Transfer Sparse Discriminant Subspace Learning for Cross-corpus Speech Emotion Recognition

Weijian Zhang, Peng Song*, *Member, IEEE*

Abstract—**Cross-corpus speech emotion recognition has attracted much attention due to the widespread existence of various emotional speech in life. It takes one corpus for training and another corpus for testing, and generally involves the following two basic problems: the corpus-invariant feature representation and relevance across different corpora. To deal with these two problems, we propose a novel transfer learning method called transfer sparse discriminant subspace learning (TSDSL) in this paper. Specifically, to solve the first problem, we learn a common feature subspace of different corpora by introducing the discriminative learning and $\ell_{2,1}-$norm penalty, which can learn the most discriminative features across different corpora. To address the second problem, we construct a novel nearest neighbor graph as the distance metric, in which the similarity between different corpora can be measured simultaneously. Extensive experiments are carried out on cross-corpus speech emotion recognition tasks, and the results show that our method can achieve competitive performance compared with state-of-the-art algorithms.**

Index Terms—**Speech emotion recognition, subspace learning, transfer learning, linear discriminant analysis.**

I. Introduction

SPEECH emotion recognition, as a vital branch of affective computing, has become a hot research topic due to its wide applications, such as human interaction, online education, intelligent tutoring systems and mental health diagnosis [1], [2], [3], [4]. The main purpose of speech emotion recognition is to recognize emotions from speech into ***one of the predefined emotion categories, e.g., anger, disgust, fear, happiness, neutral, sadness and surprise.***

In speech emotion recognition, an important problem is extracting the features that can efficiently express the emotional content of speech. A large number of speech features have been explored for speech emotion recognition, which can be grouped into four categories: continuous features, qualitative features, spectral features, and TEO (Teager energy operator)-based features [2]. Some of the most commonly used features in speech emotion recognition include fundamental frequency, energy, Mel-frequency cepstral coefficient (MFCC), and linear prediction cepstral coefficients (LPCC), perceptual linear prediction (PLP) [2]. In addition, with the rapid development of deep learning techniques [5], many studies have been made to extract the high level features from the raw speech or low-level acoustic features for emotion classification. In [6], [7], the convolutional neural networks (CNN) model is used to learn affect-salient high-level features. In [8], Lee et al. introduce a bi-directional long short-term memory (BLSTM) model to extract high-level emotional representations. In [9], the representation learning is performed on the variable length spectrograms in an end-to-end manner.

Another problem in speech emotion recognition is emotion classification. In recent years, many algorithms popular in pattern recognition field have been proposed for speech emotion classification, e.g., support vector machine (SVM) [10], hidden Markov model (HMM) [11], neural networks (NN) [12], Gaussian mixture model (GMM) [13], sparse representation [14], deep neural networks (DNN) [15]. These algorithms can perform well in traditional speech emotion recognition, where the training and testing are conducted on the same corpus. However, in reality, the speech data are often collected in different scenarios such as different languages, noises and speakers, in which the training and testing speech data will follow different distributions. The recognition performance of these algorithms will drop significantly [16].

To solve the cross-domain problem mentioned above, many transfer learning algorithms have been developed in recent years [17], where the rich labeled data in source domain are used to classify the unlabeled data in the target domain. For example, transfer component analysis (TCA) is a popular transfer learning method [17], ***which tries to learn some transfer components across domains in the reproducing kernel Hilbert space (RKHS)*** by using maximum mean discrepancy (MMD) [18]. In [19], a new transfer learning framework is presented, where the source and target domains are represented by a common subspace described by eigenvector matrices [19]. In [20], Li et al. propose a novel unsupervised multi-source domain adaptation approach, named as coupled local-global adaptation (CLGA) [20]. In [21], Wang et al. present a manifold embedded distribution alignment (MEDA) approach, which attempts to perform dynamic distribution alignment for manifold domain adaptation.

With the development of transfer learning techniques, many researches have been focused on developing new transfer learning algorithms to deal with the cross-corpus speech emotion recognition problem in the past few years [22], [16]. In [23], Hassan et al. introduce several transfer learning algorithms, e.g., kernel mean matching (KMM), unconstrained least-squares importance fitting (uLSIF) and Kullback-Leibler importance estimation procedure (KLIEP), to compensate the channel and speaker differences. In [24], Zong et al. propose a domain-adaptive least-squares regression (DaLSR) model for cross-corpus speech emotion recognition. In [25], [26], Deng et al. present an autoencoder based transfer learning

W. Zhang is with the School of Computer and Control Engineering, Yantai University, Yantai 264005, China. (e-mail: wjzhang231@foxmail.com)

P. Song is with the School of Computer and Control Engineering, Yantai University, Yantai 264005, China. (e-mail: pengsong@ytu.edu.cn)

*Corresponding author.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2955252, IEEE/ACM Transactions on Audio, Speech, and Language Processing

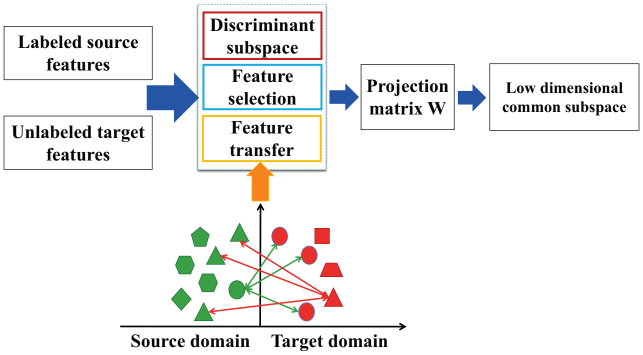


Fig. 1: The basic idea of our proposed TSDSL method. In the neighbor graph, different colors represents different domains, and different shapes represent different labels. For each sample in one domain, we can find its neighbors in another domain.

method for speech emotion recognition. In [27], Huang et al. introduce a new feature transfer approach with PCANet to facilitate the performance improvement of speech emotion recognition. In [16], we have presented a unified linear transfer subspace based learning framework for robust speech emotion recognition. These algorithms employ the MMD as a distance metric to measure the difference between different domains, and can achieve promising performance. However, MMD suffers from a major limitation, i.e., the non-linear kernel mapping used to calculate the distribution mismatch hinders the applicability of MMD, where kernel functions may not be optimal for kernel learning machines [28]. Another critical disadvantage of these existing methods is that most of them are unsupervised, which are prone to learn irrelevant feature representations, and are unfavorable to the following classifier learning.

In this paper, inspired by recent development of manifold learning and subspace learning [29], we propose a novel cross-corpus speech emotion recognition method, called transfer sparse discriminant subspace learning (TSDSL), which aims to transfer the knowledge from the source corpus to the target corpus. Different from the abovementioned transfer algorithms, we construct a novel neighbor graph to measure the ***feature distribution difference*** across source and target datasets, which can also simultaneously preserve the local geometrical structure of the data. Obtaining the desired subspace is of vital importance for our model, Hence, we introduce a discriminative term to expand the spacing between different classes and reduce the spacing in the same class, which can maintain the global structure of the data. In addition, we apply a $\ell_{2,1}-$norm to constrain the projection matrix, which aims to select the most discriminative features. Furthermore, to hold the main energy of the original source and target data by an orthogonal reconstruction matrix, we introduce a new form of representation of PCA into our framework. Finally, we can map the features from different corpora into a common discriminative subspace. The basic idea of our method is illustrated in Figure 1.

It is worthwhile to highlight the main contributions of this work, which are summarized as follows.

- The proposed TSDSL elegantly combines discriminative subspace learning, graph based distance metric and feature selection into a joint framework. Experimental results on cross-corpus speech emotion recognition tasks show its superiority.
- We construct a novel neighbor graph to minimize the ***feature distribution divergence*** between different corpora, which can efficiently transfer the knowledge gained from the labeled source corpus to the unlabeled target corpus.
- We aim to learn a corpus-invariant projection simultaneously while knowledge transfer, which can align the features from different corpora, into a new common subspace.
- TSDSL can simultaneously perserve the global and local information of data when learning projection matrix, which guarantees a more discriminative subspace to boost the recognition performance.

The rest of this paper is organized as follows. In Section II, we give a brief review of related works. Section III presents the proposed TSDSL method and optimizing algorithm, respectively. In Section IV, we conduct a series of experiments on three public emotional datasets and discuss the experimental results. Finally, we draw conclusion of our paper in Section V.

II. Related work

In this section, we briefly review the related works, i.e., subspace learning and transfer learning, and further highlight the difference between the existing works and our proposed algorithm.

A. Subspace learning

During past decades, subspace learning has been shown an efficient technique for many pattern classification problems [30], [29], and has also been shown very successful for speech emotion recognition **[31], [32], [33]**. It attempts to project the original data from high dimensional space into a lower dimensional space, where the important properties can be well preserved [30]. ***Representative*** subspace learning algorithms include principal component analysis (PCA), linear discriminant analysis (LDA), locality preserving projection (LPP) [30]. PCA aims to project the data points along the directions of maximal variances [34]. LDA [35] employs the label information to increase the between-class distance and reduce the within-class distance. LPP [36] preserves local relationships within the data and clue its essential manifold structure. In addition, locally linear embedding (LLE) [37], ISOMAP [38], laplacian eigenmaps (LE) [39] are also commonly used manifold algorithms. Most of these algorithms can be interpreted as a general graph embedding framework [29].

Subspace learning methods have been extensively used in the context of speech emotion recognition [31], [2]. In [31], You et al. develop a nonlinear manifold algorithm for speech emotion recognition. In [32], Gangeh et al. propose a multiview subspace learning to recognize four dimensional affects, i.e., arousal, expectation, power, and valence, from speech signals.

In [33], Xu et al. present a spectral regression-based subspace learning for speech emotion recognition. These algorithms can obtain good recognition performance. ***However, in practice, the training and testing data are often from different corpora, which may follow different feature distributions, and these subspace learning algorithms cannot be directly employed.*** Thus, in this work, to cope with the cross-corpus speech emotion recognition problem, we develop a novel transfer learning framework for efficient subspace learning.

B. Transfer learning

More recently, a number of transfer subspace learning methods have been proposed to solve cross-domain pattern recognition problems. As far as we know, Si et al. first present a transfer subspace learning algorithm, in which the Bregman divergence is employed to reduce the difference across source and target domains in the selected subspace [40]. Yang et al. have presented a transfer sparse subspace learning framework by introducing an effective sparse regularization, which can reduce time and over-fitting problem [41]. In [42], Shao et al. propose a framework to solve the knowledge transfer problem via a low-rank representation constraint [42], in which the structures of source and target data can be well preserved. To cope with the cross-dataset facial expression recognition problem, in [43], a transfer subspace learning method is proposed to learn a common subspace, which transfers knowledge obtained from the source domain to the target domain to improve the facial expression recognition performance. In [44], Razzaghi et al. obtain a common subspace from the source and target domains via low-rank and reconstruction matrix. However, to the best of our knowledge, these abovementioned transfer subspace learning methods have the following disadvantages: a) they do not pay much attention to the discriminative learning, which utilizes the information of data-class labels that are useful for supervised subspace learning [45]; b) they often use MMD for distance measurement, in which the kernel functions may not be optimal for kernel learning machines [28]; c) ***they neglect the redundancy observed in the high-dimensional features*** [46]. To overcome these problems, in this paper, we present a new transfer sparse discriminant subspace learning method, which aims to obtain a common discriminative subspace by combining discriminative subspace learning, graph based distance metric and feature selection into a joint framework.

Nowadays, deep neural network based transfer learning methods achieve state-of-the-art results ***[47], [48], [49], [50], [51]***. These methods can be roughly divided into two categories, i.e., discrepancy based methods and adversarial based methods ***[51], and have been successfully applied to speech emotion recognition tasks***. For example, in [47], Gideon et al. introduce the progressive neural networks to investigate how knowledge can be transferred between different emotional datasets. In [48], Deng et al. present semisupervised autoencoders to improve the recognition of cross-corpus speech emotion recognition. In [49], Abdelwahaba et al. propose a novel framework based on domain-adversarial neural network (DANN) for cross-corpus speech emotion recognition. In [50], instead of learning to transfer between different datasets, Lu et al. develop a meaningful speech front-end network to cope with the different emotion contexts, e.g., languages, domains. Generally, the deep models show better performance than conventional shallow ones. Note that in deep transfer learning, the deep models are either used as feature extractor or in an end-to-end fashion (i.e., the transfer learning module is integrated into the deep model). However, the deep models may have the following disadvantages: ***a) it is still unclear which deep learning method would perform better***; b) the computational cost of deep models is much higher than that of many shallow based models; c) some of the shortcomings of GAN, e.g., hard training and unclear stop criterion, still exist in adversarial based transfer learning algorithms. In practice, many deep and shallow based models use the same distance ***metrics***, e.g., MMD, Wasserstein distance, to measure the ***distribution divergence between*** different corpora [52], [53], [54]. ***However, these distance metric algorithms ignore the data local structure, which might cause the negative transfer to some extent***. Thus, in this work, we focus on developing the new distance metric under the transfer subspace learning framework.

III. The proposed method

In this section, first, we introduce the preliminary knowledge. Second, we present the transfer sparse discriminant subspace learning (TSDSL) for cross-corpus speech emotion recognition. Third, we describe the optimization of TSDSL algorithm in detail.

A. Preliminary

We begin with a brief introduction of notations used here, which are listed in Table I, Given n_s labeled source samples $X = [x_1, x_2, \ldots, x_{n_s}] \in R^{m \times n_s}$ and n_t unlabeled target samples $Y = [y_1, y_2, \ldots, y_{n_t}] \in R^{m \times n_t}$, we aim to learn a common feature subspace by the projection matrix $W \in R^{m \times d} (d < m)$, which can project the features from the source and target datasets into a common subspace.

TABLE I: Notations and descriptions used in this paper.

Notation	Description
$X \in R^{m \times n_s}, Y \in R^{m \times n_t}$	Labeled source/unlabeled target features
$S_b \in R^{m \times m}, S_w \in R^{m \times m}$	Inter-class/intra-class scatter matrices
β	The controlling parameter of S_b
$W \in R^{m \times d}, P \in R^{m \times d}$	Projection/orthogonal matrices
k	Number of the nearest neighbors
$D, G_1, G_2 \in R^{m \times m}$	Diagonal matrices
λ	Feature selection regularization parameter
γ_1, γ_2	Graph regularization parameters
α	Energy preserving regularization parameter
Be,Ba,e	The Berlin/Baum-1a/eNTERFACE dataset

For a matrix W, its ℓ_1-norm and Frobenius norm are defined as $||W||_1 = \sum_{j=1}^{d}\sum_{i=1}^{m}|w_{ij}|$, $||W||_F = \sqrt{\sum_{j=1}^{d}\sum_{i=1}^{m}w_{ij}^2}$, respectively. Let $||\cdot||_{2,1}$ denote the $\ell_{2,1}-$norm, which is calculated as

$$\left\|W\right\|_{2,1} = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{d}w_{ij}^2} \tag{1}$$

The $\ell_{2,1}-$norm of the matrix is first introduced as rotational invariant ℓ_1-norm in [55]. Here we give a brief explanation of $\ell_{2,1}-$norm. First we compute the ℓ_2-norm of the row vector w_i, the $i-$th row of data can be represented by $||w_i,\cdot||_2$. If the $i-$th row of W is equal to zero, the features corresponding to the $i-$th row can be considered as the unimportant or redundant features [56]. Then we can use the ℓ_1-norm constraint to compute $W = [||w_1,\cdot||_2, ||w_2,\cdot||_2, \ldots, ||w_m,\cdot||_2]^T$ to remove the unimportant or redundant features. The following constraint problem $||W||_1 = d$ indicates that we can select the d most discriminative features from m features.

1) ***Linear discriminant analysis: LDA aims to seek the most effective discriminative direction via maximizing the ratio between the inter-class and intra-class scatter matrices, and the objective function is written as follows:***

$$\arg\max_W \frac{W^TS_bW}{W^TS_wW} \tag{2}$$

where S_b ***and*** S_w ***are the inter-class and intra-class scatter matrices, respectively. Suppose we have*** c ***classes.*** $n = \sum_{i=1}^{c}n_i$ ***is the total number of all samples.*** n_i ***is the number of the*** $i-$***th class.*** S_b ***and*** S_w ***can be obtained using the following formula:***

$$S_b = \sum_{i=1}^{c}n_i(\mu^{(i)}-\mu)(\mu^{(i)}-\mu)^T \tag{3}$$

$$S_w = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_j^{(i)}-\mu^{(i)})(x_j^{(i)}-\mu^{(i)})^T \tag{4}$$

where μ ***is the overall mean vector,*** $\mu^{(i)}$ ***is the average vector of the*** $i-$***th class,*** $x_j^{(i)}$ ***is the*** $j-$***th sample in the*** $i-$***th class.***

The goal of LDA is to maximize the inter-class scatter and minimize the intra-class scatter. As shown in the graph embedding framework [29], LDA can be converted into the following trace ratio problem:

$$\arg\max_W \frac{Tr(W^TS_bW)}{Tr(W^TS_wW)} \tag{5}$$

where $Tr(\cdot)$ ***denotes the trace of a matrix. According to [57],*** (5) ***can be further expressed as the following ratio trace problem:***

$$\min_W Tr(W^T(S_w-\beta S_b)W) \tag{6}$$

where β ***is a small positive constant to balance the importance of*** S_b ***and*** S_w***.***

B. Objective function

Obtaining a good performance in cross-corpus speech emotion recognition tasks always involves two basic problems, i.e., a good feature representation and a good distance metric across different corpora. To tackle these two problems, in this work, we present a TSDSL method to learn approximate feature representations in a latent subspace. The objective function of TSDSL is expressed as

$$\min_W F(W) + \gamma R(W) \tag{7}$$

where $F(W)$ is the objective function of subspace learning, the function $R(W)$ ***is a distance metric, which is used to measure the feature distribution divergence between different corpora,*** and $\gamma > 0$ is a regularization parameter.

1) Sparse discriminant subspace learning: To address the first problem, we aim to learn a common discriminative feature subspace between different corpora. Over the past decades, conventional supervised subspace learning methods, e.g., LDA, marginal Fisher analysis (MFA) [29], local discriminant embedding [58], discriminative locality alignment (DLA) [59], have been successfully applied to extract the discriminative information from the original data. Among which, LDA is one of the most popular algorithms.

In reality, the dimensionality of emotional features is often very high, which has redundant and irrelevant features. Thus, how to select the relevant and discriminative features is very important. Accordingly, we simultaneously perform the $\ell_{2,1}-$norm regularization on the projection matrix. which can make the learned projection have better interpretability for features by its good row-sparsity property [60]. That is, we can obtain $F(W)$ with the following form:

$$F(W) = Tr(W^T(S_w-\beta S_b)W) + \lambda\left\|W\right\|_{2,1} \tag{8}$$

where $\lambda > 0$ is a regularization parameter.

2) Graph based distance metric: To address the second problem, we need to develop an efficient distance metric to measure the ***feature distribution divergence*** between different corpora. In this work, we focus on dealing with the cross-corpus recognition problem. Hence, how to transfer the information from source domain to target domain is the focus of our attention. To achieve this goal, enlightened by the LLE algorithm [37], in which each data point is reconstructed from its neighbors, we present a novel discrepancy metric to measure the difference between training and testing corpora by a nearest neighbor graph. We specify the graph as the following optimization function:

$$\min_W \sum_{i=1}^{n_s}\left\|W^Tx_i - W^T\sum_{j=1}^{k}p_{ij}y_{ij}\right\|^2 + \sum_{i=1}^{n_t}\left\|W^Ty_i - W^T\sum_{j=1}^{k}q_{ij}x_{ij}\right\|^2 \tag{9}$$

where $y_{i1}, y_{i2}, \ldots, y_{ik}$ are the k-nearest neighbors of x_i, which means that we can find k neighbors of each source data from the target data, in the same way, $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the k-nearest neighbors of y_i, which means that we can find k neighbors of each target data from the source data, and p_{ij} and q_{ij} are the corresponding local reconstruction coefficients. ***First, we find the nearest neighbors of each data point. Then, by using the algorithm of computing weights in LLE [37], we can estimate the values of*** p_{ij} ***and*** q_{ij}***, respectively.*** Eq. (9) means that the samples in one corpus can be reconstructed with a linear combination of samples in another corpus, which is simplified as the following form:

$$R(W) = Tr(W^TG_1W) + Tr(W^TG_2W) \tag{10}$$

where $G_1 = \sum_{i=1}^{n_s}(x_i - \sum_{j=1}^{k} p_{ij}y_{ij})(x_i - \sum_{j=1}^{k} p_{ij}y_{ij})^T$, $G_2 = \sum_{i=1}^{n_t}(y_i - \sum_{j=1}^{k} q_{ij}x_{ij})(y_i - \sum_{j=1}^{k} q_{ij}x_{ij})^T$.

3) The objective of TSDSL: Since the graphs G_1 and G_2 play different roles in the objective function, here we set two different weighting parameters γ_1 and γ_2 for the graphs, respectively. Moreover, to make the minimization problem with respect to W well-posed, we impose an orthogonal constraint $W^TW = I$, which can make the problem tractable. Then, by combining Eq. (8) and Eq. (10) into a joint learning framework, we can obtain the following optimization problem:

$$\begin{aligned}&\min_{W} Tr(W^T(S_w - \beta S_b)W) + \lambda\left\|W\right\|_{2,1} + \gamma_1 Tr(W^T G_1 W)\\ &\qquad +\gamma_2 Tr(W^T G_2 W)\\ &\text{s.t. } W^TW = I\end{aligned} \tag{11}$$

By solving Eq. (11), we can select d eigenvectors by using the projection matrix, where $d \leq c - 1$ (c is the number class). However, these d eigenvectors cannot preserve enough discriminative information since the dimensionality of original features is very high [61]. In addition, problem (11) is sensitive to the selection of reduced dimensions. To address these problems, inspired by the PCA plus LDA technique in face recognition [62], [61], we introduce a variant of PCA, which can hold the main energy of data, into our optimization function, and Eq. (11) becomes as

$$\begin{aligned}&\min_{W,P} Tr(W^T(S_w - \beta S_b)W) + \lambda\left\|W\right\|_{2,1} + \gamma_1 Tr(W^T G_1 W)\\ &\qquad +\gamma_2 Tr(W^T G_2 W) + \alpha\left\|X - PW^TX\right\|_F^2\\ &\text{s.t. } W^TW = I, \quad P^TP = I\end{aligned} \tag{12}$$

where $P \in R^{m\times d}$ is an orthogonal reconstruction matrix, the constraints $\left\|X - PW^TX\right\|_F^2$ and $P^TP = I$ can be seen as a variant of PCA to some extent, α is a regularization parameter, which indicates the significance of the corresponding term. In this way, we can make low dimensional data hold the main energy of the original source and target data as much as possible. Thus, our joint framework can select the number of dimension more flexibly and minimize the loss of information of data. Finally, we can obtain a better discriminative common subspace by Eq. (12).

C. Optimization

The objective function (12) involves the $\ell_{2,1}-$norm, which is non-smooth and hard to optimize directly [60]. Consequently, we propose an iterative optimization algorithm. According to [60], $\left\|W\right\|_{2,1}$ can be expressed as

$$\left\|W\right\|_{2,1} = 2Tr(W^TDW) \tag{13}$$

where $D = [d_{ii}] \in R^{m\times m}$ is a diagonal matrix, and d_{ii} can be represented as $d_{ii} = \frac{1}{2\|w^i\|_2}$, in which w^i means the $i-$th row of W. Note that $\left\|w^i\right\|_2$ may be close to zero [63], which will make the objective function non-differentiable. To overcome this problem, as [63], we rewrite d_{ii} as

$$d_{ii} = \frac{1}{2\sqrt{\left\|w^i\right\|_2^2 + \upsilon}} \tag{14}$$

where υ is a term with small constant.

To facilitate the optimization, by defining $G = \lambda D + \gamma_1 G_1 + \gamma_2 G_2$, we can rewrite the problem (12) as minimizing the following problem:

$$\begin{aligned}&O = Tr(W^T(S_w - \beta S_b)W) + Tr(W^TGW) + \alpha\left\|X - PW^TX\right\|_F^2\\ &\text{s.t. } W^TW = I, \quad P^TP = I\end{aligned} \tag{15}$$

In the following, we introduce the optimization steps in brief as follows:

- **Update** W: Fix P and update W by minimizing the Eq. (15), we employ the derivative of O w.r.t. W, then we can have

  $$\frac{\partial O}{\partial W} = (S_w - \beta S_b)W + GW + \alpha(XX^TW - XX^TP) + \phi W \tag{16}$$

  where ϕ is a Lagrange parameter. Note that D is also unknown and depends on W, thus, we develop an iterative algorithm to compute D and W. With fixed D, by setting Eq. (16) to zero, we can obtain

  $$W = \alpha(S_w - \beta S_b + G + \phi + \alpha XX^T)^{-1}XX^TP \tag{17}$$

  Then, with fixed W, D can be computed according to Eq. (14).
- **Update** P: Fix W and update P, we will obtain the following optimization problem:

  $$\min_{P^TP=I} \left\|X - PW^TX\right\|_F^2 \tag{18}$$

  Eq. (18) can be further expressed as follows:

  $$\min_{P^TP=I} \left\|X - PW^TX\right\|_F^2 = \min_{P^TP=I} Tr(X^TX - 2X^TPW^TX) \tag{19}$$

  Hence, we need to maximize the $Tr(X^TPW^TX)$. This problem is an orthogonal Procrustes problem [64]. We should realize that $Tr(X^TPW^TX) = Tr(P^TXX^TW)$, and with the operation of singular value decomposition (SVD): $\text{SVD}(XX^TW) = USV^T$, it becomes as

  $$Tr(P^TXX^TW) = Tr(P^TUSV^T) = Tr(V^TP^TUS) \tag{20}$$

  It is clear that the maximum is achieved by setting $V^TP^TU = I$, so we can obtain $P = UV^T$ [65].

Therefore, the objective function (12) can be solved by updating the above steps iteratively until convergence or the maximum number of iterations has been reached. The details of our algorithm are summarized in Algorithm 1.

IV. Experiments and results

To evaluate the effectiveness of our proposed TSDSL, we conduct extensive cross-corpus speech emotion recognition experiments with several state-of-the-art methods on three benchmark datasets, i.e., Berlin, eNTERFACE and Baum-1a. The statistics of each dataset are summarized in Table II.

Algorithm 1 The proposed TSDSL algorithm

```
Input: The feature matrix of labeled source corpus and unlabeled target corpus $X \in R^{m \times n_s}, Y \in R^{m \times n_t}$, regularization parameter $\beta$, $\lambda$, $\gamma_1$, $\gamma_2$, $\alpha$, the number of nearest neighbors $k$.
Output: The projection matrix $W \in R^{m \times d}$
  a). Initialize $D \in R^{m \times m}$ as $D = I$;
  b). Initialize $P = \arg\min_{P} Tr(P^T(S_w - \mu S_b)P) \quad$ s.t. $P^T P = I$;
  c). Calculate $S_w$ and $S_b$;
  d). Construct the $k$ nearest neighbor graphs to calculate $G_1$ and $G_2$;
  repeat
    1.With $P$, the optimal solution $W$ by solving Eq. (17);
    2.With current $W$, $P$ is obtained by Eq. (19);
  until Convergence criterion satisfied.
```

A. Experimental settings

1) Data preparation: The first dataset is Berlin [66], which is one of the most popular speech emotion datasets. It consists of 494 speech utterances in German, and the utterances are recorded by 10 actors in seven kinds of basic expressions including anger, boredom, disgust, fear, happiness, neutral and sadness.

The second dataset is eNTERFACE [67], which is a public audio-visual emotion dataset. It contains 1287 video samples from 43 subjects in English. Six types of emotions are collected, i.e., anger, disgust, fear, happiness, sadness and surprise.

The last dataset is Baum-1a [68], which is a collection of audio-visual acted and spontaneous affective expressions. The audio-visual clips are recorded by 31 subjects, who express a rich set of emotional and mental states in Turkish. It contains eight basic categories, i.e., anger, boredom, disgust, fear, interest, happiness, sadness and unsure.

We design six types of cross-corpus speech emotion recognition as follows:

- Be-e: Berlin is the labeled training dataset and eNTERFACE is the unlabeled testing dataset.
- e-Be: eNTERFACE is the labeled training dataset and Berlin is the unlabeled testing dataset.
- Be-Ba: Berlin is the labeled training dataset and Baum-1a is the unlabeled testing dataset.
- Ba-Be: Baum-1a is the labeled training dataset and Berlin is the unlabeled testing dataset.
- e-Ba: eNTERFACE is the labeled training dataset and Baum-1a is the unlabeled testing dataset.
- Ba-e: Baum-1a is the labeled training dataset and eNTERFACE is the unlabeled testing dataset.

In all cases, we choose five common emotion categories of these databases, i.e., anger, disgust, fear, happiness and sadness, for evaluation. In our experiments, each corpus is divided into 10 parts. Among which all the source dataset and 7/10 of the target dataset are used for training, while the others are used for testing.

TABLE II: The statistics of databases.

Database	Language	Size	Category
Berlin	German	494	Seven
eNTERFACE	English	1287	Six
Baum-1a	Turkish	264	Eight

2) ***Baseline algorithms:*** To verify the efficacy of our method for cross-corpus speech emotion recognition, we compare it with a number of related state-of-the-art (transfer) subspace learning methods, which are listed as follows:

- Transfer component analysis (TCA) [17]: It tries to learn some transfer components across domains in the reproducing kernel Hilbert space (RKHS).
- Transfer linear discriminant analysis (TLDA) [46]: It combines LDA and MMD into a joint framework.
- Joint distribution adaptation (JDA) [69]: It aims to jointly adapt both the marginal distribution and conditional distribution in a principled dimensionality.
- Transfer joint matching (TJM) [70]: It tries to jointly perform feature matching and instance reweighting across domains in a principled dimensionality reduction procedure.
- Semi-supervised discriminant analysis (SDA) [71]: It aims to find a projection subspace which respects the discriminant structure inferred from the labeled data, and the intrinsic geometrical structure inferred from both labeled and unlabeled data.
- Linear discriminant analysis (LDA) [29].
- Principal component analysis (PCA) [30].

3) Implementation details: For our experiments we employ the open source openSMILE toolkit [1] to extract the efficient emotional features. We adopt the feature set of INTERSPEECH 2010 Paralinguistic challenge [72], which contains 1582 dimensional features. It includes 34 basic low-level descriptors (LLDs), i.e., Mel-frequency cepstral coefficient (MFCC), line spectrum pair (LSP), loudness and 34 corresponding delta coefficients. Based on these LLDs, 21 statistical functions are applied to obtain 1428 features. In addition, 19 statistical functions are applied to the 4 pitch-based LLDs and related corresponding delta coefficients to obtain 152 features. Finally, the onset of pitch and durations of utterances are added into the last two features. The descriptions of LLDs used are listed in Table III.

Under our experimental setup, we choose the linear SVM as classifier due to its efficiency in speech emotion classification. It is trained on the labeled source dataset and tested on the unlabeled target dataset. Since the labeled and unlabeled data follow different distributions, it is impossible to tune the parameters using the cross-validation strategy [70]. Thus, we evaluate our method by regulating parameters space to find the optimal parameters, then we can obtain the best result of each experimental case. There are six main hyper-parameters in our experiment. The value of β is empirically set to 0.1. ***We set the number of nearest neighbors*** k ***by searching the range*** $3 \sim 9$. The weighting parameters λ and α are set

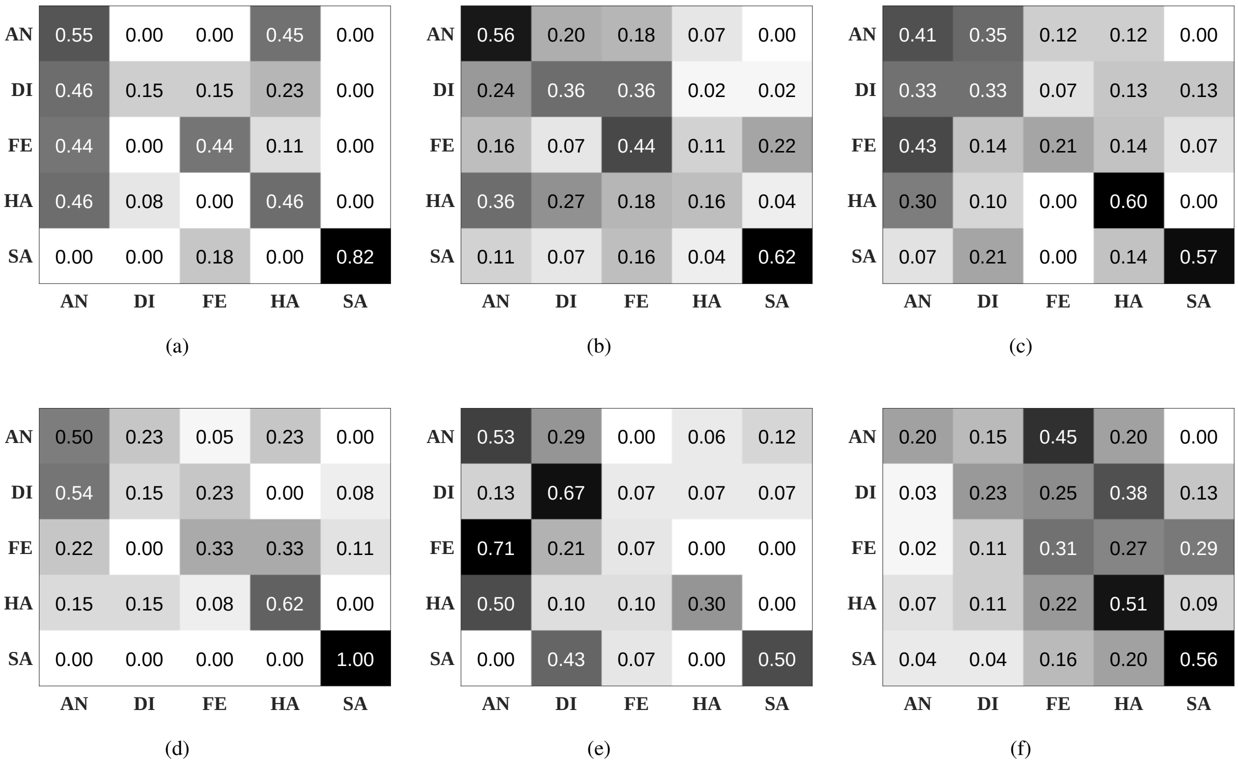
[1] https://www.audeering.com/opensmile/

TABLE IV: Recognition accuracy (%) with standard deviation of different methods under different cases.

Cases	Subspace learning			Transfer learning				TSDSL
	PCA	LDA	SDA	TCA	JDA	TLDA	TJM	
e-Be	30.88 ±0.69	32.35 ±0.20	35.29 ±1.46	36.03 ±2.61	40.59 ±1.72	43.53 ±0.76	45.40 ±0.28	**47.35** ±1.16
Be-e	38.08 ±1.06	38.00 ±0.23	39.11 ±1.75	38.89 ±1.11	41.56 ±0.79	39.82 ±1.17	42.18 ±0.39	**42.44** ±0.52
Be-Ba	33.86 ±1.35	34.42 ±1.05	34.29 ±2.02	33.58 ±1.05	34.00 ±0.90	36.00 ±1.31	37.28 ±1.31	**42.42** ±1.35
Ba-Be	30.59 ±1.35	32.64 ±0.62	36.77 ±1.36	44.11 ±0.83	45.58 ±1.72	41.17 ±0.95	45.01 ±0.14	**49.26** ±2.60
e-Ba	29.71 ±1.76	30.85 ±2.04	31.43 ±1.22	38.43 ±1.42	38.57 ±1.90	38.57 ±1.03	36.42 ±0.75	**42.14** ±0.75
Ba-e	29.30 ±1.49	30.56 ±1.89	33.49 ±1.37	31.02 ±1.42	31.90 ±1.03	33.95 ±0.82	35.85 ±0.45	**36.46** ±0.85
Average	32.07	33.14	35.06	37.01	38.70	38.84	40.36	**43.34**

TABLE V: ***F1-score (%) of different methods under different cases.***

Cases	Subspace learning			Transfer learning				TSDSL
	PCA	LDA	SDA	TCA	JDA	TLDA	TJM	
e-Be	34.02	38.76	35.07	39.49	47.18	45.82	46.10	**52.77**
Be-e	39.12	40.51	42.32	42.59	42.44	42.06	43.05	**43.06**
Be-Ba	33.04	35.21	38.84	33.39	35.13	37.56	35.81	**43.95**
Ba-Be	38.19	43.67	37.53	49.56	52.39	46.60	48.68	**52.40**
e-Ba	26.69	30.24	32.92	38.44	40.85	41.65	33.06	**43.52**
Ba-e	29.25	30.49	27.31	31.14	32.51	33.72	**39.16**	36.31
Average	33.38	36.48	35.66	39.10	41.75	41.23	40.98	**45.33**



	AN	DI	FE	HA	SA
AN	0.55	0.00	0.00	0.45	0.00
DI	0.46	0.15	0.15	0.23	0.00
FE	0.44	0.00	0.44	0.11	0.00
HA	0.46	0.08	0.00	0.46	0.00
SA	0.00	0.00	0.18	0.00	0.82

(a)

	AN	DI	FE	HA	SA
AN	0.56	0.20	0.18	0.07	0.00
DI	0.24	0.36	0.36	0.02	0.02
FE	0.16	0.07	0.44	0.11	0.22
HA	0.36	0.27	0.18	0.16	0.04
SA	0.11	0.07	0.16	0.04	0.62

(b)

	AN	DI	FE	HA	SA
AN	0.41	0.35	0.12	0.12	0.00
DI	0.33	0.33	0.07	0.13	0.13
FE	0.43	0.14	0.21	0.14	0.07
HA	0.30	0.10	0.00	0.60	0.00
SA	0.07	0.21	0.00	0.14	0.57

(c)

	AN	DI	FE	HA	SA
AN	0.50	0.23	0.05	0.23	0.00
DI	0.54	0.15	0.23	0.00	0.08
FE	0.22	0.00	0.33	0.33	0.11
HA	0.15	0.15	0.08	0.62	0.00
SA	0.00	0.00	0.00	0.00	1.00

(d)

	AN	DI	FE	HA	SA
AN	0.53	0.29	0.00	0.06	0.12
DI	0.13	0.67	0.07	0.07	0.07
FE	0.71	0.21	0.07	0.00	0.00
HA	0.50	0.10	0.10	0.30	0.00
SA	0.00	0.43	0.07	0.00	0.50

(e)

	AN	DI	FE	HA	SA
AN	0.20	0.15	0.45	0.20	0.00
DI	0.03	0.23	0.25	0.38	0.13
FE	0.02	0.11	0.31	0.27	0.29
HA	0.07	0.11	0.22	0.51	0.09
SA	0.04	0.04	0.16	0.20	0.56

(f)

Fig. 2: Confusion matrices of our method for cross-corpus speech emotion recognition: (a) The confusion matrix under the e-Be setting; (b) The confusion matrix under the Be-e setting; (c) The confusion matrix under the Be-Ba setting; (d) The confusion matrix under the Ba-Be setting; (e) The confusion matrix under the e-Ba setting; (f) The confusion matrix under the Ba-e setting (AN: anger, DI: disgust, FE: fear, HA: happiness, SA: sadness).

TABLE III: The LLDs used in this paper.

Descriptors	Number of features
PCM loudness	42
MFCC [0-14]	630
Log mel freq band [0-7]	336
LSP frequency [0-7]	336
F0 envelope	42
Voicing prob.	42
F0	38
Jitter local	38
Jitter consec. frame pairs	38
Shimmer local	38
F0 number of onsets	1
Turn duration	1

by searching from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, while for graph regularization parameters γ_1 and γ_2, the optimal values are chosen from $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$. For our TSDSL and PCA related algorithms, 98% energy is kept. ***We set the values of hyper-parameters of the baseline transfer learning methods by searching from*** $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$**.** We set different dimensions of reduced feature subspace due to their different solutions. For PCA, TCA and JDA, the optimal dimension is set as 100, for LDA, SDA and TLDA, the optimal reduced feature dimension is set as 4, while for TJM and our proposed TSDSL, the reduced feature dimension is set as 25. ***The experiments are repeated 10 trials to cover the possible cases of training set and testing set. After the experiments, we use the average recognition accuracies and F1-score values over the 10 trials to evaluate the emotion classification performance. We use the t-test for evaluation, and assert the statistical significance at p-value*** ≤ 0.05**.**

B. Results analysis

1) cross-corpus speech emotion recognition ***results****:* In this subsection, we compare our method with other transfer learning or subspace learning methods in terms of recognition accuracy and ***F1-score*** of cross-corpus speech emotion recognition. The recognition results are described in the Table IV ***and Table V***.

Firstly, from ***Table IV***, we can clearly observe that our TSDSL method obtains better performance than compared methods. The average classification accuracy of our method on the six settings is 43.34%, which gains a significant improvement of 2.98% compared with TJM. It can be also seen that our method can achieve the highest recognition rates on all experimental settings. ***From Table V, we can find that our TSDSL method achieves the highest average F1-score value, and outperforms the baseline methods in all settings except Ba-e.*** These results verify that TSDSL can obtain more effective common subspace for cross-corpus speech emotion recognition.

Secondly, we can observe that TSDSL and other transfer learning methods achieve better results than conventional subspace learning methods, i.e., PCA, LDA and SDA, in all experimental situations. A major limitation of these conventional subspace learning algorithms is that they assume that the training and testing samples are independent and identically distributed. In reality, these assumptions do not hold, which will result in poor recognition performance.

Thirdly, we can find that our method significantly outperforms other transfer subspace learning algorithms. The reasons might be that, on one hand, different from MMD similarity measurement used in these algorithms, we construct a new neighbor graph as distance metric, which can find similar features between source and target corpora and simultaneously maintain the local geometric structure. On the other hand, our proposed TSDSL provides a joint learning framework, which can achieve the most discriminative features for transfer learning.

Lastly, compared with conventional discriminant subspace learning algorithms, i.e., LDA, SDA, our proposed TSDSL can significantly achieve better recognition performance. This might be attributed to be that, TSDSL takes into account feature selection, graph regularization, energy preserving together, which can efficiently transfer the knowledge between training and testing data. In addition, we can notice that SDA performs better than LDA, which also verifies the effectiveness and importance of the graph to some extent.

The confusion matrices of our proposed method under all experimental settings are shown in Fig. 2. From the figure, we can see that the sadness expression obtains the highest recognition accuracy under the e-Be and Ba-Be settings. This result is coincided with that on single Berlin corpus [7]. From Figs. 2 (c) and (f), we can observe that the fear expression performs worse compared with the other expressions on Baum-1a, which coincides with the findings of Ref. [7], in which the fearness in Baum-1a dataset performs worst. Moreover, from these confusion matrices, we can also find that, in most cases, the sadness expression is much easier to be recognized than the other emotions. Overall, compared to the recognition performance on single corpus, the performance of cross-corpus recognition is still far from satisfactory. The reason might be that the divergence between these datasets, e.g., different languages, different recording styles (simultaneous versus spontaneous), is very large [73].

2) Effectiveness verification: In this section, we verify the effectiveness of our method by inspecting the influence of the regularization terms, i.e., feature selection, graph based distance metric and the effect of energy preserving. We consider the following three special cases of TSDSL:

- **TSDSL$_1$**: We remove the feature selection regularization term with $\lambda = 0$ in Eq. (12).
- **TSDSL$_2$**: We set the graph regularization parameters $\gamma_1 = 0$ and $\gamma_2 = 0$ in Eq. (12).
- **TSDSL$_3$**: We set the energy preserving regularization parameter α as 0 in Eq (12).

Fig. 3 shows the average recognition rates of each method. From the figure, we observe that if we set the graph regularization term as 0, TSDSL$_2$ performs much worse than other cases, which verifies that the nearest neighbor graph plays a important role in our method. We can also find that TSDSL performs better than TSDSL$_1$, which indicates the effectivenss

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2955252, IEEE/ACM Transactions on Audio, Speech, and Language Processing

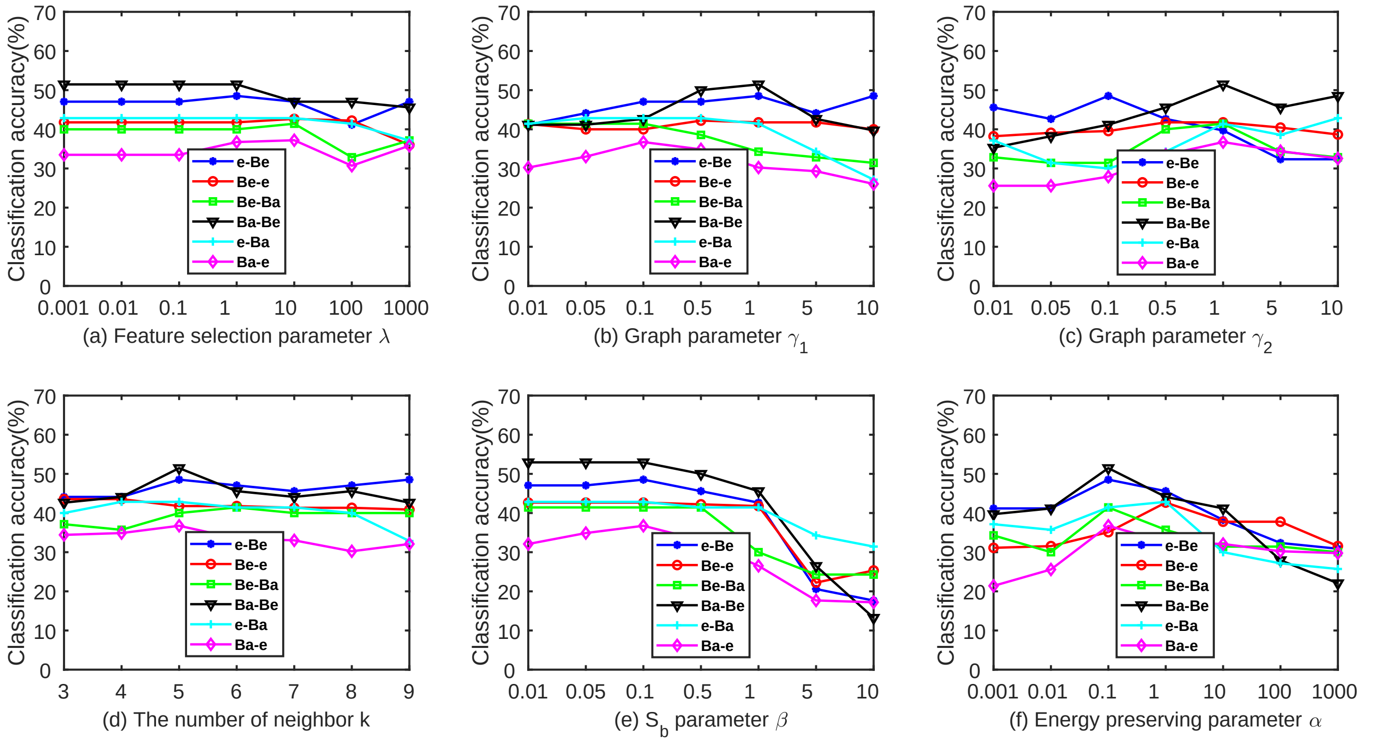


Fig. 4: Parameter sensitivity analysis of TSDSL with different parameters: (a) The recognition performance of TSDSL w.r.t. λ; (b) The recognition performance of TSDSL w.r.t. γ_1; (c) The recognition performance of TSDSL w.r.t. γ_2; (d) The recognition performance of TSDSL w.r.t. k; (e) The recognition performance of TSDSL w.r.t. β; (f) The recognition performance of TSDSL w.r.t. α.

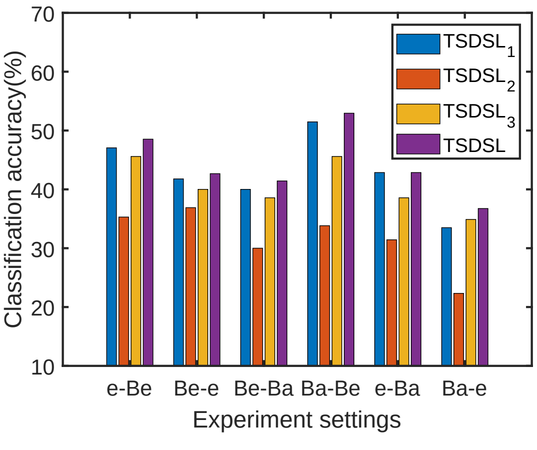


Fig. 3: Recognition results of TSDSL and three special cases, i.e., TSDSL_1, TSDSL_2 and TSDSL_3, under all settings.

of feature selection. Meanwhile, we notice that the recognition accuracies of TSDSL are higher than TSDSL_3, which proves that the energy preserving term plays a positive role.

3) Parameter sensitivity: We conduct empirical parameter sensitivity analysis, which demonstrates that TSDSL can obtain optimal performance under a wide range of parameter values. We select six main hyper-parameters to conduct sensitivity analysis, which are the feature selection regularization λ, the graph regularization parameters γ_1, γ_2, the number of nearest neighbors k, the S_b parameter β, and the energy preserving regularization parameter α. We obtain the recognition accuracies by using different values of these hyper-parameters.

We run TSDSL with different values of the feature selection regularization λ. Theoretically, λ controls the weight of feature selection regularization. Larger values of λ will make the feature selection more important in TSDSL. We plot the recognition accuracies w.r.t. different values of λ in Fig. 4 (a).From the figure, we can observe that $\lambda \in [0.001, 1]$ can be the optimal values, where TSDSL generally obtains good recognition performance.

We run TSDSL with varying values of graph regularization parameters γ_1 and γ_2. Theoretically, γ_1 and γ_2 control the weight of graph regularization. When $\lambda_1 \rightarrow 0$ and $\lambda_2 \rightarrow 0$, TSDSL will degenerate to traditional subspace learning, where the distance metric is discarded. In Fig. 4 (b) and Fig. 4 (c), we plot the recognition accuracies of our method w.r.t. different values of γ_1 and γ_2, respectively. By observing the figures, we choose the optimal values in a wide range $\gamma_1, \gamma_2 \in [0.1, 1]$.

We run TSDSL with varying values of k, which means the number of nearest neighbors in graph. Theoretically, k controls the complexity of the graph. We plot the recognition accuracies w.r.t. different numbers of nearest neighbors k in Fig. 4 (d). By observing the figure, we choose $k = 5$, where TSDSL obtains the highest recognition accuracies.

We run TSDSL with varying values of β and α. Theoretically, β balances the importance of S_b and S_w, and α controls the weight of energy preserving. In Fig. 4 (e) and Fig. 4 (f), we plot the recognition accuracies of our method w.r.t. different values of β and α, respectively. From the figures, we observe that TSDSL achieves the best recognition performance when $\beta = 0.1$ and $\alpha = 0.1$, and set the optimal values of β and α as 0.1. In addition, we investigate the recognition performance of different reduced dimensions, which is illustrated in Fig. 5. By observing the figure, we set the optimal dimension as 25 in our experiment. Furthermore, we also investigate the influence of different numbers of the target training data, which is shown in Fig. 6. From the figure, we can find that the recognition accuracies steadily increase with the number of the target training data increasing.

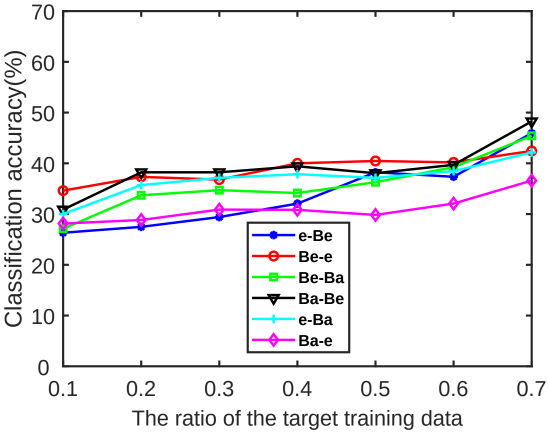


Fig. 6: Recognition results of TSDSL with different ratios of the target training data.

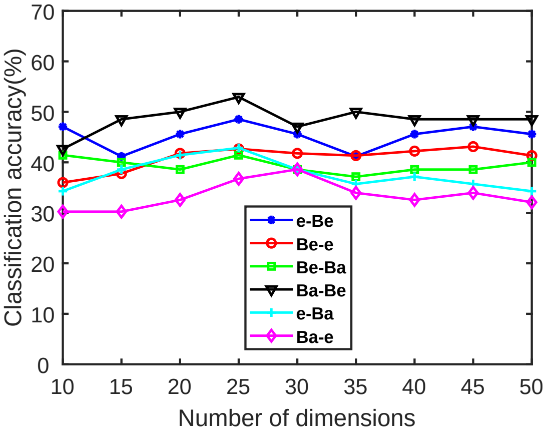


Fig. 5: Recognition results w.r.t. different dimensions under six settings.

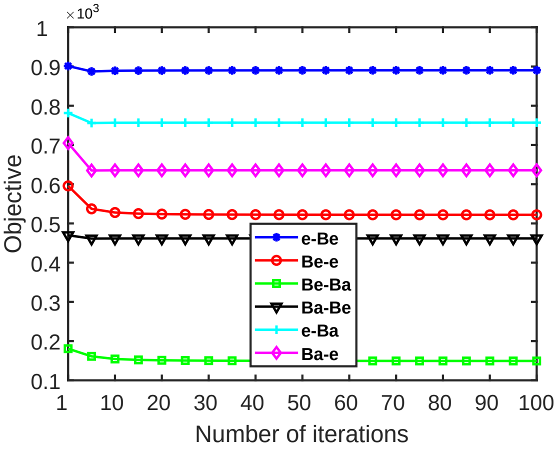


Fig. 7: Convergence study of TSDSL under six settings.

4) Convergence analysis: ***Since the objective of TSDSL in*** (12) ***involves $\ell_{2,1}-$norm, which is non-smooth and cannot obtain a closed form solution [74], [60], we have developed an iterative algorithm***. By using the proposed iterative algorithm, a local optimum can be achieved. We empirically check its convergence property on the six cross-corpus recognition settings. Fig. 7 gives the recognition accuracies w.r.t. different iterations. From the figure, we observe that the objective values decrease steadily with more iterations and can converge after 10 iterations. This means that our method can quickly converge.

V. Conclusion

In this paper, to address the cross-corpus speech emotion recognition problem, we have presented a novel transfer sparse discriminant subspace learning (TSDSL) method, which aims to extract the corpus-invariant feature representations. In TSDSL framework, a discriminant projection subspace is learned to map the features from different corpora into a common subspace, and a graph based distance metric is presented to measure the divergence between different corpora. Besides, we adopt the $\ell_{2,1}-$norm to learn the relevant and discriminative features simultaneously in the projection procedure. Extensive experiments on three benchmark datasets demonstrate the superior performance of our proposed method.

Since our model aims to learn the corpus-invariant feature representation and a novel graph based distance metric, it is possible to expand this model to more comprehensive tasks, e.g., a large-scale corpus based/multi-source cross-corpus emotion recognition problem. In addition, unlike conventional adaptation algorithms, e.g., vocal tract length normalization (VTLN), cepstral mean normalization (CMN) [75], our model has not been developed for real-time applications. We plan to make our model suitable for real-time applications by investigating conventional adaptation algorithms. Moreover, recent studies reveal that deep neural networks can learn much more transferable features [76], [49]. ***Also, the three datasets used in our experiments are too small for practical situations. Thus, it is interesting to involve larger emotion datasets, e.g., IEMOCAP, MSP-Improv [77], [78], and integrate our model into deep models to further boost the recognition***

performance in future.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China under Grants 61703360 and 61572419, and the Fundamental Research Funds for the Central Universities under Grant CDLS-2019-01.

References

[1] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[2] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[3] Diane Litman and Kate Forbes. Recognizing emotions from student speech in tutoring dialogues. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 25–30. IEEE, 2003.

[4] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[5] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[6] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213, 2014.

[7] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, 2018.

[8] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *INTERSPEECH*, pages 1537–1540, 2015.

[9] Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng. Learning discriminative features from spectrograms using center loss for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7405–7409. IEEE, 2019.

[10] Yixiong Pan, Peipei Shen, and Liping Shen. Speech emotion recognition using support vector machine. *International Journal of Smart Home*, 6(2):101–108, 2012.

[11] Shuiyang Mao, Dehua Tao, Guangyan Zhang, PC Ching, and Tan Lee. Revisiting hidden markov models for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6715–6719. IEEE, 2019.

[12] Joy Nicholson, Kazuhiko Takahashi, and Ryohei Nakatsu. Emotion recognition in speech using neural networks. *Neural Computing and Applications*, 9(4):290–296, 2000.

[13] Hao Hu, Ming-Xing Xu, and Wei Wu. Gmm supervector based svm with spectral features for speech emotion recognition. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages 413–416. IEEE, 2007.

[14] Diana Torres-Boza, Meshia Cédric Oveneke, Fengna Wang, Dongmei Jiang, Werner Verhelst, and Hichem Sahli. Hierarchical sparse coding framework for speech emotion recognition. *Speech Communication*, 99:80–89, 2018.

[15] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017.

[16] Peng Song. Transfer linear subspace learning for cross-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*, 2017.

[17] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[18] 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*, 22(14):e49–e57, 2006.

[19] Nassara Elhadji-Ille-Gado, Edith Grall-Maes, and Malika Kharouf. Transfer learning for large scale data using subspace alignment. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1006–1010. IEEE, 2017.

[20] Jingjing Li, Ke Lu, Zi Huang, Lei Zhu, and Heng Tao Shen. Transfer independently together: a generalized framework for domain adaptation. *IEEE Transactions on Cybernetics*, 49(6):2144–2155, 2018.

[21] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 402–410. ACM, 2018.

[22] Peng Song, Yun Jin, Li Zhao, and Minghai Xin. Speech emotion recognition using transfer learning. *IEICE TRANSACTIONS on Information and Systems*, 97(9):2530–2532, 2014.

[23] Ali Hassan, Robert Damper, and Mahesan Niranjan. On acoustic emotion recognition: compensating for covariate shift. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1458–1468, 2013.

[24] Yuan Zong, Wenming Zheng, Tong Zhang, and Xiaohua Huang. Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. *IEEE Signal Processing Letters*, 23(5):585–589, 2016.

[25] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 511–516. IEEE, 2013.

[26] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller. Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 24(4):500–504, 2017.

[27] Zhengwei Huang, Wentao Xue, Qirong Mao, and Yongzhao Zhan. Unsupervised domain adaptation for speech emotion recognition using pcanet. *Multimedia Tools and Applications*, 76(5):6785–6799, 2017.

[28] Mingsheng Long, Jianmin Wang, Jiaguang Sun, and S Yu Philip. Domain invariant transfer kernel learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(6):1519–1532, 2015.

[29] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

[30] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[31] Mingyu You, Chun Chen, Jiajun Bu, Jia Liu, and Jianhua Tao. Emotional speech analysis on nonlinear manifold. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 91–94. IEEE, 2006.

[32] Mehrdad J Gangeh, Pouria Fewzee, Ali Ghodsi, Mohamed S Kamel, and Fakhri Karray. Multiview supervised dictionary learning in speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(6):1056–1068, 2014.

[33] Xinzhou Xu, Jun Deng, Eduardo Coutinho, Chen Wu, Li Zhao, and Björn W Schuller. Connecting subspace learning and extreme learning machine in speech emotion recognition. *IEEE Transactions on Multimedia*, 21(3):795–808, 2018.

[34] **Pearson, Karl. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.**

[35] **Fisher, Ronald A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.**

[36] Xiaofei He and Partha Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems*, pages 153–160, 2004.

[37] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[38] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[39] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[40] Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, 2010.

[41] Shizhun Yang, Ming Lin, Chenping Hou, Changshui Zhang, and Yi Wu. A general framework for transfer sparse subspace learning. *Neural Computing and Applications*, 21(7):1801–1817, 2012.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2955252, IEEE/ACM Transactions on Audio, Speech, and Language Processing

[42] Ming Shao, Dmitry Kit, and Yun Fu. Generalized transfer subspace learning through low-rank constraint. *International Journal of Computer Vision*, 109(1-2):74–93, 2014.

[43] Haibin Yan. Transfer subspace learning for cross-dataset facial expression recognition. *Neurocomputing*, 208:165–173, 2016.

[44] Parvin Razzaghi, Parisa Razzaghi, and Karim Abbasi. Transfer subspace learning via low-rank and discriminative reconstruction matrix. *Knowledge-Based Systems*, 163:174–185, 2019.

[45] Shuicheng Yan, Jianzhuang Liu, Xiaoou Tang, and Thomas S Huang. A parameter-free framework for general supervised subspace learning. *IEEE Transactions on Information Forensics and Security*, 2(1):69–76, 2007.

[46] Peng Song and Wenming Zheng. Feature selection based transfer subspace learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 2018.

[47] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. Progressive neural networks for transfer learning in emotion recognition. *arXiv preprint arXiv:1706.03256*, 2017.

[48] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller. Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):31–43, 2017.

[49] Mohammed Abdelwahab and Carlos Busso. Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2423–2435, 2018.

[50] Chih-Chuan Lu, Jeng-Lin Li, and Chi-Chun Lee. Learning an arousal-valence speech front-end network using media data in-the-wild for emotion recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 99–105. ACM, 2018.

[51] **Zhang, Jing and Li, Wanqing and Ogunbona, Philip and Xu, Dong**. **Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective**. ***ACM Computing Surveys (CSUR)***, **52**(**1**):**7**, **2019**.

[52] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[53] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2724–2732, 2018.

[54] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4058–4065, 2018.

[55] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R 1-pca: rotational invariant l1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd International conference on Machine learning*, pages 281–288. ACM, 2006.

[56] Jie Wen, Xiaozhao Fang, Jinrong Cui, Lunke Fei, Ke Yan, Yan Chen, and Yong Xu. Robust sparse linear discriminant analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(2):390–403, 2019.

[57] Haifeng Li, Tao Jiang, and Keshu Zhang. Efficient and robust feature extraction by maximum margin criterion. In *Advances in Neural Information Processing Systems*, pages 97–104, 2004.

[58] Hwann-Tzong Chen, Huang-Wei Chang, and Tyng-Luh Liu. Local discriminant embedding and its variants. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 846–853. IEEE, 2005.

[59] Tianhao Zhang, Dacheng Tao, and Jie Yang. Discriminative locality alignment. In *European Conference on Computer Vision*, pages 725–738. Springer, 2008.

[60] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint l2, 1-norms minimization. In *Advances in Neural Information Processing Systems*, pages 1813–1821, 2010.

[61] Jian Yang and Jing-yu Yang. Why can lda be performed in pca transformed space? *Pattern Recognition*, 36(2):563–566, 2003.

[62] Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[63] Ran He, Tieniu Tan, Liang Wang, and Wei-Shi Zheng. l 2, 1 regularized correntropy for robust feature selection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2504–2511. IEEE, 2012.

[64] John C Gower, Garmt B Dijksterhuis, et al. *Procrustes problems*, volume 30. Oxford University Press on Demand, 2004.

[65] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

[66] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[67] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.

[68] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. Baum-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313, 2017.

[69] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2200–2207, 2013.

[70] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1410–1417, 2014.

[71] Deng Cai, Xiaofei He, and Jiawei Han. Semi-supervised discriminant analysis. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–7. IEEE, 2007.

[72] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S Narayanan. The interspeech 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[73] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

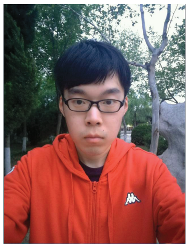
[74] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[75] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40, 2010.

[76] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017.

[77] **Busso, Carlos and Bulut, Murtaza and Lee, Chi-Chun and Kazemzadeh, Abe and Mower, Emily and Kim, Samuel and Chang, Jeannette N and Lee, Sungbok and Narayanan, Shrikanth S**. **IEMOCAP: Interactive emotional dyadic motion capture database**. ***Language resources and evaluation***, **42**(**4**):**335**, **2008**.

[78] **Busso, Carlos and Parthasarathy, Srinivas and Burmania, Alec and AbdelWahab, Mohammed and Sadoughi, Najmeh and Provost, Emily Mower**. **MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception**. ***IEEE Transactions on Affective Computing***, **8**(**1**):**67–8**, **2016**.



Weijian Zhang received the B.S. degree in Computer Science from Yantai University, Yantai, China, in 2018. He is currently pursuing the M.S. degree in Computer Science at Yantai University under the supervision of Prof. Peng Song. His current main research interests include speech signal process affective computing and pattern recognition.

Peng Song is currently an associate professor with the school of computer and control engineering, Yantai University, China. He received the B.S. degree in EE from Shandong University of Science and Technology, China in 2006, the M.E. and P.h.D degrees in EE both from Southeast University, China in 2009 and 2014, respectively. From 2007 to 2008, he was a research intern at Microsoft Research Asia. From 2009 to 2011, he worked as a software engineer at Motorola. His current main research interests include affective computing, speech signal processing and pattern recognition.