EMOTION RECOGNITION FROM SPEECH SIGNALS COMBINING PCA AND LDA

$MingyuYou^1$, $ChunChen^1$, $JiajunBu^1$, $JiaLiu^1$, $JianhuaTao^2$

{roseyoumy, chenc, bjj, liujia}@zju.edu.cn, jhtao@nlpr.ia.ac.cn
[1] College of Computer Science, YuQuan Campus, ZheJiang University, Hangzhou, CHINA, 310027
[2] National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, CHINA, 100080

ABSTRACT

Dimensionality reduction is an important issue in pattern recognition. Two popular methods used in this field are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). In this paper, detailed comparisons are performed among PCA, LDA and PCA+LDA considering the lack of similar studies. It will show that no particular method is optimal across all emotion categories. Based on this analysis, a new framework combining PCA and LDA is proposed. An appropriate dimensionality reduction method is employed for every emotion category in the new framework. Experiment results show that our approach achieves a better overall performance compared with PCA, LDA or PCA+LDA.

1. INTRODUCTION

Recognizing human emotions by computers has been an active research area. Accurate detection of emotions from speech signals will benefit the design of more natural human-machine interface. It has broadly potential applications in areas such as education, consumer service and entertainment.

The general process of emotion recognition from speech signals can be formulated as follows: extracting acoustic features like those low-level features[1], reducing feature dimensionality to an appropriate range for less computational complexity and recognizing emotions with k-nearest neighborhood (KNN), support vector machine (SVM)[2] or other classifiers.

Dimensionality reduction methods can be grouped into two categories: feature selection methods and feature extraction methods. The feature selection methods select the features by devising a figure of merit that reflects the goodness of an individual feature in the recognition task. The F-ratio (ratio of between-class and within-class variances) is often used in the feature selection methods[1]. In this paper, we mainly focus on the feature extraction methods. Feature extraction methods reduce the dimensionality by projecting the original feature space into a smaller subspace through a transformation. Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are two major methods used to extract new features in different areas.

Zhao[3] and Beveridge[4] employed PCA and LDA to reduce the feature dimensionality for face recognition. Delac[5] even made a detailed comparative study of these statistical methods in face recognition. PCA and LDA were also utilized in speech recognition area[6]. Some of the studies in speech emotion recognition adopted PCA to analyze the feature sets[2][7], but few of them referred to LDA. Actually, LDA performs better than PCA in many applications. Besides, less work has been done on the performance comparison between PCA and LDA for emotion recognition. In this paper, we compare these two methods in completely equal working conditions and propose a new framework for emotion recognition from speech signals which takes advantage of both PCA and LDA.

The remainder of the paper is organized as follows. First, we give a brief description for PCA and LDA in section 2. Then a detailed performance comparison between PCA and LDA is presented in section 3. Section 4 focuses on a new framework combining PCA and LDA. Conclusion is given in section 5.

2. PCA AND LDA

PCA tends to find a t-dimensional subspace whose basis vectors correspond to the maximum variance direction in the original s-dimensional space. This new subspace is normally lower dimensional $(t \ll s)$ and its basis vectors are named principal components of original data set. These t principal components can be found by computing the covariance matrix of the data set and then finding t eigenvectors corresponding to the largest t eigenvalues. PCA is achieved by projecting original data set into the t-dimensional subspace with projection matrix W_{PCA}.

LDA finds the subspace to best discriminate among classes. The between-class scatter matrix S_B and the within-class scatter matrix S_W are defined by:

$$S_B = \sum_{i=1}^{c} M_i (x_i - \mu)(x_i - \mu)^T$$

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T$$

where M_i is the number of samples in class i, c is the number of distinct classes, μ_i is the mean vector of samples belonging to class i and X_i represents the set of samples belonging to class i with x_k being the kth sample of that class. LDA is to maximize S_B while minimizing S_W, in other words, maximize the ratio S_B / S_W. This ratio is maximized when the column vectors of the projection matrix (W_{LDA}) are the eigenvectors of $S_W^{-1}S_B$. Data set is projected into lower-dimensional subspace using matrix W_{LDA} at the final step. Algorithm described above is the main idea of classical Fisher LDA which has two limitations: (1) dimensionality of the samples is constrained by the number of classes; (2) the basis vectors are not guaranteed to be orthogonal. Duchene and Leclercq[8] gave a direct analytic solution for calculating the optimal set of orthogonal discriminant vectors, which can be as many as the dimension of the original feature space.

3. PERFORMANCE COMPARISON

3.1. Speech Data Corpus

The speech data used in the experiment are obtained from Chinese Academy of Sciences. It's an emotional speech corpus in Chinese Mandarin. The corpus is collected from four Chinese native speakers including two men and two women. Everyone expresses 300 sentences in six emotions involving angry, fear, happy, sad, surprise and neutral. So the total amount of sentences is 300*6*4=7200. The speech corpus is sampled at 16kHZ frequency and 16 bits resolution with monophonic Windows PCM format.

3.2. Acoustic Features

In this study, we extract 48 prosodic and 16 formant frequency features. Prosody is mainly related to the rhythmic aspects of speech, and believed to be the primary indicator of speakers' emotion state. The extracted prosodic features includes: *max, min, mean, median of Pitch (Energy); mean, median of Pitch (Energy) rising/ falling slopes; max, mean, median duration of Pitch (Energy) rising/ falling slopes; mean, median of Pitch (Energy) plateaux at maxima/ minima; max, mean, median duration of Pitch (Energy) plateaux at maxima/ minima*. Here, if the first derivative is approximately zero and the second derivative is positive, the point belongs to a plateau at a local minimum. If the second derivative is negative, it belongs to a plateau at a local maximum. We also investigate formant frequency features which are widely used in speech processing applications. Statistical properties including *max, min, mean, median of the first, second, third, and fourth formant* are extracted.

3.3. Dimensionality Reduction

At speech processing step, 64 acoustic features are extracted for every sentence. Three methods are employed to reduce the feature dimensionality: PCA, LDA and PCA+LDA. In the experiment, speaker independent emotion recognition is investigated within the same gender and 10-fold cross-validation method is adopted. So, 3240(90%*3600) 64-dimensional vectors are used to train PCA. We use a 3240 * 64 matrix X to represent these vectors. After X is normalized and mean-subtracted, we get matrix Y. Y^TY forms a covariance matrix M which is 64*64. Eigenvalues and eigenvectors are computed for M. Eigenvectors corresponding to the largest t eigenvalues are selected to create the PCA projection matrix W_{PCA}. t is the number of eigenvalues which guarantee energy E is greater than 0.9. Here energy E is defined by:

$$E_t = \sum_{j=1}^{t} \lambda_j / \sum_{j=1}^{64} \lambda_j$$

where λ_j is the jth eigenvalue. In our experiment, t equals to 27 in PCA of male model and 29 in female model. 3240 training data sentences and 360 testing sentences are both projected into subspace using W_{PCA}.

While in LDA, algorithm introduced in [8] is adopted to compute the discriminant vectors of matrix Y. In order to compare with PCA in an equal situation, the same number of eigenvectors as W_{PCA} are collected in W_{LDA}. So the dimensionality of W_{LDA} is 27 in male model and 29 in female model.

In the method of PCA+LDA, we obtain a combining projection which maps the pre-processed matrix Y into subspace S_1 first, and then into the subspace S_2. The first projection is a PCA transformation and the second is an LDA action. We don't compress data in PCA, in other words, we keep all the eigenvectors into W_{PCA}. Instead, we reduce feature dimensionality to 27 for male (or 29 for female) in LDA projection.

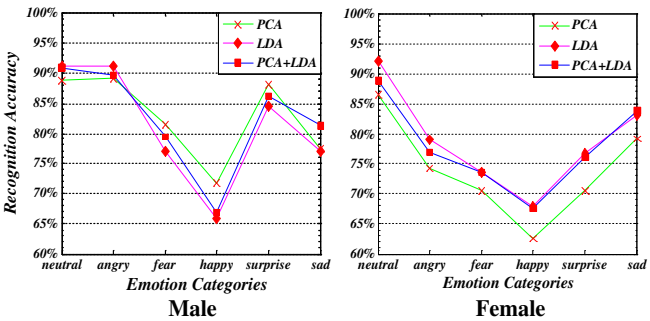

Fig. 1. Recognition Accuracy Distribution of Male and Female

3.4. Classification by Support Vector Machine

SVM(Support Vector Machine)[9], a powerful tool for classification developed by Vapnik, is introduced to classify six

emotions in our experiment. It has originally been proposed for two-class classification and is formed as solving the quadratic programming problem. In our approach, SVM using RBF(Radial Basis Function) kernels below is adopted to be the component of the classifier.

$$R_i(P) = exp[-\frac{\|P - C_i\|^2}{\sigma_i^2}]$$

Then 15 (C_6^2) one-to-one SVMs are formed into an MSVM (Multi-SVM) system in which every SVM is established to distinguish one emotion from another. Final classification result is determined by all the SVMs with majority rule.

In order to evaluate the classification results using PCA, LDA and PCA+LDA, 10-fold cross-validation method which is shown to be the best choice to get an accurate estimate is employed. Figure 1 shows the recognition accuracy of six emotions with PCA, LDA and PCA+LDA. The best recognition rate of 91.7% is obtained for male and also 92% is achieved for female.

As we can see from Figure 1(a), for male model, PCA, LDA and PCA+LDA have their advantages respectively at different emotions. For neutral and angry, LDA achieves the best recognition results. However, PCA is the winner at emotion fear, happy and surprise. At emotion sad, PCA+LDA outperforms the other two methods by 5 percents. Female model in figure 1(b) presents different phenomena from male model. We can figure out that LDA does an excellent job at almost all of the emotion categories in female model.

4. A NEW APPROACH

Based on the performance comparison stated in section 3, a new framework (in Figure 2) is proposed to achieve the best overall performance. For an input speech signal, 64 acoustic features are extracted after signal pre-processing. Speech corpus of male is separated from those of female based on pitch analysis. Feature vectors of female speech corpus are directly projected into 29-dimensional sub-space using LDA because of the outstanding behavior of LDA at all emotions in female model.

However, the question is which dimensionality reduction method we should choose for male model. Every method has its advantage at some of the emotions. Our solution is motivated by achieving the best overall performance. We divide speech corpus into several subsets and then use the data compression method suitable to each subset. Analyzing the comparison results above, neutral and angry are selected to form subset A which prefers LDA method. Under the same principle, fear, happy and surprise should be grouped into a subset and sad would be another subset alone. However, as we know, unbalanced training data is a serious problem for SVM. With this consideration, subset B will just include happy and surprise and employ PCA as its data compression method. Emotion fear together with sad form subset C which

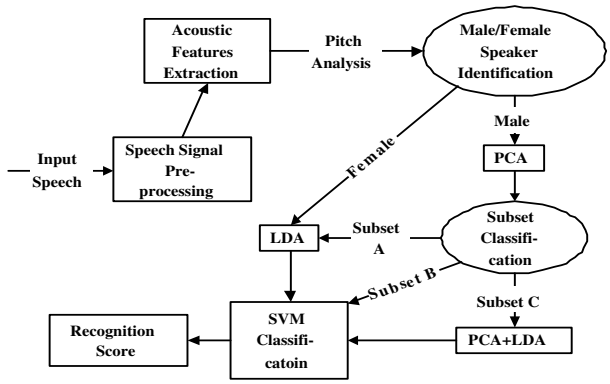

Fig. 2. An Integrated Framework for Emotion Recognition of Speech Signals

Table 1. Confusion Matrix for Classifying Subsets A, B and C

Subset	A	B	C
A	0.898	0.085	0.017
B	0.126	0.869	0.005
C	0.044	0.016	0.940

adopts PCA+LDA. So, in our framework, features of male are firstly projected into 27-dimensional sub-space with PCA. And then these compressed features are classified into three subsets with classifier $MSVM_1$. Table 1 shows the confusion matrix for classification of the three subsets. If the test sentence is classified as subset A, the 64-dimension feature data before PCA will be projected into 27-dimension subspace using LDA. In other words, after three subset classification, we need not save the PCA projection result for subset A again. But it's different for subset B which prefers to be compressed by PCA. So we just simply keep the PCA projection result conducted before three subsets classification for subset B. Processing of subset C is something similar to subset A which will make projection by PCA+LDA. After the dimensionality reduction using appropriate method, $MSVM_2$ is employed to classify the test data into six emotion categories.

We calculate the rate of emotion neutral as an example to show how to get the recognition accuracy of our approach. Neutral belongs to subset A, so from Table 1, 89.8% are classified as A, 8.5% as B and 1.7% as C. When classified as A, LDA is adopted which has 91.2% (can be observed from Figure 1(a)) recognition accuracy for emotion neutral. PCA is used when classified as B and it's recognition rate is 88.8%. There's also 1.7% are judged as subset C which employs PCA + LDA and has 91% recognition accuracy for neutral. Based on the analysis above, the accuracy for emotion neutral in our approach is $0.898 * 0.912 + 0.085 * 0.888 + 0.017 * 0.91 =$

$0.9100(91.0\%)$. Similarly, the recognition accuracy for angry(91.0%), fear(79.4%), happy(70.5%), surprise(87.7%) and sad(81.1%) are computed. From Figure 3, we can find out that our new approach has comparable performances with the best dimensionality reduction method at every emotion category. In other words, our new approach takes the recognition accuracy of all emotion categories into consideration and achieves the best overall performance.

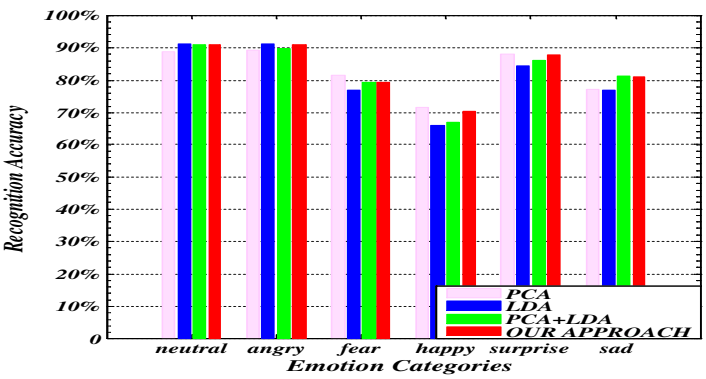

Fig. 3. Performance Comparison between our approach and the other three methods(PCA, LDA and PCA+LDA) in male model

Summarizing the performance of our new framework, the average recognition rate for six emotion categories of male model is 83.4% and that of female model is 78.7%.

5. CONCLUSION AND FUTURE WORK

This paper presents an independent, comparative analysis of three dimensionality reduction algorithms (PCA, LDA and PCA+LDA) used in speech emotion recognition. It is found that none of the three algorithms is the state-of-the-art for all emotion categories. A new approach is introduced to integrate the respective advantage of each method. In the new approach, LDA is adopted as feature extraction method for female speech corpus. Male corpus is classified into three subsets and every subset choose a suitable algorithm for dimensionality reduction from PCA, LDA and PCA+LDA. The overall performance of our framework is better than those of the three algorithms.

As we find from Figure 1 that emotion recognition rate of male model outperforms that of female. In our opinion, the reason might be that female convey more emotional information by facial expression and body gestures than male. So in feature work, we will focus on multi-modal emotion recognition including speech, facial expression and other bio-information.

6. ACKNOWLEDGEMENT

The work is partly supported by National Natural Science Foundation of China (60203013). And We thank Qi Wu, Mingli Song, Cheng jin and Weiguang Wang for their generous help to our experiment and paper.

7. REFERENCES

[1] D. Ververidis, C. Kotropoulos, I. Pitas, ”Automatic emotional speech classification”, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, Pages 593-596, May 2004.

[2] C.M. Lee, S.S. Narayanan, R. Pieraccini, ”Classifying emotions in human-machine spoken dialogs”, in Proc. IEEE International Conference on Multimedia and Expo, Volume 1, Page 737-740, Aug. 2002.

[3] W. Zhao, R. Chellappa, A. Krishnaswamy, ”Discriminant Analysis of Principal Components for Face Recognition”, in Proc. the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, Page 336-341, April 1998.

[4] J.R. Beveridge, K. She, B. Draper, G.H. Givens, ”A Nonparametric Statistical Comparison of Principal Component and Linear Discriminant Subspaces for Face Recognition”, in Proc. IEEE Conference on Computer Vision and Pattern Recognition, Page 535-542, December 2001.

[5] K. Delac, M. Grgic, S. Grgic, ”A Comparative Study of PCA, ICA and LDA”, in Proc. the 5th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services, Page 99-106, June 2005.

[6] X. Wang, K.K. Paliwal, ”Using minimum classification error training in dimensionality reduction ”, in Proc. IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing, Volume 1, Page 338-345, Dec. 2000.

[7] Z.J. Chuang, C.H. Wu, ”Emotion Recognition using Acoustic Features and Textual Content”, in Proc. IEEE International Conference on Multimedia and Expo, Volume 1, Page 53-56, June 2004.

[8] J. Duchene and S. Leclercq, ”An Optimal Transformation for Discriminant Principal Component Analysis”, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.10, No 6, November 1988

[9] N. Cristianini and J. Shawe-Taylor, ”An Introduction To Support Vector Machines”, Cambridge University Press, 2000