RESEARCH ARTICLE OPEN ACCESS

A Survey: Pre-processing and Feature Extraction Techniques for Depression Analysis Using Speech Signal

Dipti Patil[1], Shamla Mantri[2], Ria Agrawal[3], Shraddha Bhattad[4],
Ankit Padiya[5], Rakshit Rathi[6]
Department of Computer Engineering
MAEER's MIT College of Engineering,
Kothrud, Pune-411038
Maharashtra, India

ABSTRACT
Clinical depression has become a major factor causing suicides and is associated with high mortality rates making it the leading causes of death worldwide. Thus, early detection of depression is of primary importance as there has been a dramatic increase in the depressive symptoms and disorders in recent years. From a psychological point of view, one of the signs of a person being depressed is the way emotions are expressed in his/her speech. Depressive disorder affects the acoustic qualities of their speech; hence depression can be detected through an analysis of acoustical properties of speech. Features are extracted from the preprocessed speech signal and accordingly using non-linear classifiers speech is classified as depressed or controlled.

Keywords:- Acoustic features, Clinical depression, linear technique, nonlinear technique.

I. INTRODUCTION

Speech is a natural form of communication for human beings and has been recognized as one of the potential sources for detection of depression. Doctors use the term "clinical depression" to describe the more severe form of depression also known as major depression [1].When a depressed person pretends to be well , listeners can predict whether the person is depressed through his voice tones even though he is speaking proper linguistic contents. With the help of the voice tone and word emphasis we can predict person's emotional state, attitude, moods.

The evolution of information related to speech processing and speech recognition technologies in computers has made it possible to develop objective measures that can measure these speech cues.The acoustic qualities of a person's speech are affected by emotional state of a depressed person and thus we can detect this depressive disorder through analysis on changes in acoustical properties of speech [2].

The increase prevalence of clinical depression has been linked to of serious outcomes such as an increase in the number of suicides and deaths [1].The key to suicide prevention is early intervention for depressive disorder. The remaining paper consists of work done previously. It includes brief review on the architectural blocks: database, preprocessing of speech signal, feature extraction of preprocessed speech signal and the classifiers that are used for classification of controlled and depressed speech.

II. PREVIOUS WORK

A. Database

Database is used for storing patient's utterances which are used in speech information processing and speech recognition technologies [3]. The speech database was of key importance used for validation of recorded speech data. Following is the details of the method that was followed for collecting the data:

Participants: Database was created by collecting the speech data within the span of two years. This data was devised by researchers of Oregon Institute (ORI), USA.152 participants were considered for the research. Out of which 52 were male and 100 females. The entire participant's age was within the range of 14-18 years. Out of the 152 participant 75 were depressed whereas 77 participants were healthy. The depressed participants were not given any kind of medication. Depressed and controlled participants were compared and it was found that depressed participants were from low socio economic background and they had depressive history.

Observation: The next step was observation. The behaviour of the participants with their parents was observed through their interactions. Three kind of interactions were considered –

1) Event Planning Interaction (EPI): This includes discussion on planning of event such as planning for trip etc.

2) Problem-solving Interaction (PSI): At the start of the interaction both the participants and their parents were given a topic which had the mutual disagreement of both the parents and the participants and they were asked to discuss the problem and come to a solution

.

3) Family Consensus Interaction (FCI): It includes imagination of writing a book chapter which reflects the shared perspective by both the participants and their parents.

Each of these interactions was conducted for 20 minutes. These interactions were audio and video recorded. The equipment used for recording were: Video Bank TM System which was used for video recording, Audio Technical (model: ATW-831-w-a300) lapel wireless microphones used for audio recording, OS-3 hand-held microcomputers (observations systems, Inc.) which was used by the observers to observe the system.

Apart from this participants were also outfitted with other sensors that measure the physiological signals such as electrocardiograph (ECG), impedance cardiogram (ICG), skin conductance, respiratory and blood pressure.

The Living in Family Environment system (LIFE) was used to code the behavior during interaction. This system describes the specific timeline of various emotions (called affect codes) and verbal content (called content codes).

From all the recorded data the audio recordings from microphones of participants were considered. These recordings were then segmented. On an average these segments were 2 to 3 seconds long. Anti-aliasing filter was used for segmentation. It was followed by down sampling of audio signal from 44.1 kHz to 11.025 kHz sampling. Out of the 152 participants 139 participants were only considered for further analysis. It included 46 males and 93 females. Among 138 participants 68(19males and 49 females) were depressed. The average number of utterances for each participant was 278, 251 and 240 for EPI, PSI and FCI from the speech corpus. The ratio of the participant's to parent's speech duration was 0.73 for EPI, 0.71 for PSI and 0.67 for FCI.

B. *Pre-processing*

Pre-processing is basically used for removing the unwanted signal from the speech signal. In pre-processing gaps are removed from the speech signal to make it continuous signal [1]. Firstly, the speech signal was normalized and then it was segmented into 25 msec with 50% overlapping frames using a rectangular window [4]. If the last frame is shorter than 25 msec then it is appended with random noise of 30db. Segmented frames are then filtered with a zero phase filter to remove any low-frequency drift. The 13th order linear prediction coefficients (LPCs) per frame were calculated. Energy of the prediction error and the first reflection coefficient r1 were calculated and a threshold was empirically set to detect the voiced frames where N is the number of samples in analysis frame and s (n) is speech sample.

$$r1 = \frac{\frac{1}{N}\sum_{n=1}^{N-1} s(n)s(n+1)}{\frac{1}{N}\sum_{n=1}^{N} s(n)s(n)} \qquad (1)$$

C. Feature Extraction

Elements present in speech sound i.e. Acoustic features that are associated with human speech production mechanism leads to classification speech into controlled and depressed one. Acoustic features that represented objective measurements in the human speech production were extracted.

The acoustic features were grouped into five main feature categories that represented the TEO-based, cepstral (C), prosodic (P), spectral (S), and glottal (G) features. Out of which TEO-based feature category is non- linear and remaining are based on linear speech production model.

These categories define what elements are functional with regards to speech and understands how human listeners recognize speech sounds and use this information to understand spoken language i.e. Physiological and perceptual components of speech.

*1) **Glottal category :*** Glottal features category are vital component of vocal affect analysis because they refer to emotional expression of speech and its relation to overall state of speaker. Glottal feature was extracted by inversing the estimated vocal tract and lip radiation filters from the source of speech signal [1].

For glottal flow extraction TTK Aparat glottal inverse filtering toolbox was used [5].Using this flexible software package and a graphical user interface: a) effects of vocal tract and lip radiation are removed from a signal. b) And it is based on iterative adaptive inverse filtering (IAIF) that is able to give a fairly accurate estimate for glottal flow [7]. A discrete All Pole Modelling (DAP) was used instead of linear predictive filter to model vocal tract. In this method, the number of formants (i.e. Resonance) to model the vocal tract was set to 13(Fs/1000+2) so that there is atleast one formant for every kHz band in vocal tract transfer function.

When glottal flow estimation was done then quantitative analysis of glottal flow pulses was performed in time and frequency domain. Glottal Timing (GLT) and Glottal Frequency (GLF) were used to represent glottal flow parameters.

Glottal Timing (GLT) : Glottal flow is divided into few phases that describe the glottal pulse shape. These are closed, opening, closing, closed. Opening phase is subdivided into two timing instances i.e. primary opening and secondary opening whose durations are given by T01 and T02 respectively. Duration of closing phase is given by Tc and period of glottal cycle is given by T. Using this parameters several timing and frequency parameters can be calculated. Several timing parameters in GLT are :Open Quotients (OQ1, OQ2), Approximation of Open Quotients (OQa), Quasi Open Quotient (QOQ), Speed Quotient (SQ1 and SQ2) ,Closing Quotient (CIQ), Amplitude Quotient (AQ), Normalized Amplitude Quotient (NAQ) .

Glottal Frequency (GLF) : Parameters in GLF are [8] difference of 1st and 2nd harmonics in decibels of glottal flow power spectrum, harmonics richness factor and parabolic spectral parameter (PCB).

Glottal domain features have been linked to the overall voice quality of speech and more sensitive to subtle voice changes that other features can miss. It is equivalent of analysing the voice at its source of production which is the onset of all affective expressions.

*2) **TEO-based category**:* TEO is a key feature in recognizing stress. Vortex flow varies accordingly to emotional state of anger or stressed speech because there is fast air flow that causes vortices located near the false vocal folds that gives additional excitation signals other than pitch. To measure this energy which is produced by such a non-linear process, Teager developed an energy operator which is known as Teager energy Operator (TEO) given as follows

$$\Psi[x\,(n)] = x2\,(n) - x\,(n+1)\,x\,(n-1) \qquad (2)$$

Where, ψ [.] is the TEO and x (n) is the nth speech sample point.

As suggested by Zhou et al [9], if the speech signals broken into smaller bands also called as critical bands(CB) and TEO parameters are calculated for each band, presence or absence of additional harmonic components can be easily obtains which can be further used for processing. For this, smaller bands of speech spectrum are obtained by Gabor filters before calculation of TEO profile for each band. The area under normalized auto correlation environment is then calculated in each critical band to give TEO-CB-Auto-Env features. Same frequency range was followed for 15 critical bands [9]. TEO-CB-Auto-Env method is used for emotional stress classification.

TEO-CB-Auto-Env feature extraction process is carried out as shown in figure [10].

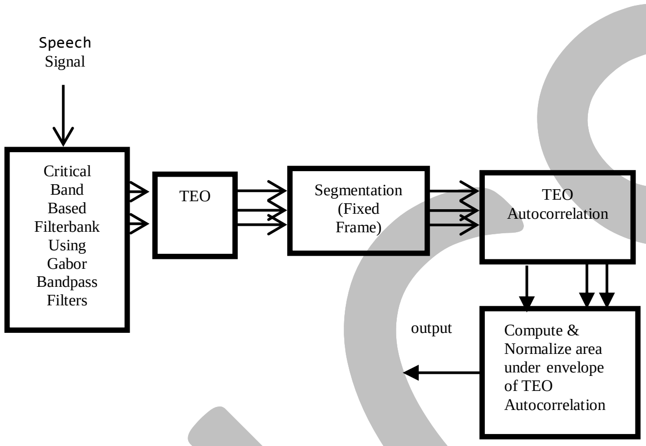
Speech
Signal
Critical Band Based Filterbank Using Gabor Bandpass Filters
TEO
Segmentation (Fixed Frame)
TEO Autocorrelation
output
Compute & Normalize area under envelope of TEO Autocorrelation

Fig 1 TEO-CB-Auto-Env feature extraction process

3) Prosodic: It involves variation in timing as well as subjective perception in loudness and pitch. It may reflect the emotional state of the speaker[6]. Prosodic have been widely used while studying the speaking behavior, emotion and stress recognition as we know depression which is associated with monotonous speech, subtle changes in expression and pitch, the prosodic feature can be expected to provide strong cues for depression. Its characteristics features are:

Fundamental frequency (f0) and log energy(log E): In this, auto-correlation method is compared with the cepstrum method and the average magnitude difference method using Roger Jangs audio box. All the above method gives near about same results but autocorrelation method is used .This method is used to determine the changes in speaking behavior with respect to factors relating to stress, intonation and emotional changes[6]. The values of f0 were calculated by using autocorrelation method.

$$acf(lag) = \sum_{n=0}^{N-1-lag} s(n)s(n+lag) \qquad (3)$$

The value of log E is also calculated for each frame to determine the changes in stress, emotion and pitch. The log E is short term energy.

$$Es(m) = \log \sum_{n=m-N+1}^{m} s^2\,(n) \qquad (4)$$

Formants (fmts) and formants bandwidth (fbws): A 13th order LP filter was used to calculate formant frequency. First three formants (FMT1-FMT3) and formant bandwidths (FBW1-FBW3) below its Nyquist frequency were taken[11].

Jitter and Shimmer: Jitter is nothing but a fluctuation in pitch. It is calculated using fundamental frequency of each cycle, subtracting from previous value of f0 and then divide it by f0.

$$Jitter = \frac{\frac{1}{N-1}\sum_{i=1}^{N=1}|F_{oi}-F_{oi+1}|}{\frac{1}{N}\sum_{i=1}^{N}F_{oi}} \qquad (5)$$

Where *i* is the frame number and *N* is the total number of frames.

Shimmer is similar to jitter but the little difference is in shimmer we calculate peak to peak amplitude of a signal.

$$Shimmer = \frac{\frac{1}{N}\sum_{i=1}^{N-1}|A_i-A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \qquad (6)$$

Where *i* is the frame number and *N* is the total number of frames [12].

4) Cepstral: In this we calculate the Mel Frequency Cepstral Coefficient (MFCC). MFCC is widely used in speech processing [6]. It uses filter banks based on human auditory system that transform linear frequency to logarithmic scale. It also gives the relationship between linear frequency scale (flinear) and mel-scale (fmel). Mel is the unit of perceived speech. In this filter banks are more important. This is more important since our whole understanding of speech is through our ears [11].

5) Spectral: Spectral centroid indicates the center of signals spectrum power distribution. It is derived from weighted mean of frequencies present in the signal, with magnitude as weight. Spectral flux is a measure of how quickly the power spectrum of a signal changing. It is calculated by comparing power spectrum of one frame against another frame. Spectral entropy is the means of measuring the amount of information based on the Shannon's information theory. It is used to emotion recognition in speech. Spectral roll-off is the point where frequency that is below some percentage of the power spectrum resides. Power spectral density (PSD) has been used to discriminate between speech of control and depressed [13].

III. CONCLUSION

In this paper we have studied various stages involved in detection of depression using speech signal: signal preprocessing which involves preprocessing of the speech data obtained from the database. Feature Extraction of preprocessed speech using various features of speech such as Prosodic, Glottal, Cepstral, Spectral and TEO. From this study we concluded that TEO is a linear feature while others are non- linear features of speech signal. And accordingly we found combination of TEO with Prosodic and Glottal features provide us with an accurate output in classifying the state of a person.

ACKNOWLEDGMENT

We wish to express our sincere gratitude to Prof. Dr. V. M. Wadhai, Principal, Prof. Dr.Prassana Joeg H.O.D of Computer Engineering Technology, MITCOE PUNE for guiding us in this survey. We also thank to our friends and other MITCOE staff members for guidance and encouragement in carrying out this work. We would like to thank them for their valuable guidance throughout the preparation of this.

REFERENCES

[1] L. S. Low, N. C. Maddage , M. Lech, L. B. Sheeber, and N. B.Allen,“Detection of clinical depression in Adolescents' speech during family interactions,” *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586,Mar. 2011.

[2] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, “Acoustical properties of speech as indicators of depression and suicidal risk,” *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul. 2000.

[3] H. Hops, A. Biglan, J. Arthur, L. Sherman, A. Tolman, and N. Longoria. (2003). Living in family environments (LIFE) coding[14] system: Reference manual for coders. [Online]. Available:

[4] D. G. Childers, *Speech Processing and Synthesis Toolboxes.* Chichester, New York: Wiley, 2000.

[5] M.Airas, “TKKAparat: An environment for voice inverse filtering and parameterization,” *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, pp. 49–

[6] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete Time Proc. SpeechSignals*. Upper Saddle River, NJ: Prentice Hall PTR, 1999.

[7] P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering,” *Speech Commun.*, vol. 11, no. 2–3, pp. 109–118, 1992.

[8] E. Moore, M. A. Clements, J.W. Peifer, and L.Weisser, “Critical analysis of the impact of glottal features in the classification of clinical depression in speech,” *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 96–107, Jan.2008.

[9] ZHOU, G. J., HANSEN, J. H. L., and KAISER, J. F., Nonlinear feature based classification of speech under stress,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 201-216, 2001.

[10] J. F. Kaiser, “On a simple algorithm to calculate the ‘energy’ of a signal,” in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process*, 1990, pp. 381– 384.

[11] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Deillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, “The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals,” in *Proc. Interspeech*, 2007, pp. 2253–2256.

[12] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, “Telephony-based voice pathology assessment using automated speech analysis,” *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 468–477, Mar. 2006.

[13] D. J. France, R. G. Shiavi, S. Silverman, M.silverman, and D.M.Wilkes, “Acoustical properties of speech as indicators of depression and suicidal risk,” *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul.2000.