ISSN 0971-3034

Sponsored by
ALL INDIA COUNCIL
FOR TECHNICAL EDUCATION

Special Issue of the
First National Conference in the Emerging Vistas of Technology in 21st Century

Eco-Friendly Communication

THE INDIAN JOURNAL OF TECHNICAL EDUCATION

Organized By
Gujarat Technological University

Supported By
Parul Institute of Engineering & Technology

Promoted by
INDIAN SOCIETY FOR TECHNICAL EDUCATION
Near Katwaria Sarai, Shaheed Jeet Singh Marg,
New Delhi - 110 016

Analytical Change in Power Spectrum Density in Speech for Normal and Depressed Person

Navin Khambhala
M.E. student, E & C Dept., Parul institute of Engineering & Technology, Limda, Waghodia
newnavin009@gmail.com

Hardik Mewada
Asst. Professor, E & C Dept., Parul institute of Engineering & Technology, Limda, Waghodia
sjbv.hardik@gmail.com

Anil C. Suthar
Asst. Professor, Department of Electronics and communication engineering, LCIT, Bhandu, Mahesana
acsuthar@yahoo.co.in

ABSTRACT

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 normal speech. In this research, these questions are addressed empirically using classifies configurations employed in emotion recognition from speech, evaluated on depressed/neutral speech database. Results demonstrate that detailed spectra features are well suited to the task; speaker normalization provides benefits mainly for less detailed features and dynamic information appears to provide little benefit. In this research first voiced part is separated from whole speech. The Spectral feature was extracted from the collected voiced speech samples using the method of the spectrum analysis.

KEY WORDS: ***Depression; Spectral power spectrum density; Speech; Voiced/Unvoiced.***

INTRODUCTION

More recent work has seen the first steps towards automatic analysis of speech. 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. 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. Spectral centroid based methods including the sub-band spectral centroid features have recently shown promise in other applications, and other work shows that these newer measures potentially include information useful in the classification of depression. As Proposed by (France, Shiavi, Silverman, and Wilkes ,2000), Although systems for the classification of depressed speech have been proposed, there is considerable further research to be conducted, particularly in light of the extensive speech-based emotion recognition literature, from which insight can be gained towards the classification of depressed speech. In this research, we examine the effect of segment selection and the choice of different speech characterization methods on the automatic classification of depressed and neutral speech. Our life under the pressure of social competition today may cause some people to feel stressful, having a hard time catching up on things every day of their lives rushing. Some have thought to approach the edge of hard-

survival time which may affect their feelings, spirit and mind temporarily and / or permanently. Without being properly supervised by the care taker or to obtain appropriate treatment in time, such damage could be more serious if the person suffers from depression. How to know when this strike would be the first symptoms to start with the first step. People at risk of such a disorder may feel unhappy and depressed, isolated, lack of energy, loss of an aperitif, and even though no hope in life and finally give their lives. Suicide could be the chance for someone severely depressed at the end-their life unpleasant and treatment effort could be powerless to that mindset. Identify depressive people among others who are normal is a critical problem in clinical practice. To assess the psychiatric status of the subject, the practitioner should involve the collection of information about the balance sheet profile, visit the hospital records, crime-related reports and oversees health care hotline. Clinical judgment is another important procedure that is based largely on information data processing and interpretation of results were analyzed in the clinical perspective. This task to acquire clinical data and psychiatric diagnosis is the procedure time and should be treated with clinical expertise. Another in some way to address this important clinical practice must be discussed and proposed. The rapid assessment of symptoms of the subject by the psychiatrist is necessarily required for decision making on the subject real serious category is part of an appropriate treatment to be accorded to this subject.. In the point of view of speech, the height of high and low speech sounds depends on a change in fundamental frequency over the time interval to meet the glottal wave generated directly from the dynamic supply system in the speech production. Change in the range of fundamental frequency, energy shift and the spectral energy ratios on the frequency bands were studied and reported as predictors of the mental state associated with depression. Methods for estimating the energy of the speech signal are varied and diverse, as decided by the researchers to choose and implement. An estimation method known is the estimate of PSD of speech and how to estimate this function is convenient for implementation. This speech is the main objective to be described on its relevance to the classification of diagnostic symptoms.

METHODOLOGY

Database and Preprocessing

The database consists of patients who were categorized as either depressed or normal by a clinician who was not involved in the research project. All speech recordings were made on mono channel. All voice recordings were digitized with a higher sampling rate of 16 kHz for better sound quality. These recordings were then edited using the NCH wave pad audio digital editor by removing the interviewer's voice, removing long pauses that are present for more than 0.5 second and removing background noises such as door slams, sneezing and paper rustling sounds.

Speech signals are comprised of voiced, unvoiced, and short silence segments that are mixed and combined together. According to (Ozdas,2004), voiced, unvoiced and silence speech samples can be estimated by segmenting the sampled signals based on their energy values at different levels of the Wavelet Transform (WT). Voiced speech samples exhibit a quasi-stationary behavior and are composed of low frequency characteristics. On the other hand, unvoiced speech samples exhibit noise-like behavior and contains more high frequencies. The sampled signals were separated into segments and for each segment; the energy was calculated for each of several different band levels.

Table 1: Frequency range for each band levels

Band level	-3dB band pass filter limits
1	2500 – 5000
2	720 – 2340
3	320 – 1080
4	160 – 540
5	80 – 260

Energy for each band was obtained by calculating the Discrete Wavelet Transform (DWT) at each frequency band levels. These bands represent a set of band pass filters as shown in Table 1. Lower bands allow higher frequency content and filter out low frequency information while higher bands capture the lower frequency information.

For this study, a similar method was used for the voiced, unvoiced and silence classification but instead of using the WT to determine the energy bands, a set of third order band-pass filters was applied to each segment of the sampled signal. If a filtered speech segment has maximum energy equal to the total energy in band one, it was classified as unvoiced. The median of the total energy in band three was set as the threshold for separating voiced and silence where energy higher than or equal to the threshold was categorized as voiced otherwise it was categorized as silence.

All unvoiced and silence terms were removed and

only the voiced terms were collected and concatenated into one new speech signal for further analysis.

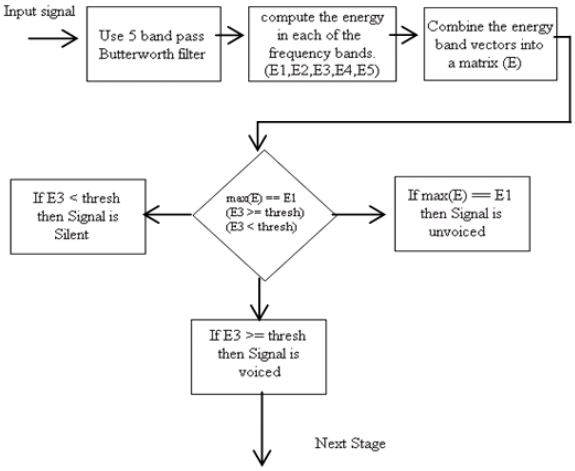

Fig. 1 Block diagram for voiced, unvoiced and silent detection

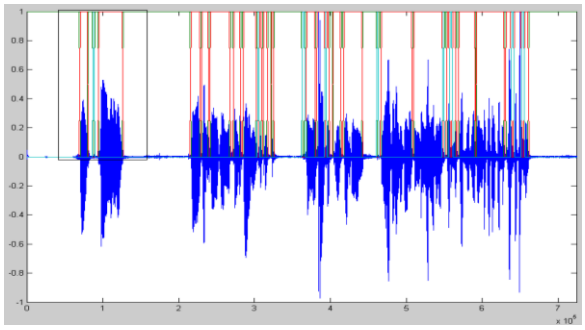

Fig. 2 voiced, unvoiced and silent part of speech

Feature Extraction

In this section of paper, only the voiced speech is further analysed for feature extraction. The procedure to obtain PSD features is described as:

- Perform the voiced/unvoiced detection on each patient's speech sample to obtain only the voiced segments of speech samples.
- Detrend and normalize the voiced speech by subtracting mean, dividing by standard deviation, and then separate into 10-second segments.
- Divide each 10-second segment into 40-msec frames (250 frames).
- Estimate PSD of each speech frame using with a non-overlapping window and a 512-point FFT.
- Divide the spectral region within a frequency range of 0-2,000 Hz into eight equal 250 Hz bands.
- Calculate the total area (energy) under the spectral curve within a 0-2,000 Hz range and sub-area in each 250Hz band by using a built-in MATLAB function, called "Trapz function."
- Calculate the ratios of energy in each 250 Hz band to total energy over a frequency range of 0-2,000 Hz.
- Repeat all procedures starting until all 40-msec frames of speech have been analyzed.
- Calculate means of energy ratios for the present 10-second speech segment and then store all average energy parameters for further statistical analyses.

Feature Classification

The Support Vector Machine (SVM), which outperforms most other classification systems in a wide variety of applications, has been used in this study for performance validation was discovered by (Yingthawornsuk and Thanawattano (2010). It achieves relatively robust pattern recognition performance using well established concepts in optimization theory. SVM separates an input $x_i \in R^d$ to two classes. A decision function of SVM separates two classes by $f(x) > 0$ or $f(x) < 0$. The training data which is used in training phase is $\{x_i, y_i\}$, for $i = 1,\ldots, l$ where $x_i \in R^d$ is the input pattern or the ith sample and $y_i \in \{-1, +1\}$ is the class label. Support Vector Classifier maps x_i into some new space of higher dimensionality which depends on a nonlinear function $\phi(x)$ and looks for a hyperplane in that new space. The separating hyperplane is optimized by maximization of the margin. Therefore, SVM can be solved as the following quadratic programming problem,

$$\max_{\alpha_i}\{\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum\sum\alpha_i\alpha_j\, y_i y_j K(x_i x_j)\}$$

Eq. (1)

$$\text{Subject to } 0 \le \alpha_i \le C \text{ and } \sum_{i=1}^{l}\alpha_i\, y_i \quad \text{Eq. (1)}$$

Where C is a parameter to be chosen by user, a larger C corresponding to assigning a higher penalty to errors, and
$\alpha \ge 0$ are Lagrange multipliers.
When the optimization problem has solved, system provides many a ≥ 0 which are the required support vector. Note that the Kernel function $K(x_i x_j) = \phi^{\wedge}{}_T(x_i)\phi(x_j)$ where $\phi(\cdot)$ is a nonlinear operator mapping input vector $x_i \in R^d$ to a higher dimensional space. In addition, other kernels can also be applied. Classification consists of two steps: training and testing. In the training phase, SVM receives some

feature patterns as input. These patterns are the extracted speech features represented by N feature parameters that can be seen as points in N-dimensional space. In this study, eight features comprised of four means and four SD's were formed as input feature matrix which is multi-dimensional. Therefore, the classifying machine becomes able to find the labels of new vectors by comparing them with those used in the training phase. In training state, the 50% of all feature samples was randomly selected as the input feature model to SVM. Then, the rest 50% of randomized samples was used in validating state. For every feature model, the cross validations have been completed for approximately 100 times. Each feature model as input feature to SVM was formed by adding one more ranked feature every time when new feature model is validated. By using the same validating procedure, the performance of validation on same feature models by linear classifier was determined as well. The tendencies of performance evaluated from both SVM and linear classifiers are provided in next section.

RESULTS AND DISCUSSION

Estimated PSD can evidently represent two distinguishing sections in speech signal: the energy concentrated voiced speech segment and the flat less energy unvoiced speech segment. The separating boundaries between voiced and unvoiced segments in speech signal can be obviously observed at the shifts in PSD value, reflecting as the end-points of voiced and unvoiced segments in speech signal.

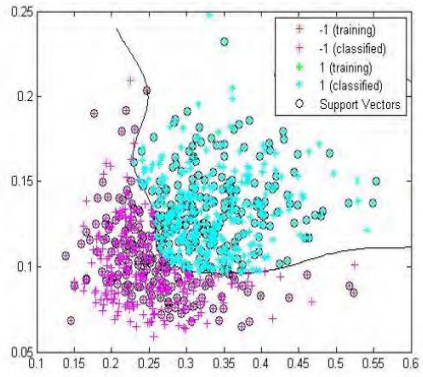

Fig. 3 Identified speech samples between classes separated by decision boundary drawn by SVM

In addition, the class discriminant distance was determined on the basis of Fisher's discriminant function, which helps project the dimensional feature set onto a single feature dimension. When the dimension of PSD feature model is increased, the distance between discriminant scores from two classes is, consequently, longer. Fig. 3 indicates the decision boundary drawn on two classes of the identified samples via SVM classifier.

CONCLUSION

The characterization of sub-band power spectrum density of speech spectrum for classifying male depressed patients and remitted subjects is proposed in this paper. Both analysis and experimental results showed that the studied entropies efficiently achieved in class separation with high correct classification percentages via SVM. More dynamic frequency sub-bands associated with the formant trajectory have to be reassigned for better performance of classification between diagnostic speech classes.

REFERENCES

Cummins, Epps, Breakspear and Goecke (2007) "An Investigation of Depressed Speech Detection: Features and Normalization" School of Elec. Eng. and Telecomm, the University of New South Wales, Sydney, Australia.

France, Shiavi, Silverman, and Wilkes (2000), " Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk*" IEEE transactions on biomedical engineering*, vol. 47, NO. 7.

Greenwood and Kinghorn, "Automatic Silence /Unvoiced/Voiced Classification Of Speech" University of Sheffield Regent Court, 211Portobello St, Sheffield S14DP, UK

Keskinpala, Yingthawornsuk, Wilkes, Shiavi, and Salomon, (2007)"Screening For High Risk Suicidal States Using Mel-Cepstral Coefficients And Energy In Frequency Bands" *15th European Signal Processing Conference, Poznan, Poland*, September 3-7.

Ozdas, Shiavi, Silverman, Wilkes, (2004) "Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk" *IEEE transactions on biomedical engineering*, vol. 51, no. 9.

Ozdas, Shiavi, Wilkes and Silverman (2004) "Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment" Vanderbilt University, Nashville, TN, USA,

Yingthawornsuk and Thanawattano. (2010) " Characterizing Sub- band Spectral Entropy Based Acoustics as Assessment of Vocal correlate of Depression " *International Conference on Control, automation and Systems*.

Yingthawornsuk and Shiavi (2008) "Distinguishing Depression and Suicidal Risk in Men Using GMM Based Frequency Contents of Affective Vocal Tract Response*" International Conference on Control, Automation and Systems.*

Y. Qi, and Hunt(1993), "Voiced-Unvoiced-Silence Classifications of Speech using Hybrid Features and a Network Classifier," *IEEE Trans. Speech Audio Processing*, vol. 1 No. 2, pp. 250-255.