Speech Communication 83 (2016) 34–41

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization

Peng Song[a,*], Wenming Zheng[b], Shifeng Ou[c], Xinran Zhang[b], Yun Jin[d], Jinglei Liu[a], Yanwei Yu[a]

[a] School of Computer and Control Engineering, Yantai University, Yantai 264005, P.R. China
[b] Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing 210096, P.R. China
[c] School of Science and Technology for Opto-electronic Information, Yantai University, Yantai 264005, P.R. China
[d] School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221116, P.R. China

ARTICLE INFO

Article history:
Available online 1 August 2016

Keywords:
Speech emotion recognition
Transfer learning
Non-negative matrix factorization
Dimension reduction

ABSTRACT

Automatic emotion recognition from speech has received an increasing amount of interest in recent years, and many speech emotion recognition methods have been presented, in which the training and testing procedures are often conducted on the same corpus. However, in practice, the training and testing speech utterances are collected from different conditions or devices, which will have adverse effects on recognition performance. To address this problem, in this paper, a novel cross-corpus speech emotion recognition method, called transfer non-negative matrix factorization (TNMF) is proposed. Specifically, the NMF approach, which is popular in computer vision and pattern recognition fields, is utilized to obtain low dimensional representations of emotional features. Meanwhile, the discrepancies between source and target data sets are considered, and the maximum mean discrepancy (MMD) algorithm is used for similarity measurement. Then, the TNMF method, which jointly optimizes the NMF and MMD algorithms, is presented. Moreover, to further improve the recognition performance, two variants of TNMF, called transfer graph regularized NMF (TGNMF) and transfer constrained NMF (TCNMF), are proposed, respectively. Several experiments are carried out on three popular emotional databases, and the results demonstrate the effectiveness and robustness of our scheme.

1. Introduction

Speech emotion recognition, which aims at predicting emotional states from his or her speech, has been a hot research topic in speech signal processing field. With the development of computer technologies, the demands for emotion recognition in new spoken dialogue systems are very urgent. It has been proven very useful in many real applications (Cowie et al., 2001; El Ayadi et al., 2011; Ververidis and Kotropoulos, 2006). For example, in health care field, the intelligent robots, which monitor the patients' emotional states, can help doctors diagnose the mental illness. In intelligent vehicle, the emotion recognition system can monitor the drivers' emotion variations to avoid accidents. It can be also deployed in many human-computer interaction (HCI) based entertainment systems.

In speech signal processing and affective computing fields, speech emotion recognition plays a very important role. Researchers have long sought robust feature representations and classification algorithms. As shown in Fig. 1, a classic speech emotion recognition system can be divided into two parts, i.e., feature extraction versus emotion classification. The goal of feature extraction aims to achieve useful emotional features from speech signal while the main task of emotion classification is to obtain the emotional categories for a testing sample. Over the past decades, many classification approaches, popular in pattern recognition and machine learning, have been developed to implement the classification function, e.g., support vector machine (SVM), neural network (NN), Gaussian mixture model (GMM) and hidden Markov model (HMM) (El Ayadi et al., 2011; Ververidis and Kotropoulos, 2006). Besides, the extreme learning machine (ELM) (Han et al., 2014) and deep neural network (DNN) (Amer et al., 2014; Zheng et al., 2015) approaches are also introduced for speech emotion recognition. All these methods can achieve satisfactory performance to

* Corresponding author.
E-mail addresses: pengsong@ytu.edu.cn, pengsongseu@gmail.com (P. Song), wenming_zheng@seu.edu.cn (W. Zheng), ousfeng@ytu.edu.cn (S. Ou), 230139080@seu.edu.cn (X. Zhang), jiny@jsnu.edu.cn (Y. Jin), jinglei_liu@sina.com (J. Liu), yuyanwei@ytu.edu.cn (Y. Yu).

http://dx.doi.org/10.1016/j.specom.2016.07.010
0167-6393/

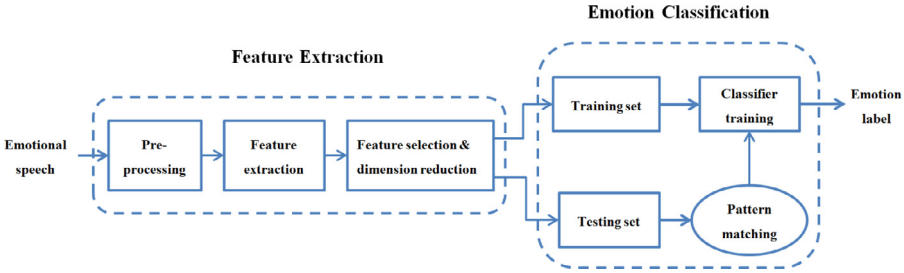

Fig. 1. Flowchart of speech emotion recognition.

some extent. However, it should be noted that they are performed on the assumption that the training and testing data are obtained under the same condition. In practice, the speech utterances are often collected under different conditions. As a result, the recognition rates will obviously drop when the training and testing data are from different corpora.

To solve the above mentioned problem, various researchers have considered the case when the speech utterances are drawn from different scenarios, e.g., languages, noises, ages, genders. Recently, a considerable amount of studies have been made in speech community. Some algorithms popular in speech and speaker recognition, e.g., maximum a posteriori (MAP) (Hu et al., 2007), factor analysis (FA) (Mariooryad and Busso, 2014; Song et al., 2015), nuisance attribute projection (NAP) (Sanchez et al., 2010), have been successfully applied to speech emotion recognition. In Xia et al. (2014), Xia et al. propose to model gender information to obtain robust emotional representations. In Schuller et al. (2011), Schuller et al. evaluate the performance of cross-corpus emotion recognition on six different emotional data sets, in which two novel voting strategies are investigated to improve the cross-corpus recognition rates. In Jeon et al. (2013), Jeon et al. conduct a preliminary study on three different languages to investigate the effects of cross-lingual emotional data on human perception and automatic recognition. To realize cross-corpus speech emotion recognition, Deng et al. (2014) introduce an adaptive denoising based domain adaptation method. Abdelwahab and Busso (2015) explore a supervised domain adaptation algorithm to reduce the mismatch problems between training and testing conditions. In Mao et al. (2016), Mao et al. present a new domain adaptation method where the priors of source and target classes are considered.

All these previously studies focus on reducing the difference between different data sets. However, they fail to consider the divergence between feature distributions of different corpora (Pan and Yang, 2010; Song et al., 2014). Recently, NMF algorithms (Jeong et al., 2009; Kim et al., 2009) have been studied on speech emotion recognition, in which robust feature representations can be obtained to boost the recognition performance. However, they do not take into account the differences between the training and testing data. Inspired by recent progress in matrix factorization and transfer learning, in this paper, we propose a novel cross-corpus speech emotion recognition algorithm, called transfer non-negative matrix factorization (TNMF), which explicitly considers the difference between feature distributions of training and testing data. Our goal is to obtain common robust feature representations for both labeled source and unlabeled target data sets. To achieve this, two types of NMF algorithms, namely graph regularized NMF (GNMF) (Cai et al., 2011) and constrained NMF (CNMF) (Liu et al., 2012), are employed to learn robust low-dimensional feature representations. Meanwhile, the maximum mean discrepancy (MMD) approach (Borgwardt et al., 2006) is adopted for similarity measurement. Then two novel transfer NMF approaches, called transfer GNMF (TGNMF) and transfer CNMF (TCNMF) are proposed, respectively, and the corresponding optimization schemes are also presented to solve the objective functions. This paper is an extended version of our work presented at ICASSP 2016 (Song et al., 2016). New contributions include the newly proposed TCNMF algorithm, analysis of the TGNMF and TCNMF approaches, and extensive experimental results. Meanwhile, Different from our previous work on transfer learning based speech emotion recognition (Song et al., 2014), instead of using traditional unsupervised dimensionality reduction algorithms, in this work, the NMF is employed to learn robust feature representations, and two novel transfer NMF algorithms, i.e., TGNMF and TCNMF, are presented.

The remainder of this paper is organized as follows. In Section 2, we briefly review the NMF method and introduce the idea of TNMF. In Section 3, Two extensions of TNMF methods and their corresponding optimization algorithms are provided in detail. Experimental results are presented in Section 4. Finally, Section 5 provides some conclusion remarks.

2. Transfer non-negative matrix factorization

2.1. Non-negative matrix factorization

Non-negative matrix factorization (NMF) is an unsupervised learning algorithm, solving many real-world problems with non-negative data (Lee and Seung, 1999). It aims to find two non-negative matrices whose product is an approximation of the original matrix. It has been successfully used in widespread tasks (Cai et al., 2011; Lee and Seung, 1999; Liu et al., 2012), e.g., face recognition, gene expression, text mining and document representation.

Given a data matrix $X = [x_1, \ldots, x_N] \in R^{M \times N}$, NMF aims to seek an approximation of X via the product of dictionary matrix $U = [u_{ik}] \in R^{M \times K}$ and the corresponding coding matrix $V = [v_{kj}] \in R^{K \times N}$, which minimizes the objective function as follows:

$$\min_{U,V} \|X - UV\|_F^2 \tag{1}$$

where $U, V \geq 0$, $\|\cdot\|_F$ is a Frobenius norm and $K \ll \{M, N\}$.

Although the above objective function is not convex when optimizing U and V together, it is convex in U and V only. In Lee and Seung (2001), Lee et al. propose an iterative algorithm to solve this problem, and the update rules are given as

$$u_{ik} \leftarrow u_{ik} \frac{(XV)_{ik}}{(UV^TV)_{ik}} \tag{2}$$

$$v_{kj} \leftarrow v_{kj} \frac{(X^TU)_{kj}}{(VU^TU)_{kj}} \tag{3}$$

where ^{T} refers to the transposition of a matrix.

2.2. Minimizing the distribution divergence

By NMF algorithm, the latent low dimensional coding matrix V can be obtained. One may expect that this coding matrix

can capture the commonality underlying labeled source and unlabeled target corpora. However, it should be found that the difference between the feature distributions of the two corpora is often very large, even in the K low dimensional feature space (Pan and Yang, 2010). In this paper, following (Long et al., 2013; Pan et al., 2008; Pan and Yang, 2010), the maximum mean discrepancy (MMD) as a nonparametric distance measurement (Borgwardt et al., 2006), which compares the distributions in reproducing kernel Hilbert space (RKHS), is employed for similarity measurement. Let V_{src} and V_{tar} be the coding representations of labeled source and unlabeled target emotional features, and n_l and n_u be the corresponding feature numbers, respectively, the distance between V_{src} and V_{tar} is written as

$$dist(V_{src}, V_{tar}) = \left\| \frac{1}{n_l}\sum_{i=1}^{n_l} v_i - \frac{1}{n_u}\sum_{j=n_l+1}^{N} v_j \right\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{N} v_i^T v_j m_{ij} = Tr(VMV^T) \tag{4}$$

where $\mathcal{H}$ refers to a universal RKHS, $Tr(\cdot)$ is the trace of a matrix, $N = n_u + n_l$ and $M = [m_{ij}]_{i,j=1}^{N}$ with

$$m_{ij} = \begin{cases} \frac{1}{n_l^2} & v_i, v_j \in V_{src} \\ \frac{1}{n_u^2} & v_i, v_j \in V_{tar} \\ \frac{-1}{n_l n_u} & \text{otherwise} \end{cases} \tag{5}$$

2.3. The transfer NMF approach

The transfer learning techniques have been proven very successful in many applications, e.g., image classification, text categorization and sentiment analysis (Pan and Yang, 2010; Song et al., 2016). Meanwhile, the NMF algorithm can learn robust low dimensional feature representations. To perform robustly across the labeled source and unlabeled target data, a transfer NMF (TNMF) approach is presented. With TNMF, we aim to learn robust representations for emotional speech from different data sets. In this way, a classifier trained on labeled source corpus can generalize better on unlabeled target corpus.

By combining Eqs. (1) and (4), the objective function of TNMF can be written as follows:

$$\min_{U,V} \|X - UV\|_F^2 + \lambda Tr(VMV^T) \quad \text{s.t. } U, V \geq 0 \tag{6}$$

where λ is a regularization parameter to balance feature representation and distribution matching.

3. Extensions of the transfer NMF

Recently, some improved methods have been proposed under the framework of NMF. For example, Cai et al. (2011) present a graph regularized NMF (GNMF) approach, in which a nearest neighbor graph is constructed to encode the geometric information of the data space. In Liu et al. (2012), Liu et al. propose a constrained NMF (CNMF) method, which considers the label information as a hard constraint. Motivated by these recent progresses, in this paper, we will extend the TNMF algorithm and present two novel methods, called transfer GNMF (TGNMF) and transfer CNMF (TCNMF), respectively. In this section, the objective functions of TGNMF and TCNMF approaches are firstly constructed, and then the iterative algorithms are developed to optimize them.

3.1. Transfer graph regularized NMF

Many previous studies (Belkin and Niyogi, 2001; Cai et al., 2011; Roweis and Saul, 2000) have shown that naturally occurring data often reside on or close to an underlying low dimensional submanifold, so a graph regularized NMF, called GNMF, is presented in Cai et al. (2011), where a graph structure is used as a regularization of NMF. Given a set of M-dimensional data points $X = [x_1, \ldots, x_N] \in R^{M \times N}$, we can construct a graph G with N vertices, in which each vertex represents a data point. Let $W = [w_{ij}] \in R^{N \times N}$ be the weight matrix of G, the most commonly and simplest $0-1$ weighting algorithm is adopted, which is written as

$$w_{ij} = \begin{cases} 1 & \text{if } x_j \in N_p(x_i) \text{ or } x_i \in N_p(x_j) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $N_p(x_i)$ and $N_p(x_j)$ are the p nearest neighbors of x_i and x_j, respectively.

Let $d_i = \sum_{j=1}^{N} w_{ij}$ be the degree of x_i, and $D = \text{diag}(d_1, \cdots, d_N)$ be a diagonal matrix. Considering the problem of mapping the graph G to the coding representations V, a reasonable solution (Belkin and Niyogi, 2001) to find a good map is to minimize the following objective function

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(v_i - v_j)^2 w_{ij} = Tr(VLV^T) \tag{8}$$

where $L = D - W$ is the graph Laplacian matrix. By combining Eq. (1) with Eq. (8), the objective function of GNMF can be written as

$$\min_{U,V} \|X - UV\|_F^2 + \gamma Tr(VLV^T) \quad \text{s.t. } U, V \geq 0 \tag{9}$$

where $\gamma \geq 0$ is a regularization parameter. By incorporating the MMD algorithm into Eq. 9, the objective function of our proposed transfer GNMF is

$$\min_{U,V} \|X - UV\|_F^2 + \gamma Tr(VLV^T) + \lambda Tr(VMV^T) \quad \text{s.t. } U, V \geq 0 \tag{10}$$

Given $T = \gamma L + \lambda M$, the above objective function can be modified as

$$\min_{U,V} \|X - UV\|_F^2 + Tr(VTV^T) \quad \text{s.t. } U, V \geq 0 \tag{11}$$

As conventional NMF, the above equation is not convex in computing both U and V together (Lee and Seung, 2001). Therefore it is unrealistic to find a global minimization of the objective function. In the following, an iterative alternating algorithm is introduced.

The Eq. 11 can be rewritten as

$$\min_{U,V} \quad Tr(XX^T) + Tr(UVV^TU^T) - 2Tr(XV^TU^T) + Tr(VTV^T) \quad \text{s.t. } U, V \geq 0 \tag{12}$$

Let $\beta = [\beta_{ik}] \in R^{M \times K}$ and $\sigma = [\sigma_{kj}] \in R^{K \times N}$ be the Lagrange multiplier matrices, the Lagrange function $\mathcal{L}$ is

$$\mathcal{L} = Tr(XX^T) + Tr(UVV^TU^T) - 2Tr(XV^TU^T) + Tr(VTV^T) + Tr(\beta U) + Tr(\sigma V) \tag{13}$$

The partial derivatives of $\mathcal{L}$ with respect to U and V are given as

$$\frac{\partial \mathcal{L}}{\partial U} = -2XV^T + 2UVV^T + \beta = 0 \tag{14}$$

$$\frac{\partial \mathcal{L}}{\partial V} = -2U^TX2U^TUV + 2VT + \sigma = 0 \tag{15}$$

According to the Karush-Kuhn-Tucker (KKT) conditions $\beta_{ik}u_{ik}=0$ and $\sigma_{kj}v_{kj}=0$ (Bishop et al., 2006), the following equations for u_{ik} and v_{kj} will be obtained as

$$(UVV^T)_{ik}u_{ik}-(XV)_{ik}u_{ik}=0 \tag{16}$$

$$(U^TU)_{kj}v_{kj}+(VT)_{kj}v_{kj}-(U^TX)_{kj}v_{kj}=0 \tag{17}$$

The above equations lead to the following update rules:

$$u_{ik}\leftarrow u_{ik}\frac{(XV)_{ik}}{(UVV^T)_{ik}} \tag{18}$$

$$v_{kj}\leftarrow v_{kj}\frac{(U^TX+VT^-)_{kj}}{(VU^TU+VT^+)_{kj}} \tag{19}$$

where T^+ and T^- are the positive and negative parts of T, respectively.

3.2. Transfer constrained NMF

The above mentioned NMF and GNMF are unsupervised algorithms. That is, they are not very applicable to many real-world classification problems where some labeled data are provided. Many previous studies have shown that a small amount of labeled data combined with unlabeled data can significantly improve the learning performance (Cai et al., 2007; Yao et al., 2015). This can naturally give rise to a semi-supervised extension of NMF.

Recently, Liu et al. (2012) present a semi-supervised NMF approach, called constrained NMF (CNMF). Given a feature set $X=\{X_{src},X_{tar}\}$, where $X_{src}=\{x_i\}_{i=1}^{n_l}$ and $X_{tar}=\{x_i\}_{i=1}^{n_u}$ denote the labeled source and unlabeled target data, respectively, a $c\times n_l$ label indicator matrix L is built, where $l_{ij}=1$ if x_i is labeled with the j-th class, $l_{ij}=0$ otherwise. Then, a label constraint matrix A is defined as

$$A=\begin{pmatrix}L_{c\times n_l} & 0\\ 0 & \mathbf{I}_{n_u}\end{pmatrix} \tag{20}$$

where $\mathbf{I}_{n_u}$ is an $n_u\times n_u$ identity matrix.

To impose the label constraint, an auxiliary matrix D is introduced, and the coding matrix V can be represented as

$$V=DA \tag{21}$$

In CNMF, the NMF is extended to semi-supervised NMF with the label constraint. The goal of CNMF is to find two non-negative matrices U and D, and the objective function becomes

$$\min_{U,D}\|X-UDA\|_F^2 \tag{22}$$

As TGNMF, by introducing the MMD constraints, a transfer CNMF (TCNMF) approach is also presented. By combining Eqs. 4 and 22, the objective function is represented as follows:

$$\begin{aligned}&\min_{U,D}\|X-UDA\|_F^2+\lambda Tr(DAM(DA)^T)\\ &\text{s.t.}\quad U,\ D\geq 0\end{aligned} \tag{23}$$

where λ is a regularization parameter. As traditional NMF algorithms, the objective function of TCNMF is not convex to jointly optimize U and D, and an iterative algorithm is also proposed.

With the matrix properties $Tr(AD)=Tr(DA)$ and $Tr(A)=Tr(A^T)$, the Eq. 23 can be rewritten as

$$\begin{aligned}&\min_{U,D}\ Tr(XX^T)+Tr\big(U(DA)(DA)^TU^T\big)-2Tr\big(X(DA)^TU^T\big)\\ &\quad+\lambda Tr\big((DA)M(DA)^T\big)\\ &\text{s.t.}\quad U,\ D\geq 0\end{aligned} \tag{24}$$

Table 1
The statistics of the databases.

Database	Language	Size	# of classes	# of features
Emo-DB	German	494	7	1582
eNTERFACE	English	1170	6	1582
FAU Aibo	German	48401	11	1582

The Lagrange function $\mathcal{L}$ is given as

$$\begin{aligned}\mathcal{L}=&\ Tr(XX^T)+Tr\big(UDA(DA)^TU^T\big)-2Tr\big(X(DA)^TU^T\big)\\ &+Tr\big(\lambda DAM(DA)^T\big)+Tr(\beta U)+Tr(\gamma DA)\end{aligned} \tag{25}$$

where $\beta=[\beta_{ik}]\geq 0$ and $\gamma=[\gamma_{kj}]\geq 0$ are the Lagrange multiplier matrices.

Requiring the derivatives of $\mathcal{L}$ with respect to U and D vanish, we will obtain

$$\frac{\partial\mathcal{L}}{\partial U}=2UDAA^TD^T-2XA^TD^T+\beta=0 \tag{26}$$

$$\frac{\partial\mathcal{L}}{\partial D}=2U^TUDAA^T-2U^TX+2\lambda DAMA^T+\gamma=0 \tag{27}$$

By the KKT conditions $\beta_{ik}u_{ik}=0$ and $\gamma_{kj}d_{kj}=0$ (Bishop et al., 2006), the equations for u_{ik} and d_{kj} are given as follows:

$$\big(UDAA^TD^T\big)_{ik}u_{ik}-(XDA)_{ik}u_{ik}=0 \tag{28}$$

$$(U^TU)_{kj}d_{kj}+\lambda(DAMA)_{kj}d_{kj}-(U^TX)_{kj}d_{kj}=0 \tag{29}$$

These equations lead to the update rules as

$$u_{ik}\leftarrow u_{ik}\frac{(XDA)_{ik}}{\big(UDAA^TD^T\big)_{ik}} \tag{30}$$

$$d_{kj}\leftarrow d_{kj}\frac{(U^TX+DA\lambda M^-)_{kj}}{(DAU^TU+DA\lambda M^+)_{kj}} \tag{31}$$

where M^+ and M^- correspond to the positive and negative parts of M, respectively.

4. Experimental results

In this section, several experiments are carried out to evaluate our proposed transfer NMF approaches for cross-corpus speech emotion recognition.

4.1. Descriptions of data sets

Three popular emotional databases are employed for our experiments, i.e., Emo-DB (Burkhardt et al., 2005), eNTERFACE (Martin et al., 2006) and FAU Aibo (Schuller et al., 2009). The important statistics of each corpus are summarized below (see also Table 1).

The Emo-DB database (Burkhardt et al., 2005) is one of the most popular and earliest emotional data sets. It consists of seven types of basic emotions, i.e., happiness, anger, boredom, disgust, fear, sadness and neutral. The sentences are uttered by 10 German professional actors with predefined content. Finally, as shown in Table 2, 494 utterances in which emotions can be clearly recognized are obtained.

The eNTERFACE database (Martin et al., 2006) is an audio-visual English emotional data set. It includes six types of basic emotions, i.e., happiness, anger, disgust, fear, sadness and surprise. With predefined English content, the utterances are recorded by 42 subjects from 14 countries. Hence, total 1170 video samples are collected as shown in Table 3.

Table 2
The number of each emotion category in the Emo-DB database.

Database	Happiness	Anger	Boredom	Disgust	Fear	Sadness	Neutral	Sum
Emo-DB	64	127	79	38	55	53	78	494

Table 3
The number of each emotion category in the eNTERFACE database.

Database	Happiness	Anger	Disgust	Fear	Sadness	Surprise	Sum
eNTERFACE	205	200	189	189	195	192	1170

Table 4
The number of each emotion category in the FAU Aibo database.

Database	Neutral	Non-neutral	Sum
FAU Aibo	39169	9232	48401

The FAU Aibo database (Schuller et al., 2009) is another commonly used emotional speech database, which is chosen for Interspeech 2009 emotion challenge. The emotional speech utterances are recorded by 51 children (21 boys and 30 girls) from two schools, who are interacted with a sony's robot Aibo for recording. Then five professional labelers are employed to annotate the emotion label for each recording. Finally, total 48401 words with 11 types of emotions are recorded, among which, over 80% are neutral, while the others are non-neutral. As shown in Table 4, in our experiments, only two types of emotions, i.e., neutral versus non-neutral, are considered for evaluation.

4.2. Experimental setup

In our experiments, three types of cross-corpus speech emotion recognition schemes are considered for evaluation, i.e., *case1, case2* and *case3*. It should be noted that, in all cases, the source corpus is labeled, while the target corpus is unlabeled. In *case1*, the eNTERFACE database is used as the source database, while the Emo-DB is used for testing. Meanwhile, in *case2* and *case3*, the Emo-DB corpus is used as the training corpus, while the eNTERFACE and FAU Aibo are chosen as the target databases, respectively. In *case1* and *case2*, five common emotional categories, i.e., happiness, anger, disgust, fear and sadness are used for evaluation. Meanwhile, in *case3*, two types of emotions, i.e., neutral versus non-neutral, are chosen for evaluation. Each corpus is divided into five parts, among which, in each test, random 4/5 of the source and target data are used for training, while the others are chosen for testing. And among the target training data, the 5% are labeled for CNMF and TCNMF algorithms. The tests are repeated 10 times to cover all the possible cases for training and testing databases.

The openSMILE toolkit (Eyben et al., 2010) is chosen to extract the acoustic features. And the baseline feature set of Interspeech 2010 paralinguistic challenge (Schuller et al., 2010) is used for our tests. As shown in Table 5, it consists of 34 basis low level descriptors (LLDs). 21 functionals are applied to the above 34 LLDs and their corresponding delta coefficients, while 19 functional are applied to 4 F0 related LLDs and their corresponding delta coefficients. In addition, the durations and F0 onsets are also considered and included into the feature set. Thus, the dimension of the emotional feature vector is 1582.

To demonstrate how the recognition performance can be improved by our proposed approach, the following 9 methods are compared:

Table 5
LLDs for our tests.

LLDs	Number
Loudness	1
MFCC [0–14]	15
Log Mel frequency band [0–7]	8
LSP [0–7]	8
F0	1
F0 envelope	1
Voicing probability	1
Jitter local	1
Jitter consecutive frame pairs	1
Shimmer local	1

- Baseline method (*Baseline*), in which the training and testing procedures are carried out on the same single database.
- Traditional method (*Automatic*), in which the classifier trained in source corpus is directly applied to emotion recognition of target corpus.
- Dimension reduction based transfer learning method (DR) (Pan et al., 2008; Song et al., 2014), one of the classic transfer learning algorithms.
- Transfer component analysis method (TCA) (Pan et al., 2011), one of the classic transfer learning algorithms.
- Conventional NMF method (NMF) (Lee and Seung, 2001).
- Graph regularized NMF method (GNMF) (Lee and Seung, 2001).
- Constrained NMF method (CNMF) (Liu et al., 2012).
- Our proposed transfer graph regularized NMF method (TGNMF).
- Our proposed transfer constrained NMF method (TCNMF).

In our experiments, the support vector machine (SVM) is chosen as the baseline algorithm since it is simple and very powerful on classification problems. It is also important to choose suitable model parameters via cross validations. The test values for the size of dictionary K are {16, 32, 64, 128, 256, 512}, and finally the optimal value of K is set to 128. Besides, the number of nearest neighbors p in GNMF approach is set to 5, by searching $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, the MMD trade-off parameter λ and the regularization parameter γ are optimized as 1 and 100, respectively, and the number of iterations is set to 100.

4.3. Experimental results

The experimental results of different approaches in *case1* and *case2* are depicted in Table 6 and Table 7, in which the recognition rates of each emotion category and the overall average performance are summarized. From the two tables, we have the following observations.

Firstly, in both cases, compared with the *Baseline* method carried out on single corpus, the recognition rates of cross-corpus *Automatic* scheme drop sharply.

Secondly, it can be easily found that, the DR, TCA, TGNMF and TCNMF methods significantly improve the recognition rates. One

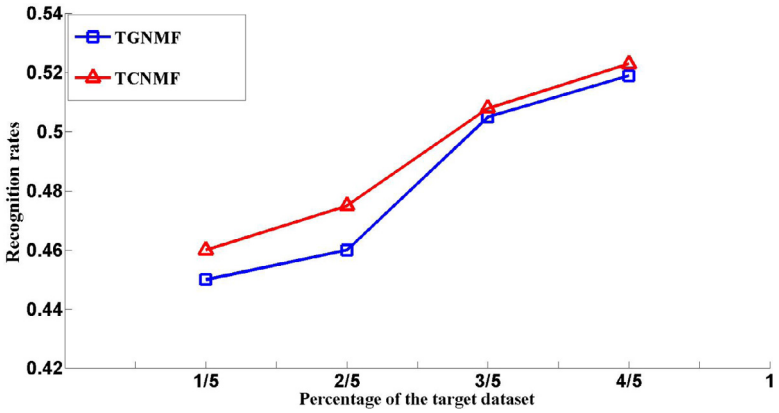

Fig. 2. Recognition rates with different percentages of target data in *case1*.

Table 6
The average recognition results using different methods in *case1* (eNTERFACE database for training, Emo-DB database 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	56.12	17.03	22.31	50.01	38.19
GNMF	33.65	68.20	17.14	22.42	50.94	39.05
CNMF	33.51	68.59	17.15	22.41	50.98	39.12
TGNMF	**36.14**	**74.52**	**19.22**	**26.69**	**71.54**	**51.98**
TCNMF	**36.81**	**74.81**	**19.54**	**27.06**	**71.68**	**52.10**

Table 7
The average recognition results using different methods in *case2* (Emo-DB database for training, eNTERFACE database 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
CNMF	39.40	21.58	20.41	27.15	30.62	28.86
TGNMF	**52.58**	**29.53**	**37.62**	**47.01**	**44.71**	**44.02**
TCNMF	**52.71**	**29.68**	**38.25**	**47.36**	**45.16**	**44.86**

possible reason is that they are all transfer learning based algorithms, which consider the difference between feature distributions of different corpora.

Thirdly, compared to other algorithms except *Baseline*, our proposed TGNMF and TCNMF methods always obtain higher recognition rates. This can be attributed to that these two approaches take advantages of both non-negative matrix factorization and transfer learning, and optimize them together. Meanwhile, in both cases, the recognition rates of TGNMF are slightly higher than those of TCNMF. The reason may be that TCNMF is a semi-supervised algorithm while the TGNMF considers the geometrical information, and the label information is more important than geometric structure to our cross-corpus speech emotion recognition.

Finally, it can be also found that the recognition rates of *case2* are lower than those of *case1*, which are consistent with the results on conventional single corpus (Jin et al., 2014; Zheng et al., 2014).

In addition, the recognition performance with different numbers of target data is also investigated, and the results are given in Fig. 2 and Fig. 3. It can be naturally thought that the recognition rates will become higher with the increase of target training data, and from the two figures, it can be found that the best recognition rates are achieved when 4/5 of unlabeled data are used. However, it should be also observed that, in *case2*, the recognition rates of 3/5 are a little lower than that of 2/5. The reason may be that although more unlabeled data are used for training, some of them have adverse effects on our proposed methods. In the future, we will investigate how to select the useful unlabeled subsets.

Table 8
The average recognition results using different methods in *case3* (Emo-DB database for training, FAU Aibo database for testing).

Methods	Recognition rates (%)		
	Neutral	Non-neutral	Average
Baseline	73.17	55.02	66.39
Automatic	42.61	29.22	34.68
DR	54.25	45.01	46.08
TCA	54.71	46.12	46.62
NMF	43.52	29.78	25.05
GNMF	44.01	30.48	25.72
CNMF	44.08	31.1 4	25.83
TGNMF	**56.08**	**46.21**	**46.97**
TCNMF	**56.24**	**47.30**	**47.25**

Table 9
Comparisons between TCA and TNMF in *case1* and *case2*.

Cases	Methods	Recognition rates (%)					
		Anger	Disgust	Fear	Happiness	Sadness	Average
case1	TCA	35.43	72.97	19.01	25.95	69.75	49.62
	TNMF	35.81	73.05	19.02	26.18	70.16	50.63
case2	TCA	50.18	28.90	34.57	45.34	44.04	40.92
	TNMF	50.68	29.01	34.89	46.12	45.13	41.57

The recognition rates of different approaches in *case3* are given in Table 8. From the table, it can be found that, to classification of two emotion categories, the recognition results show the same trends as those given in Table 6 and Table 7. It can be also observed that our proposed transfer NMF approaches obtain the best recognition rates, which are consistent with those in *case1* and *case2*.

As discussed above, our proposed TGNMF and TCNMF achieve better performance than TCA, which is a state-of-the-art transfer learning algorithm for feature extraction. In order to better investigate whether the improvements are obtained by NMF, the TCA algorithm is compared with the TNMF approach, which can be seen as a special case of TGNMF with $\gamma = 0$ in Eq. 10. From Table 9 and

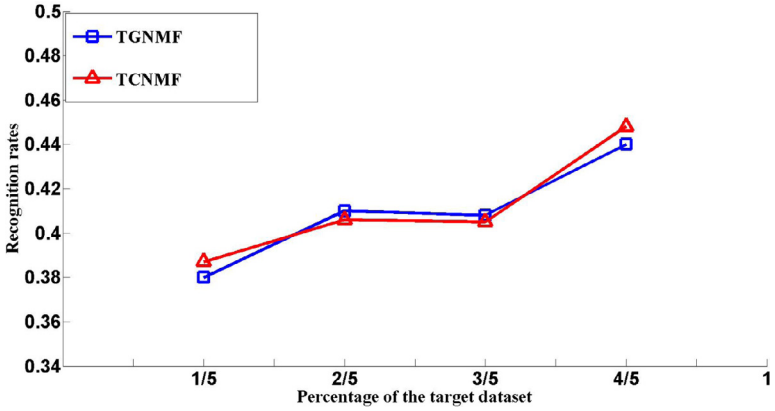

Fig. 3. Recognition rates with different percentages of target data in *case2*.

Table 10
Comparisons between TCA and TNMF in *case3*.

Methods	Recognition rates (%)		
	Neutral	Non-neutral	Average
TCA	54.71	46.12	46.62
TNMF	54.82	46.25	46.69

Table 10, it can be observed that TNMF performs better than TCA in all cases. The reasons might be that, compared with the traditional dimensionality reduction algorithms used in TCA (Pan et al., 2011), the NMF algorithm can learn more robust feature representations.

5. Conclusions

In this paper, we have presented a novel cross-corpus speech emotion recognition method, called transfer non-negative matrix factorization, which makes use of both NMF and transfer learning techniques. In this approach, the NMF approach is used to learn robust representations of the acoustic emotional features. Meanwhile, the differences between feature distributions of two corpora, described by MMD algorithm, are considered and used as a regularization term of NMF. Moreover, two types of NMF approaches, i.e., graph regularized NMF (GNMF) and constrained NMF (CNMF), are introduced, and two novel transfer NMF algorithms, called transfer GNMF (TGNMF) and transfer CNMF (TCNMF) are presented, respectively. The experimental results on three popular emotional databases demonstrate the effectiveness of our approach. Both TGNMF and TCNMF outperform the existing state-of-the-art approaches.

Acknowledgments

This paper is supported by the Natural Science Foundation of Shandong Province under Grant ZR2014FQ016, the National Natural Science Foundation of China under Grants 61231002, 61403328 and 61572419, and the Fundamental Research Funds for the Southeast University under Grant CDLS-2015-04.

References

Abdelwahab, M., Busso, C., 2015. Supervised domain adaptation for emotion recognition from speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5058–5062.
Amer, M.R., Siddiquie, B., Richey, C., Divakaran, A., 2014. Emotion detection in speech using deep networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3724–3728.
Belkin, M., Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing Systems (NIPS), 14, pp. 585–591.
Bishop, C.M., et al., 2006. Pattern Recognition and Machine Learning, 1. springer New York.
Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.-P., Schölkopf, B., Smola, A.J., 2006. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22 (14), e49–e57.
Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., 2005. A database of german emotional speech. In: INTERSPEECH, 5, pp. 1517–1520.
Cai, D., He, X., Han, J., 2007. Semi-supervised discriminant analysis. In: IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 1–7.
Cai, D., He, X., Han, J., Huang, T.S., 2011. Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. Pattern Anal. Mach. Intell. 33 (8), 1548–1560.
Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. 18 (1), 32–80.
Deng, J., Zhang, Z., Eyben, F., Schuller, B., 2014. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. IEEE Signal Process. Lett. 21 (9), 1068–1072.
El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognit. 44 (3), 572–587.
Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In: The International Conference on Multimedia. ACM, pp. 1459–1462.
Han, K., Yu, D., Tashev, I., 2014. Speech emotion recognition using deep neural network and extreme learning machine. In: INTERSPEECH. ISCA, pp. 223–227.
Hu, H., Xu, M.-X., Wu, W., 2007. GMM supervector based SVM with spectral features for speech emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 413–416.
Jeon, J.H., Le, D., Xia, R., Liu, Y., 2013. A preliminary study of cross-lingual emotion recognition from speech: automatic classification versus human perception. In: INTERSPEECH, pp. 2837–2840.
Jeong, K., Song, J., Jeong, H., 2009. Nmf features for speech emotion recognition. In: Proceedings of the 2009 International Conference on Hybrid Information Technology. ACM, pp. 368–374.
Jin, Y., Song, P., Zheng, W., Zhao, L., 2014. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4808–4812.
Kim, D., Lee, S.-Y., Amari, S.-i., 2009. Representative and discriminant feature extraction based on nmf for emotion recognition in speech. In: International Conference on Neural Information Processing. Springer, pp. 649–656.
Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–791.
Lee, D.D., Seung, H.S., 2001. Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems (NIPS), pp. 556–562.
Liu, H., Wu, Z., Li, X., Cai, D., Huang, T.S., 2012. Constrained nonnegative matrix factorization for image representation. IEEE Trans. Pattern Anal. Mach. Intell. 34 (7), 1299–1311.
Long, M., Ding, G., Wang, J., Sun, J., Guo, Y., Yu, P., 2013. Transfer sparse coding for robust image representation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 407–414.
Mao, Q., Xue, W., Rao, Q., Zhang, F., Zhan, Y., 2016. Domain adaptation for speech emotion recognition by sharing priors between related source and target classes. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2608–2612.
Marioryad, S., Busso, C., 2014. Compensating for speaker or lexical variabilities in speech for emotion recognition. Speech Commun. 57, 1–12.

Martin, O., Kotsia, I., Macq, B., Pitas, I., 2006. The enterface'05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops. IEEE. 8–8

Pan, S.J., Kwok, J.T., Yang, Q., 2008. Transfer learning via dimensionality reduction. In: AAAI, 8, pp. 677–682.

Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q., 2011. Domain adaptation via transfer component analysis. IEEE Trans. Neural Networks 22 (2), 199–210.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Sanchez, M.H., Tür, G., Ferrer, L., Hakkani-Tür, D., 2010. Domain adaptation and compensation for emotion detection. In: INTERSPEECH. ISCA, pp. 2874–2877.

Schuller, B., Steidl, S., Batliner, A., 2009. The interspeech 2009 emotion challenge. In: INTERSPEECH, 2009, pp. 312–315.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A., Narayanan, S.S., 2010. The interspeech 2010 paralinguistic challenge. In: INTERSPEECH, pp. 2794–2797.

Schuller, B., Zhang, Z., Weninger, F., Rigoll, G., 2011. Using multiple databases for training in emotion recognition: To unite or to vote? In: INTERSPEECH, pp. 1553–1556.

Song, P., Jin, Y., Zha, C., Zhao, L., 2015. Speech emotion recognition method based on hidden factor analysis. Electron. Lett. 51 (1), 112–114.

Song, P., Jin, Y., Zhao, L., Xin, M., 2014. Speech emotion recognition using transfer learning. IEICE Trans. Inf. Syst. 97 (9), 2530–2532.

Song, P., Ou, S., Zheng, W., Jin, Y., Zhao, L., 2016. Speech emotion recognition using transfer non-negative matrix factorization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5180–5184.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: Resources, features, and methods. Speech commun. 48 (9), 1162–1181.

Xia, R., Deng, J., Schuller, B., Liu, Y., 2014. Modeling gender information for emotion recognition using denoising autoencoder. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 990–994.

Yao, T., Pan, Y., Ngo, C.-W., Li, H., Mei, T., 2015. Semi-supervised domain adaptation with subspace learning for visual recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2142–2150.

Zheng, W., Xin, M., Wang, X., Wang, B., 2014. A novel speech emotion recognition method via incomplete sparse least square regression. IEEE Signal Process. Lett. 21 (5), 569–572.

Zheng, W., Yu, J., Zou, Y., 2015. An experimental study of speech emotion recognition based on deep convolutional neural networks. In: International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp. 827–831.