Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 49-64

Manifolds Based Emotion Recognition in Speech

Mingyu You*, Chun Chen*, Jiajun Bu*, Jia Liu*, and Jianhua Tao+

Abstract

The paper presents an emotional speech recognition system with the analysis of manifolds of speech. Working with large volumes of high-dimensional acoustic features, the researchers confront the problem of dimensionality reduction. Unlike classical techniques, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), a new approach, named Enhanced Lipschitz Embedding (ELE) is proposed in the paper to discover the nonlinear degrees of freedom that underlie the emotional speech corpus. ELE adopts geodesic distance to preserve the intrinsic geometry at all scales of speech corpus. Based on geodesic distance estimation, ELE embeds the 64-dimensional acoustic features into a six-dimensional space in which speech data with the same emotional state are generally clustered around one plane and the data distribution feature is beneficial to emotion classification. The compressed testing data is classified into six emotional states (neutral, anger, fear, happiness, sadness and surprise) by a trained linear Support Vector Machine (SVM) system. Considering the perception constancy of humans, ELE is also investigated in terms of its ability to detect the intrinsic geometry of emotional speech corrupted by noise. The performance of the new approach is compared with the methods of feature selection by Sequential Forward Selection (SFS), PCA, LDA, Isomap and Locally Linear Embedding (LLE). Experimental results demonstrate that, compared with other methods, the proposed system gives 9%-26% relative improvement in speaker-independent emotion recognition and 5%-20% improvement in speaker-dependent recognition. Meanwhile, the proposed system shows robustness and an improvement of approximately 10% in emotion recognition accuracy when speech is corrupted by increasing noise.

* College of Computer Science, YuQuan Campus, ZheJiang University, Hangzhou 310027, CHINA
E-mail: {roseyoumy, chenc, bjj, liujia}@zju.edu.cn
The author for correspondence is Jiajun Bu.
+ National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing 100080, CHINA
E-mail: jhtao@nlpr.ia.ac.cn

[Received August 10, 2006; Revised September 11, 2006; Accepted September 30, 2006]

Keywords: Enhanced Lipschitz Embedding (ELE), Dimensionality Reduction, Emotional Speech Analysis, Emotion Recognition

1. Introduction

Human-machine interaction technology has been investigated for several decades. Recent research has put more emphasis on enabling computers to recognize human emotions. As the most effective method in human-human and human-machine communication, speech conveys vast emotional information. Accurate emotion recognition from speech signals will benefit the human-machine interaction and will be applied to areas of entertainment, learning, social development, preventive medicine, consumer relations, etc. [Picard 1997].

The general process of emotion recognition from speech signals can be formulated as below: extracting acoustic features such as Mel-Frequency Cepstral Coefficient (MFCC), Linear Predictive Cepstral Coefficient (LPCC) or low-level features [Ververidis *et al.* 2004], reducing feature dimensionality to an appropriate range for less computational complexity and recognizing emotions with trained SVM, Hidden Markov Model (HMM), Neural Network (NN) or other classifiers.

Dimensionality reduction methods can be grouped into two categories: Feature Selection (FS) and Feature Extraction (FE). An FS method chooses a subset from the original features, preserving most characteristics of the raw data. Ververidis [Ververidis *et al*. 2004] used the Sequential Forward Selection (SFS) method to select the five best features for the classification of five emotional states. However, feature selection needs complex computation to evaluate all the features. How to acquire the best feature set is another tough task. An FE method projects the original features to a completely new space with lower dimensionality through linear or nonlinear affine transformation. PCA, LDA and Multidimensional Scaling (MDS) are popular feature extraction techniques. PCA finds a set of the most representative projection vectors such that the projected samples retain the most information about the original samples. Lee [Lee *et al*. 2002] used PCA to analyze the feature set in classifying two emotions in spoken dialogs. Chuang [Chuang *et al*. 2004] adopted PCA to select 14 principle components from 33 acoustic features in the analysis of emotional speech. LDA uses the class information and finds a set of vectors that maximize the between-class scatter while minimizing the within-class scatter. MDS computes the low dimensional representation of a high dimensional data set that most faithfully preserves the inner products between different input patterns. LDA and MDS have also been employed to reduce the feature dimensionality for emotion recognition [Go *et al*. 2003]. Though widely used for their simplicity, PCA, LDA and MDS are limited by their underlying assumption that data lies in a linear subspace. For nonlinear structures, these methods fail to detect the true freedom degrees of the data.

Recently, a number of research efforts have shown that the speech points possibly reside on a nonlinear submanifold [Jain *et al*. 2004; Togneri *et al*. 1992]. The classical ways of projecting speech into low dimensional space by linear methods are not suitable. Some nonlinear techniques have been proposed to discover the nonlinear structure of the manifold, *e.g*. Isomap [Tenenbaum *et al*. 2000] and LLE [Roweis *et al*. 2000]. Isomap is based on computing the low dimensional representation of a high dimensional data set that most faithfully preserves the pairwise distances between input patterns as measured along the submanifold from which they are sampled. The LLE method captures the local geometry of complex embedding manifold by a set of linear coefficients that best approximate each data point from its neighbors in the input space. These nonlinear methods do yield impressive results in some statistical pattern recognition applications [Jain *et al*. 2004]. However, they yield maps that are defined only on the training data points, so how to evaluate the maps on novel testing data points remains unclear. Lipschitz embedding [Bourgain 1985; Johnson *et al*. 1984] is another nonlinear dimensionality reduction method which works well when there are multiple clusters in the input data [Chang *et al*. 2004]. It is suitable for emotion classification whose input data can be grouped into several emotions.

Most previous work on detecting emotional states investigated speech data recorded in a quiet environment [Song *et al*. 2004; Zeng *et al*. 2005], but humans are able to perceive emotions even in a noisy background. The nonlinear manifold learning algorithms mentioned above [Tenenbaum *et al.* 2000; Roweis *et al.* 2000; Bourgain 1985] try to discover the underlying reason of how humans perceive constancy even though the raw sensory inputs are in flux. Facial images with different poses and lighting directions were also observed to make a smooth manifold [Tenenbaum *et al.* 2000]. Similarly, speech with different emotions, even corrupted by noise, could also be embedded into a low dimensional nonlinear manifold, although none of the previous work has paid attention to this area.

In this paper, an enhanced Lipschitz embedding system is developed to analyze the intrinsic manifold of both emotional speech recorded in quiet environment and those corrupted by noise. Geodesic distance is expected to reflect the true geometry of the emotional speech manifold. With geodesic distance estimation, ELE is developed to embed the extracted acoustic features into a low dimensional space. Then, a linear SVM is trained to recognize the emotional states of the embedded results. In addition, other dimensionality reduction methods such as PCA, LDA, feature selection by SFS with SVM, Isomap, and LLE are implemented for comparison.

The rest of the paper is organized as follows. Section 2 gives a brief description of the emotional speech recognition system. Section 3 presents the ELE algorithm. Experimental results are provided and discussed in Section 4. Section 5 concludes the paper and discusses future work.

2. System Overview

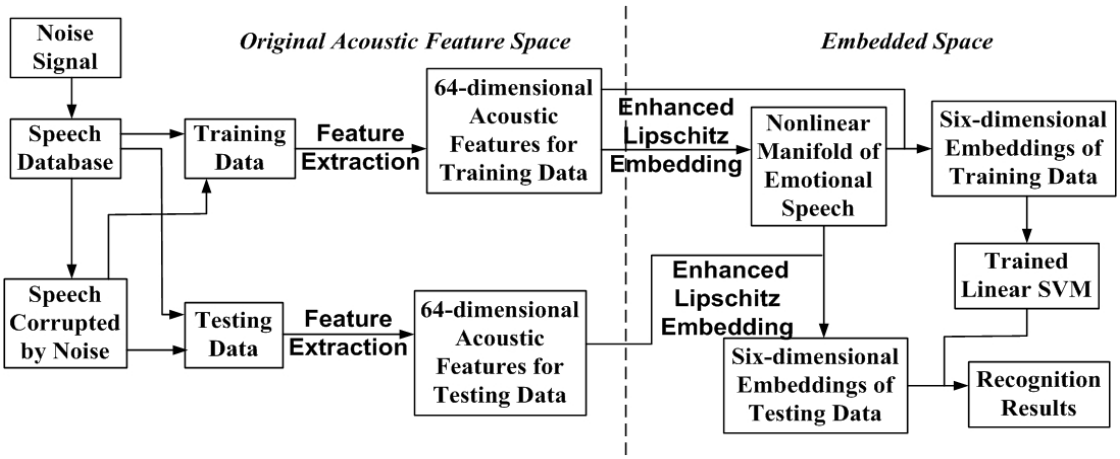


Figure 1. The Framework of Emotion Recognition from Speech

Figure 1 displays the overall structure of this system. Clean speech from the database and speech corrupted by generated noise are both investigated in the system. The emotional speech analysis is done in two phases in this system: training and testing. In the training phase, 64-dimensional acoustic features for each training utterance are obtained after feature extraction. Using ELE, a six-dimensional submanifold is then gained to embody the intrinsic geometry of the emotional training data. Finally, a linear SVM is trained by the embedded training data. In the testing phase, the feature extraction method also extracts 64-dimensional acoustic features for the testing data. The high-dimensional features are then projected into the six-dimensional manifold obtained in the training phase. The emotional state of the testing data is recognized by the trained SVM system. There are two feature spaces mentioned in the workflow: the original acoustic feature space, which is a high-dimensional space found before feature embedding and the embedded space, which is a low-dimensional space found after feature projection.

3. Enhanced Lipschitz Embedding (ELE)

A Lipschitz embedding is defined in terms of a set $R(R=\{A_1, A_2, \cdots, A_k\})$, where $A_i \subset S$ and $\bigcup_{i=1}^{k} A_i = S$. The subset A_i is termed 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 $F(o) = (d(o, A_1), d(o, A_2), \cdots, d(o, A_k))$. 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 the speech manifold, Euclidean distance fails to find the real freedom degrees of the manifold. Tenenbaum *et al.* [Tenenbaum *et al.* 2000] tried to preserve the intrinsic geometry of the data by capturing the geodesic distances between all pairs of data points, which is followed by the algorithm found in this research.

In this new approach, the speech corpus is divided into six subsets $\{A_1, A_2, \cdots, A_6\}$ according to six emotional states (neutral, anger, fear, happiness, sadness and surprise). 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) Construct a graph G connecting neighbor data points. The edge length is determined by the Euclidean distance between neighbor points. The detailed operation can be formulated as Equation (1).

Initiate element m_{ij} in matrix M:

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

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. Specifically, k is set to 10 in this method, which will be discussed further in the following section. 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 guarantees that i and j are unconnected in the graph G consisting of speech data points. Matrix M actually corresponds to the neighborhood graph G whose edge only connects neighbor data points.

(2) Reconstruct matrix M. Replace element m_{ij} with the length of the shortest path between data point i and j in graph G. The shortest path between i and j can be found by tracing through the edges in graph G.

$$m_{ij} = \min\{m_{ij}, m_{ik} + m_{kj}\} \tag{2}$$

Matrix M contains the shortest path distances between all pairs of points in graph G constructed in Equation (1).

(3) Get the coordinate values of $o(\{o_1, o_2, \cdots, o_6\})$ from matrix M. The coordinate value of object o to axis A_i is the distance from o to the closest element in A_i.

$$o_r = \min_{\mu \in A_r} m_{o\mu} \tag{3}$$

where $m_{o\mu}$ is an element of matrix *M*. In this work, object o is projected into a space with six axes $\{A_1, A_2, \cdots, A_6\}$ in accordance with the six emotional states.

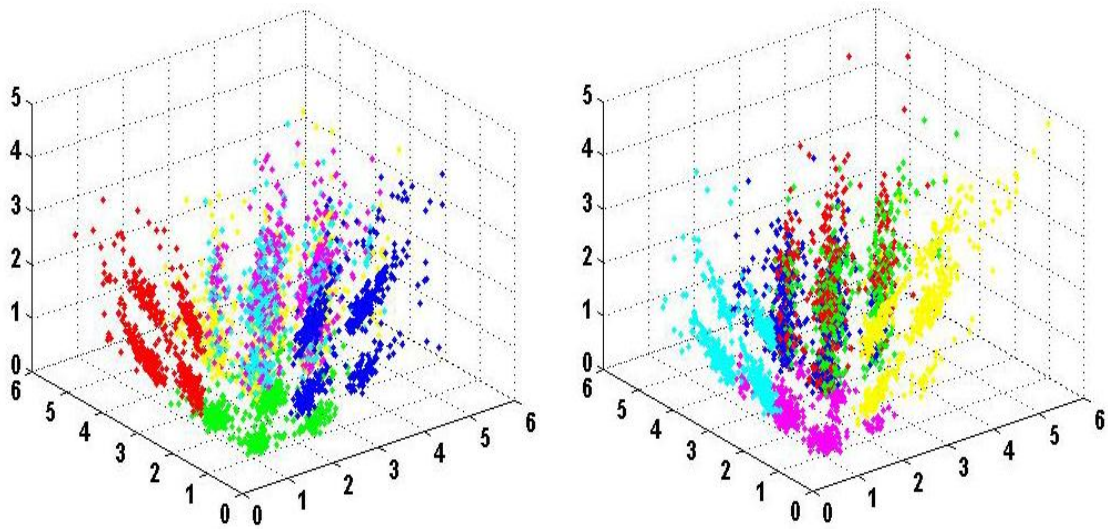


Figure 2. Training data in the embedded space. Different colors correspond to different emotions.

Figure 2 shows the six-dimensional embeddings of 64-dimensional training speech corpus in the six emotional states. Figure 2(a) reveals the first three dimensions of the embedded space and (b) displays the other three dimensions. Emotions neutral, anger and fear, denoted by points in red, green and blue, are easy to be separated in the first three dimensions. Happiness, sadness 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 highly clustered around 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 clusters.

In the proposed ELE technique, the distance matrix *M* is constructed on training data. The training data projection easily depends on the minimal distance to each emotional speech class. Similar to Isomap and LLE, how to evaluate new testing data is still unclear. It is impossible to reconstruct matrix *M* combining the testing data because it is time consuming. Based on the constructed matrix *M*, the authors propose an approach to compute the coordinate values of testing data *t* in the embedded space.

(1) Based on Euclidean distance, the *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. The k nearest neighbors come from the training data, so their coordinates can be found with the processes mentioned in the training phase.

(3) Compute the coordinate values of testing data t $(\{t_1, t_2, \cdots, t_6\})$. In this approach, the testing data t makes the shortest paths to subsets through its neighbors. Therefore, the geodesic distances of t to subsets can be approximated by averaging the sum of "short hops" to its neighboring points and the geodesic distances of its neighbors.

$$t_i = \frac{1}{k} * \sum_{\partial=1}^{k} (d_\partial + v_\partial^i) \tag{4}$$

where k is set to 10 in the proposed system. Instead of using the minimum value, average approximation defined in Equation (4) is adopted to be the distance measurement of t for a robust performance.

4. Experiments

4.1 Speech Corpus

The speech database used in the experiment is from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. The corpus is collected from four Chinese native speakers including two men and two women. Everyone reads 300 sentences in six emotions involving neutral, angry, fear, happy, sad and surprise. The total amount of sentences is thus 300*6*4 = 7200. The speech corpus is sampled at 16 kHz frequency and 16-bit resolution with monophonic Windows PCM format.

The clean speech data were also suppressed by generated noise signal. Gaussian white noise and sinusoid noise generated by LabVIEW were both added to the speech database at various signal-to-noise ratios (SNR) as determined by Equation (5). Gaussian white noise and sinusoid noise appear frequently in both real and research environments.

$$\eta = 10 \lg \frac{\frac{1}{n}(\sum_{i=1}^{n} x_i)^2}{\frac{1}{n}(\sum_{j=1}^{n} y_i)^2} \tag{5}$$

Where x_i is a sample from the speech signal and y_i is a sample from the 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 approximately 21dB, 18dB, 15dB, 11dB, and 7dB. Noisy speech with lower SNR was excluded, due to difficulty in extracting pitch from them.

4.2 Acoustic Features

In this work, 48 prosodic and 16 formant frequency features were extracted, which were shown to be the most important factors in affect classification [Song *et al.* 2004; Zeng *et al.* 2005]. 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 value of Pitch (Energy) plateau at maxima/ minima; max, mean, median duration of Pitch (Energy) plateau at maxima/ minima.

$$\textit{If } |P(x)'-0|<\varepsilon \;\;\&\&\;\; P(x)''>0 \textit{, then } x \in \textit{a plateau at minima}$$
$$\textit{Else if } |P(x)'-0|<\varepsilon \;\;\&\&\;\; P(x)''<0 \textit{, then } x \in \textit{a plateau at maxima}$$

Where $P(x)$ is the Pitch (Energy) value of point x, $P(x)'$ is the first derivative and $P(x)''$ is the second.

Statistical properties of formant frequency including max, min, mean, median of the first, second, third, and fourth formant were extracted [Ververidis *et al.* 2004]. The acoustic feature analysis tool Praat is used to extract the Pitch, Energy and Formant of speech data. All features are based on a speech sentence.

In the experiment for clean speech, speaker-independent and speaker-dependent emotion recognitions were both investigated within the same gender. On the other hand, in the experiment for noisy speech, speaker-dependent emotion recognition was investigated. 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 the six-dimensional space using the ELE method mentioned above.

4.3 Emotion Recognition in Speech

SVM, a powerful tool for classification, was introduced to classify the six emotions in this experiment. It had originally been proposed for two-class classification. In this system, 15 one-to-one SVMs were combined into an MSVM (Multi-SVM), in which each SVM was used to distinguish one emotion from another. Final classification result was determined by all the SVMs with the majority rule. After the heavy tests of polynomial, radial basis function and linear kernels with different parameters, linear SVM (C=0.1) was selected for its acceptable performance and simplicity.

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, influences made by k on male model are similar to that of female model. In both models, k =10 makes an acceptable performance with relatively low computational cost.

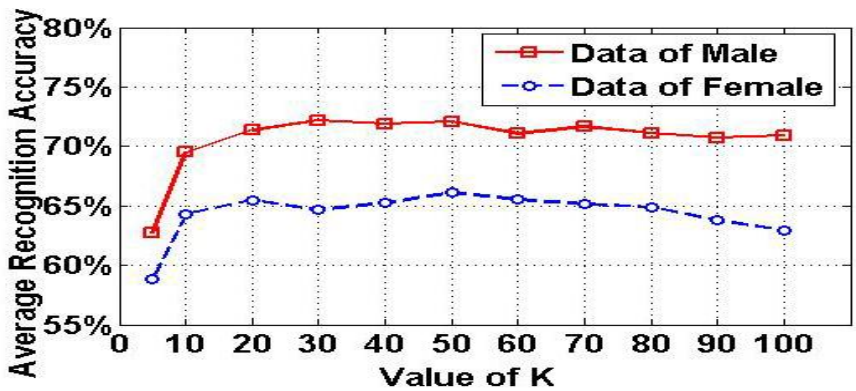
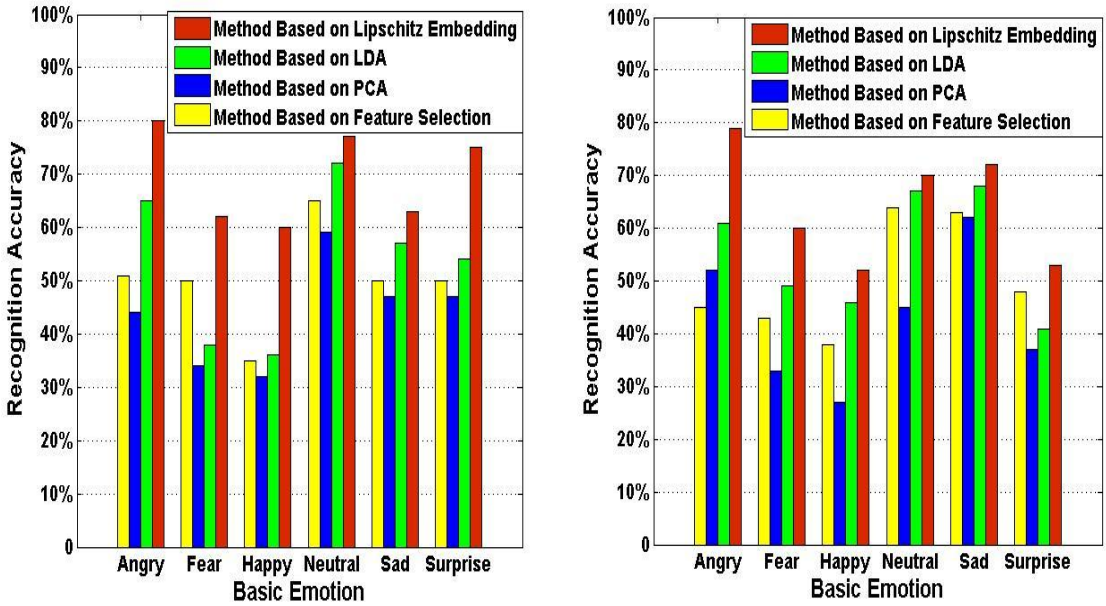


Figure 3. Distribution of recognition accuracy on different k

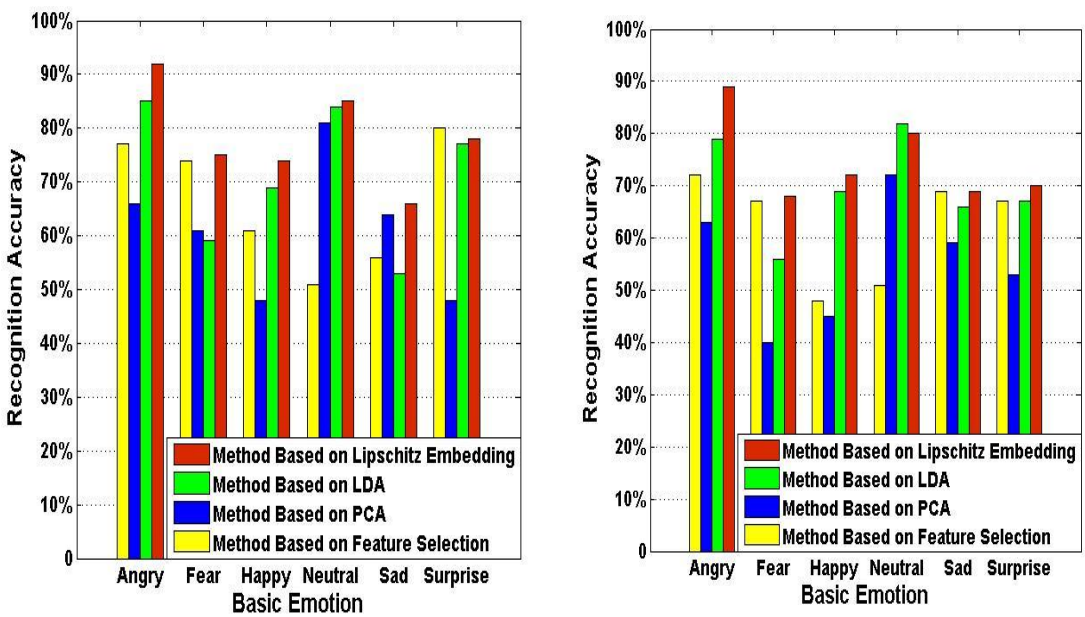
In order to evaluate the classification results of ELE, linear dimensionality reduction methods such as PCA, LDA and feature selection by SFS with an SVM classifier were also included for comparison. 64-dimensional features were projected into the six-dimensional space in every method. Figure 4 demonstrates the comparative performance of the four methods in speaker-independent emotion recognition. Speaker-dependent implementation results of the four methods are shown in Figure 5.

Due to acoustic variations that exist between different people, the average accuracy of the speaker-dependent emotion recognition (Figure 5) is about 10% higher than that of the speaker-independent (Figure 4). The classification rate of the male speaker is a little higher than the female, which probably indicates that women's facial expressions or body gestures convey more emotional information. In speaker-independent and speaker-dependent processes, the method based on ELE comes up with the best performance in almost all of the emotional situations. The relative improvement of the proposed method is 9%-26% in the speaker-independent system and 5%-20% when the system is speaker-dependent. While the classification rate of happiness is lower than other emotions in the speaker-independent system, the accuracy of happiness is comparable with the others in the speaker-dependent. What one can deduce from the results is that people express happiness in greatly varying manners.



(a) Male **(b) Female**

Figure 4. Speaker-independent performance comparison among the four methods.

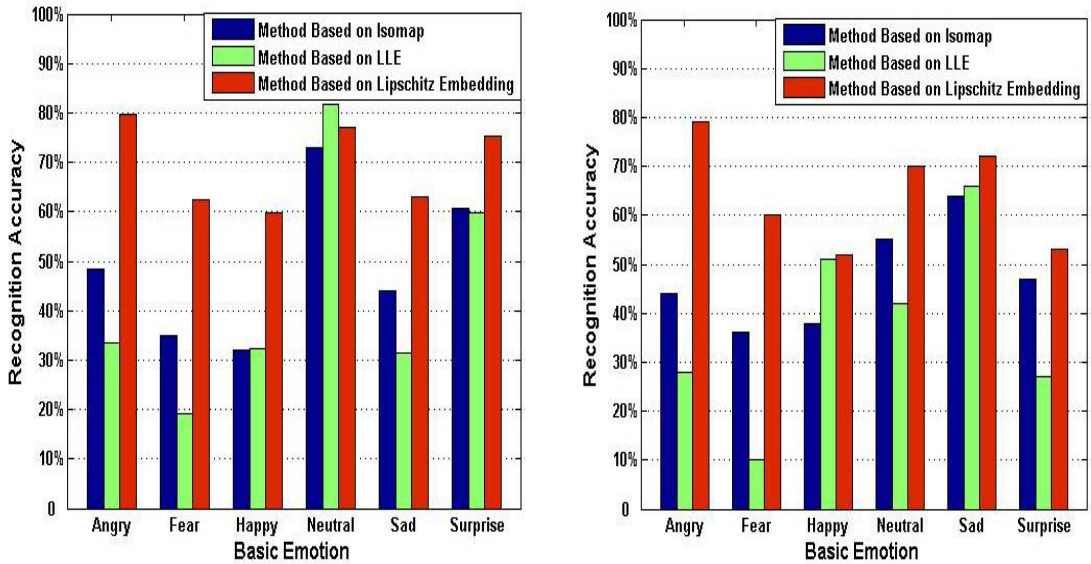


(a) Male **(b) Female**

Figure 5. Speaker-dependent performance comparison among the four methods.

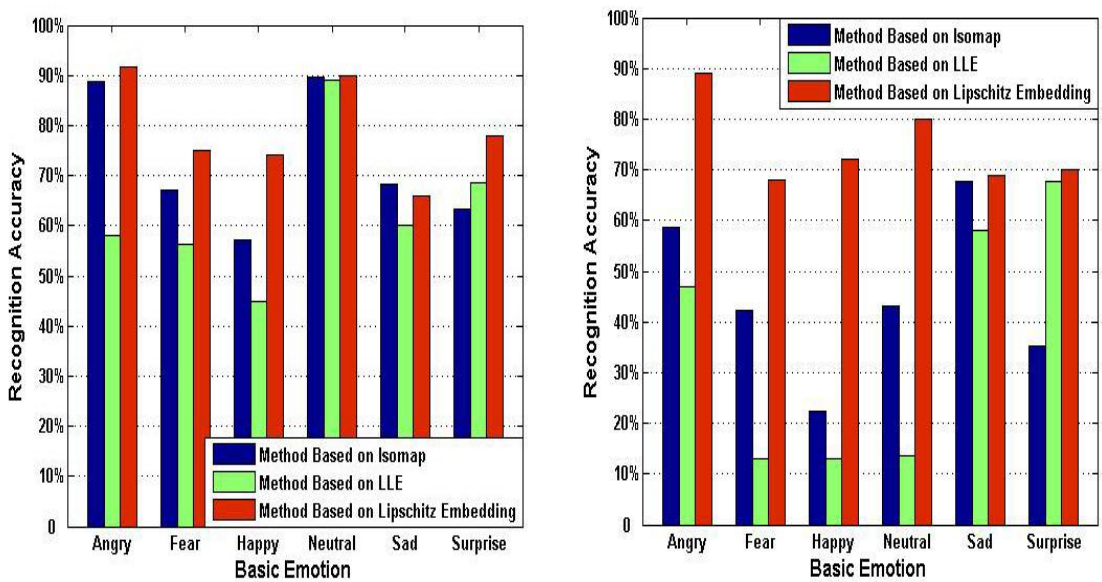
Considering the nonlinear submanifold that the ELE method involves, popular nonlinear dimensionality reduction approaches, such as Isomap and LLE, are implemented in this emotional speech recognition system for comparison. As mentioned before, Isomap and LLE only decide how to project the training data into a low dimensional manifold and leave the projection problem of novel testing data unsettled. However, in the emotion recognition system, all the training data and testing data should be embedded into low dimensional space. In the implementation of Isomap and LLE, the authors reconstruct the distance matrix M when facing the novel test data. Although it costs a lot of computation time, it will help attain Isomap and LLE's best performance. Comparison with those results gives one a solid evaluation of the proposed method.

Figure 6 and Figure 7 display the recognition accuracy of the six emotions in a speaker-independent and a speaker-dependent environment, respectively. From both figures, the method based on ELE still yields the best results in almost all of the emotional situations. In a speaker-independent environment, the proposed method outperforms the other two in the emotions angry and fear, especially. For the emotional speech recognition application, Isomap is more suitable than LLE. From Figure 6 and Figure 7, one can see that the recognition accuracy of Isomap is higher than LLE in most of the emotion states. Isomap is based on geodesic distance estimation and captures the global data structure when finding the low embeddings, while LLE focuses on preserving the local geometry of data points. ELE is somewhat similar to Isomap, which may explain why Lipschitz embedding and Isomap both outperform LLE in the experimental results. However, Isomap consumes more computation time than ELE. They both need the time-consuming operation of constructing the neighborhood graph, but the embedding step of Isomap is more complex. LLE conducts unbalanced performance when dealing with different basic emotions. For example, in Figure 6(b), LLE only attains 10% accuracy with the emotion fear, while it achieves about 65% with sad. The unbalanced recognition rate will greatly reduce the system's robustness. LLE gets a poor recognition rate for the female speaker in the speaker-dependent environment shown in Figure 7(b). Isomap and LLE behave differently between the male and the female in the speaker-dependent environment, but the performance of ELE seems stable. Comparing the results of Figure 6 and Figure 7 with those of Figures 4 and 5, nonlinear methods' performance may not be better than the linear methods', although they require more complicated computation. It is shown that the method of preserving the geometry of the data set is crucial in nonlinear manifold reduction approaches.



(a) Male **(b) Female**

Figure 6. Speaker-independent performance comparison between three nonlinear methods.

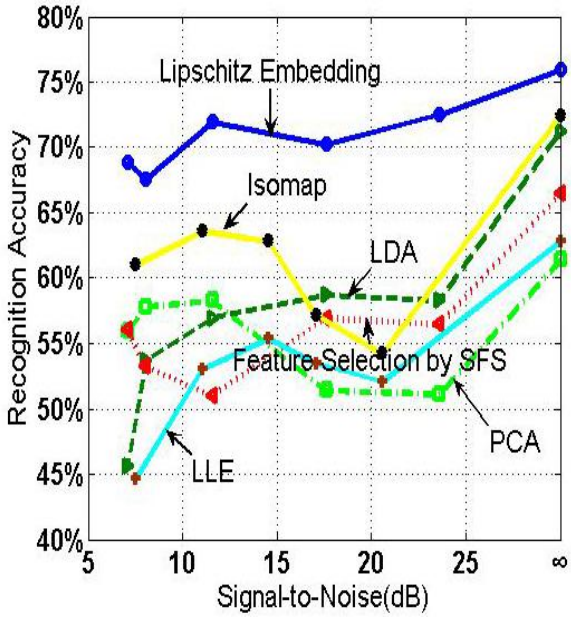


(a) Male **(b) Female**

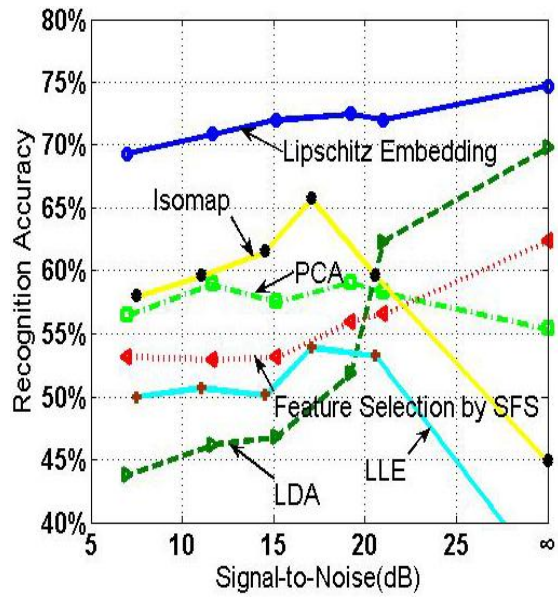
Figure 7. Speaker-dependent performance comparison between three nonlinear methods.

In order to test the perception constancy of the proposed approach, the classification performance of ELE on noisy speech is also investigated. Classical methods like PCA, LDA and feature selection by SFS with SVM were included for performance comparison. Nonlinear methods, Isomap and LLE, were also implemented. 64-dimensional features were projected into the six-dimensional space in every method. In many other applications, researchers tended to conduct noise reduction first for the noisy speech data. However, traditional noise reduction methods still face several challenges: the method using microphone array cannot avoid the problem of increasing the number of microphones; in the case of the spectral subtraction (SS) method, the musical tones arise from residual noise and processing delays also occur. With these considerations, the authors investigated emotion recognition from noisy speech directly, instead of conducting noise reduction. Since facial images with different poses and lighting directions were observed to make a smooth manifold, speech corrupted by noise may also be embedded into a low dimensional nonlinear manifold.

Figure 8 demonstrates the six methods' emotion recognition accuracy for clean speech and speech suppressed by Gaussian white noise. Performances with clean speech and speech corrupted by sinusoid noise are shown in Figure 9. Accuracies in both figures are the average recognition ratio of six emotions. From both figures, this system, based on Lipschitz embedding, shows outstanding performance with every SNR test data. Compared with the other methods, the accuracy of this method on Lipschitz embedding is stable both with speech corrupted by Gaussian white noise and with speech corrupted by sinusoid noise. Although there are differences among individuals, ELE is good at discovering the intrinsic geometry of the emotional speech manifold. The accuracy of LDA on clean speech is high, 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. The Isomap Method also achieves acceptable accuracy in different experimental environments, except for the male speaker in speech corrupted by sinusoid noise. The performance of LLE is still disappointing. Keeping the local geometry by reconstructing from neighbors seems not to be appropriate for emotional speech recognition applications. From these figures, one can see an interesting phenomenon where the recognition accuracy of noisy speech is sometimes higher than that of clean speech. Features used to distinguish the different emotion states are strengthened by mild noise.

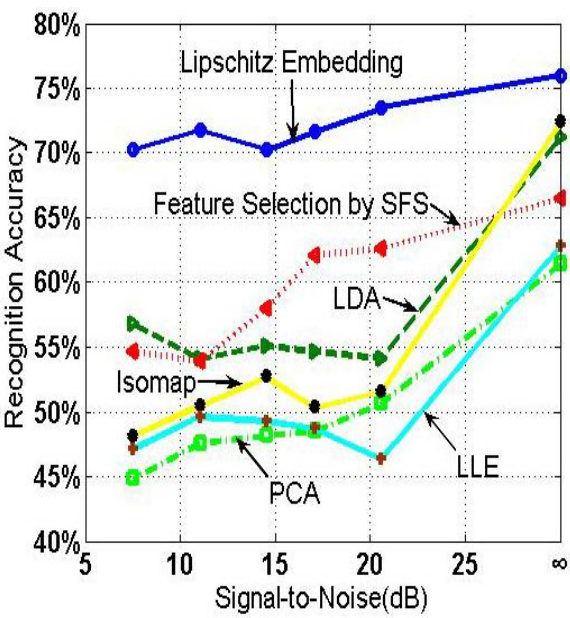


(a) Male

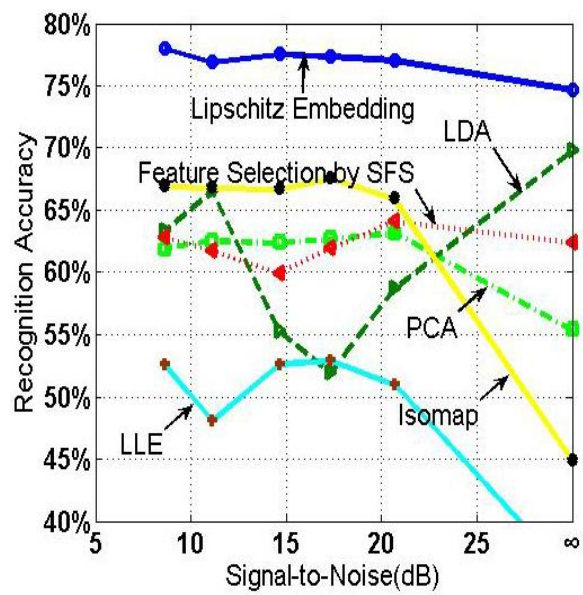


(b) Female

Figure 8. Performance comparison between linear and nonlinear methods on speech corrupted by Gaussian white noise. ∞ in the x-axis represents clean speech signal.



(a) Male



(b) Female

Figure 9. Performance comparison between linear and nonlinear methods on speech corrupted by sinusoid noise. ∞ in the x-axis represents clean speech signal.

5. Conclusion and Future Work

In this paper, the authors proposed an emotional speech recognition system based on a nonlinear manifold. Method ELE was presented to discover the intrinsic geometry of emotional speech including clean and noisy utterances. Compared with traditional approaches, including linear and nonlinear dimensionality reduction methods, this method came up with the best performance when dealing with almost all of the basic emotions in both speaker-independent and speaker-dependent processes. Even in a noisy environment, the performance of ELE was outstanding compared with the other methods and robust when different kinds of noise increase. Although LDA and Isomap also achieved plausible recognition results in the experiments, the proposed method balanced the classification rate in each emotion, which both of them lacked. The time consumption of Isomap was also higher than the proposed method. As another nonlinear method, LLE showed poor performance, meaning that preserving the intrinsic geometry of data corpus was vital. The key idea of the proposed method is to take the multiple classes of input patterns into consideration. Experimental results show that this idea is successful in emotional speech recognition applications.

With the method based on Lipschitz embedding, the average recognition accuracy of the female speaker is 5% lower than that of the male. The underlying reason should be investigated in detail and a robust algorithm is expected. Besides, the essential reason explaining the phenomenon that the accuracy of noisy speech exceeds clean speech should be investigated. In order to achieve better performance, improvement will be made to the proposed method and multi-modal emotion recognition will be included in future work.

ACKNOWLEDGEMENT

The work was partly supported by National Natural Science Foundation of China (60575032) and the 863 program (20060101Z1037). And the authors thank Cheng Jin and Can Wang for their generous help in the experiment and paper.

Reference

Bourgain, J., "On lipschitz embedding of finite metric spaces in hilbert space, " *Israel J. Math.*, 52(1-2), 1985, pp. 46-52.

Chang, Y., C. Hu, and M. Turk, "Probabilistic expression analysis on manifolds," In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004, Washington, DC, America, vol. 2, pp. 520-527.

Chuang, Z.J., and C. H. Wu, "Emotion recognition using acoustic features and textual content," In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2004, Taipei, Taiwan, vol. 1, pp. 53-56.

Duchene, J., and S. Leclercq, "An optimal transformation for discriminant principal component analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 10(6), 1988, pp. 978-983.

Go, H., K. Kwak, D. Lee, and M. Chun, "Emotion recognition from the facial image and speech signal," In *proceedings of SICE 2003 Annual Conference*, 2003, Fukui, Japan, vol. 3, pp. 2890-2895.

Jain, V., and L. K. Saul, "Exploratory analysis and visualization of speech and music by locally linear embedding," In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, Montreal, Canada, vol. 3, pp. 984-987.

Johnson W., and J. Lindenstrauss, "Extension of lipschitz mapping into a hilbert space," *Contemporary Math.*, vol. 26, 1984, pp. 189-206.

Lee, C.M., S. S. Narayanan, and R. Pieraccini, "Classifying emotions in human-machine spoken dialogs," In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2002, Lusanne, Switzerland, vol. 1, pp. 737-740.

Picard, R., *Affective computing*, The MIT Press, Cambridge, MA, 1997.

Roweis, S., and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, 2000, pp. 2323-2326.

Song, M., J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition - a new approach," In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004, Washington, DC, America, vol. 2, pp. 1020-1025.

Tenenbaum, J.B., V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, 2002, pp. 2319-2323.

Togneri, R., M. D. Alder, and Y. Attikiouzel, "Dimension and structure of the speech space," *IEEE Proceedings on Communications, Speech and Vision*, 139(2), 1992, pp. 123-127.

Ververidis, D., C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, Montreal, Canada, vol. 1, pp. 593-596.

Zeng, Z., Z. Zhang, B. Pianfetti, J. Tu, and T. S. Huang, "Audio-visual affect recognition in activation-evaluation space," In P*roceedings of IEEE International Conference on Multimedia and Expo*, 2005, Amsterdam, Netherlands, pp. 828-831.