Speech Based Assessment of Depression

Amandeep kaur
M.tech (ECE)
BCET,GSP

Anita Suman
Assistant Professor (ECE)
BCET,GSP

Abstract-Depression is considered as a psychosomatic state associated with the soft biometric features. People who are suffering from depression always behave abnormal. Depression is a clinically proven disorder that can overwhelm a person and his ability to perform even a simple task. Soft biometric provides important information about a person without being enough for their verification because they lack uniqueness. This statement comprises of features which are associated with the psychosomatic state of a person such as feelings, sentiments or brain related disorders like depression. In this paper we have estimated the depression level of each speech signal using I-Vector technique.

In our proposed approach first of all we have removed silence from the speech signal then we have extracted features from audio using I-Vector after that split overlapping function is applied to evaluate overlapped audio beats. In the end we have evaluated depression using relationship matrix. We have estimated the depression level of each speaker. This technique has better performance as compared with existing techniques. The overall result has shown that the I-Vector technique has good accuracy to detect depression in audios.

Index Terms: Depression, Audios, I vector, Fuzzy Logic

1. Introduction

Depression is a serious mental health disorder that affects mood, thoughts, feelings, and the ability to function in everyday life. Some of its characteristics are prolonged feelings of extreme sadness, guilt and hopelessness, and thoughts of death. Major depression is the leading cause of disability and is the cause of more than two-thirds of suicides each year [1]. Depression is a common mental disorder that presents persistent feelings of sadness, intrusive negative thoughts and, cognitive difficulties such as poor concentration, leading to functional impairment [3]. A later study of 28 sufferers of depressive illness found similar evidence for a reduced variability in fundamental frequency and prosody [2]. In experiments by Stassen [3], speech features such as fundamental frequency and speech pause duration were found to be sufficiently highly correlated with the HAMD-17 depression score that simple speech analysis methods were trialled as an objective measure of patient recovery during a course of antidepressants. Flint [4] studied the effect of psychomotor retardation in depressed people and found that patients with a major depressive disorder had decreased second formant (F2) measurements when compared with a control group. More recent work has seen the first steps towards automatic analysis of speech [5]. A range of acoustic features have already been identified for suitability in the classification of depression. Speech production cues such as pitch and formant measures are useful due in part to the effects of increased tension in the vocal tract associated with depression [4,5]. Spectral and energy based measures are also useful in classifiers, as depressive speech can contain more information in the higher energy bands when compared with neutral speech [5, 6]. Spectral centroid based methods including the sub-band spectral centroid features have recently shown promise in other applications [7], and other work [8] shows that these newer measures potentially include information useful in the classification of depression.

2 Acoustic Characterization Of Depression In Speech

2.1. Segment Selection

Beyond the usual isolation of speech-active regions in the signal using a voice activity detector (VAD), a question of interest is the accurate selection of speech segments that provide maximal depressed/neutral speech discrimination. To our knowledge, the decision between voiced-only, unvoiced only or mixed-voicing speech is without empirical support. In this paper, we employ an energy-based VAD and confine ourselves to investigating the relative merits of voiced and unvoiced segments, using short term energy and F0 information to estimate the degree of voicing. Based on emotion recognition results and, then expected to find that voiced segments provide the most effective discrimination.

2.2. Feature Extraction

The motivation is similar, namely to understand the relative contribution of speech production cues, detailed spectral information and broad spectral information to depressed/neutral speech discrimination. Among speech production cues, pitch, energy and formants are obvious feature choices, for which insights into depressed speech. Broad spectral information is of interest since detailed spectral information can be expected to contain substantial variability due to the phonetic composition of an utterance and the speaker identity. Typically, the former source of variability is accounted for by employing a Gaussian mixture model with many mixtures, which represent different acoustic regions within which depressed/neutral speech discrimination can be more effectively characterized. The latter source of variability can be mitigated using feature normalization methods. In this work, we investigate energy slope (ES – extracted by using least squares analysis of the magnitude spectrum to calculate the linear slope coefficient for a given frame), zero-crossing rate (ZC), and spectral centroid (SC – extracted across the full speech bandwidth).

2.3. Feature Normalization

As explained above, feature normalization can be applied to reduce the mismatch in feature distributions between different speakers. Research has shown that speaker variability is a stronger confounder for emotion recognition than phonetic variability [11], and we adopt a similar hypothesis for depressed/neutral speech classification. An important distinction between emotion recognition and depressed/neutral speech classification must be made: unlike emotion, which can be transient and hence easily elicited from a single speaker in a variety of forms, depression is a condition that is sustained for weeks to months. Finding speakers able to produce both depressed and neutral speech, even over a long period of time, is very challenging. Hence, the depressed and neutral speaker utterances might often be mutually exclusive with respect to speaker identity. Normalizing on a speaker basis for depressed/neutral speech will prove beneficial if the speaker variability is substantially larger than the depressed/neutral speech variability (this seems likely [11]). On the other hand, if the depressed/neutral speech variability is larger than the speaker variability, then per-speaker normalization might prove detrimental to depressed/neutral classification.

2.4. Modeling of Depressed Speech

Following the approach of many paralinguistic speech classification systems based on acoustic features, including some focused on depressed speech recognition, here employ Gaussian mixture models (GMMs) to model depressed and neutral speech. In contemporary systems, researchers often make use of multiple subsystems and score-level fusion, to combine the benefits of individual systems and advance the state of the art.

3. I-Vector Feature Extraction

The main idea in traditional JFA, introduced by Kenny[3], is to find two subspaces which represent the speakerand channel-variabilities, respectively. Dehak's experiments [1] show that JFA is only partially successful in separating speaker and channel variabilities. He found that the channel space contains some information that can be used to distinguish between speakers. For this reason, Dehak proposed a single space that models the two variabilities and named it the total variability space.

As mentioned previously, robust estimation of the total variability matrix T in (1) requires a large amount of data, whereas we have relatively little microphone data at our disposal. The main contribution is to show how to use telephone data in addition to microphone data to estimate a new total variability matrix (we name it N) that is suitable for speaker recognition on microphone speech. We adopt a modified version of the method proposed in [3] to deal with the cross-channel condition of the 2006 NIST SRE. So it was shown how to estimate supplementary eigen channels on microphone development data and append them to eigen channels estimated on telephone speech. Section IV in Kenny's article [3] presents results obtained with this method. First, we estimate a gender dependent matrix T of rank Rtel using only telephone data, by assuming the supervector M associated with a telephone recording has the form represented in (1). Second, we estimate a gender dependent matrix T0 of rank Rmic using only microphone data, by assuming the supervector M associated with a microphone recording has the form

$$M = m + T \wedge w + T0w0 \quad (2)$$

where w0 is a standard normally distributed random vector.

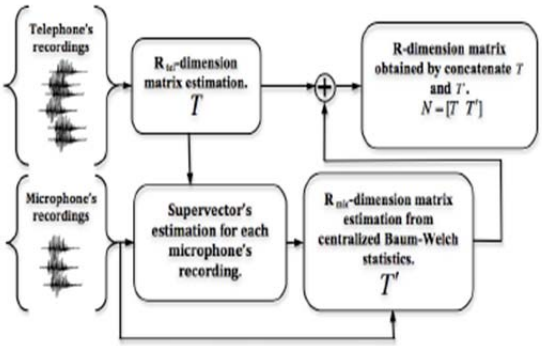
Telephone's recordings
R_{tel}-dimension matrix estimation. T
R-dimension matrix obtained by concatenate T and T'. $N=[T\ T']$
Microphone's recordings
Supervector's estimation for each microphone's recording.
R_{mic}-dimension matrix estimation from centralized Baum-Welch statistics. T'

Fig 1.3: The block diagrams showing the estimation of the matrix N.

The estimation procedure of the total matrix N is illustrated in Figure 1.3.

Finally, we combine (1) and (2) to produce a feature extractor which can be used for microphone speech as well as telephone speech. This representation was also tested on the telephone speech using probabilistic methods. That is, we assume that the supervector M associated with a recording has the form

$$M = m + Nx \quad (3)$$

where N = [T T0]1 is the new total variability matrix of rank R = Rtel + Rmic, x is a random vector of dimension D = R having a standard normal distribution. So the new feature vector associated with a recording is the MAP estimate of x.

4. Fuzzy Logic

Fuzzy logic idea is similar to the human being’s feeling and inference process. Unlike classical control strategy, which is a point-to-point control, fuzzy logic control is a range-to-point or range-to-range control. The output of a fuzzy controller is derived from fuzzifications of both inputs and outputs using the associated membership functions. A crisp input will be converted to the different members of the associated membership functions based on its value. From this point of view, the output of a fuzzy logic controller is based on its memberships of the different membership functions, which can be considered as a range of inputs. Fuzzy ideas and fuzzy logic are so often utilized in our routine life that nobody even pays attention to them. For instance, to answer some questions in certain surveys, most time one could answer with 'Not Very Satisfied' or 'Quite Satisfied', which are also fuzzy or ambiguous answers. Exactly to what degree is one satisfied or dissatisfied with some service or product for those surveys? These vague answers can only be created and implemented by human beings, but not machines. Is it possible for a computer to answer those survey questions directly as a human beings did? It is absolutely impossible. Computers can only understand either '0' or '1', and 'HIGH' or 'LOW'. Those data are called crisp or classic data and can be processed by all machines.

To implement fuzzy logic technique to a real application requires the following three steps:

1. Fuzzification – convert classical data or crisp data into fuzzy data or Membership Functions (MFs)
2. Fuzzy Inference Process – combine membership functions with the control rules to derive the fuzzy output
3. Defuzzification – use different methods to calculate each associated output and put them into a table: the lookup table. Pick up the output from the lookup table based on the current input during an application As mentioned before, all machines can process crisp or classical data such as either '0' or '1'. In order to enable machines to handle vague language input such as 'Somehow Satisfied', the crisp input and output must be converted to linguistic variables with fuzzy components. For instance, to control an air conditioner system, the input temperature and the output control variables must be converted to the associated linguistic variables such as 'HIGH', 'MEDIUM', 'LOW' and 'FAST', 'MEDIUM' or 'SLOW'. The former is corresponding to the input temperature and the latter is associated with the rotation speed of the operating motor. Besides those conversions, both the input and the output must also be converted from crisp data to fuzzy data. All of these jobs are performed by the first step – fuzzification.

In the second step, to begin the fuzzy inference process, one need combine the Membership Functions with the control rules to derive the control output, and arrange those outputs into a table called the lookup table. The control rule is the core of the fuzzy inference process, and those rules are directly related to a human being’s intuition and feeling. For example, still in the air conditioner control system, if the temperature is too high, the heater should be turned off, or the heat driving motor should be slowed down, which is a human being’s intuition or common sense. Different methods such as Center of Gravity (COG) or Mean of Maximum (MOM) are utilized to calculate the associated control output, and each control output should be arranged into a table called lookup table. During an actual application, a control output should be selected from the lookup table developed from the last step based on the current input. Furthermore, that control output should be converted from the linguistic variable back to the crisp variable and output to the control operator. This process is called defuzzification or step 3.

In most cases the input variables are more than one dimension for real applications, and one needs to perform fuzzification or develop a Membership Function for each dimensional variable separately. Perform the same operation if the system has multiple output variables. Summarily, a fuzzy process is a process of crisp-fuzzy-crisp for a real system. The original input and the terminal output must be crisp variables, but the intermediate process is a fuzzy inference process. The reason why one needs to change a crisp to a fuzzy variable is that, from the point of view of fuzzy control or a human being's intuition, no absolutely crisp variable is existed in our real world.

5. Related Work

Michel Valstar et al. [6] Mood disorders are inherently related to emotion. In particular, the behaviour of people suffering from mood disorders such as unipolar depression shows a strong temporal correlation with the affective dimensions valence and arousal. In addition, psychologists and psychiatrists base their evaluation of a patient's condition to a large extent on the observation of expressive facial and vocal cues, such as dampened facial expressions, avoiding eye contact, and using short sentences with flat intonation. It is in this context that we present the third Audio-Visual Emotion recognition Challenge (AVEC 2013). The challenge has two goals logically organised as sub-challenges: the first is to predict the continuous values of the affective dimensions valence and arousal at each moment in time. The second sub-challenge is to predict the value of a single depression indicator for each recording in the dataset. This paper presents the challenge guidelines, the common data used, and the performance of the baseline system on the two tasks.

L.J.Rodríguez-Fuentes et al. [7] presents the main features (design issues, recording setup, etc.) of KALAKA-2, a TV broadcast speech database specifically designed for the development and evaluation of language recognition systems in clean and noisy environments. KALAKA-2 was created to support the Albayzin 2010 Language Recognition Evaluation (LRE), organized by the Spanish Network on Speech Technologies from June to November 2010. The database features 6 target languages: Basque, Catalan, English, Galician, Portuguese and Spanish, and includes segments in other (Out-Of-Set) languages, which allow to perform open-set verification tests. The best performance attained in the Albayzin 2010 LRE is presented and briefly discussed. The performance of a state-of-the-art system in various tasks defined on the database is also presented. In both cases, results highlight the suitability of KALAKA-2 as a benchmark for the development and evaluation of language recognition technology.

Daniel Povey et al. [8] describe the design of Kaldi, a free, open-source toolkit for speech recognition research. Kaldi provides a speech recognition system based on finite-state transducers (using the freely available OpenFst), together with detailed documentation and scripts for building complete recognition systems. Kaldi is written is C++, and the core library supports modeling of arbitrary phonetic-context sizes, acoustic modeling with subspace Gaussian mixture models (SGMM) as well as standard Gaussian mixture models, together with all commonly used linear and affine transforms. Kaldi is released under the Apache License v2.0, which is highly nonrestrictive, making it suitable for a wide community of users.

Niels Rosenquist et al. [9] The etiology of depression has long been thought to include social environmental factors. To quantitatively explore the novel possibility of person-to-person spread and network-level determination of depressive symptoms, analyses were performed on a densely interconnected social network of 12 067 people assessed repeatedly over 32 years as part of the Framingham Heart Study. Longitudinal statistical models were used to examine whether depressive symptoms in one person were associated with similar scores in friends, co-workers, siblings, spouses and neighbors. Depressive symptoms were assessed using CES-D scores that were available for subjects in three waves measured between 1983 and 2001. Results showed both low and high CES-D scores (and classification as being depressed) in a given period were strongly correlated with such scores in one's friends and neighbors. This association extended up to three degrees of separation (to one's friends' friends' friends). Female friends appear to be especially influential in the spread of depression from one person to another.

Kua et al. [10] Most conventional features used in speaker recognition are based on spectral envelope characterizations such as Mel-scale filterbank cepstrum coefficients (MFCC), Linear Prediction Cepstrum Coefficient (LPCC) and Perceptual Linear Prediction

(PLP). The MFCC's success has seen it become a de facto standard feature for speaker recognition. Alternative features, that convey information other than the average subband energy, have been proposed, such as frequency modulation (FM) and subband spectral centroid features. In this study, we investigate the characterization of subband energy as a two dimensional feature, comprising Spectral Centroid Magnitude (SCM) and Spectral Centroid Frequency (SCF). Empirical experiments carried out on the NIST 2001 and NIST 2006 databases using SCF, SCM and their fusion suggests that the combination of SCM and SCF are somewhat more accurate compared with conventional MFCC, and that both fuse effectively with MFCCs. We also show that frame-averaged FM features are essentially centroid features, and provide an SCF implementation that improves on the speaker recognition performance of both subband spectral centroid and FM features.

Low L. S. A et al. [11] With suicidal behavior being linked to depression that starts at an early age of a person's life, many investigators are trying to find early tell-tale signs to assist psychologists in detecting clinical depression through acoustic analysis of a patient's speech. The purpose of this paper was to study the effectiveness of Mel frequency cepstral coefficients (MFCCs) in capturing the overall mental state of a patient through the analysis of their various vocal emotions displayed during 20 minutes of problem-solving interaction sessions. We also propose both gender based and gender independent clinical depression models using Gaussian mixture models. Experiments on 139 adolescents subject corpus indicates that incorporation of both first and second time derivatives of MFCCs can improve the overall classification accuracy by 3%. Gender differences proved to be a factor in improving clinical depressed subject detection, where gender based models outperformed the gender independent models by 8%.

Sethu et al. [12] Spectral and excitation features, commonly used in automatic emotion classification systems, parameterise different aspects of the speech signal. This paper groups these features as speech production cues, broad spectral measures and detailed spectral measures and looks at how they differ in their performance in both speaker dependent and speaker independent systems. The extent of speaker normalisation on these features is also considered. Combinations of different features are then compared in terms of classification accuracies. Evaluations were conducted on the LDC emotional speech corpus for a five-class problem. Results indicate that MFCCs are very discriminative but suffer from speaker variability. Further, results suggest that the best front end for a speaker independent system is a combination of pitch, energy and formant information.

McIntyre et al. [13] associate the measurements with action units (AU) groups from the facial action coding system (FACS). However, we use the neologism region units (RU) to describe regions of the face that encapsulate AUs. In contrast to Ell-gring's approach, we automatically generate the measurements and provide both prototypical expression recognition and RU-specific activity measurements. Latency between expressions is also measured and the system is conducive to comparison across groups and individual subjects. By using active appearance models (AAM) to locate the fiduciary facial points, and multiboost to classify prototypical expressions and the RUs, we can provide a simple, objective, flexible and cost-effective means of automatically measuring facial activity.

E. Moore et al. [14] The motivation for this work is in an attempt to rectify the current lack of objective tools for clinical analysis of emotional disorders. This study involves the examination of a large breadth of objectively measurable features for use in discriminating depressed speech. Analysis is based on features related to prosodics, the vocal tract, and parameters extracted directly from the glottal waveform. Discrimination of the depressed speech was based on a feature selection strategy utilizing the following combinations of feature domains: prosodic measures alone, prosodic and vocal tract measures, prosodic and glottal measures, and all three domains. The combination of glottal and prosodic features produced better discrimination overall than the combination of prosodic and vocal tract features. Analysis of discriminating feature sets used in the study reflect a clear indication that glottal descriptors are vital components of vocal affect analysis.

Yingthawornsuk et al. [15] Research has shown that the voice itself contains important information about immediate psychological state and certain vocal parameters are capable of distinguishing speaking patterns of speech signal affected by emotional disturbances (i.e., clinical depression). In this study, the GMM based feature of the vocalt tract system response and spectral energy have been studied and found to be a primary acoustic feature set for separating two groups of female patients carrying a diagnosis of depression and suicidal risk.

Brierley et al. [16] It is now well established that emotion enhances episodic memory. However, it remains unclear whether the same neural processes underlie enhancement of memory for both emotional stimuli and neutral stimuli encoded in an emotive context. We designed an experiment that specifically attempted to separate these effects and that was validated on 30 participants. We then used functional magnetic resonance imaging (fMRI) to examine the neural correlates of encoding and retrieval of the two classes of stimuli in 12 healthy male volunteers. We predicted that aversive emotional context would enhance memory regardless of content and that activation of anterior cingulate would be inversely related to retrieval of aversive items. Both predictions were supported. Furthermore we demonstrated apparent asymmetrical lateralisation of activation in the hippocampal/parahippocampal complex during recognition of words from aversive sentences: more left-sided activation for neutral words from aversive contexts, and more right-sided activation for aversive content words. These findings, if applicable to the wider population, may have application in a range of psychiatric disorders where interactions between emotion and cognition are relevant.

A.Rush et al. [17] The 16-item Quick Inventory of Depressive Symptomatology (QIDS), a new measure of depressive symptom severity derived from the 30-item Inventory of Depressive Symptomatology (IDS), is available in both self-report (QIDS-SR(16)) and clinician-rated (QIDS-C(16)) formats. This report evaluates and compares the psychometric properties of the QIDS-SR(16) in relation to the IDS-SR(30) and the 24-item Hamilton Rating Scale for Depression (HAM-D(24)) in 596 adult outpatients treated for chronic nonpsychotic, major depressive disorder. Internal consistency was high for the QIDS-SR(16) (Cronbach's alpha =.86), the IDS-SR(30) (Cronbach's alpha =.92), and the HAM-D(24) (Cronbach's alpha =.88). QIDS-SR(16) total scores were highly correlated with IDS-SR(30) (.96) and HAM-D(24) (.86) total scores. Item-total correlations revealed that several similar items were highly correlated with both QIDS-SR(16) and IDS-SR(30) total scores. Roughly 1.3 times the QIDS-SR(16) total score is predictive of the HAM-D(17) (17-item version of the HAM-D) total score. The QIDS-SR(16) was as sensitive to symptom change as the IDS-SR(30) and HAM-D(24), indicating high concurrent validity for all three scales. The QIDS-SR(16) has highly acceptable psychometric properties, which supports the usefulness of this brief rating of depressive symptom severity in both clinical and research settings.

Ozdas, A et al. [18] sample consisted of ten high-risk near-term suicidal patients, ten major depressed patients, and ten nondepressed control subjects. As a result of two sample statistical analyses, mean vocal jitter was found to be a significant discriminator only between suicidal and nondepressed control groups ($p<0.05$). The slope of the glottal flow spectrum, on the other hand, was a significant discriminator between all three groups ($p<0.05$). A maximum likelihood classifier, developed by combining the a posteriori probabilities of these two features, yielded correct classification scores of 85% between near-term suicidal patients and nondepressed controls, 90% between depressed patients and nondepressed controls, and 75% between near-term suicidal patients and depressed patients. These preliminary classification results support the hypothesized link between phonation and near-term suicidal risk. However, validation of the proposed measures on a larger sample size is necessary.

France, D. J et al. [19] Acoustic properties of speech have previously been identified as possible cues to depression, and there is evidence that certain vocal parameters may be used further to objectively discriminate between depressed and suicidal speech. Studies were performed to analyze and compare the speech acoustics of separate male and female samples comprised of normal individuals and individuals carrying diagnoses of depression and high-risk, near-term suicidality. The female sample consisted of ten control subjects, 17 dysthymic patients, and 21 major depressed patients. The male sample contained 24 control subjects, 21 major depressed patients, and 22 high-risk suicidal patients. Lustberg L et al. [20] Physiological changes in sleep related to depression correlate with the likelihood of response to psychotherapy alone and may also identify which patients are unlikely to do well with psychosocial treatment and, therefore, to need somatic therapy in order to preserve recovery. Electroencephalographic (EEG) sleep changes also correlate with the speed of response and with the brittleness or durability of response (i.e probability of relapse or recurrence).

6. Depression Level In Audios Using I-Vector Technique

The technique used in our analysis depicts audio data using I-Vector method. We have used I-Vector technique to make the method durable against various other sources in the audios. First of all the audios will be loaded into the workspace. Our proposed method is divided into four different parts: In the first part we have trained the audio signals and then silence has been removed from the audio signals. In the second part features were extracted from audios using I-Vector. In the third part split overlapping function is applied to evaluate the overlapped audio beats. In the fourth part we have evaluated depression using relationship matrix. Table I refers to nomenclature of metrics used in equations.

TABLE I. Nomenclature of metrics used in equations

Metric	Full Name
temp	A
Amp	Amplitude
F	Frequency
Spt	Split
Ovp	Overlapping
Ado	Audio
Len	Length
Inc	Increment
Abs	Absolute
Min	Minimum
Max	Maximum
Coef	Coefficient
Fft	Fast Fourier transform
Dist	Distance
Corr	Correlation
Tp	True Positive
tn	True Negative
Fp	False Positive
Fn	False Negative
Sum	∑
Ft	Filter
Frame Len	Fl
Enframe	B

i. Silence Removal

In this part we have removed silence from the audio signals using the following equations as given below. In equation 1 we calculated the sum of amplitude frequency of signal. In equation 2 and 3 we calculated the minimum amplitude frequency of signal to remove silence from the signal.

$$\mathrm{amp_f} = \sum \left(\mathrm{abs}\left(\mathrm{spt_{ovp}}(\mathrm{ado}([1 - 0.9375], 1, x), \mathrm{adoLen}, \mathrm{adoInc})\right), 2\right); (1)$$

$$\mathrm{amp_{f1}} = \min\left(\mathrm{amp_{f1}}, \frac{\max(\mathrm{amp_f})}{4}\right); (2)$$

$$\mathrm{amp_{f2}} = \min\left(\mathrm{amp_{f2}}, \frac{\max(\mathrm{amp_f})}{8}\right); (3)$$

Code for silence removal

```
if amp_fq(n) > amp_fq1
x1 = max(n-count-1,1);
elseif amp_fq(n) > amp_fq2 OR zcount(n) > zcount2
silent = silent+1;
if silent < peak_silent
silent = 0;
end
```

ii. Feature Extraction Using I-Vector

The In the second part we have extracted features from audio signals using I-Vector method using the following algorithm:

a. Initialize vector variable with Melvectorm function using equation 4.

$$\mathrm{vector} = \mathrm{melvectorm(audio)};\ (4)$$

b. Calculate frequency to bit-ratio and store it in random variable l_r using equation 5.

$$l_r = \frac{\log\left(\frac{f_{zero} + fh}{f_{zero} + f1}\right)}{p + 1};\ (5)$$

c. Convert l_r value to fast fourier transform bin numbers using equation 6, 7, 8, 9 and 10.

$$b1 = n * ((f_{zero} + f1) * \exp([0 \quad 1 \quad p \quad p + 1] * l_r) - f_{zero});\ \ (6)$$

$$\mathrm{p_f} = \frac{\log\left(\frac{\mathrm{f_{zero}} + \frac{\mathrm{b2:b3}}{\mathrm{n}}}{\mathrm{f_{zero}} + \mathrm{f1}}\right)}{\mathrm{l_r}};\ (7)$$

$$r = [ones(1, b2)fp \quad fp + 1 \quad P * ones(1, frq_{n2} - b3)];\ \ (8)$$

$$c = [1{:}b3 + 1 \ \ b2 + 1{:}frq_{n2} + 1;];\ (9)$$

$$\mathrm{v} = 2 * \left[0.5\backslash(1, \mathrm{b2} - 1)1 - \mathrm{p_f} + \mathrm{f_p} \ \ \mathrm{p_f} - \mathrm{f_p}(1, \mathrm{frq_{n2}} - \mathrm{b3} - 1\ )0.5\right]; (10)$$

d. Using equation 11 and 12 we calculated the value of melvectorm function.

$$\text{vector} = 1 - \frac{0.92}{1.08} * \cos\left(\text{v} * \frac{\text{pi}}{2}\right);\ (11)$$

$$vector = \frac{vector}{\max\big(vector(:)\big)};\ (12)$$

e. In equation 13 and 14 we calculated the vector value of signal and store it in variable w.

$$w = 1 + 6 * \sin\left(pi * \frac{[1{:}\,12]}{12}\right);\ (13)$$

$$w = \frac{w}{\max(w)};\ (14)$$

f. Store the result.

iii. Evaluate Overlapped Audio Beats

In this part we have evaluated overlapped audio beats using split overlapping function using the following algorithm:

a) Read the audio data.
b) Multiply the audio data with the hamming distance.
c) Apply the fast fourier transform function on the above data.
d) Calculate the distance vector and store it in distance vector variable.
e) Calculate correlation matrix.
f) Evaluate the overlapped audio beats using split overlapping function.
g) Store the result.

iv. Evaluate Depression Using Relationship Matrix

In this part we have evaluated depression using the Following equation:

$$relation(i,j) = sum\left(\big(t(i,:) - r(j,:)\big).^{2}\right);\ (15)$$

In this equation here t is the trained signal and r is the input signal. For given signal t, if the matched value of r signal is high then depression will be the depression value stored in the trained data set.

4.2 Estimation of Depression Level in Audios Using Fuzzy Membership Functions

The technique used in our analysis depicts audio data using fuzzy membership function. We have used fuzzy membership functions to make the method durable against various other sources in the audios. First of all the audios will be loaded into the workspace. Our proposed method is divided into five different parts: In the first part we have normalize the audio signals. In the second part we have evaluated amplitude by applying audio filtering. In the third part we have evaluated changes in amplitude audio. In the fourth part we have defined fuzzy rules. In the fifth part it will return membership values.

- Normalize Filter

$$x = \frac{x}{\max\big(abs(x)\big)};\ (16)$$

$$x1 = numel(x);\ (17)$$

$$\alpha 1 = \beta(x(1{:}\,end - 1),\, fl, inc);\ (18)$$

$$\alpha 2 = \beta(x(2{:}\,end)\ ,\, fl, inc);\ (19)$$

$$signs = (\alpha 1.*\, \alpha 2), 0;\ (20)$$

$$diffs = (\alpha 1 - \alpha 2) > 0.02;\ (21)$$

$$zcr = sum(signs.*\, diffs, 2);\ (22)$$

- Evaluate amplitude by applying audio filtering

$$\text{amp} = \sum\left(\left(\text{abs}\big(\beta(\text{ft}([1\ \ -0.9375], 1, \text{x}), \text{fl}, \text{inc})\big)\right), 2\right);\ (23)$$

$$\text{av}_{\text{zcr}} = \frac{\text{sum(zcr)}}{\text{len(zcr)}};\ (24)$$

$$\text{av}_{\text{amp}} = \frac{\text{sum(amp)}}{\text{length(amp)}};\ (25)$$

$$\max \text{amp} = \max\big(\text{abs(amp)}\big);\ (26)$$

$$\min \text{amp} = \min\big(\text{abs(amp)}\big);\ (27)$$

```
Case 2
if (Fvzcr(n) 0.5+NFvzcr(n) 0.3+NFvamp(n) 0.2)>0.8
  count = count + 1;silence=0;
  else
  silence = silence+1;
  if silence < maxsilence %
  count = count + 1;
  else
  status = 3;
  end
  end
  Case 3,
  break;
  end
end
count = count-silence;
x2 = x1 + count -1;
```

ii. Return Membership Values

$$\text{final}_{\text{f}} = \frac{\text{abs(x2+count)}}{\text{abs(x2*count)}}; \quad (28)$$

7. Visual Analysis

Figure1, Figure2 and Figure 3 are showing the input or actual signal as well as the test or matched signal and both the signals are very similar. After that we will calculate the value of accuracy, specificity, peak signal to noise ratio, f-measure and balanced classification rate for test signal. After calculating the values of these parameters we will estimate the depression level in each signal.

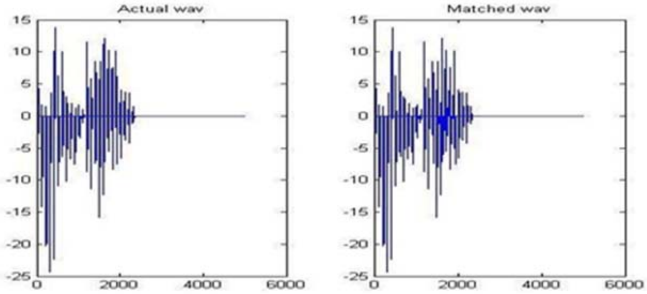

Fig.1 Diagram of actual waveform and matched waveform

8. Performance Evaluation

This section contains the comparison table and graphs of the existing and proposed techniques. Some well-known image performance evaluation parameters for audio signals have been selected to prove that the performance of the proposed algorithm is quite better than the existing method.

1. MAE – In statistics, the mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad (1)$$

The values of Mean absolute error are shown below in the comparison Table 5.1

Table 5.1: Mean absolute error comparison table

Input Audio Signals	Existing Result	Proposed Result
1	17.5924	0.3093
2	22.7627	1.5122
3	19.3833	1.5283
4	29.1466	8.1854
5	20.5230	0.6322
6	22.5241	1.2103
7	23.0225	1.2151
8	28.8334	7.3896
9	17.2772	1.5283
10	25.4083	4.4296

Figure 5.1 has shown the comparison table of the mean square error of different audio signals by Existing value in (Blue line) & proposed values in (Red lines). It is very clear from the graph that there is decrease in MAE value of signals with the use of proposed method over existing methods.

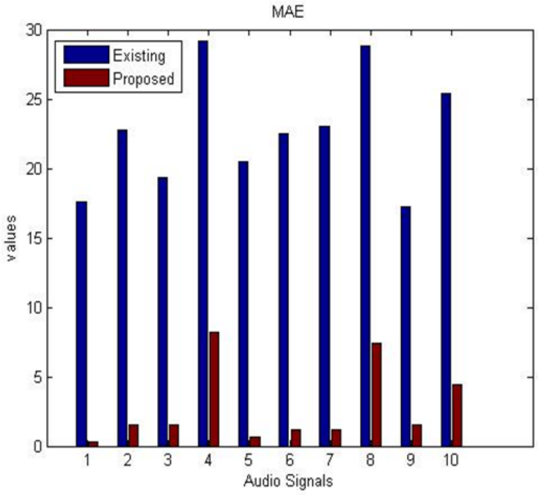

Figure 5.1: Mean Absolute error graph

This accuracy shows the accurate value of the input signal in case of the proposed algorithm as well as in existing algorithm.

Table 5.4: Accuracy comparison table

Input Audio Signals	Existing Result	Proposed Result
1	0.9569	0.9980
2	0.9533	0.9788
3	0.9565	0.9932
4	0.9467	0.9693
5	0.9557	0.9932
6	0.9441	0.9836
7	0.9499	0.9836
8	0.9293	0.9645
9	0.9617	0.9932
10	0.9305	0.9788

The values of accuracy are shown below in the comparison Table 5.4.

Figure 5.4 has shown the quantized analysis of the Accuracy. It is very clear from the plot that the value of Accuracy is getting maximized in every case with the use of proposed method over other methods.

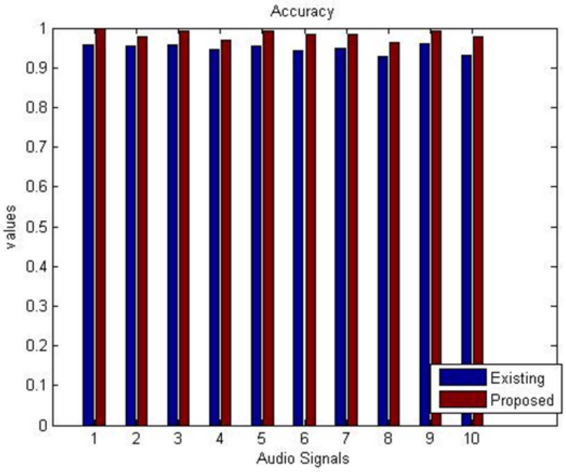

Figure 5.4: Accuracy graph

CONCLUSION

Depression has recently been attracting the attention of speech researchers, with audio/visual emotion challenge (AVEC) 2013 and 2014 organized to encourage researchers to develop approaches to accurately estimate speaker depression level. This dissertation has focused on speaker dependence of an I-Vector based depression level estimation system. I-Vector based depression level estimation system has better performance than existing techniques. In the existing depression level estimation techniques silence has not been removed from the input signal and fuzzy membership functions are also not used. In order to improve the accuracy of I-Vector based depression level estimation system we will design a fuzzy membership function and silence removal function in audio signals. The comparisons have clearly indicates that the proposed technique outperforms over the available techniques in terms of accuracy.

This work has not utilised any evolutionary optimization technique to match the trained depression audios with testing one so in near future we will utilize different evolutionary approaches to enhance the results further.

REFERENCES

[1] Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo "Assessing speaker independence on a speech-based depression level estimation system" Pattern Recognition Letters (2015) Elsevier, pp. 1-8.

[2] P.Ghahremani, B.Baba Ali, D.Povey, K.Riedhammer, J.Trmal, S.Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition", in:Proceedings of ICASSP, 2014, pp.2494–2498.

[3] S.Alghowinem, R.Goecke, M.Wagner, J.Epps, G.Parker, M.Breakspear, "Characterising depressed speech for classification", in:Proceedings of Inter speech. ISCA, 2013, pp.2534–2538.

[4] N.Cummins, J.Joshi, A.Dhall, V.Sethu, R.Goecke, J.Epps, "Diagnosis of depression by behavioural signals: a multimodal approach", in: Proceedings of AVEC'13, 2013, pp.11–20.

[5] J.R.Williamson, T.F.Quatieri, B.S.Helfer, R.Horwitz, B.Yu,D. D.Mehta, "Vocal biomarkers of depression based on motor incoordination", in: Proceedings of AVEC' 13, New York, NY, USA, pp.41–48, 2013.

[6] M.Valstar, B.Schuller, K.Smith, F.Eyben, B.Jiang, S.Bilakhia, S.Schnieder, R.Cowie, M.Pantic, AVEC2013- "The continuous audio/visual emotion and depression recognition challenge", in: Proceedings of AVEC'13, 2013.

[7] L.J.Rodríguez-Fuentes, M.Penagarikano, A.Varona, M.Díez, G.Bordel, KALAKA 2: "a TV broadcast speech database for the recognition of Iberian languages in clean and noisy environments", in: Proceedings of LREC, pp.99–105, 2012.

[8] D.Povey, A.Ghoshal, G.Boulianne, L.Burget, O.Glembek, N.Goel, M.Hannemann, P.Motlicek, Y.Qian, P.Schwarz, J.Silovsky, G.Stemmer, K.Vesely,"The Kaldi speech recognition toolkit" , in: IEEE Workshop on Automatic Speech Recognitionand Understanding. IEEE Signal Processing Society, 2011.

[9] Niels Rosenquist, J., Fowler, J. & Christakis, N. "Social Network Determinants of Depression". Molecular Psychiatry 16 (3): 273–281, (2011).

[10] Kua, J. M. K., "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition." Proc. Odyssey: Speaker and Lang. Rec. Workshop, 2010, pp. 34 - 39.

[11] Low, L. S. A., N. C. Maddage, et al. "Mel frequency cepstral feature and Gaussian Mixtures for modeling clinical depression in adolescents." in Proc. IEEE Int. Conf. on Cognitive Informatics, 2009, pp. 346-350.

[12] Sethu, V., et al., "Speaker dependency of spectral features and speech production cues for automatic emotion classification", in Proc. IEEE ICASSP, 2009, pp. 4693-4696.

[13] McIntyre, G., R. Göcke, et al., "An approach for automatically measuring facial activity in depressed subjects", in Proc. Int. Conf. on Affective Computing and Intelligent Interaction and Workshops, 2009.

[14] E. Moore, et al. “Critical analysis of the impact of glottal features in the classification of clinical depression in speech,” IEEE Trans. Biomed. Eng., vol. 55, 2008, pp. 96-107.

[15] Yingthawornsuk, T., et al., "Direct Acoustic Feature Using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Speech." in Proc. Interspeech, 2007.

[16] Brierley, B.N. Medford., "Emotional memory for words: Separating content and context", Cognition & Emotion, 2007. 21(3): p. 495-521.

[17] Ozdas, A., “Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk”, IEEE Trans. Biomed. Eng., vol. 51, no. 9, 2004, pp. 1530- 1540.

[18] A.Rush, M.Trivedi, H.Ibrahim, T.Carmody, B.Arnow, D.Klein, J.Markowitz, P.Ninan, S.Kornstein, R.Manber, M.Thase, J.Kocsis, M.Keller, “The 16 item quick inventory of depressive symptomatology (QIDS) clinician rating (QIDS-C) and self-report (QIDS-SR):a psychometric evaluation in patients with chronic major depression”, Biol. Psychiatry 54, pp.573–583, 2003.

[19] France, D. J. “Acoustical properties of speech as indicators of depression and suicidal risk”, IEEE Trans. Biomed. Eng., vol. 47, no. 7, July 2000, pp. 829-837.

[20] Lustberg L, & Reynolds CF, “Depression and insomnia: questions of cause and effect. Sleep Medicine Reviews 4 (3): 253–262, (2000).