IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 May 2023, accepted 1 June 2023, date of publication 5 June 2023, date of current version 12 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3282982

RESEARCH ARTICLE

Speech Emotion Recognition Based on Transfer Emotion-Discriminative Features Subspace Learning

ZHANG KEXIN AND LIU YUNXIANG
Department of Computer Science, Shanghai Institute of Technology, Shanghai 201418, China
Corresponding author: Zhang Kexin (1277767811@qq.com)

ABSTRACT Cross-corpus speech emotion recognition(SER) is a hot topic in emotion classification. Cross-corpus SER includes these four issues:feature selection, differences constraint, label regression and preservation of discriminative emotion features. Seldom literature can solve these four issues jointly in previous studies.In this work,we propose the transfer emotion-discriminative features subspace learning(TEDFSL) method.Acoustic features are extracted by the OpenSMILE in the source and target data. Then the extracted features are sent into CNN+BLSTM to learn higher-level global features and time series. The common low-dimensional subspace of the source data and target data is learned by Linear Discriminant analysis (LDA) to reduce the dimension and Maximum Mean Discrepancy (MMD) and Graph Embedding (GE) to constraint the differences between source data and target data. The common low- dimensional subspace is combined with the label regression matrix to learn the relationship between labels and features,after which the, DNN is selected as the final classifier to preserve emotion-discriminative features, emotion-aware center loss(l_c) is added and extensive experiments are carried out on cross-corpus SER tasks and the results demonstrate that our proposed method is superior to state-of-art cross-corpus SER.

INDEX TERMS Cross–corpus speech emotion recognition, maximum mean discrepancy, graph embedding, label regression matrix, emotion-aware center loss.

I. INTRODUCTION

With the development of artificial intelligence, making computers interact with humans in the most natural way has become a research hotspot [1], [2], [3]. The earliest research on SER was conducted in the 1980s. In 1999, Moriyama realized an interface that could recognize users' emotions at an e-commerce interface [4]. Since the beginning of the 21st century, an increasing numbers of scholars have conducted research on speech emotion recognition, including creating SER competitions, creating a journal of emotion classification, researching open-source tools related to emotion feature extraction, and conducting academic conferences on SER. In 2005, the International Conference on Affective Computing and Intelligence Interaction was held [5]. The INTERSPEED EMOTION Challenge [6] was established in 2009. In 2011, the Audio/Visual Emission Challenge and Workshop (AVEC) was established. In 2010, the IEEE Transactions on Affective Computing, a journal related to emotion classification, was established. Eyben developed the openSMILE and openEAR tools to extract acoustic features [7].

Speech emotion classification initially uses machine learning classifiers, such as support vector machine (SVM), decision tree, and random forest [8], [9]. However, machine learning has a poor ability to extract complex features. Deep learning models, such as convolution neural network (DNN), deep confidence neural network (DBN), and bidirectional cyclic memory neural network (BLSTM), can extract deeper features through deeper neural networks and the nonlinear transformation of activation functions on input features [10], [11], [12], [13]. Therefore, the deep learning

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

classifier has gradually replaced the machine learning classifier.

The experimental results of these algorithm models were better when the source data and the target data were the same. However, in reality, databases are often recorded in different environments. There are differences in social background, sex, and age between the different databases. The influence of these factors leads to a decline in the generalization ability of the models. Therefore, this problem has prompted some scholars to develop cross-corpus SER. Chen et al. [14] considered that many sutdies only considered the common knowledge of the target and source domains, while ignoring the specific information of the two domains. Therefore, they proposed dual subspace transfer learning (DSTL) to utilize the two types of information. DSTL can not only alleviate domain differences, but also makes good use of specific information. Li et al. [15] used a two-hemisphere adversarial neural networks (BIDANN). Two hemispheres, one controls the source data and the other controls the target data. The global discriminator attempts to reduce the possible domain differences between the source domain and the target domain in each hemisphere to achieve cross-corpus SER. Ocquaye et al. [16] proposed a dual exclusive attention transfer method combined with correlated alignment loss. The function of the correlation alignment is to minimize the domain offsets. Song et al. [17] proposed transfer subspace learning based on non-negative matrix decomposition to find the shared feature subspaces between the source data and target data. Only in this way can the information of the source data be transferred to the target data and the differences be eliminated.However,the limitation of these methods are as follows:(1) Feature selection, global difference constraint, local difference constraint, and the mapping relationship between features and labels are all critical components of transfer subspace learning.These methods can not solve these four issues jointly.(2) Many transfer learning methods ignore the emotion discriminative features while reducing the differences between source and target data.

In this study,we proposed the TEDFSL method. We set the projection matrix to Q and the regression matrix to P. The function of projection matrix is to find the common feature subspace of source data and target data, and the function of regression matrix is to establish the regression model, then learn the mapping relationship between features and labels.In TEDFSL method, we optimized the total loss of MMD+GE+LDA+label regression(LR) by finding optimal projection matrix Q and regression matrix P then we added the $\mathbf{l_c}$.The contributions of our paper are as follows: (1) TEDFSL method solved feature selection, differences constraint and label regression jointly in low- dimensional common transfer subspace learning,which is superior to previous transfer subspace learning methods.(2)Previous transfer learning methods narrowed the differences between source data and target data excessively and ignored the emotion discriminative features.$\mathbf{l_c}$ is considered in the TEDFSL method.

II. RELATED WORK

The basic idea of transfer subspace learning is to map the source data and target data from a high-dimensional feature space to a common low-dimensional subspace by using the difference constraint method. Transfer subspace learning typically includes three components: feature selection, label information mapping and regression, and feature distribution similarity constraints. The function of feature selection is to ensure the formation of low-dimensional subspaces and to avoid dimension disasters. Because many deep learning models require a large amount of labeled data.However, it is difficult to obtain a large amount of labeled data [18], [19], and the function of label regression is to solve the problem of unlabeled target data. Through the regression matrix, the feature is mapped to the label space, so that the label information of the source data is transferred to the target data. The function of the feature distribution differences constraint is to reduce the difference in the feature distribution between the source and target data. Common constraints include MMD constraint and GE constraint. Among them, MMD is the global difference constraint algorithm and GE is the local difference constraint algorithm. In this subspace, not only is the feature distribution of the source data and the target data being similar, but the feature dimension can also be reduced to avoid dimension disaster, so that the knowledge learned from the source dataset can be transferred to the target dataset more effectively. Song et al. [20] proposed a transfer discriminative analysis (TDA) method. The basic idea of TDA is to combine the LDA subspace dimension reduction method with the MMD algorithm to form a low- dimensional subspace with a similar feature distribution between the source target datasets. Chen et al. [21] proposed a target adaptive subspace learning (TaSL) method. TaSL uses $\mathbf{l_1}$ Normal form as a label regression, and the $\mathbf{l_{2,1}}$ normal form is used to reduce the difference between the source and target datasets.The TaSL subspace can accurately predict the feature labels. Liu et al. [22] proposed the method of transfer subspace learning (TRaSL). TRaSL converts the feature space of the source and target datasets into a label space. In label space, the feature distributions are similar. In their method, the source dataset was labeled, while the target dataset was unlabeled. TRaSL can effectively use the labels of the source dataset to predict the emotional category of the target dataset by training it. Zhang and Song [23] proposed a transfer sparse discriminative subspace learning (TSDSL) method. To obtain representative features between different corpora, they use $\mathbf{l_{2,1}}$ Paradigm.To take advantage of the correlation between different corpora, they proposed a new nearest- neighbor graph as a distance metric. Compared to the MMD algorithm, the nearest neighbor graph can retain the local geometric structure of the data.

However,these methods can not include feature selection, label information mapping and regression, and feature distribution similarity constraint simultaneously.Reducing either of them will lead to poor performance of cross-corpus SER.

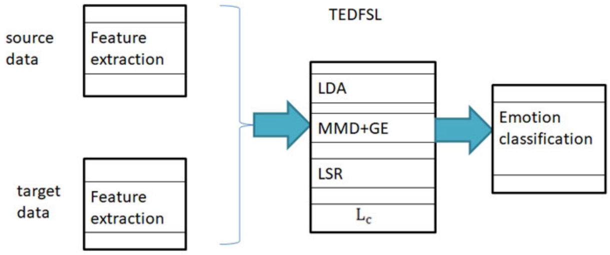

FIGURE 1. The structure of TEDFSL.

III. OUR PROPOSED MODEL

The structure of the TEDFSL is shown in Figure 1.The TEDFSL includes three components: feature extraction, lowdimensional common subspace learning and emotion classification. Low-dimensional common subspace learning includes LDA,MMD+GE,LSR and $\mathbf{l_c}$.In the stage of emotion classification stage,DNN was selected as an emotion classifier.

A. SYMBOL DESCRIPTION

D^s , D^t represents the source and target domains respectively. n_s represents the number of source domain samples,n_t represents the number of target domain samples, $X_s = [x_1, \ldots x_{n_s}]$ represents the source domain sample data,$X_t = [x_1, \ldots x_{n_t}]$ represents the target domain sample data, Y^s =[y_1^s,y_2^s, . . .y_n^s] represents the sample label of the source dataset.Q represents the projection matrix and P represents the regression matrix.

B. OBJECTIVE FUNCTION OF TEDFSL

Some traditional transfer learning methods only consider reducing the domain differences between the source data and target data, while ignoring the emotional discrimination features. In TEDFSL,we got a common subspace of the features by combining LDA,MMD+GE and LSR.In this subspace,the dimension of the features was reduced to 150 by LDA and it preserved commonality between domains by MMD+GE and it can learn the relationship between the feature representation and the corresponding label by LSR Although DNN is used as final classifier,we need label regression matrix P to describe the relationship between the feature representation and the corresponding label.If we remove label regression matrix p,the label information can not be utilized.Combining label regression matrix P with DNN is a kind of secondary classification.Moreover, secondary classification can have higher accuracy compared to primary classification.After that,$\mathbf{l_c}$ was added to preserve emotion-discriminative features in subspace. In this study,the openSMILE tool package was used to extract effective emotional features, and the feature set of the INTERSPEECH2010 Language Challenge was selected, with a feature dimension of 1582.

1) FEATURE EXTRACTION

The features of the INTERSPEECH2010 Language Challenge are shown in Table 1. We inputted the extracted speech features into the CNN, and then through CNN convolution and pooling operations [24], the output of the CNN was used as the input of BLSTM. The infrastructure of the CNN includes four convolution layers,a pooling layer, and a dense layer,which can extract global features. There were 64, 128, 256 and 512 convolution kernels in the first to fourth convolution layers. The size of the convolution layer was 3 ∗ 3, the stripe was 1 ∗ 1, the size of the pooling layer was 4 ∗ 4, and the step size was 4 ∗ 4. The BLSTM model was introduced to realize the two-way memory of information.The contextual dependency information for acoustic features can be extracted through BLSTM. Both global and local features are beneficial to speech emotion recognition.Therefore, we combined CNN+BLSTM allowing the extraction of global and local features simultaneously [25]. The number of hidden cells in the BLSTM was set to 128. A flow chart of feature extraction is shown in Figure 2.

TABLE 1. Types of features in interspeech2010 language challenge.

Feature descriptors	Number of features
PCM loudness	42
Mel spectrum coefficient(MFCC)	630
Log mel freq band	336
Linear spectrum(LSP) frequency	336
F0 envelope	42
Voicing prob.	42
F0	38
Jitter local	38
Jitter consent. frame pairs	38
Shimmer local	38
F0 number of onsets	1
Turn duration	1

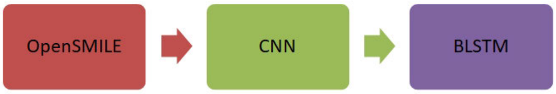

FIGURE 2. The flow chart of feature extraction.

2) LDA

We first used the LDA method to reduce the feature dimensions of the multimodal feature space in the source and target data. LDA projects the data into a subspace such that the distance between the data of the same category is the minimum, and the distance between the data of different categories is the maximum. The steps are as follows:

$$\mu^{(i)} = \frac{1}{n_i} \sum\nolimits_{x \in \text{classi}} x \tag{1}$$

$$\mu = \frac{1}{m} \sum\nolimits_{i=1}^{m} x_i \tag{2}$$

$$S_b = \sum\nolimits_{i=1}^{c} n_i \left(\mu^{(i)} - \mu\right) \left(\mu^{(i)} - \mu\right)^T \tag{3}$$

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

$$W_{opt} = \operatorname{argmax} \frac{\operatorname{tr}\left(Q^T S_b Q\right)}{\operatorname{tr}\left(Q^T S_w Q\right)} = \min \operatorname{tr}\left(Q^T \left(S_w - \beta S_b\right) Q\right) \tag{5}$$

n_i is the number of samples belonging to class i, x_i is the ith sample, $\mu^{(i)}$ is the sample mean of class i, μ is the mean of all samples, W_{opt} represens the projection matrix formed by a set of optimal discriminant features space. β is used to balance the importance between S_w and S_b.

3) MMD

We designed $X = [X_s, X_t] \epsilon R^{m*n}$ as a feature matrix. $X_s = [x_1, \ldots x_{n_s}]$ are the features of the source data, $X_t = [x_{n_s+1}, \ldots x_n]$ are the features of the target data. We calculated the MMD values between the source data and the target data. The MMD was calculated as:

$$G\left(C^s, C^t\right) = ||\frac{1}{n_s}\sum\nolimits_{i=1}^{n_s} C_s^s - \frac{1}{n_t}\sum\nolimits_{j=1}^{n_t} c_j^t||^2 \\ = tr\left(Q^T X M X^T Q\right) \tag{6}$$

Tr() is a trace of a matrix and M is the MMD matrix. M was calculated as:

$$m_{i,j} = \begin{cases} \frac{1}{n_s^2} x_i,\ x_j \epsilon X_s \\ \frac{1}{n_t^2} x_i,\ x_j \epsilon X_t \\ \frac{-1}{n_s n_t} \text{otherwise} \end{cases} \tag{7}$$

c_i^s represents the sample of the source data,c_j^t represents the sample of the target dataset, C^srepresents the common features of the source data after feature subspace mapping, C^t represents the common features of the target data after subspace mapping.

4) GE

Although MMD can reduce the differences between the source and target data, it neglects the data's geometric information. GE was used to maintain this geometric information. GE considers the similarity of samples in the neighborhood as a distribution difference constraint. For each sample vector, we can determine its p-nearest neighbors in terms of Euclidean distance. A 0-1 matrix was used in this study. The weight value of the adjacent points was 1, and that of the non-adjacent points was 0. The 0-1 weight matrix $W = w_{ij}$ of the GE can be calculated as equal (8)

$$w_{i,j} = \begin{cases} 1 \text{if} x_i^s \in N_p\left(x_j^t\right) & \text{or } x_j^t \in N_p\left(x_i^s\right) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

x_i^s represents the feature of the source data, x_j^t represents the feature of the target data, $N_p\left(x_j^t\right)$ represents the p nearest neighbors of x_j^t, $N_p\left(x_i^s\right)$ represents the p nearest neighbors of x_i^s.

The function of GE was calculated as:

$$G(Q) = \frac{1}{2}\sum\nolimits_{i,j=1}^{N} ||c_i - c_j||^2 w_{i,j} \\ = \frac{1}{2}(\sum\nolimits_{i=1}^{N} c_i^2 \sum\nolimits_{j=1}^{N} w_{ij} + \sum\nolimits_{j=1}^{N} c_j^2 \sum\nolimits_{i=1}^{N} w_{i,j} \\ - 2\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N} c_i c_j w_{i,j}) \\ = \sum\nolimits_{i=1}^{N} c_i^2 D_{ii} - \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N} c_i c_j w_{i,j} \\ = tr(Q^T X L X^T Q) \tag{9}$$

L=D-W is a Laplacian matrix, D is a diagonal matrix, and each element on its diagonal is the sum of the corresponding columns of W. c_i and c_j are low dimensional representations of the two data points.

5) LSR

TEDFSL combines subspace learning and regression methods in a unified framework. In the subspace formed by LDA+MMD+GE, we introduced a regression coefficient matrix, and used the least-squares regression method to describe the relationship between the feature representation and the corresponding label,which makes the model more discriminative and can better predict the label information of the target domain test data. Therefore, we introduced a regression matrix P to achieve this goal. The proposed regression model can be described as follows:

$$\min||Y - PQ^T X||_F^2 \tag{10}$$

6) l_c

Some traditional transfer learning methods only consider reducing the domain differences between the source data and target data while ignoring the emotion-discriminative features.Therefore, l_c is introduced. Some traditional transfer learning methods only consider reducing the domain differences between the source data and target data while ignoring the emotion-discriminative features. To solve this issue, we combined l_c with MMD and GE. The core idea of l_c is to learn emotion-discriminative and domain-invariant feature representations simultaneously. Since emotions in different domains have distant centers of emotion classes, we introduced the prior knowledge of emotion categories into deep feature learning to maintain the emotion discrimination of speech features [26]. l_c can be calculated as follows:

$$l_c = \sum\nolimits_{i=1}^{n_s} \max\left(0, \left\|f_k^i - c^i\right\|_2^2 - \alpha_1\right) \\ + \sum\nolimits_{p,q=1p\neq q}^{c} \max\left(1, \alpha_2 - \left\|c_p^b - c_q^b\right\|_2^2\right) \tag{11}$$

n_s represents the number of source samples, $f_k^{s,i}$ represents the ith speech sample in the common space of the source dataset, c^i represents the feature center of the emotion category corresponding to the ith speech sample in the entire source data, α_1 and α_2 are the thresholds for adjusting the distances, c_p^b represents the mini-batch feature centers of the pth emotion category, c_q^brepresents the mini-batch feature centers of the qth emotion category. c_q^bcan be calculated as follow:

$$c_q^b = \frac{1}{n_b^q}\sum\nolimits_{1\leq i\leq n_b^q} f_k^{s,i} \tag{12}$$

Feature representation and corresponding labels of TEDFSL
DNN to learn deeper information and final classification

FIGURE 3. Flow chart of emotion classification.

n_b^q is the number of samples corresponding to the qth emotion category.

The objective function of TEDFSL can be written as:

$$\begin{aligned} L_{total} &= \min\|Y - PQ^T X\|_F^2 + tr(Q^T(S_w - \beta S_b)Q) \\ &\quad + \mu tr(Q^T XMX^T Q) + \gamma tr(Q^T XLX^T Q) + L_c \end{aligned} \tag{13}$$

C. OPTIMIZATION

TEDFSL firstly searches for the optimal transfer subspace by selecting the optimal P and Q values through an iterative algorithm. Secondly, based on the formation of the optimal transfer subspace, $\boldsymbol{l_c}$ is added to minimize $\boldsymbol{l_{total}}$.

(1) Update Q: Minimize $\boldsymbol{l_{total}}$ by fixing P and updating Q. Take the partial derivative of Q and set it equal to 0

$$\begin{aligned} \frac{\boldsymbol{\partial total}}{\boldsymbol{\partial Q}} &= (S_w - \beta S_b)Q + \mu XMX^T Q + \gamma XLX^T Q \\ &\quad + XX^T Q - XY^T P = 0 \end{aligned} \tag{14}$$

$$Q = (S_w - \beta S_b + \mu XMX^T + \gamma XLX^T + XX^T)^{-1} XY^T P \tag{15}$$

(2)Update P: Fix Q and determine the minimum value for min $\|\boldsymbol{Y} - \boldsymbol{PQ^T X}\|_F^2$ =min tr($\boldsymbol{Y^T Y} - \boldsymbol{2Y^T PQ^T X}$). Using a singular value solution for $\boldsymbol{YX^T}$Q, SVD($\boldsymbol{YX^T}$Q)= U$\Lambda\boldsymbol{V^T}$, could be obtained.

$$tr(Tr(P^T YX^T Q) = tr(P^T U\Lambda V^T) = tr(V^T P^T U\Lambda) \tag{16}$$

Therefore,the optimal P is:

$$P = UV^T \tag{17}$$

(2) After solving Steps (1) and (2), $\boldsymbol{l_c}$ is added to make minimize $\boldsymbol{l_{total}}$.

D. EMOTION CLASSIFICATION

A flow chart of emotion classification is shown in Figure 3. In TEDFSL,the dimension of features was reduced from 1582 to 150 and the relationship between the feature representation and the corresponding labels was formed,then we sent the features and corresponding labels formed by TEDFSL into DNN to make the final emotion classification.The label regression matrix P is not sufficient for accurate classification.Therefore, we added a DNN to learn deeper information about TEDFSL for the final classification.

IV. EXPERIMENT

To validate the effectiveness of the TEDFSL algorithm, this paper conducted cross-corpus SER on the IEMOCAP, YouTube and MOUD datasets. It was also compared with traditional principal component analysis (PCA), LDA, and TRaSL [22], TSDSL [23], TDLR [27] transfer subspace SER.

A. DATASETS

The YouTube dataset consisted of 47 videos along with transcriptions. The YouTube database is a video and recording of different topics such as politics and electronic product reviews. These emotions included happiness, anger, sadness, neutrality, fear and surprise. The speakers are included 20 women and 27 men with different backgrounds.

The IEMOCAP database is a multimodal database, that includes text and speech data. Ihis includes anger, happiness, sadness, neutrality, depression, fear and surprise. The IEMOCAP database supports video and audio, as well as text transcription of all words. The database contains 5331 audio and text transcriptions.

The AVEC database was recorded by 16 men and 16 women in a natural environment. These emotions included fear, surprise, disgust, happiness, neutrality, sadness and anger.

The three datasets selected for this study were divided into six sets of cross-corpus SER experiments:

IE-Yo: IEMOCAP as source dataset, YouTube as target dataset.

Yo- IE: YouTube as source data, IEMOCAP as target data.

IE-AV: IEMOCAP as source data, AVEC as target data.

AV- IE: AVEC as source data, IEMOCAP as target data.

Yo- AV: YouTube as source data, AVEC as target data.

AV- Yo: AVEC as source data, YouTube as target data.

The source and target data were each divided into 10 parts, with all the source data and 7/10 of the target data used for training and 3/10 of the target data used for testing.We selected happiness, anger, sadness, neutrality, fear and surprise these emotional types for analysis. The collected datasets had the problem of data imbalance. To solve the problem of data imbalance, the method of data balance proposed in [28] was used. For example, for these two emotions, happiness and sadness, down sampling was performed to reduce the number of samples because they had the largest number of samples.There were 700 samples in source datasets and 300 samples in target datasets.

B. EXPERIMENTAL SETUP

The running environment of the experiment was Windows system. The programming language was python, the framework of deep learning were tensorflow and keras. And the hardware environment of the experiment was PC. In all of these methods, dropout was set to 0.5, and the learning rate was set to 0.001. Adam was chosen as the optimizer. In this paper, the value of β was set to 0.1, the parameters μ and γ were chosen from {0.001, 0.01, 0.1, 1, 10, 100} using a grid search method, and the number of neighborhood points was set to 8. The values of the parameters of the transfer learning method for the comparison were also chosen from {0.001, 0.01, 0.1, 1, 10, 100}. Common evaluation metrics included accuracy, recall and F1. Accuracy (P) is the ratio of the classifier's predictions of the positive samples and predicted samples to all predicted samples. Recall (R) was the ratio of samples for which the classifier had positive and correct predictions to all samples with true predictions.

The value of recall can be calculated as:

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{18}$$

TP represents the number of predicted positive samples and actually positive samples.FN represents the number of predicted negative samples and actual positive samples.F1 is the harmonic average based on accuracy and recall. In this paper,the accuracy and F1 were chosen as the evaluation indicators. F1 was calculated as follow:

$$F1 = \frac{2*P*R}{P+R} \tag{19}$$

C. RESULTS

Based on Tables 2 and table 3, the following conclusions could be obtained:

(1) Among the traditional methods, LDA is superior to PCA.This is because LDA selects the projection direction with the best classification performance, whereas PCA selects the direction with the greatest variance in the projection of the sample points. LDA allows for a better consideration of category information in a subspace dimensionality reduction approach.

(2)The methods that used a transfer subspace improved the accuracy and F1 values by between approximately 10% and 17% respectively compared to the methods that did not use a transfer subspace.The traditional approach is only suitable for situations in which the source and target datasets are identical. It was not possible to form subspaces that could constrain the differences as the transfer learning subspaces did.The differences between the various databases had a negative impact on the effectiveness of SER.

(3) TEDFSL achieved an improvement in accuracy and F1 values of approximately 10-20% compared to the comparison transfer subspace method. The TEDFSL algorithm considered feature selection, global disparity constraints, local disparity constraints and feature-label mapping relationships when solving for a common low-dimensional subspace.In addition, the differences between the source and target datasets were reduced without neglecting the features with sentiment discrimination.However, other transfer subspace methods cannot effectively combine these factors into consideration.

(4) In the comparison of the TRaSL, TSDSL and TDLR transfer subspace methods, TDLR was the best, followed by TSDSL, and TRaSL was the worst. It was due to the fact that TRaSL only considered feature-label regression mapping and MMD, which lacked representative feature selection as well as local disparity constraints compared to TDLR. TSDSL only considered feature selection and local disparity constraints, compared to TDLR which lacked global disparity constraints and label regression. It was thus demonstrated that knowledge transfer can only be effectively achieved by considering an integrated approach to feature selection, global disparity constraints, local discrepancy constraints and feature-label mapping relationships.

TABLE 2. Accuracy(%) for different comparison methods.

Experimental setup	Traditional methods		Other transfer subspace methods			TEDFSL
	PCA	LDA	TRaSL	TSDSL	TDLR	
IE-Yo	61.48	66.29	73.69	76.34	77.32	88.02
Yo-IE	62.25	69.33	75.31	77.09	78.67	90.36
IE-AV	66.36	68.14	78.16	79.22	80.32	90.23
AV-IE	64.11	67.45	74.04	75.66	76.89	92.43
Yo-AV	64.55	65.56	70.43	74.90	77.95	97.89
AV-Yo	63.30	64.85	72.85	73.79	76.76	88.98

TABLE 3. F1 (%) for different comparison methods.

Experimental setup	Traditional methods		Other transfer subspace methods			TEDFSL
	PCA	LDA	TRaSL	TSDSL	TDLR	
IE-Yo	64.08	69.79	75.84	78.13	79.21	87.99
Yo-IE	63.12	70.61	77.59	79.06	80.05	90.07
IE-MO	67.04	70.21	77.38	78.57	81.81	90.94
MO-IE	64.19	69.67	73.56	75.09	77.68	91.41
Yo-MO	60.50	65.52	71.13	76.72	79.51	95.31
MO-Yo	60.28	64.64	72.65	73.74	76.59	89.33

D. ABLATION EXPERIMENTS

This section of the ablation experiment was conducted in order to analyze the degree of importance of each TEDFSL component. The experiments were divided into the following 5 groups, with each group removing only one portion of the objective function expression of the TDAFSL and leaving the rest unchanged. The accuracy of TEDFSL was also compared among the following five groups and the results of the comparison are shown in Figure 3:

(1) TEDFSL$\boldsymbol{L_1}$:remove feature label regression.

(2) TEDFSL$\boldsymbol{L_2}$:remove feature reduction

(3) TEDFSL$\boldsymbol{L_3}$:remove MMD constraint.

(4) TEDFSL$\boldsymbol{L_4}$:remove GE constraint.

(5) TEDFSL$\boldsymbol{L_5}$:remove $\mathbf{l_c}$

As shown in Figure 4, compared to TEDFSL, the accuracies of the three groups TEDFSL$\boldsymbol{L_1}$, TEDFSL$\boldsymbol{L_3}$, and .TEDFSL$\boldsymbol{L_4}$ decreased to a greater extent than those of TEDFSL$\boldsymbol{L_2}$ and TEDFSL$\boldsymbol{L_5}$. This result demonstrated that the two most important factors for cross-corpus SER were the MMD, GE disparity constraint and the label regression matrix P. The superiority of TEDFSL over TEDFSL$\boldsymbol{L_2}$ proves the importance of feature dimensionality reduction in removing redundant high-dimensional features. TEDFSL outperformed TEDFSL$\boldsymbol{L_5}$ proved that the importance of selecting features with emotion discriminative features.

E. PARAMETRIC ANALYSIS

We need to choose suitable parameters for the target expression min $||\mathrm{Y}-\mathrm{PQ}^{\mathrm{T}}\mathrm{X}||_{\mathrm{F}}^{2}+$ $\mathrm{tr}(\mathrm{Q}^{\mathrm{T}}(\mathrm{S_w}-\beta \mathrm{S_b})\mathrm{Q})$ $+$ $\mu\mathrm{tr}(\mathrm{Q}^{\mathrm{T}}\mathrm{XMX}^{\mathrm{T}}\mathrm{Q})+\gamma$ $\mathrm{tr}(\mathrm{Q}^{\mathrm{T}}\mathrm{XLX}^{\mathrm{T}}\mathrm{Q})+\mathrm{L_c}$ of the TEDFSL to

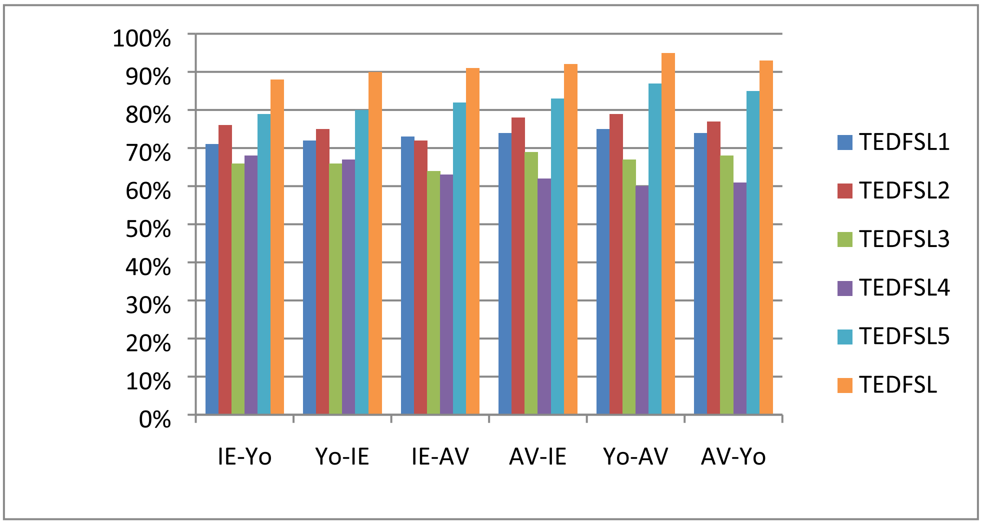

FIGURE 4. Accuracy of ablation experiments.

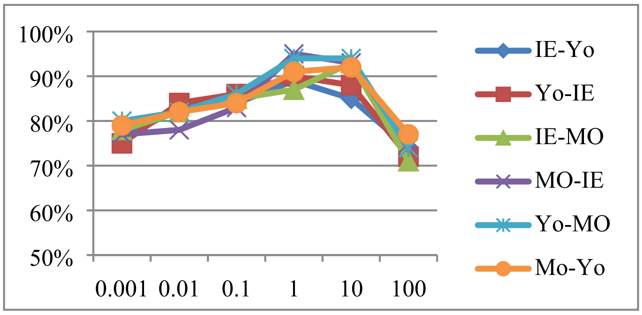

FIGURE 5. The accuracy of TEDFSL with different μ.

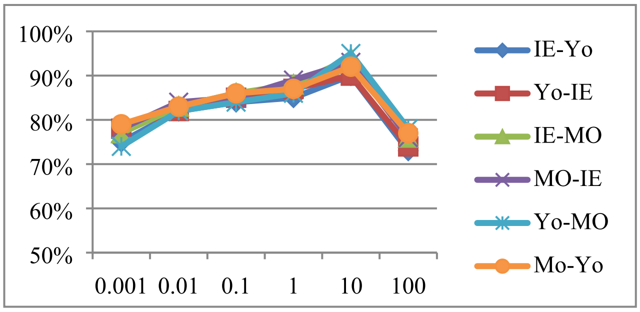

FIGURE 6. The accuracy of TEDFSL with different γ.

achieve the best SER results. Figures 5 and Figure 6 showed the accuracies for different values of μ, γ, respectively.

As shown in Figure 5, μ increased gradually between [0.0011, with the best results between [1,10] and mostly a decreasing trend between [10,100]. μ was worst at 100 and best at 1. Therefore $\mu = 1$ was chosen. The μ value is neither too small nor too large. Too small is similar to traditional subspace learning, with insignificant disparity constraints. An excessively large would compromise the emotion- discriminative features. As shown in Figure 6, γ performed best at a value of 10. γ decreased gradually at [10,100]. This is because a small value of γ would weaken the effect of the local disparity constraint, and a large value would ignore information about the category structure.

V. CONCLUSION

In this study,we proposed the TEDFSL method to address the low accuracy of cross-corpus SER. This method combines LDA, MMD, GE and label regression (LSR) to form a low-dimensional transfer subspace by jointly optimizing projection matrix Q and regression matrix P. The LDA algorithm was used to project the features into a low dimensional subspace to reduce the feature dimension. Then, the MMD was used to constrain the global difference between the source data and the target data in the low dimensional subspace, and GE was used to constrain the local difference between the source data and the target data. In addition, the label regression matrix was used to learn the relationship between labels and features, so as to achieve the transfer from the source dataset to the target dataset.Moreover,$\mathbf{l}_\mathbf{c}$ was designed to preserve emotion-discriminative features.The TEDFSL method solves the feature selection, differences constraint and label regression jointly in low-dimensional common transfer subspace learning,which is superior to previous transfer subspace learning methods.

The limitations of this study are as follows:(1) This paper only used audio modality, and did not fuse with other modalities, such as facial expression, blood pressure and heart rate.(2) The SER datasets contain a small number of emotions, which are all basic emotions. However, the emotions in real life are rich and colorful, not limited to happiness, sadness, anger, neutrality and surprise in the selected data set. These problems need to be addressed in the future study.

REFERENCES

[1] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy, "StreamAR: Incremental and active learning with evolving sensory data for activity recognition," in *Proc. IEEE 24th Int. Conf. Tools With Artif. Intell.*, vol. 1, Nov. 2012, pp. 1163–1170.

[2] I. Alnujaim, H. Alali, F. Khan, and Y. Kim, "Hand gesture recognition using input impedance variation of two antennas with transfer learning," *IEEE Sensors J.*, vol. 18, no. 10, pp. 4129–4135, May 2018.

[3] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *Proc. 23th Int. Conf. Archit. Comput. Syst.*, Feb. 2010, pp. 1–10.

[4] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, Nov. 2006.

[5] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digit. Signal Process.*, vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[6] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[7] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, Feb. 2015.

[8] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 45–55, Jan. 2020.

[9] S. Dhondge, R. Shewale, M. Satao, and J. Jagdale, "Impact of lightweight machine learning models for speech emotion recognition," in *Proc. Int. Conf. Innov. Comput. Commun.* Singapore: Springer, 2022, pp. 249–261.

[10] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: State of the art and research challenges," *Multimedia Tools Appl.*, vol. 80, pp. 23745–23812, Jan. 2021.

[11] H. Aouani and Y. B. Ayed, "Speech emotion recognition with deep learning," *Proc. Comput. Sci.*, vol. 176, pp. 251–260, Oct. 2020.

[12] S. Rajendran, S. K. Mathivanan, P. Jayagopal, M. Venkatasen, T. Pandi, M. Sorakaya Somanathan, M. Thangaval, and P. Mani, "Language dialect based speech emotion recognition through deep learning techniques," *Int. J. Speech Technol.*, vol. 24, no. 3, pp. 625–635, Apr. 2021.

[13] J. Parry, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 1656–1660.

[14] Y. Chen, Z. Xiao, X. Zhang, and Z. Tao, "DSTL: Solution to limitation of small corpus in speech emotion recognition," *J. Artif. Intell. Res.*, vol. 66, pp. 381–410, Oct. 2019.

[15] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 494–504, Apr. 2021.

[16] E. N. N. Ocquaye, Q. Mao, H. Song, G. Xu, and Y. Xue, "Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition," *IEEE Access*, vol. 7, pp. 93847–93857, 2019.

[17] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Commun.*, vol. 83, pp. 34–41, Oct. 2016.

[18] D. Phan, Y. Matsumoto, and H. Shindo, "Autoencoder for semisupervised multiple emotion detection of conversation transcripts," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 682–691, Jul. 2021.

[19] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.

[20] P. Song, S. Ou, Z. Du, Y. Guo, W. Ma, J. Liu, and W. Zheng, "Learning corpus-invariant discriminant feature representations for speech emotion recognition," *IEICE Trans. Inf. Syst.*, vol. E100.D, no. 5, pp. 1136–1139, 2017.

[21] X. Chen, X. Zhou, C. Lu, Y. Zong, W. Zheng, and C. Tang, "Target-adapted subspace learning for cross-corpus speech emotion recognition," *IEICE Trans. Inf. Syst.*, vol. E102.D, no. 12, pp. 2632–2636, Dec. 2019.

[22] N. Liu, B. Zhang, B. Liu, J. Shi, L. Yang, Z. Li, and J. Zhu, "Transfer subspace learning for unsupervised cross-corpus speech emotion recognition," *IEEE Access*, vol. 9, pp. 95925–95937, 2021.

[23] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 307–318, 2020.

[24] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," *Inf. Fusion*, vol. 46, pp. 184–192, Mar. 2019.

[25] X. Lan, Z. Li, and Z. Wang, "The influence of returnee technology executives on enterprise innovation: The innovation patent data of global exchange market listed companies," *Econ. Res.-Ekonomska Istraživanja*, vol. 36, no. 1, pp. 1361–1376, Dec. 2023.

[26] C. Lu, C. Tang, J. Zhang, and Y. Zong, "Progressively discriminative transfer network for cross-corpus speech emotion recognition," *Entropy*, vol. 24, no. 8, p. 1046, Jul. 2022.

[27] S. Li, P. Song, and W. Zhang, "Transferable discriminant linear regression for cross-corpus speech emotion recognition," *Appl. Acoust.*, vol. 197, Aug. 2022, Art. no. 108919.

[28] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Interspeech*, Aug. 2017, pp. 1103–1107.

ZHANG KEXIN received the bachelor's degree from the Dalian Institute of Science and Technology, in 2021. She is currently pursuing the master's degree with the Shanghai Institute of Technology. Her research interests include natural language processing and speech emotion recognition.

LIU YUNXIANG received the Ph.D. degree. He is a professor, the master's supervisor, and a leader of key disciplines of the institute. He is mainly engaged in the research in the fields of artificial intelligence, computer software and theory, information fusion, and intelligent information processing. He has achieved a series of important results in the theory and application of fuzzy set, the theory and application of rough set, intelligent decision support systems, data fusion system testing technology, and the research and development of intelligent instruments. He is a Senior Member of the China Computer Society. He is a reviewer of *Computer Software* and an Editorial Board Member of *Computer Measurement and Control*.