Characterizing Mental State of Depressive Speaker Using Sub-Band Spectral Energy

Thaweesak Yingthawornsuk

Abstract— **The acoustical parameters of spoken sound have been previously reported that its characteristics can possibly indicate the status of mental state in depressive speakers. Work proposed in this paper is an expansion of investigation if the certain vocal outcome in frequency domain can represent the severity level of the mental status and related by the Beck Depression Index, (BDI). Male and female speech samples were analyzed and compared with the evaluated BDI scores for predicting the mental state of speakers. The experimental results reveal that in the lowest frequency range the sub-band spectral energy can significantly predict the mental state and highly correlated to BDI score of females reading speech samples.**

Keywords— **classification, depression index, reading speech, sub-band spectral energy,**

I. Introduction

Suicide is a common outcome in persons with serious mental disorders. However, it remains a phenomenon that is still under-researched and poorly understood. Moreover, methods to help to identify persons who are at elevated risk are sorely needed in clinical practice. This study represents an attempt to identify characteristic vocal patterns in persons with imminent suicidal potential which could lead to the development of new technology to aid in the assessment of suicidal potential. This research has been studied for the vocal acoustic properties in suicidal state. Three groups of following speakers are contrasted: high-risk suicidal, depressed, and remitted.

In published pilot studies [1-2], [4], the analytical techniques have been developed to determine if subjects were in one of mental states associated with non-suicidal depressed or severe suicidal. The initial sets of recordings used for these published analyses were made in a wide variety of clinical and technical conditions, without the advantages of an acoustically controlled environment and high-technology recording equipment. Most were recorded in the 1960's through '80's by a clinical practitioner (the late Dr. S. Silverman), who routinely taped his therapy sessions. He assembled this set of tapes for just such studies of acoustical characteristics of suicidal speech, which he strongly believed could produce a clinical tool for detection of high-risk individuals. Each selected tape predated a known subsequent suicide attempt with high lethality or completed suicide. Comparison of subjects' speech was taken from the same tapes or from recordings made later in more controlled environments. All individual subjects were clinically diagnosed and assigned to the categorized groups of either healthy controls or non-suicidal depression. In the early studies using these clinical tapes, analysis focused on segments in the high-risk recordings selected by Dr. Silverman as evocative of suicidal speech sounds. The recorded tapes were randomly sampled. With this method, the diagnostic subject groups were successfully separable using parameters of vocal acoustics.

Vocal cues have been studied as indicators in diagnosing the syndrome underlying a person's abnormal behavior or emotional state by experienced clinicians [5-6], but these skills are not in widespread clinical use. The considerable evidence suggests that the emotional arousal can produce changes in the speech production scheme by affecting the respiratory, phonatory, and articulatory processes that in turn are encoded in the acoustic signal. Emotional arousal produces a tonic activation of striated musculature, and the sympathetic, and parasympathetic nervous systems. Changes in heart rate, blood pressure, respiratory patterns, muscle tension, and motor activity transiently alter respiratory, phonatory, and articulatory functions in speech production in an acutely state-related fashion, directly tied to emotions. Consequently, emotional disturbances can be expected to cause measurable changes in speech parameters. Certain changes in speech parameters may be specific to near-term suicidal states.

Emotional content of the voice can be associated with acoustical variables such as the level, range, contour, and perturbation of the fundamental frequency, the vocal energy, the distribution of energy in the frequency spectrum, the location, bandwidth and intensity of the formant frequencies, and a variety of temporal measures. Research has shown that depression has a major effect on the acoustic characteristics of voice as compared to normal controls. Prosody is slower and the energy in the speech is distributed differently over the frequency range between 0 and 2,000 Hz [1].

The following sections are organized: Section II provides methods in details of database, the subject populations involved, the acoustical feature extraction, and statistical analyses. Section III presents the results of regression modeling, classification respective to types of recording sessions and gender among three clinically diagnostic subject groups. Sections IV concludes work.

T. Yingthawornsuk is with King Mongkut's University of Technology Thonburi, Thailand.

https://doi.org/10.17758/URUAE14.UL05211008

II. Methodology

A. Database

Each studied subject from each diagnostic group has two types of speech samples recorded. They are speech samples one from an interviewing session with a therapist and another from a session which subject reads a predetermined part of a book. The unedited speech was randomly extracted from the interview session and reading passage session to represent each subject. During the reading session, each subject read the standardized text, the "Rainbow Passage" [8], which is used in speech science since it contains all the normal sounds in spoken English and phonetically balanced.

All recorded speech samples include 13 depressed males, 10 high-risk suicidal males and 9 remitted males. The subjects' age ranges from 25 to 65 years. All speech signals were digitized by using a 16-bit analog to digital converter with a sampling rate of 10 kHz with an anti-aliasing filter. The background noise and the unwanted sound rather than the subject' s voice were removed by an audio editor. This editing software was also used to remove the silences which were longer than 0.5 seconds for getting a continuous speech record. The preprocessing is finished by dividing the edited continuous speech into 20-seconds segments. The beginning and ending points of each segment were selected at zero crossings in the edited continuous speech. The different lengths of speech recording between interviewing and reading sessions are approximately eight minutes and two minutes respectively for each subject.

In preprocessing state two steps were applied. First the voiced segments of each speech sample were detected regarding the highest energy contained in that segment found and then segment collected. Second, detrending and normalization were applied to have a variance of 1 before analysis to compensate for possible differences in recording level among subjects.

To evaluate the diagnostic emotional state, each subject first completed the Beck Depression Inventory, BDI, [7] before participating in interviewing sessions. This is a standard, brief, self-rated inventory used as a measure for mood.

B. Feature Extraction

Power spectral densities (PSD's) of the voiced speech were obtained by using the classical method of PSD estimation based on Welch method with non–overlapping 100–point Hamming windows. The algorithm was written in MATLAB with using 1024-point fast Fourier transforms (FFT) to estimate the spectra with 40-ms windowing over each 20-second segment. Six features were calculated. Four were the power in the four different frequency ranges: from 0 Hz to 500 Hz, 500 Hz to 1000 Hz, 1000 Hz to 1500 Hz, and finally from 1500 Hz to 2000 Hz. The other two were the value of peak power and the frequency of the peak power. For each of the 500 Hz sub-bands (PSD_1, PSD_2, PSD_3, PSD_4) the percentages of total power were calculated and stored as a set of input parameters to state of classification. The other two features are the value of peak power and the frequency location of the peak power were also collected.

C. Regression Analysis

Because the percentages of power were used as features, only the power in the first three sub-bands can be independent and were analyzed. The BDI and the acoustical features were stored in matrix form for linear regression analysis. The BDI is the dependent variable formulated from a linear combination of sub-band energies. Its equation model is shown in equation 1.

$$bdi\_est = a_0 + a_1 psd_1 + a_2 psd_2 + a_3 psd_3 \tag{1}$$

Where the estimate of the BDI score is *bdi_est*, the weighting coefficients, a_i and the sub-band energies, psd_i respectively.

A step-wise procedure was used to determine if all or any of the three independent variables were significant for the relationship [9].

D. Statistical Analysis and Classification

All PSD parameters were arranged and stored in a matrix form for statistical analysis. Each output matrix contained N rows and M columns (N x M matrix), where N is a number of voiced speech frames and M is a number of acoustical features. Mathematically, all PSD parameters representing the suicidal, depressed, and remitted speech classes were defined into three large matrices. The parameter matrix representing each class were imported and implemented in matlab. In a state of classification of the between-groups were designed, (i.e. suicidal/depressed (**SU/DP**), depressed/remitted (**DP/RM**), suicidal/remitted (**SU/RM**). Classification accurate scores and performance of selected classifiers were tested and evaluated with using the hold-one-out method, and 95% confidence interval were used in statistical analyses. The hold-one-out method was used in this discriminant analysis to compensate for the small size of speech databases used in this study.

First, the feature samples were randomly selected for a 35% set to train classifiers, and the rest 65% of sample to validate the classification accuracy. The K-fold cross-validation on several trials on random sample selection for training and testing approximately hundred times are employed for the average performance of classifications.

III. Results

Means and standard deviations of the set of features, peak power, peak location (in Hz), and the percentages of total power from each categorized subject group during the interview were summarized in Table I. Suicidal speech was characterized by peak location which was significantly lower frequency when compared to remitted speech type. These can be also seen in case of reading passage session summarized in Table II.

TABLE I: PSD STATISTICS FOR MALE (INTERVIEW SESSION)

Class	Suicidal	Depressed	Remitted
Peak Energy	20.88, 2.70	20.63, 3.34	20.92, 1.70
Peak Location	284.47, 84.89	292.02, 58.55	331.17, 67.98
PSD_1	0.79, 0.08	0.79, 0.08	0.74, 0.05
PSD_2	0.19, 0.07	0.18, 0.08	0.23, 0.04
PSD_3	0.01, 0.01	0.03, 0.02	0.02, 0.01

Mean and standard deviation values are presented.

TABLE II: PSD STATISTICS FOR MALE (READING SESSION)

Class	Suicidal	Depressed	Remitted
Peak Energy	21.52, 2.22	20.80, 1.59	21.53, 2.08
Peak Location	298.83, 112.84	296.10, 66.11	351.65, 74.24
PSD_1	0.78, 0.08	0.82, 0.06	0.75, 0.09
PSD_2	0.19, 0.08	0.16, 0.05	0.23, 0.09
PSD_3	0.01, 0.01	0.02, 0.01	0.01, 0.01

The depressed speech exhibited elevated PSD_1, PSD_3 and reduced PSD_2 for interview session when compared to the remitted speech. This trend can be also noticed for the reading passage session. For the results of the suicidal-remitted pairwise study, the percentage of the total power in the 0 to 500-Hz sub-band (PSD_1) was reduced for remitted speech while the percentage of power in the higher sub-bands (PSD_2 and PSD_3) increased. These can be also noticed for the reading passage session except PSD_3. It indicated no significant difference for the percentage of total power between groups.

TABLE III: SENSITIVITY, SPECIFICITY, POSITIVE PREDICTIVE (PPV), AND NEGATIVE PREDICTIVE (NPV) VALUES FOR MALE PAIRWISE (INTERVIEW SESSION) AND CLASSIFICATION ANALYSES

Pairwise Group	%ACC.	SV	SP	PPV	NPV
SU/DP	77	0.89	0.63	0.76	0.80
DP/RM	94	0.94	0.94	0.94	0.94
SU/RM	85	0.91	0.76	0.84	0.86

The results of pairwise discriminant analyses performed on the male study populations were summarized in Tables III and IV for interview and reading passage session, respectively. For interview speech session the depressed subjects were well differentiated from suicidal ones (77%). However, the depressed subjects and suicidal patients were effectively (i.e. 94%, 85%) differentiated from remitted subjects. These discriminant analyses were primarily on a basis of PSD's. Table III also summarized the cumulative classification scores obtained for each statistical analysis using the discriminating features and the performance characteristics of discriminant function with scores such as Sensitivity (SV), Specificity (SP), PPV, and NPV. To calculate a measure of SV, the clinical measurement of suicidal speech from the confusion matrix of classification was selected as the conditional parameter for suicidal/depressed pairwise study. In opposite direction, when a measure of SP was calculated, the clinical measurement of depressed speech was selected as the conditional parameter.

Table IV summarized all statistical scores and performances of classification for reading passage session. Interestingly, the classification results from the reading showed just the opposite trends. The differentiation between suicidal speech and depressed speech was correctly classified at 82%. Whereas the correct classification scores among the other comparisons were lower.

In Fig. 1 some shifting in spectral energy ratios from sub-band 1 to 2 can be obviously notified as a crossing over each other between a blue line (remitted energy) and a red line (depressed energy) while a suicidal energy (in yellow) always shows up in a middle for both sub-bands 1 and 2.

TABLE IV: SENSITIVITY, SPECIFICITY, POSITIVE PREDICTIVE (PPV), AND NEGATIVE PREDICTIVE (NPV) VALUES FOR MALE PAIRWISE (READING SESSION) AND CLASSIFICATION ANALYSES

Pairwise Group	%ACC.	SV	SP	PPV	NPV
SU/DP	82	0.73	0.88	0.82	0.82
DP/RM	73	0.85	0.58	0.71	0.76
SU/RM	75	0.73	0.76	0.73	0.76

The overall results of the regression analysis are shown in Table V. Noticed that the speech samples for both reading and interview sessions for all subjects could not be always obtained. The stepwise regression analysis indicated that only PSD_1 was sufficiently needed for the regression analysis, indicating that the energies in the other two bands were correlated to it. Thus, a univariate relationship was sufficient.

TABLE V: SUMMARY OF REGRESSION ANALYSIS

Gender	Session	Population	a_0	a_1	P-Value
Female	Reading	13	71.02	-21.78	0.01
	Interview	12	21.76	20.10	0.50
Male	Reading	10	-27.88	54.41	0.64
	Interview	11	-80.08	98.04	0.11

The range of the values of the data used for the regression is shown in Table VI. The range of the PSD values is the same for all speech groups but the range of the BDI scores differs. The maximum values are the same but the minimum values for the females are higher.

TABLE VI: RANGES OF BDI SCORE AND PSD_1

Gender	Session	BDI Score	PSD_1
Female	Reading	22-51	0.65-0.96
	Interview	23-51	0.61-0.95
Male	Reading	9-51	0.62-0.90
	Interview	9-51	0.60-0.92

The significance of the regressions as indicated by the p-values shows that only the speech during reading for females is a significant predictor of BDI, and hence severity of mental state shown in Fig.2. The relationship for male speech during interviews is marginally significant.

https://doi.org/10.17758/URUAE14.UL05211008

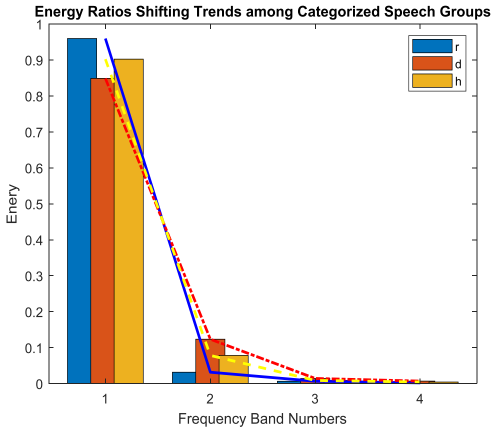

Fig. 1. Energy ratios shifting in categorized female speech samples.

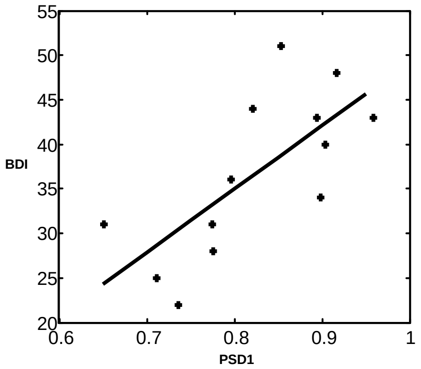

Fig. 2. Significant PSD_1 predictor of female reading BDI.

IV. Conclusion

The PSD_1 in frequency sub-band (0-500 Hz) was found as a significant predictor of mental state in female during reading speech session. This is consistent that the energy below 500 Hz increases during depression. It is known that the different mental pathways/cognitive processes are used during different interview and reading recording sessions, which is perhaps a reason why dissimilarity in result appeared for two different recording types. In male subjects no significance of PSD_1 related to BDI found means a strong relationship between these variables not in general. Inconsistencies above need more study on a larger database. Remitted group needs to be considered in as their BDI score is lower than depressed and suicidal groups.

References

[1] France, D.J., et al., *Acoustical properties of speech as indicators of depression and suicidal risk.* IEEE transactions on Biomedical Engineering, 2000. 47: p. 829-837. https://doi.org/10.1109/10.846676

[2] Scherer, K., Nonlinguistic Vocal Indicators of Emotion and Psychopathology, in Emotions in Personality and Psychopathology, C.E. Izard, Editor. 1979, Plenum Press: New York. p. 493-529. https://doi.org/10.1007/978-1-4613-2892-6_18

[3] Scherer, K.R., Vocal correlates of emotional arousal and affective disturbance, in Handbook of social psychophysiology, H. Wagner and A. Manstead, Editors. 1989, Wiley: New York.

[4] Darby, J.K., *Speech and voice studies in psychiatric populations*, in *Speech Evaluation in Psychiatry*, J.K. Darby, Editor. 1981, Grune & Stratton, Inc.: New York.

[5] Ozdas, A., et al., Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment. Methods of Information in Medicine, 2004. 43: p. 36-38. https://doi.org/10.1055/s-0038-1633420

[6] Ozdas, A., et al., Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk. IEEE Transactions on Biomedical Engineering, 2004. 51: p. 1530-1540. https://doi.org/10.1109/TBME.2004.827544

[7] Beck, A.T., et al., *An inventory for measuring depression*. Arch Gen Psychiatry, 1961. 4: p. 561-571 https://doi.org/10.1001/archpsyc.1961.01710120031004

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

[9] Afifi, A.A. and S.P. Azen, *Statistical Analysis: A Computer Oriented Approach*. second ed. 1979, New York: Academic Press.