HUMAN AND MACHINE RECOGNITION OF THE VOCAL CHARACTERISTICS OF SUICIDE

BY

Abhraneel Sinha

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

In the partial fulfillment of the requirements

For the degree of

MASTER OF SCIENCE

In

Electrical Engineering

December 2013

Nashville, Tennessee

Approved:

Associate Professor D. Mitchell Wilkes, Ph.D.

Associate Professor Ronald M. Salomon, M.D.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my adviser, Dr. Mitch Wilkes for his continuous encouragement, motivation, ideas, advices, and patience to help me understand the material and helping me with finishing the thesis. The enthusiasm he has for his research and teaching was contagious and motivational for me. I could have never imagined having a better adviser than him.

Besides my adviser, my sincere thanks also go to Dr. Salomon and The Vanderbilt University Department of Psychiatry for providing valuable database for us to use in this study. I am also very grateful that Dr. Salomon was willing to spend his time to read this thesis even on such a short notice.

I thank my fellow lab mate, Nik Nur Wahidah Nik Hashim for the wonderful support and stimulating discussions we had all throughout these two years.

My greatest appreciation goes to my parents, who have raised me with love and care until I became what I am today. To other members of my family especially my sibling, thank you very much for being my inspiration.

Lastly, to all my friends at Vanderbilt and elsewhere that I could not mention their names one by one here, your support and encouragement were a driving force for me to finish my work. Your presence makes my life meaningful and enjoyable. Every bit of your help in any aspect of my life is deeply appreciated.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Even with advances in treatment over the past decades, suicide remains a major public health problem worldwide .While other causes of death and disability are slowly declining in prevalence ,overall suicide rates have not improved, so that it has now become one of the country's worst public health problems. Referring to suicide statistics for the year 2007, there were 34,598 reported suicide deaths in the United States. Suicide is the fourth leading cause of death for adults between the age of 18 and 65 years with a total of 28,628. Between the years 2000 to 2007, a significant increase of 8.7% percent occurred [1]. It has been reported that there are an estimated 8-25 attempted suicides for every suicide death. Among the suicide deaths, 60% are caused by major depression and this number rises to over 75% if depression caused by alcoholism is included. These percentages showed that untreated depression plays a major role in the cause of suicide [2]. The statistics show the need for an analysis to identify and detect patients with near term suicidal risk.

Recognizing that suicide has a significant effect on public health, a scientific strategy for evaluating suicide is an important preventive measure that needs to be developed. Current estimate of risk assessment is through a clinical judgment process that includes some objectifiable risk factors, and often also includes a subjective "gut feeling" of the clinicians regarding the patient's potential state of suicidal thoughts. Clinical judgment requires an evaluation of comprehensive information regarding a patient's demographic profile, health record, history, family evaluation, prior mental health. treatments, as well as suicidal ideation, behavior, planning, desire, and intent [3]. Personality psychometric instruments or suicide-specific scales also have a wide range of use for estimating suicide potential and treatment decisions. But these measures are generally not suitable for making judgments regarding a patient's level of imminent risk for suicide [4].

A lot of information regarding the psychological state is accessible through the human voice. Several studies have been conducted since 1984 on the effects of emotional arousal on speech production based on indication of speech rate, voice articulation and respiration. Some

methods based on acoustic features that were investigated are estimation of fundamental phonation or pitch, features based on a nonlinear model of speech production, relation of vocal tract features to emotional speech and speech energy estimation [10]. Research findings show evidence of specific vocal characteristics among patients at the level of near term suicide [6], [7], [8], [9]. The investigation of correlation between vocal characteristics and psychological suicidal state was proposed by Drs. Stephen and Marilyn Silverman. In particular, the Silvermans reported hearing a particular sound or tonality in the voice of high risk patients. This sound reportedly would come and go in the patient's voice lasting for perhaps one or two minutes at a time. The relationships between voice parameter and psychological state are extremely complex and require thorough research in obtaining acoustic measurements that are robust in distinguishing near term suicide with ideation or major depression [11].

CHAPTER II

BACKGROUND AND SIGNIFICANCE

Voice is sound produce by humans using their lungs and vocal folds located in larynx while speech is thoughts, ideas, and feelings that are being transferred orally, triggered by the muscle articulation that alters the tone produced by voice into a known decodable sound. Speech plays a significant role in everyday life .Humans use speech to express complex or abstract information such as emotions, thoughts and ideas. Communication between human beings involves not just spoken words but also other features corresponding to intonation, accent loudness and speed.

Consider the system to be comprised of the three parts shown in the figure below. The system is driven by the energy supplied by the lungs during expiration. When inspiration takes place, the lungs, a pair of elastic, balloon like structures is filled with air. Since the walls of the lungs are elastic they will tend to deflate to a neutral condition, and air can, in addition be forced from them by the action of those muscles that control expiration. The larynx acts as a valve that can be moved into place over the expiratory air stream to convert the relatively steady expiratory air flow into puffs of air. The sequence of air, the volume velocity waveform, excites the upper vocal tract, the oral and nasal air spaces, and the single or branched tubes lying above the larynx. When the velum is lowered, air flows out of the mouth and nose ; when it is elevated ,the nasal air space is closed off so that the air flows out of the mouth only. The upper vocal tract acts as a variable acoustic filter on the output of the larynx; its properties of a filter depend upon the shape. The shape is controlled by the movement of the articulators. The velum is a muscle mass that acts to connect or shut off the nasal branches of the upper vocal tract from the air space below. The nasal air space has an almost constant shape. On the other hand the oral air space has a shape that varies considerably depending on the position of the various articulators, such as the tongue, jaw, lips. The position of the articulators in turn, is controlled chiefly by the complex interactions of a large number of muscles.

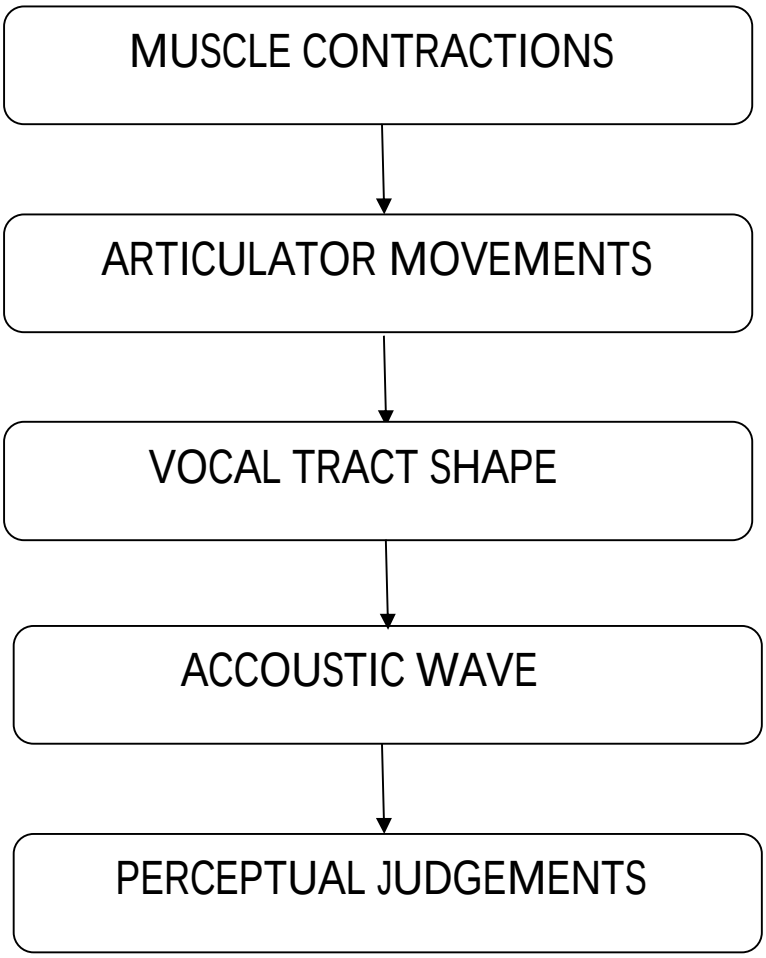

Fig 2.1: Schematic Diagram of the Articulatory System

2.1 Mechanism and Physiological Aspects of Speech Production

Figure 2.2 shows the cross sectional view of the upper portion of a human anatomical structure involved in the production of speech. Speech production can be viewed as four separate components, the respiratory system, laryngeal system, resonance and articulatory system.

The Respiratory System consists of lungs, bronchi, trachea and other associated muscles. It acts as the main energy source by supplying air, and is also responsible for the amplitude of the sound as displacement of vocal chords changes with respect to air flow energy. The diaphragm is the most powerful muscle use in producing the voice.

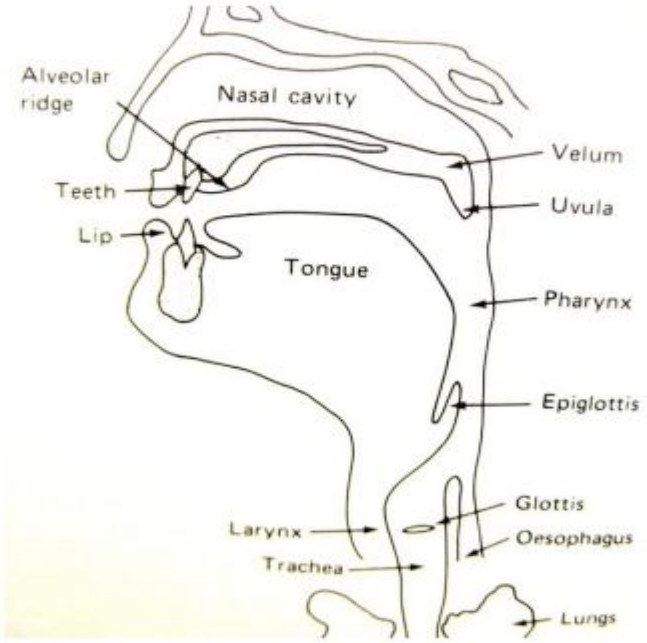

Figure 2.2: Cross-Sectional view of an anatomy structure of human vocal production [5]

Figure 2.3 shows the simplified block diagram of the speech production process. The lungs provide airflow and muscle force pushes air through trachea, bronchi and through the glottis located between the vocal chords and larynx into three main cavities consisting of the vocal tract ,the pharynx and the oral and nasal cavities. The laryngeal system consists of the larynx and the vocal chords. The larynx is a tube consisting of cartilages and muscles , commonly called the voice box involved in opening and closing of the glottis. The resonance system is the pharyngeal cavity where it is made of pharynx, oral cavity and nasal cavity. Pulses of sound are manipulated here into a recognizable voice.

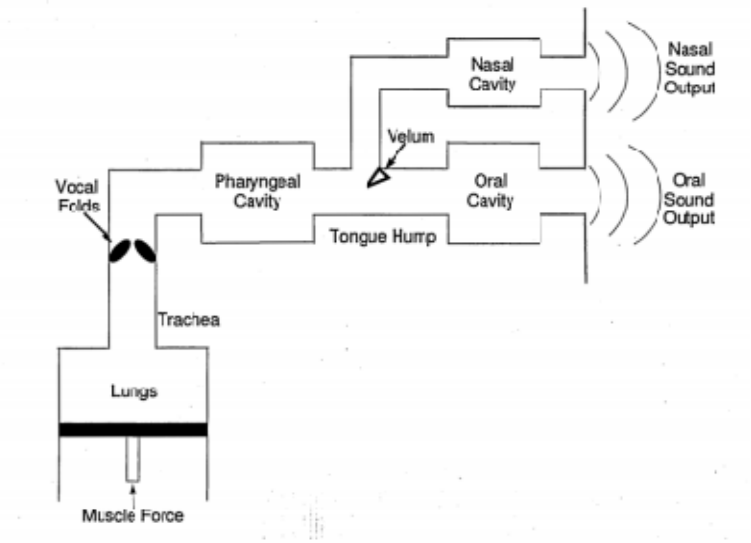

Figure 2.3: Simplified Speech Production Model [12]

The velum ,jaw, lip ,tongue ,teeth and other structures that are mostly visible outside the human form make up the articulatory system. They alter the speech into comprehensible utterance called speech. Air flow exits through the mouth and nose to become voiced and unvoiced sounds.

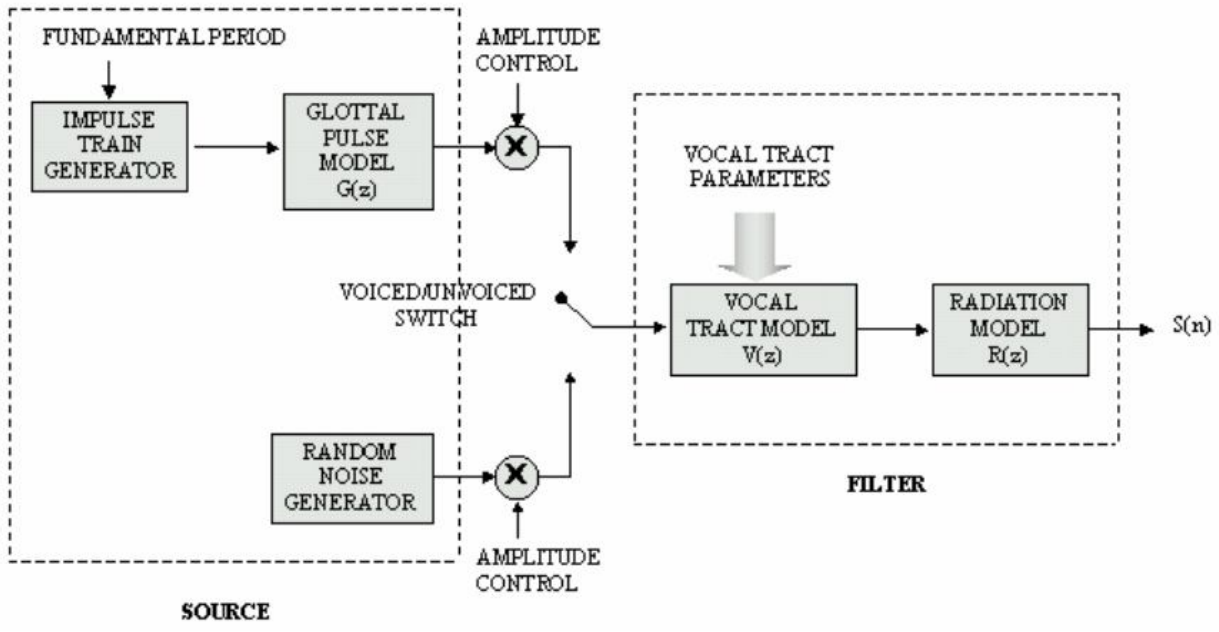

Figure 2.4: General Discrete Time Model of Speech Production [13]

The complete z –transform of the process is represented as:

$$S(z) = G(z)\, V(z)\, R(z) \qquad (2.1)$$

where S(z)= Speech waveform , G(z)= Glottal Pulse Train ,V(z)=Upper Vocal Tract, and R(z)=Lip Radiation

This model assumes that the excitation components can be separated from the vocal tract and radiation components, and the entire system is linear. It is obvious from this model that the speech signal is non stationary. Since the vocal tract articulators move slowly as relative to speech ,the system is assumed to be short-time stationary which means the general properties of the vocal tract and excitation remain fixed for a short period of time (10ms-30ms).

2.1.2.1 The Source Excitation Model

When modeling the source of voice production, there is a difference between the acoustic model and the source-filter model. In the acoustic model, the glottal flow is dependent on the vocal tract shape due to the acoustic load above the glottis that is defined by the output of the vocal tract. Meanwhile, assuming that the source-filter model is independent of the vocal tract shaping variations, the glottal source is defined as a non-interactive signal description of the voice source [24]. Implementation of this model inputs a random white noise for unvoiced sound and a discrete-time periodic impulse train with a certain fundamental period between each pulse that acts as the source excitation signal for voiced sound. Voiced speech is considered to be non-stationary over a large interval of time but the characteristics and information in the voiced speech can be measured to be relatively constant over a short period of time. Similarly, the glottal pulse can be represented by Equation 2.2 [16] where g[n] represents the discrete-time impulse train pulses and T_o is the fundamental period

$$g[n] = \sum_k \delta[n - kT_o] \qquad (2.2)$$

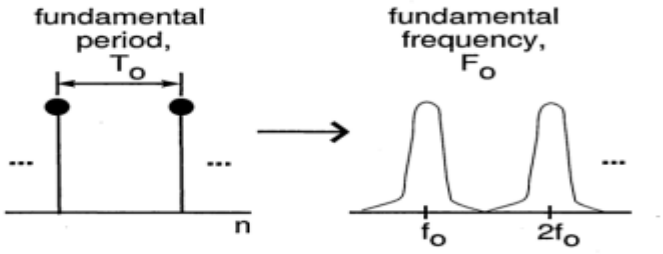

Figure 2.5: Time and Frequency Domain Representation of glottal pulses [14]

2.1.2.2 The Vocal Tract Filter

The actual model of the vocal tract consists of varying the cross-sectional area based on the position across the tract as the wave propagates over time. These variations are caused by the alteration of the frequency content of the excitation signal. The continuous-time model of a vocal tract can be conveniently represented as a discrete-time model by transforming it into a concatenation of uniform lossless tubes of varying diameters. These tubes are considered "lossless" due to the assumption that no sound energy is absorbed by the walls. For an arbitrary shape of vocal tract, the area would vary with respect to time, A(x, t). Referring to Figure 2.6, assuming that the vocal tract exhibits a uniform tube-like shape, the constant cross sectional area $\{A_k\}$ and length $\{l_k\}$ of N-sections are chosen to approximate the total area of the vocal tract, A(x).

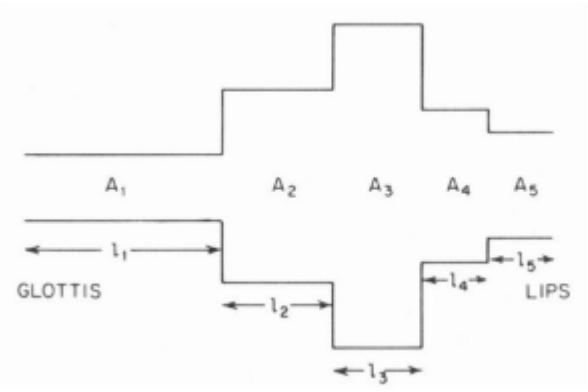

Figure 2.6: Concatenation of lossless tubes for N=5 [15]

The output speech is related to the relationship between pressure and volume velocity which are determined by the cross-sectional area of the tube and the speed of air. At the joint of two tubes, continuity must be obtained in order to keep constant pressure on both sides as the waves traveling from one tube to the other. The excitation propagates through the series of tubes with some partially reflected and some waves partially propagated across the two joint tubes. Besides the joint of two tubes, boundary conditions at the lips and glottis must also be taken into account [15]. A linear prediction (LP) analysis involves the prediction of signal parameters based on the previous values and is a technique that is used to model the vocal tract as an all-pole filter called an inverse filter as shown in Equation 2.2 where $\{ a_k, 1 \}$ and the predictor coefficient and the N order of the filter (number of poles).

$$V(z) = 1/A(z) \quad \textit{and} \quad A(z) = 1 - \sum a_k z^{-k} \text{ where } k=1 \text{ to } N. \qquad (2.3)$$

When modeling the vocal tract by an all-pole filter, the nasal and unvoiced sounds are not taken into account. According to [25], inclusion of nasal and unvoiced sound into the current all-pole model can be achieved by including more poles rather than including zeros. All poles will remain inside the unit circle considering the areas of the concatenated tubes to be positive.

2.1.2.3 Lips Radiation

The opening of the lips marks the end of vocal tract tubes. The lip opening is modeled as an orifice in a sphere where the lips are represented as radiating sound waves and the head is represented by a spherical baffle that refracts the sound waves. If the opening of the lips is small enough compared to the size of the sphere, the radiating surface can be thought of as a radiation from an infinite plane baffle. Pressure is measured from a given distance, l from the mouth and is proportional to the time-derivative of the lips flow.

2.2. The Source Filter Model of Speech Production

Speech production can be explained with a simple source-filter model. At the most primitive level, the source-filter model can take the form shown in Figure 2.7.

Figure 2.7: A Basic Speech Model

An example to illustrate this model is when a person blows a saxophone. The air pressure from the mouth is the source and the saxophone itself is a filter. The sounds made are equivalent to speech in the speech model. Figure 2.8 shows a more refined model of speech production. The voiced speech and unvoiced speech are produced by the impulse train generator and the random noise generator respectively. The switch position points to the normalized excitation source depending on the characteristics of the voice (voiced/unvoiced).

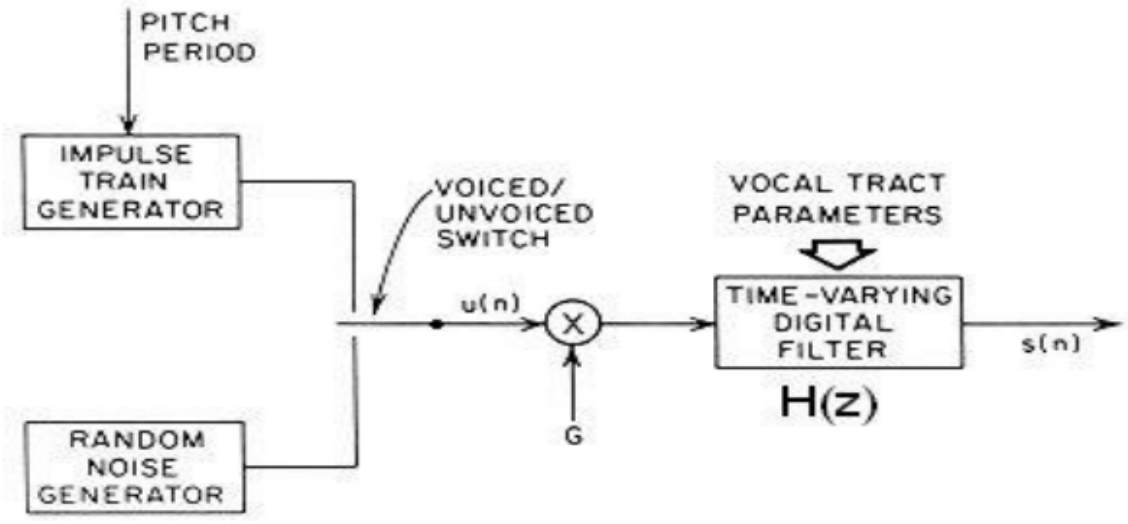

Figure 2.8: A more refined source-filter model tied to Linear Predictive Coding [16]

A gain factor (G) is measured from the speech signal and is used to scale the normalized excitation source, u(n). The scaled source then goes through the time-varying digital filter. The filter represents the vocal tract which has varying cross-sectional area throughout the tract. In the source-filter model, this varying size corresponds to the different values of the vocal tract parameters. The output of the digital filter is the speech signal, s (n).

The use of the source-filter model to represent speech production is often linked with the linear predictive coding (LPC) model. LPC is a source-filter analysis-synthesis technique that estimates the generation of sound as an excitation source that goes through an all-pole resonant filter. The details of LPC are explained in many places such as in Rabiner's textbooks [16, 17], Bradbury's paper [18] and in Howitt's Otolith [ ]homepage. Here is the basic idea behind the LPC model is described based on Rabiner's textbook.

$$S(n) = \sum a_i s(n- i) + Gu(n) \quad \text{where } i= 1 \text{ to } p. \qquad (2.4)$$

With Figure 2.8 as reference, equation (2.1) shows that a speech sample at time n is equal to a linear combination of the previous *p* speech samples with an added excitation term, *Gu(n)*, where *u(n)* is a normalized excitation and G is the excitation gain. The a_i terms are assumed to be constant coefficients over the speech analysis frame. Speech signals vary with time, so this process is conducted on short segments of the speech signal called frames. 30 to 50 frames per second is normally used as that is enough to give intelligible speech with good compression [26]. Converting the equation using the Z-transform, equation is obtained:

$$S(z) = = \sum a_i z^{-i} S(z) + GU(z) \quad \text{where } i= 1 \text{ to } p. \qquad (2.5)$$

Re arranging the terms from the equation above produces a transfer function
H (z) shown in the equation below.

$$H(z) = S(z)/GU(z) = 1/ (1- \sum a_i z^{-i}) = 1/A(z) \text{ where } i= 1 \text{ to } p. \qquad (2.6)$$

The transfer function H (z) represents the time-varying digital filter shown in Figure 2.8. This is basically an all-pole, autoregressive (AR) model of speech production, where the vocal tract is represented by non-uniform cylindrical tubes concatenated together as shown in Figure 2.7.The terms a_i are the filter coefficients that can be calculated using LPC analysis and p is the number of poles.

2.3 The Effects of Emotion on the Physiological Structure of Speech Production

The respiratory, phonatory, and articulatory movements involved in speech production are mainly controlled by the respiratory organs, the laryngeal muscles and the various articulators. The neocortex is the part of the brain that mostly controls specific motor commands producing the corresponding muscle movements leading to the desired speech sequence [31]. On the other hand, the effects of emotional arousal that can influence the speech production mechanism, even against the speaker's will, are controlled mainly by the limbic system. Emotional arousal also affects speech production via the activation of the somatic nervous system and the autonomic nervous system. The latter consists of the sympathetic and parasympathetic nervous systems (SNS and PNS) [13, 14, 32]. Changes in the activation of SNS and PNS result in variations in blood pressure, heart rate, muscle tension, respiratory patterns, and motor coordination. All these variations eventually modify the respiratory, phonatory, and articulatory systems involved in speech production [33]. Figure 2.8 shows a simplified version of the emotional arousal effect on speech production. Clearly the physiology of speech production can be altered by changes in emotions.

The activation of SNS and PNS increases the possibility of changes happening in speech acoustic characteristics, and these changes can be captured by extracting some speech parameters. Modification in respiratory patterns can cause differences in sub-glottal pressure and this, together with changes in muscle tension can alter the pattern of vocal cord vibrations and the articulation process. Besides that, disturbances in the coordination of

muscular activity involved in producing speech can also result in variations which can be reflected by measurable speech parameters [34].

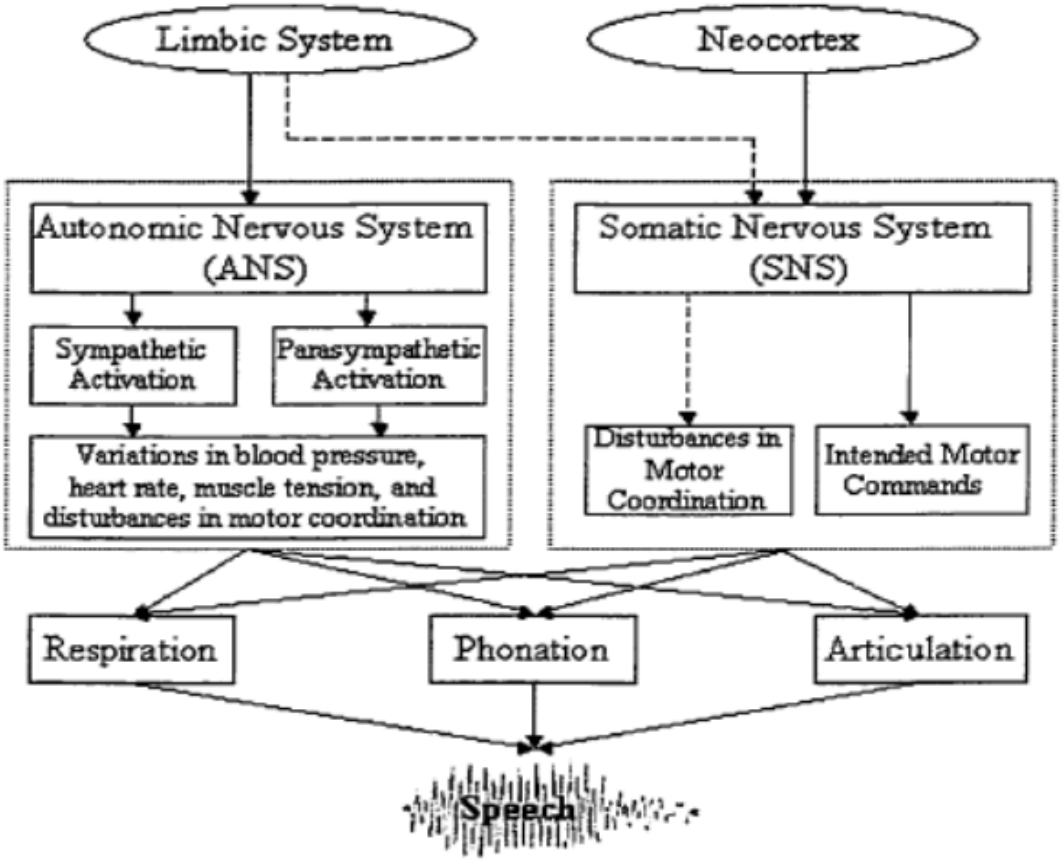

Figure 2.9: The emotional Arousal Effect on Speech Production [19]

The respiratory, phonatory, and articulatory systems are handled by neuromuscular control that has to be tuned accordingly to ensure smooth vocal cord vibration and seamless adjustments between articulatory positions. Changes in respiratory muscles, coordination, and laryngeal musculature can alter the shape of glottal flow waveform. Disturbances in coordination and phonatory muscles could also lead to changes in fundamental frequency, irregularities in the successive glottal cycle durations (vocal jitter), and variations in intensity (shimmer). Changes in articulatory musculature such as increased muscle tone would cause tenseness in the structure of the vocal tract (such as vocal tract resonance walls) and articulators which would eventually affect the resulting frequency spectrum of

the speech. These, together with increased tension in the laryngeal musculature, were suggested to cause higher energy in the upper harmonics. Lack of coordination in the articulatory structures on the other hand would decrease the precision of articulation thus producing relatively narrower formant ranges. This is due to the inability of articulators to reach their targets smoothly [19, 21, 22]. The validity of this research is supported by the known effects of emotional arousal on speech production physiology. Serious suicidal thoughts represent a major change in a human's mental condition. This change includes a wide range of complex emotions and thus, the suicidal vocal patterns are expected to be different from non-suicidal [19].

CHAPTER III

LITERATURE REVIEW

The idea of recognizing distinctive patterns and tone of voice in patients with high risk of suicide was introduced by two clinical psychologists, Drs. Stephen and Marilyn Silverman. Both had experience in treating patients with near term suicidal risk. They began research in the 1980s by collecting and analyzing suicidal tape recordings obtained through therapy sessions in an uncontrolled environment, and notes and interviews made shortly before suicide attempts [20]. They describe the similarity of vocal speech between depressed and suicidal patients but notice changes occur considerably in the tonal quality and acoustical characteristics when the patient enters the suicidal state [8]. Several other researchers continue to study the relation of vocal tract characteristic to depression and suicidal risk.

France et al [6] began the research by extracting and analyzing fundamental frequency (Fo), amplitude modulation (AM), the formants and power distribution (PSD) on speech samples. Among these perceptual qualities, formant and PSD features appeared to be distinguishing vocal features when discriminating between suicidal and major depressed patients compared to the ones collected from control groups. Linear Predictive Coding (LPC) was preferred over Long-term-average (LTAS) to calculate the formant frequencies and bandwidths due to the volume of speech analyzed which made it computationally expensive even though the LTAS approach provides a more accurate representation of the formant properties. The classical Welch method was used when extracting the PSD. The energy spectrum was investigated on the percentages of total energy in frequency sub-bands with a bandwidth of 500 Hz over the frequency range of 0-2000 Hz. It was reported that most energy are distributed in the range of 0-2000 Hz. Features were integrated when performing classification in order to obtain the best parameter combination in distinguishing between the suicidal and non-suicidal groups. Multi-parameter classification was shown to be more effective than classification of single parameters. Ozdas et al [7] studied the discriminating power of lower order mel-cepstral coefficients (MFCC) among

suicidal, major depressed and non-suicidal patients. Vocal tract characteristization using a non-model based approach for near term suicidal risk assessment was the focus of this study. The effects of source (excitation) and filter (vocal tract) on suicidal state were the two domains examined. Vocal jitter and slope of the glottal flow spectrum were two other excitation features that were further investigated related to the excitation signal, whereas in the filter domain, speech features are investigated through cepstral analysis. There were variations among different psychological states with the use of mel-cepstral filter bank coefficients and the results suggested that the use of MFCC features could provide useful measurements for identification of a possible suicidal state. The use of Gaussian mixture models yielded better class approximation when performing classification for individual diagnostic group.

Yingthawornsuk [8] continued the PSD-based study where he also used features extracted from a new proposed method of GMM inspired spectral modeling in his analysis. In the male reading speech PSD ratio only analysis, four 500 Hz PSD ratios were used to build the classifier and a result of 82% correct classification was obtained between depressed and high risk suicidal patients. When the PSD ratio features were combined with the features from the GMM model, 86% classification accuracy was obtained in depressed-suicidal analysis for both male and female interview speech. Reading speech classification produced 88.50% and 90.33% for male and female subjects respectively. These accuracy rates obtained in the analysis of integrated features were obtained by the statistical cross validation approach.

Keskinpala et al [23] centers her work on analyzing the vocal characteristic of high risk suicidal and depressed patients on both male and female speech samples using mel-cepstral coefficients and using energy in the frequency bands. Part of the study examined the method of cepstral mean normalization for compensating spectral variability due to differences in recording environments. The importance of environmental compensation was tested by performing classification with and without compensation and results demonstrated that using no compensation provided better results. Text-dependent speech samples were shown to provide better discrimination analysis in distinguishing suicidal patients compared to the interview speech sample. The cross-validation and testing with all data training were two

methods of resampling used in attaining the classification measurements. The cross validation classifier based-method was demonstrated to perform well as an assessment approach in identifying high risk suicidal patients. Other studies involved changing the number of energy bands extracted from the spectral density, extending the frequency band from 2000 Hz to 3000 Hz and varying the bandwidth of band pass filters for each spectral energy band.

Tolkmitt et al [24] analyzed patients during the course of recovery by obtaining speech samples before and after receiving treatment by comparing fundamental frequency, spectral energy distribution and the formant frequencies of vowels found in identical phonetical content. The fundamental frequency reflects the tension in the muscle tone of the vocal chords and the patient‟s speech samples tend to show decrease in the fundamental frequency when going from the depressive state to recovery. Patients with depressed speech samples experience a decrease in spectral energy below 500 Hz and an increase in spectral energy between 500-1000 Hz after receiving treatments. Moore et al [25] gathered multiple features from three main categories of prosodics, vocal tract and glottal measures. An optimal set of classifiers were chosen based on discriminant analysis techniques and features were selected according to a set of classifiers that were observed to be the most optimized classifier in separation of depressed and control groups. Formant structures and power spectral density measures were found to be the most prominent discriminators in the creation of statistical separation between patient groups. The best discriminators identified for male depressed patients were related to the vocal tract with regards to glottal and formant features. On the other hand, female depressed patients exhibited vocal tract feature related to glottal and energy features as the best discriminators.

CHAPTER IV

SIGNIFICANCE OF THE PAPER

Suicide is a major health problem that has caused many deaths in United States as discussed and demonstrated by the facts in the introduction above. In order to save lives and protect human beings, greater efforts are needed to prevent suicide. The conventional method to assess a patient's suicidal risk is highly subjective, urging a need for additional ancillary tools that might aid the clinical recognition of elevated risk and flag the clinician to perform additional explorative interviewing and therapeutic interaction with the patient. Previous studies have shown that vocal characteristics can reflect the psychological state of patients and can be used as cues for determining suicidal risk. Therefore, studying the acoustic features extracted from the speech of depressed and suicidal patients could lead to a development of an objective diagnostic tool that can assess suicidal risk in a short amount of time. This tool can be used to aid physicians in making appropriate clinical judgments on potentially suicidal patients.

This study represents a small but significant effort in the development of the desired diagnostic tool. The main focus of this paper is the analysis of acoustic features extracted from the speech of patients to determine whether the high-risk suicidal patients can be distinguished from the depressed patients based on harmonic structure of voice speech.

CHAPTER V

METHODOLOGY

The audio recordings were obtained from an ongoing research project in the Vanderbilt University Department of Psychiatry. Recordings were made during a one-to-one treatment session with a psychiatrist. Patients were between 25 to 65 years of age and are divided into High-Risk Suicidal and Depressed categories. Recordings are gathered from the "reading session" where the patients were asked to read from a standardized "rainbow passage". The "rainbow passage" was used due to fact that it contains every sound in the English language and is considered phonetically balanced [19]. We analyzed recordings of 7 High Risk Suicidal patients and 10 Depressed Patients. The recordings were made in a closed room using a TASCAM DR-1 digital field recorder with a sampling rate of 44.1Khz. Before the interview session, patients were asked to count from one to thirty while the interviewer manually adjusted the volume intensity for the recordings. All the recordings were stored as .wav files.

5.1 Voiced and Unvoiced detection

Speech signals are comprised of voiced, unvoiced, and short silence segments that are mixed and combined together. Keskinpala [23] and Yingthawornsuk [8] used the method by Ozdas [7] for voiced/unvoiced detection. According to Ozdas, voiced, unvoiced and silence speech samples can be estimated by segmenting the sampled signals based on their energy values at different levels of the Wavelet Transform (WT). Voiced speech samples exhibit a quasi-stationary behavior and are composed of low frequency characteristics. On the other hand, unvoiced speech samples exhibit noise-like behavior and contains more high frequencies. The sampled signals were separated into segments and for each segment, the energy was calculated for each of several different band levels.

1	2500-5000
2	720-2340
3	320-1080
4	160-540
5	80-260

Table 5.1: Frequency range for each band levels

For this study, a similar method was used for the voiced, unvoiced and silence classification but instead of using the WT to determine the energy bands, a set of third order band-pass filters was applied to each segment of the sampled signal. If a filtered speech segment has maximum energy equal to the total energy in band one, it was classified as unvoiced. The median of the total energy in band three was set as the threshold for separating voiced and silence where energy higher than or equal to the threshold was categorized as voiced otherwise it was categorized as silence. All unvoiced and silence terms were removed and only the voiced terms were collected and concatenated into one new speech signal for further analysis. Finally, the collected voiced signals were split into 20-seconds segments.

Wan Ahmad Sanadi [26], NikHashim [27], and Yingthawornsuk [8] attempted to investigate some human vocal features to analyze their ability to distinguish between depressed and high-risk suicidal patients. A main vocal feature in their work was the Power Spectral Density (PSD) of human speech. Yingthawornsuk analyzed the effectiveness of the vocal features including the PSD in identifying high-risk and suicidal patients.Previous work has showed some success in discriminating between high risk and depressed subjects, however, it is not clear that the

defining characteristics of the particular sound or tonality reported by the Silvermans have been found. This paper attempts to evaluate whether the PSD features used in [8], [17] and [18] correlate with such a sound.

CHAPTER VI

DATA ANALYSIS

First, we separated the sections of the speech (based on 40 ms frames) from the reading passage for each of the reading passage for each of the patients, and the result is divided into 10 second sections of voiced speech. Next a 6 band PSD is calculated for each of the 10 second sections using periodograms as follows. 10 second section was broken into non-overlapping 40ms frames and a periodogram was computed for each frame. The periodograms for each frame were averaged over the entire 10 second section to obtain the average for the section. Six equal bands were extracted in the 0-2000Hz range (each band was 333 Hz wide) using trapezoidal integration. Thus PSD_1 was the power in 0 – 333 Hz, PSD_2 was the power in the next 333 Hz wide band, etc. Each PSD band power was then divided by the total power in bands 1 through 6 resulting in normalized percentage energies so that the sum of PSD_1 through PSD_6 was 1. PSD_6 was removed due to the fact that the contained information is linearly dependent on the other five spectral bands. The resulting PSD bands (PSD_1-PSD_5) were then used as the feature vector for the 10 second section.

Linear Discriminant Analysis has been used herein for training a linear classifier for the high risk and depressed subjects based on the feature vectors just described. The method used is based on Gaussian distributions with differing means and equal covariance matrices [20] and was implemented using the Matlab classify command. The result, then, is a hyperplane that attempts to separate the two populations.

Now that we have a separating hyperplane, we can measure the distance, *d*, from each feature vector to the hyperplane. If the vector appears on the side of the hyperplane corresponding to Depressed subjects we assign a negative sign to distance, and conversely, if it is on the High Risk side it is assigned a positive value.

We examined the distances obtained for each of the High Risk subjects, and based on them attempted to predict to what extent the observed 10 second sections contained the particular tonality observed by psychiatrists. Our hypothesis was that a larger positive distance from the hyperplane would correspond to a greater likelihood of the tonality being present. To get an

idea of the total range of distances from all 7 seven subjects, Figure 1 contains a plot of the distances from all the subjects. The first 4 are from subject 1, the next 9 from subject 2, the next 6 from subject 3, the next 11 from subject 4, the next 7 from subject 5, the next 7 from subject 6, and the last 7 from subject 7.

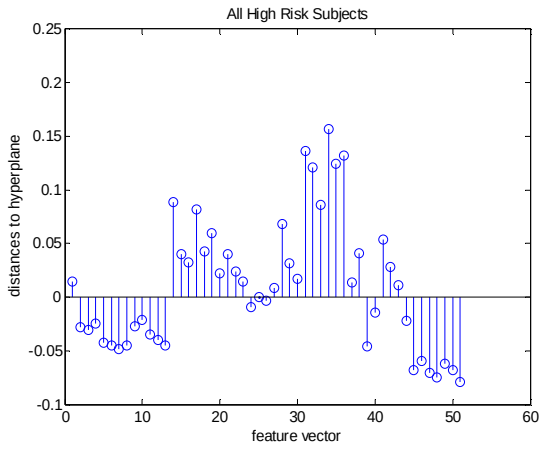

Figure 6.1 Feature Vector vs Distance to Hyperplane Graph

6.1 Predictions

High Risk Subject 1 (points 1-4 in Figure 1) yielded only 4 feature vectors, since the subject completed reading only half the passage before stopping. The first feature value is slightly positive but the rest are negative. We predicted that it would likely that the tonality would not be present for this subject.

High Risk Subject 2 (points 5-13 in Figure 1) yielded only negative values for the distances. Thus we predicted the tonality was absent in this subject.

High Risk Subject 3 (points 14-19 in Figure 1) yielded all positive distances, with half of them greater than 0.05 and a third greater than 0.075. We predicted that the tonality would be present in a substantial portion, but not all of the recorded passage

High Risk Subject 4 (points 20-30 in Figure 1) contains a mix of positive and negative values, with the positive values not tending to be particularly large. We predicted the tonality would be present in a few places in the recorded passage.

High Risk Subject 5 (points 31-37 in Figure 1) contained a large number of large positive values. All the distance values were positive. We predicted the tonality would be substantially present throughout the passage.

High Risk Subject 6 (points 38-44 in Figure 1) yielded a collection of distances qualitatively similar to Subject 4, and thus made the same prediction.

High Risk Subject 7 (points 45-51 in Figure 1) yielded only negative distances and thus we predicted the tonality would be absent.

6.2 Observations

Dr.R. Salomon, is a psychiatrist in the Vanderbilt Department of Psychiatry, and is an expert on depression and suicide. He listened to each recording of the passage and indicated the places where the tonality occurred. We then compared these observations to our predictions.

High Risk Subject 1: exhibited the tonality in about half of the portion of the passage that was recorded. This contradicts our prediction of the tonality being absent.

High Risk Subject 2: exhibited the tonality in a few spots about one fourth of the way into the passage but otherwise did not exhibit the tonality. This can be viewed as a partial agreement with the prediction. Interestingly, the portions where the tonality occurred did not correspond to the largest distance values.

High Risk Subject 3: exhibited the tonality throughout the passage, interspersed with larger portions without the tonality. This somewhat agrees with our predictions.

High Risk Subject 4: exhibited the tonality in the first fourth of the passage, the last fourth of the passage, and at one small spot about halfway through the passage. Which somewhat agrees with our prediction. Additionally, the larger positive values tended to agree with where the tonality appeared.

High Risk Subject 5: exhibited only one small duration of the tonality about three fourths of the way into the passage. This substantially contradicts our prediction.

High Risk Subject 6: did not exhibit the tonality in the passage with partially agree with our prediction.

High Risk Subject 7: exhibited the tonality at one spot about one fourth of the way into the passage, but nowhere else. This partially agrees with our prediction.

CHAPTER VII

FEATURE EXTRACTION, ANALYSIS and CLASSIFICATION

After completing the previous part of Discriminant Analysis, we moved forward to check the features based on calculating the Harmonics and Amplitude of the recordings. Yingthawornsuk [13], Ozdas [7], Nik Hashim did not use this method of determining the features and hence starting from the scratch was the only possible option.

For each patient, k number of 40 milliseconds segments was extracted and a Hamming window was applied, the total number of k small segments varies from one patient to another. Features extracted from the harmonics and amplitude calculation analysis for each patient were collected into a matrix of k rows by P columns where P represents the five harmonics calculated for each patient. These feature vectors were saved as a .mat file and separated into high risk suicidal and depressed groups by an indication of the letter "h" and "d" that were attached at the beginning of the file name.

The discriminant analyses performed on the acquired features were done on the basis of pair-wise analysis classification consisting of high risk/depressed, high risk/remitted and depressed/remitted groups. The decision boundaries for the two-class classification were obtained using a quadratic classifier and a linear classifier. Because of having small data sets, a resampling method was necessary when performing linear and quadratic classifications. The resampling methods that were adopted in this research were Equal Test-Train, Jackknife (Leave-One-Out) and Cross-Validation. The discriminant functions were applied using the "classify" command provided in the MATLAB statistical toolbox.

The "classify" command in MATLAB that was used in this study requires four
Input parameters; test, train, class label and type of discriminant function. In the two-class classification, the class labels are made of a vector of zeros and ones that are stacked

together in a column vector. The size of the column vector depends on the number of training samples. The training data consists of n rows by p columns matrix where the number of training vectors ,n, varies depending on the training sample size. P represents the five harmonics calculated for each 40 milliseconds of voiced speech. Each training sample is associated with a class label, where in this case, the class label would be either zero or one. The test data are assumed to have an "unknown" class label. The classify command will output an estimated class label for the test data according to the training data distribution and classification. The two types of discriminant functions that were used in this classification are the linear and quadratic classifiers.

Initially we normalized the amplitudes into a set of column vectors for every 40 millisecond segments. With this set of column vector matrices we calculated the percentage accuracy of determining whether a 40millisecond segment is classified as high risk or depressed. The following table gives the accuracy of detection for each segment the patients.

h020806nt1	80.6122%
h030905nt2	83.8384%
h032405nt2	73.3970%
h072604nt2	92.6923%
h083004nt1	86.1842%
h091704nt1	69.5297%

Table 7.1: *High Risk Patients (40 ms Segments) Accurate Classifications*

d011106nt1	65.3941%
d020106nt1	39.9299%
d021605nt3	72.7910%
d040505nt2	43.8111%
d061405nt1	50.3606%
d092403hnf	6.4748%

Table 7. 2: Depressed Patients (40 ms Segments) Accurate Classifications

7.1 Quadratic and Linear Classifier

Quadratic Discriminant Analysis (QDA) assumes the observed feature vector follows a Gaussian conditional density distribution. A multivariate Gaussian density function is given by

$$p(\underline{x}|\omega_i) = \frac{1}{(2\pi)^{-\frac{p}{2}}|\Sigma_i|^{-\frac{1}{2}}} e^{\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right]}$$

where {p, ω_i , μ_i and $\sum_i$ } respectively represent the dimensional of the feature space, i^{th} class, i^{th} class mean vector and covariance matrix. QDA assumes that the covariance matrices and mean vectors are not identical for each class. The Quadratic Discriminant function may be formulated as

$$g_i(x) = \ln\left(p(\underline{x}|\omega_i)\, p(\omega_i)\right)$$

And for the Gaussian Distribution case , the discriminant function becomes

$$g_i(x) = -\frac{1}{2}\left(\underline{x}^T \Sigma_i^{-1} \underline{x} - 2\mu_i^T \Sigma_i^{-1} \underline{x} + \mu_i^T \Sigma_i^{-1} \mu_i\right) + \ln\left(p(\omega_i)\right) + c_i$$

$$\text{where } c_i = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_i|)$$

The classification rule for choosing class ω_i is when the estimated $g_i(x)$ is the largest.

In comparison, Linear Discriminant Analysis (LDA) depends on the sample covariance matrix of the training data where a special case occurs when the class covariance matrices are assumed to be identical but with different mean values for every class. For this case, $\Sigma = \Sigma_i$ and c_i becomes a constant, independent of i. In LDA, different classes are assumed to exhibit the same pdf shape but are shifted according to their mean values.

A linear classifier often performs better than a quadratic classifier for small data sets because of the pooled covariance matrix assumption. Averaging covariance matrices on the entire class can produces a higher quality estimate even if it may not be an accurate

assumption. A quadratic classifier may have more flexibility when fitting the data but estimating the covariance matrix for each class increases the variance of parameter estimation. The problem for QDA arises when having small data sets compared to increase in the number of dimensions and classes, because this may cause instability in the parameter estimation.

Nonetheless, there is a trade-off between having the best fit for the data and having a simpler model to work with. A simpler model may not fit the data as well as a complex model but the simpler method might perform better due to its robustness. Classifications using all resampling methods were performed on all possible combinations of features.

7.2 Methods of Resampling

Equal Test-Train Data

Classifications were first performed using the quadratic and linear classifier with the resampling method of Equal Test-Train data where all data in the training set are also used for testing. In other words, this would be an optimistic estimate due to the fact that the testing data are duplicates of the training data. Therefore, to verify the accuracy of the classifier model, a resampling method was applied on the data when performing classification.

Jackknife (Hold-One-Out)

The jackknife method resamples data without replacement and so the training sample will not be duplicated when performing classification. The overall set consists of N patients from all classes to be classified and each patient has different number of row vectors. The implementation of the jackknife method in this research is on the basis of leave-one-out patient instead of per vector. The procedure involves leaving out one patient for testing data set and develops a training data set with the remaining N-1 patients. For the purpose of this study, the class label for testing data set is assumed to be unknown. The classifier output will be a vector of all ones, all zeros or a mixture of both depending on how well the data is classified. This process is repeated by excluding the next

patient from the overall set of data until all patients have been chosen as testing data. This is an optimum method for resampling in the sense that it uses the most amounts of data as training when doing classification.

Cross -validation

When performing classification, it is best to have as much training data as possible to prevent instability and high variance in parameter estimation. Cross-validation is an effective resampling method without replacement for the problem of small data sets. When performing classification, the data sets are partitioned into two sets of samples for testing and training. For this study, the partitioned samples were chosen to be 30% testing data and 70% training data. The testing data are chosen randomly from the original data sets. Similar to the jackknife method, the sample data were chosen according to patients instead of by vectors. According to the available database in this study, each group of patients exhibits an approximately similar number of patients. Therefore, when performing a two-class classification, a 30-70 would result in around 6 patients chosen for testing data and the remainder N-6 patients for training data. Using a random pick, 3 patients were chosen from class 1 (ω_1) and another 3 patients from (ω_1).

By using a cross-validation resampling, the output of the classifier will differ for every run. Therefore, this method was performed iteratively and the averages for all outputs were computed in order to obtain more accurate and stable parameter estimation. If the iteration run is too low, some patients may be randomly picked multiple times and some may not be picked at all. If the iteration run is too high, the computation time will increase. Thus, an optimum iteration of 100 runs was picked for implementing the analysis for this study.

Based on the data from Table 7.1 and Table 7.2 a new algorithm was developed and using cross validation technique a threshold value was found after 100 trials and eventually calculated the true positive from that.

In the next part both the amplitudes and harmonics are taken into a single m file from which feature vectors are put into a single cell array. Extracting the feature vectors from the cell array and for depressed and high risk people cross validation resampling technique has been used over 100 trials to calculate the true positive.

Both the techniques gave us almost same percentage accuracy with a tolerance of around 5%.

7.3 Results using Resampling Techniques

While working we calculated the percentages based on the three different techniques.

1. Using Normalized Harmonics
2. Using Normalized Amplitudes
3. Using Normalized Harmonics and Amplitudes as Feature Vector.

In the first technique the Harmonics of the segments for each patient has been calculated and the normalization based on the basis of first element of the row vector. We calculated five harmonics for each 40 millisecond segment and went on to normalize all of them based on the first frequency and create the m file to store each group of segments for a patient. The equal test train resampling techniques were used on the whole data set consisting of High Risk and Depressed Suicidal Subjects.

Percentage %	**All**	**High Risk**	**Depressed**
Equal Test Train (Linear)	56.5595	30.1	82.33
Equal Test Train (Quadratic)	58.79	21.39	93.75

Table 7.3 : Equal Test-Train Results in Normalized Harmonics Case

Next we proceeded in working with the normalized Amplitudes. Since we calculated five harmonics for each 40 millisecond segment, hence we will have five Amplitudes and went on to normalize all of them based on the first Amplitude of each segment and create the m file to store the segments for a patient. The equal test train resampling techniques were used on the whole data set consist of High Risk and Depressed Suicidal Subjects.

Percentage %	**All Data**	**High Risk**	**Depressed**
Equal –Test- Train (Linear)	64.29	82.62	46.14
Equal –Test- Train (Quadratic)	57.61	91.7983	23.97

Table 7.4: Equal Test-Train Results in Normalized Amplitude Case

Next we moved on using the calculations of Table 1 and Table 2 and applied cross validation resampling technique over 20 trials.

Percentage %	**True Positive**	**False Positive**
Cross Validation	61.23	12.50

Table 7.5: Cross Validation over 20 trials in Normalized Amplitude Case

In the Final One we took all the normalized amplitude vectors and the normalized harmonics and put them in a single m file for every patient. The Feature Vectors for each patient are calculated and classified using Cross Validation resampling technique over 20 trials as this is the one which gives out the most appropriate results above all the resampling techniques.

Percentage %	True Positive	True Negative	Percent Accuracy
Cross Validation	79.12	47.64	63.51

Table 7.6: Cross Validation over 20 trials in Normalized Amplitude & Frequency Case

By checking all the methods we concluded that the feature vectors consisting of normalized harmonics and normalized frequency when used for classification yields somewhat better results than using either of the feature vectors.

CHAPTER VIII

DISCUSSION AND CONCLUSION

During the data pre-processing stage, speech recordings were sampled at 44.1 kHz compared to previous publications where the speech recordings were sampled at 10 kHz. Human hearing is in the range of 20 kHz and human speech frequency is in the range of 200 to 7000 Hz for typical speech activity such as talking, singing, laughing and crying [29]. The signal was sampled at least two times the highest frequency to satisfy the Nyquist Theorem.According to Katz [30] increasing sampling rates will also automatically provide signal-to-noise advantage. Therefore, speech recordings that are digitized with high sampling rate are able to effectively represent the information contained in the waveform.

An interview session would be considered as spontaneous speech because the patient is creating what he/she is saying whereas reading is speech that is controlled and the patient does not create the content of what is being said. Therefore, based on these properties, information contained in the energy spectrum might be distributed differently across all bands. The time when we used the Linear Discriminating Analysis we already knew that the presence of the tonality and chosen feature vectors both exhibited correlation with the high risk state. It was expected to obtain at least partial agreement between predictions and observations even if our hypothesis about distance were not true. If we focus in particular on the results from Subjects 1 and 5 we see two somewhat extreme cases. The values for Subject 1 are either negative or somewhat small, while those for Subject 5 are all positive and quite large compared to the others. In both cases the predictions and observation very substantially disagreed. The other subjects exhibited partial agreement but since this was expected, it is not particularly compelling. We interpret the result as mostly negative, suggesting that these PSD-based features do not clearly capture the tonality reported by mental health professionals.

Secondly at the time when we collected the feature vectors we tried with three other possible options and the one with the combination of both the normalized harmonics and the amplitudes was found to give the best results. Since we used small amount of data we cannot

claim the system to be robust. Also the results might have been different if the process been implemented in the interview sessions. Since the results we found are reasonable there was no need to find and remove the outliers.

If the patients' data removed in spontaneous speech classification were not considered as outliers, the overall result from the classification of both types of speech indicates that automatic speech simply has a stronger ability to discriminate between depressed and high-risk female patients rather than spontaneous speech. However, it is highly possible that the removed data were outliers, because including them in the classification yielded result which are below expectation. The possibility of them being outliers is also supported by the fact that there is one patient whose audio file was present in the spontaneous speech dataset but not in the automatic speech dataset, while most of the other patients have audio files in both datasets. The below expectation initial result yielded from spontaneous speech classification might be also caused by patients sometimes switching between depressed vocal patterns and high-risk vocal patterns during the same interview session. This is based on the observations of Drs. Marilyn and Stephen Silverman [28] where they explained that a patient may not always be in near-term high-risk suicidal state all the time during his or her speech, and they do change between states during the same session. These changes of states might also influence the classification result, as demonstrated in the initial outcome of spontaneous speech analysis.

REFERENCES

[1] JiaquanXu, Kenneth D.K, Betzaida T., "Deaths: Final Data for 2007", National Vital Statistics Reports Volume 58, Number 19, May 2010

[2] AFSP (American Foundation for Suicide Prevention), January 2011, website: http://www.afsp.org

[3] R.Maris, A.Berman, J.Maltsberger, R.Yufit, "Assessment and Prediction of suicide", (pp. 183-201), New York, Guilford Press, 1992.

[4] James R.Rogers, "Suicide Risk Assessment", Wiley e book (pp 259-263), 2001.

[5] Yannakoudakis E. J., Hutton P.J., "Speech Synthesis and Recognition Systems", pp.16-26, 1987

[6] France D.J., "Acoustical properties of Speech as indicators of depression and suicidal risk, PhD Thesis, Vanderbilt University August 1997.

[7]Ozdas, A., "Analysis of Paralinguistic Properties of Speech for near term Suicidal Risk Assessment, PhD Thesis, Vanderbilt University, May 2001

[8] Yingthawornsuk, T., "Acoustic Analysis of Vocal Output Characteristics for Suicidal Risk Assessment, PhD Thesis, Vanderbilt University December 2007.

[9]Louis A.G., Goldine D.G., "The Measurement of of Psychological States through the Content Analysis of Verbal Behavior", University of California Press

[10]VerveridisD. ,Kotropoulos C., "Emotional Speech Recognition : Resources , Features and Methods" , Aristotle University of Thessaloniki , April 2006.

[11]France D.J., Shiavi R.G., Silverman S, Silverman M., Wilkes D.M., "Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk", Vol. 47, July 2000.

[12] Flanagan J.L., "Speech Analysis, Synthesis, and Perception, 2nd ed., Springer-Verlag, New York, 1983.

[13] Rabiner, L.R., Schafer, R.W., "Digital Processing of Speech Signals", New Jersey: Prantice-Hall, 1978.

[14] Moore II E., "Evaluating Objective Feature Statistics of Speech as Indicator of Vocal Affect and Depression", Ph.D Thesis, Georgia Institute of Technology, November, 2003.

[15] Rabiner L. R., Schafer R.W., "Digital Processing of Speech Signals", Prentice-Hall Signal Processing Series, pp38-105, 1978.

[16] Rabiner, L., Juang, B.H., Fundamentals of Speech Recognition, Prentice Hall, New Jersey, 1993

[17] Rabiner, L., Schafer R., Digital Processing of Speech Signals, Prentice Hall (Signal Processing Series), 1978.

[18]Bradbury,J.,"LinearPredictiveCoding",Dec.5,2000,
<http://my.fit.edu/~vkepuska/ece5525/lpc_paper.pdf>.

[19] Ozdas, A., "Analysis of paralinguistic properties of speech for near-term suicidal risk assessment", Ph.D. Thesis, Vanderbilt University, May 2001.

[20] Salisbury D.F., "Researchers measure distinct characteristics in speech of individuals at high risk of suicide", The online research journal of Vanderbilt University, October, 2000.

[21] Scherer, K., "Nonlinguistic vocal indicators of emotion and psychopathology", in C.E. Izard (Ed.), "Emotions in personality and psychopathology", pp. 493-529, Plenum Press, New York, 1979.

[22] Scherer, K. R., "Vocal affect expression: A review and a model for future research", Psychological Bulletin, vol. 99, no. 2, pp. 143-165, 1986.

[23] Keskinpala H.K., Yingthawornsuk T.,Wilkes D.M., Shiavi R.G., Solomon R.M., "Screening for High Risk Suicidal States usinf Mel-Cepstral Coefficients and Energy in Frequency Bands", 15th European Signal Processing Conference, September, 2007.

[24] Tolkmitt F., Helfrich H., Standke R., Scherer K.R., "Vocal Indicator of Psychiatric Treatment Effects in Depressive and Schizophrenics", Journal of communication disorders, Vol.15, pp.209-222, 1982.

[25] Moore II E., "Evaluating Objective Feature Statistics of Speech as Indicator of Vocal Affect and Depression", Ph.D Thesis, Georgia Institute of Technology, November, 2003.

[26] Wan Ahmad Sanadi , "Acoustical Analysis of Speech Based on Power Spectral Density features in detecting Suicidal Risk Among Female Patients" ,MS Thesis ,May 2011

[27] NikHashim, "Analysis of Power Spectrum Density of Male Speech as indicators for high risk and depressed decision", MS Thesis, May 2011

[28] Blumenthal, S.J., "Youth suicide: The physician's role in suicide prevention", JAMA, vol. 264, pp. 3194-3196, 1990.

[29] Wikipedia, Human Voice, http://en.wikipedia.org/wiki/Human_Voice

[30] K. Bob., Mastering Audio, The Art and The Science, Focal Press, pg. 63