SCREENING FOR HIGH RISK SUICIDAL STATES USING MEL-CEPSTRAL COEFFICIENTS AND ENERGY IN FREQUENCY BANDS

Hande Kaymaz Keskinpala[1], Thaweesak Yingthawornsuk[1], D. Mitch Wilkes[1], Richard G. Shiavi[1, 2] and Ronald M. Salomon[3]

[1]Department of Electrical Engineering and Computer Science, Vanderbilt University, 37235, Nashville, TN, USA
[2]Department of Biomedical Engineering, Vanderbilt University, 37235, Nashville, TN, USA
[3]Department of Psychiatry, Vanderbilt University School of Medicine, 37212, Nashville, TN, USA
email: {hande.kaymaz, thaweesak.yingthawornsuk, mitch.wilkes, richard.shiavi, ronsalomon}@vanderbilt.edu

ABSTRACT

Distinguishing high risk suicidal patients from less severely depressed patients at low risk is a critical problem. This paper describes a novel way to address this issue. The vocal characteristics of male and female speech samples from high risk suicidal and depressed patients were analyzed and distinguished using mel-cepstral coefficients and using energy in frequency bands. Two kinds of speech samples, one from an interview session and the other from a reading session, were analyzed. The results show that mel-cepstral coefficients and energy in frequency bands may be used to separate these populations and the controlled reading tended to provide better results than the interview.

1. INTRODUCTION

Suicide is a major public health problem. The number of people who died because of suicide was 32,439 in 2004 and suicide was the eleventh leading cause of death in United States that year. The overall rate was 10.9 suicide deaths per 100,000 people. [1] 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. [2] However it is also very important to evaluate risk of committing suicide, when a person is seen by care givers 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. M.K. Silverman and S.E. Silverman proposed using vocal parameters of speech for deciding if a patient is suicidal or not [3]. 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]. Tolkmitt et al. observed patients' speech during the course of recovery and compared the formant information of vowels that occurs in the identical phonetic context [5]. 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 [6]. Ozdas et al. proposed using lower order mel-cepstral coefficients among suicidal patients, major depressed patients and non-suicidal patients [7]. 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 [8].

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 [8]. Thus environmental compensation was necessary and applied by cepstral mean normalization to compensate the spectral variability introduced by possible differences in recording environments. The method of cepstral mean normalization was accomplished by subtracting the average of the feature set computed for each patient's recordings from each mel-cepstral coefficient vector. 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 this state. In this research, the classifications were performed both with and without environmental compensation in order to determine if compensation was still important. [9, 10] The results show that classification performance is better without this compensation. Since the subjects are recorded in the same environment, the environmental compensation used in the previous study is not needed, and indeed appears to negatively impact performance. The other difference between the previous study [8] 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.

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. The study of energy in voice has been previously investigated and reported by France [4] who found that the percentages of total energy in frequency bands over 0-2000 Hz were powerful features in classifying between groups of control, major depressed and suicidal subjects for both male and female studies. With the new set of data, energy in frequency bands features were reinvestigated and statistically compared between two different types of audio recording, free speech and reading text.

In this paper, only the high risk suicidal/depressed pairwise classification results are presented using different numbers of mel-cepstral coefficients and energy in frequency bands as features. The research database also contains audio recordings for remitted subjects (patients who had been depressed previously but recovered) and ideators (patients who are thinking and talking about suicide but are not yet at high risk), but they are not presented herein. The reason for presenting only this classification is that it is most important and difficult, clinically, to distinguish the high risk suicidal patients from the depressed ones.

2. METHODOLOGY

2.1 Database Information

The database consists of female and male patients who were categorized as either depressed or at high risk of suicide by a clinician who was not involved in the research project. The database had two basic type of speech samples recorded: interviews with a researcher (i.e., the interview session) and a reading of the "Rainbow Passage" (the reading session). The "Rainbow Passage" is used in speech science because it contains all of the normal sounds in spoken English, and is phonetically balanced. [11]

The numbers of patients in the database that are used for mel-cepstral coefficients study is shown in Table 1. The number of patients in the database that are used for energy in frequency bands study is shown in Table 2. The energy in frequency bands study's database has fewer subjects since this study was performed earlier than the mel-cepstral coefficients study.

The ages of the subjects were between 25 and 65 years. The speech data were gathered as part of an ongoing study within the Psychiatry Department of the Vanderbilt University School of Medicine, and was supported by the American Foundation for Suicide Prevention.

	Depressed		High Risk Suicidal	
Interview Session	**Female**	**Male**	**Female**	**Male**
	18	11	11	9
Reading Session	**Female**	**Male**	**Female**	**Male**
	16	14	9	9

Table 1 – Database for mel-cepstral coefficients study

	Depressed		High Risk Suicidal	
Interview Session	**Female**	**Male**	**Female**	**Male**
	10	8	10	8
Reading Session	**Female**	**Male**	**Female**	**Male**
	9	8	9	8

Table 2 – Database for energy in frequency bands study

A portable audio data acquisition system was used for recording. This system consisted of a Sony VAIO laptop computer with a 2 GHz Pentium IV CPU, 512 Mb memory, 60 GB hard drive, 20X CD/DVD read/write unit, 250 GB external hard drive, Windows XP OS, ProTools LE digital audio editor, a Digital Audio MBox for audio signal acquisition; and an Audix SCX-one cardioid microphone.

The preprocessing was done by digitizing all speech signals using a 16-bit analog to digital converter at a sample rate of 10 kHz with an anti-aliasing filter (i.e., 5 kHz low-pass filter). Using the GoldWave v.5.08 audio editor, voices other than the patient's voice were removed as well as silences longer than 0.5 seconds. For female patients 76 seconds of continuous speech, and for male patients 66 seconds of continuous speech, were stored for subsequent analysis in both the interview and reading sessions during the mel-cepstral coefficients study. For the energy in frequency bands study, 60 seconds of continuous speech is used for male patients in both reading and interview session and 120 seconds of continuous speech is used for interview session and 60 seconds for reading session all female patients.

In this study, the voiced/unvoiced detection is performed for all patients' speech data and only voiced segments are used.

2.2 Feature Extraction

In this research, two different types of features are used. They are mel-cepstral coefficients, and energy in frequency bands (power spectral density).

2.2.1 Mel-Cepstral Coefficients

Mel-cepstral coefficients are coefficients derived from a type of cepstral representation of the speech sample. The difference between the cepstrum and mel-cepstral coefficients is that the frequency bands are positioned logarithmically, according to the mel-scale, for mel-cepstral coefficients.

Mel-scale approximates the human auditory system's response more closely than the linearly spaced frequency bands obtained directly from the FFT. [12] The mapping between linearly spaced frequency and mel-scale frequency is shown in Figure 1.

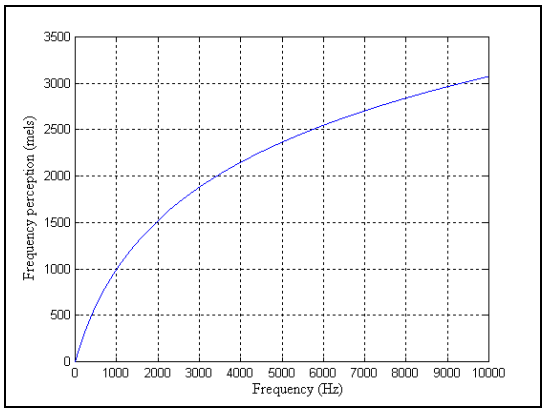

Figure 1 – Mapping between linearly spaced frequency and mel-scale frequency

Four and eight mel-cepstral coefficients for each speech sample were obtained following the steps described below:

- Divide each signal into segments of 512 points of voiced speech.
- Compute the log-magnitude spectrum (logarithm of the discrete Fourier transform (DFT)) for each voiced speech segment.
- Filter the log-magnitude spectrum by a series of 16 triangular band pass filters whose center frequencies are selected according to mel-scale.[13]
- Perform vocal tract length normalization to each patient [14].
- Calculate the inverse discrete Fourier transform (IDFT) to obtain the mel-cepstral coefficients.

The procedure for extracting mel-cepstral coefficients are also shown in Figure 2, below.

Vocal tract length normalization counteracts the effects of different vocal tract lengths among all patients in the study through linear warping of the frequency axis [14].The new frequency scale was obtained by multiply the frequency by a warping factor k_P. The warping factor k_P is the ratio between a speaker's vocal tract length (VTL) and baseline VTL. The average of third formant and fourth formant were used to represent the VTL properties for each subject.

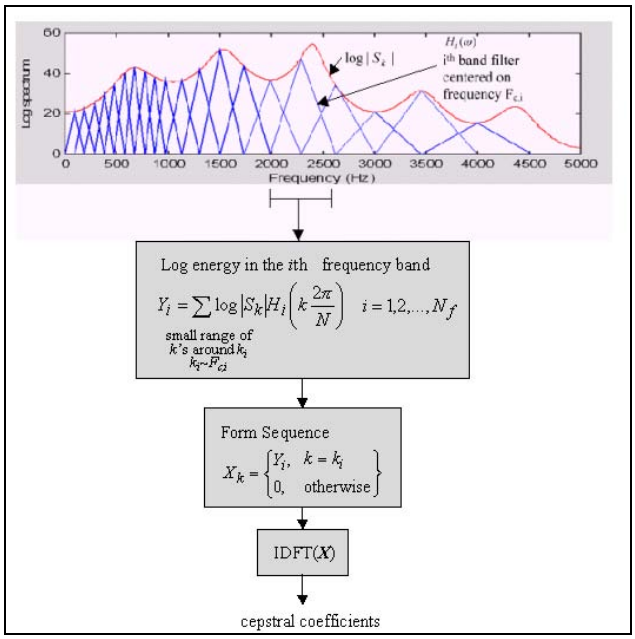

Figure 2 – Procedure for extraction of mel-cepstral coefficients

2.2.2 Energy in Frequency Bands (Power Spectral Density)

The energy in frequency bands features are extracted from 500 Hz frequency bands in the 0 to 2000 Hz range of the voice spectrum. The method used is:

- Divide each signal into segments of 512 points of voiced speech.
- Compute the PSD based on Welch method with 1024 –point FFT, no-overlapping 512-point Hamming window.
- Estimate the value of peak power, its frequency location and the percentages of the total power for each 500 Hz frequency sub-bands of 0-500 Hz (PSD_1); 500-1000 Hz (PSD_2); 1000-1500 Hz (PSD_3), 1500-2000 Hz (PSD_4).

2.2.3 Comparative Statistical Class Features

The averaged values of features over all segments were used in class feature selection. These five acoustical features (i.e., peak power, peak location, PSD_1, PSD_2, PSD_3) were formed as a vector representing each individual. The PSD_4 were not taken into this comparative statistical classification due to the property of linear dependency among sub-bands. For each subcategory, a Fisher discriminant analysis was used on each pairwise comparison between suicidal and depressed groups to find the features that have the highest separation power. The determined primary features were next used in cross-validation classification as predictor variables.

2.3 Classification

Two types of classifiers were applied in this research. The first consisted of quadratic discriminant analysis performed on the available mel-cepstral data. Quadratic discrimination

fits multivariate normal (MVN) densities with covariance estimates stratified by group. This method uses likelihood ratios to assign observations to groups. In this part of the study, different numbers of cepstral coefficients were used and the percentage of correct classification is determined for each class using unimodal Gaussian modelling with 4 and 8 features. The classification was done by using the equally weighted prior probabilities for both the high risk group and the depressed group.

In voiced energy in frequency bands study, cross-validation with the quadratic discriminant function was performed. We randomly split our measurements into two sets of samples, 65% and 35%, use the 65% of samples as a training set to estimate the classification function, and then use the resulting functions to classify 35% of samples as a testing set. To randomly select these sets of samples, we generated the uniformly random number between 0 and 1 and then assigned a weight of 1 for a random number that is less than 0.65 and a weight of 0 for a random number greater than 0.65. All numbers of 1 and 0 were stored in the weight variable as additional variable to a vector of features being used in classification.

For each pairwise study, we performed the cross-validation classification for four times to complete all 100% samples with a new randomly selected 65% training set and a 35% testing set for each time. This classification method provides us more statistically reliable analysis on an empirical measure for the success of the discrimination.

3. RESULTS

3.1 Mel-Cepstral Coefficients Results

The depressed-suicidal pairwise classification using quadratic discriminant analysis for the female patients is shown in Table 3. In this analysis, all of the data was used to train the classifier.

	Interview Session	Reading Session
4 Mel-Cepstral Coefficients	69.23%	67.82%
8 Mel-Cepstral Coefficients	72.53%	73.37%

Table 3 – Female depressed/ suicidal classification

This table shows the classification results for the interview session and the reading session. The classification performance increased from 69.23% to 72.53% in the interview session, and increased from %67.82 to 73.37% in the reading session, using 8 mel-cepstral coefficients instead of using 4 mel-cepstral coefficients. Both of the results show that using 8 mel-cepstral coefficients yielded better classification performance for both the interview and the reading sessions than using 4 mel-cepstral coefficients.

The depressed/suicidal pairwise classification using quadratic discriminant analysis for the male patients is shown in Table 4.

	Interview Session	Reading Session
4 Mel-Cepstral Coefficients	70.37%	71.28%
8 Mel-Cepstral Coefficients	77.84%	80.62%

Table 4 – Male depressed/ suicidal classification

This table shows the classification results for the interview session, and the reading session. In general, the results for the male subjects are qualitatively similar to those of the female subjects, with 8 mel-cepstral coefficients yielded better classification performance. These analyses show that the collected depressed and suicidal populations are classified with 70.37% by using 4 mel-cepstral coefficients and 77.84% by using 8 mel-cepstral coefficients in the interview session. For the reading session, the performance is increased from %71.28 to 80.62%, using 8 mel-cepstral coefficients instead of using 4 mel-cepstral coefficients. Overall, as expected, the average classification performance of the depressed/suicidal pairwise analysis increased as the number of coefficients was increased.

3.2 Energy in Frequency Bands Results

It was determined that the percentages of the total power, PSD_1, PSD_2 and PSD_3 would be the primary features for female study and PEAK, PSD_1 and PSD_2 as primary features for male study. The results are shown in Table 5 and 6. The pairwise classification from the reading session provides the highest score performance of 87% and 75.5% for female and male studies, respectively, compared to interview session. These energy features seem to be more effective discriminators in classifying the female pairwise population than in classifying the male pairwise population for both sessions of interview and reading passage.

	Interview Session	Reading Session
Optimal Feature	PSD_1, PSD_3	PSD_2
	84.25%	87%

Table 5 – Female depressed/ suicidal classification with cross validation method for energy in frequency bands study

	Interview Session	Reading Session
Optimal Feature	Peak, PSD_1	Peak, PSD_2
	74.25%	75.5%

Table 6 – Male depressed/ suicidal classification with cross validation method for energy in frequency bands study

4. CONCLUSIONS

The results of this research show that mel-cepstral coefficients are useful indicators for discriminating between high suicidal patients and depressed patients. Another observation

was that the controlled reading tended to provide better results than the interview. It gave consistently more reliable results; therefore using text-dependent speech samples may provide a good diagnostic tool for assessing suicidal patients. We observed that the suicidal and depressed populations are largely separable, which bodes well for the prospect of eventually using these features in clinical practice. For the future work, Gaussian Mixture Models of the populations may provide better classification over unimodal Gaussian models.

The studied energy features of voiced speech suggested that the power spectral density can be used as a particularly effective discriminator in differentiating depressed speech from suicidal speech which is able to perform as the helpful indicator in associated with diagnosis by clinicians. Results from discriminating analysis in reading passage study suggested that text-dependent based recording can be performed as another assessment approach of suicide risk and also provides good performance in designing an optimal classifier based on cross-validation approach. These findings illustrated that our approach of acoustical acquisition and feature extraction should be further investigated in effectiveness of classification.

REFERENCES

[1] NIMH (National Institute of Mental Health) web site, http:// www. nimh.nih.gov/publicat/harmsway.cfm

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

[3] S.E. Silverman, "Vocal parameters as predictors of near-term suicidal risk", U.S. Patent 5 148 483, Sept. 1992.

[4] D.J. France, R. G. Shiavi, S. Silverman, M. Silverman, D. M. Wilkes, "Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk", IEEE Transaction on Biomedical Engineering 2000, Vol. 47(2), pp. 829-837, 2000.

[5] 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.

[6] T. Yingthawornsuk, H. Kaymaz Keskinpala, D. France, D. M. Wilkes, R. G. Shiavi, R.M. Salomon, "Objective Estimation of Suicidal Risk using Vocal Output Characteristics", International Conference on Spoken Language Processing (ICSLP-Interspeech 2006), 2006, pp. 649-652.

[7] A. Ozdas, R. G. Shiavi, D.M. Wilkes, M.K. Silverman, S.E. Silverman, "Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment", Methods of Information in Medicine, Vol. 43, pp.36-38, 2004.

[8] A. Ozdas, "Analysis of Paralinguistic Properties of Speech for Near-Term Suicidal Risk Assessment", Ph.D. Dissertation, Vanderbilt University, 2001.

[9] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Transactions on Acoustics Speech and Signal Processing, Vol. 29, No.2, pp. 254-272, 1981.

[10] B.S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", J. Acoust. Soc. Am., Vol. 55, No.6, pp.1304-1312, 1974.

[11] G. Fairbanks, *Voice and Articulation Drillbook.* Harper &Row, New York, 1960.

[12] F. Zheng, G. Zhang, Z. Song, "Comparison of Different Implementations of MFCC", Journal of Computer Science & Technology, Vol. 16(6), pp. 582-589, September 2001.

[13] D. O'Shaughnessy, *Communication: Human and Machine,* Addison-Wesley, Massachusetts, 1987.

[14] E. Eide, H. Gish, "A Parametric Approach to Vocal Tract Length Normalization", ICASSP Proceedings, Georgia, 1996, pp. 346-348.