JOINT TRANSFER SUBSPACE LEARNING AND FEATURE SELECTION FOR CROSS-CORPUS SPEECH EMOTION RECOGNITION

Peng Song [1] Wenming Zheng [2] Shifeng Ou [3] Yun Jin [2] Wenming Ma [1] Yanwei Yu [1]

[1]School of Computer and Control Engineering, Yantai University, Yantai 264005, China
[2]Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing 210096, China
[3] School of Opto-electronic Information Science and Technology, Yantai University, Yantai 264005, China
pengsong@ytu.edu.cn, wenming_zheng@seu.edu.cn

ABSTRACT

Cross-corpus speech emotion recognition has attracted a great attention due to its widespread existence of various emotional speech. It takes one corpus as the training data to recognize emotions of another corpus, and often involves two basic problems, i.e., coupled feature matching and feature selection. Most previous studies focus on solving the first problem. In this study, we propose a general learning framework, called joint transfer subspace learning and feature selection (JTSLFS), to deal with these two problems. To address the first problem, we learn a latent common subspace by reducing the distribution difference and preserving the important properties of features, in which a shared feature representation can be discovered. Besides, we impose the $l_{2,1}$-norm on the projection matrix to deal with the second problem. A graph regularizer, which considers the geometric structure of data, is further presented to improve the recognition performance. Experimental results on cross-corpus speech emotion recognition tasks suggest that our proposed method achieves more encouraging results compared with some state-of-the-art approaches.

Index Terms— Subspace learning, feature selection, speech emotion recognition, cross-corpus, transfer learning

1. INTRODUCTION

In speech signal processing field, emotion recognition plays an important role, and has received much attention over the past decades. The objective of speech emotion recognition is to recognize emotions from speech into the following categories, e.g., happiness, sadness, disgust and surprise. It has been proven very useful in many applications [1]. Many statistical methods have been adopted to implement the classification function, such as support vector machine (SVM), Gaussian mixture model (GMM), artificial neural network (ANN), extreme learning machine (ELM), deep neural network (DNN) and regression algorithms [1, 2, 3, 4, 5]. These methods obtain satisfactory results to some extent. Unfortunately, we can notice that all these algorithms are conducted and tested on the same corpus, in which the training and testing data are drawn from the same corpus. In practical situations, since emotional speech utterances are often collected in different environments, e.g., noises, languages, devices and age groups, we have to face the cross-corpus speech emotion recognition problem. In this case, the classifier model trained in one corpus is applied to another corpus, which often degrades the recognition performance [6].

There have been reported work in automatic speaker and speech recognition, where researchers have presented many adaptation techniques to improve their systems' performance [7]. Following this idea, some adaptation algorithms, e.g., feature normalization [8, 9], maximum a posteriori (MAP), joint factor analysis (JFA), vocal tract length normalization (VTLN), have been introduced in speech emotion recognition [10, 11, 12, 13]. Meanwhile, over the past few years, with the rapid growth of deep learning techniques, much attention has been paid to developing DNN based speech emotion recognition methods [2, 3, 4], in which the common strategy is to learn high-level invariant emotional features from raw speech utterances. They can obtain better recognition performance than traditional algorithms. Nonetheless, these methods require a large amount of training data, which is hard to collect in practice, and do not take into account the "corpus bias" problem [14].

Recently, one major research direction focuses on addressing the "corpus bias" problem via domain adaptation and transfer learning algorithms, in which the differences between different feature distributions are considered. In [14], Deng et al. present an autoencoder-based unsupervised domain adaptation approach to cope with the cross-corpus speech emotion recognition problem, in which the prior knowledge from the target corpus is employed to regularize the training on the source corpus. In [15], Zong et al. present a least-squares regression based domain adaptation algorithm, in which the labeled source and unlabeled target data are jointly utilized to train the recognition model. In [12], Hassan et al. introduce the popular transfer learning algorithms to compensate the speaker and acoustic variations for cross-corpus speech emotion recognition. In [16], Zhang et al. propose a multi-task learning approach to cope with the cross-corpus emotion recognition from singing and speaking. In [6], we have presented a transfer non-negative matrix factorization (TNMF) approach to learn corpus-invariant feature representations across training and testing corpora. However, these algorithms focus on finding the common feature representations to cope with the feature matching problem, and do not consider the importance of feature selection together.

The main contribution of this work is a new learning framework that jointly performs transfer subspace learning and feature selection for cross-corpus speech emotion recognition. In this way, we learn a projection matrix to map the features of different corpora into a common low-dimensional subspace, while the $l_{2,1}$-norm is imposed on the projection matrix to perform feature selection. Moreover, a graph regularizer is further introduced to improve the recognition performance.

2. JTSLFS FOR CROSS-CORPUS SPEECH EMOTION RECOGNITION

2.1. The objective function

Let $X = [X_s, X_t] \in R^{m\times n}(n = n_l + n_u)$ be the feature matrix, with n data points in a m-dimensional feature sequence, in which $X_s = [x_1, \ldots, x_{n_l}] \in R^{m\times n_l}$ and $X_t = [x_{n_l+1}, \ldots, x_n] \in R^{m\times n_u}$ are the features of labeled source and unlabeled target corpora, respectively. Since the samples are from different corpora, our goal is to find the common representations of X in a latent low-dimensional space, denoted by $Y = [Y_s, Y_t]$, where $Y_s = [y_1, \ldots, y_{n_l}]^T \in R^{n_l\times c}$ and $Y_t = [y_{n_l+1}, \ldots, y_n]^T \in R^{n_u\times c}$. In this paper, following the idea of spectral regression [17], the dimensionality reduction algorithms, e.g., linear discriminant analysis (LDA), locality preserving projection (LPP), can be cast into a regression framework. Given the representations from the labeled source dataset Y_s, the optimal Y_s can be obtained by

$$\begin{aligned} &\min_{Y_s} \sum_{i,j} u_{ij} \left\| y_i - y_j \right\|^2 \\ &s.t.\ Y_s^T D Y_s = I \end{aligned} \tag{1}$$

where D is a diagonal matrix, whose entries are the column sums of a weight matrix $U = [u_{ij}] \in R^{n_l\times n_l}$, and u_{ij} indicates whether x_i and x_j are from the same class, which is defined as follows:

$$u_{ij} = \begin{cases} \frac{1}{n_k} & \text{if } x_i \text{ and } x_j \text{ both belong to the } k\text{-th class} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where n_k is the number of the k-th class. The constraint term $Y_s^T D Y_s = I$ removes an arbitrary scaling factor in the embeddings, where I is the identity matrix [18].

According to [17], the optimal Y_s is computed as

$$v_k = (\underbrace{0,\ldots,0}_{\sum_{i=1}^{k-1} n_i}, \underbrace{1,\ldots,1}_{n_k}, \underbrace{0,\ldots,0}_{\sum_{i=k+1}^{c} n_i})^T, \quad k = 1,\ldots,c \tag{3}$$

where c is the number of classes. Y_s can be represented as $Y_s = [v_1, \ldots, v_c] \in R^{n_l\times c}$.

Next, we want to learn a projection matrix to map the data into the latent low-dimensional space. The objective function is given as

$$\min_P \left\| X^T P - Y \right\|_F^2 \tag{4}$$

where P is the projection matrix, the superscript T denotes the transposition of a matrix, and $\|\cdot\|_F$ refers to the Frobenius norm of a matrix.

Since the training and testing data are from different corpora, and often follow different feature distributions, the difference between two feature distributions is considered. Following conventional transfer learning algorithms [6, 19], the maximum mean discrepancy (MMD) is adopted to measure this discrepancy. Thus, the objective function can be formulated as follows:

$$\min_P \left\| X^T P - Y \right\|_F^2 + \beta\Omega(P) \tag{5}$$

where β is a nonnegative regularization parameter, and $\Omega(P)$ is the MMD regularization term, which is given as

$$\begin{aligned} \Omega(P) &= \left\| \frac{1}{n_l}\sum_{i=1}^{n_l} y_i - \frac{1}{n_u}\sum_{j=n_l+1}^{n} y_j \right\|^2 \\ &= Tr(P^T X M X^T P) \end{aligned} \tag{6}$$

where $Tr(\cdot)$ denotes the trace of a matrix, and $M = [m_{ij}]_{i,j=1}^n$ is the MMD matrix, which is computed as

$$m_{ij} = \begin{cases} \frac{1}{n_l^2} & x_i, x_j \in X_s \\ \frac{1}{n_u^2} & x_i, x_j \in X_t \\ \frac{-1}{n_l n_u} & \text{otherwise} \end{cases} \tag{7}$$

Meanwhile, we perform feature selection via imposing the $l_{2,1}$-norm on the projection matrix [20]. So the objective function can be written as

$$\min_P \left\| X^T P - Y \right\|_F^2 + \beta\Omega(P) + \alpha\left\| P \right\|_{2,1} \tag{8}$$

where $\|\cdot\|_{2,1}$ refers to the $l_{2,1}$-norm, which is defined as the summarization of the l_2-norm of columns of a matrix, and α is a nonnegative regularization parameter.

2.2. The graph regularization

Motivated by recent progress in manifold learning [21], we utilize both labeled source and unlabeled target samples to design a graph, which considers the geometric structure of data, to further improve the recognition performance. Given the feature set X, we can construct a p nearest neighbor graph G with n vertices to model the relationships between the nearby data points, where edge weights encode the similarities between samples. Let $W = [w_{ij}] \in R^{n\times n}$ be the weight matrix of G. There are many choices of W, in this work, the common and simple 0-1 weighting scheme is adopted, which is written as

$$w_{ij} = \begin{cases} 1 & \text{if } x_j \in N_p(x_i) \text{ or } x_i \in N_p(x_j) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $N_p(x_i)$ is the set of p nearest neighbors of x_i. Similar to Eq. (1), a natural graph regularizer can be defined as

$$\begin{aligned} J(P) &= \min_P \frac{1}{2}\sum_{i,j=1}^{n} w_{ij} \left\| x_i^T P - x_j^T P \right\|^2 \\ &= \sum_{i=1}^{n} (x_i^T P)^T (x_i^T P) b_{ii} - \sum_{i,j=1}^{n} (x_i^T P)^T (x_j^T P) w_{ij} \\ &= Tr(P^T X L X^T P) \end{aligned} \tag{10}$$

where $L = B - W$ is called *graph Laplacian* [22], $B = [b_{ii}] \in R^{n\times n}$ is a diagonal matrix with $b_{ii} = \sum_j w_{ij}$.

Incorporating this geometrical regularization term into the objective function shown in Eq. (5), we obtain the JTSLFS model, and the objective function is given as follows:

$$\min_P \left\| X^T P - Y \right\|_F^2 + \alpha\left\| P \right\|_{2,1} + \beta\Omega(P) + \gamma J(P) \tag{11}$$

where γ is a nonnegative trade-off parameter. By combining the last two regularization terms, the objective function can be further modified as

$$\min_P \left\| X^T P - Y \right\|_F^2 + Tr(P^T R P) + \alpha\left\| P \right\|_{2,1} \tag{12}$$

where $R = X(\beta M + \gamma L)X^T$.

2.3. Optimization algorithm

The optimization problem in Eq. (12) contains the $l_{2,1}$-norm, which is non-smooth and cannot get a closed form solution. Consequently, an iterative algorithm is presented in this subsection. Given the projection matrix P, the $l_{2,1}$-norm of P is defined as

$$\|P\|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} P_{ij}^2} = 2Tr(P^T QP) \tag{13}$$

where $Q = [q_{ii}] \in R^{m\times m}$ is a diagonal matrix with $q_{ii} = \frac{1}{2\|p^i\|_2}$, p^i means the $i-$th row vector of Q.

Note that in practice, $\|p^i\|_2$ could be close to zero. Following the half-quadratic minimization [23], q_{ii} is redefined as $q_{ii} = \frac{1}{2\sqrt{\|p^i\|_2^2+\epsilon}}$, where ϵ is a very small positive constant. Consequently, we will minimize the following objective function $\mathcal{O}$ to learn the projection matrix P:

$$\mathcal{O} = \left\|X^T P - [Y_s, Y_t]\right\|_F^2 + Tr(P^T RP) + \alpha Tr(P^T QP) \tag{14}$$

The iterative algorithm is summarized as follows:

1). Update P ***as given*** Y_t. Setting the partial derivative of $\mathcal{O}$ with respect to P to zero, we obtain the following equation:

$$\begin{aligned} &\frac{\partial \mathcal{O}}{\partial P} = 0 \\ \Rightarrow\quad & 2X(X^T P - Y) - 2RP - 2\alpha QP = 0 \\ \Rightarrow\quad & (XX^T - R - \alpha Q)P = XY \end{aligned} \tag{15}$$

And left multiplying both sides of Eq. (15) by $(XX^T - R - Q)^{-1}$, we get the analytical solution of P as

$$P^* = (XX^T - R - \alpha Q)^{-1} XY \tag{16}$$

2). Update Y_t ***as given*** P. When P is fixed, Eq. (14) can be reformulated as

$$\mathcal{O} = \min_{Y_t} \left\|[Y_s, Y_t] - X^T P\right\|_F^2 \tag{17}$$

which is equivalent to the following optimization problem:

$$\mathcal{O} = \min_{Y_t} \left\|Y_t - X_t^T P\right\|_F^2 \tag{18}$$

The above optimization problem can be easily solved by the quadratic programming algorithm [24].

The detailed algorithmic procedure of learning projection matrix P is stated in Algorithm 1.

3. EXPERIMENTS

In this section, we evaluate the performance of our proposed JTSLFS approach for speech emotion recognition. In our work, the speech emotion recognition is a cross-corpus recognition task, in which the source training dataset is labeled and the target testing dataset is unlabeled.

Algorithm 1 JTSLFS algorithm

Input:

The feature matrix $X \in R^{m\times n}$ and low-dimensional representation of labeled source corpus $Y_s \in R^{n_l \times c}$;

The parameters α, β, γ and p.

Output:

The projection matrix $P \in R^{m\times c}$.

a). Compute the MMD matrix M;

b). Construct the p nearest neighbor graph G;

c). Set $k = 0$, initialize $Y_t^0 \in R^{n_u \times c}$.

repeat

1. Compute the projection matrix P according to Eq. (16): $P^k = (XX^T - R - Q)^{-1} X[Y_s, Y_t^k]$;

2. Update the low-dimensional representations of unlabeled target corpus Y_t^{k+1} according to Eq. (18);

3. $k = k + 1$;

until Convergence.

3.1. Data sets and compared algorithms

The EMO-DB and eNTERFACE emotional databases are used in our experiments, and the important statistics of these two corpora are summarized below:

- The EMO-DB[1] is a popular, acted emotional database. It contains 7 types of emotions, i.e., anger, boredom, disgust, fear, happiness, neutral and sadness. 494 speech utterances as a whole are collected by 10 actors in German, which are all used in our experiments.
- The eNTERFACE[2] is a public, acted, audio-visual emotional database. It consists of 1287 video samples with 6 emotion categories, i.e., anger, disgust, fear, happiness, sadness and surprise. These videos are recorded by 43 subjects with predefined speech content in English. In our experiments, all the audio samples are chosen for evaluation.

We adopt the openSMILE toolkit[3] to extract the acoustic features, and the 1582 dimensional standard feature set of INTERSPEECH 2010 paralinguistic challenge [25] is used in our experiments. To show the cross-corpus recognition performance, we compare our approach with other related methods. The methods that we evaluate are listed below:

- Conventional method (*Conventional*), in which the classifier trained in source corpus is directly used for recognition in the target corpus.
- Baseline method (*Baseline*), in which the training and testing procedures are conducted on the same corpus.
- Transfer sparse coding (TSC) [19], where the MMD is incorporated into the objective function of sparse coding.
- Transfer NMF method (TNMF) [6], where the MMD is incorporated into the objective function of NMF.
- Transfer subspace learning (TSL), which is a special case of our proposed JTSLFS, when α and γ are set to zero.
- Our proposed JTSLFS without graph regularization (TSLFS), which can be seen as a special case of JTSLFS, when γ is set to zero.

[1] http://emodb.bilderbar.info/docu
[2] http://enterface.net/enterface05/main.php?frame=emotion
[3] http://sourceforge.net/projects/opensmile

Table 1. The recognition performance in *test*1

Methods	Recognition rates (%)					
	Anger	Disgust	Fear	Happiness	Sadness	Average
Conventional	37.23	19.21	17.98	27.16	28.40	28.87
TSC	50.18	29.25	36.86	47.45	45.98	44.96
TNMF	50.02	29.30	36.85	47.28	46.06	43.99
TSL	47.16	26.29	32.26	46.02	45.16	40.02
TSLFS	**50.35**	**29.56**	**37.19**	**47.78**	**46.35**	**45.52**
Ours	**50.39**	**29.57**	**37.22**	**47.91**	**46.38**	**45.61**
Baseline	74.40	55.35	54.03	59.98	60.96	61.36

Table 2. The recognition performance in *test*2

Methods	Recognition rates (%)					
	Anger	Disgust	Fear	Happiness	Sadness	Average
Conventional	31.49	53.06	16.47	20.98	47.20	34.63
TSC	35.39	72.98	18.97	25.52	69.26	50.59
TNMF	36.13	73.07	19.05	25.53	69.32	51.96
TSL	37.80	72.56	18.65	25.38	69.25	50.92
TSLFS	**38.02**	**74.11**	**19.12**	**26.02**	**69.68**	**52.18**
Ours	**38.05**	**74.49**	**19.18**	**26.71**	**71.38**	**52.26**
Baseline	73.01	81.04	68.58	52.99	79.33	71.02

- Our proposed JTSLFS approach (*Ours*).

The linear SVM is used as the standard classifier for the above mentioned algorithms, and 5 common emotion categories, i.e., anger, disgust, fear, happiness and sadness, are used for evaluations.

3.2. Experimental Results

Under our experimental setup, it is impossible to select the model parameters using cross validation strategy, since the labeled source and unlabeled target data sets follow different feature distributions. Consequently, we use a search strategy by searching optimal parameters in the parameter space. Note that the objective function in Eq. (11) mainly involves three parameters, i.e., α, β and γ. α is the weighting parameter of $l_{2,1}$-norm, β is the weighting parameter of MMD matrix, and γ is the weighting parameter of graph regularization. We tune these three parameters in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. Finally, α, β and γ are set to 0.1, 1 and 100, respectively, and the number of nearest neighbors in graph is set to 5.

Two types of experiments are carried out, i.e., *test*1 versus *test*2. In *test*1, the labeled EMO-DB database is used for training, and the unlabeled eNTERFACE database is chosen for testing. Meanwhile, in *test*2, the labeled eNTERFACE database is used for training, and the unlabeled EMO-DB is used for testing. In our experiments, each database is divided into five subsets with equal size. In each test, three subsets are used for training while the others are used for testing. The tests are repeated 20 times such that they can cover all possible cases. After experiments, we use the recognition rates of each emotion and overall average to evaluate the recognition performance.

Tables 1 and 2 show the results of various methods in *test*1 and *test*2. These results reveal a number of interesting points:

- As we can see, regardless of each case, our proposed JTSLFS method achieves the best recognition rates, which demonstrates the efficacy of the joint transfer subspace learning and feature selection idea.
- The TSLFS algorithm outperforms the other three popular transfer learning algorithms, i.e., TSC, TNMF and TSL. This suggests the importance of feature selection strategy in transfer subspace learning.
- As we have described, the JTSLFS adopts a p nearest neighbor graph to capture the local data structure. In both cases, the JTSLFS obtains higher recognition rates than TSLFS. This shows that, by leveraging the power of *graph Laplacian* regularization, the JTSLFS model can obtain better feature representations.

4. CONCLUSIONS

In this paper, we have presented a novel cross-corpus speech emotion recognition method, called joint transfer subspace learning and feature selection (JTSLFS). The JTSLFS performs transfer subspace learning and feature selection in a joint framework. Specifically, a projection matrix is learned to obtain common feature representations for different data sets, while the $l_{2,1}$-norm imposed on the projection matrix is used for feature selection, and a Laplacian graph is further used to enhance the recognition performance. Experimental results on cross-corpus speech emotion recognition tasks have demonstrated that JTSLFS performs better than several relevant state-of-the-art methods.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 61703360, 61572009, 61602399 and 61773331, the Natural Science Foundation of Shandong Province under Grants ZR2016FB22 and ZR2017FM019, and the Fundamental Research Funds for the Southeast University under Grant CDLS-2017-02.

6. REFERENCES

[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *Proc. INTERSPEECH*, pp. 223–227, 2014.

[3] Jinkyu Lee and Ivan Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition.," in *INTERSPEECH*, 2015, pp. 1537–1540.

[4] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[5] Wenming Zheng, Minghai Xin, Xiaolan Wang, and Bei Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569–572, 2014.

[6] Peng Song, Wenming Zheng, Shifeng Ou, Xinran Zhang, Yun Jin, Jinglei Liu, and Yanwei Yu, "Cross-corpus speech emotion recognition based on transfer non negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.

[7] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[8] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps, "Speaker normalisation for speech-based emotion detection," in *Digital Signal Processing, 2007 15th International Conference on*. IEEE, 2007, pp. 611–614.

[9] Carlos Busso, Angeliki Metallinou, and Shrikanth S Narayanan, "Iterative feature normalization for emotional speech detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5692–5695.

[10] Hao Hu, MingXing Xu, and Wei Wu, "GMM supervector based SVM with spectral features for speech emotion recognition.," in *Proc. ICASSP*, 2007, pp. 413–416.

[11] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[12] Ali Hassan, Robert Damper, and Mahesan Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[13] Peng Song, Yun Jin, Cheng Zha, and Li Zhao, "Speech emotion recognition method based on hidden factor analysis," *Electronics Letters*, vol. 51, no. 1, pp. 112–114, 2015.

[14] Jun Deng, Zixing Zhang, Florian Eyben, and Bjorn Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[15] Yuan Zong, Wenming Zheng, Tong Zhang, and Xiaohua Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 585–589, 2016.

[16] Biqiao Zhang, Emily Mower Provost, and Georg Essi, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5805–5809.

[17] Deng Cai, Xiaofei He, and Jiawei Han, "Spectral regression for efficient regularized subspace learning," in *Proc. ICCV*, 2007, pp. 1–8.

[18] Xiaofei He, Chiyuan Zhang, Lijun Zhang, and Xuelong Li, "A-optimal projection for image representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 1009–1015, 2016.

[19] Mingsheng Long, Guiguang Ding, Jianmin Wang, Jiaguang Sun, Yuchen Guo, and Philip S Yu, "Transfer sparse coding for robust image representation," in *Proc. CVPR*, 2013, pp. 407–414.

[20] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *Proc. NIPS*, 2010, pp. 1813–1821.

[21] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, no. Nov, pp. 2399–2434, 2006.

[22] Fan RK Chung, *Spectral graph theory*, vol. 92, American Mathematical Soc., 1997.

[23] Ran He, Tieniu Tan, Liang Wang, and Wei-Shi Zheng, "$l_{2,1}$ regularized correntropy for robust feature selection," in *Proc. CVPR*, 2012, pp. 2504–2511.

[24] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[25] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, Shrikanth S Narayanan, et al., "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, 2010, pp. 2795–2798.