IEICE TRANS. INF. & SYST., VOL.E100–D, NO.5 MAY 2017

LETTER

Learning Corpus-Invariant Discriminant Feature Representations for Speech Emotion Recognition

Peng SONG[†a)], ***Member***, **Shifeng OU**[††], **Zhenbin DU**[†], **Yanyan GUO**[†], **Wenming MA**[†], **Jinglei LIU**[†], ***and*** **Wenming ZHENG**[†††b)], ***Nonmembers***

SUMMARY As a hot topic of speech signal processing, speech emotion recognition methods have been developed rapidly in recent years. Some satisfactory results have been achieved. However, it should be noted that most of these methods are trained and evaluated on the same corpus. In reality, the training data and testing data are often collected from different corpora, and the feature distributions of different datasets often follow different distributions. These discrepancies will greatly affect the recognition performance. To tackle this problem, a novel corpus-invariant discriminant feature representation algorithm, called transfer discriminant analysis (TDA), is presented for speech emotion recognition. The basic idea of TDA is to integrate the kernel LDA algorithm and the similarity measurement of distributions into one objective function. Experimental results under the cross-corpus conditions show that our proposed method can significantly improve the recognition rates.

key words: *speech emotion recognition, transfer learning, dimensionality reduction*

1. Introduction

Speech emotion recognition refers to recognizing emotions from speaker's voice, and the goal is to classify the emotions into following categories, e.g., anger, sadness, happiness, fear and disgust. It has various potential applications, e.g., helping diagnose patients' mental diseases in medical field, computer tutoring services, human computer interaction (HCI) based entertainment, and call centers [1], [2].

Over the past decades, various features have been investigated and applied for speech emotion recognition. Among these studies, the global statistics over the low level descriptors (LLDs), e.g., F0s, durations, intensities, Mel frequency cepstrals (MFCCs), achieve dominant superiority [1]. Meanwhile, all kinds of efforts have also been made for developing emotion classification methods, and many classification algorithms popular in pattern recognition and machine learning fields are employed, for example, hidden Markov model (HMM), support vector machine (SVM), Gaussian mixture model (GMM), artificial neural network (ANN), multilayer perception (MLP), decision trees, extreme learning machine (ELM), deep neural network (DNN) and some combination of these methods [1], [3]. These approaches can obtain satisfactory results in most cases. However, it should be noted that all these approaches are conducted on the assumption that the training and testing settings are the same. As aforementioned, in practice, the training data and testing data are often collected in different conditions, and this mismatch will significantly reduce the recognition performance.

To alleviate this kind of mismatch problem, a considerable amount of work has been done in speech community. Sanchez et al. [4] view the domain mismatch as a nuisance, and propose a domain adaptation and compensation method to improve the emotion detection results. By combination of multiple training corpora and classifiers, Schuller et al. [5] investigate the voting strategies for cross-corpus speech emotion recognition. Chen et al. [6] propose a linear regression based adaptation method for music emotion recognition. To reduce the discrepancy between training and testing data, Deng et al. [7] present an unsupervised domain adaptation based adaptive denoising autoencoder approach. In [8], Abdelwahab et al. introduce a supervised domain adaptation method to improve the recognition performance. These methods can achieve satisfactory results to some extents. However, there still exist many shortcomings. On one hand, these approaches need a large amount of emotional speech data, which is expensive to be collected in practice. On the other hand, they do not consider the differences between the feature distributions of training and testing datasets. In real variational conditions, this discrepancy is often very large. In [9], [10], we have proposed two kinds of transfer learning algorithms for cross-corpus speech emotion recognition, However, these approaches are unsupervised learning algorithms, which cannot efficiently utilize the label information of source corpus, and will affect the recognition performance.

Inspired by the successful applications of LDA and transfer learning techniques [9]–[11], in this letter, a new transfer discriminant analysis (TDA) method is presented to address the discrepancy between two datasets. In this work, the kernel LDA and maximum mean discrepancy (MMD) [12] algorithms are optimized together, in which the kernel LDA algorithm is used for feature dimensionality re-

Manuscript received November 15, 2016.
Manuscript revised January 11, 2017.
Manuscript publicized February 2, 2017.

[†]The authors are with the School of Computer and Control Engineering, Yantai University, Yantai 264005, P.R. China.
[††]The author is with the School of Science and Technology for Opto-electronic Information, Yantai University, Yantai 264005, P.R. China.
[†††]The author is with the Key Laboratory of Child Development and Learning Science, Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing, Jiangsu 210096, P.R. China.
a) E-mail: pengsong@ytu.edu.cn
b) E-mail: wenming_zheng@seu.edu.cn
DOI: 10.1587/transinf.2016EDL8222

duction, while the MMD algorithm is chosen for similarity measurement.

The rest of this letter is organized as follows. In Sect. 2, the transfer discriminant analysis method is presented. Experimental results are demonstrated in Sect. 3. Finally, we draw our conclusions in Sect. 4.

2. Proposed Methodology

Feature dimensionality reduction is an important part for many pattern recognition problems. Many supervised and unsupervised methods, e.g., principal component analysis (PCA), linear discriminant analysis (LDA), linear preserving projection (LPP), locally linear embedding (LLE), have been proposed for dimensionality reduction [13]. All these methods can achieve satisfactory performance to some degree, However, they assume that the training and testing processes are conducted in the same situations. In practice, the training data and testing data are often from different corpora and follow different distributions, in which the performance of traditional dimensionality reduction algorithms will drop significantly [10]. In this letter, to address this problem, a TDA approach is presented, in which the similarity between the feature distributions of source and target datasets are considered when kernel LDA algorithm is carried out.

Let $X = [X_s, X_t] \in R^{m\times n}$ be the feature sequences, where $n = n_l + n_u$, $X_s = [x_1, \ldots, x_{n_l}]$ and $X_t = [x_{n_l+1}, \ldots, x_{n_l+n_u}]$ are the feature sets of labeled source and unlabeled target corpora, respectively, the classic LDA algorithm is used for dimensionality reduction. It aims at finding directions on which the features of the same classes are close to each other while the features from different classes are far from each other [11]. The objective function of LDA is written as follows

$$\arg\max_{U} \frac{Tr(U^T S_b U)}{Tr(U^T S_t U)} \tag{1}$$

where $U \in R^{m\times c}$ is the orthogonal projection matrix, $Tr(\cdot)$ is the trace of a matrix, S_t and S_b are the total scatter matrix and between-class scatter matrix, respectively. Assuming the mean values of data are centering, S_t and S_b are given as

$$S_t = XX^T \tag{2}$$

$$S_b = \sum_{k=1}^{c} X^{(k)} W^{(k)} (X^{(k)})^T = XWX^T \tag{3}$$

where c is the number of emotion categories, $W^{(k)}$ is a $l_k \times l_k$ matrix with all elements equal to $1/l_k$, and $X^{(k)}$ denotes the data matrix of the $k-$th class.

To better perform dimensionality reduction, a nonlinear generalization algorithm called kernel LDA is further employed. Let $\phi(X)$ be a non-linear transformation, the Eq. (1) becomes as

$$\arg\max_{V} \frac{Tr(V^T KWK^T V)}{Tr(V^T KK^T V)} \tag{4}$$

where $K = \phi(X)^T \phi(X)$ is a kernel matrix, and $V \in R^{n\times c}$ is the corresponding transformation matrix. After dimensionality reduction, the new optimal feature is represented as $V^T K$.

In practice, X_s and X_t often follow different distributions, which will cause severe degradation of recognition performance. In this letter, the similarities between two distributions are considered, and the MMD [12], which is a nonparametric estimation criterion on reproducing kernel Hilbert space (RKHS), is employed for measurement, and is given by

$$\begin{aligned} Dist(X_s, X_t) &= \left\| \frac{1}{n_l} \sum_{i=1}^{n_l} V^T k_i - \frac{1}{n_u} \sum_{j=n_l+1}^{n_l+n_u} V^T k_j \right\|_{\mathcal{H}}^2 \\ &= Tr(V^T KMK^T V) \end{aligned} \tag{5}$$

where $\mathcal{H}$ is a universal RKHS, and $M = [m_{ij}] \succeq 0$ with

$$m_{ij} = \begin{cases} \frac{1}{n_l^2} & \text{if } x_i, x_j \in X_s \\ \frac{1}{n_u^2} & \text{if } x_i, x_j \in X_t \\ \frac{-1}{n_l n_u} & \text{otherwise} \end{cases} \tag{6}$$

For cross-corpus dimensionality reduction, we aim to reduce the similarity distance $Dist(\cdot)$ while efficiently making dimensionality reduction. By incorporating Eq. (4) into Eq. (5), we will obtain the TDA optimization problem as

$$\arg\min_{V} \frac{Tr(V^T KK^T V)}{Tr(V^T KWK^T V)} + \lambda Tr(V^T KMK^T V) \tag{7}$$

The above objective function is a non-linear optimization problem and is hard to find a global optimum solution. According to [14], the trace ratio problem can be relaxed as a modified form:

$$\begin{aligned} \arg\min_{V} J(V) &= Tr(V^T KK^T V) - Tr(V^T KWK^T V) \\ &\quad + \lambda Tr(V^T KMK^T V) \\ \text{s.t. } V^T V &= 1 \end{aligned} \tag{8}$$

According to the constrained optimization theory, the Lagrange function is employed to solve this problem:

$$\begin{aligned} L &= Tr(V^T KK^T V) - Tr(V^T KWK^T V) \\ &\quad + \lambda Tr(V^T KMK^T V) + \alpha(I - V^T V) \end{aligned} \tag{9}$$

where $\alpha \in R^{c\times c}$ is a Lagrange multiplier matrix. Setting $\frac{\partial L}{\partial V} = 0$, the following equation will be achieved as

$$K(I - W + \lambda M)K^T V = \alpha V \tag{10}$$

The above equation is equivalent to a generalized eigen-decomposition problem to compute V.

3. Experiments

In this section, several experiments are performed to test our

algorithm, in which the source dataset is labeled and the target dataset is unlabeled.

3.1 Experimental Setup

Two popular corpora are used for our experiments, they are the Berlin dataset [15] and the eNTERFACE dataset [16]. The Berlin dataset is one of the most popular corpora, it consists of seven kinds of emotions, i.e., neutral, anger, disgust, boredom, fear, happiness and sadness. Total 494 emotional speech utterances are recorded by 10 actors in German. The eNTERFACE is a public English audio-visual emotional database. It includes six types of emotions, i.e., anger, disgust, fear, happiness, sadness and surprise. 1170 video samples are collected by 42 subjects from 14 countries.

In our experiments, two types of scenarios are used for evaluation, called *case1* and *case2*. In *case1*, the labeled eNTERFACE dataset is used for training, while the unlabeled Berlin dataset is chosen for testing. Meanwhile, in *case2*, the labeled Berlin dataset is used for training, while the unlabeled eNTERFACE dataset is chosen for testing. Each corpus is divided into 5 parts, and in each test, random $4/5$ of the source and target datasets are used for training, while the others are used for evaluation. The openSMILE toolkit is adopted to extract the emotional features [17], and the feature set in Interspeech 2010 emotion challenge is used in our experiments, and total 1582 dimensional statistical features are chosen for our tests [18]. The SVM algorithm is chosen for feature classification, five common types of emotions including anger, disgust, fear, happiness and sadness are used, and the Gaussian kernel is chosen for K in our tests.

3.2 Results and Analysis

To evaluate the recognition performance of our proposed method, the following approaches are compared, they are the automatic recognition method (*Auto*), in which the classifier trained in source corpus is directly for emotion classification in target corpus, the baseline method (*Baseline*), in which, the training and testing processes are conducted in single corpus, the unsupervised dimensionality reduction based transfer learning method (DR) [9], in which the dimensionality reduction and MMD algorithms are performed separately, the transfer non-negative matrix factorization (TNMF) algorithm [10], and our proposed TDA method (*Ours*).

The experimental results are summarized in Table 1 and Table 2. First, it can be found that, in both cases, the DR, TNMF and our proposed TDA methods significantly outperform the automatic recognition method. This can be attributed to the power of transfer learning techniques, which can reduce the distance between the feature distributions of two datasets. Second, it can be seen that the TDA approach can obtain better recognition results in all emotions, the reasons may be that, one one hand, compared to DR algorithm, our TDA method integrates the LDA algorithm with transfer learning function, and optimizes them together. On the other hand, the TDA algorithm is a supervised transfer learning algorithm, while DR and TNMF are unsupervised algorithms. Third, it can be also observed that the recognition rates of *case1* are lower than those of *case2*, the reason is that the eNTERFACE dataset is more complex than the Berlin dataset. Last, the experimental results are still far from satisfactory, and our proposed TDA performs worse than the baseline algorithm in all cases. This phenomenon can be explained by the "negative transfer" from the source corpus, which may hurt the recognition in the target corpus [19].

Table 1 The recognition rates in *case1* (anger: A, disgust: D, fear: F, happiness: H, sadness: S).

Methods	Recognition rates (%)					
	A	D	F	H	S	Average
Baseline	74.52	55.31	53.88	60.01	61.11	61.97
Auto	37.69	19.17	17.99	27.28	28.34	23.01
DR	47.13	24.87	29.12	43.25	40.98	35.94
TNMF	52.58	29.49	37.61	47.01	44.69	44.01
Ours	53.64	30.68	38.25	47.92	45.10	45.36

Table 2 The recognition rates in *case2* (anger: A, disgust: D, fear: F, happiness: H, sadness: S).

Methods	Recognition rates (%)					
	A	D	F	H	S	Average
Baseline	73.10	81.65	68.69	52.71	79.19	74.98
Auto	31.19	52.71	17.15	19.87	47.49	34.21
DR	33.18	69.14	18.06	23.98	68.13	41.02
TNMF	36.12	74.52	19.21	26.57	71.56	51.98
Ours	36.73	75.08	20.67	27.81	71.99	52.83

4. Conclusion

In this letter, we propose a new dimensionality reduction method called transfer discriminant analysis for cross-corpus speech recognition. It makes use of both dimensionality reduction and transfer learning algorithms. The kernel LDA approach is used for dimensionality reduction, while the MMD algorithm is chosen to reduce the similarities between two feature distributions. Experimental results on two public datasets demonstrate the effectiveness of our proposed algorithm.

Acknowledgements

This work was supported in part by the Natural Science Foundation of Shandong Province (ZR2014FQ016), the National Basic Research Program of China (2015CB351704), the National Natural Science Foundation of China (61572419, 61602399), A Project of Shandong Province Higher Educational Science and Technology Program (J15LN09), and the Fundamental Research Funds for the Southeast University (CDLS-2017-02).

References

[1] M.E. Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion

recognition: features, classification schemes, and databases," Pattern Recognition, vol.44, no.3, pp.572–587, 2011.
[2] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," Proc. Speech Prosody, pp.276–289, Dresden, 2006.
[3] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," Proc. INTERSPEECH, pp.223–227, Singapore, 2014.
[4] M.H. Sanchez, G. Tür, L. Ferrer, and D. HakkaniTür, "Domain adaptation and compensation for emotion detection," Proc. INTERSPEECH, pp.2874–2877, Makuhari, Japan, 2010.
[5] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?," Proc. INTERSPEECH, pp.1553–1556, Florence, Italy, 2011.
[6] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. Chen, "Linear regression-based adaptation of music emotion recognition models for personalization," Proc. ICASSP, Florence, Italy, pp.2149–2153, 2014.
[7] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," IEEE Signal Process. Lett., vol.21, no.9, pp.1068–1072, 2014.
[8] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," Proc. ICASSP, Brisbane, Australia, pp.5058–5062, 2015.
[9] P. Song, Y. Jin, L. Zhao, and M. Xin, "Speech emotion recognition using transfer learning," IEICE Trans. Inf. & Syst., 2014, vol.E97-D, no.9, pp.2530–2532, 2014.
[10] P. Song, S. Ou, W. Zheng, Y. Jin, and L. Zhao, "Speech emotion recognition using transfer non-negative matrix factorization," Proc. ICASSP, Shanghai, China, pp.5180–5184, March 2016.
[11] A.R. Webb, Statistical pattern recognition, John Wiley & Sons, 2003.
[12] K.M. Borgwardt, A. Gretton, M.J. Rasch, H.-P. Kriegel, B. Schölkopf, and A.J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," Bioinformatics, vol.22, no.14, pp.e49–e57, 2006.
[13] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.27, no.3, pp.328–340, 2005.
[14] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," IEEE Trans. Neural Netw., vol.20, no.4, pp.729–735, 2009.
[15] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," Proc. International Conference on Data Engineering Workshops, Atlanta, USA, p.8, April 2006.
[16] 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.
[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," Proc. ACM Multimedia, Firenze, Italy, pp.1459–1462, Oct. 2010.
[18] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C.A. Muller, and S.S. Narayanan, "The interspeech 2010 paralinguistic challenge," Proc. Interspeech, pp.2794–2797, Makuhari, Japan, 2010.
[19] K. Weiss, T.M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," Journal of Big Data, vol.3, no.1, pp.1–40, 2016.