EMOTION RECOGNITION FROM NOISY SPEECH

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

{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

This paper presents an emotion recognition system from clean and noisy speech. Geodesic distance was adopted to preserve the intrinsic geometry of emotional speech. Based on the geodesic distance estimation, an enhanced Lipschitz embedding was developed to embed the 64-dimensional acoustic features into a six-dimensional space. In order to avoid the problems brought by noise reduction, emotion recognition from noisy speech was performed directly. Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) and feature selection by Sequential Forward Selection (SFS) with Support Vector Machine (SVM) were also included to compress acoustic features before classifying the emotional states of clean and noisy speech. Experimental results demonstrate that compared with other methods, the proposed system makes approximately 10% improvement. The performance of our system is also robust when speech data is corrupted by increasing noise.

1. INTRODUCTION

Enabling computers to recognize emotions is a main track of research on the human-machine interaction. As a major indicator of human emotions, speech plays an important role in detecting affective states.

The general process of speech emotion recognition can be formulated as below: extracting acoustic features from speech signal, compressing the feature set for less computational complexity and recognizing emotions with SVM, Hidden Markov Model (HMM), Neural Network (NN) or other classifiers.

Feature selection and feature extraction are two categories of methods for compressing data set. Ververidis[1] used SFS method to select five best features for the classification of five emotional states. Lee[2] and Chuang[3] both adopted PCA to analyze the maximum variance of feature set in classifying speech emotion. LDA and Multidimensional Scaling (MDS) were also popular methods of feature extraction in emotion recognition[4].

Most previous work on detecting emotional states investigated speech data which were recorded in quiet environment[5, 6], but humans are able to perceive emotions even in noisy background. Tenenbaum[7] and Roweis[8] proposed nonlinear manifold learning algorithms, trying to discover the mystery how human brain perceives constancy even though its raw sensory inputs are in flux. Different from PCA, LDA and MDS which can only learn linear manifolds, Isomap and Locally Linear Embedding (LLE) developed by Tenenbaum and Roweis attempt to learn nonlinear manifolds. As another nonlinear manifold learning algorithm, Lipschitz embedding[9, 10] works well when there are multiple clusters in the input data. It is suitable for emotion classification whose input data can be grouped into several emotions. Facial images with different poses and lighting directions were observed to make a smooth manifold[7]. Similarly, speech with different emotions, even corrupted by noise, could also be embedded into a low dimensional nonlinear manifold. Actually, several studies[11, 12] confirmed that speech signal points in the state space lay close to a nonlinear manifold of low dimensionality, though they seldom paid attention to the emotional information in speech. Although kernel PCA was used for nonlinear dimensionality reduction, it is not particularly suitable to manifold learning[13].

In this paper, an enhanced Lipschitz embedding is developed to analyze the intrinsic manifold of emotional speech including those recorded in quiet environment and corrupted by noise. Besides, other dimensionality reduction methods such as PCA, LDA and feature selection by SFS with SVM are presented for comparison.

2. EMOTION RECOGNITION SYSTEM ON ENHANCED LIPSCHITZ EMBEDDING

Figure 1 displays the overall structure of our system. Clean speech from the database and speech corrupted by generated noise are both investigated in our system. Firstly, 64- dimensional acoustic features of each utterance are obtained after feature extraction. Then using an enhanced Lipschitz embedding, the six-dimensional nonlinear manifold of emotional speech is gained. Both training and testing data are embedded into the low dimensional manifold. Finally, testing data are recognized by a trained linear SVM system.

A Lipschitz embedding is defined in terms of a set R ($R =$

1-4244-0367-7/06/$20.00 ©2006 IEEE

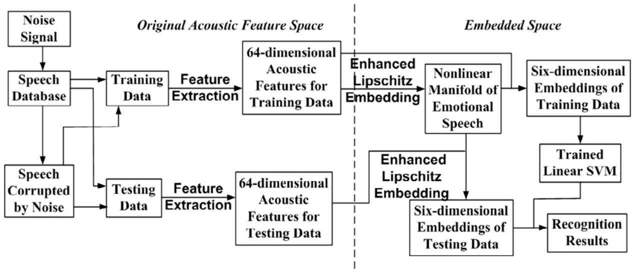

Noise Signal
Original Acoustic Feature Space
Embedded Space
Speech Database
Training Data
Feature Extraction
64-dimensional Acoustic Features for Training Data
Enhanced Lipschitz Embedding
Nonlinear Manifold of Emotional Speech
Six-dimensional Embeddings of Training Data
Speech Corrupted by Noise
Testing Data
Feature Extraction
64-dimensional Acoustic Features for Testing Data
Enhanced Lipschitz Embedding
Trained Linear SVM
Six-dimensional Embeddings of Testing Data
Recognition Results

Fig. 1. System overview

$\{A_1, A_2, \ldots, A_k\}$), where $A_i \subset S$ and $\bigcup_{i=1}^{k} A_i = S$. The subset A_i is termed as the reference set of the embedding. Let $d(o, A)$ be an extension of the distance function d from object o to a subset $A \subset S$, such that $d(o, A) = \min_{x \in A} d(o, x)$. An embedding with respect to R is defined as a mapping F such that $F(o) = (d(o, A_1), d(o, A_2), \ldots, d(o, A_k))$. In other words, Lipschitz embedding defines a coordinate space where each axis corresponds to a subset $A_i \subset S$ and the coordinate values of object o are the distances from o to the closest element in each A_i.

The distance function d in Lipschitz embedding reflects the essential structure of data set. Due to the nonlinear geometry of speech manifold, classical approaches of PCA, LDA and MDS fail to find the real degrees of freedom of the manifold. Tenenbaum et al.[7] sought to preserve the intrinsic geometry of the data by capturing the geodesic distance between all pairs of data points.

In our algorithm, the speech corpus is divided into six subsets $\{A_1, A_2, \ldots, A_6\}$ according to six emotional states (neutral, angry, fear, happy, sad and surprise) which are usually adopted. Object o of speech corpus is embedded into a six-dimensional space where the coordinate values of o are obtained from the process below.

(1) Initiate element m_{ij} in matrix M

$$m_{ij} = \begin{cases} \sqrt{\sum_{\alpha=1}^{64} (x_\alpha - y_\alpha)^2} & : \ \forall i, j \in KNN \\ C & : \ else \end{cases} \tag{1}$$

Here m_{ij} stands for the geodesic distance from point i to j. $i, j \in KNN$ means that j is among the k nearest neighbors of i. In our method, k is set to 10 for simplicity. i and j are data points in the 64-dimensional feature space, $i = [x_1, x_2, \cdots, x_{64}]$ and $j = [y_1, y_2, \cdots, y_{64}]$. C is a very large constant which represents that i and j are unconnected in the graph G consisting of speech data points.

(2) $if\ m_{ij} == C, then\ m_{ij} = \min G_{ij}$ where G_{ij} represents the length of path from i to j in graph G whose edges only connect those k nearest neighbors in (1).

(3) Get the coordinate values of o ($\{o_1, o_2, \cdots, o_6\}$)

$$o_\gamma = \min_{\mu \in A_\gamma} m_{o\mu} \tag{2}$$

where $m_{o\mu}$ is an element of matrix M.

Figure 2 shows six-dimensional embeddings of speech corpus in six emotional states. Figure 2(a) reveals the first three dimensions of embedded space and (b) displays the last three dimensions. Emotion neutral, angry and fear, denoted by points in red, green and blue, are easy to be separated in the first three dimensions. Happy, sad and surprise, denoted by light blue, yellow and pink are separable in the last three dimensions, though they are mixed in Figure 2(a). Actually, points of the same emotional state are located close to one plane in the embedded space. The distribution property of data points in the six-dimensional space indicates that they can be easily classified into six emotions.

Distance matrix M is constructed on the training data from speech corpus, which makes training data be easily embedded into low dimensional space. However, how to embed the new coming testing data into the six-dimensional space is a tough task. It's impossible to reconstruct matrix M combining the testing data. Based on the constructed matrix M, we propose an approach to compute the coordinate values of testing data t in the embedded space.

(1) Based on Euclidean distance, k nearest neighbors ($\{n_1, n_2, \cdots, n_k\}$), with distances $\{d_1, d_2, \cdots, d_k\}$, of testing data t are found in the training data set.

(2) Get the coordinate values ($\{v_n^1, v_n^2, \cdots, v_n^6\}_{n=1}^{k}$) of the k neighbors from matrix M.

(3) Compute the coordinate values of testing data t ($\{t^1, t^2, \cdots, t^6\}$)

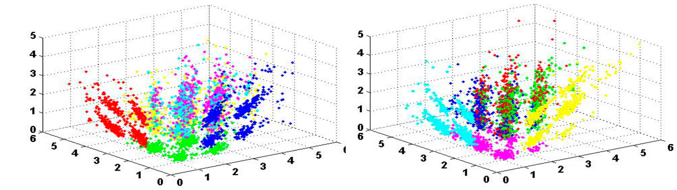
$$t^i = 1/k \times \sum_{\alpha=1}^{k} (d_\alpha + v_\alpha^i) \tag{3}$$

where k is also set to 10. In our approach, testing data t makes the shortest pathes to subsets through its neighbors. t's geodesic distances to subsets can be approximated by averaging the sum of "short hops" to neighboring points and geodesic distances of neighbors. Instead of minimum, average approximation defined above is adopted to be t's distance measurement for its robust performance.

3. EXPERIMENTAL RESULTS

3.1. Speech Corpus

The speech database used in the experiment are got from National Laboratory of Pattern Recognition, Institute of Automation,Chinese Academy of Sciences. It's an emotional speech corpus in Mandarin. The corpus is collected from four Chinese native speakers including two men and two women. Everyone expresses 300 sentences in six emotions involving neutral, angry, fear, happy, sad and surprise. 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.



(a) The first three dimensions (b) The last three dimensions

Fig. 2. Training data in the embedded space. Different colors correspond to different emotions.

The clean speech data were also suppressed by generated noise signal. Gaussian white noise and sinusoid noise were both added to the speech database at several signal-to-noise ratio (SNR) which is determined in (4). Gaussian white noise and sinusoid noise appear frequently in both actual and research environments.

$$\eta = 10 \lg \frac{\frac{1}{n}(\sum_{i=1}^{n} x_i)^2}{\frac{1}{n}(\sum_{j=1}^{n} y_j)^2} \tag{4}$$

where x_i is a sample from speech signal and y_i from noise. Due to the variations of speech signals' energy in different emotions, average SNR was measured among an individual's utterances in all emotions. The SNRs of tested noisy speech were 21dB, 18dB, 15dB, 11dB, 7dB, approximately. Noisy speech with lower SNR wasn't included, due to the difficulty of pitch extraction from them.

3.2. Acoustic Features

In this study, 48 prosodic and 16 formant frequency features were extracted, which were showed to be the most important factors in affect classification [5, 6]. The extracted prosodic features include: 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. If the first derivative of Pitch (Energy) is close to 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. Statistical properties of formant frequency including max, min, mean, median of the first, second, third, and fourth formant were extracted [1].

In the experiment, speaker-dependent emotion recognition was investigated with clean and noisy speech. 10-fold cross-validation method was adopted considering the confidence of recognition results. 90% speech data were used for training and 10% for validation. 64-dimensional vectors of all speech data were projected into a six-dimensional space using the enhanced Lipschitz embedding method mentioned above.

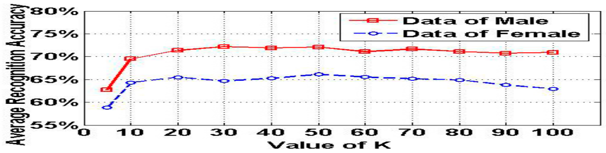
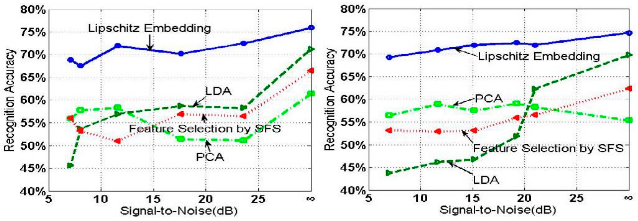


Fig. 3. Distribution of recognition accuracy on different k



(a) Male (b) Female

Fig. 4. Performance comparison between four methods on speech corrupted by Gaussian white noise. ∞ in the x-axis represents clean speech signal.

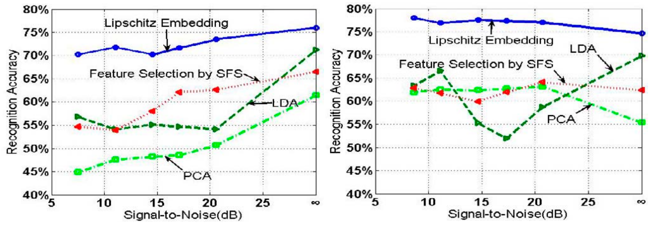
3.3. Speech Emotion Recognition

SVM, a powerful tool for classification, was introduced to classify six emotions in our experiment. It was originally proposed for two-class classification. In our system, 15 (C_6^2) one-to-one SVMs were formed into an MSVM (Multi-SVM) system in which every SVM was established to distinguish one emotion from another. Final classification result was determined by all the SVMs with majority rule. After heavy tests of polynomial, RBF and linear kernels in different parameters, linear SVM (C=0.1) was selected considering the acceptable performance and simple computation.

In the experiment mentioned above, $k = 10$ nearest neighbors were searched in constructing the distance matrix M and embedding the testing data. The impact of different k on the system performance was also investigated. Distribution of recognition accuracy from clean speech on different k is shown in Figure 3. From the curve, $k = 10$ makes an acceptable performance with relatively simple computation.

In order to evaluate the performance of our system on emotion recognition from clean and noisy speech, methods based on PCA, LDA and feature selection by SFS with SVM were also included. 64-dimensional features were projected into six-dimensional space in every method. Traditional noise reduction methods have several problems: method using microphone array cannot avoid increasing the number of microphones; in the case of spectral subtraction (SS) method, the musical tones arisen from residual noise and the processing delay also occurs. With these considerations, we investigated emotion recognition from noisy speech directly, instead of conducting noise reduction.

Figure 4 demonstrates the four methods' emotion recogni-



(a) Male (b) Female

Fig. 5. Performance comparison between four methods on speech corrupted by sinusoid noise. ∞ in the x-axis represents clean speech signal.

tion accuracy of clean speech and speech suppressed by Gaussian white noise. Performance on clean speech and speech corrupted by sinusoid noise is shown in Figure 5. Accuracy in both figures is the average recognition ratio of six emotions. From both figures, our system based on Lipschitz embedding is observed outstanding performance at every SNR test data. Compared with other methods, the accuracy of method on Lipschitz embedding is stable both on speech corrupted by Gaussian white noise and sinusoid noise. Although there are differences among persons, Lipschitz embedding is good at discovering the intrinsic geometry of emotional speech manifold.

4. CONCLUSION

In this paper, we proposed a speech emotion recognition system based on nonlinear manifold. An enhanced Lipschitz embedding method was presented to discover the intrinsic geometry of emotional speech including clean and noisy speech. Compared with the other three methods, the performance of our system is robust when different kinds of noise increase. Of the other three methods, the accuracy of LDA on clean speech is the highest, but drops quickly when noise increases. On the other hand, the accuracy of PCA can hardly be corrupted by louder noise, although its overall performance is poor. In Figure 5(b), the accuracy of noisy speech exceeds clean speech, which is an unbelievable phenomena at a glance. In the future work, more efforts will be made to investigate into the strange result.

5. ACKNOWLEDGEMENT

The work was partly supported by National Natural Science Foundation of China (60575032).

6. REFERENCES

[1] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification,” *Pro. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 593–596, May 2004.

[2] C. M. Lee, S. S. Narayanan, and R. Pieraccini, “Classifying emotions in human-machine spoken dialogs,” *Proc. IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 737–740, Auguest 2002.

[3] Z. J. Chuang and C. H. Wu, “Emotion recognition using acoustic features and textual content,” *Proc. IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 53–56, June 2004.

[4] H. Go, K. Kwak, D. Lee, and M. Chun, “Emotion recognition from the facial image and speech signal,” *SICE 2003 Annual Conference*, vol. 3, pp. 2890–2895, Auguest 2003.

[5] M. Song, J. Bu, C. Chen, and N. Li, “Audio-visual based emotion recognition - a new approach,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1020–1025, June 2004.

[6] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T. S. Huang, “Audio-visual affect recognition in activation-evaluation space,” *Proc. IEEE International Conference on Multimedia and Expo*, pp. 828–831, July 2005.

[7] J. B. Tenenbaum, V. de Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, pp. 2319–2323, December 2000.

[8] S. Roweis and L. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, pp. 2323–2326, December 2000.

[9] J. Bourgain, “On lipschitz embedding of finite metric spaces in hilbert space,” *Israel J. Math.*, vol. 52, nos. 1-2, pp. 46–52, 1985.

[10] W. Johnson and J. Lindenstrauss, “Extension of lipschitz mapping into a hilbert space,” *Contemporary Math.*, vol. 26, pp. 189–206, 1984.

[11] V. Jain and L. K. Saul, “Exploratory analysis and visualization of speech and music by locally linear embedding,” *Pro. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 984–987, May 2004.

[12] R. Togneri, M. D. Alder, and Y. Attikiouzel, “Dimension and structure of the speech space,” *IEEE Proceedings on Communications, Speech and Vision*, vol. 139, Issue 2, pp. 123–127, April 1992.

[13] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee, “Spectral methods for dimensionality reduction,” *Semisupervised Learning*, 2005.