INTECH
open science | open minds

OPEN ACCESS
ARTICLE

International Journal of Advanced Robotic Systems

Speech Emotion Recognition Using an Enhanced Kernel Isomap for Human-Robot Interaction

Regular Paper

Shiqing Zhang[1,*], Xiaoming Zhao[2] and Bicheng Lei[1]

1 School of Physics and Electronic Engineering, Taizhou University, Taizhou, P.R. China
2 Department of Computer Science, Taizhou University, Taizhou, P.R. China
* Corresponding author E-mail: tzczsq@163.com

Received 28 May 2012; Accepted 5 Dec 2012

DOI: 10.5772/55403

Abstract Speech emotion recognition is currently an active research subject and has attracted extensive interest in the science community due to its vital application to human-robot interaction. Most speech emotion recognition systems employ high-dimensional speech features, indicating human emotion expression, to improve emotion recognition performance. To effectively reduce the size of speech features, in this paper, a new nonlinear dimensionality reduction method, called 'enhanced kernel isometric mapping' (EKIsomap), is proposed and applied for speech emotion recognition in human-robot interaction. The proposed method is used to nonlinearly extract the low-dimensional discriminating embedded data representations from the original high-dimensional speech features with a striking improvement of performance on the speech emotion recognition tasks. Experimental results on the popular Berlin emotional speech corpus demonstrate the effectiveness of the proposed method.

Keywords Speech Emotion Recognition, Nonlinear Dimensionality Reduction, Human-Robot Interaction

1. Introduction

As robots take on an increasingly ubiquitous role in people's daily routines - in common places such as homes, supermarkets, hospitals, offices and so forth - they must be easy for common citizens to use and interact with. This raises the important question of how to properly interface trained humans with robots in a natural, intuitive and enjoyable manner. To solve this question, robots should be able to recognize the emotions of humans so as to provide a friendly environment. Without recognizing the emotions, it will be very difficult for robots to interact with humans in a natural way.

Affective computing, which is currently an active research area, aims to build machines that recognize, express, model, communicate and respond to user's emotion information [1]. Within this field, the recognition of emotions from human speech - i.e., speech emotion recognition - is increasingly attracting attention and has become an important issue in human-robot interaction, since human speech provides a natural and intuitive interface for interaction with humanoid robots [2-4].

The structure of the remainder of this paper is as follows. Section 2 presents the previous work in brief. In Section 3, a speech emotion recognition system in human-robot interaction is introduced. Afterwards, a speech emotion recognition system using the proposed EKIsomap is detailed in Section 4. Subsequently, Section 5 describes the experiment study in detail. In Section 6, a discussion is given. Finally, the conclusions are presented in Section 7.

2. Previous Work

In many previously reported studies [5-9], to attain as much emotional expression information as possible, a large number of speech features related to emotion expression - such as pitch-related, intensity-related, duration-related and so on - were normally extracted for speech emotion recognition. As a result of feature extraction being a high-dimensional speech feature set, it is desirable to perform feature data processing in pursuit of reducing the size of the extracted speech feature set. To achieve this goal, dimensionality reduction techniques can be used to produce few new features containing most of the valuable speech information. In terms of the characteristic of the transform between the original feature space and the compact representations, dimensionality reduction techniques can be classified into two categories: linear and nonlinear. The most well-known linear dimensionality reduction technique is principal component analysis (PCA) [10], which has been successfully used for reducing the dimensionality of emotional speech features [5-9].

In recent years, it has been found that speech data resides on or near a nonlinear submanifold embedded in a high-dimensional acoustic space [11][12]. The nonlinear manifold structure of speech data emphasizes the nonlinear relation between articulatory and acoustic spaces, and captures the essence of human speech production to a great extent. Given the nonlinear manifold structure of speech data, the traditional linear PCA method based on the linear assumption of feature data cannot effectively handle such nonlinear speech data.

In order to effectively deal with the nonlinear manifold structure of speech data, in recent years manifold learning (also called 'nonlinear dimensionality reduction') techniques, which aim to find a smooth low-dimensional manifold embedded in a high-dimensional data space, can be used to perform nonlinear dimensionality reduction and produce a few new meaningful features on the speech emotion recognition tasks [11][12]. The two representative manifold learning methods are locally linear embedding (LLE) [13] and isometric mapping (Isomap) [14]. However, these two manifold learning methods often failed to achieve satisfactory performance in the speech emotion recognition tasks, since they lack a good generalization property on new data points.

To overcome the above-mentioned drawback of manifold leaning methods, in this paper a new kernel-based nonlinear dimensionality reduction algorithm, called 'enhanced kernel Isomap' (EKIsomap) is proposed and applied for speech emotion recognition in human-robot interaction. The proposed method is used to extract the low-dimensional, discriminating, embedded data representations from the original high-dimensional speech features with a striking improvement of performance in speech emotion recognition tasks.

3. Speech Emotion Recognition in Human-robot Interaction

The basic speech emotion recognition system in human-robot interaction is shown in Figure 1. This system consists of three main steps: audio segmentation, feature extraction and processing, emotion classification. Once emotion classification is finished, the robot will provide services for the user according to the emotional states of the user.

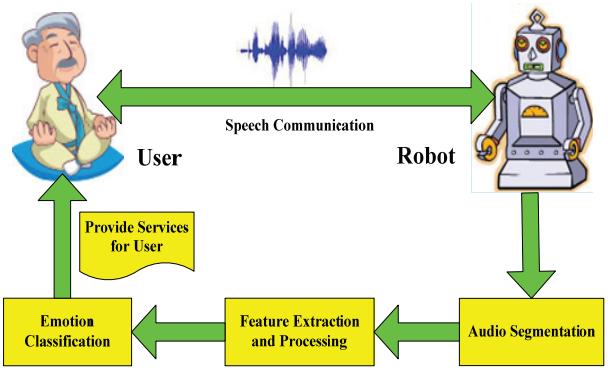

Figure 1. Speech emotion recognition system in human-robot interaction

The first step in this system is to segment the incoming speech signal into the meaningful units that can serve as emotional classification units, such as utterance. Voice activity detection (VAD) is employed to segment by pauses into signal chunks of voice activity without pauses longer than 200 ms. In this work, the long-term speech information method [15] is used to perform VAD.

The goal of feature extraction and processing is to the extract relevant features from speech signals with respect to emotions, and to reduce the size of the speech feature set to fewer dimensions. The widely used acoustic features indicating human emotion expression are prosody features and voice quality features [5-9]. In this paper, the extracted prosody features contain pitch, intensity and duration, while the extracted voice quality features include the first three formants, spectral energy distribution, harmonics-to-noise-ratio (HNR), pitch

irregularity and amplitude irregularity. The high-level statistical parameters, such as mean, standard derivations, median, quartiles, etc., are computed for each extracted acoustic feature. These extracted acoustic features are, in total, 48-dimensonal. To effectively reduce the size of speech features, we use the proposed EKIsomap method to extract the low-dimensional discriminating embedded data representations for speech emotion recognition.

Emotion classification, lastly, maps feature vectors onto emotion classes through learning by data examples. The representative emotion classification methods are linear discriminant classifiers (LDC), K-nearest-neighbour (KNN), artificial neural network (ANN) and support vector machines (SVM), etc. In this work, we use an SVM classifier to perform emotion classification.

4. Speech Emotion Recognition Using EKIsomap

4.1 System Structure

Figure 2 provides the system structure of speech emotion recognition using EKIsomap. As shown in Figure 2, it contains three principal parts: acoustic feature extraction, feature dimensionality reduction and emotion recognition. In the acoustic feature extraction stage, the original emotional speech samples from the emotional speech database are divided into two parts: training samples and testing samples. The corresponding acoustic features for training samples and testing samples, such as prosody and voice quality features, are extracted. The result of this stage is a speech data set represented by a set of high-dimensional speech features. The second stage aims at reducing the size of speech features and generating the new low-dimensional discriminating features with EKIsomap. The last stage in this system is that in the reduced low-dimensional feature space the trained SVM classifier is used to identify the accurate emotion categories, and provide recognition results.

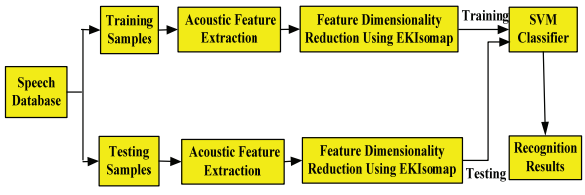

Figure 2. Speech emotion recognition system using EKIsomap

4.2 The Proposed EKIsomap Algorithm

The proposed EKIsomap is designed by integrating the kernel discriminant information extraction in a reproducing kernel Hilbert space (RKHS) with the existing kernel Isomap (KIsomap) [16] method. KIsomap effectively combines the kernel idea and Isomap and has a good generalization property on new data points, since it uses a kernel trick as is the case in kernel principal component analysis (KPCA) [17]. However, this kind of KIsomap still has two shortcomings. First, KIsomap, as an unsupervised learning method, fails to extract the discriminating embedded data representations since KIsomap does not take into account the known class label information of input data. This will heavily decrease the performance of KIsomap on classification tasks. Second, the kernel idea of KIsomap is that the geodesic distance matrix with a constant-shifting technique is referred to as a Mercer kernel matrix [16]. Therefore, as a kernel-based method, KIsomap cannot employ the characteristic of a kernel-based learning - i.e., a nonlinear kernel mapping - to explore higher-order information of input data sets, as KIsomap has no kernel function to perform a nonlinear kernel mapping. This is not a good property for KIsomap when adopted as a feature extraction method. To tackle the drawbacks of KIsomap, in this paper an enhanced variant of KIsomap, called 'enhanced KIsomap' (EKIsomap), is proposed and applied for speech emotion recognition.

The Fisher's criterion - namely, that the inter-class scatter should be maximized while the intra-class scatter should be simultaneously minimized - has become one of the most important selection criteria for projection techniques, since it endows the projected data vectors with good discriminating power. Motivated by Fisher's criterion, when using KIsomap to extract the low-dimensional embedded data representations, the inter-class dissimilarity could be maximized while the intra-class dissimilarity could be minimized in order to have improved tightness among similar patterns and better separability for dissimilar patterns. To develop the EKIsomap algorithm, a kernel matrix is first constructed by performing a nonlinear kernel mapping with a kernel function, and then a kernel discriminant distance - in which the inter class dissimilarity is maximized while the intra-class dissimilarity is simultaneously minimized - is designed to extract the discriminant information in a RKHS.

The detailed steps of EKIsomap are presented as follows:

Suppose the input data point $(x_i, L_i), i = 1,2,3,\cdots,N$, where $x_i \in R^D$ and L_i is the class label of x_i and the output data point is $y_i \in R^d$.

Step 1: Kernel nonlinear mapping for each x_i.

A nonlinear mapping ϕ is defined as:

$$\phi : R^D \to \mathcal{F}, \quad x \mapsto \phi(x) \tag{1}$$

By using the nonlinear mapping ϕ, the input data point $x_i \in R^D$ is mapped into some potentially high dimensional feature space $\mathcal{F}$. For a properly chosen ϕ,

an inner product $\langle,\rangle$ can be defined on $\mathcal{F}$, which makes a so-called reproducing kernel Hilbert space (RKHS). In a RKHS, a kernel function $\kappa(x_i,x_j)$ can be defined as:

$$\kappa(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle=\phi(x_i)^T\phi(x_j) \quad (2)$$

where κ is known as a kernel.

Step 2: Find the nearest neighbours for each $\phi(x_i)$ by using the following kernel discriminant distance.

The kernel discriminant distance used in EKIsomap is defined as follows:

$$D_\kappa(x_i,x_j)=\begin{cases}\sqrt{1-e^{\frac{-[d_\kappa(x_i,x_j)]^2}{\beta}}} & L_i=L_j\\ \sqrt{e^{\frac{[d_\kappa(x_i,x_j)]^2}{\beta}}}-\alpha & L_i\neq L_j\end{cases} \quad (3)$$

where $d_\kappa(x_i,x_j)$ is the kernel Euclidean distance matrix without class label information, whereas $D_\kappa(x_i,x_j)$ is the kernel discriminant distance matrix integrating class label information. β is a smoothing parameter related to the data 'density', and it is usually feasible to set β to be the average kernel Euclidean distance between all pairs of data points. α is a constant factor ($0\le\alpha\le1$) and gives the intra-class dissimilarity a certain probability of exceeding the inter-class dissimilarity. As shown in Eq.(3), we can make two observations. First, each dissimilarity function in $D_\kappa(x_i,x_j)$ - i.e., inter-class dissimilarity and intra-class dissimilarity - is monotone, increasing with respect to the Euclidean distance. This ensures that the main geometric structure of the original data sets can be preserved well when using EKIsomap to produce low-dimensional embedded data representations. Second, the inter-class dissimilarity in $D_\kappa(x_i,x_j)$ can always be definitely larger than the intra-class dissimilarity, conferring a high discriminating power of EKIsomap's projected data vectors. This is a good property for emotion classification.

Step 3: Compute the geodesic distances, d_{ij}, containing shortest paths for all pairs of data points by Dijkstra's algorithm and define $\mathbf{D}^2=[d_{ij}^2]$.

Step 4: Construct a matrix $\mathbf{K}(\mathbf{D}^2)$ based on the approximate geodesic distance matrix:

$$\mathbf{K}(\mathbf{D}^2)=-\frac{1}{2}\mathbf{H}\mathbf{D}^2\mathbf{H} \quad (4)$$

where $\mathbf{H}=\mathbf{I}-(1/N)ee^T$, $e=[1,\cdots,1]^T\in R^N$.

Step 5: Compute the top d eigenvectors and give an embedding of input points by:

$$\mathbf{Y}=\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^T \quad (5)$$

where $\mathbf{\Lambda}^{d\times d}$ is the eigenvalue matrix and $\mathbf{V}\in R^{N\times d}$ is the eigenvector matrix.

5. Experiment Study

To evaluate the performance of EKIsomap on the speech emotion recognition tasks, the popular Berlin emotional speech corpus [18] was used for the experiments. The Berlin emotional speech database comprises around 535 emotional utterances spoken in seven different emotions. These sentences were not equally distributed between the various emotional states: anger (127), joy (71), sadness (62), neutral (79), boredom (81), disgust (46) and fear (69). Ten professional native German-speaking actors (5 female and 5 male) simulated the emotions, producing 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication and were interpretable in all the applied emotions. The actors were advised to read these predefined sentences in the targeted seven emotions. The length of the speech samples varies from three seconds to eight seconds. The recordings were taken in an anechoic chamber with high-quality recording equipment and available at a sampling rate of 16 kHz with a 16-bit resolution and mono channel.

5.1 Experiment Setup

We used the LIBSVM package, available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, to implement the SVM algorithm with a radial basis function (RBF) kernel, kernel parameter optimization and a one-versus-one strategy for the multi-class emotion classification problem.

As was done in [5], a tenfold cross-validation scheme was employed over the emotional speech data sets for all the emotion classification experiments and the recognition results were averaged. Following [12], the number of nearest neighbours was set to 12 for manifold learning methods, including LLE, Isomap, KIsomap and EKIsomap. For kernel-based methods, including KPCA and EKIsomap, the typical Gaussian kernel $\kappa(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2)$ was adopted for its better performance when compared with linear and polynomial kernels. Moreover, the parameter σ for the Gaussian kernel was set to 1 for its satisfying performance.

5.2 Experiment 1: Emotional Data Visualization

In this section, we compare the results of two dimensional embedded mappings of both KIsomap and EKIsomap on the extracted 48-dimensional acoustic features. For simplicity, we split the original acoustic feature data into two parts: one half for training and the other half for testing. Figure 3 shows the split training

data and testing data from the Berlin speech corpus and the corresponding two-dimensional embedded mappings obtained by KIsomap and EKIsomap ($\alpha = 0.2$) on the training data and testing data. According to the distributed different colours and signs (the seven signs represent the seven emotions) of the embedding results in Figure 3, we can observe that, compared with KIsomap, EKIsomap projects training data and testing data into two embedded mappings with better tightness for similar patterns and better separability for dissimilar patterns. This is a good property for classification. In contrast, KIsomap produces the two dimensional embedded mappings with a visible overlapping of each other. The embedded results in Figure 3 clearly reveal that EKIsomap can nonlinearly extract the low-dimensional embedded data representations with a higher discriminating power in comparison with KIsomap. Therefore, EKIsomap is more suitable than KIsomap for speech emotion classification.

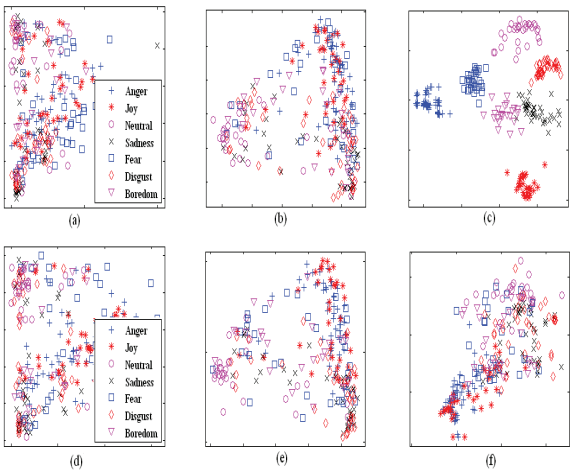

Figure 3. Comparisons of the two-dimensional embedded results obtained by KIsomap and EKIsomap. (a) Training data from the Berlin speech corpus; (b) embedded points with KIsomap on the training data; (c) embedded points with EKIsomap on the training data; (d) testing data from the Berlin speech corpus; (e) projection of the testing data with KIsomap; (f) projection of the testing data with EKIsomap.

5.3 Experiment 2: Emotion Recognition Results and Analysis

The performance of PCA, LLE, Isomap, KIsomap and KPCA as well as EKIsomap on the speech emotion recognition tasks, and the corresponding embedded feature dimension, are presented in Figure 4. In each embedded feature dimension, the constant factor α ($0 \le \alpha \le 1$) for EKIsomap can be optimized using a simple exhaustive search within a scope ($\alpha = 0, 0.1, 0.2, \cdots, 1$). The optimal α corresponds to the best performance of EKIsomap. The best results of all the used methods along with the corresponding embedded feature dimensions are listed in Table 1. Note that the "Baseline" method denotes that the recognition result was directly obtained on the original 48-dimensional acoustic features.

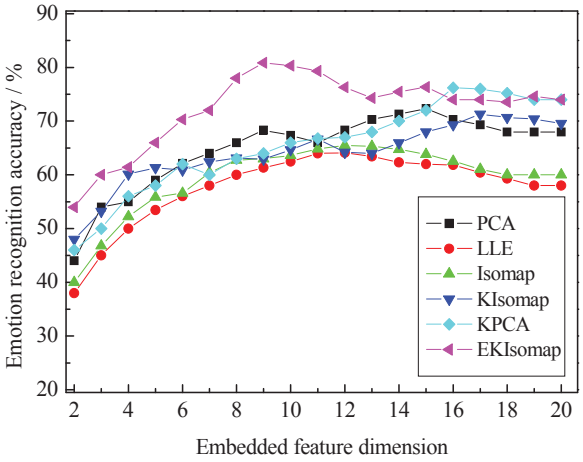

Figure 4. Emotion recognition accuracy vs. the embedded feature dimension

Methods	Feature Dimension	Recognition Accuracy
Baseline	48	76.42%
PCA	15	72.35%
LLE	12	64.10%
Isomap	12	65.48%
KIsomap	17	71.32%
KPCA	16	76.21%
EKIsomap	9	80.85%

Table 1. The best results of different methods with the corresponding embedded feature dimension

From the results in Figure 4 and Table 1, we can see that EKIsomap obtains the best accuracy, of 80.85% with 9 embedded features, outperforming the other methods used. This indicates that EKIsomap is capable of extracting the low-dimensional embedded data representations with the highest discriminating power among all the used methods. This is attributed to the fact that EKIsomap aims to make the inter-class dissimilarity definitely larger than the intra-class dissimilarity with the aid of the kernel discriminant distance used in RKHS.

	Ang	Joy	Sad	Neu	Fea	Bor	Dis
Ang	90.83	7.07	0	0	2.10	0	0
Joy	19.55	65.76	0.13	7.52	4.22	1.41	1.41
Sad	0	0	88.74	4.82	1.62	4.82	0
Neu	1.23	2.50	1.26	75.89	2.50	15.39	1.23
Fea	1.44	7.28	4.35	2.90	81.13	2.90	0
Bor	0.12	3.43	1.21	14.16	0.09	78.82	2.17
Dis	4.35	0	2.18	0	6.51	2.18	84.78

Table 2. The confusion matrix of recognition results (%) with EKIsomap (*Ang-anger, Joy-joy, Sad-sadness, Neu-neutral, Fea-fear, Bor-boredom, Dis-disgust)

To further investigate the recognition results of different emotions when EKIsomap performs best with 9 embedded features, Table 2 gives the confusion matrix of the recognition results. As can be seen from Table 2, three

emotions - i.e., anger, sadness and disgust - are identified relatively well (in detail, 90.83% for anger, 88.74% for sadness and 84.78% for disgust). Meanwhile, the other four emotions are classified with lower recognition rates (in detail, 65.76% for joy, 75.89% for neutral, 78.82% for boredom and 81.13% for fear).

6. Discussion

From the experimental results on the popular Berlin emotional speech corpus, we can see that the proposed EKIsomap method not only gives better two-dimensional embedded results than KIsomap, but also obtains the highest recognition performance among all seven of the methods used on the speech emotion recognition tasks. This can be explained by two aspects. On the one hand, as a kernel method, EKIsomap is capable of extracting the nonlinear feature information embedded in a high-dimensional data set. On the other hand, as a supervised learning method, EKIsomap can yield a high discriminating power for its low-dimensional embedded data representations for speech emotion classification, due to the used kernel discriminant distance in RKHS.

In addition, and compared with the previously published results [6][7] on the same Berlin speech corpus, the reported accuracy of 80.85% obtained by EKIsomap in our work is highly competitive. In [6], they extracted the acoustic features - such as pitch, energy, HNR and formants, etc. - and used SVM to yield the highest accuracy of 72.3%. In [7], they employed the segment-based approach method and obtained the best accuracy of 75.5% with SVM.

7. Conclusion

Dimensionality reduction is an important strategy when performing feature data processing on the speech emotion recognition tasks. In this paper, a new nonlinear dimensionality reduction method - called 'EKIsomap' - is proposed and applied for speech emotion recognition in human-robot interaction. The experimental results confirm the promising performance of the proposed method and give rise to a strong possibility of applying our speech emotion recognition system to human-robot interaction.

It is worth pointing out that many previous studies [5-9] testify to the performance of speech emotion recognition systems on the offline emotional speech database, such as the popular Berlin database as used in this work. In our future work, it will be an interesting and challenging task to develop an online real-time speech emotion recognition system for human-robot interaction.

8. Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants No.61203257 and No.61272261, and the Zhejiang Provincial Natural Science Foundation of China under Grants No.Z1101048 and No.Y1111058.

9. References

[1] Picard R (1997) Affective computing. MIT Press, Cambridge, USA.

[2] Kulic D, Croft EA (2007) Affective state estimation for human-robot interaction. IEEE Transactions on Robotics. 23 (5):991-1000.

[3] Park JS, Kim JH, Oh YH (2009) Feature vector classification based speech emotion recognition for service robots. IEEE Transactions on Consumer Electronics. 55 (3):1590-1596.

[4] Samani HA, Saadatian E (2012) A Multidisciplinary Artificial Intelligence Model of an Affective Robot. International Journal of Advanced Robotic Systems. 9:1-11.

[5] Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing. 13 (2):293-303.

[6] Schuller B, Seppi D, Batliner A, Maier A, Steidl S (2007) Towards more reality in the recognition of emotional speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07), Honolulu, Hawai'i, USA, pp. 941-944.

[7] Shami M, Verhelst W (2007) An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Communication. 49 (3):201-212.

[8] Luengo I, Navas E, Hernaez I (2010) Feature Analysis and Evaluation for Automatic Emotion Identification in Speech. IEEE Transactions on Multimedia. 12 (6):490-501.

[9] Yun S, Yoo C (2012) Loss-scaled Large Margin Gaussian Mixture Models for Speech Emotion Classification. IEEE Transactions on Audio, Speech, and Language Processing. 20 (2):585-598.

[10] Jolliffe IT (1986) Principal component analysis. Second edition. Springer, New York.

[11] You M, Chen C, Bu J, Liu J, Tao J (2006) Emotional Speech Analysis on Nonlinear Manifold. In: 18th International Conference on Pattern Recognition (ICPR 2006), Hong Kong, pp. 91-94.

[12] Kim J, Lee S, Narayanan S (2010) An exploratory study of manifolds of emotional speech In: 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2010), Dallas, Texas, USA, pp. 5142-5145.

[13] Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science. 290 (5500):2323-2326.

[14] Tenenbaum JB, Silva Vd, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science. 290 (5500):2319-2323.

[15] Ramırez J, Segura JC, Benıtez C, De La Torre A, Rubio A (2004) Efficient voice activity detection algorithms using long-term speech information. Speech communication. 42 (3):271-287.
[16] Choi H, Choi S (2007) Robust kernel isomap. Pattern Recognition. 40 (3):853-862.
[17] Scholkopf B, Smola A, Muller K (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural computation. 10 (5):1299-1319.
[18] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: Interspeech-2005, Lisbon, Portugal, pp. 1-4.