SPEECH EMOTION RECOGNITION USING TRANSFER NON-NEGATIVE MATRIX FACTORIZATION

Peng Song [1,2] *Shifeng Ou* [1] *Wenming Zheng* [2*] *Yun Jin* [2] *Li Zhao*[2]

[1]School of Computer and Control Engineering, Yantai University, Yantai 264005, China
[2]Key Laboratory of Child Development and Learning Science of Ministry of Education,
Southeast University, Nanjing 210096, China
{pengsong, wenming_zheng}@seu.edu.cn

ABSTRACT

In practical situations, the emotional speech utterances are often collected from different devices and conditions, which will obviously affect the recognition performance. To address this issue, in this paper, a novel transfer non-negative matrix factorization (TNMF) method is presented for cross-corpus speech emotion recognition. First, the NMF algorithm is adopted to learn a latent common feature space for the source and target datasets. Then, the discrepancies between the feature distributions of different corpora are considered, and the maximum mean discrepancy (MMD) algorithm is used for the similarity measurement. Finally, the TNMF approach, which integrates the NMF and MMD algorithms, is proposed. Experiments are carried out on two popular datasets, and the results verify that the TNMF method can significantly outperform the automatic and competitive methods for cross-corpus speech emotion recognition.

Index Terms— Speech emotion recognition, transfer learning, non-negative matrix factorization, maximum mean discrepancy

1. INTRODUCTION

The objective of speech emotion recognition is to classify the speech utterance into the following emotion categories, e.g., anger, fear, happiness, neutral and sadness. It can be applied to various tasks, such as diagnosing patients' mental illness, monitoring the drivers' emotion variations to avoid accidents and helping the man-machine interactions [1].

As an important branch of affective computing and speech signal processing, many studies have been made for speech emotion recognition. Various classic pattern recognition methods have been adopted to implement the classification function, such as Gaussian mixture model (GMM), hidden Markov model (HMM), support vector machine (SVM), artificial neural network (ANN), deep neural network (DNN) and extreme learning machine (ELM) algorithms [1, 2, 3]. All these methods can obtain satisfactory results to some extent, however, they are conducted and evaluated on single corpus, which requires the training data and testing data are from the same corpus.

In practice, the training and testing speech utterances are often recorded in different situations. As a result the recognition rates will obviously drop. To address this problem, recently some researches have been done for cross-corpus speech emotion recognition. Schuller et al. make a preliminary experiment on multiple datasets, in which 5 datasets are used for model training, while the remaining one is adopted for testing [4]. Deng et al. present an autoencoder based unsupervised domain adaptation method for cross-corpus emotion recognition [5]. In [6], we propose a transfer learning approach to learn better low dimensional feature representations for labeled source and unlabeled target data, which achieves better classification performance by reducing the similarity divergence between different distributions.

Inspired by the recent progress of transfer learning [7] and non-negative matrix factorization (NMF) [8], in our paper, a new transfer NMF (TNMF) method is presented for cross-corpus speech emotion recognition. Specifically, the NMF algorithm is used to learn the robust representations for labeled source and unlabeled target data. Meanwhile, the distance between the two different feature distributions is considered, and the maximum mean discrepancy (MMD) approach [9] is employed for the similarity measurement. To enrich the feature discrimination, a TNMF approach which integrates NMF and MMD algorithms is further presented. Experimental results show the effectiveness of our proposed method.

The rest of this paper is organized as follows. The NMF method is briefly reviewed in Section 2. In Section 3, first, the MMD algorithm is described in detail, respectively, then our proposed TNMF approach is presented. The experimental results and discussions are given in Section 4. Finally, our paper is concluded in Section 5.

[*] Corresponding author.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

2. NON-NEGATIVE MATRIX FACTORIZATION

NMF is an efficient algorithm that can obtain a low dimensional representation of the non-negative data [8]. It aims at finding two non-negative matrices to well approximate the original matrix data, and has been successfully applied to various fields, e.g., face recognition, document clustering and some other pattern recognition fields.

Given a non-negative feature matrix $X = [x_1, x_2, \dots, x_N] \in R^{M\times N}$, the goal of NMF is to seek a basis of latent low-rank feature space $U = [u_{ik}] \in R^{M\times K}$ and the corresponding coding matrix $V = [v_{kj}] \in R^{K\times N}$. The feature samples can be modelled as solving the following problem:

$$\min_{U,V} \|X - UV\|_F^2, \quad \text{s.t. } U,\ V \geq 0 \tag{1}$$

where $\|\cdot\|_F$ is a Frobenius norm and $K \ll \min\{M, N\}$. The above function is a non-convex problem when computing U and V together, so an iterative algorithm [10] has been presented, and the update functions of U and V are given as follows

$$u_{ik} \leftarrow u_{ik}\frac{(XV)_{ik}}{(UV^TV)_{ik}}, v_{kj} \leftarrow v_{kj}\frac{(X^TU)_{kj}}{(VU^TU)_{kj}} \tag{2}$$

In reality, many prior studies have demonstrated that the naturally occurring data may usually reside on or close to a low dimensional submanifold embedded in a high dimensional space [11], and the intrinsic geometrical information is important to the discrimination of the data. So the graph NMF method is further adopted [12], in which a graph encoding the geometrical structure is used as a regularization term of the NMF function. Given a graph with M vertices, and each vertex corresponds to a feature vector. For each vector x_i, the p nearest neighbours can be easily found. A simple 0-1 weight matrix $W = [w_{ij}] \in R^{N\times N}$ is adopted, which is given as

$$w_{ij} = \begin{cases} 1 & \text{if } x_j \in N_p(x_i) \text{ or } x_i \in N_p(x_j) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $N_p(x_i)$ and $N_p(x_j)$ denote the p nearest neighbours of x_i and x_j, respectively. Given a diagonal matrix $D = [d_{jj}] \in R^{N\times N}$, where $d_{jj} = \sum_l w_{il}$, the graph Laplacian L is written as $L = D - W$ [13]. By combining NMF and the geometrical regularizer, the objective function of GNMF can be written as

$$\begin{aligned} &\min_{U,V} \|X - UV\|_F^2 + \lambda Tr(VLV^T) \\ &\text{s.t. } U,\ V \geq 0 \end{aligned} \tag{4}$$

where $Tr(\cdot)$ and ^{T} refer to the trace and transposition of a matrix, respectively, and $\lambda \geq 0$ is the regularization parameter.

3. OUR PROPOSED TNMF METHOD FOR SPEECH EMOTION RECOGNITION

3.1. Minimizing the distribution divergence

By using the GNMF algorithm, the common latent coding vectors can be obtained for both labeled source and unlabeled target corpora. However, the differences between the distributions of coding vectors are still large in reality, which will have an adverse impact on the recognition performance. To solve this issue, following [14, 15], the empirical maximum mean discrepancy (MMD) algorithm is employed to measure the distribution differences, which can compute the similarities between the means of the labeled source and unlabeled target data in the k dimensional features as follows

$$\begin{aligned} D(V_{src}, V_{tar}) &= \left\| \frac{1}{n_l}\sum_{i=1}^{n_l} v_i - \frac{1}{n_u}\sum_{j=n_l+1}^{n_l+n_u} v_j \right\|^2 \\ &= \sum_{i,j=1}^{N} v_i^T v_j m_{ij} \\ &= Tr(VMV^T) \end{aligned} \tag{5}$$

where $N = n_l + n_u$, $V = [V_{src}, V_{tar}]$, in which V_{src} and V_{tar} are the coding representations of the labeled source and unlabeled target emotional features, n_l and n_u are the corresponding feature numbers, respectively, and $M = [m_{ij}]_{i,j=1}^N$ denotes the MMD matrix, which is given as

$$m_{ij} = \begin{cases} \frac{1}{n_l^2} & v_i, v_j \in V_{src} \\ \frac{1}{n_u^2} & v_i, v_j \in V_{tar} \\ \frac{-1}{n_l n_u} & \text{otherwise} \end{cases} \tag{6}$$

3.2. The transfer NMF method

The transfer learning methods have been successfully used in many applications, e.g., feature learning, image classification and object recognition [7]. Meanwhile, the NMF algorithm can learn robust feature representations. To enrich the feature discrimination and improve the recognition rates, a novel transfer NMF (TNMF) algorithm is presented. By regularizing Eq. (4) with Eq. (5), the objective function of TNMF can be given as

$$\begin{aligned} &\min_{U,V} \|X - UV\|_F^2 + \lambda Tr(VLV^T) + \gamma Tr(VMV^T) \\ &\text{s.t. } U,\ V \geq 0 \end{aligned} \tag{7}$$

where γ is a regularization parameter to trade off the weight between GNMF and the feature distribution matching.

Let $T = \lambda L + \gamma M$, the Eq. (7) can be rewritten as

$$\begin{aligned} &\min_{U,V} \|X - UV\|_F^2 + Tr(VTV^T) \\ &\text{s.t. } U,\ V \geq 0 \end{aligned} \tag{8}$$

Similar to the traditional NMF, the Eq. (8) is not a convex problem when optimizing U and V together. So an iterative alternating algorithm is presented, and U and V are computed separately. According to the matrix properties, the Eq. (8) can be modified as

$$\begin{aligned} &\min_{U,V} Tr(XX^T) + Tr(UVV^TU^T) - 2Tr(XV^TU^T) \\ &\qquad + Tr(VTV^T) \\ &\text{s.t. } U,\ V \geq 0 \end{aligned} \tag{9}$$

Then the Lagrange form can be written as

$$\begin{aligned} L =& Tr(XX^T) + Tr(UVV^TU^T) - 2Tr(XV^TU^T) \\ &+ Tr(VTV^T) + Tr(\beta U) + Tr(\gamma V) \end{aligned} \tag{10}$$

By using the zero gradient condition, the partial derivatives of L with respect to U and V are written as follows

$$\begin{aligned} \frac{\partial L}{\partial U} &= 2UVV^T - 2XV^T + \beta = 0 \\ \frac{\partial L}{\partial V} &= 2U^TUV - 2U^TX + 2VT + \gamma = 0 \end{aligned} \tag{11}$$

where $\beta = [\beta_{ik}] \in R^{M \times K}$ and $\gamma = [\gamma_{kj}] \in R^{K \times N}$ are the Lagrange multiplier matrices. Using the KKT conditions $\beta_{ik}u_{ik} = 0$ and $\gamma_{kj}v_{kj} = 0$ [16], the following equations will be obtained

$$\begin{aligned} &(UVV^T)_{ik}u_{ik} - (XV)_{ik}u_{ik} = 0 \\ &(U^TU)_{kj}v_{kj} + (VT)_{kj}v_{kj} - (U^TX)_{kj}v_{kj} = 0 \end{aligned} \tag{12}$$

Finally, the updating rules can be given as

$$\begin{aligned} u_{ik} &\leftarrow u_{ik}\frac{(XV)_{ik}}{(UVV^T)_{ik}} \\ v_{kj} &\leftarrow v_{kj}\frac{(U^TX + VT^-)_{kj}}{(VU^TU + VT^+)_{kj}} \end{aligned} \tag{13}$$

where T^+ and T^- denote the positive and negative parts of T, respectively. Repeat the updating functions of Eq. (13) until the maximum iteration step is reached.

4. EXPERIMENTS

In this section, the extensive experiments are conducted to evaluate our proposed TNMF approach for cross-corpus speech emotion recognition.

4.1. Data preparation

Two popular emotional speech corpora are employed for our experiments, i.e., the Berlin database [1] and the eNTERFACE database [2]. The important statistics of each dataset are simply summarized in Table 1.

Table 1. The statistics of the datasets

Datasets	Berlin	eNTERFACE
Language	German	English
Size	494	1170
Dimension	1582	1582
Category	7	6

The Berlin dataset is one of the earliest and popular emotional speech corpora. It includes seven kinds of basic emotions, i.e., anger, boredom, disgust, fear, happiness, sadness and neutral. The emotional utterances are recorded by 10 actors (5 males and 5 females) with the predefined content in German. After the listening tests by 20 professionals, total 494 utterances, which can be clearly recognized, are finally obtained and used in our experiments.

The eNTERFACE database is a public audio-visual emotional corpus. It consists of six kinds of basic emotions, including anger, disgust, fear, happiness, sadness and surprise. With predefined English contents, the utterances are acted by 42 subjects (34 males and 8 females) from 14 countries. After listening tests by 2 professionals, 1170 video samples are kept and employed for our tests.

4.2. Experimental setup

In our experiments, two types of speech emotion recognition schemes are used for evaluation, namely *case1* and *case2*. It should be noted that in both cases, the source corpus only contains labeled data, while the target corpus is unlabeled. In *case1*, the eNTERFACE corpus is used as the training dataset, and the Berlin corpus is chosen for testing. Meanwhile, in *case2*, the Berlin corpus is selected for training, and the eNTERACE corpus is used for testing. The five common emotion categories, i.e., anger, disgust, fear, happiness and sadness, are chosen for speech emotion recognition.

The open-source software openSMILE [3] is employed to extract the acoustic features, and the baseline feature set of INTERSPEECH 2010 emotion challenge [17] is chosen for our tests. It consists of 21 functionals applied to 34 basic low level descriptors (LLDs) and their corresponding delta coefficients, and 19 functions applied to 4 F0 related LLDs and the corresponding delta coefficients. Moreover, the duration of the utterance and the F0 onsets are also included into the feature set. Thus, the size of each feature vector is 1582.

In the experiments, the support vector machine (SVM) algorithm is chosen as the classifier. To evaluate the performance of our proposed approach, the following methods are compared, they are the automatic recognition method (Automatic), in which the classifier trained in the source corpus is directly applied to the target corpus, the baseline recognition method (Baseline), in which, the training and testing proce-

[1] http://emodb.bilderbar.info/docu/

[2] http://enterface.net/enterface05/main.php?frame=emotion

[3] http://sourceforge.net/projects/opensmile/

Table 2. Average recognition results of different methods in *case1* (eNTERFACE dataset for training, Berlin dataset for testing)

Methods	Recognition rates (%)					
	Anger	Disgust	Fear	Happiness	Sadness	Average
Baseline	72.98	81.09	68.54	53.01	79.34	70.99
Automatic	31.52	53.05	16.45	20.01	47.22	34.65
DR	34.75	72.13	17.88	25.32	69.07	45.83
TCA	35.43	72.97	19.01	25.95	69.75	49.62
NMF	33.42	68.20	17.03	22.31	50.01	38.19
GNMF	33.65	56.12	17.14	22.42	50.94	39.05
Ours	**36.14**	**74.52**	**19.22**	**26.69**	**71.54**	**51.98**

Table 3. Average recognition results of different methods in *case2* (eNTERFACE dataset for training, Berlin dataset for testing)

Methods	Recognition rates (%)					
	Anger	Disgust	Fear	Happiness	Sadness	Average
Baseline	74.42	55.35	54.01	59.98	60.99	61.39
Automatic	37.25	19.22	17.96	27.18	28.43	28.91
DR	46.99	25.12	29.08	44.01	41.13	37.13
TCA	50.18	28.90	34.57	45.34	44.04	40.92
NMF	39.15	21.27	20.08	26.84	30.15	28.50
GNMF	39.31	21.50	20.43	27.12	30.58	28.83
Ours	**52.58**	**29.53**	**37.62**	**47.01**	**44.71**	**44.02**

dures are conducted on the same single corpus, the dimension reduction based transfer learning method (TL) [6], the transfer component analysis method (TCA) [18]. the NMF method (NMF), the graph NMF method (GNMF), and the proposed TNMF method (Ours).

Each dataset is divided into 5 parts, and in each test, random $4/5$ data of the labeled source corpus and unlabeled target corpus is chosen for training, while the others are for testing. The experiments are conducted 10 times to cover all the possible cases of the training and testing datasets.

4.3. Experimental results and analysis

The experimental results of different methods are summarized in Table 2 and Table 3. First, it can be easily found in either case, the DR, TCA and TNMF methods outperform the other methods. The reason is that they are all transfer learning based algorithms, which consider reducing the distribution distances between different feature datasets. Then, it can be also seen that the NMF, GNMF and TNMF methods can obtain higher recognition rates than the automatic recognition method, which can be attributed to the good feature representations of the non-negative matrix factorization. Meanwhile, either GNMF or TNMF method can perform better than the NMF approach, which shows that considering the local geometrical information is also efficient to the acoustic features.

From the two tables, it can be observed that compared to DR, TCA and GNMF methods, our proposed TNMF approach always obtains higher recognition rates. The reason may be that the TNMF method takes the advantages of both non-negative matrix factorization and transfer learning, and can optimize these two factors together. It can be also seen that the recognition rates of *case2* are always lower than those of *case1*, these results are consistent with those in [19, 20].

5. CONCLUSION

In this paper, a new method called transfer non-negative matrix factorization is presented for cross-corpus speech emotion recognition. A graph based NMF approach is employed to learn the robust representations of the acoustic features, while the discrepancies of the feature distributions of two different datasets, described by the MMD matrix, are also considered and further used as a regularization term of the objective function of NMF. The experiments are carried out on two popular public emotional datasets, and the experimental evidence demonstrates the advantages of our method.

6. ACKNOWLEDGEMENTS

This work was partly supported by the National Basic Research Program of China under Grant 2015CB351704, the National Natural Science Foundation of China (NSFC) under Grants 61231002, 61403328 and 61572419, the Natural Science Foundation of Jiangsu Province under Grant BK20130020, the Natural Science Foundation of Shandong Province under Grant ZR2014FQ016, and the Fundamental Research Funds for the Southeast University under Grant CDLS-2015-04.

7. REFERENCES

[1] Dimitrios Ververidis and Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[2] 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.

[3] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *INTERSPEECH*, pp. 223–227, 2014.

[4] Björn Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?," in *INTERSPEECH*, 2011, pp. 1553–1556.

[5] Jun Deng, Zixing Zhang, Florian Eyben, and Bjorn Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *Signal Processing Letters, IEEE*, vol. 21, no. 9, pp. 1068–1072, 2014.

[6] Peng Song, Yun Jin, Li Zhao, and Minghai Xin, "Speech emotion recognition using transfer learning," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 9, pp. 2530–2532, 2014.

[7] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[8] Daniel D Lee and H Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[9] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[10] Daniel D Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[11] Jialu Liu, Deng Cai, and Xiaofei He, "Gaussian mixture model with local consistency.," in *AAAI*, 2010, vol. 10, pp. 512–517.

[12] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang, "Graph regularized nonnegative matrix factorization for data representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1548–1560, 2011.

[13] Fan Chung, *Spectral graph theory*, vol. 92, American Mathematical Soc., 1997.

[14] Sinno Jialin Pan, James T Kwok, and Qiang Yang, "Transfer learning via dimensionality reduction.," in *AAAI*, 2008, vol. 8, pp. 677–682.

[15] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu, "Transfer joint matching for unsupervised domain adaptation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1410–1417.

[16] Christopher M Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[17] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan, "The interspeech 2010 paralinguistic challenge.," in *INTERSPEECH*, 2010, pp. 2794–2797.

[18] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang, "Domain adaptation via transfer component analysis," *Neural Networks, IEEE Transactions on*, vol. 22, no. 2, pp. 199–210, 2011.

[19] Wenming Zheng, Minghai Xin, Xiaolan Wang, and Bei Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *Signal Processing Letters, IEEE*, vol. 21, no. 5, pp. 569–572, 2014.

[20] Yun Jin, Peng Song, Wenming Zheng, and Li Zhao, "A feature selection and feature fusion combination method for speaker-independent speech emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4808–4812.