ภาคผนวก ข.

ผลงานทางวิชาการที่ได้รับการเผยแพร่


ICROS
Ministry of Knowledge Economy
IEEE
ICCAS 2011
2011 11th International Conference on Control, Automation and Systems
PROCEEDINGS
October 27-29, 2011
KINTEX, Gyeonggi-do, Korea
IEEE Catalog Number: CFP1110D-CDR
ISBN: 978-89-93215-03-8 98560
ISSN: 2093-7121
In conjunction with KRC, Robot World 2011
KINTEX
Welcome Messages
Conference Organization
Reviewers
Conference Information
Plenary Lectures
Table of Contents
Author Index
Financial Contribution
E-proceeding Search
Exit

WEP : Interactive Session I (15:00~16:00)

Assessment of Vocal Correlates of Clinical Depression in Female Subjects with Probabilistic Mixture Modeling of Speech Cepstrum

Terapong Boonla[1] and Thaweesak Yingthawornsuk[1]
[1] Department of Electrical Technology Education,
King Mongkut's University of Technology Thonburi, Thailand

Abstract: The acoustical properties of speech have been reported to relate to the mental state of speaker while speaking. This proposed work describes way to address the issue of distinguishing between female depressed patients and female remitted subjects based on the measurable change in the cepstral parameters extracted from their sound record. The cepstral coefficients corresponding to the filter response characteristics, affectively mediated by the emotionally depressive illness or even in particular case of the elevated suicidal risk into the speech production system of depressed speaker, are analyzed via the speech cepstral estimation in conjunction with the GMM fitting approximation. The results of pairwise classification in combinations with SVM, cross-validation, training and testing the cepstral coefficients provide the fairly high accuracy in class separation, when evaluating the testing datasets of coefficients extracted from speech segmentations which are highly corresponding to individual female speakers.

Keywords: Clinical Depression, Automatic Speech, Vocal Filter, Cepstral Estimation, Cross-validation

1. INTRODUCTION

Clinical depression is the psychiatric disorder which can lead to the risk of suicide in person who has experienced it recursively without taking the treatment. This type of emotional illness has been popularly studied and reported to be the prominent precursor of the suicidal risk in human and suicide is the public health problem with increasing rate every year, and has an obvious impact on social, healthcare, life living and even economic growth. Therefore, preventing suicide is the most important task and has to be proceeded in earlier time by screening persons for depression and admitting that depressed person who may or may not have severe symptom of illness to the treatment program. The major objective of this study is to attempt to investigate the relation between the acoustical properties of speech and the severity of mental states in speakers who were clinically diagnosed for depression or suicidal risk by psychiatrists.

The formerly experimental studies have been proposed that the acoustical parameters estimated from speech signal can be used in observing the affection on recognizing pattern and assessing the degree of mental severity in depressive speakers. Clinical procedures to identify person who is severely depressed are in great need for practical use in healthcare centre and clinics. The most common methods to assess, if patients were at severe state of depression or even at elevated risk of suicide, are self-scored patient survey, report by others, clinical interviews and rating scales, such as the Hamilton depression rating scale [1]. Diagnosis and decision making on clinical categories patient belongs to are clinical procedure with time consuming in which practitioners must involve several steps such as information gathering, background profile checking, hospital admission and visiting records, diagnosing progress, crime related report, and hotline call-in for healthcare consultation with psychiatrist. The clinical diagnosis with simultaneous response in judging if patient were psychologically safe from suicidal risk or clinically identified for one of symptom categories, dramatically necessitates for physician to conclude the diagnosing result with the correct decision making on admission and treatment to that patient. As reported in the published studies [2-5], several analytical techniques have been proposed to achieve the way to measure the particular changes, as a result of affection from the underlying symptom of depression, in acoustics of speech of depressed patients. It has been concluded that the suicidal speech in severely depressed speaker is very similar to that of common depressive one, but the tonal quality of speech significantly changes when the symptom of near-term suicidal risk highly strikes at the moment.

Acoustical parameters extracted by several speech processing techniques, such as statistical properties of fundamental frequency, speech jitters and shimmers, formants, pitch contour, Glottal spectral tilt, and frequency distribution of PSD energies have suggested as the effective indicators in monitoring the symptom of major depression in speakers [2-3,6-7] through their sound outcomes. Another effectively discriminative parameter group is the Mel-Scale Frequency Cepstral Coefficients (MFCC) used as input to GMM classifier in attempt to identify the suicidal patients among depressed patients and out of the control subjects, previously proposed by Ozdas et. al. Result has shown the highly significant different measures in MFCC coefficients among three groups of subjects and the percentages of correct classification was considered fairly high [3]. All these research results have revealed for the possibility in use the speech parameters to determine the severity levels of illness, which represent as the vocal correlate of the speaker's emotional and mental state. The aim of our work is to determine the optimal speech parameters with the high class discriminative power and then employ them as the indicator for monitoring the symptom in speaker. In this work, an alternative way to estimate the vocal-tract parameters is explained and implemented, rather than employing the conventional method with

Linear Prediction Coding which is a model-based approach and could provide an inaccurate measure of spectral structure for formant estimates. Therefore, the probabilistic representation of the cepstral structure of the vocal tract filter response is focused and estimated from the female speech database, which is organized into comparative studies by classification validation between depressed patient group and another group of subjects with clinical diagnosis of remission (patients under clinically approval as recovered from previously being depressed) with two different feature sample models in performance evaluation, which are the testing data with frame-based feature model and another with subject-based model. In addition, the specific type of acoustically controlled audio samples collected from the post-interviews between patients and clinician in which patients read a prepared text passage. Analyzed results of classifying such speech sample cases will provide us more suggestion on the speech acoustics affecting the class separation corresponding to ways that subjects producing speech in addition to investigation of the optimal speech parameters.

2. DATABASE

The database consists of female speech samples recorded from the post-interview session in which the patient read the text-contents called “Rainbow passage". In this work any speech signal collected from this post session of interview is called “automatic speech”. The passage contains all normal sounds in spoken English with balanced phonetics [8]. All patients in this study completed this recording procedure. The categorized groups of depressed and remitted patients comprise of eighteen and fifteen females, respectively. Each subject must complete the Beck Depression Inventory-II, (BDI-II), while being interviewed by physician. BDI-II inventory is for mood measure which is recognized as a standard, brief and self-score questionnaire regarding of mental as well as physical depression related to symptom. Total 21 questions can provide numerical scores ranging from 0 to 64 for the patients' responses, where the higher scores relate to more suicidal risk [9]. The preprocessing was carried out by first digitizing all speech signals through a 16-bit analog to digital converter at a 10-KHz sampling rate and a 5-KHz anti-aliasing low-pass filter. All preprocessed speech samples were then screened over to eliminate other voices or any sound artifacts rather than the patients' only voice, as well as silences longer than set time thresholds via Audio sound editor. All edited speech samples were then stored for further analysis.

3. METHODOLOGY

3.1 Vocal Feature Extraction

Preprocessed speech samples are first detected and classified into three groups of voiced, unvoiced and silence due to energy estimated and compared to the energy-level thresholds. Only the voiced segments in speech are therefore statistically normalized to the adjustment of suitable amplitude to the group baseline for all voiced segments in database. Then four of the most dominant Gaussians associated with high probabilities from applying the GMM fitting to the individual 51.2-ms frame-based estimate of cepstral structure of voiced samples are estimated and selected in account. The following procedures are employed in our feature extraction:

- Segment the voiced speech signal into 512-sample frames
- Compute the Log-scale Cepstrum for each fram eof voiced speech signal
- Lifter the low-time section of estimated cepstrum within the detected pitch period
- Convert the estimate of low-time cepstral section to the probability density function (pdf) normalized over 0-5KHz frequency range
- Estimate the ML of GMM distribution via EM algorithm with the modification of fast iterative convergence in pdf fitting [10]
- Calculate mean, variance and weighting probability of the selected estimated GMM model corresponding to the high ratio of mixture weight and standard deviation
- Calculate means and SDs for every 200, 300 and 400 samples/frame which are used to represent as input feature vector to classifier
- Calculate F-ratio values for all means and SDs, in order to form a combination of input features, which has the increasing correct classification scores, when evaluating the features as ordered ranking by their F-ratio values.

3.2 Feature Classification

The Support Vector Machine (SVM) [11-12], which outperforms most other classification systems in a wide variety of applications, has been used in this study for performance validation. It achieves relatively robust pattern recognition performance using well established concepts in optimization theory. SVM separates an input $\boldsymbol{x} \in \boldsymbol{R}^d$ into two classes. A decision function of SVM separates two classes by $f(\boldsymbol{x}) > 0$ or $f(\boldsymbol{x}) < 0$. The training data which is used in training phase is $\{\boldsymbol{x}_i, y_i\}$, for $i = 1, \ldots, l$ where $\boldsymbol{x}_i \in \boldsymbol{R}^d$ is the input pattern for the *i*-th sample and $y_i \in \{-1, +1\}$ is the class label. Support Vector Classifier maps $\boldsymbol{x}_i$ into some new space of higher dimensionality which depends on a nonlinear function $\phi(\boldsymbol{x})$ and looks for a hyperplane in that new space. The separating hyperplane is optimized by maximization of the margin. Therefore, SVM can be solved as the following quadratic programming problem,

$$\max_{\alpha_i} \{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j \, y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \} \tag{1}$$

Subject to $0 \le \alpha_i \le C$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$

where C is a parameter to be chosen by user, a larger C corresponding to assigning a higher penalty to errors, and $\alpha \ge 0$ are Lagrange multipliers. When the optimization problem has solved, system provides many $\alpha_i > 0$ which are the required support vector. Note that the Kernel function $K(x_i, x_j) = \phi^T(x_i)\phi(x_j)$ where $\phi(\cdot)$ is a non linear operator mapping input vector $x \in R^d$ to a higher dimensional space. In this work, we choose the polynomial kernel $K(x_i, x_j) =< x_i, x_j >^d$ as the kernel function, where $d \in N$. In addition, other kernels can also be applied. Classification consists of two steps: training and testing. In the training phase, SVM receives some feature patterns as input. These patterns are the extracted speech features represented by N feature parameters that can be seen as points in N-dimensional space. In this study twelve features extracted from individual voiced frames are used to form the input feature matrix which is multi-dimensional. Then the classifying machine becomes able to find the labels of new vectors by comparing them with those used in the training phase.

In training state the 50% of all speech samples was randomly selected for the input feature model to SVM. Then, the rest 50% of randomized samples was used in validating state. For every feature model set, the cross validations have been completed for approximately 100 times. Each feature model as input feature set to SVM is formed by adding one more ranked feature at a time, when training and testing a new feature model in cross-validation. With similar procedure in validation, all speech sample sets of depressed and remitted speech databases are pairwisely classified under the similar conditions of feature models and number of random samples for all three different segmentations of 200, 300, and 400 samples per frame.

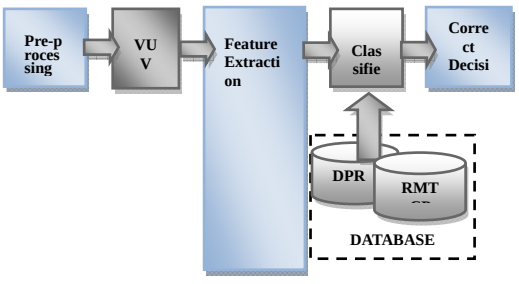

Fig. 1 Workflow for speech sample classification.

4. RESULTS AND DISCUSSION

The extracted means, variances and probabilities obtained from fitting the Gaussian mixture model to the vocal-tract cepstra are used in both training and testing states. Feature samples are first tested for their sample distribution between two speech classes clearly notified from the histograms plotted as comparison. Figure 2 shows histograms of the originally extracted mixture probability associated with the selected dominant Gaussians clearly seen for the significantly observable separation between two speech classes. In addition, the discriminant scores calculated from distributions between classes are provided as well. As a result of data transformation, the significant difference in class means is obviously determined for the separating distance, which implies us for the possibility of obtaining the effective classification between classes. Figure 3 indicates the staircase-shaped histograms in term of discriminant scores followed by employing the basis of Fisher's discriminant function with pooled covariance between class covariance matrices [13]. In order to have less complexity in calculation of discriminant function and also have reasonably high efficiency across a wide variety of feature population models which is a suitable assumption for the studied data, therefore, a single unbiased estimate of population covariance is determined for sample squared distance formulated in (2). As mentioned before in the section of feature extraction, all speech samples were processed on the basis of 51.2ms-length speech frames. Therefore, the distance scores depicted in figure 3 were calculated from all speech frames segmented from speech signal database in both categorized speech classes with blind identifying for the numbers of sample frames, which are used to represent individual subjects. Distributions of the weighting probabilities extracted from two speech classes are plotted in discriminant scores as measure of significantly different means. Sample squared distance ($\mathbf{D^2}$) between two sample means can be followed by

$$\mathbf{D^2} = (\bar{\mathbf{X}}_{\mathbf{RMT}} - \bar{\mathbf{X}}_{\mathbf{DPR}})' \boldsymbol{S^{-1}} (\bar{\mathbf{X}}_{\mathbf{RMT}} - \bar{\mathbf{X}}_{\mathbf{DPR}}) \quad (2)$$

where $\bar{\mathbf{X}}_{\mathbf{DPR}}$ is a sample mean calculated from the vocal cepstral extracts of depressed speech population, $\bar{\mathbf{X}}_{\mathbf{RMT}}$ is a sample mean of the remitted population, and $\boldsymbol{S}$ represents a sample pooled covariance matrix combined between two covariance matrices calculated from both cepstral extract populations.

As plotted in comparative performances validated by SVM all evaluated performances are increasing for all three speech segmentation cases against numbers of increasing features added in cross-validation when the subject-based samples are classified. We found that total number of nine sorted features with f-ratio ranking makes classifier's performance gradually degraded. Results from classifying samples with different frame lengths (plotted in figure 5) show that about six cepstral features in combination can differentiate the SVM performances among variety of frame lengths, but the tendency of more similarity in performance scores can be noticed at the higher combinations of features. In case of classifying the frame-based samples, the results seem to be no consistent and less significantly different among all three different segmentations. For more features formed in model combinations in case of classifying speech samples with frame length of 200 samples, the performance indicates to be the highest median score of 78% when the subject-based testing on automatic speech samples is performed with eight features in combination as input to SVM. More comparative

studies on automatic speech data set revealed that, when we used either the frame-based samples or the subject-based samples (subjects represented by an average of all frames of feature collected from that subject) in testing state of SVM cross-validation, the increasing and outperforming performances are obviously observed for classifying speech samples with 200 samples/frame compared with other cases of different sample/frame ratios at the higher number of combining features.

Figures 6 and 7 illustrate the tendencies of mean and standard error obtained from testing the classifier with three different sample segmentations against the increasing numbers of combining features in classification. Performances resulted from classifying mean and SD parameters are represented in black dot, red dash, and blue solid lines when applying 400, 300 and 200 samples/frame, respectively, in estimating the input means and SDs for classification. In plots the small standard errors (S.E.) determined when averaging all correct classification score. The smaller number can imply the overall classification results in the sense of higher reliable statistic regarding of the confidence interval of the classification performances. All these results eventually take us to the conclusion that the vocal-tract response characteristics mediated by the affection of depression represented in cepstral coefficients can be assessed and used to monitor the symptom through vocal outcome based on the statistical consistency revealed on analyzed results as formerly discussed.

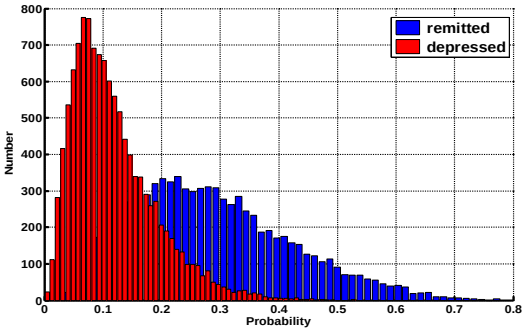

Fig. 2 Probability distribution between two classes.

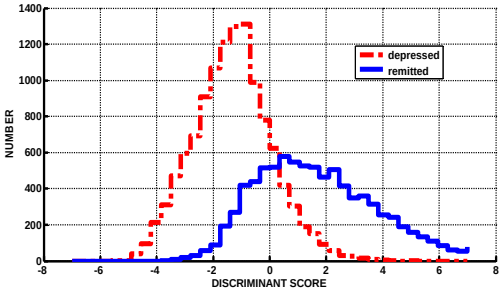

Fig. 3 Discriminant scores between remitted (in blue solid) and depressed speech samples (in red dash).

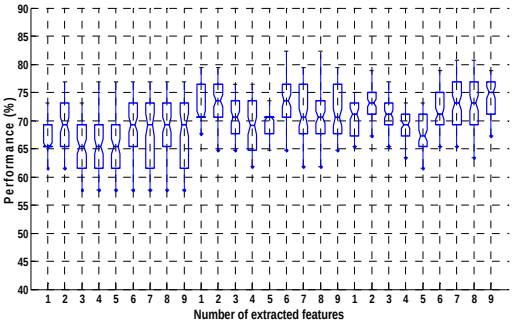

Fig. 4 Results obtained from classifying the frame-based speech samples with 400, 300 and 200 samples/frame.

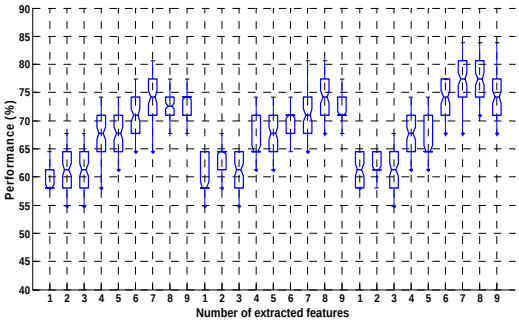

Fig. 5 Results from classifying the subject-based speech samples with 400, 300 and 200 samples/frame.

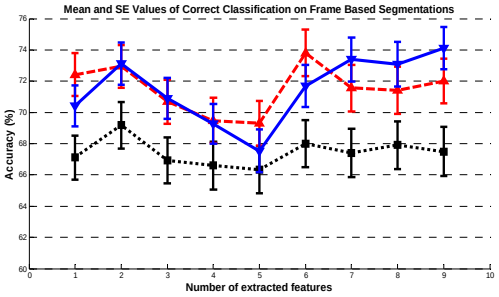

Fig. 6 Comparative accuracies in mean and s.e. values from classifying the frame-based segmentations with lengths of 400 (in dot), 300 (in dash) and 200 (in solid).

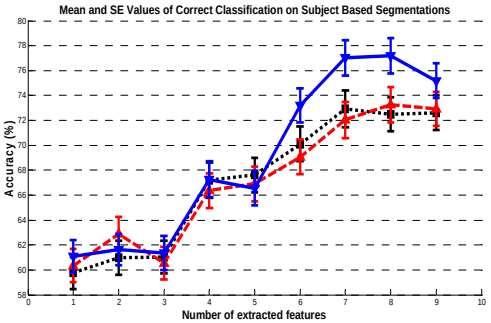

Fig. 7 Comparative accuracies in mean and s.e. values from classifying the subject-based segmentations with lengths of 400 (dot), 300 (dash) and 200 (solid).

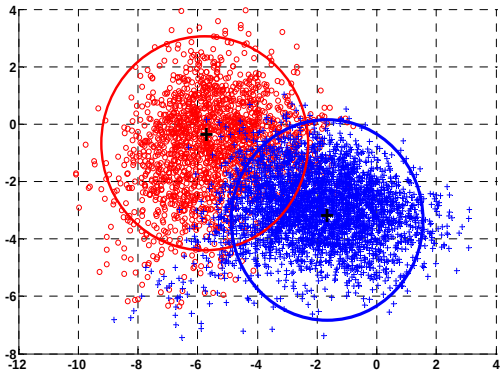

Fig. 8 Scatter plot of vocal extracts between remitted (in blue +) and depressed speech samples (in red o).

5. CONCLUSION

This paper demonstrates that the probabilistic Gaussian Mixture based vocal parameters representing characteristics of filter/vocal tract extracted from the speech samples whose acoustical properties have been changed due to depression, in terms of mean, variance and mixture probability achieve for effective features in combination capable of classifying two groups of females, one clinically diagnosed patients being depressed and patients recovered from depressed based on the vocal outcomes extracted from their speech measures. Difference in classifying performance may suggest us that speakers have different relative articulations collaborating between speech production and nervous system acting alternatively when speaking. Degradation in the classifier's performance can be identified clearly through the measure of F-ratios calculated on speech extracts. This helps organize all speech features in statistically ordered-ranking based on their class discriminating power measured and reduce the feature model dimension in validation. By concerning with the size of sample, it possibly diverts the statistical interpretation on analyzed results when the sample size of data is not adequately large to present a population of data. The experimental error may be intrusive in data interpretation on the obtained results. Results from this empirical study imply that there is an existence of the impairment in the pathway of speech production resulted from affection. Further quantifiable study could provide us more insightfully understanding about the pathophysiological impact on speech production system.

REFERENCES

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

[2] France, D.J., et al., "Acoustical properties of speech as indicators of depression and suicide", *IEEE transactions on BME*, 2000. 47:p 829-837.

[3] Ozdas, A., et al., "Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment", Meth.Info.in Medicine, 2004. 43: p. 36-38.

[4] 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 BME*, 2004. 51: p. 1530-1540.

[5] 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.

[6] T. Yingthawornsuk, et al., "Direct Acuostic Feature using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Risk", *Interspeech 2007*, Antwerp, Belguim.

[7] 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.

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

[9] A.T. Beck, st al., "An inventory for measuring depression", *Arch. Gen. Psychiatry*, 1961. 4:p. 561-571.

[10] Dempster, A.P., et al., "Maximum likelihood from incomplete data via the EM algorithm", J. *Royal Stat. Soc. Series B*, 39:1–38, 1977.

[11] V.N. Vapnik, *The Natural of Statistical Learning Theory.* 2nd ed., Springer Verlag (New York), Dec 1999.

[12] C. Cortes and V.N. Vapnik , “Support vector networks”, *Machine Learning*, vol.20, pp. 1-25, 1995.

[13] A.J. Richard, *Applied Multivariate Statistical Analysis.* 3th ed., Prentice Hall, New Jersey, 1992.