INTERSPEECH 2022
September 18 - 22 • Incheon Korea

Coupled Discriminant Subspace Alignment for Cross-database Speech Emotion Recognition

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

1. Yantai University; 2. Southeast University

Abstract

The Problem

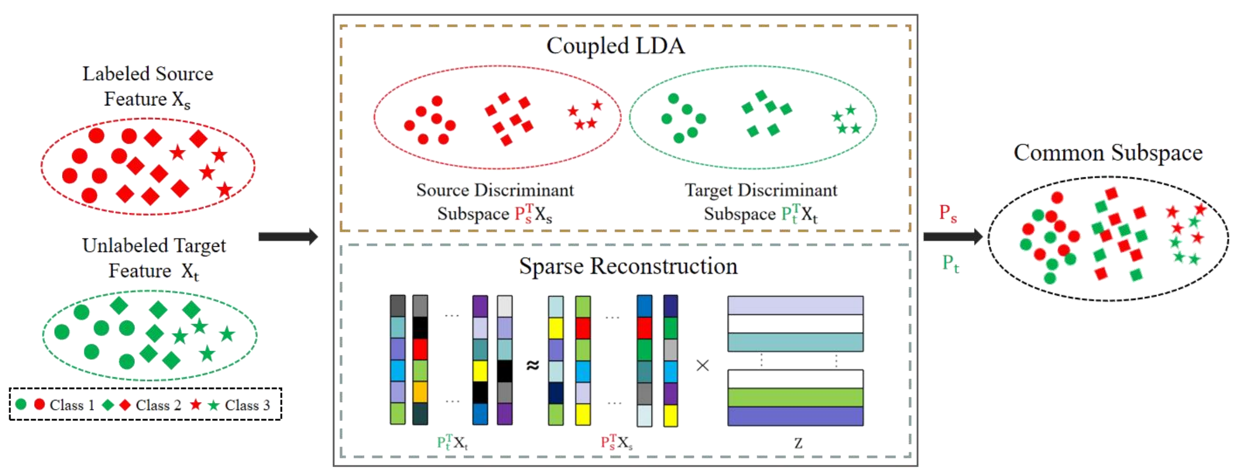
In practice, the training and test data are often collected in different scenarios, e.g., different languages, different collecting devices, which would severely degrade the recognition performance.

Our Focus

Cross-database speech emotion recognition.

The proposed method

The Framework of Coupled Discriminant Subspace Alignment



The Objective Function:

$$\min_{P_s,P_t,Z} \mathrm{Tr}(P_s^T L_s P_s) + \mathrm{Tr}(P_t^T L_t P_t) + \beta\|P_s - P_t\|_F^2$$
$$+ \alpha\|P_s^T X_s Z - P_t^T X_t\|_F^2 + \gamma\|Z\|_{2,1}$$
$$\text{s.t. } P_s^T P_s = I,\ P_t^T P_t = I$$

Optimization:

$$P_s = (L_s + \alpha X_s Z Z^T X_s^T - \beta I)^{-1}(\alpha X_s Z X_t^T P_t - \beta P_t)$$

$$P_t = (L_t + \alpha X_t X_t^T + \beta I)^{-1}(\alpha X_t Z^T X_s^T P_s + \beta P_s)$$

$$Z = (\alpha X_s^T P_s P_s^T X_s + \gamma Q)^{-1}(\alpha X_s^T P_s P_t^T X_t)$$

$$Q_{ii} = \begin{cases} 0, & \text{if } z^i = 0 \\ \frac{1}{2\|z^i\|}, & \text{otherwise} \end{cases}$$

Dataset

Four Emotional Databases

- EmoDB (E) (5 males and 5 females)
- eNTERFACE′05 (e) (34 males and 8 females)
- BAUM-1a (B) (14 males and 17 females)
- RML (R) (8 males)

Five Common Emotional Categories

Anger, sadness, disgust, happiness, and fear.

Experimental setup

Acoustic Feature

We use the openSMILE toolkit to extract the the feature set of the INTERSPEECH 2010 paralinguistic challenge (1582-dimensional).

Descriptors	Number of features
MFCC [0–14]	630
LSP frequency [0–7]	336
Log mel freq band [0–7]	336
Voicing prob	42
Loudness	42
F0 envelope	42
F0	38
Shimmer	38
Jitter	38
Jitter consecutive frame pairs	38
F0 number of onesets	1
Turn duration	1

Emotional Evaluation

Training: all source database + random 7/10 target database.

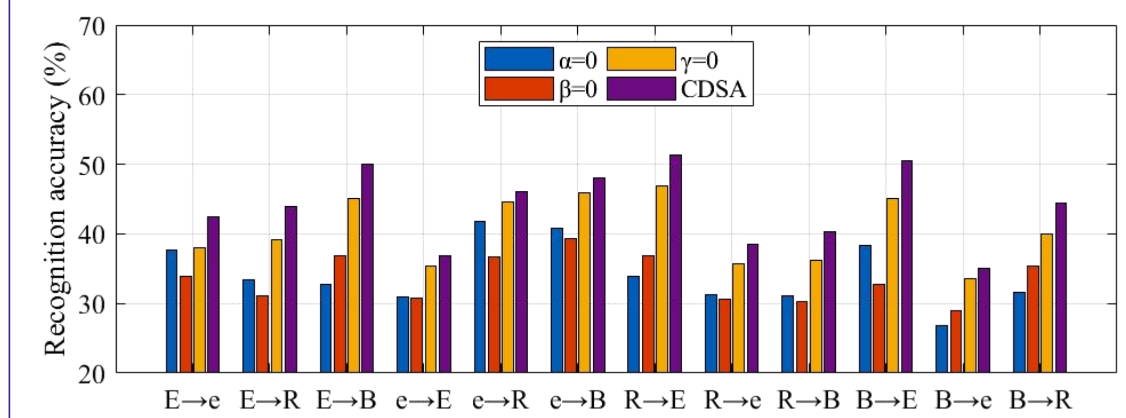
Testing: the remainder 3/10 target database.

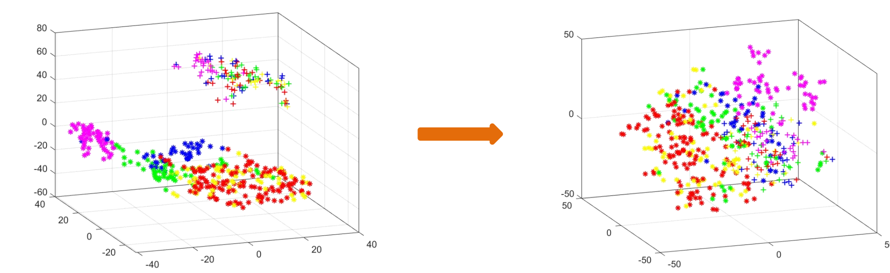
Classifier: linear SVM.

Evaluation metric: recognition accuracy.

Results

Tasks	Traditional methods		Transfer learning methods							Ours
	LDA	SDA	DRLS	TCA	JDA	DaLSR	JTSLR	JGSA	LPJT	
E→e	41.26	32.54	40.08	43.25	**42.46**	42.06	42.06	36.50	40.47	**42.46**
E→R	25.00	24.78	32.78	32.77	33.88	31.11	33.89	38.33	42.77	**43.88**
E→B	40.38	25.00	44.25	33.58	34.90	34.62	46.15	36.53	**50.00**	**50.00**
e→E	32.74	39.82	41.59	38.93	39.82	49.56	43.36	41.59	**52.21**	46.90
e→R	25.55	42.78	28.33	45.55	41.66	42.22	45.00	34.44	39.44	**46.11**
e→B	32.69	34.62	42.31	38.43	38.57	42.31	42.23	44.23	44.23	**48.07**
R→E	24.77	35.00	36.28	41.59	45.13	36.28	47.79	51.09	49.55	**51.32**
R→e	24.20	33.73	35.71	**39.68**	36.50	30.76	31.75	31.76	37.69	38.49
R→B	27.69	28.85	27.31	30.76	32.69	32.69	32.69	36.53	34.61	**40.38**
B→E	32.74	28.32	49.56	44.11	45.58	51.33	48.04	**61.94**	53.09	50.44
B→e	28.26	31.52	34.29	31.02	31.90	34.24	**36.61**	34.78	29.89	35.05
B→R	25.55	29.44	28.33	38.88	40.55	40.56	36.11	39.44	33.33	**44.44**
Average	30.06	32.19	36.73	38.21	38.63	38.97	40.47	40.59	42.27	**44.79**





Conclusion

- CSDA extends traditional LDA to a transferable manner, so that the divergence across different databases can be reduced significantly.
- Extensive experimental results show that the proposed CDSA achieves superior performance than state-of-the-art compared algorithms.