Journal of Artificial Intelligence Research 66 (2019) 381-410 Submitted 12/2018; published 10/2019

DSTL: Solution to Limitation of Small Corpus in Speech Emotion Recognition

Ying Chen YCHEN718@STU.SUDA.EDU.CN
Zhongzhe Xiao XIAOZHONGZHE@SUDA.EDU.CN
Xiaojun Zhang ZHANGXJ@SUDA.EDU.CN
Zhi Tao TAOZ@SUDA.EDU.CN
School of Optoelectronic Science and Engineering,
Soochow University, SuZhou, China

Abstract

Traditional machine learning methods share a common hypothesis: training and testing datasets must be in a common feature space with the same distribution. However, in reality, the labeled target data may be rare, so that target space does not share the same feature space or distribution as an available training set (source domain). To address the mismatch of domains, we propose a Dual-Subspace Transfer Learning (DSTL) framework that considers both the common and specific information of the two domains. In DSTL, a latent common subspace is first learned to preserve the data properties and reduce the discrepancy of domains. Then, we propose a mapping strategy to transfer the source-specific information to the target subspace. The integration of the domain-common and specific information constructs the proposed DSTL framework. In comparison to the state-art-of works, the main contribution of our work is that the DSTL framework not only considers the commonalities, but also exploits the specific information. Experiments on three emotional speech corpora verify the effectiveness of our approach. The results show that the methods which include both domain-common and specific information perform better than the baseline methods which only exploit the domain commonalities.

1. Introduction

Speech Emotion Recognition (SER), as an important branch of affective computing, has been a popular topic in Human-Computer Interaction (HCI) fields (Zhang, Zhang, Huang, & Gao, 2018; Schuller et al., 2010), aiming to identify the emotional states (*e.g.*, neutral, happiness, fear, sadness.) of human speech. SER recently has been extensively used in various areas, such as detecting the mental state of the driver and reminding him to avoid a traffic accident when needed, facilitating emotional tracking of patients with depression, and so on.

A number of Machine Learning (ML) approaches have been proposed in the field of SER, including linear methods, such as Support Vector Machine (SVM) (Vaishali & Gohokar, 2012), Artificial Neural Network (ANN) (Goldberg, 2016; Safdarkhani, Mojaver, Atieghechi, Molanoori, & Riahi, 2012), Naïve Bayes Classifier; and other non-linear methods, *e.g.*, Gaussian Mixture Model (GMM) (Vlassis & Likas, 2002), K-Nearest Neighbor algorithms (K-NN), Decision Trees, *etc.* These above mentioned methods perform well under a vital common assumption: training and testing data are in the same feature space and have the same distribution (Pan & Yang, 2010; Daumé III & Marcu, 2006). However, when the

©2019 AI Access Foundation. All rights reserved.

distribution changes, most statistical models need to be rebuilt from the very beginning using newly collected training data, which is expensive and does not scale. Another fact is that the scales of existing emotional speech corpora are usually small due to the high cost in data labeling. According to the survey (Wang & Zheng, 2015), there are more than 5000 languages around the world, but only a few languages have adequate resources (*e.g.*, speech signal, text corpus, emotional speech corpus, *etc*). The sharing of the available sources also suffers from a number of factors, such as different languages, types of emotion expressions (acted *vs.* naturalistic) (Deng, Zhang, & Schuller, 2014), ages of speakers (children *vs.* adults) (Tong, Wang, & Ma, 2017), types of recording situations. Directly applying a model trained on one corpus to another will cause severe degradation of performance for most ML approaches due to mismatch. Transfer learning methods can be a great tool to alleviate this problem.

Transfer learning, in contrast with traditional ML, allows the domains, tasks, and distributions used in training and testing to be different (Pan & Yang, 2010). Specifically, the motivation of transfer learning is to attempt to transfer the knowledge in a supervised domain (termed as source domain) to another different but related domain with only limited information (termed as target domain) to induce a better model (Gasulla et al., 2018). Transfer learning approaches have already been used in a number of applications and domains. For example, in face recognition, Kan et al. (2014) transferred the knowledge from a model trained on a source dataset to a new face dataset, which lacks data labels and possesses different lighting conditions and subjects. Also, in Natural Language Processing (NLP), labeled data for tasks like part-of-speech tagging, parsing, or information extraction are generally drawn from a limited set of document types and genres in a given language because of availability, cost, and project goals. Thus, David, Blitzer, Crammer, and Pereira (2006) adapted a classification model trained on some document sets (source domain) for a new document set (target domain). Facing the similar situation, the transfer learning approaches can also be effective in SER.

Various transfer learning approaches (Pan & Yang, 2010; Yang & Gao, 2014; Deng, Xu, Zhang, Frühholz, & Schuller, 2018) have been proposed to cope with the mismatch between training and testing datasets. Mainstream approaches are common/feature-based methods, which explore the commonalities of both source and target domains, such as common feature representation or a common subspace. A key factor in this kind of approaches is to find a good feature representation for two domains with different distributions, to preserve the discriminative properties and reduce the discrepancy as much as possible (Pan, Tsang, Kwok, & Yang, 2011; Kan, Wu, Shan, & Chen, 2014).

An early representative common/feature-based method was Maximum Mean Discrepancy Embedding (MMDE) (Pan, Kwok, & Yang, 2008), which aims to learn a latent space where the distribution differences can be reduced and the data variance can be preserved. Its limitation as not generalizing out of sample patterns was later solved in Transfer Component Analysis (TCA) framework (Pan et al., 2011), which can not only find the main components of domains, but also reduce the distribution differences. More recent methods include Transfer Non-negative Matrix Factorization (TNMF) method, where Maximum Mean Discrepancy (MMD) and Graph Embedding (GE), as regularization terms, are combined with non-negative matrix factorization method, and MMD is directly computed without non-kernel mapping (Song, Ou, Zheng, Jin, & Zhao, 2016). Furthermore, Trans-

fer Linear Subspace Learning (TLSL) method employs a novel feature grouping strategy by preserving the high transfer part and suppressing the low transfer part, to avoid the negative transfer of knowledge (Song, 2017).

Most of these above common/feature-based methods are suitable for the mismatch problems, so that the distributions of the source and target domains are close to each other. However, this kind of methods concerns only the domain-common information, while ignores the domain-specific information. Researches show that combination of common and specific information can better adapt to the problems with more extensive potential of applications (Song, 2017; Fernando, Habard, Sebban, & Tuytelaars, 2013). In our work, we propose to transfer the source-specific information to the target subspace by using a mapping matrix, to solve the lack of labeled training data.

The method proposed in this work is named Dual-Subspace Transfer Learning (DSTL) framework, which combines the commonalities in a common subspace and specific information in the target subspace. The term"dual-subspace" refers to two subspaces: common subspace for commonalities, and target subspace for specific information. The low dimensional common subspace is extracted by certain common-based methods, so as to preserve the data properties and reduce the distribution differences. To bridge the two subspaces for transferring the source-specific information to the target subspace, we propose a novel mapping strategy as SMT (Source-specific Mapping to Target subspace, to be stated in Section 2.2.2). Finally, the common and specific information are combined in DSTL to train a supervised classifier for predicting the class labels of the target data. DSTL is a general framework that can combine most common/feature-based methods with SMT method, and thus it focuses on commonalities and specific information simultaneously.

The rest of this paper is organized as follows: Section 2 describes our proposed DSTL framework in detail, including finding a latent common subspace between domains and transferring the source-specific knowledge to the target subspace. In Section 3, we verify the performance of the proposed method on three different emotional speech corpora for cross-corpus emotion recognition. In Section 4, some research issues are discussed. Finally, we give a conclusion in the last section.

2. Dual-Subspace Transfer Learning

The common space between domains dominates the discussion in the transfer learning method literature, while essential domain-specific information is ignored. Therefore, in this work, both commonalities and domain-specific information are included in our proposed DSTL framework. After finding a common space of both source and target domains, we aim to capture specific information of the source/target domains, and transfer the source-specific information to the target subspace, to maintain the source supervised knowledge at utmost.

The overall diagram of the DSTL framework is illustrated in Figure 1. The two subspaces in the term of dual-subspace are: a) the common subspace between the source and target domains; b) the specific subspace, mapped from source data to target subspace. The common subspace is derived with several baseline methods, which assembles certain distance measurements and MPCA (Modified Principal Component Analysis). According to the distance measurements, the common subspace extraction methods are MMD+MPCA

(M-MPCA), GE+MPCA (G-MPCA), and MMD+GE+MPCA (MG-MPCA) (Song, 2017). Following these methods, this paper proposes a novel method called SMT, to preserve the specific information. The xPCAs and the SMT together constitute the DSTL framework, as D-M-MPCA, D-G-MPCA, and D-MG-MPCA.

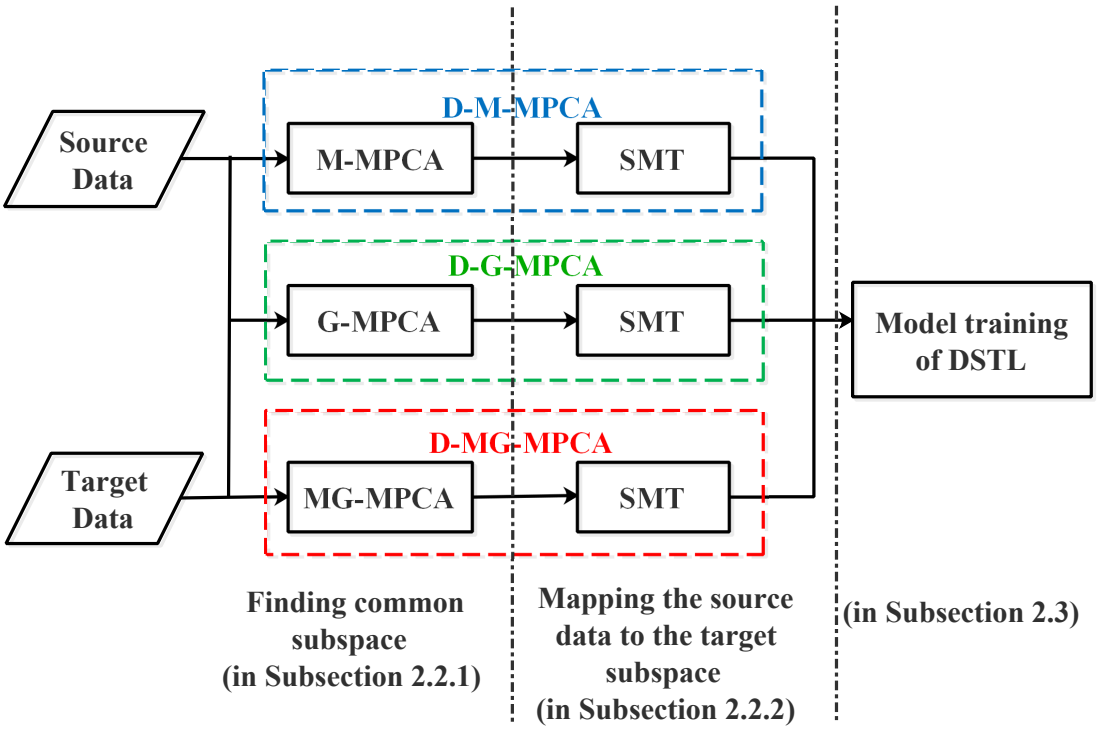


Figure 1: Overview of the DSTL framework. Three baseline methods (*i.e.*, M-MPCA, G-MPCA, MG-MPCA) for commonalities are combined with SMT to obtain the DSTL methods (*i.e.*, D-M-MPCA, D-G-MPCA, D-MG-MPCA).

2.1 Problem Description and Notation

Suppose that there are two different corpora with the same emotional classification task, where one is labeled, while the other is not. We regard the labeled corpus as the source domain, and expect to predict the emotional states for the other corpus (target domain) using the source knowledge. Thus, our proposed approach falls into the transductive transfer learning type (Pan & Yang, 2010).

The following parameters are defined for the rest of this paper. For source domain with n_s samples in c emotional classes and feature set of d dimensions, the data matrix is expressed as $X_s = \{x_1^s, x_2^s, \ldots, x_{n_s}^s\} \in R^{d \times n_s}$, where $x_i^s| \in R^{d \times 1} (i = 1, 2, \ldots, n_s)$ denotes the i^{th} sample. The label matrix is denoted as $Y_s = \{y_1^s, y_2^s, \ldots, y_{n_s}^s\} \in R^{1 \times n_s}$, $y_i^s \in \{1, 2, \ldots, c\}$, where y_i^s $(i = 1, 2, \ldots, n_s)$ is the emotional label of the i^{th} sample.

Similarly, for the target domain, which is also with feature set of d dimensions and contains n_t samples, but with limited labels, the data matrix is denoted as $X_t = \{x_1^t, x_2^t, \ldots, x_{n_t}^t\} \in R^{d \times n_t}$, where $x_i^t| \in R^{d \times 1}$ is the i^{th} sample. The data matrices of source domain X_s and target domain X_t can be concatenated as $X = [X_s, X_t]$.

In general, unless otherwise specified, the subscripts/superscripts s and t represent the source domain and target domain, respectively. Other frequently used parameters in this paper are summarized in Table 1.

Parameters	Description
X_{im}/X_{un}	original important/unimportant features
C_s/C_t	commonalities of source/target domain
F_s/F_t	specific information of source/target domain
F_{st}	specific information of source domain in target subspace
T_s/T_t	combination of C_s/C_t and F_{st}/F_t
P	projected common subspace
D_s/D_t	source/target subspace
S_{im}/S_{un}	scatter matrix of X_{im}/X_{un}
D_b	matrix of between-class distance
D_w	matrix of within-class distance
M	MMD matrix
W	weight matrix of GE
γ	regularization parameter of MMD
β	regularization parameter of GE
p	number of nearest neighbors

Table 1: Description of frequently used parameters.

2.2 General Framework

Inspired by former studies(Deng et al., 2014; Song, 2017), this work aims to build a discriminative model for the target domain, which is trained on the labeled source domain. The first step of this work is to find a latent common feature subspace between source and target domains, to retain the common discriminative features of both domains. Existent algorithms on distribution discrepancies include Maximum Mean Discrepancy (Xu, Fang, Wu, Li, & Zhang, 2016; Zhang, Provost, & Essl, 2016), Graph Embedding (Song & Zheng, 2017), Bregman Divergence (Si, Tao, & Geng, 2010), Kullback-Leibler (Noda, Yano, Doki, & Okuma, 2006), *etc.* Two of them, Maximum Mean Discrepancy and Graph Embedding, are used in this work to yield a low-dimensional common subspace by reducing distribution differences.

Besides the common subspace, the domain-specific information is also essential in transfer learning (Kan et al., 2014). Therefore, we propose a SMT (Source-specific Mapping to Target subspace) method to make use of the domain-specific information. The processing of SMT is shown in Table 2, and each processing step is marked with circled number, as shown below: 1) PCAs are applied to source and target data respectively, to capture the domain-specific information. This step results in source-specific features F_s and target-

specific features F_t, which construct a source subspace D_s and a target subspace D_t ; 2) a mapping matrix R is developed according to the relationship between the subspaces D_s and D_t; 3) the source-specific features can be mapped onto the target subspace via R as F_{st}.

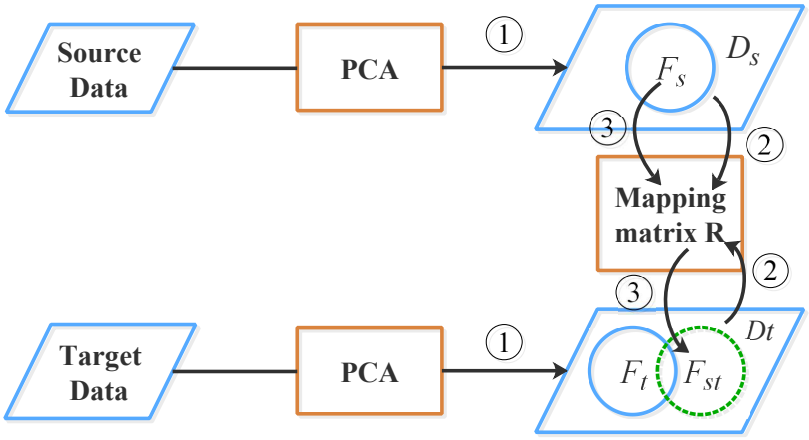


Figure 2: Overview of the SMT method. Each circled number corresponds to different processing steps: 1 PCA processing aims to reduce the feature dimension and find specific information (*i.e.*, F_s and F_t); 2) we learn the mapping matrix R by using D_s and D_t; 3) the source-specific information is transferred to the target subspace as F_{st}.

The commonalities and domain-specific information work together to achieve an effective recognition model, which is called as Dual-Subspace Transfer Learning (DSTL) framework.

The performance of the DSTL framework is evaluated in a comparison way. The approaches applying only commonalities are used as baselines: M-MPCA, G-MPCA, MG-MPCA (to be stated in Section 2.2.1). The corresponding representations in DSTL framework, as DSTL for M-MPCA (D-M-MPCA), DSTL for G-MPCA (D-G-MPCA) and DSTL for MG-MPCA (D-MG-MPCA), are exhibited to show the advantage of the proposed DSTL.

2.2.1 Find Latent Common Subspace

For two corpora which share the same recognition task, there exist both consistency and inconsistency between source and target domains. The inconsistency prevents the direct application of models trained on source domain to target domain, while the consistency provides the possibility to find a common space between the domains. A transformation matrix P is presented to project the source and target data into a common space, in which the distributions of the source and target domains are close to each other. The projected source and target samples are represented as $C_s = \{c_1^s, c_2^s, \ldots, c_{n_s}^s\}$ and $C_t = \{c_1^t, c_2^t, \ldots, c_{n_t}^t\}$ respectively, as follows:

$$c_i^s = P^T x_i^s \qquad i = 1, 2, \ldots, n_s \tag{1}$$

$$c_i^t = P^T x_i^t \qquad i = 1, 2, \ldots, n_t \tag{2}$$

or

$$C_s = P^T X_s \tag{3}$$
$$C_t = P^T X_t \tag{4}$$

Two-step methods, which reduce the feature dimension in the first step, and reduce the distribution differences between domains in the second step (Song et al., 2016; Song, 2017), are developed in this work. Modified Principal Component Analysis (MPCA) is chosen for dimensional reduction, and Maximum Mean Discrepancy (MMD) and Graph Embedding (GE) are chosen to evaluate the distribution differences.

- Step 1 - dimensional reduction: MPCA

  The classic PCA aims to find the projected subspace P with maximum variance by removing the projection direction with minimal variance, so as to preserve the principal components of data (Yan, Xu, Zhang, & Zhang, 2005). The objective function can be expressed as:

$$\begin{aligned} &\max_P tr(P^T S P) \\ &s.t.\ P^T P = I \end{aligned} \tag{5}$$

  where S is the covariance matrix of the whole data matrix as $X = [X_s, X_t]$, and I is an identity matrix; $tr(\cdot)$ means the trace of a matrix. Thus, the source and target data can be projected from the original feature space into a new subspace by P.

  In order to get optimum performance, the original features are first evaluated according to their importance before the projection by PCA (Song, 2017), and different weighting coefficients are assigned to the features. We define two categories as important feature set X_{im} and unimportant feature set X_{un} by evaluating the ratios of between-class distance $D_b = \{d_{b,1}, d_{b,2}, \ldots, d_{b,d}\} \in R^{1\times d}$ and within-class distance $D_w = \{d_{w,1}, d_{w,2}, \ldots, d_{w,d}\} \in R^{1\times d}$ in the fully labeled source domain. The between-class distance $d_{b,r}$ and within-class distance $d_{w,r}$ of the r^{th} dimensional feature are expressed as:

$$d_{b,r} = \sum_{i=1}^{c} n_i (u_{i,r} - \bar{u}_r)^2 \qquad r = 1, 2, \ldots, d \tag{6}$$

$$d_{w,r} = \sum_{i=1}^{c} \sum_{x_k^s \in class\ i} (u_{i,r} - x_{k,r}^s)^2 \qquad r = 1, 2, \ldots, d \tag{7}$$

  where c is the total number of classes; n_i is the number of source samples in the i^{th} class ($i = 1, 2, \ldots, c$), and $u_{i,r}$ is the mean value of the r^{th} dimensional feature for the source samples in the i^{th} class; $\bar{u}_r$ is the mean value of the r^{th} dimensional feature of all source samples. Both equations (6) and (7) iterate over all the $i = 1, 2, \cdots, c$ classes,

and equation (7) also iterates over all the samples of the i^{th} class in the inner loop. Larger between-class distance $d_{b,r}$ and smaller within-class distance $d_{w,r}$ indicate a feature with a high ability to distinguish between classes. This can be evaluated with the ratio σ_r between $d_{b,r}$ and $d_{w,r}$ of the r^{th} dimensional feature as:

$$\sigma_r = \frac{d_{b,r}}{d_{w,r}} \qquad r = 1, 2, \ldots, d \tag{8}$$

The features are then ranked in descending order according to σ_r. An Importance Ratio Parameter (IRP) is defined as α in the range of 0 to 1, to assign the first r_1 features with higher σ_r as important features, and the rest features as unimportant features, where $r_1 = \alpha d$ (r_1 should be an integer). Therefore, the important features of the source and target sets can be expressed as X^s_{im} and X^t_{im}, and the unimportant features are represented as X^s_{un} and X^t_{un}.

To preserve the important features and suppress the unimportant features, we modified PCA in Eq. (5) with X_{im} and X_{un} as MPCA:

$$\begin{gathered} \max_P \frac{tr(P^T S_{im} P)}{tr(P^T S_{un} P)} \\ s.t.\ P^T P = I \end{gathered} \tag{9}$$

where S_{im} and S_{un} are the covariance matrices of the important feature set $X_{im} = [X^s_{im}, X^t_{im}]$ and the unimportant feature set $X_{un} = [X^s_{un}, X^t_{un}]$, as shown below:

$$S_{im} = \sum_{i=1}^{N} (x_i^{im} - \bar{u}_{im})(x_i^{im} - \bar{u}_{im})^T \tag{10}$$

$$S_{un} = \sum_{i=1}^{N} (x_i^{un} - \bar{u}_{un})(x_i^{un} - \bar{u}_{un})^T \tag{11}$$

where N is the number of all samples; x_i^{im} is the i^{th} sample of X_{im}; $\bar{u}_{im}$ is the mean of all samples in X_{im}. Similarly, x_i^{un} is the i^{th} sample of X_{un}, and $\bar{u}_{un}$ is the mean value of all samples in X_{un}.

The MPCA technique can preserve the essential information in a low-dimensional subspace, while the inconsistency between source and target corpora due to their different origination may be still large. Thus, the distribution differences will be measured and eliminated in the next step.

- Step 2 - discrepancy measurement

  Two discrepancy-measurement methods, MMD and GE, are used in step 2 to measure the distribution differences between source and target domains.

1) Maximum Mean Discrepancy

The main idea of MMD is to measure the similarity of the two domains via computing the mean values of source data and target data in a m dimensional feature space. The distance between the two domains can be measured with a Reproducing Kernel Hilbert Space (RKHS), or without non-kernel mapping as a regularization term (Song et al., 2016). In the case without mapping, the distance of two domains can be expressed as:

$$\begin{aligned} D(C_s, C_t) &= \|\frac{1}{n_s}\sum_{i=1}^{n_s} c_i^s - \frac{1}{n_t}\sum_{i=1}^{n_t} c_i^t\|^2 \\ &= \|\frac{1}{n_s}\sum_{i=1}^{n_s} P^T x_i^s - \frac{1}{n_t}\sum_{i=1}^{n_t} P^T x_i^t\|^2 \\ &= P^T(\frac{1}{n_s^2}X_s I_s I_s^T X_s^T - \frac{1}{n_s n_t}X_s I_s I_t^T X_t^T \\ &\quad -\frac{1}{n_t n_s}X_t I_t I_s^T X_s^T + \frac{1}{n_t^2}X_t I_t I_t^T X_t^T)P \\ &= tr(P^T X M X^T P) \end{aligned} \tag{12}$$

where $X = [X_s, X_t] \in R^{d\times(n_s+n_t)}$; $I_s = [1, 1, \cdots, 1]^T \in R^{n_s\times 1}$; $I_t = [1, 1, \cdots, 1]^T \in R^{n_t\times 1}$; n_s and n_t are the number of samples in the source and target domains; C_s and C_t are the common features projected by MPCA, respectively. M is the MMD matrix with elements m_{ij} as:

$$m_{ij} = \begin{cases} \frac{1}{n_s^2} & x_i, x_j \in X_s \\ \frac{1}{n_t^2} & x_i, x_j \in X_t \\ \frac{-1}{n_s n_t} & \text{otherwise} \end{cases} \tag{13}$$

2) Graph Embedding

Graph Embedding is a dimensional reduction method by preserving the similarities of the neighboring points (Yan et al., 2005), which can be used as a distance criterion to measure the differences of two different domains. It can preserve intrinsic geometrical information that is important to the discrimination of data by giving larger weight for points with high similarity to each other. The GE method treats each vector as a vertex of a graph, and preserves the similarities of vertex pairs by calculating the graph similarity matrix, which can characterize statistical or geometrical property of a data set (Yan et al., 2005).

Given a graph with vertices, every vertex represents a sample vector. For each sample vector, we can find its p nearest neighbors in terms of Euclidean distance. We define a 0-1 weight matrix $W = [w_{ij}] \in R^{(n_s+n_t)\times(n_s+n_t)}$ as:

$$w_{ij} = \begin{cases} 1 & \text{if } x_i^s \in N_p(x_j^t) \text{ or } x_j^t \in N_p(x_i^s) \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where the $N_p(x_j^t)$ and $N_p(x_i^s)$ refer to the p nearest neighbors of x_j^t and x_i^s, respectively. The element w_{ij} of W denotes the similarity of the vertex pair. The objective function of the GE is calculated as:

$$\begin{aligned} G(P) &= \frac{1}{2}\sum_{i,j=1}^{N}\|c_i - c_j\|^2 w_{ij} \\ &= \frac{1}{2}(\sum_{i=1}^{N} c_i^2 \sum_{j=1}^{N} w_{ij} + \sum_{j=1}^{N} c_j^2 \sum_{i=1}^{N} w_{ij} - 2\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j w_{ij}) \\ &= \sum_{i=1}^{N} c_i^2 D_{ii} - \sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j w_{ij} \\ &= tr(P^T X L X^T P) \end{aligned} \tag{15}$$

where $c_i, c_j \in C$ and $C = [C_s, C_t]$; $N = n_s + n_t$; L is a Laplacian matrix which can be expressed as $L = D - W$, and D is a diagonal matrix whose entries are the column sums of W.

With the MPCA for dimensional reduction in step 1, and MMD or GE for discrepancy measurement in step 2, the two-step common subspace finding methods in this work include the following 3 assemblies of the two steps, as M-MPCA (MMD+MPCA), G-MPCA (GE+MPCA) and MG-MPCA (MMD+GE+MPCA).

(1) MMD + MPCA (M-MPCA)
In this assembly, the difference between the mean values of source and target corpora should be minimized according to the MMD principle. Thus, Eq. (12) is applied to Eq. (9) as a regularization constraint to result in:

$$\begin{aligned} &\min_P \frac{tr(P^T S_{un} P)}{tr(P^T S_{im} P)} + \gamma tr(P^T X M X^T P) \\ &s.t.\ P^T P = I \end{aligned} \tag{16}$$

where γ is a regularization parameter to control the trade-off between the direction with maximum variance and distribution differences. The trace term can be formulated as a new form using Maximum Margin Criterion (MMC) (Song, 2017) as:

$$\begin{aligned} &\min_P tr(P^T S_{un} P) - tr(P^T S_{im} P) + \gamma tr(P^T X M X^T P) \\ &s.t.\ P^T P = I \end{aligned} \tag{17}$$

The Lagrange multiplier method is employed to further optimize this formula. Let $V = XMX^T$, and λ be the Lagrange multiplier. Eq. (17) can be modified as

$$L(P, \lambda) = tr(P^T(S_{un} - S_{im} + \gamma V)P) + \lambda(I - P^T P) \tag{18}$$

Subsequently, let $\frac{\partial L(P,\lambda)}{\partial P} = 0$, we can obtain a final expression of M-MPCA as:

$$(S_{un} - S_{im} + \gamma V)P = \lambda P \tag{19}$$

This is a generalized eigendecomposition form. To minimize the objective function in Eq. (16), the eigenvalues are sorted from small to large, and the first k eigenvectors are selected to build the common subspace P.

(2) GE + MPCA (G-MPCA)
In this assembly, the objective function of GE used as a regularization constraint term, is combined with MPCA. The final objective function can be expressed as:

$$(S_{un} - S_{im} + \beta H)P = \lambda P \tag{20}$$

where β is a regularization parameter, and $H = XLX^T$, L is the Laplacian matrix as defined in Eq. (15).

(3) MMD + GE + MPCA (MG-MPCA)
In MG-MPCA, both objective functions of MMD and GE are used as constraint terms to the optimization of MPCA. The objective function of MG-MPCA can be expressed as:

$$\begin{aligned} &\min_P \frac{tr(P^T S_{un} P)}{tr(P^T S_{im} P)} + \gamma tr(P^T X M X^T P) + \beta tr(P^T X L X^T P) \\ &s.t.\ P^T P = I \end{aligned} \tag{21}$$

where the first term aims to discover the common important features of the two domains; the second term is designed to reduce the domain differences; and the last term is to preserve the local neighboring information.

Eq. (21) can be simplified as :

$$(S_{un} - S_{im} + \gamma V + \beta H)P = \lambda P \tag{22}$$

The resulted common subspace P consists of k eigenvectors corresponding to the first k smallest eigenvalues.

Among these three methods of finding common space in the cross-corpus recognition task, M-MPCA method mainly focuses on the global discrepancy by computing the mean values of the two datasets. G-MPCA method more concerns the similarities of the neighboring points, so as to preserve the local geometrical structures of data. Moreover, MG-MPCA method inherits the advantages of both methods by considering both global and local similarities.

2.2.2 SMT - Source-Specific Mapping to Target Domain

The methods in finding common space between source and target domains (*i.e.*, M-MPCA, G-MPCA, MG-MPCA) risk losing domain-specific information, thus specific information of the two domains is needed to solve this problem.

A mapping strategy called SMT is proposed in this work. First, the specific information of each domain is extracted with a coordinate transformation using PCA, as follows:

$$\max_{D_s} tr(D_s^T S_s D_s)$$
$$s.t.\ D_s^T D_s = I \tag{23}$$

$$\max_{D_t} tr(D_t^T S_t D_t)$$
$$s.t.\ D_t^T D_t = I \tag{24}$$

where S_s and S_t are the covariance matrices of the source and target data. the directions with the first h largest variances of the projected samples are selected as source subspace $D_s \in R^{d\times h}$ and target subspace $D_t \in R^{d\times h}$, which can preserve the essential specific information of each domain. Thus, the specific information F_s and F_t in each subspace can be expressed as:

$$F_s = D_s^T X_s \tag{25}$$

$$F_t = D_t^T X_t \tag{26}$$

Then, a mapping matrix R from the source subspace to the target subspace is learned to find the relationship of the two different subspaces. We expect that the source subspace mapped by R has a minimum distance with target subspace:

$$\arg\min_R \|D_s R - D_t\|_F^2$$
$$s.t.\ R^T R = I \tag{27}$$

where $\|\cdot\|_F$ refers to Frobenius norm, which is defined to be the square root of the sum of squares of the entries of the matrix (Hartley, 1997); R is the mapping matrix of the two subspaces. We expand this formula as:

$$\|D_s R - D_t\|_F^2 = tr(R^T D_s^T D_s R - 2D_t^T D_s R + D_t^T D_t) \tag{28}$$

Because D_s and D_t are orthogonal matrices, D_s and D_t can satisfy $D_s^T D_s = I$ and $D_t^T D_t = I$. Besides, $R^T D_s^T D_s R$ and $D_t^T D_t$ are both constant terms due to the constraint of $R^T R = I$. Therefore, Eq. (27) becomes:

$$\arg\max_R tr(D_t^T D_s R)$$
$$s.t.\ R^T R = I \tag{29}$$

We perform a singular value decomposition on $D_t^T D_s$ to achieve $D_t^T D_s = USV^T$, and thus Eq. (29) further becomes:

$$\arg\max_R tr(USV^T R)$$
$$s.t.\ R^T R = I \tag{30}$$

Let $Z = V^T RU$, then the objective function of Eq. (30) is converted to Eq. (31) as:

$$\arg\max_R tr(SZ) = \sum_{\tau=1}^{d} \sigma_\tau z_{\tau\tau} \tag{31}$$

where σ_τ is the τ^{th} singular value of S.

Due to $Z^T Z = U^T R^T V V^T RU = I$, Eq. (31) can obtain the maximum value when $z_{\tau\tau} = 1$, that is, $\sum_{\tau=1}^{d} z_{\tau\tau}\sigma_\tau \leq \sum_{\tau=1}^{d} \sigma_\tau$, then

$$\arg\max_R tr(D_t^T D_s R) = \sum_{\tau=1}^{d} \sigma_\tau = tr(S) \tag{32}$$

Thus, the objective function Eq. (27) has the minimum value when $Z = I$, and $R = VU^T$ can be derived.

Then, the source specific information F_s can be mapped to the target subspace as the targetized source data F_{st} iva the mapping function R. F_{st} is the source specific information represented by the target subspace.

$$F_{st} = RF_s \tag{33}$$

The mapping step in SMT can not only transfer the supervised specific information of source domain to the target subspace, but also retain discriminative information of the target subspace as much as possible. Hence, F_{st} and F_t can preserve the specific information of source and target domains, respectively.

2.3 DSTL Frame Work

In Section 2.2.2, two kinds of subspaces can be derived from different source and target domains for a same emotion recognition task:

(1) Common information subspace, which can be obtained via M-MPCA, G-MPCA, or MG-MPCA, to result in the commonalities $C = [C_s, C_t]$;

(2) Specific information subspace, which can map the source-specific information to target domain via SMT, to result in specific information $F = [F_{st}, F_t]$.

A DSTL (Dual-Subspace Transfer Learning) framework is proposed in this work to fully benefit information from both common and specific subspaces. To correspond to the common space finding methods, the approaches in the DSTL are named as D-M-MPCA (D refers to dual subspaces), D-G-MPCA, and D-MG-MPCA. The transferred source and target domains, carrying information from both subspaces, are presented as:

$$T_s = \left\{ \begin{pmatrix} C_s \\ F_{st} \end{pmatrix}, Y_s \right\} \tag{34}$$

$$T_t = \left\{ \begin{pmatrix} C_t \\ F_t \end{pmatrix} \right\} \tag{35}$$

where Y_s in Eq.(34) represents the class labels of the source samples.

In the transferred domains, the recognition model trained on the labeled data T_s can be used to predict the class labels of T_t. Algorithm 1 shows the processing steps of the DSTL framework.

Algorithm 1 : Algorithm for *DSTL*

Input:

The source domain dataset X_s with corresponding class labels Y_s;
The target domain dataset X_t.

Output:

Commonalities $C = [C_s, C_t]$;
Specific information $F = [F_{st}, F_t]$;
Class labels $Y_t \in R^{1\times n_t}$ of target samples.

1. Find a common subspace $P \in R^{d\times k}$ using *a common space finding method*;
2. Generate common features $C = P^T X$, where $X = [X_s, X_t]$;
3. Learn the specific information of the source and target domains using equations (23) - (26);
   For source domain: $F_s = D_s^T X_s$;
   For target domain: $F_t = D_t^T X_t$;
4. Find the mapping matrix R between the two domains using Eq. (27);
5. Calculate the target specific information F_{st} from the source domain: $F_{st} = RF_s$;
6. Combine the common and specific information: $T_s = \left\{ \begin{pmatrix} C_s \\ F_{st} \end{pmatrix}, Y_s \right\}$, $T_t = \left\{ \begin{pmatrix} C_t \\ F_t \end{pmatrix} \right\}$;
7. The recognition model trained on T_s is used to predict the class labels of the target dataset.

The italicized words in the table can be replaced as shown below:
DSTL: D-M-MPCA, D-G-MPCA, or D-MG-MPCA
common space finding method: M-MPCA in Eq. (19), G-MPCA in Eq. (20), or MG-MPCA in Eq. (22)

3. Experiment and Results

The DSTL framework is applied in this section on cross-corpus speech emotion recognition task. Only the discrete emotional states that are common to all the chosen corpora are considered. Although all the corpora chosen in the experiments are labeled with emotional states, the full labels are used only when the corresponding corpus is taken as the source corpus. When used as target corpus, a major part of the labels is omitted to simulate the case of partially labeled corpus. In this section, we show the recognition performance of four common emotional states on cross-corpus recognition.

3.1 Corpora Selection

Three emotional speech corpora covering two different languages are chosen in this work: MES-P (Xiao, Chen, Dou, Tao, & Chen, 2018), CDESD (Jing, Mao, Chen, & Zhang, 2015) and SAVEE (Jackson, 2011), where the first two are in Mandarin, and the third one is in English. The reasons to choose these three corpora are three fold:

(1) Cross-corpus SER can be investigated for both cases with the same language, and cross languages;

(2) There are four emotional states in common over these three corpora: neutral, happiness, anger, and sadness;

(3) Some differences exist in these three corpora, such as languages, induction means, and recording conditions, which make the three corpora have different feature distributions. Therefore, we can make full use of transfer learning for cross-corpus recognition of these three corpora to enhance the recognition performance.

The information of the three corpora is summarized and compared in Table 2.

Corpus	Language	Num	#m	#f	Style	h:mm	Ave	Emotion classes
MES-P	Mandarin	5376	8	8	acted	6:07	4.1s	moderate and intense versions of joy, anger, sadness and neutral.
CDESD	Mandarin	8400	13	7	acted	3:38	1.6s	sadness, happiness, fear, surprise, neutral, anger, disgust.
SAVEE	English	480	4	0	induced	0:51	3.8s	Anger, disgust, fear, happiness, sadness, surprise, neutral.

Table 2: The detailed information of the three corpora, where the first two are Mandarin corpora and the last one is an English corpus. The number of all utterances in each corpus (Num). Number of male (#m) and female (#f). Style of emotion recording (Style). Total audio time (h:mm) and average duration (Ave) of the utterances in each corpus. Additionally, these three corpora have four common emotions, namely neutral, happiness, anger and sadness.

Mandarin Emotional Speech Dataset Portrayed (MES-P) was built in Soochow University in 2017. It consists of 5376 portrayed utterances from 16 native speakers of Mandarin (8 males, 8 females), where 768 utterances per emotion, on a script of 16 sentences covering all phonemes of Mandarin. The total duration of male speech and female speech are 2h57m and 3h10m, with average duration of 3.9s and 4.2s per utterance, respectively. MES-P possesses two unique features. First, two sets of emotional labels as distal labels (speakers' attention) and proximal labels (listeners' perception) are developed to study the possible distortion in emotional transfer from speakers to listeners. Second, two different levels of emotional intensity ("moderate" *vs.* "intense") are considered, where the term "intensity" refers to how far a person is away from a state of pure, cool rationality, whatever the direction (Gune, Schuller, Pantic, & Cowie, 2011). The intense versions of emotions have strong typicality, which are mainly used in the evaluation of this work in Section 3.3.1. Further, the moderate versions of emotions are also used to verify the effectiveness of DSTL in Section 3.3.2.

Chinese Dual-model Emotional Speech Database (CDESD) is also an acted corpus in Mandarin, developed by Beihang University. Evoking scenes are set to help speakers to better produce utterances for the 7 corresponding emotional states, on a script of 20 sentences, by 20 native speakers of Mandarin (13 males, 7 females). Each sentence is repeated 3 times by each speaker in each emotional state, to result in totally $1200 \times 7 = 8400$ utterances, 1200 utterances for each emotional state. The total duration is 2h15m and 1h23m for male and female speech, with average utterance duration of 1.49s and 1.70s respectively.

Surrey Audio-Visual Expressed Emotion database (SAVEE) is recorded by four native English male speakers, where the emotions are induced by watching the audio-visual videos, on Ekman's 6 basic emotions (Ekman et al., 1987) as anger, disgust, fear, happiness, sadness and surprise, plus neutral. The script consists of 15 phonetically-balanced sentences for each emotion, including 3 common ones, 2 emotion-specific ones, and 10 generic ones. Besides, 3 common and $2 \times 6 = 12$ emotion-specific sentences are also recorded as neutral to obtain totally 30 neutral sentences per speaker. The total number of utterances is 480, with $30 \times 4 = 120$ neutral utterances, and $15 \times 4 = 60$ utterances for each of the other 6 emotions.

Among the 3 corpora, MES-P (768 utterances per emotion) and CDESD (1200 utterances per emotion) are balanced, while SAVEE (120 utterances for neutral, while 60 utterances for other 6 emotions) is imbalanced. For both MES-P and CDESD, the average utterance duration of male speakers is shorter than that of female speakers, indicating that male speakers tend to speak faster when expressing the same emotion. Due to the gender bias in CDESD (fewer females than males) and SAVEE (no female speakers), and the possible emotional expressive difference between the two genders, the experiments on female speech and on male speech are taken out separately in this work. As shown in Table 3, the following eight sets of experiments are designed for cross-corpus emotion recognition. Setting code for each case consists of 3 letters, where the first two capital letters correspond to training set and testing set respectively, and the last lower-case letter means gender. The form of setting code is: training corpus (M/C/S) - testing corpus (M/C/S) - gender (m/f), where "M" refers to MES-P, "C" refers to CDEDS, "S" refers to SAVEE, "m" refers to male speech, and "f" refers to female speech.

3.2 Experiment Setup

To make our work comparable to others in the sense of learning method rather than emphasizing effective features, we use the standard emotional feature sets, such as the feature sets proposed in INTERSPEECH 2009 Emotional Challenge (EC) (Schuller, Steidl, & Batliner, 2009), or in INTERSPEECH 2013 Computational Paralinguistics Evaluation (ComParE) (Schuller et al., 2013), which are widely used in the field of emotion recognition. The INTERSPEECH 2009 feature set, which consists of 384 features, is still used in some latest works (Ma, Wu, Jia, Xu, Meng, & Cai, 2017; Zong, Zheng, Zhang, & Huang, 2016). Therefore, we apply this feature set in our experiments to avoid feature redundancy caused by other larger feature sets.

These features are extracted using OpenSMILE toolkit (Eyben, Wöllmer, & Schuller, 2010), which is written in C++ and enables extraction of large audio features on a time window of 25ms, with a frame shift of 10ms using Hamming window. The 384 features of

Gender	Setting Code	Training set	Testing set
Male	M-S-m	MES-P	SAVEE
	C-S-m	CDESD	SAVEE
	M-C-m	MES-P	CDESD
	C-M-m	CDESD	MES-P
	S-C-m	SAVEE	CDESD
	S-M-m	SAVEE	MES-P
Female	M-C-f	MES-P	CDESD
	C-M-f	CDESD	MES-P

Table 3: Eight sets of experimental settings are designed in our work. In this table, each case is represented by setting code consisting of 3 letters. The first letter in setting code is the training corpus (M/C/S), the second letter is the testing corpus (M/C/S), and the third letter is the gender (m/f), where M, C, S refer to MES-P, CDESD, and SAVEE respectively, and the lower-case letters m and f denote male and female speech, accordingly. For example, M-S-m refers to MES-P as training set, SAVEE as testing set, using male speech. The detailed description about these three corpora can be found in Table 2.

the INTERSPEECH 2009 feature set come for 16 acoustic Low-Level Descriptors (LLDs) and their first order differences (denoted as Δ), such as Zero-Crossing-Rates (ZCR), Root Mean Square frame energy (RMS), Harmonics-to-Noise Ratio (HNR) by autocorrelation function, Mel-Frequency Cepstral Coefficient (MFCC), Fundamental frequency (F0), as well as statistical functions of these LLDs, including mean, standard deviation, extremes and linear regression. We can see the details in Table 4.

LLD	Number	Functionals(12)
ZCR	1	mean
ΔZCR	1	
RMS	1	Energy standard deviation
ΔRMS	1	
F0	1	kurtosis, skewness
ΔF0	1	
HNR	1	Extremes: Value, rel. position, range
ΔHNR	1	
MFCC1-12	12	linear regression: offset, slope, MSE
ΔMFCC1-12	12	

Table 4: Features in INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009): Low-Level Descriptor (LLD) and its functionals. 12 functionals are calculated based on utterance level for each LLD and its first order difference (denoted as Δ).

Other features may also be effective for SER. For example, features based on spectrogram, which can simultaneously consider both time and frequency domain information, could be a good choice in the recognition of some of the emotional states (Prasomphan, 2015). These features will be adopted in our future work when we mainly aim to improve the recognition rate.

In cross-corpus emotion recognition, both intra-corpus discrepancy and inter-corpus discrepancy exist, due to different recording settings and languages (Zhang, Weninger, Wollmer, & Shculler, 2011). Thus, a normalization method for eliminating discrepancy is essential as a pre-process. Z-normalization, (linear scaling to zero mean and unit variance) is adopted in this work.

For comparison purpose, one of the currently most widely used classifiers, Support Vector Machine (SVM), is chosen as the base classifier in all experiments of this work. The kernel function of SVM is chosen as Gaussian kernel. Because in the case when the feature dimension is smaller than the number of samples, Gaussian kernel function outperforms other kernels according to Zhou's (2016) work. In addition, the hierarchical multi-class SVM is adopted in our work, which can determine the order of classification according to the separability of classes, so as to optimize the classification performance (Cheng, Zhang, Yang, & Ma, 2008). Finally, models are trained on the source feature set T_s (Eq. (34)) and tested on the target feature set T_t (Eq. (35)) to predict the class labels on a set of parameters.

The following parameters are tested to optimize the DSTL framework:

- α - Importance Ratio Parameter (IRP) of MPCA;
- γ - MMD regularization parameter in Eq. (19);
- β - GE regularization parameter in Eq. (20);
- p - the number of nearest neighbors;
- k - feature dimension for common features;
- h - feature dimension for specific features.

Note that in our task, the training and testing sets come from different corpora, and no overlap is possible between training and testing sets, and hence it is unnecessary to choose hyperparameters for cross validations. Song's (2017) work empirically searches the parameter space for the optimal parameters to evaluate all methods. Therefore, in our work, two of the above parameters are empirically set as the number of nearest neighbors $p = 5$, and the IRP $\alpha = 0.8$. Other parameters are swept over a range of values to find out the optimal values as shown in Table 5. In this table, we find the optimal values for all parameters by searching the corresponding parameter space, to report the best results in Section 3.3. Finally, the parameters γ, β, k, and h are optimized as 100, 1000, 80, 80 respectively. The sensitivities of γ and β are further discussed in Section 3.4.

3.3 Results Analysis

The experimental results are analyzed in this subsection.

Parameters	Range	Optimal value
γ	{0.001, 0.01, 0.1, 1, 10, 100, 1000,10000}	100
β	{0.001, 0.01, 0.1, 1, 10, 100, 1000,10000}	1000
k	{40, 60, 80, 100, 120}	80
h	{40, 60, 80, 100, 120}	80

Table 5: Experimental parameters setup. According to the similar work (Song, 2017; Song & Zheng, 2017), the ranges of these hyperparameters are set. We search the optimal parameters to get the better performance by evaluating all methods.

3.3.1 Cross-corpus Speech Emotion Recognition on Four Emotional States

The performance of cross-corpus emotion recognition is evaluated in 3 groups of experiments:

(1) Common methods: Machine learning methods without transfer information between source and target domains. Two approaches are chosen, including a classic method, PCA, and a deep learning method, RNN (Recurrent Neural Network, Lee & Tashev, 2015.)

(2) Baseline transfer learning methods: transfer learning methods using only the common information, include M-MPCA, G-MPCA, and MG-MPCA.

(3) DSTL methods proposed in this work, include D-M-MPCA, D-G-MPCA, and D-MG-MPCA.

To evaluate the performance, the recalls are selected as dominant evaluation bases for each case. For the 3 chosen corpora in this work, MES-P and CDESD are balanced, while SAVEE is imbalanced. Two types of recalls are considered as weighted or unweighted average recall. Weighted Average Recall (WAR) is defined as the total number of correctly predicted test samples of all class averaged by the total number of test samples; Unweighted Average Recall (UAR) is defined as the accuracy per class averaged by total number of class. For balanced corpus as testing set, WAR and UAR are the same. However, for imbalanced corpus as testing set, both WAR and UAR are important for considering its performance. Thus, same-language cases only involving the balanced corpora (*i.e.,* MES-P, CDESD) as testing set are measured by WAR, which are listed in Table 6. However, cross-language cases involve two different aspects: balanced corpus as testing set (*i.e.,* S-M-m, S-C-m) and imbalanced corpus as testing set (*i.e.,* M-S-m, C-S-m). Therefore, to evaluate the performance of cross-language cases uniformly, WAR is first evaluated in Table 7. While for M-S-m and C-S-m cases, both WAR and UAR are then calculated and compared in Figure 3. Besides, confusion matrices of WAR for D-MG-MPCA are visualized in Figure 5, and Weighted Average Precision (WAP) is also provided.

In Table 6 and Table 7, among these approaches, the PCA and RNN methods directly use the knowledge learned from the source domain to the target domain, ignoring the differences between domains, while all the other methods belong to transfer learning methods. In most cases, transfer learning approaches show better performance than PCA and RNN, due to

	Common Methods		Transfer Learning					
			Baseline Methods			DSTL		
Setting Code	PCA	RNN	M-MPCA	G-MPCA	MG-MPCA	D-M-MPCA	D-G-MPCA	D-MG-MPCA
M-C-m	44.00%	40.93%	45.29%	49.49%	51.25%	46.31%	51.76%	**52.05%**
C-M-m	33.07%	46.61%	45.25%	43.88%	43.55%	**47.85%**	47.07%	47.59%
M-C-f	52.02%	47.85%	52.92%	52.50%	59.83%	56.25%	55.89%	**60.74%**
C-M-f	43.29%	42.14%	42.08%	46.29%	47.07%	47.33%	**47.46%**	47.39%
Average	42.42%	42.06%	45.14%	48.04%	50.43%	49.44%	50.55%	**51.94%**

Table 6: WARs of different methods (common methods *vs.* baseline methods *vs.* DSTL framework) under same-language setting. The entry in bold for each row means the best performance for corresponding case. Note that the improved methods D-X-MPCA gain the better performance than their corresponding baseline methods X-MPCA, where X refers to M, G, MG. For example, the D-M-MPCA method outperforms M-MPCA in all cases, and the same goes for the pairs D-G-MPCA *vs.* G-MPCA, and D-MG-MPCA *vs.* MG-MPCA. Besides, the D-MG-MPCA method has the best average performance as 51.94%. The precision and recall of D-MG-MPCA can be seen in Figure 5.

	Common Methods		Transfer Learning					
			Baseline Methods			DSTL		
Setting Code	PCA	RNN	M-MPCA	G-MPCA	MG-MPCA	D-M-MPCA	D-G-MPCA	D-MG-MPCA
M-S-m	55.67%	36.33%	56.00%	59.33%	59.67%	57.67%	61.33%	**61.67%**
S-M-m	38.54%	40.42%	41.35%	42.38%	42.19%	**46.88%**	46.03%	46.68%
C-S-m	53.67%	41.33%	49.67%	**56.33%**	53.33%	53.00%	50.67%	51.67%
S-C-m	43.59%	41.46%	43.95%	45.16%	45.00%	46.54%	**47.40%**	46.92%
Average	47.87%	39.88%	47.74%	50.80%	50.05%	51.02%	51.36%	**51.76%**

Table 7: WARs of different methods (common methods *vs.* baseline methods *vs.* DSTL framework) under cross-language setting. In this table, the meaning of boldface also means the best performance for corresponding case (same with Table 6). It is clear that (with one exception, *i.e.,* C-S-m) the rates of DSTL are higher than the corresponding baseline methods, especially for the rate of M-S-m that reaches 61.67% under D-MG-MPCA method. Overall, D-MG-MPCA method gains the best average WAR as 51.76%. Besides, the comparison of WARs and UARs for C-S-m and M-S-m cases is shown in Figure 3. The precision and recall of D-MG-MPCA can also be seen in Figure 5.

the ignorance of domain differences in the traditional methods. An exception appears in the case of C-S-m in Table 7, that PCA exhibits better recall than M-MPCA. A possible

reason may be that the global analysis of the MMD may have a negative impact for the imbalanced SAVEE corpus.

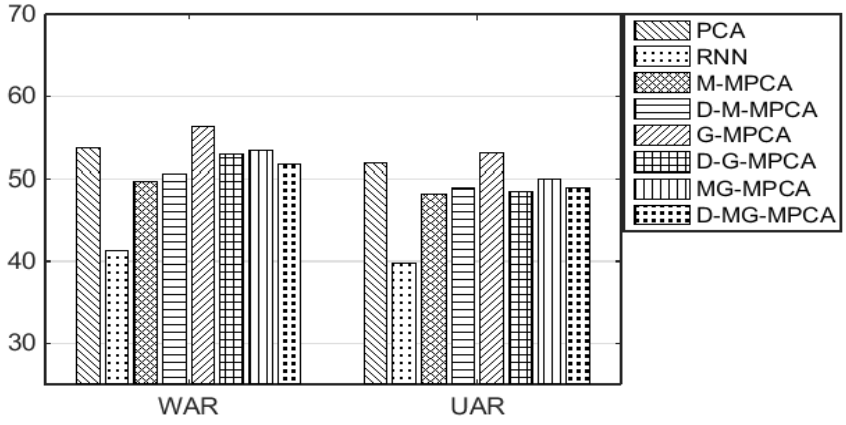
In comparison of the transfer learning methods, the DSTL approaches also significantly outperform the baseline methods in most cases. The most obvious improvement occurs to the pair M-MPCA *vs.* D-M-MPCA, with improvement of average recalls of 4.30% and 3.28% for the cases of within-language and cross-language, respectively. The best average WARs are obtained in D-MG-MPCA, as up to 51.49% and 51.76% with the utmost as 61.67% for M-S-m. This proved that the proposed DSTL framework (common-space finding + SMT) can not only consider the commonalities to reduce the distribution differences, but also make full use of the supervised specific knowledge of source domain in target subspace. Actually, DSTL, as a general framework to combine common information and domain-specific information, can also be applied to other common space finding methods, such as the methods proposed in Deng et al. (2014) and Zong et al.'s (2016) works.

There is only one case, *i.e.,* C-S-m, that the best recall does not appear in the DSTL methods. As shown in Table 7 and Figure 3(a), for C-S-m case, the best WAR and UAR both appear in G-MPCA method. This can be explained that the distribution difference between the two corpora with different languages is too huge to make effective specific information transfer from CDESD to SAVEE by SMT. Thus, it can lead to the conclusion that DSTL is more suitable for the within-language than cross-language emotion recognition, while it may risk failing in dealing with domain specific information when the corpora difference is excessively huge. The specific information transferring scheme still need to be further improved in our future work.

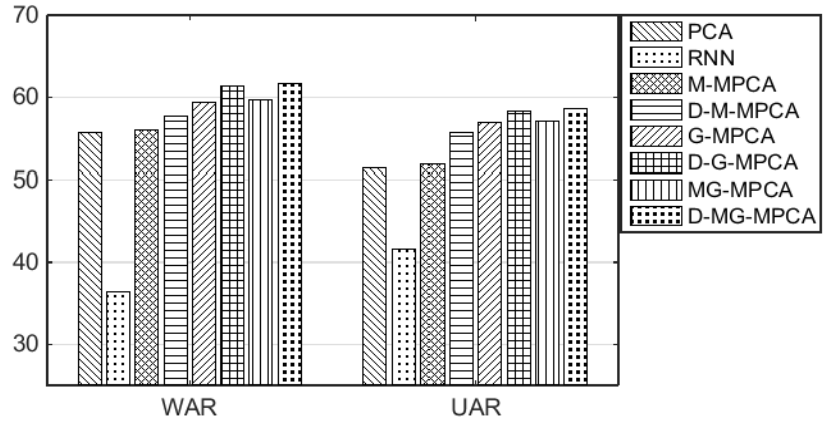
To compare the within-language and cross-language performance of DSTL methods, the WARs in the last three columns of Table 6 and Table 7 are illustrated in Figure 4. For the cases involving MES-P, the recognition performance is better when it is used as training set. For example, M-S-m performs better than C-S-m, and the same goes for M-C-f *vs.* C-M-f. The highest WARs are M-C-f and M-S-m with D-MG-MPCA, up to 60.74% and 61.97% respectively. The only exception is M-C-m *vs.* C-M-m with D-M-MPCA. This tendency indicates that MES-P corpus exhibits strong typicality and distinction in emotion expressing. Another pattern presented in Figure 4 is that there is no significant difference between the performance between within-language and cross-language recognition. Thus, the methods proposed in this work can learn the mapping relationships between different languages. The WARs of S-C-m and S-M-m are slightly lower than the other cases, probably due to the extreme small scale of the SAVEE corpus.

The confusion matrices of WAR for all eight different cases with D-MG-MPCA are shown in Figure 5, where Weighted Average Precision (WAP) and WAR are also included for each case in the caption. The diagonal numbers of each confusion matrix represent the correct WAR for each emotion under the corresponding case. The correct WAR of each emotion is higher than chance level ($1/4 = 0.25$) for all cases, and even reaches 76% for happiness emotion in M-C-f. The highest confusion occurs between neutral and sadness, even over 40% for M-C-m. The second highest confusion occurs between anger and happiness, with up to 38% of happy utterances misjudged as anger for C-S-m. We will attempt to introduce some new effective features to better categorize these highly confusing emotional pairs.

In addition, we summarize the statistical significance of the results using rank sum test (we define the significance level as $p = 0.05$). The best improved method D-MG-MPCA is



(a). C-S-m.



(b). M-S-m.

Figure 3: The WAR and UAR of C-S-m and M-S-m cases under different methods. The testing set of the two cases is imbalanced corpus SAVEE, hence WAR and UAR are both considered. In Figure 3(a), G-MPCA performs better than other methods in that SAVEE, as an imbalanced database, has a bias when using MMD. In addition, specific information may have negative impact in this case due to the huge difference. In Figure 3(b), the improved methods have significant improvement compared to the corresponding baseline methods.

compared with the best common method PCA and the baseline methods (*i.e.,* M-MPCA, G-MPCA, MG-MPCA), respectively. We observe that the D-MG-MPCA has a statistical significance at the 0.03 level compared with PCA method. Compared with the other baselines methods, *i.e.,* M-MPCA and G-MPCA, D-MG-MPCA passes the significant test at the 0.005 and 0.03 levels against the M-MPCA and G-MPCA methods, which indicates D-MG-MPCA has significant improvement. However, D-MG-MPCA shows the significance level

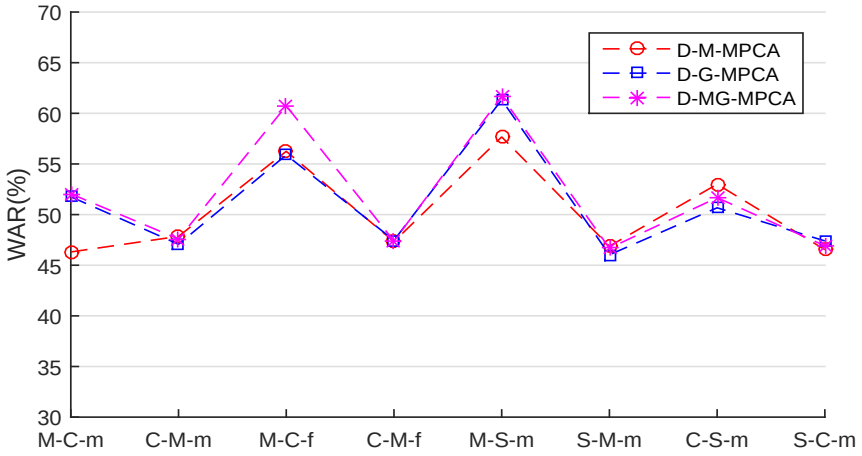


Figure 4: The WARs of three improved methods under different cases. In this figure, we can see that when MES-P is used as training set, the recognition rates are better than that of other corpora used as training set in most cases (with only one exception, *i.e.,* M-C-m *vs.* C-M-m under D-M-MPCA method). For example, M-S-m outperforms C-S-M, M-C-m outperforms C-M-m, and M-C-f outperforms C-M-f. In addition, there is no significant difference between same-language and cross-language cases in terms of performance, indicating that the proposed methods can learn the mapping relationships between different languages.

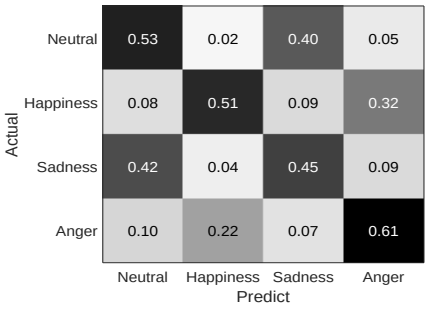
at 0.16 against the MG-MPCA, indicating that the proposed approach has improvement but fails to achieve a customary level of statistical significance compared to MG-MPCA.

3.3.2 Recognition Performance on Moderate Emotional States

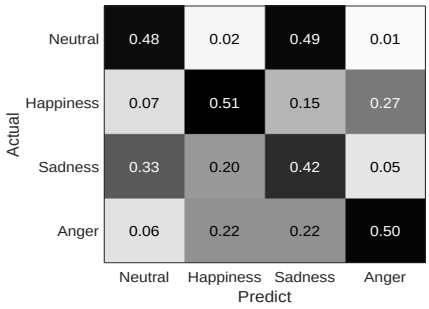
MES-P includes not only intense versions of emotions, but also moderate versions of emotions. Thus, we conduct the experiments for moderate versions of emotions to verify the validity of the DSTL framework using the best baseline method MG-MPCA and the best improved method D-MG-MPCA, as shown in Figure 6. This figure shows the WARs of all cases involving MES-P using moderate versions of emotion. From this figure, the overall WARs of moderate versions are not as high as that of intense versions of emotions which have more obvious emotional characteristics. However, it is clear to see that the D-MG-MPCA method still outperforms MG-MPCA method. The results show that the improved method does have the better performance than the baseline method in both moderate and intense versions of emotions.

3.4 Discussion on Regularization Parameter Sensitivity - γ and β

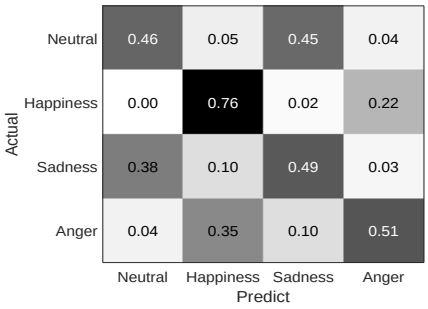
In the two methods of discrepancy measurement adopted in this work, MMD and GE, there is a regularization parameter for each of the methods: γ for MMD in Eq. (19), and β for GE in Eq. (20). The two parameters are evaluated on a series of values: γ on



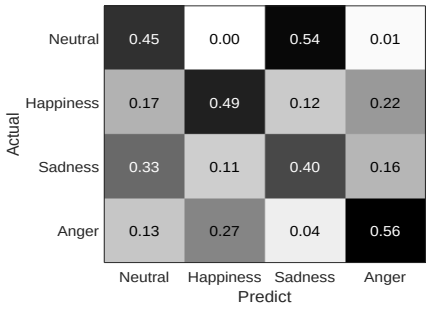
(a). M-C-m (WAP: 53.26% WAR: 52.05%)



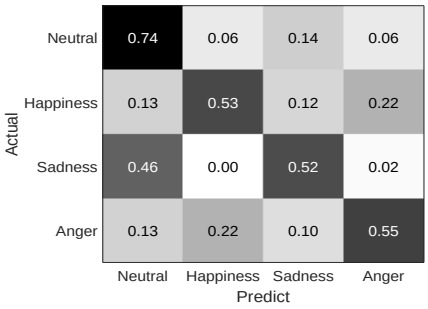
(b). C-M-m (WAP: 49.48% WAR: 47.59%)



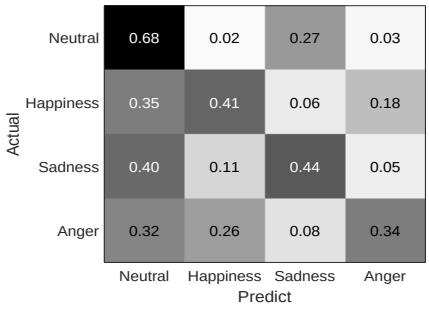
(c). M-C-f (WAP: 53.37% WAR: 60.74%)



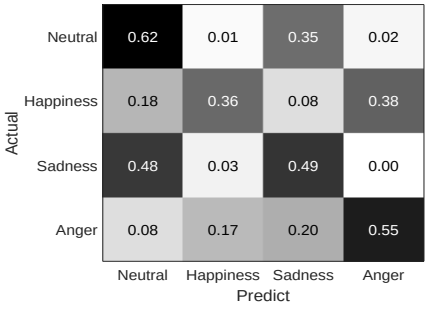
(d). C-M-f (WAP: 48.33% WAR: 47.34%)



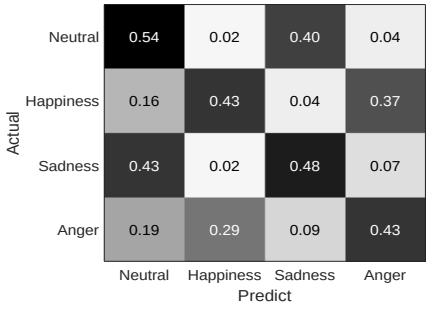
(e). M-S-m (WAP: 60.10% WAR: 61.67%)



(f). S-M-m (WAP: 49.27% WAR: 46.68%)



(g). C-S-m (WAP: 53.36% WAR: 51.67%)



(h). S-C-m(WAP: 46.29% WAR: 46.92%)

Figure 5: Confusion matrices of WAR under D-MG-MPCA for different cases, where Weighted Average Precision (WAP) and recall (WAR) are given for each case in the caption. Note that the recognition rate of each emotion (diagnoal numbers in each confusion matrix) is higher than chance level. Neutral-sadness, happiness-anger are easily confused with each other.

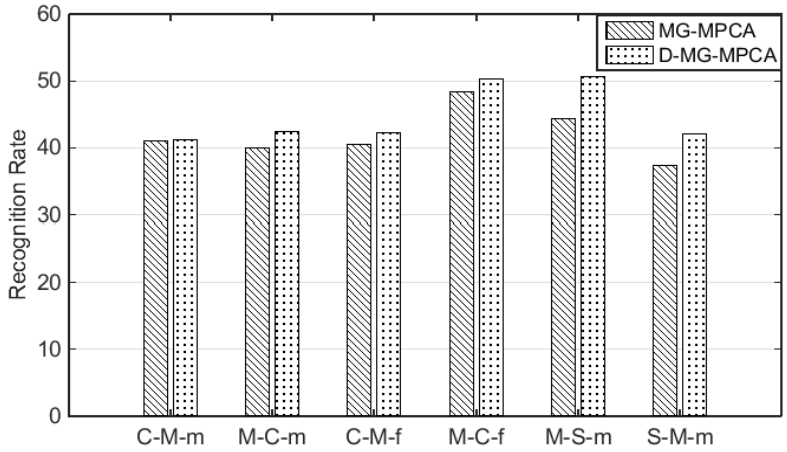


Figure 6: The WARs of all cases involving MES-P corpus using moderate versions of emotions. This figure shows that the improved method D-MG-MPCA has significant improvement compared to MG-MPCA under all cases.

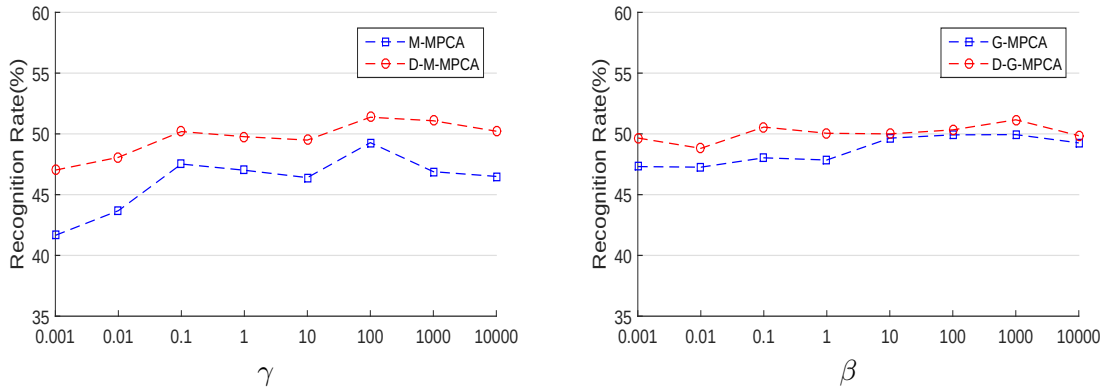
approaches M-MPCA and D-M-MPCA, and β on approaches G-MPCA and D-G-MPCA. The performance is illustrated in Figure 7.

According to Figure 7, performance from D-X-MPCA is always better than those from X-MPCA (X for M or G), with all regularization parameter settings. For γ, M-MPCA and D-M-MPCA have similar tendencies, and both reach optimal recognition at $\gamma = 100$. For β, the tendencies of G-MPCA and D-G-MPCA differ a lot with smaller βs, while become relatively stable for larger βs, and $\beta = 1000$ could be seen as an optimal value for both approaches. Thus, the two regularization parameters are optimized to $\gamma = 100$ and $\beta = 1000$ for all the experiments involving them in Section 3.3.

4. Discussion

Some of the important research issues in SER are briefly discussed below:

- Three different corpora are selected to verify the validity of our proposed methods, where two of them are acted corpora, while the remaining one is an induced corpus. However, the real challenge is to recognize emotions from natural speech. Some differences are existed between acted/induced corpora and the natural speech that people spontaneously express in real life. Compared with acted/induced corpora, natural speech is mildly expressed, thus sometimes it may be difficult to clearly recognize these emotions. To improve the recognition rate of natural speech, two possible ways can be considered: on one hand, the expression of emotions is not only included in speech, but also in other modalities like bio-signals, facial expression, and thus multimodal recognition may dominate in emotion recognition in the future; on the other hand, emotion information can be predicted from the linguistic contents of speech, and hence, the addition of textual information may be more helpful in recognizing emo-



(a). γ for M-MPCA and D-M-MPCA (b). β for G-MPCA and D-G-MPCA

Figure 7: The recognition performance of our proposed methods varies with parameters γ and β : (a) The recognition performance of M-MPCA and D-M-MPCA with respect to γ ; (b) The recognition performance of G-MPCA and D-G-MPCA with respect to β ;

tions. Besides, the proposed DSTL framework is not limited to these acted/induced corpora, but also can be used to improve the performance of spontaneous corpus.

- Most existing corpora contain only a limited number of emotional types. Research shows that about 70% databases contain only 4-5 basic emotions and few emotional speech databases contain 7-8 emotions (Koolagudi & Rao, 2012). However, human can express a wide range of emotions, the number of which far exceeds that in most corpora. Realization of this, collection of good emotional speech corpora covering a wider range of emotions is another challenge. One possible and easy-to-realize way is to capture the emotional segments from the TV dramas, which are usually performed by professional actors and convey rich and natural emotional states.

- Although our work uses DSTL framework for cross-corpus emotion recognition in terms of Mandarin and English, this framework can also be extended to resource-poor areas or other areas. For instance, there are few databases for the following languages, like Russian, Swedish, Japanese, and Indian. The DSTL framework can also help to recognize the task of these resource-poor languages based on some resource-rich language models (*i.e.,* Mandarin, English, German). In addition, other areas, like face recognition, natural language processing, can also use DSTL framework to improve recognition performance.

- The DSTL includes two completely separate steps: common space finding methods and specific information transferring method as SMT. Thus, the specific information obtained by SMT may exist some overlapping or redundant features with common features. To deal with this issue, the solution will be implemented in the future work: excluding the common features from the whole features before SMT, then obtaining the specific information from the rest features by SMT. The redundancy of common

features and specific features obtained in this way will be reduced to enhance the performance.

- The methods proposed in this paper are based on linear assumption. Firstly, the distance measure (inter- and intra-) and PCA can perform well under the assumption that the source and target spaces are linear. However, some kernel mapping will be introduced for non-linear condition, so that the source and target spaces become linear in a high dimensional space (Yan et al., 2005). Secondly, the mapping matrix learned by SMT method can transfer the source-specific information to target subspace effectively when there is a linear transformation between the two domains. However, for non-linear transformation, some deep learning methods, like autoencoders, may be helpful to learn the mapping relationships between the two domains (Deng et al., 2014).

- In this work, the problems we solved are that the source and target domains are from the same field, *i.e.*, speech emotion, but with different feature distributions. Thus, we adopt a same emotional feature set for the both datasets, resulting in the same feature dimensions for the two domains. However, in the future work, the proposed DSTL framework will be extended to solve this problem when there are different feature dimensions for the source and target domains.

5. Conclusion

In this paper, a general framework called Dual-Subspace Transfer Learning (DSTL) has been proposed for cross-corpus speech emotion recognition, which compensates for the shortcoming of most current common/feature-based methods by exploiting specific information of the two domains.

Based on several common/feature-based methods (*i.e.*, M-MPCA, G-MPCA, MG-MPCA) as baselines, we propose a mapping strategy as SMT, to transfer the source-specific information to the target subspace. The combination of the common/feature-based methods and the SMT constructs a DSTL framework, presented as improved methods as D-M-MPCA, D-G-MPCA, and D-MG-MPCA. Both the baselines and the improved methods are evaluated on three corpora upon eight settings of cross-corpus emotion recognition task, including within-language cases and cross-language cases. The results prove that the dual-subspace solution proposed in this work does present obvious advantage compared to the baselines which only concern the common space. The DSTL framework can also be assembled with other common/feature-based methods for amelioration, and can be applied to other fields, such as face recognition, natural language processing.

The results of this work also propose us several further tasks to be solved in the future as discussed in Section 4. Therefore, future work will tend to build a standard corpus with more emotions, and multimodal recognition will dominate this field to contain more effective features.

References

Cheng, L., Zhang, J., Yang, J., & Ma, J. (2008). An improved hierarchical multi-class support vector machine with binary tree architecture. In *Internation Conferernce on Internet Computing in Science and Engineering*, pp. 106–109.

Daumé III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, *26*(2006), 101–126.

David, S. B., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. In *International Conference of Neural Information Processing System*, pp. 137–144.

Deng, J., Xia, R., Zhang, Z., Liu, Y., & Shculler, B. (2014). Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 4818–4822.

Deng, J., Xu, X., Zhang, Z., Frühholz, S., & Schuller, B. (2018). Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(1), 31–43.

Deng, J., Zhang, Z., & Schuller, B. (2014). Linked source and target domain subspace feature transfer learning − exemplified by speech emotion recognition. In *International Conference on Pattern Recognition*, pp. 761–766.

Ekman, P., Friesen, W., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W., Pitcairn, T., Ricci-Bitti, P., Scherer, K., & Tomita, M. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712–717.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). opensmile - the munich versatile and fast open-source audio feature extractor. In *Association for Computing Machinery*, pp. 1459–1462.

Fernando, B., Habard, A., Sebban, M., & Tuytelaars, T. (2013). Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2960–2967.

Gasulla, D. G., Parés, F., Vilalta, A., Moreno, J., Ayguadé, E., Labarta, J., Cortés, U., & Suzumura, T. (2018). On the behavior of convolutional nets for feature extraction. *Journal of Artificial Intelligence Research*, *61*(2018), 563–592.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*(2016), 345–420.

Gune, H., Schuller, B., Pantic, M., & Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *IEEE International Conference on Automatic Face & Gesture Recognition & Workshops*, pp. 827–834.

Hartley, R. (1997). In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(6), 580–593.

Jackson, J. R. (2011). Surrey audio-visual expressed emotion (savee) database. http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE.

Jing, S., Mao, X., Chen, L., & Zhang, N. (2015). Annotations and consistency detection for chinese dual-mode emotional speech database. *Journal of Beijing University of Aeronautics and Astronautics*, *41*(10), 1925–1934.

Kan, J., Wu, J., Shan, S., & Chen, X. (2014). Domain adaptation for face recognition: Targetize source domain bridged by common subspace. *International Journal of Computer Vision*, *109*(1-2), 94–109.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology*, *15*(2), 99–117.

Lee, H., & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *INTERSPEECH*, pp. 1537–1540.

Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., & Cai, L. (2017). Speech emotion recognition with emotion-pair based framework considering emotion distribution information in dimensional emotion space. In *INTERSPEECH*, pp. 1238–1241.

Noda, T., Yano, Y., Doki, S., & Okuma, S. (2006). Adaptive emotion recognition in speech by feature selection based on kl-divergence. In *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1921–1926, Taipei, Taiwan.

Pan, S. J., Kwok, J. T., & Yang, Q. (2008). Transfer learning via dimensionality reduction. In *Association for the Advancement of Articifical Intelligence*, pp. 677–682.

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

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transaction on Knowledge and Engineering*, *22*(10), 1345–1359.

Prasomphan, S. (2015). Detecting human emotion via speech recognition by using speech spectrogram. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*.

Safdarkhani, M. K., Mojaver, S. P., Atieghechi, S., Molanoori, A., & Riahi, M. S. (2012). Emotion recognition of speech using ann and gmm. *Australian Journal of Basic and Appplied Sciences*, *6*(9), 45–57.

Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. In *INTERSPEECH 2009 BRIGHTON*, pp. 312–315.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., & Marchi, E. (2013). The interspeech 2013 computational paralinguistics challenage: social signals, conflict, emotion, autism. In *INTERSPEECH*.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, *1*(2).

Si, S., Tao, D., & Geng, B. (2010). Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(7), 929–942.

Song, P. (2017). Transfer linear subspace learning for cross-corpus speech emotion recognition. IEEE Transactions on Affective Computing, to be published.

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

Song, P., & Zheng, W. (2017). Feature selection based transfer subspace learning for speech emotion recognition. IEEE Transactions on Affective Computing, to be published.

Tong, R., Wang, L., & Ma, B. (2017). Transfer learning for children' s speech recognition. In *International Conference on Asian Language Processing*, pp. 36–39.

Vaishali, M. C., & Gohokar, V. V. (2012). Speech emotion recognition by using svm-classsifier. *International Journal of Engineering and Advanced Technology*, *1*(5), 11–15.

Vlassis, N., & Likas, A. (2002). A greedy-em algorithm for gaussian mixture learning. *Neural Processing Letters*, *15*(1), 77–87.

Wang, D., & Zheng, T. (2015). Transfer learning for speech and language processing. In *Proceeding of APSIPA Annual Summmit and Conference*, pp. 1225–1237.

Xiao, Z., Chen, Y., Dou, W., Tao, Z., & Chen, L. (2018). Mes-p: an emotional tonal speech dataset in chinese mandarin with distal and proximal labels. IEEE Transactions on Affective Computing, Submitted.

Xu, Y., Fang, X., Wu, J., Li, X., & Zhang, D. (2016). Discriminative transfer subspace learning via low-rank and sparse representation. *IEEE Transactions on Image Processing*, *25*(2), 850–863.

Yan, S., Xu, D., Zhang, B., & Zhang, H. (2005). Graph embedding: a general framework for dimensionality reduction. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 830–837.

Yang, P., & Gao, W. (2014). Information-theoretic multi-view domain adaptation: A theoretical and empirical study. *Journal of Artificial Intelligence Research*, *49*(2014), 501–525.

Zhang, B., Provost, E. M., & Essl, G. (2016). Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 5805–5809.

Zhang, S., Zhang, S., Huang, T., & Gao, W. (2018). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, *20*(6), 1576–1590.

Zhang, Z., Weninger, F., Wollmer, M., & Shculler, B. (2011). Unsupervised learning in cross-corpus acoustic emotion recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding(ASRU)*, pp. 523–528.

Zhou, Z. (2016). *Machine Learning*. Tsinghua University Press, Beijing.

Zong, Y., Zheng, W., Zhang, T., & Huang, X. (2016). Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. *IEEE Signal Processing Letters*, *23*(5), 585–589.