth motion of the surface. Also, they are inexpensive when compared with other actuators, such as cylinders or motors. This is of great benefit when thousands of actuators are required. However, they also have many disadvantages. They are very inefficient, about 1-2%.

This is mainly because the energy used to contract the wire is lost when the wire cools and expands. A full scale bed with thousands of actuators running at full capacity would generate several kilowatts of heat. Also, the small percentage stroke of SMA fibers require that very long wires be used to obtain the stroke required. Knowing these disadvantages, we must still design with SMA fibers, since they are the only technology currently available which can meet our requirements. One possible replacement actuator could be conductive polymers, currently being researched at MIT.

4. Trajectory Design

4.1 General Considerations

The shape of the trajectory of the nodes in the surface wave bed is the key to making the technology work. They must be designed to offer a smooth ride, without bumps and jerky stops and starts. They must minimize shear on the human due to velocity differentials between the nodes. They should maximize the velocity obtainable with the given shear and wave period.

The key to a smooth ride is a continuous motion of the nodes. Since the nodes must recirculate, it is critical that the nodes reconnecting and disconnecting with the body do so at the velocity of the nodes already in contact. This condition will also reduce shear forces on the patients skin.

The following analysis is for two dimensional trajectories. Generally, each node will have a two dimensional trajectory, as its goal is to move the user in a straight line. Multiple actuators will often be coordinated to produce the trajectory. Three possible wave shapes are described below, and their advantages and disadvantages are discussed.

4.2 Square Trajectory

The first is a square wave pattern. The advantage of this shape is its simplicity. Pressure data can be used to determine separations and reconnections. This shape is very easy to generate when the x, y, and z axes are decoupled. However, it does not provide a smooth ride to the user, as the separations and reconnections occur at zero velocity; therefore, acceleration and deceleration must occur every wave period. The filled dot represents the

starting and rest configuration, and also defines the origin of the trajectory's coordinate system. x_s and z_s represent the strokes in the x and z axes, respectively.

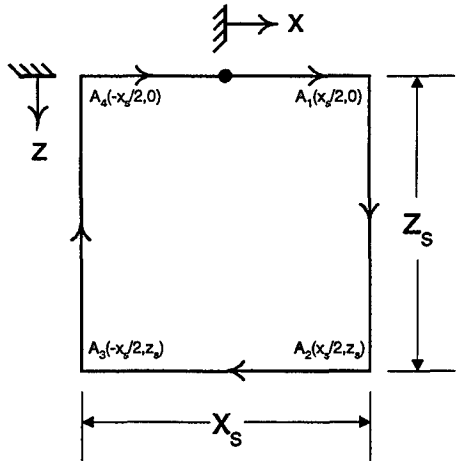


Figure 7: Square Trajectory Geometry

As shown in Figure 7, the coordinates of the corners of the square are $A_1(x_s/2,0)$, $A_2(x_s/2,z_s)$, $A_1(-x_s/2,z_s)$, $A_2(-x_s,/2,0)$. If we assume a constant velocity along each edge of the trajectory, the time velocity profile shown in Figure 8 can be established. The velocity v_{xB} represents the velocity of the body on the bed surface. v_{zd} and v_{zc} represent the speed of the z axis while disconnecting and reconnecting with the body. v_{xR} is the velocity of the x axis as the node returns while detached from the body.

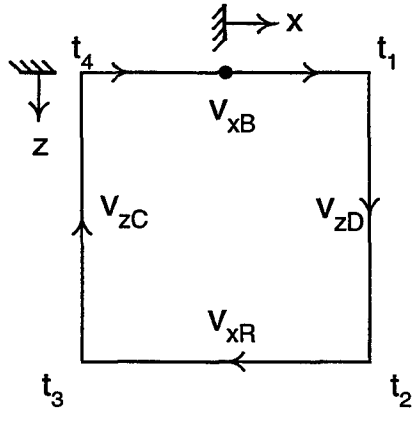


Figure 8: Square Trajectory Velocity Profile

The following equations can be developed for the times $t_1 - t_4$, assuming that t = 0 at the starting point:

$$t_1 = \frac{x_s}{2v_{xB}}$$

$$t_2 = t_1 + \frac{z_s}{v_{zD}}$$

$$t_3 = t_2 + \frac{x_s}{v_{xR}}$$

$$t_4 = t_3 + \frac{z_s}{v_{zC}}$$

The period T of the wave will be:

$$T = t_4 + \frac{x_s}{2v_{xB}} = \frac{x_s}{v_{xB}} + \frac{z_s}{v_{zD}} + \frac{x_s}{v_{xR}} + \frac{z_s}{v_{zC}}$$

If we then divide all the times by T, and multiply by 2□ radians, we obtain expressions for each point in radians:

$$\alpha_1 = 2\pi\left(\frac{\frac{1}{2}x_s v_{zD} v_{xR} v_{zC}}{x_s v_{zD} v_{xR} v_{zC} + z_s v_{xB} v_{xR} v_{zC} + x_s v_{xB} v_{zD} v_{zC} + z_s v_{xB} v_{zD} v_{xR}}\right)$$

$$\alpha_2 = 2\pi\left(\frac{\frac{1}{2}x_s v_{zD} v_{xR} v_{zC} + z_s v_{xB} v_{xR} v_{zC}}{x_s v_{zD} v_{xR} v_{zC} + z_s v_{xB} v_{xR} v_{zC} + x_s v_{xB} v_{zD} v_{zC} + z_s v_{xB} v_{zD} v_{xR}}\right)$$

$$\alpha_3 = 2\pi\left(\frac{\frac{1}{2}x_s v_{zD} v_{xR} v_{zC} + z_s v_{xB} v_{xR} v_{zC} + x_s v_{xB} v_{zD} v_{zC}}{x_s v_{zD} v_{xR} v_{zC} + z_s v_{xB} v_{xR} v_{zC} + x_s v_{xB} v_{zD} v_{zC} + z_s v_{xB} v_{zD} v_{xR}}\right)$$

$$\alpha_4 = 2\pi\left(\frac{\frac{1}{2}x_s v_{zD} v_{xR} v_{zC} + z_s v_{xB} v_{xR} v_{zC} + x_s v_{xB} v_{zD} v_{zC} + z_s v_{xB} v_{zD} v_{xR}}{x_s v_{zD} v_{xR} v_{zC} + z_s v_{xB} v_{xR} v_{zC} + x_s v_{xB} v_{zD} v_{zC} + z_s v_{xB} v_{zD} v_{xR}}\right)$$

Note that □$_1$ corresponds to t_1, □$_2$ to t_2, etc. If we have an angular frequency □ = 2□/T, we can establish a parametric function of position for the x and z directions:

$$x(\omega t)=\begin{cases}\frac{x_s}{2\alpha_1}\omega t & \{\omega t<\alpha_1\}\\ x_s/2 & \{\alpha_1\le\omega t<\alpha_2\}\\ x_s\left(\frac{1}{2}-\frac{\omega t-\alpha_2}{\alpha_3-\alpha_2}\right) & \{\alpha_2\le\omega t<\alpha_3\}\\ -x_s/2 & \{\alpha_3\le\omega t<\alpha_4\}\\ \frac{x_s}{2}\frac{(\omega t-2\pi)}{2\pi-\alpha_4} & \{\alpha_4\le\omega t<2\pi\}\end{cases}$$

$$z(\omega t)=\begin{cases}0 & \{\omega t<\alpha_1\}\\ z_s\left(\frac{\omega t-\alpha_1}{\alpha_2-\alpha_1}\right) & \{\alpha_1\le\omega t<\alpha_2\}\\ z_s & \{\alpha_2\le\omega t<\alpha_3\}\\ z_s\left(\frac{\alpha_4-\omega t}{\alpha_4-\alpha_3}\right) & \{\alpha_3\le\omega t<\alpha_4\}\\ 0 & \{\alpha_4\le\omega t<2\pi\}\end{cases}$$

4.3 Circular Trajectory

The next possible wave shape is a circular one.

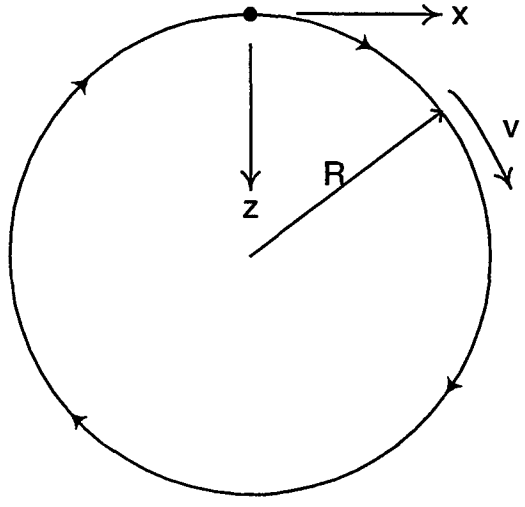


Figure 9: Circular Trajectory

R represents the radius of the trajectory, v represents the velocity of the node, which will be constant at:

$$v = \omega R$$

The velocity experienced by the body in contact will be:

$$v_b = \omega R\cos(\omega t+\phi_k)\mathbf{i} + \omega R\sin(\omega t+\phi_k)\mathbf{j}$$

where ω is the angular frequency of the wave, and ϕ_k is the phase shift of this node. Note that as the body is transferred from one set of nodes to another with a different phase, there will be a discontinuity in the velocity experienced by the body.

This shape is fairly simple to produce with actuators placed in a tripod arrangement. The advantage is that it is fairly continuous in the x and y directions; as the nodes move to reconnect, they develop a forward velocity, and as they disconnect, they lose that velocity, so that the two could be coordinated to be equal. However, the motion in the Z direction, equal in amplitude to the radius R, would be disturbing to the patient. Also, this trajectory would be difficult to generate with decoupled axes without precise feedback.

4.4 Trapezoidal Trajectory

The third choice we suggest would be a trapezoidal shape with rounded corners. This combines the advantages of the square and circular trajectories, but is difficult to generate. It requires careful regulation of the x/y velocities, as well as the z position. Figure 10 outlines the key vertices of this trajectory.

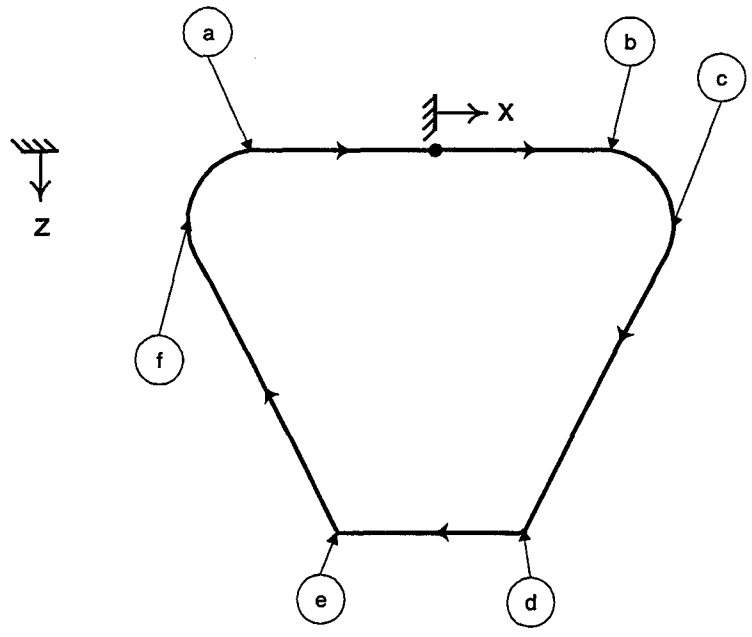


Figure 10: Proposed Trajectory

Starting at (a), the node reconnects with the body. At least one other set of nodes is still in contact at this point, and the data map of pressure is taken for the task level controller. The node is moving at velocity v_{xB}. It moves to (b), where it begins to disconnect. At (c), the pressure sensor connected to the node tells it that it has disconnected fully, and the node begins to move towards the beginning of the stroke. At (d), the node has fully retracted, and is at velocity v_{xR}.

At point (e), the node begins to reconnect. The placement of this point in the trajectory is based on a conservative estimate of how long it will take the node to cool to reconnect with the body. If the estimate is short, the node will be moving in the wrong direction when it reconnects with the body. If the estimate is too long, much of the node's stroke will be wasted before the node ever reconnects with the body. Therefore, an accurate estimate is crucial for proper operation of the device.

The data for the estimate is taken just after one of the sets of nodes has disconnected from the body. At that time, the force exerted on each contacting node is known. Since the spring constant of each node is also known, the displacement of each node can be calculated. This information is used in turn to determine how much time is required to allow the node to cool and reconnect with the body. Depending on the accuracy of the

estimate, the node will begin at (f) to reverse direction in preparation for the reconnection after 75-90% of the cooling time has passed.

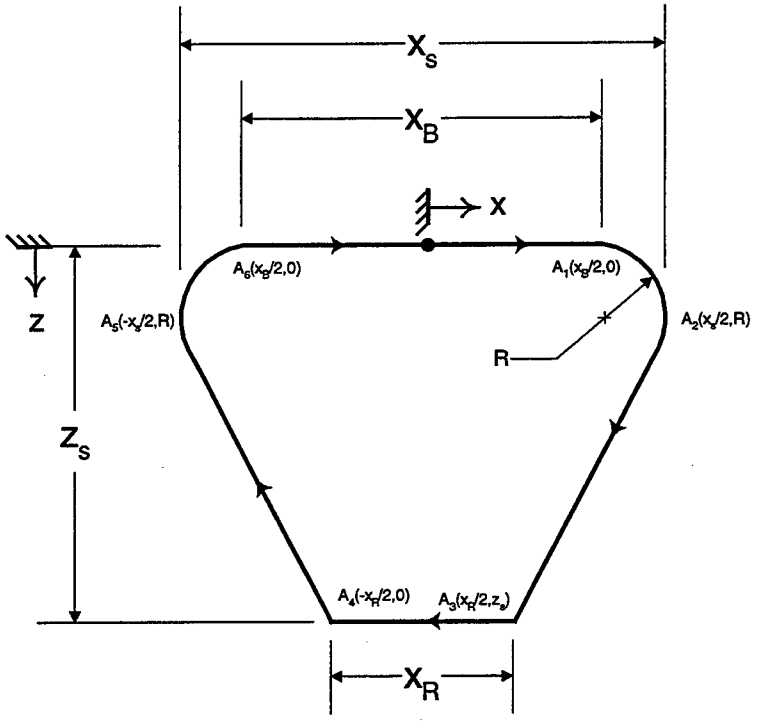


Figure 11: Trapezoidal Trajectory Geometry

The geometry of this proposed trajectory is shown above in Figure 11. x_s is the maximum stroke of the system in the horizontal direction, x_B is the stroke of the actuator while in contact with the body, x_R is the distance traveled in the return direction without z-axis motion, and z_s is the maximum z-axis stroke. There are six points of interest as shown above; their coordinates are $A_1(x_B/2,0)$, $A_2(x_s/2,R)$, $A_3(x_R/2,z_s)$, $A_4(-x_R/2,z_s)$, $A_5(-x_s/2,R)$, and $A_6(-x_B/2,0)$.

In Figure 12, we have added the velocities at each of the edges of the trajectories, and indicated the time coordinates at each significant vertex. v_{xB} is the velocity of the body riding the bed, v_{zD} is the velocity of the node as it disconnects from the body, v_{xR} is the velocity of the node as it returns to begin another stroke, and v_{zC} is the velocity of the node as it reconnects with the body. Using this information, we can derive the following equations for the time coordinates:

$$t_1 = \frac{x_B}{2v_{xB}}$$

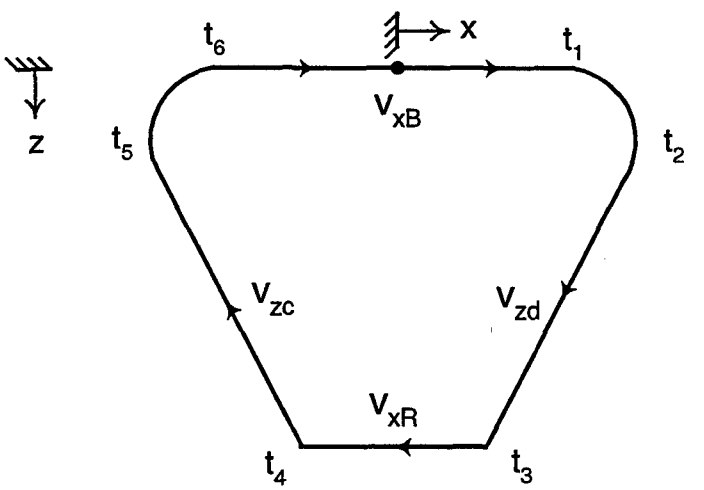


Figure 12: Velocity Profile of Trapezoidal Trajectory

$$t_2 = t_1 + \frac{R}{v_{xB}}$$

$$t_3 = t_2 + \frac{z_s - R}{v_{zD}}$$

$$t_4 = t_3 + \frac{x_R}{v_{xR}}$$

$$t_5 = t_4 + \frac{z_s - R}{v_{zC}}$$

$$t_6 = t_5 + \frac{R}{v_{xB}}$$

The period of the wave shape will then be:

$$T = \frac{x_s v_{zD} v_{xR} v_{zC} + (z_s - R) v_{xB} v_{xR} v_{zC} + x_R v_{xB} v_{zD} v_{zC} + (z_s - R) v_{xB} v_{zD} v_{xR}}{v_{xB} v_{zD} v_{xR} v_{zC}}$$

These equations are based on the assumption of instantaneous accelerations of the nodes, which is almost realizable with sufficient current through a shape memory alloy fiber. Velocity is therefore constant along all edges, except for the z velocity through the rounded corners, which will be controlled so as to gradually shift support of the from one set of nodes to another.

$$x(t)=\begin{cases}\dfrac{x_B}{2t_1}t & t\le t_1\\ \dfrac{x_B}{2}+R\sin\left(\dfrac{3\pi}{4}\dfrac{t-t_1}{t_2-t_1}\right) & t_1<t\le t_2\\ \dfrac{x_B}{2}+R\sin\left(\dfrac{3\pi}{4}\right)-\left(\dfrac{x_B}{2}+R\sin\left(\dfrac{3\pi}{4}\right)-\dfrac{x_R}{2}\right)\dfrac{t-t_2}{t_3-t_2} & t_2<t\le t_3\\ x_R\left(1/2-\dfrac{t-t_3}{t_4-t_3}\right) & t_3<t\le t_4\\ -\dfrac{x_R}{2}-\left(\dfrac{x_B}{2}+R\sin\left(\dfrac{3\pi}{4}\right)-\dfrac{x_R}{2}\right)\dfrac{t-t_4}{t_5-t_4} & t_4<t\le t_5\\ -\dfrac{x_B}{2}-R\sin\left(\dfrac{3\pi}{4}\dfrac{t_6-t}{t_6-t_5}\right) & t_5<t\le t_6\\ -\dfrac{x_B}{2}\left(\dfrac{T-t}{T-t_6}\right) & t_6<t\le T\end{cases}$$

$$z(t)=\begin{cases}0 & t\le t_1\\ R\left(1-\cos\left(\dfrac{3\pi}{4}\dfrac{t-t_1}{t_2-t_1}\right)\right) & t_1<t\le t_2\\ R\left(1-\cos\left(\dfrac{3\pi}{4}\right)\right)+\left(z_s-R\left(1-\cos\left(\dfrac{3\pi}{4}\right)\right)\right)\dfrac{t-t_2}{t_3-t_2} & t_2<t\le t_3\\ z_s & t_3<t\le t_4\\ z_s-\left(z_s-R\left(1-\cos\left(\dfrac{3\pi}{4}\right)\right)\right)\dfrac{t-t_4}{t_5-t_4} & t_4<t\le t_5\\ R\left(1-\cos\left(\dfrac{3\pi}{4}\dfrac{t_6-t}{t_6-t_5}\right)\right) & t_5<t\le t_6\\ 0 & t_6<t\le T\end{cases}$$

5. Wave Shape Design

5.1 General Considerations

In addition to trajectory design, the design of the surface wave shape must be considered. The main design parameter is the phase spacing between the nodes, □. This determines the wavelength of the wave shape:

$$\lambda=\frac{2\pi}{\phi}l$$

where l is the distance between nodes at rest. We have two main considerations for design. The first is contact area; how many nodes are in contact with the body at a given time? Let us define a contact ratio, □:

$$\alpha = \frac{a}{N}$$

where a is the number of nodes in contact with the body, and N is the number of nodes that would be in contact if the system were at rest and none of the nodes were retracted. □ will be greater than zero and less than or equal to one. An □ = 0 would correspond to no nodes in contact, an obvious impossibility. □□□□□□corresponds to all nodes in contact. This will occur at full stop, and for brief moments in a system with □□= □.

5.2 Circular Trajectory Wave Shape Design

For circular trajectories, contact time is dependant on phase. With very low phase between nodes, the body will have approximate point contact with the nodes, and contact with each node will be brief. With large phase shift, the body will spend some time riding the wave, and considerable vertical motion will result. The lowest point will be the transfer between the rising wave and the declining wave:

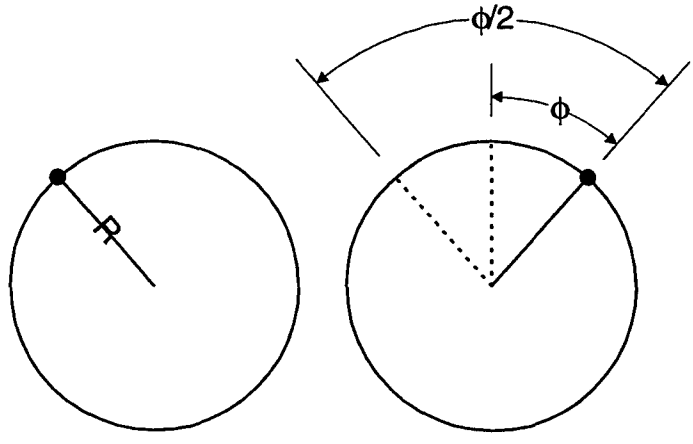


Figure 13: Circular Wave Shape Amplitude Calculation

The amplitude of vertical motion, A, will then be:

$$A = (1 - \cos(\phi/2))R$$

The contact ratio will be:

$$\alpha = \frac{\phi}{2\pi}$$

The other consideration is the velocity of the body. The velocity expression for circular trajectories was given in section 4.3.

5.3 Straight Line Trajectory Wave Shape Design

For trajectories with a straight-line contact motion, contact time is independent of phase. Since the top of the trajectory is flat, each node will contact the body during that time, regardless of other nodes. If each node is in contact with the body for q radians, than:

$$\alpha = \frac{\theta}{2\pi}$$

with

$$\phi \leq \theta .$$

For flat line trajectories, the velocity of the body will be:

$$v_b = \frac{x_s}{\theta}\omega$$

Note that □ increases with increasing □, while v_b decreases. This trade-off between body velocity and body support is a critical issue for the design of the wave shape.

5.4 Prototype Wave Shape Design

There are several specific issues regarding the implementation of the wave shape for the prototype. The system uses the trapezoidal trajectory described earlier, with eight columns of actuators for the x direction, and eight rows of actuators for the y direction. Each row or column is out of phase with its adjacent neighbors by 180 degrees. This results in an □ of 0.5. However, during part of the cycle all of the nodes are in contact, to make a measurement of the body's shape and position using the pressure sensors. The contact time for each node is slightly more than half of its cycle; □ ≅ 200°. This overlap is necessary that one set of nodes makes proper stable contact before the current set in contact detaches.

The z-axis does not have direct position feedback. Therefore, careful management of current is required to ensure that the node reconnects at the proper time. The node in a set must reconnect simultaneously for efficient operation. To this end, a calibration current is used to keep the nodes which cool and contract quickly slightly warm, slowing their contraction until it equals the slowest node. This current is determined experimentally.

Once all of the nodes in the set have detached, the set begins to return to the beginning to start another contact run. The z-axis continues to contract for a set amount of time, and then is allowed to cool. The amount of time required for cooling is dependant on the load placed on the body supporting nodes. A high load on the nearby nodes means that the springs are deflected, and therefore the node will contact the body sooner. The node must be moving at the velocity of the body before it touches it, or there will undesirable shear forces created. If the node begins moving prematurely, some horizontal stroke will be wasted, reducing system speed. The load on the nodes in contact is determined by the pressure sensors, and an estimate of the load above the node is interpolated from that data. See Figure 14.

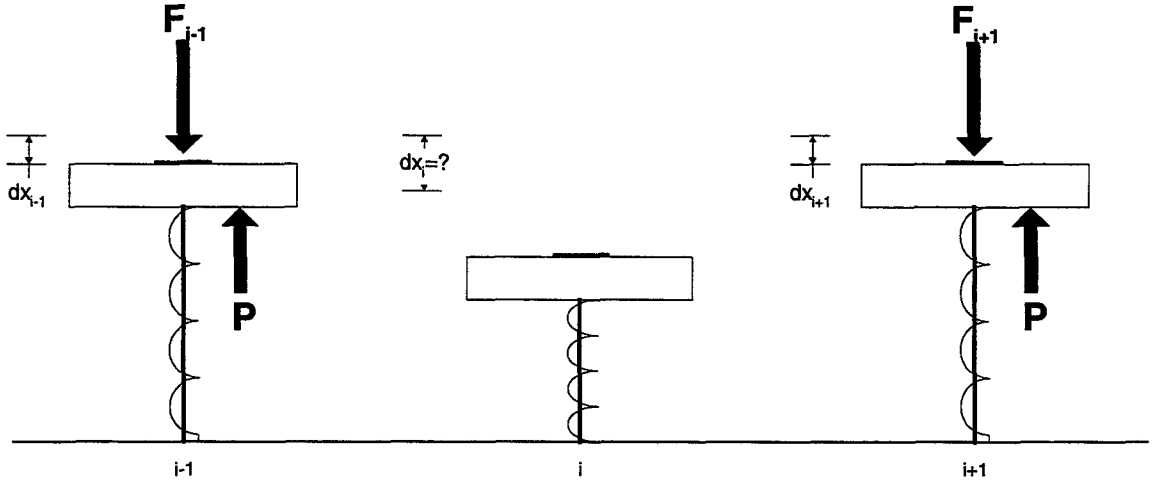


Figure 14: Estimation of Deflection

Here we are considering the i^{th} node. The i-1 and i+1 nodes are out of of phase by 180 degrees. They are in contact with the mass, which exerts a force $\mathbf{F}_{i-1}$ and $\mathbf{F}_{i+1}$ respectively on the nodes. This is balanced by the force in the spring. To keep the SMA fibers in tension when under no load, there is a pre-load in the spring **P**, which is approximately

identical for all nodes. Some force is used to overcome this pre-load, and the remainder of the force compresses the spring. The resulting expression for dx_i, the estimated displacement of the i^{th} node, is then:

$$dx_i = \frac{dx_{i-1} + dx_{i+1}}{2} = \frac{F_{i-1} + F_{i+1} - 2P}{2K}$$

6. System Design

The controller for the surface wave actuator consists of two main parts. The task-level controller determines the current velocity required of the actuators, based on the current position of the body and the desired setpoint. The actuator-level controller interacts with the actuators themselves, feeding them current to obtain the desired displacements.

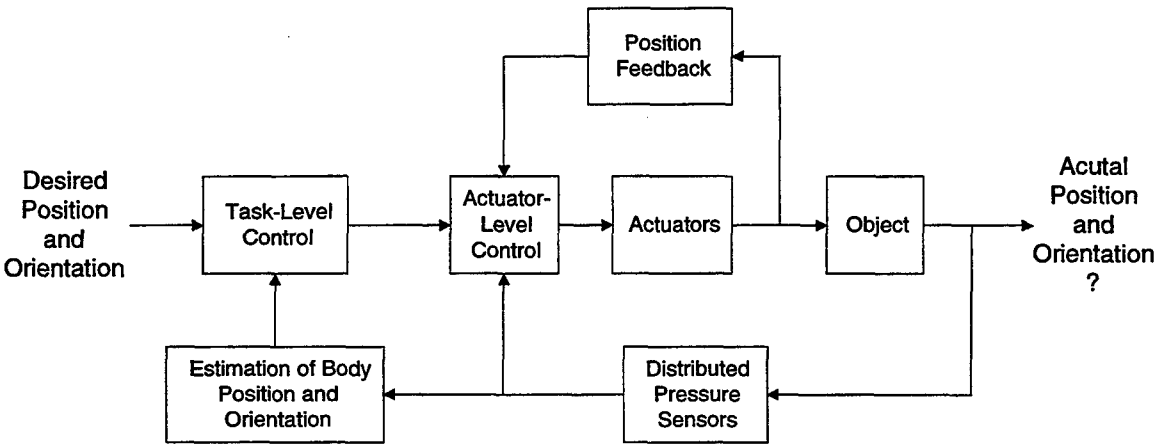


Figure 15: Closed Loop Surface Wave Control

The above figure shows a block diagram for a possible closed loop human position/posture control. Desired body position and posture are input into the upper task-level of the controller. This input could come directly from the user, or could be the automatic output of a program, such as one to combat bed sores. Such a program would generate posture setpoints every hour that require the bed to change the way the patient rests on the bed.

The bed uses distributed pressure sensors to obtain feedback of the body. Since this pressure data does not correspond with the states of position, posture and velocity, and observer must be created to generate this information. Assuming that there are only two

sets of nodes, approximately 180° out of phase, there will be a finite amount of time where all nodes under the body are in contact with it. At that time, the map of pressure data will be fed to the observer, which will extrapolate position, velocity, and posture information for the task-level controller.

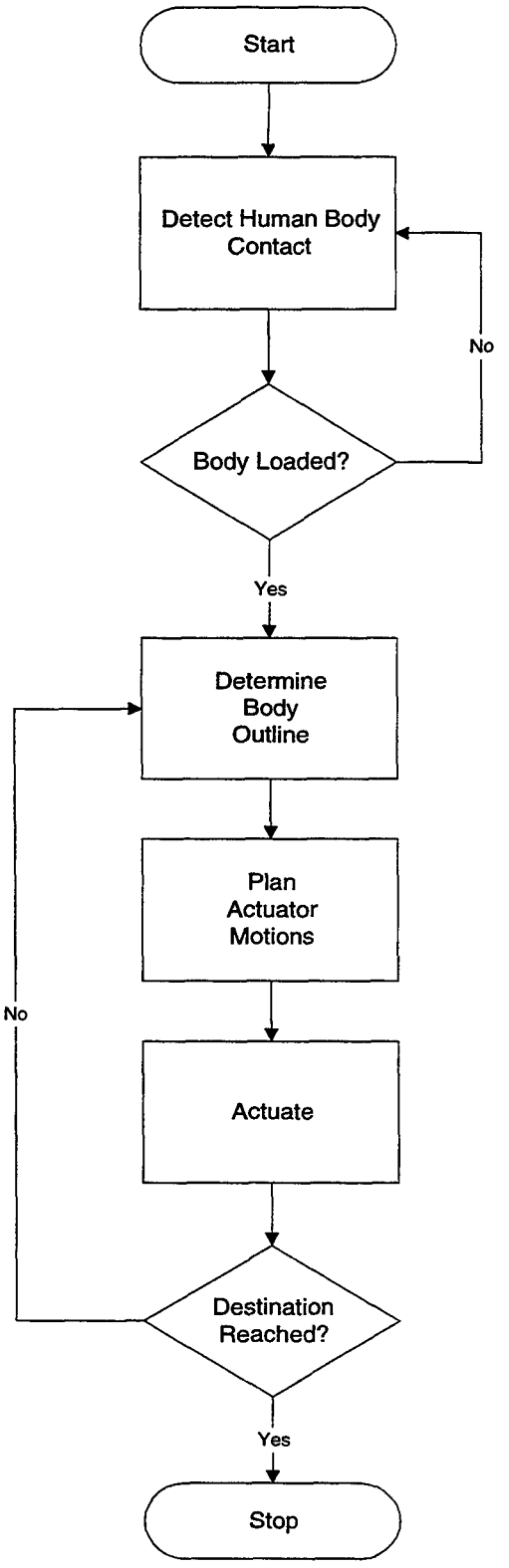


Figure 16: Flowchart of Control Algorithm

Velocity commands are then issued to the actuator-level controllers. These controllers coordinate the actuator motions to generate the required trajectories. Feedback of the actuator displacement is not directly available in an open loop design, which is likely, due to the vast number of actuators. Therefore, it must be inferred from the pressure sensor data. (See Trajectory Design, above.)

In a full bed implementation, the task-level controller would most likely be implemented in the CPU, which also controls the user interface. The actuator controllers will be distributed throughout the bed; each will handle only a few actuators. They would communicate with the main processor over a serial link.

The basic algorithm used to control the device is shown in Figure 16. The bed waits in a low energy consuming idle state for a patient to be placed on the bed. Once this has occurred, the bed makes an estimate of the body contour of the patient and his or her mass centroid, using pressure sensor data. It uses this information to

determine the position and orientation of the patient. If a desired position and orientation is fed to the bed, the actuators are engaged to gently move the patient toward this position and orientation. Once every surface wave period, another snapshot of the patients position and orientation are taken. This information is used as feedback for the actuators. When the body has been moved to the desired location, the bed returns to the rest state and awaits a new command.

7. Prototype Implementation

We have built a functional mockup table with 32 nodes. Each node consists of a spring, extracted from a commercially available mattress. An SMA actuator attached to it supplies motion in the Z-direction (orthogonal to the plane of the patient), while X and Y motion is supplied using DC servo motors. Each spring is initially compressed by tension in the SMA fiber in the z-direction.

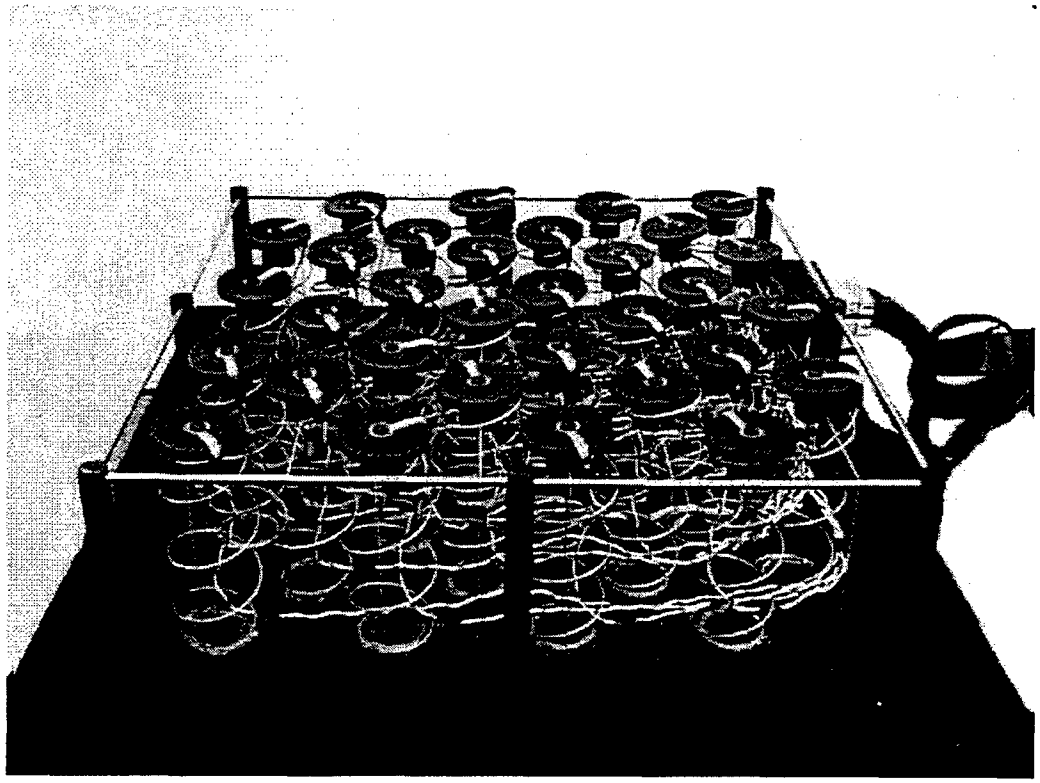


Figure 17: Top Level of Surface Wave Actuator

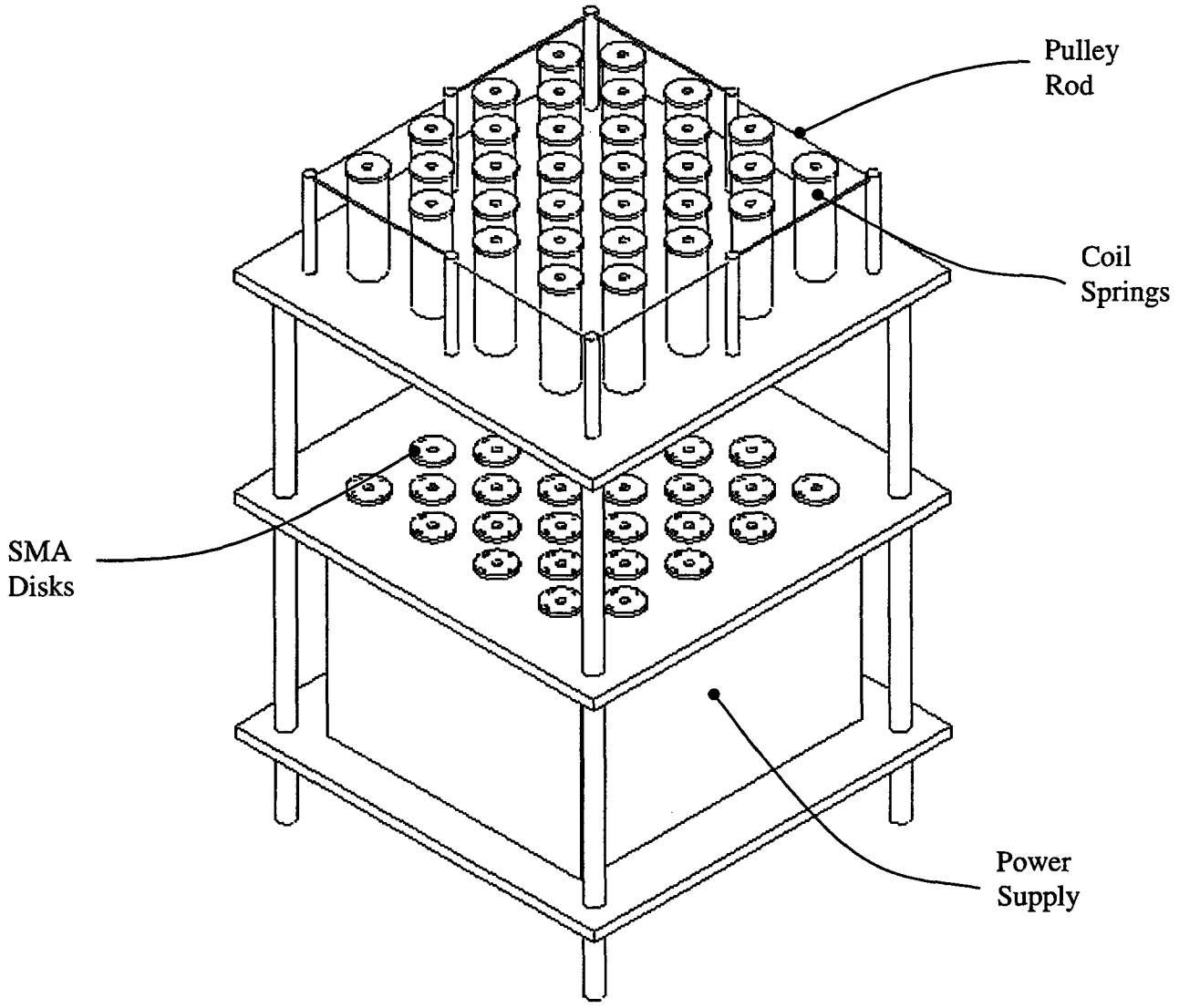


Figure 18: Bed Prototype

Figure 18 illustrates the layout of the prototype design. The table consists of three shelves. The upper shelf is used to support the coil springs and their payload. The central shelf serves as an attachment point for the SMA fibers, and the circuit boards will be mounted here. The space between the upper and central shelves will be strung with the shape memory alloy wires, and cooled by the compressed air flow. The cables to pull the springs will run through the upper shelf and attach to disks mounted to the springs. The cables for horizontal motion will first run over pulleys attached to rods which flank the upper shelf. The lower shelf will support the power supply and the PC used to control the system.

A major issue regarding this design is the lack of position feedback in the z-axis. Without this information, it is difficult to ensure that the nodes will contact the body simultaneously. If they contact prematurely, they may not be moving at the correct horizontal velocity, and will cause shear forces and friction with the human's skin. If they contact too late, some of the horizontal stroke needed to move the body will have been wasted. To ensure that the cooling occurs at an equal rate for all actuators, equalization current may be employed. This small current reduces the cooling rate of the actuator, and is used in those nodes which tend to cool faster than others.

As can be seen from Figure 19, the x axis motion (and eventually y axis motion) is provided by a pair of DC motors, each connected to an axis and four pulleys. Two motors are needed to provide 180° of phase. A cable, connected to four nodes, is wound around each pulley. PD control is used for both position and velocity control.

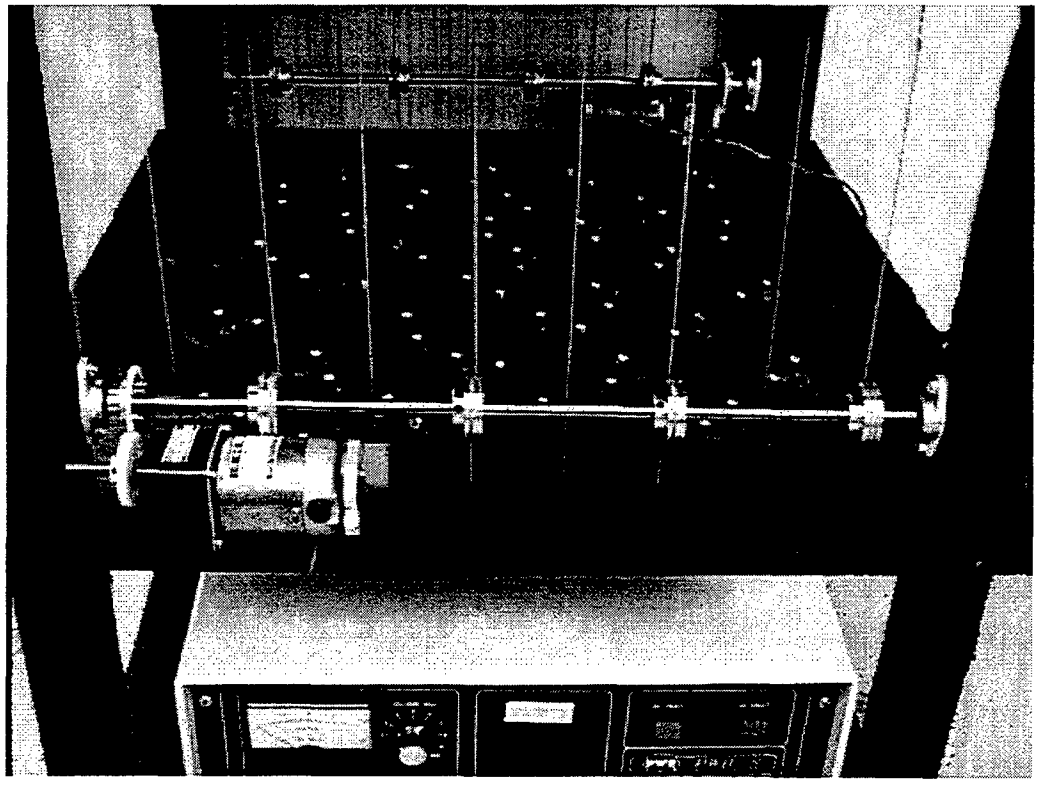


Figure 19: DC Motor Setup

Figure 20 shows the system block diagram for the prototype. The PC is the central control device in the unit. It provides the interface whereby the user can input the desired position commands, and displays position and pressure data. The PC also serves as the

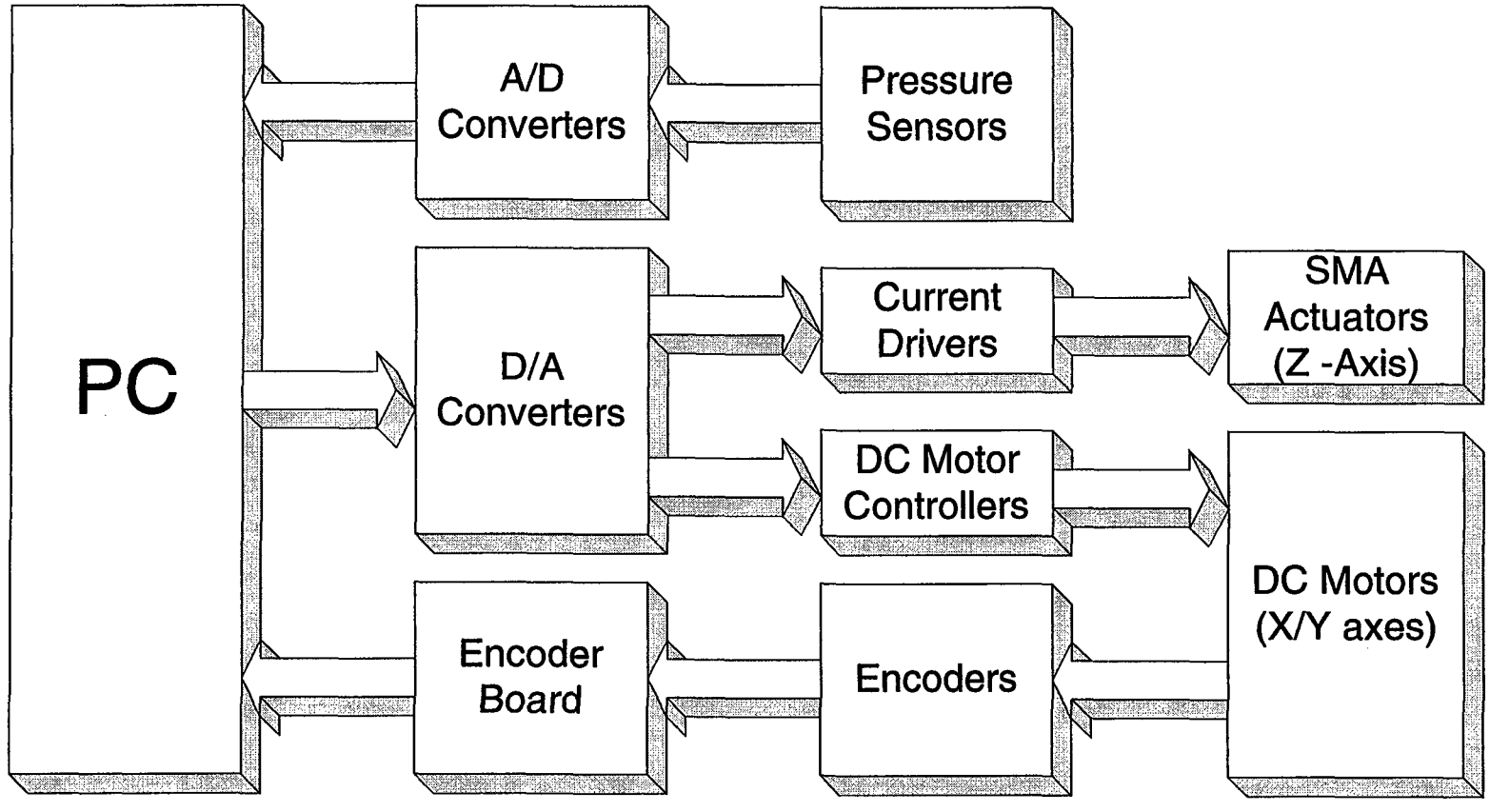


Figure 20: Surface Wave System Block Diagram

motion controller. It receives feedback from the encoders located on the DC motors, and provides the motor controllers with current commands, using PID control. It also controls the 32 shape memory alloy actuators, by providing current commands to the controllers located on the main circuit board. It receives feedback from the 32 pressure sensors located at each node of the bed, through analog to digital converters.

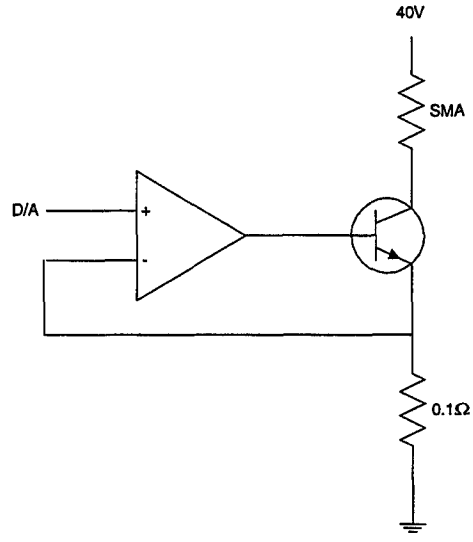


Figure 21: SMA Current Controller

Figure 21 depicts the circuit used for the current controllers for the SMA actuators. They each consist of an operational amplifier, a power transistor, and a feedback resistor of low (0.1 □) resistance. The output of the op amp is connected to the base of the transistor, the positive input is connected to a D/A output from the PC, and the negative output is connected to the feedback resistor. The transistor collector is attached to the SMA fiber and through that the 40V power supply. The emitter is attached to the feedback resistor, and then to ground. When provided a current command from the PC in the form of a voltage, the operational amplifier provides current to the base of the transistor, until the current through the transistor causes the voltage across the feedback resistor to equal the voltage input.

Appendix

A. Modeling of Shape Memory Alloys

Models of the deformation of shape memory alloys with temperature are available from many references. The Brinson model is commonly used and models SMA behavior over a wide temperature range, and is described here.

The key to the Brinson model is the separation of the martensite fraction of the material into temperature and stress induced martensite:

$$\xi = \xi_{S+\xi T}$$

The stress, strain, stress induced martensite fraction, and temperature are related by:

$$\sigma - \sigma_0 = E(\varepsilon - \varepsilon_0) - \varepsilon_L E(\xi_S - \xi_{S0}) + \Theta(T - T_0)$$

where □□□□□□□□, and T are stress, strain, martensite fraction, and temperature respectively. E is Young's Modulous, $\square_L$ is the recoverable strain limit, and □ is the thermoelastic coefficient. Variables with a zero subscript refer to initial conditions.

Refer to Figure 20 in the following equations. Note that M_s is the temperature at which the martensite transition begins, and M_f marks the end of the martensite transition. A_s and A_f mark the start and finish temperatures of the austenite transition. C_M and C_A are constatnts dependant on the material. $\square_s^{cr}$ and $\square_f^{cr}$ are the minimum and maximum stresses allowed for the martensite transition to occur. If the applied stress is below the minimum, the material remains austenite. If above it, the material is damaged during the phase transition.

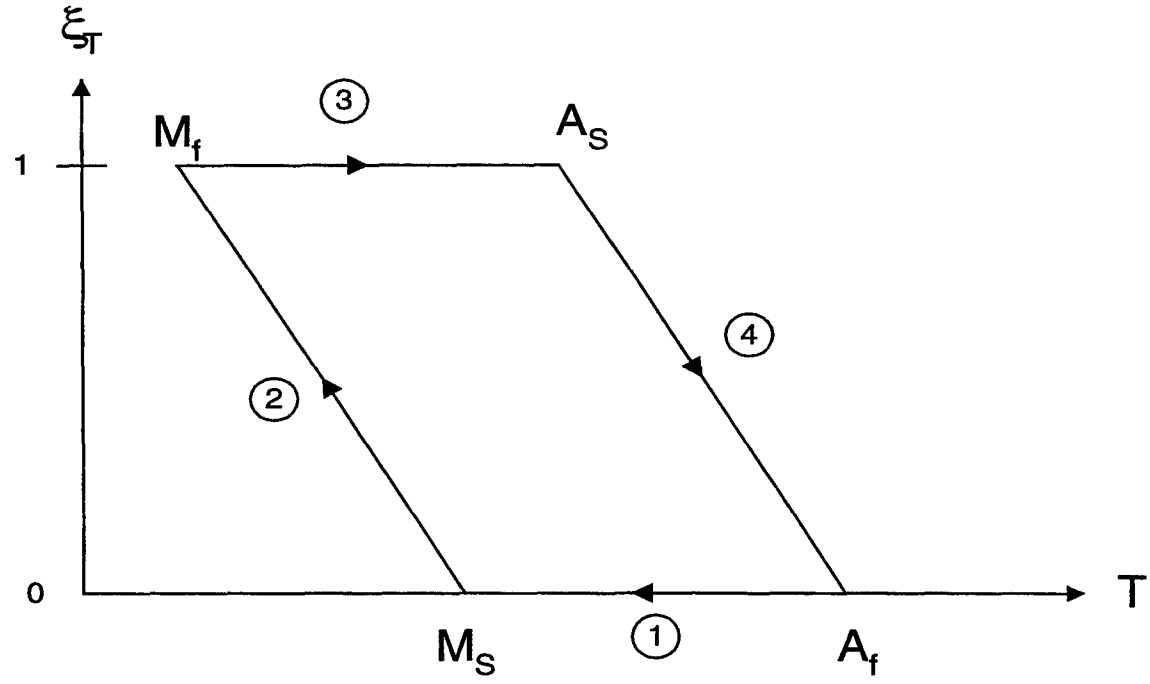


Figure 22: Martensite Fraction as a Hysteretic Function of Temperature

For conversion from austenite to martensite:

For $M_s<T<A_f$ and $\sigma_s^{cr} + C_M[T-M_s] < s < \sigma_f^{cr} + C_M[T-M_s]$, or region 1 above,

$$\begin{cases} \xi_S = \dfrac{1-\xi_{S0}}{2}\cos\left[\dfrac{\pi}{\sigma_s^{cr}-\sigma_f^{cr}}(\sigma-\sigma_f^{cr}-C_M(T-M_s))\right]+\dfrac{1+\xi_{S0}}{2} \\ \xi_T = \xi_{T0}\left(\dfrac{1-\xi_S}{1-\xi_{S0}}\right) \end{cases}$$

For $T<M_s$ and $\sigma_s^{cr}<\sigma<\sigma_f^{cr}$,

$$\begin{cases} \xi_S = \dfrac{1-\xi_{S0}}{2}\cos\left[\dfrac{\pi}{\sigma_S^{cr}-\sigma_f^{cr}}(\sigma-\sigma_f^{cr})\right]+\dfrac{1+\xi_{S0}}{2} \\ \xi_T = \xi_{T0}\left(\dfrac{1-\xi_S}{1-\xi_{S0}}\right)+\Delta_{T\xi} \end{cases}$$

where if $M_f<T<M_s$ and $T<T_0$ (region 2),

$$\Delta_{T\xi} = \frac{1-\xi_{T0}}{2}\left[\cos(a_M(T-M_f))+1\right]$$

else

$$\square_{T\square} = 0.$$

For conversion to austenite:

For $T>A_s$ and $C_A(T-A_f) < \square < C_A(T-A_s)$,

$$\begin{cases} \xi = \frac{\xi_0}{2}\left[\cos(a_A(T - A_s - \frac{\sigma}{C_A})) + 1\right] \\ \xi_S = \frac{\xi_{S0}}{\xi_0}\xi \\ \xi_T = \frac{\xi_{T0}}{\xi_0}\xi \end{cases}$$

where a_A and a_M are the following constants:

$$a_A = \square/(A_f - A_s) \qquad a_M = \square/(M_s - M_f)$$

Young's Modulus can be defined as a function of the martensite fraction:

$$E = E_A + \square(E_M - E_A)$$

where E_A is the modulus of austenite and E_M is that of martensite.

We have developed an expression for the resistance of a length of SMA wire based on its phase composition:

$$R = \frac{l}{A}(\rho_A + (\rho_M - \rho_A)\xi)$$

where l is the length of the wire, A is its cross sectional area, and $\square_A$ and $\square_M$ are the resistivity of austenite and martensite, respectively. This expression can be used to make estimations of the phase, and therefore the strain of the wire, using resistance. The hysteresis in the wire makes this very difficult at high switching frequencies, however.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Human-Machine Interface

CHAPTER 17

Human Machine-Interface and Interactive Control Part 1: Instrumented Nails and Virtual Switch Panels

H. Asada, S. Mascaro, K-W Chang

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Progress Report, March 31, 1998
Total Home Automation and Healthcare Consortium

Human Machine-Interface and Interactive Control

Part 1: Instrumented Nails and Virtual Switch Panels

H. Harry Asada
Professor, Principal Investigator

Stephen Mascaro
Graduate Research Assistant

Kuo-Wei Chang
Senior Lecturer

Abstract

A new type of touch sensor for detecting contact pressure at human fingertips is presented. Fingernails are instrumented with micro LED and photodetectors in order to measure changes in the nail color when the fingers are pressed against a surface. Unlike traditional electronic gloves, in which sensor pads are placed between the fingers and the environment surface, this new sensor allows the fingers to directly contact the environment without being impeded by any object between the finger and the environment. The finger force is detected by measuring changes in the nail color; hence the sensor is mounted on the nail side rather than the finger pad. The technique termed "photoplethysmography" is used for measuring the nail color. All the devices are miniaturized and signals are transmitted through wireless communications channels to enhance comfort of the human wearing the sensors. Using these new touch sensors, a virtual switch panel is proposed, where the switches are images on a surface rather than actual mechanisms. A prototype touch sensor is constructed and demonstrated in the context of the virtual switch panel and human-machine interaction.

d'Arbeloff Laboratory for Information Systems and Technology
MIT

1. Introduction

Electronic gloves have been extensively studied in the past decade in the robotics and virtual reality communities [1][2][3][4][5][6][7]. They all have some means of measuring finger positions to varying degrees. Some of these gloves also collect touch-force data from the human fingers as the human interacts with the environment [8][9]. To measure the forces acting at the fingers, sensor pads consisting of conductive rubber, capacitive sensors, optical detectors, and other such devices can be placed between the fingers and the environment surface. These sensor pads, however, inevitably deteriorate the human haptic sense, since the fingers cannot directly touch the environment surface.

In this paper, a new approach to the detection of finger forces is presented in order to eliminate the impediment for the natural haptic sense. Namely, the finger force is measured without having to place any sensor pad between the finger skin and the environment surface, but is detected by an optical sensor mounted on the fingernail. This allows the human to touch the environment with bare fingers and perform fine, delicate tasks using the full range of haptic sense. Furthermore, unlike the traditional finger touch-force sensors, the new sensor would last for a longer time since it does not contact the environment and has no mechanical parts to wear or become damaged due to mechanical contacts.

This new type of touch sensor opens up many new ideas in the area of human-machine interaction. Not only can they be used to replace the traditional touch sensors used in the human-robot cooperative task described in Part 2 of this report, but they can also be used to create a *virtual* switch panel for human supervisory control. The virtual switch is an image on a surface rather than an actual mechanism. The switch is triggered by the fingernail touch sensor coupled with finger position measurements from a data glove. In this way, the human can interact with computers, machines, and controls without having to affect the environment. The virtual switch panel does not wear or need repair, is not affected by the environment, can share the human's workspace, and can be reconfigured in a "soft" sense.

In the first part of the report, the principle of the new touch sensor is described. The concept of the virtual switch panel is then developed. Finally, a prototype design is demonstrated and applied to the context of the virtual switch panel and human-machine interactive control.

2. Principle of Finger Touch Sensor

2.1 Physiology

A detailed summary of the anatomy of a fingertip is given in Appendix A. A brief description of the physiology of the phenomena relating to the touch sensor is given here.

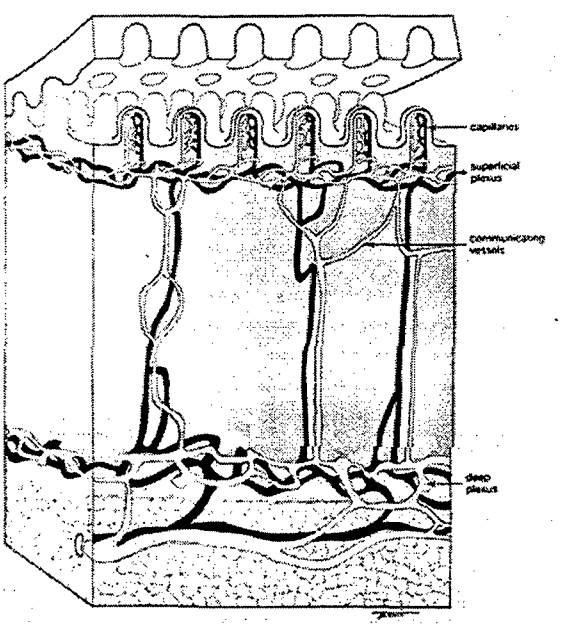


Figure 1: Dermal Vasculature

(From Moschelle, S.L. and Hurly, H.J.: Dermatology , Vol. 1, pp.35, Saunders, 1985)

As a finger is pressed on a surface with increasing force, the fingernail changes color from pale red to more intense red. As shown in Figure 1, blood flow in the capillaries, superficial plexus, communicating vessels, and the deep plexus can be disrupted by mechanical forces applied to the volar surface of the finger. Since the venous pressure is substantially lower than the arterial pressure, the venous return of blood from venules is progressively reduced by contact pressure, causing a reduction in blood supply to the affected area. As a result, digital

blood supply is diverted to poorly suffused area such as the nailbed. The pooling of arterial blood under the fingernail gives rise to a unfading reddish hue due to the perfusion of the underlying capillary loop with blood rich in oxyhemoglobin. As the contact pressure reaches a saturation point when the veins and venules are collapsed and blocked, the intensity of fingernail color stops increasing with further increase in pressure. The color change process is reversible, and as the contact pressure is reduced, the color fades out progressively with no substantial delay.

This phenomenon can be utilized to measure touching force and contact pressure by monitoring changes in fingernail color without having to put a sensor between the finger and surface. The change in color is directly related to the pooling of arterial blood and the resulting change in oxygen saturation. or oxy-hemoglobin saturation (relative concentrations of oxy- and reduced- hemoglobin). The relative concentrations of oxy- and reduced hemoglobin of blood under the fingernail can be monitored by shining a light into the fingernail and measuring the reflected light.

2.2 Photoreflective Measurements

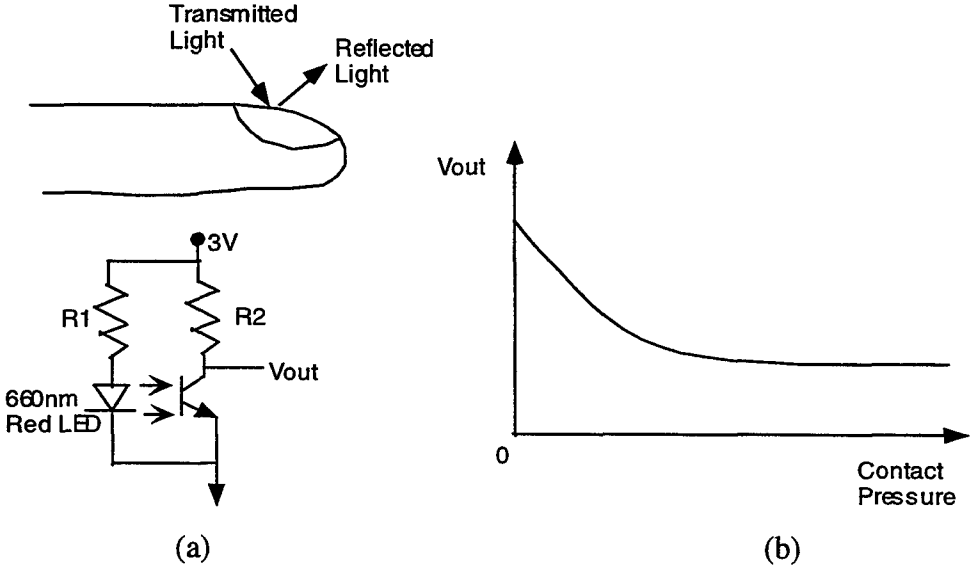


Figure 2: Photoreflective Fingernail Sensor

An example of the experimental setup is shown in Figure 2. In Figure 2a, a red LED at 660nm illuminates the nail bed with a red light. A photo-transistor is mounted on one side of the LED and catches the reflected light from the nail bed. As contact pressure increases, more arterial blood accumulates driving the oxygen saturation higher. As a result, less red light is absorbed, more red light is reflected, the impedance of the photo-transistor drops, and the output voltage, Vout, decreases. Vout reaches an asymptotic value when the veins are collapsed and closed shut.

If an infrared LED at 940nm is used instead, the output voltage increases with the contact pressure (Figure 3a). This is because the trends of the absorption curves for hemoglobin and oxy-hemoglobin reverses after crossing the isobestic point at 800nm (Figure 3b).

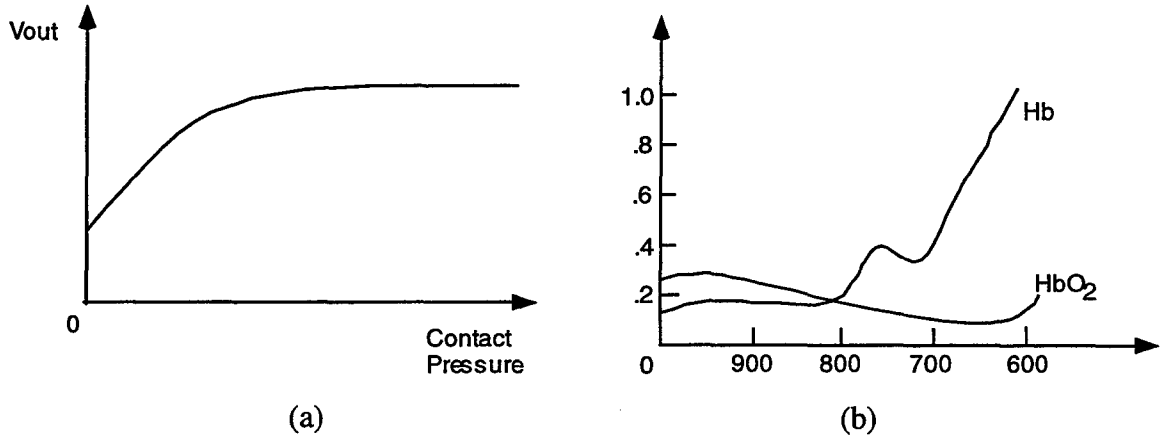


Figure 3: Effects of LED Wavelength

With high contact pressure, the fingernail turns red, the reflected IR light at 940nm is reduced due to higher absorption by the oxy-hemoglobin. The red LED at 660nm and the IR LED at 940nm can be used in the same sensor to result in a bridge type design for enhancement of the sensor sensitivity. The two types of LEDs are illuminated alternately and the reflected lights are measured by the same photo-detector with the aid of sample-and-hold circuitry.

3. Virtual Switches and Hyper Manuals

3.1 Features of the Instrumental Fingernails

The finger touch sensor using photoreflectance provides unique features for the monitoring of human behavior. These include:

- Bare Fingers

  The fingertips are not covered by any object which would impede the natural haptic sense of the human. The proposed method using photo plethysmograph has departed from the traditional force/pressure measurement that entails a sensor pad placed between the finger skin and the environmental surface. The new touch sensor allows the human to contact the environment with bare fingers and perform fine, delicate tasks by fully exploiting the keen haptic sense at the fingers.

- Miniaturized Wearable Sensors

  The proposed finger touch sensor is suited for miniaturization. Most of the key components, including LED and photo detectors, are less then 1mm square in size; they can be mounted even on fingernails. The burden on the wearer is minimum, and the sensors can be worn comfortably for a long period of time.

These features would enable us to develop novel human machine interface and interactive control methods, which would otherwise be impractical and infeasible.

3.2 Concepts of Virtual Switches

In the home as well as work environment, humans are constantly supervising, controlling, and communicating with devices, computers and machines using a multitude of switches. Switches are rudimentary means for the human to communicate his/her intention to machines. The wearable finger touch sensor would replace the traditional switches and enhance the human-machine interface. Figure 4 depicts the functionality of a traditional switch and shows how the wearable finger touch sensor provides the same functionality and replaces the physical switch. As shown in figure 4(a), the traditional switch works by means of:

1) The human movement of his/her finger to physically push some button of the switch.
2) The detection of the human intention by electrical contact in the switch, and
3) Transfer of the detected signal to a specific part of the machine to change its state.

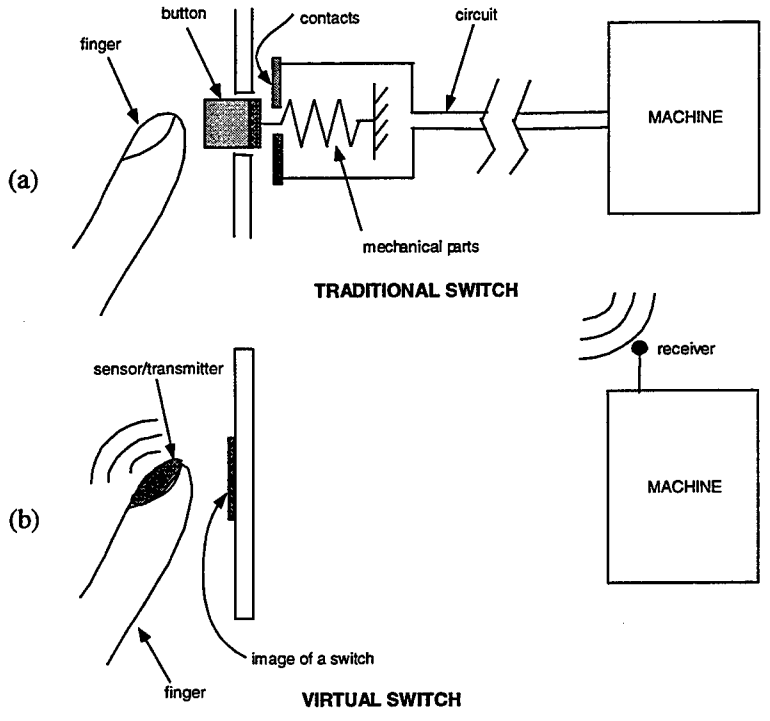


Figure 4: Traditional Switch vs. Virtual Switch

In the traditional switch, the detection of human intention is performed by the device that is attached to the machine and physically connected to a specific part of the machine. Traditional switches have the following limitations:

1. they generally cannot be reconfigured without reinstalling
2. they can be damaged in hazardous environments – chemical, mechanical, etc.
3. they take up space which could be used for some other purpose
4. they can be activated accidentally if bumped with something other than a finger

With the new wearable fingernail touch sensor, the ability now exists to measure human intention through touch without affecting a change on the environment. The detection of human intention can be performed by the device worn by the human rather than the one attached to the

machine. To this end, we propose the idea of a "virtual switch." As shown in Figure 4(b), the virtual switch is not a mechanism, but is an image on a surface that represents a switch. The intention of activating the switch is detected by the fingernail touch sensor, coupled with 3-D finger position measurements from an electronic glove. The signal is transmitted wirelessly from the human to the machine. When a touch is detected on a finger whose position measurement corresponds to a certain virtual switch, that switch is activated. This will eliminate all of the problems listed above and open up new possibilities for human-machine communication.

To summarize, the virtual switch panel offers the following advantages over traditional switches:

- virtual switches do not wear and do not need repairs/replacement
- virtual switches cannot be damaged by hazardous environments – chemical, mechanical, etc.
- virtual switches cannot be activated accidentally if bumped with something other than finger
- virtual switches can be rearranged and reconfigured without reinstalling
- virtual switches can have different functions for different fingers
- virtual switches can share the workspace and do not monopolize a work surface

The concept of the virtual switch panel opens up numerous possibilities for human-machine communication, and can be anything from a simple virtual on-off button to an entire virtual computer keyboard. Virtual switches can be placed at diverse surfaces including

- Walls, like light switches
- Control panel and remote switch box
- Tables, chairs and other furnishing, and
- The body of the machine itself.

3.3 Embodiments of Virtual Switches

Figure 5 shows a sketch of one embodiment of the virtual switch panel. In this scenario, the human is working alongside a robot to accomplish a task. The human is wearing some form of dataglove with open fingertips, which tracks the position of his fingers in 3-D space, and his

fingernails are instrumented with the photoreflective sensors to measure finger touch force. Virtual switches are painted on the surfaces around his workspace as well as the surface of his workspace. Perhaps some switches are even painted on the robot or human himself. Whenever, the fingernail sensor detects a sudden touch force, it relays the signal to the computer or robot controller along with the position of the finger which committed the touch. If the computer recognizes that the position corresponds to a certain virtual switch, then that switch is declared "activated." The function associated with the switch is performed, and the computer provides feedback to the human audibly or otherwise to confirm the activation of the switch. In this way the human can activate the robot, the computer, or other devices in his work area without affecting any change on the environment. Furthermore, the functions of each of the switches can be reprogrammed by the human at any time without having to do any work mechanically. The virtual switches can even take on different functions automatically at different stages of a task, or have different functions depending on which finger activates them. Like a computer mouse with two buttons, different actions can be recognized by using multiple finger touch sensors. Finally, the human can work over top of the virtual switches and use his desk for other tasks without the switches getting in the way.

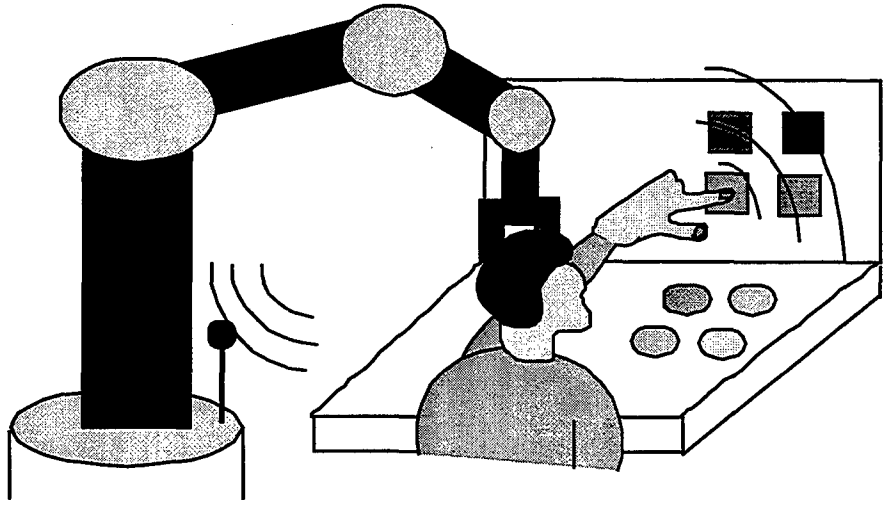


Figure 5: Virtual Switch Panel

Figure 6 shows the next level of this embodiment, which is the "totally virtual switch panel." This embodiment has all the features of the original virtual switch panel, only in this

case the switches are not even painted or drawn on the surfaces of the workspace. Instead, the switches are either projected onto the workspace, or the human wears a head-mounted, heads-up-display, which superimposes computer images of the switches on his view of the workspace. By tracking head motion, the images can be made to appear stationary on a particular surface or move around in a desired fashion. Looking in different places can cause different switch panels to be displayed. Switches can be rearranged and reconfigured completely by software.

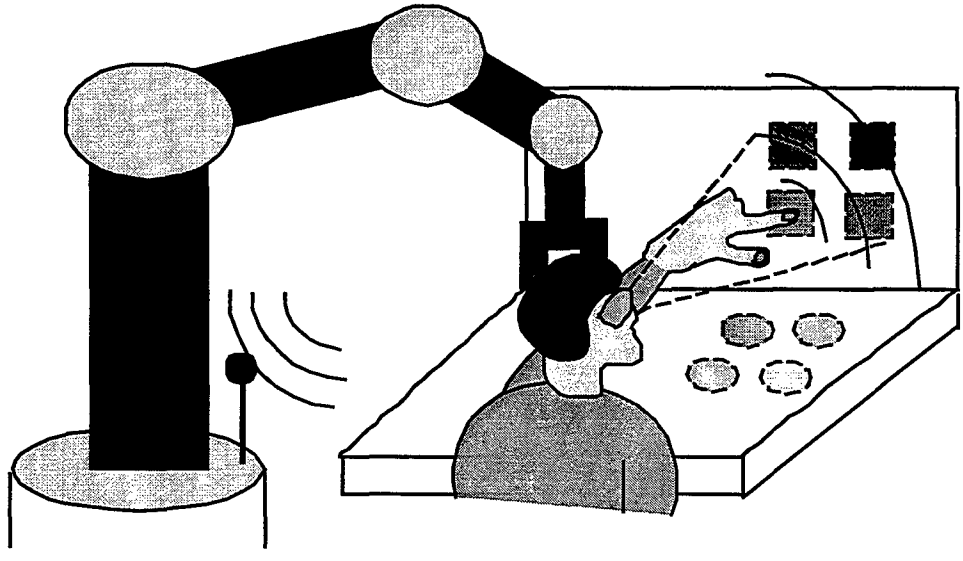


Figure 6: Totally Virtual Switch Panel

Figure 7 illustrates a distributed virtual switch system applied to healthcare and medical equipment. A hybrid bed/chair, e.g. RHOMBUS, is operated by a caregiver wearing the finger touch sensors and hand position sensor. Numerous virtual switches can be imbedded in the bed/chair surface for acquiring the care giver's intention. For example, when the caregiver wishes to raise the back leaf of the bed/chair system, he/she touches the back side of the back leaf and tends to push it upward. The virtual switch imbedded in the back leaf recognizes the human motion by detecting the hand location and the pressure increase at his/her fingers.

The detected signal is then transmitted to the powered bed/chair for activating the actuator raising the back leaf. Let us suppose that, after raising the back leaf, the caregiver wants to push the bed (now reconfigured to a wheelchair) forward. The virtual switch imbedded in the wheelchair handle detects the caregiver's intention, when he/she touches the wheelchair

handle and tends to push it forward. The position sensor recognizes that the caregiver places his/her hand on the handle, and the finger touch sensors detect that the forward button printed on the handle is pressed by his/her fingers.

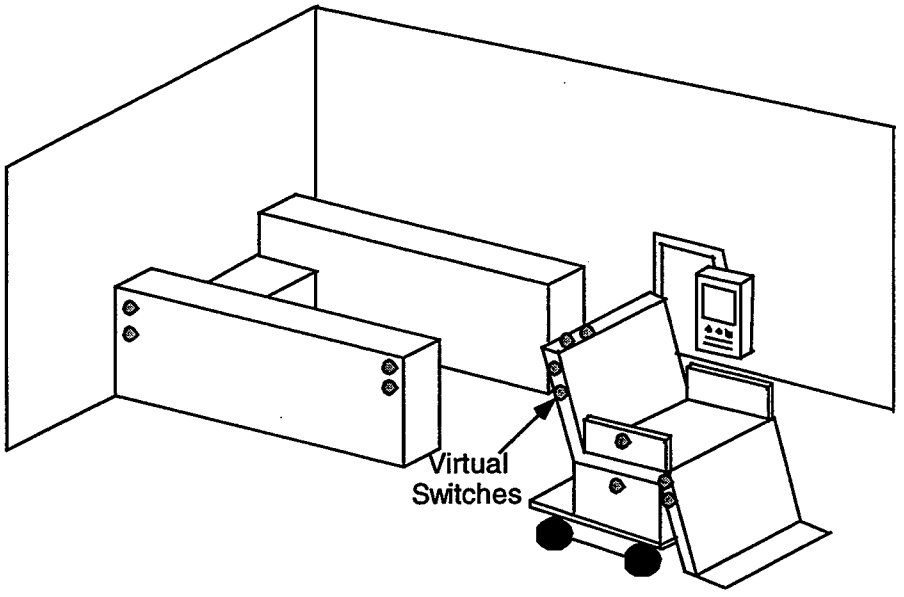


Figure 7: Distributed Virtual Switch System

The assignment of virtual switches to individual portions of the bed/chair system can be changed depending on the context, situation, and stage in the task. For example, virtual switches during the bed-mode operations and the chair-mode operations may be altered by simply changing the "map" relating sensor signals to the control actions. Although pressing the same point of the machine, different actions can be generated. Pressing the back leaf, for example, is recognized as the intention of changing the back leaf angle only when the operation is in the bed mode. Pressing the same back leaf during the wheelchair mode creates no action, thus avoiding erratic operations. This context-dependent assignment is extended to the concept of "wearable digital manuals".

3.3 A Wearable Hyper Manual

In general, a manual described step-by-step instructions for operating a machine to perform a certain task. Typically a human follows a procedure described in a manual, which includes a sequence of operations, usually pressing buttons and knobs. In the past decade, many

manuals have been computerized, i.e. digital manuals, for better service and easier use. The functionality, and features of digital manuals can be furthered by combining wearable sensors such as the finger touch sensors and hand position sensors, which monitor human behavior. Namely, combining virtual switches with a digital manual would create a powerful aid for guiding a user through a complex operational procedure. Figure 8 illustrates such a combined system, called a "wearable hyper manual".

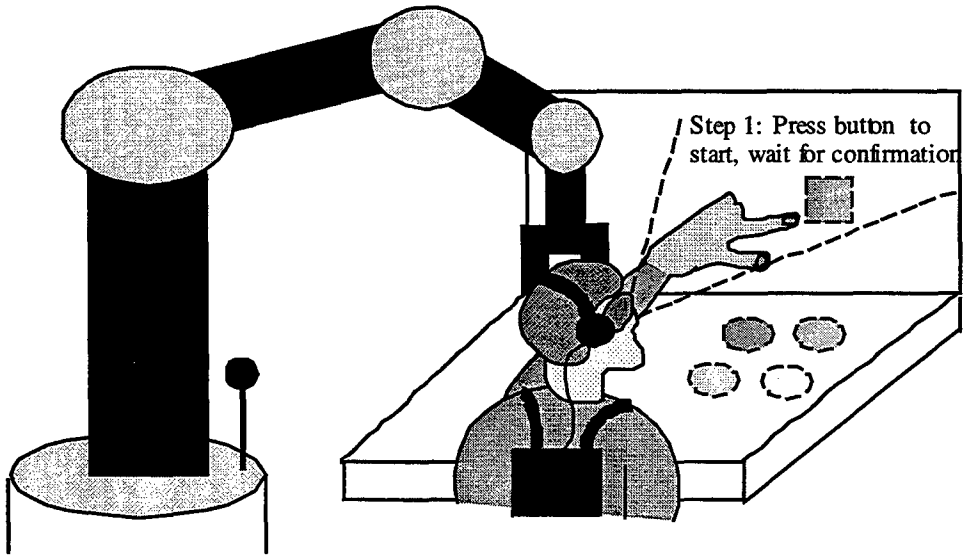


Figure 8: Wearable Hyper Manual

The wearable hyper manual consists of a wearable computer storing manual information, wearable and/or stationary sensors monitoring human behavior, and a display and/or headset for providing instructions to the human. Unlike traditional digital manuals in which the user must be able to retrieve items of information needed for each step of operation, the wearable digital manual system monitors the human behavior, identifies in which stage of procedure the human is currently involved, and provides the right items of instruction needed for that stage. Furthermore, inputs from the human are acquired from the virtual switch panel described above. The virtual switch panel presented to the human would be varied depending on the stage of the procedure and the relevance to the context of the task. The execution of this entire process entails a task programming and process control engine. Such an engine would represent the task, code the procedure, recognize each task stage, observe human behavior, retrieve manual information, present instructions, display control panels, and acquire human inputs to coordinate

a target machine with the human inputs. In the succeeding Part 2 of this report, a method based on Petri nets will be presented as a general method for administrating the human machine interactive process.

In the Wearable Hyper Manual shown in Figure 8, the virtual control panel is projected on the machine surface as well as on the table, and is varied dynamically as the task process proceeds. Moreover, the headset worn by the human provides verbal instructions retrieved from the stored manual data. Both audio and visual instructions would be coordinated with the human monitoring system using wearable sensors and the target machine to operate. Details of the control process will be discussed in Part 2.

4. Implementation

4.1 Prototype Design

A pair of prototype fingernail sensors was constructed according to the principle described in section 2, and incorporated into a free-fingered glove with magnetic tracker. Figure 9 shows the prototype system.

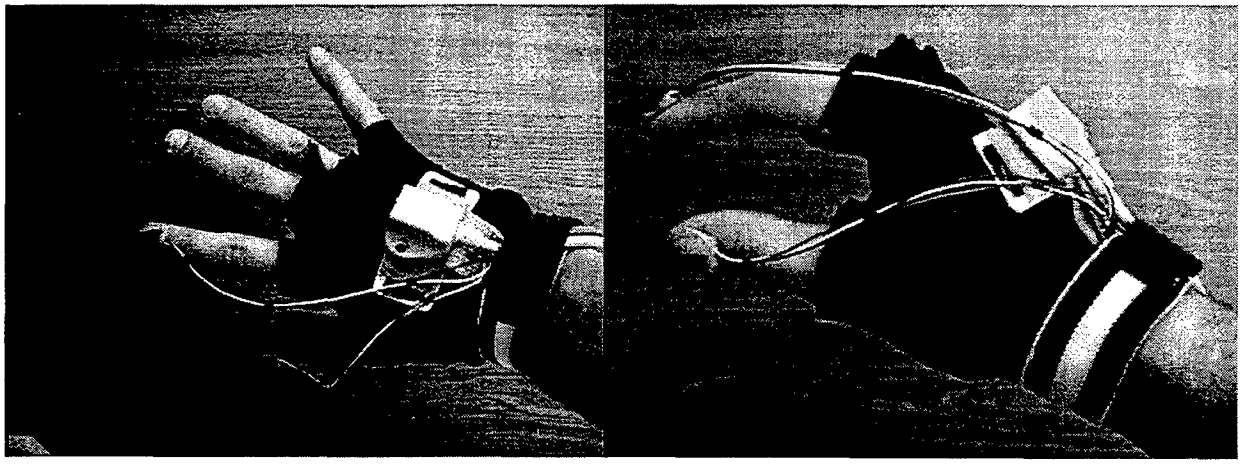


Figure 9: Free-Fingered Glove with Touch Sensing

As shown in Figure 9, the nails of the index finger and thumbs were instrumented with the photoreflective touch sensors. Each prototype touch sensor was fabricated by embedding a

single phototransistor and red LED within a prefabricated plastic fingernail. Such plastic fingernails are widely available in a variety of shapes and sizes at most drug stores, and are satisfactory for prototyping. The plastic fingernails were then attached to the fingernails of the human using a thin strip of sticky-tack around the perimeter of the nail. The sticky-tack conforms to the shape of the fingernails, provides a seal around the perimeter of the nail/sensor against ambient lighting, and allows the sensors to be affixed and detached from the fingernail as desired. In future designs, when the sensors are miniaturized and wireless, a more permanent method of attachment will be used. Figure 10 shows the design of an individual fingernail sensor and Figure 11 shows enlarged views of the actual prototype.

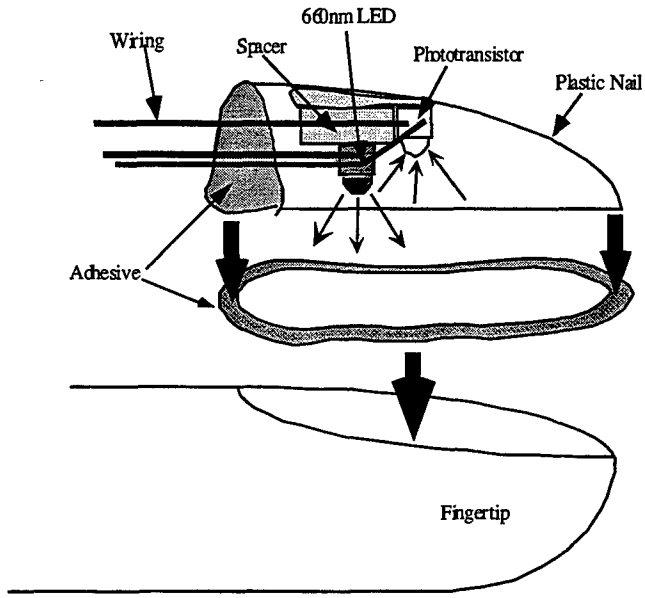


Figure 10: Prototype Design for Fingernail Touch Sensor

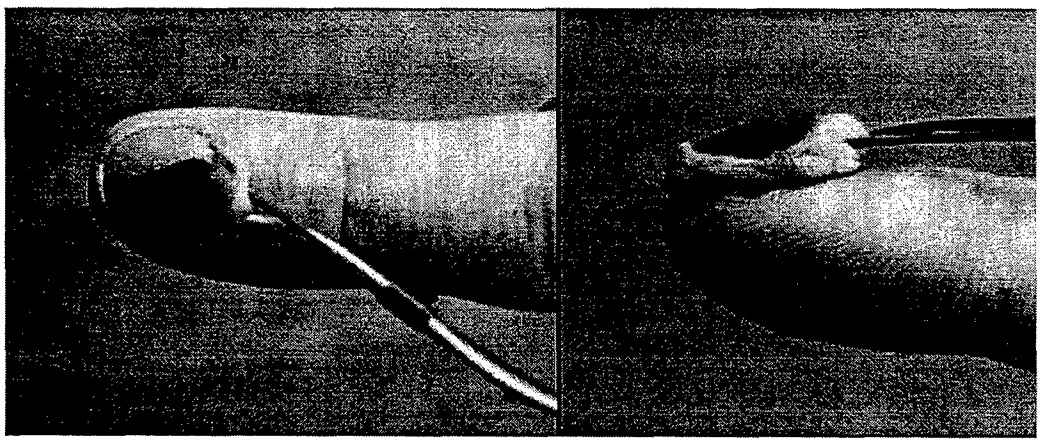


Figure 11: Prototype Fingernail Touch Sensor

4.2 Experiments

Using the prototype fingernail sensor described above, initial experiments were done to investigate the behavior and performance of the sensors. Depending on how force is applied to the fingertip, the light absorption either increases or decreases with increasing force. If the force is applied directly underneath the fingertip, causing the fingernail to redden, the amount of reflected light actually decreases. On the other hand, if the force is applied to the end of the fingertip, causing the fingernail to whiten, the amount of reflected light increases. This suggests that the photoreflective measurements are dominated by volumetric effects. In other words, the reflected light decreases when the nail reddens because of the increased volume of blood under the nail, even though the absorption coefficient of the blood may be decreased. Figure 12 shows a plot of the output of the fingernail sensor versus the output of a contact pressure sensor when the finger is pressed down on top of the contact pressure sensor. The output voltage of the fingernail increases monotonically with force and levels off at a force on the order of 1 Newton.

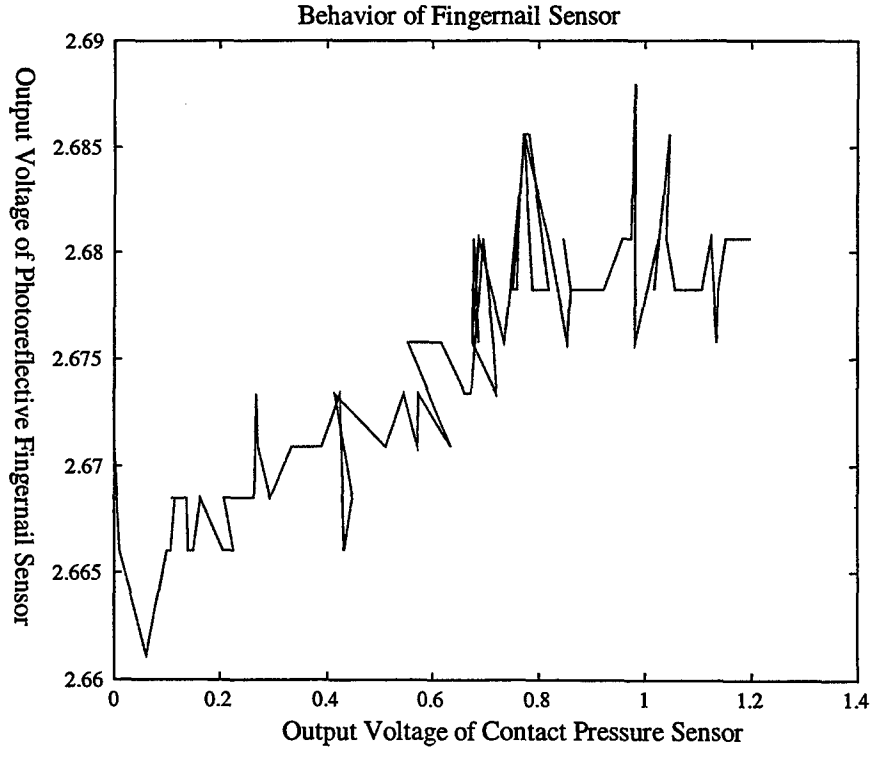


Figure 12: Experimental Data

The fingernail sensors were also tested within the context of the cable assembly task discussed in Part 2 of this report. By setting the proper threshold on the sensor outputs, the fingernail sensors were used to trigger the appropriate state transitions that are enabled by the human squeezing the screws of the cable connector. As shown by Figure 13, an initial demonstration of the virtual switch panel concept was also created by placing virtual switches on the robot that could be used to activate/deactivate the robot as well as to move the robot joints.

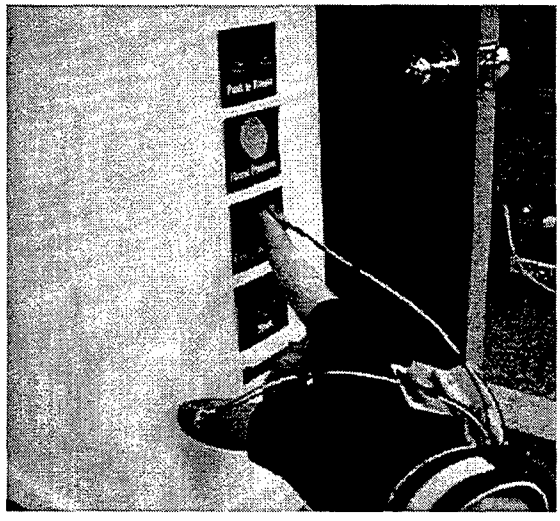


Figure 13: Virtual Switch Implementation

It was found that pushing the virtual switches consistently resulted in a change in fingernail color and a change in sensor readings. However, the manner in which the force was applied caused the output voltage to either increase or decrease. Also it was found that the sensors could be falsely triggered due to two types of motion artifact:

1. The fingernail changes color when the finger itself is bent forward or backward a significant amount
2. The amount of ambient light from the room which passes through the semi-transparent finger changes with the position and orientation of the hand with respect to light sources.

The conclusion section will discuss proposed methods of eliminating these problems.

5. Conclusion

A new type of touch sensor for detecting contact pressure at human fingertips has been presented and applications to human-machine interface have been discussed. Prototype sensors were constructed and merged with a fingerless dataglove. Initial experiments show that the concept is successful, but needs further work.

Firstly, the sensors should compensate for ambient room lighting by strobing the LEDs and subtracting out whatever light passes by the sensor shielding. Secondly, the sensors should be expanded to comprise an array of photodetectors around each fingernail. In this way, color changes caused by finger motion can be differentiated from color changes caused by touch force, and information about the location and direction of the force can be obtained from the pattern of the color change. Finally, multiple LEDs with different wavelengths should be used to differentiate the absorption change due to blood volume changes from the absorption change due to change in blood oxygen concentrations. This may also help in eliminating motion artifact. In general, by gathering more information through additional LED wavelengths and photodetectors, an accurate model of the dynamics of the fingernail blood flow can be achieved.

Applications such as the virtual switch panel and digital wearable manual will be further investigated as well, and are expected to provide significant contributions to the field of human-machine interaction.

Appendix A: Anatomy of the Human Fingernail

By **Elie Awad,** Graduate Research Assistant

The nail unit has five components: the *nail plate,* the *matrix,* the *proximal nail fold,* the *nail bed* and the *hyponochium.* [10]

1. The **nail plate** is a slightly convex semitransparent plate whose thickness increases proximodistally from 0.7 to 1.6 mm. The surface of the nail usually shows longitudinal fine

ridges but may develop transverse ridging in the case of disease or disturbance of normal growth. Minute air bubbles trapped under the nail show as white flecks. [10]

The nail plate is composed of two or three layers [10] of flattened cells containing hard keratin [11], it is thus homologous with the stratum corneum of the general epidermis. However the nail plate has a lower lipid content than the general stratum corneum [10]. Even though the water content of nail is low; its permeability is 10 times as much as that of general dermis (Baden 1970). "Elasticity of the nail plate is related to its degree of hydration" [10].

Nail plate growth is fastest - in the middle finger - at a rate of 0.1 mm per day [10], the average growth rate being 0.5 mm per week. [11]

2. The **matrix** is a "V-shaped epidermal invagination" [11] in which the proximal part of the nail plate is embedded. The matrix is referred to as *dorsal matrix* - for those cells lying *dorsal* to the nail plate - and *ventral matrix* - for the cells lying ventral to the plate. The matrix epithelium consists of typical keratinocytes. The distal extension of the nail matrix toward the nail bed results in crescentic white opaque area called the *lunule*. [10]
3. The **proximal nail fold** is composed of two epidermal layers with dermis in between.
4. The **nail bed** consists of several layers of epidermal cells [11]. It is grooved in a similar pattern to the nail itself to provide for good interlocking [10] between the two. Some references state that "except in the nail matrix, these (the nail bed) cells do not participate in formation of the nail plate" [11]. The nail bed is thus "regarded as providing a gliding surface for an already fully formed growing nail plate" [10]. Other authorities think that the nail plate consists of three layers (see Figure 14):
   - a dorsal layer produced by the dorsal matrix
   - an intermediate layer produced from the ventral matrix
   - and a ventral layer produced from the nail bed

   Beneath the epithelium of the nail bed is a layer of dermis "anchored to the periosteum of the distal phalanx without intervening subcutis" [10]. This attachment to bone is provided by strong collagen and elastic fibers. The nails are therefore "immobile over the distal phalange" [11]. Fingernails also have an important tactile function in providing "support and counterpressure for the digital pad, thereby aiding manipulation" [10].

The dermis is richly vascularized with large arteriovenous shunts [10]. Directionally the capillaries are vertical into the dermal papillae under the nail matrix, but longitudinal under the nail bed [11]. Because of its blood content, the nail bed appears pink through the translucent nail bodies. However drop in oxygen level and development of cyanosis will cause the nail bed to turn blue. [12]

5. The **hyponychium** is an area of epidermis that underlies the edge of the nail plate and continues into the general epidermis of the fingertip. Its function is to prevent bacteria from entering underneath the free end of the nail plate. [10]

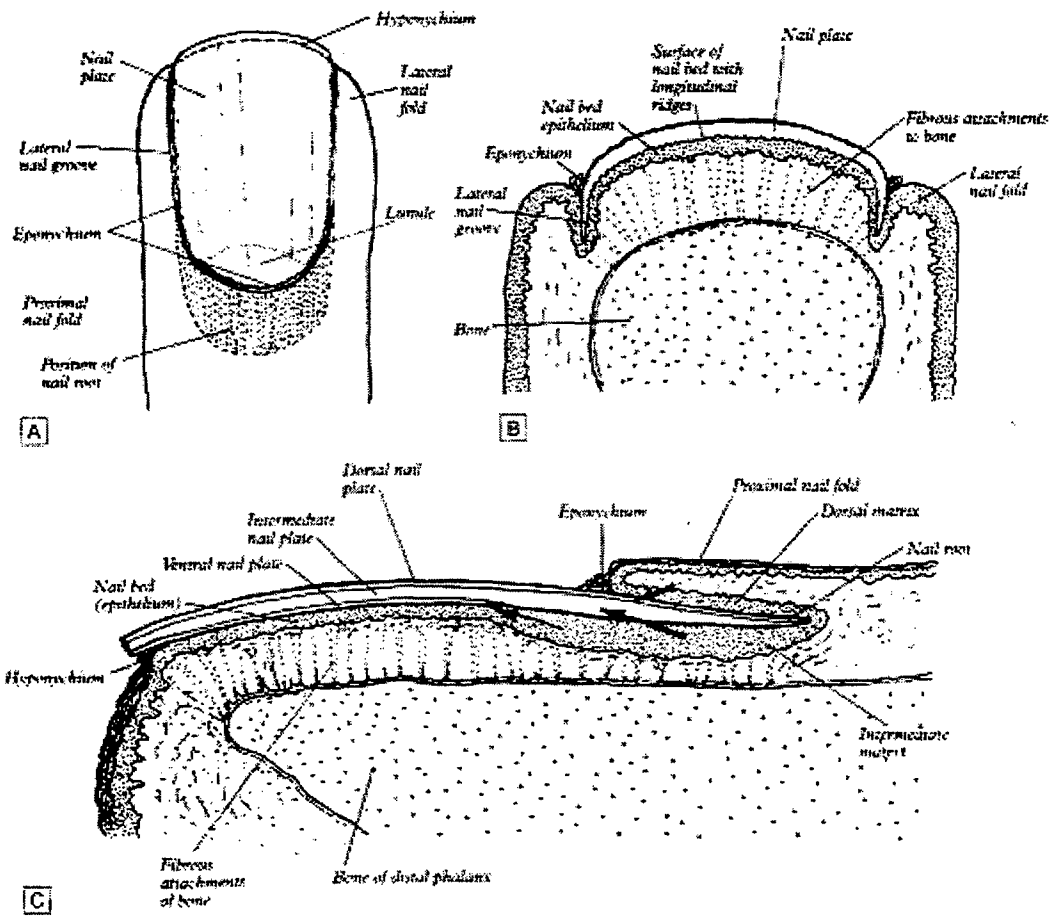


Figure 14: Nail Anatomy

(From *Gray's Anatomy: the Anatomical Basis of Medicine and Surgery*, p. 409, see [10])

References

[1] Sturman, David J. & Zelzer, David, "A Survey of Glove-base Input," IEEE computer Graphics & Applications January 1994, pp. 30-39.

[2] T. G. Zimmerman et al., "A Hand Gesture Interface Device," Proc. Human Factors in Computer systems and graphic interface, ACM Press, New York, April 1987, pp. 189-192.

[3] M. Brooks, "The DataGlove as a Man-Machine Interface for Robotics,"2nd IARP Workshop on Medical and Healthcare Robotics, Newcastle upon Tyne, UK, Sept. 1989, pp. 213-225.

[4] T.H. Speeter, "Transforming Human Hand Motion for Telemanipulation," Tech. Memorandum, AT&T Bell Laboratories, Holmdel, NJ, Sept. 19, 1989.

[5] L. Pao and T.H. Speeter, "Transformation of Human Hand Positions for Robotic Hand Control," Proc. IEEE Int. Conf. Robotics and Automation Vol. 3, IEEE CS Press, Los Alamitos. CA, 1989, pp. 1758-1763.

[6] M. Bergamasco, B. Allotta, L. Bosio, L. Ferretti, G. Parrini, G.M. Prisco, F. Salsedo, and G. Sartini, "An Arm Exoskeleton System for Teleoperation and Virtual Environments Applications", *Proceedings of the IEEE International Conference on Robotic and Automation*, pp. 1449-1454, 1994.

[7] T.B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*, MIT Press, Cambridge, MA, 1992.

[8] R.S. Kalawsky *The Science of Virtual Reality and Virtual Environments*, Addison-Wesley, Wokingham, England, 1993.

[9] R.A. Pax, J.G. Webster, and R.G. Radwin, *A Pressure Sensing Glove Using Conductive Polymer Pressure Sensors*, Madison, WI: Univ. Wisconsin Press, 1989.

[10] *Gray's Anatomy: The Anatomical Basis of Medicine and Surgery*, New York: Churchill Livingstone,1995.

[11] Kristie, *Human Microscopic Anatomy*, Springer-Verlag, 1991.

[12] Thibodeau and Patton, *Structure and Function of the Human Body*, Moseby, 1997.

Progress Report, March 31, 1998
Total Home Automation and Healthcare Consortium

Part 2: Human Machine Interactive Control Using Dual Petri Nets

H. Harry Asada
Professor, Principal Investigator

Stephen Mascaro
Graduate Research Assistant

Abstract

A new approach to interactive control of human-robot systems using dual Petri nets is presented. A human and robot(s) work side by side by sharing the task goal and part of the process state. The human is instrumented with a type of data glove so that the robot can monitor the human state and coordinate its action with the human. First, the class of interactive tasks performed by a human and robot is described as a dual-agent, concurrent, event-driven system, and represented by two Petri nets interacting to each other. One Petri net represents the human side task process, while the other represents the robot side. Both sides proceed with their assigned tasks concurrently through state transitions within each Petri net. They observe the partner's state and coordinate each step of task performance as specified by the conditional state transition involved in the Petri nets. A distributed control system is directly derived from the dual Petri net representation. As an exemplary case study, a cable connection task is considered. The task is first decomposed into two subtasks, which are assigned to the human and the robot. A data glove worn by the human is developed for measuring the hand location as well as the force and position of each finger tip. A decentralized controller is built and a proof-of-concept experiment is described.

d'Arbeloff Laboratory for Information Systems and Technology
MIT

1. Introduction

When two people work together to perform a task, e.g. preparing a meal in a kitchen, performing a surgical operation in a hospital, or assembling automobiles in a factory, they coordinate their actions by (1) observing each others actions and by (2) sharing knowledge about the task process.

Both sides have a common task goal and share all or part of the knowledge about the task. These include task procedures, assignment, coordination protocol, process status, etc. Each side observes the partner's action to recognize in which stage the current operation is involved, and which action is required. The required action may be modified depending on the partner's performance; the partner's mistake may be corrected or, at least, alerted to the partner. Although both sides work in the same workspace, collision may be avoided by observing the partner's movement.

In a fully automated manufacturing line, a robotic assembly line, for example, all the machines are highly coordinated by a centralized or decentralized controller. The task procedure is completely described by programs, including interlock and synchronization of each machine. The state of every machine is monitored, and every step of operation is completely assigned to each machine. In case a few machines are unable to perform assigned tasks, others may replace the unusable machines, or the line would be shut down for repair. Today's automated manufacturing line is highly sophisticated in coordinating numerous machines to maximize performance efficiently, robustness, and versatility.

On the other hand, when humans and machines are mixed, such a high level of coordinated operations can hardly be achieved. Today's technology does not allow humans and machines to work closely as humans alone or machines alone can do. The communication channels between the human and machines are very limited, and interactions are minimal. Each side performs assigned task separately, and the task is divided into the two such that the

interactions are minimized. In a manufacturing line, for example, robots and humans are segregated; robots are enclosed with a cage and humans cannot access the robot area. Isolation of machines from humans results in limited functionality, inflexibility, and low productivity. If humans and machines could work side by side, highly flexible and productive working relations could be established. For example, machines could correct human errors, and the machine's actions could automatically be modified depending on the human performance.

In the past decades, cooperative control of a human and robot has been studied in different contexts and application areas. Cooperative control systems have been developed for handling a large object, in which the human and robot communicate through interactive forces applied by the human and robot on the same object [1][2]. This is paralleled by the development of control and communication systems for cooperation between multiple robots [3][4]. Human supervisory control for telerobots has provided the basic framework where the human serves as the supervisor of the robot at a distance [5]. Recent progress in virtual realty such as in [6] and [7] provides a powerful tool for supervisory control so that the human can effectively interact with the environment. Human-robot interaction has also been used for robot skill acquisition and learning [8][9], leading towards greater human-robot coordination.

In this paper, a human and a robot work side by side as partners. There is neither supervisor nor hierarchical relationship. Both sides must help each other to achieve the common goal. Moreover the class of tasks considered in this paper consists of many steps of operations. As seen in many team works by humans, e.g. surgical operations and assembly of automobiles, the entire task procedure is divided into many steps, and each step is assigned to each member of the team. Many of the steps need to be performed in coordination with other members. By observing other members' task performance, each member must start off and complete assigned steps of work in the right sequence. The goal of this paper is to achieve such a team work environment between a human and a robot. In the following sections, a basic framework for designing collaborative human-robot systems and specific techniques for implementing the collaborative control system will be presented. In particular, the modeling and representation of the class of collaborative tasks by using dual Petri nets will be developed and applied to a practical task. A proof-of-concept system comprising a human wearing a data glove, a robot, and

a task process controller is developed and tested. With this system, the human becomes *hand-in-glove* with the robot.

2. Task Process Modeling Using Petri Nets

2.1 Requirements for Task Process Modeling

The human and machines must share knowledge about a given task in order to coordinate their actions. To interpret sensor information, understand human intentions, and take necessary actions at the right timing, there must be a certain description about the task process shared by the human and robot. Effective modeling and representation of interactive task processes are a critical component in building the "hand-in-glove" human-machine systems. In this section, a new method for modeling interactive task processes by using a type of Petri net is developed.

The nature of the tasks considered in this paper is summarized as follows:

- ***Interactive in Sequence:*** The human and robot must be interactive to coordinate steps of operations. The robot must perform some steps of operations, after the human has completed certain steps of his/her task and the robot has arrived at a certain state. The robot must be able to detect the human status in order to determine when to start which steps of operations. In turn, the human must know which steps of operation the robot performs in response to his/her side of task completion. The modeling tool to be used must be able to describe this interactive nature of the task sequencing.

- ***Event-Driven:*** To coordinate steps of operations at both the human and robot sides, event-driven process modeling is the rational choice [4][10]. Time-driven representation, which uses time as an independent variable and represents the entire task process as time functions and differential equations, is inappropriate; since prescribed time functions cannot be defined for interactive task processes. The time when the robot starts a particular step of task is not known a priori, but is dependent upon when the human completes some steps of tasks. Completion of steps of task is an event rather than time. In consequence, occurrence or non-occurrence of each event is the primary variable to drive the task process.

- ***Concurrency:*** Both the human and robot perform assigned tasks concurrently. Steps of operations assigned to each side proceed concurrently. Time to time they coordinate their tasks, but in essence the entire task process proceeds as two concurrent event-driven systems.

The modeling tool to be developed in this paper should meet these requirements. Namely, the task process should be described as compound, event-driven systems that interact to each other and that make state transitions concurrently.

2.2 Dual Petri Net Architecture

The Petri net, which has been extensively used for the modeling and simulation of manufacturing systems, is a powerful tool for describing concurrent, event-driven processes. Like other state networks, process states are represented by nodes, while transitions between two states are by arcs connecting them. In the Petri net, however, the overall process state is represented by combination of multiple nodes allowing for concurrent multiple transitions by moving multiple "tokens" asynchronously. This feature of Petri nets allows us to represent concurrent events occurring at both the human and robot sides simultaneously. As shown in Figure 1, two Petri nets are used for representing the two sub-tasks being performed by both the human and robot sides; one for the human side and the other for the robot side task process. The status of each sub-task process is indicated by placing a token in the corresponding node. Therefore at least two tokens are placed at all times. This dual Petri net model visualizes the concurrent nature of the human-robot systems. Moreover, the interaction and coordination can be clearly represented by conditional state transitions. In Petri nets, conditions for a state transition to fire are described in terms of the existence of a token in each specific state relevant to the conditions for the state transition. If, for example, a certain action must be taken by the robot when both the human state and the robot's state are in certain conditions, the arc indicating the firing of the action declares that two tokens must exist in the two nodes, respectively, as shown by the bars and arrows in the figure below. This allows one to describe a type-coordinated motion, or interlock between the human and machines.

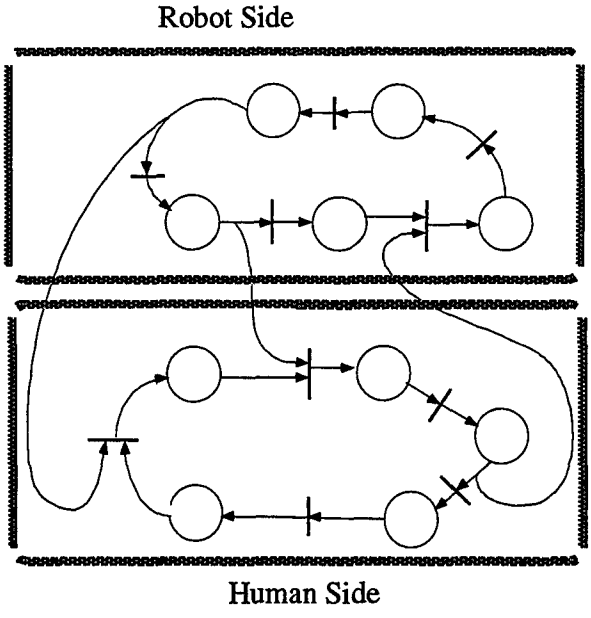


Figure 1: Dual Petri Net Example

This dual Petri net method allows not only to represent concurrent, interactive task processes, but also to *encode* the type of task knowledge needed for interpreting sensor data and estimating the process state. As will be demonstrated later in Section 4, on-line sensor data can be interpreted by referring to the Petri net process model. The system would keep track of the state of the human as well as that of the robot, and indicate the estimated states by placing tokens. Transitions would be detected by matching the profile of each sensor signal with the ones of anticipated transitions. Anticipated transitions are a group of arcs branching out of the current nodes (states), as shown in the figure. In other words, transitions other than the group of anticipated transitions are totally irrelevant to the current process state, thereby eliminated from the data interpretation process. This allows one to bring in a kind of "context" to the task process monitoring.

Within the Petri net framework, diverse conditions, states, and actions/controls can be represented. A safety check, for example, would be represented as a demon imbedded in the network, which continuously checks whether the physical distance between the human body (hands) and the robot is longer than a minimum allowable distance. When the safety condition is violated, the robot is to be moved to a certain state and/or some transitions are to be prohibited.

3. An Exemplary Case Study: Cable Harness Assembly

3.1 Task Analysis: Decomposition and Assignment

For the embodiment of the hand-in-glove concept and the dual Petri net approach, a cable assembly task has been chosen as an exemplary case study. Specifically, the task is to connect a flexible computer cable to a fixed computer port, as shown in Figure 2. In order to complete the task, the flexible cable must be picked up from a bin and inserted into the port, and two screws must be turned to a certain tightness, i.e. torque management. This task has the characteristic that it would be difficult or slow for either a human or robot to do alone. Certain parts of this task are difficult for a robot to perform but easy for a human, while other parts are difficult for the human to perform but easy for the robot. Namely, it is difficult for the robot to grasp the cable harness and insert the connector into the port, since the robot would require a specialized end effector to hold it, and/or special sensors to detect the position and orientation of the cable. However, the human can pick up such a harness and perform the insertion sub-task relatively easily. Conversely, it is relatively difficult and slow for the human to turn the screws and accommodate the torque, while this is something the robot can do quickly, easily, and consistently in terms of tightness.

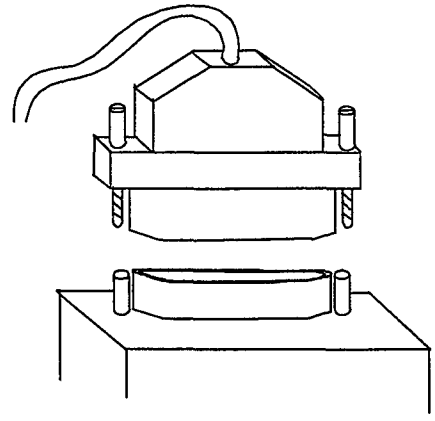


Figure 2: Cable Assembly Task

In light of this categorization of human and robot abilities, the cable assembly task is divided, assigning to the human the subtask of cable harness pick-up and connector insertion, and assigning to the robot the subtask of screw turning. Next, this sub-task assignment is expanded into separate sets of instructions for the human and robot. Figure 3 shows the

instruction manual for the cable assembly task. Notice that each step is phrased in the form of a conditional statement: wait for the set of conditions **C** to be met, then perform the set of actions **A**. On the human side, for any given step **i**, $\mathbf{C}_i$ may contain conditions that the human must meet, conditions that the robot must meet, or some combination of both. While the robot conditions are stated explicitly, the human conditions are often implicit – i.e. the human must have completed the previous step, or the human must be in some state of readiness. The actions in $\mathbf{A}_i$ must all be performed at the same time and rely on the same set of conditions $\mathbf{C}_i$. Otherwise, the step should be broken into two or more sequential or concurrent steps. For the simple cable assembly task, there are no concurrent steps within either the human or robot sub-task. The action of "place fingers on screw" can be broken down into a set of multiple actions, such as "move hand to screw" and " align fingers." However these actions can be performed in one step by the human and do not rely on separate sets of conditions to be met.

Human Side

1. Move hand into workspace and insert cable.
2. Wait for robot to reach ready position, then place fingers on a screw.
3. Wait for robot to position itself above the screw, then squeeze fingers. As bit lowers, turn screw to align slot.
4. When bit is in place, stop squeezing.
5. When screw is tight, remove fingers.
6. Repeat steps 2-5 for second screw.
7. Remove hand from workspace.

Robot Side

1. Wait for human to move hand into workspace, then assume ready position.
2. When the human places fingers on screw, move above the screw.
3. Whn human squeezes fingers, lower bit onto screw.
4. When human stops squeezing, start turning screw.
5. When human moves hand away, stop screwing, raise bit, and return to ready position.
6. Repeat steps 2-5 for second screw.
7. When human removes hand from workspace, return to home position

Figure 3: Instruction Manual

3.2 Reducing Manual Information to Dual Petri Net

Human Side			
Step (i)	**Implicit Human State (S^H_i)**	**Conditions (C^H_i)**	**Actions (A^H_i)**
1	Hand is outside workspace	Human is ready	Move hand into workspace and insert cable (goto step 2)
2	Hand is in workspace	Robot in ready position, human decides on screw #1 or screw #2	Place fingers on screw #1 or screw #2 (goto step 3)
		Human needs break	Remove hand from workspace, return to step 1
3	Fingers are on screw #N (at location **L**)	Robot positions above screw #N	Squeeze fingers and align screw (goto step 4)
		Human is finished with screw # N	Move hand from screw back to workspace, return to step 2
4	Fingers are squeezing screw # N	Robot positions bit on screw # N	Stop squeezing, return to step 3

Robot Side			
Step(j)	**Implicit Robot State (S^R_j)**	**Conditions (C^R_j)**	**Actions (A^R_j)**
1	In home position	Hand in workspace	Move to ready position (goto step 2)
2	Ready position	Fingers on screw # N (at location **L**)	Move to screw #N (goto step 3)
		Hand outside workspace	Move to home position, return to step 1
3	Position above screw # N	Fingers squeezing screw # N	Position bit on screw (goto step 4)
		Hand back to workspace	Move to ready position, return to step 2
4	Position on screw # N	Fingers on screw # N but not squeezing	Start turning screw (goto step 5)
		Hand back to workspace	Position above screw, return to step 3
5	Turning screw # N	Hand back to workspace	Stop turning screw, return to step 4

Figure 4: Formal Instruction Matrix

Once the set of instructions has been drawn up, these instructions are converted into a graphical dual Petri net model. As an intermediate step, the basic instruction manual is first transformed into a more detailed and structured form. First, each step is represented by an implicit state $\mathbf{S}_i$, a set of conditions $\mathbf{C}_i$ (both explicit and implicit), and a corresponding set of actions $\mathbf{A}_i$. These are entered into a formalized instruction matrix. Then steps with the same or like states $\mathbf{S}_i$ can be mapped into a single step with one or more variables and multiple options. For example, all the steps involving screw #1 and screw #2 can be mapped into a set of generic steps with variables **N** (number of screw) and **L** (location of screw). For each step, we can build in reversibility by specifying a separate condition and action for back-stepping. Figure 4 shows the final formalized instruction matrix for this task.

Once the formal instruction matrix has been drawn up, the conversion to a dual Petri-Net can be methodically performed. The basic rules for this conversion are as follows:

1. Draw places for each implied human state $\mathbf{S}^{\mathbf{H}}_{\mathbf{i}}$ and for each implied robot state $\mathbf{S}^{\mathbf{R}}_{\mathbf{j}}$.
2. Draw transitions for each forward action $\mathbf{A}^{\mathbf{H,F}}_{\mathbf{i}}$ and each backward action $\mathbf{A}^{\mathbf{H,B}}_{\mathbf{i}}$ on the human side. Likewise, draw transitions for each forward action $\mathbf{A}^{\mathbf{R,F}}_{\mathbf{j}}$ and for each backward action $\mathbf{A}^{\mathbf{R,B}}_{\mathbf{j}}$ on the robot side.
3. For each human step **i**, draw input arcs from $\mathbf{S}^{\mathbf{H}}_{\mathbf{i}}$ to $\mathbf{A}^{\mathbf{H,F}}_{\mathbf{i}}$ and $\mathbf{A}^{\mathbf{H,B}}_{\mathbf{i}}$. Likewise, for each robot step **j**, draw input arcs from $\mathbf{S}^{\mathbf{R}}_{\mathbf{j}}$ to $\mathbf{A}^{\mathbf{R,F}}_{\mathbf{j}}$ and $\mathbf{A}^{\mathbf{R,B}}_{\mathbf{j}}$.
4. Draw output arcs for each forward transition on the human side from $\mathbf{A}^{\mathbf{H,F}}_{\mathbf{i}}$ to $\mathbf{S}^{\mathbf{H}}_{\mathbf{i}}$. Likewise, draw output arcs for each forward transition on the robot side from $\mathbf{A}^{\mathbf{R,F}}_{\mathbf{j}}$ to $\mathbf{S}^{\mathbf{R}}_{\mathbf{j}}$.
5. Draw output arcs for each backward transition on the human side from $\mathbf{A}^{\mathbf{H,B}}_{\mathbf{i}}$ to $\mathbf{S}^{\mathbf{H}}_{\mathbf{i-1}}$. Likewise, draw output arcs for each backward transition on the robot side from $\mathbf{A}^{\mathbf{R,B}}_{\mathbf{j}}$ to $\mathbf{S}^{\mathbf{R}}_{\mathbf{j-1}}$.
6. For each action $\mathbf{A}^{\mathbf{H}}_{\mathbf{i}}$ on the human side (forward or backward), find the state $\mathbf{S}^{\mathbf{R}}_{\mathbf{j}}$ on the robot side which corresponds to the condition $\mathbf{C}^{\mathbf{H}}_{\mathbf{i}}$ (if one exists) for that action, and draw a bi-directional arc (input and output) between $\mathbf{S}^{\mathbf{R}}_{\mathbf{j}}$ and $\mathbf{A}^{\mathbf{H}}_{\mathbf{i}}$. Likewise, for each action $\mathbf{A}^{\mathbf{R}}_{\mathbf{i}}$ on the robot side (forward or backward), find the state $\mathbf{S}^{\mathbf{H}}_{\mathbf{j}}$ on the human side which corresponds to the condition $\mathbf{C}^{\mathbf{R}}_{\mathbf{i}}$ (if one exists) for that action, and draw a bi-directional arc (input and output) between $\mathbf{S}^{\mathbf{H}}_{\mathbf{j}}$ and $\mathbf{A}^{\mathbf{R}}_{\mathbf{i}}$.

7. Finally, for each condition C^H_i which corresponds to some human decision, we can draw it as an external condition or place, and from it draw an input arc to the corresponding action A^H_i. Note that these conditions give rise to the variability in progression of the task. Without them, the task would be completely deterministic.

Figure 5 shows the dual Petri Net for the cable assembly task. Each of the circles represents a place/state/condition, each of the vertical bars represents a transition/action, and each of the directional arcs represents and input or output. The bi-directional arcs indicate that a place is both an input and an output of the transition. If a condition is met, that means a token exists in that place. If tokens exist at all the input places for a given transition, then that transition is fired and tokens are placed at each of the output places. Notice that only bi-directional arcs connect the two concurrent sub-nets. This reflects the fact that the states of the robot and human are independent from each other. While the states or conditions of the human can enable certain actions of the robot, the resulting actions of the robot do not directly change the state of the human, and vice versa.

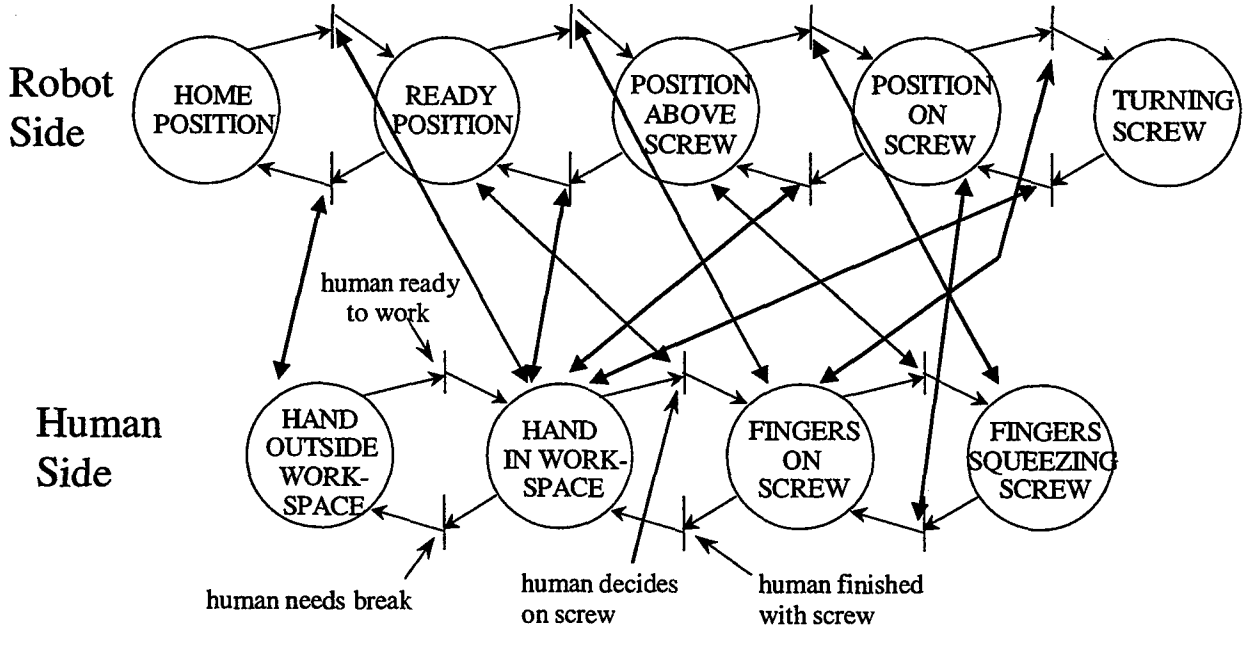


Figure 5: Petri Net Model of Cable Assembly Task

An important point to note about the Petri net in figure 5 is that while certain conditions are used in more than one step of the original instruction manual, they are represented as only one place in the Petri net. By constructing the Petri net in this manner, more flexibility is built into the task process. Actions can be reversed and errors can be corrected without starting the entire process over. Furthermore, the places labeled "above screw" or "on screw", etc, actually represent two or more screws at different locations. Certain "fuzzy" variables can be attached to places as well as tokens, which allow the robot to make decisions based on continuous information such as location of the screw or memory of the process history [12].

One of the useful features of Petri Nets in general is that the discrete system behavior can be easily examined through simulation of the Petri Net. For the cable assembly task, the dual Petri Net has been tested through simulations and has been shown to work without deadlock or failure.

4. Implementation

4.1 A Proof-of-Concept Prototype System

Once the task has been broken down and modeled, the appropriate sensors are chosen. Sensors must be chosen and designed such that all enabling conditions on the human side can be uniquely detected. For the cable assembly task, enabling conditions are based on the positions of the hand, positions of the fingers, and pressure exerted by the fingers. Therefore we must be able to measure these quantities to a level of certainty that the various conditions can be distinguished. For this initial implementation, the human is outfitted with a glove that measures palm position and orientation using a magnetic tracker, as well as pressure sensors in the tips of the thumb and index finger. Finger position can be roughly extrapolated from wrist position and orientation. Figure 6 shows a picture of the experimental sensor glove and cable assembly setup.

The magnetic tracker is capable of measuring palm position and orientation within a 48 " (1.2m) radius with accuracies of 0.07" (2mm) for translations and 0.5° for rotations. Fingertip positions can be extrapolated to accuracies of approximately 0.5" (1.3cm) by assuming the angles of bend during contact with the screws. The pressure sensors are 0.012"(0.30mm) thick,

0.2"(5.0mm) in diameter, and are capable of measuring forces up to 10kg with an accuracy of 25%. Minimum force that can be measured is about 20g.

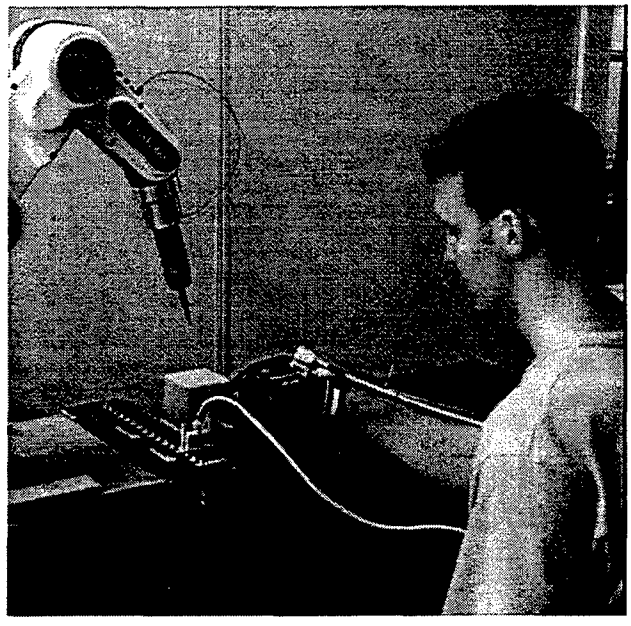


Figure 6: Experimental Setup

For this initial implementation, the position and orientation sensing is accurate enough to detect whether the fingers are on screw #1 or screw #2 of the cable, as long as the port is kept at a known, fixed location. For an unknown or variable screw location, it would be necessary to add sensors, which measure the angle of bend of the fingers, or sensors, which directly measure fingertip location. The position sensing is more than accurate enough to determine the various states of the hand in the Petri net, and the pressure sensing is sufficient to determine whether or not the fingers are squeezing the screws.

4.2 Programming

Once modeling is complete and sensors are chosen and designed, the next step is to translate the Petri net model into computer programs for monitoring the human and controlling the robot. Although it is possible to have one computer and one program to perform both these functions, it is often the case that control is distributed among multiple computers and multiple programs. Industrial robots usually come with their own controllers and programming software.

In some cases, it may be more efficient and more flexible to use existing robot controllers and software. In the initial implementation, we demonstrate this by dividing the programming between two controllers. The pre-existing robot controller is used to perform all actions pertaining to robot monitoring and control, while a separate computer is used to monitor the instrumented human. These two sub-controllers correspond directly to the two concurrent sides of the Petri net in Figure 5. Each program is completely responsible for monitoring the states and controlling transitions on one side of the dual Petri net. The two controllers have a digital data link, whereby the two programs communicate back and forth, informing each other of the states of the robot and human, respectively. This leads to a very modular control system. In case of changes of hardware for the human instrumentation, the programming on the human side can be changed without affecting the programming on the robot side. In case of changes of robot hardware, the converse is true. Figure 7 shows a diagram of the control architecture.

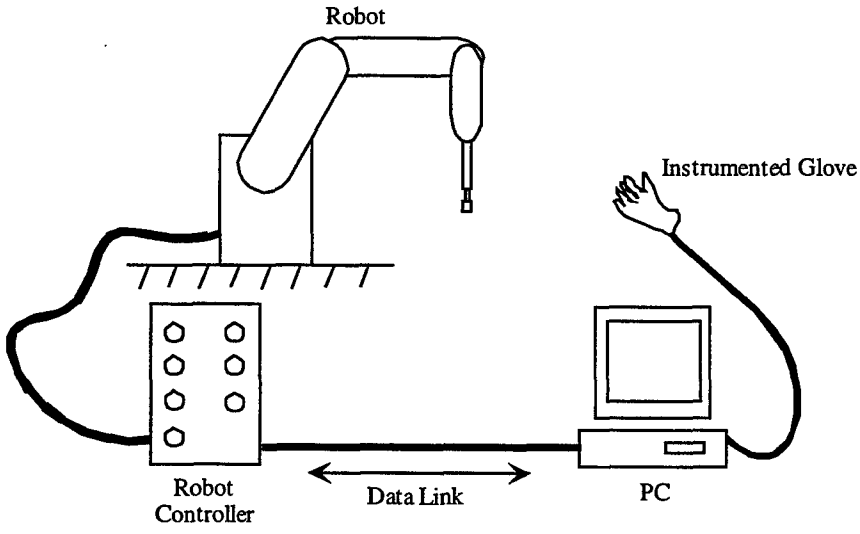


Figure 7: Control Architecture

In the initial implementation the two control programs were written in a procedural method, where each state of the human or robot is written as a separate subroutine. Each of these "state-subroutines" relies on a set of fundamental subroutines that are repeatedly called to check data from the link and sensors (in the case of the human side). Each state-subroutine then checks whether enabling conditions are met, and fires transitions by switching to another state-subroutine. This method works for a small number of states, where there is no concurrency or

loops on the human side or robot side. However if there are a large number of states, it becomes infeasible to write a subroutine for each one. Furthermore if there were concurrency, this would require multiple subroutines to be running simultaneously; also looping in the Petri net could result in an infinite nesting of subroutines. For this reason, an object oriented programming method is adopted, as in [13], where places, transitions, and tokens are represented by classes of objects, which contain pointers connecting them to each other. Member functions are used to collect data, check conditions, and fire transitions. Data records within objects are used to store information about states and events. This data can be used online to aid in decision-making, and can be compiled offline to learn about the task and to perform error analysis.

4.3 Experiments

The initial implementation was performed as described for the cable assembly task. Figure 6 shows a picture of the final experimental setup. A powered screwdriver was attached as an end effector to a 5-axis articulated robot from Panasonic, Inc. Sensors were calibrated and optimal thresholds for enabling transitions were determined experimentally. Despite small problems due to uncertainty in exact fingertip location, the human and robot were, in all cases, able to successfully work together to complete the cable assembly task. The process was reversible, and two or more screws could be turned in any order.

5. Concluding Remarks

A new approach to interactive control between humans and robots in the work place has been described in this paper. Tasks that require the skills of both humans and robots can be broken down into two sets of instructions, represented by a dual Petri net, and translated into control programs. By instrumenting the human hand with a variety of sensors, the robot is aware of the state of the human and reacts accordingly. A case study was performed and implemented for a cable assembly task.

This Hand-in-Glove approach to human-machine systems will be particularly useful for those cases where:

- Human has limited knowledge about the process as well as the functionality of the machines,
- Human error must be detected and corrected,
- High safety standards must be maintained, although humans and machines work closely,
- Human actions must be recorded together with the machine's action,
- Humans are unable to provide detailed commands to the machines.

One area for future development lies in the instrumentation. There is a need to develop a data glove and sensors that are unobtrusive, wireless and non-hindering for the human. Also, in the exemplary case study, the instrumentation is entirely one-sided. We assume that the human is fully capable of observing the robot states. Therefore, although the robot itself uses its own sensor feedback to determine its position, the human does not need this information. In some cases, however, it may be necessary to instrument the robot and provide feedback to the human.

A second area of future development is the exploration and refinement of well defined procedures for dividing the task, modeling the sub-tasks using dual Petri nets, and translating these nets into computer controls. Ideally, this can all be performed on a computer with the aid of graphical user interfaces, making the process simple and easy for the average worker.

References

[1] Y.Y. Yamamoto, E. Hiroshi, and X. Yun, "Coordinated Task Execution of a Human and a Mobile Manipulator," *Proceedings of the IEEE International Conference on Robotics and Automation*, pp.1006-1011, 1996.

[2] O.M. Al-Jarrah and Y.F. Zheng, "Arm-Manipulator Coordination for Load Sharing Using Variable Compliance Control," *Proceedings of the IEEE International Conference on Robotics and Automation*, pp.895-900, 1997.

[3] J. Szewczky, G. Morel, and P. Bidaud, "Distributed Impedance Control of Multiple Robot Systems," *Proceedings of the IEEE International Conference on Robotic and Automation*, pp. 1801-1806, 1997.

[4] N. Xi, T.J. Tarn and A.K. Bejczy, "Event-Based Planning and Control for Multi-Robot Coordination," *Proceedings of the IEEE International Conference on Robotic and Automation*, pp. 251-258, 1993.

[5] T.B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*, MIT Press, Cambridge, MA, 1992.

[6] T.H. Massie and J.K. Salisbury, "The PHANToM haptic interface: A Device for Probing Virtual Objects", *ASME Winter Annual Meeting*, Vol. 1, pp. 295-299, 1994.

[7] M. Bergamasco, B. Allotta, L. Bosio, L. Ferretti, G. Parrini, G.M. Prisco, F. Salsedo, and G. Sartini, "An Arm Exoskeleton System for Teleoperation and Virtual Environments Applications", *Proceedings of the IEEE International Conference on Robotic and Automation*, pp. 1449-1454, 1994.

[8] H. Asada and S. Liu, "Transfer of Human Skill to Neural net Robot Controllers," *Proceedings of the IEEE International Conference on Robotic and Automation*, pp. 2442-2447, 1991.

[9] Y. Xu and J. Yang, "Towards Human-Robot Coordination: Skill Modeling and Transferring via Hidden Markov Model," *Proceedings of the IEEE International Conference on Robotic and Automation*, pp. 1906-1911, 1995.

[10] C. Guo, T.J. Tarn, N. Xi, and A.K. Bejczy, "Fusion of Human and Machine Intelligence for Telerobotic Systems," *Proceedings of the IEEE International Conference on Robotic and Automation*, pp. 3110-3115, 1995.

[11] Peterson, J.L. *Petri Net Theory and the Modeling of Systems*. Prentice Hall, Englewood, N.J., 1981.

[12] T. Cao and A.C. Sanderson. *Intelligent Task Planning Using Fuzzy Petri Nets*. World Scientific, Singapore, 1996.

[13] R. Bastide, P. Palanque, and C. Sibertin-Blank, "Cooperative Objects: A Concurrent, Petri-Net Based, Object-Oriented Language," *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics,* Vol. III, pp. 286-291, 1993.

Phase 2 Progress Report: March 31, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Home Networking

CHAPTER 19

A Modular, Minimum Complexity, High-Resolution and Low Cost Field Device Implementation for Home Automation and Healthcare

S. Martel, S. Lafontaine, I. Hunter

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Home Automation Report

A Modular, Minimum Complexity, High–Resolution and Low Cost Field Device Implementation for Home Automation and Healthcare

Dr. Sylvain Martel, Dr. Serge Lafontaine and Prof. Ian W. Hunter

The field device is essential to provide the final link with the transducers/sensors and actuators. It is a critical device since deficiencies at this level will typically be reflected throughout the whole system. In this report, field devices are remote devices that link a computer to the external world. This report is particularly interested into various implementation issues related to the field devices within the home. Some of these issues are simplicity, reliability, small size, and low cost. The suggested approach to implement such devices are based on the fact that since the communication bandwidth is improving faster than the A/D and D/A conversion rate, it is expected that the centralization of the resources in a near future will become much easier, yielding the simplest field devices possible. In other words, the strategy used to simplify the electronic interfaces at various remote locations within the home is to transfer many of the functions typically executed at the remote sites to the home personal computer with special plug–in interface cards, appropriate software and the integration of fast home networks such as HANET (Home Automation Network) based on FireWire mentioned in a previous report, to provide a level of interactions sufficient to support such approach. Taking into account that the Video Electronics Standard Association (VESA) home networking group has selected FireWire as the technology most worthy consideration for home networks, such high bandwidth level of interactions necessary to support our suggested approach is much likely to become available in the home.

Initially, RS422 links have been used to test these devices. Further development will use the IEEE–1394 (FireWire) at the physical, link and transaction layers with the AV/C command sets and the proposed HANET protocol expansion at the higher levels of abstraction.

This report discusses the basic theory and describes the first implementation. Because field devices become much simplified, they are likely to be more re–usable. With this in mind, we refer in this report to such field devices as *Universal Field Device (UFD).* Two families of UFDs have been developed initially. The first family is referred to as DEV devices and are used for general simple applications within the home. The next family deals with home health care applications. An example is the EMAP interface board described later in this report.

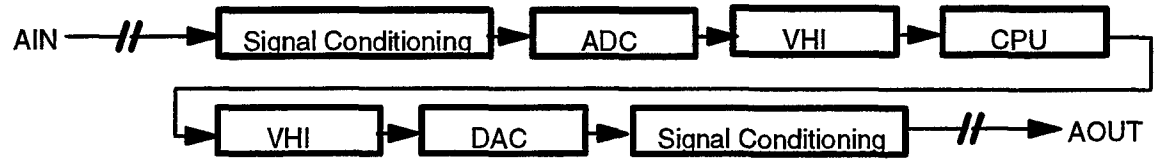
Universal Field Devices

A true *Universal Field Device (UFD)* is defined here as one which can support all possible requirements for interfacing directly with the external world. As such, a true UFD should provide A/D, D/A, and digital interfaces, with various bandwidth, resolutions, number of channels, and an appropriate signal conditioning. A true UFD is not possible because of limitations and tradeoffs, mainly imposed by analog circuits. Furthermore, even if the technology would exist, such true UFD would be very complex and hence expensive.

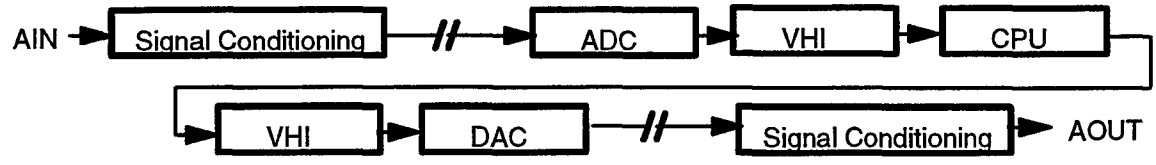
Because several functions, specially the analog functions, cannot be as universal as the functions implemented at the VHI (Virtual Hardware Interface) level for instance, the ultimate design objective when implementing UFDs is to minimize the complexity of the field device while optimizing the utilization of the hardware by transferring most of the functionality normally implemented within the field device, either within the VHI or for non time–critical functions, to a central computer. The basic idea is that if we reduce the functions on a field device that are not directly related with the functional interaction with the external world (for instance the communication protocol conversion layer), then we reduce the utilization constraints within various environments and hence, the compatibility level of the device (its universality) will increases. High universality at the remote locations can then be implemented with the simplest high–resolution A/D–based UFD implemented with only a signal conditioning block, an ADC, and a *Physical Layer (PL) communication Interface (PLI).*

UFD Hierarchy

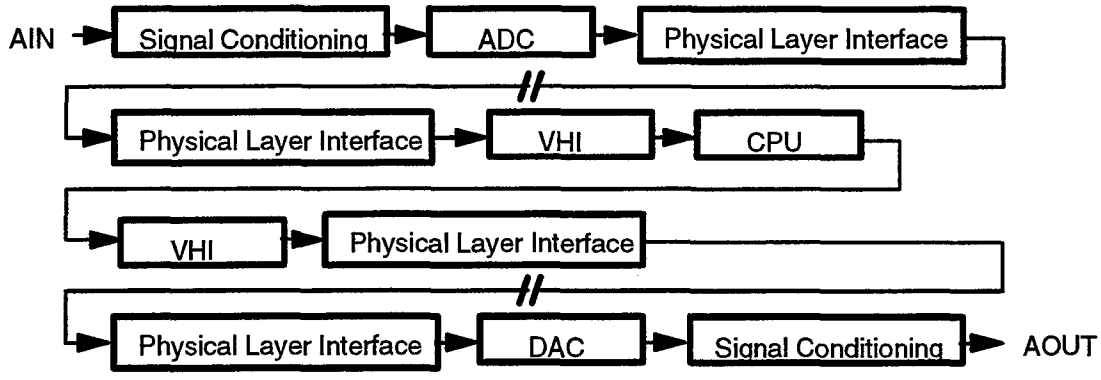
Fig. 1 shows some possible simple A/D– and D/A–based closed–loop configurations within an universal centralized home network. The configurations depicted in Figs. 1a and 1b rely on communication to and from the remote sites in analog forms. Therefore, the distance and/or accuracy are seriously limited, specially in high–resolution applications because of possible induced noises. Figs. 1c and 1d shows the level 1, 2, and 3 field devices. As opposed to the field devices on a conventional decentralized home network, the UFD always run in a deterministic manner with the central computer and is always directly synchronized with a VHI in order to minimize the amount of hardware at the remote site.



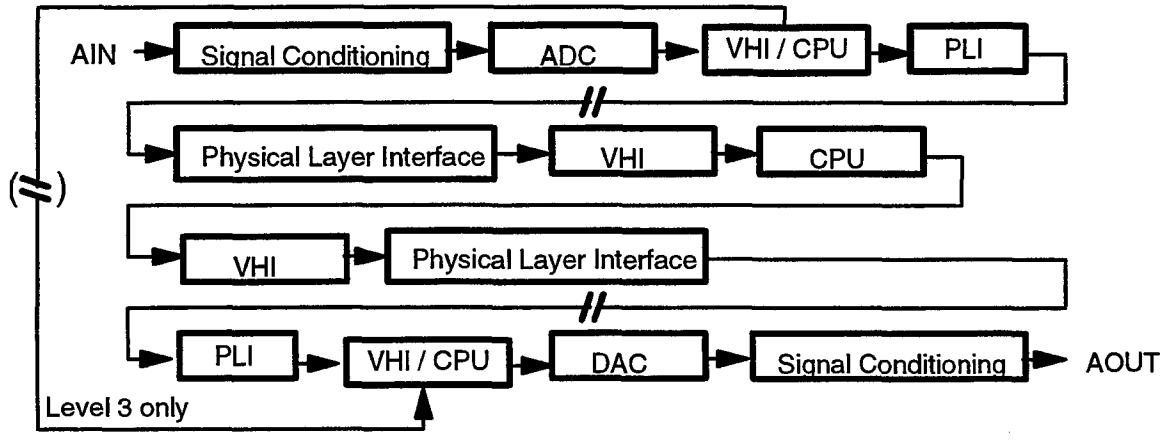
(a) – Without UFD.



(b) – UFDs with signal conditioning nodes only.



(c) – Level 1 UFDs.

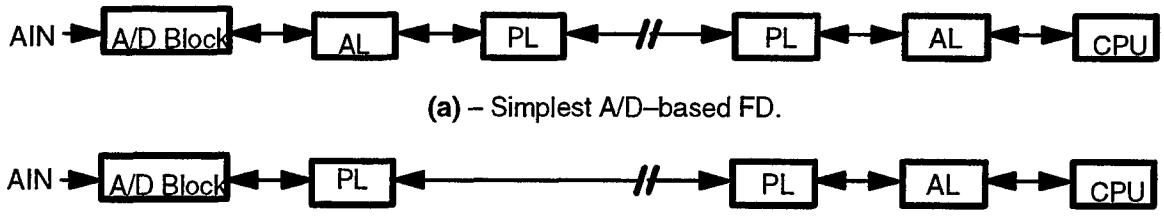


(d) – Level 2 and 3 UFDs.

Figure 1 – Main universal centralized DCS closed–loop configurations.

UFD vs. FD

The main difference between our proposed UFD and a conventional *Field Device (FD)*, either a remote smart sensor/actuator or a PLC is the absence of an *Adaptation Layer (AL)* at the remote site. This is shown in Fig. 2.



(a) – Simplest A/D–based FD.

(b) – Level 1 A/D–based UFD.

Figure 2 – Adaptation layers within the FD and UFD.

The *Physical Layer (PL)* has no intelligence and is used only to perform translation of a digital signal into a more convenient form such as differential or optical pulses prior to transmission. The adaptation layer on the other hand has intelligence and can be very complex.

The problem facing field devices with the multiplicity of communications interfaces employed in distributed instrumentation systems is likely to get worse for quite some time until industry standards are established [Gopel *et al.*, 1989]. It is not new to find that the biggest drawback to the use of digital communications in a distributed environment is the proliferation of different techniques and standards. An attempt was made by the Fieldbus committee to set a standard and most suppliers and users of Fieldbus equipment immediately expressed a commitment of using the eventual standard [Instrument Society, 1989]. It would appear that in order to achieve an agreement a compromise was required. Because of the commercial implications of standards, no one is ready to make compromise, the Fieldbus is by no means guaranteed as being a success. All attempts at defining industry wide standards such as the PROWAY [IEC, 1987] for instance would appear to have run into difficulties. Therefore, no communication standards are really optimized for any FDs such that an AL block which may consist of a communication processor and interface is necessary at the remote site. Furthermore, as the flexibility and number of options of a FD increases, the AL block becomes more complex and typically increases the latency significantly. This is further complicated by different standards for FD's within industrial users of large scale control system.

It is likely that no fixed communication standard would match without compromise the various types of FD implementations. To address this issue, a *Universal Communication Link (UCL)* [Martel and Hunter, 1995] is proposed. The proposed universal communication link has no predefined standards and is capable of matching a particular standard based entirely upon a user specified FD implementation. Because the communication link adapts to the FD, there is no compromise affecting the system's performance and the FD itself is much simplified since it does not require an AL block. This type of FD is referred to here as an UFD.

UFD vs. FD – A Simple Example

Let consider the very simple A/D–based UFD as depicted in Fig. 3. In this particular example, the UFD is built with a two–channel analog multiplexer, an ADC with a S/H amplifier, and a very simple interface IC (PL) which translates unipolar differential signals in TTL or CMOS levels.

In this example, the VHI is used as the only adaptation layer platform for time–critical functions. For instance, by implementing a virtual circuit in the fully–adaptive layer to detect a computer read ac-

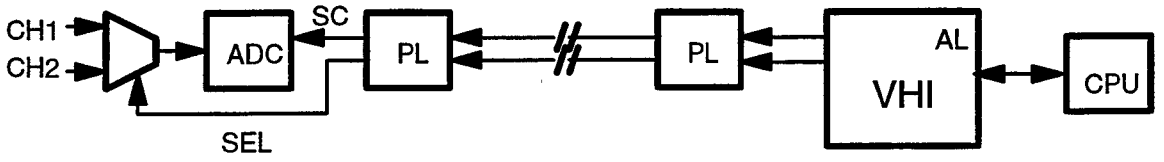
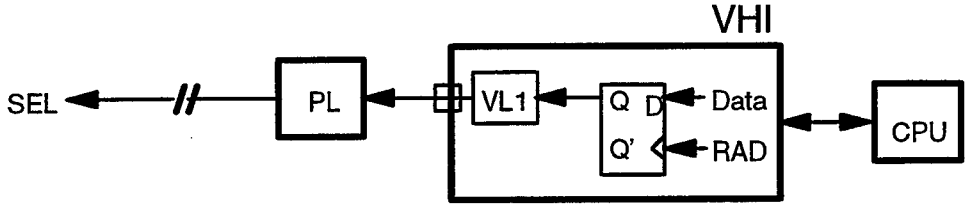
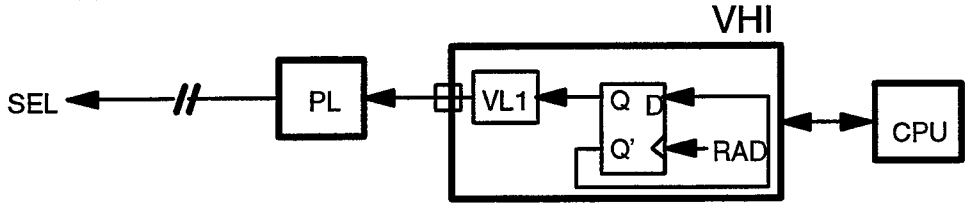


Figure 3 – Simple example of an A/D–based UFD.

cess *(Read Access Detect (RAD))* to the last converted data, a strobe can be sent automatically through one of the communication lines directly connected to a *Start Conversion (SC)* pin of an ADC. Similarly, the channel *selection (SEL)* can be done without glue logic at the remote site simply by implementing a 1–bit virtual register in the fully–adaptive layer. This is shown in Fig. below. The configuration has a communication VL assigned to an arbitrary communication line which has been intentionally assigned as the SEL line during the conception of this particular UFD.



(a) – Register–based channel select.



(b) – Alternating channel select.

Figure 4.1 – Simple example of AL reconfiguration.

In the last configuration when the sampled channel must be changed, an additional computer access is required to load a different value in the 1–bit virtual register. If for instance, the sampled channel must be changed after every conversion, the computer access overhead becomes significant. Instead, the VHI can be reconfigured with a new AL description. For instance, as shown in Fig. 4.1b, the 1–bit DFF–based virtual register is simply reconfigured as a DFF–based divide–by–two block. Notice that the throughput rate is double for this particular application. This was achieved with the same amount of gates in the AL and without modifying the UFD. We changed the adaptability level in the AL from dynamic fully–adaptive to static fully–adaptive in order to maintain the same level of complexity in term of gates. We could have implemented a dynamic fully–adaptive version to switch frequently between the two applications with an additional multiplexer controlled by a 1–bit virtual register in the AL implementation depicted in Fig. 4.1b. Fig. 4.2 is a more complete but simplified description of this particular example.

This is a very simple example that shows the possibility offered by this novel approach of a DCS based on the centralization of the hardware functions through the new VHI concept. The same principle applies for levels 2 and 3. Notice also that in a conventional FD, all possible virtual implementations in the VHI plus the additional electronics to switch between these implementations would have

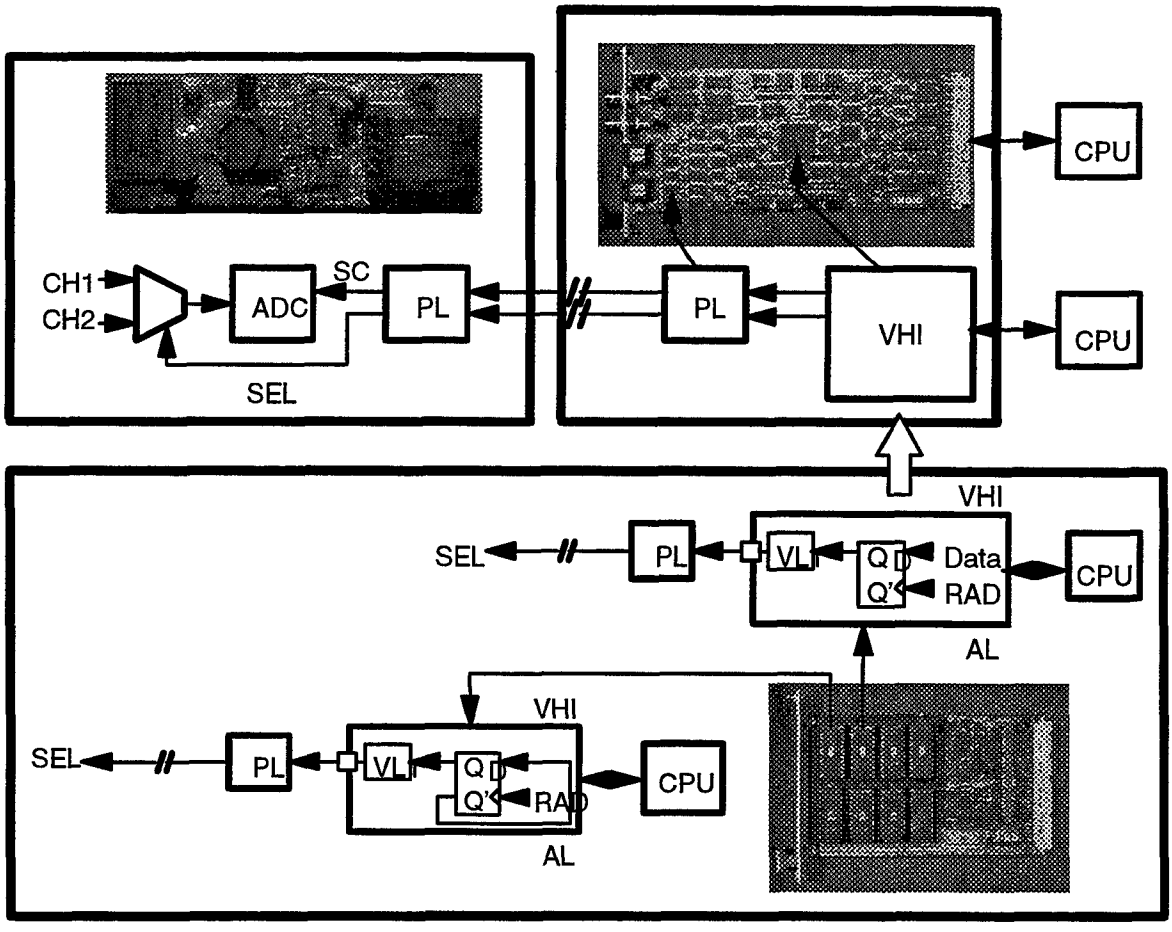


Figure 4.2 – More complete description of the example depicted in Fig. 4.1.

been implemented within the FD. This would yield a very complex FD and furthermore, rapidly increase the list of possible commands and control through the communication link, making the communication protocol complex with typically large communication overheads. Furthermore, if for instance the FD did not a priori predict an application requiring the sampled channel to switch after every conversion and did not implemented the divide–by–two DFF, then the throughput rate could be seriously affected because of the lack of a simple DFF. For these reasons, most FDs increase their flexibility with an onboard control processor. But in many cases, the throughput rate of a processor–based FD will be seriously degraded. For instance, a DFF is much simpler than a processor and can operate at a few orders of magnitude faster than a general purpose processor.

The UDEV2 shown in Fig. 5 is a good example to demonstrate how much a field device can be simplified with this concept, wherein most of the functionality is transferred to a FPGA–based central system. This device provides current output between ± 2 mA in 3.8 nA increments. It consists essentially of a few decoupling capacitors, a connector, an optional *SIP (Serial In–line Package)* termination resistor network for improved *Signal–to–Noise Ratio (SNR)* and two ICs: a DAC in a 28–pin DIP and the USO link receiver consisting of one 16–pin *Surface Outline Integrated Circuit (SOIC)* (mounted on the back of the PCB). This device can be located as far as 1200 meters without repeaters from the computer with a throughput rate exceeding 4.5 kHz, or maintain a maximum throughput

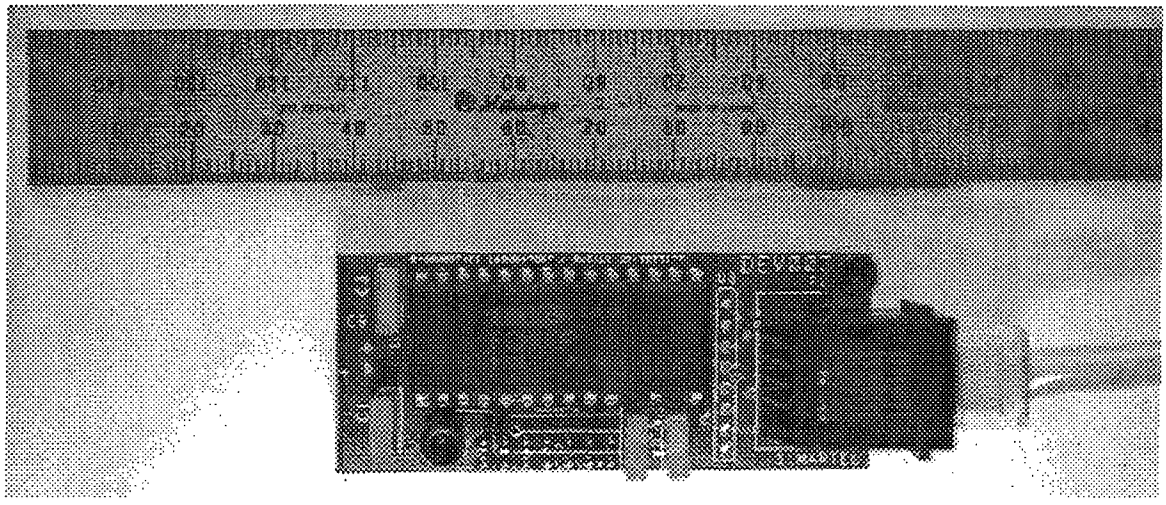


Figure 5 – The UDEV002 D/A–based UFD.

rate exceeding 450 kHz up to a maximum distance of 12 meters while maintaining a 20–bit resolution.

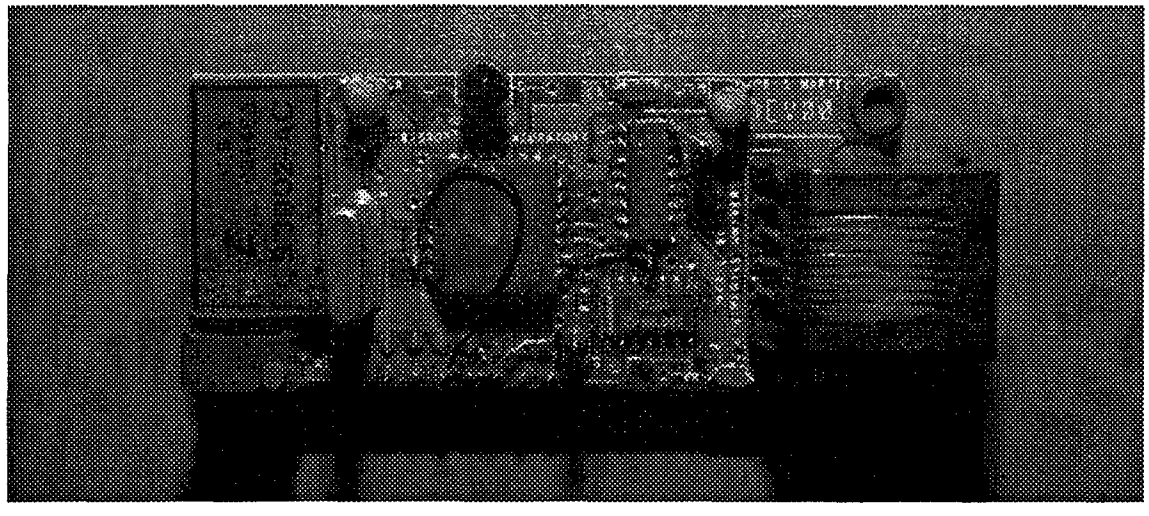


Figure 6 – The UDEV001 A/D–based UFD.

Using *Multi–Chip Module (MCM)* fabrication technology, the miniaturization of this field device could be implemented within a pen. This means that a single pen–like device linked with either radio waves, infra–red, copper–based or fiber–optic links could be used in applications where high–resolution analog signal generation is involved. Pseudo–random noise generators, wave generators or stimulators for medical applications are just a few examples of situations where the same hardware based on the VHI concept could be used. Furthermore, updates on these instruments can be done without modifying the physical hardware. An 18–bit A/D device referred to as UDEV1 has also been implemented with the same dimensions as the UDEV2 device and is shown in Fig. 6.

UFD Conversion Methods

A/D–Based UFD Conversion Methods

Several conversion methods[1] exist for signal acquisition. Details can be found in several books such as [Analog Devices, 1986][Higgins, 1983]. To help in selecting a proper ADC, [Price–1, 1992] is an excellent paper based on the principles, mode of operation, and characteristics of four types of converters. Price also in another article [Price–2, 1992] proposes such a choice with respect to speed–resolution tradeoff. But speed[2] and resolution are not the only factors to consider. The ADC needs to be optimized for the type of application [Schweber, 1991]. The issue here is to determine which A/D conversion method is better suited for a UFD for remote control applications. The main A/D conversion methods are listed below.

Review of the Main A/D Conversion Methods

• Integration–based ADC

Integration–based ADCs such as the *dual–ramp, dual–slope, quad–slope, single–ramp, and v/f converters* are better suited for applications in which a relatively lengthy time may be taken for conversion to obtain the benefits of noise reduction though signal averaging. Because of the long integration time, they are normally not well suited for hard real–time control applications but rather for applications such as digital voltmeters. Nevertheless, they are candidates for a UFD for some remote instrumentation applications. Its requirement for communication link bandwidth is very low such that multi–drop links within a centralized DCS can be easily implemented.

• Counter–comparator or tracking ADC

For *counter–comparator ADCs*, though very simple, have the disadvantage of limited speed for a given resolution since the conversion time for a full–scale change is equal to the clock frequency divided into the maximum number of counts. For example, the maximum throughput rate for 16–bit resolution ($2^{16} = 65536$ counts) is 122 S/s (Samples per second), and 1.526 kS/s with a synchronization clock frequency of 8 MHz and 100 MHz respectively. Even a variation of this converter and referred to as the *"servo" type* in which an up–down counter is used, is typically too slow to support high–performance hard real–time control applications. The counter–comparator ADCs also referred to as the *tracking ADC* can be faster than the *successive–approximation ADC* for the same resolution if the signal to be measured did not changed considerably since the last acquisition such as to involve only a small number of counts. Nevertheless, for multiplexed analog channels within a multichannel UFD with amplification set to take advantage of the ADC's dynamic range, variation in signal ampli-

1. Other alternatives which are not conventional and rather experimental can also be used such as in [Pouliquen *et al.*, 1991] where a description of a novel ADC designed to operate without any clocking circuitry is described. The design presented is a first–generation Gray–code algorithmic converter (GA–ADC) with a continuous analog transfer function. The continuous transfer function results in a digital output that is Gray–coded. Furthermore, all these conversion methods are typically implemented electronically but can also be implemented as an integrated–optic ADC [Karinskii, 1991], for example the one by [Shoop and Goodman, 1991] with 8–bit resolution and running at 1 GHz.

2. The performance of ADCs is also addressed in [Kasperovich, 1993].

tudes between successive acquired channels can be significant. Furthermore, tracking ADCs can have difficulty recording high slew rate signal since the rate at which the closed loop can follow an input signal change is defined as

$$\frac{dV}{dt} \leq 2^{-b} V_{FS} f_{sync} \qquad (1)$$

It is still a possible candidate for some remote instrument applications within a centralized DCS since its communication link bandwidth requirement is very low.

• Flash converter

The *"flash" converter* is of parallel type because 2^b comparators connected in parallel are required to support a b–bit conversion. The maximum accuracy of 'flash' ADCs is restricted since increased levels of resolution also increases exponentially the level of complexity such that the technology tends to approach the threshold of practicality at approximately 12 bits. The major advantage of this type of ADCs is a very high throughput rate. This high throughput rate is in the MHz range, which is far too fast for real–time control applications. Furthermore, distributed ADCs over a large area and connected to a central system would require I/O links and interfaces with very high bandwidth (possibly point–to–point fiber optic links) unless ADCs are not used in their full capabilities. Optimally, ADC–based UFD should be of level 2 or 3. It is thus not worthwhile to use flash ADCs in hard real–time control applications because accuracy is sacrificed for an unnecessary gain in maximum throughput rate which is well beyond the requirements of typical real–time control. This type of ADCs is better suited for instrument applications such as event detectors and recorders implemented in level 2 UFD.

• Sigma–delta or delta–sigma

Oversampled ADCs based on *sigma–delta ($\Sigma\Delta$)* modulation are attractive for *VLSI (Very Large Scale Integration)* because they are especially tolerant of circuit non–idealities and component mismatch. The sigma–delta converter is a classic demonstration of the tendency to increase the number of functions performed digitally since a sigma–delta modulator is mostly digital. Still, oversampled $\Sigma\Delta$ modulator has some points which could be improved [Sugitani *et al.*, 1993], such as inaccuracies based on the small input signal and integrator leak. Nevertheless, $\Sigma\Delta$ has some advantages. [Elektronik Praxis, 1992] discusses advantages of sigma–delta over flash and successive–approximation converters. Sigma–delta is promising for future systems, but presently, this technology is not quite mature enough to be used in high–performance control systems. For example, a relatively recent true 16–bit multibit sigma–delta ADC with digital correction [Sarhang–Nejad and Temes, 1992] achieves a maximum throughput of 20 kS/s with a signal–to–noise ratio plus distortion *(S/(N+D))* of 95 dB with an oversampling ratio of 128. This device, built approximately at the same time as our selected true 16–bit successive–approximation ADC compares relatively well except for the sampling rate which is quite inferior to its maximum rate of 100 kS/s. The sigma–delta–based UFD is a possible candidate for remote control applications provided that the capacity of the communication medium is sufficient for a given resolution. Its implementation within a centralized DCS configuration will likely become easier in the future.

• Successive–approximation

The *successive–approximation* technique offers the best alternative for high resolution hard real–time and the majority of instrument applications. It performs conversion through a binary tree–based method and resolves each *b'th* bit by comparing the bit weighted analog value with the $FSR/2^{b'th}$. The maximum throughput for true 16–bit resolution in a monolithic device is approximately 100 kS/s which is sufficiently high for a multichannel UFD in remote control applications. This is the method of conversion implemented within the EMAP system. Its feasibility within a centralized universal DCS will be demonstrated throughout the remainder of this dissertation.

D/A–based UFD Conversion Methods

The choice of D/A conversion methods for an UFD is quite restricted since the general principle of D/A conversion has remained relatively consistent during the last years. The typical approach used is briefly described in the next sections.

Theoretical Approach

The output of a DAC is adjusted digitally by selecting the appropriate resistors in the *resistor ladder* in which the current from a reference source V_{ref} will flow before reaching the amplifier. In the basic resistor ladder, each resistor is twice as large as its neighbor

$$R_n = 2^{n-1} R \ , \qquad (1)$$

such that if only switch n is closed, we will get an output voltage V_{OUT}

$$V_{OUT} = -\frac{V_{ref}}{2^n} \ , \qquad (2)$$

which can be generalized to consider more than one switch.

$$V_{OUT} = -V_{ref} \sum_{i=1}^{n} x_i \left(\frac{1}{2}\right)^i \ , \qquad (3)$$

where x_i represents the switch i and can take on the values 0 or 1.

Practical Approach – R/2R Ladder Network D/A

The previous approach is not used in practice because of limitations in resistance values, resistor adjustment, and settling time. For example, if $R = 10K$ to keep the current drain low, for a 20–bit D/A such as the one used in UDAC, $R_{20} = 2^{19}R \approx 5.2\ G\Omega$. Such large resistance values are difficult to achieve in IC's. Furthermore, for R_n to be meaningful, R_1 must be precise to one part in 2^n and would require advanced precision trimming of various resistor values. More serious is the switching time limitation set by the LSB resistor and the stray capacitance which can easily approach 100 pF. For our 20–bit DAC–based UFD, the settling time $T_{SE} \approx (5.2 \times 10^9 \Omega) \bullet (10^{-10} F) \approx 520\ ms$, which is far too slow for our target high–throughput real–time applications.

The R/2R ladder which is the most common D/A circuit and the one used in our UFDs overcomes these problems. The resistor trimming process is significantly simplified since only two resistor val-

ues are required, regardless of the number of bits. There is also no voltage changes at the switch terminals. With no voltage transients, RC settling problems are reduced. On the other hand, the throughput would still be mostly affected when the resolution increases (by the number of bits to transmit) and the time to settle to a lower error band.

Multiplying DAC

The multiplying DAC is a more specialized device which can be used in control systems. The digital inputs are isolated by MOSFET gates while V_{ref} creates current flow between source and drain of FET analog switches and can be in either direction, such that $V_{OUT} = -(A \times D)$, A being the analog input signal and D the magnitude of the binary input. This type of DAC could be useful in variable gain control systems and/or offset control within the UFD.

For simplicity, UDAC does not use multiplying DACs, and the gain is set by installing a fixed resistor accordingly to the required dynamic operating range[3]. No offset control has been implemented. The offset generated by UDAC is typically taken into account by sending output values where the offset correction has been applied. This is a possible option since the UDAC digital resolution is greater than EMAP, allowing 4 extra bits to compensate for offsets at the expense of increased channel bandwidth. Notice that the elimination of the offset control circuit at the remote site to simplify the UFD though additional bits is likely to become a serious alternative within the centralized DCS concept, specially with the advent of high–bandwidth fiber optic links.

Our system UnEmap on the other hand, makes extensive use of multiplying DACs to achieve offset corrections in analog form, so as to gain *SNR* by exploiting the available *FSR*.

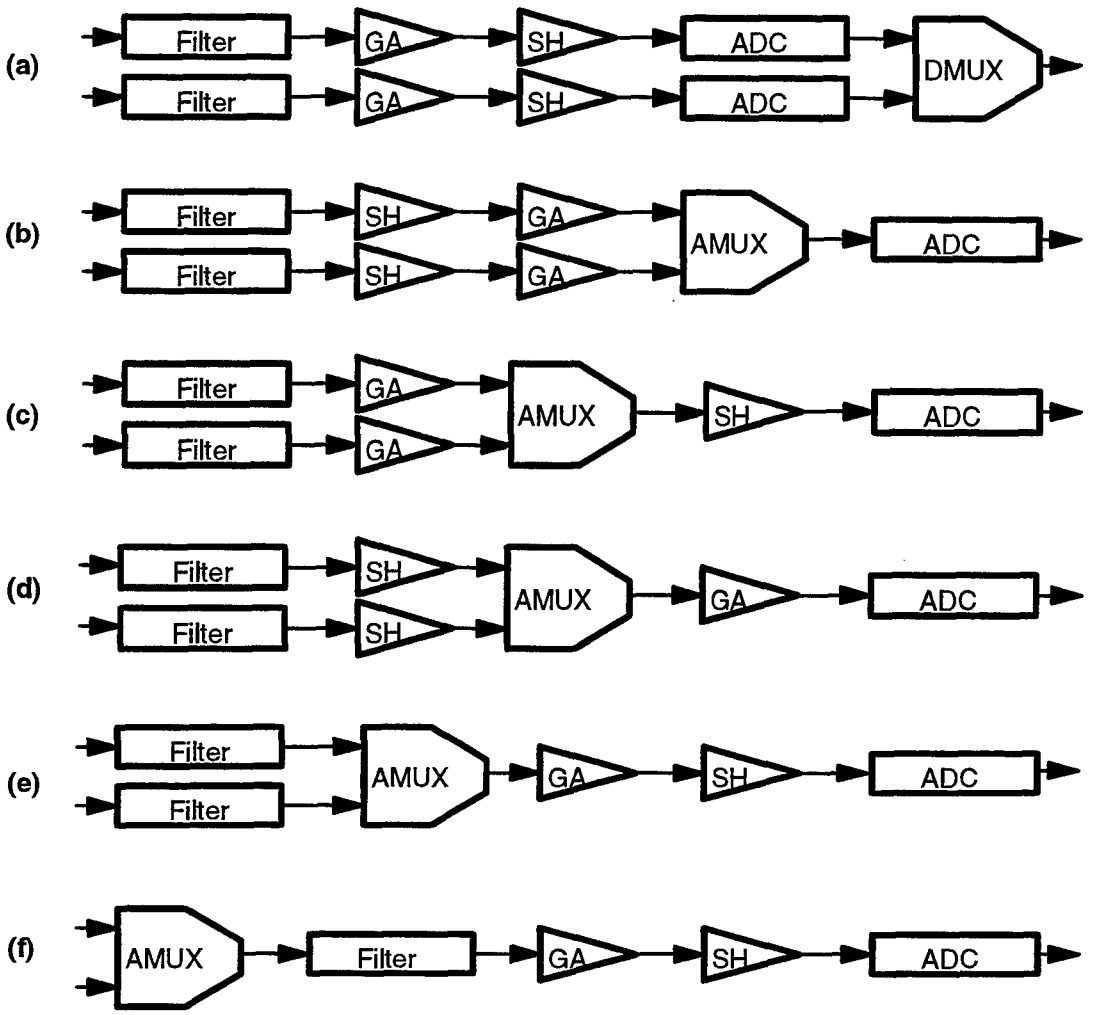
DAC Resolution

Current high–performance D/A instrumentation modules use 16–bit DACs such as [TASCO, 1991] for example. Several PC–based D/A cards use 12–bit DACs. Our system is based on 20–bit DACs. This is the best resolution commercially available. Nevertheless, efforts are still presently be in to implement a 24–bit DAC [Craven, 1993].

6.3 Multichannel A/D–Based UFDs

Although the complexity of the architecture of such UFD can increases, a high–level of integration would likely be possible in a near future. For instance, a data acquisition or D/A system can presently reside into a single IC. For example, the Lm12458 developed by National Semiconductor [Fossati, 1992] uses CMOS technology to integrate into a single chip, all the analog and digital functions of a complete 12–bit data storage system. But at present, high–performance systems must rely on several ICs. Implementing such UFDs is more than just connecting an ADC or DAC, the real challenge is the design of the peripheral circuit and whether it is placed within or outside the same die used to support the conversion [Kester–2, 1992].

3. The maximum and minimum output voltages can be changed.



Note: for configuration (C) the filters can be before or after the GAs.

Figure 7 – Main multichannel A/D–based UFD architectures.

Proposed Multichannel A/D–based UFD Architectures

Also, it could remain a sensor bus connected to HANET in case where the communication interface complexity must be minimized and the level of performance of the IEEE–1394 is not necessary for a particular field device. The schemes depicted in Fig. 7 are very simple. In reality, the data within a multichannel UFD will flow through one or more filtering blocks for antialiasing and possibly AC–coupling for baseline drift removal, gain and S/H amplifiers, ADCs, and multiplexers prior to the communication PL.

Architec-ture	Advantages	Disadvantages
A (UDEV1)	Fastest throughput with simultaneous signal tracking and independent gain and filtering per channel	Expensive approach with maximum throughput higher than required for typical real–time control applications, possible TX noise induced during acquisition and possible droop at low sampling rate
B	Very high throughput with simultaneous signal tracking and independent gain and filtering per channel	Still a relatively expensive implementation with possible TX noise induced during acquisition and possible droop at low sampling rate
C (EMAP)	High throughput with independent gain and filtering per channel, possibility of TX noise to be induced during acquisition is eliminated as well as voltage droop at low sampling rate	Average implementation complexity with maximum throughput limited by a minimum tracking period, simultaneous tracking is not possible and jitter must be minimized with a high scan rate
D	High throughput with simultaneous signal tracking and independent filtering per channel	Average implementation complexity with maximum throughput limited by the gain amplifier settling time, possible TX noise induced during acquisition and possible droop at low sampling rate
E	Simpler implementation with independent filtering per channel and no voltage droop at low sampling rate	Maximum throughput limited by the gain amplifier settling time and a minimum tracking period once the output of the amplifier is settled within the desired error band, simultaneous tracking is not possible and jitter must be minimized with a high scan rate, possibility of induced TX noises during acquisition
F	Simplest implementation with no voltage droop at low sampling rate	Maximum throughput limited by the gain amplifier settling time and a minimum tracking period once the output of the amplifier is settled within the desired error band, simultaneous tracking is not possible and jitter must be minimized with a high scan rate, possibility of induced TX noises during acquisition and filter time constant may reduce significantly the overall throughput

Table 1 – Main advantages and disadvantages of various A/D module architectures.

The architecture behind a specific analog interface system must be developed to balance the workload among all subsystems. This architecture is often a tradeoff between complexity and performance in term of accuracy, throughput, and number of channels. Fig. 7 shows some of the main A/D–based UFD architectures which could be implemented in a real–time remote control system. The list of A/D module architectures shown in Fig. 7 is not exhaustive but it demonstrates an example of tradeoff between performance and implementation complexity (cost). For simplicity, Fig. 7 only shows two channels per implementation but the number of channels will be mostly dependent upon the *sam-*

pling rate per channel ($F_{S/C}$) that we want to achieve. The main advantages and disadvantages related to each implementation are listed in Table 1.

The configuration depicted in Fig. 7a can be implemented with a minimum complexity UFD per channel such as the UDEV1 [Fig. 6] for applications requiring a very fast throughput and high accuracy. Nevertheless for several known reasons, multichannel UFDs must be considered to minimize cost. For minimum latency and for several reasons concerning the modular interchannel skew, the configurations depicted in Figs. 7b and 7d can be excluded for the implementation of the simplest universal FD. This also resolves the second issue of modular channel drift. The last major architectural issue for a multichannel UFD is concerned with the settling error.

Settling Time Issue

The minimum allowed worst case (i.e. at maximum throughput rate) settling time denoted *min* T_{SE} is obtained with $(T_H \parallel T_C)$ with the DSH configuration and can be computed as

$$\min T_{SE} = \frac{1}{\max F_{S/C}}, \; N_C > 2 \; , \tag{4a}$$

$$\min T_{SE} = \frac{2}{\max F_S}, \; N_C = 2 \; , \tag{4b}$$

since typically $T_H < T_C$. This condition holds for EMAP. With the direct or ASH configuration $(T_H \rightarrow T_C)$ [Chapter 8], we get a slightly lower *modular input throughput* ($P_{IN/M}$) but equivalent *channel input latency* ($L_{IN/C}$) since tracking and conversion processes cannot be done simultaneously on the same sampled signal but only on successive samples. In this case, the worst case settling time occurring at maximum sampling rate becomes, neglecting the very short internal card's synchronization delays,

$$\min T_{SE} = \frac{1}{\max F_S} - T_H, \; N_C = 1 \; , \tag{5}$$

which is 8 μs for EMAP operating at the maximum resolution mode[4]. Notice that Eq. 5 also holds with the DSH configuration when $N_C = 1$, in other words, there is no advantage to use the DSH when only one channel[5] is sampled on a particular module. The best case settling error which can be tolerated corresponds to 0.5 LSB of the ADC's *FSR* and can be estimated as being $e_S = 68.5$ μV for EMAP.

Since the amplifier's data sheets do not typically provide sufficient specifications about the settling time at various conditions; unless a SPICE model is well developed, a method in estimating a first order approximation settling time $\hat{T}_{SE}$ where the signal does not enter the slew rate limitation (4V/μs for EMAP) and which is typically the case when the amplifiers are located at the front–end, i.e. prior to the analog multiplexers, where the 3 dB corner frequency f_{3dB} of the amplifier is known and its roll–off is assumed to be 20 dB/decade for at least a decade above the 3 dB corner frequency is proposed in [Johnston, 1992] and is based on the following equation which yields a good approximation of T_{SE} down to e_S or 0.5 LSB.

4. Extension of the coarse charge of the S/H circuit by an additional fine charge period.
5. We consider the case where two S/H amplifiers are connected to the same signal input as two channels.

$$\hat{T}_{SE} = -\left(\frac{1}{2\pi f_{3dB}}\right) \ln \left| \left(\frac{2^b - 0.5}{2^b}\right) - 1 \right|, \qquad (6)$$

and which can be simplified to the following expression

$$\hat{T}_{SE} = \frac{0.11\ (b+1)}{f_{3dB}} \ . \qquad (7)$$

By entering the amplifier's (AMP02) characteristics used in EMAP for various gains into Eq. 7, we obtain for a gain of 1, 10, and 100 to 1000 an estimated settling time of 1.6, 6.2, and 9.3 μs respectively.

From the estimated results, with the initial gain of approximately 10, sampling rate of 100 kS/s for one channel is possible without loss of accuracy due to settling errors. A gain of approximately 100 and greater would be possible with a $(T_H \parallel T_C)$ configuration if noise contamination can be controlled, but for optimal transmission noise immunity $(T_H \rightarrow T_C)$, a reduction of throughput for slew limited signals would be necessary to support higher G.

These results suggest that for a very low value of N_C with high b and F_S, the gain should be maintained relatively low. On the other hand, several applications such as recording electrophysiological signals in the microvolt range such as EEG, requires a high input gain. The optimal configuration is then to pre–amplify as much as possible at each input such that amplification after the analog multiplexer, provided that a range amplifier is implemented, can be maintained at a low value sufficient to avoid settling errors. Ideally, such pre–amplification should be performed prior to filtering at the maximum possible value to improve the *SNR* without causing saturation.

While this is a design issue, it is obvious that the most commercially used configurations depicted in Figs. 7e and 7f are inadequate for a high–resolution high–performance UFD unless we impose restrictions in the scan list configuration. Considering that sharper signal transitions will arise after the analog multiplexer with possible high gain values, the configuration depicted in Fig. 7c, and which is the one used to implement EMAP, is much more appropriate.

EMAP Interface Module

The block diagram of an EMAP card is shown in Fig. 8 with the actual card shown in Figure 9. The cardiac mapping system had 16 of these cards in order to provide up to 512 bipolar channels. The EMAP extension card is a plug–in unit which is essential for studies involving defibrillations and/or analog filtering. Such overvoltage protection is not only required in medical studies but can also be a prerequisite in control systems [Sahle, 1994].

Each EMAP interface card has 64 inputs divided into 32 bipolar analog channels. Each channel has its own instrumentation amplifier to increase the overall system performance by reducing significantly the total required settling time. The reference of the amplifier can be set to the analog ground or set from an external source. The signal passes through an analog multiplexer controlled by a pre-

load channel scanner. The amplified signal then passes through a broadband low–pass filter and then to an analog duplexer prior to reaching the analog–to–digital conversion block. The analog duplexer is connected to an autoscan controller which can switch between the channel group 0–15 or 16–31 automatically or from direct computer control. A sample–and–hold amplifier tracks the signal prior to the start of the A/D conversion. The method of conversion is based on the successive–approximation algorithm.

Additional S/H amplifiers are available for calibration. The calibration data, which are computed automatically prior to the start the conversion process when EMAP is initialized by the UNIMA system, are stored in the calibration SRAM. This calibration is based on 18–bit resolution. A charge redistribution D/A converter uses these data and the S/H data to convert the sampled signal into digital form. All conversion steps are synchronized by a micro–controller running at 8 MHz. An external S/H signal is also available for applications that require simultaneous S/H on all channels. The conversion block sends serially the most significant 16 bits of the conversion to the UNIMA system through a communication PL consisting of optocouplers and an RS422 interface. The card supports 32 different operational modes to match the user's requirement. A control block is provided to decode and preload the scan list and the analog–to–digital conversion block. Power distribution and several voltage regulators are also provided.

An important aspect to consider is that the digital functions susceptible to decrease the *SNR* by digital noise contamination have been minimized through the COR concept which was possible with the VHI described in Chapter 5. The dimension of the A/D module as shown in Fig. 9 may appear to be large, but it is in fact much smaller than any A/D interface cards offering similar characteristics[6]. Furthermore, EMAP could have been reduced further with more advanced packaging techniques.

A/D Conversion Node

The A/D–based UFD can be built with an A/D conversion node and an optional signal conditioning node. This section deals with the main issues concerning the A/D conversion node.

Basic Architecture

The implementation of the multichannel A/D–based UFD as depicted in Fig. 9 had the functionality to test our concepts but unfortunately it had some drawbacks. For instance, because the A/D conversion node has been built initially to be conveniently inserted into a conventional crate, a portion of the signal conditioning node was also integrated to fill the space available on the card and which affected the flexibility of the system.

Ideally, the A/D conversion node should be self–sufficient and implemented with minimum electronics. In other words, it should be a single channel level 1 UFD similar to the one shown in Fig. 6. The block diagram of our proposed A/D conversion node is shown in Fig. 10. It consists of an antialiasing filter, a range amplifier, a S/H amplifier, an ADC, and the communication PL.

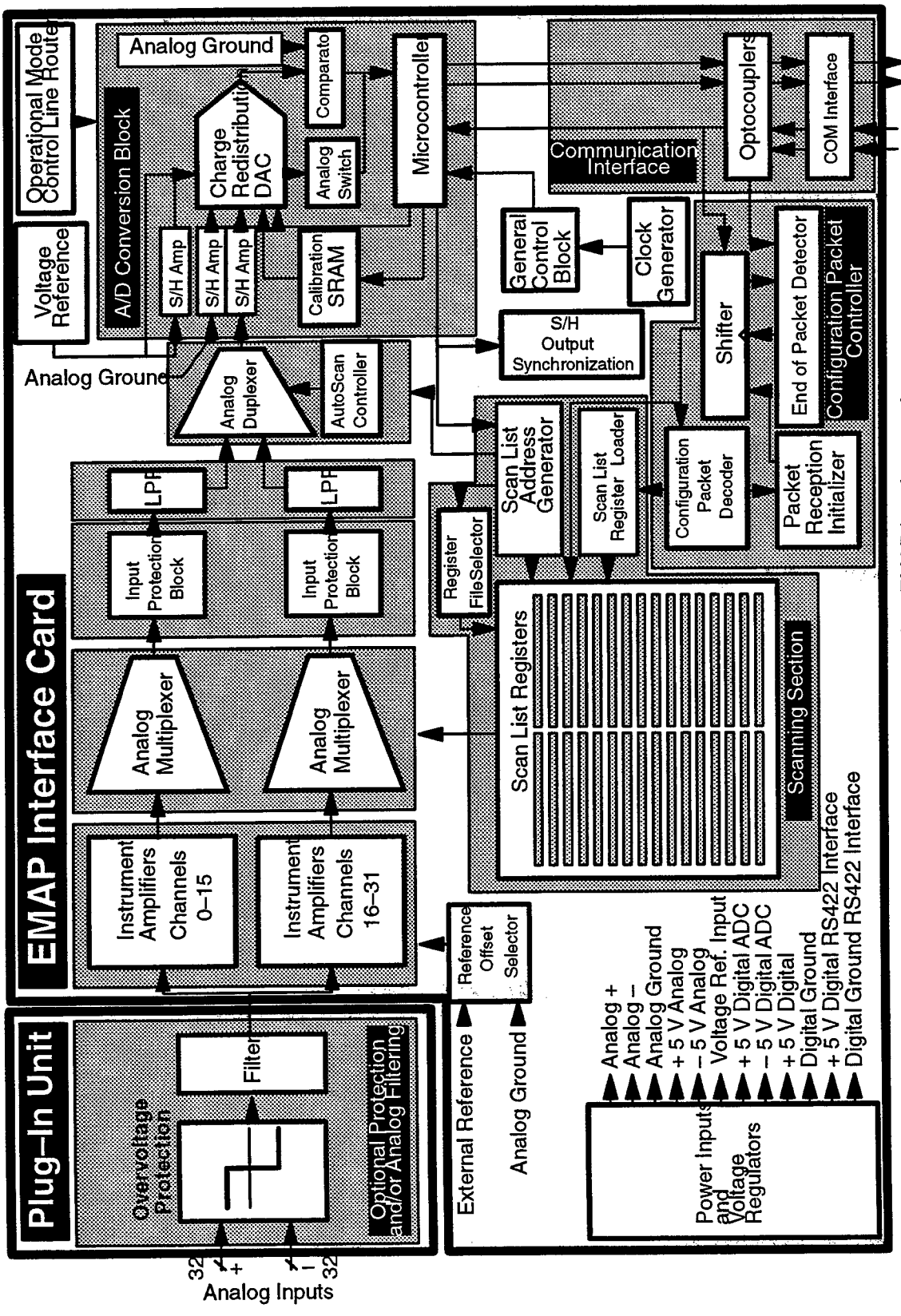


Figure 8 – Block diagram of an EMAP interface card.

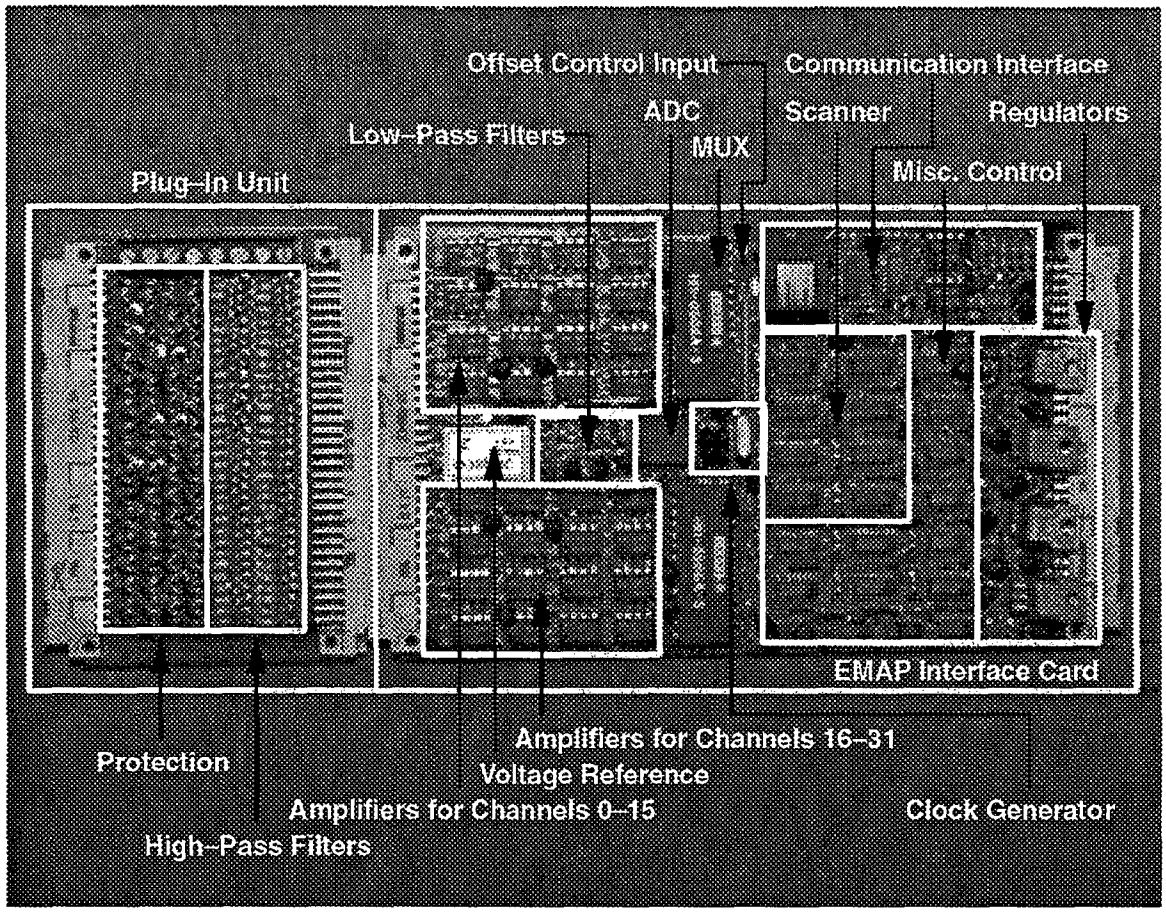


Figure 9 – The EMAP interface card.

The *Range Amplifier (RA)* has not been implemented initially in EMAP, and this was considered a drawback since the pre–amplified signal could have been adjusted closer in many cases to the ADC's *FSR* in order to improve the accuracy of the system. This has been corrected in UnEmap. Furthermore, by implementing a range amplifier as part of the A/D conversion node, one can have a self–sufficient UFD for typical low demanding applications. Nevertheless, the RA should have high bandwidth and be configured for low amplification because of settling error consideration.

There are two major issues to consider in the implementation of this high–performance A/D conversion node. First, the acquisition time must typically be minimized and secondly, because of the expected improved durability achieved with universality including the UFD, techniques should be implemented to maintain the high resolution over a possible long usable life.

Minimization of the Acquisition Time

The minimization of the acquisition time is an important issue in high–performance remote control applications for $(T_H \rightarrow T_C)$ configurations since it will have a direct impact on the throughput rate which is approximately the inverse of the sum of the acquisition and conversion times. For instance, in EMAP the acquisition time alone may contribute to approximately 20% of the minimum sampling

6. The reader is encouraged to compare the complexity with other A/D modules, even the ones that do not have an amplifier per channel.

period. When $(T_H \parallel T_C)$ is implemented with an additional S/H amplifier, the acquisition time can be extended up to the conversion time.

Figure 10 – A simple A/D conversion node.

The input signal slew rate for a given resolution is limited by the system *aperture time* which is often equivalent to the ADC conversion time. This is the case with several A/D conversion techniques. To reduce the error or allow higher slew rate signals at the inputs, the aperture time is reduced by implementing a S/H amplifier[7] prior to the ADC. This is a common technique used with successive–approximation converters. When a conversion command is issued, the A/D conversion block enters *tracking* mode. The period between the time where the UFD enters the tracking mode and the time where the S/H command is issued is referred to as the acquisition time or tracking interval. During the acquisition time, an input buffer amplifier provides the bulk of the charge on a binary–weighted capacitor array. This step is referred to as the *coarse–charge* and requires in the case of EMAP, 750 ns to get an accurate sample.

Ideally, the UFD should allow the coarse charge period to be extended to any value predefined by the input signal characteristics and the required accuracy. This is unfortunately too complex to be implemented. Instead, to increase the accuracy of the signal acquisition, following the coarse–charge, the track–and–hold circuitry bypasses the buffer amplifier and connect the input directly to the capacitor array. This adds another 1.125 μsec to allow the charge on the array to accurately settle to the input voltage. Depending on the applications and the characteristics of the signals, the user has the choice to integrate the *fine–charge* as part of the whole acquisition process.

The accurate fine–charge unfortunately increases the latency time and must not be used when the input signals have relatively high slew rates. Typically when used in cardiac mapping applications, EMAP was used with both the coarse–charge and the fine–charge in order to get accurate recording of the typically slow slew rate electrophysiological signals. The tracking mode was selected during the EMAP system configuration phase.

Maintaining Accuracy over Time

Several high–performance ADCs are available and a list of these devices with a summary of their main characteristics can be found in [Child, 1994]. The A/D conversion block in EMAP is based on the CS5101A [Crystal Semiconductor, 1991]. It was implemented in such a way that 16–bit accuracy could be obtained and maintained over time. To achieve this goal, the charge redistribution DAC is based on capacitors instead of the typical resistor–based network. The resistor materials have inadequate *thermal tracking* to achieve the required accuracy. Long–term accuracy will depend upon aging characteristics of the components and the characteristics of the resistor materials change with time. *Resistor–based A/D converter* typically uses *trimming* techniques at the factory to achieve initial ac-

7. A nice paper on sample–and–hold amplifiers can be found in [De Vittor, 1992].

curacy but lack mechanisms to maintain this accuracy over its lifetime. EMAP uses *self–calibrating capacitive–based A/D conversions* [Johnston, 1990; Johnston, 1993] to correct the problems encountered in resistor–based systems. The capacitive–based technology does not require trimming at the factory but is rather based on self–calibration managed by a micro–controller every time that the EMAP system is initialized. *On–site calibration* is performed on each capacitor taking into account the conversion block factors that could degrade the overall accuracy, and the resulting calibration information stored in the calibration SRAM provides 18–bit resolution, ideally yielding accuracy of 1/4 LSB at 16 bits.

Signal Conditioning Node

The signal conditioning node is the only node where universality cannot be achieved with a simple implementation. This is due mainly to the fact that the signal conditioning node is the front–end interface to the external world and will be subject to various requirements that must be performed using analog techniques which cannot be emulated with a VHI.

We identified two approaches to provide universal signal conditioning nodes. One approach is referred here to as resource re–use. UnEmap relies very much on this approach since it has a fair number of analog functions. The other approach is to implement them by layers, initially by functions as suggested in Fig. 7, i.e. filtering, pre–amplification, sample–and–hold, and multiplexing layers. To these layers, additional stages could be added, for instance, an input protection layer. While not optimal, interconnecting pre–defined layered signal conditioning node, will enhance in most cases the universality of the front–end interface and minimize the need to build custom interfaces.

Multichannel D/A–Based UFDs

Presently, the best architecture for a multichannel D/A–based UFD relies on one DAC per channel. This is the approach used to implement UDAC. A simplified block diagram of an UDAC interface card and the actual board are shown in Figs. 11 and 12 respectively. Each card supports up to 16 20–bit channels where each channel has its own R/2R ladder network DAC (PCM63P) [Burr–Brown, 1990].

Sixteen single–channel UFDs such as the one depicted in Fig. 5 could have been implemented instead of the implementation shown in Fig. 12. Unfortunately, such alternative would result into a relatively high number of communication links. With our experience with UDAC, we came to the conclusion that the single–channel UFD as implemented in Fig. 5 was not quite appropriate. Instead, the communication PL should be a separate node and this has been implemented in the UnEmap system. With such an implementation, a digital demultiplexer, referred to later on as the demultiplexing node, could be inserted between the PL node and the D/A conversion node in order to implement the simplest D/A–based UFD with a single communication link. The demultiplexing node.

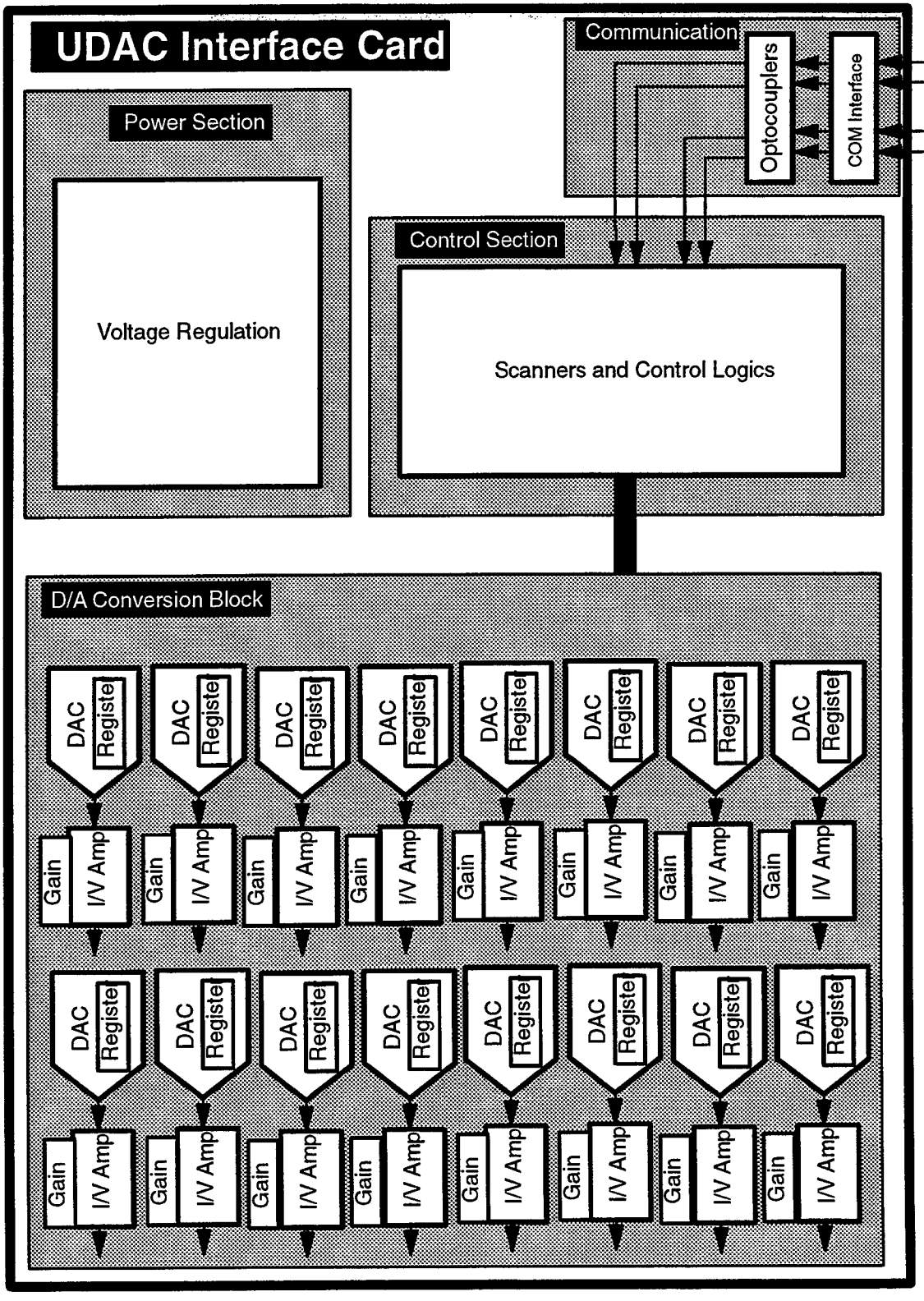


Figure 11 – Block diagram of an UDAC interface card.

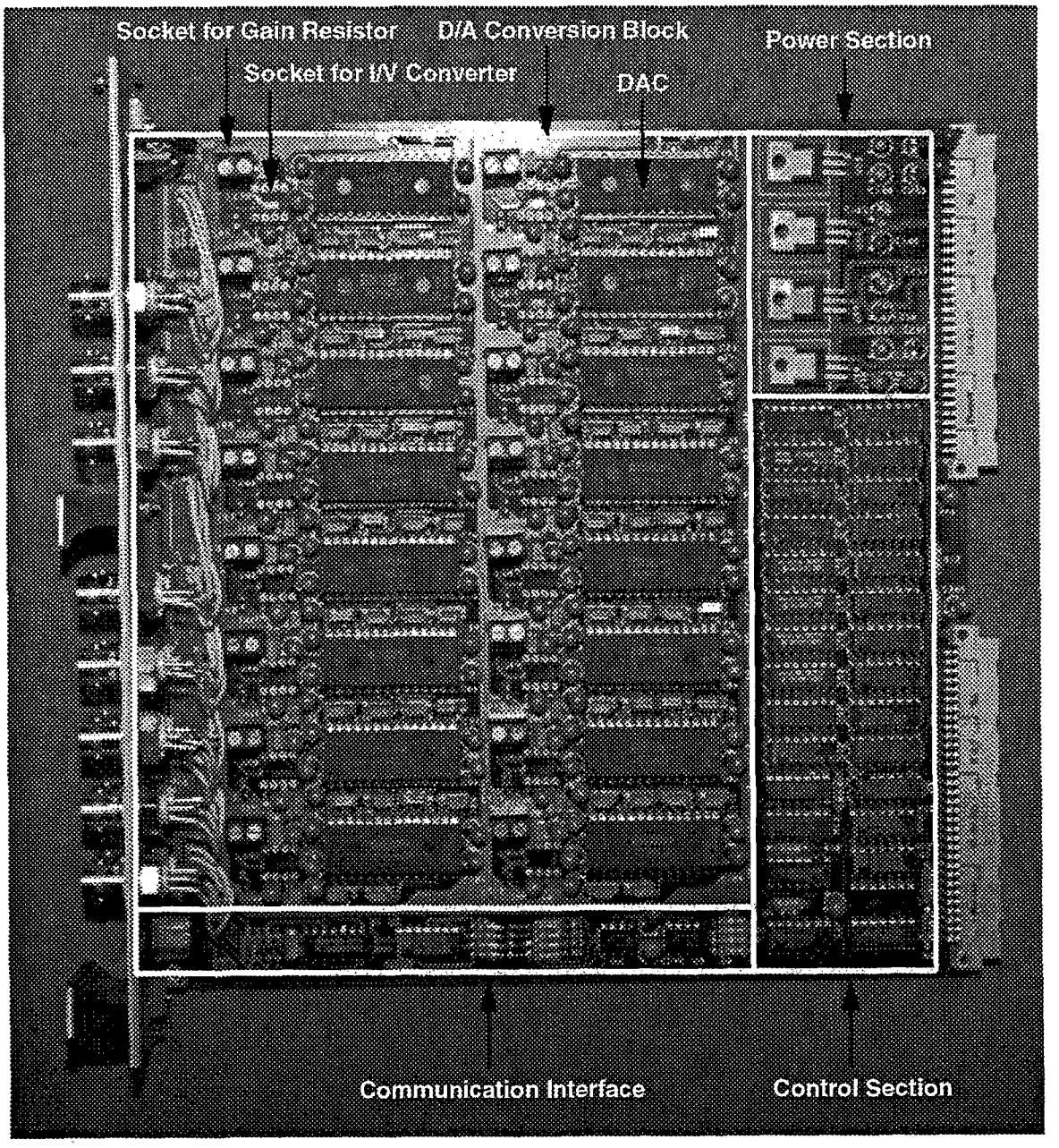


Figure 12 – The UDAC interface card.

Other Practical Issues for UFDs

Configuration and Initialization

A high–performance and flexible DCS requires that the modules can be independently configured and pre–programmed for specific requirements to minimize the number of computer accesses through the network during the real–time execution phase. The real challenge is to support configuration and initialization capability through distributed modules efficiently using minimum additional hardware. The literature mentions some approaches to configure DCS [Takura *et al.*, 1991]. Such approaches rely on intelligent FDs. The question which thus arises is whether it is possible to configure and initialize very simple FDs such as UDEV1 and UDEV2 with various configuration options and without an onboard processor to decode the configuration packets. The possibility of performing the configuration and initialization within an universal centralized DCS relying on very simple remote I/O devices is demonstrated through a simple example.

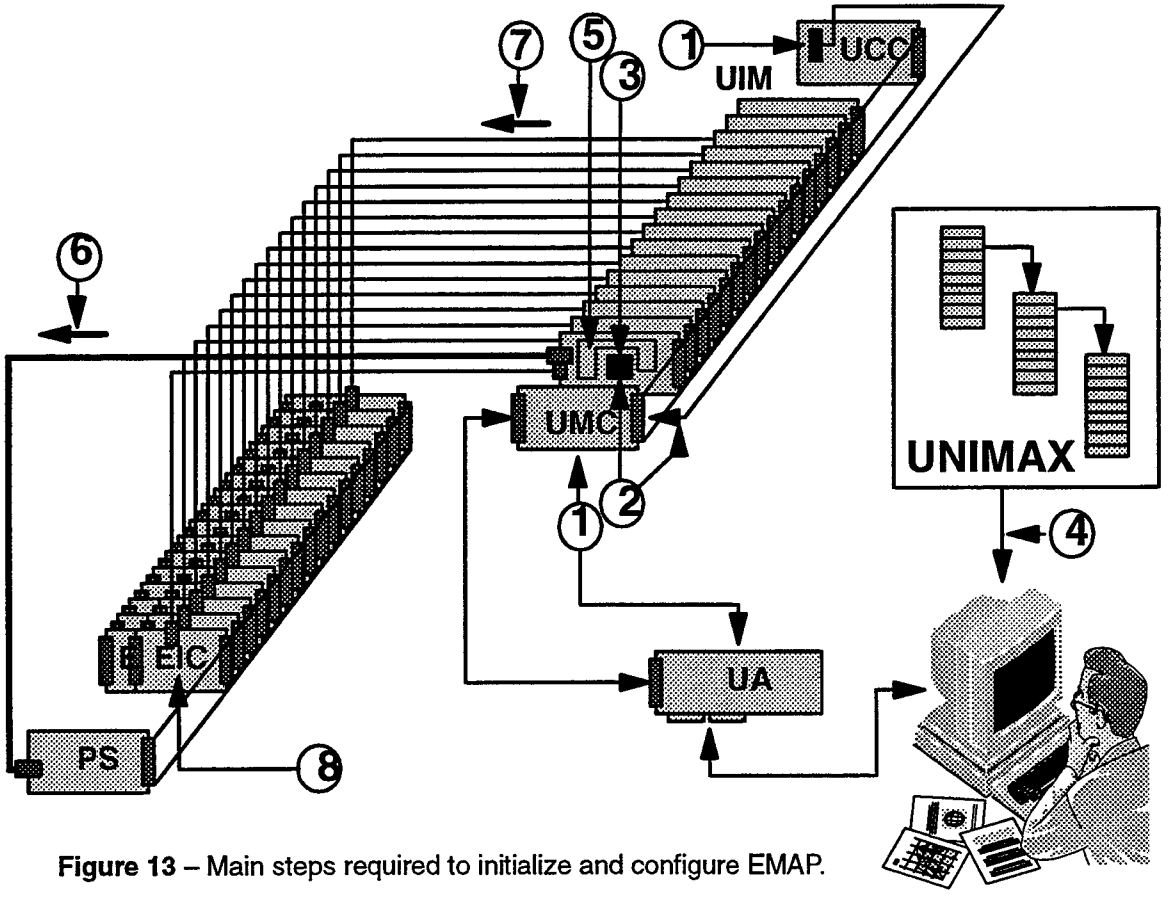


Figure 13 – Main steps required to initialize and configure EMAP.

Configuration and Initialization with UNIMA–EMAP

The UNIMA–EMAP configuration has been considered here to provide such an example of configuration and initialization since it will be the same sequences used prior to perform direct cardiac mapping. More than just switching the power on is necessary before signal acquisition starts in EMAP.

The system must perform a series of tasks prior to be coming operational. For instance, the main steps involved to initialize EMAP[8] are listed in Table 2 and are shown in Fig. 13. These steps are entirely user–transparent and require less than 2 seconds to complete.

This period of time, unlike any other systems, consists of 4 elements, mostly dedicated to constructing the required hardware functions in the VHIs, to initialize not only conventional registers but also newly implemented virtual hardware registers, to implement the most appropriate synchronization protocols between all modules involved such that the whole structure behaves like a unique system, and finally to perform calibration. All concepts discussed in this report have been implemented within this structure. It has performed very well in all applications involving not only a few channels, but also in the 512–channel cardiac mapping system.

Thermal Considerations

When the configuration packet controller releases control to the A/D conversion block, a calibration cycle begins[9]. The charge redistribution DAC consists of several capacitors in parallel which can be manipulated to adjust the overall bit weight. To achieve 16–bit accuracy, the micro–controller precisely adjusts each capacitor from the analog ground and the voltage reference, with a resolution of 18 bits. The resulted calibration coefficients are stored in the onboard SRAM. The whole process requires 1.441 seconds to complete. The voltage reference is based on a hybrid circuit providing very accurate 4.5 volts corresponding to the maximum positive dynamic range at the input of the A/D conversion block. This calibration is only valid for the A/D conversion block and does not compensate for any gain errors generated by the instrumentation amplifiers and offsets generated by both the amplifiers and the analog multiplexers. These errors are corrected through calibration tables containing data for each channel after expiration of the warm–up time. These data are added or subtracted to the acquired data through software. Recalibration of the A/D block is typically not required throughout the acquisition phase if the ambient temperature remains stable. The self–calibration process is initiated every time following the reconfiguration of EMAP which can be initiated for maximum accuracy prior to each acquisition phase.

Sequence	Step	Description
1	UNIMA–IBM communication and selection of UNIMA system implementation	The UNIMA adapter is set in the UEL (UNIMA External Link) mode, access speed controls are set and commands are sent to the UNIMA master controller to select the UNIMA system implementation required from the UNIMA configuration card
2	UNIMA ASICs configuration	The UNIMA master controller accesses the bitstream file in the configuration card and downloads it into all system's programmable ASICs in a daisy–chain fashion

8. The list of initialization steps provided in Table 6.2 is very much simplified and does not represent a complete description of the entire initialization phase. It should then be consulted as a general description of a unique system's initialization sequence.

9. EMAP uses state–of–the art ADCs with sufficiently good characteristics. In some cases, the system might require better characteristics than what are possible with the available ADCs in term of accuracy. The paper by [Lee and Song, 1993] proposes a technique which calibrates digital outputs obtained from uncalibrated ADCs after conversions.

3	UNIMA ASIC design identification number	When the ASIC configuration bitstream is downloaded and all FPGAs have sent a flag back to the master controller indicating that the configuration had been performed successfully, the UNIMA master controller releases the UNIMA bus to allow accesses by the IBM computer. The IBM system then accesses each UNIMA interface module ASIC identification number identifying the type of virtual interface implemented and the functions capabilities
4	System's description	UNIMAX accesses specific files depending upon the ASIC identifications throughout the UNIMA system and the user's selected operational modes
5	UNIMA interface module supporting block configuration	UNIMAX configures the semi–adaptive blocks (ASIC extension capabilities and supporting circuitries on each UNIMA interface module) accordingly to the ASIC design identifications, external systems connected (such as EMAP or UDAC), operational modes selected and user's requirements
6	EMAP interface cards' initialization	The first UNIMA interface module initializes all EMAP interface cards through the power sequencer
7	EMAP configuration	System's configuration packets for the scanners and the A/D operations are sent by UNIMAX through the RS422 links to all EMAP interface cards
8	EMAP system auto–calibration	When the last configuration packet has been received and decoded, EMAP starts an auto–calibration procedure for A/D conversions. When the auto–calibration is completed, EMAP is finally ready for reliable data acquisition

Table 2 – Main steps required to initialize EMAP.

Calibration can be seriously affected by change of both a device's temperature and ambient temperature. For instance, typical amplifiers have a warm up time between 2 and 3 minutes. The problem when two or more UFDs are required within the same site is that the ambient temperature may increase and it may become difficult to maintain it stable within a range that will not affect the accuracy of the system. Since both the number of channels and the resolution are likely to increase in several applications, including cardiac mapping, providing adequate remote calibration may become more difficult.

For instance, when the EMAP interface cards are distributed throughout an area, minimum warm–up time is required prior to the calibration. This is different when the number of channels implemented in a specific location is significantly high. This was the case when EMAP has been used in multichannel cardiac mapping, where a very high–density multimodule system was implemented locally, requiring a minimum warm–up time of eight minutes to compensate any self–heating effects causing voltage drifts through the very sensitive analog circuits. Fig. 6.14 shows the packaging density inside the chassis of an EMAP with 16 tightly linked interface cards for the 512–channel electrophysiological mapping. The results based on experiments to measure the variations in temperature within such a configuration are shown in Fig. 15.

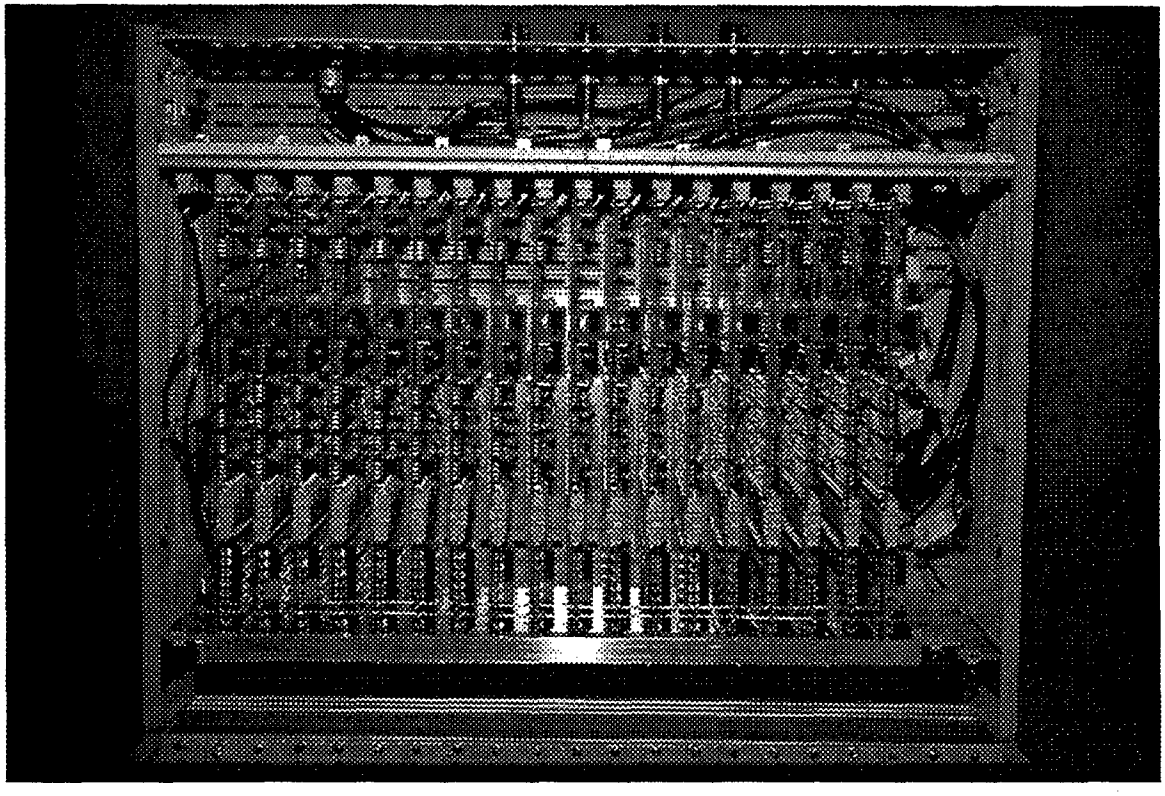


Figure 14 – Inside the 512–channel EMAP system.

After the warm–up delay of 8 seconds, the temperature stabilized within 0.5 degrees Celsius and was maintained with several fans. The cooling system was then capable to avoid a reduction in the accuracy of the measurement due to temperature by maintaining the drift variations well below the 0.5 LSB threshold since the drift of the amplifiers used in EMAP is characterized as <4μV/°C or <1/3 LSB/°C [Crystal Semiconductor, 1991].

As shown in the figure, the temperature inside the chassis could be maintained at about 6 degrees above the room temperature. As soon as the power was shut down, all 7 fans required to cool the system were also shut down. As shown in the graphic, the temperature without ventilation rises rapidly to a peak value of 10 degrees above the room temperature and then drops slowly. With the power on and no ventilation, the temperature rose quickly to a critical temperature sufficient to damage the system permanently within a few minutes. A lot of effort has been dedicated to find a configuration to cool efficiently the whole system while maintaining an air flow over the analog parts relatively constant in order to minimize the level of variations in temperature.

Such results suggest that packaging several UFDs within the same chassis is not a good approach since the cooling requirement changes with the number of modules, the times at which remote calibration can be done vary, and for high resolution UFDs, the stability control of the ambient temperature may become a real headache. Furthermore, the chassis itself is not universal since it has a limited number of slots and/or connectors. Therefore, each UFD should be enclosed within its own enclosure and with its own ventilation if required.

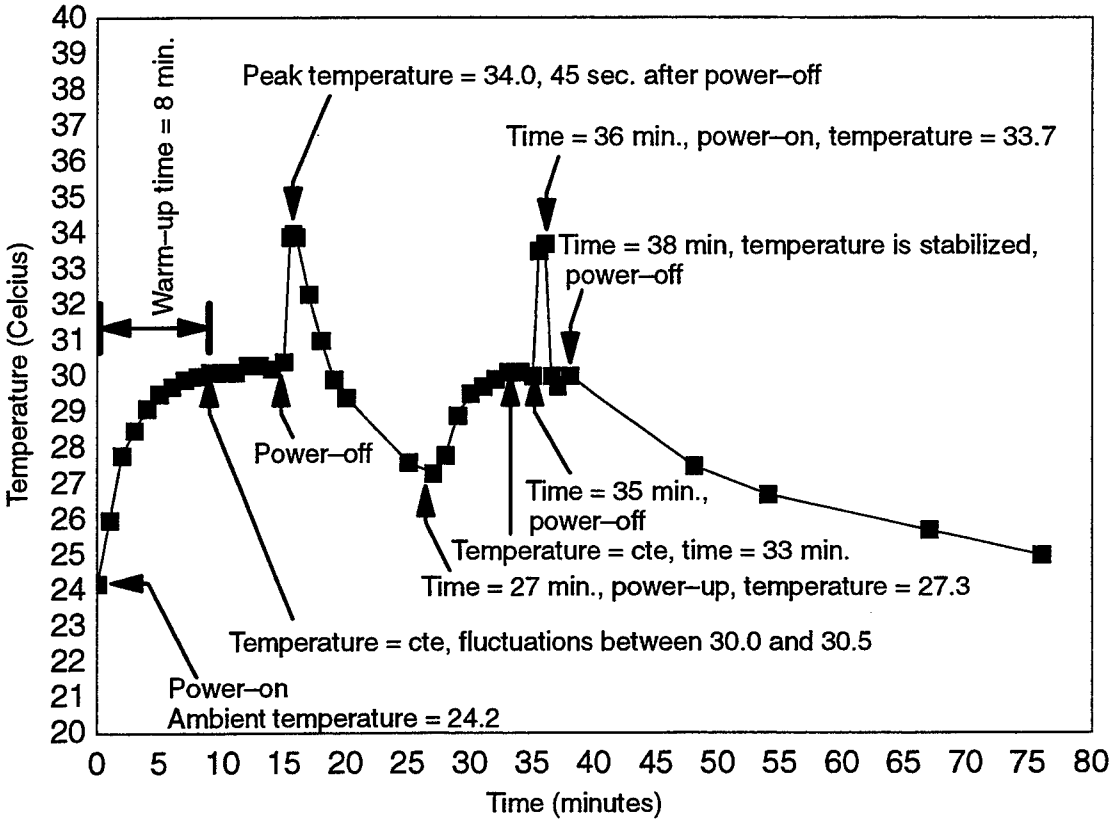


Figure 15 – Variations in temperature inside the 512–channel EMAP system.

Power Supply for UFDs

The first considerations in high–precision acquisition systems are separated digital and analog grounds as well as a good layout including ground planes. [Kester–1, 1992] examines problem of interfacing the ADC to the rest of the system and the critical issues of grounding, layout, and filtering. Furthermore, a typical power supply rejection of 80 dB for typical instrumentation amplifiers will produce a 10 μV input offset change for a common variation of 100 mV on the power supply. This means a loss of one LSB for a relatively small change in the power line alone. Consequently, the power unit should have a low output noise level, a good line and load regulation, and a good temperature stability. We found with EMAP that such requirements can be met easily with batteries[10] which have been shown to be a serious alternative for UFDs. Such alternative was also seriously considered in our next generation system UnEmap.

10. In cardiac mapping, EMAP relies on four batteries, each one assigned to a specific power requirement. Other applications use conventional low noise power–supplies. The batteries also eliminated the 60 Hz pick–up and provided maximum isolation since the system was self–powered. It is by far less expensive than power units offering the same characteristics. The disadvantages are the bulky arrangement and which requires monitoring and recharges. In EMAP the power–up and power–down sequences are done by a power sequencer board [Figure 4.4] which links the various power–sources or batteries to the acquisition modules through a power distribution backplane.

Time–Interleaved Oversampling

The idea of time–interleaved oversampling is not new [Khoini–Poorfard and Johns, 1993]. For example, 2 time interleaved operating 12–bit A/D converters with multiplexed outputs have been done [Jung *et al.*, 1993]. As another example, 4 ADCs with a sampling rate of 250 MHz were interleaved with buffer in one channel to achieve an effective sampling rate of 1 GHz [Von Walter and Rausch, 1992].

A time–interleaved ADC system can achieve superior performance, given the same implementation technology. For a given technology, there is theoretically no limit to the sample rate that can be reached using interleaved methods, although there is a limit to the bandwidth and thus the usefulness of interleaving. Real–world limitations such as power and space place a practical limits on the level of interleaving that can be achieved. Time–oversampling is theoretically possible with UNIMA but has not been studied so far.

One of the best example where high integration packaging has been used to improve the oversampling rate is the recent HP 54720 oscilloscope doing 16x500 MS/s ADCs (16 ADCs implemented in 4 hybrid modules) for a system sampling rate of 8 GS/s, sample at consistent intervals of 125 ps with a signal bandwidth of nearly 2 GHz [Montijo and Rush, 1993]. Further recent results about the performance of high resolution ADCs in time–interleave operation can be found in [Jung *et al.*, 1993]. The primary limiting factor for time–interleave in UNIMA is the packaging. Since the system is relatively large because of the quantity of modules interconnected, very high–frequency and low skew synchronization signals that would propagate through the system with almost no jitter are typically impossible to implement. Because the integration level is expected to improve, UFDs may become smaller and make time–interleaved oversampling easier. Such possibility may improve the universality of the A/D–based UFDs.

Summary

- *The level 2 and 3 UFDs exist only to compensate for the deficiencies of the communication links.*
- *An UFD is a field device without an adaptation layer. Such adaptation is performed in a central location within the VHI for time–critical functions and by a central CPU for non time–critical functions.*
- *An UFD is always connected to an universal communication link.*
- *Except for the flash converter, all conversion methods can be implemented within a level one UFD. For remote control and most instrument applications, the successive–approximation method is presently the best option to be implemented within an UFD.*
- *The only practical conversion method for D/A–based UFDs is the R/2R ladder.*
- *The A/D conversion node should consist of an antialiasing filter, a range amplifier, a S/H amplifier, and ADC, and a simple communication PL interface. The range amplifier would have high bandwidth and should be set for relatively small gain values because of the*

settling time issue. The S/H amplifier reduces the aperture time and allows high slew rate signals to be recorded.

- *For ($T_H \rightarrow T_C$) configurations, the acquisition time should be programmable. A simplest implementation relies on a coarse followed by a fine charge of the binary–weighted capacitor array.*
- *High–resolution A/D–based UFDs should rely on self–calibrating capacitive–based converters instead of resistor–based networks.*
- *Because it is analog, the signal conditioning node can hardly be universal since it cannot be implemented within a VHIs. The recommended approach, while not optimal, is to identify pre–defined types of signal conditioning nodes that could be interconnected in any order. Such types have been initially identified as filtering, pre–amplification, sample–and–hold, multiplexing, and protective nodes.*
- *Base–line drift and antialiasing cannot be done efficiently with digital techniques. Therefore, because the A/D conversion node has already an analog LPF, the filtering node should only be initially an AC–coupled analog filter. Other requirements could be performed digitally such as FIR or IIR techniques at the central location.*
- *Because AC–coupled inputs reduces the CMRR, the pre–amplifier should be FET–input instrumentation amplifiers. This should provide very high–impedance with very low bias current. Low bias current is often essential to avoid a deterioration of the CMRR with the implementation of an AC–coupled interface prior to the pre–amplification node. These characteristics are essential in an universal pre–amplification node since many applications require high input impedance differential recording of small signal with high CMRR because of a noisy environment.*
- *The multiplexing node should be a simple 8:1 or 16:1 FET–input analog multiplexer, The protective node must have very low leakage current. This also holds for the S/H amplification node.*
- *The level one D/A–based UFD should have two independent nodes, a D/A conversion node and a communication PL node. Multichannel D/A–based UFDs should be implemented with several D/A conversion nodes and a digital demultiplexing node inserted between the A/D and the PL nodes.*
- *Remote configuration, initialization, and calibration can be performed on simple UFDs.*
- *It was shown that the temperature drift is a big issue and that each UFDs should be housed in its own chassis.*