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

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

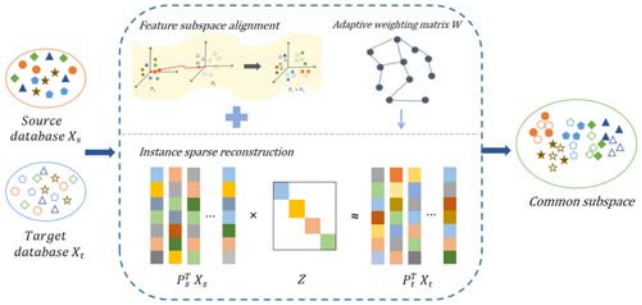
1. Yantai University; 2. Southeast University

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 co-regularized 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.

The Proposed Method

The Framework of Joint Instance Reconstruction and Feature



s

The Objective Function

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

$$\text{s.t. } Z \geq 0,\ W^T \mathbf{1} = \mathbf{1},\ W \geq 0$$

Optimization

$$w_j = \max\left(\delta_j \boldsymbol{I} - \frac{1}{\beta} v_j, 0\right),\ \delta_j = \frac{1}{d} + \frac{1}{d\beta} \sum_{i=1}^{d} v_{ij} \qquad m_{ij} = \frac{\mu h_{ij}}{\mu + 2w_{ij}}$$

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

$$C = C + \mu(P_t^T X_t - P_s^T X_s Z - E) \qquad P_t = \frac{\mu X_t Q^T}{\mu X_t X_t^T - \alpha P_s P_s^T}$$

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

Experimental Setup

Four Benchmark Datasets

- Emo-DB (Em) (5 males and 5 females)
- eNTERFACE (En) (34 males and 8 females)
- BAUM-1a (Ba) (14 males and 17 females)
- RML (Rm) (8 males)

Five Common Emotion Categories

Anger, sadness, disgust, happiness, and fear.

Acoustic Feature

We use the openSMILE toolkit to extract the 1582-dimensional feature set.

Emotional Evaluation

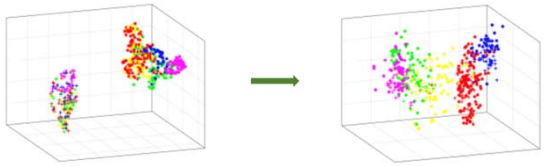
Training: all source database + random 8/10 target database.
Testing: the remainder 2/10 target database.
Classifier: linear SVM.
Evaluation metric: weighting average recall.

Results

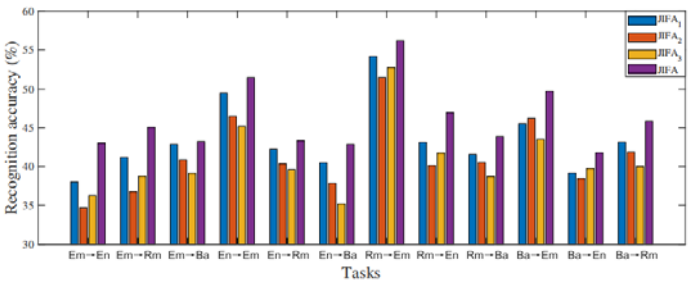
The Recognition Performance of Different Algorithms in 12 Tasks

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**

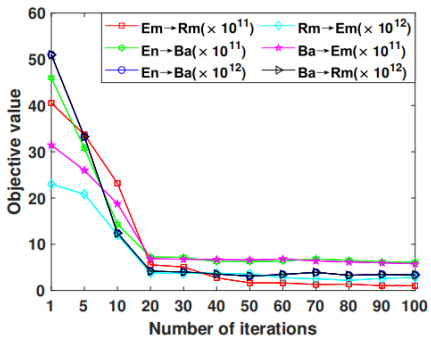
The t-SNE Visualization of Em→Ba



Results of Our Method and Three Special Tases



Convergence Curves of Our Method



Conclusion

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.