Joint Transfer Subspace Learning and Feature Selection for Cross-corpus Speech Emotion Recognition

Peng Song [1] Wenming Zheng [2] Shifeng Ou [1] Yun Jin [2] Wenming Ma [1] Yanwei Yu[1]

[1] Yantai University, Yantai, P.R. China, 264005
[2] Southeast University, Nanjing, P.R. China, 210096

ICASSP2018
The 43rd IEEE International Conference on Acoustics, Speech and Signal Processing
April 15 - 20, 2018, Calgary, Alberta, Canada

1. Overview

Goal

- To learn robust corpus invariant feature representations for cross-corpus speech emotion recognition.

Approach

- A general learning framework, called joint transfer subspace learning and feature selection (**JTSLFS**), is presented.
- To realize the coupled feature matching, we learn a latent common subspace by reducing the distribution difference and preserving the important properties of features.
- To realize the feature selection, we impose an $l_{2,1}$-norm on the projection matrix.
- A graph regularizer, which considers the geometric structure of data, is further presented to improve the recognition performance.

2. Related work

- Many statistical methods have been successfully adopted for speech emotion recognition. However, In practice, since emotional speech utterances are often collected in different environments, e.g., noises, languages, we have to face the cross-corpus speech emotion recognition problem.
- Many adaptation algorithms popular in speech/speaker recognition fields, e.g., feature normalization, maximum a posteriori (MAP), joint factor analysis (JFA), vocal tract length normalization (VTLN), have been used in speech emotion recognition. They can obtain better recognition performance than traditional algorithms. Nonetheless, these methods require a large amount of training data, which is hard to collect in practice, and do not take into account the ``corpus bias'' problem.
- Recently, one major research direction focuses on addressing the "corpus bias" problem via domain adaptation and transfer learning algorithms, However, these algorithms focus on finding the common feature representations to cope with the feature matching problem, and do not consider the importance of feature selection together.

3. Our proposed JTSLFS approach

The objective function of JTSLFS

The proposed JTSLFS algorithm aims at learning a projection matrix P to map the features of different corpora into a common low-dimensional subspace, while the $l_{2,1}$-norm is imposed on the projection matrix to perform feature selection. Moreover, a graph regularizer is further introduced to improve the recognition performance.

The objective function can given as

$$\min_{P} \left\| X^T P - Y \right\|_F^2 + \alpha \left\| P \right\|_{2,1} + \beta \Omega(P) + \gamma J(P)$$

- The 1st term: subspace learning;
- The 2nd term: $l_{2,1}$-norm;
- The 3rd term: MMD regularization;
- The 4th term: graph regularization

Optimization

The optimization problem contains the $l_{2,1}$-norm , which is non-smooth and cannot get a closed form solution. Consequently, an iterative algorithm is presented.

1). Update P as given Y_t. Setting the partial derivative of $\mathcal{O}$ with respect to P to zero, we obtain the following equation:

$$\frac{\partial \mathcal{O}}{\partial P} = 0$$
$$\Rightarrow \quad 2X(X^T P - Y) - 2RP - 2\alpha Q P = 0$$
$$\Rightarrow \quad (XX^T - R - \alpha Q)P = XY$$

And left multiplying both sides of Eq. (15) by $(XX^T - R - Q)^{-1}$, we get the analytical solution of P as

$$P^* = (XX^T - R - \alpha Q)^{-1} XY$$

2). Update Y_t as given P. When P is fixed, Eq. (14) can be reformulated as

$$\mathcal{O} = \min_{Y_t} \left\| [Y_s, Y_t] - X^T P \right\|_F^2$$

which is equivalent to the following optimization problem:

$$\mathcal{O} = \min_{Y_t} \left\| Y_t - X_t^T P \right\|_F^2$$

The above optimization problem can be easily solved by the quadratic programming algorithm [24].

4. Experiments

Data sets and compared algorithms

- The EMO-DB and eNTERFACE emotional databases are used.
- The methods that we evaluate are listed below: ***Conventional*** method, ***Baseline*** (single-corpus recognition), Transfer sparse coding (**TSC**), Transfer NMF (**TNMF**), Transfer subspace learning (**TSL**), Our proposed JTSLFS without graph regularization (**TSLFS**), Our proposed JTSLFS approach (**Ours**).

Experimental results

Table 1. The recognition performance in *test*1

Methods	Recognition rates (%)					
	Anger	Disgust	Fear	Happiness	Sadness	Average
Conventional	37.23	19.21	17.98	27.16	28.40	28.87
TSC	50.18	29.25	36.86	47.45	45.98	44.96
TNMF	50.02	29.30	36.85	47.28	46.06	43.99
TSL	47.16	26.29	32.26	46.02	45.16	40.02
TSLFS	**50.35**	**29.56**	**37.19**	**47.78**	**46.35**	**45.52**
Ours	**50.39**	**29.57**	**37.22**	**47.91**	**46.38**	**45.61**
Baseline	74.40	55.35	54.03	59.98	60.96	61.36

Table 2. The recognition performance in *test*2

Methods	Recognition rates (%)					
	Anger	Disgust	Fear	Happiness	Sadness	Average
Conventional	31.49	53.06	16.47	20.98	47.20	34.63
TSC	35.39	72.98	18.97	25.52	69.26	50.59
TNMF	36.13	73.07	19.05	25.53	69.32	51.96
TSL	37.80	72.56	18.65	25.38	69.25	50.92
TSLFS	**38.02**	**74.11**	**19.12**	**26.02**	**69.68**	**52.18**
Ours	**38.05**	**74.49**	**19.18**	**26.71**	**71.38**	**52.26**
Baseline	73.01	81.04	68.58	52.99	79.33	71.02

test1: the training dataset is EMO-DB, and the testing data is eNTERFACE,
test2: the training dataset is eNTERFACE, and the testing data is EMO-DB.

5. Key references

- 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
- 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.
- Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech communication, vol. 52, no. 1, pp. 12–40, 2010.
- 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.
- Deng Cai, Xiaofei He, and Jiawei Han, "Spectral regression for efficient regularized subspace learning," in Proc. ICCV, 2007, pp. 1–8.
- Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," in Proc. NIPS, 2010, pp. 1813–1821.
- Bjorn Schuller, Stefan Steidl, Anton Batliner, et al., "The INTERSPEECH 2010 paralinguistic challenge," in Proc. INTERSPEECH, 2010, pp. 2795–2798.