



Direct Acoustic Feature Using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Speech

Yingthawornsuk T.^{1,4}, Kaymaz Keskinpala H.¹, Wilkes D.M.¹, Shiavi R.G.^{1,2}, Salomon R.M.³

¹Department of Electrical Engineering and Computer Science, Vanderbilt University, TN, USA

²Department of Biomedical Engineering, Vanderbilt University, TN, USA

³Department of Psychiatry, Vanderbilt University School of Medicine, TN, USA

⁴Department of Electrical Technology Education, KMUTT, Bangkok, Thailand

Abstract—Research has shown that the voice itself contains important information about immediate psychological state and certain vocal parameters are capable of distinguishing speaking patterns of speech signal affected by emotional disturbances (i.e., clinical depression). In this study, the GMM based feature of the vocal tract system response and spectral energy have been studied and found to be a primary acoustic feature set for separating two groups of female patients carrying a diagnosis of depression and suicidal risk.

Index Terms: suicidal speech, depression, vocal tract, energy

1. Introduction

Suicide is a common outcome in persons with serious mental disorders. However, it remains a phenomenon that is underresearched and poorly understood. Moreover, methods to help to identify persons who are at an 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 project is to study vocal acoustic properties in suicidal states. Two study groups will be contrasted in this work: near-term suicidal and depressed. In the early 1980's the Silvermans began to collect and analyze recorded suicide notes and interviews made shortly before suicide attempts. Their results suggested that voice can provide important information about immediate psychological state. They have described that the depressed patients have the same vocal speech as suicidal patients but the tonal quality of speech changes significantly when patients become suicidal.

As reported in [1], [2], [3], the emotional arousal produces changes in the speech production scheme by affecting the respiratory, phonatory, and articulatory processes that in turn are encoded in the acoustic signal. The emotional content of the voice can be associated with acoustical variables such as the level, range, contour, and perturbation of the fundamental frequency, the distribution of energy in frequency spectrum, the location, bandwidth and intensity of formant frequencies, and a variety of temporal measures. The measurable change in vocal parameters affected by emotional disturbances is able to be evaluated by utilizing an appropriate speech processing approach associated with certain acoustic features. Researches have shown that depression has a major effect on the acoustic characteristics of voice when compared to the normal controls. Certain changes in acoustic properties of the affective speech are possibly specific to the near-term suicidal states in persons.

In the published pilot studies [4], [6], analytical techniques have been developed to determine if subjects were in one of three mental states: healthy control, non-suicidal depressed, or high-risk suicidal. Several studies have used the vocal tract (VT) measures (i.e., formants) and prosody to classify the emotional disorders. France et. al [4] found the formants and percentages of total energy in frequency spectrum over a frequency range of 0–2,000 Hz to be the most distinguishing acoustic feature set for classifying groups of control, major depressed, and suicidal subjects. These features were recently re-investigated and extracted from a new speech database recorded in a better controlled environment. The experimental results have shown that the investigated feature set was still found as powerful acoustic discriminators in distinguishing suicidal, depressed and remitted patients [1]. Ozdas et. al [6] used a set of low order mel-cepstral coefficients to identify speakers who were diagnosed to be major depressed, suicidal, and normal by a psychiatrist. Her comparative result of classification performance as a measure of group separation was significantly high. Moore et al. compared the results of speaking pattern recognition by employing the prosody, formant and glottal ratio/spectrum in classifying normal controls and depressed patients. The optimal classifiers designated by the glottal ratio/spectrum and formant performed most effectively to separate two individual groups [7].

In this work, the characterization of the vocal tract system and distribution of energy in frequency spectrum of speech signal are focused. The speech processing algorithm to solve a specific problem of extracting the vocal features representing the characteristics of the VT system response is implemented and proposed. The estimate of smoothed magnitude spectrum is determined via the cepstrum analysis and the spectral structure contained in that magnitude spectrum is modeled by a mixture of Gaussian density components whose model parameters are estimated via a well-known "Expectation-Maximization" (EM) algorithm.

This paper is organized as follows: Section 2 provides the descriptions of database, feature extraction, primary feature selection, and performance evaluation. Section 3 presents the results. Finally, section 4 concludes all findings from this work.

2. Database

The audio recordings were collected from two groups of patients; 10 suicidal females and 10 depressed females. The ages of female patients are between 25 and 65 years. The audio data acquisition was made from the ongoing research study

supported presently by the *American Foundation for Suicide Prevention*. Each subject has two types of speech samples recorded. One is a speech sample recorded from a clinical interview with a therapist and another is a speech sample recorded from a text-reading session. During the reading sessions, subjects read a predetermined part of a book which contains the standardized texts, called the "Rainbow Passage" [5]. It has been used in speech science since it contains all of the normal sounds in spoken English and it is phonetically balanced.

The same recording environment and settings were made for all interviews. This acoustically controlled environment is necessary for quantifying the relatively clean speech samples. In this work, the same audio acquisition system and preprocessing implementation as reported in [1] were employed. Additionally, two more preprocessing steps were made. First, all speech samples were tested for voicing and only voiced segments were stored for further analyses. Second, all speech samples were detrended and then normalized to compensate all possible differences in recording level among the categorized patients. In this work, the approximately 3 minutes of each patient's speech recorded during an interview session and 2 minutes of speech recorded during a text-reading session were respectively taken as speech measurements of two different session studies for further comparative analyses.

3. Acoustic Feature Extraction

3.1. Energy in Frequency Bands

The frequency spectra were determined for each 51.2ms frame of voiced speech by performing the classical Power Spectral Density (PSD) estimation based on Welch's method with the use of a 512-point Hamming window, no-overlapping sliding between the consecutive windows, and 1024-point fast Fourier transforms (FFT). Four spectral energy parameters (energy ratios) were extracted from the estimated spectrum in different frequency sub-bands: 0-500 Hz, 500-1,000 Hz, 1,000-1,500 Hz, and 1,500-2,000 Hz. These parameters are the percentages (PSD₁, PSD₂, PSD₃, and PSD₄) of the total energy in each of four 500 Hz sub-bands of a total 2,000 Hz frequency range [1].

3.2. Mixture of Gaussians for Extracting Spectral Based Feature of Vocal Tract System Response

This section outlines another approach of acoustic feature extraction involving a probabilistic model approximation for the spectral structure of the VT system that is affected by the severity of psychological state. The spectral characterization of the VT system was re-investigated utilizing cepstrum analysis and Gaussian mixture model (GMM). Fitting a GMM to the magnitude spectrum enables an extraction of the spectral based acoustic feature and this can be achieved with the use of an iterative EM algorithm. This basic learning algorithm for finding the maximum likelihood (ML) of a mixture model was introduced by Dempster, Laird and Rubin [8], and extended for superimposed signal by Schafer [9], and Feder [10]. The EM algorithm is a general method for solving a ML estimation problem with incomplete data where some of the random variables are observed and some are hidden. It is employed to estimate the Gaussian mixture distributions from a magnitude spectrum of speech signal. The EM algorithm comprises of two

main steps: Expectation step (E-step) and Maximization step (M-step). The E-step computes the expected values of data likelihood by using the current estimate of parameter and observes data by viewing the magnitude spectrum as a probability density function (strictly speaking, it is a frequency density function, the discrete equivalent to the probability density). This step requires the computations of likelihood and posterior probability for each bin in histogram resulting from each mixture component. The M-step uses a set of data after accumulating the sufficient statistics in the E-step to re-estimate the means, variances and mixture weights for all individual mixture components. It maximizes the likelihood of model parameters (ML estimates).

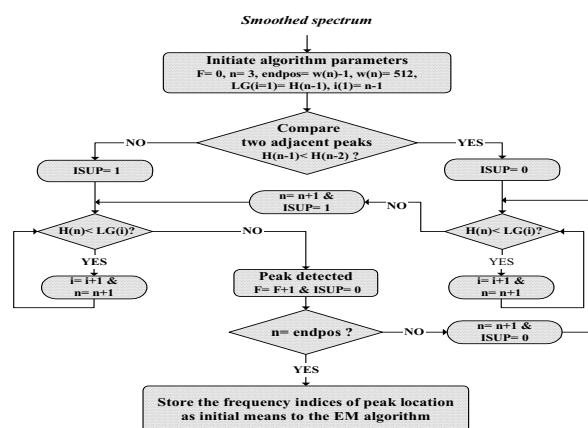


Figure 1: Flow chart of peak detection algorithm.

The total area under a histogram of the smoothed magnitude spectrum is calculated and all model parameters are initialized at this point. The mean parameters are initially determined by performing a peak detection algorithm, depicted in Figure 1, to locate the frequency indices of all spectral peaks. The obtained indices are then used as initial estimates of mean for the EM algorithm. The variances are made significant with respect to number of Gaussians in a mixture and the frequency intervals that are occupied by distributions. Additionally, the mixture weights are set equal for all individual Gaussian components. The EM algorithm can be used with most distribution types but Gaussians are generally employed in the standard applications of model approximation. The mathematical formulation for a GMM distribution can be simply

$$f_X(x; \theta) = \sum_{m=1}^M c_m \cdot \frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{1}{2}\left(\frac{x-\mu_m}{\sigma_m}\right)^2} \quad (1)$$

where the c_m is a mixture weight for each of the M mixture components, which must satisfy the following constraint that

$$\sum_{m=1}^M c_m = 1; c_m \geq 0, \quad (2)$$

and $\theta = [\sigma_m^2, \mu_m]$ represents the individual variance and mean.

The model estimates that parameterize a mixture of Gaussians can be collectively represented as

$$\lambda = \{c_m, \mu_m, \sigma_m^2\} \quad (3)$$

and the magnitude spectrum of the individual speech frame is represented by a GMM whose model estimates are referred to by λ . The procedure for extracting acoustic features that

characterize the VT frequency response, namely the energy concentration (EC) features, comprises of two main processes, depicted in Figure 2. The first process is to estimate a smoothed spectrum of the VT system response and by using the *Pseudo*-cepstrum analysis [11] a suitable window length is obtained for a lifter, low-time filter, to properly capture the low-time section of cepstrum that contributes the spectral structure information of the vocal tract. The second process is to fit a GMM to the magnitude spectrum via the iterative EM algorithm and to extract acoustic features in terms of GMM estimates (i.e., center frequency (CF), bandwidth (BW), weight coefficient (WC)).

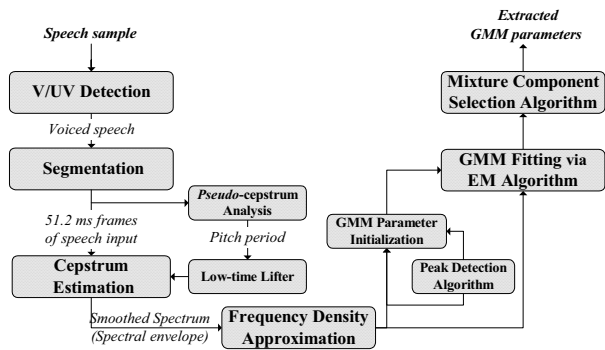


Figure 2: Procedure for extracting the spectral-based vocal tract feature via a GMM.

3.3. Dimensionality Reduction of Feature Space

There are a number of methods from the pattern recognition literatures for reducing the dimensionality of a feature space. Several of these have been used in the speech and speaker identification and recognition with good results. These methods can be grouped into two categories: feature selection method and feature extraction method. The first method reduces a dimensional feature space by selecting a subset of the original feature set. The second method reduces the dimensionality by projecting the original D -dimensional feature space on a d -dimensional subspace (where $d < D$) through a transformation.

In the feature selection method, a feature with its ability to distinguish between two classes depends on both the distance between the two classes and the amount of scatter within the classes. A reasonable measure of class discrimination must take into account both the mean and variance of the classes. One such measure of separability between two classes is the *Fisher's* discriminant ratio [12]. In equation (4) the higher discrimination is measured when the class means are farther apart and when the spread of the classes is smaller, thereby increasing separation between two classes. This measure is defined as

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \quad (4)$$

where μ_1 and μ_2 are the two means or centroids of the classes and σ_1 and σ_2 are the standard deviations of the classes.

The extension of *Fisher's* discriminant ratio which provides more ability to measure separation between multiple classes is the *F*-ratio. It is a measure used to evaluate the effectiveness of a particular feature and it has been widely used as a figure of merit for feature selection in speaker recognition application [13]. It is defined as a ratio of the between-class variance and the within-class variance. This method tries to select the feature

that maximizes the separation between different classes and minimizes the scatter within these classes. The following assumptions have to be satisfied when the *F*-ratio is employed for reducing a dimensional space of feature: 1) The feature vector within each class must have the Gaussian distribution; 2) the features should be statistically uncorrelated; and 3) the variances within each class must be equal. Since the variances within each class are generally not equal, the pooled within-class variance is used to define the *F*-ratio. The number of training feature vectors, training pattern, in the j th class of the K classes is assumed to be the same (N_j). Thus, the *F*-ratio of the i th feature can be defined as:

$$F_i = \frac{B_i}{W_i} \quad (5)$$

where B_i is the between-class variance and W_i is the pooled within-class variance of the i th feature.

3.4. Performance Evaluation

We first performed multiple runs of speech recognition on a training sample set using a subset of features selected from the original $D = 16$ features. For each run, a new subset of the $d (< D)$ features with the rank-ordered *F*-ratios was selected. The recognition accuracy as a function of size d of the reduced feature set was determined and plotted for searching for the reduced feature set that is optimal. It was then employed in the L -fold cross validation for evaluating recognition performance. In this study, we used a quadratic classifier to perform twelve repetitions of cross validation on two different sets of the randomized samples: a 75% of original samples as a training set and 25% of samples as a testing set for simulating and validating the classifier's performance. In order to observe the distinguishing power of the optimal feature set, other statistical measures were calculated such as Sensitivity (SE), Specificity (SP), Positive Predictive Value (PPV), and Negative Predictive Value (NPV).

4. Experimental Results

Figure 3 (a) presents the fitting between a GMM and a magnitude spectrum for a single frame of a female speech. Figure 3 (b) demonstrates a mixture of four individual Gaussian distributions selected into account by weighting on individual mixture probabilities well superimposed on the magnitude spectrum of the VT system response. Figure 4 illustrates the *F*-ratio measures of all studied features. The rank-ordered *F*-ratios demonstrate that the PSD_1 , PSD_2 , CF_2 and BW_2 features extracted from the text-independent speech samples appear to be the most distinguishing feature set with a high individual separation measure. The *F*-ratios measured for the same set of optimal features as described above were also found to be highest for the text-dependent speech samples, except for BW_2 with its power of group separation that diminishes. Figure 5 (a) presents the recognition accuracies as a function of size d of the reduced feature set. It can evidently be noticed that the dimensional space for feature sets can be reduced down to four without affecting the recognition performance.

Table 1 summarizes the average recognition accuracies and performance measures determined from two different types of speech samples. As a result of classification for the text-dependent speech samples, an average 90.33% recognition

accuracy was obtained by a classifier using PSD₁, PSD₂ and CF₂ as discriminating features. By using four primary features extracted from the text-independent speech samples, the classifier performed slightly worse about a 5% decrease in accuracy when compared to that of the text-dependent speech samples with less primary features used in recognition.

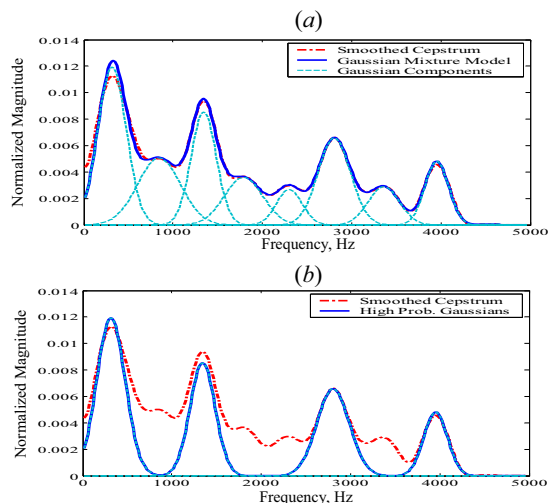


Figure 3: Plots of a mixture of Gaussians: (a) superimposed on the magnitude spectrum and (b) selected with a high probability.

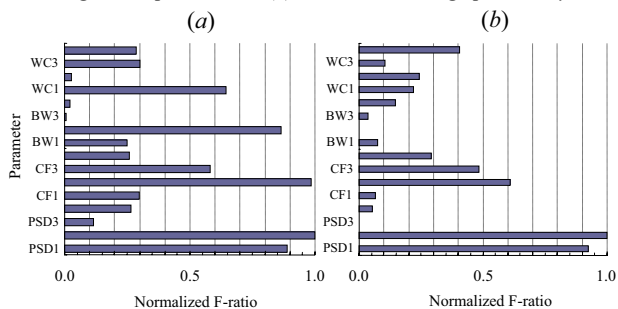


Figure 4: Normalized F -ratio plots of 16 acoustic parameters obtained from speech samples collected during: (a) the interview session study and (b) the text-reading session study

Table 1: Means of recognition accuracy and performance

Session	%Recognition	SE	SP	PPV	NPV
Interview	85.75	0.89	0.80	0.89	0.88
Text-reading	90.33	0.89	0.92	0.94	0.84

The measures of classification performance, SE (0.89), SP (0.92) and PPV (0.94) were all determined to be more effective for distinguishing the text-dependent speech samples collected from diagnostic patient groups, except for NPV (0.84) when compared to that of text-independent speech samples. These high performance measures implied that the studied acoustic feature set was distinctively powerful to be used as vocal discriminators in classifying suicidal speech and depressed speech. Due to a small sample size used in this work, the statistical power in analyses was probably decreased and the wide confidence interval for estimates was possibly introduced as well. On a larger size of samples, the results of recognition performance would probably get improved based on using the

same discriminating features in performance evaluation. By using the proposed acoustic feature extraction technique as an assisting tool for clinicians to diagnose the psychiatric disorders, patients would more correctly be identified and assigned to the exact categories of the severity of mental state with less clinical effort.

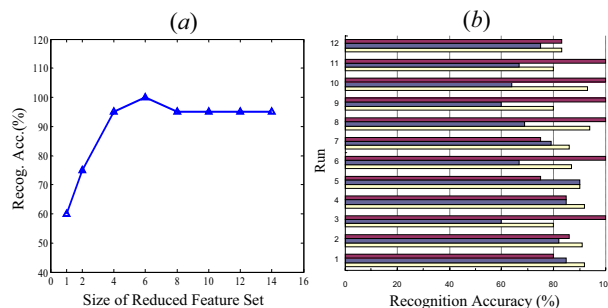


Figure 5: Recognition accuracy results (a) as a function of size d of the reduced feature set and (b) cross validations: white – training data, blue – training data with the “leave-one-out” and red – testing data.

5. Conclusions

Based on using the F -ratio, a problem of dimensional space for features was properly solved. The PSD₁, PSD₂, CF₂ and BW₂ exhibited their acoustic capability of distinguishing patients’ speaking patterns affected by depression and suicidal risk. The comparative results of classification performance demonstrated that the improvement of recognition performance can be enabled by a technique of text-dependent recording for speech samples. Such a technique can be employed for the assessment of suicidal risk in patients as an alternative of audio acquisition.

6. References

- [1] Yingthawornsuk, T., Kaymaz Keskinpala, H., France, D., Wilkes, D.M., Shiavi, R.G., Salomon, R.M., "Objective Estimation of Suicidal Risk using Vocal Output Characteristics", Int. Conf. on Spoken language Processing, 2006, pp 649–652.
- [2] Scherer, K.R., Vocal correlates of emotional arousal and affective disturbance, Handbook of social psychophysiology, Wiley, 1989.
- [3] Scherer, K.R., "Vocal affect expression: A review and a model for future research", Psychological Bulletin, vol.99, pp143–165, 1986.
- [4] France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, D.M., "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Trans. BME. , 47(7), pp829–837, 2000.
- [5] Fairbanks, G., Voice and Articulation Drillbook, New York, 1960.
- [6] Ozdas, A., Shiavi, R.G., Wilkes, D.M., Silverman, M., Silverman, S., "Analysis of Vocal Tract Characteristics for Near-term Suicide Risk Assessment", Meth. Info. Med., vol.43, 2004.
- [7] Moore, E., Et.al, "Comparing objective feature statistics of speech for classifying depression", IEEE Int. Conf. (EMBS), 2004.
- [8] Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm", J. Royal Statistical Society Series B, 39:1–38, 1977.
- [9] Schafer R. and Markel, J., Speech Analysis, IEEE Press, 1979.
- [10] Feder, M., Et.al, "Parameter estimation of superimposed signals using EM algorithm", IEEE Trans. Acous. Spch. Sig. Proc., vol.36, pp 477–489, 1988.
- [11] Lim, J.S., "Spectral Root Homomorphic Deconvolution System", IEEE Trans. Acous. Spch. Sig. Proc., vol.27, pp 223–233, 1979.
- [12] Parsons, T., Voice and Speech Processing, McGraw-Hill, 1987.
- [13] Pruzansky, S., "Talker recognition procedure based on analysis of variance", J. Acous. Soc. Am., vol. 36, pp 2041–2047, 1964.