ACOUSTIC ANALYSIS OF VOCAL OUTPUT CHARACTERISTICS FOR SUICIDAL RISK ASSESSMENT

By

Thaweesak Yingthawornsuk

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

December, 2007

Nashville, Tennessee

Approved:

Professor Richard G. Shiavi, Ph.D.

Professor A.B. Bonds III, Ph.D.

Associate Professor D. Mitchell Wilkes, Ph.D.

Associate Professor Ronald M. Salomon, M.D.

Professor Ralph N. Ohde, Ph.D.

ACKNOWLEDGEMENTS

This research was supported by the American Foundation for Suicidal Prevention. I would like to thank my advisor Prof. Richard G. Shiavi, Ph.D., D. Mitchell Wilkes, Ph.D., and Ronald M. Salomon, M.D. for their help and guidance throughout all phases of this research. I thank my outstanding committee members A.B. Bonds III, Ph.D. and Ralph N. Ohde, Ph.D. for their valuable feedbacks.

My gratitude goes out to Thai Royal Government and King Mongkut's Institute of Technology, Thonburi for my educational opportunity and financial support in United States. I would like to thank all of my friends who shared their knowledge and life experience with me at Rajamangala Institute of Technology, Bangkok Technical Campus in Thailand.

With my deepest appreciation of thankfulness, I would like to share my academic achievement with my Yingthawornsuk family and my loving and supportive wife, Pla, for their encouragement and support through this long effort. Without them, I would not be able to complete my academic goals.

TABLE OF CONTENTS

Page

Chapter

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Suicide is a major public health problem in the United States among adults and young people and is showing an increasing trend every year. It is the eleventh leading cause of deaths in the American population. The recent statistics show that approximately 32,439 people attempted suicide successfully and the overall rate was 10.9 suicide deaths per 100,000 people reported in 2004 [1]. People who committed suicide suffered from emotionally related disorders, commonly a clinical depression [2]. Depression has been reported as one of the most common precursors to suicidal risk [3], [4]. Fifty percent of persons who commit suicide are diagnosed with serious mental condition.

The work to prevent suicide involves the evaluation of degree of near-term suicidal risk in individual patients. The assessment of an individual's suicidal risk currently requires experienced clinicians in diagnosis of the psychiatric disorder based on the individual's history information, psychological testing record, self-report, reports by others, and recordings of current situation during the clinical interview with a psychiatrist. Current risk assessment is a time-consuming procedure and the data and information required in clinical diagnosis may not be available in urgent situations requiring immediate clinical judgment. Experienced clinicians and computer-based interpretation of data from all relevant sources are the recent diagnostic advancement but only a few computer-based diagnostic tools are available in hospitals at the present time.

Evaluation of risk of committing suicide in persons is the important task that can provide the decisive information in determining when a patient should be hospitalized voluntarily or involuntarily. Additional indicators of suicidal risk can help to mandate hospitalization where the risk might have been misdiagnosed and improve assessment accuracy prior to discharge of that patient.

The human voice is a source of important information regarding the physical, psychological state, and mental health of a speaker. The acoustical properties of speech representing the non-content sounds have been determined to reflect the intensity of the individual’s psychological state [5]. However, the existing link between speech acoustics and psychological state relating the suicidal behavior is very complex and hard to comprehend. Much further study is needed for proving the feasibility that some vocal features can be used to monitor perceptual changes in speech production and articulation caused by the severity of psychological state, which differ from those of normal persons. Depression has long been studied and reported to have a major impact on the speech production that consequently alters the vocal characteristic of speech. Prosody, spectral energy distribution, formants, pitch, mel-frequency cepstral coefficients (MFCC) and glottal ratio/spectrum (GLR/Spec) have previously been demonstrated to correlate with the depressive state in patients [6], [7], [8], [9], [10], [11], [12].

The purpose of this research was to study the acoustic features that can be used to identify three diagnostic groups: depressed patients, suicidal patients, and remitted patients (recovery from depression after treatment), and to demonstrate the practicality of the proposed computer-based vocal indicator that will provide the rapid assessment of the depression and suicidal risk in individual patients based on the specific acoustic features used in designing the single or multi-parameter classifier. The main aim of this dissertation was to investigate the effect of depression and suicidal risk on the acoustical properties of speech, and to determine if any significant difference exists between the identified vocal correlates of these psychiatric states.

The motivation of this task was the lack of progress in development of new objective diagnostic tool for identification of diagnostic groups of patients suffering from depression and suicidal risk. These risk indictors are not proposed to be used as the stand-alone diagnostic tool for the assessment of depression and suicidal risk, but as an additional assistive tool that can be integrated with the existing diagnostic measures and techniques to decrease the clinical effort in diagnosis of the psychiatric disorders.

The first aim was to assemble a database of audio recordings to represent three diagnostic groups: high-risk suicidal, depressed, and remitted patients. Two different audio samples recorded during interview and reading sessions were processed to form speech sample sets representing individual patients of diagnostic groups for vocal feature analyses.

The second aim was to develop computer-based algorithms to analyze and extract: first, the proportions of total energy in spectra of speech of individual patients using the Power Spectral Density (PSD) estimation based on Welch's method [13] and, second, the Gaussian Mixture Model (GMM) of the spectral structure of the vocal tract response with the use of the cepstrum estimation [15,16] and the modification of Expectation-Maximization (EM) algorithm [14].

The third aim was to statistically measure the discriminative properties of the individual features derived from a basis of the PSD-based and GMM-based spectral modeling feature extractions [17], to evaluate the recognition rates on the different sets of reduced primary features for determining the best primary feature set that provides the greatest accurate recognition rate, and to classify the diagnostic patient groups using the optimal feature set in a quadratic classifier incorporating a procedure of cross validation. The studies were performed independently on the female and male subject populations.

A study on the acoustic analysis of vocal output characteristics of high-risk suicidal, depressed, and remitted speech samples in males using solely the PSD-based vocal features in discriminant analyses is presented in Chapter III. A study on the acoustic analysis of vocal output characteristics of high-risk suicidal, depressed, and remitted speech samples in females using the PSD-based and GMM-based spectral modeling features in discriminant analyses is presented in Chapter IV. A similar study on the acoustic analysis of vocal output characteristics of high-risk suicidal, depressed, and remitted speech samples in men using the PSD-based and GMM-based spectral modeling features in discriminant analyses is presented in Chapter V. Conclusions and suggestions for the further work are present in Chapter VI.

CHAPTER II

BACKGROUND AND SIGNIFICANCE

2.1 Speech Production

This section focuses on the background information of human speech production, emotional speech, and vocal correlates of the psychiatric disorders, depression and suicidal risk.

2.1.1 Physiology of Speech Production

When dealing with speech signal analysis, understanding human speech production is essential. It is important to understand how human speech organs develop speech in the form of a sound waveform, which can be perceived by the human ear. Speech is a sequence of different sounds based on the message to be conveyed [18]. Sounds produced in speech are characterized by their articulatory gestures – the position and movement of vocal folds, tongue, lips, teeth, velum, and jaw – and their form of excitation. In traditional linguistic models, a finite number of distinguishable, mutually exclusive sounds, called phonemes, comprise a language. The speech apparatus consists of three major subsystems: respiratory, laryngeal, and articulatory subsystems [19]. Each of these subsystems plays a different role in the speech production system.

The Nervous System comprising of the Central Nervous System (CNS) and the Peripheral Nervous System (PNS) serves as a controller to coordinate the muscle groups in speech production. The respiratory subsystem serves as a power supply by providing airflow and pressure that enters the larynx. The Laryngeal System can be considered as the sound generator, since the source of most speech occurs in the larynx where vocal folds can partially or completely obstruct the airflow from the lungs. The articulatory subsystem consists of the vocal tract and all of its articulators. It is also called a sound modifier because changes in sound depend on the position, shape, and movement of the articulators.

Speech production can be viewed as a filtering operation where the speech signal is produced by exciting the vocal tract with the air forced from the lungs. This excitation is either quasi-periodic pulses of air, known as a glottal flow waveform, resulting in voiced sounds, or turbulent flow of air causing unvoiced sounds. In voiced speech, the voicing source occurs at the larynx locating at the base of the vocal tract, where puffs of air are produced by a periodic opening and closing of the glottis. The glottis is the small slit between the vocal folds and it regulates the air pressure from the lungs. Phonation (vibration of the vocal cords) occurs when the vocal cords are sufficiently elastic and close together and there is a sufficient difference between the sub-glottal pressure (below glottis) and the supra-glottal pressure (above glottis) [18]. In the case of unvoiced sound, the excitation of the vocal tract is more noise-like. The unvoiced speech makes extensive use of broadband noise, caused by turbulent air flow through a constriction in the vocal tract. This form of excitation is usually modeled as noise and can occur with or without voiced excitation. Transient excitation is generated when pressure behind a point of total closure in the vocal tract, built up by airflow, is rapidly released by removing the constriction. Finally, speech is radiated through the lips, nose, or both lips and nose. The vocal tract is a tubular passageway comprising of muscular and bony tissues, which begins at the glottis and ends at the lips and nose. Figure 2.1 shows its components. The velum, one of the articulators, acts as a valve between the nasal cavity and the oral cavity. It is open for nasalized sounds by connecting the nasal cavity to the pharyngeal and mouth cavities and it is closed for non-nasalized sounds by separating the nasal cavity from the rest of the vocal tract.

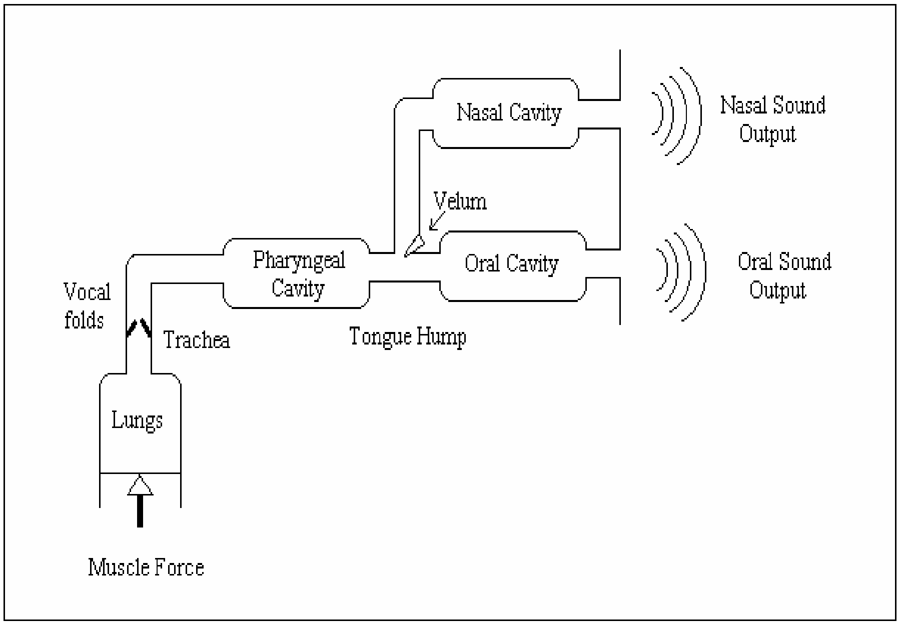


Figure 2.1 Physiological system of speech generation [20].

The vocal tract can be thought of as a filter that can be modeled as an acoustic tube with resonances called formants, and anti-resonances. Moving the articulators of the vocal tract alters its shape, which in turn changes its frequency response. As volumes of air and the corresponding sound pressure waves pass through the vocal tract, based on its shape, the energies at and near the formant frequencies are amplified, while the energies around the anti-resonant frequencies are attenuated. This phenomenon generates different sounds.

2.1.2 Model of Speech Production

Human speech production can be simplified and modeled as a source-filter system. The mathematical representations now can be formulated for main components of the speech production system. Figure 2.2 shows the overall view of the speech production process as the face model and Figure 2.3 presents a general block diagram of the speech production system. In this source-filter model, speech is defined to be the convolution of an excitation source with a time varying linear system represented by the vocal tract and radiation effects [21].

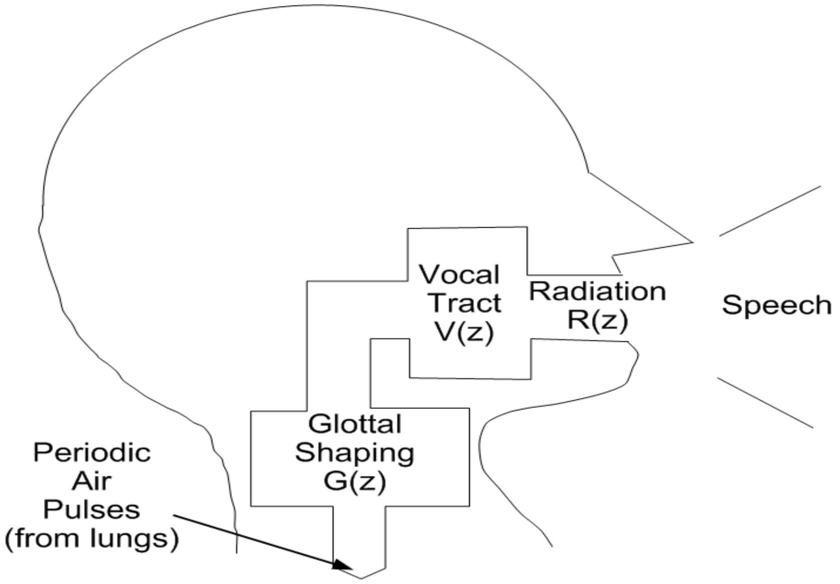


Figure 2.2 Face model of the speech production.

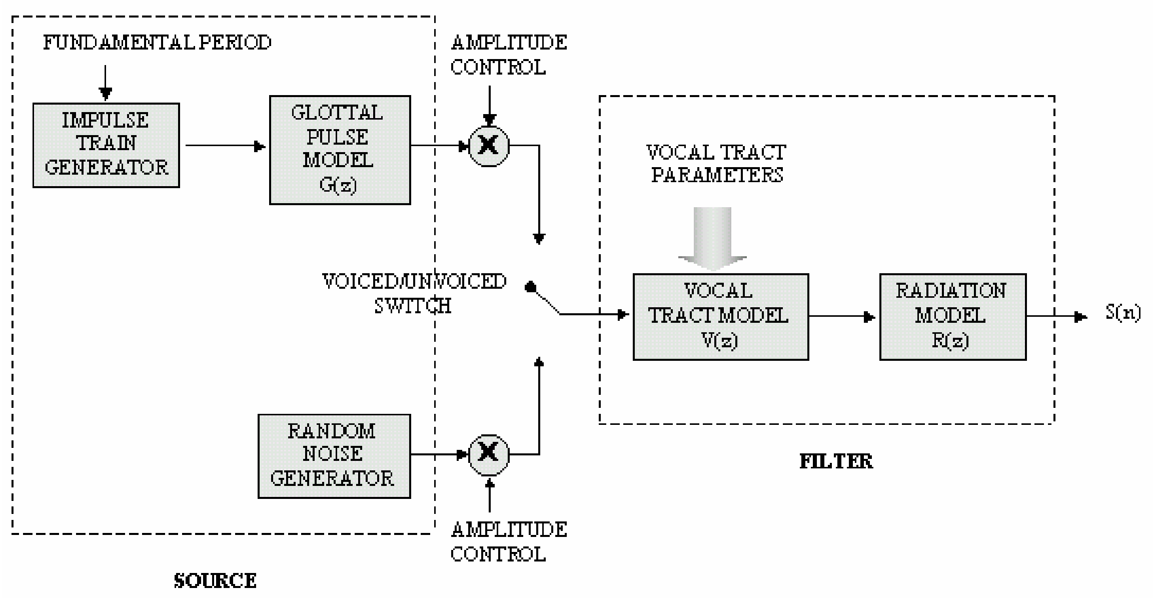


Figure 2.3 General discrete-time model of the speech production [21].

The complete z-transformation is represented as:

$$S(z) = G(z)V(z)R(z) \tag{2.1}$$

where, $S(z)$ - Speech waveform

$G(z)$ - Glottal pulse train

$V(z)$ - Upper vocal tract (Formant frequencies)

$R(z)$ - 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 Excitation Model

The excitation source generates a signal that is either a train of glottal pulses for voiced speech or random noise for unvoiced speech. The impulse train generator produces a sequence of unit impulses, spaced by the desired pitch period. It is the reciprocal of fundamental frequency, the frequency at which the vocal cords vibrate. This signal in turn excites a linear system whose output is the desired glottal wave shape. A gain parameter controls the intensity of the voiced excitation. For unvoiced sounds, the excitation model is much simpler. All that is required is a source of random noise with a gain parameter to control the intensity of unvoiced excitation.

2.1.2.2 Vocal tract Model

A widely used model for the vocal tract is based upon the assumption that the vocal tract can be represented as a concatenation of lossless tubes with different cross-sectional areas $\{A_k\}$ and lengths l, as depicted in Figure 2.4. The motivation behind this assumption is that the lossless tube models provide a convenient transition between continuous systems and discrete systems. A simple model for the vocal tract can be made by representing it as a discrete-time varying linear filter. If we assume that the variations with time of the vocal tract shape can be approximated with sufficient accuracy by a succession of stationary shapes, it is possible to define a transfer function.

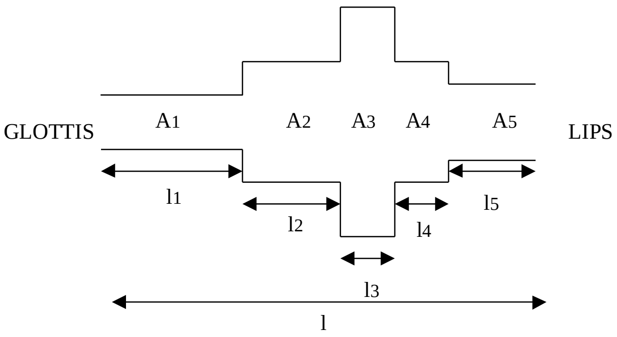


Figure 2.4 Nonuniform acoustic tube formed by cascading acoustic tubes with different cross-sectional areas and lengths.

It is well known that for non-nasal voiced sounds the vocal tract is decoupled from the nasal tract and the transfer function of the vocal tract has no zeros [22]. Thus, the vocal tract can adequately be represented as an all-pole filter, $V(z)$ which is a transfer function represented in the complex z-domain as:

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}} \tag{2.2}$$

where G, $\{\alpha_k\}$ and N are the gain, filter coefficients and number of poles, respectively and they depend upon the area function of the vocal tract. The representation of the vocal tract for unvoiced and nasal sounds should include the antiresonances (zeros) as well as resonances (poles). However, since the zeros of the vocal tract for unvoiced and nasal sounds lie within the unit circle in the z plane, each factor in the numerator of the transfer function can be approximated by multiple poles in the denominator of the transfer function [23]. Thus, an explicit representation of the antiresonances by zeros of the transfer function is not necessary. The all-pole model defined by equation 2.2 can approximate the effect of antiresonances on the speech wave in the frequency range of interest to some accuracy.

2.1.2.3 Radiation Model

The speech signal pressure wave is related to the volume velocity at the lips through the radiation impedance, $R(z)$. In reality, the vocal tract terminates with the opening between the lips. Therefore, it is necessary to model the transmission from the mouth to a given point in space (i.e., a microphone). A reasonable model for the radiation assumes the lip opening as an orifice in a sphere. In this model, at low frequencies, the opening can be considered as a radiating surface with the radiated sound waves diffracted by the spherical baffle that represents the head. If the radiating surface (lip opening) is assumed to be small compared to the size of the sphere (head), the radiation impedance can be approximated as a parallel connection of a radiation resistance and a radiation inductance (a parallel RL circuit). This radiation impedance acts like a high-pass filter reducing more energy in lower frequency range. In fact, for low frequencies, it can be argued that the sound pressure signal at a distance of l_1 from the lips is proportional to the time derivative of the volume velocity at the lips with a time delay of l_1 / c, where c is the speed of sound [24]. The ratio between the sound pressure signal at a distance of l_1 from the lips and the volume velocity at the lips can be represented in the z-transform notation as:

$$R(z) = K(1 - z^{-1}) \tag{2.3}$$

in which K is a constant related to the amplitude of the volume velocity at the lips and the distance from the lips to the recording device [21].

2.2 Physiological Effects of Emotion on Speech Production

The responses of the sympathetic nervous system (SNS) and parasympathetic nervous system (PNS) are affected directly by emotional state. Increased activation of the Sympathetic Nervous System (SNS), especially during the emotions of anger and fear, results in an increase in the heart rate, blood pressure and distribution of blood to the exterior muscles. Further effects of increased SNS activity are changes in the rate, depth and pattern of respiratory movements, and reduction in secretion from the salivary glands, leading to an increase in viscosity of saliva and drying of the mouth. Muscle tremor has also been identified as the physiological effect of the emotions such as fear, anger, and grief. Increased activation of the PNS is usually induced by the feelings of dejection, defeat, and grief, and the resultant physiological response is opposite of the SNS response. Heart rate, respiratory movements, blood pressure, and blood flow to the extremities are all reduced. The PNS is capable of selectively tuning the activity of individual structures in the body and can be subjected to some degree of voluntary control. The sympathetic nervous system is more uniform in its control mechanism and less specific [25]. The divisions of the autonomic system can significantly influence the articulatory and respiratory processes involved in speech production. The effects of increased activity in SNS and PNS can have a direct influence on the control of various respiratory, phonatory and articulatory movements and; therefore, change the vocal quality. Respiratory patterns influence the vocal intensity variations, continuity and discontinuity of speech. The mass, tension and dryness of the vocal folds modify the vibratory patterns, and thus change the spectral and frequency characteristics of the glottal pulses. Increased muscle tone and disturbed coordination also affect the articulation process. The excessive muscle tension produces tenseness or constriction in articulatory structures and causes a variation in the frequency spectrum of sound that is produced. Tense voice seems to have comparatively higher energy in the upper harmonics. Alterations in the muscle structures activated by the emotional states mediate the speech quality [25], [26], [27]. The validity of the current study relies on the conceptual link between speech and emotion that the affected acoustic properties of speech should reflect intensity of emotional state.

2.3 Vocal Correlates of Depression

The studies of the effects of psychiatric illness on free speech have long been noted on the vocal characteristics of patients. Vocal patterns that indicate reduced intensity, monotony, lack of intonation, imprecise articulation, and lack of stress were often associated with severe depression. Vocal parameters that have been investigated in terms of their relationship to depression are fundamental frequency (pitch period), distribution of spectral energy, and formants (spectral structure of speech pattern). Eldred and Price [28] studied the effects of treatment on voice pattern of one patient extensively for over a 13 month period of psychoanalysis. They found decreases in pitch, rate and volume in speech of patient during the depressive state. Roessler and Lester [29] investigated the relationship between emotions and power characteristics of the patients' free speech during psychotherapy. They demonstrated that the speech parameters were related to the different emotional states that the patient experienced during psychotherapy and the high correlations were observed between and speech parameters and perceptual ratings assigned by a group of judges.

Newman and Mather [30] performed a perceptual study on the speech samples collected from a group of 40 depressed patients. Their study identified the distinguishable vocal qualities in the speech of patients suffering from different forms of depression. In their study, patients suffering from classical depression showed "dead, listless" voice qualities. The pitch range was narrow with infrequent stepwise pitch changes. Speech tempo was slow with frequent pauses and hesitations. Emphatic accents were lacking. For patients suffering from chronic states of gloom, self-pity and dissatisfaction, their articulation was fairly crisp and their pitch distribution revealed long gliding intonations extended over a wide tonal range. Moses [31] supported Newman and Matter's findings in a later study. He reported the patient's depressed voice as uniform and monotonous. His results suggested that depressed patients spoke in a lower and narrower vocal range.

Hargreaves and Starkweather [32] studied the voice patterns of eight depressed patients and they attempted to track changes in mood using power spectrum analyses. They observed an anticipated pattern of increase in overall power with greater increase in higher formants for several patients. In a later study, Hargreaves, Starkweather and Blacker [33] investigated the correlation between mood ratings and power spectra before, during, and after treatment in 32 depressed patients. The voice qualities in interview speech recorded from all patients were observed for loudness, high, low, smooth and harsh voice representing as acoustical characteristics reflected in the power spectrum of the voice. The average spectrum for each consecutive 5 seconds of the patient's speech was estimated. The multiple regression of the voice was computed to develop a method of predicting the mood rating from the voice spectra of a particular patient. Predictions of mood ratings with relative to voice spectra significantly correlated for 25 patients and correlations were found to be the highest for those patients showing the considerable change in their mood.

Distribution of energy in spectra of speech has long been investigated in a number of studies [34], [35], [32], [36], [37], [7]. The results of these studies exhibited lower energy for depressive states. After treatment, most of the subjects showed a pattern of increase in the overall energy content. However, the results concerning the amounts of increased energy in high and low frequency ranges of the spectrum with respect to each mental state were not as consistent. Some of these studies reported greater increase of energy in the low frequency band (below 500 Hz) as compared to that in the high frequency band [35], [36], whereas others reported a greater increase in higher formants [32], [37] as a result of treatment. In a group study performed by France et al. [7], samples taken from the depressed patients were found to have the greater energy in a higher frequency range of the spectrum in comparison to healthy control subjects.

Tolkmitt et al. studied the formant patterns extracted from the speech samples of the recovering patient through the vowels that occur in identical phonetic context. They also investigated the relationship between the precision of articulation and severity of depression [36]. The formant frequencies measured before treatment were found to correspond more closely to the neutral formant frequencies produced when the vocal tract is in the resting position. These neutral frequencies are 500, 1500, and 2500 [22], [76]. It was described that before therapy, analyzed vowels were pronounced with less articulatory effort, since their first formants were closer to 500 Hz. As a result of treatment for depression, greater articulatory efforts are made to pronounce vowels such that vocal tract constrictions increase and; consequently, vowel formants reach the expected values. The recovering female patients were reported to have their first formant frequency of vowel “a” changed from 455 Hz to 877 Hz (in mean value). Their proposed results agreed with the predictions in that disturbances in muscular coordination of articulatory structures cause the reduction in articulatory precision, thus yielding the narrower formant frequency ranges due to a failure of articulatory movement reaching the positions to shape vowel sounds [34], [38].

Their results agree with the later study proposed by France et al. [7]. The first formant frequency of the major depressed speech was found to be higher compared to that of the healthy controls. The significant differences of the first and the second formants bandwidths found in classifying normal speech and depressed speech were also reported by France et al. The first formant bandwidth in depressed speech was determined to be larger while the second formant bandwidth was narrower [7].

Moore et al. compared the results of the speaking pattern recognition using prosody, formant and glottal ratio/spectrum as the classifying features in differentiating a non-depressed control group of individuals from a patient group suffering from a clinical diagnosis of depression. The classifiers designated by using glottal ratio/spectrum and formant bandwidths as the most powerful discriminating features were found to produce the best separation in patient groups [10]. By using only features derived from the formant bandwidths in discriminative analyses, the accurate clustering as high as 97.3 % was yielded from classification of male speech and 97.8% was yielded for female speech. While the accurate classification scores of 98.7% and 98.9% were obtained from using the glottal ratio/spectrum as discriminator in classification analyses for male speech and female speech, respectively.

2.4 Vocal Correlates of Suicidal Risk

Investigation of vocal parameters and their correlations to suicidal risk was first initiated by the work of Drs. Stephen and Marilyn Silverman, who have been treating severely depressed and suicidal patients for over forty years since 1960's through 2000. They began to collect and analyze recorded suicide notes and interviews made shortly before suicide attempts. The results from their study suggested that voice can provide the important information about immediate psychological state. They discovered the significant perceptual changes in the vocal qualities of the depressed patients when they became near-term suicidal. This leads to the hypothesis that suicidal state can be associated with changes in speech production and articulation that differs from non-suicidal persons [5]. The Silvermans, who provided tape recordings of patients' speech samples to make up the suicidal patients database as well as financial contributions and guidance, made continuing research in this area possible.

In 1995, Campbell [6] investigated the statistical properties of fundamental frequency to determine whether or not a person is at imminent risk of suicide. The speech segments collected from a group of 1 female patient and 2 male patients were analyzed for their distributions of fundamental frequency. These speech samples were recorded at the time when they were diagnosed as being suicidal at one time and not suicidal at other times. Therefore, the patients served as their own experimental and control subjects. The statistical properties such as skewness, kurtosis, and coefficient of variation were determined from the fundamental frequency distributions and used as measurements in discriminant analyses [6]. Her analysis based on linear classification yielded 22.7% apparent error rate (APER). This pilot study has been followed continuously until now by many studies dealing with seeking other paralinguistic speech parameters that would be proved to be capable of being more reliable discriminators of psychiatric disorders.

France [7] studied the vocal parameters such as fundamental frequency, amplitude modulation (AM), formants, and power spectral density (PSD) properties as various indicators of patients' suicidal risk. He investigated these parameters among diagnostic groups of high-risk suicidal, major depressed, dysthymic and control patients. The vocal parameters of fundamental frequency and amplitude modulation were statistically analyzed for range, variance, mean, skewness, kurtosis, and coefficient of variation. These statistics served as the observations in classification analyses. The first three frequencies and bandwidths of formants were estimated from speech samples of all diagnostic patient groups. The proportion of the total energy in the first four 500 Hz sub-bands of 0-2,000 Hz was estimated by the classical PSD method and the ratios of energy to total energy over frequency sub-bands were calculated and used as vocal parameters for investigation of their discriminative properties and for designing classifiers. This specific frequency range has been reported to have much more energy of speech spectrum distributed than other frequency ranges above 2,000 Hz [7].

The results of female study suggested that the vocal parameters derived from amplitude modulation, formants, and PSD features were determined to be the most effective features in distinguishing dysthymic and major depressed speech from healthy control speech. On the other hand, the results of male study demonstrated that the vocal parameters of fundamental frequency, formants, and PSD features were most effective in classifying major depressed and high-risk suicidal speech from healthy control speech. As a result of the depressed-suicidal analysis in male speech, the amplitude modulation and PSD features were found to be more powerful than the formant and fundamental frequency features in classifying between two patient groups. The fundamental frequency and amplitude modulation features in male suicidal speech appeared to be more similar to those of control speech than major depressed speech. It was concluded that the studied vocal features served well as the effective discriminators of mental states in patients [7].

Ozdas [39] studied the characteristics of source (excitation) and filter (vocal tract) domains separately for their correlations to suicidal risk. Source domain analysis involved in the investigation of two paralinguistic parameters; vocal jitter (a measure of the variations found within the successive periods of the laryngeal vibratory pattern) and slope of the glottal flow spectrum (glottal spectral slope) [39]. In a filter domain analysis, the lower order mel-frequency cepstral coefficients (MFCC) were investigated as features representing the acoustic characteristics of the vocal tract. They parameterized the spectral envelope shape by utilizing the cepstrum estimation incorporating the filter bank analysis. In this investigation, speech samples collected from a group of 30 male subjects were studied.

For the source domain analysis, the maximum likelihood (ML) estimates based on individual parameters of vocal jitter and glottal spectral slope was employed in classifying the categorized patient groups. This study conducted on the pairwise classification analyses between groups of control, depressed, and near-term suicidal patients yielded the promising results. The vocal jitter feature was found as the significant discriminator for a suicidal-control comparison with the 80% correct classification [39]. On the other hand, the glottal spectral slope feature was found to be the significant discriminator for the depressed-suicidal comparison with 75% in classification accuracy, and for the depressed-control comparison with the correct classification score of 90% [39]. By using both vocal jitter and glottal spectral slope features in the multi-parameter classification, the accuracy scores were improved to 85% for the suicidal-control comparison, 90% for the depressed-control comparison, and 75% for the depressed-suicidal comparison. Through filter domain analysis, the first four MFCCs appeared to be the best discriminating features for all diagnostic groups with the accuracies of 75% for the depressed-control comparison, 80% for the depressed-suicidal comparison, and 80% for the suicidal-control comparison [39]. In addition, the ML classification analyses based on an integration of source and excitation domain features by combining the *a posteriori* probabilities of the features at the decision-making stage yielded 88.3% correct classification performance among three diagnostic classes (i.e., near-term suicidal, major depresses and non-depressed control). The improved performance shows that better discrimination can be obtained among different diagnostic classes using the vocal features derived from both source and filter domain analyses [39].

2.5 Significance

Suicide is the major public health problem in the United States among adults and young people showing an increasing tendency every year. It suggests that much more effort is needed to be done to prevent this increasing death caused by suicide. The existing methodologies assessing the patient’s suicidal risk are time-consuming and not widely available at the present time. Previous studies have shown that changes in acoustic properties of speech reflected from the intensity of the patient’s mental state were determined as vocal affects that are possibly used in identifying the severity of the mental state of a speaker. Therefore, the study of vocal output characteristics of speech samples collected from either suicidal patients or depressed patients could lead to the development of new acceptable objective diagnostic tool that can assess the risk of committing suicide more rapidly and can assist clinicians as the additional diagnostic tool in diagnosis of psychiatric disorders.

This study proposed an application of acoustic analysis of psychologically affective speech, which can be integrated with other existing methods, techniques, or tools that are currently available. It may improve accuracy and speed in clinical diagnosis of suicide-related psychopathological disorders.

CHAPTER III

OBJECTIVE ESTIMATION OF SUICIDAL RISK USING VOCAL OUTPUT CHARACTERISTICS

Abstract

Vocal output characteristics of speech have previously been identified as possible cues to the assessment of suicidal risk, and there are evidences that certain vocal parameters may be used to evaluate high-risk suicidal states in persons. Investigation of acoustic properties of speech samples collected from male subjects in one of three diagnostic groups: depression, high-risk suicidal potential and remission, was the focus of this work. Acoustic analyses of proportions of total energy in 500 Hz bands within a 0-2,000 Hz frequency range were reinvestigated and compared among diagnostic groups.

The present results have confirmed that the vocal features derived from the estimation of spectral energy reveal particular potential as possible discriminators of the severity of the persons' mental condition affected by psychological stress, which agree with those of the previously published studies [40]. The results of the comparative classification analyses have reported that the studied vocal features extracted from two different types of audio recordings, interview and reading sessions, exhibited the promising discriminative properties of group separation for distinguishing diagnostic patient groups.

A quadratic discriminant function designed from the spectral energy features yielded the 82% correct classification in classifying the reading speech samples of

suicidal and depressed subjects, and even better for the depressed-remitted analysis (94%) when the interview (spontaneous) speech samples were used in analyses.

The high percentages of correct classification implied that different protocols for collecting speech samples influence the discriminative abilities of individual features in totally different way. It can be suggested that the vocal features derived from the reading speech could presumably be employed in classification analyses as an alternative instead of using those of the spontaneous speech for the assessment of near-term suicidal risk. High classification performance evidently supported that spectral energy features can possibly be the best effective discriminators for monitoring the mental states in persons suffering from the psychiatric disorders.

3.1 Introduction

Suicide is a common outcome in persons with serious mental disorders. However, it remains a phenomenon that is under-researched and poorly understood. Moreover, methods to help to identify persons who are at an elevated risk are sorely needed in clinical practice. This study represents an attempt to identify characteristic vocal patterns in persons with imminent suicidal potential which could lead to the development of new technology to aid in the assessment of suicidal potential. This project brings together investigators from the divergent disciplines of Psychiatry and Biomedical Engineering to study vocal acoustic properties in suicidal states. We will contrast three groups of patients diagnosed as: high-risk suicidal, depressed, and remitted. The present study of vocal acoustic properties in several mental states will use tightly controlled recruitment and recording conditions to replicate and extend findings from recordings made in previous studies to the ongoing study in acoustically controlled clinical interview settings.

In published pilot studies [40], [9], [41], analytical techniques have been developed to determine if subjects were in one of three mental states: healthy control, non-suicidal depressed, or high-risk suicidal. The initial sets of recordings used for these published analyses were made in a wide variety of clinical and technical conditions, without the advantages of an acoustically controlled environment or modern high-fidelity equipment. Most were recorded in the 1960's through '80's by a clinical practitioner (the late Dr. S. Silverman), who routinely taped his therapy sessions. He assembled his set of tapes for just such studies of acoustical characteristics of suicidal speech, which he strongly believed could produce a clinical tool for detection of high-risk suicidal potential. Each selected tape predated a known subsequent suicide attempt with high lethality or completed suicide. Samples for comparison of the subject's speech were taken from the same tapes (e.g. the interviewer's voice) or from recordings made later in more controlled environments. The subjects for comparison were clinically diagnosed and assigned to groups of either healthy controls or non-suicidal depression. In the early studies using these clinical tapes, analysis focused on segments in the high-risk recordings selected by Dr. Silverman as evocative of suicidal speech sounds. The comparison tapes were sampled at random. By this method, diagnostic groupings were successfully separable using parameters of vocal acoustics.

The vocal cues have been used by experienced clinicians as the risk indicators in diagnosing the syndrome underlying a person's abnormal behavior or emotional state [40], [34] but these skills are not in widespread clinical use. Considerable studies have evidently reported that emotional arousal produces changes in the speech production scheme by affecting the respiratory, phonatory, and articulatory processes that in turn are encoded in the acoustic signal [42]. The emotional arousal produces a tonic activation of striated musculature, and the sympathetic, and parasympathetic nervous systems [43]. Changes in heart rate, blood pressure, respiratory patterns, muscle tension, and motor activity transiently alter respiratory, phonatory, and articulatory functions in speech production directly tied to emotions in an *acutely* state-related fashion [44]. Consequently, the emotional disturbances can be expected to cause measurable changes in speech parameters. Certain changes in speech parameters may be specific to near-term suicidal states. Emotional content of the voice can be associated with acoustical variables such as the level, range, contour, and perturbation of the fundamental frequency, the vocal energy, the distribution of energy in the frequency spectrum, the location, bandwidth and intensity of the formant frequencies, and a variety of temporal measures [38]. Research has shown that depression has a major effect on the acoustic characteristics of voice when compared to those of normal controls. Prosody is slower and the spectral energy in the speech is differently distributed over a frequency range between 0 and 2,000 Hz. The spectral energy was estimated and determined to distribute dramatically in spectrum over a 0-2,000 frequency range rather than other higher frequency range above 2,000 Hz [7]. Recently, some studies have been reported to show that suicide also has some effect on vocal characteristics of speech, especially the proportions of spectrum related energy. In this work, the distribution of energy in each 500 Hz sub-bands of a 0-2,000 Hz frequency range was reinvestigated and statistically compared for descriptive measures and classification performance between diagnostic groups of patients. The comparative analyses of different types of speech recorded from clinical interview sessions (sessions in which patients were interviewed by a clinician) and text-reading sessions (sessions in which patients read the “Rainbow” passage) were also carried out to investigate the acoustical characteristics of speech effected by different protocols of audio recording.

This paper is organized as follows: Section 2 provides the detailed descriptions of speech database, procedure for extracting the PSD-based energy features and statistical analyses for verifying a multivariate normal distribution used to represent feature samples in classifications. Section 3 presents the statistics of energy features of interview speech in suicidal, depressed and remitted patient groups. Section 4 presents the statistics of energy features of reading speech in diagnostic groups. Section 5 presents the results obtained from performing feature analysis. Section 6 summarizes the results of the performance evaluation based on the pairwise classification analyses using the spontaneous speech samples. Section 7 summarizes the results of the performance evaluation based on the pairwise classification analyses using the reading speech samples. Section 8 discusses the results obtained from discriminant analyses and all findings from this work.

3.2 Methodology

3.2.1 Database

The audio recordings were obtained from three different groups of patients who were diagnosed with: high-risk suicide, depression and remission from depression. Each studied patient has two types of audio samples recorded. One was recorded during a clinical interview with a therapist, spontaneous speech, and another was recorded during a patient reading of a predetermined part of a book, automatic speech. During the reading sessions, patients read the standardized text, the "Rainbow Passage" [45], which is popularly used in speech science because it contains all of the normal sounds in spoken English and it is phonetically balanced. Two speech samples were randomly selected from the audio recordings of the same patient who participated in interview session and text-reading session. All audio recordings of diagnostic groups of patients were collected from the ongoing study supported by the American Foundation for Suicide Prevention. The patients' ages were between 25 and 65 years. The distribution of male patients is shown in Table 3.1.

Table 3.1 Sample sizes of the categorized groups of patients.

Group	Interview Session	Text-Reading Session
Suicidal	9	10
Depressed	14	13
Remitted	11	9

The portable audio acquisition system used for this study is comprised of a Sony VAIO laptop computer containing Pentium IV 2 GHz CPU, 512 Mb memory, 60 GB hard drive, 20X CD/DVD read/write unit, 250 GB external hard drive, Windows XP OS, and ProTools LE digital audio editor; Digital Audio Mbox for audio signal acquisition; and Audix SCX-one cartiod microphone. Before a clinical interview session began an experienced practitioner set-up the system to control the acoustical recording environment. Moreover, at the beginning of interview session a patient was instructed to count numbers from 1 to 20 at his/her typical speaking rate. Meanwhile the practitioner was adjusting the recording system settings to control audio intensity to possibly be in the same level for all interview sessions. All speech samples were digitized using a 16-bit analog to digital converter with a sampling rate of 10 kHz and an anti-aliasing filter (i.e., 5-kHz low-pass). The background noise and voices other than the patient's voice were removed by using the GoldWave v.5.08 audio editor. This software was also used to remove the silent periods which were longer than 0.5 second. For minimizing the introduction of spurious frequency effects resulting from the abrupt transitions in the edited speech, the segmentation points were selected at the zero crossings or at the beginning of the pauses in the edited continuous speech. The edited segments of the speech sample were tested for voicing and only voiced segments were kept for further analyses. The length of voiced speech was approximately 50% of that of an original speech sample. The voiced speech was detrended and then normalized to have a variance of one for compensating for all possible differences in the recording level. This preprocessing was finally finished by dividing the voiced speech sample into 20-second segments. Each of these 20-second voiced segments was used in spectrum analysis to determine an estimate of power spectral density. The unprocessed speech with approximately 8 minutes was extracted from the interview audio database and approximately 2 minutes extracted from the text-reading audio database to represent two different speech samples of individual patient.

3.2.2 Feature Extraction

3.2.2.1 Energy in Frequency Bands

For each 20-second segment of voiced speech, a Power Spectral Density (PSD) estimate was determined using the classical Welch method with a 400-point Hamming window and non-overlapping consecutive windows. The PSD estimation algorithm was designed and implemented in MATLAB using a 1024-point fast Fourier transforms (FFT). Individual 40-msec frames of voiced speech were analyzed for PSD estimates [7,40]. Six acoustic features were extracted from each estimated spectrum. The first two features were the magnitude value and frequency location of a maximum peak appearing within a 0-2,000 Hz range. Four other features were the percentages of the total energy (PSD_1, PSD_2, PSD_3 and PSD_4) in four 500 Hz sub-bands of total 2,000 Hz frequency range [7,40].

Table 3.2 Total numbers of the 20-second segments of voiced speech.

Classes	Interview Session	Text-Reading Session
Suicidal	139	30
Depressed	104	41
Remitted	100	33

Due to only voiced segments of speech used in analyses, the total numbers of speech segments collected from all diagnostic patient groups were different among patient groups and between two audio recording sessions as well. Table 3.2 shows the total numbers of 20-second segments of voiced speech samples. An outlined procedure for extracting the spectral energy features from frequency bands using the Welch PSD estimation method is:

1) Perform the voiced/unvoiced detection on each patient's speech sample to obtain only the voiced segments of speech samples.

2) Detrend and normalize the voiced speech by subtracting mean, dividing by standard deviation, and then separate into 20-second segments.

3) Divide each 20-second segment into 40-msec frames (500 frames).

4) Estimate PSD of each speech frame using Welch's method with a non-overlapping Hamming window and a 1024-point FFT.

5) Divide the spectral region within a frequency range of 0-2,000 Hz into four equal 500 Hz bands.

6) Calculate the total area (energy) under the spectral curve within a 0-2,000 Hz range and sub-area in each 500Hz band by using a built-in MATLAB function, called "Trapz function."

7) Calculate the ratios of energy in each 500 Hz band to total energy over a frequency range of 0-2,000 Hz.

8) Repeat all procedures starting from step #4 to steps #7 until all 40-msec frames of speech have been analyzed.

9) Calculate means of energy ratios for the present 20-second speech segment and then store all average energy parameters for further statistical analyses.

8) Go to step #3 until all 20-second segments have been analyzed for the present patient's speech sample.

9) Go to step #1 for the next patient's speech sample.

The block diagram of the spectral energy extraction algorithm is shown in Figure 3.1. A single mean vector comprising six spectral energy features was used to represent each 20-second segment of the patient's speech sample.

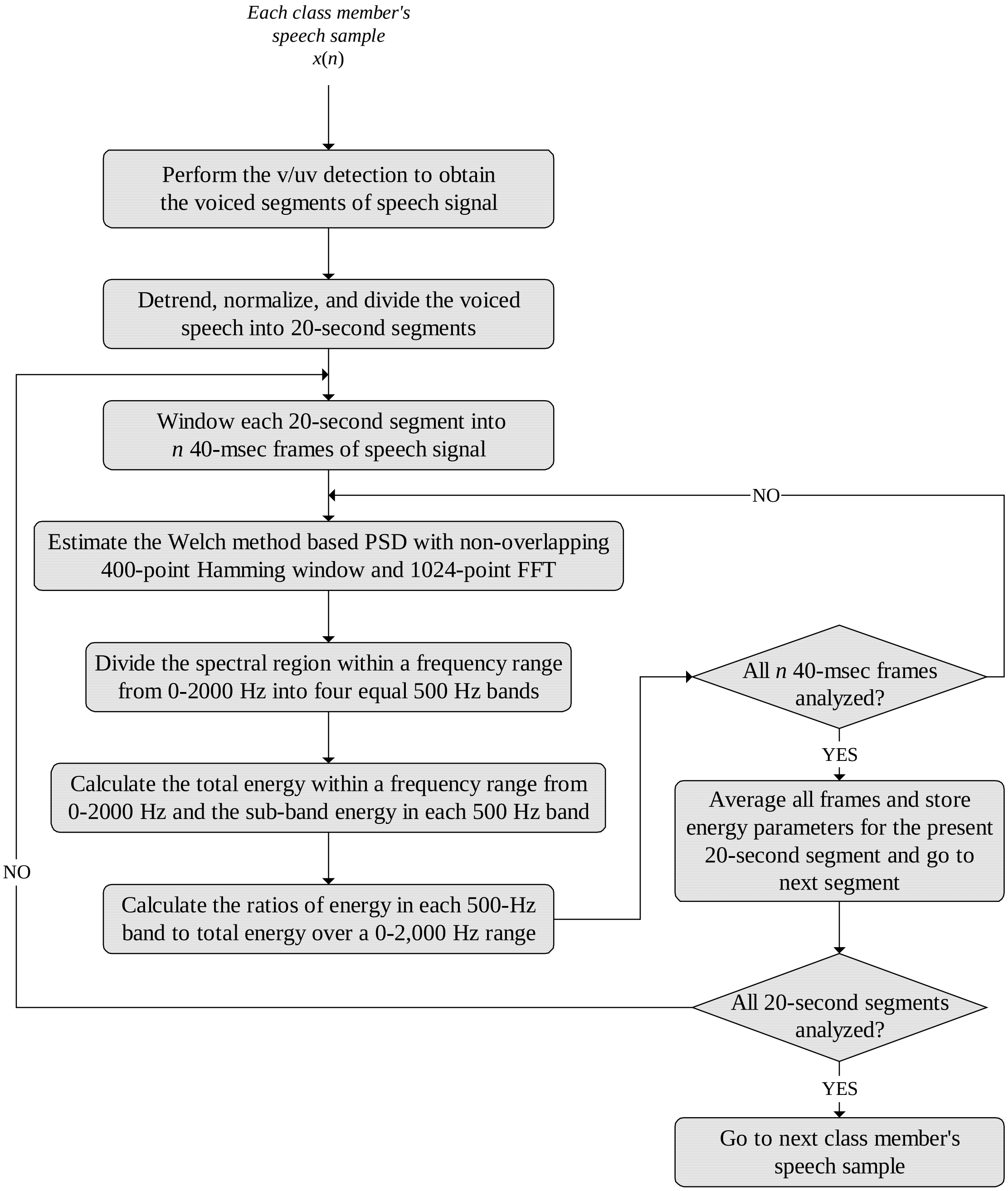

Each class member's speech sample x(n)
Perform the v/uv detection to obtain the voiced segments of speech signal
Detrend, normalize, and divide the voiced speech into 20-second segments
Window each 20-second segment into n 40-msec frames of speech signal
Estimate the Welch method based PSD with non-overlapping 400-point Hamming window and 1024-point FFT
Divide the spectral region within a frequency range from 0-2000 Hz into four equal 500 Hz bands
Calculate the total energy within a frequency range from 0-2000 Hz and the sub-band energy in each 500 Hz band
Calculate the ratios of energy in each 500-Hz band to total energy over a 0-2,000 Hz range
All n 40-msec frames analyzed?
NO
YES
Average all frames and store energy parameters for the present 20-second segment and go to next segment
All 20-second segments analyzed?
NO
YES
Go to next class member's speech sample

Figure 3.1 Flowchart of the feature extraction algorithm for spectral energy.

3.2.3 Comparative Statistical Classification of Class Features

Five acoustic features (i.e., magnitude of maximum peak, frequency location of maximum peak, PSD_1, PSD_2, PSD_3) were taken to form a matrix of energy-based vocal features for a group of patients diagnosed with the same mental disorder. The PSD_4 was not taken into account of this comparative classification due to the property of linear dependency among sub-band energy. Each matrix of features contained *N* rows and *M* columns (*N* x *M* matrix), where *N* was the number of means representing each 20-second voiced segment and *M* was the number of extracted features. The suicidal, depressed and remitted speech samples were mathematically represented by three group matrices.

Each of group matrices was imported into the On-Line Pattern Analysis System (PcOLPARS, PAR Government Systems, La Jolla, CA), and the SYSTAT (SPSS Inc., Chicago, IL) statistical package to perform feature analyses and discriminate analyses [69], [70]. Projection analyses and quantile-quantile (Q-Q) plots were employed to verify an assumption that three groups of imported feature measurements were normally distributed. In addition, Coordinate and Fisher Pair-wise projection algorithms were both employed in PcOLPARS to verify that each data set exhibited the elliptical unimodal scatter characteristic of multivariate normal distributions. The Q-Q plots were also calculated to test the marginal normality of each univariate feature.

The pairwise statistical analyses (i.e., suicidal-depressed, depressed-remitted, remitted-suicidal) were separately performed on each of the vocal features to determine which feature provided the highest power of class discrimination between two diagnostic groups. The studied vocal features were then combined to design a multi-parameter classifier. The correct classification scores and classification performance (i.e., sensitivity (S.E.), specificity (S.P.), positive predictive (PPV), negative predictive (NPV) values) were calculated and collected for all pairwise analyses as measures of effectiveness of classification.

In order to calculate a measure of sensitivity for the suicidal-depressed analysis, a clinical measurement of suicidal speech sample is chosen as a conditional parameter from a confusion matrix of classification. Conversely, a clinical measurement of depressed speech sample is chosen as the conditional parameter instead when a measure of specificity is required. The pairwise analyses involved: calculation and comparison of class covariance matrices, comparison of class features using analysis of variance (ANOVA), application of a quadratic classification. The “hold-one-out” or “Jackknife” method is required for this work to compensate for the small sample size of the speech database [71,72,73]. All discriminating analyses were performed in the SYSTAT software package.

3.3 Experimental Results of Spontaneous Speech Study

The means and standard deviations calculated from the features are summarized in Table 3.3 for all diagnostic patient groups. For the spontaneous speech of suicidal patients, the mean value of frequency location of maximum peak was found to be characterized by lower frequency compared to that of remitted and depressed speech. This feature revealed a decreasing trend of frequency as the severity of the metal state increased. The mean values of magnitude of maximum peak have not shown significant difference between diagnostic groups, but some differences in standard deviations can be observed between groups.

In a comparison of depressed speech and remitted speech, PSD_1 and PSD_3 were found to increase for depressed speech, but PSD_2 was characterized by reduced mean value.

The opposite changes in the mean values of these features can be observed for remitted speech, except for PSD$_4$ whose mean value was found to be no significantly different for both groups.

Table 3.3 Means and standard deviations of spectral energy features for spontaneous speech groups.

	Suicidal	Depressed	Remitted
Peak Magnitude	(20.88, 2.70)	(20.63, 3.34)	(20.92, 1.70)
Peak Location (Hz)	(284.47, 84.89)	(292.02, 58.55)	(331.17, 67.98)
Energy Ratio PSD$_1$	(0.79, 0.08)	(0.79, 0.08)	(0.74, 0.05)
Energy Ratio PSD$_2$	(0.19, 0.07)	(0.18, 0.08)	(0.23, 0.04)
Energy Ratio PSD$_3$	(0.01, 0.01)	(0.03, 0.02)	(0.02, 0.01)
Energy Ratio PSD$_4$	(0.01, 0.00)	(0.01, 0.01)	(0.01, 0.01)

In a comparison of suicidal and remitted speech, it can be observed that the distribution of energy in suicidal speech shifted toward lower frequency band. The proportions of the total energy in the 500-1,000 Hz and 1,000-1,500 Hz sub-bands were reduced while that in lower frequency (below 500 Hz) was increased and there was no significant difference in the mean value of PSD$_4$ determined for both diagnostic groups.

In a comparison of suicidal and depressed speech, PSD$_2$ of suicidal speech was characterized by increased mean value, but PSD$_3$ was conversely decreased. It was inferred that a small amount of proportion of energy shifted from the 1,000-1,500 Hz sub-band to the 500-1,000 Hz sub-band. PSD$_1$ characterized by the similar mean value can be noticed for these speech groups. In addition, another observation reported that the suicidal speech has a relating trend of all features nearly similar to that of the depressed speech rather than that of the remitted speech, except for slight differences in mean values of peak location, PSD$_2$ and

PSD_3 for depressed speech. However, based on a comparison of the proportions of total energy, both suicidal and remitted speech samples have less energy in a 500-1,000 Hz frequency range (sub-band #2) compared with that of the depressed speech.

3.4 Experimental Results of Reading Speech Study

The means and standard deviations measured for the PSD features are presented in Table 3.4. The feature derived from the frequency location of maximum peak was determined to be higher for remitted speech than for suicidal speech and depressed speech. It implies that the highest spectral peak of depressed speech tended to move upward a higher frequency range after the treatment.

Table 3.4 Means and standard deviations of spectral energy features for reading speech groups.

	Suicidal	Depressed	Remitted
Peak Magnitude	(21.52, 2.22)	(20.80, 1.59)	(21.53, 2.08)
Peak Location (Hz)	(298.83, 112.84)	(296.10, 66.11)	(351.65, 74.24)
Energy Ratio PSD_1	(0.78, 0.08)	(0.82, 0.06)	(0.75, 0.09)
Energy Ratio PSD_2	(0.19, 0.08)	(0.16, 0.05)	(0.23, 0.09)
Energy Ratio PSD_3	(0.01, 0.01)	(0.02, 0.01)	(0.01, 0.01)
Energy Ratio PSD_4	(0.01, 0.00)	(0.00, 0.00)	(0.00, 0.00)

In a comparison of depressed and remitted speech, results show that the proportion of total energy in spectra of remitted speech shifted from a frequency sub-band #1 (0-500 Hz) upward to a frequency sub-band #2 (500-1,000 Hz) and no further energy shift was identified above 1,000 Hz. The depressed patient seemed to speak with a greater energy in a frequency range below 500 Hz before a treatment of depression and this amount of energy was determined to decrease in spectrum of the recovering patient's speech, thus resulting in an energy shift upward higher frequency sub-band for remitted speech. The relating trend of features of the depressed speech was obtained to be similar for both spontaneous and automatic speech samples respective to that of the remitted speech.

In a comparison of suicidal and remitted speech, the vocal characteristics of suicidal speech showed that the proportion of energy slightly shifted toward the lower frequency sub-band (below 500 Hz). It can be described as increased PSD_1 in a frequency range of 0-500 Hz and reduced PSD_2 in a frequency range of 500-1,000 Hz. The maximum peak in spectrum of suicidal speech was determined to occur at the lower frequency. Moreover, the mean values in PSD_4 measured for both suicidal and remitted speech samples were characterized equally by the smallest amount of energy as compared to those of other frequency sub-bands.

In a comparison of suicidal and depressed speech, the suicidal speech exhibited the elevated PSD_2, reduced PSD_1 and PSD_3. The opposite changes can be found for depressed speech. In another word, suicidal speech revealed much broader spectrum at a higher frequency range than depressed speech. Changes in the PSD characteristics can be interpreted as the energy shift in the 500-1,000 Hz frequency range (sub-band #2). In addition, it seemed to agree with findings observed in the study of interview speech that suicidal speech has the relating trend of features almost similar to that of the depressed speech, except for PSD_1 and PSD_2 whose mean value was slightly different between diagnostic speech groups.

3.5 Experimental Results of Feature Analysis

The feature analysis of measuring the discriminative power in individual features was performed to determine which vocal feature behaved as the best discriminator. The discriminant ranking method in PcOLPARS [69] was employed for this purpose. This method ranked all five features, except for PSD_4 that we did not include in feature ranking because of the existence of linear dependency among energy proportions in frequency sub-bands and its very small energy, from best to worst based on their discriminant measures. The results of ranking features of interview speech reported that PSD_1, PSD_2 and PSD_3 were obtained to be the best rank-ordered features, whereas peak magnitude and frequency location were both ranked the fourth and the fifth features with very low group separation. In the study of reading speech, these features were also obtained to be the most powerful discriminators of diagnostic patient groups.

3.6 Performance Evaluation Results of Spontaneous Vocal Features

The ranked features comprising of PSD_1, PSD_2 and PSD_3 determined by feature analysis were used in designing a classifier. The results of pairwise discriminant analyses based on using the quadratic classification incorporating the "hold-one-out" procedure in classifying the spontaneous speech groups are summarized in Table 3.5.

Table 3.5 Recognition accuracy results of pairwise analyses of spontaneous speech based on Jackknife technique.

Pairwise Groups	**%Classification**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Suicidal/Depressed	77	0.89	0.63	0.76	0.80
Depressed/Remitted	94	0.94	0.94	0.94	0.94
Remitted/Suicidal	85	0.91	0.76	0.84	0.86

In the suicidal-depressed classification, suicidal speech was well classified from depressed speech with the 77% correct classification. The classification performance measures illustrated that sensitivity (0.89) was comparatively higher than specificity (0.63). This difference in performance measure implies that the classifier performed much better in classifying suicidal speech from depressed speech. In another word, the suicidal speech was more correctly classified than the depressed speech.

As a result of the depressed-remitted classification, the classifier yielded a 94% percentage of correct classification and it was the highest score found for the pairwise analyses of spontaneous speech. Depressed and remitted speech groups were both classified equally by the classifier. It can be inferred by the equal measures of sensitivity (0.94) and specificity (0.94).

The comparative classification analysis of remitted and suicidal speech yielded an accurate score of 85%. Performance measures suggest that the classifier performed more effectively in identifying remitted speech with 0.91 in sensitivity than in identifying suicidal speech with 0.76 in specificity.

3.7 Performance Evaluation Results of Reading Vocal Features

PSD_1, PSD_2 and PSD_3 were used in classification analyses as the most powerful discriminating features. The results of the suicidal-depressed analysis using a quadratic discriminant function demonstrated that depressed speech was most effectively classified from suicidal speech with the 82% classification rate. Depressed patients were identified more correctly by classifier than suicidal patients, as indicated by specificity (0.88) and sensitivity (0.73) in Table 3.6.

Table 3.6 Recognition accuracy results of pairwise analyses of reading speech based on Jackknife technique.

Pairwise Groups	**%Classification**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Suicidal/Depressed	82	0.73	0.88	0.82	0.82
Depressed/Remitted	73	0.85	0.58	0.71	0.76
Remitted/ Suicidal	75	0.73	0.76	0.73	0.76

As a result of the depressed-remitted analysis, the 73% accuracy of classification was obtained and this percentage was also found to be the lowest correct score for pairwise analyses using the reading speech. The effectiveness of classification was much higher in classifying depressed speech than in classifying remitted speech, as referred to by high sensitivity (0.85) respective to low specificity (0.58).

The comparative classification analysis of remitted and suicidal speech yielded a moderately accurate rate (75%) as compared to that of other pairwise analyses. The performance measures were determined to be moderate for this group comparison respective to those of other pairwise analyses. Based on the measures of sensitivity (0.73) and specificity (0.76), the classifier seemed to be equally effective in classifying both remitted patients and suicidal patients.

In order to summarize all results of pairwise classification analyses performing between diagnostic groups, and also to compare differences in classification performance obtained from using spontaneous speech and reading speech in analyses, the comparative plots are illustrated in Figure 3.2.

The largest difference in accuracy rates between different types of speech samples can be observed for a comparison of depressed and remitted speech. By using the spontaneous speech samples in pairwise analyses, the classifier yielded higher accuracies from two out of three pairwise analyses as compared to the classification results of the reading speech samples. Meanwhile the suicidal-depressed classification seemed to yield the most consistent classification accuracies between interview and reading speech studies.

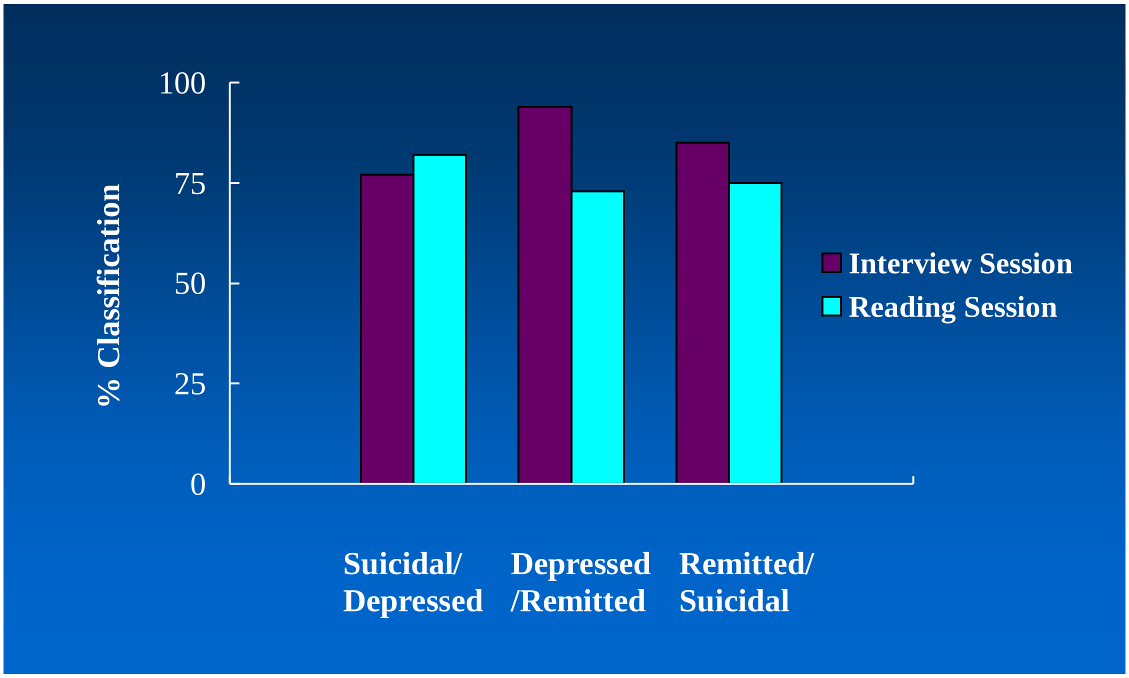


Figure 3.2 Plot of the comparative results of classification analyses of spontaneous and reading speech.

3.8 Discussion

As summarized in Table 3.5, the highest correct classification (94%) implied that the spectral energy features were effective as the best discriminators of depressed and remitted patients. Performance measures were relatively high as compared with other comparisons.

Based on the results of depressed versus remitted and remitted versus suicidal analyses using spontaneous speech samples (Table 3.5), the remitted patients were most effectively classified for both comparisons with the highest measures of specificity (0.94) determined from the depress-remitted analysis and sensitivity (0.91) determined from the remitted-suicidal analysis. It suggested that the spontaneous-speech recording may be effective as powerful acquisition of audio samples for investigating the severity of mental state in patients recovering from depressive state as comparisons with patients diagnosed with depression and suicidal risk. In addition, the acoustic properties of PSD features were more discriminative to distinguish normal speech from severe depressive and suicidal speech.

As summarized in Table 3.6, the results of the pairwise classification analyses using the reading speech revealed that the vocal features characterizing spectral energy were identified as vocal correlates of near-term suicidal risk, and were successfully used in discriminant analysis to distinguish the mental illness affected by depression and suicidal risk. The highest correct classification score of 82% indicated successful identification of different states of mental condition. As shown by performance measures of specificity (0.88) and sensitivity (0.73), the classifier was much more effective in classifying depressed patients than suicidal patients. It may infer that the suicide-related disorder and depression influence the acoustic properties of speech in different way. The intensity of proportion of energy distributions in frequency ranges and shifts of energy in spectrum was diverse, as affected by each disorder. This difference clearly makes the acoustic properties of diagnostic speech groups more separable resulting in the high accuracy of speaking pattern recognition.

As compared between studies of spontaneous speech and reading speech for a suicidal-depressed comparison, the result of the spontaneous speech classification (77%)

indicated to be less effective than that of the reading speech (82%). It may imply that the proportions of energy estimated from different speech types differently distribute over the frequency sub-bands. The inconsistency of classification performance may be resulted from variations of posture that patients had made during clinical interview. The speech production mechanism in patients may be induced physiologically by this variation and consequently affected the acoustic properties of speech.

As shown in Table 3.4, the energy features extracted from the reading speech of depressed patients showed some differences in their energy distribution as compared to those of suicidal speech. Depressed speech was characterized by the greater energy in the 0-500 Hz frequency sub-band, but much less in higher frequency sub-bands and suicidal speech illustrated a trend of energy shift from the 0-500 Hz frequency sub-band to the 500-1,000 Hz frequency sub-band. The changes of energy distribution in spectra of depressed speech can also be observed for higher frequency sub-bands, as referred to by the shift in energy upward from the 500-1,000 Hz sub-band to the 1,000-1,500 Hz sub-band.

The relating trend of PSD features of the remitted speech was more similar to that of the suicidal speech. However, the intensity of energy was much different in frequency ranges below 1,000 Hz. The frequency location of a maximum peak was also found to differ, which shifted toward higher frequency for remitted speech.

The previously published studies have reported that the speech spectra of severe depressed patients showed more energy at higher frequency sub-bands. As a result of treatment, energy distribution in speech of recovering patients was found to shift toward the lower frequencies. This phenomenon has agreed with our findings in that the energy distribution of remitted speech shifted from the higher frequency range (1,000-1,500 Hz) to the lower frequency range (500-1,000 Hz). This finding on remitted speech is consistent for our study in both studies of interview speech and reading speech. Scherer [34] and Tolkmitt [36] reported an increase of energy in the lower frequency bands as a result of treatment, which was consistent with our finding on vocal characteristics of remitted speech.

The recent study conducted by France demonstrated that the energy shift in suicidal speech can evidently be identified at higher frequency bands above 500 Hz when compared with that of major depressed speech. The maximum peak in spectrum of suicidal speech was found to locate at higher frequency as compared to that of depressed speech [7].

As presented in the result section, the shift in energy distribution of suicidal speech was determined to occur in higher frequency bands (above 500Hz) and the frequency location of maximum peak of suicidal speech located at a higher frequency as well. These findings are consistent with the previous results, but there are some contradictions in distribution of energy of remitted speech differing from that of control subjects in previous study [7]. The categorized subjects analyzed in our present study clinically differ from those subjects of previous study. The control subjects studied in previous study were normal persons comprised of licensed psychologist, psychiatrists and therapists, which were totally differed from the remitted subjects (recovering subjects from being depressed) used in our study. The inconsistency in the results of prior study as compared to other studies and our study probably relates with the difference in the acoustical quality of audio samples, which were recorded under different environmental settings and various sources. France analyzed the database of high-risk suicidal speech samples collected from therapy sessions and phone conversations between patients and Dr. Stephen Silverman, and some recordings provided by the Federal Bureau of Investigations. Moreover, the technical specifications of the tape recording equipment used in clinical interviews are unknown and recording environment was not the same for all suicidal patients. All recordings were made in a wide variety of clinical and technical conditions that lacked of an acoustically controlled environment, which possibly had the effect of equipment-based degradation on speech acoustics. These variations in sound quality can reflect in acoustic properties of features in analyses. In this work, all speech samples were recorded from the acoustically controlled environment and the same portable audio acquisition system with a tightly controlled recording condition was used for all clinical interview recordings. Therefore, the speech acoustics of the samples are considered to have similar high fidelity and no sign of equipment-based degradation exists.

Due to a problem of small sample size, the statistical power in analyses can be reduced and the wide confidence intervals of the estimated parameters determined from analyses can be introduced as well. The hold-one-out method was employed in all pairwise analyses to compensate all possible statistical effects. On the larger size of speech samples, the differentiating properties of vocal output energy extracted from the categorized groups of patients would provide more accurate classification results for discriminant analyses. Classification performance measures such as sensitivity, specificity, PPV, and NPV reveal the effectiveness of a classifier designed by the studied vocal features capable of being group discriminators of diagnostic disorders.

Improvement of classification performance may be possible by using a multi-parameter classifier designed from variety of vocal parameters (i.e., formants), feature selection method that help design a classifier, or more reliable techniques of classification such as cross validation [46].

CHAPTER IV

DIRECT ACOUSTIC FEATURE EXTRACTION USING ITERATIVE EM ALGORITHM AND SPECTRAL ENERGY FOR CLASSIFYING SUICIDAL SPEECH IN FEMALES

Abstract

Research has shown that the voice itself contains important information about immediate psychological state and certain vocal parameters can be used in distinguishing the speaking patterns of speech affected by psychological disorders (i.e., clinical depression, suicidal risk). In this study, the acoustic features extracted from the logarithm of the magnitude spectrum of the vocal tract using the Gaussian Mixture Model (GMM) and from the spectral energy in frequency sub-bands over a range of frequencies from 0 to 2,000 Hz were combined and investigated for vocal correlates of psychological state in patients. These vocal features were found to be the powerful discriminators in classifying groups of patients who have a diagnosis of suicidal risk from two other diagnostic groups: depression, and remission; that is, recovery from being depressed after treatment. In this study, two types of speech samples were collected both during clinical interview and text-reading sessions for female patients. They were analyzed and statistically compared to evaluate the effectiveness of speaking pattern recognitions for diagnostic groups. The results demonstrated that the combined vocal features in the depressed-suicidal discriminant analysis provided the highest accuracy (86%) for interview speech classification and 90% for reading speech classification. The features derived from the spectral energy and the GMM-based spectral modeling of the vocal tract exhibit strong

abilities of group discrimination which can be used to indicate the psychological states in diagnostic patients.

4.1 Introduction

The human speech carries the important information regarding the physical and metal healthy of persons. It consists not only of linguistic elements such as phonemes, but also other features which carry nonlinguistic information. These features (i.e., voice quality, speaking rate, energy, pitch, formants) occur spontaneously and are related to the speaker's emotional and/or physiological states. It has been reported that the psychomotor disturbances associated with clinical depression cause speech production mechanism to consequently produce changes in acoustical characteristics of speech. These changes can be referred to as vocal affect.

Depression is one of commonly emotional disorders [2,3,4]. It has been well established that depression is the most common precursor to risk of committing suicide. Suicidal behavior in persons diagnosed with depressive illness was reported to be associated with the serious mental disorders. Reports estimate that 50% of all patients who commit suicide suffer from major depression. However, the relationship between speech acoustics and psychological state related with suicidal behavior is extremely complex. Much research needs to be done to provide convincing evidence that the measurable acoustical parameters possibly represent the perceptual qualities which can be used to monitor the severity of psychological state. Identification of vocal correlates of the speaker's psychological or mental conditions is possibly accomplished by the acoustic analysis of human voice. Presently, methods to help identify persons who are at an elevated risk of committing suicide are sorely needed in clinical practice. This study represents an attempt to identify the characteristic of vocal patterns in persons with imminent suicidal potential which could lead to the development of new technology to aid in the assessment of suicidal risk. At the present time, there are very few accepted objective diagnostic tools, which can provide a valuable supplement to clinical judgment and a quantitative expression of the imminence of suicide risk.

In the early 1980’s the Silvermans began to collect and analyze recorded suicide notes and interviews made shortly before suicide attempts. Their results suggested that voice can provide important information about immediate psychological state. They described that the vocal speech of depressed patients was similar to that of suicidal patients but the tonal quality and acoustical characteristics of speech changed significantly when patients became suicidal.

As reported in formerly published studies [42], [38] the emotional arousal produces changes in speech production scheme by affecting the respiratory, phonatory, and articulatory processes that in turn are encoded in the acoustic signal. Emotional content of the voice can be associated with acoustical variables such as the level, range, contour, and perturbation of the fundamental frequency, the distribution of energy in frequency spectrum, the location, bandwidth and intensity of the formant frequencies, and a variety of temporal measures. The measurable change in vocal parameters affected by emotional disturbances is probably able to be evaluated by utilizing the suitable approach of speech processing with certain acoustic parameters. Research has shown that depression has a major effect on acoustic characteristics of voice compared to normal controls. Certain changes in speech acoustic parameters may be specific to the near-term suicidal state.

In published pilot studies [40], [9], analytical techniques have been developed to determine if subjects were in one of three mental states: healthy control, non–suicidal depressed or high-risk suicidal. Several studies have used vocal tract measures (i.e., formants, mel-frequency cepstral coefficients) and prosody to classify emotional disorders. The vocal tract measures relate to the spectral structure of speech that determines the sounds created and to the prosodic measures involving pitch, speaking rate and energy in voice.

France et. al [7,40] demonstrated the long-term averages of formant information, frequency and bandwidth, and the percentages of total energy in frequency sub-bands with a 500 Hz bandwidth over a frequency range of 0-2,000 Hz as a set of the most dominant acoustic features in classifying groups of control, major depressed and suicidal subjects. Most energy in spectrum was reported to distribute over a 0-2,000 Hz frequency range rather than higher frequencies above 2,000 Hz [7]. Their results have shown that the frequency location and bandwidth of the first formant estimated in major depressed and high-risk suicidal speech were significantly higher than those of normal speech in control subjects. The amount of energy proportion in a 500-1,000 Hz sub-band was determined to be greatly elevated in suicidal speech as compared to that of other diagnostic groups. In addition, the energy shift in spectra of suicidal speech was found to take place at higher frequencies (above 1,000 Hz).

Recently, the vocal features derived from a basis of the spectral energy were reinvestigated and successfully proposed for their powerful discriminative abilities in characterizing the difference of mental conditions among suicidal, depressed and remitted patients [8]. New speech database representing these populations were recorded in an acoustically controlled environment for high quality of sound. The analyses used the same implementation of speech processing as reported in previous chapter for the former study. Results of this reinvestigation confirmed that the spectral-based energy features still appeared as the very powerful discriminators of severe psychopathological conditions [40], [8].

Tolkmitt et al. studied the relationship between the precision of articulation and the severity of depression by comparing the formant patterns in the recovering patients' spoken vowels with those of the depressed patients through the identical phonetic context [36]. The formant frequencies measured for the patients' depressive speech were found to be much closer to the neutral formant frequencies [76] produced when the vocal tract is in the resting position. It was described that before therapy, the analyzed vowels are sounded with less articulatory effort. This less movement of articulation causes the first formant frequency to move closer to the neutral frequency of 500 Hz [76]. As a result of treatment for depression, the greater articulatory efforts are applied to pronounce vowels such that vocal tract constrictions increase and; consequently, vowel formants reach the expected values. The recovering female patients were reported to have their first formant frequency of vowel A went from 455 Hz to 877 Hz (mean value). Their proposed results agreed with the predictions in that disturbances in muscular coordination of articulatory structures cause the reduction in articulatory precision, thus yielding the narrower formant frequency ranges due to a failure of articulatory movement reaching the positions to shape vowel sounds [34], [38].

Ozdas et al. estimated the mel-frequency cepstral coefficients (MFCC) from speech samples and used them in classification analyses to identify the correct diagnostic categories for control, depressed and suicidal patients. In her discriminant analyses, a Gaussian mixture model (GMM) was applied to approximate a probability distribution of feature samples for individual diagnostic groups. Maximum likelihood (ML) classifier assigned a set of feature samples representing for each patient to the diagnostic group giving the highest *a posteriori* probability according to Bayes' decision rule. The GMM-based likelihood classifier designed by a set of the first four low-ordered MFCC yielded an overall accuracy score of 78.33% [9]. Results suggested that the MFCC feature set possibly serves as a good measure of psychologically different condition in suicide related persons.

Moore et al. investigated the acoustical characteristics using prosody, formants and glottal ratio/spectrum in classifying a non-depressed control group of individuals from a group of patients suffering from clinical depression. The multi-parameter classifiers designated by the glottal ratio/spectrum and formants were determined to provide the best identification performance between patient groups [10].

The paralinguistic behavior of depressed patients has long been studied and it is recognized that depression has the effect on the acoustical characteristics of the vocal tract response of patients. Formants are one of the most reliable spectral features representing the distinctive frequency response of voice corresponding to the physical movements of the vocal tract of speakers. Analysis of formants has popularly been employed to study the acoustic properties of emotionally related speech and changes in spectral pattern of formants were identified for the patients' speech during depressive state. Significant changes in the first and second formant frequencies and bandwidths in depressed speech have been reported by several research groups. Although formants were consistently found as vocal correlates of depression, the direction of change of formant frequencies was contrarily different. This inconsistency among former findings obviously relates to the difference in methodology determining estimates of formant frequencies and bandwidths.

The most popular technique to estimate formants in speech is Linear Predictive Coding (LPC). LPC is a form of parametric (model-based) analysis requiring the imposition of a model whose type and order must correspond to the signal for best results [47]. The formant estimates based on the LPC analysis is a model-based representation of the speech spectrum; therefore, the accuracy in estimation can severely be affected by the recording environment. Most LPC assume an all-pole filter model that may disregard some significant spectral characteristics for noise contaminated speech [48]. Such a technique was used formerly by France et al. to estimate the first three formant frequencies and bandwidths that were used to design the integrated classifiers. Also, Ozdas, later on, performed the LPC-based formant analysis to obtain the third and fourth formant frequencies, which were used to normalize the variations of the vocal tract length among individual diagnostic patients.

In our present study, we focused on the vocal features that are able to characterize the different spectral responses of the vocal tract in patients suffering from different psychiatric disorders, mainly depression and suicidal risk. An alternative representation of the characteristics of the vocal tract with the use of GMM was developed. In this work, GMM was purposely applied to approximate the magnitude spectrum of the affective vocal tract due to disorders. This affected spectral envelope of the vocal tract response intrinsically contains information of spectral pattern (i.e., intensity, responding frequency and bandwidth) associated with the psychological state of a speaker.

This GMM-based modeling technique was first introduced by Zolfaghari [49] for fitting a GMM to the Discrete Fourier Transform (DFT) of speech signal to determine the frequency and bandwidth parameters corresponding to resonances of speech. It has been known for the speech production system that the effect of the vocal tract produces a low frequency ripple in the logarithm of the magnitude spectrum, so-called “cepstrum”. While the periodicity of the vocal source (fundamental frequency) manifests itself as a high frequency ripple in the logarithm of the magnitude spectrum [15]. Therefore, the cepstral-domain analysis makes it possible to separate source from filter (vocal tract) characteristics in speech signal, where the vocal tract characteristic is represented by lower frequencies and the fundamental frequency information is represented by higher frequencies.

The low-frequency cepstrum is more advantageous for model approximation of the vocal tract magnitude spectrum than the DFT-based spectrum of speech, since the ripple effect of fundamental frequency (excitation in speech signal) in spectrum can be removed easily by using a simple filtering procedure, called “liftering”. By this procedure the spectral representation of the vocal tract is completely determined. Elimination of effect of fundamental frequency excitation in the DFT spectrum is more difficult to implement. The spectral peaks representing the fundamental frequency and its harmonics inherently mix into those of the vocal tract response for the DFT-based representation. In another word, the spectral characteristics of the vocal tract are more separable from the fundamental frequency characteristics when speech is represented by the cepstrum rather than by the DFT-based spectrum.

The specific features representing the characteristics of the vocal tract, called “energy concentration” in this work, are directly derived from the frequency density (probability density function) approximated from the vocal tract magnitude spectrum using the Expectation Maximization (EM) algorithm. The GMM-based features provide more robust representation of the vocal tract spectral characteristics than LPC-based features in which the desirable speech samples are corrupted by background noises generated from uncontrolled recording environment or by some filtering effects in preprocessing. According to the properties of cepstral-domain transformation, if the linear time invariant property of the vocal tract model exists, the filtering effects are cepstrally represented as constant bias terms and they can easily be removed using the mean normalization techniques [16], [50]. In addition, GMM provides a more fine-detailed modeling of the vocal tract spectrum (different mixture density distributions modeling different responding frequencies of the vocal tract spectrum) and a robust approach for speech processing applications [51].

Significance of discriminative properties in vocal features derived from a basis of spectral energy and energy concentration is main focus of this study. The speech samples collected from diagnostic groups comprising near-term suicidal, depressed and remitted subjects will be analyzed for specific features, then statistically compared to observe the significant differences in their acoustic properties between groups, and finally classified among groups in pairwise manner. Spectral energy and GMM-based vocal tract features will be jointly studied to design the multi-parameter classifiers for individual pairwise discriminating analyses.

This paper is organized as follows: Section 2 provides the detailed descriptions of database, acoustic feature extractions based on the PSD method and the GMM-based spectral modeling, dimensionality reduction of feature space using the discriminant measure and performance evaluation. Section 3 presents the results of feature extraction, GMM fitting, pairwise classifications for interview and reading vocal features. Section 4 discusses results and findings for studies of interview speech and reading speech.

4.2 Methodology

4.2.1 Database

The speech samples were collected from three groups of female subjects carrying a diagnosis of suicidal risk, depression, and remission. Each categorized group consists of 10 subjects. The ages of patients were between 25 and 65 years. This work is part of an ongoing research study supported by the *American Foundation for Suicide Prevention*. Each subject has two types of audio data recorded. The first type of audio sample was collected during a clinical interview with a therapist and the second type was collected during a session of reading a predetermined part of a book. During the post–session of clinical interview, each subject read the standardized text, the "Rainbow Passage" [45], which has been used in speech science since it contains all of the normal sounds in spoken English and it is phonetically balanced. The recording environment and settings were acoustically controlled to be the same for all clinical interviews. This acoustical controlled environment is necessary for quantifying the clean audio recordings from clinical interviews. The audio acquisition system and preprocessing steps as same as reported in previous study [8] were used in this work. Two additional preprocessing steps were made before acoustic analyses. First, all speech samples were tested for voicing and only voiced segments of speech were kept for further analysis. The voiced/unvoiced detection algorithm based on weighting energy of speech proposed by Ozdas et al. [41] was used in this work to decide which section of the patient's speech is voiced, unvoiced or silent. The length of voiced speech was approximately 50% of that of an original speech sample. This percentage seemed to be consistent for most of analyzed speech samples. Second, all voiced segments were detrended and normalized to compensate all possible differences in recording level among the categorized patient groups, resulting in having a variance of speech sample equal to one. In this work, the unprocessed speech with approximately 6 minutes was extracted from database of interview audio recordings and approximately 2 minutes from reading audio recordings to represent individual patients.

4.2.2 Spectral Energy Feature Extraction

A Power Spectral Density (PSD) was estimated from the voiced speech using the classical PSD method based on Welch's theorem with a 512–point Hamming window and non-overlapping consecutive windows. The PSD estimation algorithm based on the 1024-point fast Fourier transforms (FFT) was written and implemented in MATLAB. The frequency spectrum of each 51.2ms frame of voiced speech was calculated. In each estimated spectrum, four energy parameters (energy ratios) were extracted in four different frequency ranges: 0-500 Hz, 500-1,000 Hz, 1,000-1,500 Hz, and 1,500-2,000 Hz. These parameters are the percentages (PSD$_1$, PSD$_2$, PSD$_3$ and PSD$_4$) of the total energy calculated in individual 500 Hz sub-bands of the 0-2,000 Hz frequency range. The individual proportions of energy obtained from all 51.2ms segments were averaged. A single mean vector containing four PSD features were used to represent each patient [8].

By using only voiced segments in analyses, total numbers of voiced segments representing for individual patients were different within group, among groups, and also between two types of audio recording. The detailed procedure for estimating the spectral energy features is similar to that reported in Chapter III, except for the length of processing window that was changed to be longer for this study. A reason for this change is to match the length of window used in the voiced/unvoiced detection algorithm and also to that of the GMM-based feature extraction algorithm for the best experimental results.

4.2.3 GMM Approach to Estimation of Cepstral-Based Features of Vocal Tract Characteristics

This section outlines a new approach of acoustic feature extraction involving cepstrum estimation, pseudo-cepstrum analysis, peak detection algorithm, and model approximation of magnitude spectrum using the GMM probabilistic distribution. All methods and techniques were applied together as a novel approach to extract the acoustic features that are capable of capturing the distinctive acoustical characteristics of the vocal tract corresponding to changes in psychological state.

Our task of modeling the spectral structure of the vocal tract can be achieved with the use of EM algorithm. This basic learning algorithm to find a maximum likelihood (ML) of a mixture model was introduced by Dempster, Laird, and Rubin [14] and extended for a superimposed signal by Schafer [52] and Feder [53]. The EM algorithm is a general method to solve a problem of ML estimation and incomplete data, where some of random variables can be observed and some are hidden. This algorithm is employed to estimate ML parameters for Gaussian mixture distributions, which are approximated from the magnitude spectrum or the DFT magnitude of speech signal.

As depicted in Figure 4.2, a procedure for extracting the GMM-based features from the vocal tract spectrum comprises of two main processes. The first process is to estimate the smoothed cepstrum for each speech frame utilizing the pseudo–cepstrum estimation [54] to determine a suitable length of window for designing a lifter (low–time filter) to capture only the low–time portion of cepstrum (cepstral representation of the frequency response of the vocal tract). The second process is to fit a GMM to the magnitude spectrum via EM algorithm that iteratively estimates ML parameters of the mixture model, and to obtain the specific parameters of the mixture model properly representing the vocal tract spectral characteristics. The vocal tract features comprise of center frequencies (CF's), bandwidths (BW's) and mixture weight coefficients (WC's), which are referred to as means, variances and probabilities of the ML model parameters derived from the best fitted GMM that is successfully approximated from the magnitude spectrum.

4.2.3.1 Cepstrum Estimation of Spectral Envelope

The real cepstrum of the speech signal is defined as the Inverse Fourier Transform (IDFT) of the logarithm of the magnitude spectrum. A block diagram of analysis of the speech cepstrum is illustrated in Figure 4.3. The procedure for estimating the cepstrum of speech signal can be described as follows. First, the log-magnitude spectrum is computed using a 512-point Discrete Fourier Transform for each 51.2ms voiced speech frame and;

then, a computation of the Inverse Fourier Transform of log-magnitude spectrum is performed. Second, the slowly varying component of the cepstrum is determined by liftering the Inverse Fourier Transform of the log-magnitude spectrum with a low-time window whose length must obligatorily shorter than a pitch period of the analyzed frame of speech. By this restricted window, we will obtain the absolute pitch-free cepstrum, as referred to by $c_{vt}(n)$ shown in Figure 4.3. Third, the smoothed spectrum with a slowly varying frequency response is computed by performing the Discrete Fourier Transform on the liftered proportion of cepstrum. This whole process is called “cepstral smoothing” of the vocal tract spectrum. The real cepstral-domain transformation of speech signal can be mathematically written by

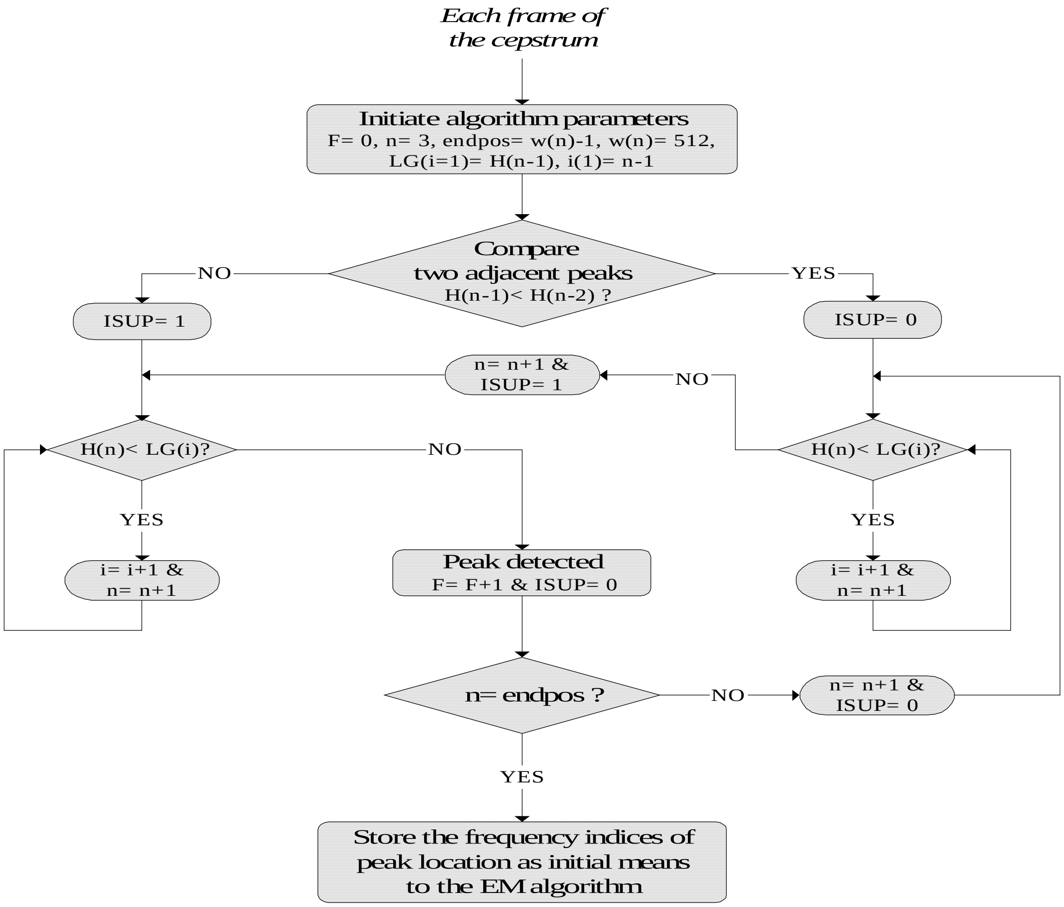
$$c_s(n) = IDFT\{\log(DFT\{s(n)\})\} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|S(\omega)|\cdot e^{j\omega n}d\omega \tag{4.1}$$

$$c_s(n) = IDFT\{\log|P(\omega)|\} + IDFT\{\log|H(\omega)|\} \tag{4.2}$$

$$c_s(n) = c_{ex}(n) + c_{vt}(n) \tag{4.3}$$

in which $s(n)$ is the speech signal, $P(\omega)$ is the Discrete Fourier Transform of the excitation (pulse train), $H(\omega)$ is the Discrete Fourier Transform of the vocal tract response, $c_{ex}(n)$ and $c_{vt}(n)$ are the high-quefrency (a quickly varying component of $c_s(n)$) and low-quefrency portions (a slowly varying component of $c_s(n)$) corresponding to the spectrum of the excitation (pulse train) and the vocal tract impulse response, respectively. The word “quefrency” is introduced by Bogert et al. in 1962 [75] to avoid confusion with “frequency” used for spectrum. The most prevalent terms, cepstrum and quefrency, are the classical paraphrased terms according to the syllabic interchange rule: spectrum (cepstrum) and frequency (quefrency).

Once the cepstral estimate of each 51.2ms speech frame is determined, $c_{vt}(n)$ can be completely separated from $c_{ex}(n)$ by liftering of a low-time cepstral portion. This liftered portion obtained from a whole cepstrum of speech is generally known as “cepstral representation of the vocal tract characteristics.” Due to the inconsistent appearance of fundamental frequency as the largest peak (pitch peak) and its harmonics as a peak train on cepstrum, some speech frames may have a very low peak of pitch period hidden in the background noisy ripples of cepstrum of speech signal. This can cause a problem in designing a low-time window. In this situation, it is very difficult to locate that hidden cepstral peak based on the estimation of classical homomorphic deconvolution. Thus, another method to detect such a peak needs to be considered. The generalized homomorphic signal analysis, namely pseudo-cepstrum analysis, is then employed to overcome this difficulty, which is crucial for obtaining perfectly suitable window length for a low-qrefuency lifter. This window length plays an important role in deconvolving of the vocal tract impulse response and excitation input. The vocal tract system response can severely affected from the spectral distortion caused by the pitch periodicity, when an inappropriate length of window is assigned for liftering. Generally, the most perfect window length of a lifter is often defined as a function of pitch period.



Each frame of the cepstrum
Initiate algorithm parameters
F= 0, n= 3, endpos= w(n)-1, w(n)= 512,
LG(i=1)= H(n-1), i(1)= n-1
Compare two adjacent peaks
H(n-1)< H(n-2) ?
NO
YES
ISUP= 1
ISUP= 0
n= n+1 & ISUP= 1
NO
H(n)< LG(i)?
NO
H(n)< LG(i)?
YES
YES
i= i+1 & n= n+1
Peak detected
F= F+1 & ISUP= 0
i= i+1 & n= n+1
n= endpos ?
NO
n= n+1 & ISUP= 0
YES
Store the frequency indices of peak location as initial means to the EM algorithm

Figure 4.1 Flowchart of peak detection algorithm.

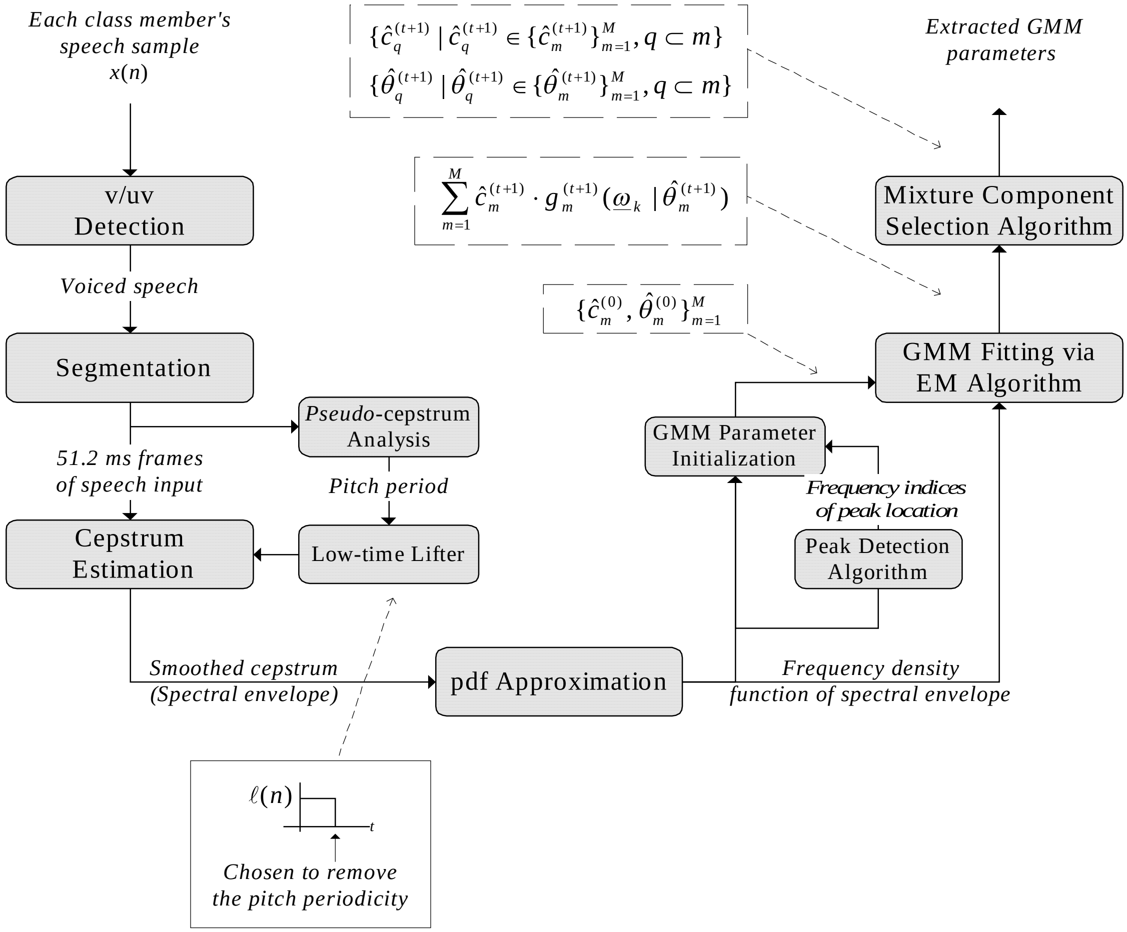

Each class member's speech sample $x(n)$
v/uv Detection
Voiced speech
Segmentation
51.2 ms frames of speech input
Pseudo-cepstrum Analysis
Pitch period
Low-time Lifter
Cepstrum Estimation
Smoothed cepstrum (Spectral envelope)
$\ell(n)$
t
Chosen to remove the pitch periodicity
pdf Approximation
Frequency density function of spectral envelope
Peak Detection Algorithm
Frequency indices of peak location
GMM Parameter Initialization
$\{\hat{c}_m^{(0)}, \hat{\theta}_m^{(0)}\}_{m=1}^{M}$
GMM Fitting via EM Algorithm
$\sum_{m=1}^{M} \hat{c}_m^{(t+1)} \cdot g_m^{(t+1)}(\underline{\omega}_k \mid \hat{\theta}_m^{(t+1)})$
Mixture Component Selection Algorithm
$\{\hat{c}_q^{(t+1)} \mid \hat{c}_q^{(t+1)} \in \{\hat{c}_m^{(t+1)}\}_{m=1}^{M}, q \subset m\}$
$\{\hat{\theta}_q^{(t+1)} \mid \hat{\theta}_q^{(t+1)} \in \{\hat{\theta}_m^{(t+1)}\}_{m=1}^{M}, q \subset m\}$
Extracted GMM parameters

Figure 4.2 Procedure of the energy concentration extraction algorithm based on a GMM fitting of the magnitude spectrum of the vocal tract.

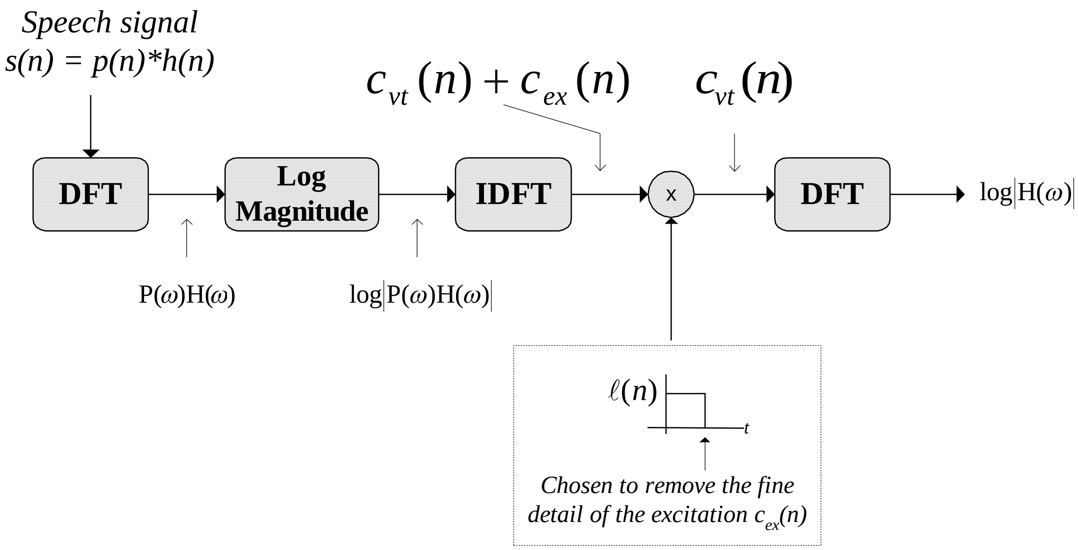


Figure 4.3 Cepstrum analysis of the speech signal.

4.2.3.2 Pseudo-Cepstrum Estimation for Pitch Periodicity

The pseudo-cepstrum analysis is a generalized homomorphic deconvolution [54]. It is used to transform a convolution of two signals (an excitation input in the form of a train of pulses and a model impulse response) into another convolution in which the transformed impulse response is shorter than the original one and better separated from the excitation. A block diagram of pseudo-cepstrum analysis is illustrated in Figure 4.4. The procedure to estimate the pseudo-cepstrum of the speech signal is described as follows. First, for each 51.2ms voiced speech frame, the root power of the magnitude spectrum is computed using the 512-point Discrete Fourier Transform. Then, the computation of the Inverse Fourier Transform of the root-magnitude spectrum is performed. This kind of signal transformation could be called "spectral root cepstra." This specific cepstral-domain transformation can be mathematically described as follows:

The Fourier transform of the speech convolution leads to

$$S(\omega) = P(\omega) \cdot H(\omega) \tag{4.4}$$

$S(\omega)$ represents the speech signal spectrum, $P(\omega)$ is the spectrum of the excitation signal and $H(\omega)$ is the frequency response of the vocal tract model. In real cepstrum computation, the function $f(S(\omega)) = \log | S(\omega) |$ is used. This can be modified for the new computation of the real pseudo-cepstrum with the defined spectrum transformation function $f(S(\omega)) =| S(\omega) |^{\gamma}$, $-1 < \gamma < 1$. By this nonlinear transformation, it converts the spectrum of speech equation (4.4) into another convolution,

$$\breve{S}_{\gamma}(\omega) =| S(\omega) |^{\gamma} = | P(\omega) |^{\gamma} \cdot | H(\omega) |^{\gamma} = \breve{P}(\omega)_{\gamma} \cdot \breve{H}(\omega)_{\gamma} \tag{4.5}$$

The symbols $\breve{S}_{\gamma}(\omega)$, $\breve{P}_{\gamma}(\omega)$ and $\breve{H}_{\gamma}(\omega)$ are introduced for the Fourier Transforms of the magnitude spectra transformed with the parameter γ. By taking the Inverse Fourier Transform, the spectra root cepstra are determined with new sequences of $\breve{s}_{\gamma}(n)$, $\breve{p}_{\gamma}(n)$ and $\breve{h}_{\gamma}(n)$ as the following convolution,

$$\breve{s}_{\gamma}(n) = \breve{p}_{\gamma}(n) * \breve{h}_{\gamma}(n) \tag{4.6}$$

This can be considered as an invertible system that maps $s(n) = p(n) * h(n)$ into $\breve{s}_{\gamma}(n) = \breve{p}_{\gamma}(n) * \breve{h}_{\gamma}(n)$ such that $\breve{p}_{\gamma}(n)$ remains a train of impulses with similar spacing as $p(n)$ but $\breve{h}_{\gamma}(n)$ is more time-limited than $h(n)$. The components in this new convolutional vector space are more easily separated by a simple time-gating procedure. The transformed sequences of the real pseudo cepstra correspond respectively to speech signal $s(n)$, excitation $p(n)$ and vocal tract response $h(n)$. The pseudo-cepstrum analysis has been examined with a male vowel “a” to demonstrate how well it can handle some speech frames in which the pitch periodicity is difficult to find from their cepstra. The male vowel “a” with a sampling frequency rate f_s = 16 KHz was analyzed through the 512-point Discrete Fourier Transform (DFT) and its real pseudo cepstrum was computed with varying root powers of parameter γ. The fundamental frequency of a male speaker is approximately 86 Hz (pitch period = 186). Figure 4.5 shows the male vowel “a” sound. By using a classical homomorphic deconvolution, the causal part of corresponding real cepstrum of vowel “a” was estimated (Figure 4.6). For Figures 4.7-4.10, the cepstral magnitude of the unidentified pitch peak was increasingly magnified by increasing γ value and it now can be identified more easily for its frequency location with a certain value of γ.

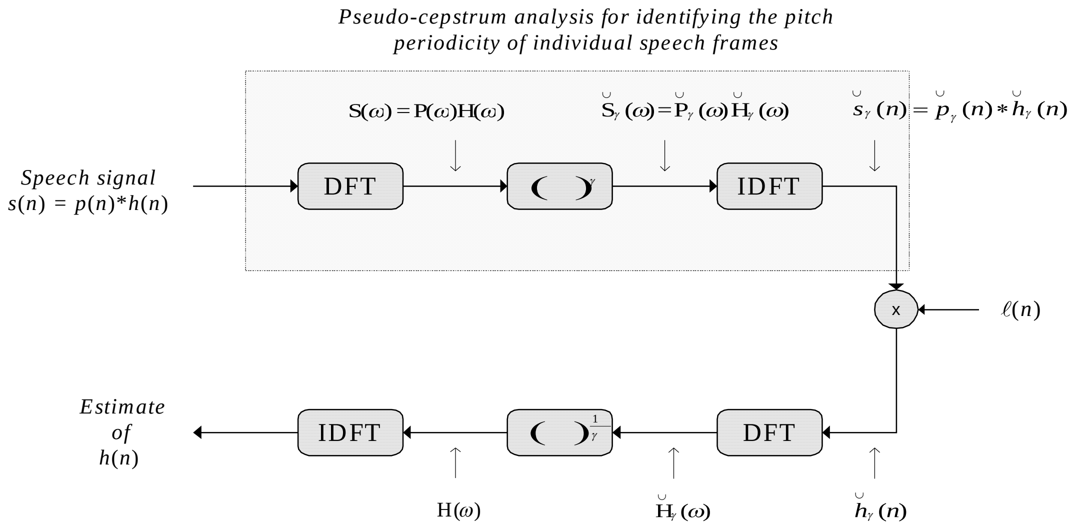

Pseudo-cepstrum analysis for identifying the pitch periodicity of individual speech frames
$S(\omega)=P(\omega)H(\omega)$
$\breve{S}_\gamma(\omega)=\breve{P}_\gamma(\omega)\breve{H}_\gamma(\omega)$
$\breve{s}_\gamma(n)=\breve{p}_\gamma(n)*\breve{h}_\gamma(n)$
Speech signal
$s(n) = p(n)*h(n)$
DFT
$(\ )^\gamma$
IDFT
x
$\ell(n)$
Estimate of $h(n)$
IDFT
$(\ )^{\frac{1}{\gamma}}$
DFT
$H(\omega)$
$\breve{H}_\gamma(\omega)$
$\breve{h}_\gamma(n)$

Figure 4.4 Block diagram of the pseudo-cepstrum analysis.

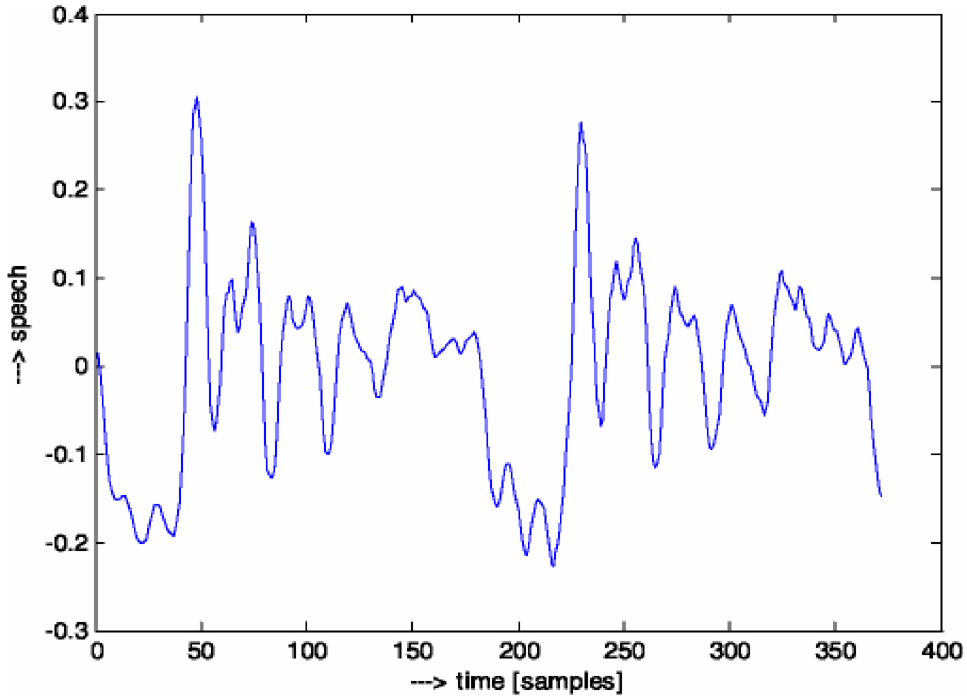


Figure 4.5 Vowel “a” of a male speaker.

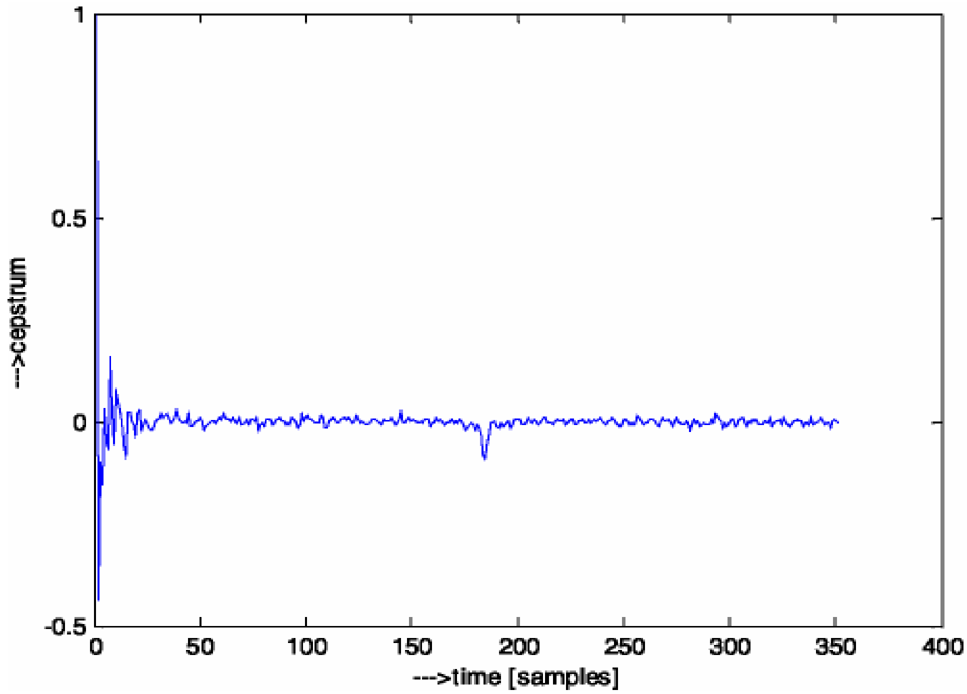


Figure 4.6 Real cepstrum of the male vowel “a”

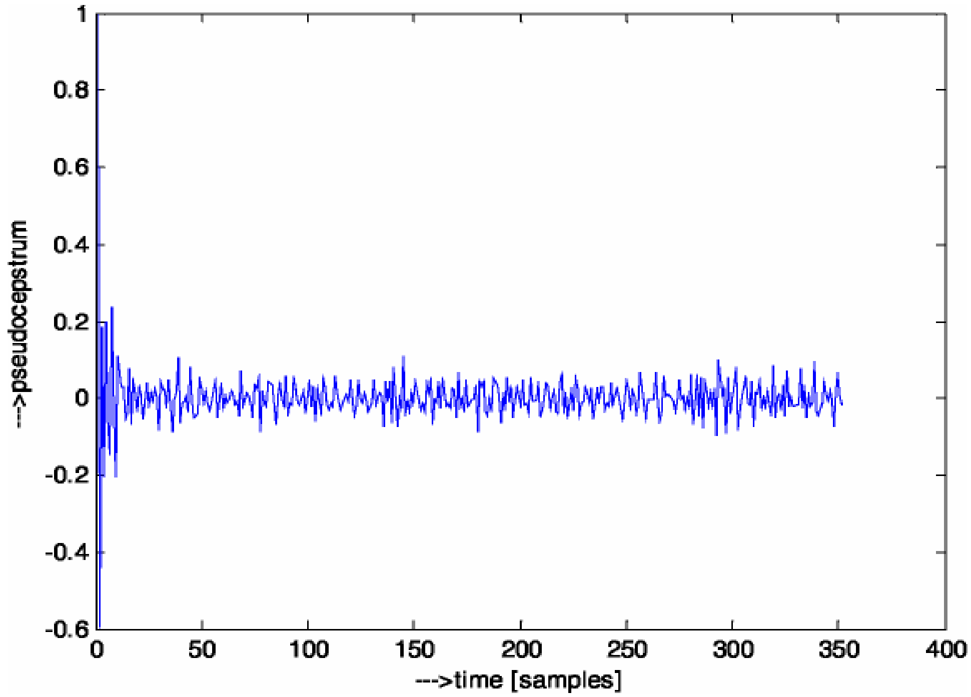


Figure 4.7 Real pseudo-cepstrum for $\gamma = -1$ of male vowel “a”

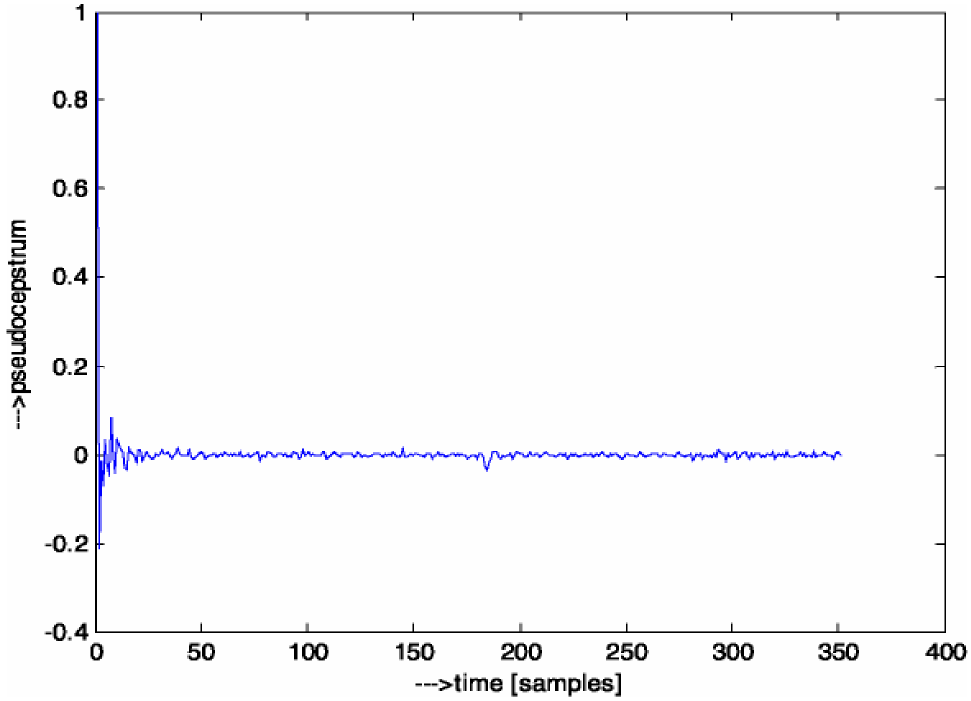


Figure 4.8 Real pseudo-cepstrum for $\gamma = -0.2$ of male vowel “a”

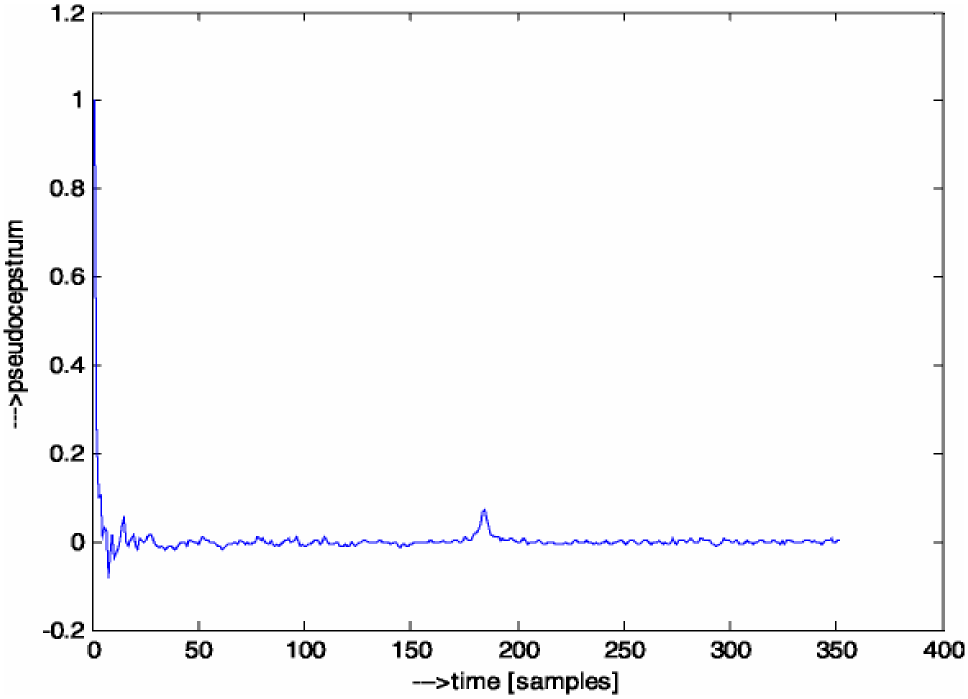


Figure 4.9 Real pseudo-cepstrum for γ = +0.2 of male vowel “a”

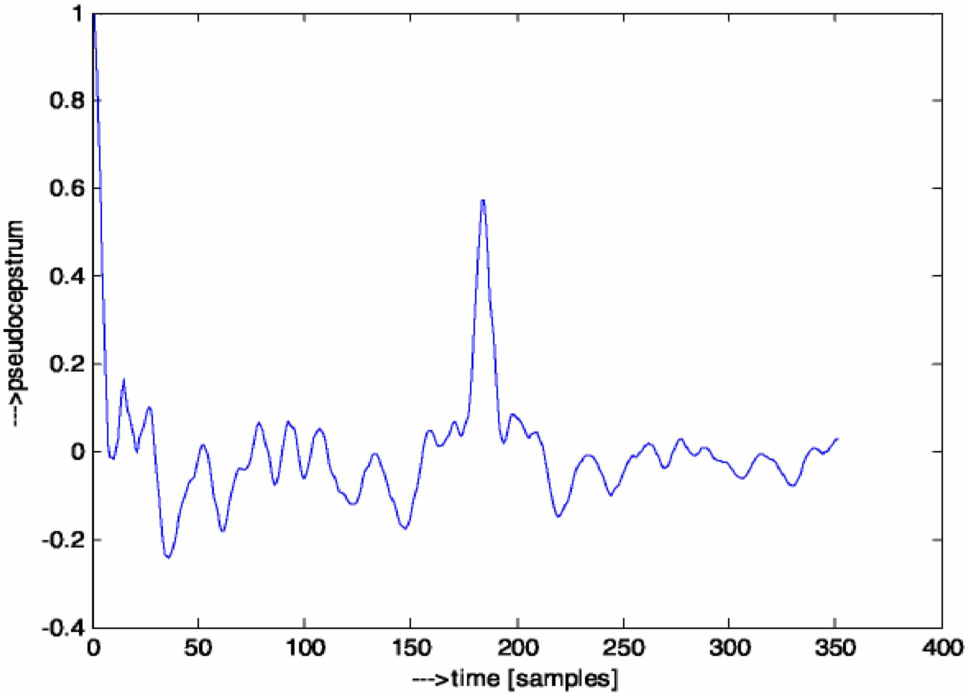


Figure 4.10 Real pseudo-cepstrum for γ = +1 of male vowel “a”

4.2.3.3 General Fundamentals of Expectation Maximization Algorithm

The Expectation Maximization (EM) algorithm is a general approach that solves a problem of the maximum likelihood (ML) estimation with incomplete data. It was discovered and employed independently by different researchers until Dempster brought their ideas together, proved convergence, and proposed it with the term “EM algorithm” in 1977 [14].

Assume the probability density function of an observable random variable Y is given by $f_{\mathbf{Y}}(y;\theta)$ and X is referred to as the “complete data” random variable, which can not directly be observed but only by the means of Y. θ is the set of parameters corresponding to the model being approximated. The relationship between these two random variables can be represented by a noninvertible and unknown mapping function T,

$$\mathbf{T}(\mathbf{X}) = \mathbf{Y} \tag{4.7}$$

Thus, the probability density function of the complete data can be expressed as

$$f_{\mathbf{X}}(x;\theta) = f_{\mathbf{X|Y=y}}(x;\theta) \cdot f_{\mathbf{Y}}(y;\theta) \tag{4.8}$$

with the complete support of T. By rearranging and taking the logarithm for both sides we have

$$\log(f_{\mathbf{Y}}(y;\theta) = \log(f_{\mathbf{X}}(x;\theta)) - \log(f_{\mathbf{X|Y=y}}(x;\theta))\,\mathbf{.} \tag{4.9}$$

If we apply the conditional expectation with respect to X given Y = y for a parameter θ on both sides, then the right-hand side of equation 4.9 is reformulated as

$$\log(f_{\mathbf{Y}}(y;\theta) = E[\log(f_{\mathbf{X}}(x;\theta)) \mid \mathbf{Y} = y;\theta'] - E[\log(f_{\mathbf{X|Y=y}}(x;\theta)) \mid \mathbf{Y} = y;\theta'] \tag{4.10}$$

For convenience, all terms are defined by

$$L(\theta) = \log(f_{\mathbf{Y}}(y;\theta)) \tag{4.11}$$

$$V(\theta,\theta') = E[\log(f_{\mathbf{X}}(x;\theta)) \mid \mathbf{Y} = y;\theta'] \tag{4.12}$$

$$W(\theta,\theta') = E[\log(f_{\mathbf{X|Y=y}}(x;\theta)) \mid \mathbf{Y} = y;\theta'] \tag{4.13}$$

The equation 4.10 can neatly be rearranged as

$$L(\theta) = V(\theta,\theta') - W(\theta,\theta') \tag{4.14}$$

This is a special expression of the log likelihood that we want to maximize. By applying Jensen's inequality to the $W(\theta,\theta')$ term [55, 56], this yields

$$W(\theta,\theta') \le W(\theta',\theta') \tag{4.15}$$

Inferring that if $V(\theta,\theta') > V(\theta',\theta')$, then

$$L(\theta) > L(\theta') \tag{4.16}$$

An iterative EM algorithm generally starts from an arbitrary initial point $\theta^{(0)}$ at time t = 0. Subsequently, the estimates of algorithm parameters are computed as follows.

- Expectation-step: Compute

$$V(\theta, \hat{\theta}^{(t)}) \tag{4.17}$$

- Maximization-step:

$$\hat{\theta}^{(t+1)} = \max_{\theta}(V(\theta, \hat{\theta}^{(t)})) \tag{4.18}$$

- The iterative algorithm is repeated, until

$$\left\| \hat{\theta}^{(t)} - \hat{\theta}^{(t+1)} \right\| \le \varepsilon \tag{4.19}$$

4.2.3.4 Fundamentals of Gaussian Mixture Models (GMM)

Gaussian mixture models are intended to express a more general multimodal probability density function (i.e. multiple peaks) as a superposition of individual Gaussian pdfs called “a mixture.” In most cases, the GMM formulation can be used as a good approximation to a real distribution, even if the mixture components are not really Gaussian.

However, Gaussians are generally used in standard applications of the EM algorithm. The mathematical formulation for a Gaussian mixture distribution [74] is simply

$$f_{\mathbf{X}}(x;\theta) = \sum_{m=1}^{M} c_m \cdot N(x;\theta_m) \tag{4.20}$$

where the c_m is a mixture weight for each of the M mixture components, which must satisfy the following constraints: that $\sum_{m=1}^{M} c_m = 1; c_m \geq 0,$ and $\theta_m = [\sigma_m^2,\ \mu_m]$ represents the variance and mean of component *m*, as found in a Gaussian distribution function

$$N(x;\sigma^2,\mu) = \frac{1}{\sqrt{2\pi}\,\sigma_m} \cdot \exp^{-\frac{1}{2}\left(\frac{x-\mu_m}{\sigma_m}\right)^2} \tag{4.21}$$

Thus, the complete Gaussian mixture is parameterized by means, variances and mixture weights of all component densities and these model parameters are collectively represented by the following notation:

$$\lambda = \{c_m, \mu_m, \sigma_m^2\} \tag{4.22}$$

4.2.3.5 EM Algorithm for GMM Parameter Estimation

The general EM algorithm can be adapted to work on a GMM problem for finding the maximum likelihood of a mixture model. The “missing data” are the means, variances, and mixture weights (prior probabilities) of the mixture components. These parameters have to be estimated from the limited number of samples [57].

- The expectation step for a GMM yields

$$\hat{p}^{(t)}(m \mid x_i, \hat{\lambda}^{(t)}) = \frac{\hat{c}_m^{(t)} \cdot N(x_i; \hat{\sigma}_m^{2(t)}, \hat{\mu}_m^{(t)})}{\sum_{m=1}^{M} \hat{c}_m^{(t)} \cdot N(x_i; \hat{\sigma}_m^{2(t)}, \hat{\mu}_m^{(t)})} \qquad (4.23)$$

which is the probability (actually *a posterior* probability at iteration step *t*) that a sample x_i belongs to the Gaussian component *m*.

- The maximization step re-estimates

$$\hat{c}_m^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \hat{p}^{(t)}(m \mid x_i, \hat{\lambda}^{(t)}) \qquad (4.24)$$

$$\hat{\mu}_m^{(t+1)} = \frac{1}{\sum_{i=1}^{N} \hat{p}^{(t)}(m \mid x_i, \hat{\lambda}^{(t)})} \sum_{i=1}^{N} x_i \cdot \hat{p}^{(t)}(m \mid x_i, \hat{\lambda}^{(t)}) \qquad (4.25)$$

$$\hat{\sigma}_m^{2(t+1)} = \frac{1}{\sum_{i=1}^{N} \hat{p}^{(t)}(m \mid x_i, \hat{\lambda}^{(t)})} \sum_{i=1}^{N} (x_i - \hat{\mu}_m^{(t)})^2 \cdot \hat{p}^{(t)}(m \mid x_i, \hat{\lambda}^{(t)}) \qquad (4.26)$$

- A termination condition as shown in equation 4.19 is used in this implementation.

4.2.3.6 Reinterpretation of Magnitude Spectrum as Spectral Density

The EM algorithm has been formulated to work on model approximation of the magnitude spectrum based on the GMM formulation as presented in section 4.2.3.5. However, we assumed that samples x_i were available from the unknown distribution when the EM algorithm was developed. In our application of modeling magnitude spectrum, we are not supplied with these observable samples. The definition of the Discrete Fourier Transform does not provide any information for some samples to form a frequency density. This means that a single sample does not convey an inherent frequency since the DFT is deterministic in nature. The formulated EM presented in section 4.2.3.5 can be employed to work only with the probability density (probability mass function) of samples. In this application, the input density for EM algorithm is the magnitude spectrum of a single frame of speech signal. This is a deterministic function modeling problem and it is possibly solved by a probabilistic artifice.

4.2.3.7 Modification of EM Algorithm for Frequency Density

The EM algorithm presented in this section is the modified algorithm that will accept the “frequency density” of magnitude spectrum as input density. Let this input to EM algorithm denoted by $\phi(\omega_i)$ whose supported points (ω_i) are frequencies at which the DFT of individual speech frames is computed. In order to have the magnitude spectrum as a valid probability density function, the area under a curve of magnitude spectrum is required to equal to one, $\sum_{i=1}^{\Omega}\phi(\omega_i) = 1$.

- Expectation-step:

$$\hat{p}^{(t)}(m \mid \underline{\omega}_k, \hat{\lambda}^{(t)}) = \frac{\hat{c}_m^{(t)} \cdot N(\underline{\omega}_k; \hat{\sigma}_m^{2(t)}, \hat{\mu}_m^{(t)})}{\sum_{m=1}^{M} \hat{c}_m^{(t)} \cdot N(\underline{\omega}_k; \hat{\sigma}_m^{2(t)}, \hat{\mu}_m^{(t)})} \tag{4.27}$$

This is a distribution curve of a single Gaussian component that is normalized by a sum of all mixture component densities. This expression provides a likelihood measure of each abscissa value of the density graph, whereas the probabilities in a former equation 4.23 were computed for each such value and even for its multiple occurrences. The number of occurrences is represented as multiplication with $\phi(\underline{\omega}_k)$ in maximization step, where real positive value can now be assumed.

- Maximization-step:

$$\hat{c}_m^{(t+1)} = \sum_{k=1}^{\Omega} \phi(\underline{\omega}_k) \cdot \hat{p}^{(t)}(m \mid \underline{\omega}_k, \hat{\lambda}^{(t)}) \tag{4.28}$$

$$\hat{\mu}_m^{(t+1)} = \frac{1}{\hat{c}_m^{(t+1)}} \sum_{k=1}^{\Omega} \underline{\omega}_k \cdot \phi(\underline{\omega}_k) \cdot \hat{p}^{(t)}(m \mid \underline{\omega}_k, \hat{\lambda}^{(t)}) \tag{4.29}$$

$$\hat{\sigma}_m^{2(t+1)} = \frac{1}{\hat{c}_m^{(t+1)}} \sum_{k=1}^{\Omega} (\underline{\omega}_k - \hat{\mu}_m^{(t)})^2 \cdot \phi(\underline{\omega}_k) \cdot \hat{p}^{(t)}(m \mid \underline{\omega}_k, \hat{\lambda}^{(t)}) \tag{4.30}$$

Equations 4.27, 4.28, 4.29 and 4.30 give us a better understanding of how the EM algorithm can operate through the fitting application; multiplication of each mixture component with observed sample data enables iterative ML estimation to move toward its next estimate and eventually converge. However, this algorithm heavily depends on the initial values of model parameters due to its nature as a steepest-descent algorithm.

Summary of procedure to estimate GMM parameters of each frame of speech spectrum by the modified EM algorithm is depicted in Figure 4.11, as presented with low complexity in GMM formulations. In model fitting of GMM, each 51.2ms frame of speech is represented by the frequency density and associated model parameters λ. This kind of prior input information generally supplies the sufficient statistics required by the EM algorithm for working on the estimation of ML parameters. The algorithm first starts with a computation of a mixture of Gaussians (equations 4.20 and 4.21) based on available information at the initial state. Then, the expectation step (E–step) computes the expected values of data likelihood using the current estimates of model parameters (means, variances, and mixture weights) and the observed data (frequency density). Likelihoods and posterior probabilities for individual mixture component densities are also computed in the E-step. Finally, the maximization step (M–step) uses data set from accumulating sufficient statistics in the E–step to re–estimate means, variances and mixture weights of individual mixture components in order to maximize likelihood of model parameters (ML estimates). The total area under a histogram of the smoothed spectrum is calculated; means, variances and mixture weights are initialized at this point. Mean parameters are initialized by using the Peak Detection Algorithm (PDA) to detect all frequency locations of spectral peaks. This algorithm (Figure 4.1) finds all local maxima of a curve of magnitude spectrum. If this peak-picking algorithm works well, a problem involving how many Gaussians should be assigned to the EM algorithm will be solved as well. It means that PDA will return a set of frequency indices of peak locations equaling to a number of Gaussians in a mixture at a beginning of ML iteration. The frequency indices serve as initials of mean in approximation.

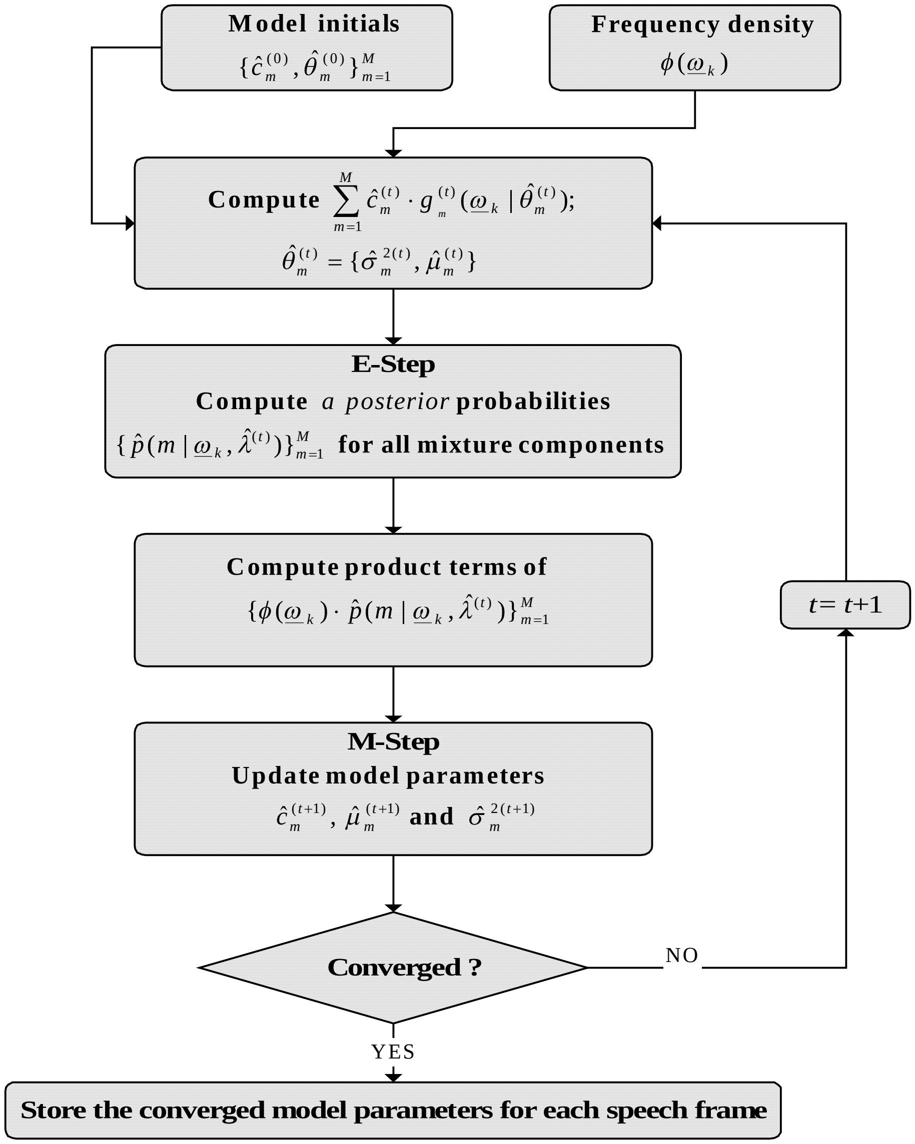

Model initials
$\{\hat{c}_m^{(0)}, \hat{\theta}_m^{(0)}\}_{m=1}^{M}$
Frequency density
$\phi(\underline{\omega}_k)$
Compute $\sum_{m=1}^{M} \hat{c}_m^{(t)} \cdot g_m^{(t)}(\underline{\omega}_k \mid \hat{\theta}_m^{(t)})$;
$\hat{\theta}_m^{(t)} = \{\hat{\sigma}_m^{2(t)}, \hat{\mu}_m^{(t)}\}$
E-Step
Compute *a posterior* probabilities
$\{\hat{p}(m \mid \underline{\omega}_k, \hat{\lambda}^{(t)})\}_{m=1}^{M}$ for all mixture components
Compute product terms of
$\{\phi(\underline{\omega}_k) \cdot \hat{p}(m \mid \underline{\omega}_k, \hat{\lambda}^{(t)})\}_{m=1}^{M}$
M-Step
Update model parameters
$\hat{c}_m^{(t+1)}$, $\hat{\mu}_m^{(t+1)}$ and $\hat{\sigma}_m^{2(t+1)}$
Converged ?
NO
$t = t+1$
YES
Store the converged model parameters for each speech frame

Figure 4.11 Illustration of the modified EM algorithm.

In addition, the parameters of mixture weights for individual Gaussian components are initially set equal to rations of individual peak's magnitude to total sum of all peaks' magnitude. This makes individual mixture weights probabilistic and sum of all weights equal to one. Variances are made significant with respective to number of Gaussians in a mixture and they strongly depend on frequency intervals that are occupied by individual Gaussians.

4.2.3.8 Mixture Component Selection Algorithm

This section presents a way to select the most dominant Gaussian components from the M-component Gaussian mixture, whose model parameters have been estimated. Total number of M-components in a mixture is equal to number of spectral peaks detected by algorithm depicted in Figure 4.1. In order to decide which individual Gaussians should be taken as one of four dominant Gaussians representing the most of vocal tract characteristics, a criterion is needed. Threshold of testing the significance of individual Gaussians is defined by the difference or distance calculated among model parameters of each Gaussian. It is called "Selective Ratio" for this work. Four Gaussians with high selective ratios (ration of mixture weight to standard deviation) are chosen automatically by algorithm. Based on our experimental determination for the threshold to select the dominant Gaussians from a mixture, such a ratio gives us the most closely equivalent to formant amplitude and; fortunately, it can obviously separate the particular Gaussians distributing with higher amplitudes (magnitude at center frequencies) and narrower bandwidths (frequency intervals occupied by Gaussians over the magnitude spectrum), preferably taken as the most significant Gaussians, from those with lower amplitudes and wider bandwidths. This threshold of selecting the dominant Gaussians is mathematically written as

$$\text{Selective Ratio} = \frac{\hat{c}_q^{(t+1)}}{\hat{\sigma}_q^{(t+1)}} \tag{4.31}$$

where the $\hat{c}_q^{(t+1)}$ and $\hat{\sigma}_q^{(t+1)}$ are the estimated mixture weights and standard deviations that belong to the mixture component *q*, where *q*<*M* and *M* is equal to total number of detected peaks found by PDA. In addition, $\hat{c}_q^{(t+1)}$ and $\hat{\sigma}_q^{(t+1)}$ are defined mathematically by $(\hat{c}_q^{(t+1)} \mid \hat{c}_q^{(t+1)} \in \{\hat{c}_m^{(t+1)}\}_{m=1}^{M}, q \subset m)$ and $(\widehat{\sigma}_q^{(t+1)} \mid \widehat{\sigma}_q^{(t+1)} \in \{\widehat{\sigma}_m^{(t+1)}\}_{m=1}^{M}, q \subset m)$.

4.2.3.9 Conversion of Gaussian Parameters

The conversion of means, variances and mixture weights to energy concentration parameters representing the most important spectral characteristics of the vocal tract is described in this section. The center frequencies are equal to mean parameters of the selected components and amplitudes are equal equivalently to the heights of component densities at their mean (center frequency). The 3-dB bandwidth is calculated from the corresponding Gaussians in a mixture, which represents the distance between two points in frequency domain where the signal is $\frac{1}{\sqrt{2}}$ of the maximum Gaussian amplitude (half power) [49]. By using the 3-dB log ratio, bandwidth (BW) of Gaussian distribution can be computed as

$$\text{BW} = 2\hat{\sigma}_q \sqrt{\ln(2)} \tag{4.32}$$

Based on four selected Gaussian components taken as the most distinctive spectral representation of the vocal tract response, a set of twelve parameters comprising of three groups of four center frequencies, four bandwidths and four mixture weights was used to represent an individual 51.2ms frame of speech. The feature samples extracted from all speech frames were averaged to represent each patient.

4.2.4 Dimensionality Reduction of Feature Space

There are a number of methods from the pattern recognition literatures for reduction of dimensionality of feature space. Several of these have been used in speaker identification and speech recognition with good results. These methods can be grouped into two categories: feature selection method and feature extraction method. The first method reduces the dimensionality of feature space by selecting a subset of the original feature set. The second method (also known as the "transformation" method) reduces the dimensionality by projecting the original D–dimensional feature space on a d–dimensional subspace (d<D) through a transformation. In this work, we employed the feature selection method to reduce the original D–dimensional feature space, which equals to 16 (12 GMM-based parameters combining with 4 energy ratios).

In the feature selection, a feature with its ability to distinguish between two classes depends on both the distance between the two classes and the amount of scatter within the classes. A reasonable measure of class discrimination must be taken into account both the mean and variance of the classes. One such measure of separability between two classes is the Fisher’s discriminant ratio [58]. In an equation 4.33, the higher discrimination is measured when the class means are further apart and when the spread of classes is smaller, thereby increasing separation between two classes. This measure is defined as:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{4. 33}$$

where μ_1 and μ_2 are the two means or centroids of the classes and σ_1 and σ_2 are the standard deviations of the classes.

The *Fisher*'s ratio can measure the separation which exists between two classes. The extension of *Fisher*'s discriminant ratio which provides more ability to measure separation between multiple classes is the F–ratio. In this paper, we use this type of statistical measure for determining the subset of primary features selected from the original features, which provides the highest discrimination power between classes. The F–ratio is a measure that can be used to evaluate the effectiveness of particular features in group classification. It has been widely used as a figure of merit for feature selection in speaker recognition application [59], [60]. It is defined as the ratio of the between–class variance and the within–class variance. This method tries to select the feature that maximizes the separation between different classes and minimizes the scatter within these classes. The following assumption must be satisfied when using the F–ratio as a figure of merit for reducing a dimensionality: 1) The feature vector within each class must have the Gaussian distribution; 2) the features should be statistically uncorrelated; and 3) the variances within each class must be equal. Since the variances within each class are generally not equal, the pooled within–class variance is used to define the F–ratio. The number of training feature vectors, training pattern, in the *j*th class of the K classes is assumed to be the same (N_j). Thus, the F–ratio of the ith feature can be defined as:

$$F_i = \frac{B_i}{W_i} \tag{4.34}$$

in which B_i is the between–class variance and W_i is the pooled within–class variance of the ith feature, which can be mathematically defined by

$$B_i = \frac{1}{K}\sum_{j=1}^{K}(\mu_{ij} - \mu_i)^2 \tag{4.35}$$

$$W_i = \frac{1}{K}\sum_{j=1}^{K} W_{ij} \tag{4.36}$$

where μ_{ij} and W_{ij} are the mean and variance of the ith feature, respectively, for the jth class, and μ_i is the overall mean of the ith feature. These are given by

$$\mu_{ij} = \frac{1}{N_j}\sum_{n=1}^{N_j} x_{ijn} \tag{4.37}$$

$$W_{ij} = \frac{1}{N_j}\sum_{n=1}^{N_j} (x_{ijn} - \mu_{ij})^2 \tag{4.38}$$

$$\mu_i = \frac{1}{N_j} \sum_{j=1}^{K} \mu_{ij} \qquad (4.39)$$

where x_{ijn} is the *i*th feature of the *n*th training feature vector, from the *j*th class.

To evaluate whether the between-class variance is larger relative to the within-class variance, it is necessary to take into account the number of independent scores, or degrees of freedom (d.f.) that contribute to each of those variances. For the between-class variance, *d.f.* = *K-1* where *K* is the number of comparison classes. For the within-class variance, *d.f. = N-K* where *N* is the total number of samples in all classes. In addition, for testing the null hypothesis that sample means of features of several classes are not different significantly from class to class. The F-ratio computation is concerned with comparing the variances among the means to the variances within the samples. What it takes to be "large enough" for the difference to be statistically significant depends on the sample size and the amount of certainty that we desire in our testing (that is *p* values or levels of statistical significance). The decision of whether or not to reject the null hypothesis that the sample means are similar to each other requires the value for the F-ratio to be compared with a critical value. The critical value of F needed to reject the null hypothesis at any given level of significance (e.g. .05, .01, or .001) varies with two rather than only one indicator of degrees of freedom. It depends on both the between-class and the within-class degrees of freedom.

4.2.5 Performance Evaluation

We first performed multiple runs of patient classification on a training data set with a subset of features selected from the original sixteen features. For each run, a new subset of top ranked d (< 16) features with the F–ratios ranked in order from high to low was selected to evaluate its recognition rate. The table of size d of reduced feature sets versus recognition accuracies was determined as comparison of all recognition rates for seeking a primary feature set providing the highest recognition rate.

The primary feature set was used in the L–fold cross validation [65] as predictor variables for classifier performance. The quadratic classifier was used in performing twelve repetitions of cross validation with different randomized training and testing data sets. The average recognition accuracy obtained from analysis of cross validation was taken into account instead of accuracy of individual runs of cross validation due to having smaller variance in performance estimates. This technique of validation provides us more statistically reliable analysis for classification on an empirical measure for success of discrimination.

In individual runs of cross validation, samples were randomly split into two subsets: 75% of original data comprised a training set; and 25% comprised a testing set. In order to randomly separate an original data set into training sets and testing sets; first, we uniformly generated random numbers between 0 and 1 and then assigned weighting factor of “1” to random numbers that are less than 0.75 and weighting factor of “0” to random numbers greater than 0.25. Samples with weight of “1” were taken as members of a training set and those with “0” were taken for a testing set.

In order to evaluate classification performance, several performance measures (i.e., Sensitivity (SE), Specificity (SP), Positive Predictive Value (PPV), and Negative Predictive Value (NPV)) were experimentally determined. Not only separabilities in features were measured, but statistical analyses including comparison of class covariance matrices, comparison of class features using analysis of variance (ANOVA), and calculation of 95%

confidence interval were all performed for investigating the significant discriminative power in features as well.

4.3 Experimental Results

4.3.1 GMM Fitting Results

We used several window lengths in liftering to obtain the low-time portion of cepstrum and found that the window length of low-time filter setting within the 85%-95% range of average pitch period produced the best result in smoothing the vocal tract spectrum. The ripple effect caused by the fundamental frequency was considered to be fairly eliminated from the vocal tract spectrum based on our experiment.

The result of initialization of GMM parameter by performing the peak detection algorithm (PDA) on one frame of the smoothed spectrum is shown in Figure 4.12. Nine peaks and eight saddles were effectively identified for their locations on spectrum, but only eight peaks and seven saddles can visually be noticed from the plot. However, the frequency indices of peak locations of all detected peaks were automatically determined by PDA and then supplied to EM algorithm. The estimated magnitude spectrum of 51.2ms frame of female speech showed that the PDA technique worked well as presented in Figure 4.12. All peaks and saddles appearing on the magnitude spectrum were precisely located for their frequency locations.

Figure 4.13 presents a mixture of nine Gaussians superimposed on the magnitude spectrum. Individual Gaussians properly distributed over small frequency intervals within a 0-5,000 Hz frequency range. In addition, a relationship among mean, variance and mixture weight of individual Gaussians was explored. Figure 4.14 shows another representation of the estimated Gaussian mixture of the same frame of spectrum, which combined three relating model parameters of individual Gaussians together and presented them as the circles with a dot at their centers (Figure 4.13). The radius of circle represents the standard deviation of individual Gaussians. The center of individual circles (x-y coordinate) refers to as a coordinate specifying mean and mixture weight of individual Gaussians. As expected for the lowest peak locating at 4,555 Hz (Figures 4.12 and 4.13), this tiny spectral peak was now represented by a circle with a mixture weight that was very close to zero, as compared with that of other circles (other higher magnitude peaks). By comparing Figures 4.12, 4.13 and 4.14, the relationship between the peak magnitude and the value of mixture weight can be observed for individual Gaussian densities. The appearance of a tiny peak at 4,555 Hz confirmed the existence of positive correlation between magnitude of spectrum and mixture weight of Gaussian distribution.

Figure 4.15 shows a plot of normalized selective ratios revealing significant Gaussian components. The vertical scale now is unit value of normalized ratios of mixture weights to standard deviations. As a result of selective ratios for individual Gaussians, the circles now have moved to new coordinates corresponding to calculated ratios. Circles seem to be more separable into two groups of circles locating above 0.4 and below 0.3. Significant difference can obviously be observed between plots in Figure 4.14 and Figure 4.15. In our experiment, four Gaussian components with high selective ratios were taken in account and then their model parameters were converted to vocal tract spectral parameters. In Figure 4.16, the circles highlighted and marked with the “x” letters at their centers represent individual dominant Gaussians that were chosen automatically by algorithm based on equation 4.31.

As compared to Figure 4.14, the first, third, sixth, and eighth Gaussians densities were now selected from the mixture as the most significant components, as referred to by Figure 4.17. Figure 4.18 shows a comparison of the original vocal tract spectrum and a new mixture of four dominant Gaussians. A goodness of model fitting is obviously visualized. Most of the dominant peaks appearing on spectrum were captured successfully by this technique.

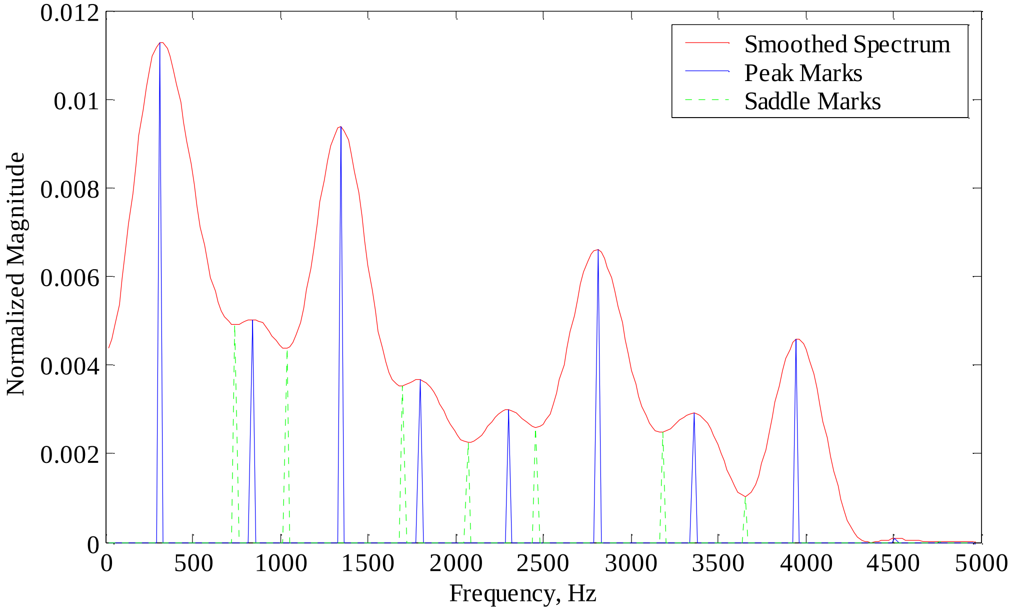


Figure 4.12 Illustration of the frequency locations of spectral peaks detected by Peak Detection Algorithm.

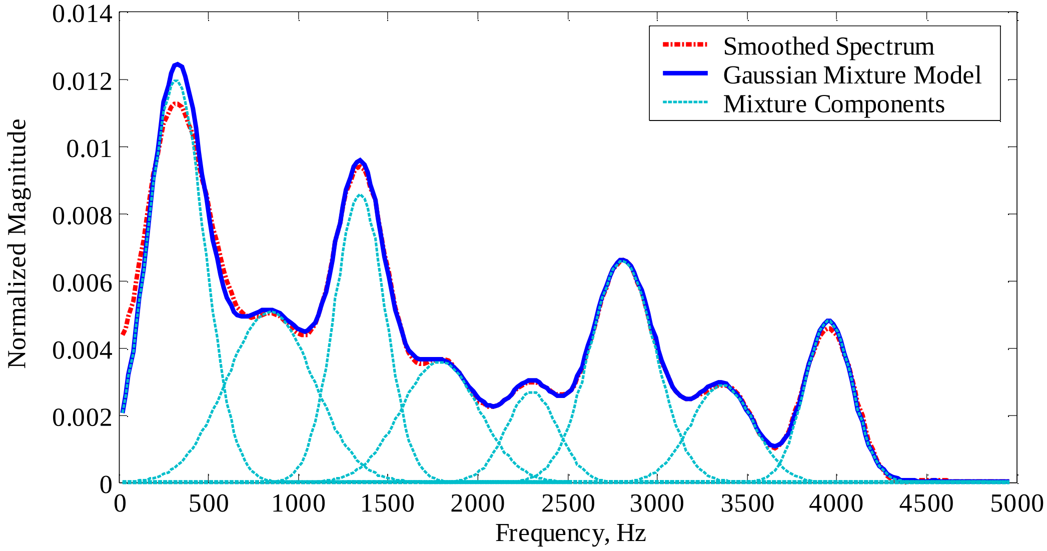


Figure 4.13 Plot of a mixture of the Gaussians superimposed on the magnitude spectrum.

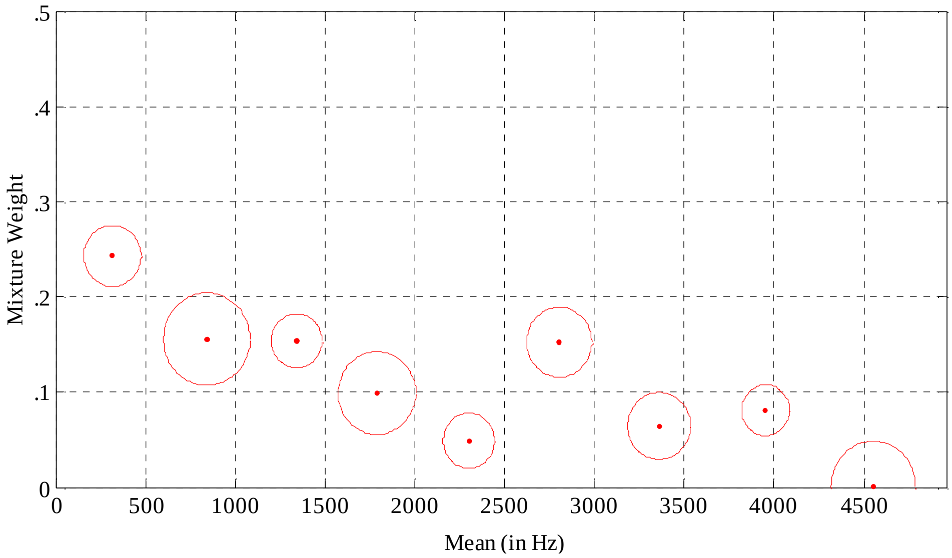


Figure 4.14 Illustration of a mixture of individual Gaussian components. For individual Gaussian components, a coordinate of circle represents the mean and mixture weight and a radius of circle represents the standard deviation.

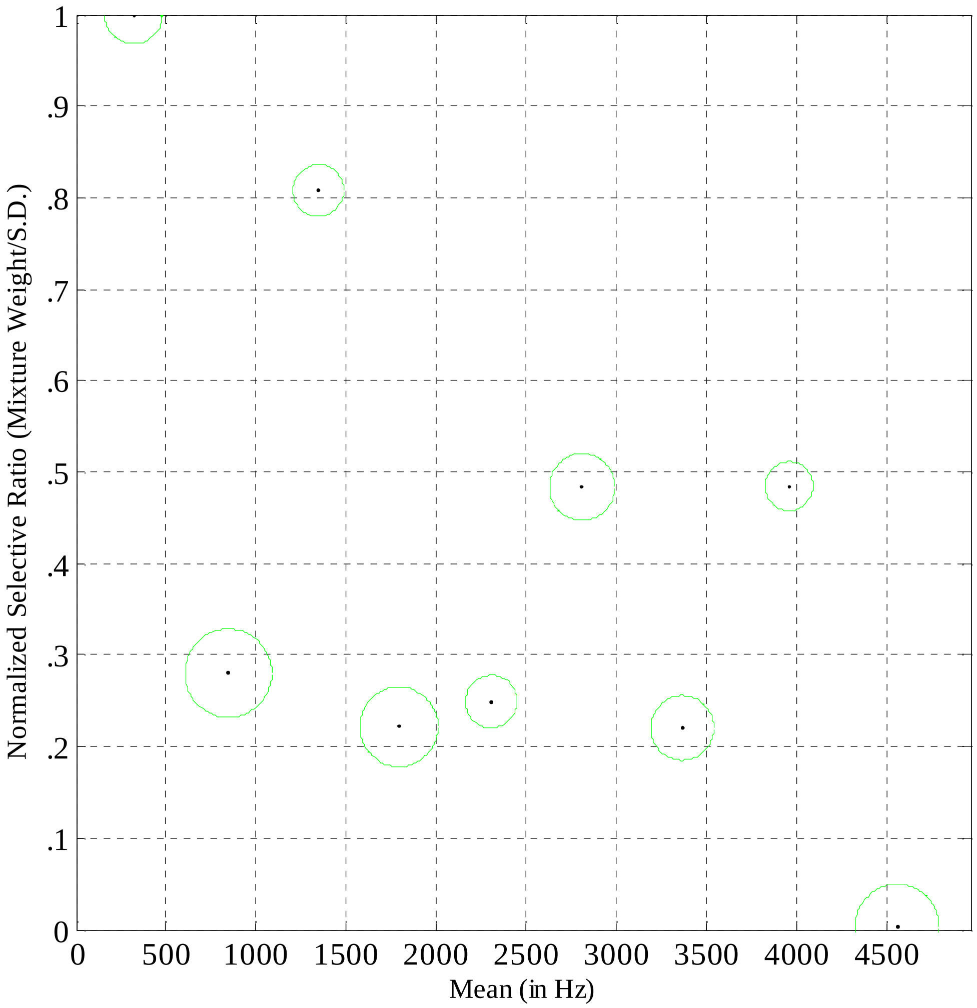


Figure 4.15 Illustration of individual Gaussian components vertically responding to their normalized selective ratios.

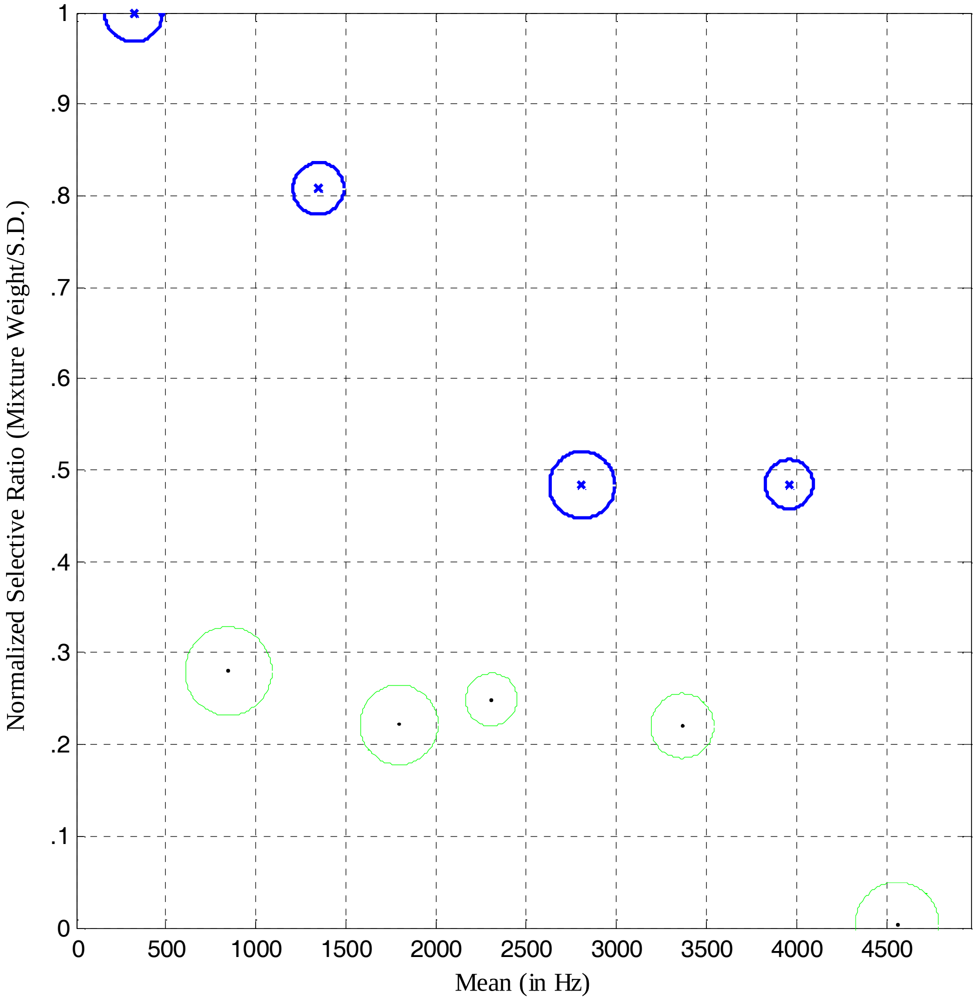


Figure 4.16 Illustration of individual Gaussian components automatically chosen by an algorithm of the mixture component selection based on the competitively high values of selective ratios.

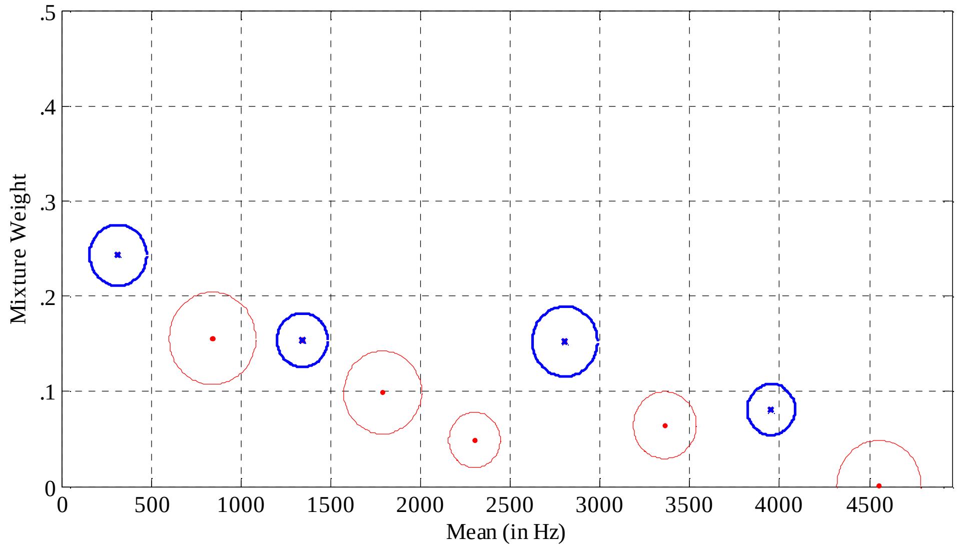


Figure 4.17 Illustration of a mixture whose four components were selected as the distinctive acoustic parameters, as referred to by the highlighted circles marked with a letter “x” right at their center.

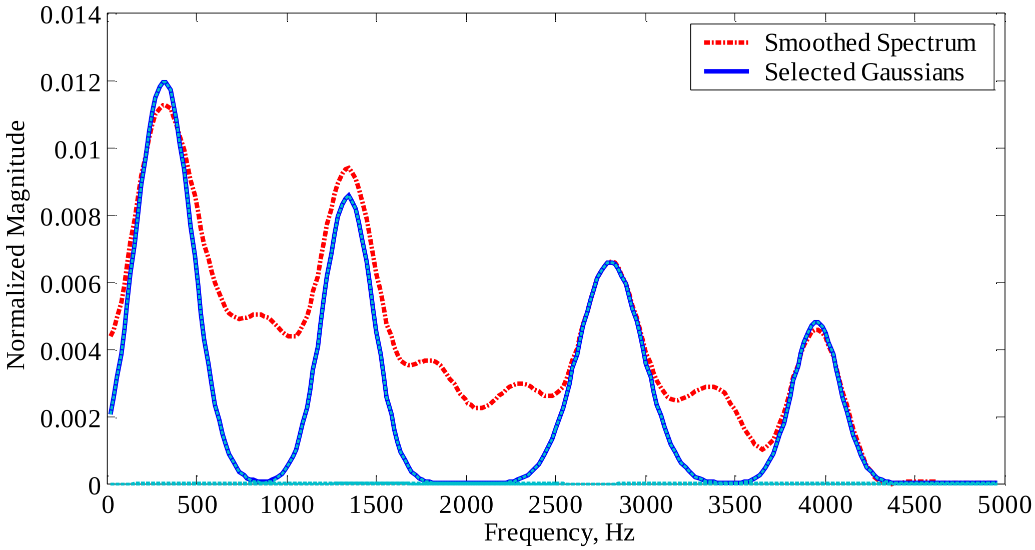


Figure 4.18 Plot of a mixture of the selected Gaussians.

4.3.2 Comparative Results of Interview Vocal Features of Depressed and Suicidal Speech

The means and standard deviations of the vocal features calculated from diagnostic groups of suicidal, depressed and remitted speech samples are summarized in Table 4.1.

Table 4.1 Summary of the feature means and standard deviations of the categorized groups of interview speech.

Feature	**Suicidal**	**Depressed**	**Remitted**
PSD_1	(0.88, 0.05)	(0.84, 0.10)	(0.78, 0.09)
PSD_2	(0.10, 0.04)	(0.14, 0.09)	(0.19, 0.07)
PSD_3	(0.01, 0.01)	(0.01, 0.01)	(0.02, 0.02)
PSD_4	(0.01, 0.01)	(0.01, 0.01)	(0.01, 0.01)
CF_1	(262.04, 45.25)	(247.57, 50.26)	(257.89, 53.39)
CF_2	(1241.55, 176.53)	(1151.59, 150.05)	(1181.34, 140.19)
CF_3	(2246.02, 250.32)	(2152.75, 186.62)	(2195.43, 159.51)
CF_4	(3338.58, 260.19)	(3273.12, 201.96)	(3371.97, 195.57)
BW_1	(261.00, 40.92)	(248.61, 48.06)	(253.37, 46.97)
BW_2	(535.87, 61.23)	(511.46, 27.56)	(492.89, 39.32)
BW_3	(363.14, 50.37)	(365.02, 41.39)	(373.34, 39.27)
BW_4	(333.32, 32.24)	(335.96, 32.43)	(320.84, 24.89)
WC_1	(0.24, 0.04)	(0.22, 0.04)	(0.21, 0.04)
WC_2	(0.25, 0.03)	(0.25, 0.02)	(0.24, 0.03)
WC_3	(0.11, 0.02)	(0.12, 0.02)	(0.13, 0.02)
WC_4	(0.09, 0.01)	(0.09, 0.01)	(0.09, 0.01)

Several vocal parameters characterized by the statistical properties show the trends relating to the severity of the mental state affected by depression and suicidal risk. Particularly, the first energy ratio (PSD_1) in 0-500 Hz frequency sub-band, bandwidth of the second Gaussian component (BW_2), and mixture weight coefficient of the first Gaussian component (WC_1) were all determined to increase as the severity of the mental state increased. Conversely, PSD_2, BW_3 and WC_3 exhibited a negative trend with the degree of the mental state.

The mean values measured for PSD_1, CF_1, CF_2, CF_3, BW_1, BW_2 and WC_1 were found to be the highest for suicidal speech compared with those of other speech groups. Conversely, PSD_2, BW_3 and WC_3 obtained from suicidal speech were characterized by the lowest mean values. In depressed speech, the mean values of CF_1, CF_2, CF_3, CF_4 and BW_1 appeared to be the lowest as compared with those of suicidal speech and remitted speech. Nevertheless, the described trends or patterns of the vocal features do not necessarily infer the statistically significant differences between classes which may improve the performance of classification.

The rank-ordered features obtained by calculating the F-ratio statistics for feature samples of interview speech are presented in Tables 4.2-4.4. In order to compare the discriminative properties of individual vocal features, the normalized F-ratios measured for the depressed-suicidal, remitted-depressed and remitted-suicidal speech analyses are plotted in Figures 4.19, 4.22, and 4.23.

Table 4.2 Ranked interview vocal features of female depressed-suicidal comparison.

Rank	Feature	Rank	Feature
1	PSD_2	9	WC_4
2	CF_2	10	PSD_4
3	PSD_1	11	CF_4
4	BW_2	12	BW_1
5	WC_1	13	PSD_3
6	CF_3	14	WC_2
7	WC_3	15	BW_4
8	CF_1	16	BW_3

By observing the results of the F-ratio pairwise analyses, we found that two features of the set of PSD_2, PSD_3 or PSD_1 always showed up as high F-ratio features approaching to one. It may imply that the PSD-based features revealed more discriminative abilities as compared to the GMM-based features.

In Figure 4.19, the plot of normalized F-ratios measured from feature samples of depressed and suicidal speech illustrates that PSD_2 and PSD_1 were not the only two features characterized by very high F-ratios, but CF_2 and BW_2 also exhibited very high separabilities.

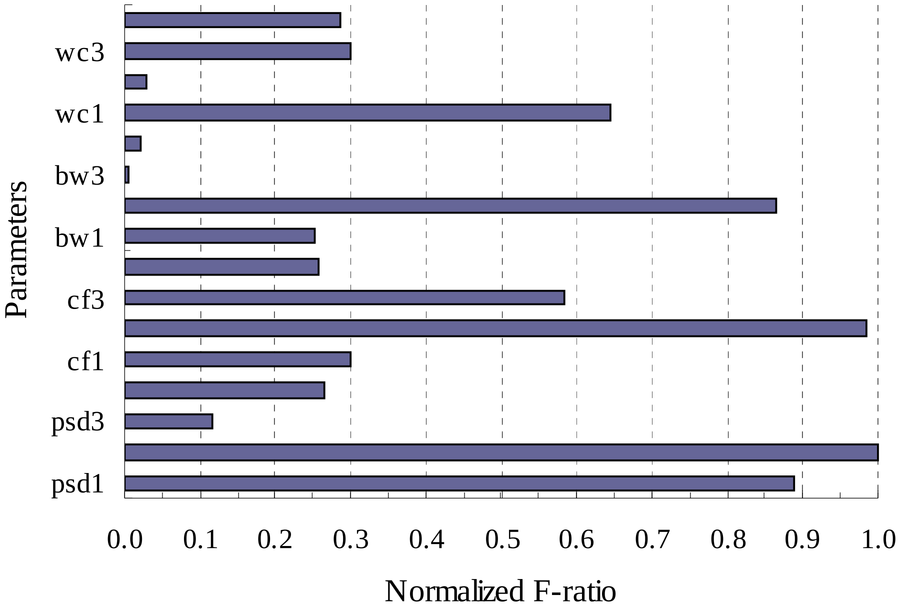


Figure 4.19 Normalized F-ratios measured for interview vocal features of female depressed-suicidal comparison.

Generally, the recognition accuracy can be used to predict the effectiveness of vocal features as discriminators in classification analyses. Multiple runs of speech recognition with the different feature sets were performed for accumulating all recognition rates. As a result of this procedure, the feature set providing the most accurate recognition rate was taken as the primary feature set.

According to practice in speech recognition, when two or more reduced feature sets provide the similar recognition accuracy, a smaller feature set is preferably taken as the definitively primary feature set rather than the larger feature set. More statistical reliability in recognition performance needs to be considered, when speech recognition involves with a small sample size of database. Thus, a smaller set of primary features is more preferred.

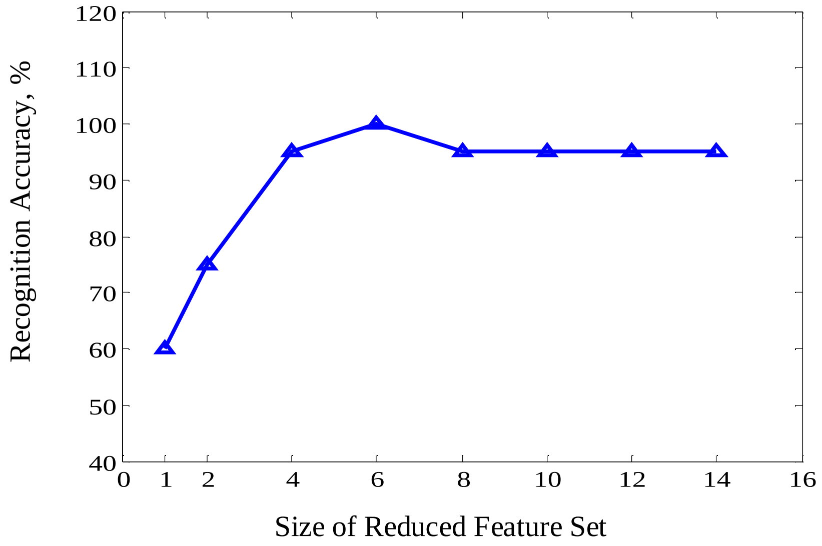


Figure 4.20 Plot of the recognition rates as a function of the size, d, of reduced feature set.

Figure 4.20 presents a plot of recognition rates as a function of size, d, of the reduced feature sets. Sixteen vocal features extracted from depressed and suicidal speech groups were tested for their recognition performance. It can obviously be seen that dimensionality of feature space can be reduced down to four without affecting recognition performance. Thus, the first four rank-ordered features were taken as a primary feature set for designing the depressed-suicidal pairwise classifier. In recognition experiment, we observed that when number of features increases to eight or even larger, the degradation of recognition accuracies occurred. This probably implies that some features causing the deconstructive performance overlapped from each other and distributed across all diagnostic groups of patients, which made the recognition less accurate.

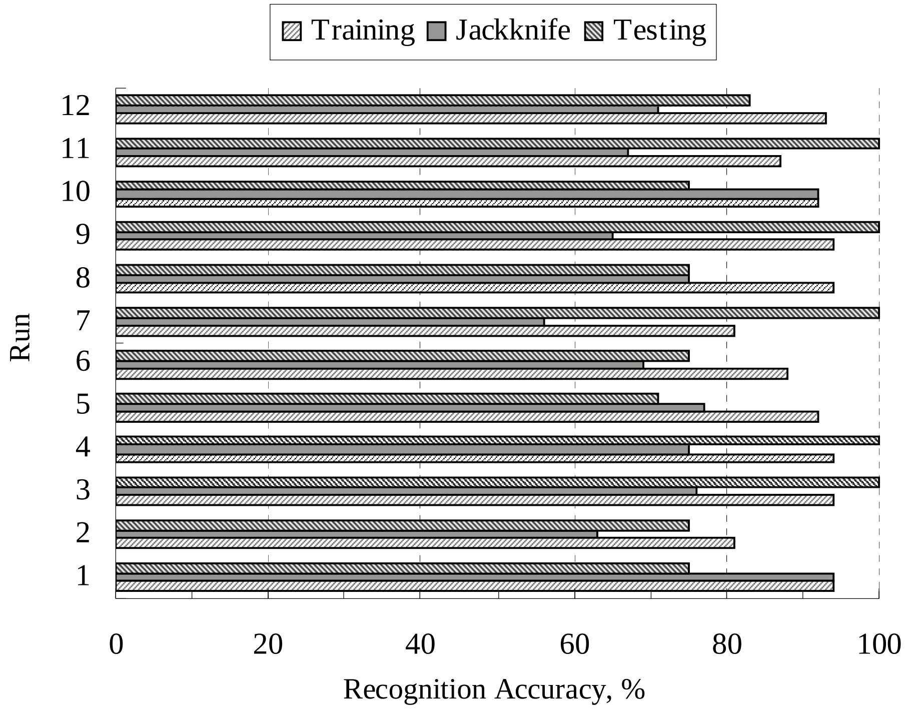


Figure 4.21 Plot of the recognition rates based on the 12-fold cross validation comprising recognition results obtained from training, training with the Jackknife procedure and validating classifier.

As shown in Figure 4.21, the results of individual classification rates were obtained from the quadratic discriminant analysis incorporating the 12-fold cross validation. Each run of cross validation produced three different scores of correct classification. They were determined from three different procedures: 1) Training a quadratic classifier with 75% of the original data; 2) performing the Jackknife classification with 75% of the original data; and 3) validating classification with a testing set (the hold out 25% from the original data set).

As a result of the 12-cross validation (Figure 4.21), we observed that most of the testing scores were found to be around 72%-100%. These accuracy scores are considered to be effectively high. However, some of these testing scores were slightly lower as compared with those training scores for some individual runs of recognition. The degradation of classification performance is probably caused by the inconsistency from selecting numbers of samples to form the 25% testing set comprising of samples taken from the depressed speech group and some taken from the suicidal speech group. In each run of cross validation, the weighting numbers between 0 and 1 generated randomly by a computer were used to weight all feature samples and separate them into groups of training samples and testing samples. Due to this uncontrolled randomizing of weighting factors, numbers of testing samples taken from depressed speech and suicidal speech for the 25% testing set are not consistently equal for runs of cross validation, thus resulting in unstable and inconsistent classification performance. Another source of this degradation possibly relates with the discriminative abilities in features itself as well.

Table 4.5 summarizes the average classification scores and performance measures obtained from the quadratic classification analyses incorporating the 12-fold cross validation. As results of dimensionality reduction based on the F-ratio measure in determining the best discriminating features, the rank-ordered PSD_2, CF_2, PSD_1 and BW_2 were statistically selected for the depressed-suicidal speech comparison. The classifier designed by these vocal features yielded a cumulative classification score of 85.75% and this score was the highest found among pairwise analyses of interview speech. As indicated by sensitivity (0.89) and specificity (0.80), it suggested that the classifier was slightly more effective in classifying depressed patients than in classifying suicidal patients.

4.3.3 Comparative Results of Interview Vocal Features of Remitted and Depressed Speech

In remitted speech, PSD_2, PSD_3, CF_4, BW_3 and WC_3 were all characterized by the increased mean values compared to those of depressed speech and suicidal speech. PSD_3 and WC_3 were found to be slightly different between remitted speech and depressed speech. Conversely, reductions of PSD_1, BW_2, BW_4, WC_1 and WC_2 were determined for remitted speech. For the depressed group, none of vocal features were characterized by the highest average values when compared with other groups. However, the reduced mean values can be observed in depressed speech for CF_1, CF_2, CF_3, CF_4 and BW_1. Another observation is that the mean values of PSD_4 and WC_4 were determined to be very low and similar for all diagnostic groups. It implied the appearance of very low-energy magnitude spectrum at very high frequency range.

Table 4.3 Ranked interview vocal features of female remitted-depressed comparison.

Rank	Feature	Rank	Feature
1	PSD_3	9	WC_2
2	PSD_1	10	CF_3
3	WC_3	11	WC_4
4	BW_2	12	CF_2
5	PSD_4	13	BW_3
6	BW_4	14	CF_1
7	PSD_2	15	BW_1
8	CF_4	16	WC_1

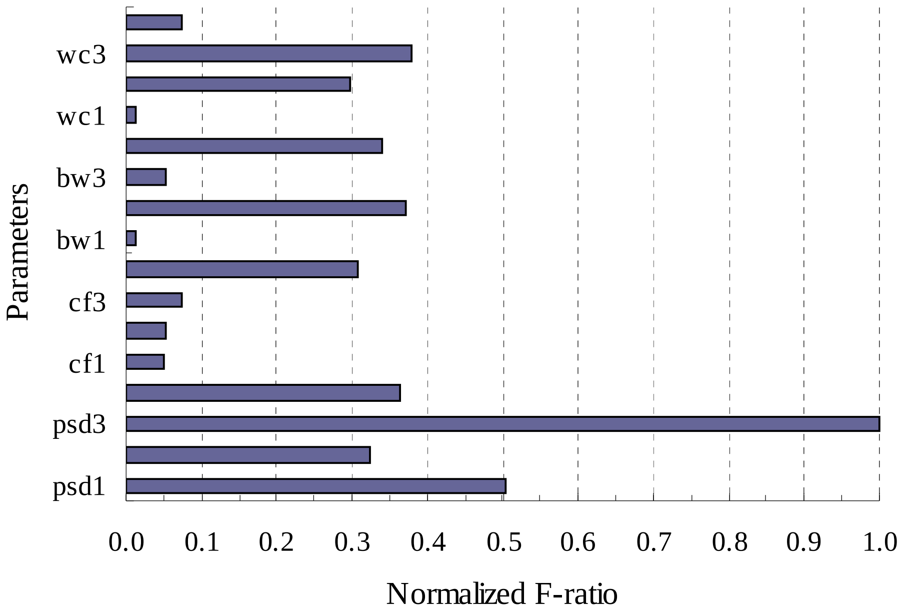


Figure 4.22 Normalized F-ratios measured for interview vocal features of female remitted-depressed comparison.

As illustrated in Figure 4.22, the normalized F-ratios show that PSD$_3$ solely emerged with the greatest normalized F-ratio for this pairwise comparison. The procedure of reducing the dimensionality of feature space found PSD$_3$, PSD$_1$ and WC$_3$ as the primary feature set providing the best classification performance. Integrated quadratic discriminator designed by this feature set was performed to classify remitted speech and depressed speech. An average classification score of 81% (Table 4.5) was obtained with the fairly high measures of sensitivity (0.75) and specificity (0.87). This average score was also found to be the lowest accuracy of classification for discriminating analyses of reading speech. The depressed patients were classified more correctly than the remitted patients, as referred to by a higher measure of specificity.

4.3.4 Comparative Results of Interview Vocal Features of Remitted and Suicidal Speech

As presented in Figure 4.23, PSD_1 and PSD_2 were ranked the most powerful discriminating features for a group comparison of remitted speech and suicidal speech. Although PSD_1 and PSD_2 were both characterized by the comparatively high F-ratios, WC_3 was also determined to be another feature that can be combined with others to form the primary feature set for the multi-parameter classifier. Dimensionality reduction determined these features as the best discriminators providing the most effective recognition.

Table 4.4 Ranked interview vocal features of female remitted-suicidal comparison.

Rank	Feature	Rank	Feature
1	PSD_1	9	CF_2
2	PSD_2	10	WC_2
3	WC_3	11	CF_3
4	PSD_4	12	BW_3
5	BW_2	13	BW_1
6	PSD_3	14	CF_4
7	WC_1	15	CF_1
8	BW_4	16	WC_4

The 82.42% correct classification was obtained by the integrated classifier using PSD_1, PSD_2 and WC_3 as discriminators to define a boundary between these groups. The performance measures of sensitivity (0.83) and specificity (0.83) imply that the classifier was equally effective to classify both remitted and suicidal speech. All of determined sets of primary vocal features are summarized in Table 4.6.

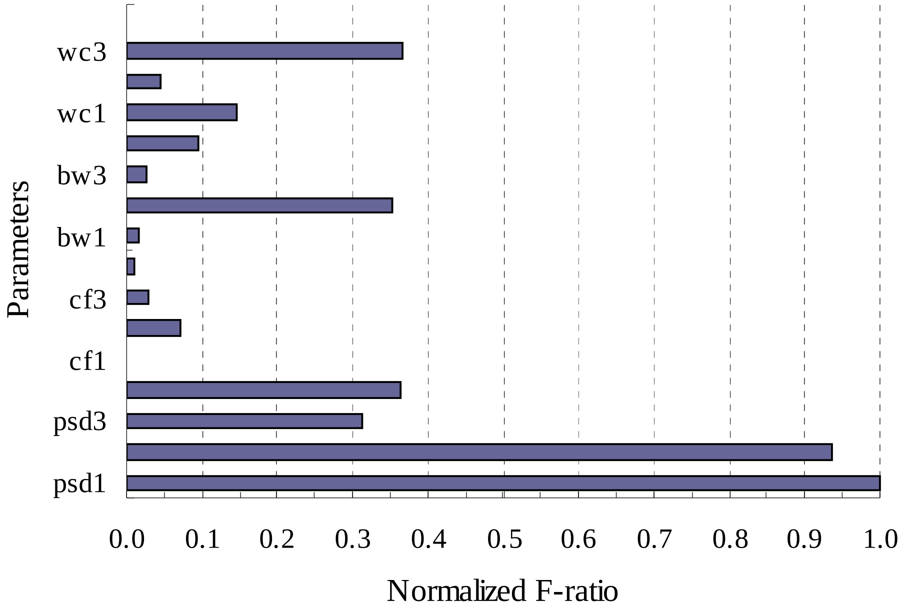


Figure 4.23 Normalized F-ratios measured for interview vocal features of female remitted-suicidal comparison.

Table 4.5 Summary of comparative classification performances based on pairwise analyses of interview speech.

Pairwise Group	**%Classification**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Depressed/Suicidal	85.75	0.89	0.80	0.89	0.88
Remitted/Depressed	81.08	0.75	0.87	0.86	0.83
Remitted/Suicidal	82.42	0.83	0.83	0.83	0.86

Table 4.6 Summary of primary acoustic features maximizing group separation and classification performance for interview speech groups.

Pairwise Group	**Primary Feature Set**
Depressed/Suicidal	PSD_2, CF_2, PSD_1, BW_2
Remitted/Depressed	PSD_3, PSD_1, WC_3
Remitted/Suicidal	PSD_1, PSD_2, WC_3

4.3.5 Comparative Results of Reading Vocal Features of Depressed and Suicidal Speech

The mean and standard deviation values measured for acoustical features extracted from the reading speech of suicidal, depressed and remitted patient groups are summarized in Table 4.7. Several features characterized by the difference in acoustic properties evidently exhibited the changing trends relative to the degree of psychological state in patients. Specifically, CF_2, BW_1, BW_2, BW_4 and WC_1 appeared to increase as the severity of the psychological state increased. Conversely, BW_3 and WC_3 negative correlate to psychological severity.

Table 4.7 Summary of the feature means and standard deviations of the categorized groups of reading speech.

Feature	**Suicidal**	**Depressed**	**Remitted**
PSD_1	(0.87, 0.06)	(0.77, 0.08)	(0.77, 0.09)
PSD_2	(0.11, 0.05)	(0.21, 0.07)	(0.20, 0.07)
PSD_3	(0.01, 0.01)	(0.01, 0.01)	(0.02, 0.01)
PSD_4	(0.01, 0.01)	(0.01, 0.01)	(0.01, 0.00)
CF_1	(289.46, 41.42)	(274.09, 37.62)	(278.47, 60.44)
CF_2	(1356.83, 127.88)	(1203.29, 133.87)	(1158.99, 164.03)
CF_3	(2373.47, 223.94)	(2177.16, 143.63)	(2177.95, 169.73)
CF_4	(3465.71, 270.38)	(3294.62, 127.35)	(3294.82, 193.99)
BW_1	(286.15, 37.96)	(270.97, 35.62)	(267.63, 52.91)
BW_2	(536.23, 62.00)	(535.65, 72.14)	(470.36, 50.46)
BW_3	(380.12, 45.84)	(394.50, 54.28)	(395.80, 36.74)
BW_4	(358.57, 46.75)	(334.72, 35.41)	(329.65, 34.41)
WC_1	(0.25, 0.04)	(0.23, 0.03)	(0.22, 0.03)
WC_2	(0.23, 0.03)	(0.26, 0.04)	(0.22, 0.02)
WC_3	(0.12, 0.02)	(0.13, 0.02)	(0.14, 0.01)
WC_4	(0.10, 0.01)	(0.09, 0.01)	(0.09, 0.01)

PSD_1, CF_1, CF_2, CF_3, CF_4, BW_1, BW_2, BW_4 and WC_1 estimated from suicidal speech were all characterized by the highest averages as compared with those of other diagnostic groups. Conversely, the mean values measured for PSD_2, BW_3 and WC_3 in suicidal speech were found to be decreased and the similar decreasing trends for these features were also identified for suicidal speech of the interview session study.

In depressed speech, the mean values of PSD_2 and WC_2 appeared to highly increase as compared against to those of the suicidal speech. CF_1 was found to conversely reduce for depressed speech when compared to that of suicidal speech and it was also identified as the lowest center frequency among those of diagnostic groups. The trends or patterns of the vocal features based on observations made are not necessary to suggest any statistically significant differences between speech groups that may correlate with the improvement of classification performance.

Table 4.8 Ranked reading vocal features of female depressed-suicidal comparison.

Rank	Feature	Rank	Feature
1	PSD_2	9	BW_4
2	PSD_1	10	WC_3
3	CF_2	11	BW_1
4	CF_3	12	CF_1
5	WC_4	13	PSD_4
6	CF_4	14	BW_3
7	WC_2	15	PSD_3
8	WC_1	16	BW_2

The rank-ordered features of the reading speech groups based on the F-ratio pairwise analyses are presented in Tables 4.8-4.10. Plots of the F-ratio measures for all pairwise analyses are depicted in Figures 4.24, 4.25 and 4.26. As shown in Figure 4.24 for the F-ratio discriminant analysis of measuring group separability between depressed and suicidal speech, PSD_2, PSD_1 and CF_2 were determined to rank the first, second and third powerful discriminating features.

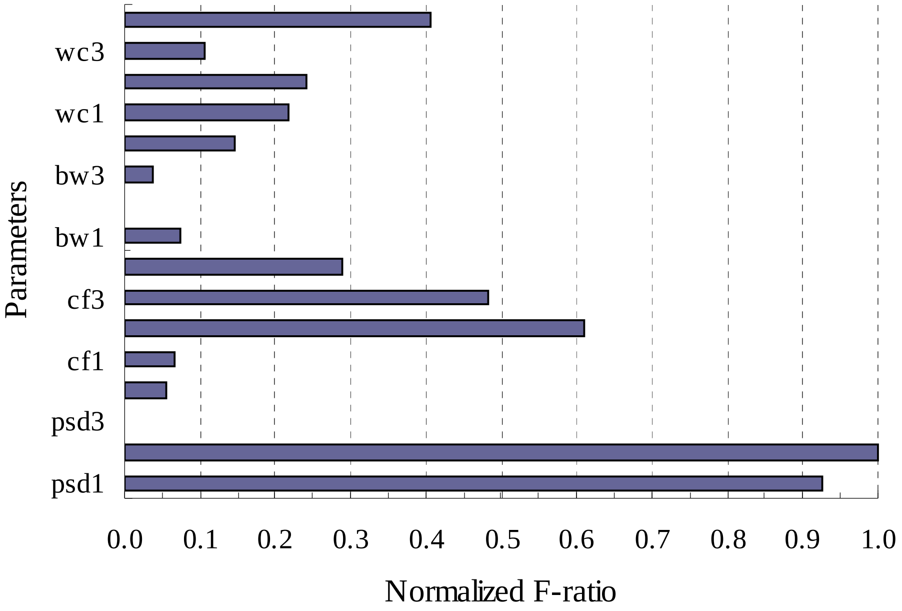


Figure 4.24 Normalized F-ratios measured for reading vocal features of female depressed-suicidal comparison.

By comparing with the F-ratio results of interview speech for the same comparison, these three rank-ordered features of reading speech were identically determined as same as those three features of interview speech analysis with very high discriminative power. The correct classification scores and performance measures resulted from performing the pairwise quadratic discriminating analyses on the randomized 25% of reading vocal feature samples are summarized in Table 4.11. The 90.33% accurate classification was obtained by the multi-parameter classifier designed using PSD_2, PSD_1 and CF_2 as the powerful discriminators in classifying depressed patients and suicidal patients. The integrated classifier performed much more effectively in classifying suicidal patients than depressed patients, as referred to by measures of specificity (0.92) and sensitivity (0.89) respectively.

4.3.6 Comparative Results of Reading Vocal Features of Remitted and Depressed Speech

In remitted speech, PSD_3, BW_3 and WC_3 were all characterized by the increased mean values as compared against with those of depressed speech and suicidal speech. Conversely, the mean values of CF_2, BW_1, BW_2, BW_4, WC_1 and WC_2 were determined to be the lowest in remitted speech as compared among diagnostic speech groups. Based on the result of the remitted-depressed feature analysis presented in Figure 4.25, WC_2, BW_2 and PSD_3 exhibited to be the most discriminating features whose normalized F-ratios were noticeably higher as compared to those of other features.

Table 4.9 Ranked reading vocal features of female remitted-depressed comparison.

Rank	Feature	Rank	Feature
1	WC_2	9	PSD_2
2	BW_2	10	BW_4
3	PSD_3	11	CF_4
4	WC_3	12	BW_1
5	PSD_4	13	PSD_1
6	WC_4	14	BW_3
7	CF_2	15	CF_3
8	WC_1	16	CF_4

The procedure to determine a set of the optimal features by observing accurate recognitions found WC_2, BW_2 and PSD_3 as the most powerful discriminating features for group discrimination. These feature set were used to investigate the effectiveness of an integrated classifier in distinguishing depressed and suicidal patient group. The 82.67% correct classification accuracy was obtained as the lowest found among the results of pairwise analyses of reading speech. As a result of high specificity (0.89) summarized in Table 4.11, the designed classifier performed much better to identify depressed patients than remitted patients.

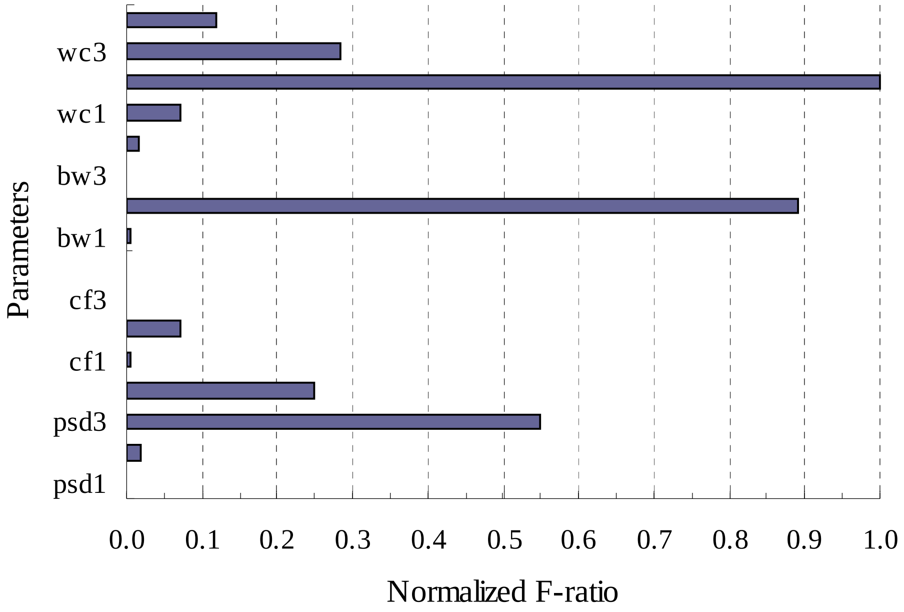


Figure 4.25 Normalized F-ratios measured for reading vocal features of female remitted-depressed comparison.

4.3.7 Comparative Results of Reading Vocal Features of Remitted and Suicidal Speech

As shown in Figure 4.26, it can be observed that the PSD_1, PSD_2, and CF_2 features were not only three parameters characterized by the comparatively high measures of group separation, but WC_3 and BW_2 were also determined to indicate high F-ratios. As result of dimensionality reduction, CF_2 and PSD_1 were identified by the highest group separating abilities as the most powerful discriminative features.

Table 4.10 Ranked reading vocal features of female remitted-suicidal comparison.

Rank	Feature	Rank	Feature
1	CF_2	9	PSD_3
2	PSD_1	10	CF_4
3	PSD_2	11	BW_4
4	WC_3	12	WC_4
5	BW_2	13	BW_1
6	WC_1	14	BW_3
7	CF_3	15	WC_2

The classifier using CF_2 and PSD_1 performed equally best in classifying both groups of remitted patients and suicidal patients with nearly identical measures of sensitivity (0.95) and specificity (0.96). The average correct classification of 94.17% was the highest score found in discriminating analyses of using the reading speech samples. In addition, as seen in Figure 4.26, there are five features, CF_2, PSD_1, PSD_2, WC_3 and BW_2 appearing as the most promising discriminators. However, all of these features have been tested for their corresponding classification rates. The experimental results demonstrated that the reduced feature set comprising CF_2 and PSD_1 provided the best correct classification rate.

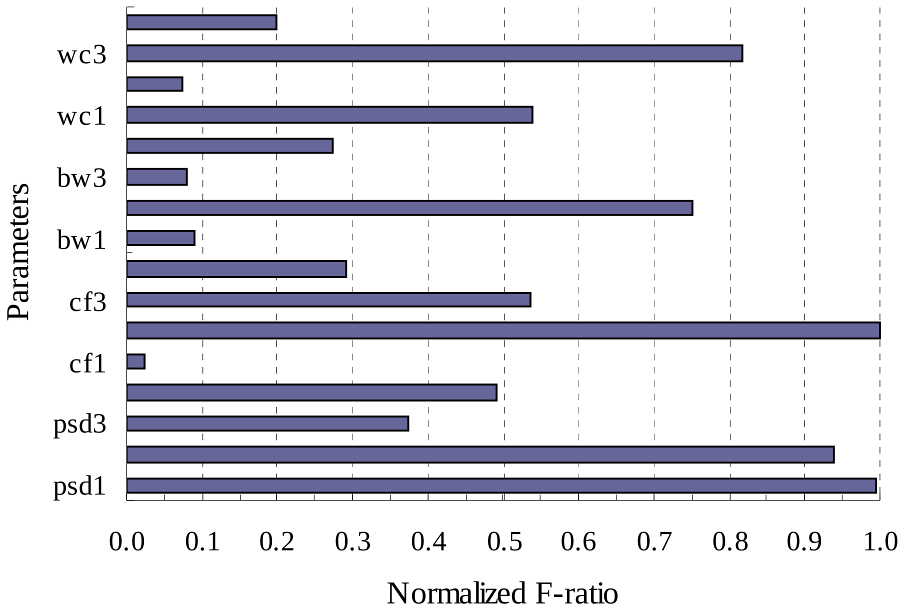


Figure 4.26 Normalized F-ratios measured for reading vocal features of female remitted-suicidal comparison.

Table 4.12 summarizes all primary acoustic feature sets that were statistically and experimentally determined to maximizing group separation and classification performance. It can be observed that primary feature sets determined from all pairwise discriminant analyses were often characterized by either one or two of the following GMM-based features: CF_2, BW_2 or WC_2 in their discriminative feature set rather than the PSD-based features.

Table 4.11 Summary of comparative classification performances based on pairwise analyses of reading speech.

Pairwise Group	**%Classification**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Depressed/Suicidal	90.33	0.89	0.92	0.94	0.84
Remitted/Depressed	82.67	0.72	0.89	0.84	0.84
Remitted/Suicidal	94.17	0.95	0.96	0.94	0.93

Table 4.12 Summary of primary acoustic features maximizing group separation and classification performance for reading speech groups.

Pairwise Group	**Primary Feature Set**
Depressed/Suicidal	PSD_2, PSD_1, CF_2
Remitted/Depressed	WC_2, BW_2 PSD_3
Remitted/Suicidal	CF_2, PSD_1

4.4 Discussion

4.4.1 Discussion on Comparative Results of Interview Speech Study

Features characterizing energy ratios of PSD_1, PSD_2 and $PSD_{3,}$ and GMM-based energy concentrations of CF_2, BW_2 and WC_3 were identified as correlates of psychological state, and were successfully used in pairwise classification analyses for distinguishing diagnostic groups of patients. The results from this study show that depressed and suicidal speech samples were effectively differentiated from remitted speech by incorporating PSD_1, PSD_2, PSD_3 and WC_3 in integrated classifiers. The classifier performed slightly better on suicidal speech (82.42%) than on depressed speech (81.08%) as comparisons to remitted speech. These high correct classifications imply that the studied feature set is capable of identifying the normality of mental condition in the recovering patients from the severities of mental state in patients suffering from depression and suicidal risk. The boundaries seem to well separate depression and suicidal risk from normality.

The highest accumulative performance (85.75%) was obtained using PSD_2, CF_2, PSD_1 and BW_2 determined as the most discriminating features for classifying depressed and suicidal patients. As shown in Table 4.1, the suicidal speech exhibited significant increases in

PSD_1, CF_2 and BW_2, and reduction in PSD_2 as compared to those of the depressed speech. Moreover PSD_1, CF_2 and BW_2 exhibited the positive correlation with the severity of the mental state. Oppositely, only PSD_2 was decreased as the severity of the mental condition increased. Classifier performance was found to be dramatically improved when these four features were used in classification instead of using a single feature or a reduced feature set with a number of features less than four (Figure 4.19). One of observations on the measured power of group separation for the depressed-suicidal speech comparison is that a number of vocal features with high normalized F-ratios (>0.5) is greater than that observed from the remitted-depressed comparison and remitted-suicidal comparison. In another word, the vocal features obtained from the depressed-suicidal discriminating analysis revealed more promisingly discriminative properties than those of other comparisons.

The results of the pairwise analyses suggest that mental conditions involving with depression and suicidal risk influence acoustical properties of speech in much different way. There appears to be differences in the intensity of vocal energy, responding frequencies and bandwidths of resonant peaks appearing on spectrum introduced by each psychiatric disorder and these differences make speech acoustic properties much more separable and they are supportive of group separation for these groups. The classification performance measures of sensitivity (0.89) and specificity (0.80) suggest that the classifier is slightly more effective to characteristics of depressed speech than that of suicidal speech.

Another finding is that the CF_2 and BW_2 of the GMM-based feature membership have emerged as the strongest vocal correlates of psychological states affected by depression and suicidal risk. These two features are model parameters of the second density component of a GMM. The second probability parameter (WC_2) did not exhibit acoustic characteristic of being a good discriminator. The differentiating power of WC_2 was measured and found to be the fourteenth rank-ordered parameter based on its F-ratio value. In addition, the speech samples of depressed and suicidal patients were both characterized by nearly identical mean value for WC_2 (0.25) as presented in Table 4.1.

Previous studies of the depressed speech have shown that the significant changes in the first and second formant frequencies were determined to correlate with depressive severity. The first formant frequency was reported to decrease in speech of patients recovering from a depressive disorder, while a reduction in the second formant frequency was identified as the strong vocal affect of depressed patients [26,36,61].

France [40] has reported formants as the most effective and consistent indicators of depressed state in patients, but his findings were not completely supportive of other previous studies. He demonstrated that depressed speech was characterized by higher average measures of the first formant frequency and bandwidth, and lower average measures of the second and third bandwidths as compared to other groups of control and dysthymic speech. The second and third formant bandwidths were investigated and reported to provide the best classification performance for all pairwise analyses. However, there was some uncertainty in his investigation on the formant bandwidth features concerning its definition. In his study, the LPC-based formant analysis was used to estimate the formant bandwidths from speech samples, which were calculated from the roots of the predictor polynomial. These estimates of bandwidth undoubtedly differ from the actual formant bandwidths due to the arbitrary imposition of model order made in analysis of formants. Moreover, the ambient variability during recordings may involve with the results of acoustic analyses in his study, including background noise and phone/microphone characteristics.

As compared with our results, the change in CF_2 of depressed speech was found to be supportive of other previous studies in that reduction in the second formant frequency has been identified as the vocal characteristic of depressed individuals. As reported by France et. al, the narrower bandwidth of the second formant characterizing depressed speech was determined and it agreed with our experiment that the reduced BW_2 was identified in depressive speech. However, our average values of CF_2 and BW_2 totally differ from those of other studies, but the changing trends in CF_2 and BW_2 is generally the same. Even though the difference in methodology exists, the vocal characterization of persons in the depressive state consistently revealed the similar acoustical properties for the same type of affective speech that has been studied.

The formant information can not be considered to make a comparison of analyzed results obtained from our present analyses and the previous studies [7] for female suicidal speech, since this kind of speech has never been investigated before for affective pattern of formants. Our investigation was the pilot study of comparing suicidal speech group with other groups of depressed speech and remitted speech.

The significant differences in the proportion of energy can be observed for the comparison of depressed and suicidal speech. The classifier designed by using PSD_1 and PSD_2 incorporating two other primary GMM-based features yielded the highest classification performance for the depressed-suicidal comparison. As compared to the distribution of energy in depressed speech, a trend of energy shift can clearly be identified for the suicidal speech in the 0-500 Hz frequency sub-band (Table 4.1). The energy of suicidal speech shifted from the sub-band #2 (500-1,000 Hz) to sub-band #1 (0-500 Hz). In depressed speech, the opposite shifting trend of energy from the 0-500 Hz sub-band to the 500-1,000 Hz sub-band can evidently be observed as a basis for comparison with suicidal speech. In addition, the remitted speech exhibited the energy shifts from the sub-band #1 (0-500 Hz) to the sub-band #2 (500-1,000 Hz) and even toward the sub-band #3 (1,000-1,500 Hz), as compared with distributions of energy in suicidal speech and in depressed speech. These energy shifts illustrate that the energy spectrum of remitted speech was flatter than those of both depressed speech and suicidal speech. It can imply that a female patient who recovered from being depressed spoke with more energy at high frequency above 500 Hz.

Our findings are consistent with the previously published results and also supportive of, especially, powerful discriminative features derived from formants. The statistical results imply the existence of acoustical variations in the characteristic of the vocal tract due to psychological stress, which inherently mediate into the spectral structure that can directly be captured by our approach. The method of feature extraction based on fitting of the vocal tract spectral structure is much closely equivalent to analysis of formants than other acoustic features of speech signal in this study. Although some detailed differences in methodologies exist, formants can be used as a close basis of comparison for our GMM-based vocal tract features.

Even though the GMM-based energy concentrations and formants both characterize the spectral structure of the vocal tract, our GMM-based vocal feature extraction approach is more advantageous as compared to the LPC-based formant estimation in that our approach is not model based and the determined vocal parameters directly represent the authentic characterizations of the vocal tract, whose acoustical properties correlate with psychological stress. In another words, the GMM-based energy concentration features are more robust representation of the vocal-tract frequency response than the popular LPC formant technique.

Changes in formant patterns in depressed speech can physiologically be explained. The muscle tone and rigidity in the vocal tract are increased during the depressive states and; consequently, the narrowing bandwidth in formants and shifting in proportions of energy toward higher frequencies can be found as the effects of more rigidity in the vocal tract for depressed patients as compared with non-depressed controls. The increased muscle tone induces the coordination of laryngeal, pharyngeal, and faucal constriction to cause a shift in energy spectrum. The harsh, metallic, and piercing voice in depressed patients is as a result of variations in formant bandwidths and distribution of energy [26,38]. The trends of speaking patterns observed from the specific vocal samples extracted from diagnostic patient groups on a basis of energy concentrations of vocal tract and spectral energy are distinguishable and they certainly correlate with variations in mental conditions caused by emotional disorders. Our findings suggest that the acoustical properties of studied features need to be investigated more for a success of medical technology development. This work could lead to develop the methodology relying on speech processing to aid clinical effort in diagnosis of psychological illness.

4.4.2 Discussion on Comparative Results of Reading Speech Study

Several vocal features among diagnostic groups exhibited significant differences in their acoustic properties. Features characterizing the spectral energy and vocal tract energy concentrations were determined to be vocal correlates of depression and suicidal risk, and were successfully employed in discriminant analyses for a comparison of depressed patients and suicidal patients. As seen in Table 4.12, depressed and suicidal patients were effectively discriminated from remitted patients by incorporating PSD_1, PSD_3, CF_2, BW_2 and WC_2 in classification analyses.

The integrated classifiers designed by these primary features effectively performed in assigning suicidal patients to their group with approximately 11% more correct classification accuracy than in identifying depressed patients, when both diagnostic groups compared with remitted patients (Table 4.11). The greater score of classification imply that acoustic characteristics of such primary features were affected by the near-term suicidal potential more significantly than severe depression. The results suggest that PSD_1, PSD_3, CF_2, BW_2 and $WC_{2v}$are capable of differentiating normality from severity of psychiatric disorders.

The results of the depressed-suicidal speech classification imply that there appeared to be some differences in measurable amount of energy in spectra and changes in spectral pattern of speech induced by each psychiatric disorder. It seems to be these differences that made the within-group scatter of speech samples becoming smaller and the between-group difference greater. The integrated classifier designed by the following features, PSD_2, PSD_1 and CF_2, yielded the considerably high discriminating performance of 90.33% in performing the depressed-suicidal speech classification as presented in Table 4.11. Sensitivity (0.89) and specificity (0.92) also infer that the classifier performance was slightly sensitive to characteristics of suicidal speech rather than to that of depressed speech. Nevertheless, they both represented the great effectiveness in identifying two groups of psychopathological disordered patients.

As presented in Table 4.7 for reading speech, suicidal speech exhibits significant increases in PSD_1 and CF_2 and a reduction in PSD_2 as compared against to those of depressed speech. It was also observed that CF_2 exhibited positive correlation with increasing severity of mental state. The testing results of recognition rates using various set of primary features were found to be most dramatically improved, when PSD_1, CF_2 and PSD_2 were used in classification as the best discriminators instead of using other feature sets.

Additional observation on group separation shows that the differences in acoustic properties of depressed speech and suicidal speech were characterized distinctly by the PSD features (Figure 4.24). The vocal features derived from the proportions of total energy still confirmed and were supportive of being the most powerful acoustical discriminators of identifying the severe conditions of mental illness in patients. The GMM-based CF_2 feature has emerged as a member of the optimal feature set as well, which was consistently determined to be the same as that found in previous depressed-suicidal classification of interview speech (Figure 4.19). However, the F-ratio in CF_2 of the reading speech was observed to be significantly lower than that of the interview speech. It may imply some physical changes taking place in the vocal tract area to shape the reading sound, while a patient was trying to utter that sound. The acoustic properties of speech were possibly altered by these changes.

As shown in plots of the F-ratios for all pairwise analyses of interview speech depicted in Figures 4.19, 4.22, 4.23 and for those of reading speech depicted in Figures 4.24, 4.25 and 4.26, we found that the comparisons illustrating more features with values of F-ratio greater than 0.5 tended to yield the most effective classification rate. In study of reading speech, for example, the remitted-suicidal comparison was found to agree with what we observed by getting the most accurate classification score, 94.17% and having seven features characterized by the normalized F-ratios higher than 0.5 (see Figures 4.26 and Table 4.11). It is also true for the study of interview speech, as observed for the depressed-suicidal pairwise with the highest classification score of 85.75% (see Figures 4.19 and Table 4.5). We may conclude for the study of female speech that the F-ratio is effective as discriminating measure to predict how accurate classification results would be relating to number of features showing high group separation.

In reading speech (Table 4.7), the relating trends of individual PSD_2, PSD_1 and CF_2 were determined to be consistent to those of the interview speech classification for depressed-suicidal comparison (Table 4.1), except that the means of individual features that were slightly different. The energy shifts were found to be similar to those of interview speech study for all diagnostic groups. However, a slight difference in amount of shifted energy can be observed for both PSD features of depressed speech and suicidal speech. CF_2 was characterized by a higher frequency for both depressed and suicidal speech as compared to those of study of interview speech.

The significance differences in pairwise classifications between the recording protocols of interview speech and reading speech were statistically determined by using two-sample test for hypothesis testing [77]. In the depressed-suicidal comparison, there was no significant difference ($p<0.05$) between the correct classifications for interview speech (85.75%) versus reading speech (90.33%). The remitted-depressed comparison was also determined to be no significant ($p<0.05$) between the classification scores of 81% for interview speech and 82.67% for reading speech. For the remitted-suicidal comparison, significant classification difference ($p<0.001$) was found between interview speech (82.42%) and reading speech (94.17%). These comparative statistics indicated that both interview speech and reading speech exhibited the similar acoustical properties in distinguishing depressed patients from suicidal patients, and remitted patients from depressed patients.

However, the reading speech revealed much more effective than the interview speech in classifying suicidal patients and remitted patients.

One of the limitations of this work is the limited number of patients with acceptable quality of sound. Even though the cross validation procedure was employed to get more reliable estimation of classification accuracy as compared to other techniques such as the “hold-one-out” technique, a larger database will evidently yield more accurate estimation of classification. Although speech samples used in this study were collected from acoustically controlled recording environment with modern high-fidelity equipment, it is still difficult to acquire recording samples of good acoustic quality for objective voice analysis. In addition, speech samples of suicidal patients are rare to collect during clinical interviews. Patients may not reveal suicidal behavior or at peak of suicidal risk during the interview recording.

The suggestive directions of the future work to improve classification performance for this ongoing research should involve in more controlled environment of audio recording, standardized procedure for individual interview sessions, preprocessing implementation before feature extraction procedure such as filtering of background noises and sound artifacts and even alternative method of voice detection for speech samples.

Other vocal parameters may be worth an investigation, such as the spectral entropy as acoustical features relative to information source. The spectral entropy is a measure of disorganization of speech spectrum (or uncertainty of random variables) and it has been used to capture formants or the peakiness of a spectral distribution [62,63,64]. As demonstrated in this research, formant information and distribution of spectral energy as vocal correlates of mental illness, the application of spectral entropy to identification of severity of near-term suicidal states is possible. The entropy concept for the speech classification is based on assumption that speech spectrum is more organized for voicing segments than unvoiced or noise segments. Spectral peak seems to be more robust to noise. Thus, individual voiced segments of speech would induce low entropy since there are clear formants in that region, while spectra for unvoiced or noise segments characterized by a flatter distribution would have higher entropy. This state-of-art entropy feature may contribute a great deal of benefits for developing diagnostic tools that can assist clinicians and reduce the clinical effort in diagnosis of psychopathological disorder.

CHAPTER V

DIRECT ACOUSTIC FEATURE EXTRACTION USING ITERATIVE EM ALGORITHM AND SPECTRAL ENERGY FOR CLASSIFYING SUICIDAL SPEECH IN MEN

Abstract

Acoustic analyses of the energy distribution and the GMM-based spectral modeling of the vocal tract response were performed on male speech samples extracted from audio recordings of 8 suicidal patients, 8 depressed patients, and 8 remitted patients. Individual patients of these diagnostic groups were represented by the vocal features derived from particular acoustics analyses performing on their speech samples which comprise of interview speech and reading speech. The pairwise discriminant analyses were performed on feature samples of patient groups and results revealed that the acoustical features of the reading speech provided more separability than those of the interview speech. The classification accuracies determined from the discriminant analyses confirmed the investigated vocal features capable of being powerful discriminators of severity of near-term suicidal risk. An 88.5% accurate classification was obtained from the depressed-suicidal comparison on the reading speech and 85.58 % on the interview speech. These high performances were supportive of the promising abilities of group discrimination in that the features derived from spectral energy and vocal–tract spectral modeling revealed specific changes in their acoustic properties correlated with serious mental states, known as vocal affects which can provide diagnostic cues of psychiatric disorders.

5.1 Introduction

Suicide is the major public health problem in the United States among adults and young people and is showing the increasing tendency every year. It is the eleventh leading cause of deaths in American population. The recent statistics have shown that approximately 32,439 people successfully committed suicide attempts. The overall rate was 10.9 suicide deaths per 100,000 people reported in 2004 [1]. Most of persons who committed suicides carried a diagnosis of emotional disorder, clinical depression [2].

Prevention of suicides involves evaluation of severity of near-term suicidal risk in patients diagnosed as being severely depressed. This task requires the patient's history information, psychological testing records, clinical reports, and current situation recorded during clinical interview. Gathering these kinds of information and diagnostic procedure for assessing risk of committing suicide in patients are time-consuming for seeking the immediate diagnostic conclusion and data required during diagnosis may not be available at time of clinical judgment in the urgent situation. In addition, the present diagnosis of this psychiatric disorder, suicidal risk, requires the experienced and skillful psychiatrist to make a decision on assigning a patient to the correct category of disorder based on degree of their mental condition and there is no formulation or predictive trend that allows a psychiatrist to empirically combine the information from various sources to make clinically certain judgment [66].

In present time, there are very few acceptable objective diagnostic tools available in clinics, which can provide some quantitative expressions of imminence of suicidal risk and assist clinicians to diagnose disorder more correctly. It has been reported that depression is the most clinically emotional disorder relating to suicidal behaviors. The

80% of completed suicides were committed by people who were depressed [66]. Thus, the degree of depression affecting mental condition of persons possibly suggested as an important factor for evaluation of suicidal potential. Since the number of depressed people who successfully committed suicide was reported to increase [66].

Research has suggested that voice can provide important information about the immediate psychological state of a speaker. The acoustic properties of speech at the non-content sound reflect the intensity of the patient's mental state [5]. The suicidal state can be associated with significant perceptual changes in speech production and articulation that alter acoustic properties of vocal parameters, which differ from that of non-suicidal persons. The published studies have demonstrated that speech of emotionally disturbed patient exhibited a significance of irregularity in vocal patterns based on the spectral energy distribution, formants, mel-cepstral coefficients [40,8,9,36].

In the past, diagnosis of severe depression and suicidal risk solely depended on the clinician's expert experience for assessing the symptoms. The perceived impressions that lead to the diagnosis of near-term suicidal risk are believed to be influenced by some certain vocal parameters. Studies have shown that observer listening to voice samples is able to describe emotional states of a speaker with some accuracy [67,68]. However, the exact vocal parameters are rarely named by clinician for their characteristics as potential indicators of depression, and attribution and impression of parameters depend upon methodology used in quantitative analysis.

The objective of this study is to reinvestigate the acoustic properties of high-risk suicidal, depressed and remitted speech based on the vocal parameters derived from a basis of spectral energy and vocal-tract spectral modeling. The goal is to test the hypothesis that there are significant acoustical differences among speech samples representing different diagnostic groups and; by these differences, an effective classifier can successfully be designed and discrimination of diagnostic patient groups can be accomplished through speech analysis.

This paper is organized as follows: Section 2 provides the detailed descriptions of database, acoustic feature extractions. Section 3 presents the experimental results of the pairwise discriminant analyses performing on interview and reading speech samples of diagnostic patient groups. Section 4 discusses the results and findings from this work.

5.2 Methodology

5.2.1 Database

This work is a part of an ongoing research study supported by the *American Foundation for Suicide Prevention*. The speech samples were collected from three different groups of subjects who carried a diagnosis of either suicidal risk, depression, or remission from depression. Each categorized group has 8 male subjects. The ages of patients were between 25 and 65 years. Each subject has two types of audio recordings. The first type of audio was recorded during a clinical interview with a therapist, spontaneous speech, and another was recorded during a session of reading a predetermined part of a book, reading speech. During a text-reading session, each subject read the standardized text, the "Rainbow Passage" [45], which has been used in speech science since it contains all of the normal sounds in spoken English and it is phonetically balanced. Two speech samples were randomly selected from the audio recordings of the same patient participating in both clinical sessions.

The same recording environment and settings were made for all clinical interviews. This acoustically controlled environment is necessary for acquiring the clean speech samples. The same audio data acquisition system and preprocessing procedure as reported in Chapter IV were also used in this work. Two more preprocessing steps were made before further acoustic analyses. First, all speech samples were tested for voicing and only voiced segments of speech were kept for further analysis. The voiced/unvoiced detection algorithm based on energy weighting proposed by Ozdas et al. [39] was used in this work to decide which section of the patient's speech is voiced, unvoiced or silent. The length of voiced speech was approximately 50% of that of an original speech sample. This percentage seemed to be consistent for most of analyzed speech samples. Second, all voiced segments were detrended and normalized to compensate all possible differences in recording level among categorized patient groups, resulting in obtaining a variance of speech sample equal to one. In this work, the unprocessed speech with approximately 6 minutes was extracted from database of interview audio recordings and approximately 2 minutes from reading audio recordings to represent each patient.

5.2.2 Acoustic Feature Extraction

The complete description and implementation of the acoustical and statistical methods used in this work are identical to those presented in the methods sections of Chapter III and Chapter IV.

5.3 Experimental Results

5.3.1 Comparative Results of Interview Vocal Features of Depressed and Suicidal Speech

The mean and standard deviations of the vocal features determined from the suicidal, depressed and remitted speech samples are summarized in Table 5.1.

Table 5.1 Summary of the feature means and standard deviations of the categorized groups of interview speech.

Feature	**Suicidal**	**Depressed**	**Remitted**
PSD_1	(0.76, 0.09)	(0.81, 0.10)	(0.80, 0.07)
PSD_2	(0.22, 0.09)	(0.16, 0.10)	(0.18, 0.07)
PSD_3	(0.02, 0.02)	(0.02, 0.02)	(0.02, 0.02)
PSD_4	(0.01, 0.01)	(0.01, 0.00)	(0.01, 0.01)
CF_1	(280.16, 25.33)	(272.41, 31.28)	(280.63, 47.25)
CF_2	(1263.59, 64.24)	(1199.51, 146.56)	(1203.68, 142.28)
CF_3	(2249.56, 98.88)	(2124.99, 174.57)	(2176.11, 168.49)
CF_4	(3430.67, 140.15)	(3226.89, 189.33)	(3248.27, 159.58)
BW_1	(281.53, 23.56)	(268.51, 28.61)	(275.30, 41.83)
BW_2	(501.95, 39.47)	(520.30, 64.14)	(500.65, 27.43)
BW_3	(384.52, 45.37)	(376.29, 25.62)	(382.72, 47.80)
BW_4	(357.13, 54.70)	(336.65, 25.59)	(356.06, 25.39)
WC_1	(0.24, 0.03)	(0.23, 0.03)	(0.23, 0.03)
WC_2	(0.24, 0.02)	(0.24, 0.03)	(0.24, 0.02)
WC_3	(0.13, 0.02)	(0.12, 0.02)	(0.13, 0.02)
WC_4	(0.10, 0.02)	(0.09, 0.01)	(0.10, 0.01)

Several features characterized by acoustic properties reveal the trends relating to the degree of mental state in patients suffering from depression and suicidal risk. Particularly, the ratio of energy in a 500-1,000 Hz frequency range (sub-band #2), CF_1,

CF_2, CF_3, CF_4, BW_1, BW_3 and BW_4 were increased in suicidal speech. However, reductions in PSD_1 and BW_2 of the suicidal speech can noticeably be observed as compared against with the depressed speech.

In suicidal speech, the means values calculated for PSD_2, CF_2, CF_3, CF_4, BW_1 and WC_1 were found to be the highest compared to those of the other two speech groups. Conversely, PSD_1 is only feature that was the least. Thus the PSD features show that the distribution of energy in suicidal speech is shifted from the 0-500 Hz band (sub-band #1) to the 500-1,000 Hz band (sub-band #2) relative to the other two groups.

In depressed speech, the mean values calculated for PSD_1 and BW_2 are the lowest compared to those of other groups. In addition, the lowest mean values can be noticed as well for the following parameters: PSD_2, CF_1, CF_2, CF_3, CF_4, BW_1, BW_3, BW_4, WC_3, and WC_4. Nevertheless, the trends of vocal features that were just described do not necessarily infer any significant differences in acoustic properties between class features which may contribute to increasing the performance of classification.

Table 5.2 Ranked interview vocal features of male depressed-suicidal comparison.

Rank	Feature	Rank	Feature
1	CF_4	9	PSD_3
2	CF_3	10	WC_2
3	PSD_2	11	WC_1
4	PSD_1	12	BW_2
5	CF_2	13	WC_3
6	WC_4	14	CF_1
7	BW_1	15	BW_3
8	BW_4	16	PSD_4

The rank-ordered features obtained from the F-ratio pairwise analyses of depressed and suicidal speech groups are presented in Table 5.2. The plots of normalized F-ratios calculated from vocal parameters of depressed and suicidal speech are illustrated in Figure 5.1. The F-ratio results show that the vocal parameters derived from the GMM-based spectral modeling (CF_4, CF_3, PSD1, PSD2) exhibited the highest discriminative power. CF_4 and CF_3 appeared to rank the first and second discriminating features with obvious difference in F-ratio values. The results of comparative F-ratios also showed that spectral energy features (PSD_2 and PSD_2) ranked third and fourth in power of group separation.

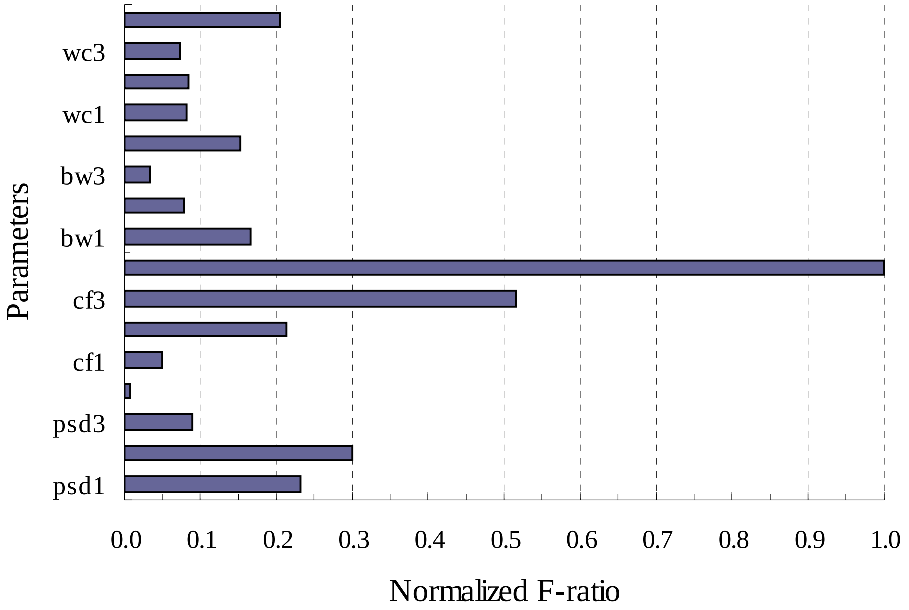


Figure 5.1 Normalized F-ratios determined from interview vocal features of male depressed-suicidal comparison.

The greatest recognition rate from the experiment of dimensionality reduction was obtained when CF_4, CF_3 and PSD_2 were used as primary features. As summarized in Tables

5.5 and 5.6, the classifier designed by these features yielded 85.58% classification accuracy. This accurate score was also found as the highest classification accuracy for pairwise analyses of interview speech. The results of classification performance indicated that suicidal patients were identified by classifier more correctly than depressed patients, as referred to by higher specificity (0.88) with respective to sensitivity (0.81). In this depressed-suicidal comparison, the depressed speech was chosen as conditional group in calculation of classifier performance.

5.3.2 Comparative Results of Interview Vocal Features of Remitted and Depressed Speech

In remitted speech, PSD_2, CF_1, CF_2, CF_3, CF_4, BW_1, BW_3, and BW_4 were characterized by the increased mean values as compared to those of depressed speech. Conversely, only BW_2 was found to decrease in remitted speech.

Table 5.3 Ranked interview vocal features of male remitted-depressed comparison.

Rank	Feature	Rank	Feature
1	WC_4	9	CF_1
2	BW_4	10	BW_1
3	PSD_3	11	BW_3
4	BW_2	12	PSD_1
5	WC_2	13	PSD_4
6	WC_3	14	CF_4
7	CF_3	15	WC_1
8	PSD_2	16	CF_2

As a result of F-ratios calculated for a comparison of remitted speech and depressed speech plotted in Figure 5.2, the first three rank-ordered features are: WC_4, BW_4 and PSD_3.

The spectral energy features again exhibited much less discriminative power than the GMM-based spectral features.

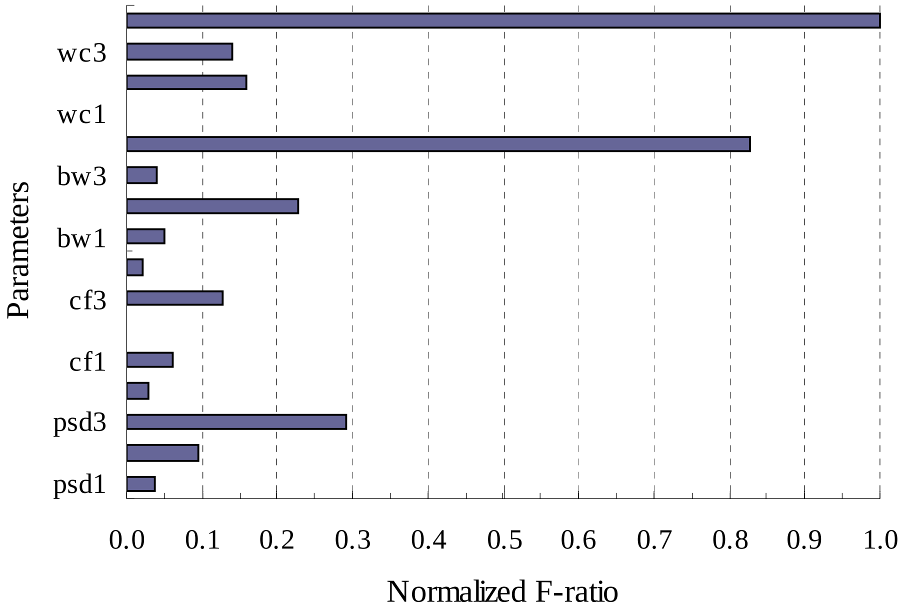


Figure 5.2 Normalized F-ratios determined from interview vocal features of male remitted-depressed comparison.

Based on the recognition rates, WC_4, BW_4 and PSD_3 were determined to be the best discriminating features providing the greatest recognition accuracy summarized in Table 5.6. This set of primary features was used in cross validation to evaluate the effectiveness of pairwise classification of remitted speech and depressed speech. The 72.25% average classification accuracy was obtained and it was the lowest score of correct classification found for pairwise analyses of classifying interview speech. As shown in Table 5.5 for calculated specificity and sensitivity, the classifier was slightly more effective in identifying depressed patients (0.75) than remitted patients (0.70) in which remitted speech was taken as conditional group for calculation of performance.

5.3.3 Comparative Results of Interview Vocal Features of Remitted and Suicidal Speech

Features characterizing by PSD_4, CF_1, BW_2, BW_3, BW_4, WC_2, WC_3 and WC_4 were identified to be similar for remitted speech and suicidal speech (Table 5.1). As plotted in Figure 5.3, CF_4 obviously appeared to be a sole feature emerging with the strongest discriminative power for comparison of remitted speech and suicidal speech. However, we found CF_2 and CF_3 ranking the second and third powerful discriminating features, which is probably combined with CF_4 as a feature set for discriminant analysis. It was also observed that PSD_2 and PSD_1 exhibited the F-ratios which were nearly equal to those for CF_2 and CF_3. These PSD features ranked the fourth and fifth discriminating features, as summarized in Table 5.4.

Table 5.4 Ranked interview vocal features of male remitted-suicidal comparison.

Rank	Feature	Rank	Feature
1	CF_4	9	PSD_3
2	CF_2	10	WC_4
3	CF_3	11	BW_2
4	PSD_2	12	BW_3
5	PSD_1	13	WC_2
6	WC_1	14	BW_4
7	PSD_4	15	CF_1
8	BW_1	16	WC_3

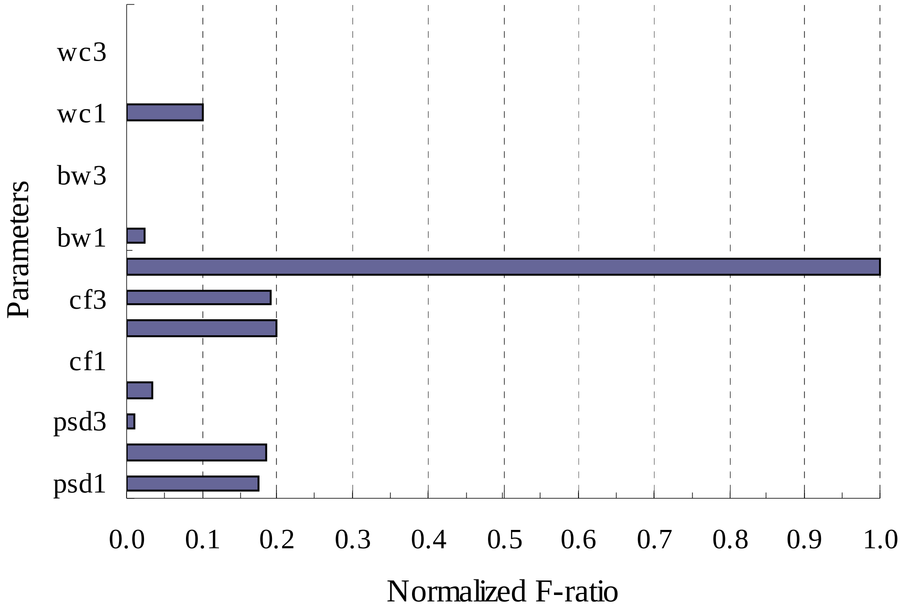


Figure 5.3 Normalized F-ratios determined from interview vocal features of male remitted-suicidal comparison.

As a result of dimensionality reduction, CF_4, CF_2 and CF_3 were determined as the most powerful discriminators for classification of remitted speech from suicidal speech. The classifier using these discriminators performed much more effectively in identifying remitted patients than suicidal patients, as indicated by sensitivity (0.90) and specificity (0.65). The average accuracy of pairwise classification was determined to be 81.08%.

Table 5.5 Summary of comparative recognition performances based on pairwise analyses of interview speech.

Pairwise Group	**%Classification**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Depressed/Suicidal	85.58	0.81	0.88	0.89	0.85
Remitted/Depressed	72.25	0.70	0.75	0.63	0.85
Remitted/Suicidal	81.08	0.90	0.65	0.86	0.83

Table 5.6 Summary of primary acoustic features maximizing group separation and recognition performance for interview speech groups.

Pairwise Group	**Primary Feature Set**
Depressed/Suicidal	CF_4, CF_3, PSD_2
Remitted/Depressed	WC_4, BW_4, PSD_3
Remitted/Suicidal	CF_4, CF_2, CF_3

5.3.4 Comparative Results of Reading Vocal Features of Depressed and Suicidal Speech

The means and standard deviations of the acoustic features determined from the reading speech samples of diagnostic groups of patients are summarized in Table 5.7.

Table 5.7 Summary of the feature means and standard deviations of the categorized group of reading speech.

Feature	**Suicidal**	**Depressed**	**Remitted**
PSD_1	(0.75, 0.08)	(0.83, 0.08)	(0.76, 0.11)
PSD_2	(0.22, 0.07)	(0.14, 0.08)	(0.22, 0.10)
PSD_3	(0.02, 0.01)	(0.02, 0.01)	(0.01, 0.00)
PSD_4	(0.01, 0.01)	(0.01, 0.00)	(0.01, 0.00)
CF_1	(287.75, 42.32)	(283.42, 36.37)	(303.02, 59.32)
CF_2	(1263.09, 86.09)	(1227.4, 152.88)	(1246.24, 190.67)
CF_3	(2266.55, 121.53)	(2152.9, 200.66)	(2199.23, 191.41)
CF_4	(3430.89, 132.61)	(3233.99, 208.45)	(3246.27, 131.86)
BW_1	(286.35, 40.34)	(277.11, 33.46)	(293.08, 48.80)
BW_2	(529.44, 63.82)	(521.56, 41.21)	(512.83, 37.38)
BW_3	(393.34, 54.61)	(367.97, 33.63)	(395.86, 33.99)
BW_4	(369.33, 77.15)	(339.06, 32.08)	(349.98, 39.02)
WC_1	(0.24, 0.04)	(0.24, 0.04)	(0.25, 0.05)
WC_2	(0.24, 0.02)	(0.24, 0.01)	(0.23, 0.03)
WC_3	(0.13, 0.02)	(0.12, 0.02)	(0.13, 0.02)
WC_4	(0.10, 0.02)	(0.10, 0.01)	(0.10, 0.02)

Many vocal parameters exhibited trends relating to the severity of mental state. Particularly, PSD_2, CF_1, CF_2, CF_3, CF_4, BW_1, BW_2, BW_3, BW_4 and WC_3 were characterized by the increased means for suicidal speech with respective to those of depressed speech. Conversely, only mean value of PSD_1 in suicidal speech was found to be less. In depressed speech, the distribution of energy shifted from the 500-1,000 Hz band (sub-band #2) toward lower frequency band (below 500 Hz). The trends or patterns of features found by observing from Table 5.7 are not necessary to suggest any statistically significant differences between diagnostic groups that may correlate with the improvement of classification performance.

Table 5.8 Ranked reading vocal features of male depressed-suicidal comparison.

Rank	Feature	Rank	Feature
1	CF_4	9	PSD_3
2	PSD_2	10	BW_1
3	PSD_1	11	WC_3
4	CF_3	12	BW_2
5	BW_3	13	WC_2
6	BW_4	14	CF_1
7	WC_4	15	CF_2
8	CF_2	16	WC_1

The rank-ordered features of the reading speech groups based on the F-ratio pairwise analyses are presented in Table 5.8. As plotted in Figure 5.4, the F-ratio analysis yielded CF_4, PSD_2, PSD_1 and CF_3 as the first, second, third, and fourth powerful discriminative features. As a result of dimensionality reduction in feature space, the rank-ordered CF_4, PSD_2, and CF_3 were statistically taken as the best optimal feature set giving the highest recognition rate for the depressed-suicidal comparison. An integrated classifier designed by these features yielded average classification score of 88.50% (Table 5.11). This integrated classifier performed slightly better in identifying depressed patients than suicidal patients. This effectiveness of classification was indicated by sensitivity (0.91) and specificity (0.88), respectively.

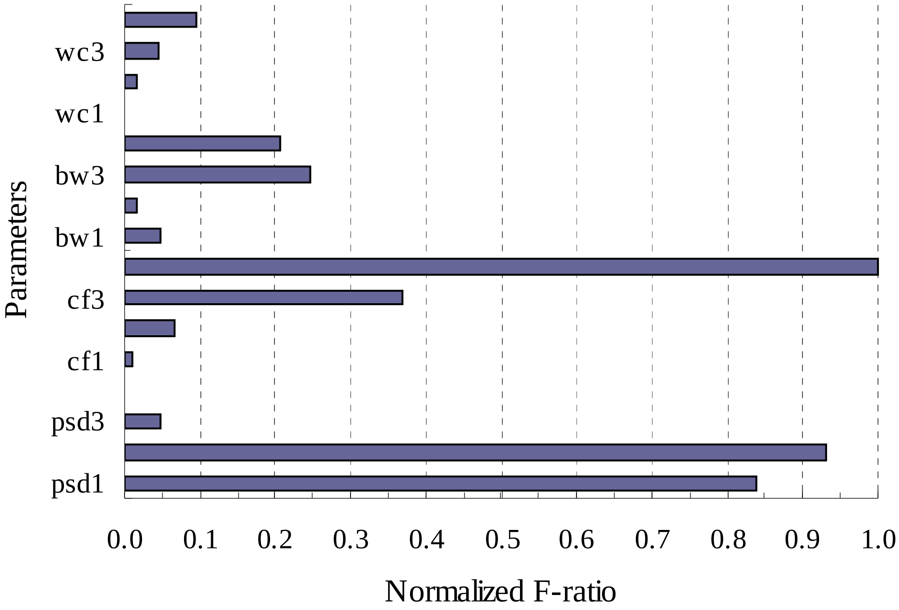


Figure 5.4 Normalized F-ratios determined from reading vocal features of male depressed-suicidal comparison.

5.3.5 Comparative Results of Reading Vocal Features of Remitted and Depressed Speech

In remitted speech, CF_1, BW_1, BW_3 and WC_1 were all characterized by having greater mean values. Conversely, the mean values of PSD_3, BW_2 and WC_2 were determined to be the lowest in remitted speech as compared against those of depressed speech (Table 5.7). Based on the F-ratio feature analysis presented in Figure 5.5, PSD_3, PSD_2, BW_3 and PSD_1 were statistically determined to be the most powerful discriminating features whose normalized F-ratios were greater than 0.5. The procedure to determine a set of the optimal features based on recognition rates found PSD_3, PSD_2 and BW_3 to be the best feature set.

Table 5.9 Ranked reading vocal features of male remitted-depressed comparison.

Rank	Feature	Rank	Feature
1	PSD_3	9	PSD_4
2	PSD_2	10	BW_4
3	BW_3	11	CF_3
4	PSD_1	12	BW_2
5	WC_3	13	WC_4
6	CF_1	14	WC_1
7	BW_1	15	CF_2
8	WC_2	16	CF_4

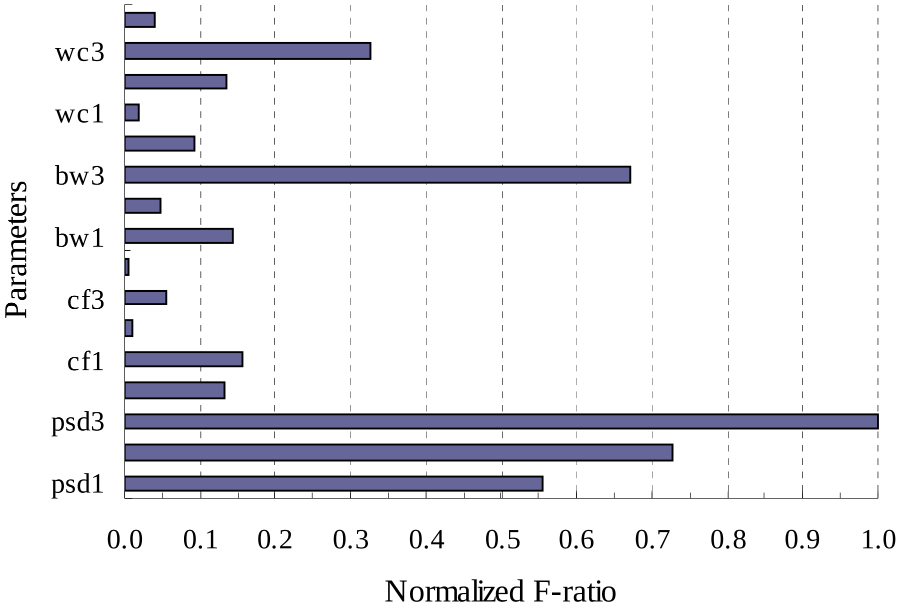


Figure 5.5 Normalized F-ratios determined from reading vocal features of male remitted-depressed comparison.

The 92% classification accuracy was obtained for this pairwise and found to be the highest correct score for pairwise classifications performing on the reading speech samples. As a result of specificity (0.97) and sensitivity (0.86) presented in Table 5.11, the classifier performed much more effectively in classifying depressed patients than remitted patients. This specificity measure was also obtained as the most effective performance among that of all pairwise analyses using the reading speech in classification and even better than that of the remitted-depressed classification using the interview speech (Table 5.5).

5.3.6 Comparative Results of Reading Vocal Features of Remitted and Suicidal Speech

As shown in Figure 5.6, it clearly showed that CF_4 and PSD_3 were only two features revealing the powerful discriminative properties and they exhibited much greater F-ratios as compared to other features in the same comparison of remitted speech and suicidal speech.

Table 5.10 Ranked reading vocal features of male remitted-suicidal comparison.

Rank	Feature	Rank	Feature
1	CF_4	9	WC_2
2	PSD_3	10	BW_1
3	CF_3	11	WC_4
4	BW_2	12	WC_1
5	BW_4	13	PSD_1
6	PSD_4	14	CF_2
7	CF_1	15	BW_3
8	WC_3	16	PSD_2

As a result of dimensionality reduction, CF_4, PSD_3 and CF_3 were identified to be the strongest discriminators for classification. The classifier designed by these primary features yielded the 90.25% accuracy and performance measures of specificity (0.93) and sensitivity (0.89). The classification performance obtained from classifying the suicidal patients was comparatively greater than that from classifying the remitted patients.

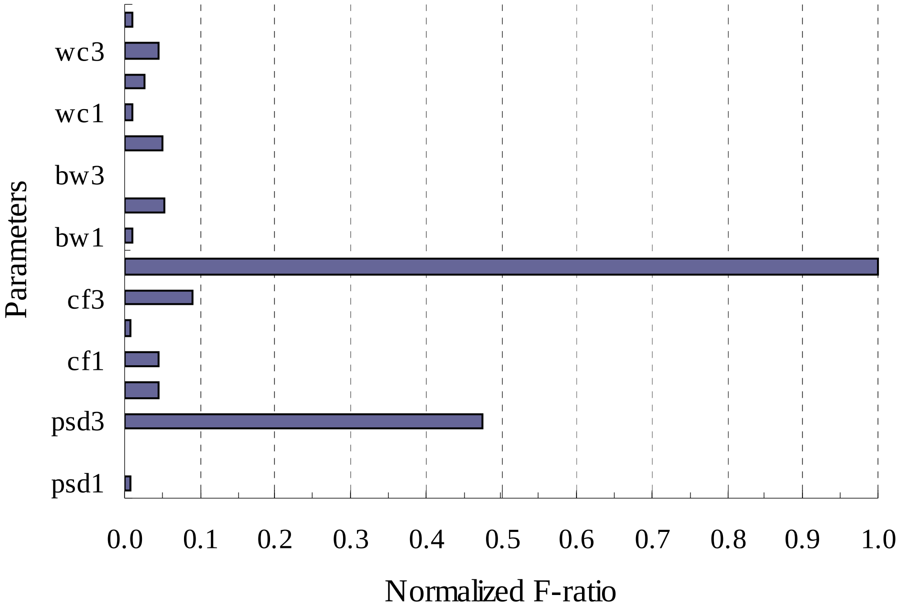


Figure 5.6 Normalized F-ratios determined from reading vocal features of male remitted-suicidal comparison.

Table 5.11 Summary of comparative recognition performances based on pairwise analyses of reading speech.

Pairwise Group	**%Classification**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Depressed/Suicidal	88.50	0.91	0.88	0.90	0.93
Remitted/Depressed	92.00	0.86	0.97	0.98	0.89
Remitted/Suicidal	90.25	0.89	0.93	0.93	0.92

Table 5.12 Summary of primary acoustic features maximizing group separation and recognition performance for reading speech groups.

Pairwise Group	**Primary Feature Set**
Depressed/Suicidal	CF_4, PSD_2, CF_3
Remitted/Depressed	PSD_3, PSD_2, BW_3
Remitted/Suicidal	CF_4, PSD_3, CF_3

5.4 Discussion

5.4.1 Discussion on Comparative Results of Interview Speech Study

Results show that PSD_2, PSD_3, CF_2, CF_3, CF_4, BW_4 and WC_4 have their acoustical properties affected significantly by the psychological state and these acoustical variations of speech possibly represent that state. Multi-parameter classifiers designed by PSD_3, CF_2, CF_3, CF_4, BW_4 and WC_4 yielded the high accurate classification scores from differentiating depressed speech (72.25%) and suicidal speech (81.08%) from remitted speech (Table 5.5). These high accuracies imply that such a primary feature set can delineate the fine boundaries to separate normality of mental state from severely depressive state and even better to separate normal state from suicidal risk state in patients.

The highest classification performance (85.58%) was obtained using CF_4, CF_3 and PSD_2 for classifying depressed patients and suicidal patients. As observed from Table 5.1, the speech features represented by PSD_2, CF_3 and CF_4 have statistical characteristics increased as referred to by the shifting in energy distribution from a lower frequency range below 500 Hz (sub-band #1) toward a higher frequency range (500-1,000 Hz) for the suicidal group with respective to the depressed group. The frequency features of CF_3 and CF_4 in suicidal speech were identified to increase as the severity of mental state increased. In depressed speech, we found a trend of energy shift taking place at a low frequency band below 500 Hz with respective to the distribution of suicidal energy. In remitted speech, the energy spectrum seems to distribute in frequency bands as same as that of suicidal speech, but some different amount of energy in frequency sub-bands lower than 1,000 Hz can be observed between remitted speech and suicidal speech.

The effectiveness of the depressed-suicidal classification was very high on the average observed from the results of three pairwise classifications and; also, it seems that the classifier performed with nearly equal effectiveness in identifying depressed patients and suicidal patients. It illustrated the discriminative properties of primary feature set in classification, which almost evenly distributed for depressed speech and suicidal speech. As summarized in Table 5.5, we can conclude from our experimental results on the correct classification percentages as the follows: Suicidal speech exhibited the highest separation (85.58%) from depressed speech, in term of class discrimination based on acoustical measures, as compared to other pairwise analyses. Remitted speech indicated high separation (81.08%) from suicidal speech and depressed speech revealed moderately high separation (72.25%) from remitted speech, which was the lowest for study of interview speech.

The previous studies have reported formants of speech as significant and consistent features that can correlate with severely depressive state in patients. The first and second frequency formants have been found to change in vocal patterns of depressed speech as compared to those of normal speech. In speech of remitted patients, a reduction in the first formant frequency has been identified, while lower frequency of the second formant has been observed as significant vocal affect in depressed speech [26,36,61]. In recent study of diagnostic speech in male patients [7], it has been demonstrated for comparison of major depressed and high-risk suicidal speech that the first formant frequency in major depressed speech was decreased, but its bandwidth was observed to conversely increase. By these significant changes in frequency and bandwidth, both formant features turned out to be the best discriminators of differentiating depressed patients and suicidal patients.

The recent investigation [7] also found that the distribution of energy in spectra of speech can be the key indicators of vocal changes in depressed speech and suicidal speech. As compared to the suicidal speech, the proportions of energy in depressed speech were increased in the first frequency sub-band (0-500 Hz), but less energy for higher frequencies (above 1,000 Hz). In suicidal speech, the distribution of energy was determined to shift from the 0-500 Hz sub-band to higher frequency sub-bands (above 1,000 Hz). The frequency location of maximum peak in spectrum was also reported as another significant feature that can be used to classify depressed patients and suicidal patients.

As compared to our results, we have found that CF_1 and BW_1 in depressed speech were decreased in frequency and bandwidth when compared to those in suicidal speech. The decreasing trends of the first formant frequency reported in prior study for depressed speech as compared to suicidal speech and that of CF_1 from our study seem to be directionally consistent based on our observation made on their appearances within a frequency range of 0-500 Hz, which corresponds closely to the first neutral formant frequency (500 Hz) as a referring frequency. However, it shows disagreement between the first bandwidth feature obtained from prior work and that from our study. In comparison of depressed and suicidal speech, we obtained narrowing bandwidth (BW_1) as distinct vocal pattern in depressed speech, while the wider bandwidth of the first formant was formerly reported as the significant vocal affect correlated with depressive state. This inconsistency of the changing trend of bandwidth parameter possibly implies that the differences exist in defining the bandwidth estimate of the actual formant in speech and in methodologies to extract that bandwidth parameter. Although our finding was partially supportive and consistent with that reported from prior study, these vocal features were not experimentally determined to be the best discriminating features as summarized in Table 5.6 for the depressed-suicidal comparison. The depressed speech and suicidal speech were both distinctively characterized by CF_3 and CF_4. Based on the discriminant analysis and test of recognition rate using different optimal feature sets to design different classifiers, CF_3 and CF_4 were determined to be the best GMM-based features with the strongest discriminative properties (see Figure 5.1) that distinguished suicidal patients from depressed patients with the most accurate classification (85.58%).

The significant difference in the distributions of energy was determined to exist between depressed speech and suicidal speech. As compared to suicidal speech, the increased energy in the 0-500 Hz band and the reduced energy in the 500-1,000 Hz band were evidently identified for depressed speech. These varying trends of spectral energy in lower frequency bands are consistent with the previously published results of prior study in that patients suffering from severe depression spoke with greater energy in low frequency band (less 500 Hz). In addition, the spectra estimated from suicidal speech illustrated the shifting trend of energy starting from a sub-band #1 toward a sub-band #2 and no further shifts were found in higher frequencies (above 1,000 Hz). Shift in energy of suicidal speech was found to be completely consistent with the results of the prior study for the frequency sub-bands below 1,000 Hz.

As a result of our findings, the frequency and bandwidth parameters representing the vocal tract characteristics and proportion of energy in spectrum reveal the possibility of being the powerful distinguishing features that have long been suggested by prior studies. The objective of this study to reinvestigate these promising acoustical features in their discriminative properties for identification of correct psychological states was met and the encouraging effectiveness and performance of spectral energy and GMM spectral modeling features in distinguishing among diagnostic patient groups should be further investigated on larger sample populations for more statistical consistency.

5.4.2 Discussion on Comparative Results of Reading Speech Study

As summarized in Tables 5.11 and 5.12, the multi-parameter classifier using PSD_2, PSD_3, CF_3, BW_3 and CF_4 performed effectively to classify depressed patient (92%) and suicidal patients (90.25%) from remitted patients, and approximately 2% difference in classification accuracy was obtained between the discriminant analyses of remitted versus depressed speech and remitted versus suicidal speech.

As compared to the results of the pairwise classification analyses performing on the interview speech samples presented in Table 5.5, the much more improvements can clearly be noticed as comparisons with the results of reading speech classifications shown in Table 5.11. The difference in correct classification score (20%) was obtained to be highly significant ($p<0.001$) between classifying interview speech (72.25%) and classifying reading speech (92%) for the remitted-depressed comparison.

In the remitted-suicidal comparison, the significant difference in classification accuracy ($p<0.001$) was found between interview speech (81%) and reading speech

(90%). These noticeable improvement in classification suggested that the vocal features extracted from the reading speech samples provided much more acoustically discriminative properties than those of the interview speech samples and these properties helped improve group separation in classification analyses between diagnostic groups of male patients.

As plotted in Figures 5.1 and 5.4 for the F-ratios in interview vocal features and reading vocal features for the depressed-suicidal comparison, we observed that the results of pairwise F-ratio analyses using the reading speech (Figure 5.4) exhibited a larger number of features with high separabilities (F-ratios>0.5) and the accurate classification score obtained by using CF_4, CF_3 and PSD_2 as optimal features in discriminant analysis improved from 85.58% (Table 5.5) to 88.50% (Table 5.11), which was significantly improved ($p<0.05$) to be approximately 3% for classification of reading speech. This finding on the relationship between the number of high F-ratio features and the improvement of classification score is consistent with the former discriminating results of the female reading speech analyses as presented in Chapter IV. The reading speech samples provide more improvement of classification. It is strongly supportive of high reliable F-ratio as very powerful statistical measure for class separation that can help predict the success of classification.

In this study of reading speech, the primary feature set consisting of CF_4, CF_3 and PSD_2 was statistically determined to be the same as that found in study of interview speech for the same comparison of depressed and suicidal speech. The relating trends of individual CF_4, PSD_2 and CF_3 consistently showed the similarity to those of the interview speech study, except for the means of individual features that were slightly different. The distribution of energy in suicidal speech was more similar to that of remitted speech than depressed speech. Our results on studies both interview speech and reading speech are similarly consistent with the results of prior study on the distributing trends of sub-band energy in spectrum.

Acoustical properties of vocal measures appeared to change significantly with the severity of the psychological states. Our results provided the solid believe of possibility in that the vocal features derived from the percentages of total spectral energy based on the classical PSD method and the GMM-based spectral modeling of the vocal tract response are capable of indicating the different identities of the psychologically disordered speech. They were successfully employed in the comparative discriminant analysis of identifying depressed and suicidal patients.

Based on the results of separate experiments for female and male, the gender factor affects differently in acoustics properties of speech samples and the classification performance as well. The primary feature sets between different genders were found to comprise of different features, as summarized in Tables 5.6 and 5.12 for interview and reading speech in male study, and in Tables 4.5 and 4.12 for prior studies of female interview and reading speech. The separate studies of speech acoustics in different genders are suggested for further investigation of vocal cues of suicidal risk assessment.

Our results provided the solid believe that the studied acoustical features are very promising acoustical indicators of the potential of committing suicide in patients and they can assist the diagnostic tasks of psychological disorders.

CHAPTER VI

SUMMARY AND CONCLUSIONS

Vocal affect identified as possible cues for clinicians to use in diagnosing the syndrome underlying a person's emotional state appears to be the unique perceptual indicator of emotional expression in speech and its relation to the overall state of a speaker. It is very important for a psychologist to be able to accurately assess the risk of persons killing themselves. This study serves as another step in the development of an algorithm capable of assisting psychologists, psychiatrists and clinicians in objectively determining whether a person is suicidal. The nonlinguistic content of a patient's speech provides the information relative to symptoms of psychomotor disturbances associated with affective disorders. By progressively investigating this nonlinguistic information in diagnostic speech samples, a measure of objectivity is reached.

This dissertation investigates acoustical properties of speech as promising vocal features that represent the severity of psychological state changed with depression and suicidal risk. Analyses of acoustic features and between-group discriminations were performed on diagnostic populations to determine if the acoustical properties of the vocal features exhibit the differences in their qualitative measures relative to the severity of the psychiatric disorders and if these differences can be used to distinguish the mental condition of individual subjects.

The first manuscript in Chapter III is titled "Objective Estimation of Suicidal Risk Using Vocal Output Characteristics". The proportions of energy in 500 Hz bands within

0-2,000 Hz spectra of male speech samples collected from the interview and reading audio recordings were estimated to represent individual groups of depressed, high-risk suicidal and remitted patients. The comparative reading speech classification of suicidal and depressed subjects yielded the 82% correct rate as an acceptable level of accuracy for the performance of the designed classifier using the PSD features. The highest correct classification (94%) was determined for the depressed-remitted comparison in the discriminant analysis of interview speech. The vocal features derived from reading speech samples can possibly be used to classify diagnostic patient groups as relatively effective as those of interview speech samples.

The second manuscript in Chapter IV is titled “Direct Acoustic Feature Extraction Using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Speech in Females”. In this study, the acoustical features representing the vocal tract spectral characteristics of female speech samples derived from a new proposed method of GMM-based spectral modeling were analyzed and then combined with those derived from the PSD-based estimation to represent interview and reading speech samples for each categorized groups, depressed, suicidal and remitted females.

In the discriminant analyses of interview speech, depressed and suicidal females appear to be effectively classified from remitted females on the basis of PSD_1, PSD_2, PSD_3 and WC_3. The depressed-suicidal discriminant analysis yielded the greatest classification accuracy of 86% by incorporating of PSD_1, PSD_2, CF_2 and BW_2 in cross validation for acoustical analysis of interview speech. As compared to suicidal patients, depressed patients exhibited reduced PSD_1, CF_2 and BW_2 which have been reported previously as significant characteristic features responding to depressive state in patients.

The increase in PSD_2 was determined as a result of a definite trend of energy shift from lower to higher frequencies (above 500 Hz) for depressed speech with respective to that of suicidal speech. The results suggest that the depressed and suicidal speech differ significantly in the terms of PSD_1, PSD_2, CF_2 and BW_2 and theses features were the powerful discriminators of degree of mental state.

In the discriminative analyses of reading speech, depressed and suicidal females were effectively distinguished from remitted females on the basis of PSD_1, PSD_3, CF_2, BW_2 and WC_2. The depressed-suicidal classification yielded high accuracy of 90.33% using PSD_1, PSD_2 and CF_2 in cross validation. In this study of reading speech samples, depressed speech exhibited similar trends of PSD_1, PSD_2 and CF_2 to those of interview speech of the same depressive patients, as compared to suicidal patients. The significant differences in vocal characteristics of these features helped improve the classifier performance, as compared with classification analyses of interview speech based on classifying the same feature samples.

The third manuscript in Chapter V is titled “Direct Acoustic Feature Extraction Using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Speech in Men”. The PSD-based and GMM-based features derived from the interview and reading speech samples of individual diagnostic groups of depressed, suicidal and remitted men. In the discriminant analyses of interview speech, depressed and suicidal men appear to be effectively classified from remitted females using PSD_3, CF_2, CF_3, CF_4, BW_4 and WC_4 as discriminators. The depressed-suicidal classification of the interview speech yielded the best accurate score of 86% by employing CF_4, CF_3 and PSD_2 in cross validation of quadratic discriminative analysis. As compared to suicidal speech, reduced PSD_2 and increased PSD_1 were determined for depressed speech, which were consistent with the previously published results of prior study [7] in that the depressed patients spoke with greater energy in lower frequencies (0-500 Hz). While the suicidal speech consistently revealed the distribution of energy shifting from lower to higher frequencies, but less than 1,000 Hz. The experimental results of testing recognition rates demonstrated CF_4, CF_3 and PSD_2 as the strongest discriminating features in differentiating depressed speech and suicidal speech.

In the discriminative analyses of reading speech for the depressed-suicidal comparison, more accurate classification score (88.50%) was obtained based on feature samples of CF_4, CF_3 and PSD_2 which were the same as those of interview speech and their changing trends were consistently similar as well. The energy distribution of suicidal speech was more similar to that of remitted speech than depressed speech. The improvement of classification performance can obviously be indicated for employing the reading speech samples in analyses rather than the interview speech samples.

This dissertation provides the promising methodology of acoustic analysis of psychiatric speech for the future of research in developing the computer-based algorithm capable of assisting clinicians in diagnosis of depression and suicidal risk. More robust and reliable extraction approach for the acoustic features correlating with symptom of emotional disorders should be the main task of next step for this ongoing research to gain more convincing evidence to prove the hypothesis in that specific acoustic parameters in speech can use to identify mental health conditions. The encouraging vocal feature, spectral entropy, has recently been proposed in the literatures of speech processing [62,63,64] and demonstrated improvement in the recognition accuracy and robustness against additive noise. It measures the power spectral flatness of the spectrum of speech. This additional type of feature, in term of spectral based feature, may better characterize the psychologically disordered speech and increase the effectiveness of classifier.

REFERENCES

[1] NIMH (National Institute of Mental Health) website, http://www. nimh.nih.gov/publicat/harmsway.cfm

[2] Conwell, Y. and Brent, D., "Suicide and aging I: patterns of psychiatric diagnosis", International Psychogeriatrics, 7(2):149-164, 1995.

[3] Brent, D.A., Correlates of the medical lethality of suicidal attempts in children and adolescents, *J Am Acad Child Adolesc Psychiatry*, 1987, 26, pp. 437-438

[4] Guze, S.B., Robins, E., Suicide and primary affective disorders. *British Journal of Psychiatry*, 1970, 117, pp. 437-438

[5] Silverman, S.E., "Vocal parameters as predictors of nearterm suicidal risk", U.S. Patent 5 148 483, Sept. 1992.

[6] Campbell, L., “Statistical Characteristics of fundamental frequency distributions in the speech of suicidal patients”, Masters Thesis, Vanderbilt University, 1995.

[7] France, D.J., "Acoustical properties of speech as indicators of depression and suicidal risk", Ph.D. Thesis, Vanderbilt University, August, 1997.

[8] Yingthawornsuk, T., Kaymaz Keskinpala, H., France, D.J., Wilkes, D.M., Shiavi, R.G., Salomon, R.M., "Objective Estimation of Suicidal Risk using Vocal Output Characteristics”, Interspeech 2006, Int. Conf. on Spoken language Processing, pp. 649-652, 2006.

[9] Ozdas, A., Shiavi, R.G., Wilkes, D.M., Silverman, M.K. and Silverman, S.E., "Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment", Methods of Information in Medicine, vol. 43, pp. 36-38, 2004.

[10] Moore, E., Clements, M., Peifer, J. and Weisser, L., “Comparing Objective Feature Statistics of Speech for Classifying Clinical Depression”, IEEE Int. Conf. (EMBS), pp. 17-20, 2004.

[11] Moore, E., Clements, M., Peifer, J. and Weisser, L., “Analysis of Prosodic Variation in Speech for Clinical Depression”, IEEE Int. Conf. (EMBC), pp. 2925-2928, 2003.

[12] Moore, E., Clements, M., Peifer, J. and Weisser, L., “Investigating the Role of Glottal Features in Classifying Clinical Depression”, IEEE Int. Conf. (EMBC), pp. 2849-2852, 2003.

[13] Welch, P.D. “The use of the fast Fourier transform for the estimation of Power Spectra: A method based on time averaging over short modified periodograms”, IEEE Trans. Audio Electroacoust., vol. AU-15, no. June, pp. 70-73, 1967.

[14] Dempster, A.P., Laird, N.M. and Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm", J. Royal Statistical Society Series B, 39:1-38, 1977.

[15] Noll, A.M., “Cepstrum Pitch Determination”, J. Acoust. Soc. Am., vol. 41, pp. 293-309, Feb 1967.

[16] Furui, S., "Cepstral analysis technique for automatic speaker verification", IEEE Trans. Acoust. Speech Sig. Proc., vol. 29, no. 2, pp. 254-272, 1981.

[17] Yingthawornsuk, T., Kaymaz Keskinpala, H., Wilkes, D.M., Shiavi, R.G., Salomon, R.M., "Direct Acoustic Feature Using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Speech”, submitted to Interspeech 2007 – Eurospeech, 2007.

[18] O’Shaughnessy, D., *Speech Communication: Human and Machine*, Addison-Wesley, Massachusetts, 1987.

[19] Hollien, H., The Acoustics of Crime: The new science of forensic phonetics, Plenum Press, New York, 1990.

[20] Flanagan, J.L., "*Speech Analysis, Synthesis, and Perception*", New York, NY: Springer Verlag, 1983.

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

[22] Fant, C., "*Acoustic Theory of Speech Production*", The Hague: Mouton, 1970

[23] Atal, B.S., Hanauer, S.L., "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Am., vol 50, pp. 637-655, 1971.

[24] Wong, D.Y., Markel, J.D., and Gray, A.H., “Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform”, IEEE Trans. Acoust. Speech Sig. Proc., vol. 27, no. 4, pp. 350-355, 1979.

[25] Kent, R., Read, C., The Acoustic Analysis of Speech, San Diego, CA.: Singular Plublishing Group, pp. 14-16, 1992.

[26] Scherer, K., Zei, B, "Vocal Indicators of Affective Disorders", Psychotherapy and Psychosomatic, vol. 49, pp. 179-186, 1988.

[27] Gravell, R., France, J., Speech and Communication Problems in Psychiatry, San Diego, CA.: Singular Publishing Group, 1992.

[28] Eldred, S.H., Price, D.B., "A linguistic evaluation of feeling states in psychotherapy", Psychiatry, vol. 21, pp. 115-121.

[29] Roessler, R., Lester, J., "Voice patterns in anxiety", In W.E. Fann, A.D. Pokorny, I. Koracau, R. Williams (Eds.), Phenomenology and treatment of anxiety. New York:Spectrum, 1979

[30] Newman, S., Mather, V., "Analysis of spoken language of patients with affective disorders", American Journal of Psychiatry, vol. 94, pp. 912-942, 1938.

[31] Moses, P.J., The voice of Neurosis, New York: Grune & Stratton, 1954

[32] Hargreaves, W.A., Starkweather, J.A., "Voice quality changes in depression", Language and Speech, vol. 7, pp. 84-88, 1964.

[33] Hargreaves, W.A., Starkweather, J.A. and Blacker, K.H., "Voice quality in depression", Journal of Abnormal Psychology, vol. 70, pp. 218-220, 1965.

[34] 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.

[35] Ostwald, P.F., *Soundmaking: The Acoustic Communication of Emotion*. Charles C. Thomas, Springfield: Illinois, 1963.

[36] Tolkmitt, F., Helfrich, H., Standke, R., Scherer, K., "Vocal indicators of psychiatric treatment effects in depressives and schizophrenics", Journal of Communication Disorders, vol. 15, pp. 209-222, 1982.

[37] Kuny, S., Stassen, H.H., "Speaking behavior and voice sound characteristics in depressive patients during recovery", J. Psychiat. Res., vol. 27, no.3, pp. 289-307, 1993.

[38] Scherer, K., Vocal Affect Expression: A review and a model for future research, Psychological Bulletin, vol. 99, pp. 143-165, 1986.

[39] Ozdas, A., “Analysis of Paralinguistic Properties of Speech for Near-Term Suicidal Risk Assessment”, PhD. Thesis, Vanderbilt University, 2000

[40] France, D.J., Shiavi, R.G., Silverman, S., Silverman, M. and Wilkes D.M., "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Trans. Biomed. Eng., 47(7):829-837, 2000.

[41] Ozdas, A., Shiavi, R.G., Wilkes, D.M., Silverman, M.K. and Silverman, S.E., "Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk", IEEE Trans. Biomed. Eng., vol. 51, pp. 1530-1540, 2004.

[42] Scherer, K., Vocal correlates of emotional arousal and affective disturbance, in H. Wagner and A. Manstead, eds., Handbook of social psychophysiology, Wiley, New York, 1989.

[43] Darby, J.K., Speech and voice studies in psychiatric populations, in J. K. Darby, ed., Speech Evaluation in Psychiatry, Grune & Stratton, Inc., New York, 1981.

[44] Scherer, K.R., Speech and emotional states, in J. K. Darby, ed., Speech Evaluation in Psychiatry, Grune and Stratton, Inc., New York, 1981.

[45] Fairbanks, G., Voice and Articulation Drillbook, Harper & Row, New York, 1960.

[46] Martinez, W.L., Martinez, A.R., "*Computational Statistics Handbook with MATLAB* ", Chapman & Hall/CRC, 2002.

[47] O'Shaughnessy, D., "Linear predictive coding", IEEE Potential, pp. 29-32, 1988.

[48] Tierney, J., "A study of LPC analysis of speech in additive noise", IEEE Trans. Acoust. Speech Sig. Proc., vol. 28, no. 2, pp. 389-397, 1980.

[49] Zolfaghari, P. and Robinson T., "Formant Analysis Using Mixtures of Gaussians", Interspeech 1996, Int. Conf. on Spoken language Processing, 1996.

[50] Atal, B.S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J Acoust. Soc. Am., vol. 55, no. 6, pp. 1304-1312, 1974.

[51] Reynolds, D.A., Rose, R.C., "Robust text-independent speaker identification using gaussian mixture speaker models", IEEE Transactions on Speech and Audio Processing", vol. 3, no. 1, pp.72-83, 1995.

[52] Schafer R. and Markel, J., Speech Analysis, IEEE Press, 1979.

[53] Feder, M., Et.al, "Parameter estimation of superimposed signals using the EM algorithm", IEEE Trans. Acoust. Speech Sig. Proc., vol. 36, no. 4, pp 477-489, 1988.

[54] Lim, J.S., "Spectral Root Homomorphic Deconvolution System", IEEE Trans. Acoust. Speech Sig. Proc., vol. 27(3), pp 223-233, 1979.

[55] Chandler, D., "*Introduction to Modern Statistical Mechanics*", Oxford, 1987.

[56] Rudin, W., "*Real and Complex Analysis*", McGraw-Hill, 1987.

[57] Moon, T.K., "The Expectation-Maximization Algorithm", IEEE Signal Processing Magazine, vol. 13(6), pp. 47-60, 1996.

[58] Parsons, T., Voice and Speech Processing, McGraw-Hill, 1987.

[59] Pruzansky, S., "Talker recognition procedure based on analysis of variance", J. Acoust. Soc. Am., vol. 36, pp. 2041-2047, 1964.

[60] Wolf, J.J., "Efficient acoustic parameters for speaker recognition", J. Acoust. Soc. Am., vol. 51, pp. 2044-2056, 1972.

[61] Whitman, E., Flicker, D., A potential new measurement of emotional state: A preliminary report. *Newark Beth-Israel Hospital*, 1966, 17, pp. 167-172.

[62] Toh, A.M., Togneri, R., Nordholm, S., “Spectral Entropy As Speech Features For Speech Recognition”, 2007.

[63] Misra, H., Ikbal, S., Bourlard, H. and Hermansky, H., “Spectral entropy based feature for robust asr”, in *Proc. ICASSP*, May 2004, pp. 193-196.

[64] Toh, A.M., Togneri, R., Nordholm, S., “Investigation of robust features for speech recognition in hostile environments”, in *Proc. APCC*, May 2005.

[65] Godino-Llorente J. I., Gomez-Vilda P., and Blanco-Velasco M., "Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short Term Cepstral Parameters", IEEE Transaction on Biomedical Engineering, 53(10):1943-1953, 2006.

[66] Fremouw, W.J., de Perczel, M., Ellis,T.E., “Suicide Risk assessment and response guidelines”, Pergamon Press, 1990.

[67] Stassen, H.H, “Affective State and Voice: The Specific Properties of Overtone Distributions”, Methods of Information in Medicine, vol. 30, no.1, pp. 44-52, 1991.

[68] Walbott, H.G., “Vocal Behavior and Psychopathology”, Pharmacopsychiatry vol. 22, pp. 13-16, 1985.

[69] Par Government System Corporation, OLPARS user’s Manual, version 8.01, 1996.

[70] SYSTAT Software, Inc., SYSTAT user’s Manual, version 11, 2004.

[71] Foley, D., Considerations of sample and feature size. IEEE Transaction on Information Theory, vol. IT-18, no. 5, September, 1972

[72] Johnson R. and Wichern, D., Applied Multivariate Statistical Analysis, Third edition, Prentice Hall, Englewood Cliffs, NJ, 1992.

[73] Lachenbruch, P. and Mickey, M., Estimation of error rates in discriminant analysis, *Technometrics*, 10, no. 1, pp. 1-11, 1968.

[74] Mclachlan, G.J. and Basford, K.E., *Mixture Models*, Marcel Dekker, Inc., New York, 1988.

[75] Bogert, B.P., Healy, M.J. and Tukey, J.W., “The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking”, in *Time Series Analysis*, M. Rosenblatt, Ed. New York: Wiley, 1963, Chap. 15, pp. 209-243.

[76] Borden, G.J., Harris, K.S. and Raphael, L.J., Speech Science Primer: Physiology, Acoustics, and Perception of Speech, Lippincott Williams & Wilkins, 2002, pp. 81-150.

[77] Rosner, B., Fundamentals of Biostatistics, Sixth Ed., Thomson Brooks/Cole, 2006, pp. 385-463.