UNSUPERVISED CROSS-CORPUS SPEECH EMOTION RECOGNITION USING DOMAIN-ADAPTIVE SUBSPACE LEARNING

Na Liu[1,4,3]*, Yuan Zong*[2]*, Baofeng Zhang*[3,1]*, Li Liu*[4]*, Jie Chen*[4]*, Guoying Zhao*[4]*, Junchao Zhu*[3*]

[1]School of Computer Science and Engineering, Tianjin University of Technology, China
[2] Research Center for Learning Science, Southeast University, China
[3]School of Electrical and Electronic Engineering, Tianjin University of Technology, China
[4]Center for Machine Vision and Signal Analysis, University of Oulu, Finland

ABSTRACT

In this paper, we investigate an interesting problem, i.e., unsupervised cross-corpus speech emotion recognition (SER), in which the training and testing speech signals come from two different speech emotion corpora. Meanwhile, the training speech signals are labeled, while the label information of the testing speech signals is entirely unknown. Due to this setting, the training (source) and testing (target) speech signals may have different feature distributions and therefore lots of existing SER methods would not work. To deal with this problem, we propose a domain-adaptive subspace learning (DoSL) method for learning a projection matrix with which we can transform the source and target speech signals from the original feature space to the label space. The transformed source and target speech signals in the label space would have similar feature distributions. Consequently, the classifier learned on the labeled source speech signals can effectively predict the emotional states of the unlabeled target speech signals. To evaluate the performance of the proposed DoSL method, we carry out extensive cross-corpus SER experiments on three speech emotion corpora including EmoDB, eNTERFACE, and AFEW 4.0. Compared with recent state-of-the-art cross-corpus SER methods, the proposed DoSL can achieve more satisfactory overall results.

Index Terms— Cross-corpus evaluation, speech emotion recognition, domain adaptation, subspace learning

1. INTRODUCTION

Speech emotion recognition (SER) aims at providing computers the ability to recognize the human beings' emotional states such as happy, fear, and disgust from their speech signals [1]. It has become a very hot research topic among affective computing, pattern recognition, and human-computer interaction (HCI). Generally speaking, a conventional SER task is to learn a classifier based on the labeled training speech signals and then predict the emotion labels of the unlabeled testing samples via the learned classifier while the training and testing samples come from a same corpus. In the practical scenarios, however, the training and testing samples may belong to different speech corpora. For example, they are recorded by different equipments or collected under different environments. This thus creates a more difficult yet interesting problem than conventional SER, i.e., unsupervised cross-corpus SER. To distinguish the training and testing speech corpora in cross-corpus SER problem, these two corpora can be referred as source corpus and target corpus, respectively. In this paper, we will investigate the unsupervised case of cross-corpus SER, in which the training and testing speech signals come from two different speech emotion corpora. Meanwhile, the training speech signals are labeled, while the label information of the testing speech signals is entirely unknown. Due to this setting, the training and testing speech signals may have different feature distributions.

In recent years, many researchers have focused on this challenging problem and proposed lots of effective methods. Schuller et al. [2] attempted to use various normalization schemes to deal with cross-corpus SER problem, which may be the first research about cross-corpus SER. Subsequently, more diverse cross-corpus SER methods are in sequence proposed. For example, Deng et al. [3–5] proposed an autoencoder based domain adaptation framework to cope with cross-corpus SER, in which autoencoder networks are used to learn the new representations for source and target speech samples. In the work of [6], Hassan et al. proposed an importance-weighted support vector machine (IW-SVM) for cross-corpus SER tasks. IW-SVM leverages three domain adaptation methods, i.e., kernel mean matching (KMM) [7], Kullback-Leibler importance estimation procedure (KLIEP) [8], and unconstrained least-squares importance fitting (uLSIF) [9], to learn a set of importance weights for target speech samples such that the feature distribution mismatch between source and target speech samples is relieved. Recently, a transfer non-negative matrix factorization (TNNMF) method is proposed by Song et al. [10] for cross-corpus SER tasks. In TNNMF, the maximum mean discrepancy (MMD) [11] is

*Corresponding author

used to balance the feature distribution difference between the originally distinct source and target speech signals. More recently, Zong et al. [12, 13] proposed a novel domain adaptation method called domain-adaptive least squares regression (DALSR) model to handle cross-corpus SER. DALSR aims at learning a regression coefficient matrix to bridge the source and target speech corpora.

In this paper, we propose a novel method called domain-adaptive subspace learning (DoSL) to deal with the unsupervised cross-corpus SER problem. The basic idea of DoSL is to learn a projection matrix which transforms the source and target speech signals from the original feature space to a common subspace. In such common space, the source and target speech signals are enforced to obey the similar feature distributions and hence we can train a classifier, e.g., support vector machine (SVM), based on the labeled source speech signals such that it can accurately predict the emotional states of the target speech signals. Motivated by the works of [12, 13], we construct a label space based on the label information provided by the source speech corpora to serve as the predefined common subspace for DoSL.

2. PROPOSED METHOD

2.1. DoSL model

Suppose we have two different speech corpora to serve as source and target corpus, respectively. Their corresponding feature matrices are denoted by $\mathbf{X}^s \in \mathbf{R}^{d\times N_s}$ and $\mathbf{X}^t \in \mathbf{R}^{d\times N_t}$, where d is the dimension of the speech feature vector and N_s and N_t are the numbers of the source and target speech signals, respectively. Since the label information of source speech signals is available for us in the unsupervised case of cross-corpus SER, we denote their label information as the vector form, which is followed by the works of [12, 13]. Specifically, let $\mathbf{L}^s \in \mathbf{R}^{c\times N_s}$ be the label matrix corresponding to the source feature matrix $\mathbf{X}^s$, where c is the number of speech emotion states, and the j^{th} element $l_{i,j}$ of its i^{th} column $\mathbf{l}_i^s$ is defined as:

$$l_{i,j} = \begin{cases} 1, \text{if } \mathbf{x}_i^s \text{ belongs to the } j^{th} \text{ emotion states;} \\ 0, \text{otherwise.} \end{cases}$$

By using these source label vectors, we are thus able to construct a new subspace as the predefined common subspace. Note that our DoSL aims at learning a projection matrix $\mathbf{U}$ to project the source speech feature matrix $\mathbf{X}^s$ from the original feature space to such common subspace spanned by the columns of $\mathbf{L}^s$, which can be formulated as the following optimization problem:

$$\min_{\mathbf{U}} \|\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s\|_F^2, \tag{1}$$

Meanwhile, with the projection matrix $\mathbf{U}$, the target speech feature matrix $\mathbf{X}^t$ can also be mapped to the predefined common subspace, where the projected source and target speech features will be enforced to share the similar distributions. To achieve this goal, following the works of MMD criterion [11] and TNNMF [10], we minimize the distance difference between mean projected source speech feature vectors and mean projected target speech feature vectors, which is formulated as follows:

$$\min_{\mathbf{U}} \|\frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{U}^T\mathbf{x}_i^s - \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbf{U}^T\mathbf{x}_i^t\|^2, \tag{2}$$

By minimizing the combination of the above objective functions in Eqs. (1) and (2), we can arrive at the new optimization problem as the following formulation:

$$\begin{aligned} &\min_{\mathbf{U}} \|\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s\|_F^2 \\ &+\lambda_1\|\frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{U}^T\mathbf{x}_i^s - \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbf{U}^T\mathbf{x}_i^t\|^2 + \lambda_2\|\mathbf{U}^T\|_{2,1}, \end{aligned} \tag{3}$$

where λ_1 and λ_2 are the trade-off parameters to control the balance among three terms in the objective functions. It should be also noted that besides previously described combination, we introduce a $L_{2,1}$ norm term with respect to the transpose matrix of $\mathbf{U}$ to serve as the regularization to select the important features contributing to SER [12] during the feature projection. Eq. (3) is namely our DoSL model.

2.2. Optimization

DoSL model is solved by using inexact augmented Lagrange multiplier (IALM) method [14]. More specifically, by introducing a auxiliary variable $\mathbf{Q}$ which satisfies $\mathbf{U} = \mathbf{Q}$, we convert the optimization problem of DoSL to a constrained one which can be expressed as:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{Q}} \|\mathbf{L}^s - \mathbf{Q}^T\mathbf{X}^s\|_F^2 \\ &+\lambda_1\|\frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{Q}^T\mathbf{x}_i^s - \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbf{Q}^T\mathbf{x}_i^t\|^2 + \lambda_2\|\mathbf{U}^T\|_{2,1}, \\ &\text{s.t. } \mathbf{U} = \mathbf{Q}. \end{aligned} \tag{4}$$

Subsequently, the Lagrange function of Eq. (4) can be obtained as follows:

$$\begin{aligned} L(\mathbf{U},\mathbf{Q},\mathbf{T},\mu) &= \|\mathbf{L}^s - \mathbf{Q}^T\mathbf{X}^s\|_F + \lambda_1\|\mathbf{Q}^T\bar{\mathbf{x}}_i^{st}\|^2 \\ &+\lambda_2\|\mathbf{U}^T\|_{2,1} + tr[\mathbf{T}^T(\mathbf{U}-\mathbf{Q})] + \frac{\mu}{2}\|\mathbf{U}-\mathbf{Q}\|_F^2, \end{aligned} \tag{5}$$

where $\bar{\mathbf{x}}_i^{st} = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{x}_i^s - \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbf{x}_i^t$, $\mathbf{T}$ is the Lagrange multiplier, and $\mu > 0$ is the regularization parameter.

Finally, to achieve the optimal solution of $\mathbf{U}$, we only need to iteratively minimize the Lagrange function of Eq. (2) with respect to one of the variables fixing the others until convergence. More specifically, perform the following four steps:

1. Update $\mathbf{Q}$: In this case, the optimization problem would become as below:

$$\min_{\mathbf{Q}} \|\mathbf{L}^s - \mathbf{Q}^T\mathbf{X}^s\|_F^2 + \lambda_1\|\mathbf{Q}^T\bar{\mathbf{x}}_i^{st}\|^2 + tr[\mathbf{T}^T(\mathbf{U}-\mathbf{Q})] + \frac{\mu}{2}\|\mathbf{U}-\mathbf{Q}\|_F^2,$$

which results in

$$\mathbf{Q} = (\frac{\mathbf{X}\mathbf{X}^T}{\mu} + \frac{1}{2}\mathbf{I})^{-1}(\frac{2\mathbf{X}^s\mathbf{L}^{sT} + \mathbf{T}}{\mu} + \mathbf{U}),$$

where $\mathbf{X} = [\mathbf{X}^s, \sqrt{\lambda_1}\bar{\mathbf{x}}_i^{st}]$.

2. Update $\mathbf{U}$: The optimization problem can be rewritten as the following formulation:

$$\min_{\mathbf{U}} \frac{\lambda_2}{\mu}\|\mathbf{U}^T\|_{2,1} + \frac{1}{2}\|\mathbf{U}^T - (\mathbf{Q}^T - \frac{\mathbf{T}^T}{\mu})\|_F^2.$$

According to Lemma 4.1 in [15], the optimal $\mathbf{U}$ can be obtained as follows:

$$\mathbf{u}_i = \begin{cases} \frac{\|\mathbf{q}_i - \frac{\mathbf{t}_i}{\mu}\| - \frac{\lambda_2}{\mu}}{\|\mathbf{q}_i - \frac{\mathbf{t}_i}{\mu}\|}(\mathbf{q}_i + \frac{\mathbf{t}_i}{\mu}), \text{ if } \lambda_2 < \|\mathbf{q}_i - \frac{\mathbf{t}_i}{\mu}\|; \\ 0, \qquad \text{otherwise.} \end{cases}$$

where $\mathbf{q}_i$, $\mathbf{t}_i$, and $\mathbf{u}_i$ are the i^{th} row of $\mathbf{Q}$, $\mathbf{T}$, and $\mathbf{U}$, respectively.

3. Update $\mathbf{T}$ and μ:
$\mathbf{T} = \mathbf{T} + \mu(\mathbf{U}-\mathbf{Q})$, $\mu = \max(\mu_{max}, \rho\mu)$,
where ρ is a scaled parameter.

4. Check convergence: $\|\mathbf{U}-\mathbf{Q}\|_\infty < \epsilon$

2.3. Cross-corpus SER using DoSL

By using the above solving method in Section 2.2 to learn the optimal $\mathbf{U}_*$, we have following method to predict the emotion states of the target speech samples. It is to assign the emotion labels to the target speech signals according to the criterion: $emotion\_labels = \arg\min_k\{[\mathbf{U}_*^T\mathbf{X}^t](k,:)\}$, where $[\mathbf{U}_*^T\mathbf{X}^t](k,j)$ means the k^{th} element of the j^{th} column (target speech signal) of the projected matrix $\mathbf{U}_*^T\mathbf{X}^t$.

3. EXPERIMENTS AND DISCUSSION

In this section, we conduct extensive cross-corpus SER experiments to evaluate the performance of the proposed DoSL method. Three popular speech emotion corpora including EmoDB [16], the audio dataset of eNTERFACE [17], and the audio dataset of AFEW 4.0 [18] are employed. EmoDB covers seven emotion categories: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Suprise. 10 (5f) professional actors speak 10 German emotionally undefined sentences. The eNTERFACE database is composed of 1287 emotion videos from 43 subjects and they are categorized into six basic emotions including Anger, Disgust, Fear, Happy, Sadness, and Surprise. The AFEW 4.0 dataset include three subsets: Train(578 samples), Val(383 samples) and Test(407 samples). Following the experimental protocol of [12], we select any two datasets of three speech corpora each time and select the samples belonging to the common emotion states from these two datasets (e.g., Angry, Disgust, Fear, Joy/Happy and Sad, EmoDB versus eNTERFACE), which are served as source and target corpus, alternatively. Therefore, there are finally six groups of experiments. For convenience, these six experiments are denoted by No.1, No.2, $\cdots$, No.6, respectively, whose detailed source and target speech corpora are illustrated in following Tables 1, 2, and 3.

We use the INTERSPEECH 2009 feature set that consists of 384 elements, i.e., 32 acoustic low-level descriptors (LLDs) and their 12 functions [19], as speech feature representation. As to the evaluation metrics, we employ the weighted average recall (WAR) and the unweighted average recall (UAR) [2] to report the performance of all the method, where WAR is the normal recognition accuracy, while UAR is the mean accuracy of each class. The comparison methods in the experiments are SVM without domain adaptation, KMM [7], KLIEP [8], uLSIF [9], and DALSR [12]. Note that their experimental results are directly taken from Tables I, II, and III in [12] since our experiment setting is exactly same as that of [12]. Finally, the parameters (λ_1, λ_2) of our DoSL are empirically fixed at (1, 5), (63, 29), (4, 18), (9, 6), (14, 4), and (2, 8) for No.1, No.2, $\cdots$, No.6 experiments, respectively. Meanwhile, we use the method described in Section 2.3 for DoSL to predict the emotion labels of target speech samples.

The experimental results in terms of UAR and WAR of all the methods for all six experiments are depicted in Tables 1, 2, and 3.The normal numbers are the recognition rate and the subscript numbers are the relative rank of UAR and WAR in each method. From the results, something interesting can be obtained.

Firstly, it can be found that our DoSL achieves both best UAR and WAR among all the methods in No.1 and No.4 experiments.

Secondly, our DoSL outperforms all the other methods in term of WAR in No.3 experiment and in term of UAR in No.5 experiment. Despite of this, it is clear to see that the UAR of DoSL in No.3 experiment and the WAR of DoSL in No.5 experiment are very competitive against the highest results in respective experiments, which is shown in the comparison between KMM and DoSL in No.3 experiment (30.39% v.s. 29.10%) and the comparison between DALSR and DoSL in No.5 experiment (26.70% v.s. 26.20%).

Finally, although in the remaining experiments (No.2 and No.6), our DoSL does not perform best in terms of both UAR and WAR among all the methods, we can from the results achieved by DoSL and DALSR (highest), observe that their differences of UAR and WAR are actually not large. In

Table 1. Results of the No.1 and No.2 cross-corpus speech emotion recognition experiments in terms of UAR and WAR, where the common emotion states (5 classes) are Angry, Disgust, Fear, Joy/Happy and Sad.

#	Source Corpus	Target Corpus	SVM		KMM		KLIEP		uLSIF		DALSR		DoSL	
			UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
1	EmoDB	eNTERFACE	30.06_3	30.08_3	23.08_5	23.14_5	21.79_6	21.82_6	25.75_4	25.75_4	36.36_2	36.40_2	$\mathbf{37.49}_1$	$\mathbf{37.51}_1$
2	eNTERFACE	EmoDB	27.83_6	24.27_6	40.18_4	44.69_3	28.58_5	27.01_5	40.42_3	42.27_4	$\mathbf{44.41}_1$	$\mathbf{52.27}_1$	44.25_2	52.00_2

Table 2. Results of No.3 and No.4 cross-corpus speech emotion recognition experiments in terms of UAR and WAR, where the common emotion states (6 classes) are Angry, Disgust, Fear, Joy/Happy, Neutral and Sad.

#	Source Corpus	Target Corpus	SVM		KMM		KLIEP		uLSIF		DALSR		DoSL	
			UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
3	EmoDB	AFEW 4.0	26.07_4	25.99_4	$\mathbf{30.39}_1$	29.78_3	25.47_6	25.57_6	25.75_5	25.93_5	27.51_3	30.19_2	29.10_2	$\mathbf{31.00}_1$
4	AFEW 4.0	EmoDB	29.87_5	35.02_5	38.17_2	46.81_3	27.41_6	31.37_6	36.25_4	44.38_4	37.33_3	47.80_2	$\mathbf{39.66}_1$	$\mathbf{50.00}_1$

Table 3. Results of No.5 and No.6 cross-corpus speech emotion recognition experiments in terms of UAR and WAR, where the common emotion states (6 classes) are Angry, Disgust, Fear, Happy, Sad and Surprise.

#	Source Corpus	Target Corpus	SVM		KMM		KLIEP		uLSIF		DALSR		DoSL	
			UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
5	eNTERFACE	AFEW 4.0	20.80_5	18.39_6	23.79_3	25.72_3	18.66_6	18.60_5	22.61_4	21.21_4	24.67_2	$\mathbf{26.70}_1$	$\mathbf{24.83}_1$	26.20_2
6	AFEW 4.0	eNTERFACE	18.68_4	18.72_4	19.75_3	19.75_3	17.48_6	17.47_6	18.10_5	18.11_5	$\mathbf{21.93}_1$	$\mathbf{21.96}_1$	21.64_2	21.66_2

these two cases, the UAR and WAR of DALSR are (44.41%, 52.27%) and (21.93%, 21.96%), while the results of our DoSL are (44.25%, 52.00%) and (21.64%, 21.66%). Additonally, based on our results, it is convincing that the limited label information provided by a small number of samples in source database will lead to low recongtion rate, the data imbalance between source and target databases is an important factor which will affect the cross-cropus speech emotion recognition tasks.

4. CONCLUSIONS

In this paper, we have proposed a domain-adaptive subspace learning (DoSL) model to deal with the unsupervised cross-corpus speech emotion recognition (SER) problem. By using DoSL model, we can learn a projection matrix to transform the source and target speech samples from the original feature space, in which the feature distributions of the source and target speech samples have large difference, into the label space, where the transformed source and target speech samples would obey the similar feature distributions. Therefore, the classifier learned based on the transformed labeled source speech samples are then utilized to predict the speech emotion category of the unlabeled target speech samples. Extensive cross-corpus SER experiments based on various speech emotion corpora are conducted to evaluate the performance of the proposed DoSL method. Experimental results show that our DoSL achieves more overall promising results than recent state-of-the-art cross-corpus SER methods. Since the addition of neural networks is desirable in cross-corpus speech emotion recognition tasks, we will introduce the convolution neutral network into our DoSL method in the future.

5. ACKNOWLEDGEMENTS

This research was supported by the Natural Science Foundation of China under Grants 61172185 and 61602345, the Application Foundation and Advanced Technology Research Project of Tianjin, the Academy of Finland, Tekes Fidipro program and Infotech Oulu.

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] Bjorn 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.

[3] Jun Deng, Zixing Zhang, Erik Marchi, and Bjorn Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2013, pp. 511–516.

[4] 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.

[5] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, 2017.

[6] 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.

[7] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola, "Correcting sample selection bias by unlabeled data," in *Advances in Neural Information Processing Systems*, 2006, pp. 601–608.

[8] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in Neural Information Processing Systems*, 2008, pp. 1433–1440.

[9] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama, "A least-squares approach to direct importance estimation," *The Journal of Machine Learning Research*, vol. 10, pp. 1391–1445, 2009.

[10] 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.

[11] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[12] 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.

[13] Yuan Zong, Wenming Zheng, Xiaohua Huang, Keyu Yan, Jingwei Yan, and Tong Zhang, "Emotion recognition in the wild via sparse transductive transfer linear discriminant analysis," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 163–172, 2016.

[14] 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.

[15] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[16] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of german emotional speech," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[17] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, "The enterface05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006, pp. 8–8.

[18] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Karan Sikka, and Tom Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 461–466.

[19] Björn W Schuller, Stefan Steidl, Anton Batliner, et al., "The interspeech 2009 emotion challenge.," in *Interspeech*, 2009, vol. 2009, pp. 312–315.