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

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/© 2016 Elsevier B.V. All rights reserved.

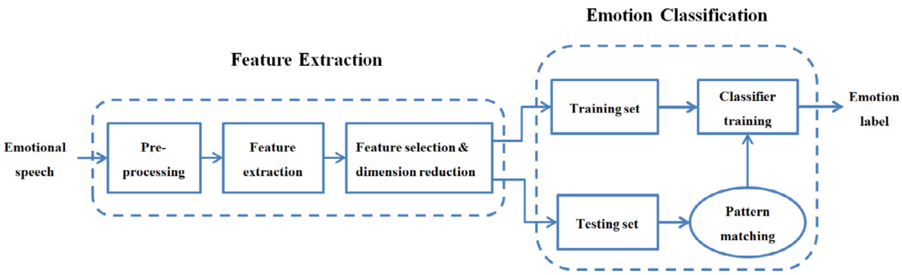

Fig. 1. Flowchart of speech emotion recognition.

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