LETTER

A novel Adaptive Weighted Transfer Subspace Learning Method for Cross-Database Speech Emotion Recognition*

Keke ZHAO[†], ***Nonmember***, **Peng SONG**[†a)], ***Member***, **Shaokai LI**[†], **Wenjing ZHANG**[†], ***and*** **Wenming ZHENG**[††], ***Nonmembers***

SUMMARY In this letter, we present an adaptive weighted transfer subspace learning (AWTSL) method for cross-database speech emotion recognition (SER), which can efficiently eliminate the discrepancy between source and target databases. Specifically, on one hand, a subspace projection matrix is first learned to project the cross-database features into a common subspace. At the same time, each target sample can be represented by the source samples by using a sparse reconstruction matrix. On the other hand, we design an adaptive weighted matrix learning strategy, which can improve the reconstruction contribution of important features and eliminate the negative influence of redundant features. Finally, we conduct extensive experiments on four benchmark databases, and the experimental results demonstrate the efficacy of the proposed method.

key words: *speech emotion recognition, subspace learning, adaptive weighted matrix, transfer learning*

1. Introduction

Speech is an important vehicle for human communication, which can reflect human's emotional states and semantic information. The goal of SER is to identify emotions from speech signals, such as anger, disgust, fear, happiness, and sadness [1]. In practical situations, due to the difference in speakers, recording devices, languages, and environments, the training and test data often follow different distributions, which would lead to inferior recognition performance [2].

To solve the above-mentioned problem, many transfer learning algorithms have been proposed. Comprehensive surveys can be found in Refs. [3], [4]. Recently, various transfer learning algorithms have been presented for cross-database SER. In [5], Schuller et al. conduct extensive experiments to investigate the cross-database SER problem. In [6], Hassan et al. introduce three types of transfer learning algorithms, i.e., kernel mean matching (KMM), Kullback–Leibler importance estimation process (KLIEP), and unconstrained least-squared importance fitting (uLSIF), for cross-database SER. In [7], Liu et al. propose a domain adaptive subspace learning approach, which learns a projection matrix to transform the original feature space to the label space. In [8], Song et al. develop a feature selection based transfer subspace learning framework. In [9], Zhang et al. present a transfer sparse discriminant subspace learning (TSDSL) method for cross-database SER. Note that these algorithms do not fully consider the contribution of different features in the process of knowledge transfer, which is very important for transfer learning.

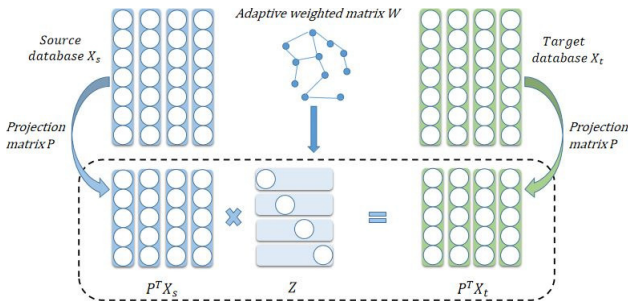

Fig. 1 The diagram of the proposed method.

Motivated by the above discussions, in this letter, we propose a novel adaptive weighted transfer subspace learning (AWTSL) approach for cross-database SER, in which the adaptive weighted subspace learning, transfer learning, and feature selection are integrated into a joint framework. Our method aims to learn a common feature subspace across databases by utilizing a novel feature reconstruction strategy. By introducing an adaptive weighted matrix, the contribution of important features is strengthened. Compared with other transfer subspace learning algorithms, our approach can learn more transferable feature representations. The framework is illustrated in Fig. 1.

2. Methodology

We first introduce the main notations used in this letter. Let $X_s \in R^{m\times n_s}$ and $X_t \in R^{m\times n_t}$ be the source and target feature matrices, respectively, where m denotes the feature dimensionality, and n_s and n_t are the corresponding numbers of samples. $W \in R^{d\times n_t}$ is the adaptive weighted matrix, $P \in R^{m\times d}$ is the projection matrix, $Z \in R^{n_s\times n_t}$ is the reconstruction matrix, and d is the dimensionality of the common subspace. Given a matrix M, $\|M\|_F$ denotes the Frobenius norm of M, $\|M\|_{2,1}$ denotes the $\ell_{2,1}-$norm of M, which is

Manuscript received March 1, 2022.
Manuscript revised May 12, 2022.
Manuscript publicized June 9, 2022.

[†]The authors are with the School of Computer and Control Engineering, Yantai University, Yantai 264005, China.
[††]The author is with the Key Laboratory of Child Development and Learning Science, Ministry of Education, and School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China.

*The work is partly supported by the National Natural Science Foundation of China (Grant No.61703360) and the Fundamental Research Funds for the Central Universities (Grant No. 2242021k30014 and 2242021k30059).

a) E-mail: pengsong@ytu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2022EDL8021

the sum of ℓ_2−norm of rows of M, and $\mathrm{Tr}(M)$ denotes the trace operation of M.

2.1 The Proposed Method

To reduce the divergence across databases, we aim to learn a projection matrix to find a common subspace, in which the source and target features are efficiently merged. Here we assume that each target sample is linearly reconstructed by using the source samples in the learned subspace. The objective function is formulated as

$$\min_{P,Z} \left\| P^T X_t - P^T X_s Z \right\|_F^2 \tag{1}$$

where $P \in R^{m \times d}$ is the projection matrix, d is the dimensionality of the common subspace, and $Z \in R^{n_s \times n_t}$ is the reconstruction matrix.

Note that the above feature reconstruction strategy treats all features equally. To further improve the feature transferable ability, we develop an adaptive weighted matrix strategy to regularize the reconstruction error, which adaptively assigns larger weights to important features. Thus, the objective function in Eq. (1) can be reformulated as

$$\min_{P,Z,W} \left\| W^{\frac{1}{2}} \odot (P^T X_t - P^T X_s Z) \right\|_F^2 \tag{2}$$

where $\odot$ denotes the element-wise multiplication, $W \in R^{d \times n_t}$ is an adaptive weighted matrix, and $W^{\frac{1}{2}}$ is the square root of W with all positive elements.

To make the values of W within a reasonable range, we impose a constraint, i.e., $W^T \mathbf{1} = \mathbf{1}$, on W. Meanwhile, to avoid the trivial solution, we add a regularization term $\|W\|_F^2$. Hence, Eq. (2) can be reformulated as

$$\begin{aligned} &\min_{P,Z,W} \left\| W^{\frac{1}{2}} \odot (P^T X_t - P^T X_s Z) \right\|_F^2 + \alpha \|W\|_F^2 \\ &\text{s.t. } W \geq 0,\ W^T \mathbf{1} = \mathbf{1} \end{aligned} \tag{3}$$

where $\mathbf{1} \in R^{d \times 1}$ is a vector that all elements are 1, and α is a regularization parameter.

In practice, the original features might be redundant and contain noises. Thus, it is necessary to select useful features for knowledge transfer. Here, we impose an $\ell_{2,1}$−norm constraint on the projection matrix P and the data reconstruction matrix Z to make the rows sparse. Then, we can obtain the following objective function:

$$\begin{aligned} &\min_{P,Z,W} \left\| W^{\frac{1}{2}} \odot (P^T X_t - P^T X_s Z) \right\|_F^2 + \alpha \|W\|_F^2 + \beta \|Z\|_{2,1} + \gamma \|P\|_{2,1} \\ &\text{s.t.} W \geq 0,\ W^T \mathbf{1} = \mathbf{1} \end{aligned} \tag{4}$$

where β and γ are the regularization parameters.

Additionally, to avoid trivial solution, we introduce a constraint $P^T X H X^T P = I$. Then, we can obtain the following objective function:

$$\begin{aligned} &\min_{P,Z} \left\| W^{\frac{1}{2}} \odot (P^T X_t - P^T X_s Z) \right\|_F^2 + \alpha \|W\|_F^2 + \beta \|Z\|_{2,1} + \gamma \|P\|_{2,1} \\ &\text{s.t.} W \geq 0,\ W^T \mathbf{1} = \mathbf{1},\ Z \geq 0,\ P^T X H X^T P = I \end{aligned} \tag{5}$$

where $H = I - \frac{1}{n}\mathbf{1}$ is the centering matrix, I is an identity matrix, and $\mathbf{1}$ is a matrix whose elements are all 1, $X = [X_s, X_t] \in R^{m \times n}$ is the total data matrix, and $n = n_s + n_t$ is the number of all samples. Here we impose a non-negative constraint on Z to ensure good interpretability of the samples, which can directly reflect the similarity relationship of two samples.

2.2 Optimization

To solve the objective function in Eq. (5), we develop an iterative alternative optimization algorithm by utilizing the alternating direction method of multipliers (ADMM) [10]. Assume $E = P^T X_t - P^T X_s Z$, we can obtain the following augmented Lagrangian function:

$$\begin{aligned} L(W, E, Z, P, C) = &\left\| W^{\frac{1}{2}} \odot E \right\|_F^2 + \alpha \|W\|_F^2 + \beta \|Z\|_{2,1} + \gamma \|P\|_{2,1} \\ &+ \frac{\mu}{2} \left\| P^T X_t - P^T X_s Z - E + \frac{C}{\mu} \right\|_F^2 \end{aligned} \tag{6}$$

where C is the Lagrange multiplier and μ is the penalty regularization parameter.

In the iterative optimization method, when one variable is solved, the other variables are fixed. It mainly includes five main sub-processes as follows:

(1) Update W by fixing the other variables, we can obtain

$$\min_{W} \left\| W^{\frac{1}{2}} \odot E \right\|_F^2 + \alpha \|W\|_F^2 \quad \text{s.t.} W \geq 0,\ W^T \mathbf{1} = \mathbf{1} \tag{7}$$

Suppose w_j is the j−th column of W, $F = E \odot E$, and f_j is the j−th column of F, the above equation is equivalent to the following minimization problem:

$$\min \sum_{j=1}^{n_t} \left\| w_j + \frac{1}{\alpha} f_j \right\|_2^2 \quad \text{s.t. } w_j \geq 0,\ w_j^T \mathbf{1} = 1 \tag{8}$$

Then, Eq. (8) can be transformed into the following Lagrangian form:

$$L(w_j, \eta_j, \lambda_j) = \frac{1}{2} \left\| w_j + \frac{1}{\alpha} f_j \right\|_2^2 - \eta_j (w_j^T \mathbf{1} - 1) - \lambda_j^T w_j \tag{9}$$

where η_j and λ_j are the Lagrangian multipliers.

By computing the partial derivative of L w.r.t. w_j, we can obtain

$$\frac{\partial L(w_j, \eta_j, \lambda_j)}{\partial w_j} = w_j + \frac{1}{\alpha} f_j - \eta_j \mathbf{1} - \lambda_j \tag{10}$$

According to the Karush-Kuhn-Tucker (KKT) condition $\lambda_j \odot w_j = 0$ and the constraint $w_j^T \mathbf{1} = 1$, we can obtain

Table 1 The recognition accuracy (%) of different methods in different tasks.

Tasks	Traditional methods		Transfer learning methods						AWTSL
	PCA	LDA	JDA	TLDA	TJM	LSDT	GSL	TSDSL	
$E \rightarrow e$	36.02	38.60	38.14	37.21	39.53	37.67	33.05	**43.25**	40.47
$E \rightarrow R$	22.22	26.39	25.93	32.87	29.63	32.50	36.19	38.09	**44.17**
$E \rightarrow B$	40.00	34.29	34.29	38.57	37.14	37.14	39.15	37.28	**42.86**
e→E	32.35	39.71	45.59	32.55	41.18	30.88	35.71	50.00	**57.35**
$e \rightarrow R$	31.48	28.24	28.24	34.24	31.02	45.00	32.86	41.01	**47.50**
$e \rightarrow B$	25.71	31.43	28.57	26.71	20.00	28.57	28.23	37.14	**40.00**
$R \rightarrow E$	23.65	22.06	38.24	29.12	29.41	32.06	39.29	41.17	**60.29**
$R \rightarrow e$	27.95	31.16	31.63	26.05	29.77	33.49	34.07	33.48	**36.74**
$R \rightarrow B$	23.14	26.43	27.14	30.00	22.86	37.14	35.38	42.57	**51.43**
$B \rightarrow E$	32.35	44.12	44.12	41.18	45.59	30.88	35.04	42.64	**54.41**
$B \rightarrow e$	33.49	28.37	26.98	**36.98**	36.28	33.95	31.63	35.53	35.35
$B \rightarrow R$	40.43	37.50	24.17	40.83	34.17	**43.33**	36.11	37.67	42.50
Average	30.73	32.36	32.75	33.85	33.05	35.22	34.72	39.98	**46.09**

$$w_j = \max\left(\eta_j \boldsymbol{1} - \frac{1}{\alpha} f_j, 0\right) \tag{11}$$

$$\eta_j = \frac{1}{d} + \frac{1}{d\alpha} \sum_{i=1}^{d} f_{ij} \tag{12}$$

where f_{ij} denotes the $j-$th element of vector f_i, and d represents the total number of elements of vector f_i. When η_j is computed, the optimal solution for w_j can be obtained by Eq. (11), which in turn gives the optimal solution for the adaptive weighted matrix W.

(2) Update E by fixing the other variables, we can obtain

$$\min_{E} \left\| W^{\frac{1}{2}} \odot E \right\|_F^2 + \frac{\mu}{2} \left\| P^T X_t - P^T X_s Z - E + \frac{C}{\mu} \right\|_F^2 \tag{13}$$

Let $G = P^T X_t - P^T X_s Z - E + \frac{C}{\mu}$, we can obtain

$$\sum_{i=1}^{d} \sum_{j=1}^{n_t} \min_{e_{ij}} \left(e_{ij} - \frac{\mu g_{ij}}{\mu + 2w_{ij}} \right)^2 \tag{14}$$

where g_{ij} and e_{ij} denote the $i-$th row and $j-$th column elements of G and E, respectively. It can be deduced that the solution for e_{ij} is

$$e_{ij} = \frac{\mu g_{ij}}{\mu + 2w_{ij}} \tag{15}$$

(3) Update Z by fixing the other variables, we can obtain the following objective function:

$$L(Z) = \beta \|Z\|_{2,1} + \frac{\mu}{2} \left\| P^T X_t - P^T X_s Z - E + \frac{C}{\mu} \right\|_F^2 \tag{16}$$

Since the objective function contains the $\ell_{2,1}-$norm, which is non-smooth and difficult to be solved directly. According to [11], we introduce an iterative optimization algorithm to solve it. $\|Z\|_{2,1}$ can be expressed as $\|Z\|_{2,1} = \mathrm{Tr}(Z^T BZ)$, where $B \in R^{n_s \times n_s}$ is a diagonal matrix, $B = diag(\frac{1}{2\|z_1\|^2}, \frac{1}{2\|z_2\|^2}, \ldots, \frac{1}{2\|z_{n_s}\|^2})$, in which z_i means the $i-$th row of Z. Let $M = P^T X_t - E + \frac{C}{\mu}$, by taking the derivative of $L(Z)$ w.r.t. Z, we obtain

$$\frac{\partial L(Z)}{\partial Z} = \beta BZ - \mu X_s^T PM + \mu X_s^T PP^T X_s Z \tag{17}$$

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

$$Z = \frac{\mu X_s^T PM}{\beta B + \mu X_s^T PP^T X_s} \tag{18}$$

(4) Update P by fixing the other variables, we obtain

$$\begin{aligned} L(P) =& \gamma \|P\|_{2,1} + \frac{\mu}{2} \left\| P^T X_t - P^T X_s Z - E + \frac{C}{\mu} \right\|_F^2 \\ &+ \phi \,\mathrm{Tr}(P^T XHX^T P - I) \end{aligned} \tag{19}$$

Let $T = E - \frac{C}{\mu}$, $V = X_t - X_s Z$, $\|P\|_{2,1} = \mathrm{Tr}(P^T AP)$, and $A = diag(\frac{1}{2\|p_1\|^2}, \frac{1}{2\|p_2\|^2}, \ldots, \frac{1}{2\|p_m\|^2})$, in which p_i means the $i-$th row of P. Calculating the derivative of $L(P)$ w.r.t. P, we get

$$\frac{\partial L(P)}{\partial P} = \gamma AP + \mu VV^T P + \mu VT^T + \phi XHX_T P \tag{20}$$

By setting $\frac{\partial L(P)}{\partial P} = 0$, we can obtain

$$P = \frac{\mu VT^T}{\gamma A + \mu VV^T + \phi XHX^T} \tag{21}$$

(5) Update C and μ by fixing the other variables, we obtain

$$C = C + \mu(P^T X - P^T XZ - E) \tag{22}$$

$$\mu = \min(\mu_{\max}, \rho\mu) \tag{23}$$

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

The above five steps are repeated until convergence or the maximum number of iterations is reached.

3. Experiments

3.1 Experimental Setup

In the experiments, we use four popular emotional databases, including Emo-DB (E), eNTERFACE (e), RML (R), and BAUM-1a (B) [1]. Based on these databases,

we conduct 12 different settings of experiments for cross-database SER (source → target), i.e., $E \rightarrow e$, $E \rightarrow R$, $E \rightarrow B$, $e \rightarrow E$, $e \rightarrow R$, $e \rightarrow B$, $R \rightarrow E$, $R \rightarrow e$, $R \rightarrow B$, $B \rightarrow E$, $B \rightarrow e$, and $B \rightarrow R$. we select five common emotional categories for evaluation, namely anger, disgust, fear, happiness, and sadness. In our experiments, we assume the source database is labeled and the target database is unlabeled, and all the source data and 4/5 of the target data is used for training, while 1/5 of the target data is used for testing. A 1582 dimensional standard feature set in INTERSPEECH 2010 Paralinguistic challenge [12] is used for evaluation.

To evaluate the efficacy of the proposed method, we compare it with several popular subspace learning algorithms, including principal component analysis (PCA), linear discriminant analysis (LDA), joint distribution adaptation (JDA) [13], transfer linear discriminant analysis (TLDA) [8], transfer joint matching (TJM) [14], latent sparse domain transfer Learning (LSDT) [15], guide subspace learning (GSL), [16], and transfer sparse discriminant subspace learning (TSDSL) [9]. The two main trade-off parameters α and β are tuned in the range of $\{10^{-4} \sim 10^{4}\}$. The maximum iteration number is set to 20. We choose the popular linear support vector machine (SVM) as the classifier and use the average recognition accuracy for evaluation.

3.2 Results and Discussions

Table 1 shows the recognition results of different methods. From the table, we have the following observations.

First, the proposed method achieves much better average recognition performance in comparison with the baseline algorithms. This demonstrates that our method can effectively solve the database mismatch problem for cross-database SER.

Second, the recognition performance of transfer subspace learning methods is better than that of traditional subspace learning methods. The reason is that the transfer subspace learning algorithms can solve the database mismatch problem to some extent, which is neglected in traditional subspace learning algorithms.

Third, our method significantly outperforms the second best algorithm TSDSL in most tasks. The reason might be that, on one hand, TSDSL utilizes a graph distance metric, which only considers the local similarity and cannot well discover the relationship between different databases. On the other hand, TSDSL equally treats different features, and neglects the contribution of important feature. By contrast, we develop a global sparse reconstruction matrix, which ensures that the target samples are better represented by the source samples. Moreover, we develop an adaptive weighted matrix strategy, which effectively increases the contribution of important features.

4. Conclusion

In this letter, we present an AWTSL approach for cross-database SER. Specifically, a common subspace is learned in which the target features can be linearly represented by the source features. Meanwhile, an adaptive weighted matrix is learned to enhance the role of important features. In addition, an $\ell_{2,1}-$norm is used to constrain the projection matrix and reconstruction coefficients to make the model more robust. Experimental results on four popular databases show that our method can significantly outperform the state-of-the-art transfer subspace learning methods.

References

[1] M.B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," Speech Commun., vol.116, pp.56–76, Jan. 2020.
[2] P. Song, Y. Jin, L. Zhao, and M. Xin, "Speech emotion recognition using transfer learning," IEICE Trans. Inf. & Syst., vol.97, no.9, pp.2530–2532, Sept. 2014.
[3] J. Zhang, W. Li, P. Ogunbona, and D. Xu, "Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective," ACM Comput. Surv., vol.52, no.1, pp.1–38, Jan. 2019.
[4] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," Proc. IEEE, vol.109, no.1, pp.43–76, Jan. 2020.
[5] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," IEEE Trans. Affective Computing, vol.1, no.2, pp.119–131, July–Dec. 2010.
[6] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," IEEE Trans. Audio, Speech, Language Process., vol.21, no.7, pp.1458–1468, July 2013.
[7] N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, and J. Zhu, "Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning," 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp.5144–5148, IEEE, 2018.
[8] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," IEEE Trans. Affective Computing, vol.11, no.3, pp.373–382, July–Sept. 2018.
[9] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," IEEE/ACM Trans. Audio, Speech, Language Process., vol.28, pp.307–318, 2020.
[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "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.
[11] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," Advances in Neural Information Processing Systems, vol.23, pp.1813–1821, 2010.
[12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," Proc. INTERSPEECH 2010, pp.2794–2797, 2010.
[13] M. Long, J. Wang, G. Ding, J. Sun, and P.S. Yu, "Transfer feature learning with joint distribution adaptation," Proc. IEEE Int. Conf. Comput. Vis., pp.2200–2207, 2013.
[14] M. Long, J. Wang, G. Ding, J. Sun, and P.S. Yu, "Transfer joint matching for unsupervised domain adaptation," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.1410–1417, 2014.
[15] L. Zhang, W. Zuo, and D. Zhang, "LSDT: Latent sparse domain transfer learning for visual adaptation," IEEE Trans. Image Process., vol.25, no.3, pp.1177–1191, March 2016.
[16] L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, and C.L.P. Chen, "Guide subspace learning for unsupervised domain adaptation," IEEE Trans. Neural Netw. Learn. Syst., vol.31, no.9, pp.3374–3388, Sept. 2019.