INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Joint Instance Reconstruction and Feature Subspace Alignment for Cross-Domain Speech Emotion Recognition

Keke Zhao[1], *Peng Song*[2*], *Shaokai Li*[2], *Wenming Zheng*[3]

[1] The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Tibetan information processing and Machine Translation Key Laboratory of Qinghai province
[2] School of Computer and Control Engineering, Yantai University, China
[3] School of Biological Science and Medical Engineering, Southeast University, China

pengsong@ytu.edu.cn

Abstract

Speech emotion recognition is a popular research branch of speech signal processing. Many previous studies have proven that the generalization ability of the emotion recognition model across domains can be improved by using transfer learning methods. To solve the cross-domain speech emotion recognition problem, this paper proposes a novel transfer learning method, which simultaneously performs the instance reconstruction and subspace alignment. Firstly, we conduct the instance transferring based on coupled projection, which utilizes a weighting reconstruction strategy to exploit the intrinsic information of cross-domain samples and improve the contribution of essential features through an adaptive weighting matrix. Then, we conduct the feature transferring through a novel coregularized term, which can make the source and target subspace be well aligned. Finally, extensive experiments indicate that our method is superior to several state-of-the-art methods.

Index Terms: cross-domain, speech emotion recognition, instance reconstruction, subspace alignment

1. Introduction

Speech emotion recognition (SER) has gained much attention in recent years. Many machine learning methods have been introduced for SER [1], which aim to recognize the emotional states from speech signals, e.g., happiness, anger, sadness, and fear. However, these traditional SER methods always assume that the training and test samples follow the same distribution, which is not applicable in real-world scenarios. Due to several factors, including tagging schemes, linguistic environments, the type of domains (e.g., evoked, performed, and natural), the degree of spontaneity, and even the recording devices, the training and test samples often follow different distributions, which would lead to the degradation of the emotion recognition performance [2]. To address the above problem, the transfer learning technique has been introduced, which can transfer useful information from one or multi-source domains to a related target domain [3]. Many previous works have shown that transfer learning can significantly improve the generalization ability of the classification model, especially when only a small amount of data is available in the target domain [4]. Therefore, in this paper, we focus on designing a transfer learning algorithm to solve the cross-domain SER problem.

This work was supported by the National Natural Science Foundation of China under Grant U2003207, Jiangsu Frontier Technology Basic Research Project under Grant BK20192004, the Natural Science Foundation of Shandong Province under Grant ZR2022MF314, and the Fundamental Research Funds for the Central Universities under Grants 2242021k30014 and 2242021k30059.

In recent years, many transfer learning works have been presented to deal with the cross-domain SER problem. For instance, different normalization strategies on six standard domains are used to reduce the variation between domains. In [5], three types of transfer learning algorithms are introduced for the cross-domain SER tasks. In [6], the feature selection strategy is integrated into a general transfer learning framework to deal with the challenging cross-domain SER problem. In [7], Zhang et al. design a cross-domain graph as the transfer metric to reduce the discrepancy between domains. In [8], Liu et al. propose an unsupervised transfer subspace learning (TRaSL) model for cross-database SER. In [9], Padi et al. utilize the transfer learning and spectrogram augmentation strategy to improve the SER performance. In [10], Li et al. present a novel coupled discriminant subspace alignment (CDSA) method for cross-database SER. In [11], Ghriss et al. propose a multi-task pre-training method for SER, which pre-trains the SER model simultaneously on the automatic speech recognition (ASR) and sentiment classification tasks to make the acoustic ASR model more "emotion aware". The above algorithms have achieved satisfactory results, but they ignore the contribution of different features in the process of knowledge transfer, which would directly affect the performance of emotion recognition. Moreover, they do not fully consider the relationship between the source and target features.

Based on the above analysis, in this article, we propose a novel transfer learning algorithm for cross-domain SER, named joint instance reconstruction and feature subspace alignment (JIFA). Our method takes into account the contribution of essential features by using a new instance reconstruction strategy, which introduces an adaptive weighting matrix on the reconstruction term to narrow the divergence gap across domains. Moreover, the feature subspace is aligned by minimizing the two projection matrices, which can make the source and target domains closer. Comprehensive experimental results demonstrate that our method can learn better feature representations than other transfer subspace learning algorithms. For better illustration, the schematic of our method is given in Fig. 1.

2. Methodology

2.1. Adaptive weighting instance reconstruction

Let $X_s \in R^{m\times n_s}$ and $X_t \in R^{m\times n_t}$ be the labeled source and unlabeled target feature matrices, respectively, where m denotes the dimension of features, where n_s and n_t are the corresponding numbers of samples. We first conduct a linear reconstruction on coupled subspace, in which each target sample can be linearly reconstructed by the source samples in their subspace. The data reconstruction between different domains can effectively reflect the intrinsic information of the data by a

10.21437/Interspeech.2023-275

reconstruction matrix. Also, to eliminate the redundant features and noise in the data, we impose an $\ell_{2,1}-$norm on the data reconstruction matrix to control its row sparsity. The problem can be formulated as

$$\min_{P_s,P_t,Z} \|P_t^T X_t - P_s^T X_s Z\|_F^2 + \gamma \|Z\|_{2,1} \tag{1}$$

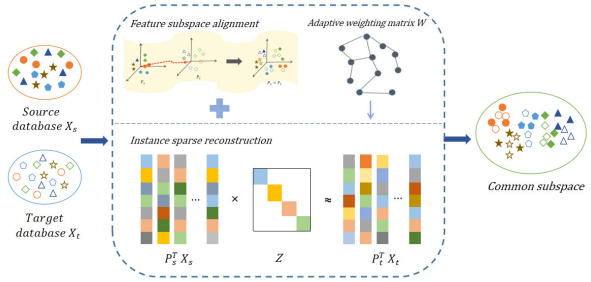


Figure 1: *The schematic of JIFA.*

where $P_s \in R^{m\times d}$ and $P_t \in R^{m\times d}$ are the source and target projection matrices, $Z \in R^{n_s\times n_t}$ is the reconstruction matrix, $||\cdot||_F$ is the Frobenius norm, $||\cdot||_{2,1}$ is the $\ell_{2,1}-$norm, and $\gamma \geq 0$ is used as a regularization parameter to control the sparsity of matrix Z.

The discrepancy between the source and target domains can be reduced by minimizing Eq. (1), but the discriminant information between cross-domain samples is not well discovered. In other words, the important features and redundant features are considered equally important. To overcome this problem, we use an adaptive weighting strategy to penalize the reconstruction error, which can adaptively assign larger weights to the essential features and assign lower weights to the unimportant features. Thus, Eq. (1) can be further written as

$$\min_{P_s,P_t,Z,W} \|W^{\frac{1}{2}} \odot (P_t^T X_t - P_s^T X_s Z)\|_F^2 + \beta \|W\|_F^2 + \gamma \|Z\|_{2,1} \tag{2}$$

where $\beta \geq 0$ is a regularization parameter, $W \in R^{d\times n_t}$ is the adaptive weighting matrix, and the operator $W^{\frac{1}{2}}$ represents its square root. We impose a constraint $W^T \boldsymbol{1} = \boldsymbol{1}$ to limit the range of the element values of W, where $\boldsymbol{1} \in R^{d\times 1}$ is a vector whose elements are 1. Hence, Eq. (2) can be reformulated as

$$\begin{aligned} &\min_{P_s,P_t,Z,W} \|W^{\frac{1}{2}} \odot (P_t^T X_t - P_s^T X_s Z)\|_F^2 + \beta \|W\|_F^2 + \gamma \|Z\|_{2,1} \\ &\text{s.t. } Z \geq 0,\ W^T \boldsymbol{1} = \boldsymbol{1},\ W \geq 0 \end{aligned} \tag{3}$$

where $Z \geq 0$ and $W \geq 0$ are the non-negative constraints to ensure good interpretability of the samples, in which the similar relationships between these cross-domain samples can be directly reflected.

2.2. Feature subspace alignment

Transfer learning aims to enable two domains to learn from each other to improve overall performance. Hence, we need to integrate the compatible and complementary information from the two domains. To this end, we might align the feature subspace through various standard normalization terms to reduce the divergence between the low-dimensional representations, which would facilitate the knowledge transfer of information across domains.

As shown in Eq. (3), two projection matrices are used to reconstruct the low-dimensional representations. Each column of the projection matrices can be regarded as the coding of the original features. Therefore, JIFA attempts to minimize the divergence between each pair feature from source and target projection matrices as follows:

$$D(P_s, P_t) = \| \frac{L_{P_s}}{\|L_{P_s}\|_F^2} - \frac{L_{P_t}}{\|L_{P_t}\|_F^2} \|_F^2 \tag{4}$$

where $L_{P_s} = P_s P_s^T, L_{P_t} = P_t P_t^T$, and L_{P_s} and L_{P_t} represent the graphs that contain the relationships between all features in the source and target domains, respectively. In other words, L_{P_s} and L_{P_t} are the adjacency matrices that is typically the linear kernel matrices. For L_{P_s} and L_{P_t}, the similarity weight of the features of each pair of nodes is calculated by using the inner product of every two features. Minimizing Eq. (4) encourages the feature subspace to learn from each other and bridge the gap between them. Furthermore, $D(P_s, P_t)$ can be replaced with a trace form through mathematical deductions [12] as follows:

$$D(P_s, P_t) = -\operatorname{Tr}(P_s P_s^T P_t P_t^T) \tag{5}$$

Combining Eqs. (3) and (5), we can get the objective function of the proposed method as follows:

$$\begin{aligned} \min_{P_s,P_t,Z,W} &\|W^{\frac{1}{2}} \odot (P_t^T X_t - P_s^T X_s Z)\|_F^2 \\ &- \alpha \operatorname{Tr}(P_s P_s^T P_t P_t^T) + \beta \|W\|_F^2 + \gamma \|Z\|_{2,1} \\ \text{s.t. } &Z \geq 0,\ W^T \boldsymbol{1} = \boldsymbol{1},\ W \geq 0 \end{aligned} \tag{6}$$

where $\alpha \geq 0$ is a regularization parameter.

3. Optimization

In this section, we use the alternating direction method of multipliers (ADMM) [13] strategy to facilitate the optimization of Eq. (6). Let $M = P_t^T X_t - P_s^T X_s Z$, we get the following Lagrangian function:

$$\begin{aligned} L = &\|W^{\frac{1}{2}} \odot M\|_F^2 - \alpha \operatorname{Tr}(P_s P_s^T P_t P_t^T) + \beta \|W\|_F^2 \\ &+ \frac{\mu}{2} \|P_t^T X_t - P_s^T X_s Z - M + \frac{C}{\mu}\|_F^2 + \gamma \|Z\|_{2,1} \end{aligned} \tag{7}$$

where $\mu \geq 0$ is a penalty regularization parameter, and C is a Lagrange multiplier. We update one variable by fixing other variables. The procedures are given as follows.

(1) Update W: Denote w_j as the $j-$th column of the matrix W, $V = M \odot M$, and v_j is the $j-$th column of V, we can obtain

$$\min \sum_{j=1}^{n_t} \|w_j + \frac{1}{\beta} v_j\|_2^2 \quad \text{s.t. } w_j \geq 0,\ w_j^T \boldsymbol{1} = 1 \tag{8}$$

The above equation can be transformed into the following Lagrangian form:

$$L(w_j, \delta_j, \kappa_j) = \frac{1}{2} \|w_j + \frac{1}{\beta} v_j\|_2^2 - \delta_j (w_j^T \boldsymbol{1} - 1) - \kappa_j^T w_j \tag{9}$$

where $\delta_j \geq 0$ and $\kappa_j \geq 0$ are the introduced Lagrangian multipliers. By computing the derivative of $L(w_j, \delta_j, \kappa_j)$ w.r.t. w_j, we get

$$\frac{\partial L(w_j, \delta_j, \kappa_j)}{\partial w_j} = w_j + \frac{1}{\beta} v_j - \delta_j \mathbf{1} - \kappa_j \tag{10}$$

By adding a constraint, i.e., $w_j^T \mathbf{1} = 1$, and using the Karush-Kuhn-Tucker (KKT) condition, i.e., $\kappa_j \odot w_j = 0$, we can get

$$w_j = \max\left(\delta_j \mathbf{1} - \frac{1}{\beta} v_j, 0\right), \ \delta_j = \frac{1}{d} + \frac{1}{d\beta} \sum_{i=1}^{d} v_{ij} \tag{11}$$

where d is the number of elements of the vector v_i. The closed-form solutions of w_j and δ_j are obtained by iteratively optimizing Eq. (11). In this way, we can get the optimal adaptive weighting matrix W.

(2) Update M: Let $H = P_t^T X_t - P_s^T X_s Z - M + \frac{C}{\mu}$, we can obtain

$$\sum_{i=1}^{d} \sum_{j=1}^{n_t} \min_{m_{ij}} \left(m_{ij} - \frac{\mu h_{ij}}{\mu + 2w_{ij}}\right)^2 \tag{12}$$

where h_{ij} and m_{ij} represent the elements of H and M, respectively. The optimal m_{ij} can be expressed as follows:

$$m_{ij} = \frac{\mu h_{ij}}{\mu + 2w_{ij}} \tag{13}$$

(3) Update Z: According to [14], we use an iterative optimization algorithm to solve the $\ell_{2,1}-$norm. Define $\|Z\|_{2,1} = \mathrm{Tr}(Z^T G Z)$, where $G \in R^{n_s \times n_s}$ is a diagonal matrix, i.e., G=$diag(\frac{1}{2\|z_1\|^2}, \frac{1}{2\|z_2\|^2}, \ldots, \frac{1}{2\|z_{n_s}\|^2})$, in which z_i is the $i-$th row of Z. Let $T = P_t^T X_t - E + \frac{C}{\mu}$. By calculating the derivative of L w.r.t. Z, we get the following equation:

$$\frac{\partial L}{\partial Z} = \gamma G Z - \mu X_s^T P_s T + \mu X_s^T P_s P_s^T X_s Z \tag{14}$$

By setting $\frac{\partial L}{\partial Z} = 0$, we can get

$$Z = \frac{\mu X_s^T P_s T}{\gamma G + \mu X_s^T P_s P_s^T X_s} \tag{15}$$

(4) Update P_s and P_t: By taking the derivative of L w.r.t. P_s, we can obtain

$$\frac{\partial L}{\partial P_s} = -\alpha P_t P_t^T P_s + \mu X_s Z Z^T X_s^T P_s - \mu X_s Z T^T \tag{16}$$

By setting $\frac{\partial L}{\partial P_s} = 0$, we can get

$$P_s = \frac{\mu X_s Z T^T}{\mu X_s Z Z^T X_s^T - \alpha P_t P_t^T} \tag{17}$$

Similarly, let $Q = P_s^T X_s + E + \frac{C}{\mu}$, we can obtain

$$P_t = \frac{\mu X_t Q^T}{\mu X_t X_t^T - \alpha P_s P_s^T} \tag{18}$$

(5) Update C and μ: we can obtain

$$C = C + \mu(P_t^T X_t - P_s^T X_s Z - E) \tag{19}$$

$$\mu = \min(\mu_{\max}, \rho\mu) \tag{20}$$

where $\mu_{\max}$ and ρ are constants.

Algorithm 1 JIFA (solving Eq. (6))

Input: Source feature matrix X_s and target feature matrix X_t; the regularization parameters α, β, γ; and a small threshold value ε.
Output: The source projection matrix P_s, the target projection matrix P_t, the reconstruction matrix Z and adaptive weighting matrix W.
Initialize: Initialize P_s and P_t via PCA; Initialize Z and set $t = 0$.
repeat
1. Fix other variables and update W by using Eq. (11);
2. Fix other variables and update M by using Eq. (13);
3. Fix other variables and update Z by using Eq. (15);
4. Fix other variables and update P_s and P_t by using Eq. (17) and Eq. (18);
5. Fix other variables and update C and μ by using Eq. (19) and Eq. (20);
6. $t = t + 1$;
7. Check the convergence condition: if $t > 2$ and $\Delta L = L^{(t)} - L^{(t-1)} < \varepsilon$, where $L^{(t)}$ is the objective value in the t-th iteration;

until Convergence condition is satisfied or the maximum number of iterations is reached.
return P_s, P_t, Z, W.

4. Experiments

4.1. Experimental setup

In this section, we evaluate the effectiveness of the proposed algorithm on four benchmark datasets: Emo-DB (*Em*) [15], eNTERFACE (*En*) [16], RML (*Rm*) [17], and BAUM-1a (*Ba*) [18]. Two of the above datasets are randomly used as the source and target domains, so we can get 12 sets of cross-domain SER (source→target), i.e., *Em→En, En→Em, Em→Rm, Rm→Em, Em→Ba, Ba→Em, En→Rm, Rm→En, En→Ba, Ba→En, Rm→Ba*, and *Ba→Rm*. In the experiments, we consider five common emotion categories: anger, sadness, disgust, happiness, and fear. For training and test data, the target dataset are split into ten parts and 8/10 of them with the source dataset are used for training and the rest are used for testing.

To evaluate the efficacy of our method, we select several popular transfer subspace learning algorithms as the baseline methods, including joint distribution adaptation (JDA) [19], transfer joint matching (TJM) [20], latent sparse domain transfer learning (LSDT) [21], feature selection based transfer subspace learning (FSTSL) [6], transfer sparse discriminant subspace learning (TSDSL) [7], guide subspace learning (GSL) [22], and coupled discriminant subspace alignment (CDSA) [10]. Additionally, we select two popular traditional subspace learning methods, i.e., principal component analysis (PCA) [23] and linear discriminant analysis (LDA) [24], for comparison. We choose the linear SVM as the basic classifier for all the methods and use the weighting average recall (WAR) as the experimental evaluation metric.

4.2. Results and discussions

We report the recognition performance of different algorithms in Table 1. From this table, we obtain the following observations.

Firstly, the recognition performance of JIFA outperforms that of all the baseline methods in most cases. This demonstrates that our method can effectively solve the cross-domain SER problem.

Secondly, the performance of all the transfer learning methods, including ours, is significantly better than that of traditional subspace learning methods. This indicates that the transfer

Table 1: *Recognition performance (WAR%) of different algorithms in 12 tasks. The best performance is shown in bold.*

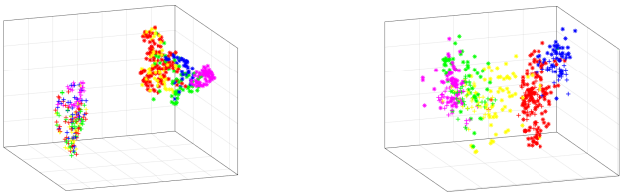
Tasks	Traditional methods		Transfer learning methods							JIFA
	PCA	LDA	JDA	FSTSL	TJM	LSDT	GSL	TSDSL	CDSA	
Em→En	36.02	38.60	38.14	37.21	39.53	37.67	33.05	**43.25**	42.65	43.02
En→Em	32.35	39.71	45.59	32.55	41.18	30.88	35.71	50.00	47.10	**51.47**
Em→Rm	22.22	26.39	25.93	32.87	29.63	32.50	36.19	38.09	44.08	**45.36**
Rm→Em	23.65	22.06	38.24	29.12	29.41	32.06	39.29	41.17	51.52	**56.18**
Em→Ba	40.00	34.29	34.29	38.57	37.14	37.14	39.15	37.28	**50.20**	43.21
Ba→Em	32.35	44.12	44.12	41.18	45.59	30.88	35.04	42.64	**50.56**	49.71
En→Rm	31.48	28.24	28.24	34.24	31.02	45.00	32.86	41.01	**46.31**	43.33
Rm→En	27.95	31.16	31.63	26.05	29.77	33.49	34.07	33.48	38.70	**46.95**
En→Ba	25.71	31.43	28.57	26.71	20.00	28.57	28.23	37.14	**48.18**	42.86
Ba→En	33.49	28.37	26.98	36.98	36.28	33.95	31.63	35.53	35.05	**41.77**
Rm→Ba	23.14	26.43	27.14	30.00	22.86	37.14	35.38	42.57	40.38	**43.86**
Ba→Rm	40.43	37.50	24.17	40.83	34.17	43.33	36.11	37.67	44.54	**45.84**
Average	30.73	32.36	32.75	33.85	33.05	35.22	34.72	39.98	44.93	**46.13**

learning algorithms can efficiently solve the domain mismatch problem, whereas the traditional subspace learning algorithms do not consider this problem.

Thirdly, in most tasks, our method is superior to the transfer subspace learning methods, i.e., TSDSL and CDSA. The reason might be as follows. We introduce a novel instance reconstruction strategy, which considers the contribution of essential features to better reflect the relationship between different domains. Meanwhile, we align the feature subspace to facilitate the information transfer between two domains.

4.3. t-SNE visualization

In this subsection, we give the visualization results of t-SNE [25] in Fig. 2. Here we take the Em→Ba task as an example. Fig. 2 (a) and (b) show the original data and the data projected by the proposed method, respectively. From the figure, we can notice that our method can make the source and target data follow similar distributions, and the samples of the same category are close to each other.



(a) Original data (b) JIFA

Figure 2: *The t-SNE visualization of Em→Ba. The "+" and "∗" indicate the source and target samples, respectively, and different colors indicate different emotion categories.*

4.4. Ablation study

To further prove the effectiveness of our method, we conduct an ablation study. By setting α, β and γ to zero, we can get the following three special cases, i.e., JIFA$_1$, JIFA$_2$ and JIFA$_3$. Fig. 3 shows the comparison results of JIFA and its three special cases. From the figure, we can notice that JIFA achieves significantly better performance than the three special cases, which proves that both the weighting instance reconstruction and feature subspace alignment items can improve the recognition results.

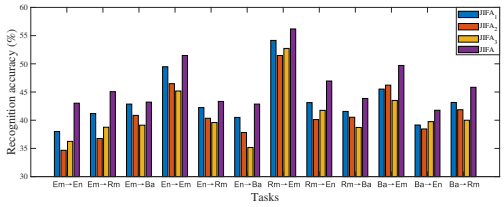


Figure 3: *Results of our method and three special cases.*

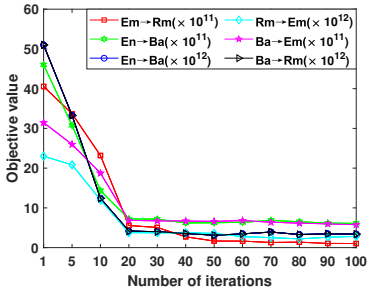


Figure 4: *Convergence curves of our method.*

4.5. Convergence analysis

As discussed in Sec. 3, an iterative optimization algorithm is developed to solve our method. In this section, we study the convergence property of our method. Fig. 4 shows the results under six settings. From the figure, we find that the target values decrease steadily with the increase of the number of iterations and converge after 20 iterations. This result demonstrates that our method has good convergence properties.

5. Conclusions

In this paper, we propose a new joint instance reconstruction and feature subspace alignment method for cross-domain SER. To be specific, we first develop an adaptive instance reconstruction strategy to reduce the divergence across domains. In this way, the target samples can be linearly reconstructed by the target samples. In addition, we consider the contribution of the essential features through an adaptive weighting matrix learning strategy. Furthermore, we develop a feature subspace alignment strategy to align the source and target subspace. Extensive experimental results on four benchmark datasets verify the efficacy of our method.

6. References

[1] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[2] E. N. N. Ocquaye, Q. Mao, H. Song, G. Xu, and Y. Xue, "Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition," *IEEE Access*, vol. 7, pp. 93 847–93 857, 2019.

[3] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[5] A. Hassan, R. Damper, and M. 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.

[6] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 373–382, 2018.

[7] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 307–318, 2020.

[8] N. Liu, B. Zhang, B. Liu, J. Shi, L. Yang, Z. Li, and J. Zhu, "Transfer subspace learning for unsupervised cross-corpus speech emotion recognition," *IEEE Access*, vol. 9, pp. 95 925–95 937, 2021.

[9] S. Padi, S. O. Sadjadi, R. D. Sriram, and D. Manocha, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 645–652.

[10] S. Li, P. Song, K. Zhao, W. Zhang, and W. Zheng, "Coupled Discriminant Subspace Alignment for Cross-database Speech Emotion Recognition," in *Proc. Interspeech 2022*, 2022, pp. 4695–4699.

[11] A. Ghriss, B. Yang, V. Rozgic, E. Shriberg, and C. Wang, "Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7347–7351.

[12] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[14] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint L2, 1-norms minimization," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1813–1821, 2010.

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[16] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[17] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2016.

[18] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[19] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2200–2207.

[20] ——, "Transfer joint matching for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1410–1417.

[21] L. Zhang, W. Zuo, and D. Zhang, "LSDT: Latent sparse domain transfer learning for visual adaptation," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1177–1191, 2016.

[22] L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, and C. P. Chen, "Guide subspace learning for unsupervised domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3374–3388, 2019.

[23] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[24] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[25] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2559–2566.