International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 31061-31090
© Research India Publications
http://www.ripublication.com

# Timing Patterns of Speech As Potential Indicators of Near-Term Suicide Risk

**Nik NurWahidah N. H.*[1], Mitchell D. Wilkes[2], Ronald M. Salomon[3]**

*[1,2] Department of Electrical Engineering,*
*Vanderbilt University, Nashville, TN 37235 USA*
*[3] Department of Psychiatry and Behavioral Sciences,*
*Northwestern University, Chicago, IL, 60611, USA.*
*Email: Nik NurWahidah N. H.* - niknurwahidah@iium.edu.my;*
*Wilkes D. M. – mitch.wilkes@vanderbilt.edu; Salomon R. M. –rsalomon@nmff.org*
**Corresponding author*

## Abstract

An important clinical problem remains the differentiation between non-suicidal and more lethal episodes of depression. In an effort to find a reliable method that could assist clinicians in risk assessment, information in the speech signal has been found to contain characteristic changes associated with high risk suicidal states. This paper addresses the questions of (1) Does information contain in the speech timing-based measures able to discriminate between high risk suicidal (HR) speech from the depressed (DP) speech. (2) How well do speech features, specifically the timing-based measures can predict the ratings from a well-known medical diagnostic tool known as the Hamilton Depression Rating Scale (HAMD). In the first study, using the leave-one-out procedure as a means to measure a classifier performance for all-data classification revealed a single speech timing-based measure to be a significant discriminator with 79% overall correct leave-one-out classification in male (MR) and female (FR) reading speech from Database A. For male patients, using the trained features on Database A and testing on Database B1 successfully demonstrated up to 100% detection of high risk speech in Database B1. In the second study, the acoustic measurements were shown to effectively predict the HAMD score with less than 5% mean absolute error using only combinations from the timing-based measures and eliminating all spectrum-based measures.

**Key Words:** Speech pauses, voiced, silence, classification, prediction, regression, suicidal, depression

## 1.0 INTRODUCTION

Suicide continues to be a major concern to public health worldwide. In the United States, the current available summary by the National Center for Health Statistics reported an age-adjusted increase of 2.4 percent from 2008 to 2009 followed by an increase in 2010 by another 2.5% [1]. The 2009 list of cause of death revealed that suicide ranks in 10th with 36,909 suicides for all age groups, ranks 3rd for the age group of 15-24 with 4,371 suicides [2]. A surge in the military suicide rate with an increase of 18 percent in the year 2012 compared to the statistic reported at the first half of the previous year [3].

Despite decades of research, accurate prediction of suicide and imminent suicide attempts still remains elusive. Clinical interview tools remain the standard of care, but are often highly sensitive and commonly reveal “high” risk in Emergency Department evaluations. Current suicide risk assessment requires clinicians to use specific interviewing approaches, sometimes relying on their intuition, and to deploy specialized skills they develop through formal education, clinical training and clinical experience. This traditional method is time- and energy-intensive, and still frequently misses true positives. A detection of biometric characteristics known to be associated with imminent suicidal risk at an early stage may prompt the clinician to propose and promote a more intensive treatment plan. No objective tool is available to assess (or to assist with) the false negative finding where a person at risk denies suicidal thoughts [4, 5]. Negative screenings can bring the clinician to believe that persons who are actually at imminent risk of committing suicide are experiencing a less severe (often depressive) disorder. Misdiagnosis in such situations may result in an untoward, unfortunate outcome. Even for specialists, and especially for non-specialists, an objective metric might signal a critical need for more extensive interviewing and other precautions.

Besides performing clinical based assessment, diagnosis of psychological disorders, particularly in depression, has also been examined by means of visual based expression and speech-based measurements. Results from the analyses performed in [33]-[37] suggest that patients experiencing psychological disturbances exhibit particular facial expressions, behavior patterns and physical movements. Another study in [29] has explored multimodal fusion for detecting depression by combining visual and verbal cues based on the hypothesis that the information from individual cues complements each other thus the multiple feature fusion will improve the performance of the system.

A number of studies have also focused on using only the speech based measurements to reflect the psychological states [6]-[29].Among the distinctive speech patterns associated with depression are decreases in intonation, stress, loudness, inflection, intensity and speech rate, sluggishness in articulation, monotonous, and lack in vitality [12]-[14]. These characteristic correlates with the changes occurring in the speech production mechanism by affecting the respiratory, laryngeal, resonance and articulatory subsystem that in turn are encoded in the acoustical signal. Studies on vocal characteristics in terms of their relationship to depression include speech prosody (e.g., pitch, energy, and speaking rate) [6, 13], [14]-[16], spectral features (e.g., power spectral density, formants and their associated bandwidth) [6, 8, 9], glottal features [7, 17] and MFCCs [7, 9]. This study is built upon literatures in the field of speech based measurements in identifying between voiced production in depressed speech with and without an elevated risk of suicidal behavior.

Patients undergoing major depressive disorder often experience psychomotor retardation and cognitive disturbances [13], [30]-[32]. One of the effects observed is abnormalities in verbal productivity such as slower speech rate and increase pause time in between responses. Psychomotor retardation occurs due to the condition in which the brain has difficulty in communicating with the rest of the body, thus increasing the response time and radically reducing muscle activity. The disturbances of the interactions between numbers of neuromuscular systems thus affect the motor execution and production of speech. On the other hand, cognitive function relates to impairment in attention, information processing, working memory and decision-making processes. Cognitive impairment might also affect the number and duration of speech pauses as opposed to articulation due to hesitancy and reduction in attentiveness. These effects of psychomotor retardation and cognitive disturbances are related to the vocal characteristics known as prosody and speech rate. According to Monrad-Krohn [18], the definition of prosody consists of the normal variation of pitch, stress and rhythm which includes silent intervals of pauses. Alpert [15] on the other hand separated speech productivity and pausing under the term fluency and defined prosody as emphasis and inflection. Speech rate comprises a combination of phonation length (voiced), frequency of short pauses and the duration of pauses.

This paper attempts to address two research questions. The first question asks whether the information contain in the speech timing-based measures are able to discriminate high risk suicidal (HR) speech from depressed (DP) speech. The study reported herein focuses on the use of certain features related to the rhythm, fluency of speech and speech rate in an attempt to capture information related to voiced segments (or phonation) and silent pauses. We introduce a new approach to representing the pauses and vocalization using Markov model and also constructing a histogram using the voiced, unvoiced and silent sections in a speech signal. We refer to these features as Transition Parameters and Interval pdf, respectively. Secondly, we try to address the question of how well do speech features, specifically the timing-based measures predict the ratings from a well-known medical diagnostic tool known as the Hamilton Depression Rating Scale (HAMD).Previous research mainly investigated the correlation between the clinical ratings and speech measurements. Correlation merely describes the relationship between two variables whereas regression predicts the value of a dependent variable (i.e., clinical scores) using one or more measurements of independent variables (i.e., acoustic features).

This distinction and prediction would be highly useful for use in real-world applications, especially when using a method that is unobtrusive to patients and practicable for the use of researchers and clinicians. Our focus is to address concerns from the clinical side of the problem and to find viable acoustic features in speech that have good reliability and can make the clinically critical separations between DP and HR.

### *Suicide assessment by the Hamilton Depression Rating Scale (HAMD)*

A common interview scale and administered diagnostic tool to measure the severity of depression in an inpatient population is the Hamilton Depression Rating Scale (HAMD). The HAMD assessment has also been considered as the primary standard for determining suicidal risk for this database. It contains 17-items questionnaires including one item on suicidal thoughts with rating scales that can be evaluated only by trained clinicians. Clinicians rely on their intuitions during evaluation and determining the ratings for the provided questionnaires. Generally accepted opinions by clinicians on interpretation of the total HAMD scores is that score between 0 to 7 shows no presence of depression, 8 to 13 indicates mild depression, 14 to 18 indicates moderate depression, 19 to 22 indicates severe depression and score over 22 indicates very severe depression [43]. For a single suicide item, patients scoring 2 or higher were found to be 4.9 times more likely to die by suicide [44].

## 2.0 PREVIOUS WORK

Investigations on the acoustic features for identifying depression and imminent suicidal risk are small in number and often revolve around spectrum-based measures of speech. France [6] examined speech features that are characterized by the long-term Fundamental Frequency statistics (mean, variance, skewness and kurtosis), Amplitude Modulation, Formants (including bandwidth and ratios) and Power Distribution. Ozdas [7] employed the low order Mel-Frequency Cepstral Coefficient (MFCC), small cycle-to-cycle variations of fundamental frequency known as voice jitter, and glottal flow spectral slope as discriminating features among the near-term suicidal, depressed and remitted groups. Yingthawornsuk [8] extracted features based on the Power Spectral Density (PSD) and Gaussian mixture model (GMM) based spectral modeling of the vocal tract which contains information on spectral pattern (intensity, responding frequency and bandwidth). Keskinpala [9] proposed an optimization study of multiple MFCC coefficients and performed an extensive study on different numbers, ranges and edges of the spectral energy bands.

In the effort to address the first research question, we initially identified that the summation of pauses and the summation of vocalizations in previous studies were collected manually. The work reported in [28, 29] extracted the switching pauses (silence between turns) by manually transcribing the recordings and forced-aligning the speech in order to obtain start time, stop time and utterance. Other methods include recording from a non-spontaneous speech where a patient counts from 1 to 10 [20], readings of standardized text [21, 25] such as the “grandfather passage” and collecting pauses that occurred in between a series of questions and answers during an interview with the clinician [15, 22]. The outcomes were consistent from one study to another where they reported that during the period of improvement, the patients exhibited a decrease in pause time and displayed no significant changes in phonation time. Although the use of pauses within counting and text reading revealed a positive relationship with depression, the effect of shorter pause time after improvement might also be connected to the practice effects from repeated events of measurements. Thus according to [13], it should be considered with some caution. On the other hand, a constant length of pause time that was observed in the control and healthy patients throughout a period of improvement might suggest that pauses in speech are independent of the practice effect [20].

Investigation on the correlation between the pauses in speech and the clinical rating evaluations has recently attracted the interest of researchers in this field. The method of Pearson product-moment [24] and Spearman correlation [26] coefficients were adapted in these studies. According to a study in [23], speech pause time was shown to be significantly correlated to the Retardation Rating Scale for Depression (ERD) as opposed to the HAMD. On the contrary, [24] reported moderate correlation and [13], [25]-[27] reported significant correlation between speech features that are related to pauses and the HAMD scores. Among these highly correlated features are the total recording duration, total pause time, variability of pauses, vocalization to pause ratio, speaking rate and minimal fundamental frequency. However, [13] only demonstrated the correlation for their female subjects but not on the male subjects which was most likely due to small number of samples. A study performed by [26] investigated acoustic features within the phonemes of speech signals and their relationship with individual symptom sub-topic ratings of each 17 HAMD score. The paper claimed that changes in speech patterns correlate with different HAMD symptom sub-topic ratings. In [28], they studied naïve listeners with no experience in making clinical judgment to predict the participant's and interviewer's HAMD ratings. This method reported moderate predictability of the HAMD ratings.

## 3.0 DATABASE COLLECTION AND PRE-PROCESSING

All recording sessions were conducted at the Vanderbilt Emergency Department or Psychiatric Hospital with patient's documented informed consent. Patients who volunteered were made aware of the aim of the study with assurance of maximum identity protection procedures. Patients under the influence of alcohol, toxicity or experiencing respiratory problems such as shortness of breath were excluded. All recordings were made in a standard, empty psychiatric room without the benefit of soundproof or acoustically ideal environment, mimicking the real-world clinical environment.Group assignment was made according to assessment made by experienced clinicians using the Hamilton Depression Rating Scale (HAMD), Beck Depression Inventory (BDI-II), MINI International Neuropsychiatric Interview and Pierce Suicide Intent Scale (SIS) [38]. Patients were interviewed by trained clinicians and asked to read from a standardized "rainbow passage" (known as reading speech) which contains every sound in the English language and is considered to be phonetically balanced with the ratios of assorted phonemes similar to the ones in normal speech [39]. For this analysis, only the reading recordings were used.

Two types of databases were collected for this study. All speech samples were digitized using a 32-bit analog to digital converter at 44.1 kHz sampling rate. Table 1 shows the information regarding the two databases. In the first database (Database A), recordings were collected once per patient and each recording was categorized as either high risk suicidal or depressed. Audio acquisitions were made using a high-quality Audix SCX-one cardioid microphone with a frequency response of 40Hz to

20kHz, Sony VAIO laptop with Pentium IV 2GHz CPU 512 Mb memory, Windows XP, a Digital Audio MBox for digital audio interface and recording software PROTools LE for the digital audio editor.

**TABLE 1:** Information on Database A and B1

| **Database A** | **Male** | | **Female** | |
|---|---|---|---|---|
| Total number of patients | HR | DP | HR | DP |
| | 7 | 12 | 10 | 18 |
| Total number of patients with HAMD score | 9 | | 14 | |
| **Database B1** | **Male** | | **Female** | |
| Total number of HR patients | 7 | | 12 | |

**HR - High Risk Suicidal*DP - Depressed*

In the second database (Database B1), entry criteria restricted inclusion to patients who were labeled as high risk suicidal. Audio acquisitions for Database B1 were made using a portable high-quality field recorder, a TASCAM DR-1, with a frequency response of 40Hz to 20kHz, Samsung Q40 laptop with Intel Core i5 2.4GHz 4G memory and Windows 7.

In the preprocessing stage, recordings were edited using a free audio digital editor called Audacity 2.0.1 to remove any identifying information, to preserve patients' privacy. Undesirable sounds such as the interviewer's voice, voices other than the patient, sneezing, coughing and door slams were removed from the de-identified recordings.

The HAMD scores database was only available for nine male patients and 14 female patients. The regression analysis allows for these scores to be gathered from a pooled category of high risk suicidal, depressed, ideation and remitted.

## 4.0 METHODOLOGY

Speech signals comprise a mixture of voiced, unvoiced, and silence intervals. Voiced, unvoiced and silence speech samples can be estimated by segmenting the sampled signals according to their energy values. Voiced speech samples exhibit the quasi-stationary behavior and are composed of low frequency characteristics. On the other hand, unvoiced speech samples exhibit noise-like behavior and contain higher frequencies. A voiced/unvoiced/silence decision was made for each frame based on the method in [7]. The analysis was then divided into two categories called classification analysis and regression analysis.

### 4.1 Classification Analysis

#### 4.1.1 Feature Extraction

#### A. Transition Parameters

A sampled signal contains a combination of voiced, unvoiced and silence frames that were represented as three different states labeled 1, 2, and 3 respectively. The words spoken in all recordings were the same (reading the “rainbow passage”) but the variation occurs in the timing pattern of the speech. The idea is to capture the variations in the form of transition from one state to another. These states can interchange with each other or perhaps return to the same state according to a set of probabilities that pertains to the states. The probabilities were estimated with a method of an observable discrete-time Markov process [40] implemented using the statistical toolbox available in MATLAB. The state and sequence are initially known with the emission probabilities set to be the matrix identity. One of the output parameters given is the estimated transition matrix ($T$) which in this case is a three-by-three matrix where$t_{ij} = \Pr(X_{k+1} = j | X_k = i)\ fori = 1,2,3\ andj = 1,2,3$. Each row $i$ is a conditional probability given that you are in state $i$at time k and column $j$ is the possible next state at time k+1. For example, $t_{13}$ denotes the conditional probability of going from a voiced frame to a silent frame (Voiced-to-Silence).

$$TransitionMatrix, T = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{bmatrix}$$

The nine features were concatenated into a row vector representing each patient as $\{t_{11}, t_{12}, t_{13}, t_{21}, t_{22}, t_{23}, t_{31}, t_{32}, t_{33}\}$.

### B. Interval Length Probability Density Function

The idea behind this feature is to observe any variations in the distribution of the number of frames per interval for each collection of voiced, unvoiced and silence intervals. Some questions that arise regarding the distribution of intervals are whether a person in high risk suicidal or depressed holds their vowels longer? Do they slur their speech and end up producing longer unvoiced segments or do they have longer silences? The shape of the pdf describes where the variability mostly occurred. The pdf is estimated by counting the number of occurrence for a consecutive number of 40ms frames per interval, that belongs to voiced, unvoiced or silence, that are mixed within a sampled signal. The implementation procedure to obtain the Interval Length pdf for a sampled signal is as follows:

1) For the Interval Length pdf of voiced intervals, find all the voiced intervals in the signal. Figure 1 shows a voiced interval (denoted by 1’s) of the length three
2) Count all the intervals of length one (40ms) and divide by the total number of voiced intervals for normalization
3) Do the same for voiced interval of lengths two (80ms) through 24 (0.96 sec) and normalize
4) Count all the intervals of length 25 (one sec or longer) and normalize. At this point, you have a vector of interval length percentages, i.e., a histogram
5) Repeat step 1-4 for unvoiced (labeled ‘2’) and silence (labeled ‘3’) for a maximum of 0.24s (six frames per interval) and 2.0s (50 frames per interval) respectively

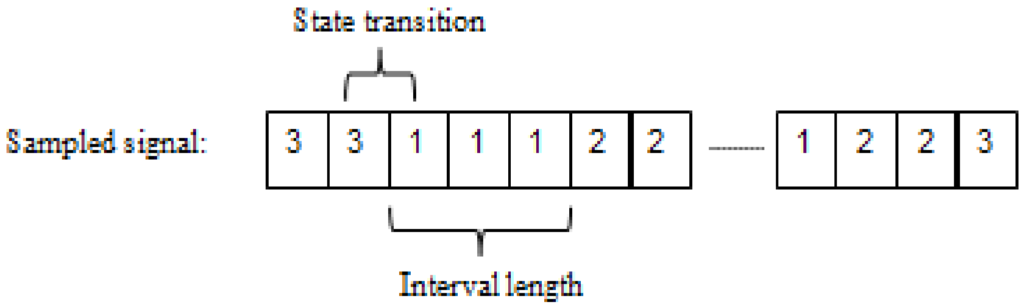


**FIGURE 1:** Graphical representation of the state transition and interval length pdf in a sampled signal.

Examples of the resulting pdfs are shown in Figure 2. Each bin is treated as a feature. For the silence interval distribution, every five consecutive interval ratios were combined in order to reduce the number of features from 50 to 10. Therefore, each bin in the silence interval distribution represents multiple numbers of interval lengths that occurs within an increment of 0.2s.

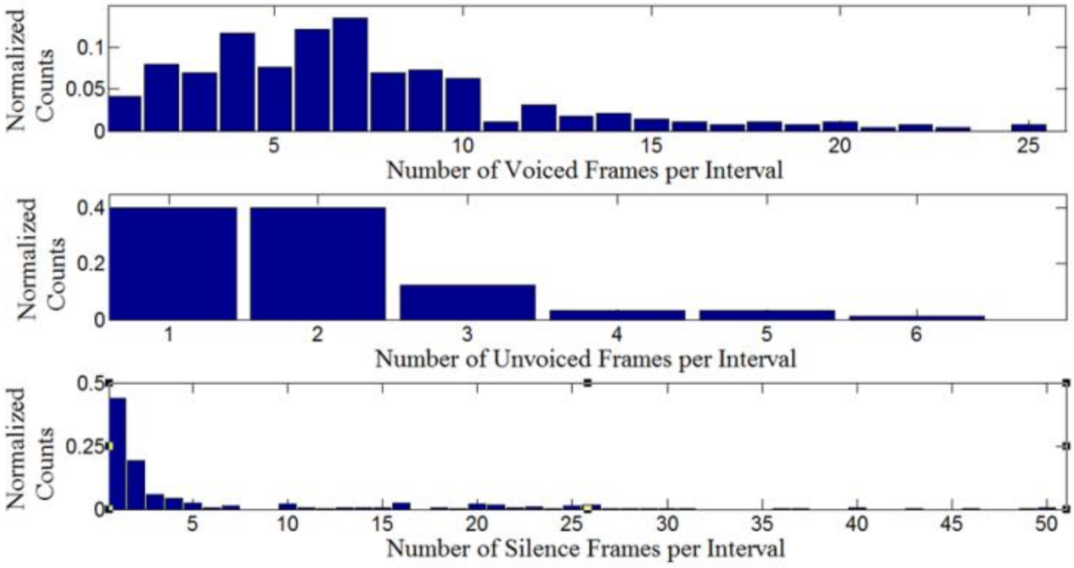


**FIGURE 2:** Examples of the voiced, unvoiced and silence interval pdf distributions

### 4.1.2 Classifier and Resampling Methods

The discriminant analyses performed on the acquired features were done on the basis of pairwise analysis classification of high risk suicidal and depressed. The decision boundaries for the two-class classification were obtained using a quadratic classifier and a linear classifier. The resampling methods that were adopted in this research were Equal Test-Train, Jackknife (Leave-One-Out) and Cross-Validation. Equal-Test-

Train uses an equal data on the training and testing database. For Cross-Validation, the partitioned samples were divided into 30% testing data and 70% training data and the resampling was iterated 100 times.

### 4.1.3 Two Stage Classification Analysis

The classification analysis was divided into two stages where in stage 1, the classification analysis was performed within the Database A, to determine how well the groups of high risk suicidal and depressed speech were able to be separated when using the proposed features. The classifications were performed on a single and multiple combinations of features using the three methods of resampling within each feature category (i.e., Transition Parameters, voiced, unvoiced and silence interval pdf). In the search for good performance in classification, features that yielded the best classification result within each category were then combined. Similar analysis was also performed on the PSD and MFCCs in order to demonstrate a comparison between the classifier performances.

For stage 2, features that were recognized to perform well according to analysis results in stage 1 were then used as the features to identify recordings of high risk suicidal in Database B1. The identification analysis was implemented by treating all patients in Database A as the training data and each recording in Database B1 as the test data.This process is repeated until all recordings in Database B1 have been chosen as the test data. This method will determine how well the information from Database A translates to Database B1, and to see if it is possible to classify patients from Database B1 with prior knowledge from a different subpopulation (Database A).

## 4.2 Regression Analysis

### 4.2.1 Voice Acoustic Features

So far there is no common agreement on which feature contains the most distinguishable information for the identification of psychological state. Common approaches include a large number of features and then applying a feature selection algorithm for dimensionality reduction. Considering the small size of the dataset, it is beneficial to include relevant features that are expected to predict well according to results from other studies on feature classification. Redundant, correlated or irrelevant features may negatively influence the performance of the predictor. The initial set consists of the following 67 acoustic features:

1) Seven equal bands of Power Spectral Density (PSD) from 0 to 1750 Hz
2) 13 Mel-Frequency Cepstral Coefficients (mfcc)
3) Transition parameters; voiced-to-voiced($t_{11}$), voiced-to-silence($t_{13}$), unvoiced-to-voiced($t_{21}$), unvoiced-to-unvoiced($t_{22}$), silence-to-voiced($t_{31}$), silence-to-silence($t_{33}$)
4) Interval Length pdfs; 25 voiced bins (voi), 6 unvoiced bins (unv), 10 silence bins (sil)

Features in items (1) and (2) are related to spectrum-based measures while features in items (3) and (4) are associated with time timing-based measures. Procedure for obtaining the PSD is similar to the work done in [8, 9] except that for this analysis, the Periodogram was applied instead of the Welch method. Also, the procedure for obtaining the MFCC is similar to the work done in [7] with the exception of implementing the code provided by Slaney [45].

### 4.2.2 Multiple Linear Regression Model

The method of multiple linear regressions using least squares was applied to examine the relationship and to obtain the model coefficients of the features (independent variables) to the clinical HAMD score (dependent variable). The general regression model equation for $P$ number of features is$\underline{h} = D\underline{a}$, where $\underline{h}$is a column vector of the actual clinical scores associated with the training data set that matches the number of rows in the matrix$D$. Matrix $D$ consists of the input vectors where the number of rows represents observations and each column represents the independent variables. The dimension of columns for $D$ is represented by: $D$ = [PSD, Transition parameters, Interval pdfs, MFCC]. $\underline{a}$is a matrix that consists of a column vector of the model coefficients resulting from the multiple regression process and having the same size as matrix $\underline{h}$. The error minimization was performed in two steps:

**Step 1:**

(Estimation) The least square solutions minimize the sum of the squared error for each prediction which is the differences between each test sample's actual clinical score and the predicted score using the estimated model coefficients from the training data set.

**Step 2:**

(Evaluation) For$\boldsymbol{N}$ number of patients, the mean sum of absolute error is calculated by summing the absolute values of the error and then dividing the total error by $\boldsymbol{N}$ which can be written as,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|e_i|\, and e_i = h_i - [D_i \ 1]a_{-i}$$

Then, using methods of feature selection, the combinations of features that give the minimum sum of absolute errors were calculated.

### 4.2.3 Feature Selection

In order to improve the jackknife method results, the next step is to identify which combination of features would provide the best predictability for the left-out patient. Essentially, the feature selection procedure attempts to find a set of features that are generalizable from the rest of the populations. Besides improving the generalization capability, by having a smaller number of features will also reduce complexity and run-time. Also, one particular problem with the linear regression is that when the number of features exceeds the number of observations, the least square solution will not be unique. Too many features may lead to a bad prediction due to the method finding a way to fit itself not just to the underlying structure but also to the irrelevant information in the training data set as well. One way to solve this is by finding out which features are relevant and eliminating features that contribute less to the prediction without involving a transformation. This approach was carried out analytically by applying two common sequential search algorithms which are called the Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS).

### *A. Sequential Forward Selection (SFS)*

SFS starts with an empty set $Z_0$ = {Ø} and let X = [$x_1x_2 \ldots x_f$] be the feature matrix. Form a linear regression estimator using exactly one feature and continue to evaluate until each feature has been chosen. Select one feature that produces the minimum mean sum of absolute error and add it to the empty set. Next, add each of the remaining features X = [$x_1x_2 \ldots x_{f-1}$] one at a time to the new subset and evaluate the two features combination that yields the best performance. Repeat the process until all features are chosen, X = {Ø}.

### *B. Sequential Backward Selection (SBS)*

SBS is a process of sequential discarding bad features. SBS starts with all features $\boldsymbol{Z}_0$ = $\boldsymbol{X}$. Repeatedly, remove one feature at a time and form a linear regression estimator using the remaining features.Discard the one feature that when remove from the set, yields the minimum mean sum of absolute error. The process is repeated until there is only one feature left in the set.

## 5.0 RESULTS

### 5.1 Results for Classification Analysis

***Stage 1: Analysis on Subpopulation in Database A***

Table 2 displays the estimated means and standard deviations of the nine Transition Parameters and the Interval pdf of voiced, unvoiced and silence in units of interval frames collected from recordings in Database A. Since each row in the transition matrix adds up to one, the probabilities in a row vector interrelate with each other, implying that if silence-to-silence is larger, silence-to-voiced and silence-to-unvoiced will also be affected. The mean and standard deviation of the Interval pdf are given in units of frames which can be translated into time by multiplying the number of interval frames by 40ms.

**TABLE 2:** Mean and Standard Deviation of the Nine Transition Parameters and the Interval pdf of Voiced, Unvoiced and Silence for Recordings

| Transition Parameter | Mean and Standard Deviation | | | |
|---|---|---|---|---|
| | Male | | Female | |
| | HR | DP | HR | DP |
| $t_{11}$ | 0.8450 ± 0.0285 | 0.8011 ± 0.0298 | 0.8167 ± 0.0292 | 0.7955 ±0.0413 |
| $t_{12}$ | 0.0098 ± 0.0072 | 0.0161 ± 0.0122 | 0.0066 ± 0.0053 | 0.0039 ± 0.0028 |
| $t_{13}$ | 0.1451 ± 0.0326 | 0.1828 ± 0.0309 | 0.1767 ± 0.0296 | 0.2006 ± 0.0403 |
| $t_{21}$ | 0.1385 ± 0.0960 | 0.1979 ± 0.0600 | 0.2418 ± 0.0416 | 0.2226 ± 0.0569 |

| | | | | |
|---|---|---|---|---|
| $\mathbf{t}_{22}$ | 0.4511 ± 0.1975 | 0.5020 ± 0.0915 | 0.4360 ± 0.0451 | 0.4032 ± 0.0882 |
| $\mathbf{t}_{23}$ | 0.2855 ± 0.1908 | 0.3001 ± 0.0501 | 0.3222 ± 0.0465 | 0.3742 ± 0.0946 |
| $\mathbf{t}_{31}$ | 0.1524 ± 0.0320 | 0.2020 ± 0.0363 | 0.1790 ± 0.0313 | 0.2050 ± 0.0446 |
| $\mathbf{t}_{32}$ | 0.0286 ± 0.0197 | 0.0553 ± 0.0141 | 0.0390 ± 0.0134 | 0.0359 ± 0.0210 |
| $\mathbf{t}_{33}$ | 0.8190 ± 0.0428 | 0.7428 ± 0.0486 | 0.7820 ± 0.0345 | 0.7591 ± 0.0575 |
| **Interval pdf** | **Mean and Standard Deviation (Interval of Frames)** | | | |
| **Voiced** | 5.7143 ± 1.1127 | 4.5000 ± 0.6742 | 4.6364 ± 0.9244 | 4.2778 ± 1.1785 |
| **Unvoiced** | 1.8571 ± 0.3780 | 1.9167 ± 0.5149 | 1.4545 ± 0.5222 | 1.4444 ± 0.5113 |
| **Silence** | 2.1429 ± 0.3780 | 2.0000 ± 0 | 2.0909 ± 0.5394 | 2.2778 ± 0.5745 |

Both male and female speech revealed a higher voice-to-voice ($t_{11}$) and silence-to-silence ($t_{33}$) mean transition probability in the high risk suicidal group compared to depressed group. The high mean transition probability of silence-to-silence ($t_{33}$) indicates that the silence pauses are longer and the high mean transition probability of voiced-to-voiced demonstrated that patients were inclined to hold their vowels longer. These behaviors were also demonstrated by the larger mean value of voiced and silence intervals in high risk suicidal speech when compared to the depressed speech, with the exception of female silence.

For unvoiced-to-unvoiced ($t_{22}$), male depressed speech and female high risk speech exhibited a higher occurrence of unvoiced, which may indicate that male depressed and female high risk suicidal patients experienced more sluggishness in speech. The same trends were observed in the Interval pdf of unvoiced speech. Although the mean and standard deviation does show some correlation or a redundancy in the information between the Transition Parameters and Interval pdf, the shape of the overall Interval pdf does contain information that is distinct from the information conveyed through the Transition Parameters.

### Transition Parameters

Table 3 presents the most effective classification performance for discriminating between groups of high risk suicidal from the depressed in male and female reading speech that were obtained by an analysis of a single and multiple combinations of features from Transition parameters. The all-data percentage indicates the percentage of vectors that are correctly classified over both groups. High risk and depressed percentages denote the percentage of vectors that are correctly classified within each group respectively.

**TABLE 3:** Results of Classification Performances using Transition Parameters for Male and Female Reading Speech in Database A

| **Male Reading Speech** | | | |
|---|---|---|---|
| **TransitionParameter** | Feature: **Silence-to-Voiced ($\mathbf{t}_{31}$)** | | |
| | All-Data % | High Risk % | Depressed % |
| Equal-Test-Train | 74 | 71 | 75 |
| Jackknife | 74 | 71 | 75 |

| | | | |
|---|---|---|---|
| Cross-Validation | 73 | 73 | 72 |
| **Female Reading Speech** | | | |
| **TransitionParameter** | Feature: **Voiced-to-Silence ($t_{13}$)** | | |
| | All Data % | High Risk % | Depressed % |
| Equal-Test-Train | 72 | 73 | 72 |
| Jackknife | 72 | 73 | 72 |
| Cross-Validation | 71 | 72 | 70 |

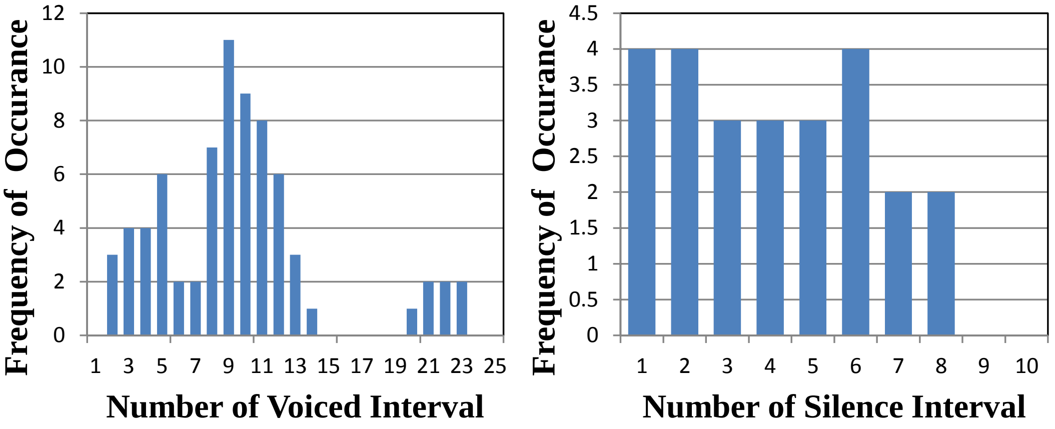
Classification using a linear discriminant analysis (LDA) on a single feature of Silence-to-Voiced ($t_{31}$) for male patients and Voiced-to-Silence ($t_{13}$) for female patients from the Transition Parameters works equally well in identifying both high risk and depressed patients. Similar results were also demonstrated by all methods of resampling.

**Interval Length pdf**

Analysis of classification using Interval pdfs were divided into three parts; voiced, unvoiced and silence. Unvoiced features did not yield any good performance for both male and female patients. For voiced and silence intervals, the analysis was performed with every single feature (i.e., a histogram bin) from the collection of 25 voiced bins and 10 silence bins. The analysis proceeded for all possible combinations of two and three features. Due to a number of combinations that yielded good classifier performances for male patients, the amounts of occurrences that a histogram bin contributes to a classification performance within the range of 75% to 100% using the jackknife procedure were collected. The histogram distribution suggests which portions of the pdfs contain significant information.

Referring to the histogram in figure 3(a) and 3(b), the discriminative information in the male reading speech occurred when patients hold their vowels for a range of time intervals from 0.16s (eight consecutive frames) to 0.48s (12 consecutive frames) with a peak at an interval of 0.36s (nine consecutive frames). On the other hand, silence pauses that occurred within an approximately 40ms (one frame) to 1.2s (30 consecutive frames) time interval contained most of the information relating to the variability characteristics in the speech of high risk and depressed.

Table 4 displays classification using a quadratic discriminant function with a single feature of 16 voiced frame per interval (Voiced16) and a combination of 16 to 20 voiced frames per interval (Voiced16:20) produced the best classification performance in identifying between high risk suicidal and depressed female patients. The classifier performed equally effective on Voiced16:20 and Voiced16 using the jackknife method. Cross-validation on the other hand effectively classified depressed speech using voiced16:20 and high risk speech when using voiced16. However, the higher percentage of correctly classified high risk suicidal patients is more preferable because identifying high risk is more critical than depressed. Therefore, if we are required to choose between the two features, Voiced16 is preferable because of the higher percentage of correctly classified high risk suicidal.



(a) (b)

**FIGURE 3:** Histogram of the individual (a) 25 voiced interval ratios and (b) 10 silence interval ratios that contributed 75% to 100% correct jackknife classification using a single and/or combination of features for male high risk and depressed speech in Database A

**TABLE 4:** Optimal Results for high Risk Suicidal and Depressed Female Reading Speech Classification using Interval pdf in Database A

| **Interval pdf** | Feature: **Voiced16: 20** | | |
|---|---|---|---|
| | All-Data % | High Risk % | Depressed % |
| Equal-Test-Train | 93 | 82 | 100 |
| Jackknife | 79 | 82 | 78 |
| Cross-Validation | 75 | 65 | 84 |
| **Interval pdf** | Feature: **Voiced16** | | |
| | All-Data % | High Risk % | Depressed % |
| Equal-Test-Train | 83 | 91 | 78 |
| Jackknife | 79 | 82 | 78 |
| Cross-Validation | 76 | 83 | 69 |

### Combined Timing-Based Features

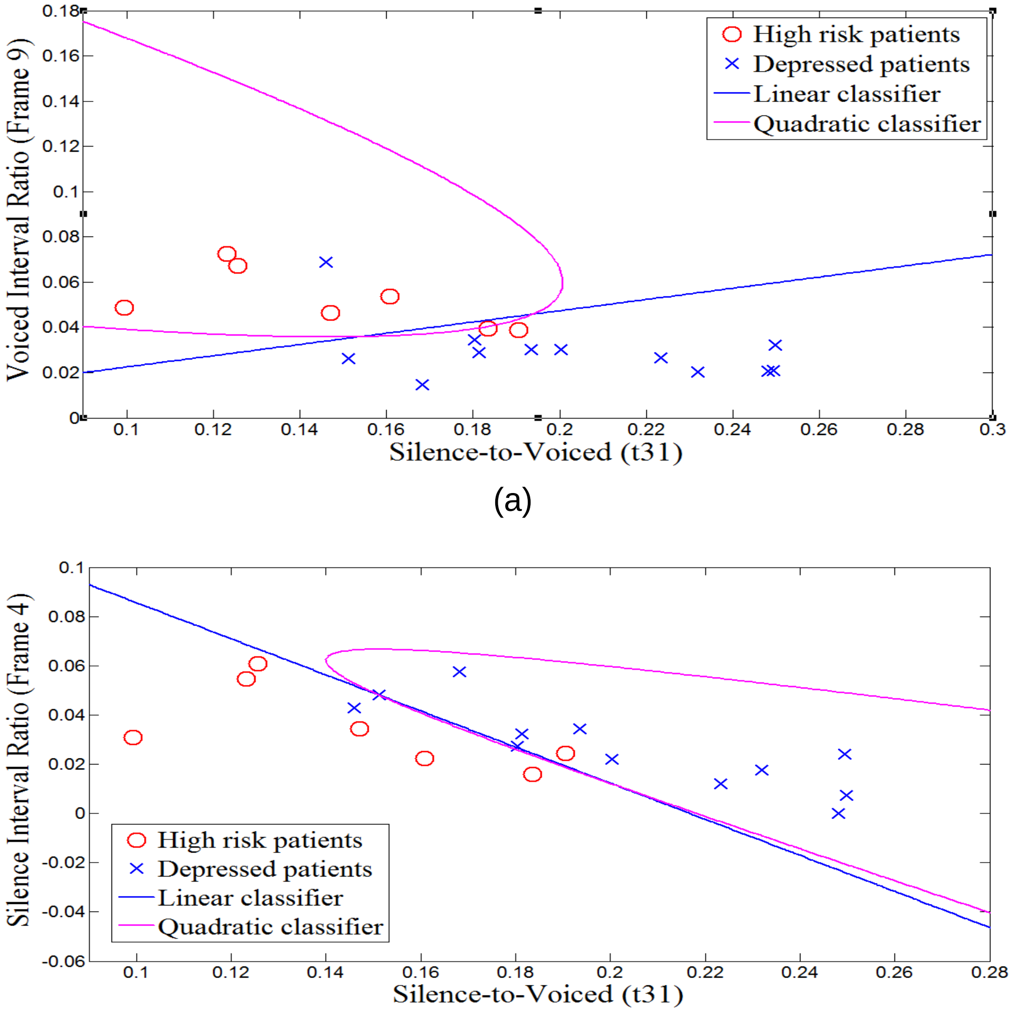
A combined feature set can affect the performance of the classifier depending on the variability of the information. The variability that exists within each feature can either complement or nullify each other. Even if the variability is high (i.e., high classification performance from each set), both features can still be carrying similar information and maintain the performance as it is.

Results of the classification analysis on female reading speech using the combination of Voiced-to-Silence ($t_{13}$) with the 16th to the 20th voiced bins and another combination with the only the 16th voiced bin are demonstrated in table 5. However, we observe that classification using the Interval pdf feature by itself performed remarkably better compared to the single Transition Parameter and the combined feature set

**TABLE 5:** Results of the Combined Features Sets Classification for High Risk and Depressed Male and Female Reading Speech in Database A

| | | | | |
|---|---|---|---|---|
| **Male Reading Speech** | | | | |
| | Feature: $\mathbf{t_{31}}$ **+ Voiced9** | | | |
| | All % | HR % | DP % | Classifier |
| Equal-Test-Train | 84 | 71 | 92 | LDA/QDA |
| Jackknife | 84 | 71 | 92 | LDA |
| Cross-Validation | 79 | 71 | 88 | LDA |
| | Feature: $\mathbf{t_{31}}$ **+ Silence4** | | | |
| | All % | HR % | DP % | Classifier |
| Equal-Test-Train | 89 | 86 | 92 | LDA/QDA |
| Jackknife | 74 | 71 | 75 | LDA |
| Cross-Validation | 79 | 72 | 85 | LDA |
| **Female Reading Speech** | | | | |
| | Feature: $\mathbf{t_{13}}$ **+ Voiced16: 20** | | | |
| | All % | HR % | DP % | Classifier |
| Equal-Test-Train | 93 | 82 | 100 | QDA |
| Jackknife | 76 | 64 | 83 | QDA |
| Cross-Validation | 65 | 47 | 84 | QDA |
| | Feature: $\mathbf{t_{13}}$ **+ Voiced16** | | | |
| | All % | HR % | DP % | Classifier |
| Equal-Test-Train | 79 | 82 | 78 | QDA |
| Jackknife | 72 | 64 | 78 | QDA |
| Cross-Validation | 71 | 66 | 76 | QDA |

For male patients, we performed a classification analysis using combinations of Silence-to-Voiced ($t_{31}$) with each single feature of eighth to 12th voiced bin and the first to sixth silence bin. The ninth voiced interval (Voiced9) and fourth silence interval (Silence4) produced the best classification when combined with Silence-to-Voiced ($t_{31}$) as shown in table 5. The overall results demonstrated that the classifier performed better on the depressed speech compared to the high risk suicidal speech. Figure 4(a) and 4(b) plot the distribution of high risk and depressed patients using the combined feature set. By observation, the distributions of high risk patients and depressed patients were distinct from each other and vectors that are misclassified were fairly close to the boundary except for one of the depressed patients as shown in figure 4(a).



(a)

(b)

**FIGURE 4:** Plot of the high risk and depressed patient distribution for the combined feature set of Silence-to-Voiced ($t_{31}$) with (a) voiced interval ratios in frame 9 and with (b) silence interval ratios in frame 4 using linear and quadratic discriminant classifier

### Power Spectral Density and Mel-Frequency Cepstral Coefficients (Spectrum-Based Measures)

An error histogram was generated by performing 100 iterations on cross-validation results to identify outliers that were affecting the PSD classifier performance. Two male patients and a female patient were identified as outliers in the high risk suicidal group. Also, a depressed male patient was identified as outlier. Table 6 demonstrates an effective classification percentage using a linear classifier on the PSD and MFCC features for male and female reading speech in Database A. Using the Jackknife procedure as a measure of performance, the classifier performed equally well in identifying between the pairwise groups.

**TABLE 6:** Results of the Suboptimal LDA Jackknife Classification using PSD and MFCC for Male and Female Reading Speech in Database A

| **Male Reading Speech** | | | |
|---|---|---|---|
| **Features** | All % | HR % | DP % |
| 4PSD$_1$, 4PSD$_2$ | 90 | 83 | 96 |
| MFCC 5, MFCC 7 | 82 | 76 | 88 |
| **Female Reading Speech** | | | |
| 8PSD$_1$, 8PSD$_2$, 8PSD$_3$, 8PSD$_4$ | 78 | 80 | 76 |
| MFCC 6, MFCC 13 | 82 | 76 | 88 |

***Stage 2: Analysis of Classification between Two Populations***

In the second stage, we test the ability of the classifier to identify high risk recordings in Database B1 using the trained features and classifier that we analyzed in stage 1 on Database A. As shown in table 7, using the feature of Silence-to-Voiced ($t_{31}$) alone, seven out of eight male speech recordings that were labeled as high risk suicidal were successfully identified. The results improve significantly to a perfect identification of the high risk suicidal speech when Silence-to-Voiced ($t_{31}$) was combined with Voiced9 or Silence4. Remarkably, the classifier that was trained using only Voiced9 produced similar result as the combined features. On the other hand, the spectrum-based measures, PSD and MFCC poorly identified the high risk male patients with only a correct classification of 30% and 52%, respectively.

**TABLE 7:** Results of the Tested Classifier for the Identification of High Risk Suicidal Recordings in Male Patients in Database B1

| **Feature Combination** | **High Risk %** | **Classifier** |
|---|---|---|
| $t_{31}$ | 86 | QDA |
| Voiced9 | 100 | LDA/QDA |
| $t_{31}$ + Voiced9 | 100 | LDA/QDA |
| $t_{31}$ + Silence4 | 100 | QDA |

Although classification results of the trained classifier in male reading speech using the timing-based measures were outstanding, features from female reading speech did not translate well between the two populations. Among all the features that performed well in stage 1, the highest performance of the classifier was only able to achieve 58% correct identification of female high risk suicidal patients in Database B1 which was using the Voiced-to-Silence($t_{13}$) feature.

## 5.2 Results for Regression Analysis

Table 8 displays the estimated number of features, Mean Sum of Absolute Error (MAE), Standard Deviation Sum of Absolute Error (SDAE) and Median Sum of Absolute Error (MdAE), maximum error (MaxE) and percentage of absolute error above MAE (%>MAE) using methods of SFS and SBS for predicting the patient's clinical scores according to groups of male and female interview and reading speech in Database A. Plus or minus sign on maximum error signify direction of error. A negative error indicates that the clinical score prediction is less than the actual score and vice versa. The numbers α(β,γ) in row labeled %>MAE represents percentage of the absolute errors that are above the MAE (α), percentage of errors that are above +MAE (β), and percentage of errors that are below–MAE (γ).

**TABLE 8:** Statistical Comparison on the Application of the SFS and SBS Procedure using the Reading Speech from Male and Female Patients in Database A for the Prediction of HAMD Scores

| | | **Male Reading** | **Female Reading** |
|---|---|---|---|
| **SFS** | **# features** | 4 | 9 |
| | **MAE** | 1.4282 | 1.6893 |
| | **SDAE** | -0.0064 | -0.1341 |
| | **MdAE** | 1.8551 | 2.3055 |
| | **MaxE** | -4.5200 | 7.1571 |
| | **% >MAE** | 38 (14,24) | 36 (18,18) |
| **SBS** | **# features** | 14 | 13 |
| | **MAE** | 1.3389 | 2.1558 |
| | **SDAE** | 0.1629 | 0.1750 |
| | **MdAE** | 2.9023 | 2.7834 |
| | MaxE | -9.5037 | -6.3280 |
| | % > MAE | 10 (0,10) | 48 (21,27) |

The MAE value directly expresses the measure of how close the prediction scores are to the actual HAMD scores without considering the direction of error. If for example a patient has an actual HAMD score of 23, an error of minus four could misplace the patient in the lower level category of severe depression considering 23 is the threshold score for the high risk suicidal. By observation and from a clinical perspective, an error of approximately three or less is considered to be insignificant.

The analysis revealed an effective prediction of the HAMD scores by means of speech features as demonstrated by the minimal MAEs in both male and female categories. Apart from the application of SBS method in the male reading speech, the rest of the groups achieved a minimal MAE using a total number of features that is less than the total number of patients which are nine for male and 14 for female patients (see table 1). Thus, indicating that the suboptimal combinations of features are generalizable.

The plots of MAE of the HAMD scores predictions with respect to the total number of features obtained by the SFS and the SBS procedure for the male and female reading patients are illustrated in figures 5 and 6.

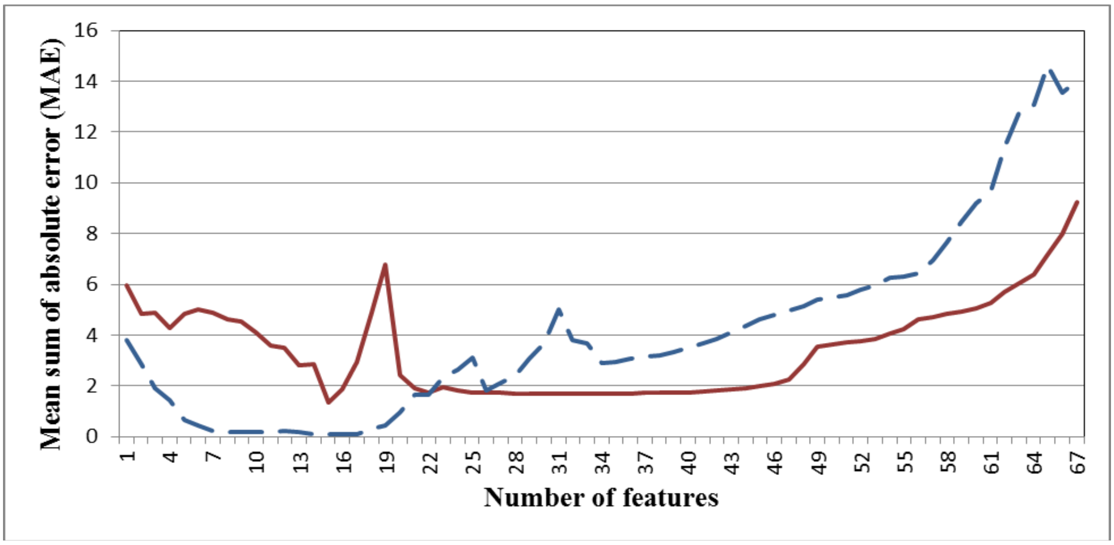


**FIGURE 5:** Characteristic plot of the SFS (blue line '--') and the SBS (red line) methods using the male reading speech from Database A to predict the HAMD scores

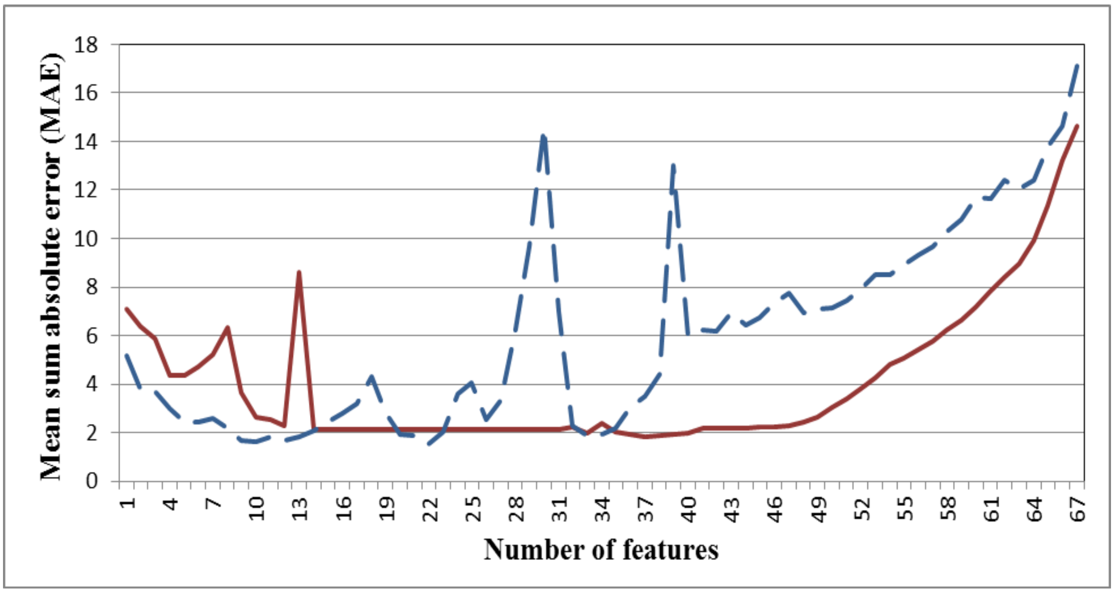


**FIGURE 6:** Characteristic plot of the SFS (blue line '--') and the SBS (red line) methods using the female reading speech from Database A to predict the HAMD scores

The SBS graph reads from right to left because the initial set contains all 67 features and each feature was then discarded one at a time. The graph originally exhibits a decreasing trend which demonstrate that the discarded features were irrelevant thus improving the prediction performance. After reaching a certain point, the discarded features do not change the MAE significantly thus the approximately straight line was obtained for a number of feature combinations. After removing all irrelevant features, the graph exhibit an increment in MAE. This shows that the discarded features contains significant information thus removing these features decreases the performance of the prediction. For the female reading speech, the minimum MAE was obtained after removing all spectrum-based measure. For the male reading speech, the feature combination that yielded the minimum MAE contains a mixture of the spectrum- and timing-based measures.

Based on figure 5, the characteristic plot of the SFS method using male reading speech initially demonstrated a decreasing trend when adding one feature at a time. This shows that each added feature contains information that increases the accuracy of the prediction thus reducing the MAE. Beginning at the 5th added feature, the MAE remains less than one until the 20th added feature. This demonstrates that the added features contain no significant information. The graph then exhibits an increase in MAE indicating that the additional information reduces the performance of the prediction. The combination of the first five features yielded a MAE of 0.6720. The 5th feature that was added to the combination is a spectrum-based measure. However, removing the 5th feature produced a MAE of 1.4282 which is still considered insignificant error prediction and plus, all four features are the timing-based measure.On the other hand, the combination of features using the SFS method that produced the minimum MAE (shown in figure 6) for the female reading speech consists of a combination of the spectrum- and timing-based measures.

Figures 7 to 8 display comparison plots of the actual and the predicted HAMD scores obtained using the selected SFS and SBS feature combination from the male reading speech. The predicted HAMD scores for the male patients using a feature combination obtained through the method of SBS were more accurate compared with the SFS method. The prediction errors were mostly concentrated on the first patient (the first two 20 second segments).

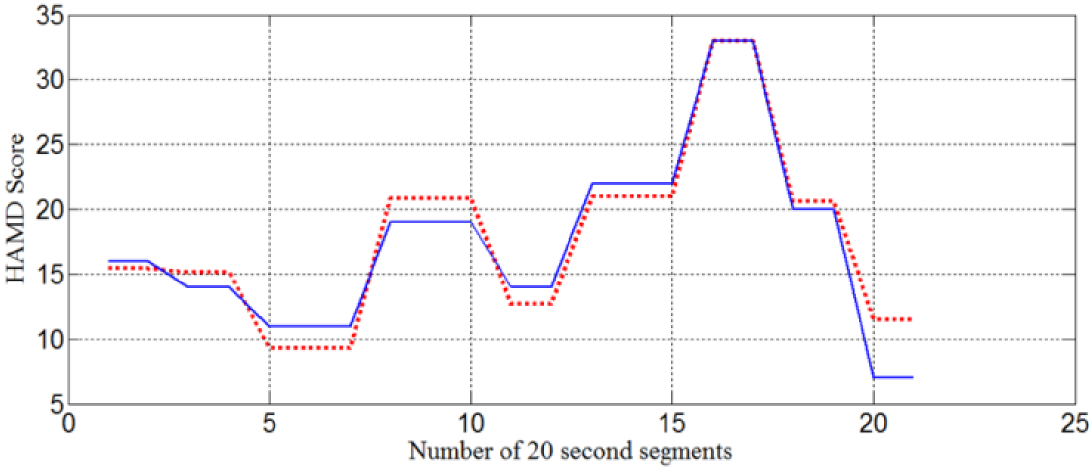


**FIGURE 7:** The actual (blue '—') and the predicted (red '--') HAMD scores for male reading patients in Database A using the SFS

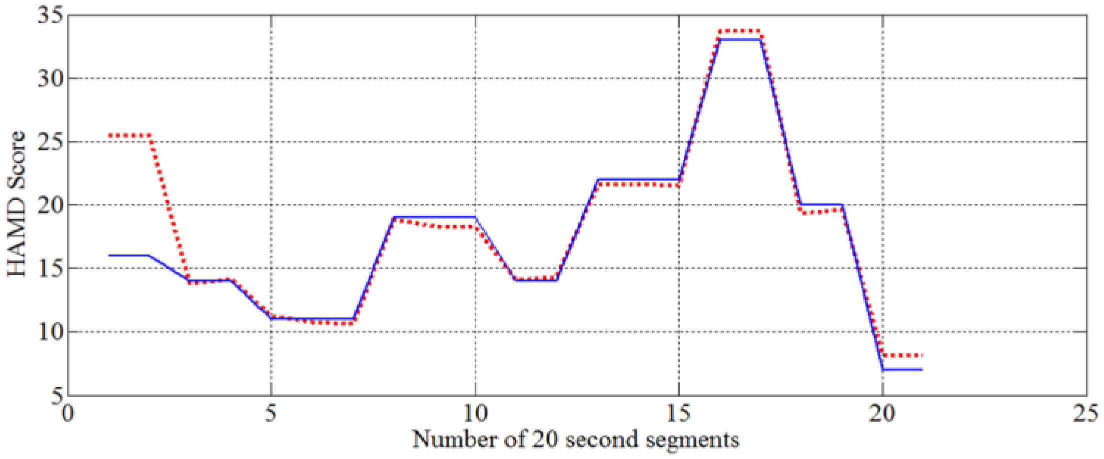


**FIGURE 8:** The actual (blue '—') and the predicted (red '--') HAMD scores for male reading patients in Database A using the SBS

Figures 9 to 10 display comparison plots of the actual and the predicted HAMD scores obtained using the selected SFS and SBS feature combination from the female reading speech. Comparison between the predicted HAMD score using a combination of features obtained by the method of SFS and SBS on female speech exhibited a similar characteristic in the direction of the errors. Overall, the prediction errors caused by the selected feature combination using the SBS method are slightly higher than the SFS method.

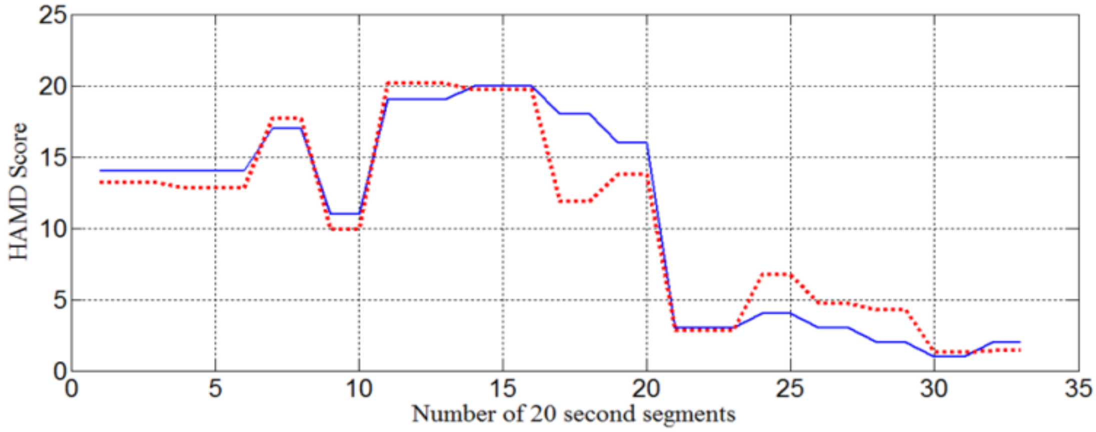


**FIGURE 9:** The actual (blue '—') and the predicted (red '--') HAMD scores for female reading patients in Database A using the SFS

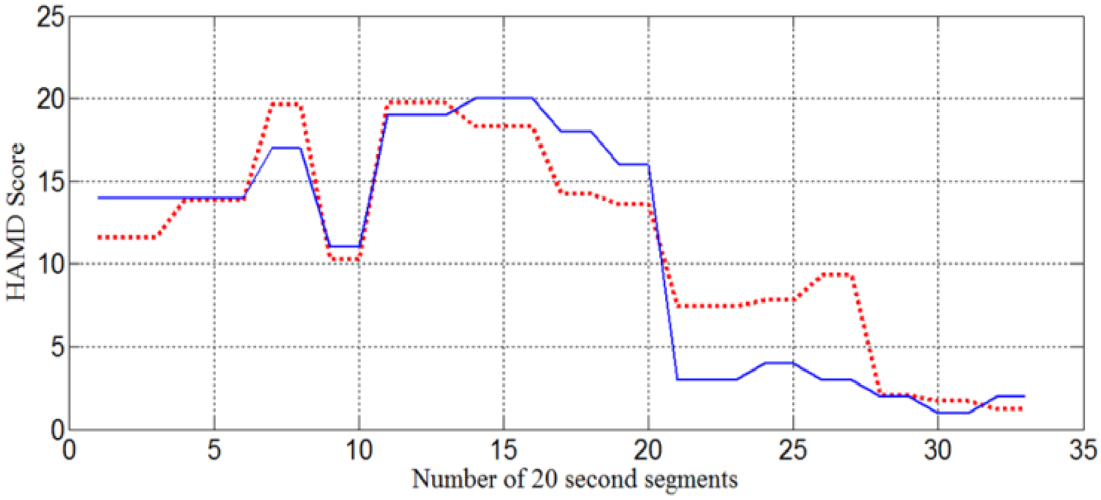


**FIGURE 10:** The actual (blue '—') and the predicted (red '--') HAMD scores for female reading patients in Database A using the SBS

Table 9 displays the comparison between methods of SBS and SFS based on the percentage of the 20 second segment of speech that produced a HAMD score with prediction error of less than one, two and three for male and female reading speech. Error larger than three is considered to be significant. Both sequential methods that were applied to the male reading speech successfully identified combination of features that effectively predicted the HAMD score with a performance of 90.48% error less than three. For the female patients, the method of SFS slightly outperformed the SBS in determining a combination of speech feature that could predict the HAMD score with a better accuracy and a higher percentage of prediction error less than three.

**TABLE 9:** Percentage of Patients with an Error Prediction of the HAMD Score of Less Than One, Two and Three for the Male (MR) and Female (FR) Reading Speech by Methods of SFS and SBS

| | **SFS** | | | **SBS** | | |
|---|---|---|---|---|---|---|
| **% Error** | **< 1** | **< 2** | **< 3** | **< 1** | **< 2** | **< 3** |
| **MR** | 28.57 | 90.48 | 90.48 | 80.95 | 90.48 | 90.48 |
| **FR** | 39.39 | 72.73 | 87.88 | 42.42 | 51.52 | 72.73 |

## 6.0 DISCUSSION

### Part 1: Discussion on the Classification Analysis

This first part of the paper presented new methods of extracting features based on the timing patterns of speech using the Markov Transition Matrix and the Interval pdf of voiced, unvoiced and silence for the analysis of vocal characteristics for high risk suicidal and depressed detection. The results of this investigation correlate with the previous findings where it was shown that features relating to voice and silence from Transition Parameters and Interval pdf provided prominent results in classifying the two groups. Besides the preliminary process of separating voiced, unvoiced and silence segments, the process of obtaining the Transition Parameters and Interval pdf are not related to the spectrum-based measures. In the reading (controlled) speech, variations in phonemes and articulation can almost be eliminated because each patient was reading from the same passage.According to Ellgring [13], the use of controlled speech disregards the involvement of complex cognitive planning processes and variation of the pause time is emphasized.

The Transition Parameters represents the decision or transition probabilities between speech frames. Results demonstrated that information on the distinguishing characteristics of high risk suicidal and depressed in both male and female are mostly embedded in the transition probabilities of silence and voiced speech. Silence-to-Voiced($t_{31}$) was found to be the most significant features in distinguishing between high risks suicidal and depressed male patients and the Voiced-to-Silence($t_{13}$) to female patients. However, the strong consistency of this feature can be observed at least for the male patients and thus will be discussed in further detail. The probability that the current frame is silence and the next frame is voiced is affected by the length of the silent pauses because a row in the transition matrix sums to one. If silent pauses are longer, the probability of silence-to-silence will increase and will force other probabilities to decrease. The probability of the observed significant feature that revolves around the interaction between silent and voiced frames can also be affected by either longer voiced sections or longer silences. Therefore, four features of Voiced-to-Voiced ($t_{11}$), Voiced-to-Silence ($t_{13}$), Silence-to-Voiced ($t_{31}$) and Silence-to-Silence ($t_{33}$) can be used to analyze the characteristic between high risk suicidal and depressed. Referring back to table 2, the means of the inter-transitions between silence and voice for depressed patients were higher than the high risk patients thus signifying a more frequent and active start and stop in depressed speech. The higher means in the intra-transitions within silence and voiced for high risk patients indicate a slower speech rate, holding out vowels longer and taking longer time for pauses.

The Interval Length pdf describes the overall shape of the distribution within voiced, unvoiced and silence where longer intervals are expected to have more variability. The overall pdf for voiced, unvoiced and silence exhibit similar characteristic of an asymmetrical right-skewed distribution.The distribution's peak is off centered with a tail stretching in the opposite direction away from it. Most variability occurs in the tail end shape of the pdf which are demonstrated by the results obtained from both male and female categories. For male interval distribution, the significant feature of the nine voiced frame per interval (Voiced9) and the four silence frame per interval (Silence4) were located in the tail direction away from their mean. Similarly for female interval distribution, the significant feature of 16 to 20 voiced frames per interval (Voiced16: 20) was nearly close to the tail end shape of the pdf.

Only a small feature set was used to generate a strong performance, developed using two completely different databases. One database was used to train the paradigm, and it was tested on the second dataset. We were able to find a single

Transition Parameter Silence-to-Voiced ($t_{31}$) and the ninth bin of voiced interval pdf (Voiced9) that produced 86% and 100% separation on high risk recordings, respectively. Also, using two combined parameters of Silence-to-Voiced ($t_{31}$) with the ninth bin of voiced interval pdf (Voiced9) and combination of Silence-to-Voiced ($t_{31}$) with the fourth bin of silence interval pdf (silence4), both revealed 100% separation on high risk recordings. The fact that only one or two parameters were able to produce the quality of discrimination and also perform across two datasets recorded in different environments (a variety of clinical interview rooms) using different high-quality devices strengthens the argument that these results are not coming from over-modeling or from spurious environmental factors. It is a strong indication that there is significant information within these parameters. Additionally, the parameters are easily calculated.

Reading speech was used in this study due to the consistency in the spoken words. Each patient was saying the exact same words. Patients were given the standard “rainbow passage” essay that contains all phonemes found in the English language. However, the difference in the transition probability might have existed because of the variability in the decision made between voiced, unvoiced and silence. For the same word, some that are marked as voiced in one patient might have been marked as something else in another patient. This can contribute to a low classification result when a trained classifier is tested on a different population as demonstrated by the female group. Aside from that, female speech has been reported to be more breathy than male. Breathy voice quality occurs because of the incomplete closure of vocal folds that allows air to flow through the glottis and thus presents an existence of noise in the higher frequency spectrum, domination of harmonic excitation by aspiration noise and alternations in vocal tract which are shown by the extra poles and zeros in the vowel spectrum [41]. Therefore, a word that should have ended with a full voicing might be influenced by the aspiration.

When discussing the issue of small sample size, it is important to keep the dimensionality of studied features low. High dimensional feature spaces often lack generalization and can lead to over-modeling the limited dataset. Results could become highly questionable when using huge dimensional spaces on a very small amount of data because it could easily be modeling things that are not the characteristics of general population but just the individuals of the small data sets. Thus, it might work well for the corresponding dataset but failed to generalize well to novel datasets. According to a rule of thumb for an adequate sample size, an appropriate number of samples per estimated feature are of the 5: 1 ratio [42]. Nevertheless, this study only used one and/or two features to obtain effective classifier performance within a small number of sample set. This suggests that there is valuable information embedded in the small number of features relative to the data.

The two databases were recorded at different times, during collection intervals that used two different types of high-quality recording devices (Audix SCX-one cardioid and TASCAM DR-1). The fact that the trained classifier performed so well when tested on the male participants Database B1, demonstrated that the features were not affected by different recording devices. Based on the poor results demonstrated by the PSD and MFCC classification on Database B1, we derived two hypotheses that were affecting the analysis, (1) different recording devices or (2) different people chosen from the two populations. However, the timing-based measures efficiently translated the information within Database A and Database B1. So, the latter hypothesis was rejected. To validate the former hypothesis, we then performed a separate experiment by recording a speech using both devices (SCXone and DR1) at the same time, keeping the environment and the speaker constant. Interestingly, we observed the extracted PSD values for both speech samples to be considerably different. Thus, the analysis concludes that the spectrum-based measures were sensitive to the use of different recording devices.

**Part 2: Discussion on the Regression Analysis**

In the second part of the analysis, we demonstrated the effectiveness of using acoustic measurements as a possible means to predict the clinical HAMD score. The results are based on method of linear regression and applying a feature selection procedure to increase the performance of predictions with fewer numbers of features. Backward and forward feature selection methods are popular choices for the use of dimensionality reduction of feature space and removing redundant, irrelevant or noisy data because the algorithms are simple and easy to implement. Also, feature selection selects a subset of features without transformation thus retaining the original physical interpretation whereas feature extraction reduces dimensionality by projecting onto a new dimension. Even though both methods performed well in this study, these algorithms have a tendency to produce error that will move in a downward direction and can sometimes become trapped in local minima. The drawback for the SFS procedure is the inability of replacing or eliminating selected features that have become redundant after the inclusion of new features while the SBS method does not allow the discarded features to be reexamined once removed from the set and thus eliminate the likelihood of finding a feature that works best on its own.

In this study, we implemented the jackknife analysis using a heuristic search algorithm which is the sequential feature selection method for searching a set of features that are close to an optimal solution. This procedure is simple but it only explores a limited number of structures. In contrary, the brute force technique searches for all possible outcomes and it is also capable of producing an optimal solution, however, this procedure is considered unreasonable to be applied in this study. The reason is because of the extensive number of possible combinations to search for the 67 features. For example, finding all possible combinations of 10 features out of 67 features will require approximately 1.28x$10^{12}$ jackknife analyses to be executed.

Evaluation of model performance error was based on the measure of mean absolute error (MAE). This quantity was preferred over mean squared error (MSE) because the nature of its calculation clearly describes the results. MAE weighs all individual differences in an equally manner and clearly measures how close prediction are to the actual value without considering their direction. Although MSE is a conventional method that has been used commonly, the error measures are considered unreliable for this study because the interpretations of the results are abstruse. MSE quantifies the variance of errors by measuring the difference between the prediction and the actual value. Also, by squaring the errors, greater emphasis is being put on to large errors thus allowing the total square error to be affected by a possible existence of outliers. Another error measurement that was also reported in this study is the maximum error (MaxE) which expresses the maximum estimated differences between the prediction and the actual value. The results act as a bound that describes the maximum margin of an error for a certain set of predictions.

Essentially, by looking at the results, we found that we were able to identify features that do generalize by demonstrating the ability of a certain population to predict the behavior of an individual through their clinical scores. However, the regression and the feature selection methods are design selections and may require further analysis by taking into account the different error trade-offs.

## 7.0 CONCLUSION

Features that relate to the timing pattern of speech and are not affected by the acoustic content of the speech were investigated in this report. Classifications using only a single and/or two combined features were shown to achieve the best classification performances for all methods of resampling (equal-test-train, jackknife and cross-validation). The advantage of having a small number of features demonstrated generalizabilty. Analysis of classification using male reading speech displayed consistent results within and throughout populations. Also, the features are robust across data sets despite the less than ideal recordings conditions and different equipment used during the recordings sessions. A consistent pattern of significant and predictive validity of the regression model was demonstrated through the feasibility of using speech features as a potential means to predict the clinical HAMD scores. Application of Multiple Linear Regression was effective for generating the model prediction with this study database. The encouraging performance of the predictors makes this research worthy of further investigation on a larger sample population. Above all, this procedure is practical for use in real applications and can be used during a standard clinical interview by having the recordings done in a normal closed room and without strict control on the recording environment.

## REFERENCES

[1] S. L. Murphy, J. Q. Xu, K. D. Kochanek, “Final Data for 2010. National Vital Statistics Reports”, Hyattsville, MD: National Center for Health Statistics, vol. 61, no. 4, 2013.

[2] M. Heron, “Deaths: Leading Causes for 2009: National Vital Statistics Reports”, Hyattsville, MD: National Center for Health Statistics, vol. 61, no. 7, 2012.

[3] R. Burns, “An Impact: Suicides are Surging among US Troops”, *The Associated Press*, 2012.

[4] K. A. Busch, J. Fawcett, D. G. Jacob, “Clinical Correlates of Inpatient Suicide”, *J. Clin. Psychiatry*, vol. 64, no. 1, pp. 14-19, 2003.

[5] J. C. Fowler, “Suicide Risk Assessment in Clinical Practice: Pragmatic Guidelines for Imperfect Assessment”, *American Psychological Association*, vol. 49, no. 1, pp. 81-90, 2012.

[6] D. J. France, R. G. Shiavi, S. E. Silverman, M. K. Silverman, D. M. Wilkes, “Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk”, *IEEE Transaction on Biomedical Engineering*, vol. 47, no. 7, 2000.

[7] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, D. M. Wilkes, “Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk”, *IEEE Transaction on Biomedical Engineering*, vol. 51, no. 9, 2004.

[8] T. Yingthawornsuk, H. K. Keskinpala, D. M. Wilkes, R. G. Shiavi, R. M. Salomon, “Direct Acoustic Feature using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Speech”, *INTERSPEECH*, pp. 766-769, 2007.

[9] H. K. Keskinpala, T. Yingthawornsuk, D. M. Wilkes, R. G. Shiavi, R. M. Salomon, “Screening for High Risk Suicidal States using Mel-Cepstral Coefficients and Energy in Frequency Bands”, *In European Signal Processing Conf.*, pp. 2229-2233, 2007.

[10] K. R. Scherer, “Vocal affect expression: A Review and Model for the Future Research”, *Psychological Bulletin*, vol. 99, no. 2, pp. 143-145, 1986.

[11] D. Ververidis, C. Kotropoulos, “Emotional Speech Recognition: Resources, Features and Methods”, *Speech Communication*, vol. 48, no. 9, pp. 1162-1181, 2006.

[12] A. Askenfelt, S. Nilsonne, “Voice Analysis in Depressed Patients: Rate of Change of Fundamental Frequency Related to Mental State”, *STL-QPSR*, vol. 21, no. 2-3, pp. 71-84, 1980.

[13] H. Ellgring, K. R. Scherer, “Vocal Indication of Mood Change in Depression”, *J. of Nonverbal Behavior*, vol. 20, no. 2, pp. 83-110, 1996.

[14] J. K. Darby, H. Hollien, “Vocal and Speech Patterns of Depressive Patients”, *Int. J. of Phoniatrics, Speech Therapy and Communication Pathology*, vol. 29, no. 4, pp. 279-91, 1997.

[15] M. Alpert, E. R. Pouget, R. R. Silva, “Reflections of Depression in Acoustic Measures of the Patient’s Speech”, *J. of Affective Disorders*, vol. 66, no. 1, pp. 59-69, 2001.

[16] E. Szabadi, C. M. Bradshaw, J. A. O. Besson, “Elongation of Pause-Time in Speech: A simple, Objective Measure of Motor Retardation in Depression”, *The British J. of Psych.*, vol. 129, no. 7, pp. 592-597, 1976.

[17] E. Moore, M. A. Clements, J. W. Peifer, L. Weisser, "Critical Analysis on the Impact of Glottal Features in the Classification of Clinical Depression in Speech", *IEEE Trans. On Biomed. Eng.,*vol. 55, no. 1, pp. 96-107, 2008.

[18] G. H. Monrad-Krohn, "The Third Element of Speech: Prosody in the Neuro-Psychiatric Clinic", *The British J. of Psych.,* vol. 103, no. 431, 326-331, 1957.

[19] A. Ghozlan, D. Widlocher, "Decision Time and Movement Time in Depression: Differential Effects of Practice Before and After Clinical Improvement", *J. of Perceptual and Motor Skills,* vol. 68, no. 1, pp. 187-92, 1989.

[20] G. M. Hoffman, J. C. Gonze, J. Mendlewicz, "Speech Pause Time as a Method for the Evaluation of Psychomotor Retardation in Depressive Illness", *The British J. of Psych.,* vol. 146, pp. 535-538, 1985.

[21] A. Nilsonne, "Acoustic Analysis of Speech Variables During Depression and After Improvement", *Acta. Psychiatr. Scand*, vol. 76, no. 3, pp. 235-45, 1987.

[22] A. Nilsonne,: Speech Characteristics as Indicators of Depressive Illness", *Acta. Psychiatr. Scand*, vol. 77, no. 3, pp. 253-63, 1988.

[23] P. Hardy, R. Jouvent, D. Widlocher, "Speech Pause Time and the Retardation Rating Scale for Depression (ERD)", *J. of Affective Disorders*, vol.6, pp. 123-127, 1984.

[24] M. Cannizaro, B. Harel, N. Reilly, P. Chappell, P. J. Snyder, "Voice Acoustical Measurement of the Severity of Major Depression", *Brain and Cognition*, vol. 56, no. 1, pp.30-35, 2004.

[25] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, D. S. Geralts, "Voice Acoustic Measures of Depression Severity and Treatment Response Collected Via Interactive Voice Response (IVR) Technology", *J. of Neurolinguistic*, vol. 20, pp. 50-64, 2007.

[26] A. C. Trevino, T. F. Quatieri, N. Malyska, "Phonologically-based Biomarkers for Major Depressive Disorders", *EURASIPJ. on Advances in Signal Processing,* vol. 42, 2011.

[27] J. C. Mundt, A. P. Vogel, D. E. Feltner, W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response", *J. of Biological Psychiatry*, vol. 72, no. 7, pp. 580-587, 2012.

[28] Ying Yang, F. Catherine, J. F. Cohn, "Detecting Depression Severity from Vocal Prosody", *IEEE Trans. On Affective Computing*, vol. 4, no. 2, 2013.

[29] J. F. Cohn, T. S. Kruez, I. Matthews, Ying Yang, Minh Hoai Nguyen, M. T. Padilla, Feng Zhou, F. De la Torre, "Detecting Depression from Facial Actions and Vocal Prosody", *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference,* pp. 1-7, 2009.

[30] R. M. Lane, J. F. O'Hanlon, "Cognitive and psychomotor effects of antidepressants with emphasis on selective serotonin reuptake inhibitors and the depressed elderly patient", *German J. of Psych*, 1999.

[31] D. Schrijvers, W. Hulstijn, B. G. C Sabbe, "Psychomotor symptoms in depression: A diagnostic, pathophysiological and therapeutic tool", *J. of Affective Disorders*, vol. 109, no. 1-2, 2008.

[32] D. Marazziti, G. Consoli, M. Picchetti, M. Carlini, L. Faravelli, "Cognitive Impairment in Major Depression", *European J. of Pharmachology*, vol. 626, pp. 83-86, 2010.

[33] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, D. P. Rosenwald, "Social Risk and Depression: Evidence from Manual and Automatic Facial Expression Analysis" *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops*, pp. 1-8, 2013.

[34] J. Joshi, A. Dhall, R. Goecke, J. F. Cohn, "Relative Body Parts Movement for Automatic Depression Analysis", *Affective Computing and Intelligent Interaction (ACII), 2013 HumaineAssociation Conference* pp. 492-497, 2013.

[35] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, M. Breakspear, "Head Pose and Movement Analysis as an Indicator of Depression" *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference,* pp. 283-288, 2013.

[36] J. Joshi, R. Goecke, G. Parker, M. Breakspear, "Can Body Expressions Contribute to Automatic Depression Analysis?",*Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops,* pp. 1-7, 2013.

[37] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, "Automatic Behavior Descriptors for Psychological Disorder Analysis", *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops,* 2013.

[38] R. M. Salomon, H. K. Keskinpala, M. H. Sanchez, T. Yingthawornsuk, N. H. Nik Wahidah, W. S. Hasan, N. Taneja, D. Vergyri, B. H. Knoth, P. E. Garcia, D. M. Wilkes, R. Shiavi, "Analysis of Voice Speech Indicators in Suicidal Patients", Manuscript submitted for publication, 2012.

[39] International Phonetic Association, Phonetic Description and the IPA Chart, Handbook of the International Phonetic Association: A Guide to the Use of International Phonetic Alphabet, Cambridge University Press, 1999

[40] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE,* vol. 77, no. 2, pp. 257-286, 1989.

[41] D. H. Klatt, L. C. Klatt, “Analysis, Synthesis and Perception of Voice Quality Variations Among Female and Male Talkers”, *J. Acoust. Soc. of America*, vol. 87, no. 2, 1989.

[42] H. M. Kalayeh, D. A. Landgrebe, “Predicting the required number of training samples, Pattern analysis and machine intelligence”, *IEEE transaction on Pattern Analysis and Machine Learning*, vol. 5, no.6, pp. 664-667, 1983.

[43] M. Hamilton, A Rating Scale for Depression, Journal Neurol. Neurosurg. Psychiat., 23, 56, 1960.

[44] G. K. Brown, A Review of Suicide Assessment Measures for Intervention Research with Adults and Older Adults, Technical report submitted to NIMH Bethesda, MD: National Institute of Mental Health, 2002.

[45] M. Slaney, “Auditory Toolbox Version 2”, Interval Research Corporation, Technical Report #1998-010.