Acoustical Properties of Spectral Tilt & Cepstral Coeff^{n} in Speech for Classifying Depressed Speakers

Thaweewong Akkaralaertsest, and Thaweesak Yingthawornsuk

Abstract— This paper has described the study on estimation of Spectral tilt, known as Glottal Spectral Slope (GSS) in other research reports, and Male-Frequency Cepstral Coefficients (MFCC) in voiced segment of speech for the possibility of identifying the degree levels of the severe depression in speakers clinically diagnosed as Depressed in studying categories with two other groups of Remitted (recovering speakers) and High-risk suicidal. A speaker with severely depressed tends to be possibly at risk of committing suicide when the symptom strikes, unless that person is properly assigned by hospital admission and treatment in time by the experienced therapist. In this work the combination of both source and filter responses as two main parts in speech production system model has been studied for their property used in classifying between different speaker groups.

The GSS originated in source domain and MFCC in filter domain of production system of three categorized speech database were extracted, statistically tested and used in pairwise classification based on empirically selected LS and RBF techniques. The minimum error of classification was found to be 0.335 based on LS classifier with 20% of combined features in testing phase and the corresponding 80% obtained for 0.062 by the same classifier when classifying between depressed and suicidal speech groups. The LS performed slightly better than the RBF for all pairwise cases. Based on numbers of feature combination compared to former study of individual feature in classification it suggests for the higher performance of classifiers comes with more diversity in a combination of high dimensional features.

Keywords— Speech, MFCC, Glottal Spectral Slope, Depression

I. Introduction

SUICIDE is a major public health problem. The number of people who died because of suicide was climbing up every days and suicide was one of the leading cause of death in United States and other countries. As can be seen from the statistics, suicide remains a frequent but preventable cause of death in United States. Therefore, it is very important to evaluate a patient's risk of committing suicide. When the patient is seen by psychiatrists, the psychiatrists evaluate the patient's risk of committing suicide as a part of the clinical interview. Researchers and psychiatrists also assess mood by different techniques, such as the Hamilton depression rating scale. [4] However it is also very important to evaluate risk of committing suicide, when a person is seen by care takers who are not psychiatrists. People who are suicidal often come to a physician's office or emergency room because of another illness. Our work, described in this paper, could be a very useful in helping these physicians to evaluate the risk of suicide. It is widely known that psychological state affects the human speech production system. For this reason the vocal characteristics of speech have been recognized as potential indicators for the assessment of suicide risk. The former researchers proposed using vocal parameters of speech for deciding if a patient is suicidal or not [5]. They describe suicidal speech as being similar to depressed speech but when the patient becomes at high risk of suicide, he/she exhibits significant changes in the tonal quality of the speech. Several researchers have studied vocal tract characteristics related to depression and suicidal risk. France et al. used long term averages of the extracted formant information and compared them among patients to distinguish near term suicidal groups from depressed and control groups [4]. Yingthawornsuk et al. used the percentages of the total power, the highest peak value and its frequency location at which the percentages of the total power (PSD_1, PSD_2 and PSD_3) found to be the primary features effectively distinguish between groups of individuals carrying diagnoses of suicide risk, depression and remission [1]. Ozdas et al. proposed using lower order Mel-Cepstral Coefficients and Glottal Spectral Slope among suicidal patients, major depressed patients and non-suicidal patients [10]. They used Gaussian mixture models and unimodal Gaussian models for depressed/suicidal, control/suicidal, and control/depressed pairwise classification and compared classification performance. The work presented in this paper is a follow-up to work described in [10].

This study is different from the previous study due to greater control of the recording environment. In the previous study, the database included recordings of suicide notes left on tapes and interviews of patients who attempted suicide and failed, therefore the recording environments were different for each patient [10]. Thus environmental compensation was necessary and applied by cepstral mean normalization to

Thaweewong Akkaralaertsest is with Division of Electronics & Telecommunication Engineering, Faculty of Engineering, Rajamangala University of Technology Krungthep, Thailand.

Thaweesak Yingthawornsuk is with Department of Media Technology, Faculty of Industrial Education and Technology , King Mongkut's University of Technology Thonburi, Bangkok, Thailand.

http://dx.doi.org/10.15242/IIE.E0814548

compensate the spectral variability introduced by possible differences in recording environments. In this study, audio recordings were made during clinical interviews in the same environment for all patients. It is very difficult to get this type of data, since it is recorded in a specific environment from patients who do not always exhibit the high risk suicide state. The recording must be made when the patient is in such state of mind. The other difference between the previous study [10] and this study is the size of the patient database. The number of the patients was increased in this study over the previous one. Additionally, each patient provided two kinds of speech samples recorded in the same environment: an interview session and a reading session. In this work only database from interview session was studied. As reported in [6] the emotional arousal produces changes in the speech production scheme by affecting the respiratory, phonatory, and articulatory processes that in turn are mediated in the acoustic signal. This affective speech carrying emotional disturbances naturally has vocal characteristics associated with measurable changes which are able to be extracted by approaches of speech processing by utilizing such features as prosody (pitch, energy, speaking rate), spectral characteristics (formants, power spectral density) of the acoustic speech signal. All following sections are organized and their details are provided in individual sections. 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

All speech recordings were intentionally collected from three different categorized groups of depressed, high-risk suicidal and remitted subjects. Database consists of thirty females evenly grouped into clinical categories. Each subject in each group has two different types of speech samples recorded. One is collected from main interviewing session with psychiatrist and another from the post session that patient reads a predetermined part of book. The passage used in post session is composed of the standardized texts generally used in speech science since it contains all of normal sounds in spoken English and it is phonetically balanced. Only interview speech is studied in this work for further investigation on discriminative property of focused acoustical features, MFCC and GSS. In prior processing state before classification, the raw speech samples of individual subject was randomly extracted from our database and then used to represent that subject. All speech samples are off-line analyzed and processed throughout the entire analysis procedure. First each speech signal was digitized via a 16-bit analog-to-digital (A/D) converter at a 10-kHz sampling rate with an anti-aliasing filter (i.e., 5-kHz low-pass). The background noise and voice artifact not belonged to patient are removed via manually monitoring with an audio editor software.

B. Speech Segmentation

C. 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;

$$UV = (n|\delta_i = 2^1) \quad ; \quad n = 1, \dots \dots, N \qquad (1)$$

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

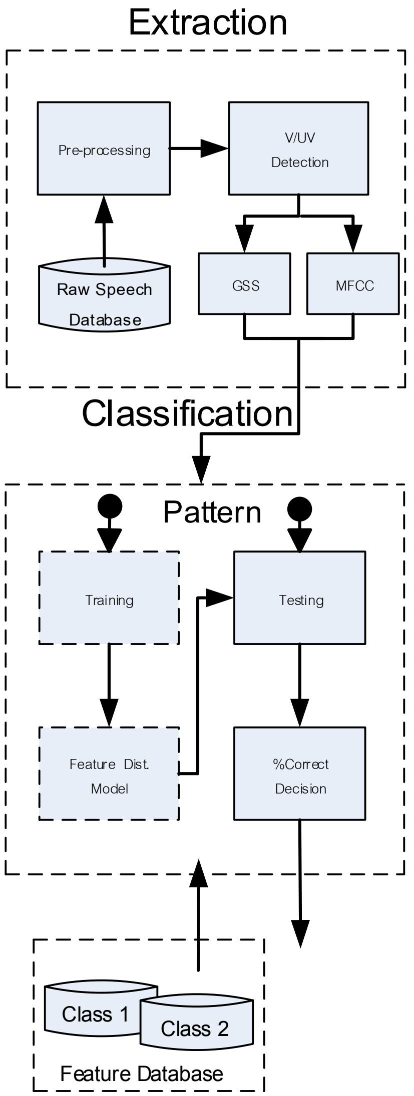

Fig. 1 Speech processing and classification procedure

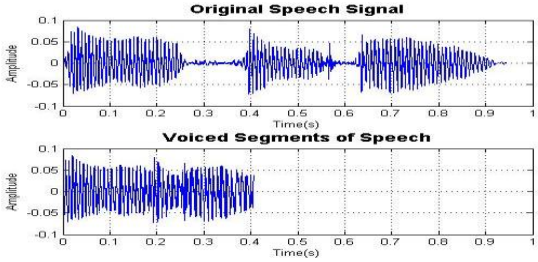

Fig. 2 Original speech signal (upper) and voiced segment of speech (lower)

C. *Feature Extraction*

Voiced segments of all speech signals in database are processed for Mel-Scale Frequency Cepstral Coefficients (MFCC) based on the same technique reported in former studies [7-8,10-12]. The estimation procedure of studied features can be described as follows: Windowing speech signal into smaller 25.6 ms-length segments, estimating the logarithm of the Discrete Fourier Transform (DFT) for all segments, computing the IDFT of the log-magnitude spectrum filtered out from a 16-triangular BPF bank with center frequencies at Mel-frequency scale frequency response, applying two dimensional PCA analysis on all obtained MFCC parameters, and finally classifying with LS and RBF classifiers. In additional to MFCC parameter, based on the model the vocal properties of near-term suicidal patients are separately investigated in Source and Filter domains of speech production mechanism. The variations reflected on the source domain properties were investigated by the excitation-based speech feature, Glottal Spectral Slope (GSS) and those reflected on the filter domain properties, Mel-Cepstral Filter Bank Coefficients (MFCC) as mentioned above how to estimate it. This proposed work has been strictly focused on the Glottal Flow Slope and MFCC for one-on-one parameter domain analysis in term of evenly less bias multi-integrated parameter input to multivariate classifier.

D. *Principal Component Analysis*

The PCA technique has been applied to both combined MFCC and GSS as a set of features in order to analyze for the most significant components of it. 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 LS and RBF are selected to train and test on two dimensional combined MFCC and GSS dataset and make comparison among three different subject groups for determining the performances of individual classification. In this study three groups of combined feature samples are arranged into pairwise manners which are RMT/DPR, RMT/SUI and DPR/SUI. First, feature samples are randomly selected for 20% from sample dataset, and then used to train classifier, and other 35%, 50% from same dataset for training the same classifier. The investigation on sample size has been carried out in that its affection may have impact on classification. Several trials on random selection of samples for training and testing approximately hundred times are further proceeded to find the average classification performance.

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 3-5 for all cases of various selected training samples, summarized averages from DPR/SUI classification found to have the least pairwise error, when the 80% of randomizedly selected samples used in classification with Least Squares (LS) classifier.

The comparative errors obtained from several trials on selections of combined sample in classification are graphically depicted in Figures for cases of training and testing with LS vs and RBF classifiers. As seen in box-and-whisker diagrams, sampled MFCC+GSS represented in Figures 3-11for any groups vs depressed group provided very less classification for all 20%, 35% and 50% of training and testing samples for mostly LS classifier case. The slightly greater errors can be seen as well for mostly RBF classification for all cases of randomizedly selected percentages of combined feature sample.

Based on our focused combination of the extracted acoustical parameters, the fairly high correct classifying scores can be obtained 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 source and filter-domain parameters to gain more accurate classification and improvement of research result toward the goals of research work.

TABLE 1
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 2
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 3
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 4
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 5
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 6
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

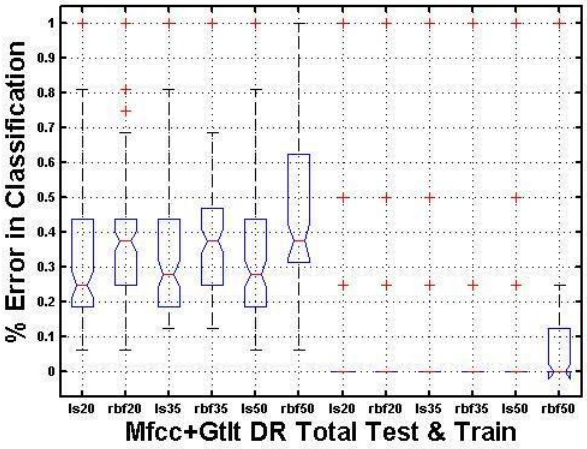

Fig. 3 Comparison of Box plots between LS and RBF classification overall with 20%, 35% and 50% of samples in depressed and remitted

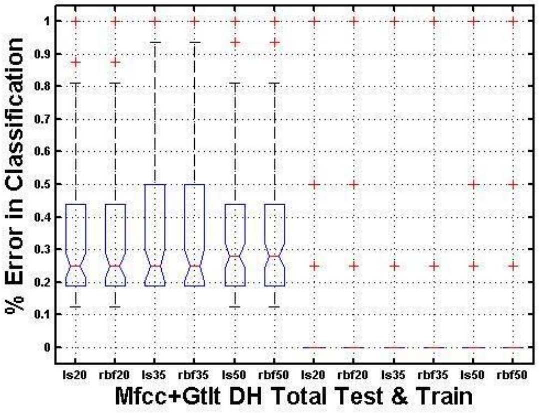

Fig. 4 Comparison of Box plots between LS and RBF classification overall with 20%, 35% and 50% of samples in depressed and suicidal

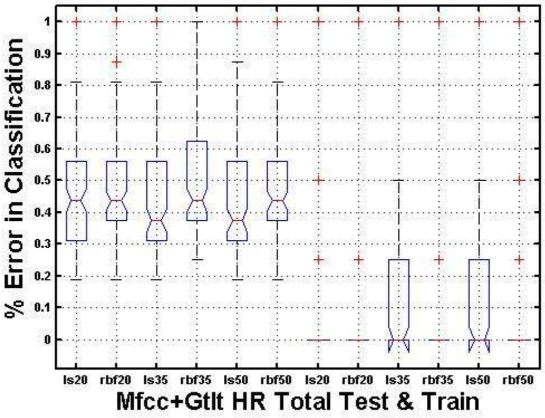

Fig. 5 Comparison of Box plots between LS and RBF classification overall with 20%, 35% and 50% of samples in suicidal and remitted

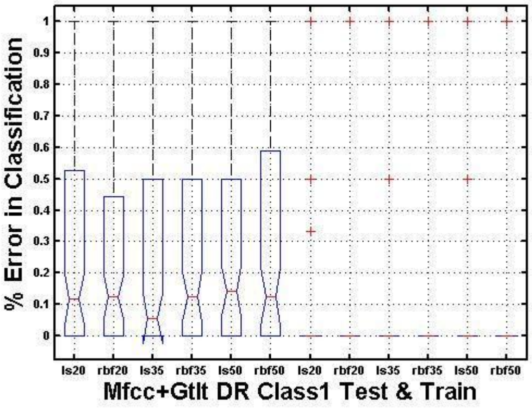

Fig. 6 Comparison of Box plots between LS and RBF classification on class 1 with 20%, 35% and 50% of samples between depressed and remitted

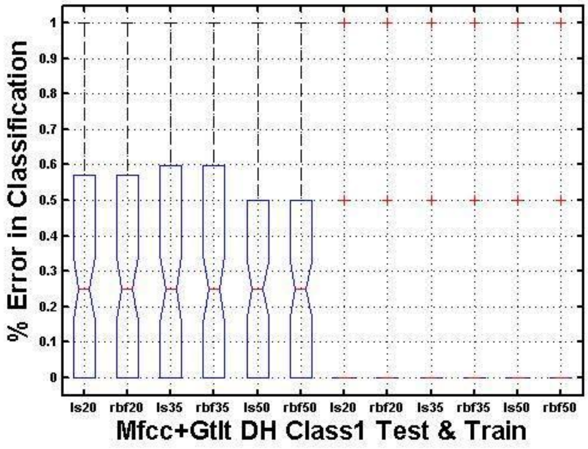

Fig. 7 Comparison of Box plots between LS and RBF classification on class 1 with 20%, 35% and 50% of samples in depressed and suicidal

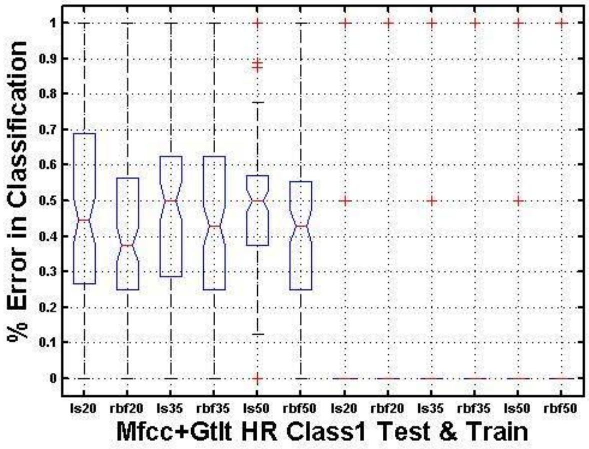

Fig. 8 Comparison of Box plots between LS and RBF classification on class 1 with 20%, 35% and 50% of samples in suicidal and remitted

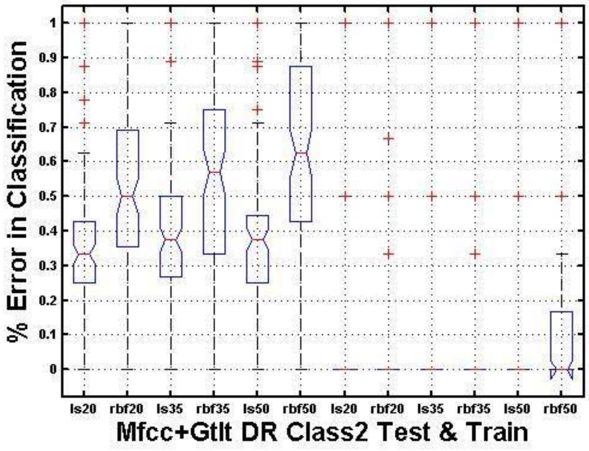

Fig. 9 Comparison of Box plots between LS and RBF classification on class 2 with 20%, 35% and 50% of samples in depressed and remitted

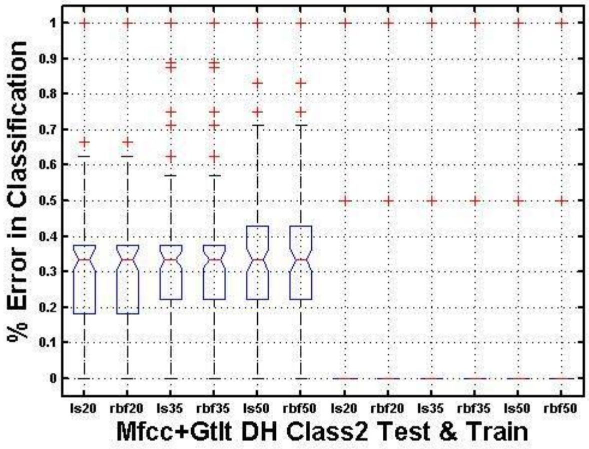

Fig. 10 Comparison of Box plots between LS and RBF classification on class 2 with 20%, 35% and 50% of samples in depressed and suicidal

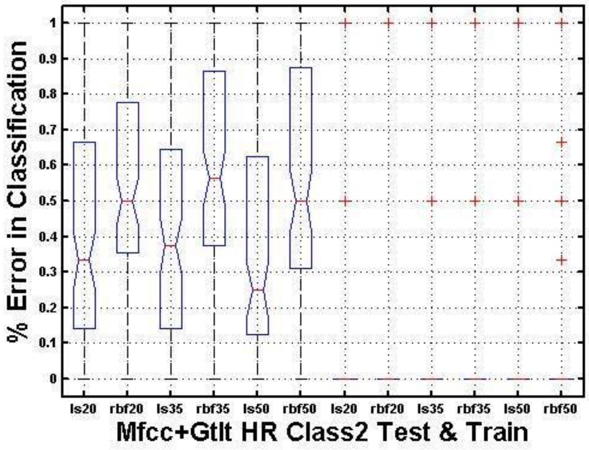

Fig. 11 Comparison of Box plots between LS and RBF classification on class 2 with 20%, 35% and 50% of samples in suicidal and remitted

IV. Conclusion

Results from study reveal that the discriminative property of the focused acoustical parameters are able to assess the speaker's psychiatric state as class separation based on what we found from this work among categorized speaker groups. Different sampling percentages also investigated in this study surely affect the classification scores in some classifiers selected to evaluate the vocal samples, but not dramatically. Further direction will focus on other more effective acoustics able to be assistive to presently proposed study with the highly significantly statistical difference and larger size of database required

V. Acknowledgement

This work was financially granted by National Research Council of Thailand.

References

[1] T.Yingthawornsuk, “Comparative Study on Vocal Cepstral Emission of Clinical Depressed & Normal Speaker”, ICCAS’11, Korea.

[2] T.Yingthawornsuk & et al., “Comparative Study of Pairwise Classification by ML & NN on Unvoiced Segments in Speech Sample”, (ICSEE’12), Phuket, Thailand, 2012.

[3] T.Yingthawornsuk, “Classification of Depressed Speakers Based on MFCC in Speech Sample”, (ICEEE’12), Thailand, 2012.

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

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

[6] F. Tolkmitt, et al., "Vocal Indicators of Psychiatric Treatment Effects in Depressives and Schizophrenics", J. Comm. Disorders, Vol.15, pp.209-222, 1982. http://dx.doi.org/10.1016/0021-9924(82)90034-X

[7] Godino-Llorente J.I. & et al.,“Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short Term Cepstral Parameters”, IEEE Trans. on BME, 53(10):1943-1953, 2006. http://dx.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, 2009.

[9] T. Yingthawornsuk, R.G. Shiavi, “Distinguishing Depression and Suicidal Risk in Men Using GMM Based Frequency Contents of Affective Vocal Tract Response”, ICCAS 2008, Korea, 2008.

[10] Ozdas, A.& et al.,"Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment", Meth. Info. Med., vol. 43, pp 36-38, 2004.

[11] Koeing, W., "A new frequency scale for acoustic measurements", Bell Telephone Laboratory Record", Vol. 27, pp. 299-301, 1949.
S. Furui, “Speaker-independent isolated word recognition based on emphasized spectral dynamics,” Proc. ICASSP, 1986.