SunFesT

(www.seas.upenn.edu/~sunfest)

2005

SUMMER UNDERGRADUATE FELLOWSHIPS
IN
SENSOR TECHNOLOGIES

TECHNICAL REPORT
TR-CST01DEC05
Center for Sensor Technologies
University of Pennsylvania
Philadelphia, PA 19104

SUNFEST 2005

SUMMER UNDERGRADUATE FELLOWSHIP IN SENSOR TECHNOLOGIES
Sponsored by the National Science Foundation (EEC-0244055)

PREFACE

This report is the result of 11 undergraduate students' research efforts during the summer of 2005. From May 23 through July 29, 2005, students from Penn and other colleges participated in the SUNFEST program, which is organized by the Center for Sensor Technologies of the School of Engineering and Applied Science at the University of Pennsylvania. This unique "Summer Experience for Undergraduates in Sensor Technologies" program was initiated in 1986 and has grown considerably in size. It is now recognized as one of the most successful summer programs for undergraduates in the country. I would like to express my sincere gratitude to the National Science Foundation for their continued support for this REU Site, as well as Microsoft Corporation for sponsoring two of our students.

The purpose of the SUNFEST program is to provide bright, motivated undergraduate students with the opportunity to become involved in active research projects under the supervision of a faculty member and his graduate student(s). The general area of research concentrates on sensor technologies and includes projects such as materials and technology for sensors, microstructures, smart imagers, and neural networks for sensory processing and robotics. By providing the students with hands-on experience and integrating them with a larger research group where they can work together with other students, the program intends to guide them in their career choices. By exposing the students to the world of research, we hope they will be more inclined to go on for advanced degrees in science and engineering.

The students participated in a variety of hands-on workshops in order to give them the tools to do first-rate research or enhance their communication skills. These included "Giving Effective Presentations", "Ethics in Science and Engineering", "Use of Electronic Databases" and "Writing Technical Reports". Students also had plenty of opportunity for social interactions among themselves or with faculty and graduate student advisors.

This booklet contains reports from this year's projects, the quality of which testifies to the high level of research and commitment by these students and their supervisors. I would like to express my sincere thanks to the students for their enthusiastic participation; the help of the faculty members, graduate students and support staff is very much appreciated. I would also like to thank Ms. Delores Magobet, Shelley Brown, Denice Gorte, Sid Deliwala, Scott Slavin, and the rest of the ESE staff for their invaluable help in making this program run smoothly.

Jan Van der Spiegel, Director
Center for Sensor Technologies

FINAL REPORT
2005 SUMMER UNDERGRADUATE FELLOWSHIP
IN SENSOR TECHNOLOGIES
Sponsored by the National Science Foundation

http://www.seas.upenn.edu/~sunfest/pastProjects/Projects05.html

TABLE OF CONTENTS

HANDHELD DEVICE FOR REMOTELY MEASURING BRAIN FUNCTION
Adam Wang (Electrical Engineering Department) University of Texas, Austin
Advisor: Dr. Britton Chance **173-190**

LIST OF STUDENT PARTICIPANTS
IN THE SUNFEST PROGRAM SINCE 1986

Summer, 2005

Robert Callan	University of Pennsylvania
David Cohen	University of Pennsylvania
Louie Huang	University of Pennsylvania
Roman Geykhman	University of Pennsylvania
An Nguyen	University of Pennsylvania
Olga Paley	University of Caliofornia at Berkeley
Miguel Perez Tolentino	University of Puerto Rico at Humacao
Ebenge Usip	University of Southern California
Adam Wang	University of Texas, Austin
Kejia Wu	University of Pennsylvania

Summer, 2004

Benjamin Bau	Massachusetts Institute of Technology
Alexander H. Chang	University of Pennsylvania
Seth Charlip-Blumlein	University of Pennsylvania
Ling Dong	University of Rochester
David Jamison	Johns Hopkins University
Dominique Low	University of Pennsylvania
Emmanuel U. Onyegam	University of Texas, Dallas
J. Miguel Ortigosa	Florida Atlantic Unversity
William Rivera	University of Puerto Rico, Mayaguez
Matthew Sauceda	Texas A&M University, Kingsville
Olivia Tsai	Carnegie Mellon University

Summer, 2003

Emily Blem	Swarthmore College
Brian Corwin	University of Pennsylvania
Vinayak Deshpande	University of Virginia
Nicole DiLello	Princeton University
Jennifer Geinzer	University of Pittsburgh
Jonathan Goulet	University of Pennsylvania
Mpitulo Kala-Lufulwabo	University of Pittsburgh
Emery Ku	Swarthmore College
Greg Kuperman	University of Pennsylvania
Linda Lamptey	University of Pennsylvania
Prasheel Lillaney	University of Pennsylvania
Enrique Rojas	University of Pennsylvania

Summer, 2002

Christopher Bremer	Colorado School of Mines
Aslan Ettehadien	Morgan State University
April Harper	Hampton University
Catherine Lachance	University of Pennsylvania
Adrian Lau	University of Pennsylvania
Cynthia Moreno	University of Miami
Yao Hua Ooi	University of Pennsylvania
Amber Sallerson	University of Maryland/Baltimore County
Jiong Shen	University of California-Berkeley
Kamela Watson	Cornell University
John Zelena	Wilkes University

Summer, 2001

Gregory Barlow	North Carolina State University
Yale Chang	University of Pennsylvania
Luo Chen	University of Rochester
Karla Conn	University of Kentucky
Charisma Edwards	Clark Atlanta University
EunSik Kim	University of Pennsylvania
Mary Kutteruf	Bryn Mawr College
Vito Sabella	University of Pennsylvania
William Sacks	Williams College
Santiago Serrano	Drexel University
Kiran Thadani	University of Pennsylvania
Dorci Lee Torres	University of Puerto Rico (Humacao)

Summer, 2000

Lauren Berryman	University of Pennsylvania
Salme DeAnna Burns	University of Pennsylvania
Frederick Diaz	University of Pennsylvania (AMPS)
Hector Dimas	University of Pennsylvania (AMPS)
Xiomara Feliciano	University of Turabo (Puerto Rico)
Jason Gillman	University of Pennsylvania
Tamara Knutsen	Harvard University
Heather Marandola	Swarthmore College
Charlotte Martinez	University of Pennsylvania
Julie Neiling	University of Evansville, Indiana
Shiva Portonova	University of Pennsylvania

Summer, 1999

David Auerbach	Swarthmore
Darnel Degand	University of Pennsylvania
Hector E. Dimas	University of Pennsylvania
Ian Gelfand	University of Pennsylvania
Jason Gillman	University of Pennsylvania
Jolymar Gonzalez	University of Puerto Rico – Mayaguez
Kapil Kedia	University of Pennsylvania
Patrick Lu	Princeton University
Catherine Reynoso	Hampton University
Philip Schwartz	University of Pennsylvania

Summer, 1998

Tarem Ozair Ahmed	Middlebury University
Jeffrey Berman	University of Pennsylvania
Alexis Diaz	Turabo University, Puerto Rico
Clara E. Dimas	University of Pennsylvania
David Friedman	University of Pennsylvania
Christin Lundgren	Bucknell University
Heather Anne Lynch	Villanova University
Sancho Pinto	University of Pennsylvania
Andrew Utada	Emory University
Edain (Eddie) Velazquez	University of Pennsylvania

Summer, 1997

Francis Chew	University of Pennsylvania
Gavin Haentjens	University of Pennsylvania
Ali Hussain	University of Pennsylvania
Timothy Moulton	University of Pennsylvania
Joseph Murray	Oklahoma University
O'Neil Palmer	University of Pennsylvania
Kelum Pinnaduwage	University of Pennsylvania
John Rieffel	Swarthmore College
Juan Carlos Saez	University of Puerto Rico, Cayey

Summer, 1996

Rachel Branson	Lincoln University
Corinne Bright	Swarthmore College
Alison Davis	Harvard University

Rachel Green	Lincoln University
George Koch	University of Pennsylvania
Sandro Molina	University of Puerto Rico-Cayey
Brian Tyrrell	University of Pennsylvania
Joshua Vatsky	University of Pennsylvania
Eric Ward	Lincoln University

Summer, 1995

Maya Lynne Avent	Lincoln University
Tyson S. Clark	Utah State University
Ryan Peter Di Sabella	University of Pittsburgh
Osvaldo L. Figueroa	University of Puerto Rico-Humacao
Colleen P. Halfpenny	Georgetown University
Brandeis Marquette	Johns Hopkins
Andreas Olofsson	University of Pennsylvania
Benjamin A. Santos	University of Puerto Rico-Mayaguez
Kwame Ulmer	Lincoln University

Summer, 1994

Alyssa Apsel	Swarthmore College
Everton Gibson	Temple University
Jennifer Healy-McKinney	Widener University
Peter Jacobs	Swarthmore College
Sang Yoon Lee	University of Pennsylvania
Paul Longo	University of Pennsylvania
Laura Sivitz	Bryn Mawr College
Zachary Walton	Harvard University

Summer, 1993

Adam Cole	Swarthmore College
James Collins	University of Pennsylvania
Brandon Collings	Hamilton University
Alex Garcia	University of Puerto Rico
Todd Kerner	Haverford College
Naomi Takahashi	University of Pennsylvania
Christopher Rothey	University of Pennsylvania
Michael Thompson	University of Pennsylvania
Kara Ko	University of Pennsylvania
David Williams	Cornell University
Vassil Shtonov	University of Pennsylvania

Summer, 1992

James Collins	University of Pennsylvania
Tabbetha Dobbins	Lincoln University
Robert G. Hathaway	University of Pennsylvania
Jason Kinner	University of Pennsylvania
Brenelly Lozada	University of Puerto Rico
P. Mark Montana	University of Pennsylvania
Dominic Napolitano	University of Pennsylvania
Marie Rocelie Santiago	Cayey University College

Summer, 1991

Gwendolyn Baretto	Swarthmore College
Jaimie Castro	University of Puerto Rico
James Collins	University of Pennsylvania
Philip Chen	University of Pennsylvania
Sanath Fernando	University of Pennsylvania
Zaven Kalayjian	University of Pennsylvania
Patrick Montana	University of Pennsylvania
Mahesh Prakriya	Temple University
Sean Slepner	University of Pennsylvania
Min Xiao	University of Pennsylvania

Summer, 1990

Angel Diaz	University of Puerto Rico
David Feenan	University of Pennsylvana
Jacques Ip Yam	University of Pennsylvania
Zaven Kalayjian	University of Pennsylvania
Jill Kawalec	University of Pennsylvania
Karl Kennedy	Geneva
Jinsoo Kim	University of Pennsylvania
Colleen McCloskey	Temple University
Faisal Mian	University of Pennsylvania
Elizabeth Penadés	University of Pennsylvania

Summer, 1989

Peter Kinget	Katholiek University of Leuven
Chris Gerdes	University of Pennsylvania
Zuhair Khan	University of Pennsylvania
Reuven Meth	Temple
Steven Powell	University of Pennsylvania
Aldo Salzberg	University of Puerto Rico
Ari M. Solow	University of Maryland
Arel Weisberg	University of Pennsylvania
Jane Xin	University of Pennsylvania

Summer, 1988

Lixin Cao	University of Pennsylvania
Adnan Choudhury	University of Pennsylvania
D. Alicea-Rosario	University of Puerto Rico
Chris Donham	University of Pennsylvania
Angela Lee	University of Pennsylvania
Donald Smith	Geneva
Tracey Wolfsdorf	Northwestern University
Chai Wah Wu	Lehigh University
Lisa Jones	University of Pennsylvania

Summer, 1987

Salman Ahsan	University of Pennsylvania
Joseph Dao	University of Pennsylvania
Frank DiMeo	University of Pennsylvania
Brian Fletcher	University of Pennsylvania
Marc Loinaz	University of Pennsylvania
Rudy Rivera	University of Puerto Rico
Wolfram Urbanek	University of Pennsylvania
Philip Avelino	University of Pennsylvania
Lisa Jones	University of Pennsylvania

Summer, 1986

Lisa Yost	University of Pennsylvania
Greg Kreider	University of Pennsylvania
Mark Helsel	University of Pennsylvania

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

LIQUID FLOW MEASUREMENTS USING A PYROELECTRIC ANEMOMETER

NSF Summer Undergraduate Fellowship in Sensor Technologies
Robert Callan, Electrical Engineering, University of Pennsylvania
Advisor: Dr. J.N. Zemel

ABSTRACT

The transport of fluids through pipes is an essential part of many industrial processes, requiring an accurate measurement of the flow. The measurement of natural gas flow in pipelines and the precise proportions of reactants flowing into a reaction chamber are but two examples. Many flow meters have been developed based on a wide variety of operating principles. Properties, such as physical size, speed, measurement accuracy, cost, reliability, and difficult recalibration often limit the usefulness of these meters in different applications.

The device investigated in this study is the Pyroelectric Anemometer (PA). Its operation is based on the convective heat loss from the device to the fluid in gas flows and standard heat transfer theory accounted nicely for the experimental data in previous research. This project addresses the pyroelectric anemometer's response to liquid flows. A test system was built and data were collected. These measurements were compared to gas flows and a previously derived fluid flow model.

Table of Contents

1. INTRODUCTION

The movement of fluids and their accurate measurement is important in many industrial processes. As a result, numerous sensors have been developed which rely on a number of different operating principles, including pressure differences, mechanical methods, and heat transfer characteristics to determine flow rates. The purpose of this report is to present the results of a study of the operation of a thermal flow sensor based on the pyroelectric effect. Previous research had established that the Pyroelectric Anemometer (PA) is remarkably precise, and may be used in flows whose Reynolds numbers (Re) range over 5 orders of magnitude. These studies showed that the PA met and/or exceeded the performance of commercially available sensors for a wide range of gas flows. [1]

This paper presents information on the PA's performance and behavior in liquid flows. Two absolute measurement systems were developed and an electronic signal processing system was developed to take measurements of the flow rate. Several differences between the responses to the liquid and gas flows are demonstrated, and their causes explored.

2. THEORY OF OPERATION

Pyroelectricity arises in certain types of crystalline materials as a result of ion motions as temperature changes. The result is a change of the surface charge that is directly proportional to the change in the temperature. In convective heat flow, the net heat flow depends on the difference in the temperature between the fluid and solid. The change in temperature of the solid is then a measure of the convective heat flow, and .a measure of the fluid motion. This is common experience where a breeze is a welcome coolant on a hot day.

The PA is used to determine flow rate by relating changes in convective heat loss from the sensor to the moving liquid. The heat loss is transduced to electrical signals by the pyroelectric effect with a 3x4mm $LiTaO_3$ crystal shown in Figure 1.

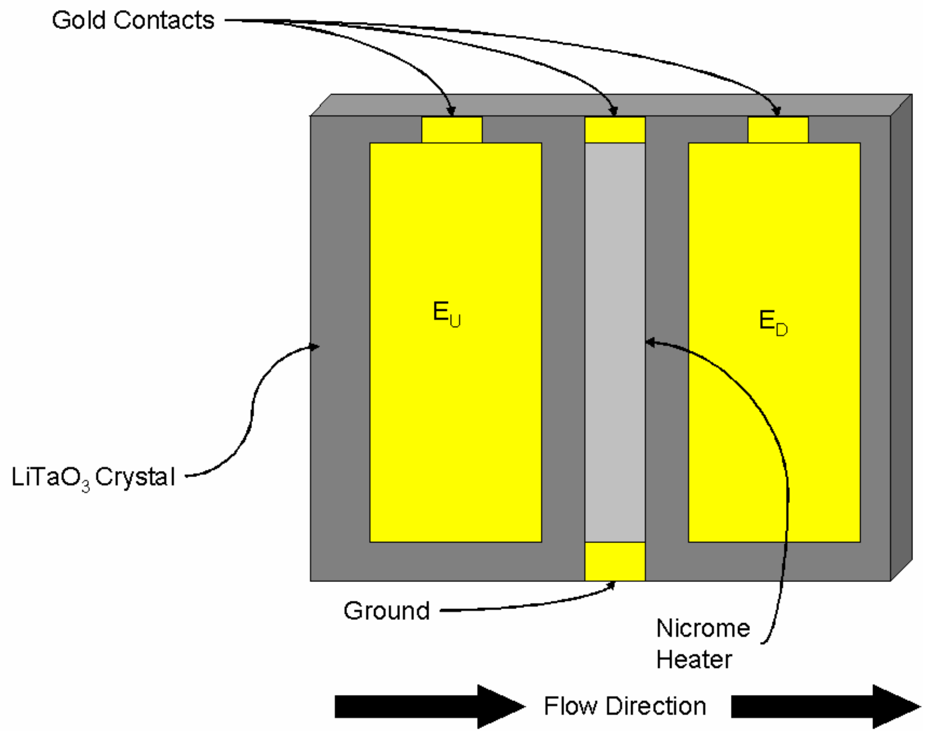


Figure 1: Layout of PA Crystal

The physical properties of the $LiTaO_3$ crystal can be used as in [1] to develop an equivalent circuit of the PA, which consists of a current source, capacitance, and resistance in parallel.

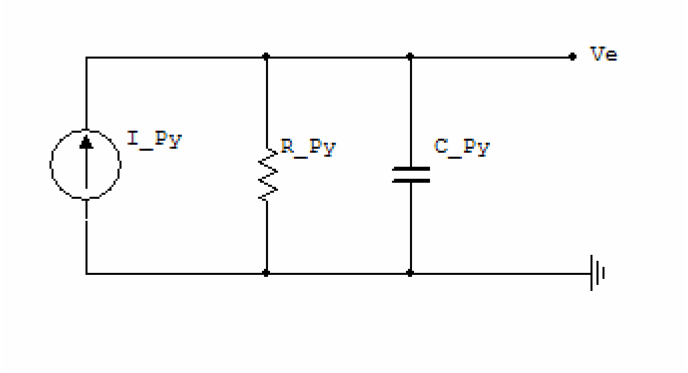


Figure 2: An equivalent circuit for a PA electrode

Because the $LiTaO_3$ crystal is a good dielectric, there is a capacitance C_py between the electrodes and ground. A resistance R_py models the leakage current that flows from the electrodes through the crystal to ground. The current source I_py is used to model the pyroelectric effect, where $I_{py} = \tilde{p} \cdot A_e \cdot \frac{d\Phi}{dt}$ where $\tilde{p}$ is the pyroelectric coefficient, A_e is the area of the electrode, and $\frac{d\Phi}{dt}$ is the change in temperature of the crystal (under the electrode) with respect to time.

The heater of the PA is driven by a sinusoidal voltage, adding an amount of heat proportional to the power dissipated in the heater resistor equal to

$$I^2 R= I_M^2 \cos(wt)^2 R = I_M^2 (1-\cos(2wt)) R/2.$$

This causes Joule heating which then changes the temperature (and heat content) of the device and then produces a varying surface charge on the crystal. This varying charge induces an external detectable current. Any net charge accumulation that arises as a result of the DC change in heat is neutralized over time, and all that remains is the AC component at twice the driving frequency. It is this signal which is processed and is related to the flow rate.

The two electrodes of the PA measure the changing charge on either side of the PA in a one dimensional flow. Heat flows from the heater, both through the crystal and through the fluid around the crystal. A change in the velocity of the fluid flowing over the crystal causes a change in the heat exchange between the crystal and the fluid. This heat exchange provides the connection between the velocity of the fluid flowing over the crystal and the change in the surface charge of the crystal. Previous gas flow research has demonstrated that the difference in the amplitudes of the response signals at the two electrodes is directly proportional to Reynolds numbers ≤ 10 and is proportional to the square root of the flow for a region of flows above Reynolds numbers of ~10 [1].

2.1 Electrical Response of the PA

The characteristics of interest in the system are the frequency, amplitude, and phase of the thermal input to the system called Φ_h (which is given by I^2R as above), as well as the frequency, amplitude, and phase of the thermal content of the crystal below each electrode, given as Φ_d, Φ_u (subscripts meaning upstream and downstream). All three of these signals are of the same frequency. Φ_h is determined by the voltage across the heating element, and its resistance and is the heat dissipated in the resistive element. The charge that develops on each of the electrodes as a result of the pyroelectric effect is proportional to Φ_d, Φ_u. Therefore $\frac{dQ_d}{dt}, \frac{dQ_u}{dt}$ are proportional to $\frac{d\Phi_d}{dt}, \frac{d\Phi_u}{dt}$. $\frac{dQ_d}{dt}, \frac{dQ_u}{dt}$ are currents, and are proportional to voltages named V_d, V_u which are the quantities actually measured. The frequency of the voltage across the heater was set to 3Hz, resulting in a thermal response at 6Hz. Therefore, the quantities to be measured are the amplitude and phase of the 6Hz component of each of Φ_h, V_d, V_u (called $A_h, A_d, A_u, \theta_h, \theta_d, \theta_u$) as well as the flow rate (given as the Reynolds Number of the flow). These measurements were taken over a wide range of flow rates.

3. MEASUREMENT SYSTEM

3.1 Overview

A diagram of the complete measurement system is given in figure 3 below.

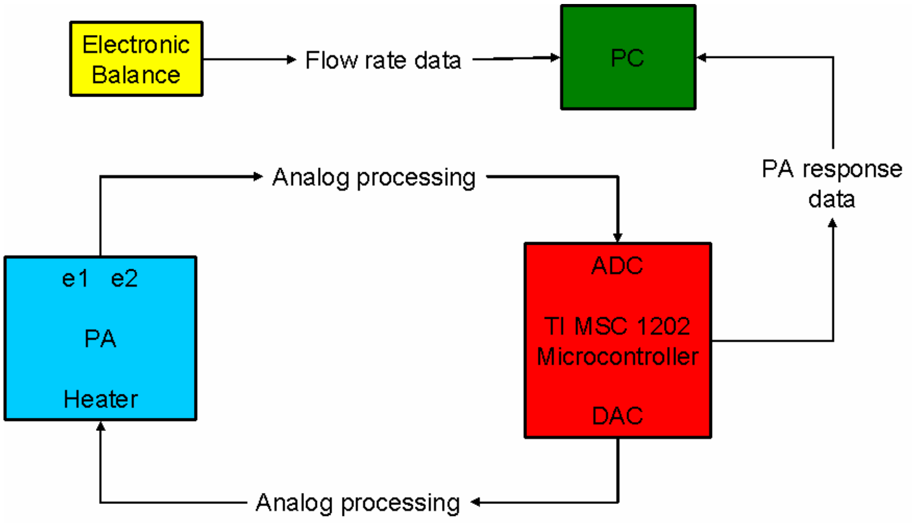


Figure 3: Diagram of the complete measurement system

The measurement system consists of 5 separate components: the PA, an analog processing section, an electronic balance, a PC, and a TI MSC1202 Microcontroller. The microcontroller outputs the driver signal on its DAC, which is filtered and amplified to excite the PA. The signals from each electrode of the PA, E_u, E_d are filtered and then each is sampled by the ADC of the microcontroller. The microcontroller converts these voltages to a digital representation, and sends them via an RS232 line to a PC for processing. Simultaneously, an electronic balance weighs a collection container to determine how rapidly the liquid is being collected. It sends this weight data over a second RS232 line to the PC for processing. The PC extracts the necessary data from these two sources, and determines the values of $A_u, A_d, \theta_u, \theta_d, flow$ and outputs them to the PC's screen. The process is repeated continuously as the flow rate is varied.

3.2 Electronics Onboard the PA

The PA sensor has the electronics shown in figure 4, directly connected to the electrodes on the crystal of the PA (at the point Ve in the figure). The raw voltage present at the point Ve is in the range of 2-20mV peak-to-peak. Each electrode has the buffer and amplifier shown below. The second op amp gives an amplification of 410 to bring the signal to usable levels.

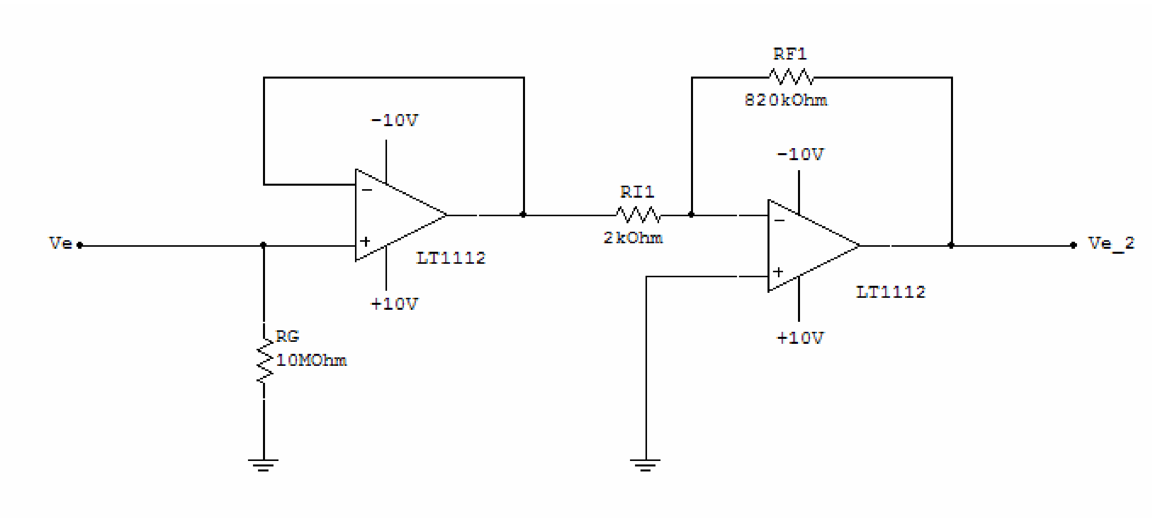


Figure 4: PA Onboard Electronics

3.3 Analog Filtering of the Reponse Signals

The PA is connected to a protoboard containing the electronics shown in figure 5. These electronics allow adjustment of the signal level and the addition of an offset of 1.25V. The ADC of the microcontroller has an input range of 0-2.5V, so it is necessary to adjust the signal level in the first stage, and to add the 1.25 offset.

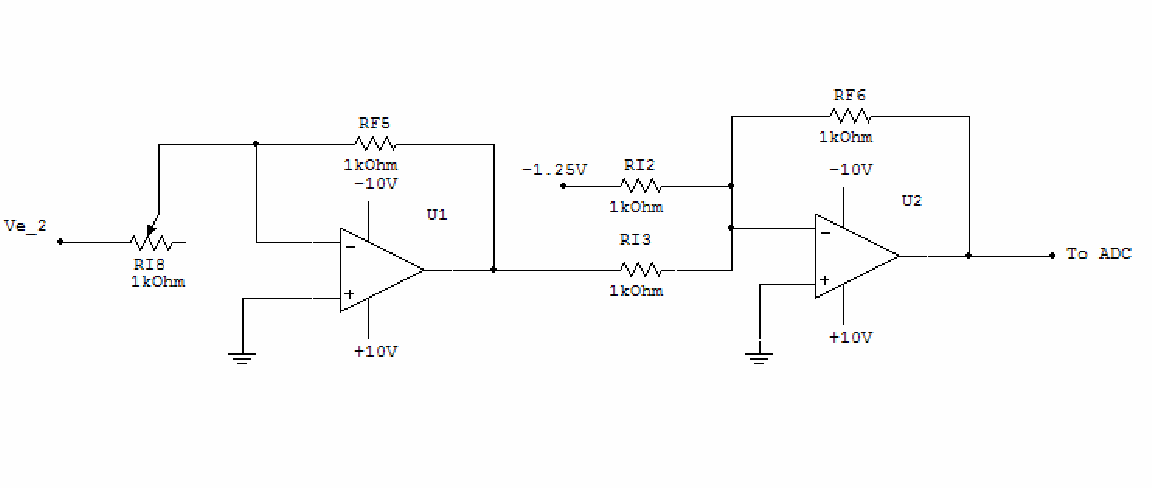


Figure 5: Analog Filtering of Response

3.4 Analog Filtering of the Heater Signal

The output of the DAC is a high impedance signal, and must be conditioned and amplified in order to drive the PA heater at the required energy level (20mW[1]). The resistance of the heating element is 800 Ohms, so an RMS voltage of 4V is needed. In addition, an offset of -0.5 V must be added to the signal because the output of the DAC is positive only. If there is an offset present in the input signal to the heater, this can have a negative effect on the response signals. A low pass filter is also necessary to smooth the heater signal.

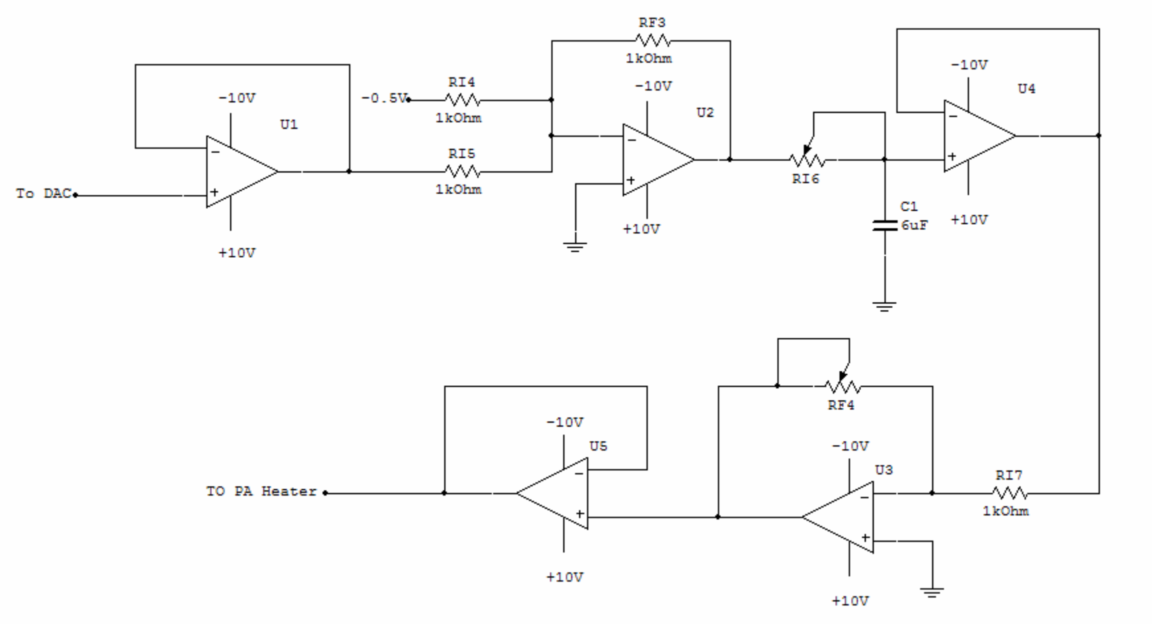


Figure 6: Analog filtering of the heater signal

3.5 TI MSC1202 Microcontroller

As shown in the System Diagram in Figure 3, the microcontroller generates the driver signal, samples the filtered responses of both electrodes, and sends this information to the PC for processing. The microcontroller chosen to do this is the TI MSC, an 8051 class Microcontroller which has a 16 bit Delta Sigma ADC, and an 8 bit IDAC. The ADC offers 16 Bits of resolution with an input voltage range of 0-2.5V, and the IDAC has an output current range of 0-1mA. Driving this current through a 1k resistor, results in an output voltage in the 0-1V range.

The microcontroller's ADC sampling rate controls the timing of the system. The signal of interest is at 6 Hz, and was sampled at 16 points per wave. The ADC is switching between two channels (one per electrode) so it must throw away every other measurement (due to the continuous sampling of the Delta-Sigma architecture). This means that the sampling rate of the ADC must be (2 Channels)*(16*6 Samples Per Second/Channel)*2(for thrown away measurements)= 384 Samples Per Second.

The microcontroller completes the following cycle every ~2.6 mSec (1/384):

1) Output the appropriate voltage on the IDAC to create a 3Hz sine wave
2) Record the voltage of the ADC conversion for this cycle
3) If cycle number is odd:
   a. Switch the Mux between the two input channels
   b. Output the raw ADC data through the RS232 link to the PC

If it were necessary to embed the measurement system such that the microcontroller was the only processor in the system, the microcontroller would also have to perform the amplitude calculations. This however was not desirable, because it was much easier to develop, monitor, and update code running on the PC.

4. FLOW SYSTEM

4.1 Use of the Electric Balance to Measure Flow

An AND electronic balance was used in these experiments to determine the flow rate. This was accomplished by relating a change in mass of the liquid collected to a change on volume of liquid (by knowing the density of the liquid) and then converting this to a velocity by knowing the area of the tubing at the PA element. The density was determined by using a 500mL standard flask to determine the weight in grams of 500mL of liquid. The density of the liquid tested, Wolf's Head Automatic Transmission Fluid, is .857 g/ml (g/cm^3), and the radius of the tubing is .292 cm^2. Therefore the conversion factor from grams/sec to cm/sec is grams/sec*1/.857 cm^3/gram*1/(.292 cm^2)= 3.996cm/sec. The accuracy of the scale is ± .01 gram, so the accuracy of velocity measurement is ~± .04 cm/sec.

4.2 Flow System and Absolute Measurement of the Flow

Two flow systems were used to generate a steady flow. They had in common the PA sensor element, and a collection container resting on a balance. The flow rate was determined by measuring the difference in masses over a period of 2 2/3 seconds. This was related to the velocity of the flow as described in "Use of the Electronic Balance to Measure Flow."

The measurement system designed for low flows is shown in Figure 7. It uses the pressure developed by a column of liquid to create a flow past the sensor. The large tube at the left of the figure is filled with liquid, and measurements of the PA response and flow rate are simultaneously collected as the liquid drains from the tube. This system requires no intervention from the experimenter.

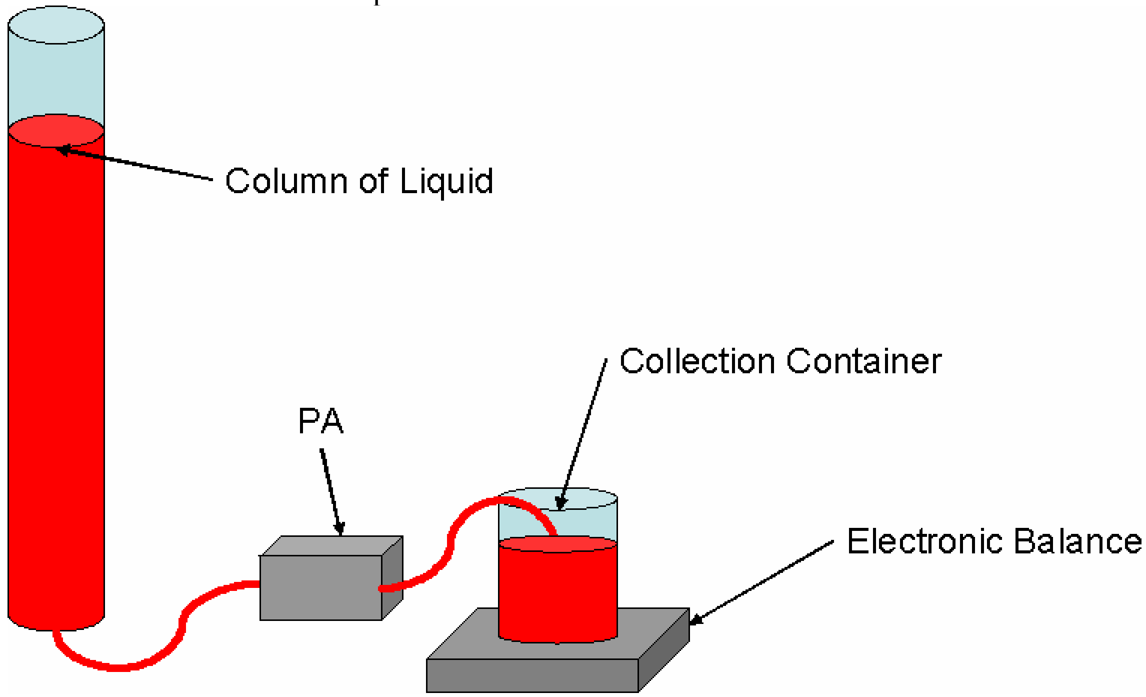


Figure 7: Flow system for lower flow rates

The second measurement system, which is used for higher flow rates, is shown in Figure 8. It uses a pump to develop a flow past the sensor. This allows for higher flow rates, but requires more time to take measurements because the speed of the pump must be adjusted gradually by hand. The higher flow rates also require more frequent emptying of the collection container, which complicates the process of taking accurate measurements.

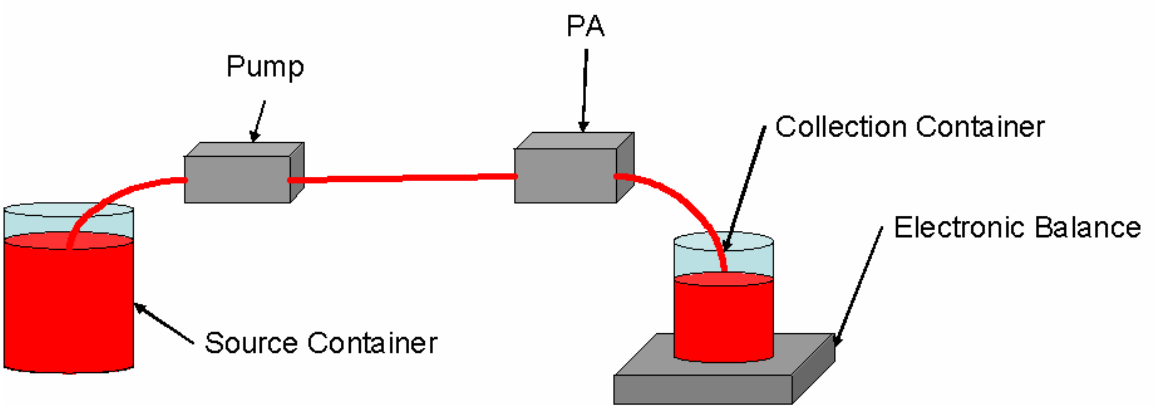


Figure 8: Flow system for higher rates

5. ROLE OF THE PC

The PC combines and analyzes all the sensor data and displays the end result: the PA amplitude and phase data $A_d, A_u, \theta_d, \theta_u$, as well as the flow rate past the sensor in cm/sec. Each of these measurements is averaged over approximately 2.6 seconds (256 sampling points per each channel) and then output and recorded continuously.

All the software used to achieve this task is written in Java. The RS232 ports are controlled using an API from Sun called javacomm. Java was chosen for two reasons: 1) it could perform all the necessary calculations within a single synchronized process, and 2) It provides excellent error reporting and handling.

Figure 9 shows a functional flow diagram of the program used to process the flow data. The yellow boxes are devices linked to the PC through RS232 ports and the blue boxes are java classes.

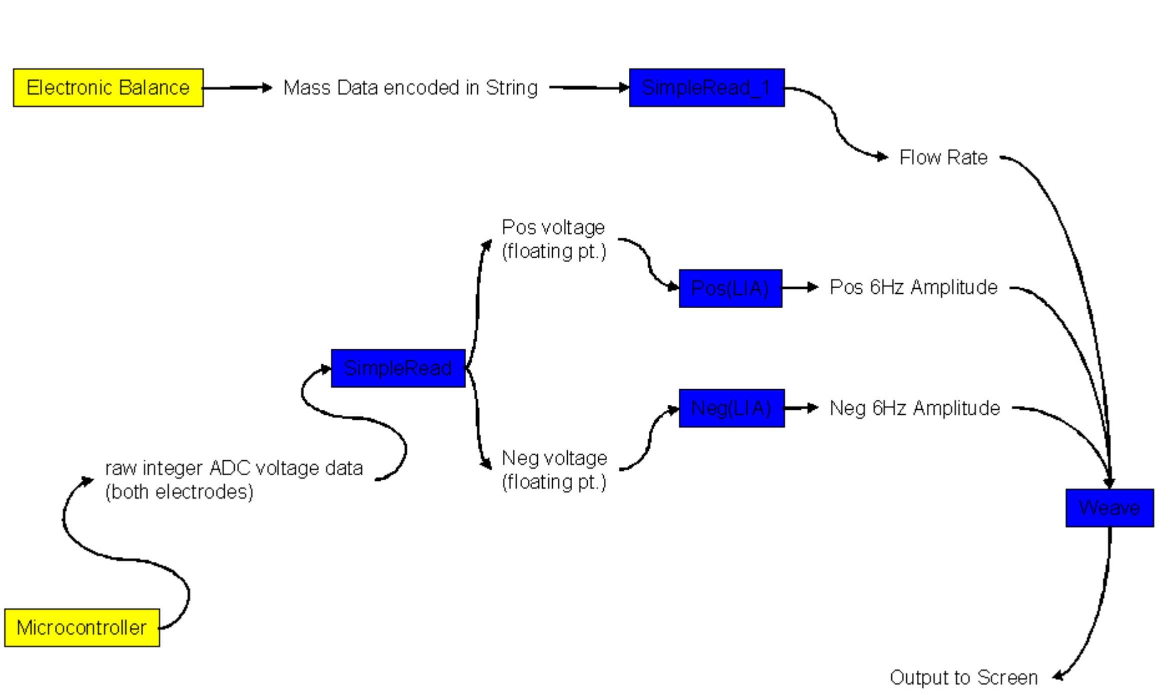


Figure 9: Diagram of the software used to determine PA response amplitudes and flow rate

The data sent from the microcontroller and the balance is converted into a usable form first. Two instances of the LIA class each collect 256 voltage measurements from the ADC, and then perform the LIA procedure described below.

5.1 The Lock-In-Amplification (LIA) Procedure

The purpose of this algorithm is to calculate the amplitude and phase of a sinusoid of a particular frequency. This is accomplished by multiplying the signal by $e^{-12j\pi t}$ and then averaging over a period of time. The complex amplitude of the 6Hz component(X) of the input signal s(t) is given by:

$$X = 2\,\frac{\int_0^T s(t)\,\mathbf{e}^{(-12j\pi t)}\,dt}{T}$$

The result is multiplied times 2 because the integral gives only the amplitude of the positive frequency component $\mathbf{e}^{(j\,12\,\pi\,t)}$ of $\cos(12\,\pi\,t) = \frac{\mathbf{e}^{(j\,12\,\pi\,t)} + \mathbf{e}^{(-j\,12\,\pi\,t)}}{2}$ and divided by T so that all the other frequency components sum to zero. The magnitude of the component X is given by $|X| = \sqrt{\Im(X)^2 + \Re(X)^2}$ and the phase is given by $\text{angle}(X) = \arctan\left(\frac{\Im(X)}{\Re(X)}\right)$. This is implemented by taking

$XReal = \int_0^T s(t)\cos(12\pi t)\,dt$ and $XComplex = \int_0^T s(t)\sin(12\pi t)\,dt$. This gives the real and complex parts of X. These equations hold because $\mathbf{e}^{(j\theta)} = \cos(\theta) + j\sin(\theta)$.

The code below (Taken from the class LIA given in appendix A) shows how this process is actually performed. The array pts[] contains the voltage data from the ADC over a period of 2 2/3 seconds. The double[] sines and cosines contain 6Hz sine and cosine waves respectively.

```
public double[] pts= new double[256];
public double[] sines= new double[pts.length],
                cosines= new double[pts.length];
 public double calculateAmplitude(){
     double sineSummation=0, cosineSummation=0;
     for(int i= 0; i<pts.length; i++){
         sineSummation+= pts[i]*sines[i];
         cosineSummation+= pts[i]*cosines[i];
     }
     return Math.sqrt(sineSummation*sineSummation
+cosineSummation*cosineSummation)*2./pts.length;
}
public double calculatePhase(){
    double sineSummation=0, cosineSummation=0;
    for(int i= 0; i<pts.length; i++){
       sineSummation+= pts[i]*sines[i];
       cosineSummation+= pts[i]*cosines[i];
    }
    return Math.atan(cosineSummation/sineSummation)*180/Math.PI;
}
```

6. ANALYSIS OF DESIGN

6.1 Analog vs. Digital Implementation

Research previously done with the PA in gases used an analog *differential* of the two input signals $(V_d - V_u)$ and then sampled the signals, and found the amplitude of this differential signal. This method assumes that the phase difference between the two electrodes is zero, and more importantly, that the two signals remain completely in phase over all measurements. These two assumptions were true in gases. However, it was observed in the oil tested that this was not the case. The phase between the two electrode signals at zero flow differed by ~10 degrees and the difference changed by more than 20 degrees at high flow rates. Therefore it would not be feasible to use an analog differential of the two signals, if a pure amplitude must be known. However, it would be possible to take the differential for practical purposes, ignoring the fact that the resulting signal would be a combination of phase differential and amplitude change. Figure X shows the resulting error surface when the differential of two signals is taken.

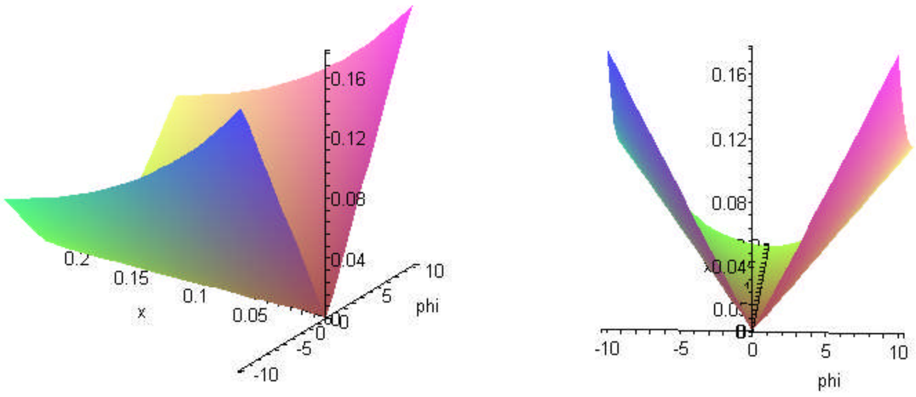


Figure 10: Error Surfaces resulting from Differences in Phase between Electrodes

These two figures show the result of $\frac{amplitude(a(t)-b(t))}{a(t)}$ where $a(t)=(1+x)\sin(wt), b(t)=\sin(wt+phi)$.

In order to eliminate this source of error entirely, a digital difference of the amplitudes is taken. The two signals $V_d(t), V_u(t)$ are each sampled separately and each of their amplitudes and phases determined. Then these two amplitudes are normalized and their difference $\frac{A_u}{A_{u0}}-\frac{A_d}{A_{d0}}$ is related to the flow rate.

6.2 Synchronization

The synchronization between the heater signal and the frequency at which the calculations take place must be as close as possible. Slight differences in frequency between the signal of interest contained in pts[], and signals sines[] and cosines[] can have a dramatic effect on the amplitude and phase data calculated using this algorithm. The results of such a difference are shown in Figure 11.

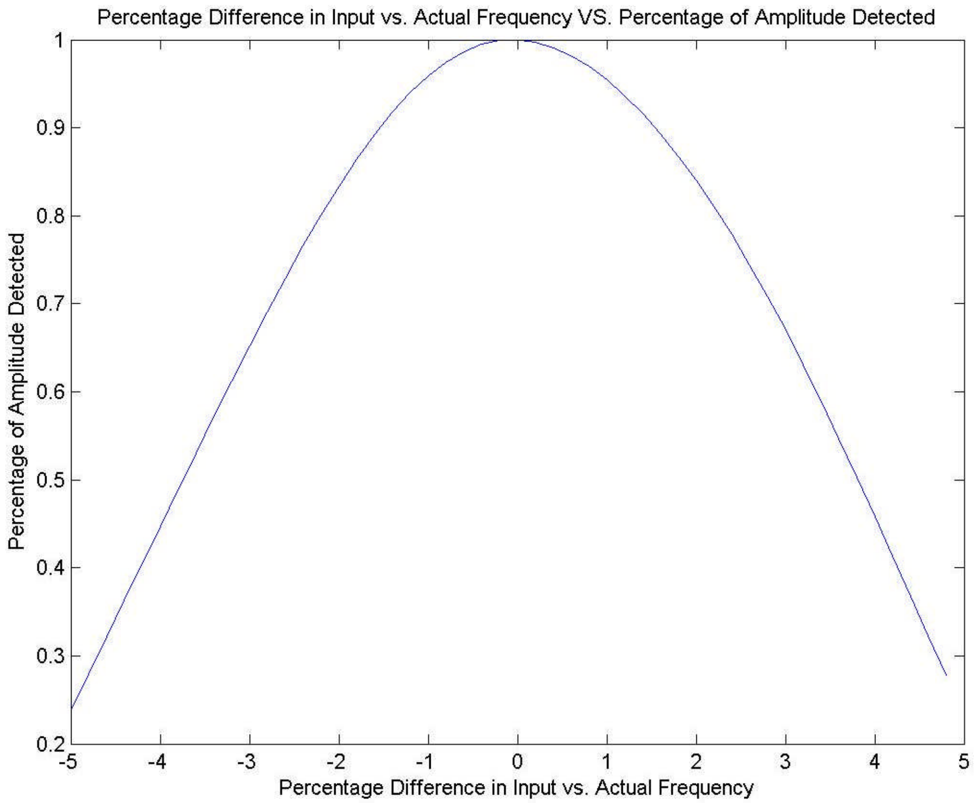


Figure 11: Change in measured amplitude as a result of differences in frequency

The results suggest that even relatively small differences in frequency can result in large errors, especially if the differences in frequency change during the course of the experiment (as was observed when using a low quality function generator). This analysis shows that the heater signal must be synchronized directly with the sampling of the ADC, which was accomplished in this system by using a microcontroller with DAC and ADC driven by the same clock.

7. EXPERIMENTAL RESULTS

7.1 Summary of Previous Research

There are three conclusions from previous research that we were able to compare to our results. The first is the theoretical models developed to model the response of the PA to gas flows. It was found that in flows of lower Reynolds number, theory predicts that the response of the PA will be linearly related to the Reynolds number. This model agreed quite well with experimental data collected in Nitrogen, Helium, and Argon gases. As the flow rate increases sufficiently, the response transitions to the square root of the Reynolds number according to theory. This result also agreed with the experimental data.

The second result is related to the phase of the responses with respect to the heater signal. It was found in the gases that as the Reynolds number increased, the phase of the differential response with respect to the heater signal remained constant up to a point, then changed quickly for a period, and then saturated again at another value. In the case of the liquid flow measurements made in this study, the phase difference between the two electrodes varied as shown in Figure 17.

The third conclusion relevant to the data collected in this study was that the single electrode response and normalized differential response of the PA were relatively insensitive to changes in temperature and pressure. While thorough testing over a wide range of temperatures was not done, it was noted that small changes in temperature did not affect the response of the device. All these conclusions were reached in [1].

7.2 Summary of Collected Data

Four sets of data were collected. The first set of data to be collected used the flow system employing a column of liquid to generate the flow. No effort was made to monitor the temperature of the oil, the phase of the signals, or the timing between trials. The second set of data was collected from the high flow system, which used a pump to generate a flow. The phase of data was collected along with the amplitudes of the responses. Again, no attempt was made to hold the temperature of the oil constant, or to monitor the amount of time passing between trials. The third set of data collected recorded the response of the PA, holding the flow rate constant, and varying the temperature, as well as the passage of time between trials. The fourth set of data collected was similar to the second, however the temperature of the oil, and the amount of time between trials was held constant. All flow rates are given in cm/sec, because the viscosity of the oil had not been determined.

7.3 Constant Flow Response

After a significant amount of variability was detected in preliminary measurements, several measurements were taken to determine whether the responses would change over time, even as the flow rate was held constant. Figure 12 shows one such set of measurements.

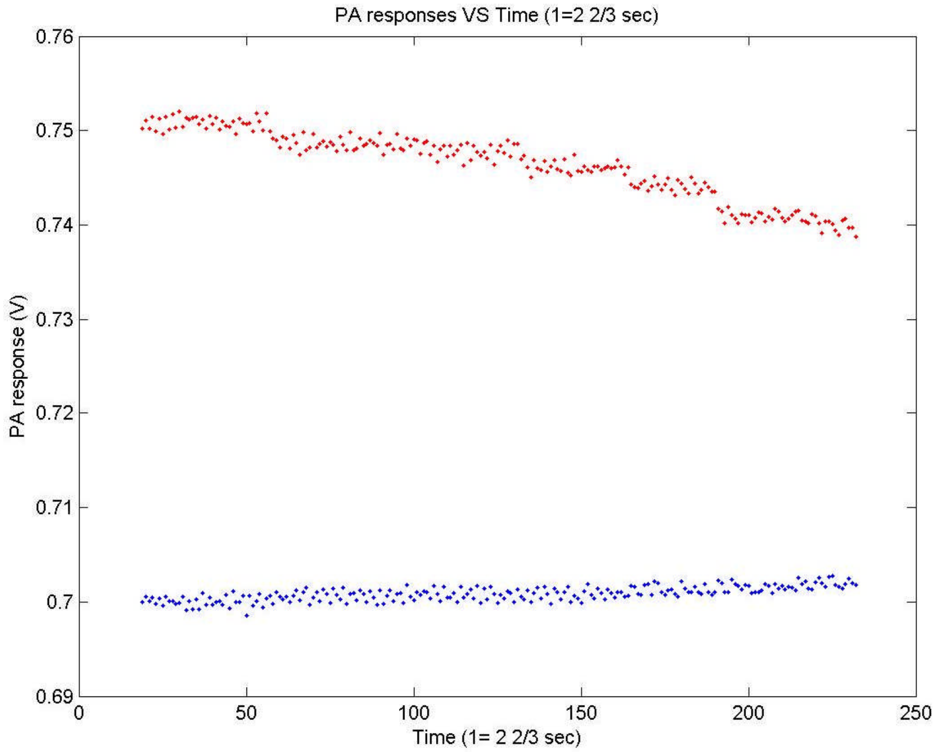


Figure 12: PA responses VS. time with pump current at .35 amps

This data was collected by setting up the pump in a closed loop system and allowing it to run as measurements were taken. Under ideal conditions, the two responses should remain constant because the flow rate should be constant. However, it is believed that neither the flow rate, nor the speed of the pump was constant over the course of this measurement as will be described later. Further measurements were taken, and it was determined that after a period of time the signal levels reached a final level (though this level was not the same between trials).

Another experiment was performed to determine the temperature sensitivity of the PA. At room temperature (20^{o}C), the zero flow responses of the two electrodes are approximately 15mV. However, if the oil is heated to ~40^{o}C, the responses drop to ~11.7mV. When the oil temperature is 0^{o}C, the responses remained at 15mV.

7.4 PA Response to Low Flow

The data collected in these trials was taken by filling the column of oil in the low flow system and allowing the liquid to drain into the collection container on the balance. This was done four times, and the results are presented in figures 13 and 14.

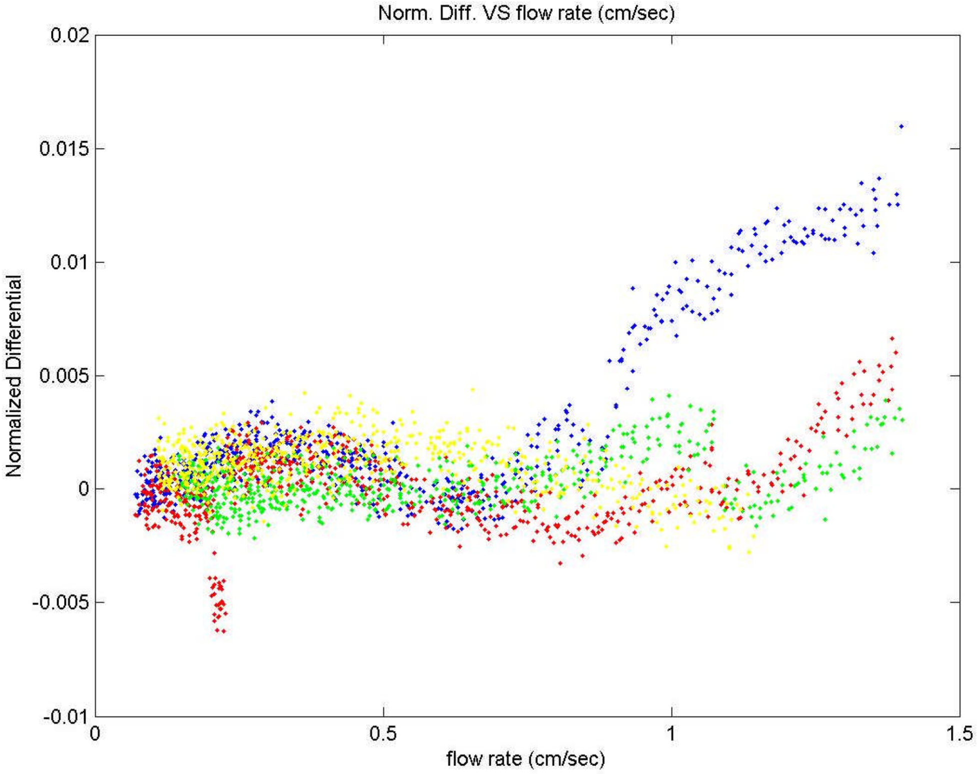


Figure 13: Normalized Differential Response at low flow rates on linear scale

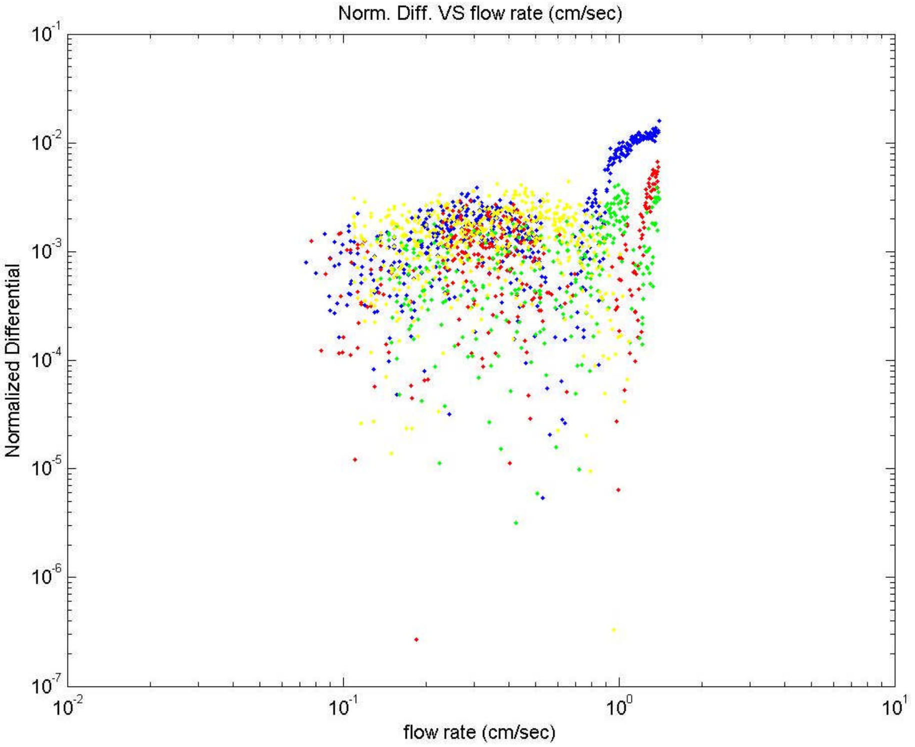


Figure 14: Normalized Differential Response in low flow rates on log log scale

The results are clearly not reproducible across trials. It appears that the response is increasing as the flow rate increases, but this varies between trials. There also appear to be long term fluctuations in the response. This is believed to be caused by a change in temperature, as will be explain later. The temperature of the oil most likely fluctuated over the course of the experiment, as the temperature of the oil and the tubing approached the same temperature. It is believed that either the measurement system is not sensitive enough to detect the response accurately, or that the PA response itself is not significant at these flow rates. The measurements made in the gases had a “noise floor” below which the response could not be measured as a result of thermal noise, as well as noise in caused by the electronics. It is possible that the situation is similar at these rates in the liquids.

The oil was initially at room temperature, as was the pump. As the pump circulates the oil, there are two occurrences which could possibly explain this data. The friction of the moving parts of the pump causes its temperature to rise. This produces two results. First, the speed of the pump, (and therefore the flow rate) decreases slightly because of the increased resistance. Second, the oil begins to heat up as it draws heat away from the pump.

7.5 PA Response to High Flow Rates

Two sets of data were collected at high flow rates. The first set of data consists of 8 separate trials during which the effects of temperature and the time elapsed between trials were disregarded. The speed of the pump was varied gradually to achieve different flow velocities. The normalized differential response vs. flow rate is shown in figure 15.

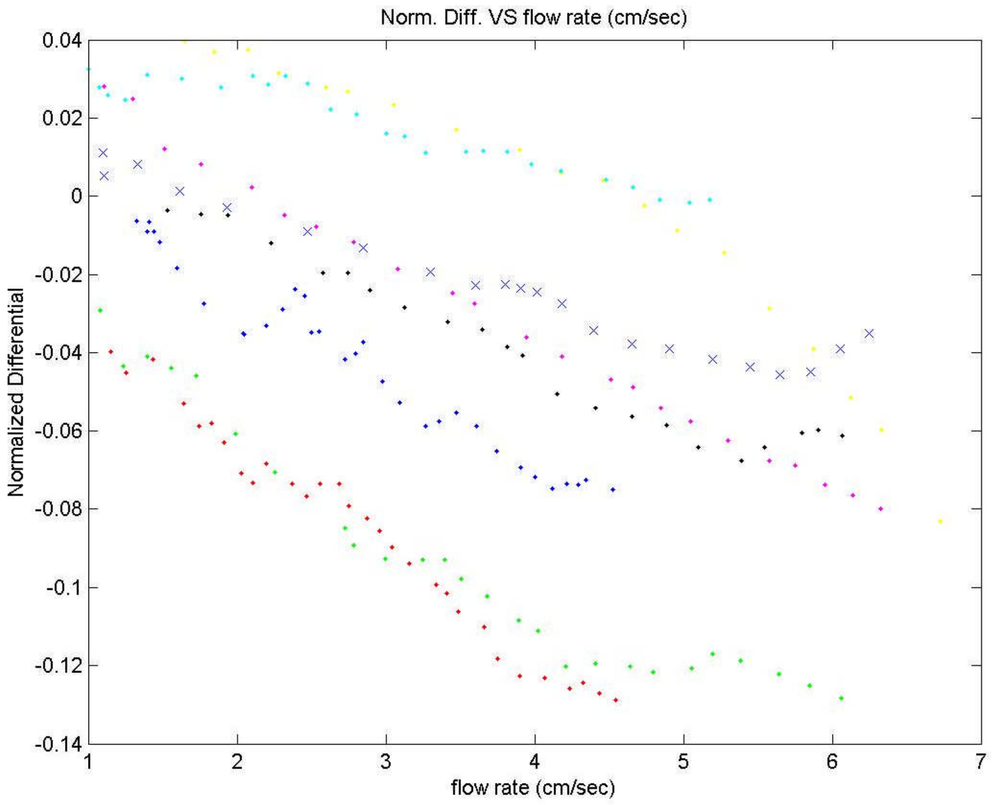


Figure 15: Norm. Diff. vs flow rate, neglecting effects of temperature

The second set of data was collected in a way to minimize the effects of temperature and the passage of time. The pump was run in a closed loop system for 30 minutes with the PA electronics on to allow the temperature of the system to reach an equilibrium. Then eight trials were run, with approximately 30 seconds between trials. The trials appeared to be more repeatable, as shown in figure 16.

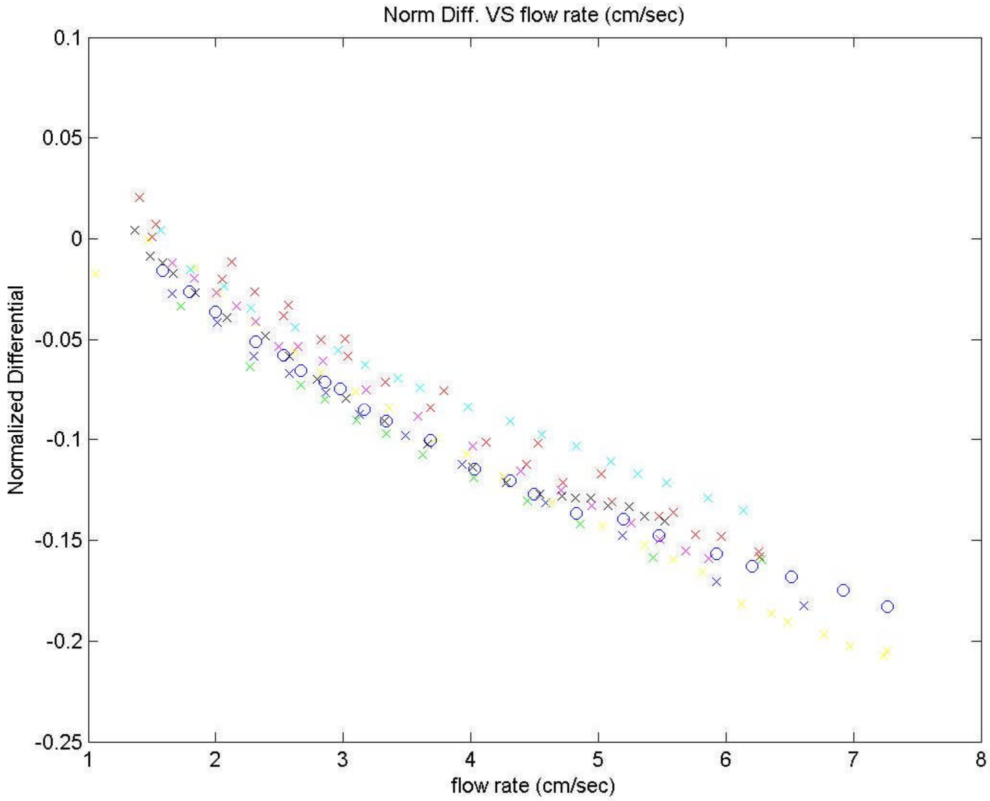


Figure 16. Normalized Differential vs flow rate under steady temperature conditions in linear scale

The phase for the second set of high flow measurements is shown in figure 17. It is clear that a direct comparison to previous research will not be possible because the phase of the two electrodes relative to one another changes with flow rate. This did not happen in measurements taken in gases.

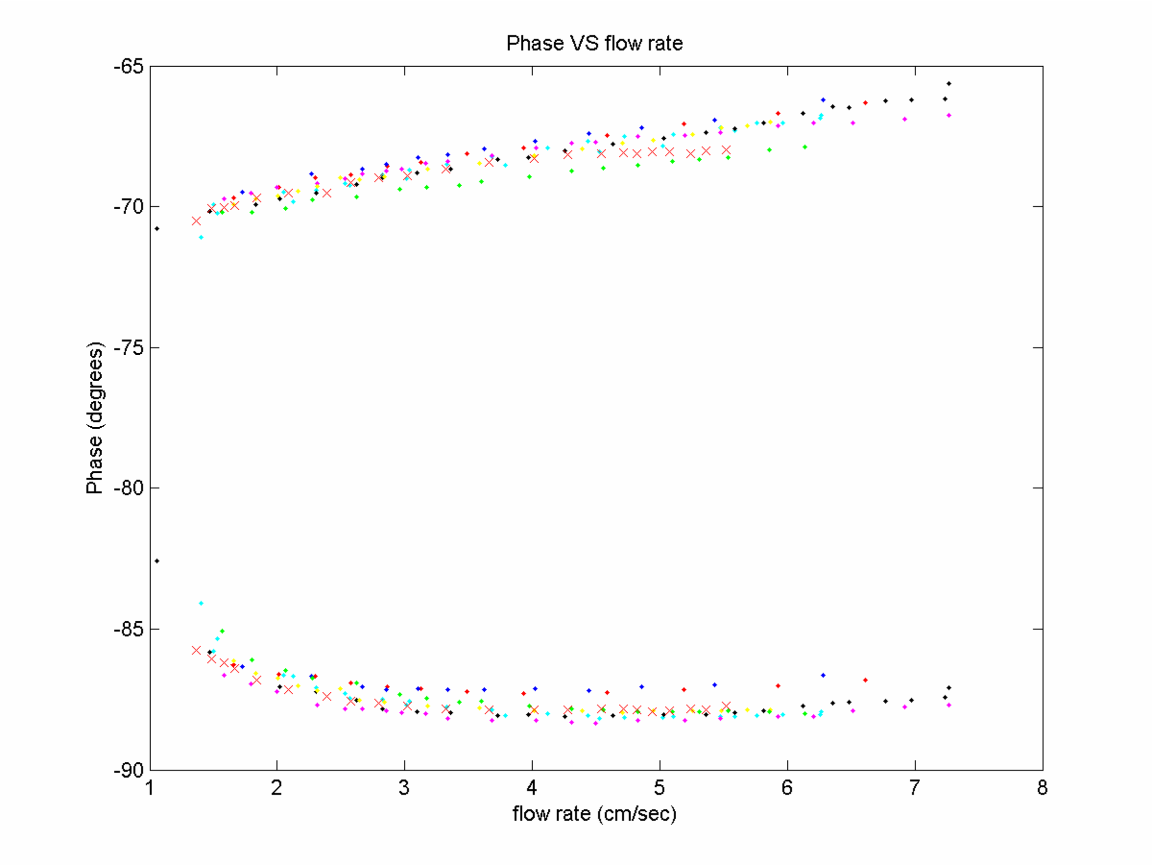


Figure 17: Phases of the two electrodes versus flow rate (phase of upstream electrode greater than -70, downstream electrode less than -80)

8. DISCUSSION AND CONCLUSIONS

8.1 The Effects of Temperature

The experimental results indicate that there are similarities, as well as differences between the results of previous research and this study. There are several major differences between the properties of helium, nitrogen, and argon gases and the properties of the oil tested. First, the thermal conductivity of these gases is much lower than that of the PA and of the oil. Second, the viscosity of the oil was much more sensitive to changes in temperature than that of the gases. This is important because the Reynolds number of the flow is given by $\mathrm{Re} = \dfrac{V \cdot D}{v}$ where V is the velocity of the flow, D is the diameter of the tubing, and v is the kinematic viscosity of the oil. This means that as the viscosity of the oil changes, the Reynolds number (and therefore the PA response) will change as well, even if the volumetric velocity of the oil remains constant. Since the PA response is related to the Reynolds number, it is important to examine in detail how changes in temperature affect the viscosity, the resulting Reynolds number and the PA response.

Several observations were made of the temperature over the course of the measurements, and it was found that the temperature of the oil varied in the range 20-25^{o}C. Room temperature was ~20^{o}C. It is believed that as the oil circulates through the pump, it is heated, depending on how fast the pump is running, and how long it has been on. When the oil is collected on the balance, it cools down again, approaching room temperature. A viscosity-temperature curve of the oil that was used (transmission fluid) was not available, but one was available for SAE 10W oil, which is believed to have similar characteristics. Over the observed temperature range, the kinematic viscosity of 10W oil varies from .0001 to .00006 m^2/sec. It is also noted that the specification of 10W oil allows for a 50% variation in these figures “especially at low temperatures”.[2] It is not known what the specification for viscosity of the transmission fluid is for these temperatures. This would result in a range of error of ± 30% in the Reynolds number (and therefore PA response, assuming a linear relationship) for a constant volumetric flow.

Further complicating the effects of temperature on the PA’s behavior is the observation that the single electrode response of the PA appears to decrease with an increase in temperature. It was observed that the amplitude of the single electrode response at zero flow decreased from 15mV to 11.7mV when the temperature was changed from 20^{o}C to ~40^{o}C. This results in a single electrode response of 14.2mV at 25^{o}C assuming linearity. The exact signal-temperature relationship was not determined. This change in signal level can have a profound effect on the normalized differential response if the temperature change occurs during the trial. If the sensor’s zero flow response is measured when the temperature is 20^{o}C, and a normalized difference measurement is taken over temperatures ranging from 20-25^{o}C (as is believed to be the case), a significant amount of variability would be expected.

The exact effect of temperature is difficult to determine because the temperature of the liquid was not recorded during any of the flow measurements. However, these preliminary calculations seem to suggest that the effects of a relatively minor fluctuation in temperature (+5^{o}C) could result in relatively large errors.

8.2 Comparisons to Previous Research

There were three main points to be compared to the previous research. The first was that the PA response is linearly related to the flow rate for lower Reynolds numbers. While the Reynolds numbers of the flows used could not be determined, there appears to be a linear relationship between the volumetric flow rate of the oil and the PA response. Assuming the viscosity of the oil remained relatively constant, the PA response is found to be linearly related to the Reynolds number as had been observed in gases. However, the significant fluctuation in signals between trials, and within the trials themselves, suggest that the most important factor affect the PA response may be the temperature.

The second result of previous research was that the phase between the two electrodes of the PA with respect to each other remained constant regardless of the flow rate. This can not be said for the data collected for flows in oil. The phase of both electrodes clearly changes with respect to the heater signal and with respect to each other. This makes it difficult to make a comparison to the phase changes observed in the gases.

The third result of the research conducted previously was that the PA response (both single electrode and normalized differential) was not significantly affected by small changes in temperature. The effect of temperature was not under study in previous research, but it appears as though the changes in temperature (if they occurred) did not lead to variability in signal levels of the PA. At room temperature, the viscosity of gases is much less dependent on temperature than oil. [2]

9. RECOMMENDATIONS

This study uncovered several characteristics of liquid flows that could possibly warrant changes in the current design of the measurement and flow systems. The most obvious is the apparent effect changes in temperature have on the PA response. Future research should attempt to keep the temperature of the liquid as constant as possible during and between trials, and also to quantitatively determine the effect temperature has on the response of the PA. The temperature should be monitored during all trials. Also, switching the order of the PA and pump in the tubing (so that any heat generated by the pump would not effect the temperature as it passes the PA) might improve the variability of the measurements. It also might be beneficial to develop a system that does not require a pump, such as a larger version of the low flow system developed, as it is likely the temperature of the oil in this type of system would remain more constant. It might also be worthwhile to consider the temperature vs. viscosity curve when selecting future liquids to be tested, as it would be desirable to keep the viscosity as constant as possible.

It would also be beneficial to develop a system that could handle gas as well as liquid flows. This would allow more direct comparisons to previous research and would ensure that the measurement electronics themselves are not contributing to the uncertainty and error in the measurements.

10. ACKNOWLEDGMENTS

I would like to thank Dr. Zemel for his help, advice, support and encouragement in completing this project. I would also like to thank the National Science Foundation for their support through an NSF-REU grant and Microsoft Corporation for their financial support.

11. REFERENCES

1. Hsieh,Hsin-Yi, . "Pyroelectric Anemometers. A Dissertation in Electrical Engineering" 1993.
2. F.M. White, *Fluid Mechanics*, Mc Graw-Hill, New York, 3rd ed., 1994, p. 700-702.

APPENDIX

JAVA CODE

This code was compiled an run on a PC using the javacomm API available on sun.com. These java classes read the data sent over the rs232 line from the microcontroller and from the electronic balance and determine the amplitude of each electrode as well as the flow rate.

(start LIA.java)

```
/*
 * LIA.java
 *
 * Created on June 16, 2005, 11:22 AM
 */

package commapi.samples.Simple;

/**
 *
 * @author  Rob
 */
public class LIA {
    public Weave weave;
    public String name;
    public int ptsIndex=0,ptsPerWave=16;
    public double[] pts= new double[256];
    public double[] sines= new double[pts.length], cosines= new double[pts.length];
    public double amplitude,phase;
    public boolean full= false;
    public double calculateAmplitude(){
        double sineSummation=0, cosineSummation=0;
        for(int i= 0; i<pts.length; i++){
            sineSummation+= pts[i]*sines[i];
            cosineSummation+= pts[i]*cosines[i];
        }
        return
Math.sqrt(sineSummation*sineSummation+cosineSummation*cosineSummation)*2./pts.length;
    }
    public double calculatePhase(){
        double sineSummation=0, cosineSummation=0;
        for(int i= 0; i<pts.length; i++){
            sineSummation+= pts[i]*sines[i];
            cosineSummation+= pts[i]*cosines[i];
        }
        return Math.atan(cosineSummation/sineSummation)*180/Math.PI;
    }
    public void printAll(){
        System.out.println(name+"pts: ");
        for(int i= 0; i<pts.length; i++){
            System.out.println(pts[i]);
        }System.out.println();

    }
    public void printIDAC(){
        System.out.println(name+"IDAC: ");
        for(int i= 0; i<pts.length; i++){
            System.out.println(IDAC[i]);
        }System.out.println();

    }
    public final double LSB= 3.8147e-5;
    /** Creates a new instance of LIA */
    public LIA(String name,Weave weave) {
        this.name= name;
        this.weave= weave;
        for(int i= 0; i<pts.length; i++){
            sines[i]= Math.sin((i*2*Math.PI)/ptsPerWave);
            cosines[i]= Math.cos((i*2*Math.PI)/ptsPerWave);
        }
    }
    public int[] IDAC= new int[pts.length];
    public void getNextInt(int next,int IDAC){
        if(full) {System.out.println("ERROR: LIA was full, tried to rewrite."); System.exit(0);}
        this.IDAC[ptsIndex]= IDAC;
        pts[ptsIndex++]= LSB*next;

        if(ptsIndex==pts.length){

            //System.out.println(name+" amp: "+calculateAmplitude());
            weave.LIADone(this);
            ptsIndex=0;
        }

    }
    double phaseX= Math.random(), phase2= Math.random();

    public double signalSource(int i, double percentDiff){
        return Math.sin(i*2*Math.PI/(ptsPerWave)*(1+percentDiff));
    }
    public void fill(int i,double percentDiff){
        phase= Math.random();
        for(int j= 0; j<pts.length; j++){
            pts[j]= signalSource(i+j,percentDiff);
        }
    }
     public static void main(String[] args) {
        LIA l= new LIA("NONE",null);
        for(int i= 0; i<50; i++){
            l.fill(0,(i-25)*10./50./100.);
            System.out.println((i-25)*10./50+"\t"+l.calculateAmplitude()+"\t"+l.calculatePhase()+";");
        }


    }
}
```

(end LIA.java)

(start SimpleRead.java)

```
/*
 * @(#)SimpleRead.java    1.12 98/06/25 SMI
 *
 * Copyright (c) 1998 Sun Microsystems, Inc. All Rights Reserved.
 *
 * Sun grants you ("Licensee") a non-exclusive, royalty free, license
 * to use, modify and redistribute this software in source and binary
 * code form, provided that i) this copyright notice and license appear
 * on all copies of the software; and ii) Licensee does not utilize the
 * software in a manner which is disparaging to Sun.
 *
 * This software is provided "AS IS," without a warranty of any kind.
 * ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES,
 * INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A
 * PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE HEREBY EXCLUDED. SUN AND
 * ITS LICENSORS SHALL NOT BE LIABLE FOR ANY DAMAGES SUFFERED BY
 * LICENSEE AS A RESULT OF USING, MODIFYING OR DISTRIBUTING THE
 * SOFTWARE OR ITS DERIVATIVES. IN NO EVENT WILL SUN OR ITS LICENSORS
 * BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR DIRECT,
 * INDIRECT, SPECIAL, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES,
 * HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING
 * OUT OF THE USE OF OR INABILITY TO USE SOFTWARE, EVEN IF SUN HAS BEEN
 * ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.
 *
 * This software is not designed or intended for use in on-line control
 * of aircraft, air traffic, aircraft navigation or aircraft
 * communications; or in the design, construction, operation or
 * maintenance of any nuclear facility. Licensee represents and
 * warrants that it will not use or redistribute the Software for such
 * purposes.
 */
package commapi.samples.Simple;
import java.io.*;
import java.util.*;
import javax.comm.*;

public class SimpleRead implements Runnable, SerialPortEventListener, CommPortOwnershipListener {
    static CommPortIdentifier portId;
    static Enumeration portList;
    LIA posElectrode,negElectrode;
    InputStream inputStream;
    OutputStream outputStream;
    SerialPort serialPort;
    Thread readThread;



    public static void printBytes(byte[] b){
        for(int i= 0; i<b.length; i++){
            System.out.print(Integer.toBinaryString((b[i]&0x1FF)|0x100)+" ");

        }
        System.out.println();
    }
    public SimpleRead(int num,LIA pos, LIA neg) {
```

```
  try{
    portId = CommPortIdentifier.getPortIdentifier("COM"+num);
    portId.addPortOwnershipListener(this);
  }catch(Exception e){
    System.out.println("There is no such port");
  }
  try {
    serialPort = (SerialPort) portId.open("SimpleReadApp", 2000);
  } catch (PortInUseException e) {}
  try {
    inputStream = serialPort.getInputStream();
    outputStream = serialPort.getOutputStream();
  } catch (IOException e) {}
     try {
    serialPort.addEventListener(this);
     } catch (TooManyListenersException e) {}
  serialPort.notifyOnDataAvailable(true);
  try {
    serialPort.setSerialPortParams(115200,
      SerialPort.DATABITS_8,
      SerialPort.STOPBITS_1_5,
      SerialPort.PARITY_NONE);
  } catch (UnsupportedCommOperationException e) {}
  String s;


  posElectrode= pos;
  negElectrode= neg;
  try{
    s= "\r";outputStream.write(s.getBytes());
    System.out.println("Hit enter for PA electrodes.");
    while(System.in.available()==0)  ;
    s= "\r";outputStream.write(s.getBytes());
  }catch(Exception e){System.out.println("Couldn't write");}

  readThread = new Thread(this);
  readThread.start();
}

public void run() {
  try {
    Thread.sleep(20000);
  } catch (InterruptedException e) {}
}

public void serialEvent(SerialPortEvent event) {
  switch(event.getEventType()) {
  case SerialPortEvent.BI:
  case SerialPortEvent.OE:
  case SerialPortEvent.FE:
  case SerialPortEvent.PE:
  case SerialPortEvent.CD:
  case SerialPortEvent.CTS:
  case SerialPortEvent.DSR:
  case SerialPortEvent.RI:
  case SerialPortEvent.OUTPUT_BUFFER_EMPTY:
        break;
    case SerialPortEvent.DATA_AVAILABLE:
        byte[] readBuffer = new byte[200];
        int numBytes=0;
        try {
            while (inputStream.available() > 0) {
                numBytes = inputStream.read(readBuffer);

            }
            //displayText(readBuffer,numBytes);
            groupBytes(readBuffer,numBytes);

        } catch (IOException e) {}
        break;
    }
}
public int byteI=0;
public byte[] twoInts= new byte[5];
public void groupBytes(byte[] bytes, int byteCount){
    for(int i= 0; i<byteCount; i++){
        if(byteI==5){
            //printBytes(twoInts);

            posElectrode.getNextInt(((twoInts[1]<<8)&0xFF00)|(twoInts[2]&0xFF),twoInts[0]);
            negElectrode.getNextInt(((twoInts[3]<<8)&0xFF00)|(twoInts[4]&0xFF),twoInts[0]);
            byteI= 0;
        }
        twoInts[byteI]= bytes[i];
        byteI++;
    }
}
private String displayText(byte[] bytes, int                           byteCount)
{
    String          str;
    int  i,
            idx;
    byte[]          nb;



    nb = new byte[byteCount * 4];

    for (i = 0, idx = 0; i < byteCount; i++)
    {
            /*  Wrap any control characters      */



                    nb[idx++] = bytes[i];

    }
    str="";
    try{
            byte[] newa= new byte[idx];
            System.arraycopy(nb,0,newa,0,idx);
            System.out.print("REC: ");
            printBytes(newa);
            str = new String(nb, 0, idx);

            //System.out.println("DISPLAYING: "+str);
            //printBytes(str.getBytes());
            //this.text.append(str);
        }catch(Exception e){System.out.println("OOPS");}
        return(str);
    }
    public void ownershipChange(int param) {
    }



}
```

(end SimpleRead.java)

(start SimpleRead_1.java)

```
/*
 * @(#)SimpleRead_1.java           1.12 98/06/25 SMI
 *
 * Copyright (c) 1998 Sun Microsystems, Inc. All Rights Reserved.
 *
 * Sun grants you ("Licensee") a non-exclusive, royalty free, license
 * to use, modify and redistribute this software in source and binary
 * code form, provided that i) this copyright notice and license appear
 * on all copies of the software; and ii) Licensee does not utilize the
 * software in a manner which is disparaging to Sun.
 *
 * This software is provided "AS IS," without a warranty of any kind.
 * ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES,
 * INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A
 * PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE HEREBY EXCLUDED. SUN AND
 * ITS LICENSORS SHALL NOT BE LIABLE FOR ANY DAMAGES SUFFERED BY
 * LICENSEE AS A RESULT OF USING, MODIFYING OR DISTRIBUTING THE
 * SOFTWARE OR ITS DERIVATIVES. IN NO EVENT WILL SUN OR ITS LICENSORS
 * BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR DIRECT,
 * INDIRECT, SPECIAL, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES,
 * HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING
 * OUT OF THE USE OF OR INABILITY TO USE SOFTWARE, EVEN IF SUN HAS BEEN
 * ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.
 *
 * This software is not designed or intended for use in on-line control
 * of aircraft, air traffic, aircraft navigation or aircraft
 * communications; or in the design, construction, operation or
 * maintenance of any nuclear facility. Licensee represents and
 * warrants that it will not use or redistribute the Software for such
 * purposes.
 */
package commapi.samples.Simple;
import java.io.*;
import java.util.*;
import javax.comm.*;

public class SimpleRead_1 implements Runnable, SerialPortEventListener, CommPortOwnershipListener
{
    static CommPortIdentifier portId;
    static Enumeration portList;
    static LIA posElectrode,negElectrode;
    InputStream inputStream;
    OutputStream outputStream;
    SerialPort serialPort;
   Thread readThread;

    public static void main(String[] args) {
        //portList = CommPortIdentifier.getPortIdentifiers();
        SimpleRead_1 x= new SimpleRead_1(4);


    }
    public static void printBytes(byte[] b){
        for(int i= 0; i<b.length; i++){
            System.out.print(Integer.toBinaryString((b[i]&0x1FF)|0x100)+" ");
   }
   System.out.println();
}
public SimpleRead_1(int num) {
   try{
      portId = CommPortIdentifier.getPortIdentifier("COM"+num);
      portId.addPortOwnershipListener(this);
   }catch(Exception e){
      System.out.println("There is no such port");
   }
   try {
      serialPort = (SerialPort) portId.open("SimpleReadApp", 2000);
   } catch (PortInUseException e) {}
   try {
      inputStream = serialPort.getInputStream();
      outputStream = serialPort.getOutputStream();
   } catch (IOException e) {}
        try {
      serialPort.addEventListener(this);
        } catch (TooManyListenersException e) {}
   serialPort.notifyOnDataAvailable(true);
   try {
      serialPort.setSerialPortParams(2400,
         SerialPort.DATABITS_7,
         SerialPort.STOPBITS_1,
         SerialPort.PARITY_EVEN);
   } catch (UnsupportedCommOperationException e) {}
   String sir= "SIR\r\n";
   try{outputStream.write(sir.getBytes());}catch(Exception e){}
   readThread = new Thread(this);
   readThread.start();
}

public void run() {
   try {
      Thread.sleep(20000);
   } catch (InterruptedException e) {}
}
double[] masses= new double[600];//5*120secs
int massIndex= 0;

public boolean isSteady(){
   return true;
}
public double getRateAndReset(){
   //for(int i= 1; i<massIndex; i++) System.out.print((masses[i]-masses[i-1])*4+"  ");
   //System.out.println();

   double returnD= (masses[massIndex-1]-masses[0])*4/(massIndex-1);//-1?
   massIndex= 0;
   return returnD;
}
public byte[] line= new byte[17];//the maximum size of a line xxxnnnnnnnnnxxx\n
public int lineIndex=0;
public void serialEvent(SerialPortEvent event) {
    switch(event.getEventType()) {
    case SerialPortEvent.BI:
    case SerialPortEvent.OE:
    case SerialPortEvent.FE:
    case SerialPortEvent.PE:
    case SerialPortEvent.CD:
    case SerialPortEvent.CTS:
    case SerialPortEvent.DSR:
    case SerialPortEvent.RI:
    case SerialPortEvent.OUTPUT_BUFFER_EMPTY:
      break;
    case SerialPortEvent.DATA_AVAILABLE:
      try {
        while (inputStream.available() > 0) {
          inputStream.read(line,lineIndex,1);
          if(line[lineIndex]=='\n'){
            byte[] number= new byte[9];
            for(int i= 0; i<9; i++){
              number[i]= line[(lineIndex+17-13+i)%17];//circular
            }
            String s= new String(number);

            masses[massIndex++]= Double.parseDouble(s);

          }
          lineIndex= (lineIndex+17+1)%17;//circular
        }
        //displayText(readBuffer,numBytes);
        //groupBytes(readBuffer,numBytes);

      } catch (Exception e) {System.out.println("Missed a measurement");}
      break;
    }
  }
  public void ownershipChange(int param) {
  }



}
```

(end SimpleRead_1.java)

(start SimpleWrite.java)

```
/*
 * @(#)SimpleWrite.java   1.12 98/06/25 SMI
 *
 * Copyright (c) 1998 Sun Microsystems, Inc. All Rights Reserved.
 *
 * Sun grants you ("Licensee") a non-exclusive, royalty free, license
 * to use, modify and redistribute this software in source and binary
 * code form, provided that i) this copyright notice and license appear
 * on all copies of the software; and ii) Licensee does not utilize the
 * software in a manner which is disparaging to Sun.
 *
 * This software is provided "AS IS," without a warranty of any kind.
 * ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES,
 * INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A
 * PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE HEREBY EXCLUDED. SUN AND
 * ITS LICENSORS SHALL NOT BE LIABLE FOR ANY DAMAGES SUFFERED BY
 * LICENSEE AS A RESULT OF USING, MODIFYING OR DISTRIBUTING THE
 * SOFTWARE OR ITS DERIVATIVES. IN NO EVENT WILL SUN OR ITS LICENSORS
 * BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR DIRECT,
 * INDIRECT, SPECIAL, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES,
 * HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING
 * OUT OF THE USE OF OR INABILITY TO USE SOFTWARE, EVEN IF SUN HAS BEEN
 * ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.
 *
 * This software is not designed or intended for use in on-line control
 * of aircraft, air traffic, aircraft navigation or aircraft
 * communications; or in the design, construction, operation or
 * maintenance of any nuclear facility. Licensee represents and
 * warrants that it will not use or redistribute the Software for such
 * purposes.
 */
package commapi.samples.Simple;
import java.io.*;
import java.util.*;
import javax.comm.*;

public class SimpleWrite {
    static Enumeration portList;
    static CommPortIdentifier portId;
    static String messageString = "\n";
    static SerialPort serialPort;
    static OutputStream outputStream;

    public static void main(String[] args) {
        portList = CommPortIdentifier.getPortIdentifiers();

        //while (portList.hasMoreElements()) {
            try{
            portId = (CommPortIdentifier) CommPortIdentifier.getPortIdentifier("COM4");
            }catch(Exception e){System.out.println("NO");}
            if (portId.getPortType() == CommPortIdentifier.PORT_SERIAL) {
                // if (portId.getName().equals("COM1")) {
                if (true){//portId.getName().equals("/dev/term/a")) {
                    try {
                        serialPort = (SerialPort)
```

```
              portId.open("SimpleWriteApp", 2000);
          } catch (PortInUseException e) {System.out.println("IN USE");}
          try {
            outputStream = serialPort.getOutputStream();
          } catch (IOException e) {}
          try {
            serialPort.setSerialPortParams(9600,
              SerialPort.DATABITS_8,
              SerialPort.STOPBITS_1,
              SerialPort.PARITY_NONE);
          } catch (UnsupportedCommOperationException e) {}
          try {
            outputStream.write(messageString.getBytes());
          } catch (IOException e) {}
        }
      }
    //}
  }
}
```

(end SimpleWrite.java)

(start Weave.java)

```
/*
 * Weave.java
 *
 * Created on July 5, 2005, 4:44 PM
 */

package commapi.samples.Simple;

/**
 *
 * @author  Rob
 */
public class Weave {
    SimpleRead_1 sr1;
    SimpleRead sr;
    LIA pos,neg;
    int cycleCount=0;
    public Weave(){
        pos= new LIA("pos",this);
        neg= new LIA("neg",this);
        sr= new SimpleRead(5,pos,neg);
        sr1= new SimpleRead_1(4);

    }
    public void LIADone(LIA lia){
        if(lia.full) {System.out.println("ERROR: LIA completed twice"); System.exit(0);}
        //System.out.println(lia.name+" is full.");
        lia.amplitude= lia.calculateAmplitude();
        lia.phase= lia.calculatePhase();
        lia.full=true;
        if(pos.full&&neg.full) endCycle();
    }

    public void endCycle(){
        //check if rate is steady enough
        //pos.printIDAC();
        //if(System.in.available()>0){
            //tell SimpleRead_1 to finish, give the flow rate, and restart
            //print all three
            double phaseAdjust= 360./64;

System.out.println(cycleCount+"\t"+pos.amplitude+"\t"+neg.amplitude+"\t"+sr1.getRateAndReset()+"\t\t\t
"+(pos.phase-phaseAdjust)+"\t"+(neg.phase-360/16.-phaseAdjust));
            pos.full= false;
            neg.full= false;
            //set pos and neg to false;
            //else error
            cycleCount++;
        //}

    }
    /**
     * @param args the command line arguments
     */
    public static void main(String[] args) {
        // TODO code application logic here


        Weave w= new Weave();


    }

}
```

(end Weave.java)

```
}*/
```

8051 MICROCONTROLLER CODE

This code was compiled using Raisonance IDE which was supplied along with the TI MSC1202 Evaluation module. It creates an executable image that is loaded into the microcontroller over the RS232 line using TI's supplied bootloader.

(Start adc.c)

```
//**********************************************************************
// File name: adc.c
//
// Copyright 2003 Texas Instruments Inc as an  unpublished work.
//
// Version 1.0
//
// Compiler Version (Keil V2.38), (Raisonance V6.10.13)
//
// Module Description:
//  ADC Example Program
//
//**********************************************************************
//#include "legal.c"        //Texas Instruments, Inc. copyright and liability
#include <reg1200.h>       // The header file with the MSC register definitions
#include <stdio.h>         // Standard I/O so we can use the printf function
#include <math.h>
extern signed long bipolar(void);
extern unsigned long unipolar(void);
extern int getSpi();

#define autobaud()                                      ((void (code *) (void)) 0xFBFA) ();// MSC1200
#define sendByte(BYTE)                                  ((void (code *) (char)) 0xFBEA) (BYTE);    //
MSC1200

#define LSB (3.8147e-5)
#define SET 1
#define CLEAR 0

sbit RedLed = P3^4;
sbit YellowLed= P3^5;

int ticks= 0;
unsigned int voltage;

sbit SCLK=P3^6;
sbit SDA=P1^2;
sbit SSN=P1^1;
sbit ADC_CS = P1^0;
long tCosSum,tSinSum;
/*void liad(int result){
                        //tCosSum+= (result-(65535/2))*(cos(ticks/64.*2*3.1416)));
                        tSinSum+= (result-(65535/2));//*(sin(ticks/64.*2*3.1416)));
                       if(ticks==0){
                        tSinSum= sqrt(tCosSum*tCosSum+tSinSum*tSinSum);
                        tCosSum= 0; tSinSum=0;
                       }
}*/
```

```
void main(void) {
         int adc2; //the better one
         //IDAC= 0x00;
         char is=0;
  RedLed = !RedLed;
         CKCON = 0x10; // MSC1200 Timer1 div 4
         TCON = 0;               // MSC1200 Stop TR1
  autobaud();
  /*while(1){
        scanf("%c",&is);
        //printf("ABCDE");
        sendByte(0x55);
        sendByte(0xf0);
        //sendByte(0x55); sendByte(0x0f);

        }*/
  //printf("ADC Test, ACLK2\n");
  //printf("ADC Test, ACLK3\n");
  //printf("ADC Test, ACLK4\n");
  //printf("ADC Test, ACLK5\n");
  //while(1);

  PDCON = 0x75;        // Turn on the A/D
  PDCON&= 0xBF;
  SYSCLK= 0x00;
  ACLK =3;//3          // ACLK freq. = XTAL Freq./(ACLK +1) = 0.9216 MHz
               // 0.9216 Mhz/64 = 14,400 Hz
  DECIMATION = 150;//150    // Dtrata Rate = 14,400/1,440 = 10 Hz
  ADMUX = 0x10;         // AINP = AIN7, AINN = AIN6
  ADCON0 = 0x30;        // Vref On, 2.5V, Buffer Off, PGA=1
  ADCON1 = 0x51;        // unipolar, auto, self calibration, offset, gain

  //TCON= 0x50;//6 is run 1, 4 is run 0
  //CKCON|= 0x18;//4  is 1 div by (0-12,1-4),3 is 0 div by...
  //TMOD= (TMOD&0xF0)|0x01;//0x1 for correct mode


  //#define time 65536-x

  ADMUX= 0x20;
  //IDAC= 0x00;
  YellowLed= !YellowLed;
  //for(is= 0; is<128; is++){
  //while(1){
  //int i= 0;
  //printf("ENTER DECIMATION\n"); scanf("%i\n",&adc2);
  //DECIMATION= adc2;
  //printf("ENTER ACLK\n"); scanf("%i\n",&adc2);
  //ACLK= adc2&0xFF;
  //printf("ACLK %i,DEC %i\n",ACLK,DECIMATION);

        //for(i= 0; i<50; i++){
        while(1){
          char tock=ticks%4;

          while(!(AIPOL&0x20))  ;
```

```
       //voltage= bipolar();
       IDAC= ((char)(127*sin(ticks/(128.)*2*3.1416)))+127;

       voltage= unipolar();
       //while(!TF0);
       //TH0= 256-(256-179)/2-2;//(65536-5000)>>8;
       //TL0= 0x00;//(65536-5000)&0xFF;

       //TF0=0;
       ticks++;
  //while(ticks>127) ;
  ticks= ticks&(128-1);

  //sendByte(is&0xFF);
  if(ticks%128==1) {
              //printf("%f\n",voltage*LSB);
  }
  if (tock==0){
      ADMUX= 0x10;//
      sendByte(ticks&0xFF);
      sendByte((voltage>>8)&0xFF);
      sendByte(voltage&0xFF);

  }
  else if(tock==2){
               ADMUX= 0x20;
               sendByte((voltage>>8)&0xFF);
      sendByte(voltage&0xFF);

       }







}
/*while(1){
  while(!(AIPOL&0x20))  ;

  voltage= LSB*bipolar();//unipolar();
  //adc2= getSpi();
  if(ticks%4==0){
   ADMUX= 0x10;
  }
  if(ticks%4==2){
               ADMUX= 0x23;
       }


  if(ticks==2);//; //ticks++;
    //printf("%f\n",adc2*2.5/65536);
    //printf("%f\n",voltage);
    IDAC= ((char)(127*sin(ticks/(2*256.)*2*3.1416)))+127;
    //printf("V=%li\n",sendSpi(1028));
    ticks= ticks%(2*256);
    //nop
    //nop
    //nop
    //IDAC+= 0x80;
    ticks++;
  }*/



        /*while(1){
          TL0= (0xA4); TH0=(0x98); TF0= 0;
          //printf("!");
           i++; j=i;
           while(!(AIPOL&0x20))  ;
           j=j%2;
    //if(j==0){ ADMUX=0x23;}
    //else if(j==1){
    ADMUX =0x10;//}

    result=unipolar();     // Save Results
    voltage= result*LSB;

    if(i%65==0){

    //if(j==1)printf("23-1\t");
    //if(j==1)printf ("%f\t",voltage);
    //if(j==3)printf("12-1\t");
    //if(j==0)
    printf ("%f\n",voltage);

    }
    IDAC= ~IDAC;
        //IDAC= ((char)(127*sin(j/128.*2*3.1416)))+127;
           //while(!(AIPOL&0x20))  ;
           //result=unipolar();     // Save Results

          while(!(TF0));
        }/*
        j=1;
  while(1){
     // Waiting for conversion
    if((j%4)==0||(j%4)==1) ADMUX=0x23;
    else ADMUX =0x10;
    IDAC= ~IDAC;
    result=unipolar();     // Save Results
    lastResult= result;
    j++;

  }
  //long getSysTime(){
  //   long time*/
  while(1);
}
```

(end adc.c)

(start utilities.a51)

```
;*********************************************************************
; File name: utilities.a51
;
; Copyright 2003 Texas Instruments Inc as an unpublished work.
; All Rights Reserved.
;
; Revision History
;           Version 1.0
;
; Assembler Version  (Keil V2.38), (Raisonance V6.10.13)
;
; Module Description:
; ADC routines to read 24-bit ADC and return the value as a long integer.

;*********************************************************************
;$include (legal.a51) ; Texas Instruments, Inc. copyright and liability
$include (reg1200.inc)

;*********************************************************************
PUBLICunipolar, bipolar

adc_sub SEGMENT  CODE
            RSEG  adc_sub

;;;;;;;;;;;;;;;;;;;;
; unsigned long unipolar(void)
; return the 3 byte adres to R4567 (MSB~LSB)
; unsigned long int with R4=0
unipolar:
            mov     r4,#0
            mov     r5,adresh
            mov     r6,adresm
            mov     r7,adresl
            ret

;;;;;;;;;;;;;;;;;;;;
; signed long bipolar(void)
; return the 3 byte adres to R4567 (MSB~LSB)
; return signed long int with sign extendsion on R4
bipolar:
            mov     r4,#0
            mov     a,adresh
            mov     r5,a
            mov     r6,adresm
            mov     r7,adresl
            jnb     acc.7,positive
            mov     r4,#0ffh
positive:
            ret

;;;;;;;;;;;;;;;;;;;;
; signed long read_sum_regs(void)
; return the 4 byte sumr to R4567 (MSB~LSB)
; return signed long int, sign extension done by hardware
read_sum_regs:
```

```
mov	r4, SUMR3;
mov	r5, SUMR2;
mov	r6, SUMR1;
mov r7, SUMR0;
ret

end
```

```
$NOMOD51
$include (REG1200.INC)

;PUBLIC _spim_send_recv_byte
;.org 0x00
PUBLIC getSpi

spim_routines       SEGMENT  CODE
            RSEG    spim_routines
delay equ 32d

waitShort macro insts;( ;must be 6 atleast)
   add A,#insts-4d
   rrc  A
   mov r0,A
  jnc $+2
   nop
   djnz r0,$
endm
   ;mov r,
off macro
anl P3,#10111111b
endm
on macro
orl P3,#01000000b
endm

wait macro insts
   mov A,#((insts)-9d)
   call waitsub
endm

toggle macro      ;starts high
             off
             add A,#0
   wait delay
   on
endm
waitsub:;the number in A+1(so put the number in a and call)
   ;;MUST BE AT LEAST 12
   rrc  A
   mov r0,A
  jnc $+2
   nop
   djnz r0,$
   ret
getSpi:
            orl P1DDRL,#11000000b;set P1.3 as input
            anl P1,#11111110b ; CS to low
            mov r7,#25d
            mov r7,#0x00
            mov r6,#0x00
            mov r1,#22d
            mov r2,#02
lp:
            toggle
            mov A,r7
            add A,#0
            rlc A   ;;carry?
            mov r7,A
            mov A,r6
            rlc A
            mov r6,A
            mov A,P1;get result
            anl A,#00001000b
            add A,#11111000b
            mov A,#0
            addc A,r7
            mov r7,A
            wait delay-15d;+1d;1 for clearing carry
            djnz r1,lp
            djnz r2,store_result
            orl P1,#01b ;CS to high
            mov A,r4
            mov r7,A
            mov A,r5
            mov r6,A
            ;jmp getSpi
            ret
store_result:
            mov A,r7
            mov r4,A
            mov A,r6
            mov r5,A
            mov r1,#10d
            jmp lp

;garbage
;           call spiSub;
;           mov r2,r6
;           mov r3,r7
;           call spiSub
;           mov A,r3
;           add r7,A
;           mov A,r2
;           addc r6,A
;           add A,#0
;           mov A,r6
;           rr A
;           mov r6,A
;           mov A,r7
;           rrc A
;           mov r7,A
;           ret

;DSEG
;    bit: 1

end
```

(end utilities.a51)

(start reg1200.inc)

```
;*******************************************************************
; File name: reg1200.inc
;
; Copyright 2004 Texas Instruments Inc as an  unpublished work.
; Created By:  Ritu Ghosh - Russell Anderson
;
; Version 1.0 Initial Version 12/03/2003
;
; Compiler Version (Keil V2.38), (Raisonance V6.10.14)
;
; Module Description:
; Header file for TI MSC1200 microcontroller
;
;*******************************************************************

$NOMOD51
$SAVE
$NOLIST
;  BYTE Registers


SP      DATA    081H ;STANDARD 8051
DPL     DATA    082H ;STANDARD 8051
DPL0    DATA    082H ;STANDARD 8051
DPH     DATA    083H ;STANDARD 8051
DPH0    DATA    083H ;STANDARD 8051
DPL1    DATA    084H
DPH1    DATA    085H
DPS     DATA    086H
PCON    DATA    087H ;STANDARD 8051
TCON    DATA    088H ;STANDARD 8051
TMOD    DATA    089H ;STANDARD 8051
TL0     DATA    08AH ;STANDARD 8051
TL1     DATA    08BH ;STANDARD 8051
TH0     DATA    08CH ;STANDARD 8051
TH1     DATA    08DH ;STANDARD 8051
CKCON   DATA    08EH
MWS     DATA    08FH
P1      DATA    090H ;STANDARD 8051
EXIF    DATA    091H
CADDR   DATA    093H
CDATA   DATA    094H
SCON    DATA    098H ;STANDARD 8051
SCON0   DATA    098H ;STANDARD 8051
SBUF    DATA    099H ;STANDARD 8051
SBUF0   DATA    099H ;STANDARD 8051
SPICON DATA 09AH
I2CCON  DATA    09AH
SPIDATA DATA 09BH
I2CDATA DATA    09BH
AIPOL   DATA    0A4H
PAI DATA 0A5H
AIE DATA 0A6H
AISTAT DATA 0A7H
IE      DATA    0A8H ;STANDARD 8051
```

```
P1DDRL DATA 0AEH
P1DDRH DATA 0AFH
P3      DATA    0B0H ;STANDARD 8051
P3DDRL DATA 0B3H
P3DDRH DATA 0B4H
IDAC    DATA    0B5H
IP      DATA    0B8H ;STANDARD 8051
EWU DATA 0C6H
SYSCLK DATA    0C7H
PSW     DATA    0D0H ;STANDARD 8051
OCL DATA 0D1H
OCM DATA 0D2H
OCH DATA 0D3H
GCL DATA 0D4H
GCM DATA 0D5H
GCH DATA 0D6H
ADMUX DATA 0D7H
EICON   DATA    0D8H
ADRESL DATA 0D9H
ADRESM DATA 0DAH
ADRESH DATA 0DBH
ADCON0 DATA 0DCH
ADCON1 DATA 0DDH
ADCON2 DATA 0DEH
ADCON3 DATA 0DFH
ACC     DATA    0E0H ;STANDARD 8051
SSCON   DATA 0E1H
SUMR0 DATA 0E2H
SUMR1 DATA 0E3H
SUMR2 DATA 0E4H
SUMR3 DATA 0E5H
ODAC DATA 0E6H
LVDCON DATA 0E7H
EIE     DATA    0E8H
HWPCO DATA 0E9H
HWPC1 DATA 0EAH
HWVER DATA 0EBH
FMCON DATA 0EEH
FTCON DATA 0EFH
B       DATA    0F0H ;STANDARD 8051
PDCON   DATA    0F1H
PASEL   DATA    0F2H
PLLL    DATA    0F4H
PLLH    DATA    0F5H
ACLK    DATA    0F6H
SRST    DATA    0F7H
EIP     DATA    0F8H
SECINT DATA 0F9H
MSINT DATA 0FAH
USEC DATA 0FBH
MSECL DATA 0FCH
MSECH DATA 0FDH
HMSEC   DATA    0FEH
WDTCON  DATA    0FFH
```

```
;  BIT Registers
; *** TCON ***
TF1     BIT     08FH
TR1     BIT     08EH
TF0     BIT     08DH
TR0     BIT     08CH
IE1     BIT     08BH
IT1     BIT     08AH
IE0     BIT     089H
IT0     BIT     088H

; *** P1 ***
INT5    BIT     097H
INT4    BIT     096H
INT3    BIT     095H
INT2    BIT     094H
DIN     BIT     093H
DOUT    BIT     092H

; *** SCON0 ***
SM0_0   BIT     09FH
SM0     BIT     09FH
SM1     BIT     09EH
SM1_0   BIT     09EH
SM2     BIT     09DH
SM2_0   BIT     09DH
REN     BIT     09CH
REN_0   BIT     09CH
TB8     BIT     09BH
TB8_0   BIT     09BH
RB8     BIT     09AH
RB8_0   BIT     09AH
TI      BIT     099H
TI_0    BIT     099H
RI      BIT     098H
RI_0    BIT     098H

; *** IE ***
EA      BIT     0AFH
ES      BIT     0ACH
ES0     BIT     0ACH
ET1     BIT     0ABH
EX1     BIT     0AAH
ET0     BIT     0A9H
EX0     BIT     0A8H

; *** P3 ***
CLKS    BIT     0B6H
T1      BIT     0B5H
T0      BIT     0B4H
INT1    BIT     0B3H
INT0    BIT     0B2H
TXD     BIT     0B1H
TXD0    BIT     0B1H
RXD     BIT     0B0H
RXD0    BIT     0B0H
```

```
; *** IP ***
PS      BIT     0BCH
PS0     BIT     0BCH
PT1     BIT     0BBH
PX1     BIT     0BAH
PT0     BIT     0B9H
PX0     BIT     0B8H

; *** PSW ***
CY      BIT     0D7H
AC      BIT     0D6H
F0      BIT     0D5H
RS1     BIT     0D4H
RS0     BIT     0D3H
OV      BIT     0D2H
F1      BIT     0D1H
P       BIT     0D0H

; *** EICON ***
;SMOD1   BIT     0DFH
EAI     BIT     0DDH
AI      BIT     0DCH
WDTI    BIT     0DBH

; *** EIE ***
EWDI    BIT     0ECH
EX5     BIT     0EBH
EX4     BIT     0EAH
EX3     BIT     0E9H
EX2     BIT     0E8H

; *** EIP ***
PWDI    BIT     0FCH
PX5     BIT     0FBH
PX4     BIT     0FAH
PX3     BIT     0F9H
PX2     BIT     0F8H

; *** Reg ***
Reg0 Data 00H
Reg1 Data 01H
Reg2 Data 02H
Reg3 Data 03H
Reg4 Data 04H
Reg5 Data 05H
Reg6 Data 06H
Reg7 Data 07H
RegB Data 0F0H
$RESTORE
```

(end reg1200.inc)

(start rom.a51)

```
PUBLIC_put_string                ; void put_string(char code *string);
PUBLIC_page_erase                ; char page_erase (int faddr, char fdata, char fdm)
PUBLICwrite_flash                ; char write_flash (int faddr, char fdata, char fdm)
PUBLIC_write_flash_chk ; char write_flash_chk (int faddr, char fdata, char fdm)
PUBLIC_write_flash_byte          ; char write_flash_byte (int faddr, char fdata, char fdm)
PUBLIC_faddr_data_read ; char faddr_data_read(char);
PUBLIC_data_x_c_read             ; char data_x_c_read(int addr);
PUBLIC_tx_byte         ; void tx_byte(char);
PUBLIC_tx_hex                    ; void tx_hex(char);
PUBLIC  putok                    ; void putok(void);
PUBLICrx_byte                    ; char rx_byte(void);
PUBLICrx_byte_echo               ; char rx_byte_echo(void);
PUBLICrx_hex_echo                ; char rx_hex_echo(void);
PUBLICrx_hex_double_echo         ; char rx_double_echo(void);
PUBLICrx_hex_word_echo           ; char rx_word_echo(void);
PUBLIC  autobaud                 ; void autobaud(void);
PUBLICputspace4                  ; void putspace4(void)
PUBLICputspace3                  ; void putspace3(void)
PUBLICputspace2                  ; void putspace2(void)
PUBLICputspace1                  ; void putspace1(void)
PUBLIC  putcr                    ; void putcr(void);


;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;
; Interface to C compatible Boot ROM Build In Routines
;
        CSEG AT 0FFD5H
_put_string:                     ; void put_string(char *string);

        CSEG AT         0FFD7H
_page_erase:                     ; char page_erase (int faddr, char fdata, char fdm)

        CSEG AT 0FFD9H
write_flash:                     ; DPTR = address, acc = data (Not callable by C programs)

        CSEG AT 0FFDBH
_write_flash_chk:                ; char write_flash_chk (int faddr, char fdata, char fdm)

        CSEG AT 0FFDDH
_write_flash_byte:               ; char write_flash_byte (int faddr, char fdata, char fdm)

        CSEG AT 0FFDFH
_faddr_data_read:                ; char faddr_data_read(char);

        CSEG AT 0FFE1H
_data_x_c_read:                  ; char data_x_c_read(int addr);

        CSEG AT 0FFE3H
_tx_byte:                        ; void tx_byte(char);

        CSEG AT 0FFE5H
_tx_hex:               ; void tx_hex(char);

        CSEG AT 0FFE7H
```

```
putok:                           ; void putok(void);

        CSEG AT 0FFE9H
rx_byte:                 ; char rx_byte(void);

        CSEG AT 0FFEBH
rx_byte_echo:                    ; char rx_byte_echo(void);

        CSEG AT 0FFEDH
rx_hex_echo:                     ; char rx_hex_echo(void);

        CSEG AT 0FFEFH
rx_hex_double_echo:                      ; char rx_double_echo(void);

        CSEG AT 0FFF1H
rx_hex_word_echo:                        ; char rx_word_echo(void);

        CSEG AT 0FFF3H
autobaud:                        ; void autobaud(void);

        CSEG AT 0FFF5H
putspace4:                       ; void putspace4(void)

        CSEG AT 0FFF7H
putspace3:                       ; void putspace3(void)

        CSEG AT 0FFF9H
putspace2:                       ; void putspace2(void)

        CSEG AT 0FFFBH
putspace1:                       ; void putspace1(void)

        CSEG AT 0FFFDH
putcr:                           ; void putcr(void);

        END
```

(end rom.a51)

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

LEARNING LEGGED LOCOMOTION OVER EXTREME TERRAIN

NSF Summer Undergraduate Fellowship in Sensor Technologies
David Cohen (Mechanical Engineering) – The University of Pennsylvania
Advisor: Dr. Daniel Lee

ABSTRACT

Currently, legged locomotion over obstacles remains a great challenge in robotics. The equations describing so-called “extreme” walks are highly complex, and the contact forces between the robot and its environment are not easily modeled. A major research goal in the field of robotics is to develop capacity for “extreme” walks while keeping computations tractable.

This study was designed to create a method for making a Sony Aibo walk up a 1-inch step consistently. Project activities focused on finding ways to apply simple models for generating walks onto the step. Once the study identified a class of “extreme” walks in lower dimensions, parameters were hand-tuned in an effort to assess whether a prospective class of walks was appropriate or not.

As a result of the work completed in this study, both a 35 mm and a 50 mm step were scaled successfully by the Sony Aibo. Avenues for future work include: 1) the development of an automatic system for determining footfall order and step

displacements, 2) a method for automatic primitive identification and switching, 3) refinements to the fields employed, and 4) extensions to extreme dynamic locomotion.

Table of Contents

1. INTRODUCTION

A major research goal in the field of robotics is to develop capacity for “extreme” walks while keeping computations tractable. In order to create a class of agile legged robots capable of assisting troops in the field, DARPA (Defense Advanced Research Projects Agency) has issued solicitation BAA 05-25, which has as its primary objective the development of learning techniques that will permit a quadruped to walk across an obstacle-strewn terrain.

This study was designed to create a method for making a Sony Aibo walk up a 1-inch step consistently, in preparation for the associated DARPA project. The Sony Aibo was selected because, although the DARPA-supplied robot is not currently available, the Aibo has a similar size and structure. Project activities focused on finding ways to apply simple models for generating walks onto the step. The rationale behind developing the simple models was that the models would eventually be machine learned. To address the present objective, hand-tuning alone was sufficient.

As a result of the work completed in this study, both a 35 mm and a 50 mm step were scaled successfully by the Sony Aibo. The method used will be explained fully in the following sections. Areas requiring future work -- such as footfall order determination and primitive switching -- will also be described.

2. OVERVIEW OF STEP METHOD

The method used for scaling the obstacles mixed a potential field formulation for torso placement with pre-set geometries for footpaths during steps. The torso was subjected to rigid-body translation and rotation from fields that were determined by leg extension, leg orientation, and body balance. At each frame, the step function would attempt to find a local minimum in the potential field by iteratively “riding” the field, and the torso would move toward this minimum as quickly as possible. The feet were pre-sequenced to place themselves at certain end positions in turn and, as such, were not affected by the fields.

Interspersed between the footfalls were periods in which the robot would shift its center of balance over the three feet that were to be stationary over the next cycle. This insured that the robot did not immediately fall over when the next foot was raised.

A set of primitives (namely, front-up move-forward and rear-up) were found that allowed the robot to more effectively negotiate large obstacles. By separating the larger task of obstacle-scaling into these primitives, field parameters could be tuned more specifically and different footfall sequences could be determined for different situations.

Part 3 will explain the types of potential fields exerted on the torso. Part 4 will explain the rigid-body dynamics used to evaluate the effect of these fields. Part 5 will briefly explain the implementation of trapezoidal step paths. Part 6 will discuss the various primitives’ parameters and the footfall sequences for these primitives.

3. EXPLANATION OF POTENTIAL FIELDS

Early on, it was recognized that it would be necessary to have the torso displace and pitch in order to keep the robot's feet stable and within their configuration spaces. The original study proposal [1] called for the use of explicit primitives, like "pitching" and "climbing", to accomplish these tasks. This, however, would have required complex tuning and switching. The use of potential fields allowed vast simplification.

There were three main types of fields applied to the torso: radial leg fields, angular leg fields, and a balance field. The leg fields were applied to the hips of each leg, and resulted in both translation and rotation for the torso. The balance field was applied at the center of mass of the robot, and thus allowed for only translation. As will be explained in section 4, these fields determined instantaneous momentum, not force.

The radial leg and angular leg potential fields were modeled to keep the feet within their configuration spaces. The balance field was added to keep the dog from falling over. These fields are explained in greater detail below.

3.1 RADIAL LEG FIELDS

The radial leg fields, depicted in Figure 1 as springs, were nicknamed "shock-absorber fields" since their effect was similar to attaching shock absorbers between the hip and foot of each leg. Stretching a leg near the limit of its configuration space tugged the torso after it; likewise, bringing a leg close in to the torso tended to push the torso away. The purpose of these fields was to avoid the imaginary angles and odometry problems associated with attempts to position the leg beyond the singularities.

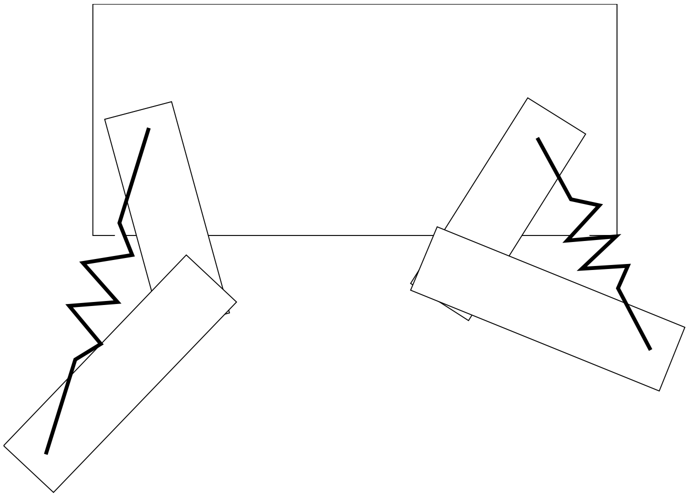


Figure 1 – Radial Leg Fields

The characteristic force-displacement response, though, was quite dissimilar to the classical linear spring. The study showed that it was advantageous to create a large "dead zone" within which the robot was unaffected by the shock-absorber field. Therefore, the field was modified to be the product of a line with the sum of two exponentials:

$$\text{Force} = A \cdot (x \cdot \text{natural\_length}) \cdot (\exp(B \cdot (x - \text{max\_length})) + \exp(C \cdot (\text{min\_length} - x)))$$

In the function above, 'x' is the extension between the hip and foot. Note that, by modulating the constants B and C, the operator can make the force-displacement profile asymmetric – an advantage not afforded by other functions with "dead zones" (such as hyperbolic sine or high-degree polynomials). An example of the force-displacement profile of this function, with real parameters taken from the code, is given in Figure 1. Note that, since the constants B and C are equal, the asymmetry is not fully employed.

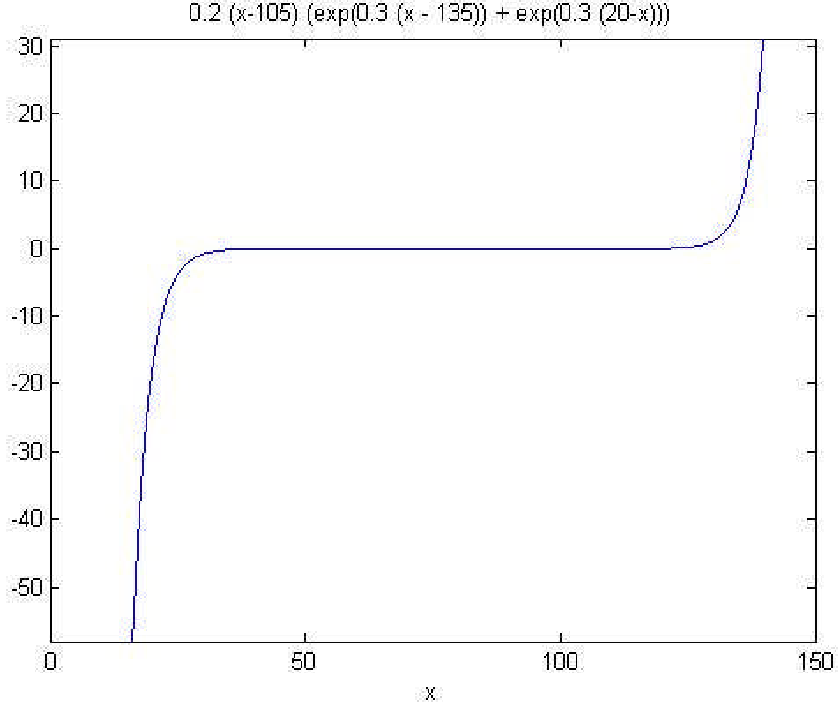


Figure 2 – Force-Displacement profile of Radial Leg Field

3.2 ANGULAR LEG FIELDS

The radial leg fields served to keep the legs within their extension limits. However, using radial fields alone led to situations in which the robot would try to position its legs outside their angular limits. Therefore, it was necessary to add in angular leg fields, so that extreme angles could be avoided.

The angular fields were decomposed into two fields for each leg: a "flap" field, and a "swing" field. Figure 3 shows the axis convention employed for the robot, and Figures 4 and 5 show the "swing" and "flap" angles, respectively. The axis convention is consistent across the three figures.

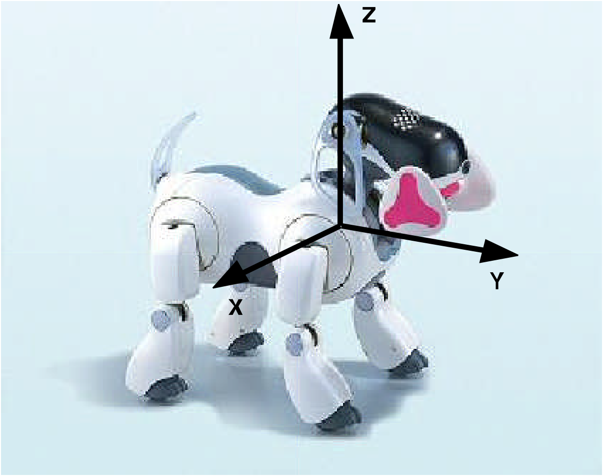


Figure 3 – Sony Aibo Axis Convention

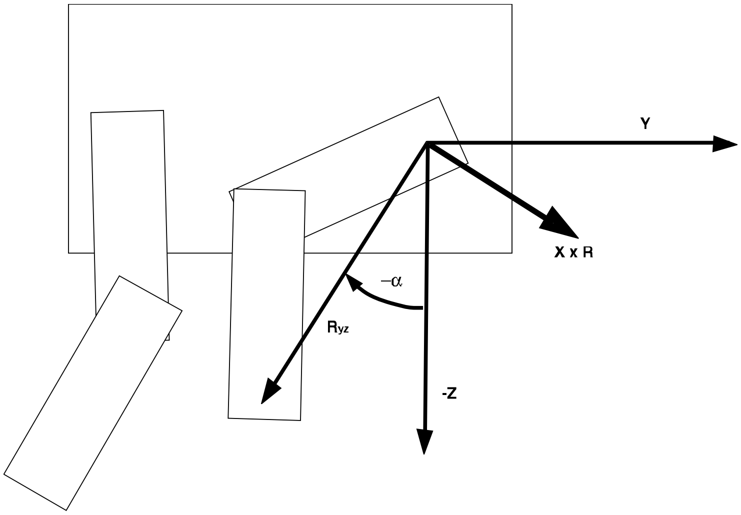


Figure 4 – Y-Z Plane of Robot with "Swing" Angle Indicated (Front Right Leg)

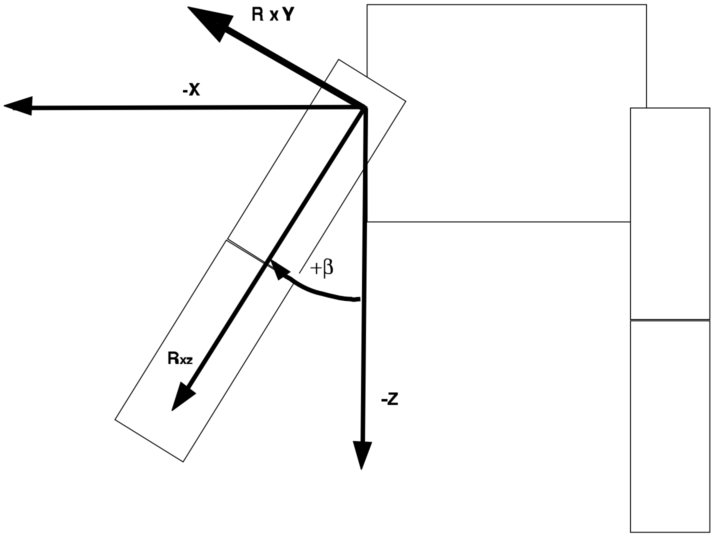


Figure 5 – X-Z Plane of Robot with "Flap" Angle Indicated (Front Right Leg)

The "swing" angle, marked α on Figure 4, is the angle between the projection of **R** (the vector from the hip to the foot) on the Y-Z plane and the **–Z** vector. The sign convention used made the angle depicted negative, so it is marked as such. Likewise, the "flap" angle, marked β on Figure 5, is the angle between the projection of **R** on the X-Z plane and the **–Z** vector (the sign convention makes the angle depicted positive).

For each field a force function similar to those used by the radial fields was implemented, since it was found that a "dead zone" was advantageous with the angular fields as well. Thus, both "flap" and "swing" had a natural angle, a minimum angle, and a maximum angle associated with it.

The direction along which the force was applied to the torso, however, was different between the two fields and distinct, obviously, from the radial fields as well. The direction for the force application due to the "swing" field was along the cross product of the **X** vector and **R**, and the direction for the force application due to the "flap" field was along the cross product of **R** and the **Y** vector. Both directions are denoted on their respective figures with the heavier arrows. The rationale behind using the cross products for the directions of application was that they would provide the most efficient angular change without radial change, given that the foot position remained constant.

3.3 BALANCE FIELD

The final potential field force implemented on the torso was a balance field. In order to keep the robot statically stable (the gaits generated were all static, not dynamic, gaits) it was necessary both to keep at least three legs on the ground at a time and to ensure that the robot's center of mass was within the polygon determined by the feet on the ground. Therefore, a field was implemented that drew the torso towards the centroid of the polygon determined by the planted feet.

Figure 6 shows a schematic of the balance field for the case where the front right foot is off the ground. The planted feet are shown and labeled (no legs are shown, since they are unnecessary for the calculation of the force or direction). The force runs from the centroid of the torso to the centroid of what is, in this case, a triangle whose vertices are the positions of the planted front left foot, rear left foot, and rear right foot.

The magnitude of the force is calculated using a hyperbolic tangent:

$$\text{Force} = A \cdot \tanh(B \cdot x)$$

Variable 'x' is the magnitude of the vector from the torso centroid to the foot centroid.

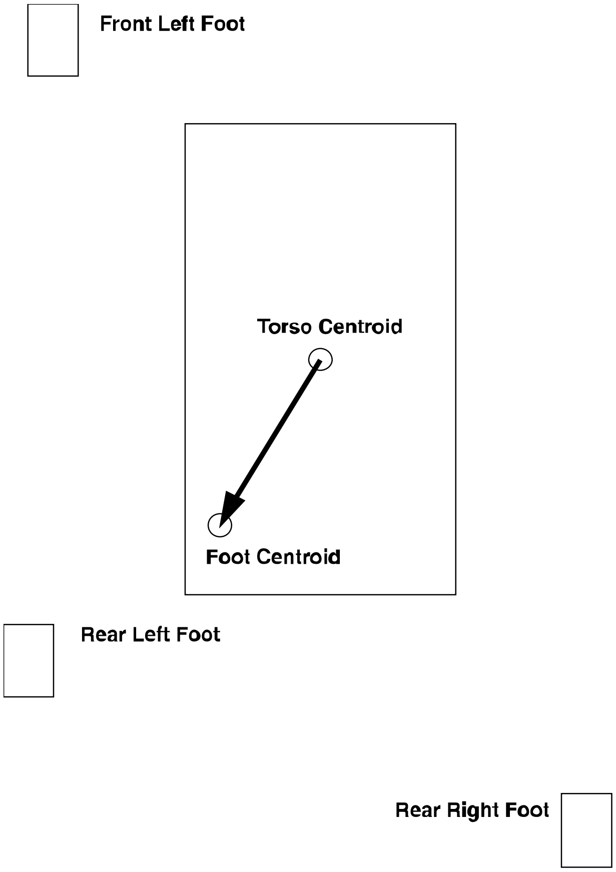


Figure 6 – The Balance Field (Bird's Eye View)

4. EXPLANATION OF THE RIGID-BODY REACTIONS

The torso acted as a 4-DOF rigid body, limited to x-y-z translation and pitching. The "forces" at the hips of the four legs and at the center of mass of the torso were summed by standard techniques to provide a net resultant force and a net moment (it was assumed that the center of geometry of the torso was also the center of mass). The way the force and the moment were applied, however, were different from the standard convention:

$$\text{Force} = \text{mass} \cdot \text{velocity}$$

$$\text{Moment} = \text{moment of inertia} \cdot \text{angular velocity}$$

Thus, the fields implemented are probably most appropriately called "momentum fields". In the interest of simplicity, these "momentum fields" were implemented instead of "force fields". Since the function was aiming for the local minimum, it didn't make a difference whether velocity was integrated from accelerations or not.

The mass and moment of inertia were defined as parameters in the step function. Since the system was 4-DOF, the moments of inertia corresponding to yaw and roll were effectively set to infinity. There were also caps on the maximum angular and translational velocities that could be achieved.

5. EXPLANATION OF THE LEG PATHS

The footpaths used were inspired by the trapezoidal steps simultaneously developed by Newcastle University and the University of New South Wales for the Robocup 2003 competition [2, 3]. The rationale behind using a trapezoidal step in robot soccer was the speed advantage. Disengaging the claw of the Sony Aibo from the field material yielded a speed increase of about 10%. While, for the case at hand, speed was not important, it was crucial that the robot's feet not be hindered by the material. Consequently the trapezoidal method was employed.

A schematic of the trapezoidal step is provided in Figure 7. There was no bottom stroke to the step. Traversal was accomplished by torso repositioning across different foot positions. The parameters ϕ, γ, and 'minimum clearance' determine the shape of the step. Additional parameters determine the speed with which the step is executed in the Rise, Traverse, and Lower Stages.

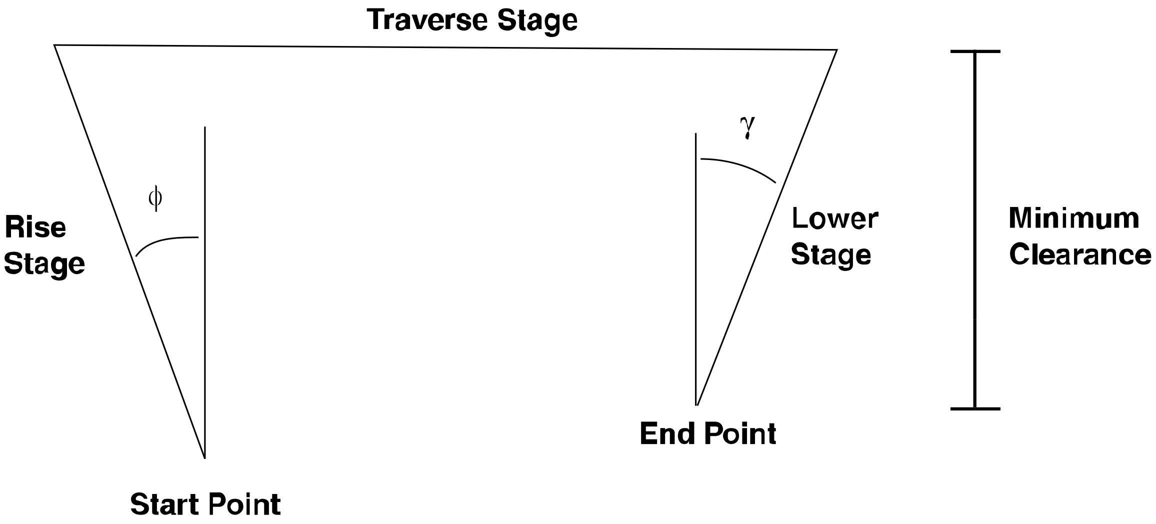


Figure 7 – Trapezoidal Step Path (Side View)

In the case that there is an x-displacement in the step, the path must skew in the X-Y plane. In this case, both the Rise and Lower Stages have no component in the x-direction. This is to ensure that there is still a clean disengagement from the field material. Figure 8 gives a top down view of what a skewed step might look like.

6. PRIMITIVE PARAMETERS AND FOOTFALL SEQUENCES

Once all the fields were in place, the final challenges were to find a suitable sequence of footfalls to climb the step and to tune the field parameters.

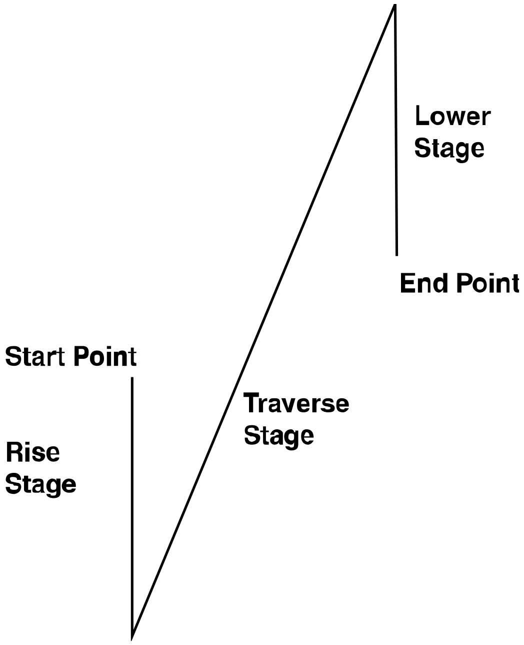


Figure 8 – Skewed Trapezoidal Step Path (Bird's Eye View)

The footfall sequence was a list of 4 element vectors, which denoted the x, y, and z displacements of the step as well as an index of the leg that was stepping. The step function assumed that the foot was just resting on a surface at the end of each footfall. The function was completely open-loop and this "foot resting" condition was necessary so that the robot's orientation could be accurately calculated. Unfortunately, it was the general case that for the scaling footfalls some experimentation was required to find the exact z-displacement that would allow the foot to just set down. This was because the Sony Aibo has large plastic casings around its forepaws, making it very difficult to get exact measurements for the point of contact.

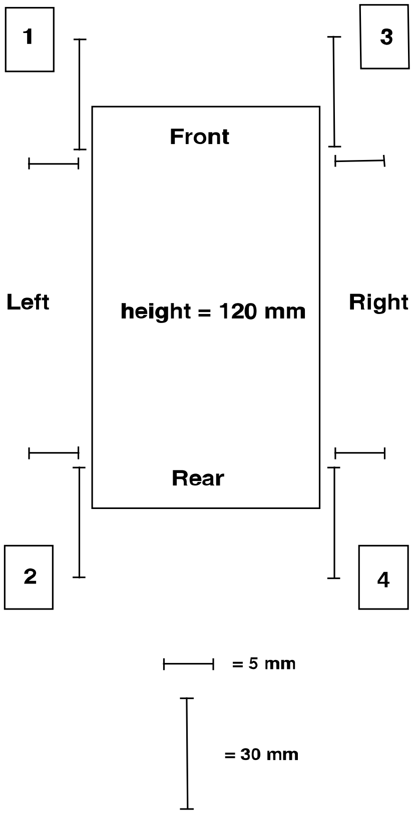
The method for parameter tuning, likewise, was trial and error: any set of parameters that led the robot to try to effect positions outside its configuration space was deemed unsuitable, as was any set that resulted in the robot's losing its balance. Tuning the parameters thus proved exceptionally difficult, as the robot seemed to inevitably violate at least one of the two criteria. This was remedied, however, by the development of a shifting foot-cycle: while traditional non-extreme static gaits use 4 step foot-cycles, it was found that an irregular cycle led to a large qualitative improvement in the robot's performance. Once the irregular cycle was implemented, it was not long before the parameters were tuned appropriately, and the robot was able to successfully scale a 35 mm step.

The next step after scaling a 35 mm, or 0.25 L (L = leg length) step was to scale a 50 mm (0.35 L) step. Unfortunately, the advances that resulted in the scaling of the smaller step were not quite sufficient to permit scaling the larger one. In general, parameters that allowed good performance for one portion of the step led to significant failure in other portions.

The solution was to simply use different parameters for different parts of the step. The three primitives that resulted (front-up, move-forward, and rear-up) each had its own footfall sequence and parameters – thus the irregular cycle used to scale the 35 mm step became a concatenation of three distinct cycles.

Once the primitives were set in place, it was a simple matter of re-tuning the parameters and footfalls to scale the new step. The footfall sequence remained unchanged from the 35 mm step to the 50 mm step. Indeed, the ease with which the robot was retuned suggests the robustness of the field method, although it should be noted that the motion returned from the front-up primitive had to be smoothed and sped up. This was, however, a problem for finding the minimum of the force field and not a problem inherent in the fields themselves. Figures 9 and 10 give information on the primitive parameters and footfall sequences employed for the 50 mm step.

	Front-Up	Move-Forward	Rear-Up
k_radial	1	1	1
coeff_radial	0.3	0.3	0.3
m	1	1	1
l	50	50	50
max_v_ride	1	2	2
max_v_trap	1	1.5	1.5
max_o	0.1	0.01	0.01
b	0.5	0.5	0.5
b_coeff	0.05	0.05	0.05
phi	0.15	0.15	0.15
gamma	-0.15	-0.15	-0.15
front_h	25	20	10
rear_h	25	25	25
rise_speed	2	4	4
traverse_speed	2	4	4
lower_speed	2	4	4
natural_flap	0.2	0.2	0.2
max_flap	1	1	1
min_flap	-0.05	-0.05	-0.05
natural_swing	0	0	0
max_swing	1.4	1.4	1.4
min_swing	-1.4	-1.4	-1.4
k_flap	1	1	1
k_swing	5	5	5
coeff_flap	5	5	5
coeff_swing	5	5	5



	1	2	3	4
natural_length	105	110	105	110
max_length	135	140	135	140
min_length	60	65	60	65

Figure 9 – 50 mm step Parameter Summary

The picture in the upper right of Figure 9 identifies the four legs by the numerical convention employed in the footfall sequence. The lower table in Figure 9 uses this convention. Figure 9 also identifies the initial stance of the robot.

The natural_length, max_length, and min_length given in the lower table refer to the associated radial field values (all in mm). The radial parameter k_radial serves as A and coeff_radial serves as both B and C in the following expression (given earlier in section 3.1):

$$\text{Force} = A \cdot (x - \text{natural\_length}) \cdot (\exp(B \cdot (x - \text{max\_length})) + \exp(C \cdot (\text{min\_length} - x)))$$

Likewise, k_flap and k_swing are both A in their respective functions, and coeff_flap and coeff_swing are each both B and C. The angles natural_flap, max_flap, etc. are all given in radians.

The parameter b serves as A and b_coeff serves as B in the following expression (given earlier in section 3.3):

$$\text{Force} = A \cdot \tanh(B \cdot x)$$

The parameter max_v_ride refers, in mm/frame, to the maximum speed the torso is allowed to move during while "riding" the potential field (i.e. with all four feet on the ground); likewise, max_v_trap is the maximum speed the torso is allowed to move while stepping (three feet on the ground). Similarly, max_o is the maximum angular speed the robot is allowed to pitch (in rad/frame).

Phi and gamma are as noted in Figure 7. Rise_speed, traverse_speed and lower_speed are the maximum mm/frame a stepping foot is allowed to move in each of the three stages (noted in Figures 7 and 8).

	Front-Up			
x displacement	0	0	0	0
y displacement	65	70	70	60
z displacement	0	50	50	0
leg index	2	1	3	4

	Move-Forward			
x displacement	0	0	0	0
y displacement	70	70	70	60
z displacement	0	0	0	0
leg index	1	3	2	4

	Rear-Up			
x displacement	0	0	0	0
y displacement	85	75	85	60
z displacement	0	30	-10	30
leg index	1	2	3	4

Figure 10 – 50 mm step Footfall Summary (displacements in mm)

Certainly, a quick examination of Figure 10 reveals that the three primitives are quite distinct, not only in footfall order but in the characteristics of their steps. Note that the Rear-Up z-displacements for the rear left and rear right legs are much less than 50 mm. This factor is attributable to the previously mentioned forepaw casings. Also, across all three primitives the y displacements of leg 4 (rear right) seem to lag behind the others. It is possible that this asymmetry makes it necessary to bring foot 3 down 10 mm in the Rear-Up primitive.

7. DISCUSSION AND CONCLUSIONS

Extreme Legged Locomotion is an important and very difficult problem. The huge number of dimensions faced by a researcher attempting to solve the problem makes most solution techniques worthless. In the interest of limiting the problem so that part of it could be solved using traditional (gradient based) learning techniques, a field/pre-set geometry system was set up to create an appropriate class of extreme steps. The system was tested on two obstacles (a 35 mm step and a 50 mm step), and was successful at negotiating both.

Problems remain, however, with determining the proper footfall sequence as well as foot displacements during these sequences. Currently, these sequences and displacements are tailored to specific primitives and are associated with specific obstacles. There is no algorithm in place to find a footfall sequence or set of displacements for an arbitrary obstacle field. In addition, the primitive switching employed here will have to be codified if it is to be learned automatically. Also, refinements to the field system used, including allowances for balance improvements, anticipatory fields, 6-DOF motion, and computational optimization might be necessary. Finally, the matter of extreme dynamic stepping must be addressed. Future studies will be required to solve these crucial problems (see section 8).

8. RECOMMENDATIONS

The problems listed in the closing of section 7 will now be expanded upon, and possible solutions will be suggested.

8.1 FOOTFALL ORDER AND STEP DISPLACEMENTS

The largest unaddressed problem in the extreme step approach outlined here is that the footfall order and step displacements were pre-set by the operator. Any system that endeavors to provide locomotion over an arbitrary obstacle field cannot use this technique.

One possible way to solve this problem is to use the fields themselves to plan appropriate sequences of extension and recovery. By integrating the fields to create a scalar potential field (or by creating a new, simplified scalar field) important information can be gleaned about the “comfort” of certain poses. It may be possible to use a state machine to decide when to extend into “uncomfortable” positions and when to relax into “comfortable” ones (a naïve approach of only executing “comfortable” positions would greatly hinder the efficacy of the walk). Future footfalls could be planned in advance, and appropriate “set-up” steps could be sequenced to allow low potential and good balance.

This footfall code was not put in place due to its sheer computational complexity. Given that the step sequences themselves took upwards of 2 minutes to calculate, cycling through hundreds of potential future moves was not feasible. In addition, since all steps used in the study were handmade, the limited amount of terrain available for testing was likely to make any experimental system unreasonably condition-dependent. Finally, and perhaps most importantly, there was not enough good feedback to test out different footfall patterns. The DARPA set-up will probably solve the last two problems. Significant re-tweaking might solve the first (see section 8.3).

8.2 PRIMITIVE SWITCHING

The next significant problem with the previously outlined extreme step approach is the arbitrary creation of primitives, with their own field parameters and footfall sequences. The footfall sequence problem might be solved by the method described in section 8.1, but the field parameter switching problem remains.

Enumerating primitives is a possible solution. Such primitives could be identified by the pitch angle and roll angle (see section 8.3) as well as the planned trajectory of the stepping foot. A large, but certainly non-infinite, number of primitives would result, and each could be tuned in turn. One difficulty with this method would be learning the border conditions between primitives.

A promising alternative would be to make the field parameters a function of pitch, roll, and intended step trajectory. A simple linear function might suffice. In this way, the problem of primitive creation and switching might be sidestepped. Also, the parameters that define the functions could themselves be learned with traditional gradient-based techniques. The number of dimensions to be learned, however, would at least double, and most likely triple, under this system.

8.3 FIELD REFINEMENTS

The first clear area for field refinement is the balance field. The radial and angular foot fields enjoyed a “dead zone”, which tended to enhance their performance. The balance field, on the other hand, had no such property. Indeed, the field, with its sole dependence on the distance between the torso centroid and planted-foot-triangle centroid, neglects what often are extreme aspect ratios in the triangle. Given additional time for implementation, the simple fix shown in Figure 11 might be applied. The lines are force directions, leading directly in from each face of the triangle. Within the triangle, there is a much smaller pull towards the center.

The inclusion of anticipatory fields is a second proposed improvement for future studies. Currently, all anticipatory action is hard-coded in: in between footfalls, the robot shifts its weight to the feet that will be planted during the next step. A set of anticipatory fields could accomplish this task automatically, making the entire process more coherent.

The third area for improvement is more obvious: the 4-DOF system should be extended to 6-DOF. This was not implemented in the current project, since it was decided that the symmetric nature of the step made the ability to roll and yaw unnecessary. For arbitrary obstacle fields, however, roll and yaw will require substantial refinement. To this end, a quaternion-based representation of the torso's orientation could be implemented, as described by Mirtich [4].

Finally, the fourth area for improvement is the method for determining field equilibrium. In the present state, the field is evaluated up to 100 times per frame to determine in which direction the torso must move. With proper tuning, it might be possible to eliminate this inefficiency all together, or at least limit the cycles to a more reasonable number. This would allow the robot to execute its moves in real time, reacting to outside information. Computational overhead limited the current approach to being open-loop.

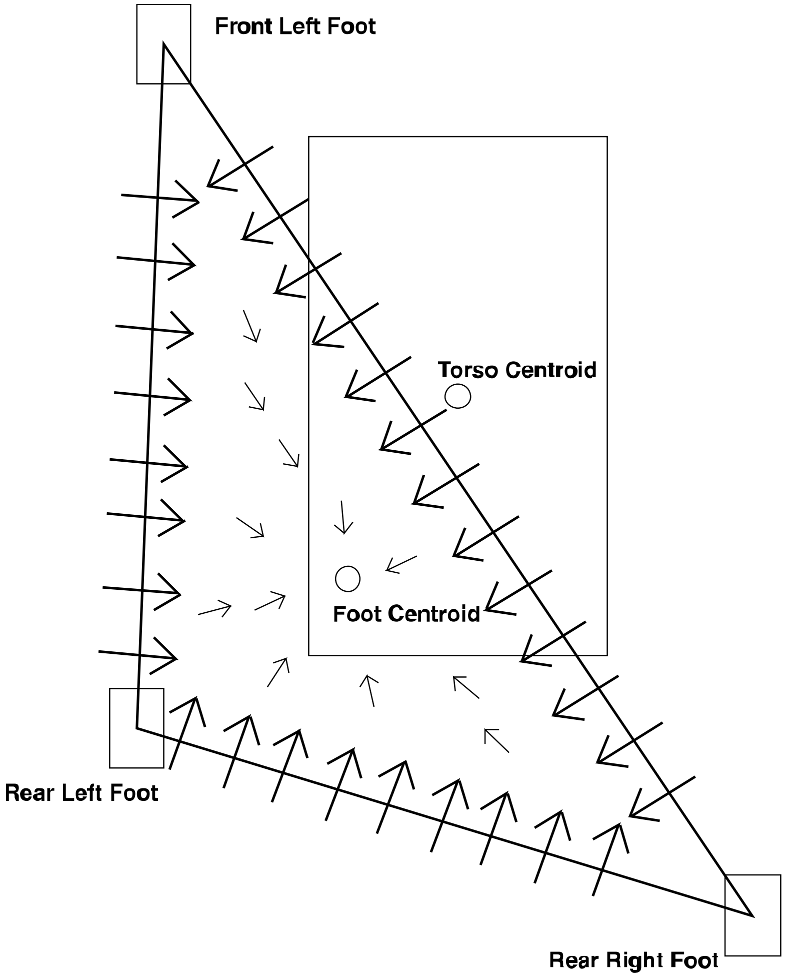


Figure 11 – Refined Balance Field

8.4 EXTREME DYNAMIC LOCOMOTION

The final problem to be addressed is most certainly the most interesting: the difficulty of developing dynamic steps for extreme conditions. Dynamic steps would not only provide speedier traversal of obstacles, they are also likely to offer improved performance, allowing the robot to cross obstacles which static steps alone cannot surmount.

The same field principles used in the execution of static steps could be used. But they would require significant and complex modifications. In particular, the balance fields and heretofore non-existent anticipatory fields would have to be extremely well tuned to deal with the unpredictable conditions encountered. The balance field would probably require some strategic imbalance as well. More likely, a completely different system will have to be developed to deal with so much additional complexity.

9. ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Daniel Lee, as well as Gilad Buchman and Paul Vernaza at the GRASP lab. I would also like to thank the National Science Foundation for the NSF-REU grant and the Microsoft Corporation for their financial support. It is a pleasure doing significant research as an undergraduate, and I would like to thank all of the people whose tireless efforts make the SUNFEST program possible.

10. REFERENCES

1. D. D. Lee, *Technical Proposal: Learning Low-Dimensional Controllers for High-Speed Quadruped Locomotion*, The University of Pennsylvania, Philadelphia, PA, 2005.

2. J. Bunting et al, *Return of the NUBots*, Newcastle Robotics Laboratory, The University of Newcastle, Australia, Oct. 2003.

3. J. Chen et al, *Rise of the Aibos III – AIBO Revolutions*, The University of New South Wales, Australia, Nov. 2003.

4. B. V. Mirtich, *Impulse-based Dynamic Simulation of Rigid Body Systems*, Dissertation for Ph.D. in Computer Science, The University of California at Berkeley, California, 1996.

APPENDIX: INVERSE KINEMATICS

Inverse kinematics refer to the calculations required to convert an (x, y, z) triple to a set of angles (θ_1, θ_2, θ_3). In the case of the Sony Aibo (and for the DARPA robot, since they share the same limb geometry) there are typically multiple solutions, except at the singularity. Beyond the singularity, any angles returned by the inverse kinematics function will have imaginary components. Presented below is a MATLAB function that takes **XYZ**, an (x, y, z) triple relative to the right front shoulder, and returns **angles**, the angle vector (shoulder swing, shoulder flap, and knee swing) required by the front right leg to effect the position.

The variable ‘lone’ is the length of the thigh on the robot, and ‘ltwo’ is the length of the shin. The variable ‘upperoffset’ is the y-offset from the hip to the knee when the leg is pointing straight down, and ‘loweroffset’ is the y-offset from the knee to the foot pad under the same conditions. All lengths are in millimeters.

```
function angles = getanglesfrontright(XYZ)

lone = 69.5;
upperoffset = 9;
ltwo = 71.5;
loweroffset = -9;

P = 2*(upperoffset*loweroffset + lone*ltwo);
Q = 2*(ltwo*upperoffset - lone*loweroffset);
R = loweroffset^2 + ltwo^2 + upperoffset^2 + lone^2 - (sum(XYZ.^2));

angles(3) = atan2(Q/sqrt(P^2 + Q^2),P/sqrt(P^2 + Q^2)) - acos(-
R/sqrt(P^2 + Q^2));

if (angles(3) < 0)      %So that we don't snap the dog's knees
    angles(3) = atan2(Q/sqrt(P^2 + Q^2),P/sqrt(P^2 + Q^2)) + acos(-
R/sqrt(P^2 + Q^2));
end

P = loweroffset*cos(angles(3)) + ltwo*sin(angles(3)) + upperoffset;
Q = lone + ltwo*cos(angles(3)) - loweroffset*sin(angles(3));

angles(2) = asin(XYZ(1)/Q);

Q = cos(angles(2))*Q;

components = inv([Q P;P -Q])*[XYZ(2);XYZ(3)];

angles(1) = atan2(components(1), components(2));
```

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

WORKING TOWARD A BETTER VISION-BASED OBSTACLE DETECTION METHOD

NSF Summer Undergraduate Fellowship in Sensor Technologies
Roman Geykhman (Dept. of Electrical and Systems Engineering) - University of Pennsylvania
Advisor: Dr. Dan Lee

ABSTRACT

Obstacle detection is a vital component of any autonomous mobile robotics application. Vision-based systems for obstacle detection offer the advantage of using relatively inexpensive and readily available video cameras to supply a mobile robot with information about its environment. The key challenge, however, is that cameras supply too much information and complicate the efficient, automated extraction of meaningful features from raw images. Fast and effective obstacle acquisition from this input is still an unsolved problem in robotics.

This paper documents the experimental development of an accurate and efficient vision-based obstacle detection method using the Learning Applied to Ground Robotics (LAGR) experimental platform. The LAGR platform is designed to test algorithms for

the successful autonomous navigation of an unmanned vehicle through rural terrain, relying almost entirely on vision to collect information about the environment. The platform is equipped with two pairs of stereo cameras, each with a dedicated Pentium-M computer committed to processing its input and converting it into meaningful information about obstacles present in the platform's environment.

This project focuses on the use of low-level filtering and curve-fitting techniques to enable mobile robots to extract enough information about surrounding obstacles to successfully avoid them. The study is focused on developing and testing new and preexisting algorithms that will also enable the robot to avoid false detection without expending too much computation time in the process. The algorithms will be implemented in C and MATLAB code and tested on board the LAGR platform using both real-time and prerecorded stereo image pairs.

Table of Contents

1. INTRODUCTION

Effective obstacle detection is a vital component of any mobile robotics application. Several methods exist for obtaining information about a robot's environment, using various sensors such as RADAR, SONAR, and laser range-finders. These types of sensors are generally quite accurate, and can be calibrated to provide very precise range and direction data about obstacles in the environment. However, despite any possible accuracy and range benefits these methods offer, they suffer from the fact that they each require specialized and expensive equipment to implement.

Vision-based obstacle detection methods offer the significant advantage of using relatively inexpensive, off-the-shelf video cameras as the chief method for obtaining information about the environment. It also offers the guarantee that each frame of video carries enough information about the environment to sufficiently detect all visible obstacles. Whereas typical SONAR receivers and laser range finders are usually configured to receive data from only one direction or only one plane, cameras are able to see and record a significant portion of the mobile robot's environment and get the “whole picture.” The problem with vision-based obstacle detection methods is that, while they provide sufficient information about obstacles, they in fact provide too much information, making the effective recognition of features a relatively involved process.

Numerous existing techniques can be used to extract meaningful features from images, and to classify those features as obstacles in the robot's environment. This paper will focus on the development of a relatively fast and accurate low-level system to detect obstacles from pairs of stereo images in order to facilitate the autonomous navigation of the Learning Applied to Ground Robotics (LAGR) experimental platform. These techniques and methods will focus mainly on modeling the low-level geometry of the robot's environment, using both commercial and custom-coded vision software.

2. PRIOR WORK IN ROBOT VISION

2.1 Stereo Vision Basics

One of the simpler obstacle detection methods involves the use of fixed pairs of cameras in order to triangulate the locations of obstacles. As Figure 1 illustrates, two images are captured using neighboring cameras pointed at the same object. Each camera has a set of pixel coordinates for each point on this object. The triangle outlined in red lies on the epipolar plane. By definition, this plane passes through the point in space under observation and the optical centers of both cameras. It is evident that corresponding pixels in the two cameras must lie on the same epipolar plane.

It is a fact that this plane does not always intersect horizontal scanlines. However, if the relative orientation and displacement of the two cameras' optical centers is known, a coordinate transformation can be performed on both images in order to map epipolar lines in the raw image to horizontal scanlines in the transformed image. This process is known as rectification and reduces the search for matching image features to a one-dimensional problem, as matching image features are, by construction, mapped to the same horizontal scanline in both images.

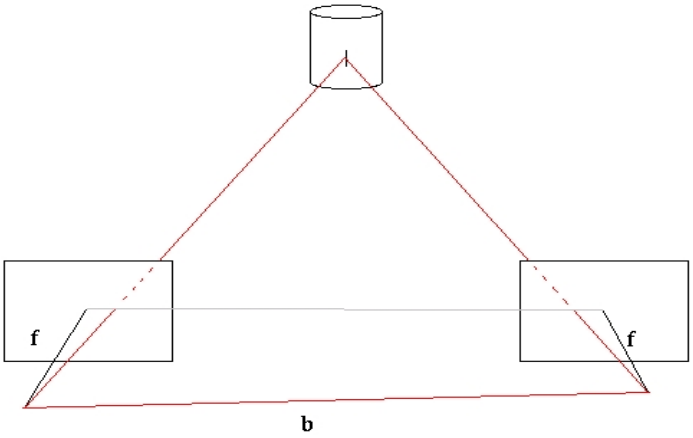


Figure 1: Geometry of Stereo Vision.
Epipolar plane is in red. Epipolar line is in grey.

The triangulation equations are illustrated in Figure 2, and the equations for the real-world coordinates of the object under observation are enumerated in Table 1. By inspection, the equations are derived from manipulations of similar triangles in Figure 2.

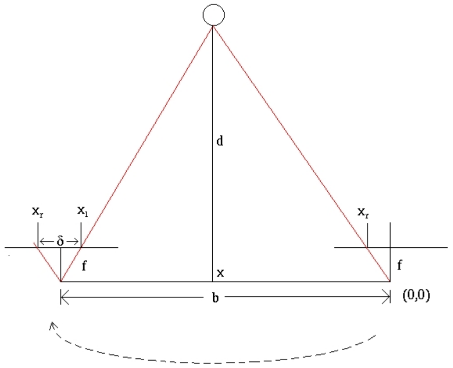


Figure 2: Triangulation Geometry in Stereo Vision

$$\frac{x}{d} = \frac{x_r}{f}$$

$$\frac{b + x}{d} = \frac{x_l}{f} = \frac{x_r + \delta}{f}$$

$$\ldots$$

$$d = \frac{fb}{\delta}$$

$$x = \frac{x_r d}{f}$$

$$y = \frac{y_r d}{f}$$

Table 1: Triangulation Equations in Stereo Vision

The equations in Table 1 reveal that in three dimensional space, the x-, y-, and z-coordinates of the image feature are all proportional to the inverse of disparity. Consequently, the uncertainties in these coordinates are proportional to the inverse square of the disparity. This error is nearly negligible for close objects, but for far objects, it can become as high as 1 meter. Given also that at large distances, the quantization of disparity values becomes noticeable, it is not reasonable to expect truly accurate data at low disparities. Subpixel interpolation methods exist to increase the precision of the feature matching algorithms at low disparity, and these algorithms may significantly improve the precision of range data obtained from the stereo image pair [1, pp. 35-38]. But there is an associated computation cost which, for the purpose of rough estimates of obstacle position (the required map resolution is only 0.5 m), is unnecessary.

2.2 The Correspondence Problem

Finding matches between features in stereo image pairs is known as the correspondence problem. The usual methods for establishing the necessary correspondence between image features involve comparing pixels from one image to possible corresponding pixels in the other image and determining a "goodness of match" criteria for the pairing.

A very popular and relatively fast method for establishing correspondence between features in grayscale images is the minimization of the sum of absolute differences of pixel values in patches of image in the left and right frames (as a function of displacement). In this method, a small patch in one image is compared pixel-by-pixel with identically sized patches in the other image which lie along the same scan line. The sum of each difference between corresponding pixels is then taken to be the difference between the two patches. Under ideal conditions, the minimum of this difference would occur where the two patches contain views of the same physical object in their respective images. Once this minimum is found, the coordinate difference between these matching patches is the disparity of that image feature. A pair of rectified images taken from a pair of stereo cameras is shown in Figure 3a. A disparity image, generated from this image pair with 7 x 7 pixel patches, is shown in Figure 3b. Blue indicates low disparity and yellow indicates higher disparity. Dark red indicates no data.

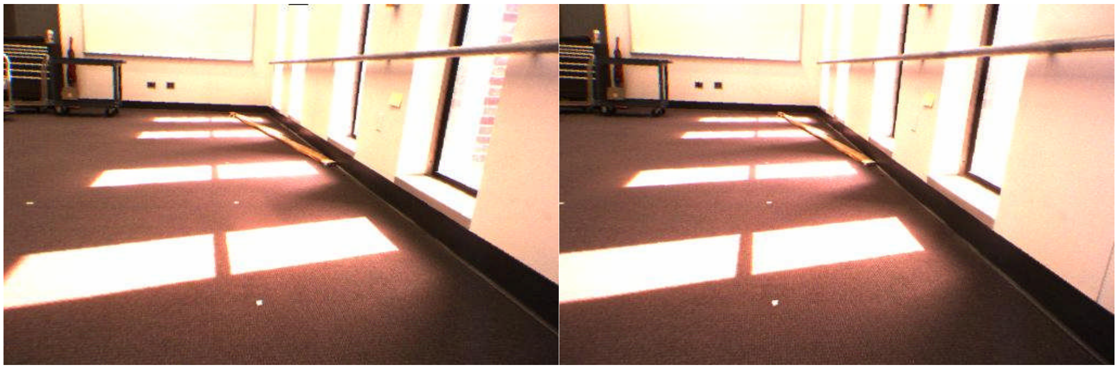


Figure 3a: Left and Right Rectified Images

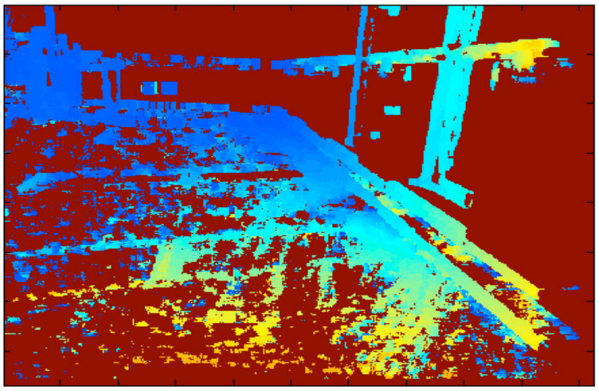


Figure 3b: Disparity Image

Correspondence matching, is, however, not an error-free process. As is evident from Figure 3a and Figure 3b, establishing an accurate correspondence between image features is not a trivial task. The disparity image in Figure 3b does not capture information about the wall to the right side, the far wall with the whiteboard, and parts of the floor. Indeed, it is seen that regions with low contrast are not captured in the disparity image calculation. Such regions offer insufficient features for making a definitive match between image patches by the sum of absolute differences algorithm operating on 7 x 7 pixel neighborhoods. This effect is illustrated in Figure 4.

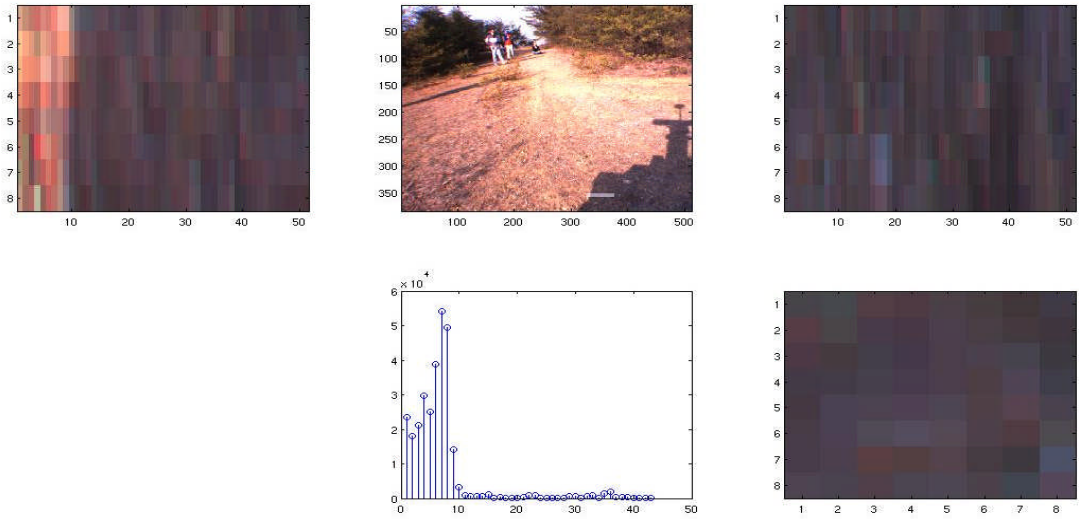


Figure 4: Sum of Absolute Differences Algorithm in Low Contrast Regions

The center top pane shows a rectified image with the region of comparison highlighted in the bottom right of the image. The top left and top right panes show a magnification of the region of interest in the left and right rectified images. The bottom right pane shows the 8 x 8 pixel magnification of the patch in the right image being compared with successive patches in the left image in the disparity calculation. The bottom center pane shows the sum of absolute differences as a function of displacement. Now, the region of interest lies on the ground directly in front of the robot, approximately 1 meter in front of the camera. With the geometry of the camera system, this would correspond to a disparity value of approximately 40 pixels. However, as can be seen in the bottom center pane, the low contrast of the region of interest causes very low sum of absolute difference values for displacements of anywhere from 10 to 45 pixels. Indeed, with this kind of difference data, no clear minimum is evident and a disparity value cannot be assigned without ambiguity. In fact, it is entirely plausible that random errors in the images determine the absolute minimum of this difference, and that the horizontal coordinate of this minimum value will have no relation to the actual disparity of the image feature being observed. The identification and elimination of these kinds of uncertain disparities is known as validation, and will be discussed in Section 4.

2.3 A Review of Obstacle Detection Methods

In order to identify obstacles in an image, it is necessary to perform some kind of segmentation. In one way or another, pixels (and their projections into 3D space) need to be grouped into obstacles and non obstacles. The search for effective segmentation processes is a topic of ongoing research in computer vision, and various methods and criteria are being studied to identify whole objects in images. Despite the precise detail and robustness that these methods promise, they are too computationally intensive for the current hardware to execute in a reasonable time frame.

Segmentation methods such as the normalized cuts approach [2] yield very good segmentation results on complex scenes. However, they take several minutes just to segment a single frame of video. While these techniques can be a springboard toward more complicated tasks such as higher level object extraction, recognition, and classification, for the purpose of simply *detecting* obstacles, they do too much work.

The precision of hi-level segmentation notwithstanding, low-level techniques can accomplish quite a bit in the area of simple detection. The majority of these methods work with disparity images generated from rectified pairs of stereo images. Successful low-level operations are possible on this kind of data, as it contains all the information about the 3D coordinates of every pixel in the camera image.

Many low-level obstacle detection techniques involve the brute-force projection of a three-dimensional point cloud into the robot's environment. Several possibilities exist for further processing the data. Some of the more successful techniques are outlined below.

One approach, as described in [3], is to take the point cloud and extract, by various statistical methods, an estimate for the ground plane. This estimate is then converted into a set expected disparity values for every point in the image. Significant deviations from this expected value are classified as obstacles, and the coordinates of each pixel comprising these obstacles are then easily calculated and projected into a map of the environment.

The current implementation of the obstacle detection algorithm on the LAGR platform uses a similar method to populate a cost map of the environment with counts of stereo points that lie above a similarly-obtained ground plane estimate. This cost map is then taken directly to the planning algorithm.

Noise due to false matches can be a big problem with both of these methods. The process of using raw stereo count data to populate a local obstacle or cost map is very susceptible to false detections brought about by spurious data.

The other, more critical problem arises from false correspondences with high disparity. This noise will generally be projected very close to the robot. Furthermore, given the projective nature of the camera, it will all be projected into the same region near the front of the camera. This will artificially inflate the stereo point count of the region directly in front of the robot and cause an artificial obstacle to appear directly in the robot's path. For purposes of navigation, this poses a significant problem. In practice, a certain region in front of the robot is always assumed to be obstacle-free and stereo data that places obstacles in it is ignored. This eliminates problems caused by false matches, but also makes it impossible to detect actual obstacles in that region.

D.R. Murray's work in stereo vision [1] brings up several key problems in the filtration and processing of disparity images. One issue is that the projective nature of the camera reduces the pixel count of objects as their distance from the camera increases, thereby reducing the number of data points for far-off objects. The other is that manipulating raw point data is subject to certain limitations. As has been mentioned, manipulations of this data are highly susceptible to noise, lose accuracy with distance, and effectively do too little work to detect obstacles.

Furthermore, low level noise reduction algorithms can only go so far. It is a self-evident fact that humans have a certain intuition about vision that allows them to make high-level descriptions of complex scenes without necessarily requiring access to the low level computations that occur in their visual cortex. One of these intuitive notions is the fact that most obstacles that may be encountered in the world are contiguous elements in three-dimensional space, and generally are smooth and thus exhibit only gradual changes in the surface normal vectors over their surfaces.

Murray's work has shown that surface orientation is indeed a good criterion to use in segmenting stereo images for use in higher-level computations and algorithms. The remainder of this paper will explore the use of surface orientation as a low-level spring-board to accurate and fast obstacle detection.

3. HARDWARE AND DESIGN REQUIREMENTS

The purpose of this project was to develop an improved vision-based obstacle detection algorithm for the Learning Applied to Ground Robotics (LAGR) platform, shown in Figure 5. This platform is equipped with two pairs of stereo cameras, a dedicated 2GHz Pentium M computer for each pair, a central planning computer, and a low level computer to interface with the robot's hardware.

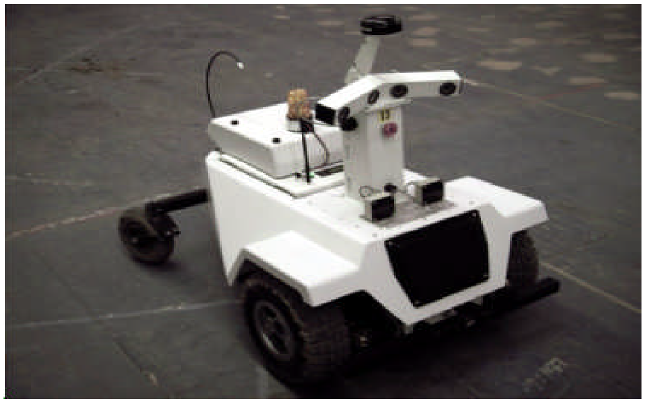


Figure 5: LAGR Test Platform

3.1 Onboard Cameras and Vision Software

The two onboard camera pairs come with the commercial Triclops/Digiclops stereo matching and image rectification software. This software includes several built-in filters and validation methods. Stereo point correspondence is established by a Sum of Absolute Differences algorithm. Several validation methods are used to eliminate false matches: a texture check, which ensures that featureless, low-contrast, regions are not scanned, a uniqueness check, which checks the "goodness" of the minimum in the sum of absolute differences compared with other minima along the scanline, and a surface size validation check. The exact details of the algorithms implemented in the Triclops libraries are not made available for inspection. Their effectiveness is discussed in Section 4.2. The image resolution used for this project was typically 512 x 384 pixels.

3.2 Design Goal

The LAGR platform is tested in a rural environment, and is expected to be able to navigate its way around reasonably textured obstacles such as trees, bushes, shrubs, rocks, and fences. It is not required to classify these objects as anything other than obstacles in order to navigate around them. The robot will be navigating mostly over dirt trails and grassy surfaces, which experience has shown, generally provide sufficient texture to be accurately recognized as contiguous surfaces for ground detection. Typical examples of the terrain are shown in Figure 6.

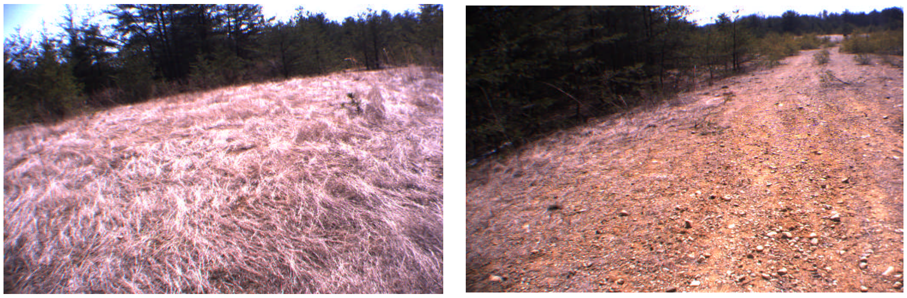


Figure 6: Typical Terrain for Autonomous Navigation

The primary need is to design a software system that will accurately detect obstacles observed through the pairs of stereo cameras and construct a map of the robot's environment as faithfully and reliably as possible, while utilizing as little computation time as possible. A desired value for processing time of one frame of video is typically less than half a second.

4. DEVELOPMENT OF A BETTER VISION ALGORITHM

4.1 Initial Vision System Status

Initially, obstacle detection on the LAGR platform was done using the projection of a 3D point cloud from a filtered disparity image. This point cloud was then broken up into a two-dimensional grid of rectangular columnar cells 0.5 m in width by 0.5 m in length. Cells in which the vertical standard deviation of stereo points was low were used in order to extract a ground plane estimate. This ground plane estimate was then used to classify the remaining stereo points as obstacles if they fell in a region from approximately 0.25 to 1.25 meters above the ground plane. A traversal cost was then calculated for each cell based on the number of stereo points in each cell classified as obstacle points. This cost was taken directly as the obstacle map. An example of such a map, constructed from an indoor scene, is shown in Figure 7. Lighter cells indicate high traversal cost and darker cells indicate lower traversal cost.

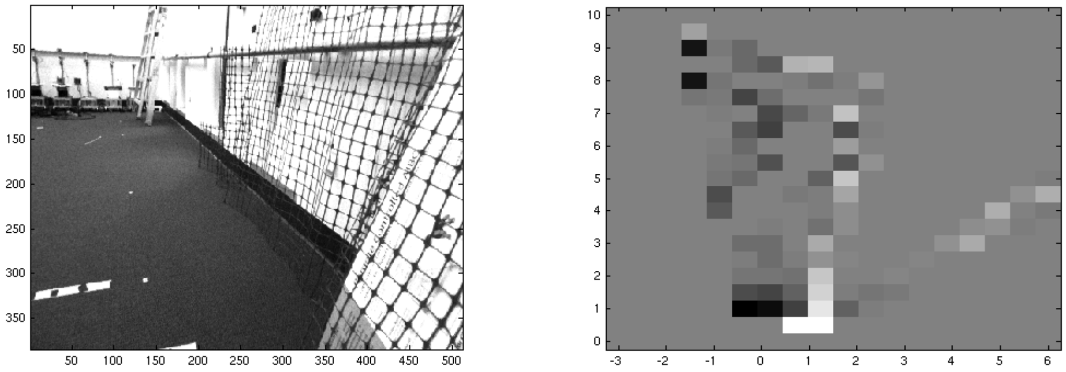


Figure 7: Indoor Scene and Cost Map Generated By Point Cloud Projection

This approach suffers from one main drawback. If a cell contains both ground points and obstacle points, it will not be used in the ground plane estimate because of the high spread in the z-coordinates of the points in the cell. Furthermore, the criterion used to classify a cell as containing ground points assumes that the ground is relatively flat in the robot coordinate system. As Figure 8 shows, a flat sloping ground surface will yield a higher vertical standard deviation, and will be more likely to be rejected for ground plane estimation. Furthermore, if the robot is facing a sloping hill, it may not get any cells with sufficiently low z-coordinate standard deviation, default to using a flat-ground assumption, and register the hill's stereo points as an obstacle, impeding the robot's ability to effectively reach its goal.

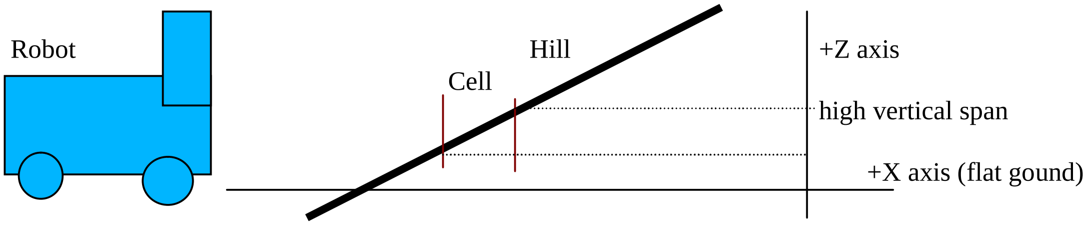


Figure 8: Problems Caused by Sloping Ground

The solution to this problem is to increase the threshold for the maximum vertical standard deviation a cell can contain to still be considered a valid ground point. Unfortunately, this will have the detrimental effect of throwing off the accuracy of the ground plane estimate. This ground plane issue, has, in fact, been a major problem in the vision system to date.

The other problem with the vision system has been the false detections caused by the incorrect matches discussed in Section 2. Specifically, incorrect matches in low-contrast regions such as the sky have resulted in false obstacles being detected directly in front of the robot, causing problems with navigation.

4.2 Initial Attempts at Improvement of the Vision System

The first attempt to improve the vision system involved experimentation with filtration and validation methods on the stereo matching algorithm. Various methods for rejecting false matches generated by the sum of absolute differences minimization were explored.

Three of these algorithms were supplied with the Triclops package that came with the LAGR platform. These were the texture validation, uniqueness validation, and surface size validation. The fourth algorithm was a back-checking confirmation, which compared left-to-right disparity values with their right-to-left counterparts. Mismatches were rejected as invalid.

The problem with the texture and uniqueness validation was that while they succeeded in eliminating most of the spurious data, they also eliminated valid obstacle and ground points. Because the obstacle detection algorithm relies heavily on being able to extract an accurate ground plane estimate from the stereo data, this posed a significant hindrance. After some experimentation on prerecorded images, it was decided that any benefit of error reduction was outweighed by the cost of loosing valuable ground and obstacle data, and consequently, these two validation methods were not employed.

The back-checking confirmation algorithm proved quite effective at eliminating false readings without rejecting valid readings. The disparity images produced by this algorithm were generally accurate and had few incidences of false data, especially when combined with relatively mild uniqueness and texture checks. A significant problem with this approach was the additional overhead and computation time required to make the comparison for every pixel in the disparity image. In practice, the generation of the disparity image takes about 33% of the total computation time for each frame of video, and the cost of increasing it by executing what amounted to only a low-level filtration step was not acceptable.

The final validation technique, surface size validation, proved extremely effective at eliminating false detections while preserving true readings. This method works on the assumption that spurious readings will be small in size in the disparity image and rejects objects that fall below a certain size threshold. This method was so effective that it was incorporated into all subsequent obstacle detection algorithms.

Yet despite any improvements gained from low level filtration techniques, spurious readings still managed to find their way into the disparity image, and the problems of sloping ground and cells occupied by both ground and obstacles remained. These problems served as the motivation for a higher level analysis of the disparity image.

4.3 Obstacle Detection Based on Disparity Image Segmentation

A cursory glance at any disparity image, such as Figure 9, will immediately suggest a method of obstacle detection. It is a fact that obstacles appear in the disparity image as large contiguous objects with gradually changing disparity values over their area. It was thought that the extraction of such large contiguous objects could be an effective method of obstacle detection.

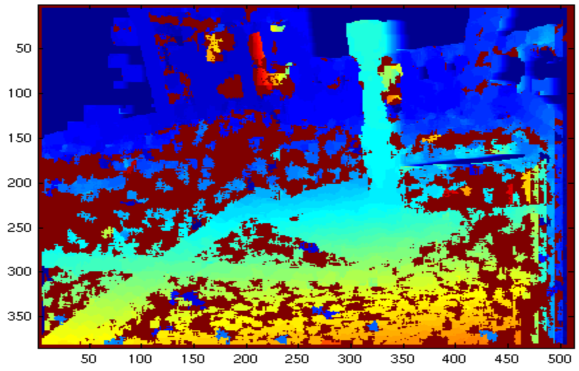


Figure 9: A Disparity Image

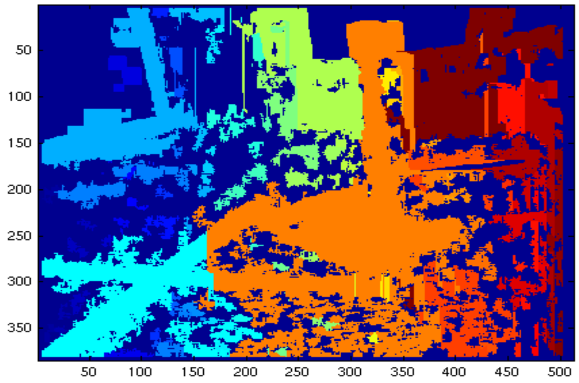


Figure 10: Segmentation of Figure 9. Regions of the same color are unified objects.

A simple union find algorithm—similiar to the one used to measure the size of contiguous objects for the surface size validation—was run on disparity images captured through the LAGR's stereo cameras. As Figure 10 shows, many obstacles seen in the disparity image in Figure 9 are indeed successfully segmented out and separated into discrete units. Once this is done, criteria such as real world size and height above the ground may be used to populate not a continuous-valued cost map susceptible to false readings, but a discrete-valued binary map showing precisely where the large obstacles are in the robot's environment.

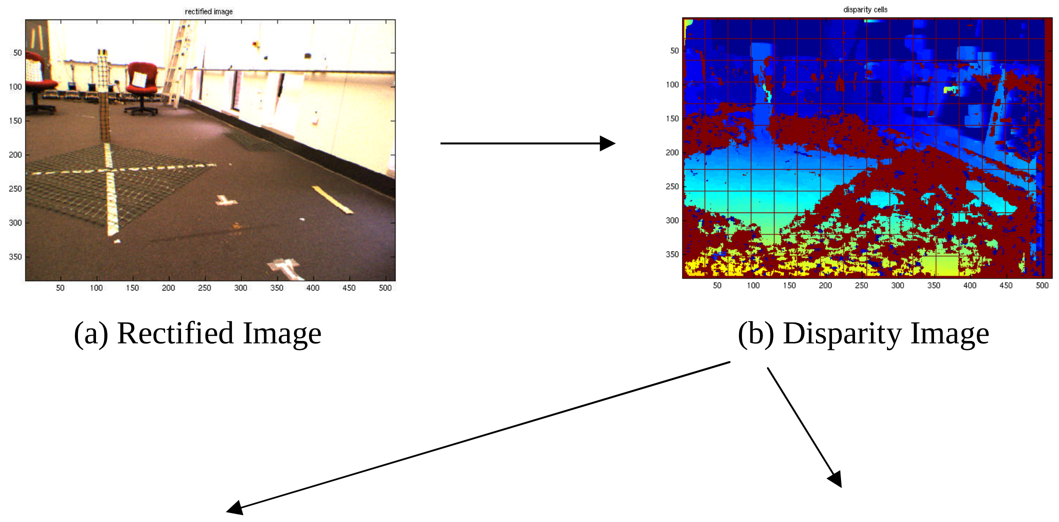
However, this simplistic approach suffers from two related and important drawbacks. First, it does not solve the problem of obtaining an accurate ground plane estimate, which, as was mentioned before, creates a problem in sloping or uneven terrain. Second, as is evident in Figure 10, objects that are sitting on the ground blend into the ground in the disparity image, and a union-find algorithm will classify both the object and the ground it is sitting on as a single unit. Potentially, entire swaths of ground and objects will be classified as a unified body, and cause the map to be populated with obstacles where none exist.

What was required was a way to separate the objects that sit on the ground from the ground itself. Murray accomplished this by computing a surface normal vector for a curve fit to each pixel's neighborhood and using it as the segmentation criterion. As seen in [1], this method is quite successful. Unfortunately, it requires quite a bit of computation time—up to 2 minutes per frame for a 320 x 240 image. Even with Murray's proposed code optimizations, his method is designed to extract highly precise information about complicated surfaces. This was not the goal of this design. Since the only required data was a fairly rough estimate of where an obstacle is located, and not any information about its structure, a slightly cruder algorithm was in order.

4.4 The Final Obstacle Detection Algorithm

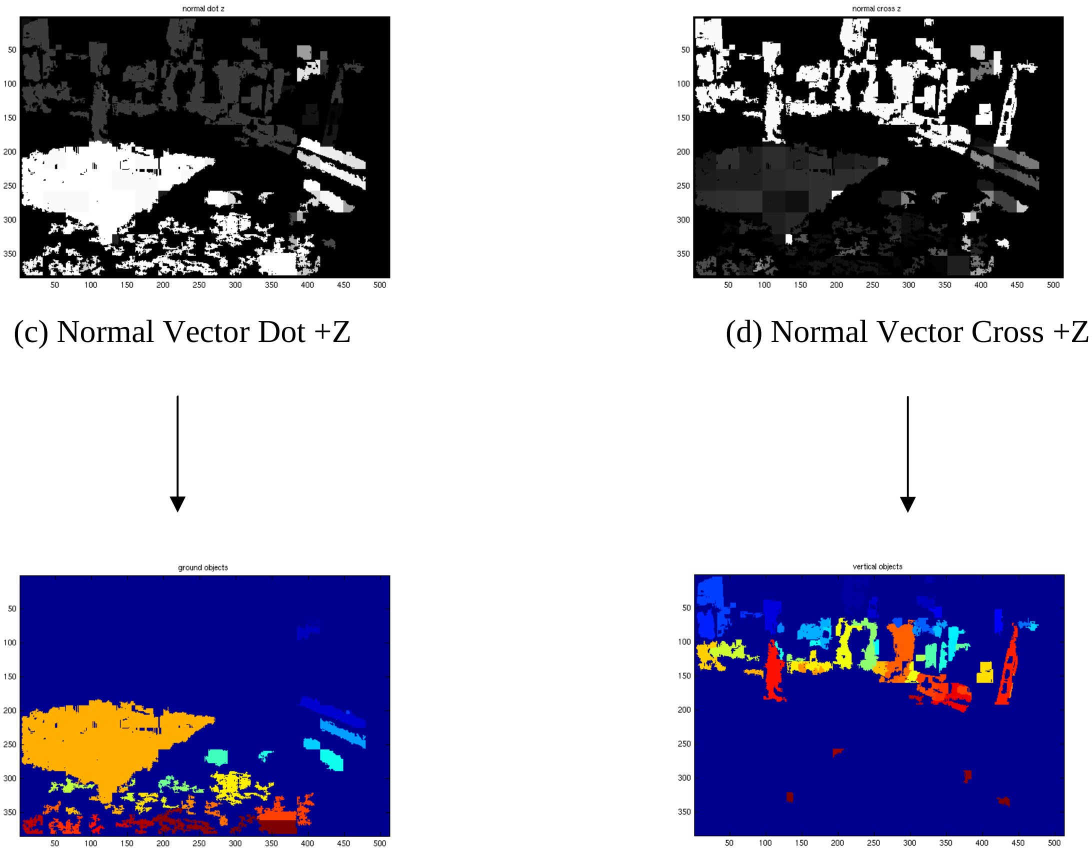
Now, intuitively, visual information contains only data about the surfaces of nearby objects. Further intuitive reasoning will reveal that most obstacles are vertical and that the ground is mostly horizontal. The final version of the obstacle detection algorithm developed over the course of this project employs these simple observations in order to classify stereo points generated from a regular disparity image as obstacles or ground points based on the orientation of the normal vector of dynamically determined surface elements in the image. A flowchart of the algorithm is given in Figure 11.

First, a disparity image is generated by the commercial Triclops software from rectified image pairs. The right rectified image is shown in 11 (a) and the corresponding disparity image in 11 (b). From (b), the distance image is calculated using the equations of Table 1. In order to save computation time, the x- and y-coordinates are also calculated for each pixel, as they will be required in future steps. Next, the distance image is split into 32 x 32 pixel nonoverlapping regions and the union-find algorithm is executed on these 32 x 32 pixel patches in order to extract contiguous surfaces for the calculation of surface normal vectors. It should be noted that the initial splitting of the image avoids the problem of having objects blend into the ground by keeping the segmentation algorithm confined to the local 32 x 32 pixel region.



(a) Rectified Image

(b) Disparity Image



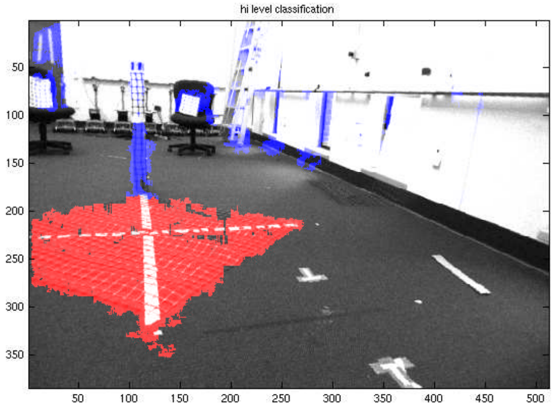
(c) Normal Vector Dot +Z

(d) Normal Vector Cross +Z

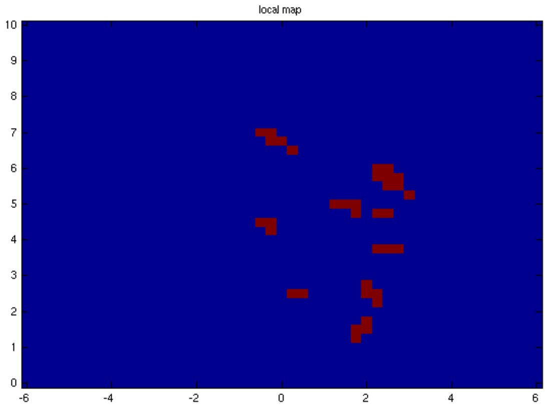
(e) Horizontal Surfaces

(f) Vertical Surfaces

Figure 11: Obstacle Detection Algorithm Flowchart (Part I)



(g) Final Classification of Stereo Points (Red = Ground, Blue = Obstacle)



(h) Bird's Eye Map of Local Environment (Robot is at (0,0); Red Dot = Obstacle)

Figure 11: Obstacle Detection Flow Chart (Part II)

For each contiguous object extracted by the union find algorithm in the 32 x 32 pixel patches, the 3D coordinates of all the pixels comprising that object are collected. From these points, a plane of best fit is calculated by means of orthogonal distance regression. As Figure 12 and the equations in Table 2 show, this problem is solved by finding the eigenvectors of the covariance matrix generated by the 3D coordinates of the points comprising the object. The eigenvector corresponding to the minimum eigenvalue gives the direction of the surface normal vector.

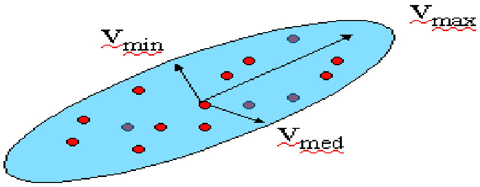


Figure 12: Orthogonal Distance Regression

$$\begin{bmatrix} \frac{1}{n}\sum dxdx & \frac{1}{n}\sum dxdy & \frac{1}{n}\sum dxdz \\ \frac{1}{n}\sum dydx & \frac{1}{n}\sum dydy & \frac{1}{n}\sum dydz \\ \frac{1}{n}\sum dzdx & \frac{1}{n}\sum dzdy & \frac{1}{n}\sum dzdz \end{bmatrix} = \begin{bmatrix} \vec{v_1} & \vec{v_2} & \vec{v_3} \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \cdot \begin{bmatrix} \vec{v_1} \\ \vec{v_2} \\ \vec{v_3} \end{bmatrix}$$

Table 2: Equations for Orthogonal Distance Regression

Once this surface normal vector is computed for each object in each 32 x 32 pixel patch, it is dotted with the +Z direction of the vehicle coordinate frame. Absolute values of this dot product near 1.0 indicate a highly horizontal surface likely to be the ground and absolute values of this dot product near 0.0 indicate vertical surfaces most likely to be obstacles. This dot product value is recorded in a new image for each pixel in each contiguous object, and the corresponding cross product value (= sqrt($1 - \text{dot}^2$)) is also recorded in a separate image. Figure 11 (c) and 11 (d) show the absolute value of the dot and cross products of the surface normal vectors with the +Z axis in the vehicle coordinate frame. It can be seen that in Figure 11 (c), ground points are brighter, representing a high value for the dot product, and in 11 (d), the cross product values are higher for vertical obstacles.

After this initial low level classification step, a ground plane estimate is obtained by searching for large contiguous regions of pixels classified as horizontal ground points with a union-find algorithm operating only on pixels belonging to sufficiently horizontal surfaces. In Figure 11, orientations within 30 degrees of absolutely horizontal were included. The results of the union-find are then processed by an arbitrarily-chosen plane-fitting algorithm. For the purposes of this experiment, orthogonal distance regression was used in this step. Similarly, vertical obstacles are extracted by running the same union find algorithm on the distance image, but now using only the pixels previously classified as belonging to vertical surfaces by the low level preliminary classification step. The horizontal and vertical results of the union-find algorithm are displaying in Figure 11 (e) and 11 (f), respectively.

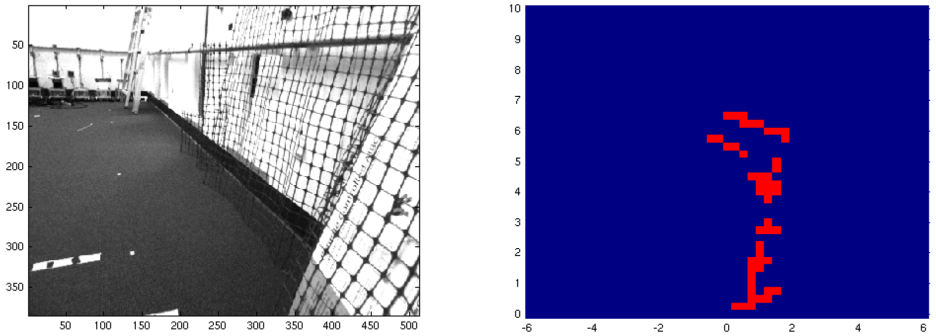
Once the ground plane estimate is obtained, and the coordinates of the vertical obstacles are extracted from the step above, the map generation step is trivial. Each vertical obstacle has a set of points with coordinates in the vehicle coordinate system. The point has a vertical span, a horizontal span, and a horizontal orientation given by the *maximum* eigenvector of the covariance matrix generated by the x- and y-coordinates of the points corresponding to the obstacle. These quantities are compared with certain thresholds for height above the ground plane, pixel count, and absolute size in the robot coordinate system. Objects satisfying these thresholds are classified as valid obstacles. Figure 11 (g) shows the final classification of pixels in the stereo image superimposed onto the rectified image. Areas shaded red are composed of pixels classified as belonging to large horizontal objects assumed to be the ground, and areas shaded blue are comprised of pixels classified as belonging to vertically-oriented objects satisfying the aforementioned criteria. Once these ground points and obstacles are classified, their coordinates are recorded in the local environment map, as shown in Figure 11 (h).

5. SUMMARY OF EXPERIMENTAL RESULTS

The results in Figure 11 and Figure 13 are typical examples of the accuracy with which the algorithm can operate. As can be seen in Figure 11, all major vertical obstacles that have sufficient contrast to be detected by the stereo system are faithfully placed into the local map in Figure 11 (h). Figure 13 (a) shows the indoor scene from Figure 7, with artificially added texture in the form of garden fencing along the otherwise featureless wall along the right side. It is seen in Figure 13 (b) that the algorithm accurately records the entirety of the wall as an obstacle, as well as the ladder and small robots toward the back of the room.

Compared with the results of the current vision algorithm (in Figure 7), the results of the new algorithm in Figure 13 seem much sharper in terms of what is an obstacle in the robot's environment. The entire wall can be seen as impassible, whereas Figure 7 shows fading traversal costs over its length.

The discontinuities in the wall seen at farther distances in the map are a consequence of the uncertainty of obstacle position increasing with lower disparity values, as discussed earlier in Section 2.1. Average execution time per frame of video peaked at 1/6 – 1/5 seconds per frame.



(a) Rectified Image

(b) Local Map (Red = Obstacle)

Figure 13: Indoor Scene and Map

6. DISCUSSION AND CONCLUSIONS

Effective obstacle detection is a vital component of autonomous navigation, and vision-based obstacle detection methods offer the advantage of cheap sensors, wide viewing angles, and a guarantee that the raw data from the cameras contains sufficient information about the environment in order to generate an accurate and reliable map. Stereo vision offers the ability to effectively locate objects in three-dimensional space with relatively simple and fast algorithms, and the ability to access object properties such as size, location, and surface orientation.

The algorithm developed over the course of this project uses these simple properties, surface orientation, in particular, to generate accurate, reliable, and stable maps, and has proved to be robust when faced with variable environmental conditions such as sloping terrain. Testing has shown it to be reasonably resistant to noisy stereo data and false detections, as compared to the obstacle detection algorithm previously implemented on the LAGR platform, and as seen in Section 7, the maps it generates are much more decisive in terms of what is and what is not traversable. Furthermore, it is sufficiently fast to be used in real time on a mobile platform. Its success is further confirmation of the effectiveness of using surface orientation as a classifying criterion for 3D stereo data.

7. RECOMMENDATIONS

After considerations of various low-level obstacle detection algorithms, it is recommended that the algorithm developed during the course of this project be implemented on the LAGR platform in order to improve its ability of effectively identify obstacles in its environment. This algorithm offers the advantages of reliability, robustness, and accuracy. Furthermore, the algorithm can be easily implemented as a starting point for more sophisticated obstacle detection techniques that rely on learning algorithms and other hi-level techniques.

8. ACKNOWLEDGEMENTS

I would like to thank the many people without whose generous support, this project would not have been possible. Professor Daniel Lee of the University of Pennsylvania Electrical and Systems Engineering Department has been a source of invaluable wisdom and guidance. His support as an advisor is much appreciated. Paul Vernaza of the University of Pennsylvania GRASP Laboratory is a master guru of robotics hardware. His assistance with technical issues and software allowed this project to proceed smoothly and efficiently. Many thanks also to the National Science Foundation, whose support for undergraduate research initiatives made this project possible, and to the University of Pennsylvania Department of Electrical and Systems Engineering Department for hosting the REU. And, of course, many thanks to Prof. Jan Van der Spiegel for capably managing this undergraduate research program.

9. REFERENCES

1. D. R. Murray, Patchlets: a method for interpreting correlation stereo 3D data, *PhD Thesis, University of British Columbia, 2004.*

2. J. Shi, and J. Malik, Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol 22, Issue 8, Aug. 2000, 888 – 905.

3.N. Molton, S. Se, J.M. Brady, D. Lee, and P. Probert, A Stereo Vision-Based Aid for the Visually-Impaired, *Image and Vision Computing Journal*, Vol 16, No 4, (1998) 251 - 263.

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

PATH PLANNING MOBILE ROBOTICS

NSF Summer Undergraduate Fellowship in Sensor Technologies
Louie Huang (Electrical and Systems Engineering) - University of Pennsylvania
Advisor: Dr. George Pappas

ABSTRACT

One of the fundamental problems in the field of robotics is path determination and motion planning. The project described in this paper focuses on path determination in a known environment. The goal is to enable a mobile robot to successfully navigate an environment according to a specified temporal logic formula. On a high level, temporal logic formulas can effectively provide the robot with directions on where and when to go. As part of this project, a program will be developed to formulate a continuous path plan that will fulfill the temporal logic formula supplied for the robot to follow.

In this project, the ActivMedia Pioneer 3-DX robot will be used. This robot model was chosen because it is preconfigured for basic navigation. Preprogrammed with algorithms for shortest path determination, obstacle avoidance, and localization, the Pioneer 3-DX is also capable of new navigation techniques that can be programmed in C and C++. Maps of known environments can be generated by the Pioneer 3-DX. A graphical user interface program will utilize these maps in conjunction with user supplied directions as expressed by a temporal logic formula to construct a path plan. The path determined by the program can then be relayed back to the robot for implementation.

Table of Contents

11. INTRODUCTION

In the constantly evolving field of robotics, path determination is a topic that attracts much interest because of its numerous potential applications. A robot which can plan its own path when given destinations and certain guidelines can be used for patrol and mobile surveillance or transport and delivery of items. Using robots for such applications can not only offer convenience to users but can also save lives when employed in military situations. Such beneficial uses validate the importance of research pertaining to path determination.

The project that this paper is based upon focuses on determining a path that fulfills a specified direction represented by a temporal logic formula. It stems from an earlier paper [1] published by two graduate students -- Georgios Fainekos and Hadas Kress-Gazit -- and Professor George Pappas of the General Robots Automation Sensing Perception (GRASP) Laboratory at the University of Pennsylvania. The paper presents the method by which continuous path plans can be generated for a robot in a known environment. The continuous path will be implemented by the robot and should satisfy temporal logic input. The overall goal behind this study is to allow users to effectively direct a robot on a high communication level that is close to natural human language in the form of temporal logic [1].

The sections that follow this introduction will further explain the details of the project and the progress made to date. Section 2 will discuss the background of the development of a continuous navigation path. Section 3 will introduce the ActivMedia Pioneer 3-DX, the specific robot that will be used in this project. This section will also present some key navigation behaviors that the Pioneer 3-DX is capable of and discuss how new behaviors can be programmed. Section 4 will cover the purpose and the development of the robot map graphical user interface (GUI) program, which will be used to ultimately generate paths for the robot. In section 5, discussions and conclusions of the overall project are presented. Section 6 covers recommendations for future work and Section 7 is dedicated to acknowledgements. All references can be found in Section 8.

12. OVERVIEW OF CONTINUOUS PATH CREATION

The following description of continuous path creation was originally presented and discussed in the research paper [1] that served as a foundation for this project. In the interest of brevity, the prior work on continuous path creation will be briefly summarized here. Readers interested in further detail should consult the original source.

The process of creating a continuous path for implementation begins with navigation directions. Directions are to be given in the form of temporal logic formulas. Temporal logic formulas can delineate multiple destinations and specify when the destinations should be reached. For example, an instance of a temporal logic formula can be used to direct the robot to visit all rooms but not to visit a certain room until all other rooms have been visited. Temporal logic formulas are close to human language, providing greater accessibility to users who would therefore not need to be concerned with the lower implementation of the robot’s movements. A direction alone is not sufficient for path generation. The robot must know its environment through a map of the environment that pertains to the directions provided.

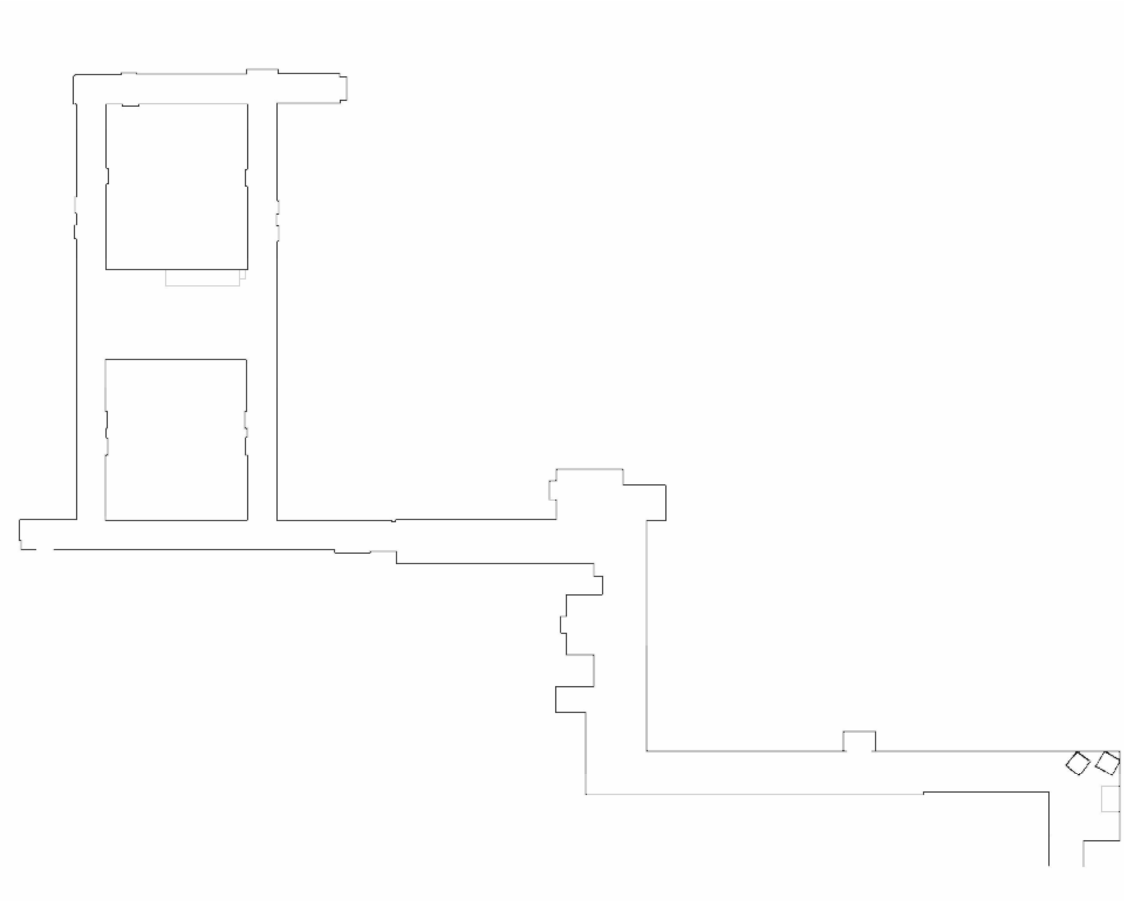


Figure 1: 2-D environment map of a floor in the Levine Building

Another necessary component for creating a path is a 2-D graphical map (as illustrated in Figure 1) delineating, from a bird’s eye view, walls and obstructions of the environment to be navigated. The next step in arriving at a continuous path is to develop a discrete path through model checkers. To develop a discrete path, the 2-D graphical map needs to be partitioned into discrete units. Although other partitioning methods could also be considered, triangulation has been recommended because of its relative ease in computation and readily available algorithms. The destinations then specified by the temporal logic formula will each be composed of at least one triangle. A discrete path can be created that visits the necessary triangles to fulfill the temporal logic formula. Upon generation of a discrete path, a continuous path can be developed. The continuous path must obviously still satisfy the original temporal logic formula. For further detail on the process of continuous path creation, please refer to the original path planning paper [1].

13. ROBOT DESCRIPTION

13.1 Pioneer 3-DX

The Pioneer 3-DX robot made by ActivMedia Robotics is an all purpose robot suitable for research [2]. The robot is readily capable of basic navigation functions. One of the most important features of the Pioneer 3-DX is its ability to localize itself fairly accurately within a known environment. In Figure 2, the Pioneer 3-DX is observed from its front. Localization is made possible by the eight sonar shown in the figure; two of the sonar are hidden from view as they are mounted on the sides of the robot. The laser rangefinder that sits on top of the robot also assists with localization. The laser rangefinder has the advantage of greater range and accuracy than the sonar. However, the sonar is necessary to detect low lying obstacles close to the ground.

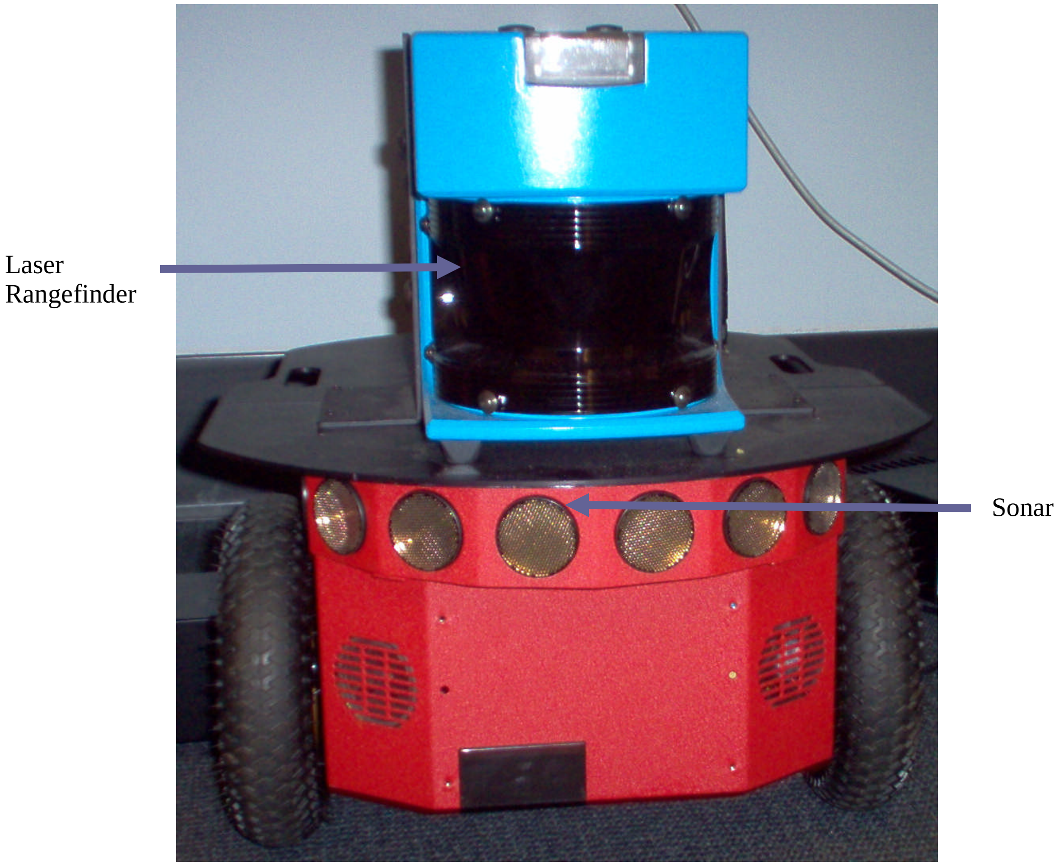


Figure 2: Front View of Pioneer 3-DX

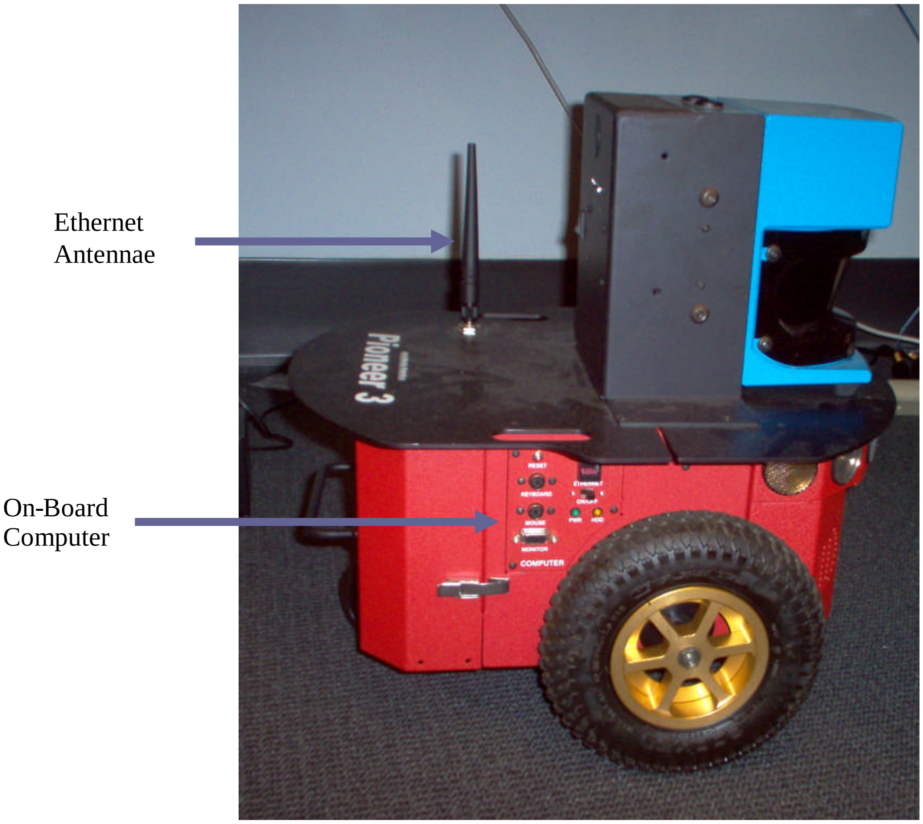


Figure 3: Side View of Pioneer 3-DX

For the Pioneer 3-DX to localize itself, it would need to know its environment. To acquaint the robot with an unfamiliar environment, a joystick can be plugged into a USB port on the robot. With the laser rangefinder activated, the robot can be driven manually with the joystick until the new environment has been fully covered. The robot has an on-board computer (see Figure 3) which can store points scanned from the laser rangefinder and create a map file similar to that shown in Figure 1. Sometimes, erroneous points can be introduced during the mapping process. Because the environment is unlikely to be static, movement by people or other mobile objects within range of the laser will be picked up and plotted on the map. The laser also lacks the ability to successfully detect transparent surfaces such as glass on windows as obstructions. See Figure 4 for the original floor map of the Levine Building. Hazy or noisy areas exist where erroneous or unintended data points are plotted. To correct these possible errors, the map file needs to be edited.

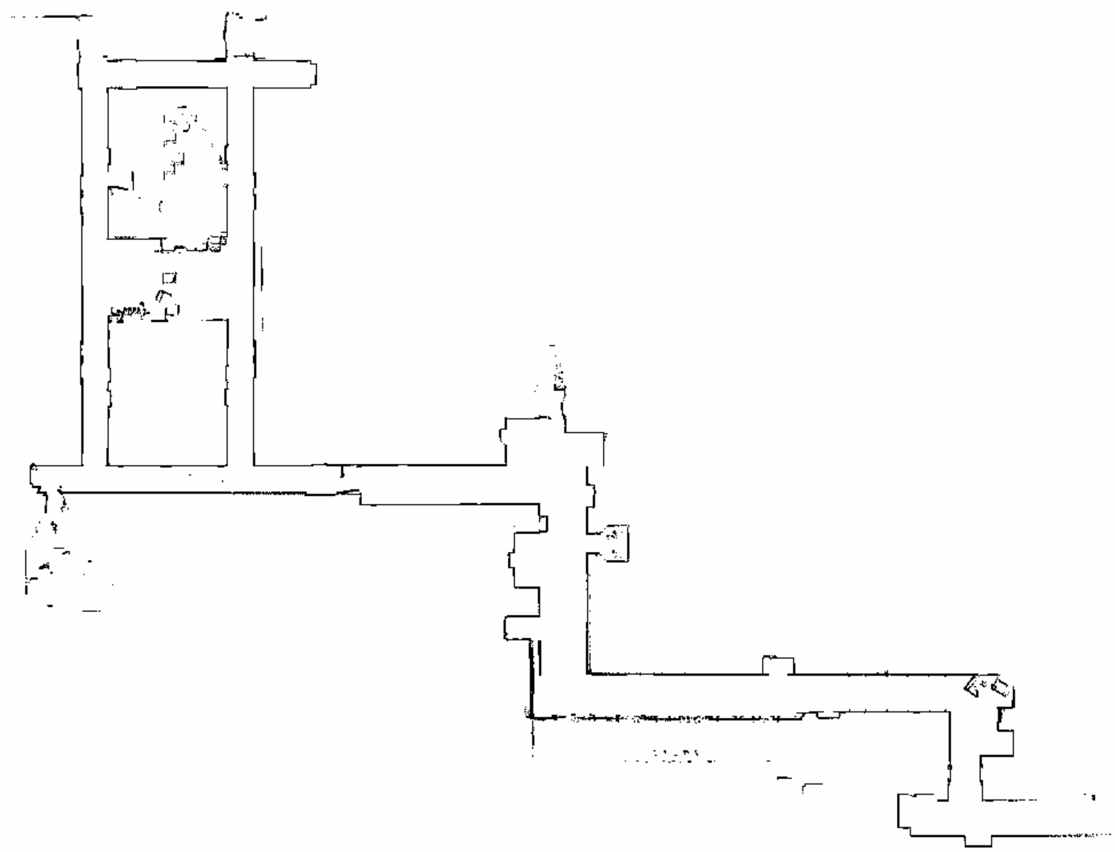


Figure 4: Robot generated map of floor in Levine Building before editing

The on-board computer of the Pioneer 3-DX is sufficient for implementing the programs required for the robot’s navigation. However, due to resource restrictions, the map editing process will be fairly slow on the on-board computer. To modify the map file more efficiently, files on the Pioneer 3-DX can be sent to a base computer via wireless Ethernet. Figure 3 shows the Ethernet antennae that can wirelessly transmit information packets to another PC. Upon transferring the map, editing can be done on the base computer and then the map can be sent back to the robot for storage and future use. Besides editing out erroneous points, the base computer will also eventually be used to implement the process by which the map is partitioned and a continuous path based on temporal logic formulas supplied is created.

13.2 Robot Behaviors

The Pioneer 3-DX comes equipped with some useful high level navigational behaviors. Two of the most important behaviors are obstacle avoidance and shortest path determination. In obstacle avoidance, the robot takes readings from its laser rangefinder and sonar to determine whether there are objects blocking its path. Upon detecting an obstacle in its path it will stop to avoid collision. There are several behaviors that can follow the stop action after obstacle detection. One consequent behavior that occurs after detecting an obstruction instructs the robot to back off after stopping and turn towards a different direction. The other important behavior with which the Pioneer 3-DX is preprogrammed is shortest path determination in a known environment. This behavior requires a map file of the environment and a home position on the map from where it will start. A destination is also specified on the map. With this map file, the robot has the ability to determine a shortest path while avoiding walls specified by the map.

The robot behaviors that the Pioneer 3-DX can exhibit are developed with the ActivMedia Robotics Interface Application (ARIA). ARIA is programmed in C++ and its classes can be used to obtain readings from the robot's sensors (sonar, laser, wheel movements, etc.) [2].The robot's basic movements -- as represented by direction -- and speed can also be manipulated by ARIA. To develop new behaviors or modify existing navigation behaviors, ARIA's classes can be used in a C++ programming environment such as Microsoft Visual Studio C++. Because the on-board computer has limited resources, a programming environment such as Visual Studio would be operated at a base computer. A behavior after compilation then can be relayed back to the robot for implementation through wireless Ethernet.

Using SRI International's Saphira is often an easier way of implementing behaviors. Saphira is built utilizing ARIA and exists as a higher level programming environment for robot behaviors. As a consequence of its higher level nature, programming in Saphira is inherently simpler than programming in C++ with ARIA. Saphira includes its own programming language, Colbert, which is based on the C programming language [3]. Colbert can be used to construct activities which provide navigational instructions to the robot upon implementation. Activities can be used to provide direct movement commands which, for instance can tell the robot to move forward 10 meters and turn 90 degrees. An activity can also be used to implement robot behaviors like obstacle avoidance by producing movement commands that depend on sensor readings [3].

Saphira also provides a simulator from which activities can be interactively tested and modified. The robot used in the simulation can be an actual Pioneer 3-DX or a virtual robot that emulates actual physical robot limits. The simulator allows multiple activities to be implemented. Of course, some activities may contradict others at times in which one activity is directing the robot forward and another is simultaneously directing it to move in reverse. To resolve such movement conflicts, Saphira uses a system by which each activity is given a priority. Direct motion directions are usually given precedence over directions derived from sensor dependent behaviors [3]. For implementation of activities that involve a known environment, map files can be imported into the Saphira simulator.

14. GRAPHICAL USER INTERFACE FOR ROBOT MAPS

14.1 Purpose

As discussed earlier, map files generated using the robot's laser rangefinder may have many points or representations of obstacles which should not be there. Also, walls that really exist may occasionally fail to appear if the wall is transparent. To correct these problems, a graphical user interface is needed to modify the map file. The map can be readily accessed from a base computer that receives the file from the robot's wireless Ethernet transmission. ActivMedia, the manufacturer of the robot already provides software to edit the map. The ActivMedia Mapper 3 program allows users to eliminate erroneous points on the map and insert lines that represent walls. It also allows users to add goal points or destinations. Other additions that can be made to the map file include forbidden regions and forbidden lines, which may not be actual obstructions in the environment, but are nevertheless areas that the user wants the robot to avoid.

The Mapper 3 is fairly useful but for the purposes of this project, it is not sufficient. To fulfill the necessary specifications that would allow for the creation of the continuous path, the Mapper 3 would need to be able to partition the map into discrete regions through triangulation or some other partitioning method. In addition, the Mapper 3 would also need to be able to accept temporal logic formulas and construct a discrete path that would ultimately lead to the continuous path for the robot to implement. Because the Mapper 3 is not open source, it was decided that the best option would be to build a graphical user interface from scratch that -- in addition to including most of the Mapper 3's features -- would also allow for triangulation and path determination based on directions given by temporal logic formulas. A crucial development requirement is that the modifications made to any maps through this robot map GUI must still maintain the map files' compatibility with the robot.

The base computer that is used for the Pioneer 3-DX in this project is an IBM compatible machine that runs Microsoft Windows. To build the GUI as a Windows application, Microsoft Visual Studio .Net is used as the programming environment. C++ is used as the programming language in Visual Studio .Net to construct the GUI as a Windows form application. The choice to use Microsoft Visual Studio .Net rather than other environments is based on the ease of use that Visual Studio .Net offers for programming GUIs using the Windows operating system. The classes provided by Visual Studio .Net particular to Windows GUI programs simplify the GUI programming significantly.

14.2 Development of the GUI

The first function to be developed for the GUI was to display map files correctly. Map files have the extension .map and can be accessed from a basic text editor. A map file generated by the Pioneer 3-DX always begins with the line "2D-Map". The three lines that follow this initial line pertain to obstacle points. Most environments would not have obstacle points since walls are represented by lines. An obstacle point is most likely noise since an obstacle point represents an object with an area of 1mm^2. The three lines that pertain to any existing obstacle points provide the minimum x and y coordinates and maximum x and y coordinates in millimeters, as well as the number of obstacle points in the map. The next three lines contain similar information for obstacle lines. Following the information on obstacle lines, the map file contains the definitions for forbidden areas, forbidden lines, goals, and home points in that order. Each line contains an individual definition represented by coordinates in millimeters and begins with the word "Cairn:" and the type of object defined. After definitions of these objects, the map file contains a list of lines each of which consists of 4 numbers and represents two coordinates in millimeters. At the beginning of this list is the word “LINES” to signify that the list consists of definitions for obstacle lines in the map. After this list is the list for obstacle points which begins with the word “DATA”. This list contains lines of 2 numbers, each pair belonging to the coordinate pair of one obstacle point.

To properly display the map file, the GUI must read the map file and utilize the information provided to construct a graphical map. Extracting the data and drawing the map lines and objects in the GUI’s main panel did not prove to be a complicated task. However, the map’s coordinate system is inherently different from that used in Windows forms programming. This can be seen in Figure 5. The + sign signifies the direction in which coordinates increase positively. The map therefore needed to be adjusted according to its maximum and minimum coordinates as they are given at the beginning of each map file, flipping them vertically due to the inverting nature of the y axis in the Windows forms coordinate system. For details, please refer to the translate function in the Form1.h class in the Appendix. Another problem arose in the display of the map because the coordinates given in millimeters were too great in magnitude to be displayed at a level for the user to view the map as a whole. Therefore, the x and y parts of each coordinate need to be reduced in magnitude through dividing by a reduction factor. The standard reduction factor when a map is loaded is 50. This number was determined through trial and error to produce a suitable viewing size for the map.

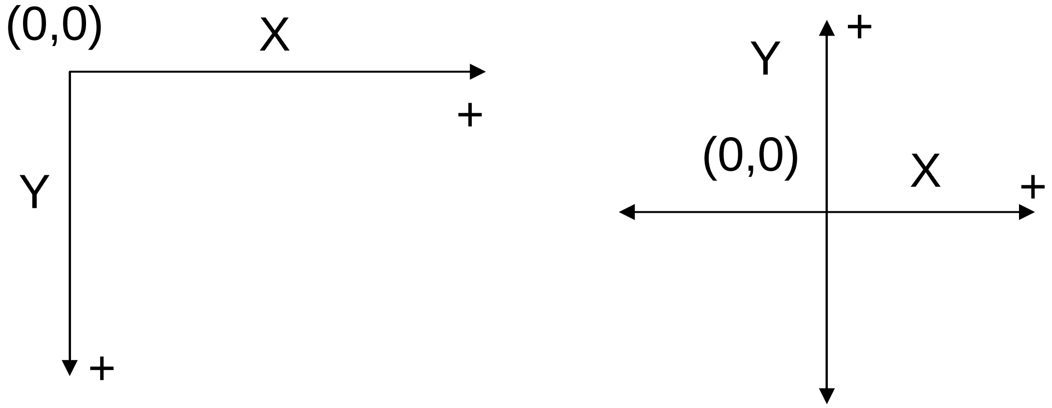


Figure 5: Differing coordinate systems: Windows form coordinate system (left) and standard map coordinate system (right).

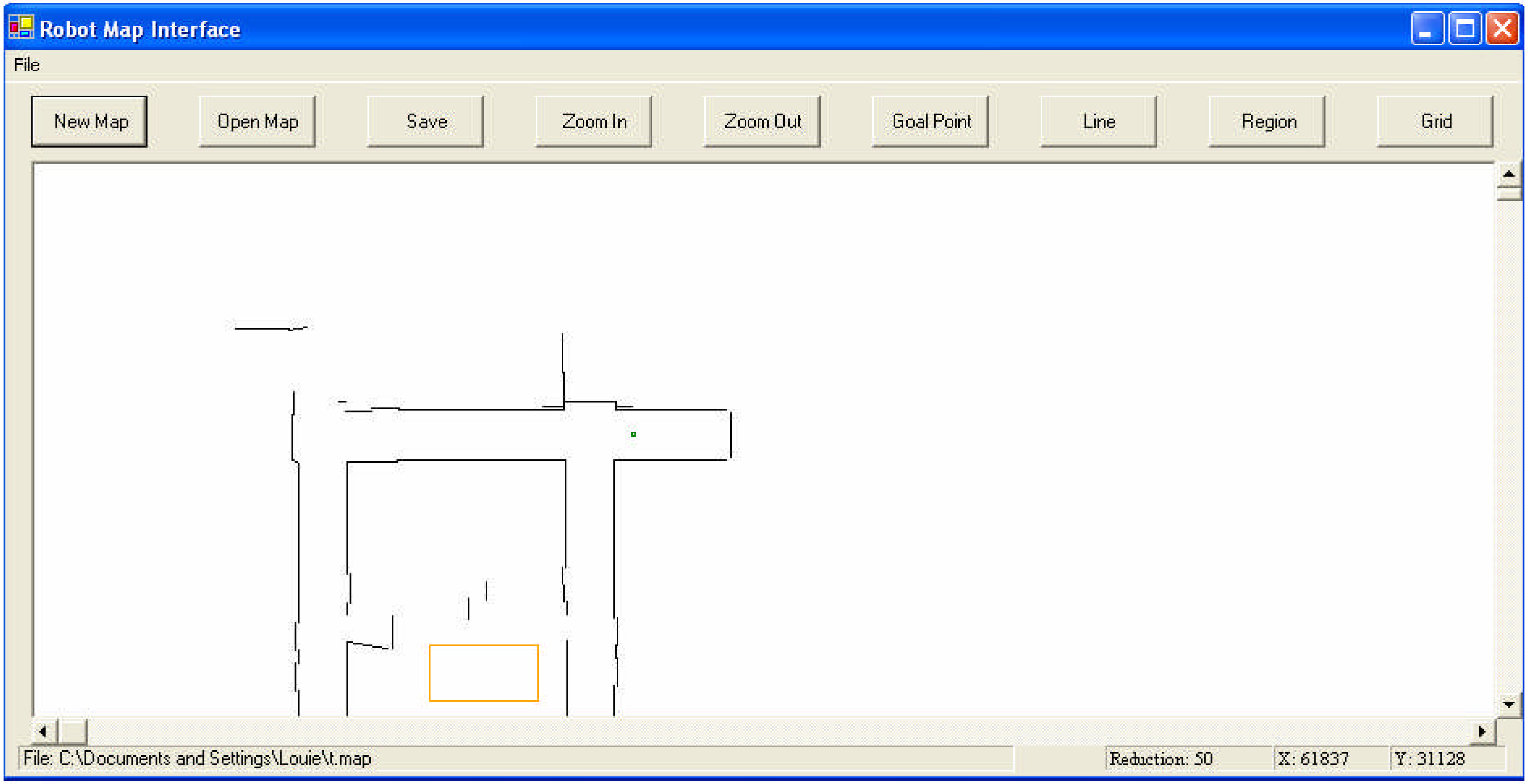
Even with the reduction factor at 50, the main panel is too small to display every map. The map shown in Figure 1 of a floor in the Levine Building could not be fully shown in the panel. To account for this problem, scrollbars were added to the GUI. The vertical and horizontal scrollbars, whose limits are determined by the maps’ maximum coordinates, successfully allow any map to be fully displayed. To allow for greater viewing versatility, in addition to the scrollbars, two buttons were added to the GUI that would control zoom. One button would zoom in on the image and essentially magnify it by reducing the reduction factor. The second button would zoom out and essentially shrink the image by increasing the reduction factor. With the addition of the zoom functions, most maps can be scrutinized up close or viewed in their entirety without scrolling on the main panel.

The next functions to be implemented would allow for the additions of lines, goal points, and regions to a map file. Three buttons are added to the GUI each to represent a type of object to be drawn. For example, if the line button is pressed, the map file will only accept the additions of lines. The reason these buttons are necessary is because each object is drawn similarly using mouse clicks. Lines are to be drawn by clicking down the mouse button at the site where the first endpoint should be. The mouse button is held down until the mouse is moved to the second endpoint. Goals are drawn by clicking any point in the map. For ease of viewing, goals are shown as small green squares. The center of each square is the actual goal point. Regions are currently general custom structures added to the map. They may constitute future forbidden regions but they can also be used to parameterize the map. Though triangulation is likely to be the parameterization method to be used, the regions' function is currently programmed to construct any polygon. To define a simple rectangle, the user only needs to click two points, which will always be the upper left and the lower right vertices. The second point must be double clicked to signify that the region will be a rectangle. Any other type of polygon can be constructed by mouse clicks. Each mouse click will define a vertex and the last vertex will require a double click signifying the end of the region definition.

The added lines, goal points, and regions are not saved to any files when drawn onto a map. To register the additions made to the map, a save function was implemented in the GUI. To maintain the integrity of the original map file and its compatibility with the robot, the definition for each new line and goal must be added to the file in the appropriate areas with the appropriate syntax. The regions defined in the GUI are not necessarily forbidden regions and therefore their definitions are stored in an auxiliary text file with a filename that is always the name of the original map followed by "_regions". In addition to saving the map files generated, the GUI will also save newly created map files. The new map function in the GUI will generate a blank map with user defined dimensions. Maps newly created from the GUI will be indistinguishable in format from robot generated maps.

To date, the last function to be added to the GUI is the grid display and the position tracker. When editing a map, it is useful to have a grid and to know the coordinates of the region being modified. The grid function can be displayed when a map file is opened by a click of the grid button. It can also be made invisible by a second click. Grid lines are 1000 mm or 1 meter away from one another. As a result, the perceived distances shrink or grow when a zoom is applied to the map. The position tracker simply tracks the mouse position over the main panel where the map is displayed. Whenever the mouse moves, the tracker display located at the bottom right corner of the GUI is updated. The coordinates are displayed in millimeters and properly match the coordinate system native to the maps as portrayed in Figure 5.

The latest robot map GUI is shown in Figure 6. Each of the 9 buttons represents a function described. The green square represents a goal point while the orange rectangle is a region. The black lines represent walls in the map. The bottom of the window shows the file currently being accessed. At the bottom right corner, the reduction factor is displayed along with the current mouse position.



File: C:\Documents and Settings\Louie\t.map

Reduction: 50

X: 61837

Y: 31128

Figure 6: Snapshot of the robot map GUI; the map file t.map is being loaded at a reduction factor of 50

15. DISCUSSION AND CONCLUSIONS

The goal of the GUI in this project was to successfully access robot generated map files and make modifications. In this aspect, the GUI in its current stage is fairly successful. With the GUI, new lines and goals can be added to existing maps or created in new maps. Robot compatibility with files modified and created by the GUI has been successfully maintained. Regions which may be later used to partition the environment can be added in the GUI. As the regions defined in the GUI are not directly useful to the robot, regions made to a map file are stored in separate text file with a similar name as the original map. The GUI, unlike the Mapper 3 offered by ActivMedia, does not allow users to add home points or forbidden areas or lines. For the project at hand, these features do not appear to be needed. However, if this should change in the future, simple modifications can be made to the robot map GUI program to allow for the addition of home points, and forbidden regions and lines.

One peculiarity of the GUI is that sometimes at large reduction factors, the coordinate display may show coordinates slightly different from their expected appearance. The scale of map coordinates stored is 1 mm per pixel on a monitor display. This scale can only display a small piece of a map at any one time. To increase the distance per pixel so that more of the map can be viewed at once on the screen without having to scroll around, a reduction factor is maintained by the GUI. The reduction factor reduces the magnitude of all coordinates in the map. For higher calculation speeds, the division of coordinates by the reduction factor is made an integer division, dropping the decimals. This rounding off results in slightly inaccurate displays in coordinates of up to a 10 mm deviation under a reduction factor of 50. However, this slight inaccuracy is insignificant because 10 mm is relatively small in any normal sized environment. Also, if accuracy is needed, the zoom function can focus in on any area of a map. Though accurate work can be done by zooming in or lowering the reduction factor, the smaller magnified portion of the map will make editing work more difficult.

One key feature which has not been implemented in the GUI is an erase function. Currently, cleaning up erroneous data points and lines requires the use of Mapper 3. The erase function is fairly important as the GUI will eventually need to function independent of the Mapper 3 in modifying maps. The only remedy for erasing lines or goals that were inadvertently added is by reopening the map in the Mapper 3 and using its erase function. An alternative method is to open the map file in a text editor. This method is painstaking, however, as the user must know at least how many lines were added to the map file. The most recent lines added are always added to the beginning of the list by the mapping GUI. Currently, the only method by which added regions can be removed from a map is to edit the region file with a text editor. A region file begins with a line that identifies the number of regions in the map, followed by region definitions that are individually identified by a number representing the order in which each was created. Each region definition also identifies the number of points in the region and all the points that pertain to the region. In order to be able to delete an erroneous region, a user would have to know either the defining characteristics of a region as determined by number of points and point location, or when the region was added in reference to all other regions.

16. RECOMMENDATIONS

The most pressing focus for future research should be the creation of an erase function. Besides providing independence from the Mapper 3, the erase function is also crucial to deleting erroneous regions generated by the mapping GUI. These regions are not accessible from Mapper 3 and currently can only be deleted through a text editor. File verification safeguards are another feature that could be added to enhance the robustness of the current GUI. Currently, when opening a file, the open file dialog filter allows only files with .map extensions to be accessed. However, the program does not go further in verifying that the actual map file is valid. If the format is invalid or the map file has been tampered with, the program in its current state will just crash. Within the map file, the maximum and minimum points which determine the size of the map to be displayed need to be verified each time a map file is opened and corrected if they are inconsistent with data. Occasionally, the maximum and minimum listed in the map file are not correct. This results in some lines not being displayed in the GUI because the lines are out of the range of the display created for the map.

The current version of the GUI is clearly not complete in the sense that it does not accept temporal logic formula input nor does it partition the environment. Without these functions, the mapping GUI is not substantially more enhanced in features than the Mapper 3. The ability to define general regions is a step towards environment partitioning since this feature can be integrated into a future partitioning function. The current region creator can generate triangles and this will be useful since triangulation will likely be the partitioning method. Upon successful implementation of map partitioning and temporal logic formula processing in the GUI, further work in regards to path plan creation and implementation will require working with ARIA classes. ARIA classes can be used to implement the robot's motion in a continuous path that is developed by the GUI. Saphira and Colbert will not be used for two reasons. The first is that ActivMedia no longer supports Saphira and will not offer future updated versions. The second is that ARIA, being on a lower programming level, will provide more detailed control in implementation.

17. ACKNOWLEDGMENTS

This project has provided an invaluable opportunity to conduct research in the developing field of robotics. The author has gained a great amount of knowledge in robotics and programming from this project. In this sense, this experience has been truly rewarding. The author is also grateful to have been given the privilege to work on such a fascinating project.

I would like to thank Professor George Pappas for entrusting this project to me for the summer and providing valuable guidance and advice. I very much appreciate having been given the opportunity to work on this project. I would also like to thank Hadas Kress-Gazit, the graduate student responsible for the project, for working closely with me throughout the summer and offering her knowledge and support.

I would like to thank Professor Jan Van der Spiegel for allowing me to be part of the SUNFEST program. The program has provided me with valuable knowledge in regards to technical writing, presentations, conducting research, and engineering ethics. Finally, I would like to thank the National Science Foundation for its support of Research Experience for Undergraduate Programs like SUNFEST at the University of Pennsylvania.

18. REFERENCES

[1] G.E. Fainekos, H. Kress-Gazit, G. J. Pappas, Temporal Logic Motion Planning for Mobile Robots, IEEE Conference on Robotics and Automation, Barcelona, Spain, April 2005.

[2] Mobile Robots, ActivMedia Robotics, LLC, 2005, www.mobilerobots.com

[3] K.G. Konolidge, Saphira Software Manual, SRI International, Menlo Park, California, 2001.

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

EFFECTS OF CROSS-LINKING ON MECHANICAL FUNCTION IN THE DEGENERATE NUCLEUS PULPOSUS

NSF Summer Undergraduate Fellowship in Sensor Technologies
An Nguyen (Bioengineering) – University of Pennsylvania
Advisor: Dawn M. Elliott

ABSTRACT

Background Context: In the United States, the most prevalent cause of disability of workers under 45 years of age is low back pain. Past research suggests that there is a link between low back pain and intervertebral disc degeneration. One of the earliest known components of disc degeneration is a decrease in proteoglycan content in the nucleus pulposus, which in turn leads to changes in mechanical properties. In this study, the effects of proteoglycan content and collagen cross-linking on swelling behavior were investigated.

Study Design: Swelling pressure of the nucleus pulposus was measured in a confined compression experiment. The effects of two injectable agents were studied. The first was Chondroitinase ABC (ChABC), which is an enzyme that breaks down proteoglycan. ChABC was used to model disc degeneration. In addition, genipin -- a cross-linking agent that promotes the cross-linking of fibers in the disc's collagen network -- was also used.

Objectives: The main objective of this study was to determine the effects of cross-linking on the mechanical properties of degenerate nucleus pulposus. The first step was to determine baseline swelling pressure for control or normal sheep discs. The second was to determine the concentration of ChABC needed to mimic the degeneration found in human discs. The final step of the study was to investigate the potential of genipin required to restore the mechanical function in degenerate discs.

Results: The average swelling pressure measured for normal sheep discs was 0.21 ± 0.13 MPa. The optimal ChABC dose was 0.5 U with an average swelling pressure of 0.082 ± 0.04 MPa. The average swelling pressure for the 0.5% genipin group and 0.5% genipin + 0.5 U ChABC were 0.15 ± 0.07 MPa and 0.083 MPa, respectively.

Conclusion: Though findings from this study were inconclusive, they do not eliminate the potential of genipin as a treatment for disc degeneration. This warrants further investigation into effects of cross-linking using genipin and the mechanisms by which genipin increases cross-linking.

TABLE OF CONTENTS

1. *INTRODUCTION*

Low back pain is a widespread health problem. Between 70% and 85% of the population will experience back pain at some point in their lives [1-3]. In the United States, it is the most prevalent cause of activity limitation in people under 45 years of age [1, 4]. Among the reasons cited for people losing time at work and seeking medical care, low back pain is second only to the common cold [5]. Further, low back pain results in at least $50 billion dollars in annual health care costs in the United States [3, 6].

Though the causes of low back pain remain unclear, approximately 75% of low back pain cases are linked to intervertebral disc degeneration [6, 7]. The intervertebral disc consists of three parts, the endplate, the nucleus pulposus and the annulus fibrosus (Figure 1). The nucleus pulposus has a high proteoglycan content. Proteoglycan is a core protein linked to glycosaminoglycans (GAGs) [8]. GAGs are chains of polysaccharide, the majority of which have a negative charge [8]. This leads to an overall negative charge in the proteoglycan and attracts cations, such as Na^+. This then results in an intake of water in order to equilibrate the ion concentrations and forms a hydrated gel [9]. Hence, the nucleus pulposus has a high water content that ranges from 60% to 80% [9]. This hydrated gel is the structure that supports the axial loads in the disc. The annulus fibrosus, unlike the nucleus pulposus, is highly organized into a fibrous structure and consists mainly of collagen. In tension, it supports load due to the swelling behavior of the nucleus pulposus. In compression, the annulus fibrosus supports load due to bulging and deformation [10].

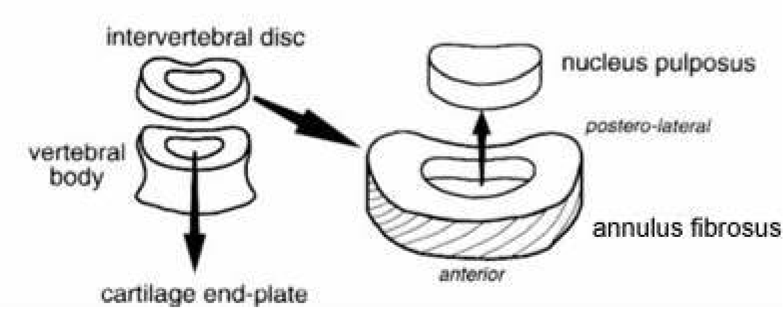


Figure 1. The structure of the intervertebral disc, which consists of the nucleus pulposus surrounded by the annulus fibrosus [11]

Disc degeneration consists of a complex process of interacting mechanisms that include chemical, biological and mechanical changes [12, 13]. The early stages of disc degeneration are characterized by a loss of proteoglycan and a decrease in water [12-16]. This leads to several changes in mechanical properties, including a significant decrease in swelling pressure [13, 17, 18].

Past research has looked at the swelling pressure of the nucleus pulposus in humans with the average pressure ranging from 0.05 to 3 MPa [11]. The swelling pressure is dependent on the loading conditions. It was also found that the nucleus pulposus is capable of swelling to more than 2 times its original volume [11].

Some studies have investigated the effects of cross-linking on the mechanical function of the nucleus pulposus. Preliminary studies done in the McKay Orthopaedic Research Laboratory looked at the effects of genipin, a natural cross-linking agent (Figure 2). In one of these studies, it was hypothesized that genipin could be used to model degeneration in sheep intervertebral discs. However, it was observed that genipin increased the swelling pressure of the nucleus pulposus. This was unexpected. It was previously believed that genipin would cause increased stiffness and decrease the swelling pressure due to the increased number of cross-links. These preliminary results suggested that genipin could be used as a treatment for disc degeneration.

Genipin is used in traditional Chinese medicine and is obtained from geniposide, its parent compound. Geniposide can be found in the fruits of *Gardenia jasminoides ELLIS* [19-21]. Previous research has focused on genipin effectiveness as a cross-linking agent in biological tissue fixation [19]. Genipin promotes cross-links by reacting with free amino groups, including lysine, hydroxylysine and arginine residues. A blue pigment forms at the site of these reactions as shown in Figure 4 [20, 21]. Past studies suggest that genipin forms intramolecular and intermolecular cross-links within collagen fibers in biological tissue (Figure 5) [20]. Collagen fibers are present in both the annulus fibrosus and the nucleus pulposus, with type II being the predominant collagen type found in the nucleus pulposus [22].

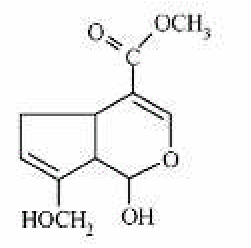


Figure 2. The chemical structure of genipin used in this study[21]

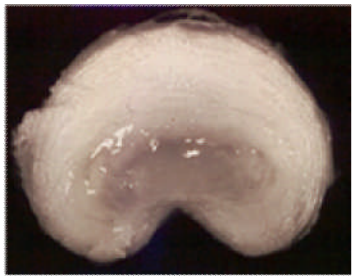


Figure 3. Isolated control intervertebral disc

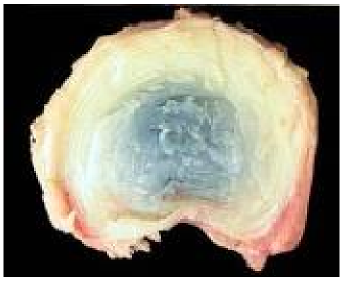


Figure 4. Isolated genipin-treated intervertebral disc

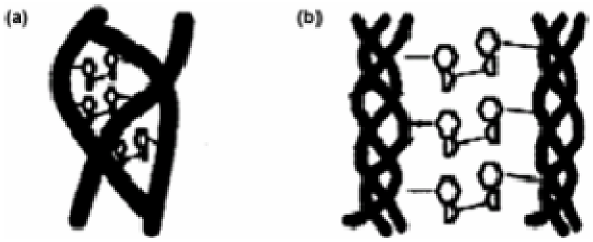


Figure 5. Illustration of (a) intramolecular and (b) intermolecular cross-links [23]

The objective of this study was to first to determine the baseline swelling pressure in the normal intervertebral disc and then to model degeneration in the nucleus pulposus using Chondroitinase ABC (ChABC), an enzyme that breaks down proteoglycan. The next objective was to determine the effect of increased cross-linking on the swelling pressure of the nucleus pulposus measured in confined compression using an in vitro model of degeneration. It was hypothesized that increasing cross-linking would increase the nucleus pulposus swelling pressure.

2. MATERIAL AND METHODS

2.1 Study Design

Studies have assessed the sheep as a model for human spine and have shown the sheep spine is a viable model due to the similarities in anatomy, size and material characteristics [24, 25]. Several studies have documented the similarities between the sheep and human lumbar discs in water content and proteoglycan content in both the nucleus pulposus and annulus fibrosus [22, 25]. However, due to the difference in mechanical loading in quadripedal animals, intervertebral discs in these types of animals do not undergo degeneration to the same extent that human discs do [26]. Chondroitinase ABC (ChABC) (Seikagaku America, USA) is an enzyme that depolymerizes proteoglycans and has been used in previous studies to induce degeneration in animals discs [27-30]. To determine the effects of cross-linking, ChABC and/or genipin was injected into sheep intervertebral discs.

A confined compression experiment was used to determine the mechanical effects of ChABC injection in non-degenerate nucleus pulposus tissue. Because little research of this kind has been done, this study first focused on determining the average swelling pressure of the sheep nucleus pulposus. The second part of the study focused on optimizing the concentration of ChABC in non-degenerative nucleus pulposus to best mimic the changes in degenerative nucleus pulposus mechanics. In particular, this study looked at the proportional changes in swelling pressure. In a previous study of human nucleus pulposus, the degenerate swelling pressure was 27% of the non-degenerate swelling pressure [31]. A dose finding study was performed using 4 different ChABC doses (0.1, 0.5, 1.0, and 2.5 U / 0.1 ml). Additional samples were tested to find the average swelling pressure of the optimal ChABC dose.

With these preliminary results, a preliminary genipin study was then performed. One disc was tested with 0.5% genipin and a second disc was tested with 0.5 U ChABC and 0.5% genipin. Finally, a preliminary study was performed using genipin. Four discs were treated with 0.5% genipin and four discs were treated with PBS. These eight discs were harvested from sheep spines with an approximate age of 6 months. All previous discs were harvested from sheep spines with an approximate age of 2.5 years.

2.2 Methods

Genipin (Challenge Bioproducts Inc., Taiwan) was prepared as a 0.5% concentration solution in 0.15 M PBS solution. The 0.5% genipin concentration was selected based on previous work [32, 33]. The combination solution of ChABC and genipin was prepared by using an aliquot of the previously prepared 0.5 U ChABC solution and adding genipin to reach the 0.5% concentration.

The cadaveric sheep discs were obtained in motion segments, where a motion segment is the intervertebral disc surrounded by the vertebrae on each side. The motion segments were thawed to room temperature and injected with 0.1 mL of the treatment solution using a 27-gauge needle. Following injection, the discs were incubated at 37° C for 18 hours in PBS and returned to a freezer at -20° C to prevent swelling of the nucleus pulposus. ChABC enzymatic breakdown of proteoglycan is optimal at 37^0C and requires between 1 to 5 hours of reaction time [34]. Genipin cross-linking requires 12 hours at 37^0C [23]. Hence, all discs were incubated at the same temperature of 37^0C for the same duration of 18 hours in order to avoid differences in mechanical testing due to variations in incubation. Discs were then dissected by removing the surrounding vertebrae (Figure 6). A sledge microtome (Model SM2400; Leica, Nussloch, Germany) with a freezing stage (Model BFS-30; Physitemp, Clifton, NF) was then used obtain a uniform thickness of 2.5 mm, measured using a non-contact micro laser sensor (LM10 Laser, Aromat, New providence NJ) (Figures 7 and 8). Average thickness was 2.50 ± 0.20 mm (n = 25). Samples were frozen at -20^0C until testing.

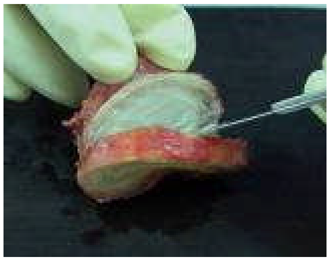


Figure 6. Intervertebral disc in the process of being dissected from the surrounding vertebrae

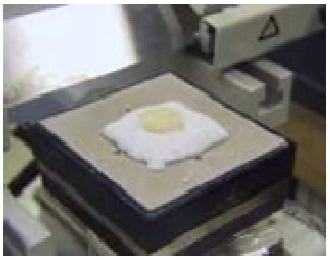


Figure 7. Intervertebral disc being microtomed to a uniform height of 2.5 mm

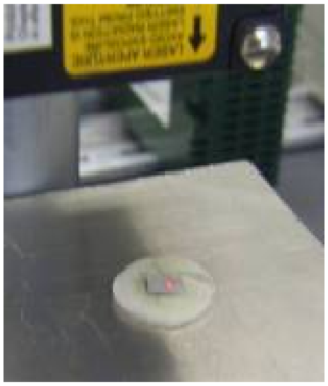


Figure 8. Micro laser sensor used to measure the thickness of samples

A 4.37 mm diameter circular punch was used to remove cylindrical plugs from the nucleus pulposus of each disc (Figure 9). These plugs were mechanically tested in a custom-built load and displacement controlled compression testing device (Figure 10). The device consists of a 10lb uniaxial load cell (Model 31; Honeywell Sensotec, Columbus, OH), LVDT (Model PR812-200, Macro Sensors, Pennsauken, NJ), linear stepper motor (Model 18512; Spectra Physics Oriel, Stratford, CT) and porous platen (diameter = 4.76, 50% porosity, 45-53 μm pore size). Displacement and load data were acquired through a LabView interface. The plug was placed into the chamber of the device (Figure 11). The porous platen was lowered at 10 μm/sec by the linear stepper motor until a contact load of 0.045 lbs was measured. The chamber was then filled with 0.15 M phosphate buffered solution (PBS). After a 5 minute wait period, a 1% compressive strain was applied followed by a 3 hour hold to measure the equilibrium swelling pressure. Swelling pressure was calculated as the reported load divided by the cross-sectional area of the device.

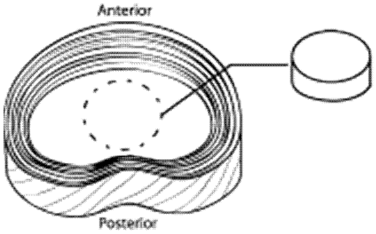


Figure 9. Schematic of a plug removed from the nucleus pulposus

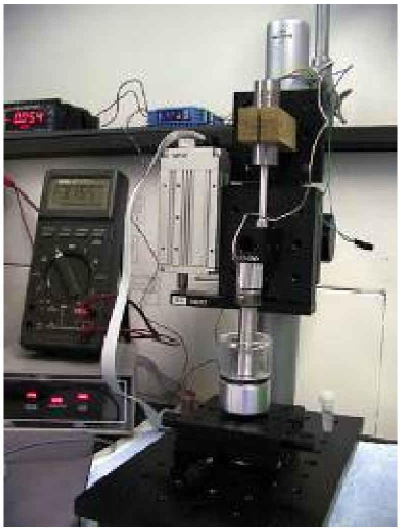


Figure 10. Custom-built confined compression testing device

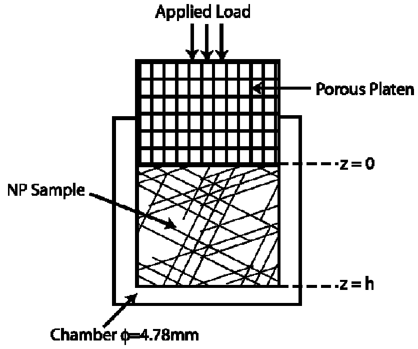


Figure 11. Schematic of chamber where the nucleus pulposus sample is placed with the porous platen applying the load from above

Data analysis consisted of two parts. First, a one-way ANOVA with Bonferroni post hoc test were used to determine if the average swelling pressure of each ChABC treatment group differed significantly between treatment groups. Second, a one-way ANOVA with the Bonferroni post hoc test were used to determine any significance between the control, ChABC, genipin and ChABC + genipin groups.

3. *RESULTS*

3.1 ChABC Dose Finding Study

The control discs with PBS injections were first evaluated. Three control discs were tested. An average swelling pressure of 0.35 ± 0.02 MPa was measured. A previous study measured the swelling pressure of non-degenerate and degenerate human nucleus pulposus in confined compression [31]. Degeneration led to about a 63% decrease in swelling pressure. A similar decrease was targeted in a ChABC dose finding study. The target swelling pressure for induced degeneration in the sheep nucleus pulposus was then 0.095 MPa. One disc was tested with 2.5 U ChABC with a swelling pressure of 0.051 MPa. A second disc was tested with 1.0 U ChABC (0.040 MPa). Four discs were tested at 0.5 U ChABC with an average swelling pressure of 0.082 ± 0.051 MPa. Seven discs were tested at 0.1 U ChABC with an average swelling pressure of 0.17 ± 0.11 MPa. Therefore, 0.5 U ChABC was selected as the optimal dose of ChABC to induce degeneration in the sheep nucleus pulposus (Figure 12). The 0.5 U ChABC group's average swelling pressure was significantly different from the control group ($p < 0.05$).

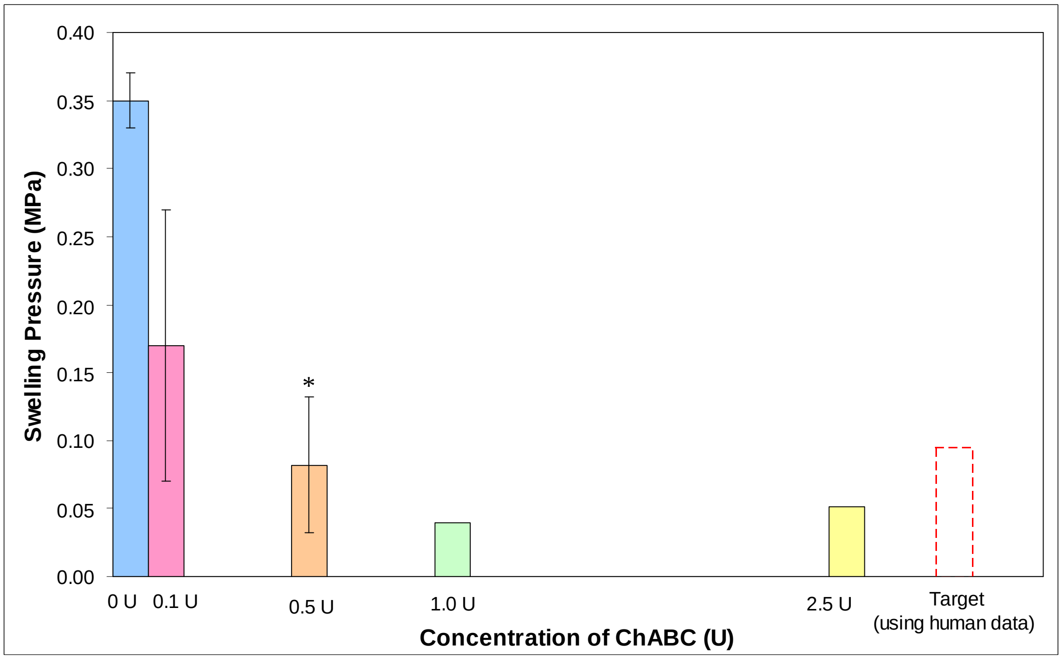


Figure 12. Confined compression results for ChABC dose finding study. Significant differences between the control and the 0.5 U ChABC group are indicated by *.

3.2 Preliminary Genipin Study

One disc treated with 0.5% genipin had a swelling pressure of 0.27 MPa. One disc treated with 0.5% genipin and 0.5 U ChABC had a swelling pressure of 0.083 MPa (Figure 13, Table 1). Representative swelling curves of the different treatments are shown in Figure 14. The control group's average swelling pressure was significantly different from both the ChABC and ChABC + genipin group. The ChABC was not significantly different from the ChABC + genipin group.

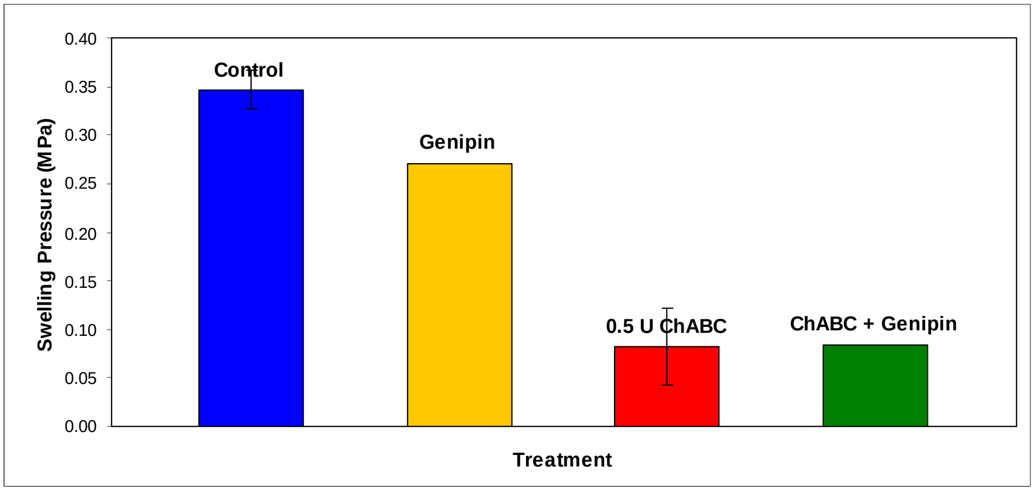


Figure 13. Comparison of swelling pressure across different treatment groups.

	Swelling Pressure (MPa)

Control *(n = 3)*	0.35 ± 0.02
0.5 U ChABC *(n = 4)*	0.082 ± 0.051
0.5% Genipin *(n = 1)*	0.27
0.5 U ChABC + 0.5% Genipin *(n = 1)*	0.083

Table 1. Summary of swelling pressure across four treatment groups

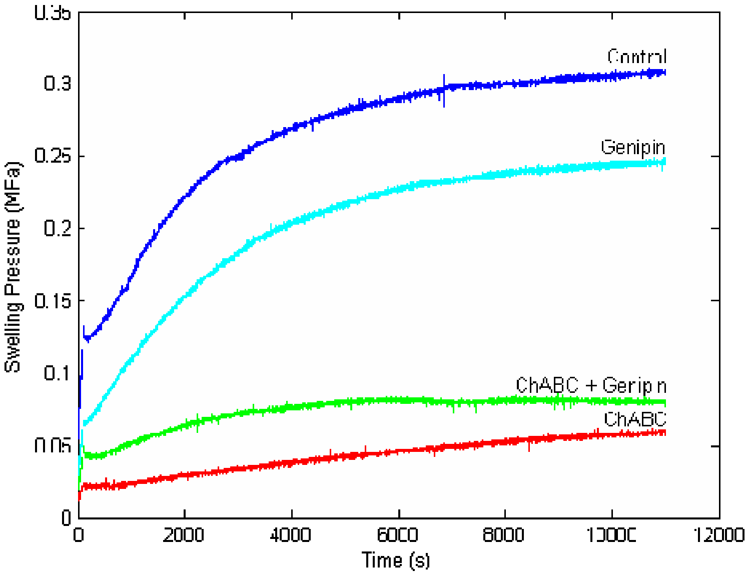


Figure 14. Representative swelling curves of 4 different groups

3.3 Additional Genipin Study

Four additional genipin treated discs and sham treated discs were tested. The average swelling pressure for the 0.5% genipin discs was 0.12 ± 0.01 MPa and the average for the control discs was 0.086 ± 0.02. There was a large difference between the swelling pressure measured for the preliminary genipin-treated disc and the swelling pressures measured for the genipin-treated discs of the young sheep (Figure 15). This difference was also observed in the control discs of the mature sheep and the control discs of the young sheep (Figure 16). A second plug was tested from the control group with mature sheep discs and was within the range of the preliminary data. Data collected using the young sheep discs showed that the average swelling pressure measured for the genipin group was significantly different from the average swelling pressure of the control group ($p < 0.05$).

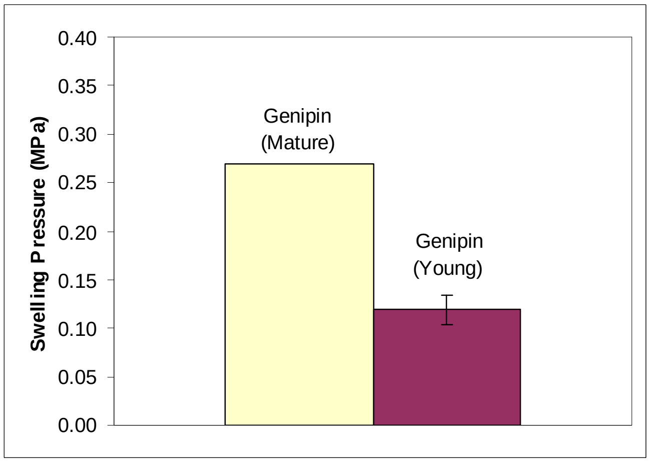


Figure 15. Comparison of average swelling pressure for 0.5% genipin groups

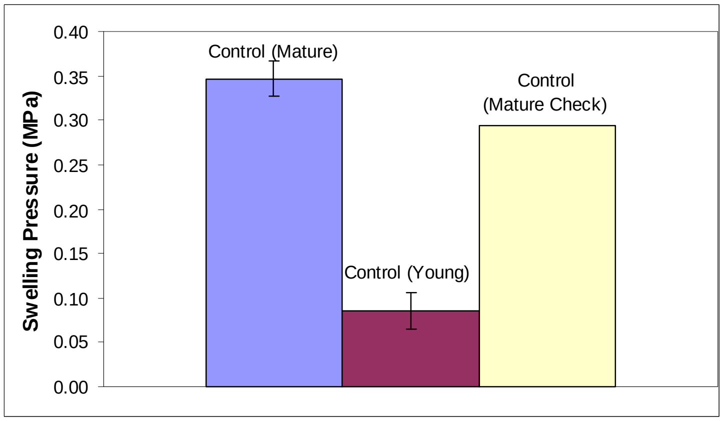


Figure 16. Comparison of average swelling pressure for control groups

4. DISCUSSION AND CONCLUSIONS

The average swelling pressure was measured in confined compression experiments in a ChABC dose finding study followed by a study of genipin cross-linking. Swelling pressure has been studied in human nucleus pulposus but not in sheep nucleus pulposus [31]. With the preliminary data on genipin, the effect of cross-linking on swelling pressure was inconclusive.

In the first part of this experiment, the ChABC dose finding study was performed. As expected, the swelling pressure decreased as the concentration of ChABC increased. With the average swelling pressure for the control group being 0.35 ± 0.02 MPa, the target swelling pressure using ChABC was 0.095 MPa. Looking at figure 12, the 2.5 U and 1.0 U ChABC doses led to average swelling pressures lower than the target swelling pressure. The target swelling pressure was within one standard deviation of both average swelling pressures measured for 0.1 U or 0.5 U ChABC groups. Due to the large variability in the 0.1 U ChABC group, 0.5 U ChABC was selected for the preliminary study with genipin.

The second part of the experiment involving genipin yielded inconclusive results. Using the mature discs, a genipin treated sample resulted in a decrease in swelling pressure when compared to the control group. Testing a sample with both genipin and ChABC yielded no difference with the ChABC only group. However, using the young discs, the genipin group had a significantly higher swelling pressure than the control group. It was expected that increasing cross-linking using genipin would increase the swelling pressure.

The variability observed in the 0.1 U and 0.5 U ChABC groups may be explained by the proteoglycan content of the cylindrical test samples measured. The changes in swelling pressure are highly correlated to the proteoglycan content. Further, delivery via injection may be a variable leading to non-uniform change in proteoglycan within each ChABC dosage group. Hence, biochemical analysis to measure proteoglycan content of the discs treated with ChABC will allow for a more accurate understanding of the changes in swelling pressure. Future work in this study includes using biochemistry to quantify the proteoglycan content.

The overall trend in the ChABC dose finding study was that as concentration of ChABC increased, the swelling pressure measured decreased. However, when comparing the 2.5 U ChABC treated disc with the 1.0 U ChABC treated disc, there was a slight decrease. This discrepancy can be explained by the variability of the data and a small sample size of only one sample per treatment. Further, data showed small changes in swelling pressure from 2.5 U to 0.5 U of ChABC. A large increase between the average swelling pressures of the 0.5 U and 0.1 U ChABC groups was observed. This suggests that there may be a threshold effect for ChABC treatment. Above this threshold, ChABC affects the tissue to such an extent that very low swelling pressures are measured. Data suggests that this threshold is between 0.5 U and 0.1 U ChABC. Further, biochemistry may reveal that there is no significant different in proteoglycan content in the 2.5 U and 1.0 U ChABC groups. It is possible that above the threshold, the ChABC has depolymerized the majority of proteoglycan in the sample; using a concentration of ChABC higher than the threshold would not increase depolymerization of proteoglycan.

The main objective of this study was to determine the effect of cross-linking on swelling pressure in sheep nucleus pulposus. Though the average swelling pressure measured for the genipin + ChABC group was not significantly different from the ChABC group, this may be explained by the small sample size and the variability in mechanical testing. Further, there was a significant difference between the genipin and control groups (Figures 15 and 16). A second plug from the preliminary discs was tested to check for a device malfunction. However, this second plug was within range of the preliminary data and the difference between the young and old sheep data was not believed to be the result of a device malfunction. The significant differences between the young and old sheep data may be attributed to the different sheep spines used. The approximate age of the sheep spines used in gathering the preliminary data was 2.5 years whereas the approximate age of the sheep spines used in the secondary genipin study was 6 months. Hence, the differences in skeletal maturation may have led to the difference in swelling pressures measured. This large difference between the young and mature discs was not expected because these animals do not experience disc degeneration.

A past study investigated collagen cross-linking in the intervertebral disc and correlated two different types of cross-linking with aging and degeneration [35]. A decrease in pyridinoline and an increase in pentosidine cross-links were found with disc aging. Pyridinoline cross-links are thought to be the most predominant cross-link in adult intervertebral discs (Figure 16) [36]. Pyridinoline cross-link levels increase from birth until skeletal maturation. Afterwards, pyridinoline cross-link levels tend to decrease slightly. Unlike pyridinoline, pentosidine cross-link levels have been observed to increase with age. Little is known about the specific type of cross-links that genipin promotes. It is possible that genipin increases the level of pyridinoline cross-links and thus counteracts the effects found with disc aging and degeneration.

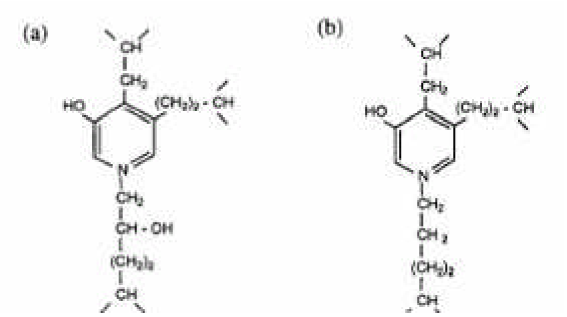


Figure 16. The structure of two different types of pyridinoline cross-links: (a) hydroxylysyl pyridinoline and (b) lysyl pyridinoline.[36]

In conclusion, data from this study showed that ChABC treatments lead to a degenerative change in the mechanical function of nucleus pulposus. The effects of genipin on this degenerative change were not fully determined because of the inconclusive data.

There were several limitations associated with this preliminary study. First, confined compression experiments are a more accurate measure of nucleus pulposus mechanics than unconfined compression experiments. *In situ* the nucleus pulposus is confined by the endplates and the annulus fibrosus but this is neither fully confined nor fully unconfined [31]. In addition, the swelling pressure was indirectly measured through a uniaxial load cell. A more direct method is to measure the interstitial fluid pressure, the pressure of the fluid within the nucleus pulposus [9]. The interstitial fluid is the fluid which the nucleus pulposus uses in pressurization under loading. Consequently it plays a key role in supporting loads in the intervertebral disc. Interstitial fluid pressure has previously been measured in articular cartilage [37].

The second study limitation involves using ChABC as model for degeneration. Disc degeneration is a complex process that is not limited to just proteoglycan depolymerization. Using ChABC, only one of the early changes in disc degeneration is modeled. The final study limitation involves genipin. Little is known about genipin's cross-linking mechanism and cross-linking was not quantified in this study. Genipin treatment may not have equally affected each sample. Correlating amount of cross-linking to swelling pressure would provide a better understanding of cross-linking and its effect on mechanical function. Moreover, a study has not been done to investigate any joint action between genipin and ChABC. There may have been interaction between the genipin and ChABC that further affected mechanical function.

In conclusion, results from this study, though inconclusive, indicate that further investigation into the effects of cross-linking is worthwhile. While various treatments for disc degeneration exist, the majority require invasive surgery. By understanding the effects of cross-linking on the mechanical function in the degenerate nucleus pulposus, an alternative, less invasive treatment for disc degeneration and low back pain may be developed. Finally, the ChABC dose finding data reported for sheep nucleus pulposus are especially valuable to other studies aiming to study degeneration using in vitro animal models.

5. RECOMMENDATIONS

This study was the first step in evaluating genipin's effectiveness as a treatment for disc degeneration. It would be useful to perform a dose response study of genipin both to gauge genipin's effect as it relates to concentration and to determine any interaction with ChABC. Genipin's cross-linking mechanism can be investigated using high-performance liquid chromatography and a detector [38]. Cross-linking can also be quantified using a ninhydrin assay [23]. In addition, a more in-depth dose response study of ChABC would be useful for further work in modeling disc degeneration. Finally, adding a pressure sensor to the current confined compression testing device would allow measurement of interstitial fluid pressure [37, 39, 40].

6. ACKNOWLEDGMENTS

I would like to thank Dr. Dawn Elliott for allowing me this research and learning opportunity for these past ten weeks. In addition to Dr. Elliott, I would like to thank Wade Johannessen for his guidance and support throughout the summer. Finally, I would like to thank Dr. Jan Van der Spiegel, the SUNFEST program, and the National Science Foundation for supporting this research.

7. REFERENCES

1. Andersson, G.B., *Epidemiological features of chronic low-back pain.* Lancet, 1999. **354**(9178): p. 581-5.
2. Brant-Zawadzki, M.N., et al., *Low back pain.* Radiology, 2000. **217**(2): p. 321-30.
3. Errico, T.J., *Lumbar disc arthroplasty.* Clin Orthop Relat Res, 2005(435): p. 106-17.
4. Boelen, E.J., et al., *Intrinsically radiopaque hydrogels for nucleus pulposus replacement.* Biomaterials, 2005.
5. Sagi, H.C., Q.B. Bao, and H.A. Yuan, *Nuclear replacement strategies.* Orthop Clin North Am, 2003. **34**(2): p. 263-7.
6. Thomas, J., A. Lowman, and M. Marcolongo, *Novel associated hydrogels for nucleus pulposus replacement.* J Biomed Mater Res A, 2003. **67**(4): p. 1329-37.
7. Benneker, L.M., et al., *2004 Young Investigator Award Winner: vertebral endplate marrow contact channel occlusions and intervertebral disc degeneration.* Spine, 2005. **30**(2): p. 167-73.
8. Alberts, B., et al., *Molecular Biology of the Cell*. 4 ed. 2002, New York: Garland Science. 1616.
9. Allen, M.J., et al., *Preclinical evaluation of a poly (vinyl alcohol) hydrogel implant as a replacement for the nucleus pulposus.* Spine, 2004. **29**(5): p. 515-23.
10. Setton, L.A. and J. Chen, *Cell mechanics and mechanobiology in the intervertebral disc.* Spine, 2004. **29**(23): p. 2710-23.
11. Iatridis, J.C., et al., *Is the nucleus pulposus a solid or a fluid? Mechanical behaviors of the nucleus pulposus of the human intervertebral disc.* Spine, 1996. **21**(10): p. 1174-84.
12. Lotz, J.C., *Animal models of intervertebral disc degeneration: lessons learned.* Spine, 2004. **29**(23): p. 2742-50.
13. Johannessen, W. and D.M. Elliott, *Effects of Degeneration on the Biphasic Material Properties of Human Nucleus Pulposus in Confined Compression.* Spine, in press.
14. Buckwalter, J.A., *Aging and degeneration of the human intervertebral disc.* Spine, 1995. **20**(11): p. 1307-14.
15. Preradovic, A., et al., *Quantitation of collagen I, collagen II and aggrecan mRNA and expression of the corresponding proteins in human nucleus pulposus cells in monolayer cultures.* Cell Tissue Res, 2005.
16. Antoniou, J., et al., *Apparent diffusion coefficient of intervertebral discs related to matrix composition and integrity.* Magn Reson Imaging, 2004. **22**(7): p. 963-72.
17. Urban, J.P. and J.F. McMullin, *Swelling pressure of the inervertebral disc: influence of proteoglycan and collagen contents.* Biorheology, 1985. **22**(2): p. 145-57.
18. Urban, J.P. and J.F. McMullin, *Swelling pressure of the lumbar intervertebral discs: influence of age, spinal level, composition, and degeneration.* Spine, 1988. **13**(2): p. 179-87.
19. Sung, H.W., et al., *Crosslinking of biological tissues using genipin and/or carbodiimide.* J Biomed Mater Res A, 2003. **64**(3): p. 427-38.
20. Sung, H.W., et al., *Mechanical properties of a porcine aortic valve fixed with a naturally occurring crosslinking agent.* Biomaterials, 1999. **20**(19): p. 1759-72.
21. Jin, J., M. Song, and D.J. Hourston, *Novel chitosan-based films cross-linked by genipin with improved physical properties.* Biomacromolecules, 2004. **5**(1): p. 162-8.
22. Demers, C.N., J. Antoniou, and F. Mwale, *Value and limitations of using the bovine tail as a model for the human lumbar spine.* Spine, 2004. **29**(24): p. 2793-9.
23. Sung, H.W., et al., *Fixation of biological tissues with a naturally occurring crosslinking agent: fixation rate and effects of pH, temperature, and initial fixative concentration.* J Biomed Mater Res, 2000. **52**(1): p. 77-87.
24. Wilke, H.J., A. Kettler, and L.E. Claes, *Are sheep spines a valid biomechanical model for human spines?* Spine, 1997. **22**(20): p. 2365-74.
25. Meakin, J.R. and D.W. Hukins, *Effect of removing the nucleus pulposus on the deformation of the annulus fibrosus during compression of the intervertebral disc.* J Biomech, 2000. **33**(5): p. 575-80.
26. Bao, Q.B. and H.A. Yuan, *New technologies in spine: nucleus replacement.* Spine, 2002. **27**(11): p. 1245-7.

27. Eurell, J.A., M.D. Brown, and M. Ramos, *The effects of chondroitinase ABC on the rabbit intervertebral disc. A roentgenographic and histologic study.* Clin Orthop Relat Res, 1990(256): p. 238-43.
28. Henderson, N., V. Stanescu, and J. Cauchoix, *Nucleolysis of the rabbit intervertebral disc using chondroitinase ABC.* Spine, 1991. **16**(2): p. 203-8.
29. Ando, T., et al., *Effects of chondroitinase ABC on degenerative intervertebral discs.* Clin Orthop Relat Res, 1995(318): p. 214-21.
30. Sasaki, M., et al., *Effects of chondroitinase ABC on intradiscal pressure in sheep: an in vivo study.* Spine, 2001. **26**(5): p. 463-8.
31. Johannessen, W. and D.M. Elliott, *Effects of Degeneration on the Biphasic Material Properties of Human Nucleus Pulposus in Confined Compression.* Spine, 2005.
32. Liu, B.S., et al., *In vitro evaluation of degradation and cytotoxicity of a novel composite as a bone substitute.* J Biomed Mater Res A, 2003. **67**(4): p. 1163-9.
33. Yao, C.H., et al., *Biocompatibility and biodegradation of a bone composite containing tricalcium phosphate and genipin crosslinked gelatin.* J Biomed Mater Res A, 2004. **69**(4): p. 709-17.
34. Hiyama, K. and S. Okada, *Crystallization and some properties of chondroitinase from Arthrobacter aurescens.* J Biol Chem, 1975. **250**(5): p. 1824-8.
35. Pokharna, H.K. and F.M. Phillips, *Collagen crosslinks in human lumbar intervertebral disc aging.* Spine, 1998. **23**(15): p. 1645-8.
36. Knott, L. and A.J. Bailey, *Collagen cross-links in mineralizing tissues: a review of their chemistry, function, and clinical relevance.* Bone, 1998. **22**(3): p. 181-7.
37. Basalo, I.M., et al., *Cartilage interstitial fluid load support in unconfined compression following enzymatic digestion.* J Biomech Eng, 2004. **126**(6): p. 779-86.
38. Sims, T.J., N.C. Avery, and A.J. Bailey, *Quantitative determination of collagen crosslinks.* Methods Mol Biol, 2000. **139**: p. 11-26.
39. Soltz, M.A. and G.A. Ateshian, *Experimental verification and theoretical prediction of cartilage interstitial fluid pressurization at an impermeable contact interface in confined compression.* J Biomech, 1998. **31**(10): p. 927-34.
40. Soltz, M.A. and G.A. Ateshian, *Interstitial fluid pressurization during confined compression cyclical loading of articular cartilage.* Ann Biomed Eng, 2000. **28**(2): p. 150-9.

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

NSF Summer Undergraduate Fellowship in Sensor Technologies
Olga Paley (Chemical Engineering) - University of California, Berkeley
Advisor: Professor Dawn Elliott

Applying Immunohistochemistry & Reverse Transcription PCR to Intervertebral Disc Degeneration in an Animal Model

ABSTRACT

STUDY DESIGN: Devise a set of protocols for immunohistochemical (IHC) and gene expression analysis that will permit the measurement of changes that occur during degeneration in the lumbar intervertebral disc of a rat model. OBJECTIVES: Quantify how the composition of the disc changes with degeneration in the animal model and prove that these changes correspond to those within a degenerate human disc, supporting the rat as a valid model of spinal disc degeneration. In the future, use the model to determine mechanisms for the degenerative process. SUMMARY OF BACKGROUND DATA: Rodents have been widely used as models to study disc degeneration[1,2,3,4]. The chemical changes within the human disc have also been extensively studied by various methods[5,6,7,8,9]. However, the research to understand the same chemical changes in the rat has been limited. METHODS: IHC staining of the lumbar intervertebral discs of healthy adult rats (as a basis) began with fixing thin slices of paraffin-embedded tissue onto glass slides. The samples were then treated with an antibody specific to the chosen antigen that was then coupled to a biotin-labeled secondary antibody. Finally, the antigens were localized using colorimetric staining. Reverse transcription-polymerase chain reaction (RT-PCR) was also employed to understand gene expression changes within the disc. An RNA extraction was first performed, then RT-PCR to create first-strand cDNA, and finally, standard PCR to amplify a desired gene. RESULTS: The nucleus pulposus on the slide experienced some degradation during the fixing and staining procedures. The collagen I staining proved the most problematic,

with staining occurring within the inner annulus and nucleus, a result contradictory with literature[7]. Collagen II primarily stained within the inner annulus, as hypothesized. Finally, aggrecan stained as expected, but there was evidence of significant background effects. The gene expression work produced results for aggrecan, collagen I, and fibronectin, proving that the protocol employed for the RNA extraction and RT-PCR was effective. CONCLUSIONS: A number of techniques were established for IHC and RT-PCR, but further development is needed. There is significant evidence that the staining was problematic due to background effects. Thus, the continuation of the study will focus on perfecting the fixation and sectioning procedure and doing regular histological staining to insure a sample which does not disintegrate and stains evenly. The preliminary gene expression work's success leads to the conclusion that more specific results may be achieved by splitting the nucleus from the annulus. These results can be quantified and compared by normalizing them to a standard gene.

Table of Contents

1. INTRODUCTION

The study of the intervertebral disc has been motivated by the high incidence of back pain. Over 100 billion dollars are spent annually in connection to back pain. This includes health care, disability payments, and other related costs. On the United States alone, over 5 million people are permanently disabled by back pain[10,11]. Previous research has associated back pain with the degeneration of the intervertebral disc, and although much work has gone into understanding that degeneration, little is known about the mechanisms by which the process occurs. By improving the understanding of how degeneration takes place, research can move forward to develop improved treatments to alleviate and cure many forms of back pain that currently debilitate people.

2. GOAL OF STUDY AND BACKGROUND

2.1 Goal of Study

The goal of the work conducted was to devise a set of protocols for immunohistochemical (IHC) and gene expression analysis which would permit the study of changes that occur during degeneration in the lumbar intervertebral disc of a rat model. These protocols were then to be tested for the quality of results obtained. This work initially focused on healthy discs in order to acquire a set of baseline data against which further tests could be compared.

2.2 Intervertebral Disc Background

2.2.1 Anatomy

The intervertebral disc is a soft tissue which exists between the vertebral bodies of the spine. The disc permits motion of the spine and dissipates energy, acting as a cushion and a support. As Fig. 1 shows, it is composed of two major sections. The first is the annulus fibrosis, and the second is the nucleus pulposus. Located between these two regions is the inner annulus, which exhibits some properties of both sections. Finally, above and below the disc there are endplates.

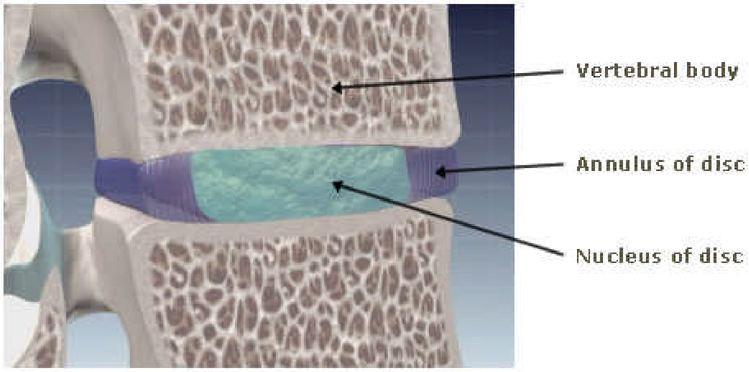


Fig. 1: Anatomy of the intervertebral disc.

The nucleus pulposus is a gel-like structure, while the annulus fibrosis is fibrous, as its name suggests. These fibers are arranged in an organized fashion within the outer annulus, but are more disorganized within the inner. Overall, the concentration of cells within the disc is low, and thus the extracellular matrix within the disc is believed to control disc function[12].

2.2.2 Chemical Composition

The chemical composition of this extracellular matrix is highly significant. Several important proteins which exist in the matrix have been examined within this study. These include: collagens type I and II, the proteoglycan aggrecan, and the fibrous protein fibronectin. Both types of collagens form fibers, although collagen I is known to be primarily present in bone and the outer annulus while collagen two is present in the inner annulus and the nucleus. Aggrecan is a proteoglycan and is known to exist in high concentration in the nucleus, the inner annulus, and at the endplates[12]. It is a relatively short protein which is negatively charged. It binds to long chains of carbon and creates a negatively charged network. This then leads to osmosis of water into the disc, which gives the disc its gel-like properties[12]. Finally, fibronectin is a protein which is associated with tissue repair and can be expected to be found both within the annulus and the nucleus[13]. Samples were stained for collagen II, but its gene expression was not analyzed. Also, the fibronectin genes were looked at with PCR, but the protein itself was not stained immunohistochemically. There are also a number of enzymes and inhibitors that play major roles within the disc and warrant future study.

2.2.3 Degeneration

The human intervertebral disc is known to undergo irreversible degeneration with age. Almost immediately after birth, the disc begins to alter. Degeneration changes the mechanical properties of the disc, its composition, and its structure. Disc degeneration has also been linked to pain[14], which is the reason it has been studied so extensively. Although much research has been done in the field, the mechanisms by which early degeneration occurs remain unknown[14,15].

It is known, however, that a drop in proteoglycan content occurs early in degeneration. There is also evidence that aggrecan is increasingly downregulated as degeneration proceeds[16,17], while collagen localization is changed[7,18] along with the regulation of significant enzymes and their inhibitors[6,9,18]. Fibronectin has also been to shown to be upregulated with increasing degeneration[13].

Due to the reasons discussed earlier, understanding these changes in detail is important. The work discussed here attempts to lay a foundation for studies which will enrich the current standard of knowledge and ultimately lead to the understanding of the mechanisms of degeneration. This may ultimately allow for the prevention, cessation, or even reversal of the process.

2.3 Staining Background

Histologic staining is generally very straightforward. This type of staining utilizes reagents that mark certain types of tissues. Cell nuclei may stain, or for example, collagen. The reason that histological staining is not sufficient to localize a desired protein is because it is not specific enough. Although histologic staining can mark collagen or proteoglycan, the stain cannot distinguish between different types of these proteins. Thus, immunohistochemistry is required for a more detailed picture of the sample. Immunohistochemical, staining, however, is a more complex process. The goal of the stain is to identify and locate some desired entity, such as the antigen, within a sample. IHC does this by labeling the desired entity with a probe. This probe must bind to the antigen in question irreversibly, avoid binding to anything else, and be detectable. Antibodies are very specific proteins, and will bind irreversibly and highly selectively to their target. Thus, antibodies have been employed for IHC. Primary antibodies are typically raised in an animal that is a different species from the sample being stained.

The sample is exposed to the primary antibody which binds to the antigen. This combined structure must be visualized. It is possible to attach some marker to the primary antibody before exposing the sample to it, but since these are so specific, it would not be practical to do so. Thus, the primary antibody-antigen complex is further exposed to a secondary antibody which is specific against the antibodies in the animal used to produce the primary antibody. This secondary antibody can be much less specific, as it only needs to recognize tissue from the primary antibody animal. It is this secondary antibody which comes conjugated to some marker which can be easily visualized by light or fluorescence microscopy.

2.4 Gene Expression Analysis and RT-PCR Background

The general procedure for RT-PCR can be split into four main sections: 1) extracting genetic information from a sample in the for of RNA, 2) converting this information into first-strand cDNA, 3) amplifying a desired gene to make a billion copies of that gene’s cDNA, and 4) finally viewing the product of the previous four steps by using gel electrophoresis. Gel electrophoresis separates DNA fragments by size. The larger segments (those with a higher molecular weight) do not travel as far down the agarose gel as smaller fragments when a potential is induced over the length of the gel. This allows for the identification of the size, and thus the identity, of DNA present in the sample.

2.5 Previous Research Done on Animal Models and Disc Degeneration

Rodents have been widely used as models to study mechanical effects of disc degeneration. The extracellular chemical changes within the human disc have also been extensively studied by various methods. However, the research to understand the same chemical changes in an animal model has been limited. The obstacle has been the difficulty in developing a good enough animal model of degeneration[1]. Animals, specifically rats, do not experience degeneration, and it must therefore be induced. The rat is a desirable model for a number of reasons: it is cheaper, easier to obtain, and most importantly, as any good model must be, the rat model is easier to control than a set of human samples. In order for this to be true, however, the degeneration induced must mirror the degeneration which naturally occurs within the human.

Earlier studies have used chemonucleolysis to model degeneration in ovine and canine samples[2], but were expensive, degraded the nucleus aggressively, and did not allow for the extraction of genetic information from the discs. Other work has shown that puncture models demonstrate degeneration, but this mechanism does not follow the natural progression of degeneration.

Our laboratory uses a moderate chemonucleolysis model in the rat which resembles human disc degeneration[20,21]. This work has aimed at developing the methods by which this model may be evaluated and applied to understanding the mechanisms and changes of degeneration.

3. METHODS

3.1 Sample Preparation

The animals used to obtain specimens were adult Sprague Dawley rats. These were sacrificed, and the spines were surgically removed. The intervertebral discs were appropriately separated into individual samples. The bone above and below each disc was partially left intact for the samples to be sectioned and stained, while those samples which would undergo the gene expression study were removed completely.

Those discs which were to be stained were fixed and then decalcified for 54 hours in a 10% formalin and formic acid solution. Some excess bone was then trimmed, and the sampled were imbedded in paraffin. After cooling, sagittal and axial sections of these were made on a microtome at a thickness setting of approximately 7 µm. The staining study primarily utilized the sagittal sections, although axial cross-sections should also be stained at a later point. These sections were then placed on glass slides and heated to assure fixation.

The discs used for the gene expression work were, as stated earlier, completely separated from bone. They were then placed into RNAlater (Ambion; Austin, TX) to prevent degradation.

3.2 Sample Staining

The sections were stained in two ways. First, two different histologic stains were performed. Then, the sections underwent IHC staining for several proteins of interest.

3.2.1 Histological

Hematoxylin & Eosin: The samples were first immersed in CitriSolv (Fisher Scientific; Pittsburgh, PA) to dissolve the paraffin, and then rehydrated using a series of baths with decreasing alcohol concentration, from 100% to 0%. These were then immersed in a hematoxylin solution (Hematoxylin Gill no. 2, Sigma; St. Louis, MO) for 10-15 minutes, and then washed in tap water. The next wash was with acid alcohol for 20 seconds, followed with a tap water wash, a wash in Scott's buffer, and another tap water wash. They were then immersed in the eosin solution (Eosin Y alcoholic, Sigma; St. Louis, MO) for 1-2.5 minutes, again water washed, and finally dehydrated by a reverse version of the rehydration process. Finally, the samples were dehydrated in CitriSolv, set under an acrylic sealant, dried, and viewed under three different magnifications on a light microscope. All stained samples were viewed at: 2.5 X, 10X, and 40X magnifications.

Alcian Blue & Picrosirius Red: The samples were rehydrated by the same procedure as described above. They were then stained in an acidic alcian blue solution (Alcian Blue 8GX, Sigma; St. Louis, MO) for

30 minutes, washed with tap water, stained with picrosirius red (Sirius Red, Aldrich; St. Louis, MO) (Picric Acid, Fisher Scientific; Hampton, NH) for 45 minutes, washed with acidified water, dehydrated, and sealed. The samples were imaged by the method described above.

3.2.2 Immunohistochemical

Step	Purpose
1. De-wax	Remove paraffin from tissue to allow rehydration
2. Rehydrate Specimen	Condition the fixed specimen to aqueous reagent penetration
3. Antigen Retrieval	Improve epitope exposure from fixed tissues
4. Endogenous Enzyme Block	Inhibit any endogenous enzyme activity that could non-specifically develop a colored reaction
5. Protein Block	Limit any non-specific protein binding to specimen
6. Primary Antibody	Specific binding to antigen
7. Secondary Antibody	Amplify the antibody-antigen reaction
8. Enzyme Complex	Label the immune complex with an enzyme
9. Color Development	Visualize antigen expression by a colored precipitate
10. Dehydrate Specimen	Set stain and prepare for storage
11. Seal	Set sample permanently

Table 1: General steps for immunohistochemical staining.

Each immunohistochemical stain performed followed the same general steps, which are presented in Table 1.

The three proteins studied in this way were collagen I, collagen II, and the proteoglycan aggrecan. The primary antibodies used were as follows: a 1:100 dilution of a polyclonal rabbit antibody against collagen I (Chemicon; Temecula, CA), a 1:4 dilution of a monoclonal mouse antibody for collagen II (DSHB; Iowa City, IA), and a 1:100 dilution of a polyclonal rabbit antibody against aggrecan (Abcam; Cambridge, UK). Control slides were included in every run. These were treated with everything except the primary antibody, and were instead exposed to non-immune horse serum. The secondary, biotin-labeled, antibodies used were 1:100 rat anti-mouse monoclonal and 1:200 goat anti-rabbit monoclonal (BD Pharmigen; San Diego, CA), choice depending on the primary antibody used. The enzyme complex used was a Vestastain ABC and the colored precipitate agent was DAB kit (Vector Laboratories; Burlingame, CA). Once sealed, the slides were viewed in the same way as those in basic histology.

3.3 Understanding Gene Expression

All materials, containers, and solutions used were assured to be RNAse and DNAse free. The samples were removed from RNAlater, and homogenized. (Two complete discs were included in each tube.) This was done by adding Trizol (Invitrogen; Carlsbad, CA) and grinding the sample with a mortar and pestle. Once the sample had completely disintegrated, the solution was homogenized with a power homogenizer for 1 minute. The sample was then allowed to separate before being homogenized again. This was repeated until complete blending was achieved. The sample was then centrifuged to remove RNA from the heavier components of the solution. The supernatant was removed and chloroform was added, then centrifuged. This formed several layers, with the top layer containing the RNA. This layer was drawn off and transferred to another container. Finally, isopropyl alcohol was added, the tube centrifuged, and the supernatant drawn off. Left in the tube was a pellet of RNA. This was then dissolved in DEPC water and stored at -80º C.

The concentration of RNA in the stored solution was determined by a spectrophotometer. RT-PCR was performed following the guidelines presented in the Background section. Everything was kept RNAse and DNAse free. Enough RNA solution was added to assure 5 micrograms of RNA in each PCR tube. Water and OligoDT primer were added. The samples were exposed to 70°C for 10 minutes. To this, buffer, DTT, dNTPs, and Superscript II were added and the mixture was incubated at 42°C for fifty minutes and then at 70°C for fifteen minutes. This was now a solution of first strand cDNA.

This cDNA was then used in PCR to amplify a gene of interest. The cDNA was combined with buffer, dNTPs, forward primer, reverse primer, water, and Taq polymerase. PCR was performed for three genes: aggrecan, fibronectin, and collagen I. Each required a different annealing temperature and number of cycles: aggrecan required 56° C and 36 cycles, fibronectin required 60° C and 32 cycles, and collagen I required 60° C and 32 cycles.

Finally, gel electrophoresis was performed to test for PCR products.

4. RESULTS AND DISCUSSION

4.1 Section Staining

4.1.1 Histological

Basic histology was performed on the healthy disc samples as described within the methods section. The staining by Hematoxylin and Eosin (H&E) produced expected results. The hematoxylin stained nuclei a dark blue while the eosin stained the background tissues red. Fig. 2 presents this staining at three different magnifications.

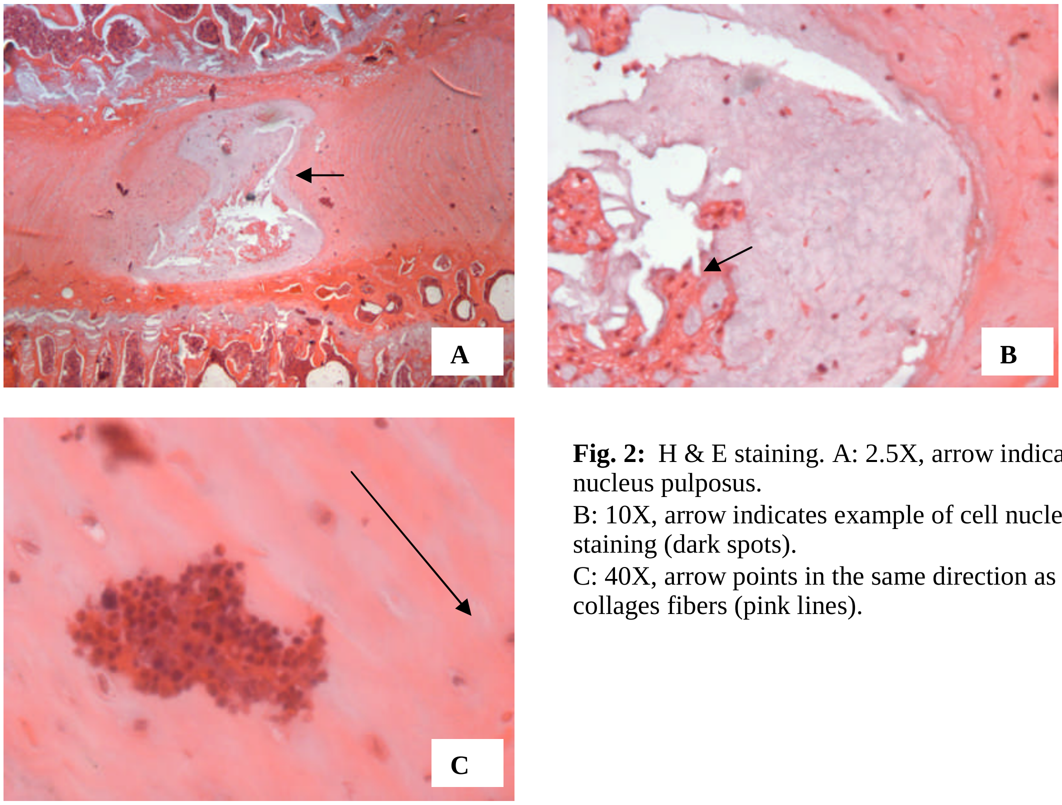


Fig. 2: H & E staining. A: 2.5X, arrow indicates nucleus pulposus.
B: 10X, arrow indicates example of cell nucleus staining (dark spots).
C: 40X, arrow points in the same direction as collages fibers (pink lines).

The make-up of the disc is clearly visible in the first image. The porous sections above and below the disc are bone, while the disc itself is composed of the nucleus, which stained lightly, and the annulus, which stained more heavily. The second image, taken within the nucleus, shows more detail. Specifically, the cell nuclei are visible as dark dots. They are scattered throughout the disc, at a relatively low concentration, as expected. The third image reveals the fibrous structure of the annulus fibrosis. The pink lines are these fibers, believed primarily to be made up of collagen I in the outer annulus (shown). Located between the outer annulus and nucleus is a less organized inner annulus, which is somewhat difficult to see in these images but is known to contain a high concentration of collagen II.

An Alcian Blue & Picrosirius Red stain was also performed, again, as described in the methods section. This stain labels proteoglycan blue, and collagen red. This protocol also produced expected results (see Fig. 3).

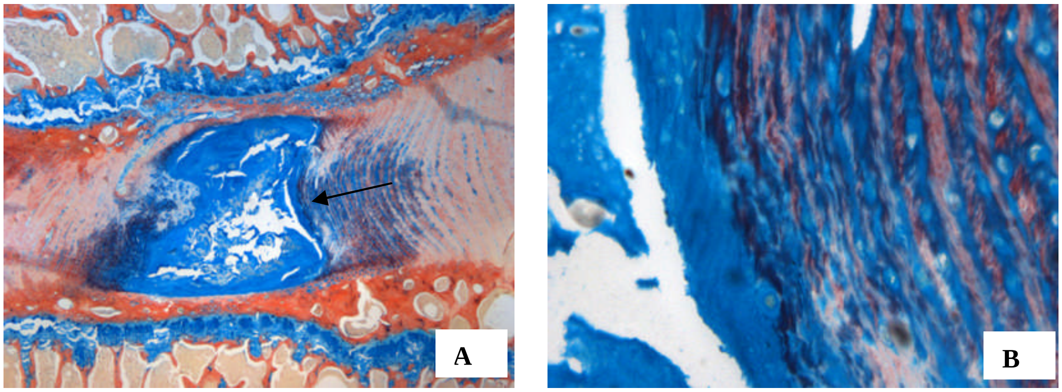


Fig. 3: Alcian Blue & Picrosirius Red staining. A: 2.5X B: 10X arrow in A points to location of magnification.

As expected, the nucleus pulposus stained a solid blue, supporting the fact that healthy discs are primarily made up of proteoglycan. A significant amount of proteoglycan is also visible at the endplates, along with some collagen. The annulus appears to have stained correctly as well. The inner annulus has a high concentration of proteoglycan and some collagen, in a somewhat disorganized structure. The structure becomes more organized further away from the nucleus, and much of the blue staining is replaced by red. This transition in color and structure supports the expected transition from the inner to outer annulus. This transition is accompanied by a more organized structure and an increase in the ratio of collagen to proteoglycan.

Thus, the basic histological stains were overall successful. They could, however, be improved. Due to some loss of detail and solidity of color, it is possible that these tissues were overstained (see also aggrecan IHC stain). The proposed solution to this problem is presented below. It is likely that the tissue sample was too thick. The nucleus pulposus also experienced disintegration on the slide. This problem occurs due to the original high concentration of water in the nucleus, which results in a relative weakness and fragility of this tissue.

4.1.2 Immunohistochemical

Immunohistochemical staining was performed as presented in the methods section for collagen I, collagen II, and the proteoglycan aggrecan. Fig. 4, 5, and 6 present the results.

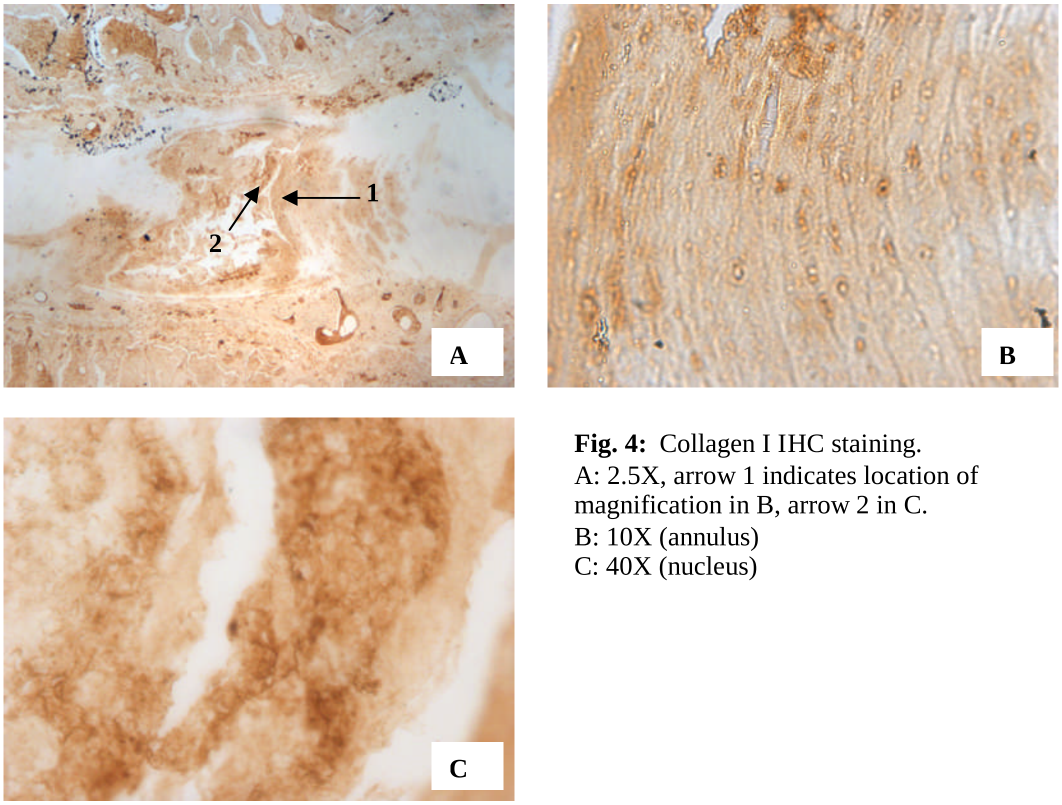
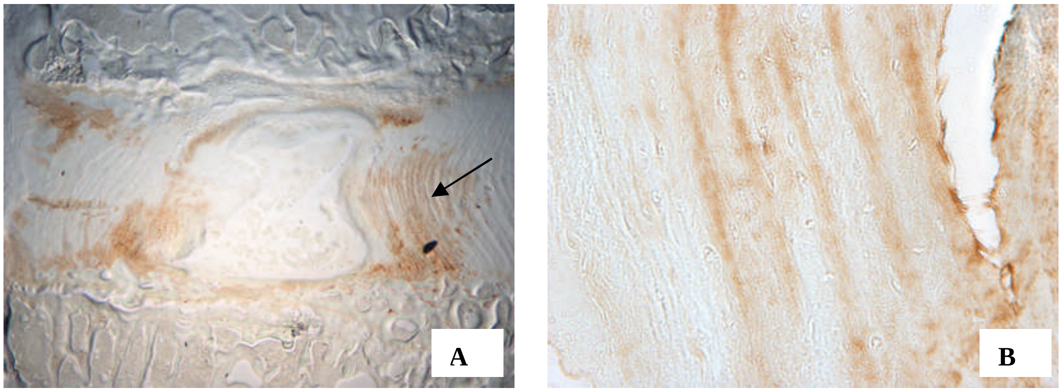


Fig. 4: Collagen I IHC staining.
A: 2.5X, arrow 1 indicates location of magnification in B, arrow 2 in C.
B: 10X (annulus)
C: 40X (nucleus)

Collagen I staining proved to be the most problematic. The bone, endplates, and some of the outer annulus stained for this protein, as expected. The nucleus and inner annulus, however, also stained, in some regions significantly more than the outer annulus. By comparison to previous work done in both humans and animals, these results were deemed incorrect. Although the second image in Fig. 4 contains the expected fibrous structure on the right (outer annulus region) the amount of staining within the more irregular inner annulus (left) is comparable. The third image is a high magnification of the nucleus, which reveals highly irregular staining.

The presence of staining in the inner annulus and nuclear region, and the overall brightness of the stain in all locations suggests the presence of background staining. Published data suggests that staining for collagen I in the disc should be very faint overall. The material on the slide is therefore believed to be trapping the reagents and causing a false positive colorimetric response. There is also a possibility of some cross reactivity within the disc of the antibodies and something other than the target antigen.



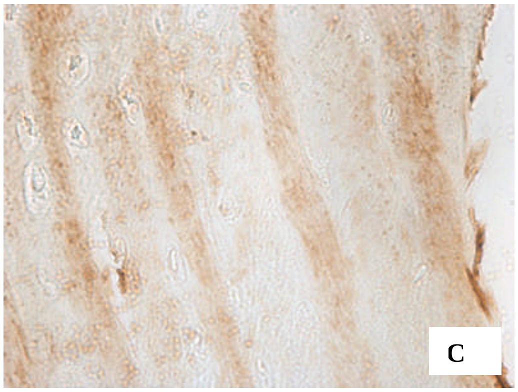
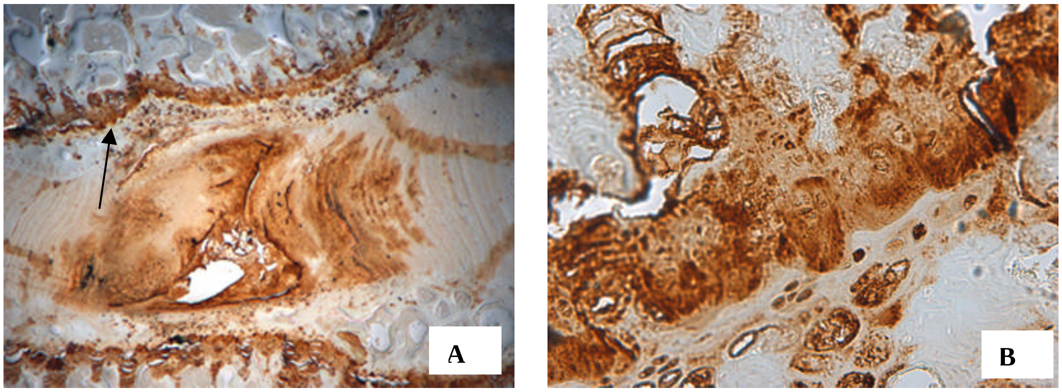


Fig. 5: Collagen II IHC staining.
A: 2.5X, arrow indicates location of magnification in B and C.
B: 10X (annulus)
C: 40X (annulus)

The stain for collagen II proved more successful (see Fig. 5), A. Although the nucleus is thought to contain this protein, the concentration of it is low in a healthy disc and thus little staining was expected there. The samples obtained from this stain support that fact. Unlike the nucleus, the inner annulus did stain for collagen II, as expected. All three images also show some staining of the outer annulus. Although collagen I is known to dominate the outer annulus in a healthy disc, this stain proves that some collagen II exists in this region as well.

Aggrecan, which is a proteoglycan (see Fig. 6), stained successfully when compared to the Alcian Blue stain from routine histology. The locations of staining by IHC correspond to the blue regions in the Alcian Blue stain. The end plate, included here as an example in the second two images, stained quite dramatically.

The drawback to such vivid staining is the high probability of significant background effects. Reagents may become trapped within the tissue and stain more than is actually there. Since the staining is so dark in this case, it is quite probable that this phenomenon explains the results. The goal of IHC staining is to localize proteins on a cellular level. General background noise in the data must be minimized to make that possible.



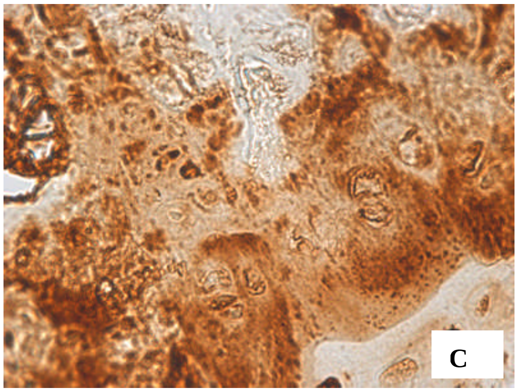


Fig. 6: Collagen II IHC staining.
A: 2.5X, arrow indicates location of magnification in B and C.
B: 10X (endplate)
C: 40X (endplate)

In all three cases the immunohistochemical staining would improve significantly if background staining and cross-reactivity effects were minimized. To assure such a minimization, the protocols for staining could be adjusted by: 1) lowering the concentration of reagents- such as the antibodies, 2) increasing the number of washes between each reagent, 3) changing the amount of time a reagent spends on the sample, and 4) any combination of these.

The most effective way to minimize background staining, however, is to minimize the amount of tissue that is being stained. In order to do this, the thickness of the sample on the slide must be minimized. In theory, samples can be sliced to a size as thin as 4 or 5 µm (on the particular equipment available). Currently, the thinnest slices accomplished have been 6 µm. The goal is to be able to slice 5 µm sections which have an intact nucleus. This can be done by adjusting the fixation procedure and the physical slicing methods. When the appropriate thickness is achieved, basic histology can provide information about how well that sample actually stains. Only when fixation and sectioning, tested by basic histology, have been optimized should immunohistochemistry be attempted again.

4.2 Understanding Gene Expression

As described in the methods section, RNA was extracted from healthy rat lumbar discs. The RNA was then converted to cDNA by RT-PCR, and finally, appropriate primers were added and PCR performed for three proteins: aggrecan, fibronectin, and collagen I. Fig. 7 shows the results. The leftmost column in each image is the ladder which specifies the size of the product. The brightest band visible (closest to the major bands) is at 600 base pairs. Each band is 100 base pairs away from its neighbor. The next column to the right is the product from the L2 level of the spine, while the one to the right of that is from the L3 level.

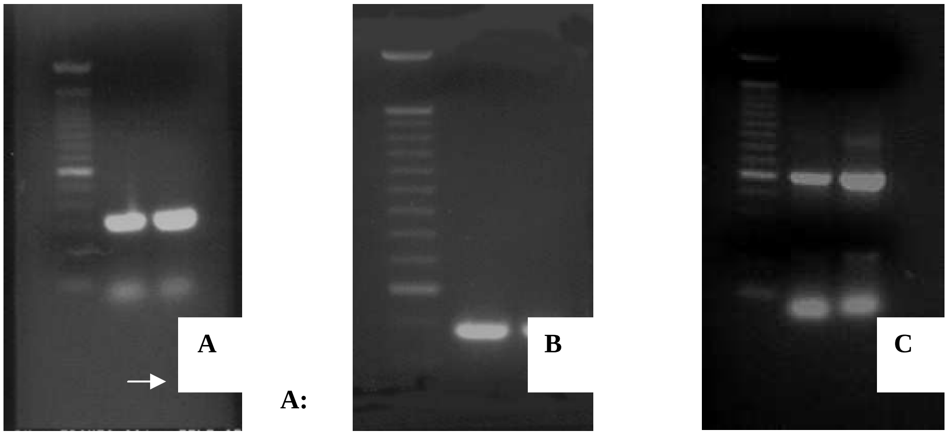


Fig. 7: **A:** Agarose gel of product from RT-PCR for aggrecan (322bp) **B:** Agarose gel of product from RT-PCR for fibronectin (481bp) **C:** Agarose gel of product from RT-PCR for collagen I (599bp)

Fig. 7, A is the product of the PCR for aggrecan. The presence of bands clearly means that the RNA extraction was successful. These bands are also in the location we expect, at 322 base pairs. Thus, the gene for the correct protein has been amplified. The only real problem with the bands lies in their brightness. Such brightness suggests over-amplification which could be fixed by reducing the number of cycles during PCR.

Fig. 7, B is the product of the PCR for fibronectin. Again, the presence of bands in the correct locations suggests a successful PCR. It is known that fibronectin content increases with degeneration of the disc in humans, so the presence and detection of it in a healthy rat disc is significant.

Fig. 7, C is the product of the PCR for collagen I. Bands for collagen I are clearly present, but so are a number of other bright bands which we do not expect. This suggests that the primer designed for collagen I was not as successful as it should have been and fragments of cDNA which were not coding for the protein of interest were amplified along with the collagen I fragments. A different primer must be tried to assure that the dominant bands are those for collagen I only.

Future work will split the nucleus from the annulus to study the two separately. It will also apply this protocol to more proteins and enzymes. Finally, and most importantly, the PCR will be quantified. This must be done by including a control sample for each sample of interest when doing PCR. This control sample would amplify a gene that is always present in the disc and is not affected by degeneration. Thus, by measuring the intensity of the band for the desired gene and normalizing it by the intensity of the band for the control, comparable information can be obtained for each sample.

Another possibility for a more quantitative result is real time PCR. In the future, the gene expression work will move to incorporate this tool. In real time PCR, the concentration of DNA is measured by the instrument as the reaction progresses. This process removes the need to perform gel electrophoresis- other than to check on the identity of the product. Real time PCR, however, requires a different set of primers than standard PCR. Consequently these will have to be redesigned.

5. CONCLUSIONS

The work conducted this summer has shown that immunohistochemistry and RT-PCR are powerful tools that can be applied to understanding the intervertebral disc. Collagen II and aggrecan stained as expected under the application of IHC. All three proteins studied by RT-PCR supported the success of the protocol. There is, however, still a significant amount of development that must be done before these techniques can be applied to an animal study. Each staining process showed some results, suggesting that the general protocol used is a good first step, but most had background staining which must be minimized in order to make the results meaningful. The nucleus must, for the most part, remain intact. Accomplishing this will require significant effort to optimize the fixation and slicing protocol of the samples. The gene expression work was successful, but must be quantified before a study can be attempted.

Once the protocols have been perfected, a study can be devised. A number of rat samples should be used with each animal at a different time point and at a different level of degeneration. IHC and RT-PCR can then monitor the expression and localization of various proteins, enzymes, and inhibitors of interest within the disc. By understanding these changes, we may understand the mechanisms of degeneration and ultimately be able to better diagnose, prevent, halt, and even reverse this process. This work may ultimately lead to back pain relief for the millions who suffer with this ailment.

6. RECOMMENDATIONS

As noted previously, the immunohistochemical staining would improve significantly if background staining and cross-reactivity effects were minimized. To assure such a minimization, the protocols for staining should be adjusted by lowering the concentration of reagents, such as the antibodies, increasing the number of washes between each reagent, changing the amount of time a reagent spends on the sample, and most importantly, minimizing the thickness of the tissue by optimizing the fixation and slicing protocols. Basic histologic stains should be used to test the quality of the sections, and immunohistochemistry only attempted when no background staining is apparent.

Future work in understanding gene expression should split the nucleus from the annulus to study the two separately. It should also apply this protocol to more proteins and enzymes. Finally, and most importantly, the PCR must be quantified. This must be done by including a control sample for each sample of interest when doing PCR. This control sample would amplify a gene that is always present in the disc and is not affected by degeneration. Thus, by measuring the intensity of the band for the desired gene, and normalizing it by the intensity of the band for the control, comparable information can be obtained for each sample.

7. ACKNOWLEDGEMENTS

Thank you very much to Professor Dawn Elliott, Ph.D. for providing the opportunity for me to work on this interesting project, John Boxberger for helping to guide my efforts, and Spencer Lake, Stephanie Perry, and Miltiadis Zgonis, M.D. for providing protocols and answering questions about techniques. The Elliott Lab for everyone's varied aid, and the McKay Orthopaedic Research Laboratory in general. A great thank you to NSF for providing funding and to SUNFEST and the University of Pennsylvania for involving me with this project.

8. REFERENCES

1. Lotz, J. C. (2004). "Animal models of intervertebral disc degeneration: lessons learned." Spine **29**(23): 2742-50.

2. Lu, D. S., Y. Shono, et al. (1997). "Effects of chondroitinase ABC and chymopapain on spinal motion segment biomechanics. An in vivo biomechanical, radiologic, and histologic canine study." Spine **22**(16): 1828-34.

3. Ogawa, T., H. Matsuzaki, et al. (2005). "Alteration of gene expression in intervertebral disc degeneration of passive cigarette- smoking rats: separate quantitation in separated nucleus pulposus and annulus fibrosus." Pathobiology **72**(3): 146-51.

4. Perry, S. M., S. E. McIlhenny, et al. (2005). "Inflammatory and angiogenic mRNA levels are altered in a supraspinatus tendon overuse animal model." J Shoulder Elbow Surg **14**(1 Suppl S): 79S-83S.

5. Boos, N., A. G. Nerlich, et al. (1997). "Immunolocalization of type X collagen in human lumbar intervertebral discs during ageing and degeneration." Histochem Cell Biol **108**(6): 471-80.

6. Kanemoto, M., S. Hukuda, et al. (1996). "Immunohistochemical study of matrix metalloproteinase-3 and tissue inhibitor of metalloproteinase-1 human intervertebral discs." Spine **21**(1): 1-8.

7. Nerlich, A. G., N. Boos, et al. (1998). "Immunolocalization of major interstitial collagen types in human lumbar intervertebral discs of various ages." Virchows Arch **432**(1): 67-76.

8. Roberts, S., B. Caterson, et al. (2000). "Matrix metalloproteinases and aggrecanase: their role in disorders of the human intervertebral disc." Spine **25**(23): 3005-13.

9. Xi, Y. M., Y. G. Hu, et al. (2004). "Gene expression of collagen types IX and X in the lumbar disc." Chin J Traumatol **7**(2): 76-80.

10. Elliott, D. M., P. S. Robinson, et al. (2003). "Effect of altered matrix proteins on quasilinear viscoelastic properties in transgenic mouse tail tendons." Ann Biomed Eng **31**(5): 599-605.

11. Elliott, D. M. and J. J. Sarver (2004). "Young investigator award winner: validation of the mouse and rat disc as mechanical models of the human lumbar disc." Spine **29**(7): 713-22.

12. Boden, Scott D., Buckwalter, Joseph A., Eyre, David R., Mow, Van C., Weidenbaum, Mark, et al: Intervertebral Disc Structure, Composition, and Mechanical Function: *Orthopaedic Basic Science.* Park Ridge, IL, American Academy of Orthopaedic Surgeons, 1994, pp 547-555.

13. Oegema, T. R., Jr., S. L. Johnson, et al. (2000). "Fibronectin and its fragments increase with degeneration in the human intervertebral disc." Spine **25**(21): 2742-7.

14. Buckwalter, J. A. and J. Martin (1996). Intervertebral disc degeneration and back pain. Low Back Pain: A Scientific and Clinical Overview. J. N. Weinstein and S. L. Gordon. Rosemont, IL, American Academy of Orthopaedic Surgeons**:** 607-623.

15. Buckwalter, J. A., S. D. Boden, et al. (2000). Intervertebral disk aging, degeneration, and herniation. Orthopaedic Basic Science. J. A. Buckwalter, T. A. Einhorn and S. R. Simon, American Acaemy of Orthopaedic Surgeons**:** 557-566.

16. Cs-Szabo, G., D. Ragasa-San Juan, et al. (2002). "Changes in mRNA and protein levels of proteoglycans of the anulus fibrosus and nucleus pulposus during intervertebral disc degeneration." Spine **27**(20): 2212-9.

17. Ishihara, H., D. S. McNally, et al. (1996). "Effects of hydrostatic pressure on matrix synthesis in different regions of the intervertebral disc." Journal of Applied Physiology **80**(3): 839-846.

18. Yokota, A., J. A. Gimbel, et al. (2005). "Supraspinatus tendon composition remains altered long after tendon detachment." J Shoulder Elbow Surg **14**(1 Suppl S): 72S-78S.

19. Weiler, C., A. G. Nerlich, et al. (2002). "2002 SSE Award Competition in Basic Science: expression of major matrix metalloproteinases is associated with intervertebral disc degradation and resorption." Eur Spine J **11**(4): 308-20.

20. Boxberger, J., S. Sen, et al. (2005). "Glycosaminoglycan content affects intervertebral disc neutral zone mechanics in axial loading." Transactions of the Biomedical Engineering Society.

21. Yerramalli, C., A. Chou, et al. (in review). "The effect of nucleus pulposus crosslinking and glycosaminoglycan degradation on disc mechanical function." Biomechanics and Modeling in Mechanobiology.

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

NSF Summer Undergraduate Fellowship in Sensor Technologies
Miguel Pérez Tolentino (Physics and Electronics)-University of Puerto Rico at Humacao
Advisor: Jorge Santiago Aviles

Study on the Implementation of sintered LTCC and Graphite as a sacrificial material for the fabrication of Microcombustors

ABSTRACT

The micro-combustor is a compact, sub-millimeter device that burns hydrocarbon fuels homogeneously as a source of power. It efficiently converts heat generated by combustion into electric power, and has the potential to replace batteries in portable applications that require long-term power. The possible benefits of these devices include their ability to provide greater energy and power density, higher temperatures and greater efficiency as a heat source. Also this technology can have many applications such as military portable systems, consumer portable system, and chemical control reaction systems.

The problem we are addressing is the fabrication of a gas fuel micro-combustor for a compact, portable electric power generator using thermoelectric elements. This device has been fabricated [1] using a competing technology, which is more complex, time consuming and therefore more expensive.

The materials to be used for the construction of this device are fundamentally Low Temperature Co-Fired Ceramic (LTCC) and Graphite. The fabrication of this device will rely essentially on a thermal process (sintering of the LTCC tapes). The instruments that will be use for the fabrication / characterization include: a furnace for sintering the ceramics, a heated press for the ceramics lamination, and a thermal laser and a numerically controlled milling machine for the patterning and machining of the tape.

In order to obtain a sense of the flow behavior within the device, numerous simulations have been made using a commercial program call FEMLAB. This program will take into consideration a diversity of parameters to measure such as the speed, pressure, fluid Reynolds Number, among others.

The main objective of this project is to complete the fabrication of a small combustor that contains fundamentally three inputs, one output and a combustion area. In one of its inputs a combustible gas (hydrogen) is injected, and oxygen from the air as an oxidizer flows through the other two inputs. The gases are mixed in the combustion area. A flame is initiated in the combustion area to burn the fuel / oxidizer mixture.

It is hoped that the combustor fabrication will be completed as designed. The parameters that characterize its combustion and power are expected to be consistent with its application as an electrical generator by means of the thermoelectric effect.

We are using FEMLAB, a commercially available numerical analysis software package for the simulation process.

As means to practice the lamination process, where the LTCC and graphite are bonded by means of uni-axial forces and heat, an experiment was carried out using pencil leads (0.5mm and 0.7mm in diameter) as source of graphite. Three parameters, force, temperature and time, were used to control the pencil lead lamination to LTCC. The parameters were 1,100 -1,300 pounds, 100º to 200ºF, and 15 to 20 minutes, respectively. By sintering the LTCC (after the lamination with pencil leads), micro-channels were formed, taking the characteristics of graphite leads morphology. In other words, graphite serves as a sacrificial material in the formation of channels or conduits

Contents

1. Introduction

When the air is mixed with a combustible and ignited to form a flame producing high temperatures, the process is known as combustion. The combustion is a chemical reaction in which a fuel (element or component) is combined with an oxidizer (generally oxygen in form of gaseous O_2), giving off heat and producing an oxide. Frequently used types of element for the combustion are the carbon and hydrogen. The combustion process happens as often in living beings as in devices used as sources of energy.

When this process happens inside a device, is known as a combustor. Combustors are commonly seen in mechanical motors such as in cars, airplanes, boats, etc. As one knows these are made to move and climb, as they are designed for the displacement of great weights that require enormous amounts of energy, which implies the consumption of great amounts of fuel. Nevertheless components exist that do not require large amounts of energy. These in their majority are electronics systems, which are designed to consume energy supplied by means of batteries and electricity.

To make a combustor at a small scale, one that will work for devices requiring lower energy levels, it has been proposed that one must construct a combustor of proportionally smaller dimension. This is known as a micro-combustor. The micro-combustor is a compact, millimeter length device that burns hydrocarbon fuels homogeneously as a source of power.

The main objective of this project is to complete the fabrication of a micro-combustor out of LTCC tapes, which contains fundamentally three inputs, one output and a combustion area. In one of its inputs a combustible gas (probably hydrogen) is injected, and air flows through the other two inputs. The gases are mixed in the combustion area. A flame is initiated in the combustion area to burn the fuel / oxidizer mixture by means of a capacitor discharge or a piezoelectric element.

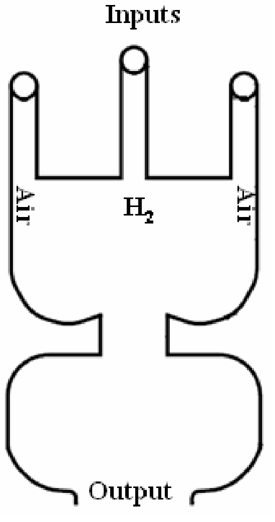
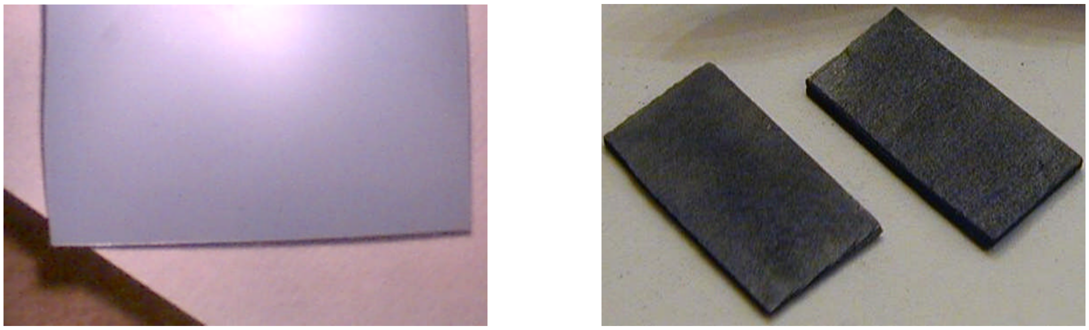


Fig1. Two-dimensional micro-combustor

The materials to be used for the construction of this device are Low Temperature Co-Fired Ceramic (LTCC) tape and Graphite. The LTCC represent an important alternative to be used as substrates for machining in the meso and micro scale. They provide several advantages including: electronic circuits can be integrated because of their hybrid nature, tapes of different compositions can be formulated to obtain desired layer properties (e.g. magnetic permeability), possibility of fabrication of hybrid structures consisting of ceramics, silicon, metals and/or some other suitable materials, layer count can be high, possibility of self-packaging, fabrication techniques are relatively simple, inexpensive and environmentally benign. Graphite is one of the two allotropic phases of carbon. This material has uses in many applications such as: electrodes, pistons, pencils, washers and diverse applications in engineering. It is of black color metallic, refractory brightness and it is easily worn away by means of heat (gasify).



a) b)

Fig2. a)LTCC and b)Graphite

The fabrication of this device will rely essentially on a thermal process (sintering of the LTCC tapes). The instruments that will be use for the fabrication / characterization include: a furnace for sintering the ceramics, a heated press for lamination of the ceramics, and a thermal laser combined with a numerically controlled milling machine for the patterning of the tape and machining of the graphite block.

Sacrificial material (graphite) insert.

The materials for the fabrication of the structure in figure 1 were graphite and LTCC. Even though the combustor is fundamentally two dimensional, it has a cavity where the fuel – oxidizer pre-mixed flame will be ignited. To form this cavity, a graphite insert (a sacrificial material, as it will disappear during high temperature sintering of the LTCC tape) will be used to keep the cavity from collapsing during sintering.

For graphite, a computer numerical control (CNC) was used. The CNC is a versatile system that allows controlled motion of the tools and parts through a computer programs that use numeric data. The numerical data that was used for the CNC in this work are the result of exporting data from a commercial program called Autocad 2000 i. This program is a useful tool in the field of engineering since it allows for an excellent quality draft of the architecture for any design. The detailed structure of the micro-combustor of a symmetrical form was designed using such CAD software. After exporting the data of the design to the CNC machine, a grinder was used as a tool to define the structure in graphite. The following figure schematically illustrates the process of definition of the structure morphology with the CNC.

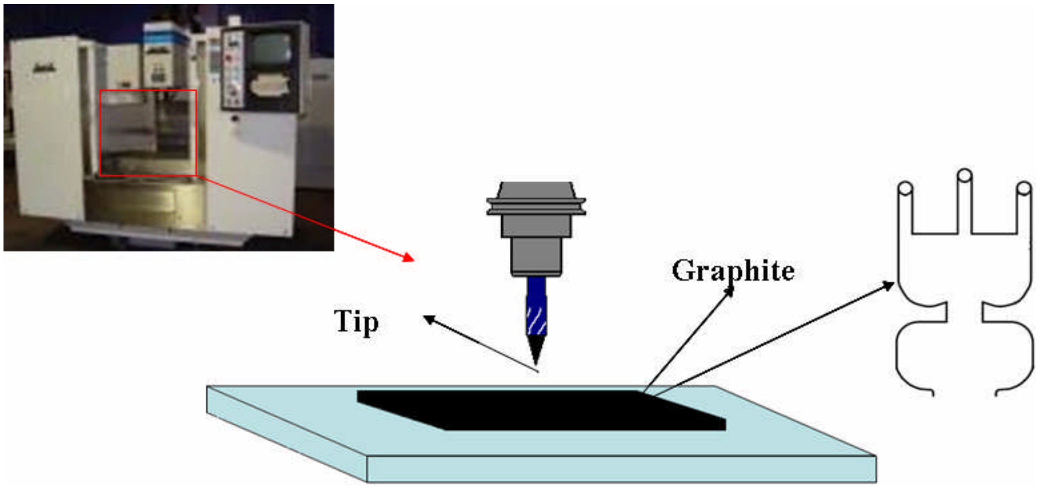


Fig.3 Schematic of Process with the CNC for the Microcombustor

When finished, the CNC machined graphite insert has the follow aspect.

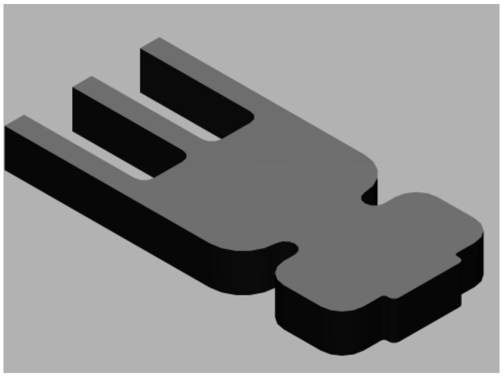


Fig.4 Scheme of the Graphite after of the CNC process

The LTCC tape combustor structure.

To define the combustor structure in LTCC tape a thermal laser was used to transfer the device pattern from an Autocad file (DXF file) to the ceramic tape. The laser system consists of a platform where the LTCC sample lies and a moving laser head shinning IR (10.6 μm) on the tape. A large piece of LTCC tape was placed in the laser platform and multiple units of the same pattern were serially machined.

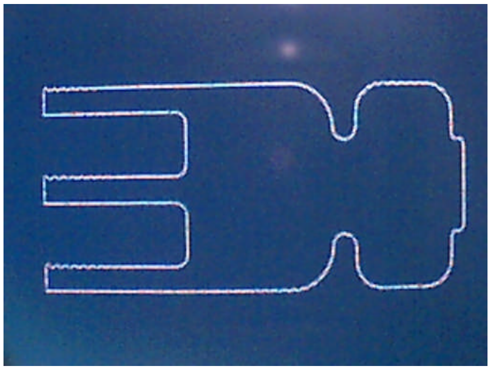


Fig.5 LTCC micro-combustor pattern

Once the pattern was cut into the LTCC tape by the laser, one remove the material outside of the pattern outline and utilize the resulting units for lamination.

The resulting cavity in the LTCC tape unit, served as the combustor cavity. Multiple tape units were laminated and the resulting cavity filled with the graphite sacrificial insert. Lamination is the method or process utilized to bond all the LTCC tape sheets as to construct a monolithic 3-D structure upon heating under a stress. When laminating LTCC tapes, it is important to keep all the sheets consistently with the same side up, that is, the LTCC tapes are fabricated over a Mylar sheet. To facilitate the release of the LTCC from the Mylar, a lubricant is utilized. It is important that the "shinny" side (side facing the Mylar) is always up or down.

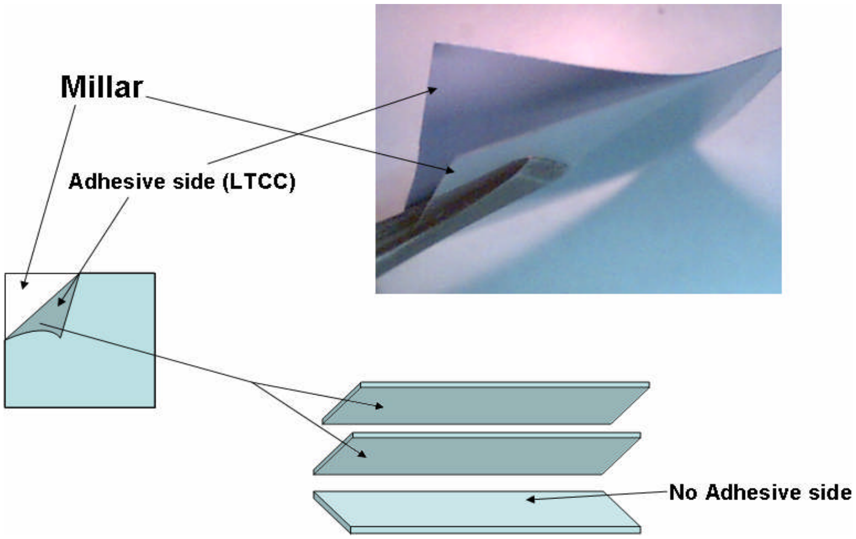


Fig.6 Consistent orientation of LTCC sheets before lamination.

The LTCC sheets at both ends of the laminate (top and bottom) are protected from the hot plattens in the hydraulic press by Mylar sheets. During lamination, the stress and time are controlled for best results. In our case we laminated at 1000 psi for 20 minutes at a platten temperature of 80 C.

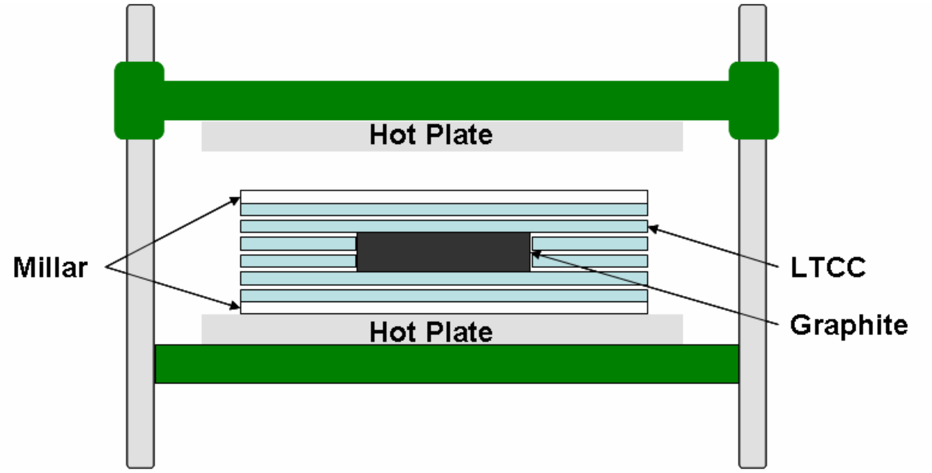


Fig.7 Scheme of the Lamination Process for the combustor

LTCC sintering

Sintering of the LTCC is the process in which green tape changes from a compliant plastic to a rigid solid and from a clear blue color to a darker hue. Sintering occurs after the structure has been laminated and it consists in placing the laminated structure in a furnace (LTCC and Graphite) and heat treat the composite to 850 C following a programmed sequence of heating steps. The structures to be sintered are placed on a alumina substrate as to avoid deformations conformal to the furnace substrate.

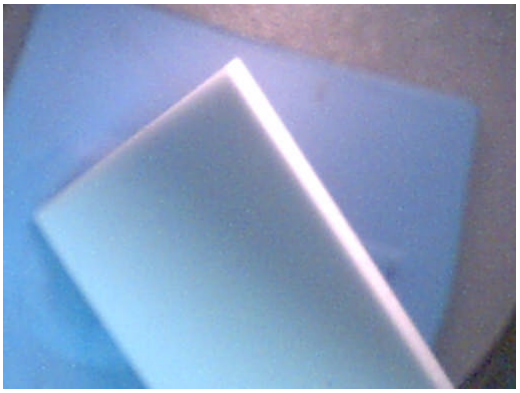


Fig9. Alumina substrate.

After the structure has been sintered one should observe that the graphite has disappeared and the perforations or cavities in the LTCC structure are clean.

The picture below is an example of a practice run, where we laminated LTCC tapes and used as a sacrificial material several pencil leads (graphite). The morphology is obviously that of a cylindrical channel, as can be observed.

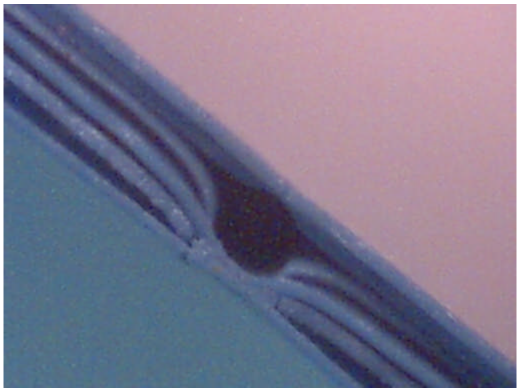


Fig10. LTCC sintering example using a pencil lead as sacrificial material.

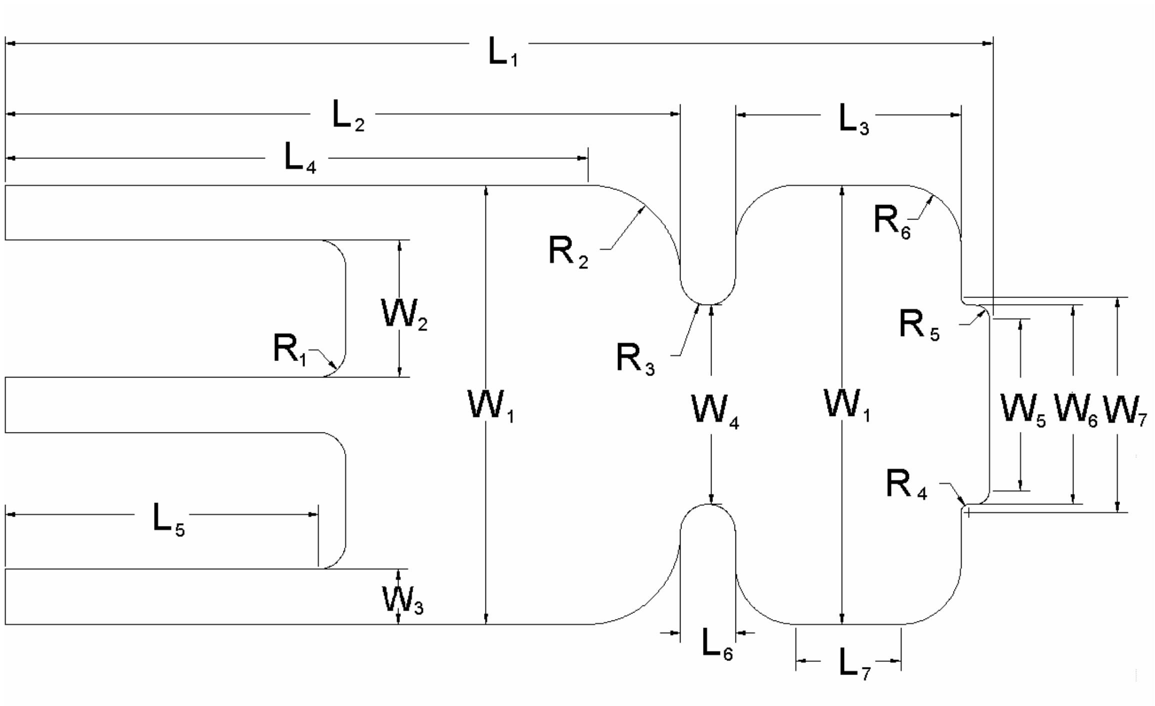
In order to obtain a preview of the flow behavior within the device, numerous simulations have been made using a commercial program call FEMLAB. This program will take into consideration a diversity of measurable parameters such as the speed, pressure, fluid Reynolds Number, and others. All simulations are two dimesional in natrue and they explore different configurations, morphologies, and conditions on which the combustor might be operated. The bulk of the simulations were concerned with the fluid mechanics. We used the Navier-Stokes formulation, preserving non-linearities and for convective and diffusive transport, the k.ε pertubatory approach. We were concerned with the effect of combustor geometry and volumetric flow rate of both fuel and oxidizer on the mixing. We would like to have a “pre-mixed” flame with a homogeneous composition near stoichiometry. All the simulations provided us with some insight into what was important concerning geometry (by indicating “dead zones” for mixing and particular configurations leading to mixing)

Combustor architecture

Below is a detailed description of the combustor morphology and configuration. There basically two competing forms, differing in length.

2. AutoCAD2000i based architecture diagrams.

Below is a detailed scheme of the combustor architecture. All the dimensions are in the SI system.



Number	Dimensions (millimiter)					
	L	W	R	1.5L	1.5W	1.5R
1	18.0641	8.0304	0.5	27.9615	12.0456	0.75
2	12.3539	2.5063	1.7	18.5309	3.7595	2.55
3	4.1274	1.0074	0.5	6.1911	1.5111	0.75
4	10.6538	3.6379	0.15	15.9807	5.4569	2.25
5	5.7291	3.1378	0.25	8.5937	4.7067	.375
6	0.9991	3.6378	1.1013	1.4987	5.4567	1.5195
7	1.9248	3.9378		2.8872	5.9067	

Some experimental details

Materials and tools

The tools and materials utilized in the fabrication of these devices were:

1. Green tape type 951AT from Dupont, Delaware, PA, U.S.A (approximately 100 µm thick).
2. Isotemp Programmable Forced-Draft Furnace (Fisher Scientific)
3. Heated press (Carver Model C)
4. Graphite (Arrow Springs, CA,U.S.A)
5. commercially available Alumina substrates
6. Computer Numerical Control (CNC) from Fadal
7. X-660 Laser Platform (Universal Laser Systems, Scottsdale, AR, U.S.A., 60 W CO_2 laser, wave length of 10.6 µm)

Computer Numerical Control (CNC) milling machine processing

Using a band saw a piece of graphite of 2mm thickness was obtained. A face mill tool in the CNC was used to smooth the graphite facing the LTCC tapes. The FACE Mill is a 2 in diameter cylinder with 3 carbide tips capable of polishing the graphite to less than 2 mils.

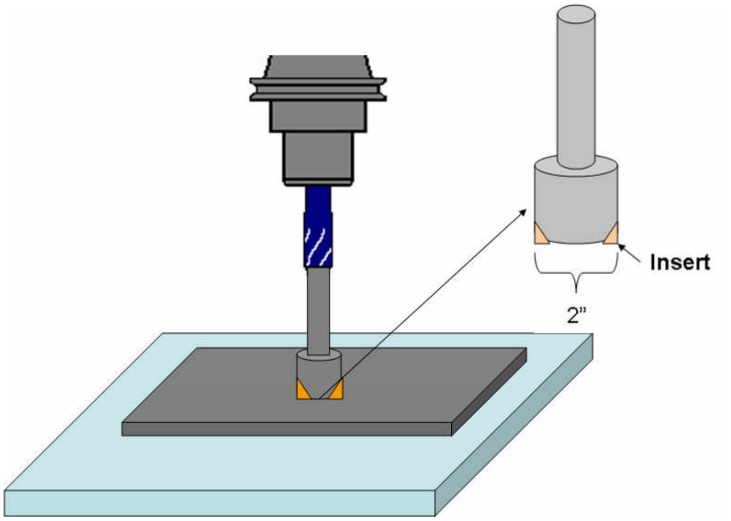


Fig11. Smoothing the graphite faces.

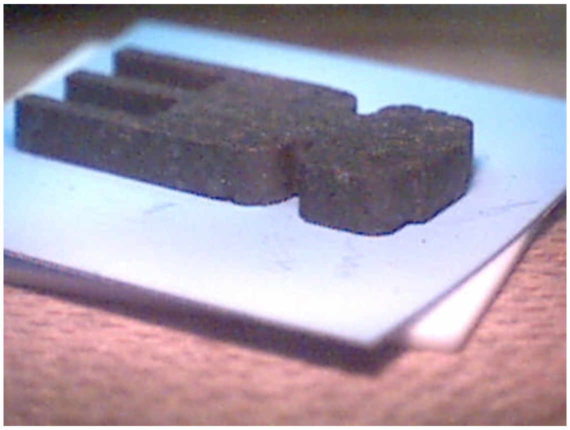


Fig13. Finished graphite insert

Some further details of the LTCC processing

Before attempting the fabrication of the combustor, we calculated the number of LTCC tape sheets needed for the laminated structure. The DuPont 951 LTCC is about 4 mils thick (around 100 μm) and for one of the possible combustor insert thickness (2mm), at least 20 layers were needed. The schematic below depict some of the process details.

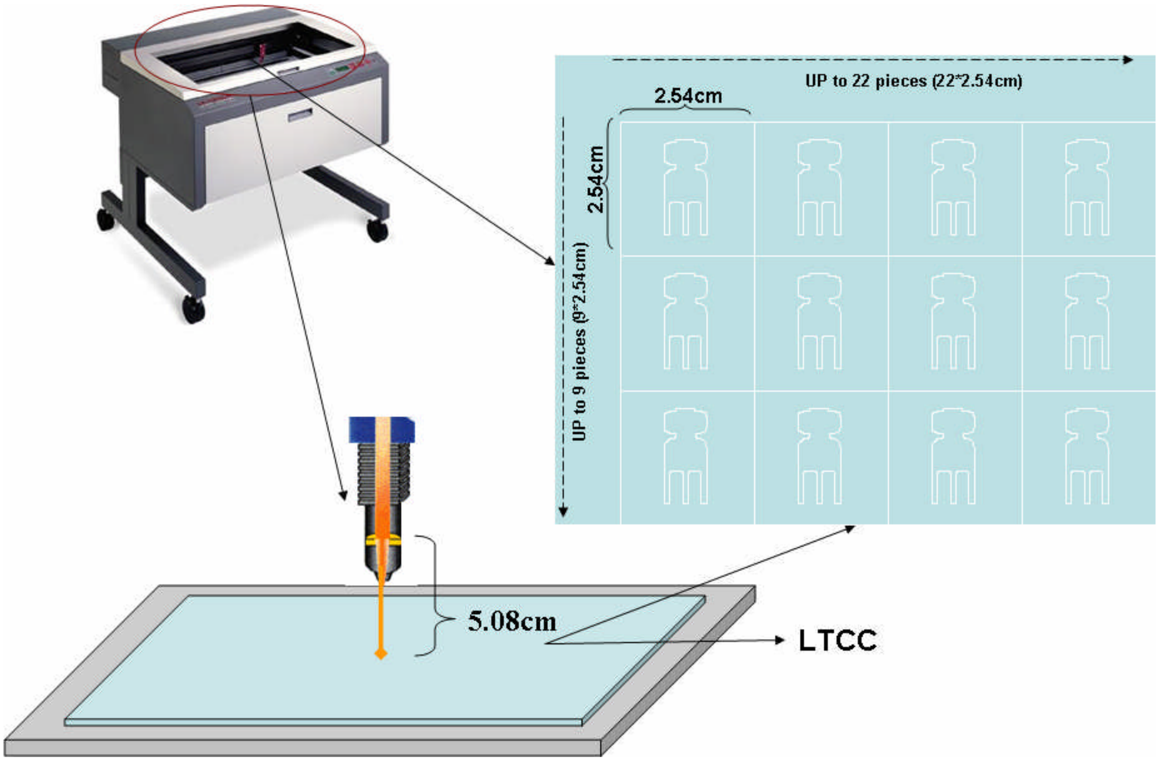


Fig14. Laser Process

The laser head velocity and power were the parameters best utilized for the LTCC tapes cutting. The cutting power was a strong function of where in the tape sheet you wanted to cut. If you wanted to cut far from the periphery (inside), then the power applied was 3% of the maximum power and the velocity 5 % of max. The outside (near the periphery) was 0.05 P_{max} at 0.05 V_{max} as depicted below.

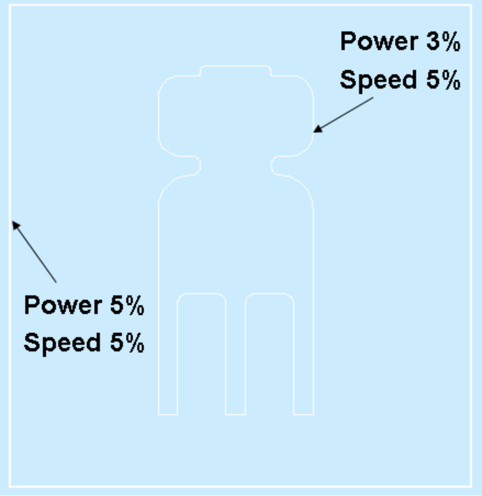


Fig15. Power (%) and Speed (%) for cut of LTCC

Some further details of the lamination Process

Again, twenty layers of LTCC were laminated around the graphite insert. Below one can note the near plastic behavior of the LTCC before firing (green tape), and stacked before lamination with the 20 LTCC layers.

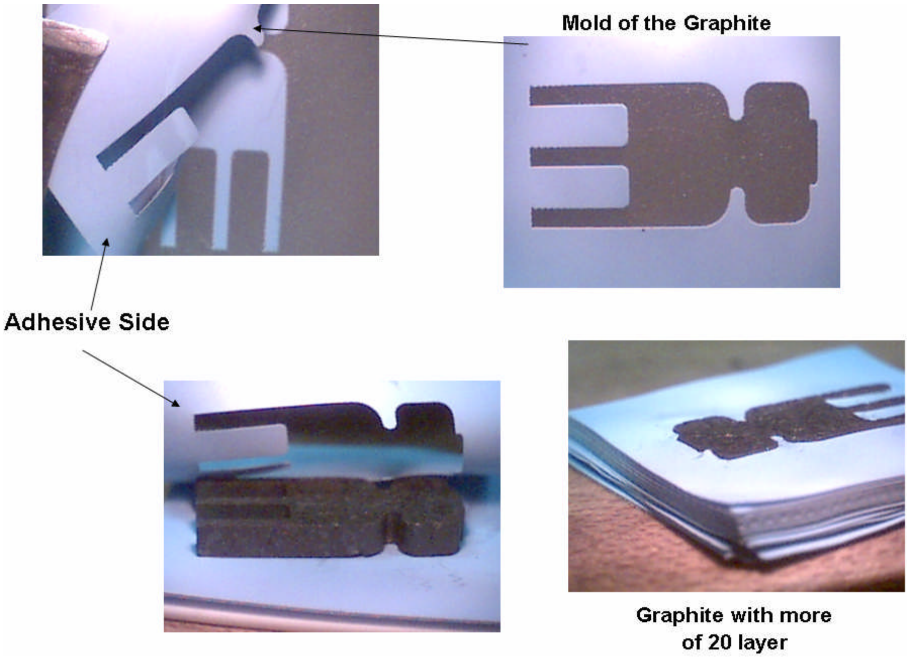


Fig16. Lamination process before pressing.

For the nominal dimensions structure a uniaxial pressure of 1000 to 1300 psi was utilized. For the 1.5 scale, the uniaxial pressure was incremented to 3300 psi. Both structures were exposed to temperatures in the range from 80 to 120C for a period ranging from 15 to 30 minutes.

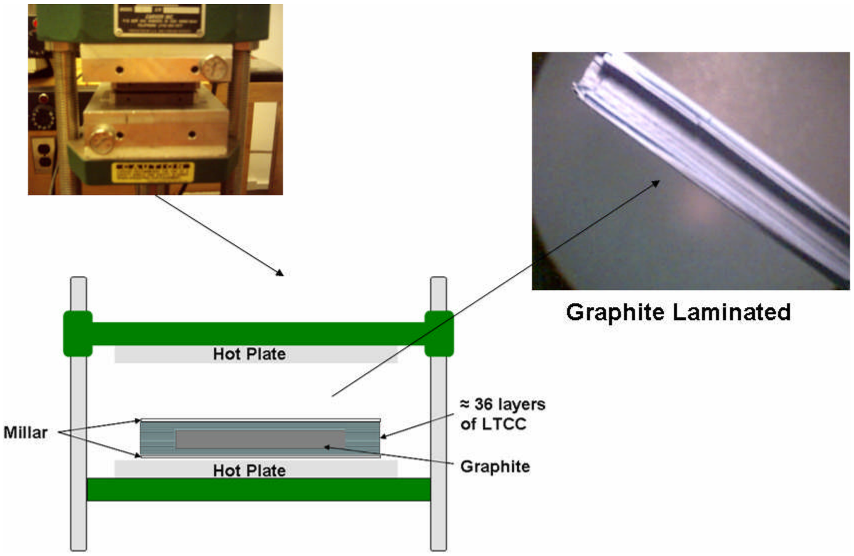


Fig.17 The lamination press

Sintering Process

The laminated structure was placed on a fully sintered Alumina substrate and heat treated as described below.

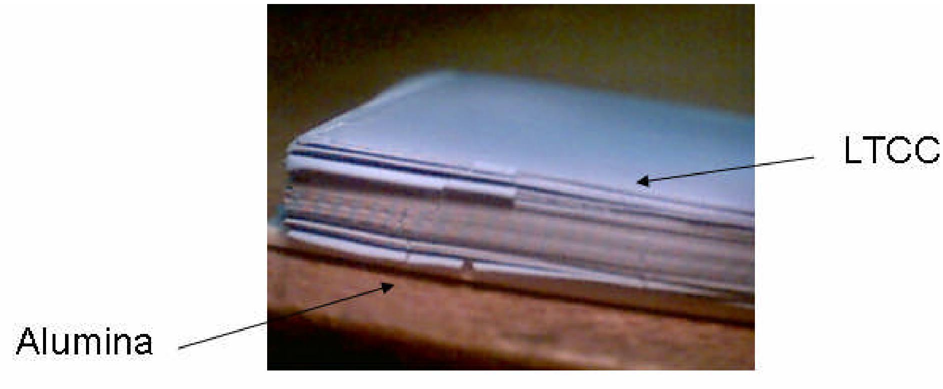


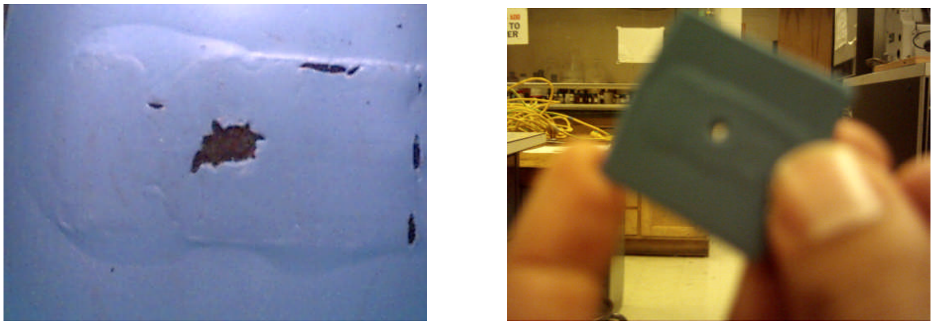
Fig18. Alumina and LTCC

The heating schedule as programmed in the resistance furnace:

- From room temperature to 300º C at a rate of 10º C/min.
- Kept at 300 C for 30 minutes
- Ramp from 300 C to 850º C at a rate of 10º C/min
- Kept at 850º C for 1 hour and 45 minutes.
- Turn furnace off and let it cool to room temperature.

Overall results

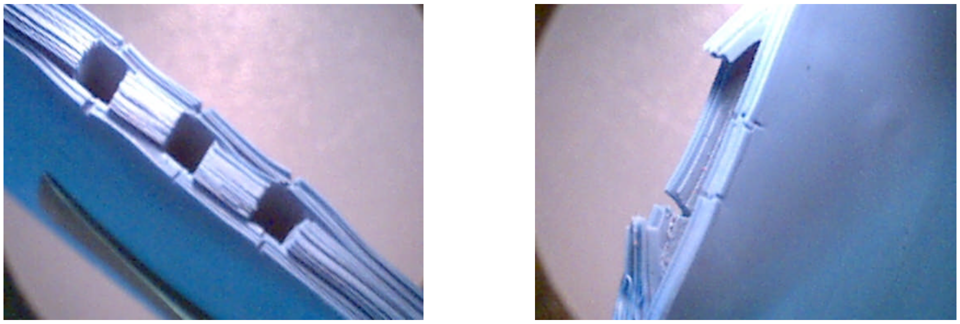
After sintering, the structures were examined for termination details. For example to what extent the sacrificial material has been utilized (oxidized, disappear)



a) Simple Scale b) 1.5 Scale

Fig19. Cavity in the combustor surface to evaluate graphite gasification.

The gas inlets and outlets of the initial trial I participated yield the following results.



a) b)

Fig20 Micro-combustor inlets (a) and outlet (b).

The top laminate inlet / outlet orifices after sintering looked, as depicted below. Both the lateral and top laminate I /O ports were considered marginally usable.

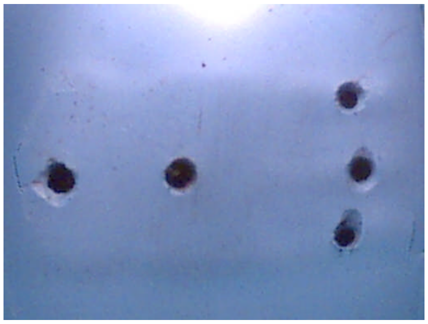
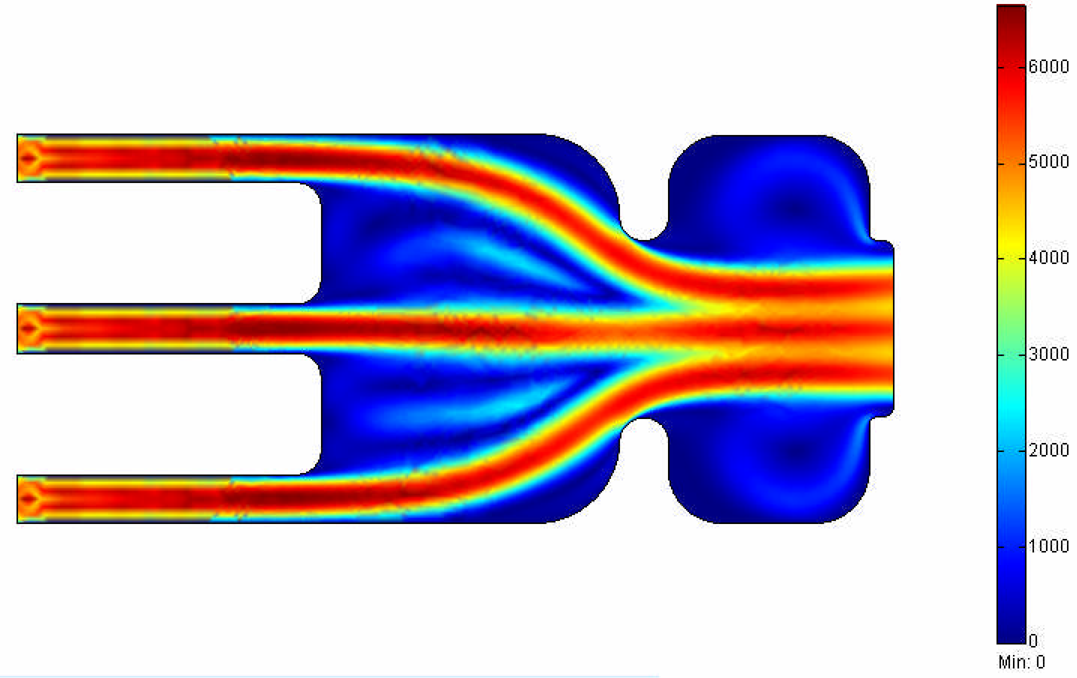
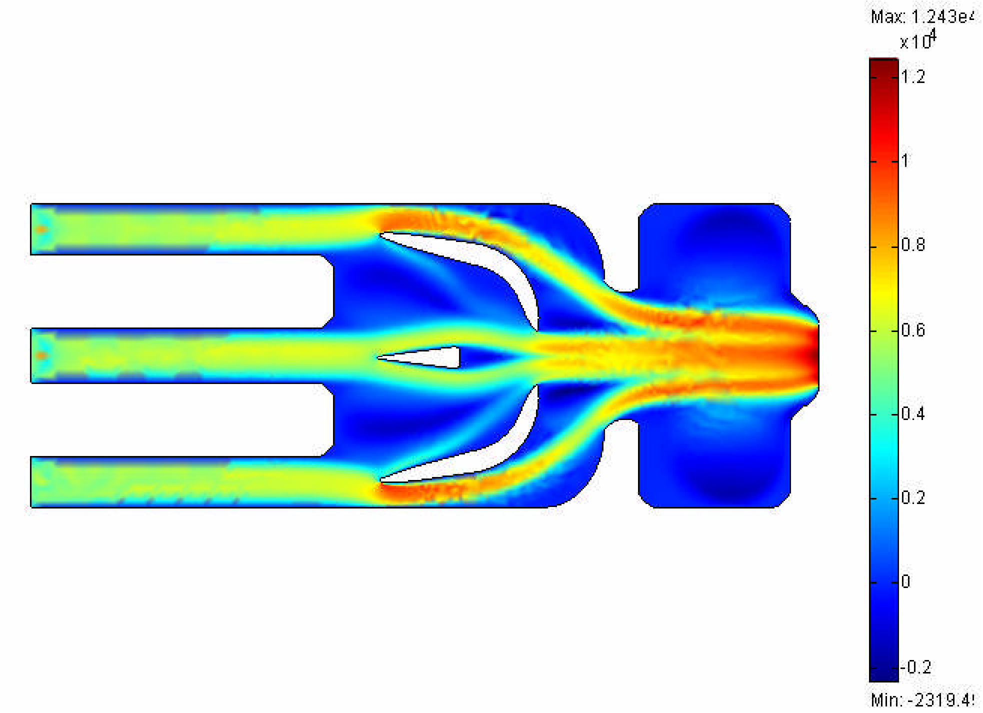


Fig.21 Inlet / outlets on the top laminate of a sintered micro-combustor.

Numerical Simulations.

To obtain some insight that might help us in the design of the combustor, multiple numerical simulations using the commercial package FEMLAB were realized. The fluid mechanic simulations yield surface velocity (mm/s) on a 2-D structure. The inlet velocity (Reynolds number) was varied as well as some structural and configuration parameters.





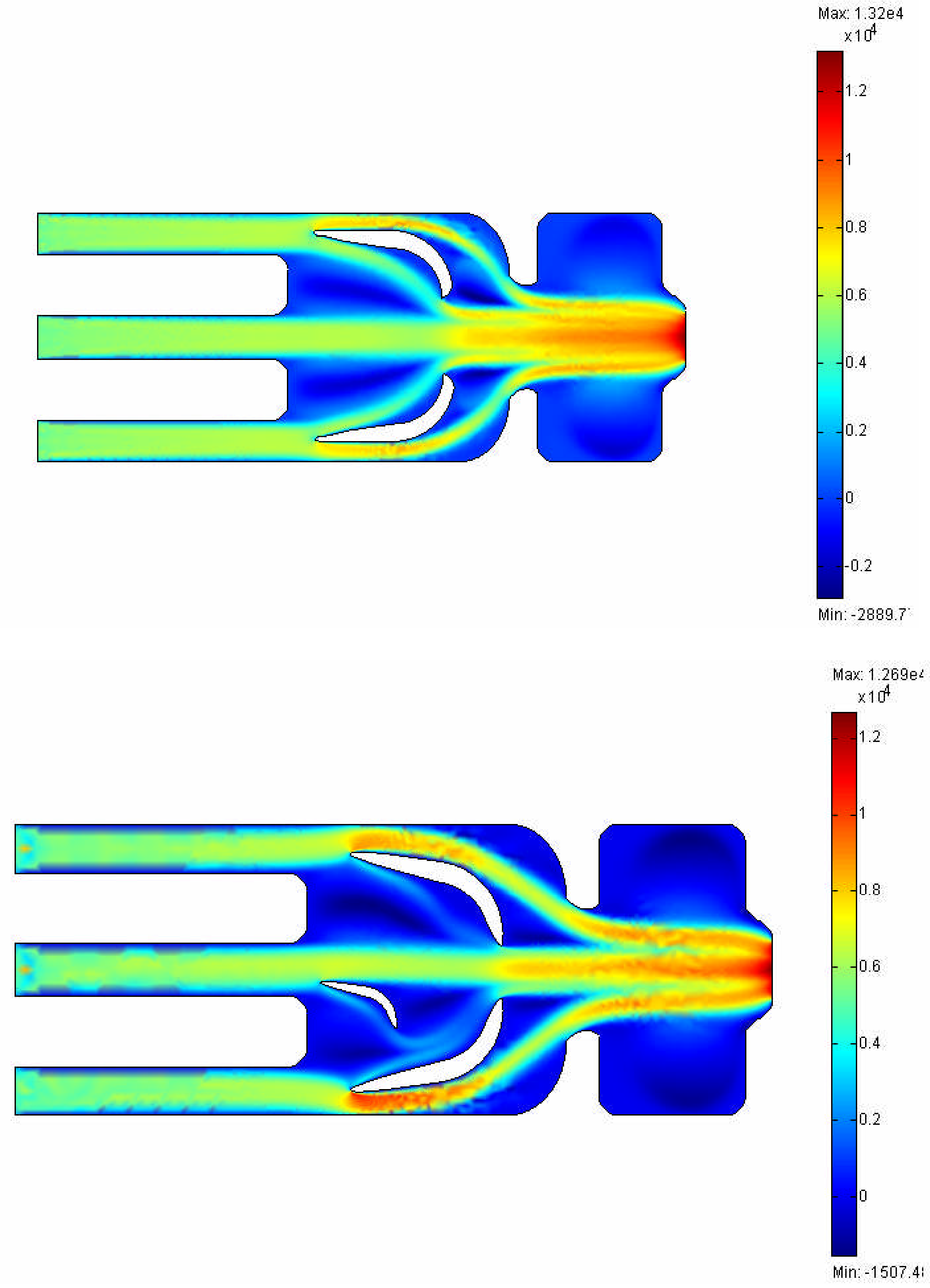


Fig22 Samples of simulation results for different flow rates and configurations of the combustor geometry. The false color map gives the surface velocity as depicted in the scale in the right of the figure.

Conclusions

One of the first things we learned in processing the combustor is that the sacrificial material, in this case graphite, might be difficult to gasify if the process is not done carefully. A larger outlets for the gasification products (CO and CO_2), drilling holes in the graphite insert and increasing the sintering temperature and time. Since we did only two runs, it was difficult for us to find a processing window for the gasification. Another structural issue was that the number of layers in both the top and bottom has to be increased to avoid stress induced de-laminations and cracking.

Our experiments showed that having at least 8 LTCC tape layers in both top and bottom circumvent the stress induced mechanical failures.

The combustor edges easily deformed when sintered without lateral support, suggesting that perhaps we need a sacrificial mold in the outside of the combustor structure to avoid sagging and deformation.

No doubt that LTCC tapes with graphite as a sacrificial material form an inexpensive and easy to manipulate materials for combustor applications as compared to metals and nitrides as previously reported in the literature.

Recommendations

An inexpensive and convenient material as a lamination aid (glue) is honey. Common honey burns all its volatiles before 300C and possesses the viscosity necessary for lamination. It was used by us, in the formation of a cylindrical channel with a pencil lead as sacrificial material and works like a champ. One can see in the figures below how difficult it is to laminate around a cylindrical structure. Once honey was used to enhance the lamination, de-lamination problems were minimized.

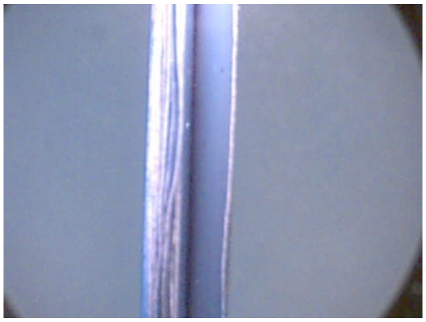


Fig23. Lamination failure (de-lamination) using pencil-lead as sacrificial material.

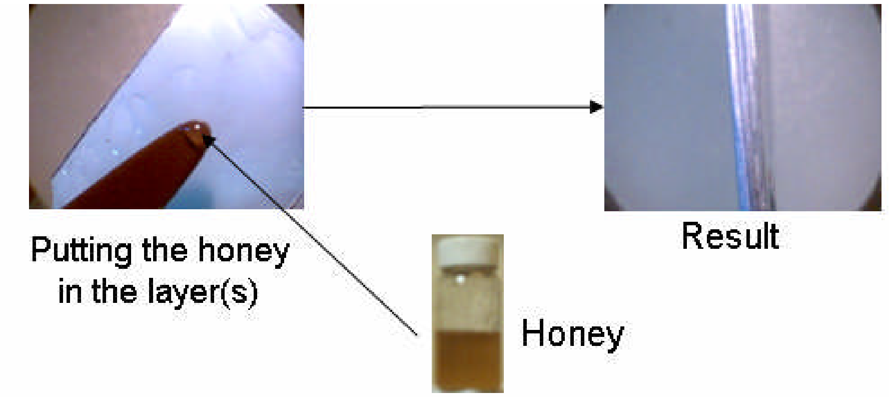


Fig24. Well laminated channel using pencil-lead as a sacrificial material.

Another significant improvement in preserving symmetry and surface smoothness is achieved by placing the laminated structure between two fully fired alumina sheets during sintering.

Another recommendation borne by the simulation results is to elongate the combustor outlet to avoid de-laminations and cracking, as suggested by the figure below.

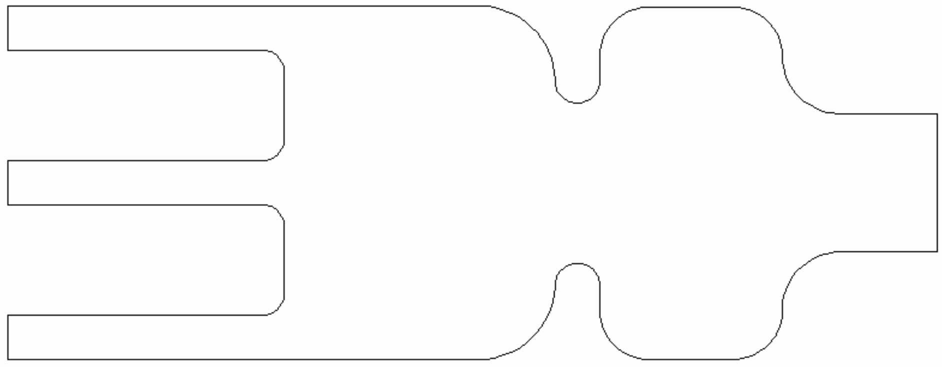


Fig. Proposed configuration and morphology to avoid thermal stresses damage during sintering and perhaps operation.

Acknowledgements

I wish express the thanks, first for the Professors Jorge Santiago and Rogerio Furlan for having recommended me to apply to the Sunfest Program. I also want to acknowledge the advice, help and guidance offered to me by Brett Ian Balough. I also thank to Dr. Jan Van Spiegel, the University of Pennsylvania, and University of Puerto Rico at Humacao, Sunfest Program and NSF.

Reference

[1] Lesley Millar, “Novel Microcombustor for Highly Efficient Generation of Electric Power from Fuel”, *Office of Technology Management*, University of Illinois at Urbana.

[2] Shuji Tanaka, Takashi Yamada, Shinya Sugimoto, Jing-Feng Li, and Masayoshi Esashi “Silicon nitride ceramic-based two-dimensional microcombustor”, *Journal of Micromechanics and Microengineering*, 13 (2003) 502–508.

[3] Lars Sitzki, Kevin Borer, Ewald Schuster and Paul D. Ronney, “Combustion in Microscale Heat-Recirculating Burners”, *The Third Asia-Pacific Conference on Combustion*, Seoul, Korea, June 24-27, 2001

[4] D. G. Norton, K. W. Voit, T. Brüggemann, and D. G. Vlachos, “Portable Power Generation Via Integrated Catalytic Microcombustion-Thermoelectric Devices”, Department of Chemical Engineering,University of Delaware.

[5] Miguel Perez Tolentino, Rogerio Furlan, Idalia Ramos and Jorge J. Santiago Aviles “Study of Laser Milling Of Sintered LTCC and Quartz Substrates for Microfluidic Applications”, *The National Conference On Undergraduate Research (NCUR) 2005*,Washington and Lee University Virginia Military Institute Lexington, Virginia, April 21– 23, 2005.

[6] M.R. Gongora-Rubio, P. Espinoza-Vallejos, L. Sola-Laguna, J.J. Santiago-Aviles, “Overview of low temperature co-fired ceramics tape technology for meso-system technology (MsST)”, *Sensors and Actuators A 89*, 2001, pp. 222-241.

[7] E.W. Simoes, I. Ramos, L. Garcia, R. Furlan, J.J. Santiago-Aviles, and M.T. Pereira, “Numerical modeling of a process for definition of features in low temperature co-fired ceramics”, *in Electrochemical Society Proceedings*, Volume 2002-8, Porto Alegre (Brazil), SBMicro 2002, September 9-14, 2002, pp. 244-252.

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

Thin-film Polypropylene Capacitors

NSF Summer Undergraduate Fellowship in Sensor Technologies
Usip, Ebenge (Physics) - University of Southern California
Advisor: Dr. Jorge Santiago-Aviles

ABSTRACT

The focus of this project was to determine the dielectric constant of polypropylene doped with various ethynol porphyrins at different concentration levels in order to determine if the degree of conjugation within ethynol porphyrin oligomers would increase the capacitance of naked polypropylene films. The research procedure for this study began by forming a capacitor with a film of polypropylene as the dielectric material. The capacitance was then measured and the dielectric constant was calculated. In order to deposit the film, polypropylene was spin-cast. Over the course of this study the capacitance values for polypropylene films doped with the compounds ZnTPP, ZnO1, ZnO3, Zn 3,5 and Zn 2,6 were studied. The capacitance values have shown a significant increase, roughly one million times greater than naked polypropylene although this value is inherently flawed due to inaccurate measurements of film thickness.

Table of Contents

1. INTRODUCTION

As the world attempts to wane its dependence on fossil fuels, the flexibility of electricity as an energy source has become increasingly important. The most efficient way for small amounts of high-voltage energy to be quickly stored and released is through the use of a capacitor. Of the many available capacitor designs those using thin-film technology are considered the best because they are capable of storing energy at high density levels. Amongst the many available films, polypropylene is considered the dielectric of choice because….. [1] Nonetheless, there are still many properties of polypropylene films which, with fine tuning, could be improved upon in order to make such thin-film capacitors more readily usable in many possible applications. One way of adjusting the properties of polypropylene is by using dopants, additives that are added in small amounts to a pure substance to change its physical properties. [2] The dopant used in this experiment is porphyrin. Porphyrins are tetrapyrrolic conjugated macrocycles with large π–conjugated ring systems and heteroatoms that give rise to porphyrin-porphyrin π–interactions. [7] This effectively makes for a highly polarizable substance which is ideal for this experiment.

This paper will cover many aspects of the thin-film polypropylene capacitor technology including its fundamental design and response to various doping agents. Section Two will cover a brief background of the classic capacitor and explain the transition from such capacitor designs to those of thin-film capacitors. Section Three will discuss design, implementation, and evaluation of the experimental thin-film capacitors. In Section Four final analyses of the experiment and results will be made.

2. BACKGROUND

2.1 Modern Day Applications

As mentioned earlier, thin-film polypropylene capacitors have immediate use in various industrial and military applications. For instance, industrial strength lasers and other pulse power equipment require high voltage power that can be quickly released. This characteristic of energy is one which capacitors can best provide. Moreover, this same kind of energy storage and release is desirable for a variety of other applications, notably, military weaponry. The most obvious of all applications would include refinement of the stun gun, which releases large doses of electrical charge in extremely short periods.

2.2 Parallel Plate Capacitor

The parallel plate capacitor is a capacitor which consists of two parallel capacitor electrodes. When these electrodes are equally and oppositely charged, they have an ability to store charge which is known as capacitance. The charge placed on the capacitor electrodes and the potential between the two plates are proportional according to the equation [3]:

$$Q = VC$$

In this equation Q represents charge, V represents voltage, and C represents capacitance. Although in this equation the constant C plays a special role since it has no dependence on Q or V. The value of capacitance is determined by the geometry of the capacitance electrodes which comprise the capacitor. In the case of a parallel plate capacitor capacitance is evaluated using the equation [3]:

$$C = \varepsilon_0 * A/d$$

In this equation C is again capacitance while ε_0 is the permittivity constant 8.85 E^{-12} F/m, A is the area of the two capacitor electrodes, and d is the distance between the two plates of the capacitor.

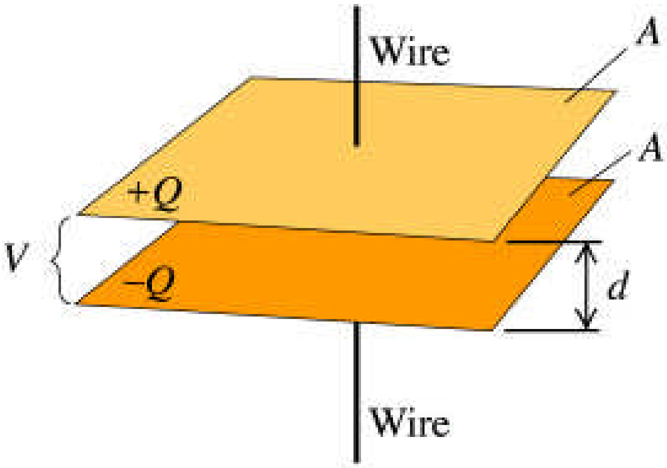


Figure 1-A charged parallel plate capacitor

These equations are of great importance for even at the thickness of microns they govern the design of thin-film capacitors.

2.3 Parallel and Series Capacitance

When connected in parallel *n* number of capacitors have an equivalent capacitance of [3]:

$$C_{equiv} = C_1+C_2+\ldots C_n$$

Capacitors in series have an equivalent capacitance of [3]:

$$1/C_{equiv} = 1/C_1+1/C_2+\ldots+1/C_n$$

The equation for equivalent series capacitance was very important in the experimentation process because this was the only method available to evaluate the polypropylene film capacitance using the capacitance bridge available.

2.4 Dielectrics

A dielectric is a material which sits between the two plates of a parallel plate capacitor. [3] Furthermore, a dielectric material must be polarizable because a capacitor stores electrical energy by polarizing its dielectric resulting in an electric field. [3]

2.5 Porphyrins

As stated earlier, porphyrins -- the core components of this research project -- are tetrapyrrolic conjugated macrocycles with large π–conjugated ring systems and heteroatoms that give rise to porphyrin-porphyrin π–interactions. [7] Each pyrrole subunit is linked to each other by way of a methine bridge. [4,5] The great appeal of porphyrins -- a natural pigment -- stems from the fact that freebase porphyrins can easily be manipulated by adding metal ions to the center of their structures. [6] This allows the properties of a porphyrin to be manipulated. In this experiment the metal ion selected was zinc. This choice was made because zinc porphyrins are commonly known to be one of the easier and more stable porphyrins to prepare.

3. EXPERIMENTAL METHODS

3.1 Shadow Mask Design

A shadow mask was employed to create an assortment of capacitors using one slide a shadow mask was used. This allowed for evaluation of the capacitance of numerous smaller capacitors with plate areas varying from .03516 cm^2 to 1.43 cm^2. The shadow mask used was designed using AutoCAD LT 2005 and fabricated with a CNC milling machine. Seen below, the mask was approximately 2.6 cm by 2.6 cm.

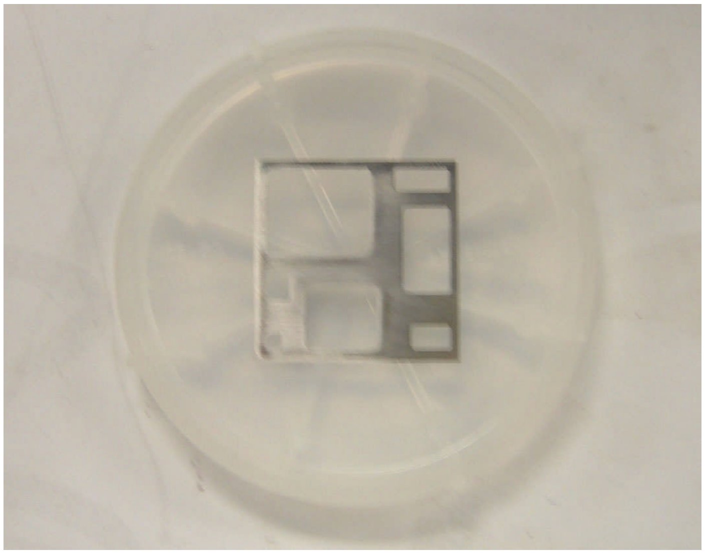
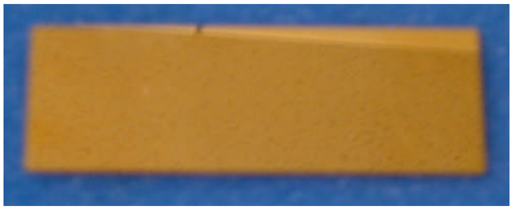


Figure 2-A picture of the shadow mask used

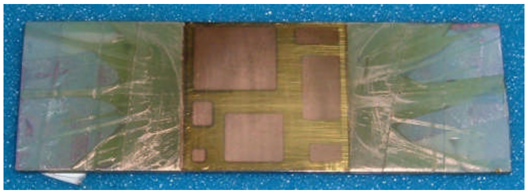
3.2 Slide Preparation

The original design proposed for this experiment called for plating glass slides with 100 nanometers of gold. This process proved too difficult to implement. Thus, it was decided that the slides would be purchased already plated from the company GenTel BioSurfaces.



3.3 Polypropylene Preparation

Preparation of the polypropylene was done with the assistance of doctoral student Paul Frail of Dr. Michael Therien's research group which is part of the Chemistry Department at the University of Pennsylvania. A variety of undoped solutions of polypropylene were prepared in order to: 1) better understand the spin-cast process, 2) determine the best concentration of polypropylene to use for the experiment, and 3) determine a background value to evaluate ethynol porphyrin oligomer's ability to increase the capacitance of polypropylene film. These solutions were heated to reflux, approximately 150 degrees Celsius. This is the approximate melting temperature of polypropylene. At this point, drops of the solution were placed upon a contained area of the slide marked off using Teflon tape. The slides were then placed in a spin-casting machine where they were spun so that the solution would form a film on top of the slide. Once the film was deposited it was taken to a sputter-coat machine where anywhere from 600 to 1000 angstroms of gold were deposited through the shadow mask. Upon completion the slide was taken to a LCZ meter where the capacitances of the areas formed by the shadow mask were measured using micromanipulating electrodes. After evaluating solutions of five different concentrations, it was determined that 1 gram of polypropylene in 50 mL of Decalin (solvent) formed the most reproducible and best measurable capacitance values.



3.4 Dopants

Dopants are additives added in small amounts to a pure substance in order to alter its properties. [2] In this experiment dopants were added to polypropylene with the intention of increasing the capacitance per area of the thin-film capacitors created. The amount of dopant added was determined via percent relative to weight. The quantity was varied in order to determine the most effective loading.

3.4.1 ZnTPP

The first and simplest dopant to be added to the polypropylene was tetraphenyl porphyrin (ZnTPP). Doped solutions were prepared from 200 mg of polypropylene in 10 mL of Decalin. Solutions with four different loadings were made including 1%, 5%, 10%, and 15% by weight. After assessment of the slides it was determined that the 1% and 5% doping levels produced films which were too thin. In some instances the capacitance produced caused measurement overflows in the LCZ meter. Thus, it was decided that according to the various films consistency, and lateral homogeneousness, 10% doping would be sufficient for taking capacitance measurements. In order to further increase film thickness it was decided that multiple layers of film would be deposited. The concept was implemented as such: first a layer of undoped polypropylene was spun on top of the slide.

Then a second layer of polypropylene containing the dopant was spun onto the slide. Lastly, a third film of undoped polypropylene was spun on top of the doped film. Based on the results of the preceding steps and the new film technique, a new set of more complex dopants were tested.

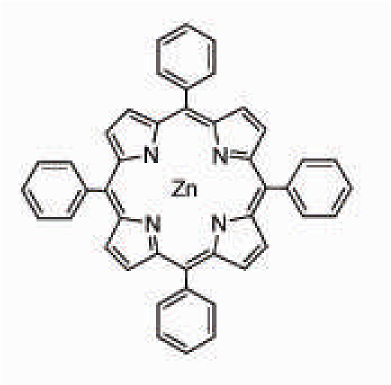


Figure 3-A diagram of the porphyrin structure ZnTPP

3.4.2 ZnO3

In the second round of trials a set of dimer porphyrins were assessed. ZnO3 was the first to be evaluated. Once again, doped solutions were prepared from a polypropylene-based solvent. In this set of slides all dopants were added at a proportion of 10% by weight. The compound ZnO3 is characterized by its relatively high level of interaction with similar porphyrins. Increased capacitance values for the film doped with this substance suggest that the relatively high level of interaction of ZnO3 may be attributed to polypropylene's increase in capacitance relative to solutions doped with ZnTPP.

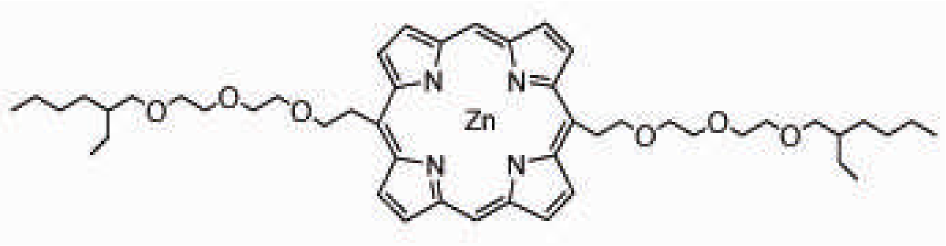


Figure 4-A diagram of the porphyrin structure ZnO3

3.4.3 ZnO1

The porphyrin ZnO1 was the next dopant evaluated. It is similar to ZnO3 yet has four less oxygen molecules. This makes for a slightly less interactive, however more bulky molecule. No fair predictions could be made about whether this characteristic would aid in increasing or decreasing the capacitance of polypropylene. Interestingly, results showed a negligible difference in capacitance values for similar capacitors at 100 Hz frequency AC current when compared to ZnO3. Nonetheless at the 1 kHz frequency the capacitance values for ZnO1-doped film were significantly less and on average, were almost half that of ZnO3-doped films.

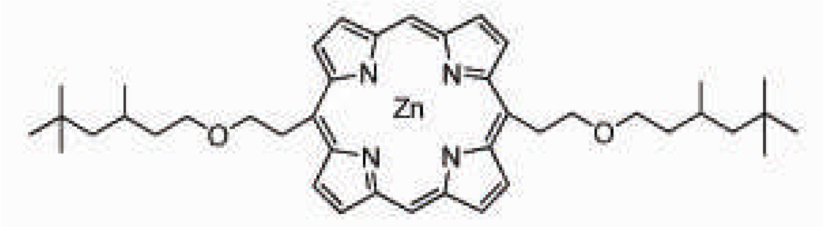
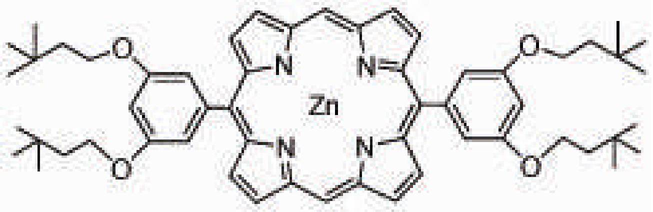


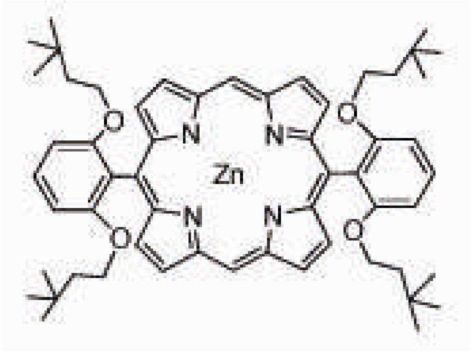
Figure 5-A diagram of the porphyrin structure ZnO1

3.4.4 Zn (3,5) & Zn (2,6)

The next two dopants that were tested have structures significantly different from the previous porphyrins. The first of these is Zn (3,5) which contains two phenyl groups on opposite sides of the structure. Furthermore, both phenyl groups have bulky alkoxy side chains at the 3 and 5 positions. Capacitance values for the film to which this dopant was added were significantly greater than films doped with ZnO1 and ZnO3. In fact, the capacitance values recorded were on the order of roughly 1.5 to 3 times greater.



The last dopant to be thoroughly evaluated was Zn (2,6). Even bulkier than Zn (3,5) this porphyrin also contains phenyl groups on opposite side of the structure. In addition alkoxy side chains are attached at positions 2 and 6. This orientation makes for a larger angle between the alkoxy side chains than the positions 3 and 5 occupied in Zn (3,5) and the result is an even bulkier molecule which takes up mores space. Nonetheless, there was a negligible difference in capacitance values for Zn (3,5) and Zn (2,6). These results suggest that positioning of the alkoxy side chains has no effect on the capacitance values of the polypropylene film.



3.4.5 Trimers

The last group of porphyrins tested was trimers. These porphyrins included ZnO3, ZnO1, Zn (3,5) and Zn (2,6). The capacitance of films doped with these compounds was significantly greater than that of films doped with any previous porphyrins. In fact, the differences were immeasurable. Even when run in series with capacitors of 5 and 6 mF the capacitance values could not be accurately measured. From the few values that could be obtained it is estimated that the capacitance of polypropylene film doped with trimers is some where between 200 and 1000 mF.

4. DISCUSSION AND CONCLUSIONS

Preliminary results have shown that porphyrins are most certainly a proper agent for increasing the capacitance of polypropylene film. With a metal-coated slide, a spin-caster, and a sputter-coating machine capacitors can be prepared to evaluate the dielectric film in question. The technique used in testing various porphyrins is still being refined yet is certainly a fruitful area for future research. It is also essential that researchers keep in mind that the technique which was used for testing requires numerous pieces of rather expensive and highly delicate machinery that are vulnerable to mechanical failure. Nonetheless, the experimental capacitor is a great way to record the capacitance values of a dielectric film and its use is only limited by the accuracy of the measuring equipment used.

5. RECOMMENDATIONS

Overall the technique used in this experiment worked reasonably well. However, there were two factors that hindered the pace of this research project. These were the availability and the precision of working equipment. Obviously these issues were of great importance when carrying out the experiment; they also became critical when doing calculations. A resulting problem was that measurements could not be taken for the later films because their capacitance values were too large for the capacitance bridge available. In many cases the meter simply read “overflow”. Additionally, no calculations could be completed on ε_0 for any of the films used because thickness could not be measured precisely. The instrument used was an ellipsometer and it is recommended that this instrument not be employed for any similar experiments because of its limited utility.

6. ACKNOWLEDGMENTS

I would like to thank Dr. Jorge Santiago and Paul Frail for their guidance throughout this project. I would also like to thank Dr. Therien’s lab for allowing me to use their materials and equipment in order to carry out this project. Sincere thanks to Scott Slavin for allowing me to use the microfabrication lab to make measurements. In addition I would like to thank the staff of the Electrical and Systems Engineering department for their help in solving many of the problems that I encountered over the course of the summer. Last but not least I would like to thank the National Science Foundation for their continuing support of the SUNFEST program through the NSF-REU grant. Without their support this whole experience would not have been possible.

7. REFERENCES

[1] 3. R. R. Anderson, Selecting the right plastic film capacitor for your power electronic applications, Proc. Industry Applications Society, IEEE-IAS Annual Meeting, 1996, pp. 1327-1330.
[2] “Dopant” Available from: http://en.wikipedia.org/wiki/Dopants, Accessed from: July 2005.

[3] Z. Popović, B.D. Popović *Introductory Electromagnetics*, Prentice Hall, Upper Saddle River, NJ, 2000, p. 84,104-110.

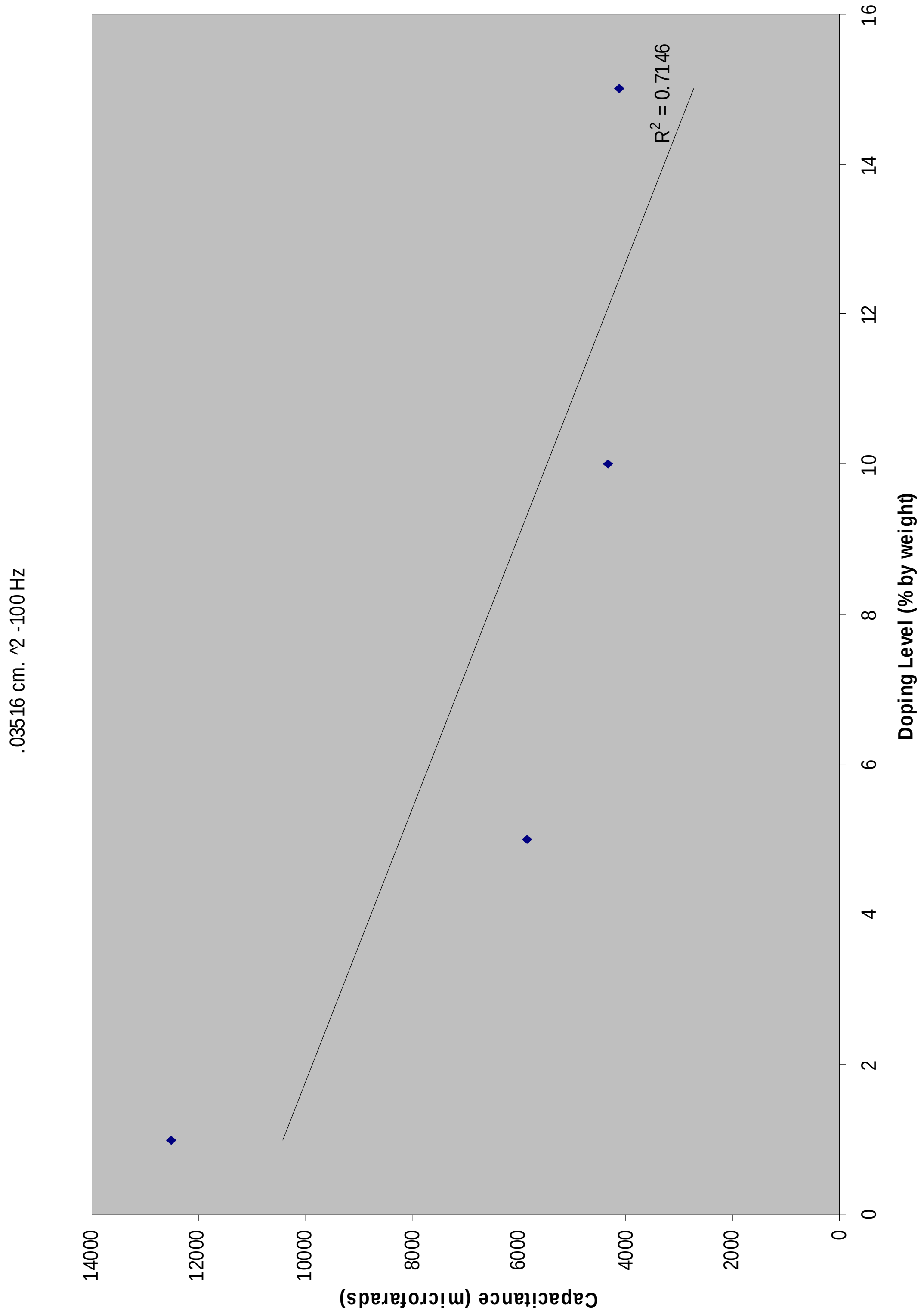
[4] “Executive Summary: Porphyrin and Metalloporphyrin Chemistry” Available from: http://www.scs.uiuc.edu/suslick/execsummporph.html, Accessed from: July 2005.

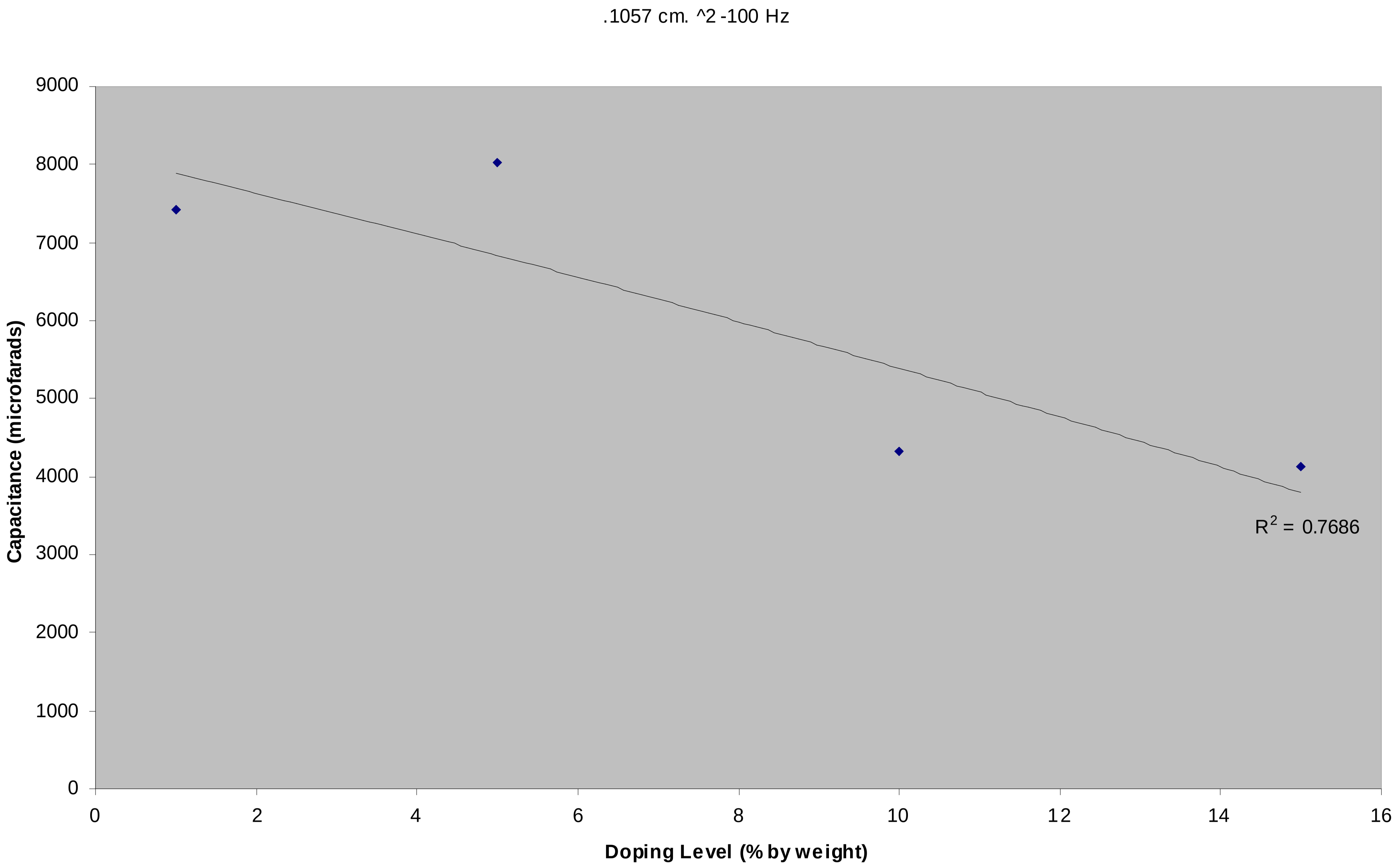
[5] “Porphyrin” Available from: http://en.wikipedia.org/wiki/Porphyrin, Accessed from: July 2005.

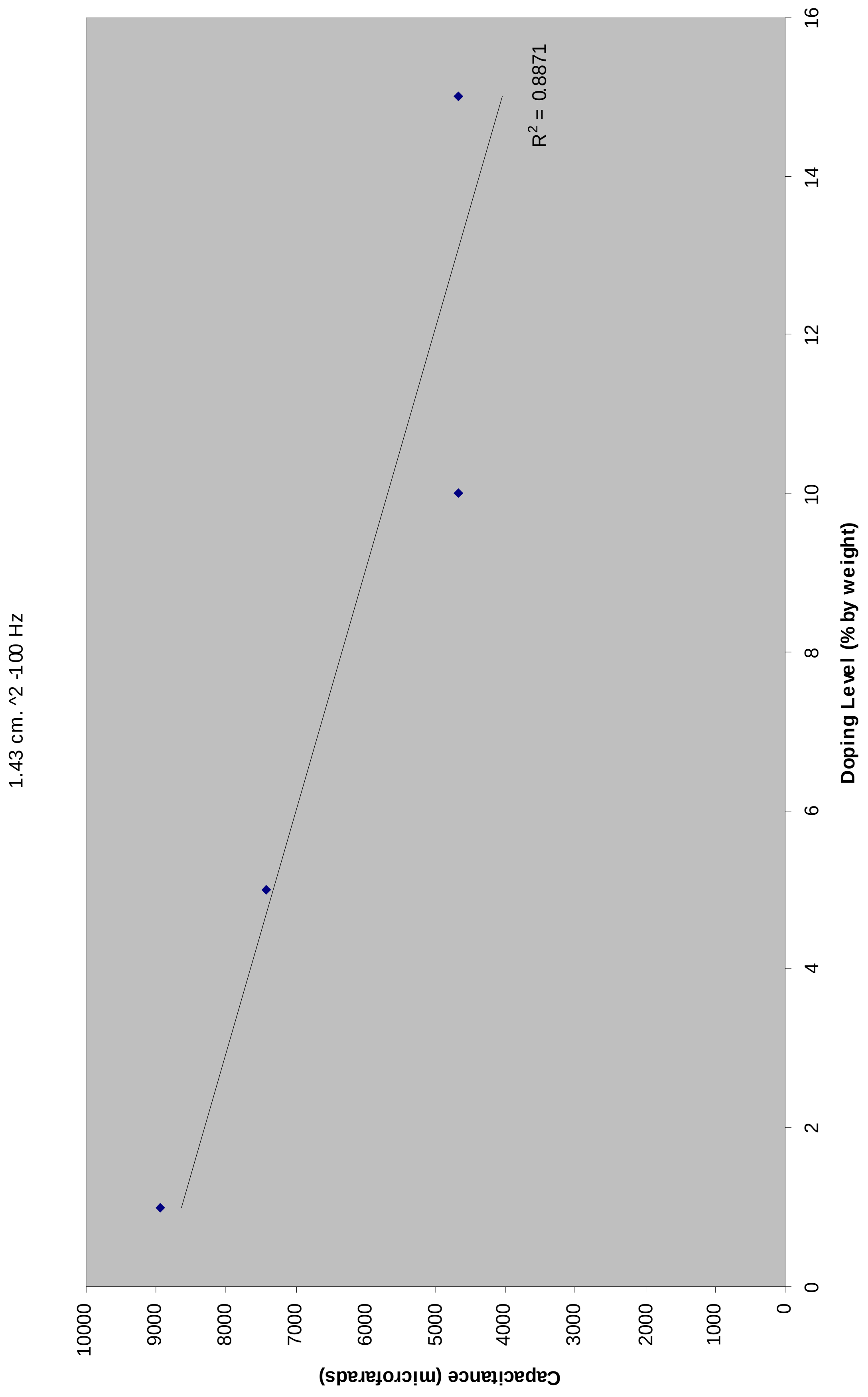
[6] “The Porphyrin Page” Avaialble from: http://www.washburn.edu/cas/chemistry/sleung/porphyrin/porphyrin_page.html, Accessed from: July 2005.

[7] P. R. Frail, K. Susumu, et al, Extraordinarily High Dark Electrical Conductivities within a New Class of Alkyl Ethynyl Porphyrins, Manuscript in preparation.

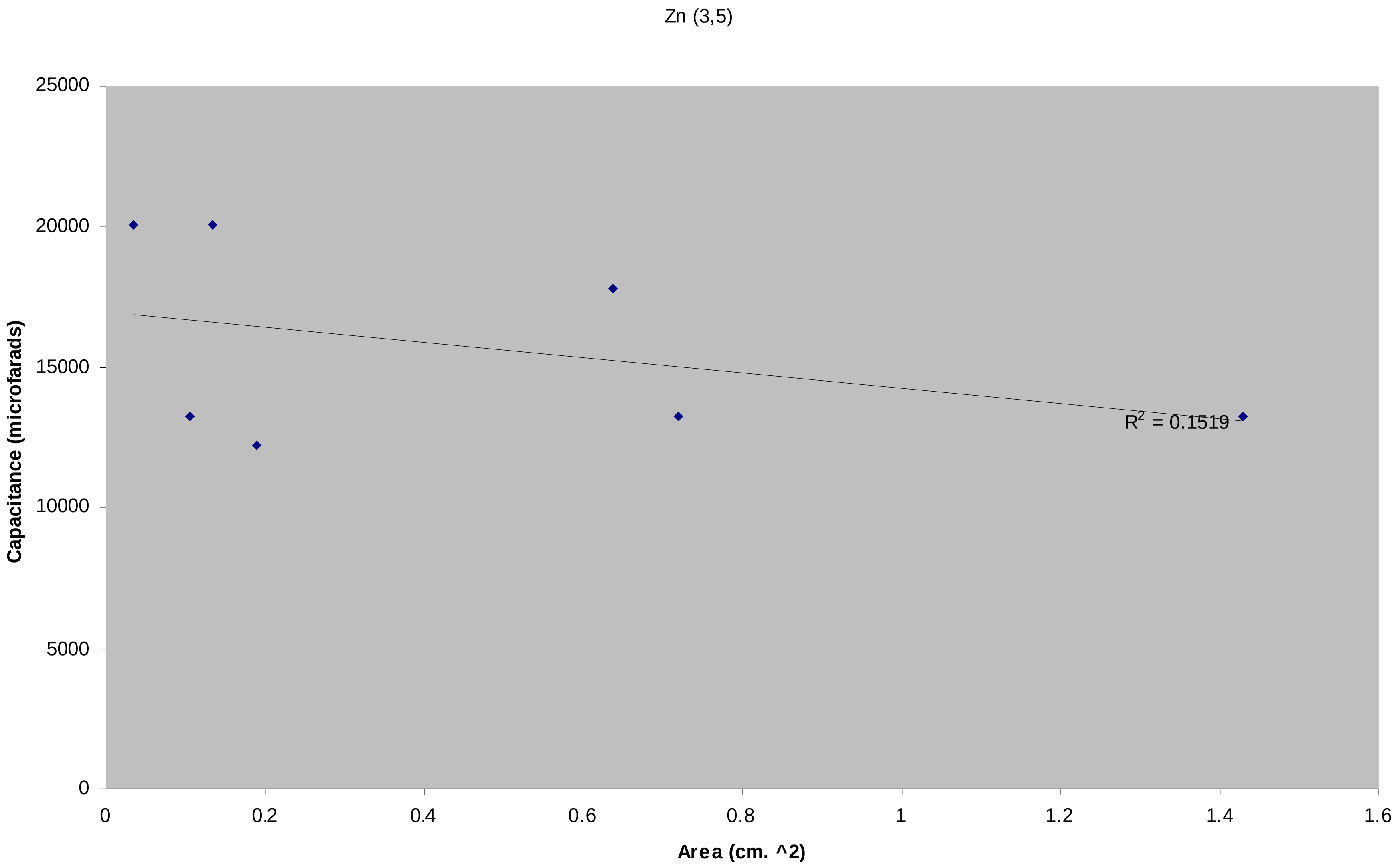
8. APPENDIX

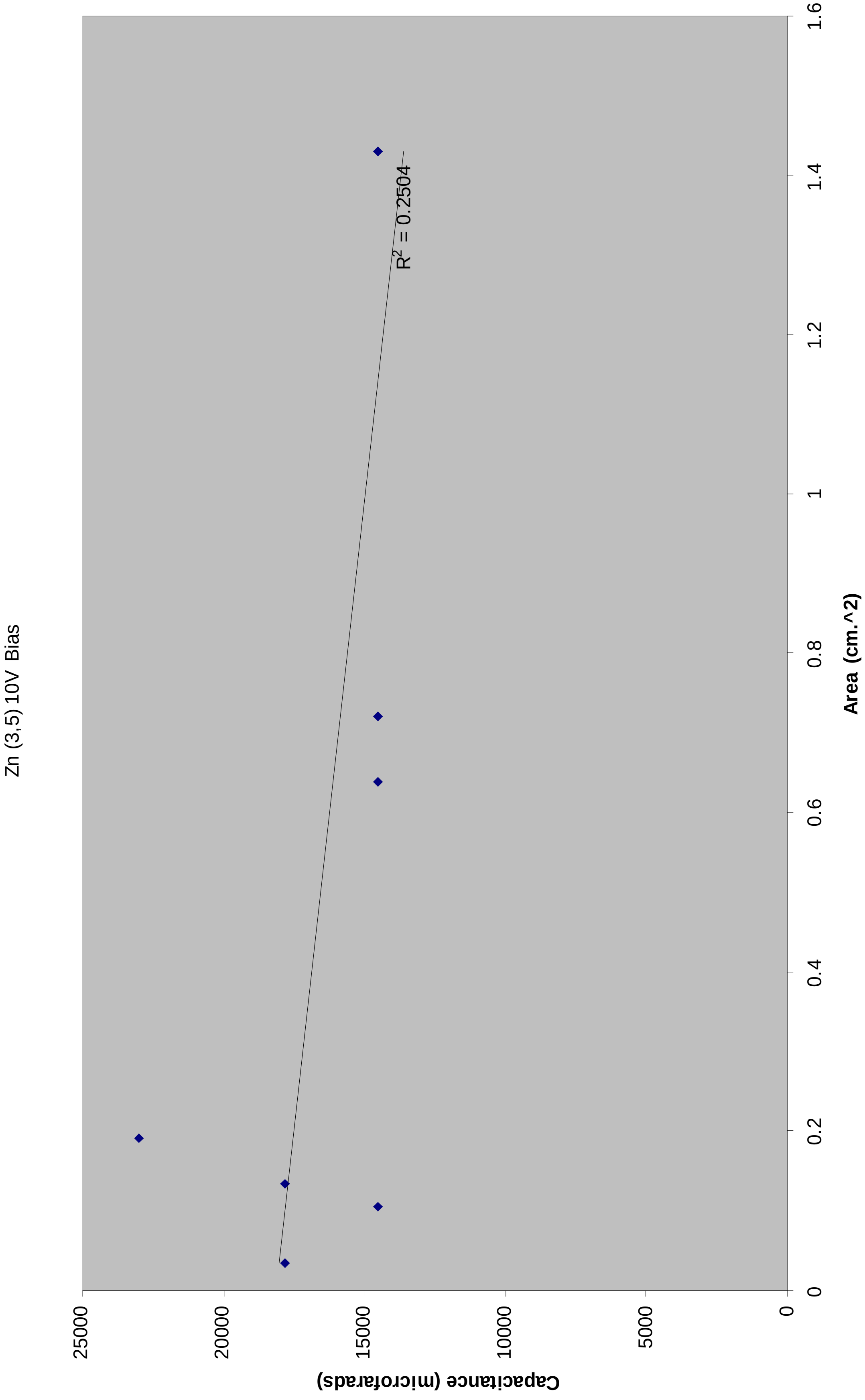

.03516 cm. ^2 -100 Hz
Capacitance (microfarads)
14000
12000
10000
8000
6000
4000
2000
0
Doping Level (% by weight)
0
2
4
6
8
10
12
14
16
$R^2 = 0.7146$


.1057 cm. ^2 -100 Hz
Capacitance (microfarads)
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
2
4
6
8
10
12
14
16
Doping Level (% by weight)
$R^2 = 0.7686$



1.43 cm. ^2 -100 Hz
$R^2 = 0.8871$
Capacitance (microfarads)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Doping Level (% by weight)
0
2
4
6
8
10
12
14
16


Zn (3,5)
Capacitance (microfarads)
25000
20000
15000
10000
5000
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Area (cm. ^2)
$R^2 = 0.1519$


Zn (3,5) 10V Bias
Capacitance (microfarads)
25000
20000
15000
10000
5000
0
Area (cm.^2)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
$R^2 = 0.2504$

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

Handheld Device for Remotely Measuring Brain Function

2005 NSF Summer Undergraduate Fellowship in Sensor Technologies
Adam Wang (Electrical Engineering) – University of Texas at Austin
Advisor: Dr. Britton Chance

ABSTRACT

Infants, and especially premature infants, are carefully monitored while in their incubators, and their brain health is of great concern. While it is currently the preferred procedure to attach a device to a patient’s forehead for monitoring pulse rate and oxygen levels in the brain, the development of a remote handheld system would make it possible to spot check patients from a distance, without having to deal with the obtrusiveness of attaching probes. Additionally, past studies have shown that measuring changes in blood volume and oxygen levels in the forebrain can be used to study brain function. Current work suggests that affordable and safe handheld devices for contact systems can be built from inexpensive components such as LEDs and photodiodes to measure these parameters.

The goals of this project were to study the aspects of, analyze the feasibility of, and suggest a design for a handheld device capable of remotely sensing brain function by employing the basic building blocks of near-infrared technology. As part of the project, experiments have been designed to simulate the uses – such as finding the arterial pulse and tracking changes in blood volume and oxygen levels in the brain – of a handheld remote sensing device for measuring brain health and function. In this study, remote sensing has demonstrated promising results for use over small distances. This paper also includes suggestions for extending remote sensing over greater distances for use in more practical, real world situations.

Table of Contents

1. INTRODUCTION

Approximately nine percent of babies are born prematurely, and many need special care in a Neonatal Intensive Care Unit (NICU) [1]. Since they are not fully developed, premature babies frequently suffer from health problems and are carefully monitored in their incubators. Common health concerns include apnea, anemia, respiratory distress syndrome, and patent ductus arteriosus, of which pulse rate, blood volume, and oxygen levels can be very significant indicators [2]. Furthermore, several recent studies have suggested that premature babies are impacted negatively by noise, light, and activity in their NICUs. Thus, many hospitals take care to maintain a quiet environment, shield the babies from light, and handle them slowly and deliberately [1]. A handheld remote sensing device for measuring a baby's brain health parameters such as blood volume and oxygen levels would provide a more discrete and desirable method for assessing these health indicators..

In addition to monitoring the brain health of babies, such a device could also be used to evaluate other brain functions. Current devices use contact methods for monitoring brain function during tasks such as problem solving or lying. A probe placed on the subject's forehead measures changes in blood volume or oxygen levels in certain regions of the brain [3]. These changes correspond to brain activity in the region of the brain measured. If, for example, a certain region of the brain displays increased blood volume after a subject tells a lie, monitoring this region for changes in blood volume can determine whether the subject tells a lie again. Moreover, determining which parts of the brain are more active, as determined by changes in blood volume or oxygen levels during certain tasks, can map out overall brain function. Past tests include determining which areas of the forebrain show activity during lying or solving anagram puzzles. A handheld device for remotely measuring brain function would offer portability and convenience to both the user and the subject.

2. BACKGROUND

Near-infrared (NIR) light has been used in recent years to non-invasively measure optical properties of tissue [4]. In particular, NIR light with a wavelength of 700-900 nm has optimal properties for measuring blood volume and oxygenation levels since most tissue – other than oxygenated and deoxygenated hemoglobin – absorb little light at these wavelengths. These qualities permit deep light penetration and backscattering from light sources such as white light, lasers, or LEDs, which can be measured by light detectors such as photomultiplier tubes or photodiodes.

In addition to its noninvasiveness, the low-cost and convenience of NIR imaging has been attracting much interest recently [5]. Some current NIR spectroscopy applications include brain functional imaging, breast cancer imaging, and muscle activity monitoring. For example, the finger pulse oximeter revolutionized hospital care with its ability to monitor arterial oxygen saturation and pulse rate while still being portable, noninvasive and capable of providing continuous real-time monitoring [6].

Given the success of contact model NIR devices, some current studies are moving toward remote sensing using NIR light. The goal of these studies is to combine the functionality of contact model NIR devices with the versatility of remote sensing, which might eliminate the subject's awareness of being tested altogether. One such contact model NIR idea of interest is brain oxygen monitoring in premature infants. The use of such a device could be extended to monitoring comatose or brain-injured patients and to measuring brain function.

Depending on the need, various NIR systems – including continuous-wave systems and time-resolution systems – can be used. A time-resolution system can measure the absorption coefficient and reduced scattering coefficient of tissue to find the absolute concentrations of deoxygenated hemoglobin (Hb) and oxygenated hemoglobin (HbO_2). From these two values, the absolute blood volume and oxygenation level can be determined [5]. However, this method relies on accurate measurement of photon arrival times that are backscattered from the tissue [7]. A much simpler method for NIR spectroscopy is a continuous-wave system, which merely emits a constant light at tissue and measures the backscattered light. An important limitation of a continuous-wave system is that it can only measure changes in Hb and HbO_2. Thus, only changes in blood volume and changes in oxygen levels can be measured [5].

All NIR methods use a NIR light source to illuminate tissue. The tissue surface reflects some of the incident light, while the rest enters the tissue. Light inside the tissue is either absorbed or scattered about until some of it reemerges. A light detector then measures this backscattered light. In the case of a continuous-wave system, the change in backscattered light is measured. Since Hb and HbO_2 are the greatest sources of absorption of the NIR light used, only these two are considered in the Beer-Lambert Law

$$I = GI_0 e^{-\left(\alpha_{Hb} C_{Hb} + \alpha_{HbO_2} C_{HbO_2}\right) \cdot L} \tag{1}$$

where I is the light intensity after absorption and backscattering; G is a constant attenuation; I_0 is the input light power; α_{Hb} and α_{HbO2} are the molar extinction coefficients of deoxygenated and oxygenated hemoglobin, respectively, and determine the amount of light absorbed; C_{Hb} and C_{HbO2} are the concentrations of deoxygenated and oxygenated hemoglobin, respectively; and L is the photon path length, which can be determined experimentally and varies based on the particular setup of the light source and detector [5]. Note that α_{Hb}, α_{HbO2}, and L are all functions of the light wavelength used. Water and other tissues have absorption coefficients that are orders of magnitude lower than Hb or HbO_2 and are not considered when dealing with NIR light [6].

A continuous-wave system needs at least two light sources of different wavelengths to operate. Let I'_{760} be the backscattered light intensity from a 760 nm light source at a predetermined "baseline state," and let C'_{Hb} and C'_{HbO2} be the unknown concentrations of deoxygenated and oxygenated hemoglobin concentrations at the baseline state. Then, for a light source of 760 nm, let the optical density (OD) be

$$
\begin{aligned}
OD_{760} &= \ln\left(\frac{I'_{760}}{I_{760}}\right) = \ln\left(\frac{GI_0 e^{-(\alpha_{Hb,760}\cdot C_{Hb} + \alpha_{HbO_2,760}\cdot C'_{HbO_2})\cdot L_{760}}}{GI_0 e^{-(\alpha_{Hb,760}\cdot C_{Hb} + \alpha_{HbO_2,760}\cdot C_{HbO_2})\cdot L_{760}}}\right) \\
&= -\left(\alpha_{Hb,760}\cdot\left(C'_{Hb} - C_{Hb}\right) + \alpha_{HbO_2,760}\cdot\left(C'_{HbO_2} - C_{HbO_2}\right)\right)\cdot L_{760} \\
&= \left(\alpha_{Hb,760}\cdot \Delta C_{Hb} + \alpha_{HbO_2,760}\cdot \Delta C_{HbO_2}\right)\cdot L_{760}
\end{aligned}
\tag{2}
$$

where I_{760} is the backscattered light intensity from the 760 nm light source at some other state, and $\Delta C'_{Hb}$ and $\Delta C'_{HbO2}$ are the changes in concentrations of deoxygenated and oxygenated hemoglobin, respectively, from the baseline state. Similarly,

$$
OD_{830} = \ln\left(\frac{I'_{830}}{I_{830}}\right) = \left(\alpha_{Hb,830}\cdot \Delta C_{Hb} + \alpha_{HbO_2,830}\cdot \Delta C_{HbO_2}\right)\cdot L_{830} \tag{3}
$$

Thus, by using two wavelengths, the amount of change in Hb and HbO_2 from the baseline state can be determined:

$$
\Delta C_{Hb} = \frac{\alpha_{HbO_2,830}\cdot OD_{760}/L_{760} - \alpha_{HbO_2,760}\cdot OD_{830}/L_{830}}{\alpha_{Hb,760}\cdot \alpha_{HbO_2,830} - \alpha_{HbO_2,760}\cdot \alpha_{Hb,830}} \tag{4}
$$

$$
\Delta C_{HbO_2} = \frac{\alpha_{Hb,760}\cdot OD_{830}/L_{830} - \alpha_{Hb,830}\cdot OD_{760}/L_{760}}{\alpha_{Hb,760}\cdot \alpha_{HbO_2,830} - \alpha_{HbO_2,760}\cdot \alpha_{Hb,830}} \tag{5}
$$

Then, ΔBV, the change in blood volume, and ΔOXY, the change in oxygenated blood, can easily be calculated:

$$
\Delta BV = \Delta C_{Hb} + \Delta C_{HbO_2} \tag{6}
$$

$$
\Delta OXY = \Delta C_{HbO_2} - \Delta C_{Hb} \tag{7}
$$

If a third light source is used with a wavelength of 800 nm, at the isosbestic point [Fig. 1], then the change in blood volume, ΔBV, could be validated by OD_{800} alone since the molar extinction coefficients of Hb and HbO_2 are the same at this wavelength.

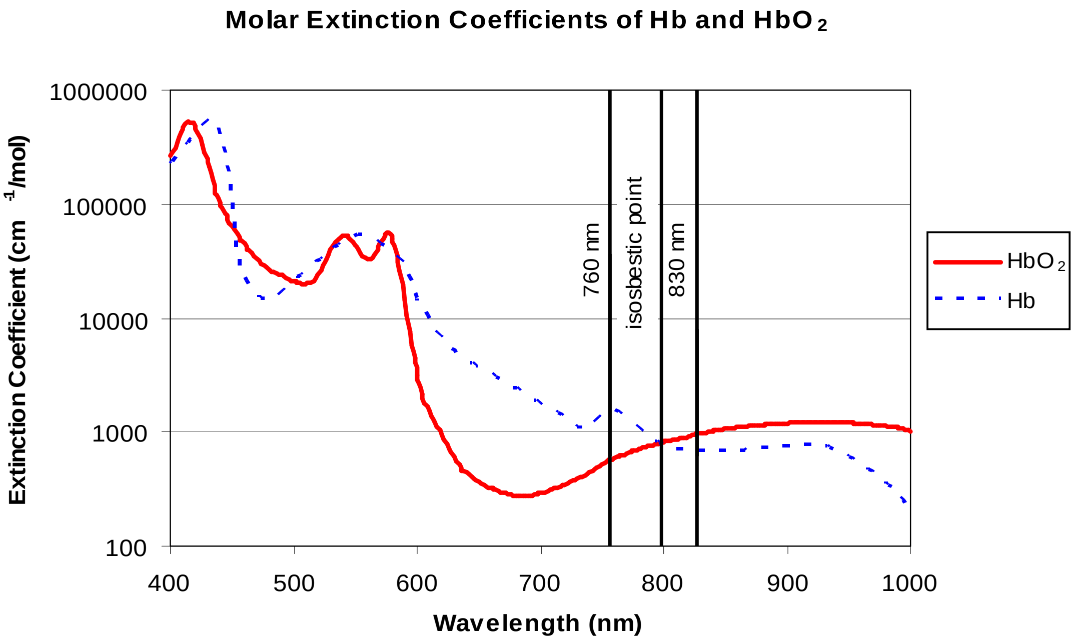


Figure 1: Molar extinction coefficients of Hb and HbO_2, as a function of wavelength, plotted on a logarithmic scale [8].

When monitoring brain health with a handheld remote sensing device, if a predetermined "healthy state" is used as the baseline state, then any changes or deviations in blood volume or blood oxygenation levels can be tracked to gauge brain health. Depending on the case, a tolerable range of fluctuation or deviation should be allowed, but once the change in blood volume or oxygenation levels from the predetermined healthy state exceeds these bounds, the device should alert the device operator. Similarly, when monitoring brain activity, a predetermined "resting state" could be used as the baseline state, and any changes in blood volume or blood oxygenation levels in a region of the brain could track brain function.

3. PROJECT GOALS

The goals of this project were to study the aspects of, analyze the feasibility of, and suggest a design for a handheld device capable of remotely sensing brain function by employing the basic building blocks of near-infrared technology. While it is currently the preferred procedure to attach a device to a patient's forehead for monitoring pulse rate and oxygen levels in the brain, the development of a remote handheld system would make it possible to spot check patients from a distance, without having to deal with the obtrusiveness of attaching probes. Current work suggests that affordable and safe handheld devices for contact systems can be built from inexpensive components such as LEDs and photodiodes. These components are small, safe, and operate at low voltage levels. A general system that can remotely measure brain function must include a light source and driver, a light detector, and a processing unit [Fig. 2].

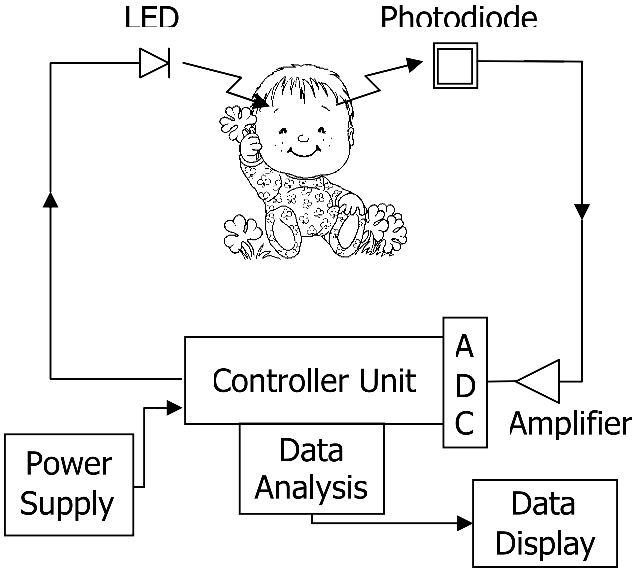


Figure 2: Block diagram of handheld device for remotely sensing brain function.

For this study, fundamental building blocks were selected to find the minimum requirements to build an effective, durable device at low cost. In addition to being small and cheap, LEDs generally have greater subject acceptance, and due to their more diffuse nature, the Food and Drug Administration does not limit their power when applied to humans. This provides LEDs with an advantage over lasers, which are subject to more federally determined limitations [5]. Due to their ease of use and versatility, photodiodes were selected as light detectors for this study. Unlike photomultiplier tubes, they do not require high voltages and can operate at light levels that would easily damage photomultiplier tubes due to the photomultiplier tubes' great sensitivity. Finally, an ordinary programmable microcontroller was selected to control the LED, sample the photodiode response with its built-in analog-to-digital converter (ADC), and process the sampled data. The object of much of this study was to determine the efficacy of these components when used together.

4. EXPERIMENTAL SETUP

Since the data measured with the device must be analyzed, there must be a way to sample, store, and interpret it. As with any digital device, an analog signal must be quantized in order to be sampled; the continuous range of the analog signal is broken up into a finite set of discrete values, to which the analog signal is mapped. In the case of the ADC selected for this study, a 0-5 V signal is uniformly quantized into a 10-bit value (1024 quantization levels), with 0 corresponding to 0 V and 1023 corresponding to 5 V. Thus, the precision of this particular ADC is $5/2^{10}$, or approximately 5 mV. The dominant problems with quantization are lack of precision and quantization error; this can be resolved by closely matching the quantized range to the range of the signal and by increasing the precision with more quantization levels. However, for a fixed range, fixed precision system, the only remedy is to use an analog system to fit the signal into the full dynamic range of the ADC.

For example, if the component of the desired signal is too small, the discrete nature of the quantization levels may not be able to detect the differences of the actual continuous, infinite precision values. In this situation, all that can be done is to extract the desired signal and amplify it to better fit the dynamic range of the ADC so that more accurate sampling can be done [Fig. 3].

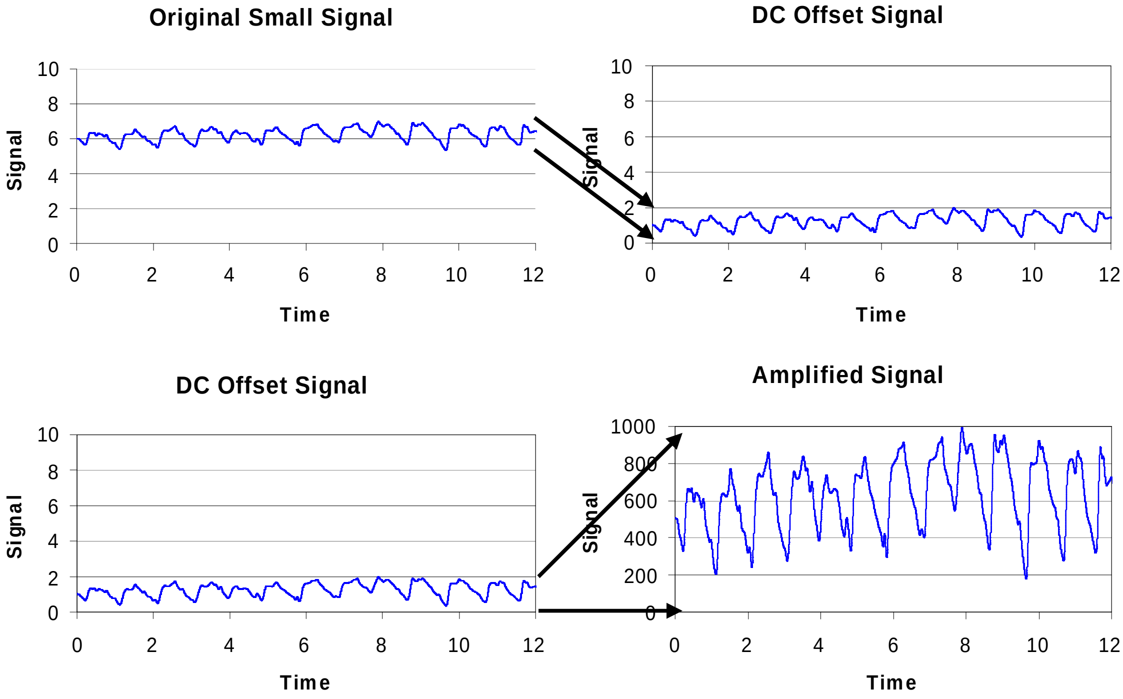


Figure 3: The original small signal's DC component is first reduced, and then the signal is amplified to better match the ADC's full dynamic range.

This must be done with analog components, which are continuous by nature. The circuit built for this project performs this exact function by first introducing a DC offset to crop off the DC component. Then the signal is amplified to better fit the dynamic range of the ADC before it is sampled. Nonetheless, due to the discrete nature of quantization, noise is introduced. Even for an ideal *Q*-bit analog to digital converter, the theoretical Signal to Noise Ratio (SNR) is [9]:

$$SNR_{ADC} = (1.763 + 6.02 \cdot Q) \text{ dB}. \tag{8}$$

Thus, for the 10-bit ADC used, the theoretical maximum SNR is 62.0 dB, which despite being an upper bound, is a very impractical bound; noise introduced from numerous other sources greatly degrades the signal, and the SNR is not limited by the ADC quantization.

Much of the circuit used was built from op-amps, a fundamental element of circuits. These conceptually simple devices have a wide variety of uses ranging from logic comparators to amplifiers to integrators. The first step necessary to take advantage of the ADC's full dynamic range is to remove the DC component so that only the component of the signal of interest is amplified. This was accomplished by subtracting a DC offset ranging from 0 V to +5 V that could be adjusted by a potentiometer. Next, another op-amp configuration amplified the signal with a gain up to 100, adjusted by a 100 kΩ potentiometer [Fig. 5].

Depending on the need, the Motorola 9S12C32 microcontroller was programmed to sample at certain frequencies or to drive two LEDs (760 and 830 nm). For simplicity, the sampled data was stored and analyzed offline. As the project progressed, the necessity of using higher quality components became apparent. A Wratten gelatin light filter #89B placed over the photodiode filtered out undesired wavelengths [Fig. 4], and greatly aided in reducing noise since photodiodes respond to a wide range of wavelengths.

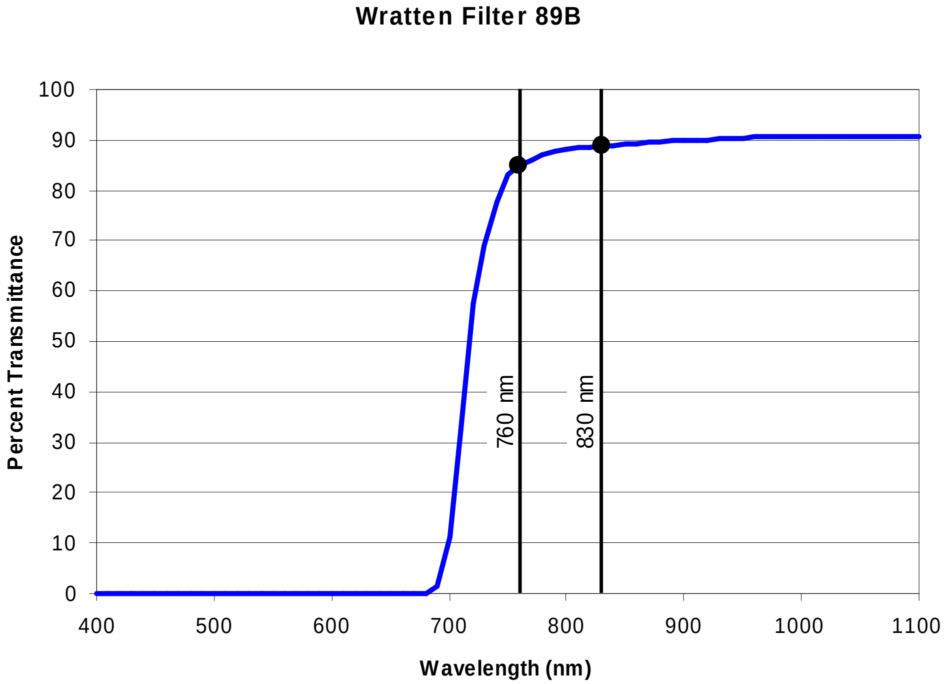


Figure 4: Wratten filter #89B’s percent transmission curve as a function of wavelength [10]. The filter minimally affects the two LEDs used while lower frequencies (visible light) are blocked.

Higher precision, lower noise OP27 op-amps replaced the more standard op-amps originally used (80 nV compared to 1000 nV peak-to-peak noise in the 0.1 to 10 Hz range), and the OPT101 photodiode was replaced by the FDS1010 Si photodiode (9.7 × 9.7 mm), which has 18 times the surface area for collecting light. When the passive FDS1010 Si photodiode was implemented, the leads were fed through a difference amplifier with a pre-amp gain of 10 [Fig. 5]. Furthermore, a first order RC low-pass filter was added to reduce high frequency noise in the signal, such as the 60 Hz AC component in power from wall outlets and room light.

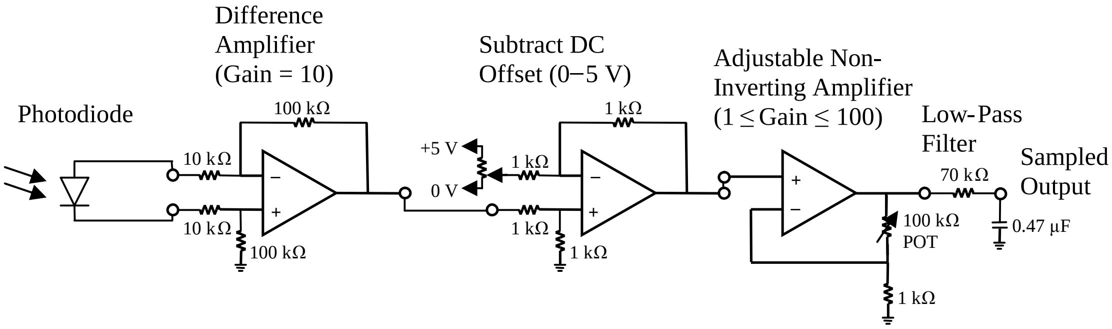


Figure 5: Circuit diagram showing op-amps (±12 V power supply) used to extract signal throughout this project.

5. EXPERIMENTAL RESULTS

Two aspects of remote sensing were explored: 1) an attempt to find the arterial pulse and 2) an attempt to track changes in light absorption of an ink test via remote sensing. The former illustrates the need for finding the pulse rate of the subject, and the latter simulates changes in concentrations of Hb or HbO_2.

5.1 Arterial Pulse

Starting with the contact model, where both the LED and photodiode are in contact with the subject, the arterial pulse is easily extracted. An LED shines light through the subject’s thumbnail, and a photodiode measures the transmitted light through the thumb. The amount of blood at the tip of the thumb fluctuates in relation to the arterial pulse. This rhythmic fluctuation is the basis for determining the pulse rate. As the blood volume increases, the amount of light transmitted decreases due to the greater absorption of NIR light by Hb and HbO_2. The raw data was run through a digital 4th order Butterworth low-pass filter to clean up the signal. Using a Discrete Fourier Transform, the pulse rate of the subject (in this case, the author) was determined over a sampling duration of 30 seconds to within 2 beats per minute to be 74 beats per minute [Fig. 6]. This was validated by counting the pulse rate on the wrist. In this setup, the presence of room light was not a problem since it merely contributed to the transmitted light and was still modulated by the arterial pulse.

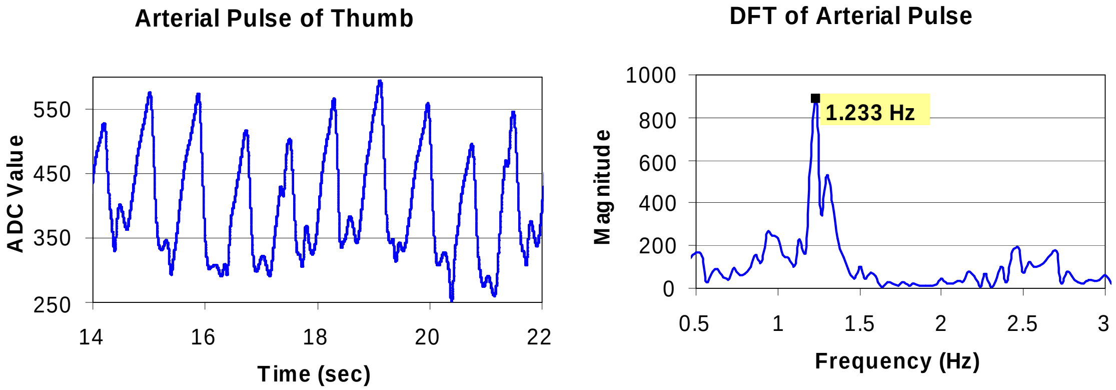


Figure 6: Measured transmitted light (left) through thumb with a contact setup, and its frequency spectrum (right), showing a pulse rate of approximately 74 beats per minute.

After showing that the contact model could be reproduced, a small-scale setup for remote sensing [Fig. 7] was built to test the feasibility of remote sensing. Neither the LEDs nor the photodiode were in direct contact with the tissue; instead, they were fixed a small distance from the tissue, as a step toward achieving distances more likely for remote sensing. A barrier prevents light from the LED from directly illuminating the photodiode and reduces reflected light from the surface of the tissue. Over the course of hundreds of tests, it was determined that extracting the arterial pulse from the measured backscattered light required: 1) higher quality op-amps, 2) low-noise OP27 chips, 3) the addition of a light filter, 4) the change to a greater surface area FDS1010 Si photodiode, and 5) the reduction of incident room light. Even with these enhancements, extracting the arterial pulse from this scaled down setup of remote sensing proved difficult because of the small signal and noise.

Small-Scale Remote Sensing

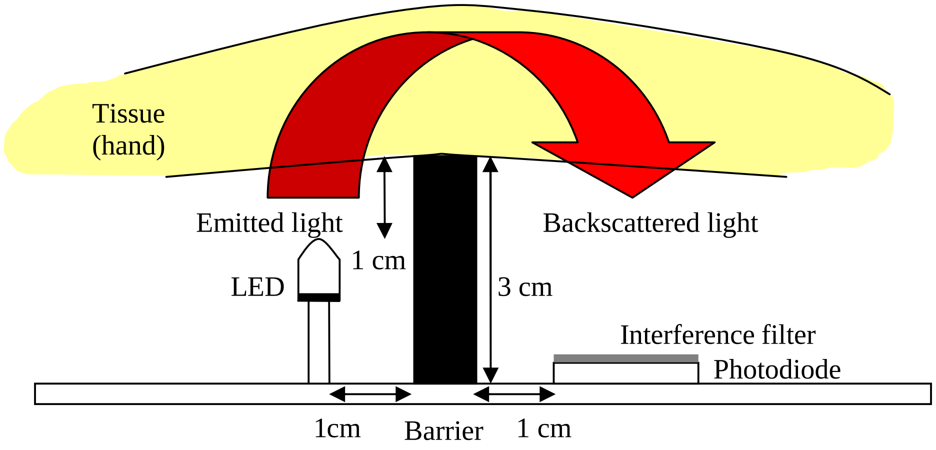


Figure 7: Small-scale remote sensing setup with neither the LED nor the photodiode in contact with the tissue.

In order to utilize Equations 2-7, which can determine the change in blood volume or oxygen levels in tissue, two LEDs are necessary. To monitor the change in the intensity of backscattered light from each wavelength, the two LEDs cannot emit light simultaneously; instead, they are time-shared and alternately emit light at a rate much greater than the rate of change of the parameter of interest. For finding the arterial pulse, the microcontroller controls a 760 nm and an 830 nm LED by flashing them alternately for a 100 ms duration each in the small-scale remote sensing setup. The average backscattered light intensity is different for each LED, but the arterial pulse modulates the backscattered light for both wavelengths [Fig. 8].

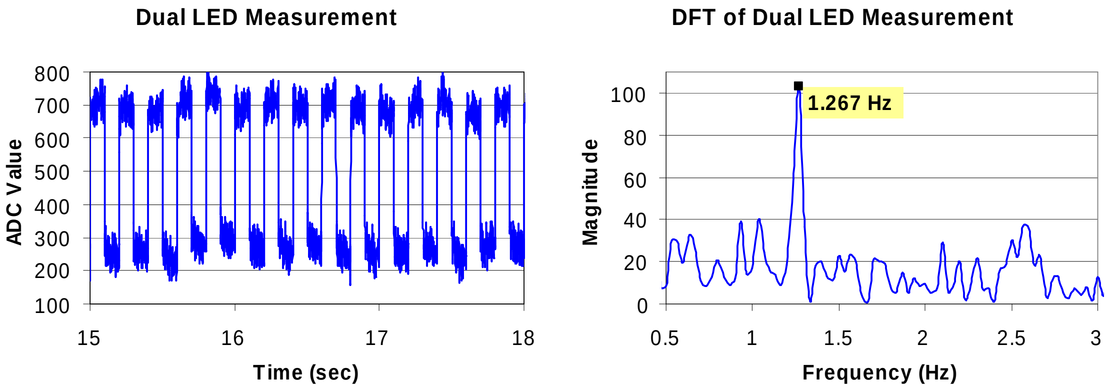


Figure 8: Measured backscattered light (left) from the author's hand with the small-scale remote sensing setup and two time-shared LEDs; the frequency spectrum of the backscattered light (right), showing a pulse rate of approximately 76 beats per minute modulating the backscattered light.

5.2 Ink Tests

In addition to remotely sensing a subject's pulse rate, another goal for the handheld device is to monitor changes in blood volume and oxygen levels in the brain. To achieve this, the handheld device must be sensitive to changes in concentrations of Hb and HbO_2, which can be simulated by an ink test. An intravenous fat emulsion solution in a large plastic beaker provided the basis for a scattering medium that allows incident light to be backscattered [Fig. 9]; a photodiode collects the backscattered light. Ink is slowly added to change the amount of light absorption, just as a change in the concentration of Hb or HbO_2 would change the amount of light absorption. Various setups were tested, with varying distances between the beaker and the light barrier, with both the 760 and 830 nm LEDs, and with and without the presence of room light.

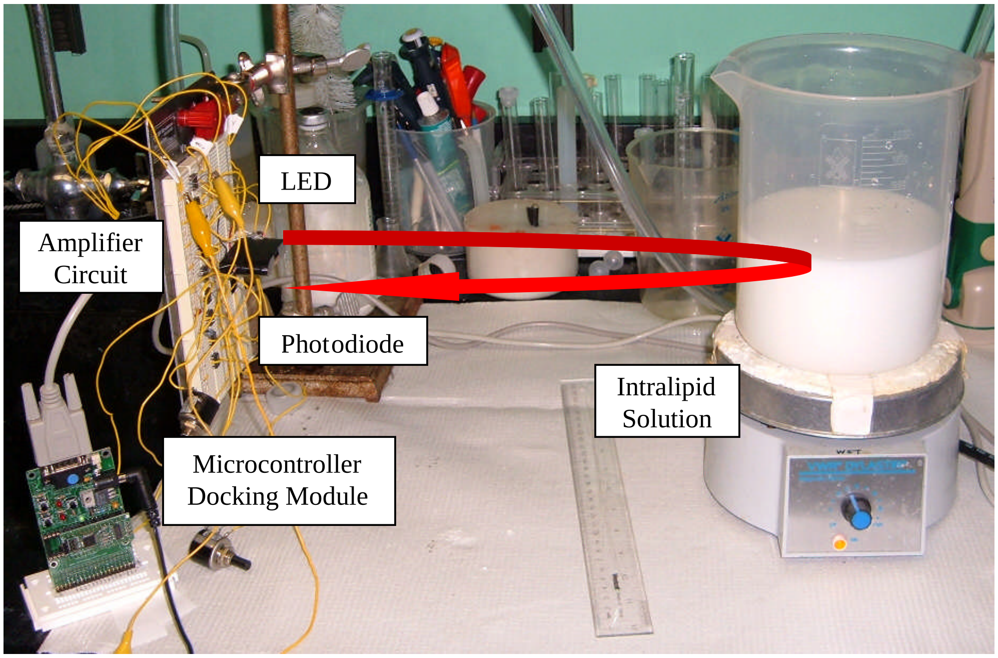


Figure 9: Remote sensing ink test setup with handheld device components.

The photodiode sampling for the ink tests was different from the arterial pulse tests since no specific frequencies were being tested for; only the general trend over a long period of time was being tested for. For example, with the beaker 10 cm from the LED, no room light, and a 760 nm LED, the measured backscattered light shows clear decreases as 1 mL of ink is added approximately every minute [Fig. 10] to the 1 L of intralipid solution. Each vertical division marks the addition of ink to the intralipid solution, which increases the absorption of light and reduces the amount of backscattered light. Before each addition of ink, the signal is stabilized and shows little change; as ink is added, a press of a button on the microcontroller's docking module records the action of adding ink.

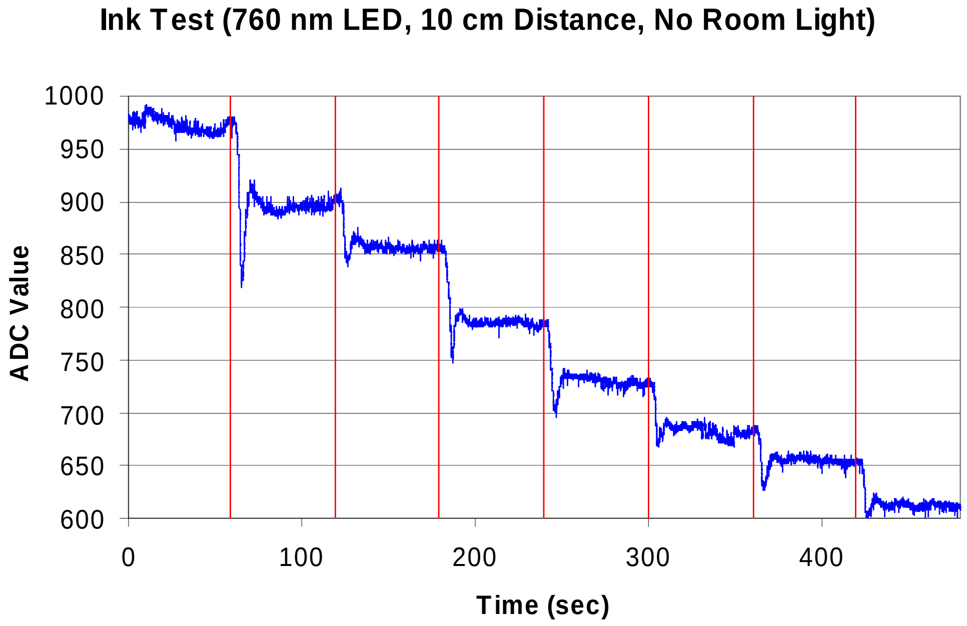


Figure 10: Ink test with a 760 nm LED at a distance of 10 cm from the beaker with no room light. The backscattered light intensity drops with each addition of ink, marked by the red vertical lines.

While the 760 nm and 830 nm LEDs showed similar effectiveness in the different ink test configurations, the distance between the beaker and LED and whether room light was present greatly affected the ability of the system built to determine the decrease in backscattered light when ink was added. Adding room light to the configuration used to obtain Figure 10 injects significantly more noise to the raw data and requires twice the gain to extract a similar signal even with the Wratten light filter; changing the distance from 10 cm to 30 cm without room light also significantly weakens the signal and requires five times the gain to extract the signal [Fig. 11]. At a distance of 30 cm between the beaker and the LED, with the presence of room light, the downward trend in backscattered light due to the addition of ink could no longer be measured; at a distance of 60 cm without the presence of room light, this could no longer be accomplished.

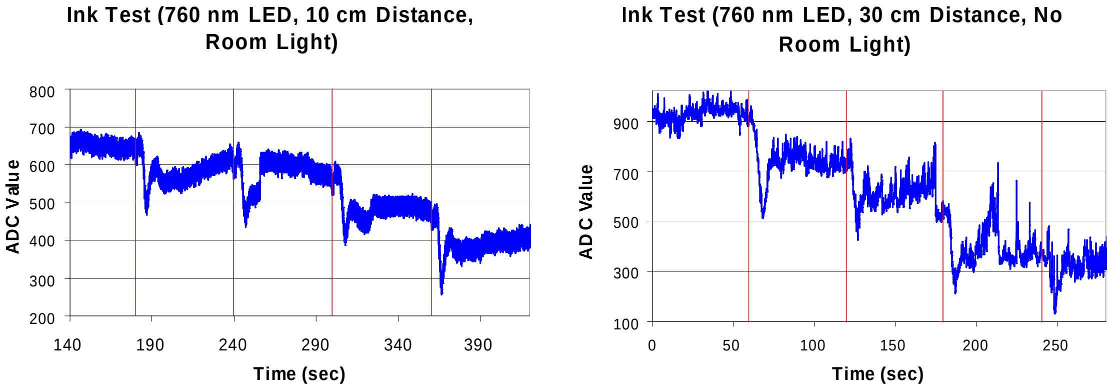


Figure 11: Ink tests with room light present (left) and at a greater distance (right), show a general degradation of the signal but still demonstrate a response to ink being added.

6. DISCUSSION

While the basic components used in this study have shown some success in finding the pulse rate and tracking changes in absorption values with remote sensing, there is much left to be desired. Even small-scale remote sensing detection of the arterial pulse required careful shielding from room light. The ink tests also showed that distance between the subject and device and the presence of room light greatly affected the ability to accurately track changes in light absorption. More realistic remote sensing scenarios for such a device might require distances of up to several feet between the subject and the device in the presence of room light. High gain circuits further exacerbate the problem of high sensitivity, requiring an extraordinarily stationary subject and operator. Furthermore, room light injects significantly greater noise, especially as the signal of interest weakens with distance, with a large 60 Hz component that must be diminished by a low-pass filter.

When two LEDs of different wavelengths are time-shared – as they must be to monitor changes in blood volume or oxygen levels – the backscattered light intensity from the two different wavelengths will differ. Therefore, if the same gain circuit amplifies the signal from both LEDs, the difference in backscattered light intensity between the two LEDs will be amplified, and both signals may not fit in the ADC's range. To solve this problem, the gain could be lowered to accommodate for the different backscattered light intensities. However, this change causes the signal from each LED to be smaller too. Either the LED power should be adjusted so that the backscattered light intensities from the two LEDs are very close in value, or, better yet, each wavelength should have its own DC offset and gain circuit that the microcontroller could sample at the appropriate time.

Since the variable volume of blood due to the arterial pulse only accounts for 5% of the total blood volume in tissue [6], finding the arterial pulse from backscattered light would be much more challenging than tracking general changes in the overall trend of blood volume and oxygen levels in the brain. The ink tests demonstrated the *qualitative* ability of the handheld device components to track changes in absorption values; however, a *quantitative* indicator is needed. To accomplish this, effective blood tests must closely simulate the scattering and absorption coefficients of human tissue and quantitatively calibrate a system with known changes in the absorption coefficients at the wavelengths used.

At the end of the project, the issue of bandwidth and shot noise arose. Shot noise, which is inherent to photodiodes, is proportionally related to the square root of the bandwidth of the photodiode [5]. Therefore, in order to limit shot noise, the bandwidth of the entire system should be limited to the signal of interest, which, at its greatest frequency, would probably not exceed a few hertz. Some tests showed that simply adding a capacitor in parallel to the negative feedback resistor of the difference amplifier could significantly reduce shot noise. The system originally built did not take shot noise into consideration and had a bandwidth of approximately 600 Hz. Adding a 47 μF capacitor that would reduce the bandwidth to approximately 1 Hz still allows for the detection of the arterial pulse but reduces shot noise by almost a factor of 25 [Fig. 12]. Further tests would be required to determine the effect of shot noise.

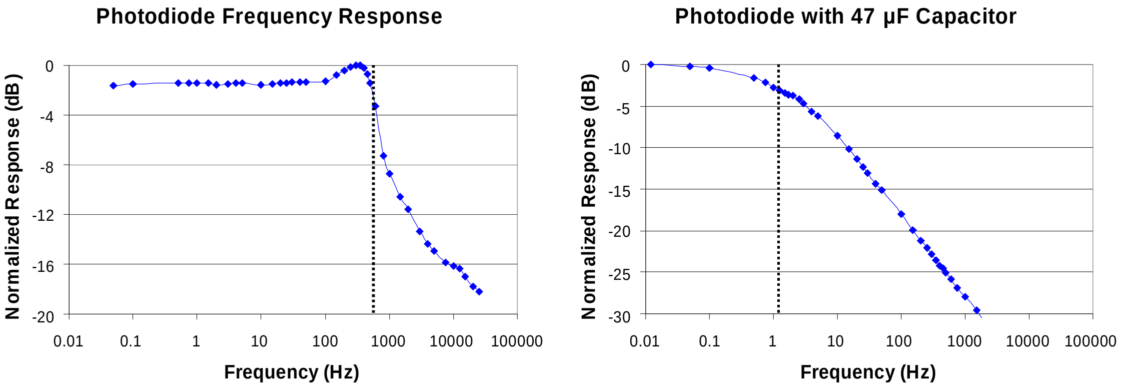


Figure 12: System built for this study (left) with a bandwidth of about 600 Hz that can be improved by filtering noise early on by limiting the bandwidth.

7. RECOMMENDATIONS

Due to the weak signal and the high level of noise inherently present in remote sensing, sources of noise should be targeted aggressively – clean power sources and low-noise components are essential. The Wratten gelatin filter used was a good start for reducing unwanted room light, but a better, more expensive solution would be a narrow bandpass lens coating that would exclusively pass the wavelengths of the LEDs. Furthermore, the gelatin filter is sensitive to cleaning, and a handheld remote sensing device should be durable and hardy, considering the more rigorous uses intended outside of the lab.

A weak signal entails the need for a high gain circuit, which also unfortunately amplifies sampled noise. Thus, various measures could be enacted to obtain a stronger signal, such as reducing ambient lighting, increasing the LEDs' power, minimizing the distance between the device and the subject, and maximizing the collecting area of the backscattered light. An array or multiple LEDs or photodiodes could be explored as a possibility to increase the effective illumination intensity or effective collecting area of the handheld device. A simple way to increase the effective collecting area of the photodiode would be to use a large lens to collect light and focus it onto the photodiode.

Since the change in blood volume or oxygen level in the brain is determined by the change of measured backscattered light intensity, care must be taken to limit any movement of the subject or the operator of the handheld remote sensing device. Slight angular movements of either the device or the subject could change the amount of incident light intensity from the LED onto the subject or backscattered light intensity from the subject onto the photodiode [Fig. 13]. Additionally, if the LED is considered as a point source, the light intensity varies inversely with the square of the distance from the source to the subject. In the presence of room light, even a passerby's shadow would dramatically change the readings due to the high gain necessary to accurately sample the photodiode.

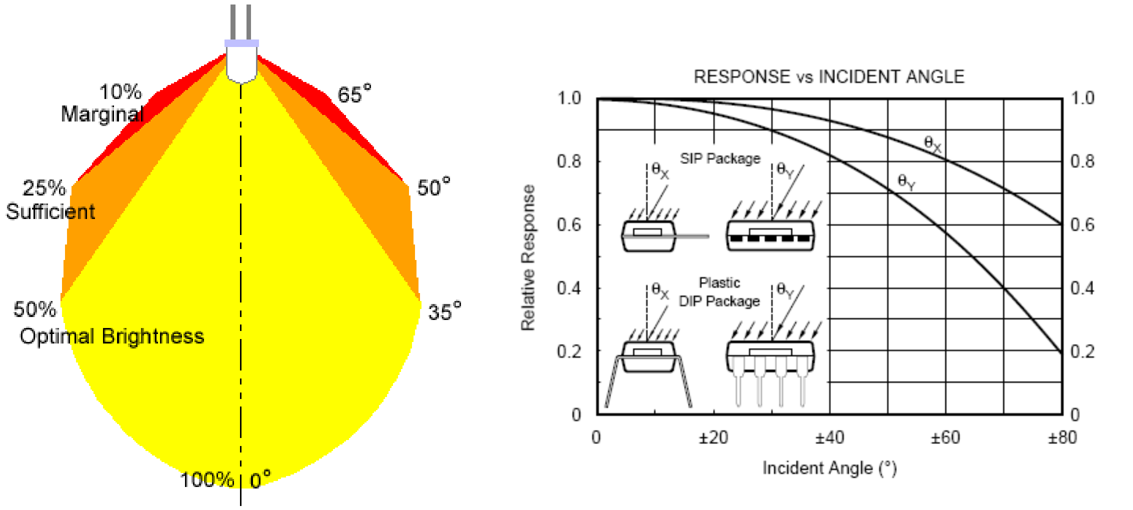


Figure 13: Relative brightness of an LED (left) [11] and relative response of a photodiode (right) [12], based on the incident angle.

The microcontroller selected for this study would not have any trouble performing the calculations demanded by Equations 2-7 to monitor changes in blood volume or oxygen levels in the brain. Its I/O ports could digitally display real-time data on an LCD, light up various indicator LEDs as to the subject's health, read user inputs, and sound audible warning alarms. Moreover, instead of the analog potentiometer knobs used for this study, the microcontroller could control more precise digital potentiometers; the gain and DC offset could be adjusted automatically based on the need to keep the signal in the ADC's range and to take advantage of the ADC's full dynamic range. All changes to the potentiometers would be recorded and taken into consideration in the microcontroller's calculations. Nonetheless, a Digital Signal Processor may be desired for its generally greater computational power, ADC range, and quantization precision.

8. CONCLUSION

This study has shown the feasibility, challenges, and limitations of remote sensing with an LED, photodiode, and microcontroller. Together, these components could provide a compact, durable, and effective handheld device if a consistent and stable signal can be extracted. Small-scale remote sensing tests have demonstrated the ability to determine the pulse rate of a subject and changes in blood volume or oxygen levels in experiments designed to simulate measuring brain health or brain function. Further study that implements some of the recommendations of this paper could very possibly bring remote sensing in the real world to reality. Even so, the exciting possibility of remote sensing with a handheld device requires much additional work before it can be an effective technology.

9. ACKNOWLEDGEMENTS

I would like to thank everyone who has helped to make this project possible for me. First, I would like to extend my deepest gratitude to Dr. Britton Chance and his lab for their continuing support and valuable ideas. Everyone working there expressed a strong interest in each other's work, and Dr. Chance continually offered advice, suggestions, and critique. In addition to the Chance lab, I truly appreciate the effort Dr. Jan Van der Spiegel put into making SUNFEST a rewarding summer experience. I am also very grateful to the University of Pennsylvania and the National Science Foundation for supporting this wonderful Research Experience for Undergraduates. Finally, and most importantly, I cannot forget to acknowledge my loving family's encouragement throughout my life.

10. REFERENCES

[1] FamilyFun. *FamilyFun: Health Dictionary: Prematurity*. <http://familyfun.go.com/parenting/child/health/childhealth/dony89enc_prem/> (24 July 2005).

[2] Nemours Foundation. *A Primer on Preemies*. July 2004. <http://kidshealth.org/parent/growth/growing/preemies.html> (24 July 2005).

[3] B. Chance, S. Nioka, and Y. Chen, Shining new light on brain function, *SPIE's OEMagazine*, 3 (2003) 16-19.

[4] Z. Zhao, X.C. Wang, and B. Chance, Remote sensing of prefrontal cortex function with diffusive light, *SPIE*, 5616 (2004) 103-111.

[5] Y. Lin, G. Lech, S. Nioka, X. Intes, and B. Chance, Noninvasive, low-noise, fast imaging of blood volume and deoxygenation changes in muscles using light-emitting diode continuous-wave imager, *Rev. Sci. Instrum.*, 73 (2002) 3065-3074.

[6] A. Zourabian, A. Siegel, B. Chance, N. Ramanujan, M. Rode, and D. Boas, Trans-abdominal monitoring of fetal arterial blood oxygenation using pulse oximetry, *J. Biomed. Opt*., 5 (2000) 391-405.

[7] D. A. Boas, D. H. Brooks, E. L. Miller, C. A. DiMarzio, M. Kilmer, R. J. Gaudette, and Q. Zhang, Imaging the body with diffuse optical tomography, *IEEE Sig. Proc. Mag.*, 18 (2001) 57-74.

[8] Prahl, Scott. *Tabulated Molar Extinction Coefficient for Hemoglobin in Water*. 4 March 1998. <http://omlc.ogi.edu/spectra/hemoglobin/summary.html> (24 July 2005).

[9] Wikipedia. *Quantization Noise*. 6 July 2005. <http://en.wikipedia.org/wiki/Quantization_noise> (24 July 2005).

[10] Molitor, Andrew. *Wratten Filters for Infrared- & UV-Photography*. <http://www.a1.nl/phomepag/markerink/irfilter.htm#89B> (24 July 2005).

[11] Grandwell Industries Inc. *LED Sign Viewing Angle and Brightness*. 4 July 2005. <http:// www.grandwell.com/vw_angle.htm> (24 July 2005).

[12] Burr-Brown Products from Texas Instruments. *OPT101: Monolithic Photodiode and Single-Supply Transimpedance Amplifier*. 23 July 2003. <http://www-s.ti.com/sc/ds/opt101.pdf> (24 July 2005).

University of Pennsylvania
Center for Sensor Technologies

SUNFEST

NSF REU Program

Summer 2005

Fabrication of micro-polarizer array with polymer thin film

NSF Summer Undergraduate Fellowship in Sensor Technologies
Kejia Wu (Electrical and System Engineering, BS '06)
University of Pennsylvania
Advisors: Jan Van der Spiegel, Viktor Gruev

ABSTRACT

In this project, significant advancements have been made in creating an array of micro-polarizers using a polymer thin film in order to extract various polarization parameters about the imaged environment. The array of micro-polarizers is to be integrated with custom-made VLSI image sensor in order to create a complete low power real-time bio-inspired polarization sensitive imaging sensor. Two different photoresists, i.e. positive and negative photoresist, are used for patterning the polarization thin-film. The advantages and disadvantages of both procedures are analyzed in details. The isotropic properties of the etching procedure of the thin-film are compared between oxygen plasma and reactive ion etching. The optimum isotropic etching is outlined as a function of temperature, pressure and gases flow. Finally, we have summarized the complete procedure for creating a micro-polarize array with sub-µm precision.

Table of Contents

1. INTRODUCTION

Two characteristics of visible light: wavelength and intensity can be easily detected by unaided human eyes and present as color and brightness. The third important characteristic of light is its polarization, which our human eyes are not able to detect directly. However, the ability to see light polarization has been found in many animals for vital purposes such as navigation and detections of both preys and predators. By definition, visible lights (beams of electromagnetic radiation) have wave properties, in which possesses three-dimensional vectors, thus the presences of these vectors essentially cause the phenomenon of polarization. In most cases, the direction of polarization is the also the direction of electrical field. There are different ways a light source can be polarized, i.e. by double refraction, reflection, scattering, etc, where different vectors are absorbed and preserved.

The electrical field in unpolarized waves can be decomposed into two perpendicular linear components. The purpose for a single polarizer is to block one of these perpendicular components. Consequently, two of such polarizers position perpendicular to each other should block 100% of the light source. Any angle other than the perpendicular (90 degrees) and parallel (0 degree) ones let pass a certain amount of original light, from minimum in perpendicular up to the maximum value in parallel configuration. All four Stokes' parameters ($S_0 \sim S_3$) in Equations (1) can be used to describe full polarization information in an object.

$$\begin{aligned} S_0 &= I_t, \\ S_1 &= 2I(0^\circ, 0) - I_t, \\ S_2 &= 2I(45^\circ, 0) - I_t, \\ S_3 &= I_t - 2I(45^\circ, \pi/2). \end{aligned} \qquad (1)$$

Equations (1) show that S_0 to S_2 can be obtained by measuring intensity of linear polarization of 0 and 45 degrees ($I(0^\circ, 0)$; $I(45^\circ, 0)$) without any phase change along with the total intensity (I_t). [1] Therefore, it's meaningful to fabricate a micro-polarizer array which can collect such information: I_t, $I(0^\circ, 0)$, $I(45^\circ, 0)$. One main future use for the array that this experiment will focus on is to be part of an overall sensor system: it will be placed on top of a custom made imaging sensor. Unlike other polarization detection systems available today, this particular system acts as a single unit, thus be able to display object's polarization, along with the conventional colors and brightness in real time settings. In addition, the small size of our array and chip (3mm by 3 mm in length and width) has low power consumption, which adds another advantage to the system.

Upon completion of this sensor system, it will be utilized in detection in light scattered environments related fields.

2. CHOOSING POLARIZER

There are two ways of obtaining polarizer in this experiment. One is to make a polarizer using procedures that are very similar to LCD monitor production, which is to align liquid molecules on rubbed surfaces. In many cases, a silica glass substrate is repeatedly rubbed toward one direction by a roller wrapped with velvet clothes. After the rubbing process is completed, a thin layer of polarizing material is then spin-coated onto the rubbed surface by an electric spinner. Molecules of the polarizing material would align themselves along the direction rubbing, thus making up the direction of light absorption. The biggest advantage for using such technique is that the thickness of polarizer can be easily controlled by spin speed and concentration of the polarizing material. Less than 1 μm layer of polarizer can be easily attained. Despite the fact that as the thickness varies, there is a trade off between the transmission coefficient and the polarization extinction ratio.

Although the expected data looks very promising, the actual rubbing process is a nontrivial task. Such process has become extremely refined over years of experience from multi-million dollars industries: high standards of rubbing pressure and uniformities are vital to the qualities of polarizer; close monitoring on the level of roller degeneration and dusts created during the process is absolutely required [2]. Therefore, with numerous factors that could affect the performances of polarizer, good polarization data and characteristics can not be guaranteed. Each individual polarizer made from this process will need to go through testing before proceeding to be used. Many sophisticated and expensive machineries have been designed just for the rubbing process, which are not practical to obtain in this experiment. However, simple rubbing process without precision involving dichroic dye solution (polarizing material) called *POLACOAT* is described below:

1) A cleaned glass substrate surface is rubbed repeatedly with a high speed drilling bits wrapped with cotton clothes (with no numerical measurements on pressure and uniformity).
2) The dichroic dye (4%) is then spin-coated on at 2000 rpm for 2 minutes
3) The surface is then heated to 140 Celsius for 14 minutes to evaporate the solvent.

Small amount of polarization effects on a layer of less than one micrometer are observed in the area that has been rubbed. This technique is proven that it can be effective if more sophisticated equipments are to be used in the process.

To get around these shortcomings, some commercially available polarizers are being considered. The advantage here is that, while very cheap and available to obtain, most of them have great proven data for the interests of this experiment. *TECH SPEC™ Linear Polarizing Laminated Film* is chosen for this experiment, it is a brand of polymer film that has shown great extinction ratio in the visible light spectrum.

3. TESTING POLARIZER

The most important characteristic to describe how polarizer performs is its extinction ratio. In ideal cases, one single polarizer should block 50% of the light intensity from an original un-polarized light source, while if two identical polarizers are to be placed at 90 degrees from each other, this resulting position should block 100% of the light. Base on these scenarios, the qualities of the actual polarizer depend on how close it can perform in such way. To measure this, a variable called *extinction ratio*, shown in Equation (2), is introduced. It's the ratio of amount of light that passes through with only one polarizer over when two polarizers that are crossed (90 degrees from each other):

$$\textit{Extinction Ratio} = \frac{\text{light intensity through one polarizer}}{\text{light intensity through crossed polarizers}} \qquad (2)$$

If the ratio unit is converted to units of dB in ideal case, it's obvious to see that the ratio will become infinite. In practical real life case, the result becomes a finite number, which the higher dB it possesses, the better the polarizer performs.

To validate the fact that the purchased polarizer is indeed able to match up data provided by manufacturer (figure 1) [3], a testing process goes as follow: light with different wavelengths are shining through the polarizer to measure the intensity. One narrow-banded green and one red LED are used as light sources. Either a single polarizer or two crossed polarizers are placed between the LED and the optical power meter. From the graph provided by manufacturer, it's clear to observe that the crossed transmission rate is significantly greater for red LED (700 nm) than to green LED (600 nm) due to higher energy emitted from red light. As result, from Equation (2), extinction ratio for the green LED is expected to be higher than red. Measured data from both LEDs are shown in table 1.

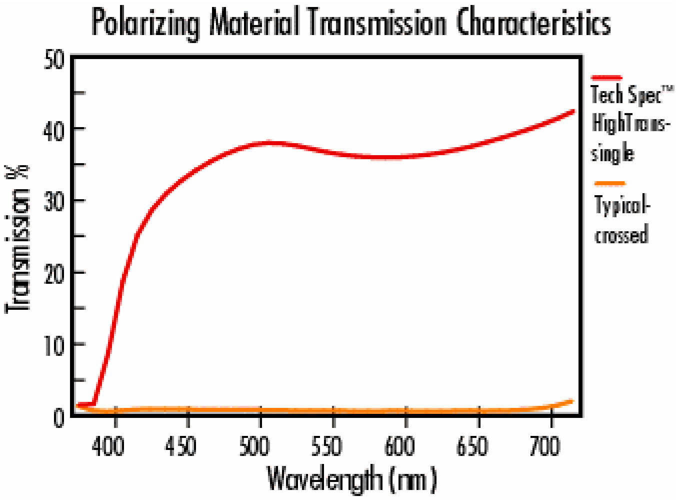


Figure 1: Extinction ratio data of the polarizer provided by TECH SPEC™

Table 1: This table shows the amount of light that has been permitted to pass through either single polarizer or cross polarizer (90 degree from each other). The Green LED shows a better extinction ratio than Red LED due to lower energy, which corresponds to the Figure 1.

	No Polarizer	**Single Polarizer**	**Cross Polarizer**	**Extinction Ratio**
Red LED (700 nm)	**480 mW/cm2**	**190 mW/cm2**	**2 mW/cm2**	**40dB**
Green LED (600 nm)	**480 mW/cm2**	**180 mW/cm2**	**0.2 mW/cm2**	**60dB**

4. POLARIZING THIN FILM STRUCTURE

The basic structure for *TECH SPEC™ Linear Polarizing Laminated Film* [4]has 5 layers: two outermost layers are protective plastic lamination layers, which can be easily peeled off. In the middle consists of a layer of doped Polyvinyl Alcohol (PVA) sandwiched between two Cellulose Acetate Butyrate (CAB) layers. PVA layer is the actual polarizing layer, and thus the main focus of this experiment. The CAB layers are hard and transparent. Their purpose is to give the polarizer a sturdy structure while attenuating a very small amount of light rays going through them. To expose the actual PVA layer, either one or both layers of CAB need to be removed.

100% pure acetone solution is used for the removal of CAB. The solution has no effect on PVA, and thus makes it an ideal chemical for CAB wet removal. When a sample film submerged into acetone solution, within 20 minutes, CAB starts to soften and peeling up from edges. After almost an hour, CAB is in gel-like state, and it turns into white solid when expose to air and water. Putting it back to acetone solution, the white residue will disappear and become gel-state again. If nothing is done to physically remove the CAB, after a period of 2-3 hours, the PVA is completely detached away from CAB acetone. However, due to small thickness of the PVA layer, it's extremely difficult to keep it undamaged.

One approach to solve film's fragility is to remove only one layer of CAB, while CAB from the other side is still thick enough for support. To prevent the CAB layer from being attacked by acetone, a brand of thermally and chemically resistant tape is used to tape one side of polarizer onto a glass substrate with the other side ready to be removed by acetone. To speed up the CAB removal process, as soon as CAB layer becomes gel-like state within an hour, lightly and gently rub the surface along with DI water rinsing. To avoid scratches on the film, most top portion of CAB is to be removed with physical rubbing, the rest is washed away with repeated DI water rinsing.

With a sample that contains one PVA layer and one CAB layer sample is obtained, a simple thickness measurement by a microscope is performed. It has been found that most PVA appear dark (Shown in figure 2). One problem arises that originally PVA is expected to be rather uniform in thickness. In reality, the layer appears to have thickness ranging from 10 to 25 μm. Variation in thickness can be explained by the way the polarizer is made and manufactured: special machines are used to stretch the polymer film to very thin. As the result, due to different level of stress applying across the surface, the thickness varies. Due to this result, the polarizing capability changes from point to point. This variable presents a major obstacle in later part of this paper on how long the sample should be put in the dry etching process.

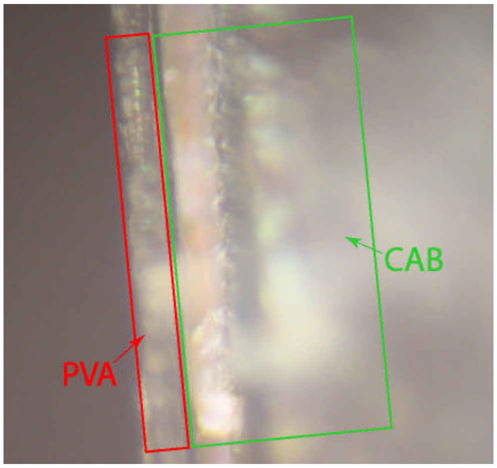


Figure 2: TECH SPEC™ Linear Polarizing Laminated Film when one layer of CAB has been stripped, it shows the clear interface between PVA and CAB layers.

5. POSITIVE PHOTORESIST MASKING

Patterns on the microscopic level need to be formed on actual PVA layer as the final product. The idea is to initially spin coat a layer of photoresist on top of PVA. The desired pattern is then put on photoresist through a conventional process called *photolithography*. This particular pattern is then transformed onto the PVA layer through designated etching procedures.

Positive photoresist is first being put to test first. It's more commonly used in semiconductor fabrication industry. *Shipley S1813* positive photoresist appears to possess a red-orange color, and it's slightly more viscous than regular water. To guarantee the uniformity of the photoresist layer, it will be spin-coated by an electric spinner. The desired thickness can be controlled by the spin speed.

An UV mask aligner is used to pattern the photoresist. The basic principle is that, when exposed under UV light for a certain period of time, energy exerted by UV would be enough to change properties of positive photoresist so to make it soluble to the photoresist developer [5]. The amount of time to be exposed is calculated by Equation (3):

$$\text{Power (wavelength)} = \text{Required photoresist energy} \times \text{Time} \qquad (3)$$

The entire photolithography process is described as follows:

1. Depending on the desired thickness, photoresist will be spin-coated with a different speed. (for 1.5 µm thickness of photoresist, spin 4000 rpm for 30 seconds, faster spin speed means thinner photoresist)
2. Undergo soft baking process to evaporate the solvent. (125 Celsius for 2 minutes)
3. UV Photolithography (with patterned masks), expose under the time that is necessary to change the exposed photoresist's structure completely with a *Kasper System™ 2001 Mask Aligner*.
4. Develop in the photoresist developer for certain period of time, depending on the expected thickness. (1 minute)
5. Hard bake for a period of time to harden the photoresist.

Figure 3 shows the positive photoresist UV photolithography process:

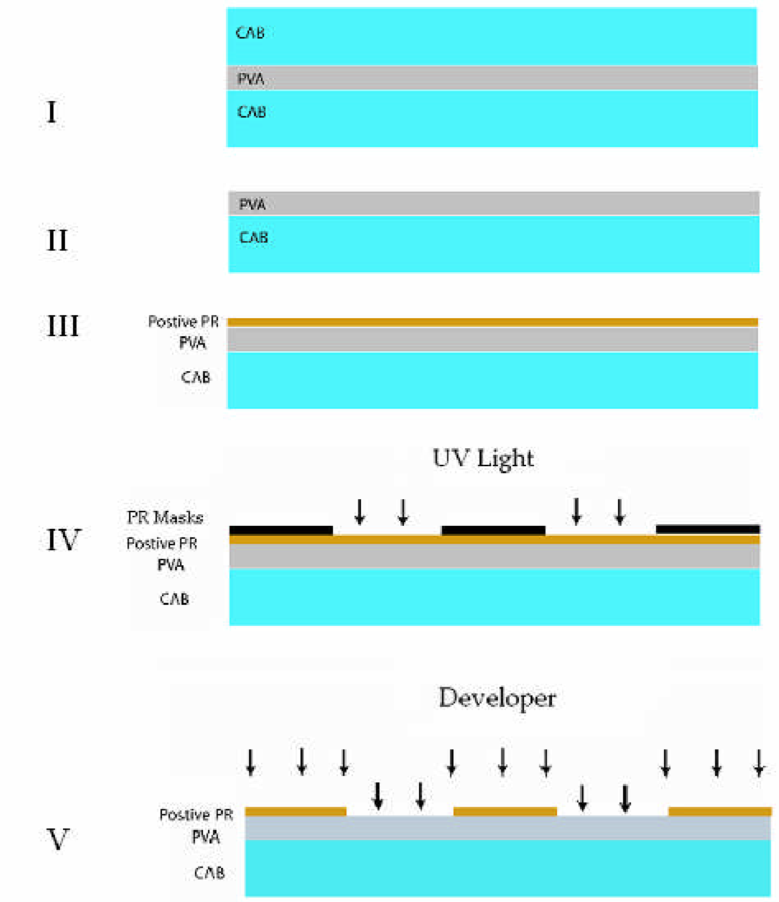


Figure 3: I: the original polymer film. II: one side of CAB is stripped by acetone. III: a layer of positive photoresist is spin-coated on top of PVA. IV: UV light shines on the masked photoresist. V: After development, the part of positive photoresist that has been exposed to UV has been dissolved.

The amount of time spent on hard and soft bake depends on the photoresist material and its thickness. The way to test relation between the spin speed and the thickness of the photoresist is to go through the UV Photolithography process described above, then use *Tencor Instruments* ™ *Profimeter* to scan through the pattern to measure thickness.

5.1 Oxygen plasma etching and positive photoresist

Once a photo resist pattern is put on top of the PVA, to transfer the pattern, an etching technique is needed to etch both photo resist and exposed PVA simultaneously in order to achieve the purpose of pattern transferring. In best possible scenario, this technique is able to etch vertically in an anisotropic fashion. With the thicknesses of both photoresist and PVA are known, by the end of process, it's expected that both exposed photoresist and PVA should disappear at the same time, while the PVA underneath the photoresist remains to retain the original pattern from photoresist. The process is shown in figure 4:

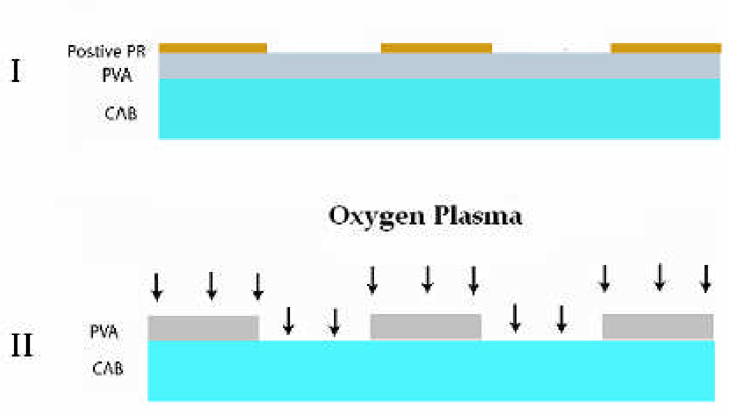


Figure 4: I: The sample after UV Photolithography. II: The oxygen plasma etching process is suppose to etch away the thickness of photoresist mask and the thickness of exposed

One commonly used dry etching method is the oxygen plasma etching (performed by *Technics PlasmaEtch™ II Oxygen Plasma instrument*). Gases such as oxygen are heated to extremely high temperature to their plasma forms. The electrons released in the process will give target surface the bombardment it needs to be etched away. In reality, it's more relevant to look at the etching rate ratio for both photoresist and PVA layers rather than the actual etching rates.

Among all the important factors that can affect the etching rate ratio, the most important one is the O_2 and CF_4 gas flow ratio [6]. Other factors such as amount of etching time and amount of power all have a linear relation with the etching rate, and therefore easy to predict. Figure 4 shows the etching rate for both PVA and photoresist layers, with power and time being fixed. One horizontal axis, O_2, CF_4 gas flows are in increments of 0.5 sccm.

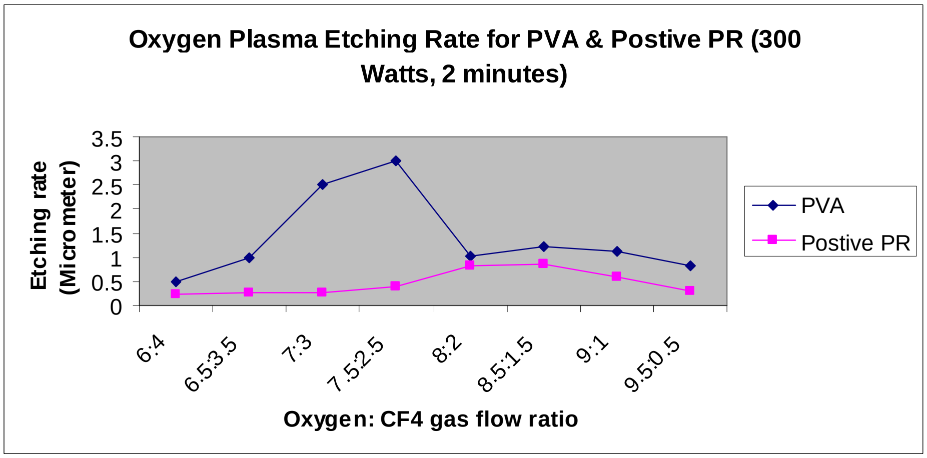


Figure 5: It's obvious to notice that the etching ratio (PVA over Positive PR) is at greatest when O_2 : CF_4 gas flow ratio is at 7.5:2.5.

As shown in figure 5, even if the right gas flow ratio to achieve the optimum etching rate ratio is attained, the maximum etching ratio (PVA over Positive photoresist) is approximately 7:1, much lower than the average thickness ratio(PVA over Photoresist) of 10:1. This indicates that at some point during the etching process, the protective photoresist layer will be long gone and PVA underneath will start to be etched away before the exposed PVA is etched away completely. As this paper emphasized earlier, changes in the thickness of the actual PVA can cause its polarizing performance to be altered, and thus undesirable in this experiment.

Besides the etching ratio, there is another important etching aspect that will directly affect how small of pattern's dimensions that can created, and it's called the isotropic effects of etching. Ideally, when both exposed PVA and photoresist pattern are etched away, an anisotropic etching fashion is being assumed. However, in real life, absolute anisotropic etching can hardly be achieved. It's only possible to find a technique with relatively less isotropic effects. In this case, with oxygen plasma etching process, it has been measured that for each PVA unit that has been etched vertically, the gas plasma also etches away half unit horizontally (see Figure 6), thus we conclude the vertical to horizontal etching ratio to be 2:1. With a simple calculation, if the PVA is at its thickest: 30 µm, by the time oxygen plasma completely etches away the exposed PVA, it will also etch approximately 7.5 µm from each side of the PVA that's under the photoresist patterns, with the desired PVA pattern to be in units of 10 µm, it will be impossible to create such patterns. A better technique is needed.

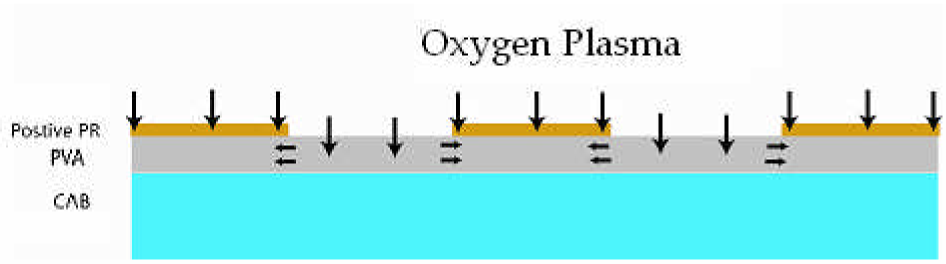


Figure 6: This is the actual Oxygen Plasma etching process. Ideally, it should only be etching vertically, but it does have the isotropic effects of etching all directions.

5.2 Wet etching

During the photoresist development process, once the developer completely dissolves the UV exposed positive photoresist, it gradually attacks the PVA layer underneath due to the similar structure shared by both positive photoresist and PVA. With this property, in theory, the developer can be used for wet etching. However, positive photoresist is etched away in a much more rapid rate than to PVA. The vertical to horizontal etching ratio (isotropic effects) for PVA could reach as low as 1:1, thus making it an undesirable etching technique also.

5.3 Problems with Positive photoresist and oxygen plasma etching

As Several shortcomings have been observed from using positive photoresist masking:

1. Because of the similar structure from both *Shipley S1813* and PVA layer in the polymer film, the developer will also etches away PVA underneath once the top photoresist layer is depleted. Due to the severe isotropic etching effect (as much as 1:1) from the positive photoresist developer, it is not desirable to use positive photoresist as the protective layer on top of PVA. Under-exposing the photoresist is tested to retain a layer of protection photoresist layer. However, the under exposed surface has high non-uniformity and roughness, and thus undesirable. See figure 7.

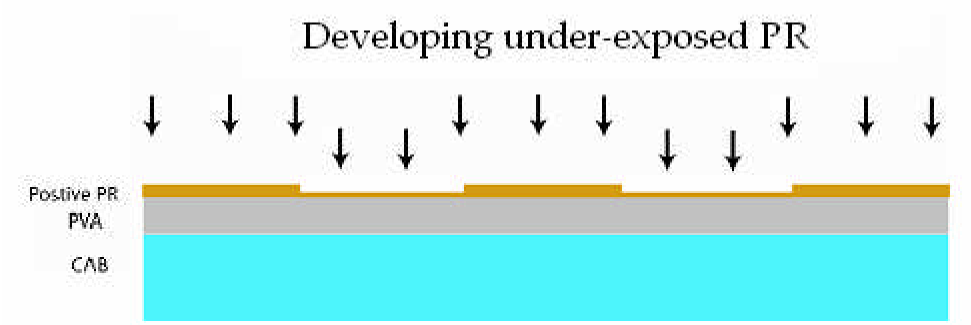


Figure7: For under-exposed PR, it would not be completely dissolved by the solution, and therefore leaves a thin protection layer, however, since this is the wet etching process, that layer is non-uniform.

2. The thickness ratio of PVA to positive photoresist could reach as high as 10:1, however, with the optimal O_2 and CF_4 gas flow ratio at 7.5 to 2.5, the oxygen plasma etching could only attain as high as 7:1 etching rate, which is much lower than the thickness ratio. In order to decrease PVA to positive photoresist thickness ratio, photoresist thickness has to increase. However, spin multiple layers of positive photoresist to increase the thickness is not possible because *Shipley S1813* is not designed for multiple spins.
3. With oxygen plasma dry etching, it shows a considerate amount of isotropic effects: as high as 2:1, which presents a problem in the compensation process of aligner mask making.

With these ideas in mind, a new type of photoresist that can be uniformly spun onto a surface with a larger thickness is needed since the thickness of PVA is fixed can't be uniformly decreases. It should also be made by much different (if not completely different) material from PVA, so that when it's undergoing development, PVA isn't to be affected at the same time. One final important improvement is that a better anisotropic etching technique is needed.

The following sections will describe how different approaches take place to solve these problems one by one.

6. NEGATIVE PHOTORESIST MASKING

Negative photoresist has opposite properties as the positive photoresist when exposed under UV light. Instead of exposed region being washed away by the developer like positive photoresist, it stays intact and the unexposed negative photoresist region is being washed away.

SU-8 is a series of negative photoresist that can be spin-coated on with a large range of thickness: from as thin as 1 µm to as thick as 1 mm.(depending on serial code of the photoresist) All products from this series have very transparent appearances and the viscosity increases with its serial number (the more viscous product is used to obtain a thicker photoresist layer). Because *SU-8* is made up by completely different material as PVA, its developer has little or no effect on PVA.

One important property to notice when applying *SU-8* onto a surface is that the photoresist is highly hydrophobic. Even the humidity in the air condensed onto the PVA surface will decrease *SU-8*'s adhesion onto PVA dramatically. Therefore, in addition to heat the PVA before applying the photoresist, an extra thin layer of adhesion promoter will be put on top of PVA. It has been found that *AP300* series chemical solutions can be used. When spun on top of PVA, it reacted with moist to form a protective layer: TiO_2. The adhesion between TiO_2 and *SU-8* photoresist becomes much stronger.

The rest of negative photoresist masking process is similar to positive except for a couple of extra steps:

1. Spin coat the adhesion promoter
2. Immediately apply and spin coat desirable thickness of *SU-8* within half of a minute after adhesion promoter *AP300*.
3. Soft bake the photoresist to evaporate the solvent so to achieve higher density for the layer. It has two levels of temperature of 75 Celsius and 115 Celsius.
4. Expose through UV-photolithography with calculated time, which is based on the amount of energy needed to expose the photoresist completely
5. Post-expose bake is needed to cross-link the exposed portions selectively, it also has both 75 and 115 Celsius temperature levels.
6. Develop with MicroChem *SU-8* Developer

6.1 Negative photoresist gradients

For one particular trial in the experiment, after AP300 adhesion promoter is spun onto PVA layer (2000 rpm for 30 seconds), immediately afterward, a layer of 40 µm *SU-8* 2015 is applied on top of promoter (500 rpm initial spin speed for 10 seconds, then 10 seconds of constant acceleration up to 2000 rpm followed by 40 seconds of 2000 rpm). Soft bake process is then followed (1 minute at 75 Celsius, and 5 minutes at 115 Celsius), photoresist needs to be checked whether it's completely solidified. The sample is then going through UV photolithography (22 seconds at 11 ms per cm^2) with a mask pattern of 96 µms wide. A hard bake process is also added after the photolithography to cross-link the photoresist (1 minute at 75 Celsius, and 5 minutes at 115 Celsius). Finally, the pattern is formed when the sample is developed in the solution for 3 minutes. The vertical cross section of the pattern appeared to be a trapezoid shape instead of rectangle, shown in figure 8.

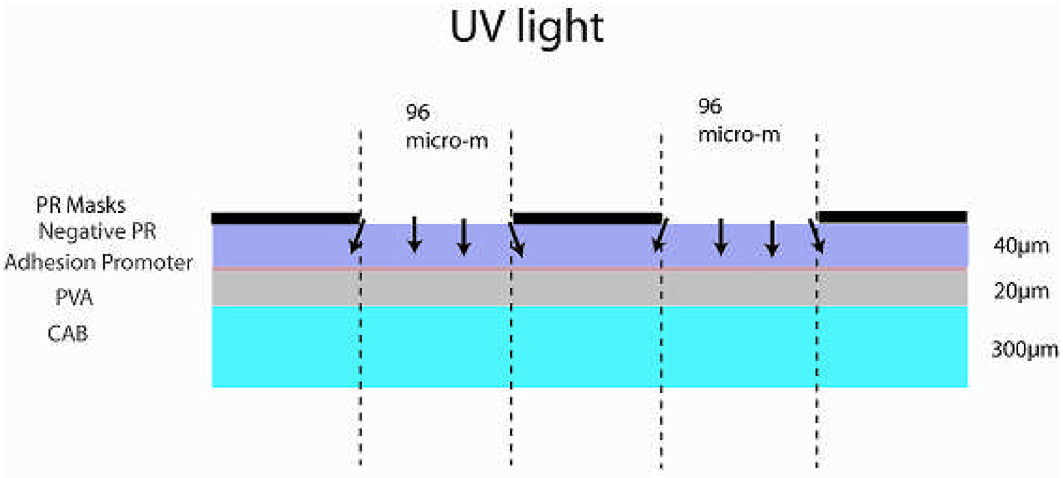


Figure 8: at the edges of the PR masks, the UV light rate does not travel straight as expected, instead, it gets refracted and penetrating PR through an angle.

Figure 9 shows the photoresist pattern formed after the UV-photolithography. The extra area UV light has covered is called the *gradient of photoresist.* In this case, due the 40 µms of thickness, *SU-8* layer has a gradient of 2 µm on each side. This is a significant length to overcome, considering the desirable size for each unit is only 10 µm. However, with a simple geometric calculation, if *SU-8* thickness can be controlled, so are gradients. Table 2 shows the size of gradients comparing to *SU-8* thickness.

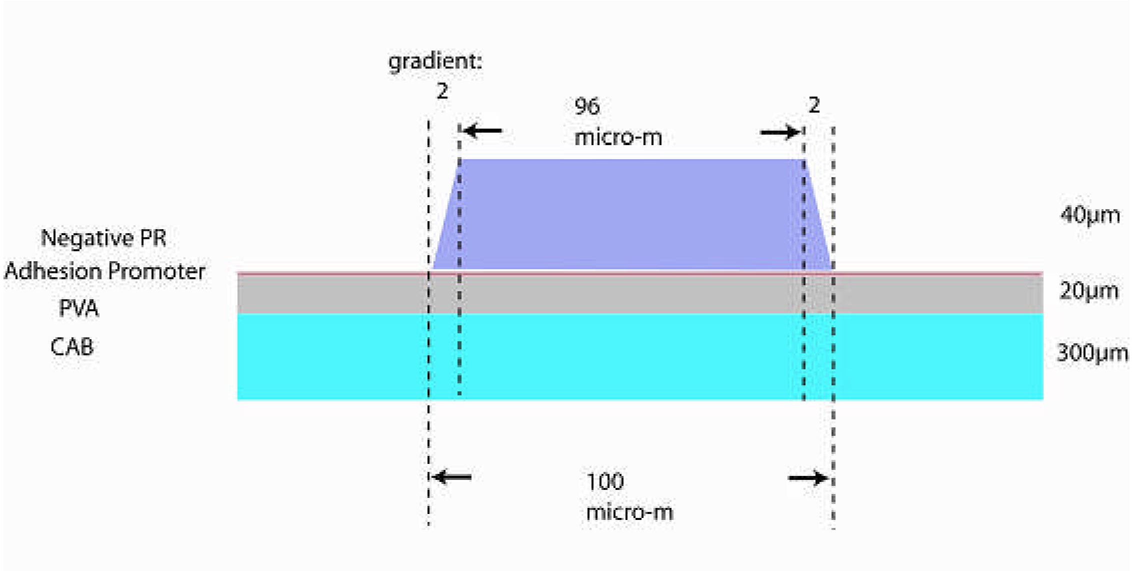


Figure 9: The gradients are shown between the dotted lines.

Table 2: it shows that when *SU-8* is around 1 μm, the gradients are negligible, which is the case when we used the positive photoresist.

SU-8 (μm)	Gradients (μm)
1	0.05
15	0.75
40	2

6.2 RIE (Reactive Ion Etching) process

It is mentioned earlier that an improved etching process with less isotropic effects is needed. One such effective etching process is called Reactive Ion Etching (Performed by *Plasma Tech. Plasmalab μp*). In addition to the mixture of O_2 and CF_4, Argon gas is added, which will provide more mechanical etching power vertically. With the optimal gas flow ratio of 7.5:2.5:2.88 (O_2 : CF_4 : Ar), the PVA over photoresist etching rate can be achieved as high as 3:1. Since *SU-8* photoresist can range from less than 5 μm up to 1 mm (in case of *SU-8 2015*, it can attain from 15 to 38 μm with 3000 rpm to 1000 rpm), *SU-8* can well accommodate the 3:1 etching ratio. The targeted photoresist will be around 15 μm ~ 20 μm to go with maximum possible PVA thickness of 40 ~ 60 μm, matching closely with etching rate of 3:1. In addition, for RIE process, the vertical VS horizontal etching ratio (isotropic effects) for PVA can reach as high as 10:1, which is a significant improvement over oxygen plasma etching. One undesirable properties of RIE is that the procedures blur both PVA and CAB surfaces, which will present problems in optical domain of the system operation.

7. CONCLUSION AND RECOMMENDATION

It has been found that *Shipley S1813* is not able to achieve the desired specifications as a mask layer for this experiment. With the thickness flexibility of *SU-8* combined with relatively good anisotropic etching from RIE process, all the comparisons and calculations are drawing the conclusion that it's feasible to fabricate polarizing array with each individual unit size at 10 μm with these procedures. The future plan for the continuation of this experiment will include:

1) Explore different ways to treat the blurry surface left by RIE etching process.
2) Use a test mask that has features with desired dimensions (close to 10 μm) and go through the fabrication procedures described in this report. Scan the entire structure profile of the test product with Scanning Electronic Microscope.
3) Two-layers configuration so to collect both I(0°, 0) ; I(45°, 0).

8. ACKNOWLEDGEMENT

I would like to thank my advisor Professor Jan Van der Spiegel and Dr. Viktor Gruev for their guidance. I would also like to thank National Science Foundation and the SUNFEST staff for giving me this great research opportunity.

9. REFERENCES

[1] J. Guo and D Brady, "Fabrication of thin-film micropolarizer arrays for visible imaging polarimetry", Applied Optics, Vol. 39, No.10, April 2000

[2] J. V. Haaren, "Wiping out dirty displays", Nature, Vol. 411, May 2001

[3] http://www.edmundoptics.com

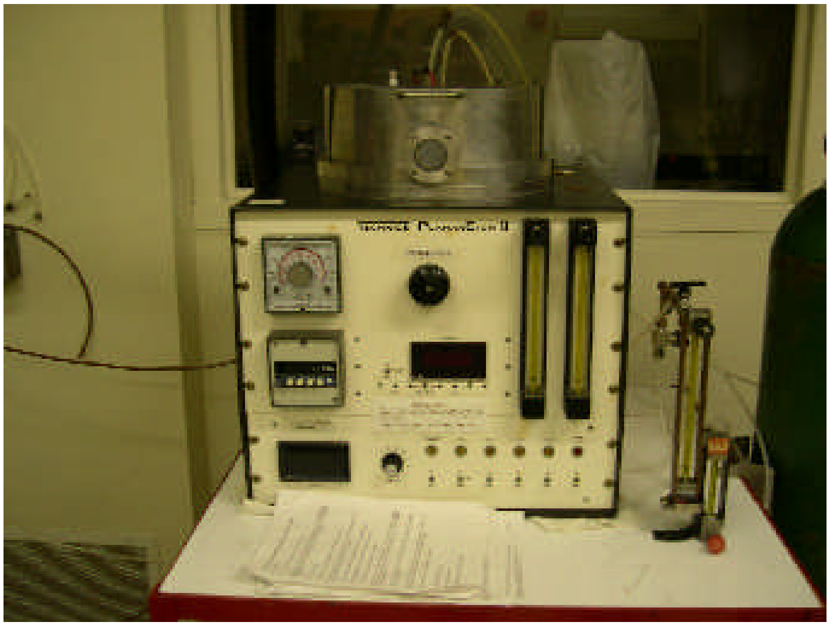
[4] http://www.microchem.com

[5] R. C. Jaeger, Introduction to microelectronic fabrication Volume V, 2nd edition. New Jersey: Prentice Hall, 2002, 1998.

[6] F. D. Egitto, "Plasma etching and modification of organic polymers", Pure & Appl. Chem, Vol. 62, No. 9, pp. 1699-1708, 1990.

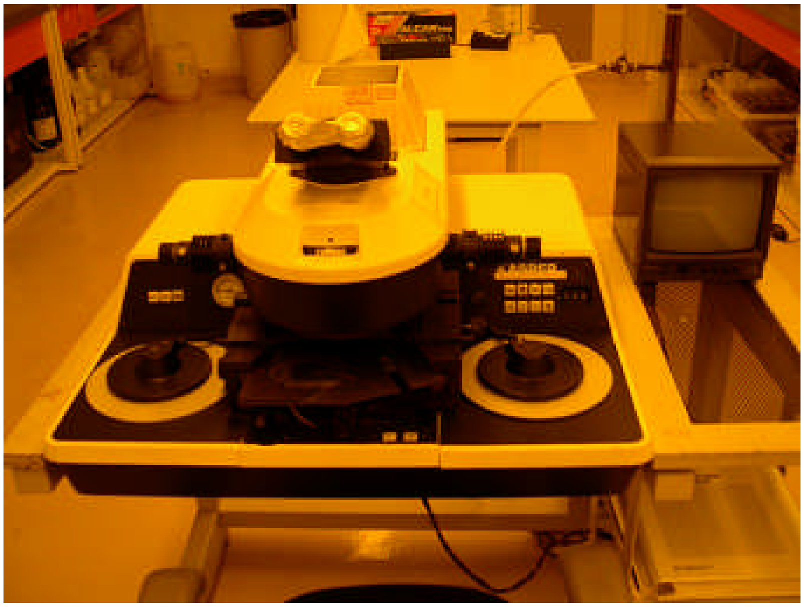
APPENDIX A

INSTRUMENTS

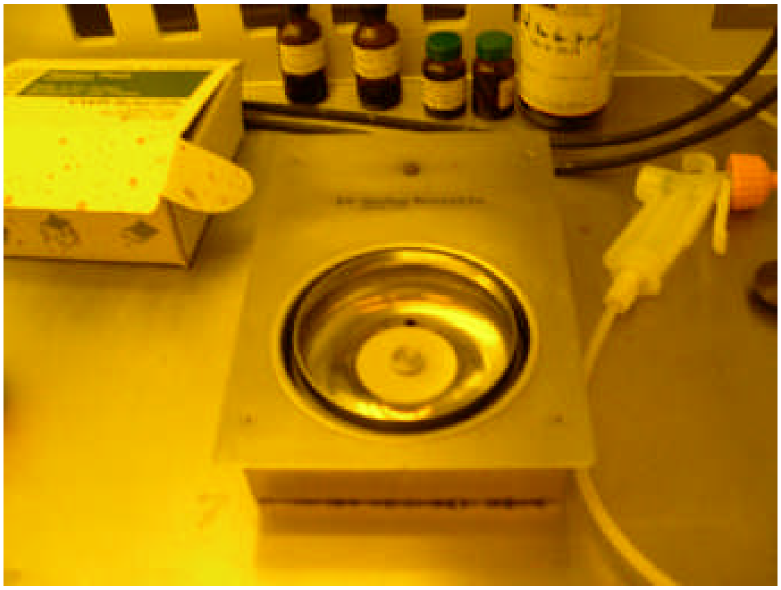


Technics PlasmaEtch™ II Oxygen Plasma instrument

Plasma Tech. Plasmalab μp



Kasper System™ 2001 Mask Aligner



Spinner

Tencor Instruments ™ Profimeter