Comparative Analysis of Vocal Characteristics in Speakers with Depression and High-Risk Suicide

Thaweewong Akkaralaertsest and Thaweesak Yingthawornsuk

***Abstract*—Evaluation of speakers who are high-risk suicidal compared to those with less clinical depression are critical when the syndrome underlying a patient's abnormal behaviour is diagnosed without expertise. This study describes a way to classify the speech samples collected from groups of depressive and suicidal speakers by employing the speech processing technique in data analysis. First, the Glottal Spectral Slope (GSS) and Mel-Frequency Cepstral Coefficients (MFCC) were computationally estimated from the voiced segments detected from the categorized speech sample database. Second, the pairwise classification was then made on the combination of those extracted vocal features respectively corresponding to the frequency response of the source and the filter in speech production system model.**

The procedure of this research was carried out in order to investigate the discriminative property of the focused vocal parameters mainly between depressed speakers and high-risk suicidal speaker groups. The result revealed that MFCC and GSS parameters are slightly high effective in term of vocal indicator corresponding to severe depression with fairly high performance in between-group separation.

***Index Terms*—Depression, glottal spectral slope, MFCC, speech.**

I. Introduction

Suicide is a major public health problem in mostly every society. The number of people who died by committing suicide is increasing up every year. This kind of tragedy has been reported to be among the leading cause of death with growing dead toll rate. As acknowledged from the statistics reported publically, suicide remains frequent but preventable cause of death with in-time recue or earlier diagnosis and admission into the care-taking program in hospitals before the lethal risk of suicide will elevate. Therefore prevention is currently a solely way to be made to save life of people from such tragedy. Screening patients who are at risk of committing suicide is very important task and only possibly completed by the psychiatrist with high expertise. In addition, the suicide prevention program is presently limited to clinical level which consumes time and bases heavily on the psychiatrist's experience and judgment. In this work what we found from data processing and speech analysis could lead to be additionally one of supplements to the suicide prevention program. In the past, many research groups have been attempting to figure out the way to identify individual categorized groups of patients with emotional disorders and the various methodologies have been conducted to reach the conclusion. The most popular technique emerging in area of speech processing has been taken in account of research procedure to accomplish the specific tasks such as data processing, information retrieval and feature extraction prior to analysis and interpretation of results.

Severely depressive persons at near-term suicidal risk exhibit the perceptual changes significantly in their vocal qualities that can distinguish them from normal ones [1]. In formerly published papers [2]-[6], the analytical techniques have been designed and developed to determine if the subjects were categorized into any of the following patient groups: Control, Non-suicidal Depressed or High-risk Suicidal. In those studies [1], [7] the vocal cues have been properly used in term of indicator as assistive tool in diagnosing the underlying symptom in patient by experienced clinicians but these skills are not widespread in clinical use. The psychological state in speaker has been widely known that it can affect human speech production system and modulate in the spoken sound which is changed in its acoustical property in term of amount of energy contained in corresponding bandwidths, shifted formant pattern, and variation in fundamental frequency contour [8]. The former researchers proposed to use the vocal parameters extracted from speech database for deciding clinically if a patient is suicidal or not. Likewise, the suicidal speech was described to be very similar to the depressed speech but when the patient becomes at high risk of suicide, he/she exhibits the significant changes in the tonal quality of speech. The characteristics of vocal tract related to depression and suicidal risk have been also investigated and reported to have a high effective performance in identification of patient categories. The long-term averages and variances of formant information were also studied for their separation power among groups of patients reported by France [1]. The percentages of total power, the highest peak value and its frequency location at which the percentages of the total power were found to be the effective features in distinguishing groups of individuals carrying diagnoses of depression, remission and high-risk suicide [2]. In addition, the lower-order coefficients corresponding to the Cepstrum of speech signal and dynamic response in associated with the duration of Glottal waveform were proposed by Ozdas *et al.* [4] to be prominent as input parameters to Maximum Likelihood classification in a form of Gaussian mixture Model approximated from the pdf estimate of the proposed acoustical parameters.

The main objective of work presented in this manuscript is to investigate first the acoustical properties of the Glottal

Manuscript received August 31, 2014; revised October 25, 2014. This work has been supported financially by National Research Council of Thailand.

Thaweewong Akkaralaertsest is with Division of Electronics and Telecommunication Engineering, Faculty of Engineering, Rajamangala University of Technology KrungThep, Bangkok, Thailand (e-mail: thaweewong.a@rmutk.ac.th).

Thaweesak Yingthawornsuk is with Media Technology, King Mongkut's University of Technology Thonburi – Bang Khuntien Campus, Bang Khuntien, Bangkok, 10150 Thailand (e-mail: thaweesak.yin@kmutt.ac.th).

DOI: 10.7763/IJCTE.2015.V7.1001

response represented as the characteristics of source domain affected from severe depression and suicidal risk, and second Cepstral response represented as that of the filter part of human speech production system. All following sections of the manuscript are organized and detailed as follows: Section II provides details in database, method, feature extraction and classification. Section III describes and discusses on the result obtained from study. Section IV summarizes work and suggests for the future direction of this ongoing work.

II. Methodology

A. Speech Database

The speech database was recorded from the categorized groups of depressed, high-risk suicidal and remitted (recovered from being depressed after treatment) speakers. Total speakers included in our database consist of thirty female speakers collected from individual groups of ten females belonging to the clinical categories of symptom in which the individual subject has been clinically diagnosed and judged by the psychiatrist. In recording speech from each volunteer speaker, speech samples were intentionally recorded into two separating sessions to a close-look monitoring and a proper adjustment of the acoustical controlling environment. First recording was made from the main interviewing session with psychiatrist and second recording was made from the post session that volunteer had to read a predetermined part of book. Both recordings were completed on the same day when volunteers visited hospital. Only speech samples of all categorized volunteers recorded from first session were studied in this work for investigating the discriminative property in the studied acoustical parameters. The high possible thought of what on the speaker's mind could be inherently moderated into their spoken speech when that speaker had been communicated verbally with physicist during interview.

In pre-processing state before feature extraction, the raw speech sample of individual volunteer was used to represent individual speaker in analysis and classification. All speech samples were 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 were monitored and manually removed via audio editor software.

B. Speech Separation for Voiced Segments

Based on the theoretical knowledge regarding study area of speech processing that the unvoiced segment of speech signal contains a very high frequency component compared to the voiced segment which is low frequency and quasi-periodic. To detect which segment of speech signal is voiced or unvoiced the energy of each segment was estimated and weighted to the energy scales determined from the calculation of the Dyadic Wavelet Transform (DWT). The DWT was then computed from each segment of 256 samples/frame. The segments of unvoiced speech signal can be simply detected by comparing the energies of DWT at the lowest scale $\delta_1 = 2^1$ and the highest energy level used in making energy comparison among scales is set high up at $\delta_5 = 2^5$. Any segment contains its largest energy level estimated at scale $\delta_1 = 2^1$ is favorably classified as an unvoiced segment, otherwise, identified as the voiced segment. The energy threshold used in energy weighting procedure to detect the unvoiced segment is defined as the following equation;

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

where the UV parameter is a speech segment classified as unvoiced at which the n segment with its energy at δ_1 scale maximized [4].

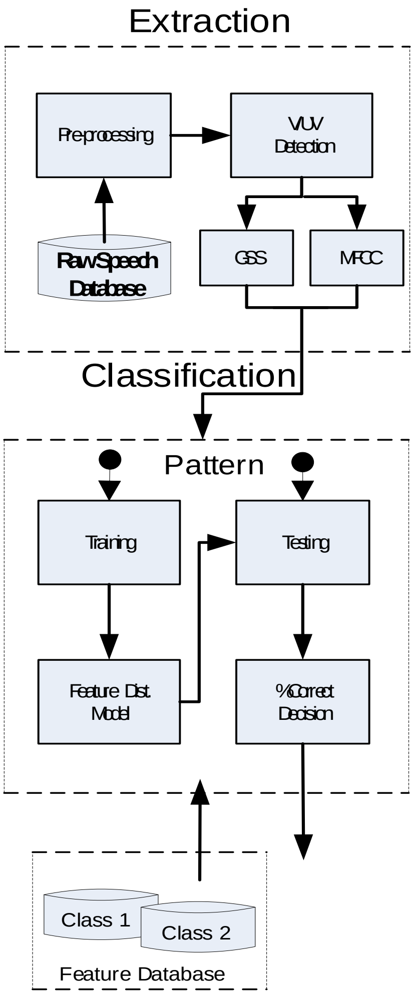

Fig. 1. Speech extraction and classification procedure.

C. Speech Feature Extraction

In this state all voiced segments with a window length of 25.6 milliseconds formerly detected from the database of all speech records are extracted for the Mel-Scale Frequency Cepstral Coefficients (MFCC) and the Glottal Spectral Slope. In this work the similar technique reported in the formerly published studies [9]-[15] has been adapted and employed to estimate for MFCC and GSS parameters. The estimation procedure of studied features can be described in details as follows: First the whole concatenated segments of voiced speech corresponding to individual speaker were windowed into the consecutive segments with a frame length of 25.6ms, estimating the Logarithm of the Discrete Fourier Transform for all segments, calculating the energy of the log-magnitude spectrum filtered out via a 16-triangular Band-pass filterbank with center frequencies respective to the Mel-frequency scale frequency response, computing the Inverse Discrete Fourier transform (IDFT) which represent for the sixteen–order cepstral coefficients corresponding to the vocal-tract response in term of frequency characteristic response of the filter, the second major part of the speech production system. Next step is a procedure of parameter estimation to determine the GSS parameter for all voiced speech segments [6]. The complete procedure of estimation of GSS parameter is depicted in Fig. 2 in which its procedure begins with estimating the periodogram of the voiced segments, averaging the normalized periodogram for less variation in the vocal-tract response, approximating the line of least squares (LS) error that best fits to the estimated glottal spectrum, therefore determining the corresponding slope to the glottal spectrum by calculating a product between inverse matrix of frequency values and matrix of amplitude values. The outcome resulted from calculation then appeared in a form of matrix contained the desirable slope and intercept values of the LS optimal fitting line.

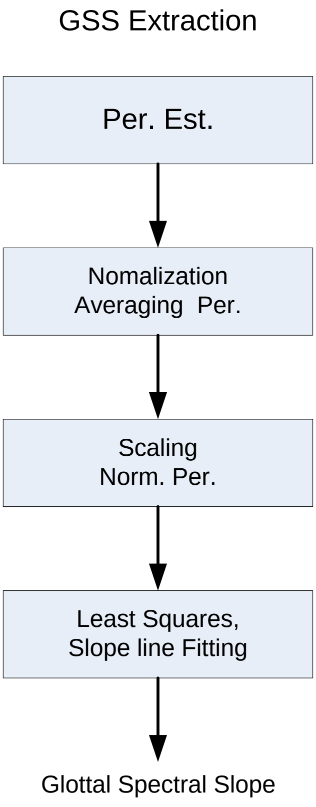

Fig. 2. Extraction procedure of Glottal Spectral Slope.

D. Principal Component Analysis

In prior to classifying all extracted parameters, the principal component analysis (PCA) was applied in prior to classification in order to reduce the multi-dimension of parameters to be lower. The reduced dimensional parameter consequently contains only major components with the most significance in term of distinct statistical characteristics which is adequate for training and testing the classifier.

E. Classification

The probabilistic model of distribution of the extracted features was first approximated. The PDF of the training feature set which represents each of the diagnostic subject group was modeled with the Unimodal Gaussian Modeling. The approximated model of extracted features obtained depends on how the feature distributes and its *a posterior* probability is maximized. Then, the classification performance was evaluated on the testing set of MFCC and GSS. The percentages in partitioning the sample database set randomly into small training sets are 20%, 35% and 50% and the rests of individual respective percentage to the total sum of 100% consequently used as testing sets respectively. These different numbers of training samples were purposely selected to observe the affection on the correct classification scores due to the various sizes of sample set used in training and testing the classifiers. Furthermore, the handout procedure was also employed in classification procedure to compensate for small database. This technique enables us to train and test on each sample of feature, but not use that same selected sample for both training and testing at a time. The procedure was repeated until all samples were completely classified. Several trials on the random selection of samples in training and testing approximately hundred times were made to obtain the average score for all individual classifications.

III. Results and Discussion

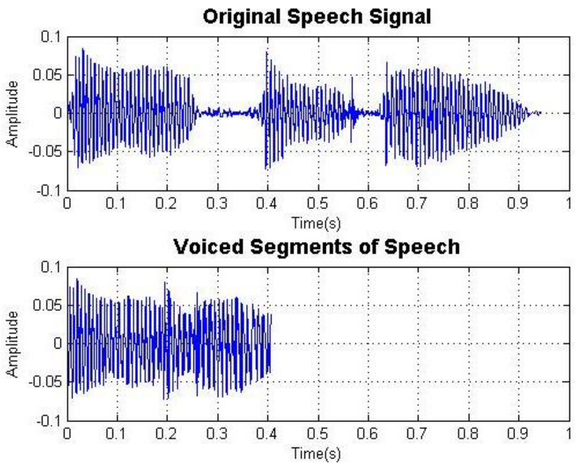

Fig. 3. Original speech signal (upper) and the detected segment of voiced speech (lower).

The procedures of whole study and detailed GSS estimation were depicted as the block diagrams shown in Fig. 1 and Fig. 2 respectively. The procedure of work shown in Fig. 1 can be separated into two main parts: extraction and classification. Fig. 3 shows the original speech waveform and the detected voiced segments obtained from the voiced- and unvoiced-segment detection based on the energy weighting threshold. As obviously notified from plots the difference in signal amplitude and time occurrence between the original speech signal and the detected voiced segment can be clearly observed in which all pauses (silent portions of spoken sound) in the original speech signal were completely eliminated shown in Fig. 3 as a plot of the voiced segment of speech signal. All voiced segments of speech in database were analyzed and then extracted for the GSS and MFCC features with the same estimation techniques throughout study for all speakers in database.

The average values of classification error were tabulated in comparison between different percentages of randomly selected training samples versus classifiers listed in Table I and Table II. The summarized averages from the pairwise classification on GSS and MFCC of depressed speech and those of suicidal speech were found to be approximately 0.335 in error when the 20% of randomly selected samples was used in training the Least Squares (LS) classifier. It seems to be not as that low error as expected from classification but these two features tended to be supportive of the possibility of being vocal psychiatric-related feature that should be further studied as a part of combination with other promising speech features. As shown in Table I, all error values found from classifying depressed and suicidal speech samples based on GSS and MFCC features as combined input to classifier with different sampling percentages of 20%, 35% and 50% were found to be slightly different in their values, which can be interpreted in that there is no any significant change in accurate scores of classification affected by the different size of samples for both cases of training and testing.

TABLE I: SUMMARIZED ERRORS FROM TESTING PHASE IN CLASSIFICATION BETWEEN DEPRESSED AND HIGH-RISK SUICIDAL GROUPS

Types of Classifier	**Sample percentages in testing classifiers**		
	20%	**35%**	**50%**
LS	*0.335*	0.358	0.355
RBF	0.368	0.405	0.533

TABLE II: SUMMARIZED ERRORS FROM TRAINING PHASE BETWEEN DEPRESSED AND HIGH-RISK SUICIDAL GROUPS

Type of Classifier	**Sample percentages in training classifiers**		
	80%	**65%**	**50%**
LS	0.062	0.052	0.075
RBF	0.090	0.082	0.104

In addition our experimental results have also revealed the outliners of classification score (see boxplots in Fig. 4 and Fig. 5). In Fig. 4 the cross markers in red located in areas above and below the upper and lower bounds of each box-and-whisker refer to the outliners of the average value of classification score found from training the classifier with different percentages of feature sample. Based on visual observation, a number of outliners appeared in Fig. 5 seems to be larger than that found in Fig. 4. These outliners came from when classifier identified either depressed or suicidal speech samples as suicidal speech, without including classification scores resulted from identifying input speech samples as depressed speech into account.

The higher number of outliners in case of identification of suicidal speech could be resulted from some suicidal speakers who might not be willing to or intentionally participate the interviewing program with their opened mind, and made themselves well corporation with psychiatrist during interview. Especially, classifying the high-risk suicidal speech sample has more outliners of error than a case of classifying depressed speech samples. This could suggest that the recording procedure has to be carefully taken with more time for subjects who participate in program to get more comfortable and peace of mind during interview to express verbally on what had been affected on their mind which probably moderates all that affection in their spoken speech during recording sound.

In future direction, analysis of new recordings of speech sample with the currently suggested consideration could be improved and supportive of the objective direction of research in a quest of developing the assistive tool which could achieve a task of evaluating the categorized subject groups with more accuracy. The data analysis applied via more various helpful techniques should be further employed into this ongoing study.

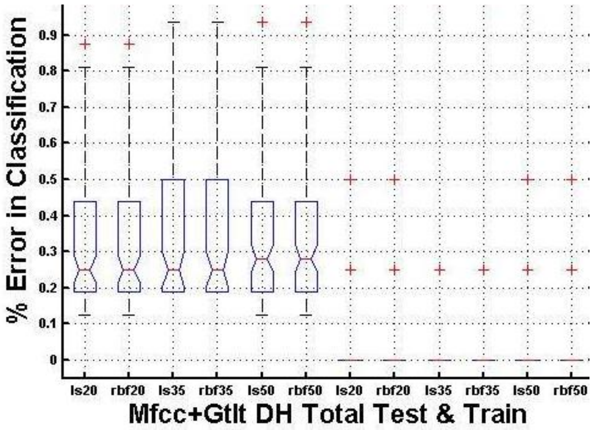

Fig. 4. Comparison of box plots of averaged classification error from identifying as depressed and suicidal speech groups with different training percentages.

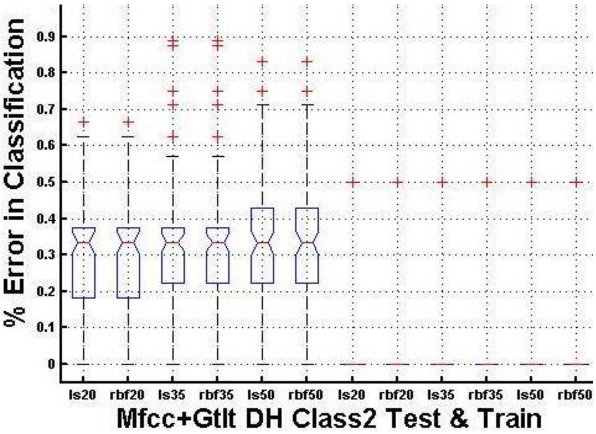

Fig. 5. Comparison of box plots of classification error from identifying as suicidal speech.

IV. CONCLUSION

The vocal parameters, MFCC and GSS extracted from speech samples collected from the focused groups of speakers during interview with psychiatrist were analyzed and classified to investigate for the feature's potential on assessment of the psychiatric level in those speaker. Results from study indicated that the separating characteristics contained in the studied MFCC and GSS parameters extracted computationally from our speech database can be fairly considered as additive measurement in term of quantitative change in acoustical parameters affected by the focused psychiatric disorders like depression and suicidal risk.

Besides, the various training percentages of the integrated speech parameters applied in classification were additionally investigated and they have shown not to have highly significant influence on the classification scores as the overall evaluation result.

Further direction of this ongoing study will be focusing on searching for more effectively acoustical parameters that can provide any improvement to research work considered as one of supplements in depression assessment with more support in term of high significant findings, and any of consideration found from this present work will be taken in account of developing more reliable analysis technique. As well, the larger speech database is also required for more desirable improvement in research work.

ACKNOWLEDGMENT

This work has been financially granted by National Research Council of Thailand.

REFERENCES

[1] D. J. France, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. on BME.*, vol. 47, pp. 829-837, 2000.

[2] T. Yingthawornsuk, "Comparative study on vocal cepstral emission of clinical depressed & normal speaker," in *Proc. ICCAS Conf.*, 2011.

[3] F. Tolkmitt *et al*., "Vocal indicators of psychiatric treatment effects in depressives and schizophrenics," *Journal of Communication Disorders*, vol. 15, pp. 209-222, 1982.

[4] A. Ozdas and R. G. Shiavi, "Analysis of vocal tract characteristics for near-term suicidal risk assessment," *Meth. Info. Med.*, vol. 43, pp. 36-38, 2004.

[5] H. Stassen, "Modeling affect in terms of speech parameters," *Psychopathology*, vol. 21, pp. 83-88, 1988.

[6] A. Ozdas and R. G. Shiavi, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Trans. BME.*, vol. 51, pp. 1530-1540, 2004.

[7] K. Scherer, *Nonlinguistic Vocal Indicators of Emotion and Psychopathology*, New York: Plenum Press, 1979, pp. 493-529.

[8] M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 23, pp. 56-62, 1960.

[9] T. Yingthawornsuk, "Comparative study of pairwise classification by ML & NN on unvoiced segments in speech sample," in *Proc. ICSEE Conf.*, 2012.

[10] T. Yingthawornsuk, "Classification of depressed speakers based on MFCC in speech sample," in *Proc. ICEEE Conf.*, 2012.

[11] J. I. Godino-Llorente, "Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short term cepstral parameters," *IEEE Trans. on BME.*, vol. 53, pp. 1943-1953, 2006.

[12] L. A. Low and N. C. Maddage, "Content based clinical depression detection in adolescents," in *Proc.17th EUSIPCO Conf.*, 2009, pp. 2362-2366.

[13] T. Yingthawornsuk and R. G. Shiavi, "Distinguishing depression and suicidal risk in men using GMM based frequency contents of affective vocal tract response," in *Proc. ICCAS Conf.*, 2008.

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

[15] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *Proc. ICASSP Conf.*, 1986.

Thaweewong Akkaralaertsest was born in Bangkok, Thailand on September 18, 1970. He received the B.Eng and M.Eng degrees in electrical engineering from Rajamangala University of Technology Thanyaburi, Thailand in 2001 and 2011, respectively. He is a lecturer in Division of Electronics and Telecommunication Engineering, Faculty of Engineering, Rajamangala University of Technology KungThep, Bangkok, Thailand. His research interests are in electronic and telecommunication engineering.

Thaweesak Yingthawornsuk was born in Bangkok, Thailand on August 30, 1970. He received the B.Sc. degree in communication engineering from King Mongkut's Institute of Technology Thonburi in 1993. He received the M.S. and Ph.D. degrees in electrical engineering from Vanderbilt University, TN, USA in 2003 and 2007, respectively. Since 2007 he has been worked as a lecturer at King Mongkut's University of Technology Thonburi, Bangkok, Thailand. He has served as the chairman of B.Sc. program in Media Technology and the head of Biomedical Media Technology Research Laboratory at the same university since 2008. His research interests are in speech processing, biomedical signal processing, physiological system identification, modeling and classification. Currently his ongoing research involves prediction of psychiatric related disorder based on applied speech processing techniques, acoustical parameter modeling and classification.