Advances in Electrical and Computer Engineering **Volume 14, Number 2, 2014**

Graph Learning Based Speaker Independent Speech Emotion Recognition

Xinzhou XU[1], Chengwei HUANG[2], Chen WU[1], Qingyun WANG[1], Li ZHAO[1,3]
[1]*Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing, 210096, China*
[2]*School of Physical Science and Technology, Soochow University, Suzhou, 215006, China*
[3]*Key Laboratory of Child Development and Learning Science (Ministry of Education), Research Center for Learning Science, Southeast University, Nanjing, 210096, China*
230129137@seu.edu.cn

***Abstract*—In this paper, the algorithm based on graph learning and graph embedding framework, Speaker-Penalty Graph Learning (SPGL), is proposed in the research of speech emotion recognition to solve the problems caused by different speakers. Graph embedding framework theory is used to construct the dimensionality reduction stage of speech emotion recognition. Special penalty and intrinsic graphs of the graph embedding framework is proposed to penalize the impacts from different speakers in the task of speech emotion recognition. The original speech emotion features are extracted by various categories, reflecting different characteristics of each speech sample. According to the experiments in speech emotion corpus using different classifiers, the proposed method with linear and kernelized mapping forms can both achieve relatively better performance than the state-of-the-art dimensionality reduction methods.**

***Index Terms*—speech emotion recognition, speaker penalty graph learning, graph embedding framework, dimensionality reduction.**

I. INTRODUCTION

The research of speech emotion recognition (SER) develops rapidly with the application needs. The applications of call center and human-computer interaction call for the new improvement of SER, to achieve natural interaction between human beings and machine or automatic emotion processing by computers. Because of the demands above, many research works have been processed[1-5]. However, the previous works are mostly focus on the choices of features based on prior knowledge or the experiments programmed manually, neglecting fully using training data. Additionally, the original extracted speech emotion features include too much redundant information, in which some acoustic features used in speakers recognition may mask helpful factors for emotion recognition. Currently, some of research works[6-8] are on the methods to reduce the impact from different speakers in the stage of feature selection, which can verify the corresponding only by the limited experimental results. To improve the recognition performance of a speech emotion recognition system, the information speaker labels is added to penalize the influence of speaker features.

Manifold learning methods are usually adopted in the stage of dimensionality reduction to show the intrinsic structure of data. The graph learning based manifold learning methods[9-15], such as LE (Laplacian Eigenmaps) or LPP (Locality Preserving Projections)[9-10], LLE (Locally Linear Embedding)[11], DM (Diffusion Maps)[12], Isomap[13] and LDE (Locally Discriminant Embedding) or MFA (Marginal Fisher Analysis)[14-15] can be represented as the unified graph embedding framework[15], least-squares framework[16] or their extensive forms. When those methods are used in the field of speech emotion recognition[17-18], the supervised informaition should be included because of the difficulties in accurately extracting appropriate speech emotion features without the help of label information. The information merely drawn from the feature vectors of training samples themselves may mislead the purpose of recognition for which we expect.

This paper is aim to construct a graph-learning-based framework to reduce the impact on speech emotion recognition caused by redundant speaker recognition features. The new effective methods, Speaker-Penalty Graph Learning (SPGL) and its linear and kernelized data mapping forms, Linear SPGL (LSPGL) and Kernel SPGL (KSPGL), are proposed in the paper to elevate the recognition rates in SER. The proposed methods can obviously raise the performance of a speech emotion recognition system with the additional speaker label of training samples. The intrinsic and penalty graphs in the methods of SPGL can inhibit the speaker-related factors which could hinder correct classification in speech emotion recognition.

The rest of this paper is organized as follows. Section 2 shows the basic theory of graph embedding learning methods. Then in this section, the proposed algorithm SPGL is described in detail. In Section 3, the experiments for recognition rates based on the proposed LSPGL and KSPGL are processed compared with some conventionally adopted dimensionality reduction algorithms using public speech emotion corpus.

II. GRAPH EMBEDDING FRAMEWORK

The framework of graph embedding was proposed in [15], in which some discriminant analysis, component analysis and manifold learning methods can be represented as the form of Graph embedding. Different from previous methods, graph embedding framework lets the dimensionality methods be a framework including 3 stages. The first one is construction of embedding graphs. Then, mapping forms

This work was supported by the Natural Science Foundation of China under Grant No. 61231002 and No. 61273266, Doctoral Fund of Ministry of Education of China under Grant No. 20110092130004.

Digital Object Identifier 10.4316/AECE.2014.02003

which connect training and test samples are worth considering. Additionally, the framework is with a relatively fixed optimization method. Although the thoughts of those different dimensionality reduction methods are not the same, the 3-stage framework above, including its extensive forms, is still enough to describe most standard ways of dimensionality reduction.

The optimization of graph embedding frameworks is shown as (1).

$$\arg\min_{y^T By=d} \sum_{\substack{i,j \\ i\neq j}} \left\| y_i - y_j \right\|^2 W_{ij} = \arg\min_{y^T By=d} y^T L y \tag{1}$$

$$(B = L^p \ or \quad B = \Lambda)$$

where the adjacency matrices of intrinsic and penalty embedding graphs are respectively W and W^p, with the Laplacian matrices of $L = D - W$ and $L^p = D^p - W^p$ respectively. d is the fixed scale-controlling constant value. y is the column vector with each element y_i indicating the feature of training sample i, where $i = 1,2,...,N$. N is the number of training samples. The diagonal elements of diagonal matrices D and D^p are the degrees of corresponding sample nodes. Λ in (1) is the diagonal matrix which controls scales of y.

In addition, some generalized forms for graph embedding learning can be proved to be able to signify more kinds of dimensionality reduction forms, such as Diffusion Maps[12] etc.. The coming up with least-squares frameworks[16] can be seen as the extension of graph embedding as well.

III. SPEAKER PENALTY GRAPH LEARNING

In this part, we start our proposed methods, Speaker Penalty Graph Learning (SPGL), with some unified definitions of variables. The training sample set is $X = [x_1, x_2, ..., x_N]$, where N is the number of training samples. Each training sample $x_i \in \Re^{n\times 1}$ ($i = 1,2,...,N$) is with two kinds of labels, which are emotion category labels $l_i^E = \{1,2,...,N_E\}$ and speaker labels $l_i^S = \{1,2,...,N_S\}$, where the numbers N_E and N_S indicate the categories of speech emotions and speakers respectively. n is the dimensionality of the feature space before the stage of dimensionality reduction. The dimensionality-reduced training set can be represented as $Y = [y_1, y_2, ..., y_N] = [y^{(1)}, y^{(2)}, ..., y^{(m)}]^T$, where $y_i \in \Re^{m\times 1}$ ($i = 1,2,...,N$) and $y^{(j)} \in \Re^{N\times 1}$ ($j = 1,2,...,m$). m is the dimensionality of Y.

First, because of the not obvious features in representing emotion factors, we hope any two samples of the training set with the same emotion label to be with small distance, while each pair of two samples with different emotion labels is far away from each other in the newly generated feature space. Suppose the adjacency matrix of the graph is:

$$W_{LDA} = \sum_{c=1}^{Nc} \frac{1}{n_c} e^c e^{cT} \tag{2}$$

which is the same as the intrinsic graph of graph-embedding-form LDA (Linear Discriminant Analysis) or FDA (Fisher Discriminant Analysis)[15,19], where $e^c \in \Re^{N\times 1}$ is the column vector with the elements which are corresponding to emotion class c being equal to 1, otherwise they are equal to 0. n_c is the number of samples in class c. N_c is the number of emotion classes.

We hope that each pair of samples in the same emotion classes are with small distances:

$$\min \sum_{i,j=1}^{N} (y_i^{(k)} - y_j^{(k)})^2 (W_{LDA})_{ij} \tag{3}$$

where $(W_{LDA})_{ij}$ is the element for row i and column j of W_{LDA}. $y_i^{(k)}$ is the corresponding element of sample i for the new dimension k.

According to the derivation in [9-10], the optimization form of (3) can be represented as (4), with the scale condition.

$$\min y^T L_{LDA} y \tag{4}$$

The Laplacian matrix of W_{LDA} obeys:

$$L_{LDA} = D_{LDA} - W_{LDA},\ (D_{LDA})_{ij} = \begin{cases} \sum_{k=1}^{N} (W_{LDA})_{ik}, & i = j \\ 0, & i \neq j \end{cases} \tag{5}$$

For the penalty part of the same speakers with different speech emotion labels, the element W_{ij}^{ps} of the embedding penalty graph is designed as (6), where the penalty element W_{ij}^{ps} are equal to 1 when the neighboring samples i and j maintain the same speaker label while they are included by different emotion categories. It means that we can constrain the speaker-related features which are simultaneously not helpful in speech emotion recognition.

$$W_{ij}^{ps} = \begin{cases} 1, & l_i^S = l_j^S, l_i^E \neq l_j^E \quad and \quad i \in N_k(j)\ or\ j \in N_k(i) \\ 0, & otherwise \end{cases} \tag{6}$$

The matrix form of W^{ps} can be consequently represented as (7).

$$W^{ps} = [\sum_{c_S=1}^{N_S} e^{c_S} e^{c_S T} - \sum_{c=1}^{N_c} \sum_{c_S=1}^{N_S} (e^{c_S} \circ e^c)(e^{c_S} \circ e^c)^T] \circ W_{kNN} \tag{7}$$

where the ' $\circ$ ' means the element-wise multiplication between two matrices. $e^{c_S} \in \Re^{N\times 1}$ is the column vector with the elements which are corresponding to speaker class c_S being equal to 1, otherwise they are equal to 0. W_{kNN} is an $N \times N$ matrix with $(W_{kNN})_{ij} = (W_{kNN})_{ji} = 1$ when sample $i \in N_k(j)$ or sample $j \in N_k(i)$, otherwise the corresponding elements are equal to 0.

Like what is described for W, the distance is expected to be large when every pair of two samples is with the same speaker label while their emotion labels are different. The original form of it can be written as (8).

$$\max \sum_{i,j=1}^{N} (y_i^{(k)} - y_j^{(k)})^2 W_{ij}^{ps} \tag{8}$$

Therefore, the standard form of the penalty part of speaker factors is:

$$\max y^T L^{ps} y \tag{9}$$

The Laplacian matrix of the embedding graph W^{ps} is show in (10).

$$L^{ps} = D^{ps} - W^{ps},\ D_{ij}^{ps} = \begin{cases} \sum_{k=1}^{N} W_{ik}^{ps}, & i = j \\ 0, & i \neq j \end{cases} \qquad (10)$$

Then, another optimization term is given to make the features related to different speakers with more attentions. Thus, the adjacency matrix W^{is} of $L^{is} = D^{is} - W^{is}$, where the diagonal element i of diagonal matrix D^{is} is $D_{ii}^{is} = \sum_{k=1}^{N} W_{ik}$, with:

$$W_{ij}^{is} = \begin{cases} 1, & l_i^S \neq l_j^S \quad and \quad l_i^E = l_j^E \\ 0, & otherwise \end{cases} \qquad (11)$$

The graph of W^{is} can be seen as an intrinsic embedding graph to penalize nonemotional factors between different speakers. The matrix form of W^{is} can be represented as:

$$W^{is} = \sum_{c=1}^{N_c} e^c e^{cT} - \sum_{c=1}^{N_c} \sum_{c_S=1}^{N_S} (e^{c_S} \circ e^c)(e^{c_S} \circ e^c)^T \qquad (12)$$

The optimization form of the intrinsic part is:

$$\min \sum_{i,j=1}^{N} (y_i^{(k)} - y_j^{(k)})^2 W_{ij}^{is} \Rightarrow \min y^T L^{is} y \qquad (13)$$

Noticing that maximizing of the distance between each two samples is adopted in PCA (Principal Component Analysis):

$$\max y^T H y \Rightarrow \max y^T (I - \frac{1}{N} e e^T) y \qquad (14)$$

where $e \in \Re^{N \times 1}$ is the column vector with all elements equal to 1. $H = I - \frac{1}{N} e e^T$.

The maximizing in (14) is also used in LDA as the penalty part. It can be seen as the form of inner product after removing mean value of samples. We consider it as a section of the final form of the proposed methods.

Combining the two forms of optimization of (4), (9), (13) and (14) together, we can obtain the form of (15) with simultaneously minimizing the numerator section and maximizing the denominator section of the objective function in (15).

$$\min \frac{(1-\gamma_1) y^T L_{LDA} y + \gamma_1 y^T L^{is} y}{(1-\gamma_2) y^T H y + \gamma_2 y^T L^{ps} y} \qquad (15)$$

To keep the similar form as the embedding graph of LDA and to make the parameters $0 \leq \gamma_1 \leq 1$ and $0 \leq \gamma_2 \leq 1$ balanced in representing the relationship between (4) and (13), as well as (9) and (14), we can also let the Laplacian matrix of the graphs as (16). When D^{ps} and D^{is} is with zero diagonal elements, the non-zero subblocks can be used to solve this problem.

$$\begin{cases} \tilde{L}^{ps} = (D^{ps})^{-\frac{1}{2}} L^{ps} (D^{ps})^{-\frac{1}{2}} = I - (D^{ps})^{-\frac{1}{2}} W^{ps} (D^{ps})^{-\frac{1}{2}} \\ \tilde{L}^{is} = (D^{is})^{-\frac{1}{2}} L^{is} (D^{is})^{-\frac{1}{2}} = I - (D^{is})^{-\frac{1}{2}} W^{is} (D^{is})^{-\frac{1}{2}} \end{cases} \qquad (16)$$

Supposing the intrinsic and penalty graphs of the proposed methods can be written as $(1-\gamma_1) L_{LDA} + \gamma_1 \tilde{L}^{is}$ and $L^p = (1-\gamma_2) H + \gamma_2 \tilde{L}^{ps}$ respectively, we can obtain the proposed SPGL as (17).

$$\min \frac{y^T L y}{y^T L^p y} = \min \frac{y^T [(1-\gamma_1) L_{LDA} + \gamma_1 \tilde{L}^{is}] y}{y^T [(1-\gamma_2) H + \gamma_2 \tilde{L}^{ps}] y} \qquad (17)$$

When the data mapping of SPGL is adopted as the linear form, the optimization form of the the proposed linear SPGL (LSPGL) with one-dimension situation in consideration is shown as:

$$\arg\min_{a} \frac{a^T X L X^T a}{a^T X L^p X^T a} \quad s.t. a^T a = 1 \qquad (18)$$

With the orthogonal constraint, (18) can be also written as (19) with multiple dimensions after dimensionality reduction.

$$\arg\min_{A} \frac{tr(A^T X L X^T A)}{tr(A^T X L^p X^T A)} \quad s.t. A^T A = I \qquad (19)$$

where the mapping direction matrix $A = [a_1, a_2, ..., a_m]$, whose column vectors are orthogonal between each other. The column vectors of A are corresponding to a in (18). Consequently, the optimization of (18) and (19) can be solved according to (20) as the generalized eigenvalue problem with eigenvalue λ.

$$\lambda X L X^T a = X L^p X^T a \qquad (20)$$

It is noticeable that to prevent the situation of small sample size problems and to improve recognition performance as well, (20) can be processed by the prior stage of PCA or SVD (Singular Value Decomposition). Then, the generalized eigenvalue problem is modified into a common eigenvalue problem which is easily solved, with orthogonalization of the separately obtained eigenvectors.

Compared with MFA[15], the proposed LSPGL method penalizes the speaker factors, which are not helpful for speech emotion classification in the original features, instead of only penalizing the neighboring marginal sample pairs. Compared with LDA[19], RDA (Regularized Discriminant Analysis)[20], SDA (Semi-supervised Discriminant Analysis)[21] and some other methods[22], weighted terms with speaker penalty information are added to improve performance, instead of other information.

For the kernelized form of SPGL, KSPGL, the prior SVD dimensionality reduction work for Gram matrix K for the reasons as what exist in LSPGL and the interference when the generalized eigenvalue problem is solved. The optimization for of KSPGL is shown as (21).

$$\arg\min_{a} \frac{\alpha^T K L K \alpha}{\alpha^T K L^p K \alpha} \quad s.t. \alpha^T \alpha = 1 \qquad (21)$$

where the Gram matrix $K = \phi^T(X)\phi(X)$. The N high-dimension training samples $\phi(X) = [\phi(x_1), \phi(x_2), ..., \phi(x_N)]$, which come from the original space X.

The computational cost of the proposed LSPGL and KSPGL is the same as Frobenius-norm based graph embedding methods.

IV. SPEECH EMOTION RECOGNITION USING SPGL

We adopt the proposed SPGL, including LSPGL and KSPGL, in the dimensionality reduction stage of speech emotion recognition. The proposed SPGL methods can mainly constrain the non-emotional features in the same speakers by the term of speaker penalty.

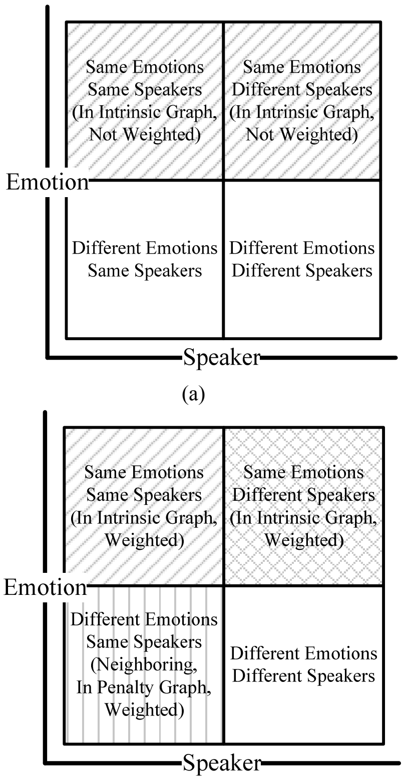
Each speech emotion sample is processed by pre-emphasizing and some basic denoising methods. Then, enframing stage using Hamming window is used to extract frame-wise information. Different kinds of original speech emotion features are obtained according to the frame-wise information above. The categories of speech emotion features adopted are pitch[1-5,18-19], zero-cross rate[3], energy[2-5,19], formant[2-3,5,19], durance[1-3,5,19] and MFCC(Mel Frequency Cepstrum Coefficient)[2-3] features. Those features are for the whole utterance of each speech emotion sample based on the statistical information of the frames in the sample.

With the features provided, the works of normalization and feature selection should be adopted to improve performance of the system. After that, the dimensionality reduction stage of LSPGL and KSPGL is designed to achieve effective factors for speech emotion classification. The kernels are selected as conventional Gaussian kernels.

The inter-embedding-graph weight parameters γ_1 and γ_2 of the proposed methods can be drawn according to cross-validation or some empirical methods.

As the stage of classifiers, kNN (k-Nearest Neighbor), SVM (Support Vector Machine) or some other effective classifiers can be used in recognition.

The overall thought of SPGL is illustrated in Figure1, where the shaded areas mean the information included in graph learning. Figure 1(a) shows the different kinds of information used in LDA while Figure 1(b) means the information adopted in SPGL.



(a)

(b)

Figure 1. The emotion and speaker factors taken into consideration in LDA and proposed SPGL respectively. (a) LDA. (b) SPGL

V. SPEECH EMOTION DATABASE

Berlin speech emotion database (EMO-DB)[23] and eNTERFACE'05 multimodal emotion corpus[24] are adopted in the experiments.

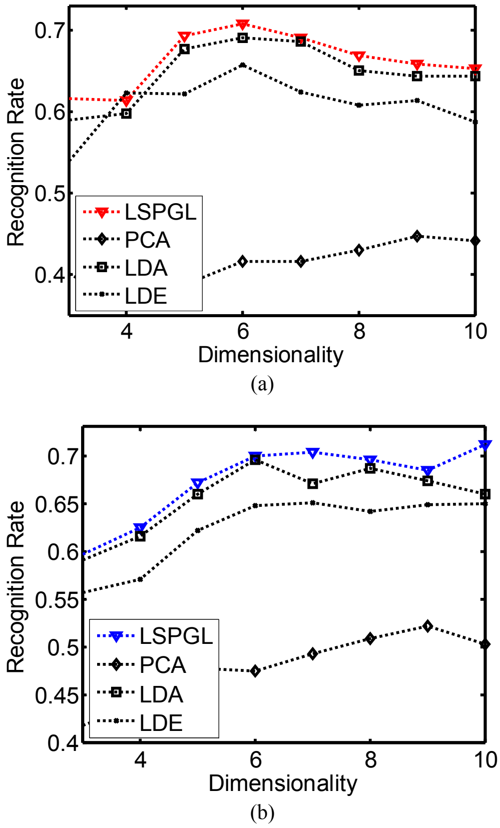
7 classes of speech emotions, which are neutral, fear, disgust, joy(happiness), boredom, sadness and anger, are in Berlin corpus. 10 speakers (5 male and 5 female) with 10 different German sentences are included in the corpus. 494 samples are chosen from the original database of EMO-DB in our experiments.

The corpus of eNTERFACE'05 provides the emotions of happiness, sadness, fear, disgust, surprise and anger. Short English sentences are spoken by 42 persons from different regions of the world. We choose samples from 15 speakers with only the parts of speech, with the whole video or face expression sections.

VI. EXPERIMENTAL RESULTS

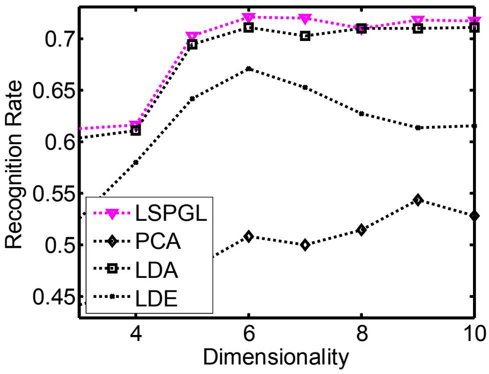
We use the experimental method of Leave One Speaker Out (LOSO)[6] to show the effectiveness of our proposed methods in the condition of speaker independent speech emotion recognition. LOSO makes the training and testing process divided by different speakers, without the impact of speaker factors connecting training and testing.

The experiments on EMO-DB is show as follows. The recognition rates corresponding to the low dimensions are represented as Figure 2, where Figure 2(a) is the recognition rates when 1NN classifiers are used and SVM and NB(Naive Bayesian) classifiers are adopted in Figure 2(b) and Figure 2(c) respectively. Linear dimensionality reduction methods, PCA, LDA, LDE and the proposed LSPGL, are compared in Figure 2.



(a)

(b)



(c)

Figure 2. The average recognition rates in EMO-DB using LOSO in different low dimensions using different classifiers. (a) 1NN. (b) SVM. (c) NB

The highest recognition rates are shown in Table I when the dimensionality is low in EMO-DB, with the classifiers of 1NN, SVM and NB respectively. The corresponding dimensions of the highest recognition rates for the methods are also attached in Table I.

TABLE I. THE MAXIMUM AVERAGE RECOGNITION RATES IN EMO-DB AND THE CORRESPONDING DIMENSIONS

Methods	EMO-DB(Berlin)		
	1NN(%) /Dimension	SVM(%) /Dimension	NB(%) /Dimension
Baseline	53.34/_	69.67/_	62.89/_
PCA	44.69/9	52.26/9	54.35/9
LDA	69.00/6	69.58/6	71.06/6
LDE/MFA	65.65/6	65.04/7	67.03/6
LSPGL	**70.77**/6	**71.14**/10	**71.94**/7
KSPGL	**71.75**/8	**72.42**/7	**72.44**/7

Like Table I, Table II shows the highest recognition rates of different methods in the experiments in the corpus of eNTERFACE’05.

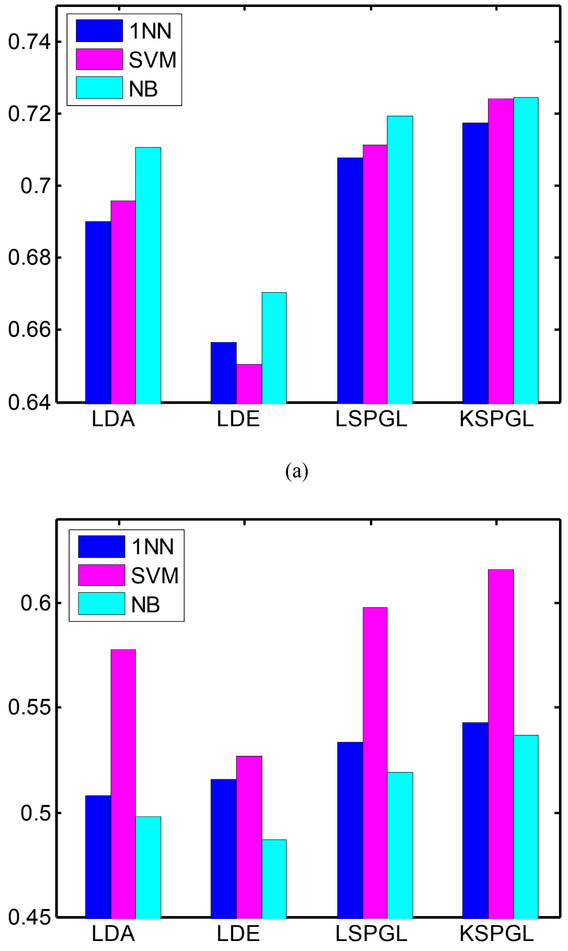
TABLE II. THE MAXIMUM AVERAGE RECOGNITION RATES IN ENTERFACE'05 AND THE CORRESPONDING DIMENSIONS

Methods	eNTERFACE’05		
	1NN(%) /Dimension	SVM(%) /Dimension	NB(%) /Dimension
Baseline	46.44/_	54.22/_	45.33/_
PCA	49.23/8	43.50/7	42.00/9
LDA	50.82/5	57.82/5	49.78/5
LDE/MFA	51.59/9	52.69/8	48.67/8
LSPGL	**53.36**/5	**59.78**/7	**51.89**/8
KSPGL	**54.33**/5	**61.64**/9	**53.67**/7

According to the recognition results in Figure 2, Table I and Table II, the proposed LSPGL can achieve better performance than the state-of-the-art graph learning based dimensionality reduction methods in most conditions. Additionally, the algorithm KSPGL can improve the performance of speaker independent speech emotion recognition by using nonlinear kernel mappings, based on the proposed LSPGL.

Figure 3 provides the recognition rates comparison of the algorithms with supervised information using the different three classifiers, 1NN, SVM and NB. According to the experimental results in Table I, Table II and Figure 3, the performance of speech emotion recognition systems turns to be better when SVM classifiers are adopted compared with using 1NN classifiers. However, computational costs may be higher for SVM classification since SMO (Sequential Minimal Optimization) is used in its iterative optimization in training procedures. The performance of the classical NB classifiers are not stable. This is most likely due to the fact of the relatively fixed model selection. Based on common experience, a more complexed and adaptive model may be valid for this kind of situation.

In conclusion, in the condition of speaker independent speech emotion recognition, the recognition rate of SPGL can achieve 72.44% in EMO-DB, while it is 61.64% in the corpus of eNTERFACE’05 according to the experiments. The recognition rates are able to be raised based on more effective original feature extraction, feature selection methods and the choices of adopting different categories of classifiers for the given features.



(a)

(b)

Figure 3. The average highest recognition rates in the corpus of EMO-DB and eNTERFACE’05 with different classifiers for different supervised dimensionality methods. (a) EMO-DB. (b) eNTERFACE’05

VII. CONCLUSIONS AND FUTURE WORK

The methods based on speaker penalty information, LSPGL and KSPGL, are proposed in this paper, where the methods provide a new perspective to constrain the useless information in speech emotion recognition. They are able to improve the performance of a speech emotion recognition system with the weighted speaker-penalty term added, according to the results of the experiments. However, the methods can only improve the performance by not a large margin, which means that some other important categories of interference may also exist in the work of speech emotion recognition, such as the features used in automatic speech recognition. Therefore, a more generalized form of graph learning methods can be proposed to reduce the influences from the features which are unfavorable for speech emotion recognition.

ACKNOWLEDGMENT

The authors thank to the anonymous reviewers.

REFERENCES

[1] F. Dellaert, T. Polzin, A. Waibel, "Recognizing emotion in speech," in International Conference on Spoken Language, Philadelphia, PA, USA, 1996, pp.1970-1973. [Online]. Available: http://dx.doi.org/10.1109/ICSLP.1996.608022.

[2] D. Ververidis, C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication, vol. ED-48, pp. 1162-1181, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2006.04.003.

[3] B. Schuller, G. Rigoll, "Timing levels in segment-based speech emotion recognition," in INTERSPEECH'2006, Pittsburgh, PA, USA, 2006, pp. 1818-1821.

[4] P. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," International Journal of Human-Computer Studies, vol. ED-59, pp. 157-183, 2003. [Online]. Available: http://dx.doi.org/10.1016/S1071-5819(02)00141-6.

[5] R. Tato, R. Santos, R. Kompe, J. Pardo, "Emotional space improves emotion recognition," in International Conference on Spoken Language, Denver, CO, USA, 2002, pp. 2029-2032.

[6] B. Schuller, R. Müller, M. K. Lang, G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in INTERSPEECH'2005, Lisbon, Portugal, 2005, pp. 805-808.

[7] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, "Speaker independent speech emotion recognition by ensemble classification," in IEEE International Conf. Multimedia and Expo(ICME), Amsterdam, The Netherlands, 2005, pp. 864-867. [Online]. Available: http://dx.doi.org/10.1109/ICME.2005.1521560.

[8] T. Kostoulas, T. Ganchev, N. Fakotakis, "Study on speaker-independent emotion recognition from speech on real-world data," in Verbal and nonverbal features of human-human and human-machine interaction, Springer Berlin Heidelberg, 2008, pp. 235-242. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-70872-8_18.

[9] M. Belkin, P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in Advances in Neutral Information Processing Systems(NIPS) 14, Vancouver, Canada, 2002, pp. 585-591.

[10] X. He, P. Niyogi, "Locality preserving projections," in Advances in Neural Information Processing Systems (NIPS) 16, Whistler, Canada, 2003, pp. 153-160.

[11] S. Roweis, L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. ED-290(5500), pp. 2323-2326, 2000. [Online]. Available: http://dx.doi.org/10.1126/science.290.5500.2323.

[12] 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. ED-28(9), pp. 1393-1403, 2006. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2006.184.

[13] J. Tenenbaum, V. de Silva, J. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol. ED-290, pp. 2319-2323, 2000. [Online]. Available: http://dx.doi.org/10.1126/science.290.5500.2319.

[14] H. Chen, H. Chang, T. Liu, "Local discriminant embedding and its variants," in IEEE Conf. Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 2005, pp. 846-853. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.216.

[15] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. ED-29(1), pp. 40-51, 2007. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2007.250598.

[16] F. De la Torre, "A least-squares framework for component analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. ED-34(6), pp. 1041-1055, 2012. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2011.184.

[17] M. You, C. Chen, J. Bu, J. Liu, J. Tao, "Emotional speech analysis on nonlinear manifold," in International Conference on Pattern Recognition(ICPR), Hong Kong, 2006, pp. 91-94. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2006.490.

[18] 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. ED-10(114), pp. 1-7, 2013. [Online]. Available: http://dx.doi.org/10.5772/55403.

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

[20] Friedman J H, "Regularized discriminant analysis," Journal of the American Statistical Association, vol. ED-84(405), pp. 165-175, 1989. [Online]. Available: http://dx.doi.org/10.1080/01621459.1989.10478752.

[21] D. Cai, X. He, "Semi-supervised discriminant analysis," in International Conference on Computer Vision(ICCV). Rio de Janeiro, Brazil, 2007, pp. 1-7. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2007.4408856.

[22] L. He, J. M. Buenaposada, L. Baumela, "An empirical comparison of graph-based dimensionality reduction algorithms on facial expression recognition tasks," in International Conf. Pattern Recognition (ICPR), Tampa, FL, USA, 2008, pp. 1-4. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2008.4761731.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, "A database of German emotional speech," in INTERSPEECH'2005, Lisbon, Portugal, 2005, pp. 1517-1520.

[24] O. Martin, I. Kotsia, B. Macq, I. Pitas, "The enterface'05 audio-visual emotion database," in IEEE Conf. Data Engineering Workshops, Atlanta, GA, USA, 2006, pp. 8-8. [Online]. Available: http://dx.doi.org/10.1109/ICDEW.2006.145.