IEICE TRANS. INF. & SYST., VOL.E97–D, NO.9 SEPTEMBER 2014

LETTER

Speech Emotion Recognition Using Transfer Learning

Peng SONG[†a)], **Yun JIN**[††], **Li ZHAO**[†], ***Nonmembers***, ***and*** **Minghai XIN**[†††], ***Member***

SUMMARY A major challenge for speech emotion recognition is that when the training and deployment conditions do not use the same speech corpus, the recognition rates will obviously drop. Transfer learning, which has successfully addressed the cross-domain classification or recognition problem, is presented for cross-corpus speech emotion recognition. First, by using the maximum mean discrepancy embedding (MMDE) optimization and dimension reduction algorithms, two close low-dimensional feature spaces are obtained for source and target speech corpora, respectively. Then, a classifier function is trained using the learned low-dimensional features in the labeled source corpus, and directly applied to the unlabeled target corpus for emotion label recognition. Experimental results demonstrate that the transfer learning method can significantly outperform the traditional automatic recognition technique for cross-corpus speech emotion recognition.

key words: *speech emotion recognition, transfer learning, cross-corpus, dimension reduction*

1. Introduction

Speech emotion recognition is an important research topic in the areas of speech signal processing. Unlike speech or speaker recognition, which considers the emotions as irrelevant noises, emotion recognition aims at recognizing the emotions of speech regardless of speaker identity or verbal content. It plays an important role in many applications, and has been successfully deployed in call center applications and mobile communications, etc.

Over the past few decades, many kinds of approaches have been presented for the task of speech emotion recognition, such as hidden Markov model (HMM), Gaussian mixture model (GMM), artificial neural network (ANN), support vector machine (SVM) and some combination methods [1]. Among these methods, the SVM is the most popular approach, and it is chosen as the classifier in this letter.

The above mentioned methods are proposed mainly for the single corpus based emotion classification, without considering the linguistic differences between speakers, and they can obtain satisfactory results to some extent. However, when training and testing corpora are different, the recognition rates will decrease significantly [2]. Meanwhile, transfer learning has successfully solved many cross-domain pattern classification and recognition problems [3]. This is our motivation for introducing the transfer learning for cross-corpus speech emotion recognition.

This letter is organized as follows. In Sect. 2, the transfer learning based speech emotion recognition method is proposed. In Sect. 3, the experimental results and discussions are presented. Finally, the conclusions are drawn in Sect. 4.

2. Transfer Learning Based Speech Emotion Recognition

A novel transfer learning based speech emotion recognition scheme is presented, and an efficient transfer learning approach via dimension reduction [4] is adopted. The interest in using transfer learning lies in that it can efficiently address the cross-domain problem. The key idea is that, although the distributions of features between labeled source and unlabeled target corpora are different, there exists some common or close latent feature space among them, and this latent feature space is utilized to learn the classifier for the unlabeled target corpus.

Figure 1 illustrates the flowchart of the proposed approach. First, the high-dimensional feature sets are extracted from the labeled source and unlabeled target corpora, respectively. Then, two close latent low-dimensional feature spaces are learned by the optimization and dimension reduction algorithms, and a SVM classifier is learned by using the low-dimensional features and labels of the source corpus. Finally, the classifier is directly applied to the target corpus to predict the unknown labels.

Let $X_s = \{x_{s_1}, x_{s_2}, \ldots, x_{s_M}\}$ and $X_t = \{x_{t_1}, x_{t_2}, \ldots, x_{t_N}\}$

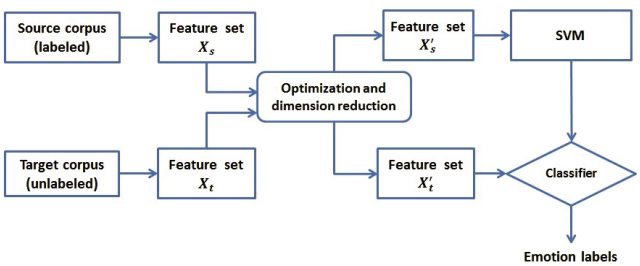

Fig. 1 Flowchart of the proposed method.

Manuscript received March 2, 2014.
Manuscript revised April 28, 2014.

[†]The authors are with the Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, School of Information Science and Engineering, Southeast University, Nanjing 210096, P.R. China..

[††]The author is with the School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221116, P.R. China.

[†††]The author is with the Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing 210096, P.R. China.

a) E-mail: pengsongseu@gmail.com

DOI: 10.1587/transinf.2014EDL8038

be the feature sets of labeled source and unlabeled target speech corpora, respectively, and $Y_s = \{y_{s_1}, y_{s_2}, \ldots, y_{s_M}\}$ and $Y_t = \{y_{t_1}, y_{t_2}, \ldots, y_{t_N}\}$ be the corresponding labels, the close latent low-dimensional feature spaces will be learned for both corpora. Let ψ be the projection mapping function to the lower-dimensional feature space, and the projected data of source and target corpora, X'_s and X'_t, will be obtained as $X'_s = \psi(X_s)$ and $X'_t = \psi(X_t)$, respectively. Assume ϕ be the mapping function to the reproducing kernel Hilbert space (RKHS), the maximum mean discrepancy (MMD) algorithm is employed to describe the distance of the projected data, and the empirical estimate of MMD $D(X'_s, X'_t)$ is written as follows

$$D(X'_s, X'_t) = \left\| \frac{1}{M} \sum_{m=1}^{M} \phi(x'_{s_m}) - \frac{1}{N} \sum_{n=1}^{N} \phi(x'_{t_n}) \right\|_H \tag{1}$$

Where H represents a universal RKHS. By employing the maximum mean discrepancy embedding (MMDE) algorithm [4], the $D(X'_s, X'_t)$ can be further written as

$$D(X'_s, X'_t) = tr(KL) \tag{2}$$

Where tr is the trace, $K = \begin{bmatrix} K_{ss} & K_{st} \\ K_{ts} & K_{tt} \end{bmatrix}$ is a composite kernel, and $L = \{l_{ij}\}$ is given as

$$l_{ij} = \begin{cases} \frac{1}{M^2} & \text{if} \quad x_i, x_j \in X_s \\ \frac{1}{N^2} & \text{if} \quad x_i, x_j \in X_t \\ \frac{-1}{MN} & \text{otherwise} \end{cases} \tag{3}$$

The embedding problem will be solved as the following formulation:

$$\begin{array}{ll} \min\limits_{K=\widetilde{K}+\varepsilon I} & tr(KL) - tr(K) \\ s.t. & K_{ii} + K_{jj} - 2K_{ij} = d_{ij}^2 \quad (x_i, x_j) \in \mathbf{N} \\ & K\mathbf{I} = \mathbf{0} \qquad\qquad\quad \widetilde{K} \geq 0, \varepsilon > 0 \end{array} \tag{4}$$

Where $(x_i, x_j) \in \mathbf{N}$ means that x_i and x_j are the nearest neighbors, I is the identity matrix, and $\mathbf{I}$ and $\mathbf{0}$ are the vectors of ones and zeros, respectively. It can be solved by the standard semidefinite problem (SDP) [5]. After the kernel matrix K is obtained, the dimension reduction algorithms are applied to obtain the low-dimensional features X'_s and X'_t, respectively.

In the training corpus, given the low-dimensional features x'_s and corresponding labels y_s, by employing the traditional SVM algorithm, a classification function $f(x'_s)$ will be learned. Then, the function is directly applied to the target corpus to predict the labels $y_t = f(x'_t)$. Finally, $(x_t, f(x'_t))$ will be used to predict the unknown labels for the new data enrolling in the target corpus.

3. Experiments

3.1 Experimental Setup

The Berlin dataset (EMO-DB) [6] is used as the source training corpus, and 5 basic emotions (anger, fear, happiness, neutral and sadness) totally with 377 utterances are chosen for training. Another Chinese emotional dataset with the 5 emotions is collected as the target testing corpus, 6 Chinese speakers (3 male and 3 female) are employed to utter each sentence with stimulated emotion states, totally resulting in 500 utterances. To avoid the over-fitting problem, 5 independent tests are conducted repeatedly. In each test, 70% of the source corpus is randomly used for training, while 70% of the target corpus is randomly chosen for testing.

To perform the speech emotion recognition, the following 26 kinds of features are extracted by the openSMILE Toolkit [7], i.e., 12 Mel cepstral coefficients (MFCCs), 8 line spectral pairs (LSFs), intense, loudness, zero-cross rate (ZCR), probability of voicing, F0, and F0 envelopes. Finally, in total 988 features, including 19 statistical functions of these features and their first order delta coefficients, are used for the experiments.

Three kinds of methods are compared, one is the automatic recognition method (*Automatic*), in which the SVM function learned from the source corpus is directly applied to the unlabeled target data, another is the baseline method (*Baseline*), in which a SVM based recognition is conducted on the labeled target data, and the other is the proposed transfer learning method, which is conducted on the labeled source and unlabeled target corpora. Two kinds of dimension reduction methods, principal component analysis (PCA) (*Proposed1*) and local preserving projections (LPP) [8] (*Proposed2*) are employed for the proposed method, respectively. The Gaussian kernel is selected, and the dimensions of the low-dimensional acoustic features are optimized as 80.

3.2 Results and Analysis

Table 1 shows the overall recognition rates (i.e., the average recognition rates for the 5 emotions in the experiments). It can be easily found that, for each independent test, our proposed method can effectively and significantly improve the recognition rates compared to the *Automatic* method. It can be also seen that the LPP dimension reduction method can obtain slightly higher recognition rates than the PCA method to some extent. Taking the overall recognition rates into account, it can be stated that our proposed scheme is efficient for the cross-corpus speech emotion recognition.

Figure 2 summarizes the average confusion matrix of the recognition results using the *Proposed2* method. As can be seen from the figure, the neutral achieves the best recognition rate with about 0.65. Meanwhile, the fear and happi-

Table 1 Comparison of overall recognition rates of different methods in 5 independent tests (%).

Methods	No. of tests				
	1	2	3	4	5
Baseline	85.6	86.1	85.9	85.8	85.5
Automatic	32.2	30.5	35.1	32.2	29.8
Proposed1	57.8	58.5	56.3	57.9	58.3
Proposed2	**58.6**	**59.3**	**57.8**	**58.9**	**59.8**

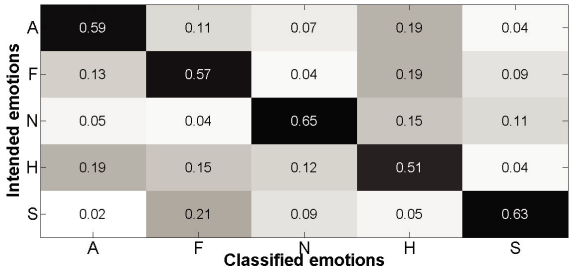

Fig. 2 The confusion matrix of the transfer learning based speech emotion recognition results (Anger: A, Fear: F, Neutral: N, Happiness: H, Sadness: S).

ness obtain the lowest recognition rates, and are more easily confused with other emotions. This result is consistent with the experimental results conducted on single corpus [9]. It can be also observed that the fear and sadness obtain the highest confused value with about 0.21, this might be attributed to the fact that the valence levels of these two emotions are similar [10].

4. Conclusion

In this letter, we have presented a novel speech emotion recognition approach based on transfer learning. For the labeled source and unlabeled target corpora, two close low-dimensional feature spaces are learned by using the MMDE optimization and dimension reduction algorithms, and a SVM classier function is trained by using the learned low-dimensional features and labels in the source corpus, and then the classier function is applied to predict the unknown emotion labels in the target corpus. The experimental results demonstrate that compared to the automatic recognition approach, the proposed method can significantly improve the recognition rates.

Acknowledgement

The work has been supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61375028 and 61273266. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the manuscript.

References

[1] E.A. Moataz, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol.44, no.3, pp.572–587, 2011.
[2] O. Toledo-Ronen and A. Sorin, "Voice-based sadness and anger recognition with cross-corpus evaluation," Proc. ICASSP, pp.7517–7521, Vancouver, Canada, May 2013.
[3] S.J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowl. Data Eng., vol.22, no.10, pp.1345–1359, 2010.
[4] S.J. Pan, J.T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," Proc. AAAI, pp.677–682, Chicago, U.S.A., July 2008.
[5] G.R. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan, "Learning the kernel matrix with semidefinite programming," J. Machine Learning Research, vol.5, pp.27–72, 2004.
[6] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss, "A database of German emotional speech," Proc. Interspeech, pp.1517–1520, Lisbon, Portugal, Sept. 2005.
[7] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," Proc. ACM Multimedia, pp.1459–1462, Firenze, Italy, Oct. 2010.
[8] X. He and P. Niyogi, "Locality preserving projections," Proc. NIPS, pp.153–160, Vancouver, Canada, Dec. 2003.
[9] Y. Jin, P. Song, W. Zheng, L. Zhao, and M. Xin, "Speaker-independent speech emotion recognition based on two-layer multiple kernel learning," IEICE Trans. Inf. & Syst., vol.E96-D, no.10, pp.2286–2289, Oct. 2013.
[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Process. Mag., vol.18, no.1, pp.32–80, 2001.