Online ISSN 1848-3380, Print ISSN 0005-1144
ATKAFF 57(1), 37–45(2016)

Xinzhou Xu, Chengwei Huang, Chen Wu, Li Zhao

Locally Discriminant Diffusion Projection and Its Application in Speech Emotion Recognition

DOI 10.7305/automatika.2016.07.853
UDK 004.934‘1.021:159.942.5

Original scientific paper

The existing Diffusion Maps method brings diffusion to data samples by Markov random walk. In this paper, to provide a general solution form of Diffusion Maps, first, we propose the generalized single-graph-diffusion embedding framework on the basis of graph embedding framework. Second, by designing the embedding graph of the framework, an algorithm, namely Locally Discriminant Diffusion Projection (LDDP), is proposed for speech emotion recognition. This algorithm is the projection form of the improved Diffusion Maps, which includes both discriminant information and local information. The linear or kernelized form of LDDP (i.e., LLDDP or KLDDP) is used to achieve the dimensionality reduction of original speech emotion features. We validate the proposed algorithm on two widely used speech emotion databases, EMO-DB and eNTERFACE’05. The experimental results show that the proposed LDDP methods, including LLDDP and KLDDP, outperform some other state-of-the-art dimensionality reduction methods which are based on graph embedding or discriminant analysis.

Key words: diffusion maps, graph embedding framework, locally discriminant diffusion projection, speech emotion recognition

Lokalno diskriminantna projekcija difuzije i njena primjena za prepoznavanje emocionalnog stanja iz govornog signala. Postojeće metode mapiranja difuzije u uzorke podataka primjenjuju Markovljevu slučajnu šetnju. U ovom radu, kako bismo pružili općenito rješenje za mapiranje difuzije, prvo predlažemo generalizirano okruženje za difuziju jednog grafa, zasnovano na okruženju za primjenu grafova. Drugo, konstruirajući ugrađeni graf, predlažemo algoritam lokalno diskriminantne projekcije difuzije (LDDP) za prepoznavanje emocionalnog stanja iz govornog signala. Ovaj algoritam je projekcija poboljšane difuzijske mape koja uključuje diskriminantnu i lokalnu informaciju. Linearna ili jezgrovita formulacija LDDP-a (i.e., LLDDP ili KLDDP) koristi se u svrhu redukcije dimenzionalnosti originalnog skupa značajki za prepoznavanje emocionalnog stanja iz govornog signala. Predloženi algoritam testiran je nad dvama široko korištenim bazama podataka za prepoznavanje emocionalnog stanja iz govornog signala, EMO-DB i eNTERFACE’05. Eksperimentalni rezultati pokazuju kako predložena LDDP metoda, uključujući LLDDP i KLDDP, pokazuje bolje ponašanje od nekih drugih najsuvremenijih metoda redukcije dimenzionalnosti, zasnovanim na ugrađenim grafovima ili analizi diskriminantnosti.

Ključne riječi: mapa difuzije, okruženje s ugrađenim grafom, lokalno diskriminantna projekcija difuzije, prepoznavanje emocionalnog stanja iz govornog signala

1 INTRODUCTION

Speech emotion recognition (SER) [1-5] is a brand new field of study in machine learning and affective information processing. Speech emotion recognition may innovate many applications in human-computer interaction (HCI). For instance, negative emotion detection may help us automatically evaluate people’s work attitude in the service industry. So far, many researchers believe that emotion space, including speech emotion feature space, can be represented by a small number of features. However, most of the traditional methods in speech emotion feature extraction are merely based on some empirical prior knowledge, without exploring the possible bias for feature extraction. That may lead to the problems of redundant features or ‘curse of dimensionality’. Thus, dimensionality reduction methods are necessary in speech emotion recognition to solve the problems above.

Currently, manifold learning methods, spectral graph learning algorithms and some classic discriminant analysis or regression methods related have been proposed to solve dimensionality reduction problems in machine learning, e.g. LE (LPP) [6-7], LLE [8], DM (Diffusion Maps) [9-12], Isomap [13], MFA (LDE) [14-15], SDE [16], MLE [17] etc. In addition, some algorithm frameworks, graph

embedding[14] and least-squares[18] frameworks, have been proposed recently, and successfully applied to nonlinear dimensionality reduction and manifold learning. In the field of speech emotion recognition, some basic researches [19-20] have been conducted by manifold learning based methods.

In speech emotion recognition, compared with the classification of faces or expressions, there are a large number of the outliers because of the original feature space. The outliers not only influence the structure of data space, but also affect recognition performance. The method of DM provides a way to constrain the outliers by random walk in the embedding graph of the data. Depending on the original solution form of DM, we propose the extended graph embedding framework in the condition of diffusion.

In this paper, the existing Diffusion Maps method is reviewed first. A generalized graph embedding framework, namely generalized single-graph-diffusion embedding framework, is then proposed. Based on this framework, we propose the method of Locally Discriminant Diffusion Projection (LDDP), which combines graph embedding framework and the process of diffusion together. Then, the linear and kernelized forms [21] of LDDP are adopted to achieve dimensionality reduction in speech emotion recognition.

The rest of this paper is organized as follows. Section 2 introduces Diffusion Maps and the basic theory of the proposed algorithm. In Section 3, the experiments for methods' comparison are shown. Section 4 is the conclusions of this paper.

2 METHODS

2.1 Diffusion Maps and Graph Embedding Framework

Diffusion Maps is originally proposed in [9-10] to improve learning performance by 'coarse-graining' structure on samples. The method of DM brings a solution in controlling noise data and it can also achieve different scales of a given data set.

Given a symmetric matrix $W \in \Re^{N\times N}$, where N is the number of training samples, related to the adjacency matrix of the graph whose elements can reflect the weights or similarity between each two samples. The elements of row i and column j in Wobey: $W_{ij} = W_{ji} \geq 0$. The degree diagonal matrix is $D = diag(d_1, d_2, ..., d_N)$, where the diagonal elements $d_i = \sum_{k=1}^{N} W_{ik}$. For the convenience of description, we assume that all the diagonal elements in D are positive. The transition probability matrix is $P = D^{-1}W$, which means the diffusion probability to a give data point from each point.

The right and left eigenvector problem of P can be respectively written as (1).

$$\begin{cases} P\psi_j = \lambda_j\psi_j, \ \ right\ eigenvector, \\ \phi_j^T P = \lambda_j\phi_j^T (P^T\phi_j = \lambda_j\phi_j), \ \ left\ eigenvector. \end{cases} \tag{1}$$

We assume that the number of diffusion(random walk) steps is $t = 1, 2, 3, ...$

According to the idea of diffusion maps in [9-10], the feature mapping of a certain sample x in the original feature space can be represented as the form of (2). In addition, the diffusion distance of two data points x and z is shown as (3).

$$x \to \Psi_t(x) = \Lambda^t \left[\psi_1(x), \psi_2(x), ..., \psi_{q(t)}(x)\right]^T = \Lambda^t\Psi(x)\,, \tag{2}$$

$$D_t^2(x,z) \approx \|\Psi_t(x) - \Psi_t(z)\|^2 = \sum_{j=1}^{q(t)} \lambda_j^{2t}[\psi_j(x) - \psi_j(z)]^2\,, \tag{3}$$

where $\Psi_t(x)$ is the mapping of x corresponding to t-step diffusion, while $\Psi(x)$ is for one-step diffusion. $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_{q(t)})$ is the diagonal matrix with its diagonal elements be $\lambda_1, \lambda_2, ..., \lambda_{q(t)}$. $\lambda_1, \lambda_2, ..., \lambda_{q(t)}$ and $\psi_1(x), \psi_2(x), ..., \psi_{q(t)}(x)$ are the first $q(t)$ maximal nontrivial eigenvalues and their corresponding eigenvectors of P respectively.

The framework of graph embedding was proposed in [14]. The main purpose of this framework is to provide a general framework for most of the manifold learning or subspace learning methods in dimensionality reduction. The graph embedding framework can be represented as (4):

$$\begin{aligned} \arg\min_{y^TBy=d} \sum_{i\neq j} \|y_i - y_j\|^2 W_{ij} = \arg\min_{y^TBy=d} y^TLy \\ (B = L^p = D^p - W^p \quad \text{or} \quad B = \Lambda, \\ L = D - W \\ D_{ij}^p = \begin{cases} \sum_{k=1}^{N} W_{ik}^p, & i = j \\ 0, & i \neq j \end{cases} \\ D_{ij} = \begin{cases} \sum_{k=1}^{N} W_{ik}, & i = j \\ 0, & i \neq j \end{cases}), \end{aligned} \tag{4}$$

where y_i is the low-dimensional feature vector of sample i after dimensionality reduction; W and W^p respectively represent intrinsic graph and penalty graph adjacency matrix; L and L^p are respectively the Laplacian matrix of W and W^p. Λ is a diagonal scaling matrix.

By graph embedding, PCA, LDA, LPP etc. can be unified or transformed into this framework. The linearization

and kernelization of them are also included in the framework. The differences of the graph embedding related algorithm typically depend on designing of graphs, involving intrinsic graph and penalty graph. Therefore, label or some other information, which reflect the relationship between training samples can be included by constructing proper graphs.

2.2 The Extended Graph Embedding Framework of Diffusion Maps

In this section, the form of DM method is modified in order to be calculated under our proposed extended graph embedding framework.

2.2.1 Preliminary Propositions

Theorem 1 and Theorem 2, as well as their proof, are firstly proposed below.

Theorem 1 *The eigenvalues of P and P^t are real, and the eigenvectors of P and P^t are only with real elements.*

Theorem 2. *P and P^t share the same eigenvectors, and P and P^t are with the eigenvalues as Ω and Ω^t respectively (suppose diagonal matrix Ω is with diagonal elements corresponding to the eigenvalues of P).*

Proof: Here we firstly show the proof of Theorem 1. The eigenvalue problem of $P = D^{-1}W$ ($W \in \Re^{N\times N}$) can be represented as (5):

$$P\psi = \lambda\psi \Rightarrow D^{-1}W\psi = \lambda\psi\,. \tag{5}$$

Then, (5) can be written as (6):

$$(D^{-\frac{1}{2}}WD^{-\frac{1}{2}})(D^{\frac{1}{2}}\psi) = \lambda\psi \Rightarrow \tag{6}$$

$$(D^{-\frac{1}{2}}WD^{-\frac{1}{2}})(D^{\frac{1}{2}}\psi) = \lambda(D^{\frac{1}{2}}\psi)\,. \tag{7}$$

Let $D^{\frac{1}{2}}\psi = \hat{\psi}$. (8) can be drawn:

$$(D^{-\frac{1}{2}}WD^{-\frac{1}{2}})\hat{\psi} = \lambda\hat{\psi}\,. \tag{8}$$

The symmetry of $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ makes $P = D^{-1}W$ maintain real eigenvalues and eigenvectors, $\hat{\psi}$. With $\hat{\psi}$ multiplied by $D^{-\frac{1}{2}}$, the eigenvectors of P are proved to be real. Let the eigenvalue diagonal matrix of P be $\Omega = diag(\lambda_1, \lambda_2, ..., \lambda_N)$. Let the eigenvalue diagonal matrix of P^t be $\Theta = diag(\rho_1, \rho_2, ..., \rho_N)$.

$$P^t\xi = \rho\xi \Rightarrow (D^{-1}W)^t\xi = \rho\xi\,. \tag{9}$$

Like what is in (6), (9) can be converted to (10):

$$D^{-\frac{1}{2}}(D^{-\frac{1}{2}}WD^{-\frac{1}{2}})^tD^{\frac{1}{2}}\xi = \rho\xi \tag{10}$$

$$\Rightarrow (D^{-\frac{1}{2}}WD^{-\frac{1}{2}})^t(D^{\frac{1}{2}}\xi) = \rho(D^{\frac{1}{2}}\xi)\,. \tag{11}$$

Being the same as the form for P, the eigenvalues and eigenvectors of P^t are accordingly real.

The proof of Theorem 2 is as follows.

Based on (8), P shares the same eigenvalues with $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. Also, based on (10), P^t shares the same eigenvalues with $(D^{-\frac{1}{2}}WD^{-\frac{1}{2}})^t$. The eigenvectors for P and P^t are the same if $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ and $(D^{-\frac{1}{2}}WD^{-\frac{1}{2}})^t$ are with the same eigenvectors. Let $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ be equal to $\hat{W}$. The relationship of eigenvalues and eigenvectors between $\hat{W}$ and $\hat{W}^t$ are easy to achieve by SVD. Therefore, Theorem 2 is proved. □

2.2.2 Step Diffusion Maps in Graph Embedding Framework

Here we only consider the right eigenvector problem of P in (1). The method of one-step diffusion maps is to calculate the first $q(t)$ maximal nontrivial eigenvalues and their corresponding eigenvectors of P, where $t = 1$ when one-step diffusion is used. According to (1), they can be written as $\lambda_1, \lambda_2, ..., \lambda_{q(t)}$ (commonly represented as λ for each one) and $\psi_1, \psi_2, ..., \psi_{q(t)}$ (commonly represented as ψ) respectively. Thus, based on Theorem 1, the eigenvalue problem of P is expressed as the form of (12).

$$D^{-1}W\psi = \lambda\psi \Rightarrow (D - W)\psi = (1 - \lambda)D\psi\,. \tag{12}$$

With the change of eigenvalue $\eta = 1 - \lambda$, (13) is consequently the form of GEP (Generalized Eigenvalue Problem) for the eigenvalue problem of P.

$$(D - W)\psi = \eta D\psi\,. \tag{13}$$

By the description in [11] or by adding premultiplication term ψ^T into (13), the graph embedding form of one-step diffusion (related to P) is shown in (14) and (15). The first $q(t)$ maximal nontrivial eigenvalues $\lambda_1, \lambda_2, ..., \lambda_{q(t)}$ (or first $q(t)$ minimal nontrivial eigenvalues $\eta_1, \eta_2, ..., \eta_{q(t)}$ in (13)) are also needed to calculate. It is noticeable that the orthogonal constraint is implicitly included in this transformation for one-step diffusion.

$$\arg\min_{\psi} \frac{\psi^T L\psi}{\psi^T D\psi}\,, \tag{14}$$

$$\arg\min_{\psi} \psi^T L\psi \quad s.t. \quad \psi^T D\psi = \varepsilon \quad , \tag{15}$$

where ε is a real constant value. The diagonal matrix D and Laplacian matrix L follows:

$$D_{ij} = \begin{cases} \sum_{k=1}^{N} W_{ik}, & i = j \\ 0, & i \neq j \end{cases} \quad \text{and} \quad L = D - W\,. \tag{16}$$

Obviously, (14) and (15) are both in the framework of graph embedding as (4), yet not containing data mapping calculation.

2.2.3 t-Step Diffusion Maps in Graph Embedding Framework

In diffusion maps, P^t is related to the t-step random walk diffusion. We can solve the diffusion maps problem by calculating the eigenvalues and eigenvectors of P^t. With the help of Theorem 1 and Theorem 2, it is easy to solve the eigenvalue problem of P^t. Therefore, finding the first $q(t)$ minimal eigenvalues and their corresponding eigenvectors for P^t is equivalent to calculating $\lambda_1^t, \lambda_2^t, ..., \lambda_{q(t)}^t$ and $\psi_1, \psi_2, ..., \psi_{q(t)}$ respectively for one-step transition probability matrix P. The eigenvalue and eigenvector, λ and ψ respectively, can be solved by GEP in graph embedding theory, according to (14) and (15). In other words, we can design embedding graphs, yet only for intrinsic graph, in the framework of diffusion maps.

Figure 1 shows the equivalent calculation method of diffusion maps. In Figure 1, the calculation of diffusion maps can be eventually represented as the form of graph embedding solution.

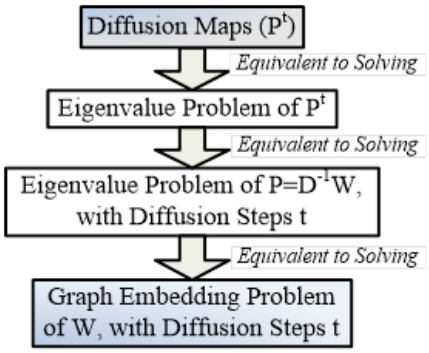

Fig. 1. The equivalent calulation forms of diffusion maps

Then, we consider the data mapping in diffusion maps. For training samples $x_i \ (i = 1, 2, ..., N)$, the mapping $\Psi(x_i)$ in (2) can be directly drawn from the eigenvectors of P. This procedure is equal to the optimizing problems of (14) and (15). For a certain testing sample x, we can attach the mapping $\Psi(\bullet)$ which is calculated based on training to x, as $\Psi(x)$. Though it is available to adopt the similar method used in Isomap[13] to solve this mapping calculation problem, it may bring redundant calculation and ignore the essence of diffusion maps.

Based on the theory of graph embedding, for a testing sample x, the linear and kernelized mapping[21] forms are respectively shown in (17).

$$\Psi(x) = \begin{cases} A^T x, \ \ linear\ mapping\,, \\ \tilde{A}^T \varphi^T(X)\varphi(x) = \tilde{A}^T K_x, \ \ kernel\ mapping\,, \end{cases} \tag{17}$$

where $\varphi(x)$ is high-dimensional mapping of x. The high-dimensional mapping for the training sample set $X = [x_1, x_1, ..., x_N]$ is $\varphi(X) = [\varphi(x_1), \varphi(x_2), ..., \varphi(x_N)]$. K_x in (17) is represented as:

$$\begin{aligned} K_x &= \varphi^T(X)\varphi(x) = [\varphi(x_1), \varphi(x_2), ..., \varphi(x_N)]^T \varphi(x) \\ &= [\kappa(x_1, x), \kappa(x_2, x), ..., \kappa(x_N, x)]^T \end{aligned}, \tag{18}$$

where $\kappa(x_i, x)$ is the kernel function between sample x_i and x. Then, the linear and kernelized mapping forms for P^t are consequently as (19).

$$x \to \begin{cases} \Lambda^t A^T x, \ \ linear\ mapping, \\ \Lambda^t \tilde{A}^T K_x, \ \ kernelized\ mapping. \end{cases} \tag{19}$$

According to the discussions above, we can draw the algorithm in solving diffusion problems using graph embedding framework.

Algorithm 1 *Generalized Single-Graph-Diffusion Embedding Framework.*

The t-step diffusion form for graph embedding with one graph (intrinsic graph [14]) can be solved by the following steps: (1) Given the adjacent matrix W of intrinsic graph, we can get its degree matrix D and Laplacian matrix L according to (16); (2) Solve the optimization problem in (15) and get the first $q(t)$ minimal nontrivial eigenvalues $\eta_1, \eta_2, ..., \eta_{q(t)}$ and their corresponding eigenvectors, which shares the similar form of LPP [7]. Thus, the SSS (Small Sample Size) problem can be solved using the same method in [7];

- For the linear mapping condition, the optimization form is:

$$\arg\min_a a^T X L X^T a \quad s.t. \quad a^T X D X^T a = 1. \tag{20}$$

- For the kernelized mapping condition, the optimization form is:

$$\arg\min_\alpha \alpha^T K L K^T \alpha \quad s.t. \quad \alpha^T K D K^T \alpha = 1. \tag{21}$$

(3) The mapping for any sample x can be represented as (22) in linear condition and (23) in kernelized condition.

$$x \to (I - \Gamma)^t A^T x, \tag{22}$$

$$x \to (I - \Gamma)^t \tilde{A}^T K_x, \tag{23}$$

where the diagonal matrix for the eigenvalues of GEP (15) is $\Gamma = diag(\eta_1, \eta_2, ..., \eta_{q(t)})$. In order to make the eigenvalues match with their corresponding eigenvectors exactly, we use Lagrange multiplier method of (20) or (21) to obtain the respective results. The linear orthogonal projection matrix $A = [a_1, a_2, ..., a_{q(t)}]$. The kernelized orthogonal projection matrix $\tilde{A} = [\alpha_1, \alpha_2, ..., \alpha_{q(t)}]$. Each column

vectors of A and $\tilde{A}$ is corresponding to $\eta_1, \eta_2, ..., \eta_{q(t)}$ respectively. □

According to Algorithm 1, diffusion in learning on a single given graph is easy to be calculated by graph embedding framework. The diffusion process means a separate term based on the original graph embedding framework. It is noticeable that each mapping form is assumed to be able to completely fit every training sample in Algorithm 1. The idea of the proposed framework is shown as Figure 2, where once the training data, mapping form, embedding graph and the number of diffusion steps are given, we can easily draw embedding graph diffusion by using Algorithm 1. The number of diffusion steps determines the combination of eigenvectors and eigenvalues of the solutions.

2.3 Locally Discriminant Diffusion Projection

In this part, the embedding graph for diffusion is designed to solve the learning problems for supervised training samples, since supervised or label information plays an essential role when the original features are not accurate enough in a certain recognition problem. To insert supervised information into spectral graph learning methods, the original form is choosing the embedding graph of LDA. The adjacency matrix of the embedding intrinsic graph of LDA [14] is shown as (24).

$$W_{LDA} = \sum_{c=1}^{N_C} \frac{1}{n_c} e^c e^{cT}, \tag{24}$$

where e^c represents the column vector with the elements, which are corresponding to class c, being equal to 1, otherwise being equal to 0. n_c is the number of samples in class c. N_c is the number of classes.

Theorem 3 *The diffusion procedure has no effect on the embedding graph of LDA.*

It is obvious that the transition probability matrix of LDA, P_{LDA}, is equal to W_{LDA}. Thus, $P_{LDA}^2 = P_{LDA}$ can be simply proved. By induction, $P_{LDA}^t = P_{LDA}$ can be proved.

Because of the reason in Theorem 3, the embedding graph of LDA can not be chosen as the diffusion graph here, since the diffusion process does not affect the embedding graph. However, as an important category of unsupervised information, the local-sample information can be added to make the embedding graph connected. Therefore, we propose the LDDP method, with the adjacency matrix of the embedding intrinsic graph expressed as:

$$W_{LDDP} = (\sum_{c=1}^{N_C} \frac{1}{n_c} e^c e^{cT}) + \tau W_{Local}, \tag{25}$$

where W_{Local} is the adjacency matrix for the local information in the embedding graph, which is chosen the same as what is in LE or LPP [6-7]. The element of row i and column j in W_{Local} is:

$$W_{Local,ij} = \begin{cases} 1, & i \in N_k(j) \quad or \quad j \in N_k(i), \\ 0, & otherwise, \end{cases} \tag{26}$$

where $N_k(i)$ and $N_k(j)$represent the k-nearest-neighbor sets of sample i and j respectively. τ is the fixed weight parameter of local-information embedding graph W_{Local}. In (26), heat kernel mapping also can be used instead of 0-1 neighboring. The Laplacian matrix of W_{Local} is $L_{Local} = D_{Local} - W_{Local}$, where the elements of D_{Local} are:

$$D_{Local,ij} = \begin{cases} \sum_{k=1}^{N} W_{Local,ik}, & i = j, \\ 0, & i \neq j. \end{cases} \tag{27}$$

By taking the embedding graph W_{LDDP} in (25) into Algorithm 1 (replacing the common term W), LDDP algorithm can be constructed. Also, with the projection methods of (20) and (21) respectively, the linear and kernelized form of LDDP, LLDDP and KLDDP can be performed to realize the process of training and test.

2.4 LDDP in Speech Emotion Recognition

It is widely admitted that compared with face recognition and speaker recognition etc., speech emotion recognition relies more on supervised information. Meanwhile, some noise sample points, which could disturb the performance of the algorithm, may also exist in the training stage during speech emotion recognition. So the proposed LLDDP and KLDDP methods are appropriate to be used in the stage of dimensionality reduction in speech emotion recognition.

With the processing of pre-emphasize and enframing (by Hamming window) for each sample, original speech emotion features are extracted. The features are generally divided into 6 categories, including energy [2-5,20], pitch [1-5,19-20], zero-cross rate [3], durance [1-3,5,20], formant [2-3,5,20] and MFCC (Mel Frequency Cepstrum Coefficient) [2-3] features. Most of these features come from statistics of frame features, while others are abstracted from significant prior knowledge which is useful to represent different speech emotion states.

To make the features attached with less redundant factors which are negative for speech emotion classification, some feature selection strategies are used to delete some redundant features in the original speech emotion features. Then, the proposed dimensionality reduction methods, LLDDP and KLDDP, are separately adopted as is discussed in Algorithm 1, using the embedding graph W_{LDDP}. The number of diffusion steps, t, can be decided by cross-validation on training set with the same classifiers which are used in test stage.

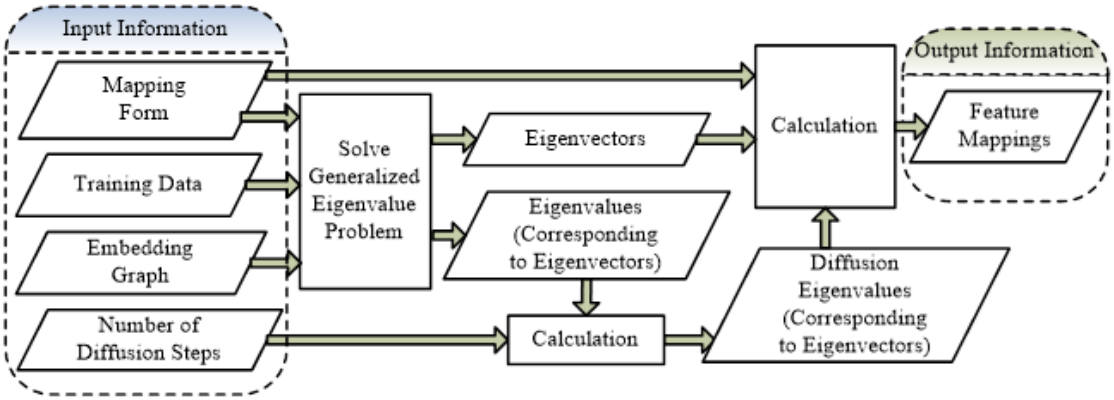

Fig. 2. The idea of Generalized Single-Graph-Diffusion Embedding Framework

3 EXPERIMENTS

3.1 Speech Emotion Corpus

In the experiments, Berlin speech emotion corpus (EMO-DB) [22] and the audio section of eNTERFACE'05 multimodal emotion corpus [23] are used.

The Berlin corpus includes 7 categories of speech emotions, which are neutral, fear, disgust, joy(happiness), boredom, sadness and anger in detail. The corpus adopts 10 persons (5 male and 5 female). Each one reads 10 German short sentences. We choose 494 samples from 900 original ones in the experiments.

In the multimodal database of eNTERFACE'05, 42 subjects from different nations speak English sentences in 6 classes of emotions, including happiness, sadness, fear, disgust, surprise and anger. We only adopt the speech part of the database, without using the expression information. 900 audio samples by 30 different speakers are selected from 1286 ones in the speech emotion experiments.

3.2 Preparations and Parameters

We use 3-fold crossvalidation, in which we partition each corpus into 3 parts with nearly the same sample size. The training and testing cross validation procedures repeat for 25 times in the experiments with random partitioning of samples, considering speaker facts. The detailed original speech emotion features are show in Table 1, where the 'statistics' refers to maximum, minimum, mean, median, standard deviation and range of an utterance formed by frames.

In the feature selection stage, 35 features with relatively low Fisher discriminant values are eliminated from the 408-dimensional original speech emotion features. The diffusion steps are chosen from 2 to 10. The kernel used in KLDDP is Gaussian kernel, with parameters chosen around the number of dimensions after the stage of feature

Table 1. Detailed original speech emotion features

Feature Categories	**Choices of Features**
Energy features	statistics, first-order and second-order flux of energy sequence; statistics of energy sequence's first-order and second-order difference sequence; statistics, first-order and second-order flux of energy sequence respectively with three different frequency bands.
Pitch(F0) features	statistics, first-order and second-order flux of pitch sequence; statistics of pitch sequence's first-order and second-order difference sequence; slope of voiced-frame sequence.
Zero-cross rate features	statistics of zero-cross rate sequence and its first-order and second-order difference sequence.
Durance features	the number of voiced and unvoiced frames and segments; the longest duration of voiced and unvoiced segments; ratio of the number of unvoiced and voiced frames; ratio of the number of unvoiced and voiced segments; speech rate.
Formant (F1-F3) features	statistics of formant frequency sequence and bandwidth sequence, as well as their first-order and second-order difference sequence; first-order and second-order flux of formant frequency sequence.
MFCC features	statistics of MFCC sequences and their first-order difference sequence.

selection, where the parameters are represented as ρ^2 in the kernels $\exp(-|x_i - x_j|^2 / \rho^2)$. x_i and x_j are the column feature vectors of data points i and j respectively.

The selection of diffusion steps and other parameters, including the between-graph parameter in (24) and the kernel parameter, are decided according to cross-validation or experimental results in training sets.

3.3 Experimental Results

The experiments are performed with different random selections of training and test data sets. We show the comparison of the recognition performance using PCA [21], LPP [7], MFA [14], LDA [21] and the proposed methods, LLDDP and KLDDP in Figure 3, with different low dimensionality in Berlin corpus. It is obvious that the proposed LLDDP outperforms the other existing methods using linear mapping especially when the dimensionality is relatively high. In addition, the kernelized form of LDDP can improve recognition rates on the basis of LLDDP. It is noticeable that the best recognition rate of LDA appears on the dimension of the number of classes because of the practical condition, trivial eigenvectors, information in null space and the preprocessing ways. In Figure 3, KLDDP

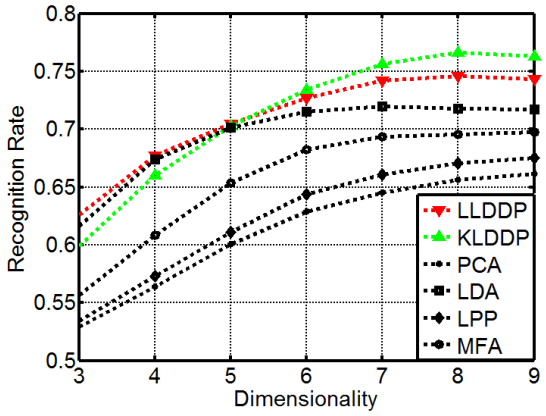

Fig. 3. Recognition rates of EMO-DB in low-dimensional conditions using different methods

turns to be with inferior recognition rates when the reduced dimensionality is low. It is caused by the character of the Gaussian kernel with specific parameter we choose in the experiments.

Then, the average maximal recognition rates of the algorithms as well as their corresponding dimensions when the dimensionality is no higher than 9 using the speech emotion corpus of Berlin and eNTERFACE'05 are represented in Table 2.

According to Tabel 2 and Figure 3, the proposed algorithms, LLDDP and KLDDP are with better performances than some common subspace learning methods in

Table 2. Average maximal low-dimensional recognition rates and their corresponding dimensions using different methods

Methods	EMO-DB (Berlin) Recognition Rates / Dimensions	eNTERFACE'05 Recognition Rates / Dimensions
PCA	66.16% / 9	41.24% / 9
LPP	67.50% / 9	43.49% / 9
LDA	71.96% / 7	53.41% / 5
MFA	69.90% / 9	47.64% / 9
LLDDP	**74.60%** / 8	**55.11%** / 7
KLDDP	**76.59%** / 8	**55.62%** / 8

speech emotion recognition. However, the experiments in eNTERFACE'05 corpus are with lower recognition rates compared with Berlin corpus mainly because of the interferences caused by more speakers and choices of classifiers.

The proposed LDDP methods add weighted eigenvalue factors of GEP to show the random-walk diffusion process. When the diffusion steps turn to increase, each two eigenvalues with larger distance get away from each other more than the pairs with smaller distances. Thus, the diffusion process makes embedding graph be divided into hierarchical blocks or communities. Different numbers of diffusion steps provide different scales of embedding graph.

By the experiments, the baseline recognition rates of Berlin and eNTERFACE'05 corpus are 72.21% and 52.73% respectively, which are similar as the recognition rates of the method of LDA. We can see that the improvement of the recognition rates seems not obvious for the feature extraction methods above. However, the choice of SVM in the stage of classifier may require much more calculation than some weak classifiers, such as k-Nearest Neighbouring and Naive Bayesian classifiers. In addition, the aim of feature extraction is not only the improvement of recognition performance, but also the reduction of transmission efficiency especially in poor conditions of communication channels.

4 CONCLUSIONS

This paper shows the algorithms of Diffusion Maps in the generalized framework of graph embedding. On the basis of this form, combining embedding graphs and diffusion procedures together, the algorithms of Locally Discriminant Diffusion Projection, as well as its linear and kernelized forms, are proposed to solve speech emotion recognition problems. Validated by the experiments, the proposed algorithms can achieve relatively better performance than some existing state-of-the-art methods in

speech emotion recognition. Although the proposed methods perform well in achieving supervised 'coarse-graining' and controlling outliers in speech emotion recognition, many problems, such as the influences from separate speakers or semantics may also have a negative impact on speech emotion recognition.

[1] F. Dellaert, T. Polzin, A. Waibel, "Recognizing emotion in speech," *International Conference on Spoken Language*, Philadelphia, PA, USA, vol. 3, pp. 1970-1973, Oct. 3-6, 1996.

[2] D. Ververidis, C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, pp. 1162-1181, 2006.

[3] B. Schuller, G. Rigoll, "Timing levels in segment-based speech emotion recognition," *International Conference on Spoken Language*, Pittsburgh, PA, USA, pp. 1818-1821, Sep. 17-21, 2006.

[4] P. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 157-183, 2003.

[5] R. Tato, R. Santos, R. Kompe, et al., "Emotional space improves emotion recognition," *International Conference on Spoken Language*, pp. 2029-2032, 2002.

[6] M. Belkin, P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," In *Advances in Neutral Information Processing Systems(NIPS) 14*, Vancouver, British Columbia, Canada, pp. 585-591, Dec. 3-8, 2002.

[7] X. He, P. Niyogi, "Locality preserving projections," In *Advances in Neutral Information Processing Systems(NIPS) 15*, MIT Press, Cambridge, 2003.

[8] S. Roweis, L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000.

[9] R. Coifman, S. Lafon, "Diffusion maps," *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5-30, 2006.

[10] S. Lafon, A. Lee, "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1393-1403, 2006.

[11] R. Socher, M. Hein, "Manifold Learning and Dimensionality Reduction with Diffusion Maps," *Seminar report, Saarland University,* 2008.

[12] J. Liu, Y. Yang, I. Saleemi, et al., "Learning semantic features for action recognition via diffusion maps," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 361-377, 2012.

[13] J. Tenenbaum, V. de Silva, J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000.

[14] S. Yan, D. Xu, B. Zhang, et al., "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, 2007.

[15] H. Chen, H. Chang, T. Liu, "Local discriminant embedding and its variants," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 846-853, 2005.

[16] K. Weinberger, L. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 77-90, 2006.

[17] R. Wang, S. Shan, X. Chen, et al., "Maximal linear embedding for dimensionality reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1776-1792, 2011.

[18] F. De la Torre, "A least-squares framework for component analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1041-1055, 2012.

[19] M. You, C. Chen, J. Bu, et al., "Emotional speech analysis on nonlinear manifold," *IEEE Conf. Pattern Recognition (ICPR)*, Hongkong, vol. 3, pp. 91-94, 2006.

[20] S. Zhang, X. Zhao, B. Lei, "Speech emotion recognition using an enhanced Kernel Isomap for human-robot interaction," *International Journal of Advanced Robotic Systems*, vol. 10, no. 114, pp. 1-7, 2013.

[21] J. Shawe-Taylor, N. Cristianini, *Kernel methods for pattern analysis*, Cambridge University Press, Cambridge, 2004.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, et al., "A database of German emotional speech," *International Conference on Spoken Language*, pp. 1517-1520, 2005.

[23] O. Martin, I. Kotsia, B. Macq, et al., "The enterface'05 audio-visual emotion database," *IEEE Conf. Data Engineering Workshops*, pp. 8-8, 2006.

Xinzhou Xu received his B.S. degree from School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, China, in 2009, and M.E. degree from School of Information Science and Engineering, Southeast University, China, in 2012. He is currently a Ph.D. candidate at Southeast University. His research interests lie in machine learning, pattern recognition, speech signal processing and speech emotion analysis.

Chengwei Huang received B.E. and Ph.D. degrees from Southeast University, China, in 2007 and 2013 respectively. He joined Soochow University, China, in 2013. His research interests include speech signal processing and machine perception.

Chen Wu received B.E. and M.E. degrees from Nanjing University of Aeronautics and Astronautics, China, in 2009 and Southeast University, China, in 2012 respectively. He is currently a Ph.D. candidate in Southeast University. His research interests include signal processing and pattern recognition.

Li Zhao received B.S., M.E. and Ph.D. degrees from Nanjing University of Aeronautics and Astronautics, China, in 1982, Southeast University, China, in 1988, and Kyoto Institute of Technology, in 1998, respectively. He is currently with School of Information Science and Engineering in Soochow University.

AUTHORS' ADDRESSES

Xinzhou Xu,
Chen Wu,
Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education,
Southeast University,
Nanjing, China
email: xinzhouxu@seu.edu.cn, 230129135@seu.edu.cn

Chengwei Huang, Ph.D.,
School of Physical Science and Technology,
Soochow University,
Suzhou, China
email: huangcwx@126.com

Li Zhao, Ph.D.,
Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education,
Key Laboratory of Child Development and Learning Science of Ministry of Education,
Soochow University,
Nanjing, China
email: zhaoli@seu.edu.cn

Received: 2014-04-30
Accepted: 2015-05-18