Abstract

Depression is a mood disorder that mostly goes undetected and untreated. There is an urgent need to build an objective system that can detect depression in time, to avoid the dire consequences of suicide. In literature, speech, speech content, images, videos are some of the modalities that have been explored independently (unimodal) as well as in combination (multimodal), to detect depression. As multimodal systems are not cost effective and are computationally intensive there is a need to strengthen the unimodal systems. Speech is a non-invasive medium which is strongly correlated with the depressed state of mind and is simple and cost effective. Though some feature extraction methods have been explored in speech based systems, an exhaustive set of possible temporal and spectral features has not been studied. We have investigated an exhaustive set of Temporal and Spectral features, along with a hybrid of the two for speech based depression detection. It is also desirable to determine a smaller set of relevant features to overcome the problem associated with the small sample size problem. To acheive an effective learning model, a relevant subset of features from these exhaustive feature set has been obtained using univariate feature selection methods (Pearson Correlation(PC), Mutual Information method(MI), Fisher Discriminant Ratio(FDR)) and four classification methods (k-Nearest Neighbour(KNNC), Linear Discriminant(LDC), Support Vector machine(SVC) and Decision Tree(DT)).

Experiments have been performed on both Gender Independent and Gender Based data of the DAICWOZ dataset. The proposed method has given high performance in terms of f1-score of depression in comparison to the previous work in the domain of speech based depression detection. Moreover, our proposed method has also outperformed some multimodal depression detection systems, indicating the strength of speech as an effective unimodal based depression detection system. The present study strengthens the importance of the hybrid of Temporal, Spectral and Spectro-Temporal speech based features set to obtain a successful depression detection system. Further, the application of the univariate feature selection techniques to obtain the relevant set of speech features, enhances the efficacy of the decision model.

Acknowledgement

It would have not been possible to undertake and complete this dissertation without the ~~illuminating~~ guidance of several people whose assistance was valuable at different times.

Firstly, I would like to express my sincere thanks to my supervisor, Prof. R. K. Agrawal, School of Computer & Systems Sciences, JNU, New Delhi for providing the necessary facilities to carry out quality research. I am thankful to him for giving me the opportunity to work on an important medical topic like depression using machine learning. His innovative ideas and suggestions were a great help in driving this research forward.

I would like to express my heartfelt gratitude to Dr. Baljeet Kaur, Swati Rathi, Snigdha and Harsh for their valuable support and for always being there when I needed their help. I would like to thank all my friends for their advice when I needed their assistance in my dissertation work.

Last but not the least, I am very grateful to the entire staff of SC & SS, JNU, for being extremely cooperative during my dissertation work.

Surbhi Sharma

Contents

1 Introduction 5

- 1.1 Depression 5
- 1.2 Objective Measures for Depression Detection 6
- 1.3 Related Work 8
- 1.4 Motivation 9

2 Acoustic Features 11

- 2.1 Temporal Features 11
- 2.2 Spectral Features 13
- 2.3 Spectro-Temporal Features 19

3 Feature Selection and Classification 21

- 3.1 Feature Selection 21
- 3.2 Classification 22

4 Experimental Results 25

- 4.1 Experimental Setup 25
- 4.2 Observations and Discussion 26

5 Conclusion 34

List of Tables

4.1 Hold out Result of Gender Independent for set S, T and ST 27
4.2 Hold out Result of Gender Dependent(Female) for set S, T and ST 28
4.3 Hold out Result of Gender Dependent(Male) for set S, T and ST 29
4.4 Hold out Result of Gender Independent for set A and C 31
4.5 Hold out Result of Gender Dependent(Female) for set A and C 31
4.6 Hold out Result of Gender Dependent(Male) for set A and C 32
4.7 Comparative table of results of the proposed method with the State of the Art 33

Chapter 1

Introduction

1.1 Depression

In recent times, depression has emerged as the most challenging health concern for our society. Physicians define depression as a psychiatric mood disorder in which an individual is unable to cope with stressful life events and is troubled with persistent feelings (for longer than a two-week period) of sadness, negativity and difficulty in carrying out everyday responsibilities. Depression has been identified as the fourth biggest cause of disability by World Health Organization (WHO) in 2012[1]. Moreover, it is predicted to be the second most leading cause of disability by 2030. In the year 2014, WHO declared that 800,000 people die from suicides every year. Out of the total global depressed population, 18% belongs to India[1]. As per report, 4.5% of India's population has been diagnosed with depression in the year 2015. ASSOCHAM reports [2] that 42.5% of corporate employees in India suffer from depression. It has been reported that depression and hectic work hours drives at least one Delhi cop to suicide every month [3]. The economic impact of depression is effected by non productivity of individuals and increasing suicide rates.

As depressed people are more prone to suicidal tendency, therefore, it becomes essential to diagnose depression at an early stage so that timely prevention can be enforced and well wishers of the depressed people can be informed. The American Psychiatric Association publishes a well acclaimed manual to diagnose mental disorders: The Diagnostic and Statistical Manual Of Mental Disorders (DSM) [4]. The DSM gives standard symptoms to identify depression, viz. psychomotor retardation or agitation, fatigue or loss of energy, insomnia or hypersomnia, significant weight loss or weight gain, recurrent thoughts of death and feeling of worthlessness.

There are some well known subjective measures to diagnose depression used by clinicians. One of the measure is the Hamilton Rating Scale for Depression (HAMD)[5] which is an interview style measure used by clinicians. HAMD consists of 21 questions. Each question has a score based on importance of symptom level. Based on their responses an average score is computed and depression is categorized as normal, mild, moderate, severe and very severe.There are some self assessment measures as well. The Beck Depression Index (BDI) [6] and the eight-item Patient Health Questionnaire depression scale (PHQ-8)[7] are the self assessment measures based on self likes and dislikes. PHQ-8 consists of 8 questions. Self assessment measure takes less time to perceive depression in comparison to clinical based questionnaire.

1.2 Objective Measures for Depression Detection

Subjective measures are biased as they are dependent upon the clinician's interpretation of the patient's response. Largely patients are not cooperative and do not express themselves honestly with the clinician, scores are also unreliable as at times the patients are familiar with the repeated questions. So there is a need to explore some objective biomarkers which are reliable and which can quantify the extent of problem that an individual is facing. In recent years, researchers have explored mental state in conjunction with speech[8], gestures [9], gait [10],video[11], body language [9] and have also used Natural Language Processing [12] techniques and found a correlation between the depression severity and the result of these measures. The depression detection systems that combine the features of two or more modalities (multimodal systems) are capable of capturing the psycho-motor as well as the psycho-cognitive impairments that are manifested in people suffering from depression. Such systems are able to capture the various aspects of behaviourial markers of depression, and they take the advantage of the individual modality. But, capturing multimodal data is tedious and expensive due to the involvement of the visual-recording equipment and the reluctance of the individual in the process where his/her identity is exposed. Futher, these methods are computationally intensive, therefore there is a need to strengthen the unimodal systems, which are cost effective and are computationally inexpensive. Among the various unimodal systems explored by researchers, speech based depression detection involves low cost of acquiring the data and low cost of the learning model. Also, this system has the added advantage of being non-invasive and does not expose the individual's identity. Therefore, unimodal system based on speech is a strong contender for depression detection. In literature, many features are extracted from speech and their correlation has been established with depression detection.

Speech is a non-invasive medium and is strongly correlated with depression symptoms. It is well understood that mental stress has an immediate impact on the tenseness of the vocal chords and the tract [13], resulting in change of the quality of speech of an individual. The vocal tract and chords get tensed due to cognitive impairments and stress exposure associated with depression [13], leading to changes in acoustics, articulation and rate of speech. Therefore, speech can be used as an objective biomarker to diagnose depression. Speech is a medium to understand the state of the mind of a person and this opens up possibilities of exploring various features that can reliably predict the severity of depression of an individual. Speech has shown positive results as a discriminative measure which can distinguish a depressed individual from the control [8, 11, 13, 14, 15, 16].

Various feature extraction methods [15] are employed to extract features like Mel Frequency Cepstral Coefficients (MFCCs), formants, Normalized amplitude quotient (NAQ), Quasi-open quotient (QOQ) and fundamental frequency (F_0). Feature extraction is the process of transforming raw speech signal to more compact representation with limited redundancy. Features from speech signal can be extracted in Time domain (Temporal features), Frequency domain (Spectral features) and Spectro-Temporal domain.

Some of the time domain features suggested in literature [8] to identify depression state are: short time energy [14], short time zero crossing count (ZCC)[8], pitch period[14], linear predictive coefficients (LPC)[17] and loudness [14].

Some of the Spectral features (frequency domain features) [17] suggested in literature are Mel Frequency Cepstral Coefficients (MFCCs) [18], Spectral-Centroid [17], Empirical Mode Decomposition (EMD) [19], Energy Slope [17], Jitter[20, 21], Shimmer [20, 21], Harmonic to Signal Noise Ratio (HNR) [22] and Power Spectral Density (PSD) [16] .

Some of the Spectro-Temporal features suggested in literature are Maxima Dispersion Quotient (MDQ) [23] and Peakslope [24] which measure both temporal as well as spectral information.

Though researchers have explored a few temporal/spectral features, no work has been done that explores an exhaustive set of features. We have explored the exhaustive set of Temporal, Spectral and Spectro-Temporal features. We have conducted experiments to understand the importance of just the temporal or the spectral featutres, and also a combination of all features, in order to establish the role of these features for depression detection. Speech based unimodal depression detection system, which is non-invasive, involves low cost and computation time in comparison to multi-modal systems. The performance of any decision system mainly depends on the choice of feature selection method and the classifier. We have investigated the combination of three well-known univariate filter methods (Fisher Discriminant Ratio, Pearson Correlation Coefficient and Mutual Information) and four well-known classifiers (k-Nearest Neighbour (KNNC), Linear Discriminant classifier (LDC), Decision Tree (DT), Support Vector Machine (SVC)) to obtain a minimal set of relevant features to improve the performance. This will speed-up the acquisition of features from speech and build the decision system with low cost and complexity.

1.3 Related Work

Speech is a natural, non-invasive evidence that has been investigated for depression over the years and has been found indicative in classifying depressed people from the control group. The tenseness of the vocal tract and the vocal chords in depressed people translates to a set of features that are well known and have been investigating over the years.

France et al. [16] worked with the fundamental frequency (F_0), Amplitude Modulation (AM), formants, and Power Spectral Density (PSD) for discriminating control, dysthymic and major depressed persons. Formant frequencies and PSD were successful with 0.94 accuracy to distinguish between the control and depressed female patients and 0.82 between the male patients.

Cummins *et al.* [8] have shown merits of voiced and unvoiced segments considering short term energy and fundamental frequency (F_0), ZCC etc. Spectral features (frequency domain features) were also used to classify depressed from control. The spectral features considered in this research are mel frequency cepstral coefficients (MFCCs) and linear predictive group delay. Features normalization was used to reduce the features's range mismatch of different speakers. The classification accuracy with Gaussian mixture model (GMM) was found to be 80 percent considering MFCC as feature for speaker dependent case and 77 for speaker independent case.

Alghowinem *et al.* [14] extracted low level descriptors features from a frame duration of 25 ms. They have evaluated linear features: fundamental frequency (F_0), intensity, loudness, voice probability, quality, jitter, shimmer, Harmonics to Noise Ratio (HNR), log energy, root mean square energy and the non-linear teager energy to understand the difference in voiced/unvoiced and mixed speech on depression detection. They suggested the suitability of teager energy operator (TEO) based features for depression detection using voiced and unvoiced speech. Voicing probability and log energy were found to be more suited for mixed speech. They investigated high jitter, lower shimmer, high HNR(Harmonic Noise to Ratio), lower vocal energy in glottal pulse of depressed subjects and lower range of fundamental frequency (F_0). The GMM was used for dimension reduction and then Support Vector Machine (SVM) was used to classify depressed and control subjects.

Scherer *et al.* [9] employed a multimodal framework in which video features along with acoustic features were used to classify depressed individual. The acoustic features that were extracted mainly focused on calculating breathiness and tenseness of voice. Normalized amplitude quotient (NAQ) and quasi-open quotient (QOQ) both are derived from amplitude measurement of glottal flow pulse. It has been deduced that both measures are inversely correlated with the tenseness of the voice. The smaller the value the more tense the voice will be and thereby are highly correlated with each other.

Ozdes *et al.* [25] explored the significance of vocal jitter and spectral slope as indicators of suicidal tendencies. The pairwise classification accuracies among control/depressed, depressed/suicide, control/suicide using jitter was 0.65, 0.60 and 0.80 respectively and using spectral slope was 0.90, 0.75 and 0.60 respectively. Using both jitter and spectral slope, the accuracies reported were 0.90, 0.75 and 0.85 respectively for the three pairs.

Sethu *et al.* [17] evaluated emotions based on pitch, energy slope and formants for speaker dependent and speaker independent studies. Mel Frequency Cepstral Coefficients (MFCCs) and Linear Prediction Coefficient (LPC) based group delay performed well in speaker dependent system, while the first three formants performed well in the speaker independent system.

The study by Scherer *et al.* [26] involved the analysis of Normalized Amplitude Quotient (NAQ), Quasi-open-Quotient (QOQ), peakslope and open Quotient Neural Network (OQNN). It was observed that these features are independent of gender 0.75 accuracy was obtained with support vector machine (SVM) as classifier.

In 2016, Pampouchidou *et al.* [11] used a fusion of low level features and discrete cosine transform (DCT) based features and high level features. They analysed two sets of size 494 and 1278 respectively using the DAICWOZ dataset and the COVAREP feature repository. The gender based results using low level descriptors are reported in terms of f1 depressed(nondepressed) as 0.45(0.85) (leave-one-out cross validation (LOOCV)) and 0.59(0.87) (testing using development set).The DCT based features for gender independent data gave 0.19(0.71) and 0.47(0.83) in terms of f1 measure respectively. In 2017, Pampouchidou et al. [27], evaluated the AVEC dataset [28] using the COVAREP based audio features and reported a precision of 0.948 while using visual OR(ed) with audio gender based features. Audio alone gave the f1 score of 0.641.

In 2017, Cummins *et al.*[29] performed depression detection using eGeMAPS, COVAREP, gender dependent VL-Formants, and a fusion combination. Performance was measured in terms of f1-score for depressed (non-depressed) classes. The overall f1 depressed was calculated to be 0.63(0.89). Results were performed on the training and development partitions of the DAICWOZ Corpus. All classifications were performed using the Liblinear package [30] using grid search.

In 2018, Takaya Taguchi *et al.* [31] investigated the second dimension of the Mel Frequency Cepstral Coefficients (MFCCs) for depression detection. They found it highly discriminatory for classifying depressed subjects from control. An accuracy of 81.9 % was obtained using second dimension of MFCCs feature as a biomarker.

1.4 Motivation

After going through various research works of speech based depression detection system, it was noted that most of the research work mainly focused only on the feature extraction techniques, [11, 14, 16, 17, 25, 26]. Some researchers have used only one or few of the Temporal features, Spectral features but none of them has exhaustively explored the Temporal features, Spectral features and combination of the two including Spectro-Temporal features. This has motivated us to investigate these features, in order to establish the role of these features for depression detection. It is also well-known that if appropriate combination of feature selection method and classifier is not used to build the decision system, the performance of the learning system may degrade. In one of the recent work by Pampouchidou et al. [11], a large set of both low and high level features from speech has been explored. The importance of individual features was noted by removing each feature or set of features based on the f1 score using decision tree classifier. However, this wrapper approach is computationally intensive. In literature, filter feature selection methods in combination with a classifier are found to be more successful to find a smaller set of relevant features to improve the performance in many domains such as microarray based cancer classification, object recognition and text analysis [32]. This combination of filter feature selection and the classifier is computationally less intensive. But, in speech based depression detection, filter feature selection method has not been investigated much to improve the performance of the depression detection system, to the best of our knowledge. Motivated by this, we have investigated the univariate filter methods (Fisher Discriminant Ratio, Pearson Correlation Coefficient, Mutual Information) to obtain a minimal set of relevant features to improve the performance of depression detection system.

In literature, researchers have investigated one or two classifiers. This has also motivated us to use four classifiers. We have used four well known classifiers (k-Nearest Neighbour, Linear Discriminant, Decision Tree classifier, Support Vector Machine) to build the learning model.

For our experiments we have used the Temporal and Spectral features individually as well as in combination. Audio features from DAICWOZ [33] repository are used along with our set of Temporal and Spectral features. The features so chosen has also performed better than many speech based system and a few multimodal systems. The f1-score is calculated to analyse the efficacy of the decision system.

Chapter 2

Acoustic Features

To work on the raw dataset is computationally intensive in the domain of machine learning therefore, it becomes essential to extract the features from raw dataset. The dataset used is DAICWOZ [33] which consists of 42 depressed and 100 non depressed subjects. Relevant features extracted from the raw dataset are categorized as Temporal features,Spectral features and Spectro-Temporal features. Our task is to work on these desired features which results in reduced memory space and hence classification of depressed subjects from non-depressed ones becomes easy. Some features have been extracted from the COVAREP [34] and additional Temporal and Spectral Features which are not given in COVAREP have also been explored as they have been found to be effective in discrimination of depressed ones from non-depressed.

2.1 Temporal Features

Time domain analysis is a way to efficiently represent the speech parameters. Many speech parameters can be quickly calculated from time domain analysis. Speech can be analyzed in time domain with easy physical interpretation. It represents the raw speech signal into more compact form with limited redundancy. Time domain analysis provides efficient storage and manipulation of the relevant speech signal where not losing the raw signal properties. Speech signal is categorized as voiced and unvoiced. Speech signal is composed of phonemes, which are produced by vocal cords and vocal tract. Voiced signal is produced when vocal cords vibrate, like in the pronounciation of vowel phonemes (/a/, /e/, /i/, /u/, /o/). However, unvoiced signal does not entail vibration of vocal cords like in pronounciation of the plosives consonants(/p/, /t/, /k/). Parameters of speech signal are sampled at lower rate to capture relevant information from it.

Several speech parameters like Energy [14], Zero Crossing Count (ZCC) [8], Normalized Amplitude Quotient (NAQ) [35], Quasi Open Quotient (QOQ) [36] etc. are calculated in time domain. Speech analysis technique presumes that properties of the signal remain stationary for 10 ms to 20 ms. Therefore, signal must be divided into successive frame to capture the relevant changes. We can achieve the splitting of signal into frames by multiplying the signal with an appropriate sized window. Therefore, choice of window becomes an important factor here. Various types of window like rectangular, Hamming window etc. impart smoothing and low filter effect to speech signal which is desirable to do the short time processing of the speech signal. The speech signal S(r) multiplied by the window function W(n) undergoes transformation (T) to give output signal H(n) at original sample rate. Short time processing formula is given as follows.

$$H(n) = \sum_{r=-\infty}^{r=\infty} T[S(r) \times W(n-r)] \tag{2.1}$$

H(n) is the smoothed version of speech signal.

.

2.1.1 Normalized Amplitude Quotient (NAQ): The normalized amplitude quotient parameter [35] is a voice quality source parameter that helps to distinguish between breathy, modal and tense voice qualities. It is estimated from the glottal speech waveform derivative. It is defined as the ratio of the largest peak-to-peak amplitude and the largest amplitude of the cycle to cycle minimum derivative.

2.1.2 Quasi Open Quotient (QOQ): QOQ [36] is a voice quality feature stating instants of the glottal opening phase . It is the duration during which the glottal flow is 50 % above the minimum flow, normalized to the pitch period. It is helpful in studying the physical emotional changes from glottal source.

2.1.3 Energy: The total squared amplitude values of each frame is called energy.

$$\begin{gathered} E = \sum_{n} [S(n) \times W(m-n)]^2 \\ W(n) = \begin{cases} 0, & \text{if } 0 \leq n \leq N \\ 1, & \text{otherwise.} \end{cases} \end{gathered} \tag{2.2}$$

Window starts at sample m where W(n) refers to window function.There is a large variation in amplitudes of voiced and unvoiced signal which makes energy an important speech parameter. The voiced signal has high energy and unvoiced signal has low energy.

2.1.4 Zero Crossing Count (ZCC): The ZCC measures number of times the speech signal S[K] crosses the time axis. Indirectly we can say that it depicts the frequency content of the frame. For voiced signal in which the vocal cord vibrates, the energy content is high, therefore, ZCC will be low. However, for unvoiced signal in which vocal cord does not vibrate, the energy content is low, therefore, ZCC will be high. The ZCC provides spectral information at low computation cost. The ZCC is computed as follows.

$$ZCC = \sum_{K=1}^{N-1} |0.5 \times (sign\ S[K] - sign\ S[K-1])| \tag{2.3}$$

2.1.5 Intensity: Intensity is the the energy carried by the sound waves per unit area.

2.1.6 Loudness: Loudness of sound (in air) is generally reported as Sound Pressure Level (SPL) in decibels. SPL (in dB) and pascals are related as

$$Pa_{ref} = 20 \times 10^{-6} \tag{2.4}$$

$$SPL_{dB} = 10 \times log_{10} \times \frac{Pa^2}{Pa_{ref}^2} \tag{2.5}$$

To get a time-smoothed SPL, the standard approach is to extract the envelope of the pressure signal, rectify and low-pass filter.

The above mentioned temporal features have been explored [8] and correlation has been found with depressed speech.

2.2 Spectral Features

The speech parameters can be analyzed more effectively in frequency domain as compared to time domain. The repeated utterances of the same speaker remain quite similar in the frequency domain as opposed to in the time domain. Therefore, spectral analysis becomes primarily important in extracting relevant speech parameters. Fourier analysis is used to represent the raw speech signal in terms of frequency components and amplitude. Several speech parameters like Energy Slope [17], Mel Frequency Cepstral Coefficients(MFCCs) [18], Spectral Centroid [17], Empirical Mode Decomposition [37], Jitter and Shimmer [20, 21], Harmonic to Signal Noise Ratio(HNR) [22] and Power Spectral Density [16] etc. are calculated in frequency domain. As we know that speech signal is non stationary and slowly varies with time, therefore, short-time Fourier analysis becomes important using a suitable window function. The short-time Fourier transform is defined as follows:

$$S_n(e^{j\omega}) = \sum_{t=-\infty}^{t=\infty} [s(t) \times e^{-j\omega m} \times W(n-t)] \tag{2.6}$$

where s(t) refers to speech signal multiplied by complex exponential frequency shift of ω radians and W(n) refers to window function and n is the starting point of window. $S_n(e^{j\omega})$ reflects the amplitude and phase of s(t) within the bandwidth of the window centered at ω. Emotional state of the person results in changes in the speech spectrum which can be effectively captured by the spectral features. Moreover, speech consists of different ranges of frequencies which can be captured in the frequency domain. Spectral features have been found discriminating for classification of depressed subjects from non-depressed subjects. The spectral features explored are given as follows:

2.2.1 Fundamental Frequency(F_0): It can be calculated in time domain as well as in frequency domain. The F_0 in time domain is calculated from autocorrelation and average magnitude difference function.

(i) Autocorrelation: It measures the similarity between two signals (S[n] and S[n+k]) where k refers to delay of samples, which is defined as:

$$\phi(k) = \frac{1}{N} \sum_{n=0}^{N-1} (S[n] \times S[n+k]) \tag{2.7}$$

For distinct values of k delay $\phi(k)$ would give different value resulting in different peaks at delay of k = 0, P, 2P ... where P is pitch period. The distance between two peaks of the samples is the pitch period. Autocorrelation is computationally more intensive as multiplication is complex operation.

(ii) Average Magnitude Difference Function(AMDF): The magnitude of the difference of speech signal S[n] and S[n+k] is taken which would give different minimum values at k = 0, P, 2P ..., where P refers to pitch period. The distance between the two minima would be the pitch period. AMDF is considerably faster than autocorrelation as subtraction is simpler operation. AMDF is defined as:

$$D(k) = \frac{1}{N} \sum_{n=0}^{N-1} (S[n] - S[n+k]) \tag{2.8}$$

In literature [34], the Summation of the Residual Harmonics (SRH) [38] method which is a pitch tracking algorithm has been used to calculate the F_0.

2.2.2 Parabolic Spectral Parameter (PSP):Parabolic Spectral Parameter (PSP) [39] is used for the quantification of the glottal volume velocity waveform. It relies on the low-frequency part of computed spectrum of the estimated glottal flow. It gives a single value that describes the spectral delay of the glottal flow.

2.2.3 Difference in amplitude of the first two harmonics of the differentiated glottal source spectrum($H1H2$): $H1$ is the amplitude of the first harmonic and $H2$ is the amplitude of the second harmonic [34] . Difference of the amplitude of two harmonic gives idea about the spectral shape that is helpful in estimating the voice quality.

2.2.4 Harmonic Model and Phase Distortion Mean (HMPDM) and Deviation (HMPDD): The adaptive harmonic model (aHM) [40, 41] represents speech through the relevant spectral information and noise. The representation of phase is not handled by the aHM. Parameterization methods tend to discard the phase information which has been taken care of by the parameters: HMPDM and HMPDD [40, 41]. The instantaneous phase from the waveform is obtained and the minimum-phase term is subtracted from the measured phases to get the Phase Distortion (PD). The short-time mean (HMPDM) and standard deviation (HMPDD) of the PD are computed in the neighbourhood of each frame. HMPDM is highly correlated to the maximum-phase component and the HMPDD is correlated to the degree of noisiness.

2.2.5 Formants: The air inside the vocal tract vibrates due to its varying cross-sectional area as well as due to the closing and opening of vocal folds of the vocal tract that results the sound to vibrate at different frequencies called formants. The vocal tract's cross sectional area is controlled by jaw, tongue,teeth and lips. Every phoneme creats a different configuration of the above articulators resulting in resonance of the sound . Therefore, formants exist for both voiced and unvoiced sounds. Generally, first three formants are found to be predominant in depression detection.

2.2.6 Mel Frequency Cepstral Coefficients(MFCCs): MFCCs [18] combine the advantage of the Cepstrum analysis with a frequency scale(Mel scale) based on human hearing perception. The Cepstrum coefficients and the Mel Frequency Scale that define the hearing perception are explained below.

2.2.6.1 Cepstral Analysis: The Cepstrum is defined as the inverse Discrete Fourier Transform (DFT) of log magnitude of the DFT of the signal (s[t]). It is defined as follows:

$$C[t] = F^{-1}(\log|F(s[t])|) \tag{2.9}$$

Here, F corresponds to Fourier Transform of the signal and F^{-1} corresponds to the inverse Fourier Transform of the signal. Log spectrum helps to reduce the amplitude differences in the harmonics in the spectral domain. The inverse DFT of the log spectrum brings the signal in Cepstral domain which is useful for separating the glottal excitation and vocal tract induced frequencies. The coefficients thus obtained in Cepstral domain are known as Cepstral coefficients, and are generally decorrelated and are widely used in speech recognition. Cepstral analysis is used to find local periodicity of the signal, in detecting voiced speech segment and unvoiced speech segment. The presence of strong peak denotes that speech segment is voiced.

2.2.6.2 Mel Frequency Scale: The human ear can not discern the difference between the two closely spaced frequencies. This phenomenon is observed as the frequency increases. To estimate the energy in the different frequency regions, a set of filter banks known as the Mel filter bank based on the Mel scale of human hearing perception are generated. In the Mel filter bank, the first filter is very narrow and indicates the energy content near zero hertz. As the frequencies increase, the filter banks get wider and an estimated energy in that band can be calculated. The perception of frequency of human auditory system is not linear. The Mel scale defines the spacing of the filterbank. Mel scale is defined as follows:

$$F_{mel} = 1125 \ln \times [1 + \frac{F}{700}] \tag{2.10}$$

where F_{mel} is the frequency measured on the Mel scale and F corresponds to frequency measured in Hz.

Using the concepts of Cepstral analysis and Mel frequency scale the steps to calculate the MFCCs are given as follows:

(i) Segment the signal into frames.

(ii) Compute the complex DFT of each frame as given below:

$$S_i(k) = \left\{ \sum_{t=0}^{N-1} s_i(t) W(t) e^{\frac{-j2\pi kt}{N}}, \quad 1 \leq k \leq K \right\} \tag{2.11}$$

where $W(t)$ refers to window function, $s_i(t)$ corresponds to speech signal and K refers to length of DFT.

(iii) Compute the power spectrum given as follow:

$$P_i(k) = \frac{1}{N} |S_i(k)|^2 \tag{2.12}$$

(iv) Compute the Mel-spaced filterbank. Mel-scale relates the perceived frequency to the actual measured frequency.

(v) Apply the set of filters to the power spectrum. Multiply each filter bank with power spectrum and add-up the coefficients.

(vi) Take log of each of filter bank energies computed above.

(vii) Cepstral analysis is performed on the log of the filter bank energies to extract Mel spaced Cepstral coefficients.

2.2.7 Energy Slope: The ratio of energy in the low frequency band to high frequency band[17] is defined as Energy Slope. The low frequency band corresponds to 0-1 kHz.

2.2.8 Spectral Centroid: Spectral Centroid[17] is the spectral measure of the speech which provides spectral magnitude information. It is defined as follows:

$$SpectralCentroid = \frac{\sum_{i=1}^{N} |S(i)| \cdot f_s \cdot i}{N \sum_{i=1}^{N} |S(i)|} \tag{2.13}$$

where N refers to frame size, *S(i)* is the Discrete Fourier Transform (DFT) of the framed signal and f_s corresponds to sampling rate.

2.2.9 Empirical Mode Decomposition (EMD): EMD [37] is used to decompose the non stationary signal into components also known as Intrinsic Mode Functions (IMFs). The signal s(t) is decomposed into Intrinsic Mode Functions(IMFs) which are characterized by the following features:

(i) The difference of number of extrema and zero crossing must be atmost one.

(ii) At any point in the envelope of maxima extrema and minima extrema the mean value should be zero.

The IMFs are computed recursively. The algorithm for generating the IMF is described below:

(i) Identify all local maxima extrema points in the signal and join them by cubic spline curve.

(ii) Identify all local minima extrema points in the signal and join them by cubic spline curve.

(iii) Compute the mean of each data point enclosed between the two envelopes and denote it as m_1.

(iv) Subtract the above computed mean m_1 from the signal

$$h_1 = s(t) - m_1 \tag{2.14}$$

(v) Continue the above steps(i-iv) and consider h_1 as input signal until it can be considered as an IMF as per the definition stated above.

(vi) h_1 is the desired IMF_1. The residue r_1 is obtained by subtracting IMF_1 from s(t) i.e. $r1 = s(t) - IMF_1$. The residual of this step becomes the signal s(t) for the next iteration to calculate the next IMF.

(vii) Iterate the steps(i-vi) on the residual r_j, j = 1, 2, 3... n in order to find all the IMFs of the signal

(viii) The algorithm stops when the value of standard deviation in the IMFs is less than a pre-assigned value.

2.2.10 Linear Predictive Coefficients (LPC): Linear Predictive analysis [42] is one of the fundamental speech analysis techniques. This method is used for estimating speech parameters such as pitch, formants, spectra, vocal tract area functions and for representing the speech signal compactly. It works on the principal that speech sample can be represented in terms of linear combination of past speech samples. It is given by the following difference equation where S[n] represents speech sample at nth sample point.

$$S[n] = \sum_{k=1}^{p} a_k S[n-k] + Gu[n] \tag{2.15}$$

where G refers to gain factor or amplification factor of impulse train and *u[n]* refers to excitation. Here, a_k are called predictive coefficients. The predictive coefficients can be determined in terms of autocorrelation coefficient by solving p simultaneous linear equation. As voiced speech is periodic, *Gu[n]* will be zero between pitch pulses of voiced impulse train. Therefore, *S[n]* can be predicted as weighted combination of past speech sample during this interval. However, *Gu[n]* factor is present in case of unvoiced speech as it is not periodic, thus speech signal to be predicted approximately. This technique has been used to find F_0 and other relevant features.

2.2.11 Jitter: Jitter[[20],[21]] is defined as the variation of the fundamental frequency from one cycle to another cycle. Jitter is defined as follows:

$$Jitter = \frac{1}{M-1} \sum_{i=1}^{M-1} |T_i - T_{i+1}| \tag{2.16}$$

M refers to the number of the extracted periods and T_i refers to the extracted period of the fundamental frequency.

2.2.12 Shimmer: Shimmer[[20],[21]] is measurement of amplitude to amplitude variability over consecutive periods being divided by average amplitude. Shimmer is defined as follows :

$$Shimmer(dB) = \frac{\frac{1}{M-1}\sum_{i=1}^{M-1}|A_i - A_{i+1}|}{\sum_{i=1}^{M} A_i} \tag{2.17}$$

Where M refers to the number of extracted periods and A_i refers to the extracted peak to peak amplitude.

2.2.13 Harmonic to Signal Noise Ratio (HNR): The Harmonic to Signal Noise Ratio [22] is the measurement of the signal to noise ratio. It quantifies the noise added to the signal. It is measured over the(dB) scale. The lower the HNR more would be the noise and vice - versa.

2.2.14 Power Spectral Density (PSD): Power Spectral Density [16] shows the variation of the energy per frequency. It shows at which frequencies energy variations are strong and at which frequencies energy variations are weak. One of the method to calculate (PSD) is to compute the Fourier transform of the sampled signal x(n) where $n = 0, 1, 2 \cdots M - 1$ which is given as follows:

$$P_{xx}(f) = \frac{|X(f)|^2}{M} \tag{2.18}$$

where

$$X(f) = \sum_{n} x(n) \exp^{-i2\pi f} \tag{2.19}$$

The another method to calculate PSD is in terms of Fourier transform of the autocorrelation given as follows:

$$P_{xx}(f) = \sum_{k} r_{xx}(k) \exp^{-i2\pi f} \tag{2.20}$$

where r_{xx} refers to autocorrelation of the sampled signal.

2.3 Spectro-Temporal Features

2.3.1 Maxima Dispersion Quotient(MDQ) : MDQ [23] is a indication of voice quality feature from breathy to tense voice. It depicts glottal source dynamics. The output signal obtained from wavelet based filtering from glottal excitation helps in capturing varying voice quality features from breathy to tense voice that would be helpful in detection of the depression.

2.3.2 PeakSlope: Peakslope [24] is also a voice quality parameter indicative of the breathy to noise scale. It includes the wavelet based decomposition of the speech signal and fitting the regression line of maxima obtained on different scales. The slope of the spectral tilt indicates the voice quality as breathy, modal and tense noise.

2.3.3 Rd and Rd conf : The common parameters that characterize the main shape of the glottal flow are the rise time, decay time and the open quotient [43]. The duration of the return phase T_a is inversely proportional to the degree of the spectral tilt.

$$Fa = \frac{1}{2\pi T_a} \tag{2.21}$$

In the LF model [43], fewer parameters retain the wave shape. U_o is the peak value of the oscillatory component of glottal flow and it is closely related to the amplitude of the voice fundamental. E_e, is the flow derivative at the point excitation and it is the basic determinant of formant amplitudes. Dependencies of E_e and U_o on F_o are different for males and females. The various properties of the glottal wave are quantified as a single shape parameter,

$$Rd = \frac{U_o}{E_e}\frac{F_o}{110} \tag{2.22}$$

Chapter 3

Feature Selection and Classification

3.1 Feature Selection

Once a temporal and spectral feature set is obtained it becomes imperative to reduce the dimensionality of the feature set to handle the curse of dimensionality [44]. Feature selection methods to reduce the dimensionality are categorized in two major approaches: Filter and Wrapper [45, 46]. Wrapper methods are computationally intensive due to the repeated training of each candidate subset. On the other hand, filter methods do not involve any learning algorithm to measure the importance of features, hence are simple and computationally less intensive. They are further subdivided into univariate and multivariate methods [45]. Univariate filter methods measure the relevance of the individual features based on the statistical characteristics of the feature and ultimately collect the top ranked relevant features. Univariate feature selection method used are Mutual Information Criterion Pearson Correlation Coefficient and Fisher Discriminant Ratio(FDR).

3.1.1 Mutual Information Criterion : It is a feature ranking method which computes the non-linear correlation between feature and class label. Mutual Information Criterion is calculated as follows:

$$MI(x,y) = \sum P(d_i, f_i) * \log \frac{P(d_i, f_i)}{P(d_i) * P(f_i)} \tag{3.1}$$

3.1.2 Fisher Discriminant Ratio(FDR) : It is a univariate feature ranking method in which mean separation between samples of two classes is high and variation within the same class is low. FDR is defined as follows:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}} \tag{3.2}$$

where μ_1 and μ_2 are respective mean of samples of two classes and σ_1 and σ_2 are standard deviation of respective classes. FDR would be high when dispersion between the two classes is high and variance within the class is minimal.

3.1.3 Pearson Correlation Coefficient(PC) : It is used to calculate the linear correlation between feature (x_i) and class label (c). Features are ranked in descending order of the scores. PC is given as follows :

$$PC(x_i, c) = \frac{n\sum x_i * c_i - n\overline{x} * \overline{c}}{\sqrt{x_i^2 - n\overline{x^2}} * \sqrt{c_i^2 - n\overline{c^2}}} \tag{3.3}$$

3.2 Classification

3.2.1 K-Nearest Neighbour(KNNC) :K-nearest neighbour classifier[47] is a non-parametric classification that classifies the new test sample by comparing against the training samples in terms of a similarity or distance measure. The class label of a test sample is determined in terms of majority of class labels of its K nearest neighbours. KNN is compu-tationally intensive when the number of training samples is large.

3.2.2 Linear Discriminant Classifier(LDC) :Linear discriminant classifier[48] assigns a feature vector x to class c with a set of discriminant functions. It generally uses the simplified linear discriminant function, which is given as

$$H_i(x) = \log P(c_i|x) \tag{3.4}$$

Where $P(c_i|x)$ represents the posterior conditional probability. The class label of a given sample x is determined in terms of $arg(max(H_i))$. LDC assumes the data should follow multivariate normal distribution.

3.2.3 Support Vector Machine (SVM) : A support vector machine [49] determines a decision boundary that maximizes the distance between samples of two classes. Generally, it works as a linear classifier by constructing an optimal hyperplane that may classify data in such a way that maximizes the margin of separation between the two classes. Let us consider a point z in the d-dimensional feature space with class labels in [-1,1]. We can define the two hyperplanes for separating the two classes given as follows:

$$w^T z_k + m \geq 1 \ \forall y_k = +1 \tag{3.5}$$

$$w^T z_k + m \leq -1 \ \forall y_k = -1 \tag{3.6}$$

We can combine the above two equations and rewrite them as follows:

$$y_k(w^T z_k + m) \geq 1 \tag{3.7}$$

For the case where the data is linearly separable, we can define two parallel hyperplanes as follows:

$$H1 : w^T z_k + m = 1 \tag{3.8}$$

$$H2 : w^T z_k + m = -1 \tag{3.9}$$

The perpendicular distance between above two hyperplanes, is represented as follows:

$$K = \frac{2}{\|w\|} \tag{3.10}$$

The main motive of svm is to maximize the distance between the two hyperplanes.This maximization problem can be stated as the minimization of the following function of SVM:

$$\phi = \frac{2}{\|w\|} \tag{3.11}$$

subject to the criterion:

$$y_k * (w^T z_k + m) \geq 1 \; \forall k = 1......n \tag{3.12}$$

For this optimization problem, to find the suitable weights Lagrange function is given as:

$$L(w, m, \alpha) = \frac{1}{2} \left\| w^2 \right\| - \sum_{t=1}^{n} \alpha_t y_k (w^T * z_k + m) - 1) \tag{3.13}$$

Where α_k are Lagrange's multipliers. The dual version of above problem can be given as

$$Maximize \; R(\alpha) = \sum_{k=1}^{n} \alpha_k - \frac{1}{2} \sum_{k=1}^{n} \sum_{j=1}^{n} \alpha_k \alpha_j y_k y_j z_k^T z_j \tag{3.14}$$

Subject to the criterion:

$$\sum \alpha_k y_k = 0 \tag{3.15}$$

$$\alpha_k \geq 0 \, \forall k \tag{3.16}$$

The optimal weight vector is given as follows:

$$w = \sum_{k=1}^{n} (\alpha_k y_k z_k) \tag{3.17}$$

Where α_k are Lagrange's multiplier coefficient. The decision function is given as

$$f(x) = w^T z_k + m \tag{3.18}$$

If g(x) >1, the sample x belongs to class with label +1 and if g(x) <-1, the sample x belongs to the class -1. If the vectors are not linearly separable then there is need to transform the data to higher dimension so that the data becomes linearly separable.

3.2.4 Decision Tree (DT): Decision Tree Classifier[47] is based on non-metric me-thods in which samples at each node are partitioned into two or more homogeneous sets based on most significant distinguisher among available set of input features. Among them entropy and chi-square are most common. Decision tree always helps to formulate a rule based decision process. Due to the visual representation of this classifier, its uses in medical field seem to be very handy for a physician. It is interesting to see, however, that when there is a large number of attributes (determining parameters), the decision to declare a patient with a certain disease becomes a complex task. In that case the classifier must be used only after pruning the branches where a little information is gained. Based on the concept learning system, Quinlan[50] gave a decision tree algorithm, ID3. ID3 is based on the entropy measure based on Shannon's Information Theory. The whole dataset is considered as the node of the tree. The node gets divided based on the attribute which decreases the total entropy (randomness) and gives maximum information gain.

Chapter 4

Experimental Results

4.1 Experimental Setup

The DAICWOZ dataset is used for experimetal setup. It consists of audio files, COVAREP features files, formants file and trancripts files for 42 depressed and 100 non-depressed subjects.

In literature, exhaustive Temporal set, exhaustive Spectral set and combination of these two exhaustive set has never been explored. This has motivated us to work in this direction.

The Exhaustive Temporal ferature set (T) extracted by us consists of Energy [14], Zero Crossing Count(ZCC)[8], Intensity, Loudness, Normalized Amplitude Quotient parameter[35] and Quasi Open Quotient(QOQ) [36].

The Exhaustive Spectral set(S) consists of Energy Slope [17], Mel Frequency Cepstral Coefficients(MFCCs) [18], Spectral Centroid [17], Empirical Mode Decomposition [37], Jitter and Shimmer [20, 21], Harmonic to Signal Noise Ratio(HNR) [22] and Power Spectral Density [16], Linear Predictive Coefficients[34] and Fundamental Frequency(F_0)[34] and Harmonic Model and Phase Distortion Mean (HMPDM) and Deviation (HMPDD)[[40] [41]], formants[34] and Difference in amplitude of the first two harmonics of the differentiated glottal source spectrum($H1H2$)[34].

The hybrid set(ST) consists of Temporal features (T), Spectral features (S) and Spectro-Temporal features. The Spectro-Temporal features consists of Maxima Dispersion Quotient(MDQ)[23],Peak Slope[24] and Rd and Rd_{conf}[43].

The Statistical features of low level descriptors of Temporal set(T), Spectral set(S) and hybrid set(ST) that have been calculated are Mean, Min, Skewness, Kurtosis, Standard Deviation, Median, Peak-magnitude to Root-mean-square ratio, Root mean square level, Interquartile range and Spectral flatness.

(i) **Mean**: Sum of observations divided by number of observations.

(ii) **Min** : Give minimum value among observations.

(iii) **Skewness** : It is a measure of symmetry. $X_1, X_2 \cdots X_N$ are univariate observations. Skewness is given as:

$$S = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^3}{N\sigma^3} \tag{4.1}$$

(iv) **Kurtosis**: It gives measure of the normal distribution in terms of the extent of whether it is rightly tailed or leftly tailed. $X_1, X_2 \cdots X_N$ are univariate observations. It is given as

$$S = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^4}{N\sigma^4} \tag{4.2}$$

(v) **Standard Deviation**: It is a measure of the dispersion of the data $X_1, X_2 \cdots X_N$.It is given as:

$$S = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1} \tag{4.3}$$

(vi) **Median**: Middle number of the sorted dataset.

(vii) **Peak-magnitude to root-mean-square ratio**: It is defined as the ratio of the largest absolute value in $X_1, X_2 \cdots T_N$ to the root-mean-square (RMS) value of $X_1, X_2 \cdots X_N$.

(viii) **Root mean square level**: It gives root mean square level of the feature vector.

(ix) **Interquartile range**: It is a measure of variability of the ordered dataset divided into first quartile (Q_1), second quartile (Q_2) and third quartile (Q_3). Q_1 is the middle value of the first half, Q_2 is the middle value. Q_3 is the middle value in the second half of the dataset. The interquartile range is Q_3 - Q_1.

(x) **Spectral flatness**: The spectral flatness is given as the ratio of the geometric mean of the power spectrum to the arithmetic mean of the power spectrum.

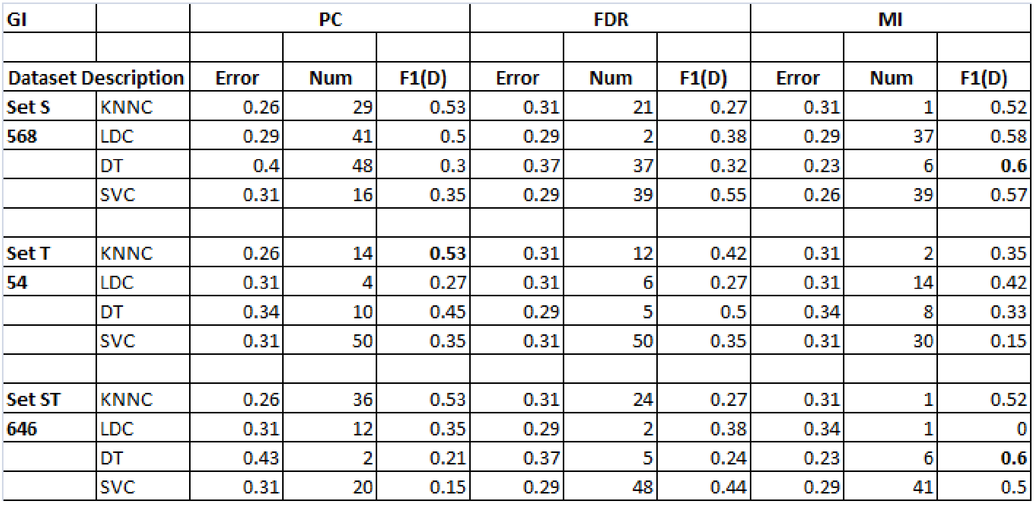
4.2 Observations and Discussion

In the proposed model, Temporal set (T), Spectral set (S) and combination of Temporal set and Spectral set including Spectro-Temporal features (ST) have been investigated. Feature selection was further applied to find a minimal set of relevant features which provides maximum f1 depressed. We use three feature ranking methods namely Mutual Information

(MI), Fisher Discriminant Ratio score (FDR) and Pearson Correlation Coefficient (PC). For each feature set obtained, feature selection by these three methods has been applied and four classifiers have been used to build the learning model in order to reduce the error of misclassification and to maximize the f1 depressed. DAICWOZ [33] consists of a training and a development set. In the hold out technique, we have used training set for training the model and development set for testing. The experiments have been performed for both gender independent and gender dependent cases.

The result for the gender independent hold out case is shown in table 4.1.

Table 4.1: Hold out Result of Gender Independent for set S, T and ST



GI		**PC**			**FDR**			**MI**		
Dataset Description		**Error**	**Num**	**F1(D)**	**Error**	**Num**	**F1(D)**	**Error**	**Num**	**F1(D)**
Set S	KNNC	0.26	29	0.53	0.31	21	0.27	0.31	1	0.52
568	LDC	0.29	41	0.5	0.29	2	0.38	0.29	37	0.58
	DT	0.4	48	0.3	0.37	37	0.32	0.23	6	**0.6**
	SVC	0.31	16	0.35	0.29	39	0.55	0.26	39	0.57
Set T	KNNC	0.26	14	**0.53**	0.31	12	0.42	0.31	2	0.35
54	LDC	0.31	4	0.27	0.31	6	0.27	0.31	14	0.42
	DT	0.34	10	0.45	0.29	5	0.5	0.34	8	0.33
	SVC	0.31	50	0.35	0.31	50	0.35	0.31	30	0.15
Set ST	KNNC	0.26	36	0.53	0.31	24	0.27	0.31	1	0.52
646	LDC	0.31	12	0.35	0.29	2	0.38	0.34	1	0
	DT	0.43	2	0.21	0.37	5	0.24	0.23	6	**0.6**
	SVC	0.31	20	0.15	0.29	48	0.44	0.29	41	0.5

(i) The overall highest f1 score for depressed has been obtained is 0.6 with set S and set ST with method MI and classifier DT .

(ii) The set S has shown the maximum f1 depressed 0.6 with MI and DT.

(iii) The set T has shown the maximum f1 depressed 0.53 with PC and KNNC.

(iv) The set ST has shown the maximum f1 depressed 0.6 with MI and DT classifier.

(v) Considering PC as feature selection the maximum f1 depressed has been found to be 0.53 with all the three sets with KNNC.

(vi) Considering FDR as feature selection the maximum f1 depressed has been found to be 0.55 with set S with SVC.

(vii) Considering MI as feature selection the maximum f1 depressed has been found to be 0.6 with set S and ST with DT.

(viii) With KNNC the maximum f1 depressed is 0.53 with all the three sets with PC.

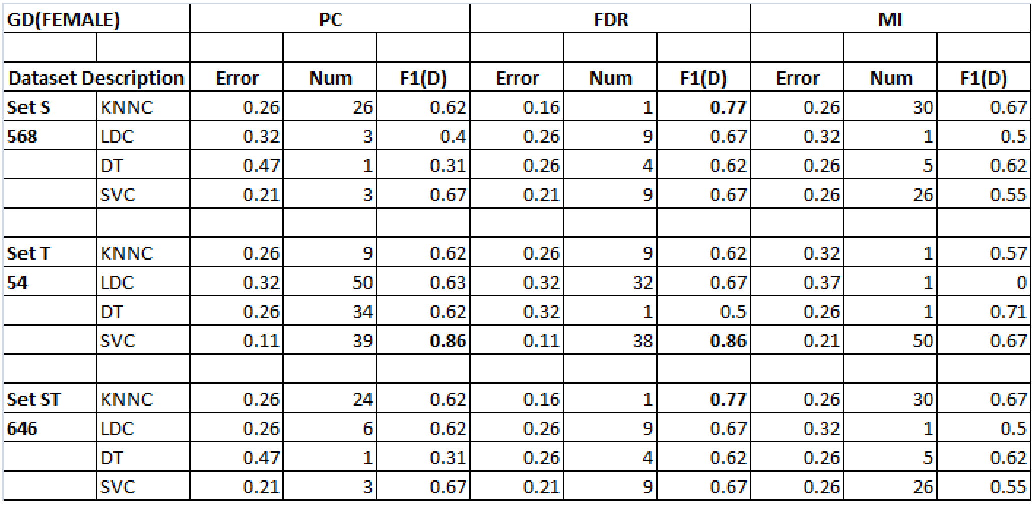
(ix) With LDC the maximum f1 depressed is 0.58 with set S with MI.

(x) DT has shown maximum f1 depressed 0.6 with two sets (S and ST) with PC.

(xi) SVC has shown maximum f1 depressed 0.57 with set S with MI.

The result for the gender dependent(Female) hold out case is shown in table 4.2.

Table 4.2: Hold out Result of Gender Dependent(Female) for set S, T and ST



GD(FEMALE)		**PC**			**FDR**			**MI**		
Dataset Description		**Error**	**Num**	**F1(D)**	**Error**	**Num**	**F1(D)**	**Error**	**Num**	**F1(D)**
Set S	KNNC	0.26	26	0.62	0.16	1	**0.77**	0.26	30	0.67
568	LDC	0.32	3	0.4	0.26	9	0.67	0.32	1	0.5
	DT	0.47	1	0.31	0.26	4	0.62	0.26	5	0.62
	SVC	0.21	3	0.67	0.21	9	0.67	0.26	26	0.55
Set T	KNNC	0.26	9	0.62	0.26	9	0.62	0.32	1	0.57
54	LDC	0.32	50	0.63	0.32	32	0.67	0.37	1	0
	DT	0.26	34	0.62	0.32	1	0.5	0.26	1	0.71
	SVC	0.11	39	**0.86**	0.11	38	**0.86**	0.21	50	0.67
Set ST	KNNC	0.26	24	0.62	0.16	1	**0.77**	0.26	30	0.67
646	LDC	0.26	6	0.62	0.26	9	0.67	0.32	1	0.5
	DT	0.47	1	0.31	0.26	4	0.62	0.26	5	0.62
	SVC	0.21	3	0.67	0.21	9	0.67	0.26	26	0.55

(i) The overall highest f1 score for depressed has been obtained is 0.86 with set T with methods PC and FDR with classifier SVC.

(ii) The set S has shown the maximum f1 depressed 0.77 with MI and DT.

(iii) The set T has shown the maximum f1 depressed 0.86 with PC and FDR with SVC.

(iv) The set ST has shown the maximum f1 depressed 0.77 with FDR and KNNC.

(v) Considering PC as feature selection the maximum f1 depressed is 0.86 with set T and KNNC.

(vi) Considering FDR as feature selection the maximum f1 depressed is 0.86 with set T and SVC.

(vii) Considering MI as feature selection the maximum f1 depressed is 0.71 with set T and DT.

(viii) With KNNC the maximum f1 depressed is 0.77 with set(S and ST) and FDR.

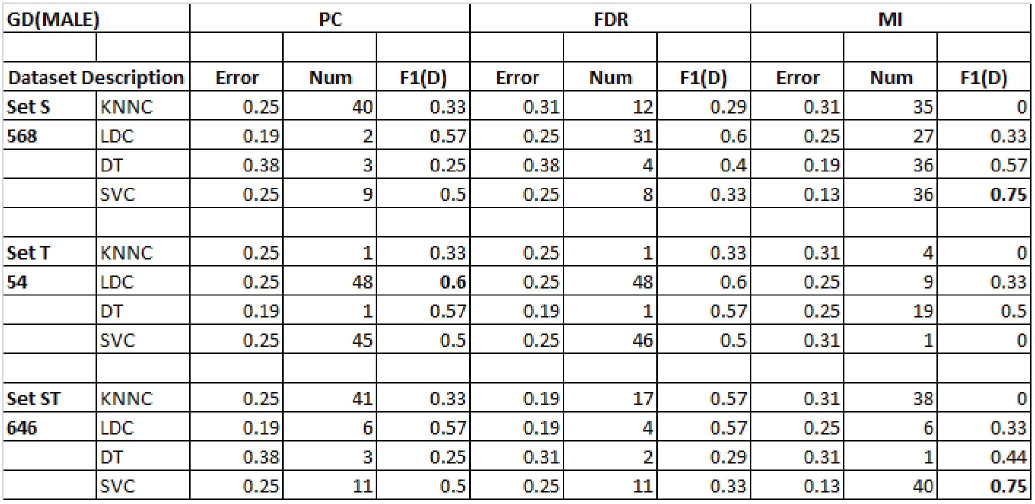
(ix) With LDC the maximum f1 depressed is 0.67 with all the three sets and FDR.

(x) With DT the maximum f1 depressed is 0.71 with set T with MI.

(xi) SVC has shown the maximum f1 depressed 0.67 with both set (S and ST) with PC and FDR.

The result for the gender dependent(Male) hold out case is shown in table 4.3.

Table 4.3: Hold out Result of Gender Dependent(Male) for set S, T and ST



GD(MALE)		PC			FDR			MI		
Dataset Description		Error	Num	F1(D)	Error	Num	F1(D)	Error	Num	F1(D)
Set S	KNNC	0.25	40	0.33	0.31	12	0.29	0.31	35	0
568	LDC	0.19	2	0.57	0.25	31	0.6	0.25	27	0.33
	DT	0.38	3	0.25	0.38	4	0.4	0.19	36	0.57
	SVC	0.25	9	0.5	0.25	8	0.33	0.13	36	**0.75**
Set T	KNNC	0.25	1	0.33	0.25	1	0.33	0.31	4	0
54	LDC	0.25	48	**0.6**	0.25	48	0.6	0.25	9	0.33
	DT	0.19	1	0.57	0.19	1	0.57	0.25	19	0.5
	SVC	0.25	45	0.5	0.25	46	0.5	0.31	1	0
Set ST	KNNC	0.25	41	0.33	0.19	17	0.57	0.31	38	0
646	LDC	0.19	6	0.57	0.19	4	0.57	0.25	6	0.33
	DT	0.38	3	0.25	0.31	2	0.29	0.31	1	0.44
	SVC	0.25	11	0.5	0.25	11	0.33	0.13	40	**0.75**

(i) The overall highest f1 score for depressed has been obtained is 0.75 with set S and set ST with method MI and classifier SVC.

(ii) The set S has shown the maximum f1 depressed 0.75 with MI and SVC.

(iii) The set T has shown the maximum f1 depressed 0.60 with PC and LDC.

(iv) The set ST has shown the maximum f1 depressed 0.75 with MI and SVC.

(v) Considering PC as feature selection the maximum f1 depressed is 0.60 with set T and LDC.

(vi) Considering FDR as feature selection the maximum f1 depressed is 0.60 with both set (S and T) and LDC.

(vii) Considering MI as feature selection the maximum f1 depressed is 0.75 with both set (S and T) and SVC.

(viii) With KNNC the maximum f1 depressed is .57 with set ST with FDR.

(ix) With LDC the maximum f1 depressed is .60 with set(S and T) with FDR.

(x) With DT the maximum f1 depressed is 0.57 with set(S and T) with all three feature selection methods.

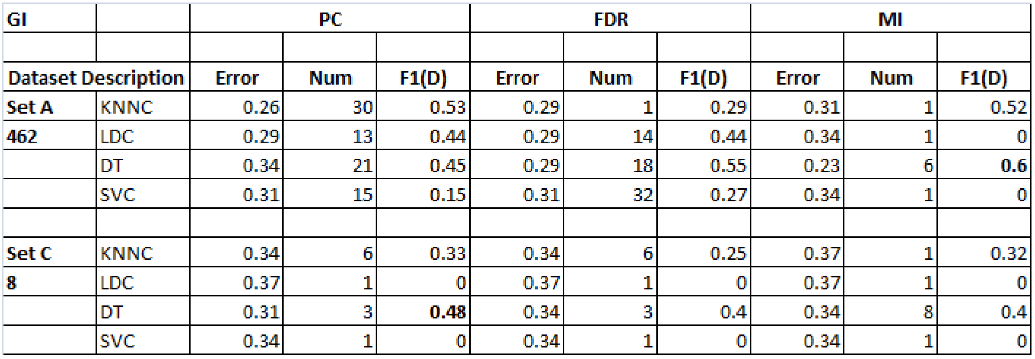
(xi) With SVC the maximum f1 depressed is .75 with set (S and ST) with MI.

Across all the experiments we see considerable improvement of gender dependent (GD) results over gender independent (GI) results. The GI maximum of 0.6 has been improved with female based experiments with 0.86 and male based with 0.75. All these results are higher than the quoted results for the individual feature performances.

In 2016, Pampouchidou *et al.* [11] used a fusion of low level features, discrete cosine transform (DCT) based features and high level features. They analysed three sets using the DAICWOZ dataset and the COVAREP feature repository. The first set (set A) comprised of statistical descriptors (refer Table 1 of ([11])) of the following low level descriptors: Fo (normalized, delta and delta-delta), NAQ, QOQ, Amplitude difference of the first two harmonics (H1H2), Parabolic Spectral Parameter (PSP), Maximum Dispersion Quotient (MDQ), peakslope, shape parameter of the Liljencrants-Fant glottal model (Rd), Rd confidence measure (Rd_{conf}), Mel Cepstral Coefficients (MCEP0-24, delta and delta-delta), Harmonic Model and Phase Distortion Mean (HMPDM 1-24), Harmonic Model and Phase Distortion Deviation (HMPDD 1-12) and 1-3 Formants. Set B comprised of DCT coefficients of the low level descriptors. The set (Set C) consists of the eight high level features such as Pause Ratio, Voiced Segment Ratio, Speaking Ratio, Mean laughter Duration, Mean Delay in Response, Mean Duration of Pauses, Maximum Duration of Pauses and the Fraction of Pauses in Overall Time.

The gender independent results stated by [11] for f1(depressed) gave the maximum value of 0.24 while the gender dependent results using low level descriptors are reported in terms of f1 depressed as 0.59 (testing using development set in the hold out method). We have also performed hold out experiments with set A (statistical features of low level descriptors) and set C (high level features). The experiments have been performed with both gender independent and gender dependent cases (both male and female). We have used the feature selection methods and four classifiers to build the decision model. Our proposed method has outperformed the state of the art results mentioned above[11]. A detailed discussion is given below. The result for the gender independent hold out case is shown in table 4.4.

Table 4.4: Hold out Result of Gender Independent for set A and C



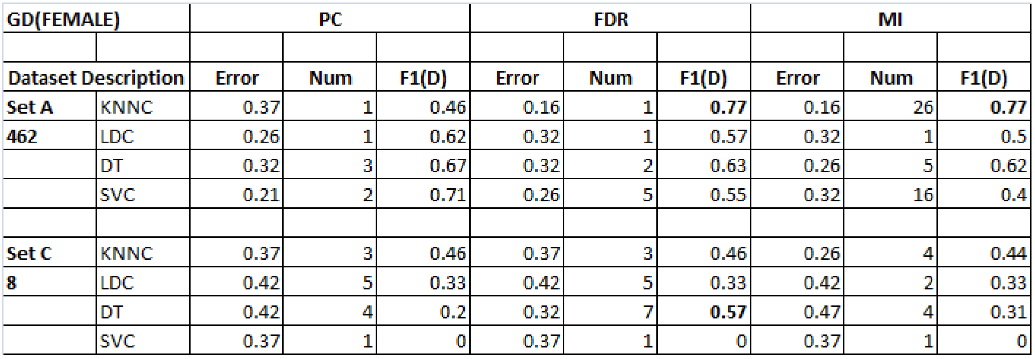
GI		**PC**			**FDR**			**MI**		
Dataset Description		**Error**	**Num**	**F1(D)**	**Error**	**Num**	**F1(D)**	**Error**	**Num**	**F1(D)**
Set A	KNNC	0.26	30	0.53	0.29	1	0.29	0.31	1	0.52
462	LDC	0.29	13	0.44	0.29	14	0.44	0.34	1	0
	DT	0.34	21	0.45	0.29	18	0.55	0.23	6	**0.6**
	SVC	0.31	15	0.15	0.31	32	0.27	0.34	1	0
Set C	KNNC	0.34	6	0.33	0.34	6	0.25	0.37	1	0.32
8	LDC	0.37	1	0	0.37	1	0	0.37	1	0
	DT	0.31	3	**0.48**	0.34	3	0.4	0.34	8	0.4
	SVC	0.34	1	0	0.34	1	0	0.34	1	0

(i) The overall maximum f1 depressed is 0.60 with A set with combination of MI and DT.

(ii) The f1 score obtained with set C is 0.48 with combination of PC and DT.

The result for the gender dependent(female) hold out case is shown in table 4.5.

Table 4.5: Hold out Result of Gender Dependent(Female) for set A and C



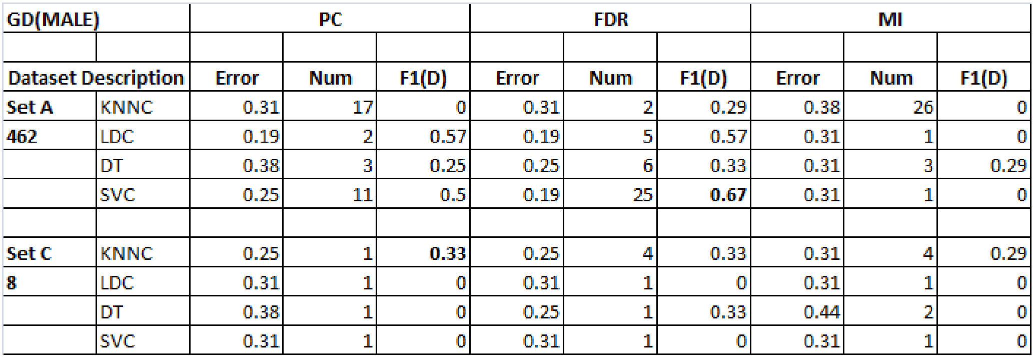
GD(FEMALE)		**PC**			**FDR**			**MI**		
Dataset Description		**Error**	**Num**	**F1(D)**	**Error**	**Num**	**F1(D)**	**Error**	**Num**	**F1(D)**
Set A	KNNC	0.37	1	0.46	0.16	1	**0.77**	0.16	26	**0.77**
462	LDC	0.26	1	0.62	0.32	1	0.57	0.32	1	0.5
	DT	0.32	3	0.67	0.32	2	0.63	0.26	5	0.62
	SVC	0.21	2	0.71	0.26	5	0.55	0.32	16	0.4
Set C	KNNC	0.37	3	0.46	0.37	3	0.46	0.26	4	0.44
8	LDC	0.42	5	0.33	0.42	5	0.33	0.42	2	0.33
	DT	0.42	4	0.2	0.32	7	**0.57**	0.47	4	0.31
	SVC	0.37	1	0	0.37	1	0	0.37	1	0

(i) The overall maximum f1 depressed is 0.77 with A set with combination of KNNC with FDR and MI.

(ii) The f1 score obtained with set C is 0.57 with combination of FDR and DT.

The result for the gender dependent(Male) hold out case is shown in table 4.6.

Table 4.6: Hold out Result of Gender Dependent(Male) for set A and C



GD(MALE)		PC			FDR			MI		
Dataset Description		Error	Num	F1(D)	Error	Num	F1(D)	Error	Num	F1(D)
Set A	KNNC	0.31	17	0	0.31	2	0.29	0.38	26	0
462	LDC	0.19	2	0.57	0.19	5	0.57	0.31	1	0
	DT	0.38	3	0.25	0.25	6	0.33	0.31	3	0.29
	SVC	0.25	11	0.5	0.19	25	**0.67**	0.31	1	0
Set C	KNNC	0.25	1	**0.33**	0.25	4	0.33	0.31	4	0.29
8	LDC	0.31	1	0	0.31	1	0	0.31	1	0
	DT	0.38	1	0	0.25	1	0.33	0.44	2	0
	SVC	0.31	1	0	0.31	1	0	0.31	1	0

(i) The overall maximum f1 depressed is 0.67 with A set with combination of FDR and SVC.

(ii) The f1 score obtained with set C is 0.33 with combination of PC and KNNC.

The result quoted by Pampouchidou *et al* in terms of f1 score is 0.24 for GI and 0.59 for GD. Our proposed method on the same dataset has obtained 0.60 for GI and 0.77 (female) and 0.67(male) which has considerably improved over their results. It can also be observed that using our proposed method, the high level features, set C, gives high f1 score for gender independent experiments.

Table 4.7 gives the comparative study of our propsed method's results along with other state of the art results given in literature for depression detection. The results show that our proposed model which is a simple feature selection method, outperforms the speech based result and also outperforms some multimodal systems. It also performs well for gender dependent systems.

Table 4.7: Comparative table of results of the proposed method with the State of the Art

GI	0.60(PROPOSED)(Speech)
	0.52[51](Speech)
	0.46 [28](Speech)
	0.50 [28] (Video)
	0.50 [28] (Ensemble)
	0.57(mean) [52](Speech)
	0.81(mean) [52](Ensemble)
	0.55 [53](Speech)
	0.57 [54](Speech)
	0.63 [54] (Video)
	0.63 [54](Ensemble)
female	0.86(PROPOSED)(Speech)
	1.0 [53](Speech)
male	0.75(PROPOSED)(Speech)
	0.53 [53] (Speech)
GD	0.62[11](Ensemble)

Chapter 5

Conclusion

For the effective detection of depression using speech, we have investigated an exhaustive set of Temporal and Spectral features, along with a hybrid of the two. As the number of features thus obtained is large, while the sample size is small, feature selection is carried out, to obtain the relevant subset of features that discriminate the depressed subjects from the control and enhance the efficacy of the decision system. We have used three well known univariate feature selection methods and four different classifiers for our study. Results suggest that the features extracted from the exhaustive set of Temporal, Spectral and the combination of the two, followed by selection of a subset of relevant features is effective in providing high performance, among gender independent as well as the gender based experiments. The best performance was achieved with the combination of MI with DT; FDR with SVC; and MI with SVC for GI, female and male based experiments respectively.

The simple and cost effective univariate feature selection method is instrumental in selecting the relevant features that can be focused upon for building an effective, non-invasive and an objective depression detection system. The proposed method has given high performance in terms of f1-score of depression in comparison to the previous work in the domain of speech based depression detection. Moreover, our proposed method has also outperformed some multimodal depression detection systems, indicating the strength of speech as an effective unimodal based depression detection system. We propose a hybrid of Temporal, Spectral and Spectro-Temporal speech based features to obtain a successful depression detection system. Further, the application of the univariate feature selection techniques to obtain the relevant set of speech features, enhances the efficacy of the decision model. There is a need to investigate other feature selection methods that can eliminate the redundancy present among the selected features and enhance the efficacy of the unimodal system.

Bibliography

[1] World Health Organization. Preventing suicide: A global imperative. *Geneva: WHO 2014*.

[2] http://www.assocham.org/newsdetail.php?id=4918.

[3] http://indiatoday.intoday.in/story/delhi-police-suicide-high-stress-depression-financial-issues/1/1069444.html.

[4] American Psychiatric Association and others revised (DSM-III-R). Diagnostic and statistical manual of mental disorders;. *Washington DG*, 1987.

[5] Albert FG Leentjens, Frans RJ Verhey, Richel Lousberg, Harro Spitsbergen, and Frederik W Wilmink. The validity of the hamilton and montgomery-åsberg depression rating scales as screening and diagnostic tools for depression in parkinson's disease. *International journal of geriatric psychiatry*, 15(7):pages 644–649, 2000.

[6] Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, 67(3):pages 588–597, 1996.

[7] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1):pages 163–173, 2009.

[8] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. An investigation of depressed speech detection: Features and normalization. 2011.

[9] Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Louis-Philippe Morency, et al. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):pages 648–658, 2014.

[10] Matthias R Lemke, Thomas Wendorff, Brigitt Mieth, Katharina Buhl, and Martin Linnemann. Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. *Journal of psychiatric research*, 34(4):pages 277–283, 2000.

[11] Anastasia Pampouchidou, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meri-

audeau, Panagiotis Simos, Kostas Marias, et al. Depression assessment by fusing high and low level features from audio, video, and text. pages 27–34, 2016.

[12] Raefel A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, pages 1–37, 2017.

[13] Klaus R Scherer. Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):pages 143, 1986.

[14] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Gordon Parker, Michael Breakspear, et al. Characterising depressed speech for classification. pages 2534–2538, 2013.

[15] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed AI-Shuraifi, and Yunhong Wang. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. pages 21–30, 2013.

[16] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):pages 829–837, 2000.

[17] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. Speaker dependency of spectral features and speech production cues for automatic emotion classification. pages pages 4693–4696, 2009.

[18] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):pages 357–366, 1980.

[19] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. Empirical mode decomposition based weighted frequency feature for speech-based emotion classification. pages pages5017–5020, 2008.

[20] Mireia Farrús, Javier Hernando, and Pascual Ejarque. Jitter and shimmer measurements for speaker recognition. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[21] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa Sheeber, and Nicholas Allen. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5154–5157. IEEE, 2010.

[22] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, and Gordon Parker. Detecting depression: a comparison between spontaneous and read speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7547–7551. IEEE, 2013.

[23] John Kane and Christer Gobl. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179, 2013.

[24] John Kane and Christer Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[25] Asli Ozdas, Richard G Shiavi, Stephen E Silverman, Marilyn K Silverman, and D Mitchell Wilkes. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, 51(9):1530–1540, 2004.

[26] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Interspeech*, pages 847–851, 2013.

[27] Anastasia Pampouchidou, Olympia Simantiraki, C-M Vazakopoulou, C Chatzaki, Matthew Pediaditis, A Maridaki, Kostas Marias, Panagiotis Simos, Fan Yang, Fabrice Meriaudeau, et al. Facial geometry and speech analysis for depression detection. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pages 1433–1436. IEEE, 2017.

[28] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2016.

[29] Nicholas Cummins, Bogdan Vlasenko, Hesam Sagha, and Björn Schuller. Enhancing speech-based depression detection through gender dependent vowel-level formant. In *Proc. of Conference on Artificial Intelligence in Medicine. Springer*, volume 5, 2017.

[30] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[31] Takaya Taguchi, Hirokazu Tachikawa, Kiyotaka Nemoto, Masayuki Suzuki, Toru Nagano, Ryuki Tachibana, Masafumi Nishimura, and Tetsuaki Arai. Major depressive disorder discrimination using vocal acoustic features. *Journal of affective disorders*, 225:214–220, 2018.

[32] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, 5(2):65–75, 2016.

[33] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The

distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Citeseer, 2014.

[34] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE, 2014.

[35] Paavo Alku, Tom Bäckström, and Erkki Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

[36] Tamas Hacki. Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie. *Folia Phoniatrica et Logopaedica*, 41(1):43–48, 1989.

[37] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 454, pages 903–995. The Royal Society, 1998.

[38] Thomas Drugman and Abeer Alwan. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[39] Paavo Alku and Erkki Vilkman. A frequency domain method for parametrization of the voice source, 1996.

[40] Yannis Stylianou. Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. *Ph. D thesis, Ecole Nationale Superieure des Telecommunications*, 1996.

[41] George P Kafentzis, Yannis Pantazis, Olivier Rosec, and Yannis Stylianou. An extension of the adaptive quasi-harmonic model. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4605–4608. IEEE, 2012.

[42] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[43] Gunnar Fant. The lf-model revisited. transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2(3):40, 1995.

[44] R Bellman. Curse of dimensionality. *Adaptive control processes: a guided tour. Princeton, NJ*, 1961.

[45] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[46] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[47] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[48] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[49] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[50] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[51] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 35–42. ACM, 2016.

[52] James R Williamson, Thomas F Quatieri, Brian S Helfer, Gregory Ciccarelli, and Daryush D Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM, 2014.

[53] Bogdan Vlasenko, Hesam Sagha, Nicholas Cummins, and Björn Schuller. Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition. In *Proc. Interspeech*, pages 3266–3270, 2017.

[54] Md Nasir, Arindam Jati, Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, and Panayiotis Georgiou. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 43–50. ACM, 2016.