Classifying Depression in Categorized Emotional Speech Samples using MFCC and Glottal Slope Tilt

Thaweesak Yingthawornsuk

***Abstract*—** The comparative study of acoustical properties in speech sample as emotional indicator based on spectral characteristics of speech signal have formerly been studied and reported for its quantitative information in association with the emotional state in persons suffering depression. This state can affect the speech production system of speaker, which modulates in spoken sound. The MFCC has been popularly reported for its characteristic change corresponding to severity of depression. The sixteenth MFCCS and Glottal Slope Tilt (GTLT) from three different speaker groups which are remitted, depressed and suicidal were extracted, statistically tested and classified by several classifiers. The best classification score can be obtained at 76% accuracy based on the combination between MFCC and GTLT samples in testing phase. Results show the dominant property of the focused speech features in separation between suicidal and recovering speakers from depression.

***Keywords*—** Classification, Speech, MFCC, Glottal Slope, Depression

I. Introduction

Now a day, our world has an increasing rate on population growth and it is climbing up every year. But natural resources inversely decrease due to our daily basis consumption. This situation can affect our living in term of healthiness. In some serious situation it can make our risk of life increased without our notification. Such risk has been known clinically as depression, or in severe case suicidality could happen. We can find many reports about the suicidal risk in person and suicide is popularly the public health problem associated with high population. In addition, so does the increasing rate of hotline call-in, which is simultaneously monitored by psychiatrist to evaluate those callers who may or may not depressed. How accurate the diagnosis made by physician could impact on their family and lethal risk to that caller if he/she is planning to commit suicide.

Apparently, if psychiatrists can diagnose about symptom for depression or suicidal risk correctly, it can help patients who have agonized due to emotional illness, depression or suicidal risk in time and get a proper treatment right away from the beginning. The formerly experimental studies [1-3, 5-10] have been proposed that the acoustical parameters estimated from speech signal can be used in association with the affection on recognizing pattern and assessment of the degree of mental severity in depressive speakers. The most common methods to assess, if patients were at severe state of depression or even at elevated risk of suicide, are self-scored patient survey, report by other, clinical interviews and rating scales [4]. Diagnosis and decision making on clinical categories patients belong to are clinical procedure with time consuming in which practitioners have to get involved in several steps such as information gathering, background profile checking, hospital admission and visiting records, diagnosing with simultaneous response in judging if patient were psychologically safe from suicidal risk or clinically identified for one of symptom categories, dramatically necessitates for physician to conclude the diagnosing result with the correct decision making on admission and treatment for that patient. As reported in the published studies, several analytical techniques have been proposed for achievement of measuring the particular changes, as a result of affection from the underlying symptom of depression, in acoustics of speech of depressed patients. It has been concluded that the suicidal speech in severely depressed speaker is very similar to that of common depressive one, but the tonal quality of speech significantly changes when the symptom of near-term suicidal risk highly strikes at the moment.

The sections in paper are organized and detailed as follows. Section II describes on method, database, feature extraction, PCA and classification. Section III deals with experimental result and discussion. Section IV provides conclusion and future research direction at the end.

II. Methodology

A. Speech Database

The database consists of speech samples recorded from interviewing session with psychiatrist. It is categorized into three groups of 10 remitted, depressed and high-risk suicidal female subjects. The pre-processing is carried out by first digitizing all speech signals through a 16-bit analog to digital converter at a sampling rate 10 KHz via a 5 KHz anti-aliasing low-pass filer. Prior to detection of voiced, unvoiced, silent

segment in speech files, the monitor and screening on any sound artifact possibly appeared during interviewing are offline implemented by using the Goldwave, including the silences longer than 0.5 seconds are manually removed. All speech signals of remitted, depressed and high-risk suicidal speakers are carefully processed under the same condition of pre-processing and the similar acoustical environment control is made during the period of recording speech sample in interviewing conversation.

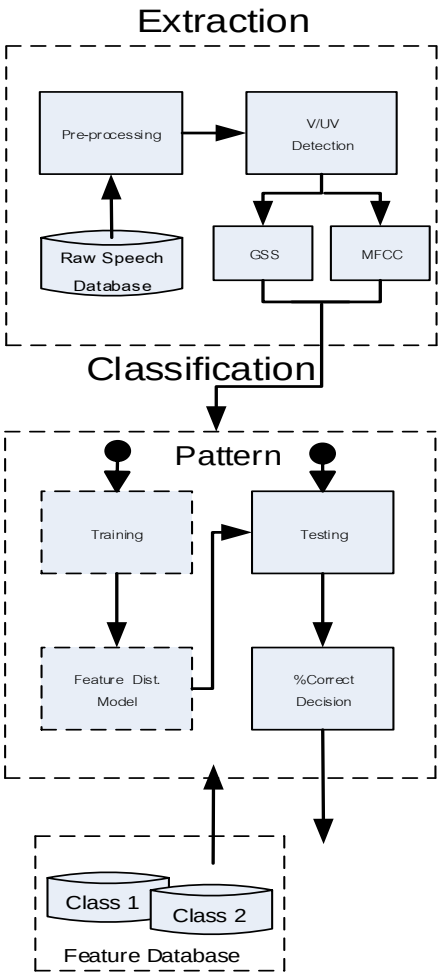

Fig. 1 Research study procedure

B. Speech Segmentation

Based on the exploiting fact that the unvoiced segments of speech signal are very high frequency component compared to the voiced speech which is low frequency and quasi-periodic. To classify which segments of speech signal based on their energy and then weighted using the Dyadic Wavelet Transform (DWT) of speech samples were computed in each segment of 256 samples/frame. The unvoiced speech segments can be readily detected by comparing the energies of DWTs at the lowest scale $\delta_1 = 2^1$ and the highest energy level is $\delta_5 = 2^5$. Any segment of speech signal with its largest energy level estimated at scale $\delta_1 = 2^1$ is favorably classified as an unvoiced segment, otherwise found voiced segments. The following equation is the energy threshold defined as unvoiced segment;

$$UV = (n|\delta_i = 2^1) \quad ; \quad n = 1, \ldots\ldots\ldots, N \tag{1}$$

Where *uv* is speech segment classified as unvoiced at which the *n* segment with energy at scale δ_1 maximized.

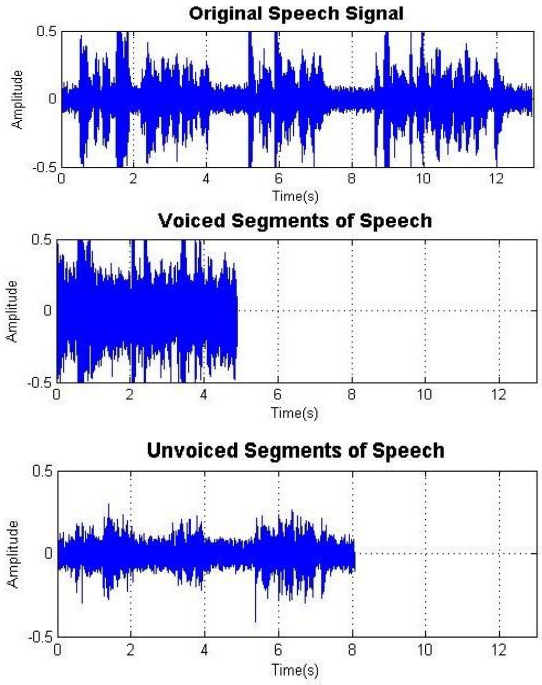

Fig. 2 Original speech signal (upper), voiced segment of speech (middle) and unvoiced segment (lower)

C. MFCC Extraction

Voiced segments of all speech signals in database are processed for Mel-Scale Frequency Cepstral Coefficients (MFCC) [7-8,10]. The estimation procedure of studied energy parameter is described.

- Windowing each concatenated voiced-segment into 25.6 ms-length frames
- Computing the logarithm of the discrete Fourier Transform (DFT) for all windowed frames of voiced speech
- Applying the log-magnitude spectrum through the 16 triangular bandpass filter bank with center frequencies corresponding to Mel-frequency scale
- Computing the inverse discrete Fourier Transform (IDFT), then calculate the 16-order cepstral coefficients
- Analyzing all extracted MFCC dataset with two dimensional PCA and then classifying with LS and RBF classifiers

The purpose of Mel-frequency scale is to map between linear to logarithmic scale for frequencies of speech signal

higher than 1 kHz. The characteristics of spectral frequency will correspond to human auditory perception. The Mel-scale frequency mapping is defined [11]:

$$f_{mel} = 2595 * LOG_{10}\left[1 + \frac{f_{lin}}{700}\right] \quad (2)$$

in which f_{mel} is the perceived frequency and f_{lin} is the real linear frequency in speech signal.

In filtering phase, a series of the 16 triangular bandpass filters, N_s = 16 is used for a filter bank whose center frequencies and bandwidths are selected according to the Mel-scale. Once the center frequencies and bandwidths of the filter are obtained, the log-energy output of each filter *i is* computed and encoded to the MFCC by performing a Discrete Cosine Transform (DCT) defined as follow:

$$C_n = \frac{2}{N'}\sum_{i=1}^{N_f} x_k cos\left(k_i \frac{2\pi}{N'} n\right) \quad ;1 \leq n \leq p \quad (3)$$

Regarding less complexity, the factor $\frac{2}{N'}$ in equation 3 is discarded from algorithm computation.

Delta-MFCC features were reported in [12] to gain more information on dynamic characteristics to the static MFCC features. They improve accuracy by adding a characterization of temporal dependencies to pattern classification, which is nominally assumed to be statistically independent of one another. For a short-time MFCC coefficient C_n, the delta-MFCC features are typically defined as

$$D_n = C_{n+m} - C_{n-m} \quad (4)$$

where n is the index of the analysis frames and in practice m is approximately number of 2.

D.Principal Component Analysis

The PCA technique has been applied to MFCC features to extract the most significant components of feature. This technique helps reduce multi-dimension of dataset down to two dimensions which is adequate for training and testing phases in classification.

E. Pairwise Classification

Several classifiers such as Least Squares (LS), LMS and RBF are selected to train and test on MFCC and GTLT datasets and compared among three different speech sample groups for performances of individual classification. In this study three groups of extracted MFCC and GTLT samples are arranged into pairwise manners which are RMT/DPR, RMT/SUI and DPR/SUI. First, MFCC and GTLT samples are randomly selected for 20% from sample dataset, and then used to test classifier, and other 35%, 50% from same dataset for testing the same classifier. The reason of doing these is to compare the performances of classification among categorized subject groups, which might be affected from sizes of sample. Several trials on random selection of samples for training and testing approximately hundred times are further proceeded to find the average performance of classification.

III. EXPERIMENTAL RESULTS AND DISCUSSION

Original speech waveform, voiced and unvoiced segments of speech signal are plotted in Figure (2). The difference in amplitude and time interval can be obviously notified between voiced and unvoiced segments. Averaged errors in classification are tabulated in categorized pairwise groups versus types of classifier listed in Table 1-6 for case of 20%, 35%, 50% of testing sample, summarized averages from RMT/SUI training as best pairwise with least error, 20% testing, and summarized averages from RMT/SUI testing.

The comparative errors obtained from several trials on selections of the focused MFCC and GTLT samples in classification are graphically depicted in Figures 3 and 4 for cases of training and testing with LS and RBF classifiers. As seen in box-and-whisker diagrams, extracted samples represented as class 2 for suicidal group provided very less error of classification approximately 0.15 for all 20%, 35% and 50% of training samples and as well for both LS and RBF classifiers. The greater errors can be seen for class1 represented for remitted group in training and testing for all classifiers and percentages of sampling approximately 0.35. More notification can be made on similar results of classification between two classifiers with different sampling percentages.

Based on the first four lower-order MFCC in conjunction with GTLT, the fairly high correct classifying scores can be obtained in this study, which are likely productive for its class discriminative property beneficial to emotional disorder assessment. More various acoustical parameters are suggested into the same account with studied speech feature for more accurate classification and improvement of research result toward same golden goals committed to research work.

TABLE I
SUMMARIZED ERRORS OF CLASSIFICATION BETWEEN DEPRESSED AND REMITTED TESTING

Classification	**Percent of sample in testing overall classifier**		
	20%	**35%**	**50%**
LS	0.353	**0.341**	0.350
RBF	0.396	0.408	0.408

TABLE II
SUMMARIZED ERRORS OF CLASSIFICATION BETWEEN DEPRESSED AND REMITTED TRAINING

Classification	**Percent of sample in training overall classifier**

	80%	65%	50%
LS	0.079	**0.042**	0.063
RBF	0.118	0.104	0.196

TABLE III
SUMMARIZED ERRORS OF CLASSIFICATION BETWEEN DEPRESSED AND HIGH-RISK SUICIDAL TESTING

Classification	Percent of sample in testing overall classifier		
	20%	35%	50%
LS	**0.335**	0.358	0.355
RBF	0.368	0.405	0.533

TABLE IV
SUMMARIZED ERRORS OF CLASSIFICATION BETWEEN DEPRESSED AND HIGH-RISK SUICIDAL TRAINING

Classification	Percent of sample in training overall classifier		
	80%	65%	50%
LS	0.062	**0.052**	0.075
RBF	0.090	0.082	0.104

TABLE V
SUMMARIZED ERRORS OF CLASSIFICATION BETWEEN REMITTED AND HIGH-RISK SUICIDAL TESTING

Classification	Percent of sample in testing overall classifier		
	20%	35%	50%
LS	0.466	0.468	**0.451**
RBF	0.499	0.517	0.501

TABLE VI
SUMMARIZED ERRORS OF CLASSIFICATION BETWEEN REMITTED AND HIGH-RISK SUICIDAL TRAINING

Classification	Percent of sample in training overall classifier		
	80%	65%	50%
LS	0.087	0.124	0.103
RBF	0.083	**0.062**	0.101

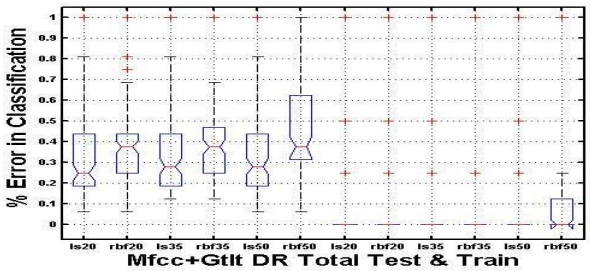

Fig. 3 Comparison of Box plots between LS and RBF classification overall with 20%, 35% and 50% of samples in depressed and remitted

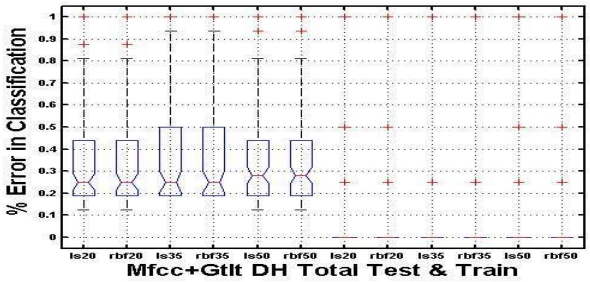

Fig. 4 Comparison of Box plots between LS and RBF classification overall with 20%, 35% and 50% of samples in depressed and suicidal

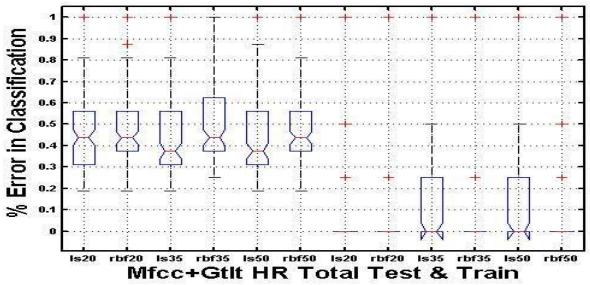

Fig. 5 Comparison of Box plots between LS and RBF classification overall with 20%, 35% and 50% of samples in suicidal and remitted

IV. CONCLUSION

Experimental results show the property of both MFCC and GTLT able to indicate speaker's psychiatric state, especially in class separation between remitted and suicidal speaker groups. Different sampling percentages investigated in this study can affect slightly the classification score in some classifiers selected to evaluate vocal samples. Further direction will focus on much more effective acoustics that can be assistive to currently studied speech parameters in class separation with highly significantly statistical difference, and larger size of speech sample database required.

REFERENCES

[1] T.Yingthawornsuk, "Comparative Study on Vocal Cepstral Emission of Clinical Depressed & Normal Speaker", Int'L Conf. On Control Automation & Systems, Korea, Oct. 26 -29, 2011.

[2] T.Yingthawornsuk & et. al, "Comparative Study of Pairwise Classification by ML & NN on Unvoiced Segments in Speech Sample", Int'L Conf. On System & Electronic Engineering (ICSEE' 2012), Phuket, Thailand, Dec. 18 -19, 2012.

[3] T.Yingthawornsuk, "Classification of Depressed Speakers Based on MFCC in Speech Sample", Int'L Conf. On Advances in Electrical & Electronics Engineering, Pattaya, Thailand, April 13 – 15, 2012.

[4] M. Hamilton, "A rating scale for depression", Journal of Neurology, Neurosurgery and Psychiatry, Vol. 23, pp. 56-62, 1960. https://doi.org/10.1136/jnnp.23.1.56

[5] France, D.J., et al., "Acoustical properties of speech as indicators of depression and suicide", IEEE transactions on BME, 2000. 47:p 829-837. https://doi.org/10.1109/10.846676

[6] F. Tolkmitt, H. Helfrich, R. Standke, K.R. Scherer,"Vocal Indicators of Psychiatric Treatment Effects in Depressives and Schizophrenics", J.Communication Disorders, Vol.15, pp.209-222, 1982. https://doi.org/10.1016/0021-9924(82)90034-X

[7] Godino-Llorente J.I., Gomez-Vilda P., and Blanco-Velasco M., "Dimensionality Reduction of a pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short Term Cepstral Parameters", IEEE Transaction on Biomedical Engineering, 53(10):1943-1953, 2006. https://doi.org/10.1109/TBME.2006.871883

[8] Lu-Shih Alex Low, et al., " Content Based Clinical Depression Detection in Adolescents", 17th EUSIPCO 2009, Scotland Aug. 24-28, 2009.

[9] T. Yingthawornsuk, R.G. Shiavi, "Distinguishing Depression and Suicidal Risk in Men Using GMM Based Frequency Contents of Affective Vocal Tract Response", International Conference on Control, Automation and System 2008, Seoul, Korea, 2008.

[10] Ozdas, A., Shiavi, R.G., Wilkes, D.M., Silverman, M., Silverman, S.,"Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment", Meth. Info. Med., vol. 43, pp 36-38, 2004. https://doi.org/10.1055/s-0038-1633420

[11] Koeing, W., "A new frequency scale for acoustic measurements", Bell Telephone Laboratory Record", Vol. 27, pp. 299-301, 1949.

[12] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," Proc. ICASSP, 1986. https://doi.org/10.1109/TASSP.1986.1164788