Journal of Southeast University (English Edition) Vol. 32, No. 2, pp. 151 - 157 June 2016 ISSN 1003—7985

Speech emotion recognition via discriminant-cascading dimensionality reduction

Wang Rugang[1,2] Xu Xinzhou[1] Huang Chengwei[1] Wu Chen[1] Zhang Xinran[1] Zhao Li[1]

([1] Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China)
([2] College of Information Engineering, Yancheng Institute of Technology, Yancheng 224051, China)

Abstract: In order to accurately identify speech emotion information, the discriminant-cascading effect in dimensionality reduction of speech emotion recognition is investigated. Based on the existing locality preserving projections and graph embedding framework, a novel discriminant-cascading dimensionality reduction method is proposed, which is named discriminant-cascading locality preserving projections (DCLPP). The proposed method specifically utilizes supervised embedding graphs and it keeps the original space for the inner products of samples to maintain enough information for speech emotion recognition. Then, the kernel DCLPP (KDCLPP) is also proposed to extend the mapping form. Validated by the experiments on the corpus of EMO-DB and eNTERFACE'05, the proposed method can clearly outperform the existing common dimensionality reduction methods, such as principal component analysis (PCA), linear discriminant analysis (LDA), locality preserving projections (LPP), local discriminant embedding (LDE), graph-based Fisher analysis (GbFA) and so on, with different categories of classifiers.

Key words: speech emotion recognition; discriminant-cascading locality preserving projections; discriminant analysis; dimensionality reduction

doi: 10.3969/j.issn.1003-7985.2016.02.004

Speech emotion recognition (SER) is a novel research direction, which is an important branch of affective computation. In the application of human-computer interaction, speech information is processed by machines automatically using appropriate SER technology. Some poor conditions also appeal for the use of SER due to the compressive advantages of speech and acoustic signals. In addition, we can formulate speech emotion recognition as a generalized problem in which the original extracted features include many "improper" factors. These factors may be useful for other recognition tasks, e.g., speaker classification, automatic speech recognition etc., but they can affect the performance of the SER system. Currently, some valid research has been processed with various subtopics[1-9]. These works used basic machine learning algorithms to explore suitable features for speech emotion recognition. However, these methods mostly rely on the experimental results (accuracy etc.), ignoring the inner structure of the training data with existing original speech emotion features. Based on the psychological hypothesis and experiments, speech emotions can be generally represented by the low-dimensional feature space[10]. For these purposes, we use dimensionality reduction methods to solve speech emotion recognition. Dimensionality reduction methods have been widely investigated recently. In the research of dimensionality reduction, the existing classical methods include principal component analysis (PCA), linear discriminant analysis (LDA), linear discriminant projections (LDP) and so on[11-13]. In these algorithms, graph learning or similar form-based methods have a large proportion. Some of other subspace learning and component analysis methods can also be represented by the graph learning framework or the framework of regression.

In the research of speech emotion recognition, due to the combination of speaker, language, speech recognition and some other types of features in the original extracted feature space, the importance of supervised information far exceeds the other information, e.g. neighboring structure, linear reconstruction etc. The unsupervised methods without label information are usually of little help in raising the performance of SER systems. Based on the analysis above and our experimental results, we propose a simple and valid method using discriminant-cascading graph construction to make neighboring information useful in SER again. The methods, namely discriminant-cascading locality preserving projections (DCLPP) and kernel DCLPP (KDCLPP), focus on adopting the new space constructed by discriminant-cascading neighbors. Compared with some related existing algorithms[9-14], the proposed methods possess the characteristics as follows. On the one hand, our methods utilize supervised information from the discriminant-cascading structure, compared with the ex-

Received 2015-10-16.

Biographies: Wang Rugang (1975—), male, doctor, associate professor; Zhao Li (corresponding author), male, doctor, professor, zhaoli@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 61231002, 61273266), the Ph.D. Program Foundation of Ministry of Education of China (No. 20110092130004), China Postdoctoral Science Foundation (No. 2015M571637).

Citation: Wang Rugang, Xu Xinzhou, Huang Chengwei, et al. Speech emotion recognition via discriminant-cascading dimensionality reduction [J]. Journal of Southeast University (English Edition), 2016, 32(2): 151 - 157. doi: 10.3969/j.issn.1003-7985.2016.02.004.

isting methods which directly combine discriminant and local information together[14]. On the other hand, compared with the supervised kernel methods[15], our methods adopt original features to avoid the loss of valid information for recognition.

1 Methods

1.1 Preliminaries

We assume that $\boldsymbol{X} = \{x_1, x_2, \ldots, x_N\} \in \mathbf{R}^{n \times N}$ is the normalized training set, where the column i is the training sample $x_i (i = 1, 2, \ldots, N)$ and N is the number of training samples; n is the dimensionality of the original feature space. $\boldsymbol{x} \in \mathbf{R}^{n \times 1}$ is the column vector standing for any of the normalized testing samples. By the processing of dimensionality reduction, the dimensionality of the new feature space turns to be m. $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_N\} \in \mathbf{R}^{m \times N}$ and $\boldsymbol{y} \in \mathbf{R}^{m \times 1}$ are consequently the eventual dimensional-reduced training and testing samples of $\boldsymbol{X}$ and $\boldsymbol{x}$, respectively.

Then, we define the label feature vector of training sample $x_i (i = 1, 2, \ldots, N)$ as a column vector $\boldsymbol{l}_i \in \mathbf{R}^{N_c \times 1}$, where N_c is the number of classes. The c-th element of $\boldsymbol{l}_i$ is equal to 1 when sample i belongs to class c, otherwise it is equal to 0. In our research, each sample belongs to only one class, so only one element of $\boldsymbol{l}_i$ is equal to 1. We further assume that the label set matrix $\boldsymbol{\mathcal{Y}} = \{l_1, l_2, \ldots, l_N\}$. For the other common variables, $\boldsymbol{e} \in \mathbf{R}^{N \times 1}$ is the column vector with all of its elements equal to 1 and $\boldsymbol{I}$ is the identity matrix.

1.2 Graph embedding framework and LPP

Based on the existing research in subspace learning and manifold learning, the graph embedding framework[8] is used to make some similar dimensionality reduction methods in a unified framework. This framework keeps nearly the same forms of optimization and solution methods, but it can provide different categories of data mapping, which is sometimes adopted to connect training and testing data sets. Compared with graph embedding framework, the latest proposed least-square regression framework is more generalized. In spite of this, the graph embedding framework is still worth researching because this framework reveals the apparent relationships between each pair of training samples by constructing relatively appropriate embedding graphs.

First, we show the original optimization form of graph embedding framework as

$$\arg\min_{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{y}=d} \sum_{\substack{i,j \\ i \neq j}} \| y_i - y_j \|^2 \boldsymbol{W}_{ij} = \arg\min_{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{y}=d} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{y} \qquad (1)$$

where d is a constant; $\boldsymbol{B} = \boldsymbol{L}^{\mathrm{p}}$ when the penalty graph is used and $\boldsymbol{B} = \boldsymbol{\Lambda}$ when the scaling factor is used; $\boldsymbol{\Lambda}$ is the diagonal matrix controlling the scale of the training data. $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ and $\boldsymbol{L}^{\mathrm{p}} = \boldsymbol{D}^{\mathrm{p}} - \boldsymbol{W}^{\mathrm{p}}$ are the the Laplacian matrices of intrinsic graph and penalty graph, respectively; $\boldsymbol{W}$ and $\boldsymbol{W}^{\mathrm{p}}$ are the adjacency matrices of intrinsic and penalty graphs. The diagonal matrices $\boldsymbol{D}$ and $\boldsymbol{D}^{\mathrm{p}}$ are the degree matrices of the two graphs. $\boldsymbol{y} \in \mathbf{R}^{N \times 1}$ is the one-dimensional sample set with the i-th row corresponding to training sample i.

For linear mapping, $\boldsymbol{y} \in \boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}$, where $\boldsymbol{a}$ is the linear mapping column vector. For kernelized mapping, $\boldsymbol{y} = \boldsymbol{K}\boldsymbol{\alpha}$, where $\boldsymbol{K}$ is the Gram matrix of training samples and $\boldsymbol{\alpha}$ is the kernelized mapping column vector.

LPP is a typical form of graph embedding. When LPP is in the framework of graph embedding, the adjacency matrix of intrinsic graph $\boldsymbol{W} = \boldsymbol{W}_{\mathrm{LPP}}$, where the element of row i and column j in $\boldsymbol{W}_{\mathrm{LPP}}$ is $(\boldsymbol{W}_{\mathrm{LPP}})_{ij} = \exp\left(-\frac{(x_i - x_j)^2}{t}\right)$ or $(\boldsymbol{W}_{\mathrm{LPP}})_{ij} = 1$ when x_i is in the neighborhood of x_j or x_j is in the neighborhood of x_i, otherwise $(\boldsymbol{W}_{\mathrm{LPP}})_{ij} = 0$, where $t > 0$ is the scaling parameter between each two samples. We assume that $(\boldsymbol{W}_{\mathrm{LPP}})_{ii} = 0$ $(i = 1, 2, \ldots, N)$. Suppose that $\boldsymbol{D}_{\mathrm{LPP}}$ is the degree matrix of $\boldsymbol{W}_{\mathrm{LPP}}$, as is described for $\boldsymbol{D}$ and $\boldsymbol{W}$. Then, the adjacency matrix of penalty graph $\boldsymbol{B} = \boldsymbol{B}_{\mathrm{LPP}} = \boldsymbol{B}_{\mathrm{LPP}}$, as a scaling factor. The mapping for LPP is linear and for KLPP is with a linear mapping matrix and a kernel mapping.

For LDA, the adjacency matrix of the intrinsic graph is

$$\boldsymbol{W}_{\mathrm{LDA}} = \boldsymbol{I} - \sum_{c=1}^{N_c} \frac{1}{n_c} \boldsymbol{e}^c \boldsymbol{e}^{c\mathrm{T}} \qquad (2)$$

where $\boldsymbol{e}^c \in \mathbf{R}^{N \times 1}$ is a column vector with the elements equal to 1 when the corresponding samples belong to class c; n_c is the number of training samples in class c. It can be seen that $\boldsymbol{\mathcal{Y}} = \{l_1, l_2, \ldots, l_N\} = \{e^1, e^2, \ldots, e^{N_c}\}^{\mathrm{T}}$. The penalty graph is $\boldsymbol{B}_{\mathrm{LDA}} = \boldsymbol{I} - \frac{1}{N}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$.

1.3 Proposed discriminant-cascading LPP

The original speech emotion features originate generally from acoustic quality and prosodic factors. Due to the existence of acoustic quality in the features, the features can represent the characteristics of different speakers. The features can also show the differences in various speech contents due to the prosodic factors. Therefore, when examining the original extracted features of speech emotion[9], we can see that the features often include speaker factors (including gender), speech information and other categories of factors. These factors in the original speech emotion features can bring mistaken information during the procedure of unsupervised learning. This means that neighboring construction or some other unsupervised learning fails to achieve a satisfying performance with the original speech emotion features.

Like the manifold structure of training samples in the original feature space, there are also nonlinear structures

in the new space generated by discriminant-cascading ways. As can be seen in Fig. 1, the training samples in the original feature space are shown in the left part, with not ideal distribution of samples. By the process of discriminant dimensionality reduction, training samples in the new feature space often obey a more reasonable distribution, but the samples may obey a nonlinear manifold distribution for each class, as demonstrated in the right part in Fig. 1. Although we can solve this problem by kernel mapping, the choice of kernels is still an existing problem for representing the structure of samples. In addition, kernelization may lead to loss of information since it also changes the original space. Therefore, local structure learning is valid in this condition.

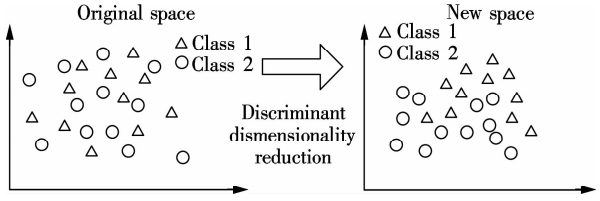

Fig. 1 Training samples in the original space (left part) and in the new space generated by discriminant dimensionality reduction (right part)

Based on the analysis above, we propose a novel method, namely discriminant-cascading LPP (DCLPP), using supervised information to construct neighboring embedding graphs. In this method, the newly generated space by supervised learning can successfully provide better local structure for classification, compared with both purely supervised learning and locality preserved learning. Meanwhile, the method can protect the original information of training data by a large margin.

First, we carry out dimensionality reduction by LDA. With the help of training sample set $\boldsymbol{X} = \{x_1, x_2, \ldots, x_N\}$ and its corresponding label information matrix $\boldsymbol{L} = \{l_1, l_2, \ldots, l_N\}$, the LDA low-dimensional features for the training set can be written as $\boldsymbol{X}^{(\mathrm{LDA})} = \{x_1^{(\mathrm{LDA})}, x_2^{(\mathrm{LDA})}, \ldots, x_N^{(\mathrm{LDA})}\}$. The dimensionality of $\boldsymbol{X}^{(\mathrm{LDA})}$ can be set to be $N_c - 1$ because of the number of minimal eigenvalues for the generalized eigenvalue problem of LDA, which can be easily proved. In order to achieve better performance when confusion of eigenvalues occurs, we use cross-validation to choose dimensionality of $\boldsymbol{X}^{(\mathrm{LDA})}$ as $N_c - 1$ or N_c due to the influence from the trivial eigenvectors.

Using LPP and $\boldsymbol{X}^{(\mathrm{LDA})}$, we formulate the optimization problem as

$$\begin{aligned} &\min \sum_{i,j=1}^{N} (y_i^{(k)} - y_j^{(k)})^2 (\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{LDA})})_{ij} \\ &\text{s. t.} \quad \sum_{i=1}^{N} (\boldsymbol{D}_{\mathrm{LPP}}^{(\mathrm{LDA})})_{ii} (y_i^{(k)})^2 = d \end{aligned} \tag{3}$$

where $\boldsymbol{L}_{\mathrm{LPP}}^{(\mathrm{LDA})} = \boldsymbol{D}_{\mathrm{LPP}}^{(\mathrm{LDA})} - \boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{LDA})}$ and the element of row i and column j for $\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{LDA})}$ is

$$(\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{LDA})})_{ij} = \exp\left(-\frac{[x_i^{(\mathrm{LDA})} - x_j^{(\mathrm{LDA})}]^2}{t} \right) \tag{4}$$

with $i, j = 1, 2, \ldots, N$ and $i \neq j$, when $x_i^{(\mathrm{LDA})}$ is in the neighborhood of $x_j^{(\mathrm{LDA})}$ or $x_j^{(\mathrm{LDA})}$ is in the neighborhood of $x_i^{(\mathrm{LDA})}$, otherwise $(\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{LDA})})_{ij} = 0$. The diagonal matrix $\boldsymbol{D}_{\mathrm{LPP}}^{(\mathrm{LDA})}$ contains the i-th diagonal element: $(\boldsymbol{D}_{\mathrm{LPP}}^{(\mathrm{LDA})})_{ii} = \sum_{j=1}^{N} (\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{LDA})})_{ij}$. Note that the optimal parameter $t > 0$ can be achieved by experience or cross-validation.

Then, we change the form of (3) to construct the new optimization form. The form can be represented as

$$\arg\min_{a} \frac{\boldsymbol{a}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{L}_{\mathrm{LPP}}^{(\mathrm{LDA})} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{D}_{\mathrm{LPP}}^{(\mathrm{LDA})} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{a}} \quad \text{s. t.} \quad \boldsymbol{a}^{\mathrm{T}} \boldsymbol{a} = 1 \tag{5}$$

The optimization of (3) and the cascading LDA can be solved by the generalized eigenvalue problem (GEP), to achieve several optimal mapping directions. The solution method of GEP can be obtained according to the way in Ref. [14].

It is notable that we do not use the low-dimensional features generated by LDA to conduct a second-time dimensionality reduction. The procedure of dimensionality reduction often brings a loss of information, though it can show most of the effective information for classification. In detail, the low-dimensional LDA features can only provide better presentation for each pair of training samples. Using this presentation, we employ original features to obtain the final results of the demensionality reduction.

Next, we propose the kernelized form of DCLPP, namely KDCLPP. We assume that the Gram matrix of the original speech emotion training samples is $\boldsymbol{K} = \boldsymbol{\varphi}^{\mathrm{T}}(X)\boldsymbol{\varphi}(X)$, where $\boldsymbol{\varphi}(X) = \{\boldsymbol{\varphi}(x_1), \boldsymbol{\varphi}(x_2), \ldots, \boldsymbol{\varphi}(x_N)\}$ is the high-dimensional RKHS (reproduced kernel hilbert space) of training set, with each column representing RKHS of its corresponding training sample. In KDCLPP, we adopt the KFD (kernel Fisher discriminant) method to obtain low-dimensional discriminant training samples. The optimization form of KDCLPP is

$$\arg\min_{\alpha} \frac{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{L}_{\mathrm{LPP}}^{(\mathrm{KFD})} \boldsymbol{K}^{\mathrm{T}} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{D}_{\mathrm{LPP}}^{(\mathrm{KFD})} \boldsymbol{K}^{\mathrm{T}} \boldsymbol{\alpha}} \quad \text{s. t.} \quad \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{\alpha} = 1 \tag{6}$$

where $\boldsymbol{L}_{\mathrm{LPP}}^{(\mathrm{KFD})} = \boldsymbol{D}_{\mathrm{LPP}}^{(\mathrm{KFD})} - \boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{KFD})}$ is the Laplacian matrix of $\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{KFD})}$. The corresponding element of $\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{KFD})}$ is

$$(\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{KFD})})_{ij} = \exp\left(-\frac{[x_i^{(\mathrm{KFD})} - x_j^{(\mathrm{KFD})}]^2}{t_0} \right) \tag{7}$$

where if $i \neq j$, $x_i^{(\mathrm{KFD})}$ is in the neighborhood of $x_j^{(\mathrm{KFD})}$ or $x_j^{(\mathrm{KFD})}$ is in the neighborhood of $x_i^{(\mathrm{KFD})}$; otherwise, $(\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{KFD})})_{ij} = 0$. $(\boldsymbol{D}_{\mathrm{LPP}}^{(\mathrm{KFD})})_{ii} = \sum_{j=1}^{N} (\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{KFD})})_{ij}$ is the degree diagonal matrix of $\boldsymbol{W}_{\mathrm{LPP}}^{(\mathrm{KFD})}$. $\boldsymbol{\alpha}$ is the projection vector for KDCLPP.

We use the same Gram matrix $\boldsymbol{K}$ in solving KFD and

KDCLPP to reduce computational costs. As when solving linear situations, it is also necessary to conduct decomposition to decrease the influence from small singular values in solving GEP. The solution of GEP is based on the new feature vectors generated by $\boldsymbol{K}$.

The computational cost of our proposed DCLPP method is equal to the cost combining LDA and LPP. However, the computational complexity of DCLPP is similar to conventional subspace learning methods, when the common solution for GEP is adopted. In the kernelized form, the computational cost results from solving KFD plus KLPP, subtracting a one-time decomposition of the Gram matrix.

2 Experiments

2.1 Speech emotion corpora and original speech emotion features

The speech emotion corpora adopted in the experiments are EMO-DB (Berlin speech emotion database)[16] and eNTERFACE'05[17]. EMO-DB is a widely used corpus including 10 German speakers and 10 different German sentences, with 7 basic emotions, which are fear, disgust, joy, boredom, neutral, sadness and anger. However, some samples are not suitable for experiments in reflecting emotional states. Therefore, we select some samples (494) and delete the other ones in our research. eNTERFACE'05 is a multimodal database containing both video and speech sections. Its speech section includes 42 English speakers. 30 samples are recorded for each speaker, with 5 different sentences in 6 basic emotions (anger, disgust, fear, happy, sadness and surprise).

In the experiments, we adopt the leave-one-subject-out to process training and testing of data samples. Some of the speakers of eNTERFACE'05 are selected from 42 in total in our experiments, with relatively more differences between each other.

We use feature selection to choose 370 original speech emotion features including the statistics of pitch, format, MFCC, zero-cross rate, energy and durance, e.g. relationships between voiced and unvoiced frames; as well as the speech rate for each speech emotion sample by using raw feature selection methods to delete a small number of features which contribute little to emotion classification. It should be clear that we have to keep most features to sufficient that there is sufficient information for the learning process.

2.2 Setting of parameters

The original speech samples, including training and testing samples, are enframed by the Hamming window to achieve a desirable processing performance. Then, original speech emotion features are extracted based on the description above. With the original speech emotion features after feature selection, our proposed method is adopted for obtaining valid features for speech emotion recognition. Finally, classifiers including kNN and NB are at the stage of classification. The number of neighbors for kNN is fixed to be 1, which means that the classifier is 1NN. We do not use the SVM classifier due to its high computational costs and convergence properties in SMO (sequential minimal optimization).

All the generalized eigenvalue problems will be solved according to the methods described in Ref. [14]. The preserved dimensionality in the experiments is set to be 100 in order to evaluate the proposed method under an equal condition. The numbers of neighboring samples in our proposed method and the existing ones (LPP, LDE, GbFA etc.) are set to be the same.

In KDCLPP, we adopt the most commonly used Gaussian kernel to achieve kernelization. The parameter of Gaussian kernel mapping is chosen according to the preserved dimensionality.

2.3 Experimental results and analysis

First, the relationship between reduced dimensionality and recognition rates or UA (unweighted accuracy) is represented in Fig. 2. We only show the recognition rates of the linear form, DCLPP. The other methods used in comparison here include: PCA, LDA, LDE, LDP and GbFA. Figs. 2(a) and (b) show the recognition rates for the corpus of EMO-DB, using 1NN and NB classifiers respectively. Similarly, Figs. 2(c) and (d) reflect the recognition rates for eNTERFACE'05. Fig. 2 indicates that our proposed method can achieve a better performance in speech emotion recognition, compared with the existing linear-mapping-based methods in the experiments. However, the recognition rates of the proposed method are not always preferable. The system containing our methods does not perform well in extremely low dimensionality.

In addition, it can be seen from Fig. 2 that the NB classifier outperforms the 1NN classifier in both databases, which means that the samples in the low-dimensional feature space follow the fixed modeling of distributions (single Gaussian distribution) relatively well. We can also see that eNTERFACE'05 corpus is relatively difficult to deal with in emotion recognition, though it includes fewer categories of emotions than EMO-DB.

Tab. 1 shows the best recognition rates for each dimensionality reduction method and its corresponding dimensionality. With the recognition rates of baseline, the emotions in EMO-DB are easier to recognize than the emotions in eNTERFACE'05. Moreover, the NB classifier seems more effective than the 1NN classifier in speech emotion recognition. It is clear in Tab. 1 that unsupervised dimensionality reduction methods, e.g. PCA and LPP, can barely outperform the classification result of the baseline in the speech emotion recognition experiments.

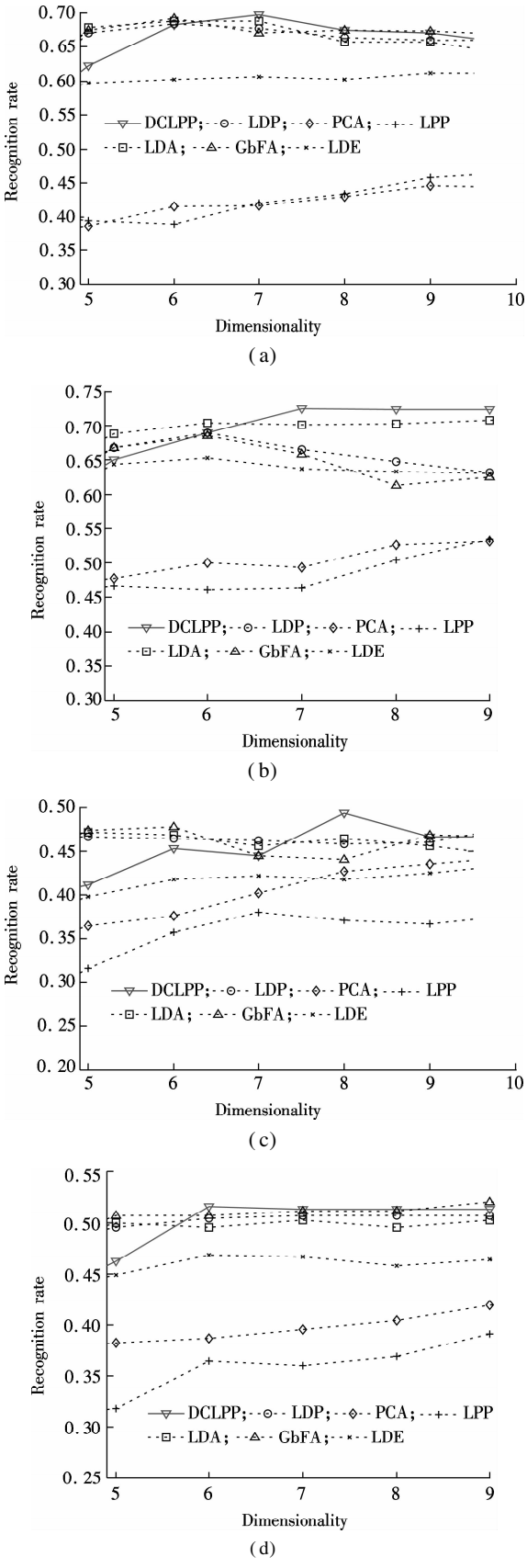

Fig. 2 Recognition rates of EMO-DB and eNTERFACE'05 in low-dimensional conditions using different methods. (a) EMO-DB with 1NN classifier; (b) EMO-DB with NB classifier; (c) eNTERFACE'05 with 1NN classifier; (d) eNTERFACE'05 with NB classifier

In addition, some methods, e. g. LPP and LDE, including neighboring information often fail in the recognition task. Some other ones, e. g. GbFA, can sometimes achieve relatively high recognition rates, but the performances are not stable enough. Although the proposed DCLPP method is not always the best choice, the method is still a desirable one based on the experiments.

Tab. 1 Recognition rates using commonly used dimensionality reduction methods and our proposed method with different classifiers in the corpora of EMO-DB and eNTERFACE'05 %

Corpora	EMO-DB		eNTERFACE'05	
	1NN	NB	1NN	NB
Baseline	54.27	62.70	43.56	45.33
PCA	50.09(13)	60.91(14)	45.33(14)	42.44(14)
LDA	68.81(6)	71.05(10)	47.11(5)	50.22(7)
LDP	68.34(6)	68.86(6)	48.89(12)	50.89(13)
LPP	50.40(14)	60.31(13)	40.67(14)	42.22(10)
LDE	61.30(10)	65.38(6)	47.11(11)	46.89(6)
GbFA	69.14(6)	68.57(6)	47.78(6)	52.00(9)
DCLPP	69.73(7)	72.49(7)	49.33(8)	51.56(6)
KDCLPP	70.32(9)	72.77(10)	50.45(6)	53.11(6)

Note: The number in ( ) is the dimensionality of the corresponding recognition rate.

Furthermore, we investigate the kernelized form of the proposed method KDCLPP which outperforms the linear mapping form. However, the better performance is based on the appropriate choices of the kernels. In other words, we can select the kernels for the proposed KDCLPP in manual or automatic ways. Cross-validation and incremental optimization only using training samples are both valid in selecting appropriate parameters of kernels here.

Then, we investigate the recognition performance of different emotions using the proposed DCLPP method. From Tab. 2, it can be seen that the emotions of sadness and anger are easy to recognize due to the extracted features in reflecting the characteristics of the two emotions. In contrast, the emotion of joy/happy is difficult to recognize. It is not certain that whether the other emotions have good or bad recognition performance based on the classifiers and databases in our experiments.

Tab. 2 Recognition rates for different emotions using classifiers of 1NN and NB in the corpora of EMO-DB and eNTERFACE'05 %

Database	EMO-DB		eNTERFACE'05	
	1NN	NB	1NN	NB
Fear	63.0	74.1	42.7	48.0
Disgust	62.2	70.3	45.3	50.7
Joy/Happy	50.8	55.4	49.3	53.3
Boredom	62.0	69.6		
Neutral	75.6	76.9		
Sadness	81.1	77.4	61.3	60.0
Anger	77.3	74.2	54.7	61.3
Surprise			42.7	36.0

The confusion matrices are shown in Tabs. 3 to 6, when different databases and classifiers are adopted. The

tables show the experimental results (classification of the testing samples) between every two emotions in the corpora. According to the experimental results of EMO-DB in Tab. 3 and Tab. 4, the emotions of joy and anger are more likely to be confused than other emotion pairs. The same conclusion holds when we use the corpus of eNTERFACE'05. In addition, the emotion pairs of anger-sadness, anger-surprise and happy/joy-sadness are easy to recognize.

Tab. 3 Confusion matrix of speech emotions for EMO-DB using 1NN classifier in the experiments

Emotion	Fear	Disgust	Joy	Boredom	Neutral	Sadness	Anger
Fear	34	0	8	2	5	2	3
Disgust	0	23	4	6	3	0	1
Joy	6	1	33	0	3	0	22
Boredom	2	10	0	49	11	7	0
Neutral	1	1	2	11	59	3	1
Sadness	3	0	0	5	2	43	0
Anger	5	0	22	2	0	0	99

Tab. 4 Confusion matrix of speech emotions for EMO-DB using NB classifier in the experiments

Emotion	Fear	Disgust	Joy	Boredom	Neutral	Sadness	Anger
Fear	40	0	6	1	2	2	3
Disgust	1	26	2	5	1	1	1
Joy	7	0	36	1	0	0	21
Boredom	8	3	0	55	3	10	0
Neutral	3	0	2	10	60	2	1
Sadness	5	0	0	7	0	41	0
Anger	7	2	23	1	0	0	95

Tab. 5 Confusion matrix of speech emotions for eNTERFACE'05 using 1NN classifier in the experiments

Emotion	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	41	7	4	13	2	8
Disgust	10	34	10	6	6	9
Fear	9	8	32	5	9	12
Happy	13	7	4	37	0	14
Sadness	4	6	9	1	46	9
Surprise	3	9	9	13	9	32

Tab. 6 Confusion matrix of speech emotions for eNTERFACE'05 using NB classifier in the experiments

Emotion	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	46	3	6	10	5	5
Disgust	8	38	8	5	6	10
Fear	7	9	36	7	9	7
Happy	16	3	5	40	0	11
Sadness	3	7	11	1	45	8
Surprise	4	10	9	12	13	27

3 Conclusion

We investigate the dimensionality reduction methods and propose a novel method to solve the neighboring-based dimensionality reduction. This method is reasonable when the originally extracted features are not sufficiently effective for reflecting the given recognition task. Speech emotion recognition is a typical application, in which the original features cannot effectively provide indications for recognition. In the proposed method, discriminant-cascading thought is combined in the framework of graph embedding and LPP. Validated by experiments on the corpora, our proposed DCLPP and KDCLPP outperform the existing dimensionality reduction and subspace learning methods in most conditions. However, the discriminant-cascading structure does not exactly follow the correct structure of training samples. Therefore, more appropriate discriminant-cascading ways are worth researching in the future.

References

[1] Alonso J B, Cabrera J, Medina M, et al. New approach in quantification of emotional intensity from the speech signal: Emotional temperature[J]. *Expert Systems with Applications*, 2015, **42**(24): 9554−9564. DOI: 10.1016/j.eswa.2015.07.062.

[2] Raptis S, Karabetsos S, Chalamandaris A, et al. A framework towards expressive speech analysis and synthesis with preliminary results[J]. *Journal on Multimodal User Interfaces*, 2015, **9**(4): 387−394.

[3] Kantrowitz J T, Hoptman M J, Leitman D I, et al. Neural substrates of auditory emotion recognition deficits in schizophrenia. [J]. *Journal of the Society for Neuroscience*, 2015, **35**(44): 14909−14921. DOI: 10.1523/JNEUROSCI.4603−14.2015.

[4] Mao Q, Dong M, Huang Z, et al. Learning salient features for speech emotion recognition using convolutional neural networks[J]. *IEEE Transactions on Multimedia*, 2014, **16**(8): 2203−2213. DOI: 10.1109/tmm.2014.2360798.

[5] Arruti A, Cearreta I, Alvarez A, et al. Feature selection for speech emotion recognition in Spanish and Basque: On the use of machine learning to improve human-computer interaction. [J]. *Plos One*, 2014, **9**(10): e108975. DOI: 10.1371/journal.pone.0108975.

[6] Ooi C S, Seng K P, Ang L M, et al. A new approach of audio emotion recognition[J]. *Expert Systems with Applications*, 2014, **41**(13): 5858−5869. DOI: 10.1016/j.eswa.2014.03.026.

[7] Yan J. Speech emotion recognition based on sparse representation[J]. *Archives of Acoustics*, 2013, **38**(4): 465−470. DOI: 10.2478/aoa-2013−0055.

[8] Xu X, Huang C, Wu C, et al. Graph learning based speaker independent speech emotion recognition[J]. *Advances in Electrical & Computer Engineering*, 2014, **14**(2): 17−22. DOI: 10.4316/aece.2014.02003.

[9] Xu X, Deng J, Zheng W, et al. Dimensionality reduction for speech emotion features by multiscale kernels[C]// *Annual Conference of International Speech Communication Association.* Dresden, Germany, 2015: 1532−1536.

[10] Zha C, Zhang X R, Zhao L, et al. Speaker-independent speech emotion recognition based multiple kernel learning of collaborative representation[J]. *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*, 2016, **99**(3): 756−759. DOI: 10.

1587/transfun. e99. a. 756.
[11] Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding[J]. *Science*, 2000, **290**: 2323 - 2326. DOI: 10. 1126/science. 290. 5500. 2323.
[12] He X, Niyogi P. Locality preserving projections[J]. *Advances in Neural Information Processing Systems* 16 (NIPS 2003). Vancouver and Whistle, Canada, 2003.
[13] Cui Y, Fan L. A novel supervised dimensionality reduction algorithm: Graph-based Fisher analysis[J]. *Pattern Recognition*, 2012, **45**(4): 1471 - 1481. DOI: 10. 1016/j. patcog. 2011. 10. 006.
[14] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C]// *Advances in Neural Information Processing Systems* 14 (NIPS 2001). Vancouver, Canada, 2001.
[15] Yu X, Wang X, Liu B. Supervised kernel neighborhood preserving projections for radar target recognition[J]. *Signal Processing*, 2008, **88**(9): 2335 - 2339. DOI: 10. 1016/j. sigpro. 2007. 11. 015.
[16] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]//*Eurospeech, European Conference on Speech Communication and Technology*. Lisbon, Portugal, 2005: 1517 - 1520.
[17] Martin O, Kotsia I, Macq B. The eNTERFACE'05 audio-visual emotion database[C]//22*nd International Conference on Data Engineering Workshops*. Atlanta, GA, USA, 2006.

基于级联降维判别的语言情感识别

王如刚[1,2] 徐新洲[1] 黄程韦[1] 吴 尘[1] 张昕然[1] 赵 力[1]

([1] 东南大学水声信号处理教育部重点实验室,南京 210096)
([2] 盐城工学院信息工程学院,盐城 224051)

摘要:为了准确地识别语音情感信息,研究了语音情感识别的降维中判别级联效应. 基于现有的局部投影算法和图形嵌入理论,提出了一种新型判别分析算法,即 DCLPP 算法. 为了能够对语音情感识别保持足够的信息,该算法利用嵌入图形为样本的内部特点保留了原始空间. 然后,为了扩展映射形式,提出了一种 kernel dCLPP (KDCLPP)的方法. 在 EMO-DB 和 eNTERFACE'05 情感语音数据库上对该算法进行了验证,结果表明,所提算法可明显地超越现有的常用主成成分分析(PCA)、线性判别分析(LDA)、局部保持投影(LPP)、局部鉴别嵌入(LDE)和图优化的 Fisher 判别分析(GbFA)等判别分析算法,这些算法都有不同类型的分类器.

关键词:语音情感识别;级联降维的保局投影算法;判别分析;降维
中图分类号:TN911. 72