Cao *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2022) 2022:32
https://doi.org/10.1186/s13636-022-00264-5

EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Cross-corpus speech emotion recognition using subspace learning and domain adaption

Xuan Cao[1], Maoshen Jia[1*], Jiawei Ru[1] and Tun-wen Pai[2]

Abstract

Speech emotion recognition (SER) is a hot topic in speech signal processing. When the training data and the test data come from different corpus, their feature distributions are different, which leads to the degradation of the recognition performance. Therefore, in order to solve this problem, a cross-corpus speech emotion recognition method is proposed based on subspace learning and domain adaptation in this paper. Specifically, training set data and the test set data are used to form the source domain and target domain, respectively. Then, the Hessian matrix is introduced to obtain the subspace for the extracted features in both source and target domains. In addition, an information entropy-based domain adaption method is introduced to construct the common space. In the common space, the difference between the feature distributions in the source domain and target domain is reduced as much as possible. To evaluate the performance of the proposed method, extensive experiments are conducted on cross-corpus speech emotion recognition. Experimental results show that the proposed method achieves better performance compared with some existing subspace learning and domain adaptation methods.

Keywords: Speech emotion recognition, Cross-corpus, Subspace learning, Domain adaption

1 Introduction

There are many ways for people to express emotions, such as through speech, actions, and facial expressions. Speech is an important way to express emotions among these ways, because it contains riches emotions, such as happy, angry, and sad. Speakers can deliver their intentions through different tones, volumes, or content. How to judge a speaker's emotion through speech becomes crucial. Therefore, speech emotion recognition (SER) is an important branch of many modal affective computing, and it is also an important part of speech recognition. With the development of SER, it has been applied in the fields of psychotherapy, human-computer interaction, etc. According to the results of SER, the machine can generate appropriate responses for the user in an interactive environment. Therefore, SER is one of the most important technologies for human-computer interaction [1–4].

The semantic-based methods are an important class of SER methods, because emotions can be expressed effectively by semantics. If the speakers use emotive words to communicate with others, then we can directly judge the emotion from the semantics of the words. Therefore, semantic-based research gradually began to develop. A multi-classifier emotion recognition model based on prosodic information and semantic labels is introduced in [5]. Similarly, the semantic labels and the non-verbal audio in speech, such as onomatopoeia such as crying, laughter, or sighing, are used in SER [6]. Subsequently, temporal and semantic coherence is introduced for SER [7]. In addition, the model of bimodal SER from acoustic and linguistic information fusion is proposed [8].

Although semantics understanding is simply for humans, it is a complex process for machines. Therefore, more research is currently aimed at speech

*Correspondence: jiamaoshen@bjut.edu.cn

[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
Full list of author information is available at the end of the article

features that are easily understood by machines, which is also important for SER. Compared with semantic information, speech features are more abstract. But they are very important for expressing the speaker's emotions. The main features used in SER are divided into acoustic features and spectral features. The acoustic features include intensity, pitch, and timbre. Features like energy, Mel-frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC), and fundamental frequency are called spectral features. The features such as pitch, MFCC, formant, intensity, and chroma are adopted for SER [9, 10]. Also, the pitch, spectrum, and formant are combined with semantic information for recognizing emotions in [5]. To improve the robustness of SER, some methods have been used to process the features. Specifically, PCA is adopted to reduce the dimensionality of the features [11], and a statistical method is utilized to find robust spectral features [12].

In practical scenarios, the speaker's emotion is very complex. The speaker may have multiple emotions at the same time, rather than a single emotion, or the emotion expressed by the speaker is inconsistent with the actual emotion. It makes SER difficult. There is also research proposed for complex emotions. A circular continuous dimensional model to describe an emotion, called valence-arousal model (VA) was proposed in [13, 14]. The model no longer regards emotions as discrete but uses two-dimensional coordinates to describe the continuous distribution of emotions. The PAD emotional model was shown in [15, 16], which has P (pleasure), A (arousal), and D (dominance) values to represent all emotional states. In addition, based on the emotional probability distribution, an ambiguous label is proposed to solve the inconsistency problem in ambiguous emotional cognition [17].

Another problem in SER is how to recognize emotions. To this end, some machine learning methods were adopted to recognize emotions, such as support vector machine (SVM) [18], hidden Markov model (HMM) [19], and Gaussian mixed model (GMM) [20]. In recent years, with the rapid development of deep learning, various neural network structures have been introduced in SER. From convolutional neural networks (CNN) [21], recurrent neural networks (RNN) [22], back propagation neural network (BPNN) [23], and deep neural network (DNN) [24] to sequential capsule networks [25] and adversarial data augmentation network [26], they are both used for SER. A segment-based iterative self-learning enhanced speech emotion recognition model is proposed in [27]. The above algorithms perform well in traditional SER, and the recognition accuracy of some algorithms can even reach more than 80% in some corpora settings. In the actual scene, the speech signals do not belong to a specific corpus, which are recorded in different scenes. The speech data is also affected by language, gender, speaking styles, and other factors. So, when the training set and the test set came from different corpus, the training and testing data often follow different feature distributions. The recognition performance will be reduced at this time.

Therefore, transfer learning is adopted to solve the problem of data cross-corpus [28]. The known corpus data is considered as the source domain, and the unknown data to be learned constitutes the target domain. Transfer learning is to transfer the knowledge of the source domain to the target domain to reduce the data distribution difference between the two domains, and in SER, the features of the source and target domains are distributed in different spaces. So, the transfer from the source domain to the target domain is a feature-based transfer, that is, a mapping relationship between two domains is established to reduce the differences in feature distributions. With the development of transfer learning, more transfer learning algorithms are applied to SER. Among them, in order to solve the cross-corpus SER problem, many researches focus on transfer subspace learning and domain adaptation, such as unsupervised transfer subspace learning [28], transfer subspace learning based on feature selection [29], transfer subspace learning based on non-negative matrix factorization [30], transfer linear subspace learning [31], and Universum autoencoder-based domain adaptation [32]. In addition, a cross-corpus speech emotion recognition based on domain adaptive least squares regression is proposed in [33], and in [34, 35], ADDoG-based and DANN-based methods are proposed according to the idea of domain adversarial. Most of the above methods involve transfer subspace learning and domain adaptation, which are important issues in transfer learning and the focus of this paper. The two parts are considered jointly in this paper. Therefore, inspired by the frame in [36], a cross-corpus speech emotion recognition method is proposed.

The contributions of the proposed method are summarized as follows:

- The proposed method combines subspace learning and mapping to realize speech emotion recognition across the corpus. The feasibility of the proposed method is proved by experimental results.
- In this paper, a subspace learning model is constructed based on the Hessian matrix, so that the

extracted features both in the source domain and the target domain have good robustness in their independent subspace, which can be adopted to improve the subsequent cross-corpus transfer ability.

- Information entropy is used to establish a domain adaption model in the proposed method. The numerical descent is used to minimize information entropy, so that a common space of source and target domains is learned, thereby the difference in features distribution between the two domains is reduced.

The rest of the paper is organized as follows. In Section 2, the specific process of the proposed method is introduced, along with some optimizations. In Section 3, the emotion recognition performance of the proposed method is analyzed on three public datasets, and the effects of different parameters on the performance are analyzed through experiments. Finally, the conclusion is drawn in Section 4.

2 The proposed method

A cross-corpus speech emotion recognition method is proposed by combing subspace learning and domain adaption. The block diagram of the proposed method is shown in Fig. 1.

Firstly, features of speech in the source corpus and target corpus are extracted to form the source domain and the target domain. Then, the Hessian-based subspace learning is performed on the feature in the source domain and the target domain to obtain low-dimensional features for forming their own independent subspace. The flowchart of the Hessian-based subspace learning part is shown in Fig. 2. Furthermore,

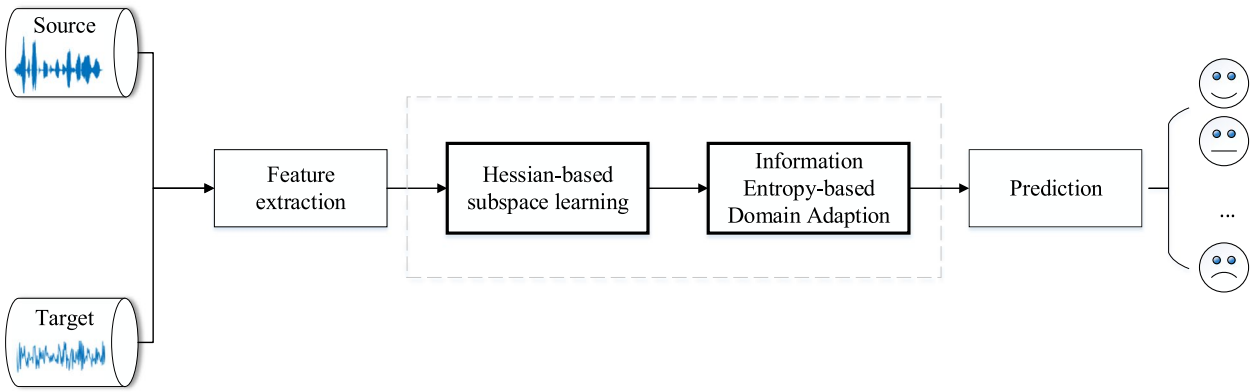

Fig. 1 The block diagram of the proposed method

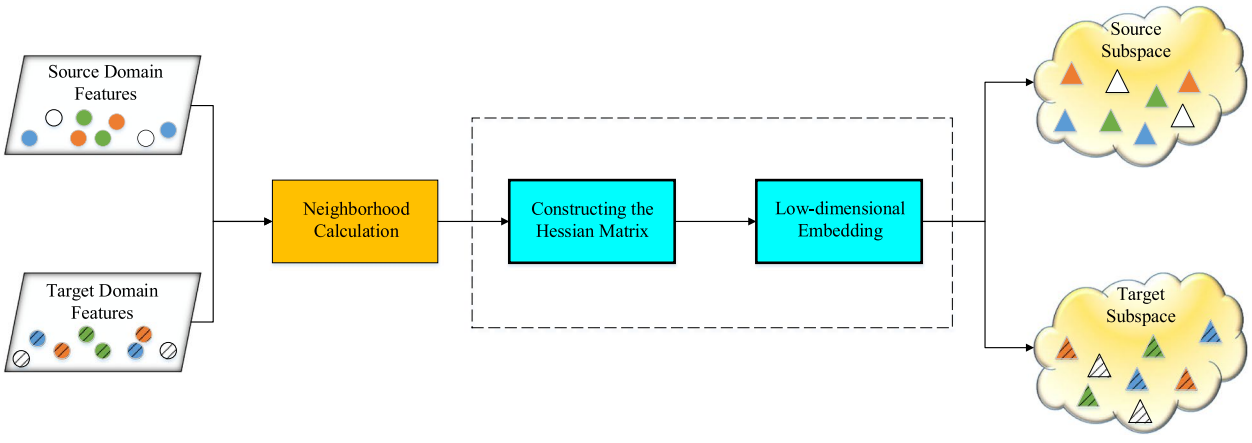

Fig. 2 Flowchart of Hessian-based subspace learning. In the flowchart, shapes with different colors represent different characteristics. The circles represent the original features in the source and target domain. Triangles represent the subspace of the source and target domain. Features in different domains are distinguished by diagonal shading

the mapping relationship between the source domain subspace and the target domain subspace is established by using information entropy, which is used for reducing the difference of feature distribution between different domains. This mapping relationship is revealed by the common space. Therefore, it is important to find the common space corresponding to the two domains in this method. The flowchart of the domain adaption part is shown in Fig. 3. Finally, emotions are predicted.

In the part of Hessian-based subspace learning, the neighboring frames of the current frame are found based on neighborhood calculation. Then, the Hessian matrix [37] is constructed for low-dimensional embedding to obtain the subspace of the source and target domain, respectively.

After obtaining the subspace of the source and target domain, the transformation matrix is obtained through correlation coefficients of the subspace. Then, the distance between the feature data of each frame in the source domain subspace with that of each frame in the target domain subspace is calculated. And the probability that a frame in the subspace of the target domain is neighborhood to each frame in the source domain is obtained according to the distance. In this way, the posterior probability that the features of each frame in the target domain subspace are estimated to be a certain class can be obtained according to the known class labels of the features of each frame in the source domain subspace. Then, the entropy between the target domain features and emotion labels and the entropy between the features and domain labels of the two domains are calculated. Finally, the two information entropies are jointly optimized by numerical descent. The mapping relationship between the source domain subspace and the target domain subspace is acquired, which is described by a common space.

Then, Hessian-based subspace learning [38] and the domain adaption based on information entropy are introduced in detail. Finally, a specific optimization method for finding the common space is given.

2.1 Hessian-based subspace learning

An input feature matrix $\mathbf{X}=(x_{mn})_{M\times N}$ is given, which is composed of the features of the speech. m and n are the feature index and the frame index, respectively. M and N are the total number of the feature dimension and the number of frames, respectively. First, the feature energy of each frame is as follows:

$$x_n^{\mathrm{e}} = \sum_{m=1}^{M} x_{mn}^2, \tag{1}$$

where x_n^{e} represents the feature energy of the nth frame, and x_{mn} represents the feature of the mth dimension in the nth frame.

Thus, an energy matrix can be formed as $\mathbf{X}^{\mathrm{e}}=[x_1^{\mathrm{e}}, x_2^{\mathrm{e}}, \ldots, x_N^{\mathrm{e}}]$. Then, two new feature energy matrices $\mathbf{A}$ and $\mathbf{B}$, which are used for calculating the distance of the feature between different frames, are defined as follows:

$$\begin{cases} \mathbf{A} = (a_{ij})_{N\times N} \\ \mathbf{B} = (b_{ij})_{N\times N} \end{cases} \tag{2}$$

where $a_{ij} = x_j^{\mathrm{e}}$, $b_{ij} = x_i^{\mathrm{e}}$, $1 \le i, j \le N$, and i and j represent the index of the row and column, respectively. In order to find the nearest K frames of each frame, the distance $\mathbf{D}_e=(d_{ij})_{N\times N}$ of the feature between different frames is calculated as follows:

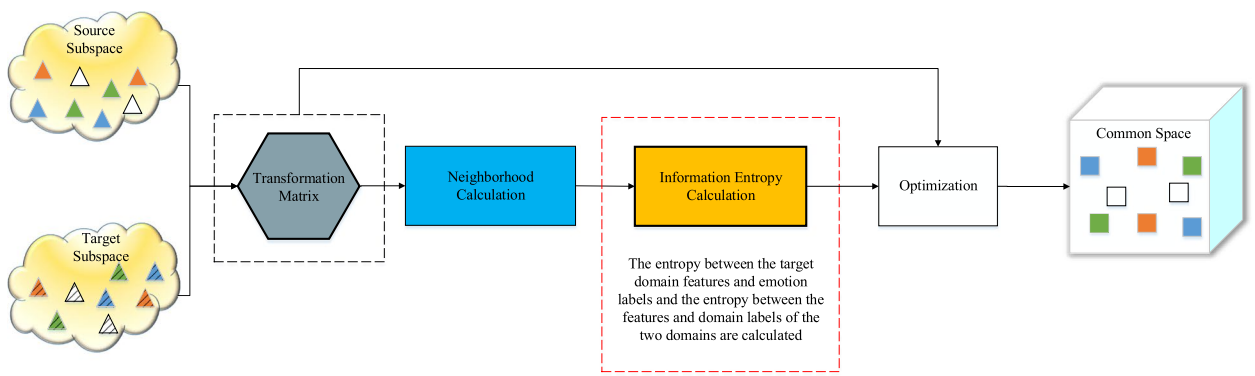

Fig. 3 Flowchart of information entropy-based domain adaption. In this flowchart, triangles represent the subspace of the source and target domain. In the common space, the squares represent the features in the source and target domain. Feature distribution differences are reduced in the common space

$$\mathbf{D}_e = \mathbf{A} + \mathbf{B} - 2\mathbf{X}^T\mathbf{X} \tag{3}$$

where d_{ij} represents the distance between the feature energy of the ith frame and the jth frame. The smaller the distance d_{ij} is, the closer the feature energies of the ith frame and the jth frame are. In fact, the definition of distance $\mathbf{D}_e$ is derived from Euclidean distance. $\mathbf{A}$ and $\mathbf{B}$ are formed by the square of the elements in the input matrix $\mathbf{X}$. According to Eqs. (1), (2), and (3), the distance defined in this paper meets the requirements of non-negativity, directness, and identity. $\mathbf{A}$ and $\mathbf{B}$ are constructed in a way that also satisfies the symmetry of the distance.

The jth column from the matrix $\mathbf{D}_e$ (i.e., $\boldsymbol{d}_j^e = \left[d_{1j}^e, d_{2j}^e, \ldots, d_{Nj}^e\right]^T$) denotes the distance vector of feature energy between the jth frame and each frame. The sorted distance matrix in ascending order is $\mathbf{d}_j^{eS} = \left[d_{S_j(1)j}^e, d_{S_j(2)j}^e, \ldots, d_{S_j(N)j}^e\right]^T$; $S_j(i)$ denotes the index of the frame sorted by the distance from the jth frame, where $S_j(1)$ represents the index with the minimum distance in d_{ij}^e; and $S_j(N)$ is the index of the maximum distance. It is worth mentioning that for each frame, d_{jj}^e is the minimum element in $\boldsymbol{d}_j^e$, i.e., $S_j(1)=j$. The 2nd to the $(K+1)$-th minimum distance from $\mathbf{d}_j^{eS}$ are selected to form the adjacent index matrix $\boldsymbol{i}_j=[S_j(2), S_j(3), \ldots, S_j(K+1)]^T$of the jth frame. K denotes the number of the largest neighbor frames. Thereby, the $K\times N$ adjacent index matrix $\mathbf{I}=[\boldsymbol{i}_1, \boldsymbol{i}_2, \ldots, \boldsymbol{i}_N]$ of N frames is obtained. Then, the elements in the input matrix $\mathbf{X}$ correspond to the indices in $\mathbf{I}$ and are selected to form a neighborhood matrix $\mathbf{Z}_n$, which is defined as follows:

$$\mathbf{Z}_n = \left(z_{mk}^n\right)_{M\times K} \tag{4}$$

where $z_{mk}^n = x_{mS_n(k+1)}$, $1 \le k \le K$, $1 \le m \le M$, $1 \le n \le N$. k, m, and n are the neighbor index, the feature index, and the frame index, respectively. $\mathbf{Z}_n$ represents the neighborhood matrix corresponding to the nth frame.

$\mathbf{E}_n$ is a centralized matrix of $\mathbf{Z}_n$, which is defined as follows:

$$\mathbf{E}_n = \left(e_{mk}^n\right)_{M\times K} \tag{5}$$

where $e_{mk}^n = \frac{1}{K}\sum_{k=1}^{K} z_{mk}^n$

The purpose of the proposed Hessian-based subspace learning is to obtain the local coordinates of the neighborhood, which are transitioned by tangent coordinates. The tangent space consists of tangent coordinates, which is regarded as a subspace of the Euclidean space. A standard orthogonal coordinate system is associated with the inner product inheritance of the Euclidean space, which can be obtained by using singular value decomposition. Therefore, $\mathbf{Z}_n - \mathbf{E}_n$ is subjected to singular value decomposition. The standard orthonormal basis $\mathbf{V}_n = \left(v_{ij}^n\right)_{K\times K}$ can be obtained by singular value decomposition as follows:

$$\mathbf{Z}_n - \mathbf{E}_n = \mathbf{U}_n\mathbf{\Sigma}_n\mathbf{V}_n^T \tag{6}$$

where $(\cdot)^T$ denotes transposition. $\mathbf{U}_n$ is the left singular vector of $\mathbf{Z}_n - \mathbf{E}_n$. $\mathbf{\Sigma}_n$ is a diagonal matrix of singular values.

First d columns of $\mathbf{V}_n$ are extracted to constitute the tangent coordinates $\mathbf{V}_n^d = \left(v_{ij}^n\right)_{K\times d}$ with dimension $K\times d$.

Next, an association Hessian matrix $\mathbf{Q}_n$ is given by using $\mathbf{V}_n^d$, which is defined as follows:

$$\mathbf{Q}_n = \left(q_{kj}^n\right)_{K\times\frac{d(d+1)}{2}} \tag{7}$$

where $q_{kj}^n = v_{kj_1}^n v_{kj_2}^n$, n is the frame index, $1 \le n \le N$. j_1 and j_2 are the dimension indexes. The corresponding relationship among j, j_1, and j_2 is given as follows:

$$j = j_2 + \sum_{l-1}^{j_i-1}\sum_{i=j_1}^{d} 1 \tag{8}$$

where $1 \le j_1 \le d$, $1 \le j_2 \le d$, $j = 1, 2, \ldots, \frac{d(d+1)}{2}$.

Furthermore, an estimation matrix $\mathbf{L}_n = \left(l_{ij}^n\right)_{K\times\left(1+d+\frac{d(d+1)}{2}\right)}$ is constructed as follows:

$$l_{ij}^n = \begin{cases} 1 & j = 1 \\ v_{ij}^n & 2 \le j \le d \\ q_{ij}^n & d+1 \le j \le \frac{d(d+1)}{2} \end{cases} \tag{9}$$

where $1 \le i \le K$, $1 \le n \le N$.

$\mathbf{G}_n = \left(g_{ij}^n\right)_{K\times\left(1+d+\frac{d(d+1)}{2}\right)}$ can be obtained by Schmitt orthogonalization of estimated matrix $\mathbf{L}_n$ [39]. The last $\frac{d(d+1)}{2}$ columns of $\mathbf{G}_n$ are taken to obtain the matrix $\mathbf{G}_n^b = \left(g_{ij}^{bn}\right)_{K\times\frac{d(d+1)}{2}}$. Then, Hessian quadratic matrix $\mathbf{H}$ can be constructed by using the matrix $\mathbf{G}_n^b$, which is formed as follows:

$$\mathbf{H} = \sum_{n=1}^{N}\mathbf{C}_n^T\mathbf{C}_n \tag{10}$$

where $\mathbf{C}_n = \left(c_{ij}\right)_{\frac{d(d+1)}{2}\times N}$ is a matrix composed of $\mathbf{G}_n^{bT}$, and it is defined as follows:

$$c_{iS_n(j)} = \begin{cases} g_{ij}^{bn}, & 1 \le j \le K \\ 0, & K < j \le N \end{cases} \tag{11}$$

where $1 \le i \le \frac{d(d+1)}{2}$, and $S_n(j)$ denotes the index of the frame sorted by the distance from the nth frame, $1 \le n \le N$.

Next, the d-dimensional subspace corresponding to the d smallest eigenvalues can be obtained by using $\mathbf{H}$, which is a null space and denotes as $\mathbf{U} = (u_{ij})_{N \times d}$. If a manifold is locally equidistant to an open subset in Euclidean space, then the mapping function from this manifold to the open subset is a linear function. The quadratic mixed derivative of the linear function is 0, so the local quadratic form formed by the Hessian coefficients is also 0. Hence, the global Hessian matrix has a $(d+1)$-dimensional null space. The first-dimension subspace of the Hessian matrix is composed of a constant function, and other d-dimensional subspaces form equidistant coordinates. Then, the embedding matrix $\mathbf{R} = (r_{ij})_{d \times d}$ can be calculated as follows:

$$r_{ij} = \sum_{l \in \boldsymbol{J}} u_{li} u_{lj} \tag{12}$$

where $\boldsymbol{J}$ represents the set of the index of the neighborhood frames, $1 \le i \le d$, $1 \le j \le d$.

Finally, the subspace $\mathbf{Y}$ is obtained according to the low-dimensional embedding:

$$\mathbf{Y} = \mathbf{R}^{\mu} \mathbf{U}^{T} \tag{13}$$

where μ is a regularization parameter, and $(\cdot)^T$ denotes transposition.

There may be a small number of outliers in the subspace $\mathbf{Y}$ after the low-dimensional embedding. In order to solve this problem, the outliers in the subspace $\mathbf{Y}$ are corrected in this paper. These outliers are characterized by a small number, with values that deviate from the distribution of most data. So, the detection thresholds are set to recognize the outliers. Then, the outliers are replaced with $2Tr(\mathbf{U}^T \mathbf{E} \mathbf{U})$ [40], where $Tr(\cdot)$ means the trace of the matrix in parentheses. $\mathbf{E} = (e_{ij})_{N \times N}$ is a diagonal matrix, where e_{ij} is defined as [41]:

$$e_{ij} = \begin{cases} \frac{1}{2\|u_i\|_2} & i = j \\ 0 & i \ne j \end{cases} \tag{14}$$

Following the above steps, the source domain subspace $\mathbf{Y}_s$ and the target domain subspace $\mathbf{Y}_t$ can be obtained.

2.2 Information entropy-based domain adaption

A domain adaption method was proposed to build the relationship between the source domain subspace and the target domain subspace. In detail, a common space with similar feature distributions in the source and target domains is constructed. Both the information entropy between the data and emotion labels and the entropy between data and domain labels are used to optimize the mapping [42]. Thereby, the difference in feature distribution in different corpora can be reduced.

After obtaining the source domain subspace $\mathbf{Y}_s = \left(y_{ij}^s\right)_{d \times N}$ and target domain subspace $\mathbf{Y}_t = \left(y_{ij}^t\right)_{d \times N}$, a principal component coefficient of the source domain $\mathbf{W}_s = \left(w_{ij}^s\right)_{d \times d}$ and the target domain $\mathbf{W}_t = \left(w_{ij}^t\right)_{d \times d}$ is calculated. In some cases, the dimension of the source domain and the target domain is different, which leads to different dimensions of the principal component coefficients. The dimension of the principal component coefficient of the target domain and the source domain with the smallest dimension should be taken as d_w. The dimensions of the source domain and the target domain are the same in this paper, so d_w is set as d. Since the transfer is carried out from the source domain to the target domain, the target domain is used as the basis for the transformation space. The transformation matrix $\mathbf{W}$ for both source domain and target domain is set as $\mathbf{W} = \mathbf{W}_t$. Features in the source domain and target domain can be mapped into a common space by $\mathbf{W}$.

First, the distance matrix $\mathbf{D} = (d_{ij})_{N \times N}$ formed by the features between different frames from the source domain subspace and the target domain subspace is given as follows:

$$\mathbf{D} = \mathbf{A}' + \mathbf{B}' - 2\mathbf{X}_s{}^T \mathbf{X}_t \tag{15}$$

where $\mathbf{X}_s = \left(x_{mn}^s\right)_{d \times N} = \mathbf{W}^T \mathbf{Y}_s$ denotes the source domain subspace features in transform space, $\mathbf{X}_t = \left(x_{mn}^t\right)_{d \times N} = \mathbf{W}^T \mathbf{Y}_t$ denotes the target domain subspace features in transform space, $\mathbf{A}' = (a_{ij})_{N \times N}, a_{ij} = \sum_{m=1}^{d} \left(x_{mj}^s\right)^2$, $\mathbf{B}' = (b_{ij})_{N \times N}, b_{ij} = \sum_{m=1}^{d} \left(x_{mi}^t\right)^2$.

The neighbor frames are detected according to the distance between the feature of each frame. Therefore, a conditional probability model is defined as follows:

$$p_{ij} = \frac{e^{-d_{ij}}}{\sum_{i=1}^{N} e^{-d_{ij}}} \tag{16}$$

where $1 \le i \le N$, $1 \le j \le N$, and p_{ij} is the conditional probability density that the jth frame in the target domain is adjacent to the ith frame in the source domain. It can describe the probability of the nearest neighbor between each frame feature in the source domain and the frame feature in the target domain.

The emotion label corresponding to the ith frame in the source domain is Label_i, $\text{Label}_i \in \mathbf{Label} = \{1, 2, \ldots, L\}$, i.e., there are a total of L types of emotion. According to formula (16), an emotion label probability estimate $\hat{p}_{lj}$ of the jth frame in the target domain is given as follows:

$$\widehat{p}_{lj} = \sum_{Label_{i=l}} p_{ij} \tag{17}$$

where $1 \le l \le L$, $1 \le j \le N$, $1 \le i \le N$, and $\hat{p}_{lj}$ express the probability that the jth frame in the target domain is discriminated as the lth type of emotion when the emotion of the source domain is known.

Since $\hat{p}_{lj}$ is a preliminary probability estimate of the emotion label of each frame feature in the target domain, the relationship between target domain features and emotion labels cannot be directly revealed by $\hat{p}_{lj}$ [43–45]. Therefore, the entropy $I(\mathbf{X}_t; \mathbf{Label})$ between the target domain features and emotion labels is calculated by using $\hat{p}_{lj}$ in this paper, which is defined as follows:

$$I(\mathbf{X}_t; \mathbf{Label}) = -\sum_{l=1}^{L} \left(\log \left(\sum_{j=1}^{N} \frac{\hat{p}_{lj}}{N} \right) \sum_{j=1}^{N} \frac{\hat{p}_{lj}}{N} \right) - \frac{\left(-\sum_{j=1}^{N} \sum_{l=1}^{L} \left(\hat{p}_{lj} \log \left(\hat{p}_{lj} \right) \right) \right)}{N} \tag{18}$$

Equation (18) is composed of two parts. In the first part, the entropy of the average probability that the feature of all frames in the target domain belongs to each emotion label is calculated. The average of the entropy of the feature in the target domain belonging to each emotion label is computed in the second part. In order to reduce the influence of incorrect labels on the feature discrimination results of each frame in the target domain, Eq. (18) needs to be optimized later. It should be noted that if only the second part is minimized, a degenerate solution will be obtained. That is, all frames in the target domain may be classified into the same type of emotion. So, the first part in Eq. (18) is necessary.

Then, the entropy $I^{st}(\mathbf{X})$ between the features and domain labels of the two domains are introduced to maximize the similarity between the two domains, which is defined as:

$$I^{st}(\mathrm{X}) = -\sum_{t=1}^{2} \left(\sum_{j=1}^{N+M} \frac{p_{tj}}{N+M} \log \left(\sum_{j=1}^{N+M} \frac{p_{tj}}{N+M} \right) \right) - \frac{\left(-\sum_{j=1}^{N+M} \sum_{t=1}^{2} \left(p_{tj} \log \left(p_{tj} \right) \right) \right)}{N+M} \tag{19}$$

where $1 \le j \le N+M$.

To calculate the entropy $I^{st}(\mathbf{X})$, firstly, the distance d'_{ij} between the ith frame feature in the source domain and the jth frame feature in the target domains is calculated according to Eq. (3), where $\mathbf{X} = (x_{ij})_{d \times (N+M)}$ denotes the feature for all frames in the source and target domains, $\mathbf{A} = (a_{ij})_{(N+M) \times (N+M)}$, $a_{ij} = \sum_{m=1}^{d} (x_{mj})^2$, $\mathbf{B} = (b_{ij})_{(N+M) \times (N+M)}$, and $b_{ij} = \sum_{m=1}^{d} (x_{mi})^2$. N and M denote the number of frames in the source domain and target domain, respectively. In this paper, the number of frames in the source domain is the same as that in the target domains, i.e., $N = M$. Then, the probability p'_{ij} of the ith frame feature and the jth frame being adjacent to each other in the source domain and the target domain is calculated according to Eq. (16) using d'_{ij}. Next, the probability p_{tj} that the jth frame in the source domain and the target domain is judged as the target domain or the source domain is calculated according to Eq. (17).

2.3 Optimization

In this subsection, an iterative optimization algorithm based on numerical descent [46] is introduced using Eqs. (18) and (19). The objective function is:

$$f = \min \left\{ \lambda I^{st}(\mathbf{X}) - I\left(\mathbf{X}_t; \mathbf{Label} \right) \right\} \tag{20}$$

where λ is the regularization parameter.

In the optimization process, the transfer coefficient matrix $\boldsymbol{g}$ is given for numerical descent in this paper, which is defined as follows:

$$\boldsymbol{g} = \lambda \boldsymbol{g}^{st}(\mathbf{X}) - \boldsymbol{g}(\mathbf{X}_t; \mathbf{Label}) \tag{21}$$

where λ is the regularization parameter.

The calculation process of $\boldsymbol{g}(\mathbf{X}_t; \mathbf{Label})$ is as follows. First, an information matrix $\mathbf{I}^C = \left(i^c_{lj} \right)_{L \times N}$ is defined using $\hat{p}_{lj}$ as:

$$i^c_{lj} = \frac{\log \left(\hat{p}_{lj} \right) - \log \left(\sum_{j=1}^{N} \frac{\hat{p}_{lj}}{N} \right)}{N} \tag{22}$$

where i^c_{lj} represents the difference between the probability that the feature of the jth frame in the target domain belongs to the emotion of the lth category and the average probability that the features of all frames in the target domain belong to the emotion of the category.

Input: $\boldsymbol{g}$ ► Transfer parameter for the first iteration
Input: f ► Information entropy for the first iteration
Input: L_{step} ► Stepsize as a function of the number of iterations
Input: $L_{maxstep}$ ► Maximum stepsize
Input: $N_{maxiter}$ ► The maximum number of iterations
Input: $\mathbf{W}$ ► Transformation matrix
Input: d ► Dimension of the transformation matrix
Output: $\mathbf{L}$ ►Common space of source and target domains

1. Initialize i_{iter} ►i_{iter} is the index of the iteration
2. Initialize L_{step}, $L_{maxstep}$, $N_{maxiter}$
3. for i_{iter} =2: $N_{maxiter}$
   if $f^{(i_{iter})}>f^{(i_{iter}-1)}$ ►$f^{(i_{iter})}$ is the Information entropy in (i_{iter})th iteration.
   $L_{step}=L_{step}\times c$, ($0<c<1$)
   else if
   $L_{step}=L_{step}\times c$, ($c>1$)
   end
   if $L_{step}\geqslant N_{maxiter}$
   $L_{step}=N_{maxiter}$
   end
4. Compute the transformation matrix: $\mathbf{W}_d = \mathbf{W} - L_{step} * \boldsymbol{g}$
5. Calculate the common space: $\mathbf{L} = \sqrt{\frac{d}{trace(\mathbf{W}_d{}^T\mathbf{W}_d)}}\mathbf{W}_d$ ► $trace\,(\cdot)$ means to trace the matrix in parentheses
   ►stop condition: $\mathrm{Min}\{|f^{(i_{iter}-3)}-f^{(i_{iter}-2)}|, |f^{(i_{iter}-2)}-f^{(i_{iter}-1)}|, |f^{(i_{iter}-1)}-f^{(i_{iter})}|\} > \epsilon \times f^{(i_{iter})}$ or $i_{iter} = N_{maxiter}$
6. Recalculate $f^{(i_{iter}+1)}$ and $\boldsymbol{g}$
7. **break**
8. **end for**
9. **return L**

Algorithm 1. Optimization method based on numerical and information entropy and calculation method of mapping space

Next, a coefficient matrix $\boldsymbol{\Gamma}=(\gamma_{ij})_{N\times N}$ is calculated from p_{ij} and i_{lj}^c as follow:

$$\gamma_{ij} = \left(\sum\nolimits_{i=1}^{N} o_{ij}p_{ij} - o_{ij}\right)p_{ij} \tag{23}$$

where $o_{ij} = i_{lj}^c$, $Label_i = l$. $\boldsymbol{g}(\mathbf{X}_t;\mathbf{Label})$ is obtained as follows:

$$\boldsymbol{g}(\mathbf{X}_t;\mathbf{Label}) = 2\left[\mathbf{Y}_s\boldsymbol{\Omega}\mathbf{Y}_s{}^T + \mathbf{Y}_t\boldsymbol{\Omega}\mathbf{Y}_t{}^T - \mathbf{Y}_s\boldsymbol{\Gamma}\mathbf{Y}_t{}^T - \mathbf{Y}_t\boldsymbol{\Gamma}\mathbf{Y}_s{}^T\right]\mathbf{W} \tag{24}$$

where $\boldsymbol{\Omega}$ is a diagonal matrix, and the main diagonal element is $\sum_{j=1}^{N}\gamma_{ij}$. $\mathbf{W}$ is the transfer matrix.

Since the calculation process of $\boldsymbol{g}(\mathbf{X}_t;\mathbf{Label})$ and $\boldsymbol{g}^{st}(\mathbf{X})$ is the same, the calculation process of $\boldsymbol{g}(\mathbf{X}_t;\mathbf{Label})$ is introduced in detail in this paper. The variables for the calculation process of $\boldsymbol{g}^{st}(\mathbf{X})$ refer to the calculation process of $I^{st}(\mathbf{X})$.

Finally, the common space $\mathbf{L}$ is obtained. So, the feature data in the source domain after mapping is $\mathbf{F}_s=\mathbf{Y}_s{}^T\mathbf{L}$, and the feature data from the target domain is $\mathbf{F}_t=\mathbf{Y}_t{}^T\mathbf{L}$.

3 Experiments and results analysis

To evaluate the effectiveness of the proposed cross-corpus speech emotion recognition method, a number of experiments are conducted with some baseline methods

on three commonly standard datasets, namely Berlin [47], NNIME [48], IEMOCAP [49], MSP-Improv [50], and MSP-PODCAST [51]. The specific statistics of each dataset are shown in Table 1.

3.1 Data preparation

Berlin dataset is a German emotional speech corpus recorded by the Technical University of Berlin. In this dataset, ten actors performed 7 emotions, including neutral, angry, fearful, happy, sad, disgusted, and bored. The sampling rate is 16 kHz. The dataset contains 233 male emotional sentences and 302 female emotional sentences saved in WAV format.

The NTHU-NTUA Chinese Interactive Multimodal Emotional Corpus (i.e., NNIME) is a multimodal dataset. In this dataset, audio, video, ECG, etc. were recorded for 44 actors during oral interactions. There are 6 emotions including anger, happy, sad, neutral, frustration, and surprise in this dataset. The audio sampling rate is 16 kHz. The dataset also contains annotation results from 49 annotators in different perspectives.

IEMOCAP, known as the Interactive Emotional Binary Motion Capture Database, is recorded by the Speech Analysis and Interpretation Laboratory at the University of Southern California. Ten emotions are shown by recording the expressions, movements, and audio of 10 actors in this dataset. Twelve hours of data are contained in this dataset. The audio sampling rate is 16 kHz. Considering the relevance and ambiguity of different types of emotions, 4 typical emotions (angry, neutral, happy, and sad) audio data were selected from the above three datasets in this paper.

MSP-Improv is an improvised multimodal emotional corpus. There are 6 sessions each session is a dyadic interaction between two speakers. Twenty target sentences are consisted in each session. In this corpus, 12 actors (six male and six female) performed 4 emotions, including neutral, angry, happy, and sad. Two actors improvise these emotion-specific situations, leading them to utter contextualized, non-read renditions of sentences that have fixed lexical content and convey different emotions. The sampling rate is 44.1 kHz. MSP-Improv is more natural than other corpora. Hereinafter referred to as MSP-Improv is MSP.

Table 1 Database statistics

Database	Language	Number of samples	Emotional kinds
Berlin	German	535	7
NNIME	Chinese	4773	6
IEMOCAP	English	10,039	10
MSP-Improv	English	8438	4
MSP-PODCAST	English	104,267	9

MSP-PODCAST, a large and natural emotional corpus. It relies on existing spontaneous recordings obtained from audio-sharing websites. The criterion to select the podcasts is to include only episodes that can be shared to the broader community. In this corpus, the types of emotions and themes are diverse, and the audio quality is very good in this corpus, because segments recorded with poor quality are removed. Segments with SNR values less than 20 dB are discarded. Phone-quality speech are also removed. Therefore, this step also removes segments that do not have significant energy above 4 kHz. Podcasts in the corpus contain 9 emotions, including angry, sad, happy, neutral, fear, surprise, disgust, others, and contempt. However, angry, happy, neutral, and sad are selected in this paper. There are also many real-world corpora like LSSED [52], and so on.

3.2 Experimental settings

In this experiment, 5 artificial audio features are used, including static MFCC and their first- and second-order dynamic differences, LPC, log amplitude-frequency characteristics, Philips Fingerprints [53], and spectral entropy. The selected audio features are listed in Table 2.

In the following, the amplitude characteristic of the frequency coefficient is described by log amplitude-frequency characteristics (LAFC).

Considering that different features contribute differently to speech emotion recognition, each feature in the source domain and the target domain is weighted before training. The weights are set by the dimensions of the features in this paper. For MFCC, LPC, LAFC, Philips Fingerprint, and Spectral Entropy, the corresponding weights are β_1, β_2, β_3, β_4, and β_5, respectively.

After subspace learning and domain adaption, the weighted features in the source domain are trained. That is, the features are used to build a training set. Similarly, the weighted features in the target domain are used to build a test set.

In the training process, a constant recognition accuracy threshold α is set in advance. Next, the test set is divided into two parts of equal amount of data, i.e., test set 1 and

Table 2 The features used in this paper

Feature	Feature dimensions
Static MFCC	12
First-order dynamic difference of MFCC	12
Second-order dynamic difference of MFCC	12
LPC	12
LAFC	129
Philips fingerprint	1
Spectral entropy	1

test set 2. Test set 1 is used for assist training, and test set 2 is used to optimize the performance of the proposed method. If the recognition accuracy of a certain type of emotion is less than α in the first training, the features corresponding to the emotion need to be re-trained in the next training. The operations repeated until one of the following conditions is met: (1) the recognition accuracy of all emotions is greater than α, and (2) the number of the emotion with recognition accuracy less than α remains unchanged in the two adjacent training.

To evaluate the performance of the proposed method in the cross-corpus condition, the Berlin, NNIME, and IEMOCAP are combined in pairs in this paper. Then, any two datasets are taken as the source domain and the target domain. Therefore, a total of 6 combination cases are designed as follows:

- N-B: NNIME is the source domain dataset, and Berlin is the target domain dataset.
- B-N: Berlin is the source domain dataset, and NNIME is the target domain dataset.
- N-I: NNIME is the source domain dataset, and IEMOCAP is the target domain dataset.
- I-N: IEMOCAP is the source domain dataset, and NNIME is the target domain dataset.
- B-I: Berlin is the source domain dataset, and IEMOCAP is the target domain dataset.
- I-B: IEMOCAP is the source domain dataset, and Berlin is the target domain dataset.

3.2.1 Parameter details

Linear SVM is chosen for training and testing. The grid search method is used to optimize the kernel function coefficients of the SVM and the independent terms of the sum function. There are four hyperparameters and five feature weight coefficients in this experiment. The recognition accuracy threshold α is set to 0.45. It is determined by an informal experiment. According to the dimension of the feature, the weight coefficient β_1, β_2, β_3, β_4, and β_5 are set as 0.3, 0.3, 0.3, 0.05, and 0.05, respectively. The complexity of the algorithm is affected by K. The larger the value of K is, the higher the algorithm complexity is, and the more features are extracted. So, the range of the neighboring value K is set as [3, 9]. For the two regularization parameters μ and λ, the range is set to $\{-1/4, -1/3, -1/2, 1, 1/2, 1/3, 1/4\}$ and {0.001, 0.01, 0.1, 1, 10, 100, 1000}, respectively. Considering that embedding regularization parameter μ is an exponent, if μ is a positive integer, the value will affect the value of the element in $\mathbf{Y}$. Nevertheless, if μ is a positive or negative fraction, it may affect the value range of the element in $\mathbf{R}$. Hence, both integer and fraction can be chosen for μ. For regularization parameter λ, it affects the importance of both parts of two information entropy. For the proposed method, the dimension of the simplified subspace feature is set to 169.

3.2.2 Traditional linear baseline

In order to evaluate the performance of the proposed method for cross-corpus speech emotion recognition, on the basis of the above 6 sets of experiments, the proposed method is compared with some related most commonly used and advanced transfer learning methods. The following is an introduction to these baseline methods:

- Principal components analysis (PCA) [54]: A dimensionality reduction method that maps data into a low-dimensional subspace through linear transformation to prevent information loss as much as possible.
- Linear discriminant analysis (LDA) [55]: In this method, the projection direction that maximizes the ratio of the inter-class distance and minimizes the intra-class distance ratio is found. The subsequent classification results are affected while reducing the dimension.
- Kernel spectral regression (KSR) [56–58]: In reproducing kernel Hilbert spaces (RKHS), the problem of learning embedding functions is transformed by SR into a regression problem.
- Geodesic flow kernel (GFK) [59]: The movement of the domain is simulated by integrating an infinite number of subspaces. The changes in geometric and statistical properties from the source domain to the target domain are described by these subspaces.
- Subspace alignment (SA) [60]: SA is a transfer learning algorithm for two subspaces by matching the feature. The core of this method is to seek linear transformation to transform and align for different data.
- Manifold embedded distribution alignment (MEDA) [61]: Taking into account the importance of both conditional and marginal distributions, a domain-invariant classifier is learned via a Grassmann manifold with structural risk minimization.
- Joint distribution adaptation (JDA) [62]: The marginal probability distribution and conditional probability distribution of the source and target domains are adapted to reduce the distribution difference between different domains.
- Transfer component analysis (TCA) [63]: The data in both domains are mapped together into a high-dimensional regenerated kernel Hilbert space. In this space, the distance of data in the source domain and target domain is minimized.
- Balanced distribution adaptation (BDA) [64]: The weights of marginal and conditional distributions are adaptively utilized on the basis of JDA.
- Transfer joint matching (TJM) [65]: The domain variance is reduced by jointly matching features and

Table 3 Weighted accuracy (%) of different methods in different cases

Case	Subspace learning				Distribution adaptation					Feature selection	The proposed method
	GFK	PCA	LDA	KSR	SA	MEDA	JDA	TCA	BDA	TJM	
N-B	35.63%	32.08%	30.63%	37.29%	36.46%	35.00%	36.88%	31.04%	37.29%	37.50%	50.42%
B-N	32.71%	37.29%	34.79%	40.42%	42.92%	38.75%	49.58%	43.04%	51.04%	45.63%	67.08%
N-I	43.13%	40.83%	47.08%	47.92%	37.29%	42.29%	43.54%	57.50%	44.17%	46.04%	61.25%
I-N	33.33%	38.33%	32.29%	37.71%	41.88%	41.86%	44.79%	41.04%	44.58%	45.00%	67.75%
B-I	41.46%	39.58%	46.67%	50.21%	37.71%	47.71%	33.96%	43.33%	38.54%	46.04%	55.83%
I-B	31.88%	38.75%	42.08%	39.58%	41.04%	41.25%	38.96%	36.67%	47.29%	49.17%	46.88%
Average	36.35%	37.81%	38.92%	42.18%	39.55%	41.14%	41.29%	42.10%	43.82%	44.90%	58.20%

Table 4 Unweighted accuracy (%) of different methods in different cases

Case	Subspace learning				Distribution adaptation					Feature selection	The proposed method
	GFK	PCA	LDA	KSR	SA	MEDA	JDA	TCA	BDA	TJM	
N-B	39.1%	33.75%	31.67%	36.67%	37.71%	30.42%	33.33%	32.71%	36.86%	38.75%	48.54%
B-N	35.28%	38.75%	31.46%	40%	44.58%	31.25%	44.79%	43.75%	42.08%	45%	65.28%
N-I	43.75%	41.86%	32.71%	27.92%	37.5%	38.96%	45.42%	50.42%	31.67%	30.63%	53.33%
I-N	31.04%	30.63%	34.38%	36.67%	42.29%	31.25%	38.33%	35%	39.58%	43.88%	64.04%
B-I	42.79%	38.54%	45%	48.96%	38.33%	39.17%	34.42%	45.42%	37.92%	42.92%	52.28%
I-B	32.91%	36.08%	40.86%	34.79%	34.58%	39.58%	36.67%	34.79%	46.25%	45.83%	46.08%
Average	37.47%	36.6%	36.01%	37.5%	39.17%	35.11%	39.33%	40.35%	39.06%	41.17%	54.93%

reweighting instances across domains in a dimensionality reduction process. The new feature representations invariant to both distributional variance and uncorrelated instances are built.

3.3 Results analysis

3.3.1 Comparison with the traditional linear baseline method

In this section, the recognition accuracy of the proposed method is compared with that of some traditional linear baseline methods. The result is shown in Tables 3 and 4.

From Table 3, it is clear that the performance of the proposed method outperforms that of other methods in most cases. Only in the case of I-B, the performance of the proposed method is slightly lower than that of BDA and TJM. For the proposed method, the average recognition accuracy reached 58.20% in the six cases. In the case of I-B, the recognition accuracy is the lowest among the six cases, which is 46.88%. In contrast, in the case of I-N, the recognition accuracy reached 67.75%, which is the highest among the six cases. Compared with TJM which has the highest recognition accuracy among the baseline methods, the average recognition accuracy of the proposed method is significantly improved by 13.3%.

Although weighted accuracy is an important indicator to evaluate the overall classification performance of the model, weighted accuracy is affected by the unbalanced distribution of sample classes. Therefore, unweighted accuracy is very important for evaluating the overall classification performance of the model when the distribution of sample classes is unbalanced. It can be seen from Table 4 that the unweighted accuracy of almost all methods is lower than the weighted accuracy. For the proposed method, unweighted accuracy is 3.27% lower than weighted accuracy. Compared with the baseline method, it still has advantages.

Furthermore, we can find that the average recognition accuracy of the proposed method, distribution adaptation method, and feature selection method is higher than that of most subspace learning. The reason is that the distribution of data in different domains is different. Therefore, the recognition performance of traditional subspace learning algorithms is poor in cross-corpus speech emotion recognition. Transfer learning can be used to improve recognition performance.

In addition, the confusion matrix of the proposed method in six cases is shown in Fig. 4. It can be seen that there are two types of emotion with more than 50% recognition accuracy in most cases. In the case of N-B and N-I, the highest recognition accuracy can be achieved for neutral. From Fig. 4b and f, it is clear that the proposed method has a good recognition ability for happy, and the highest recognition accuracy can be achieved for angry in the case of I-N and B-I. Moreover, it can be also found that sad is easier to be recognized than other emotions in most cases.

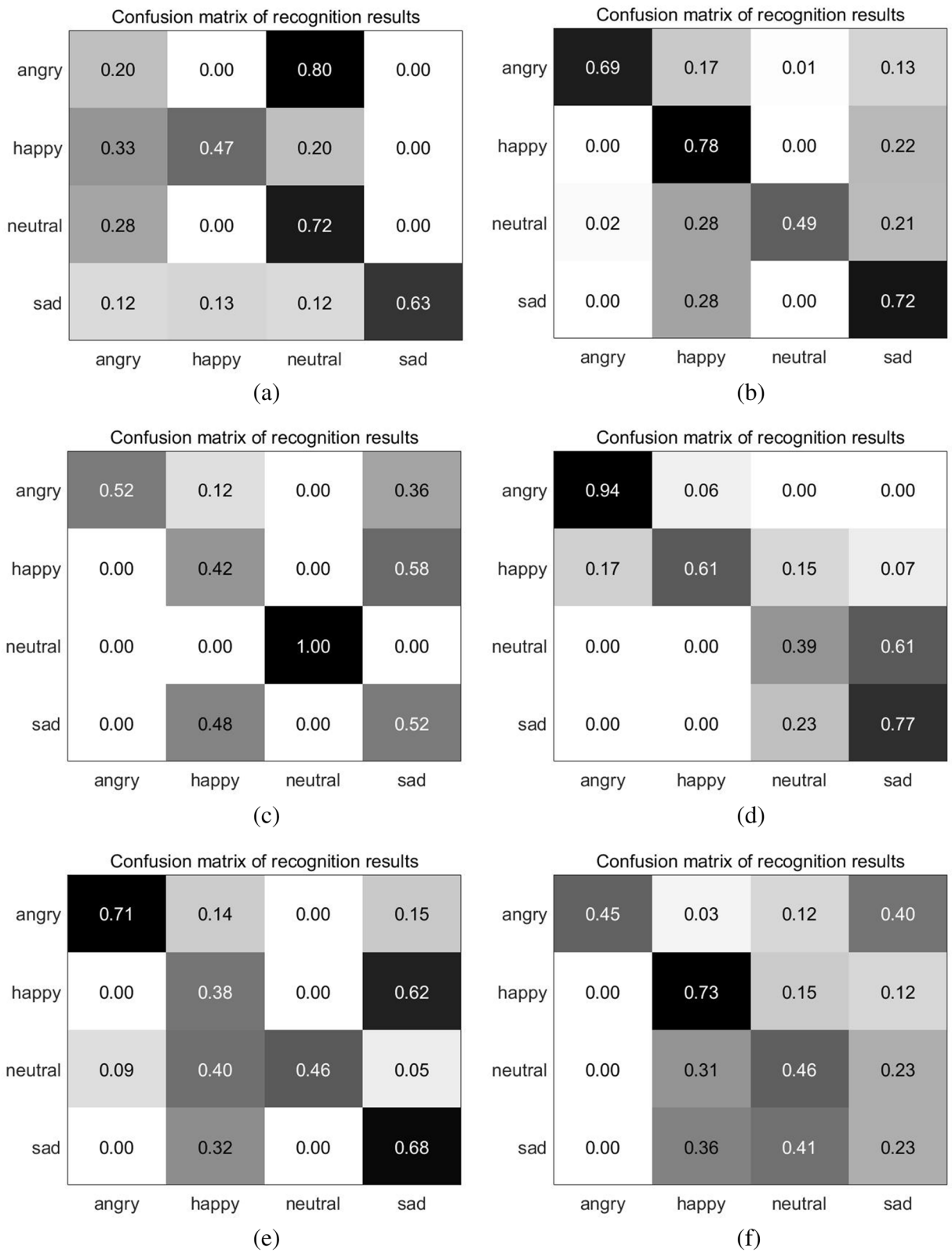

Fig. 4 Confusion matrices for cross-corpus speech emotion recognition of the proposed method in various situations. **a** Confusion matrix in the N-B case. **b** Confusion matrix in the B-N case. **c** Confusion matrix in the N-I case. **d** Confusion matrix in the I-N case. **e** Confusion matrix in the B-I case. **f** Confusion matrix in the I-B case

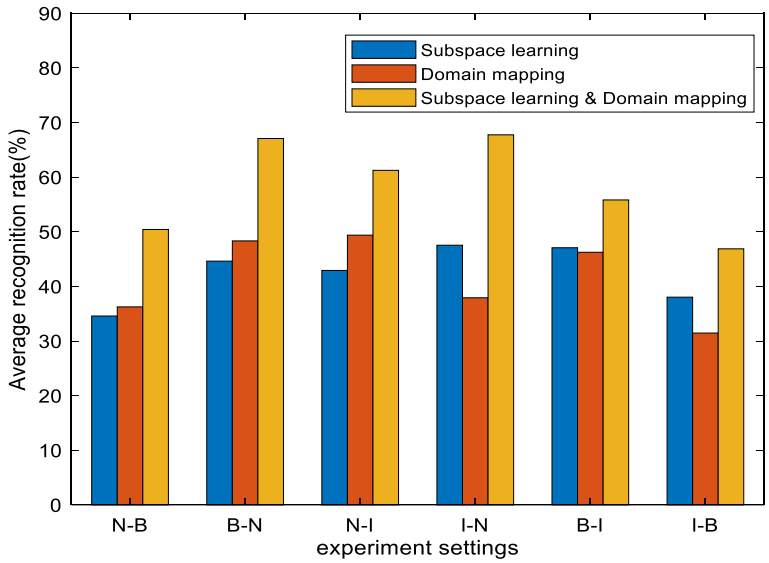

Fig. 5 Ablation experiment results

3.3.2 Ablation experiment

In this section, a set of ablation experiments is established to verify the impact of the two parts of the proposed method on the recognition performance. The specific results are shown in Fig. 5. The specific settings are as follows:

- Subspace learning: Only Hessian-based Subspace Learning is performed.
- Domain adaption: Only information entropy-based domain adaption is performed.
- Subspace learning and domain adaption: Hessian-based subspace learning and domain adaption are combined.

The average recognition accuracy of the ablation experiments is shown in Fig. 5. It can be found that the recognition performance of the combined method (i.e., the proposed method) is better than that of the method only with Hessian-based subspace learning or domain adaption. Through ablation experiments, it is clear that both Hessian-based subspace learning and domain adaption have played a positive role in cross-corpus speech emotion recognition. In the cases of N-B, B-N, and N-I, the recognition accuracy of the domain adaption method is slightly higher than that of the Hessian-based subspace learning method. On the contrary, in the cases of I-N, B-I, and I-B, the recognition accuracy of the Hessian-based subspace learning method is slightly higher than that of the domain adaption method.

3.3.3 Comparison with deep learning-based method

In this section, IEMOCAP and MSP-Improv are used for cross-corpus speech emotion recognition. ADDoG-based method and CNN-based method [34] are chosen as reference methods. The recognition accuracy of the proposed method is compared with these reference methods. The result is shown in Fig. 6:

It can be seen from Fig. 6 that when MSP-Improv is the source domain and IEMOCAP is the target domain, the unweight accuracy of the proposed method is better than that of the CNN-based method but slightly lower than that of the ADDoG-based method. However, in the corpus reverse experiment, the unweight accuracy of the proposed method is slightly higher than that of the CNN-based method and ADDoG-based method. It can be clearly seen that the performance of the ADDoG-based method is the most stable among the three methods. In general, the proposed method can achieve well performance compared with traditional linear methods and deep learning methods.

3.3.4 Experiment of real-world corpus

In order to verify that the method proposed in this paper is also effective in the real world, in this section, a real-world corpus MSP-PODCAST and several corpora in controlled experimental environments are used for cross-corpus speech emotion recognition. The experimental setup of this paper is to set MSP-PODCAST as the source corpus and target corpus respectively for experiments with other corpora. The recognition accuracy of the proposed method using MSP-PODCAST as

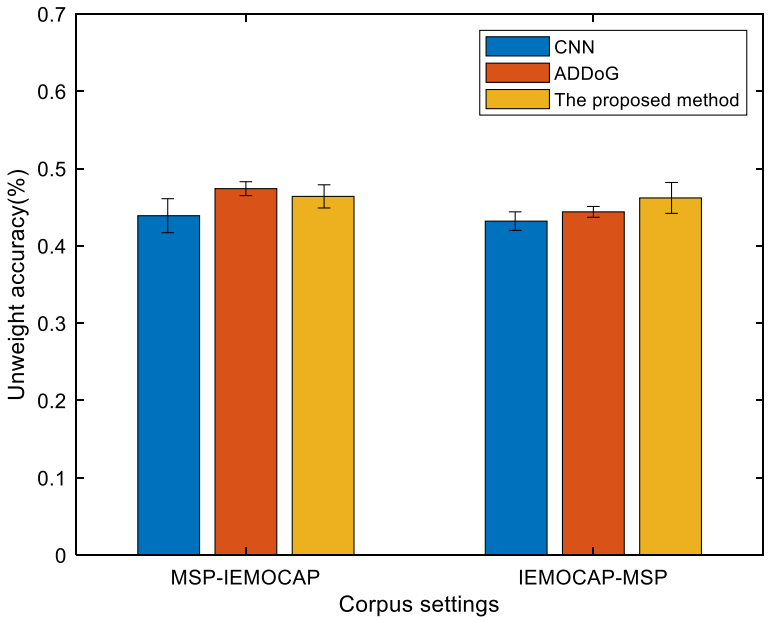

Fig. 6 Results of the unweight accuracy with IEMOCAP and MSP-Improv

the target corpus is shown in Fig. 7, and Fig. 8 shows the recognition accuracy of the accuracy of MSP-PODCAST as the source corpus:

It can be seen from Figs. 7 and 8 that the recognition performance of the proposed method using MSP-PODCAST as the target corpus is better than that using MSP-PODCAST as the source corpus. When MSP-PODCAST is used as a source corpus, the transferable knowledge is limited due to the influence of complex acoustic conditions. It can be seen that the performance of speech emotion recognition is indeed affected by the corpus environment. In addition, it is clear that the recognition

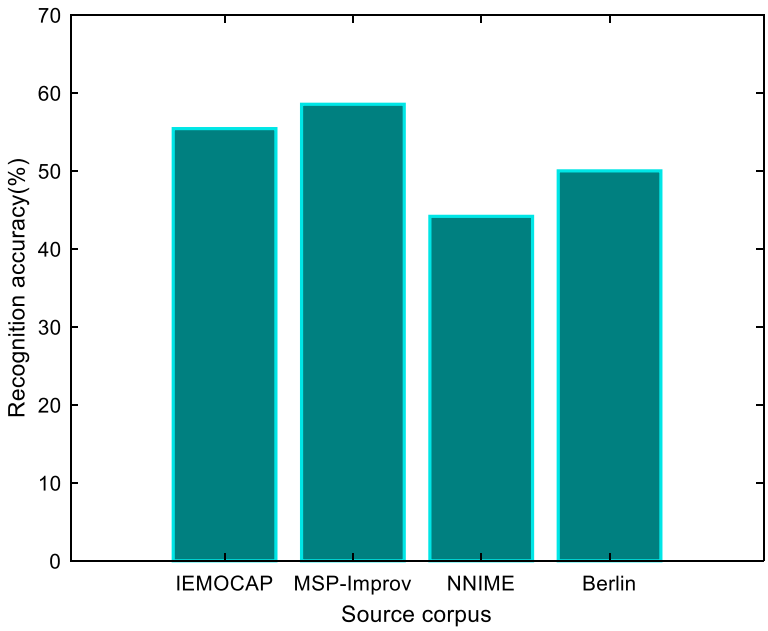

Fig. 7 Recognition accuracy of MSP-PODCAST as target corpus

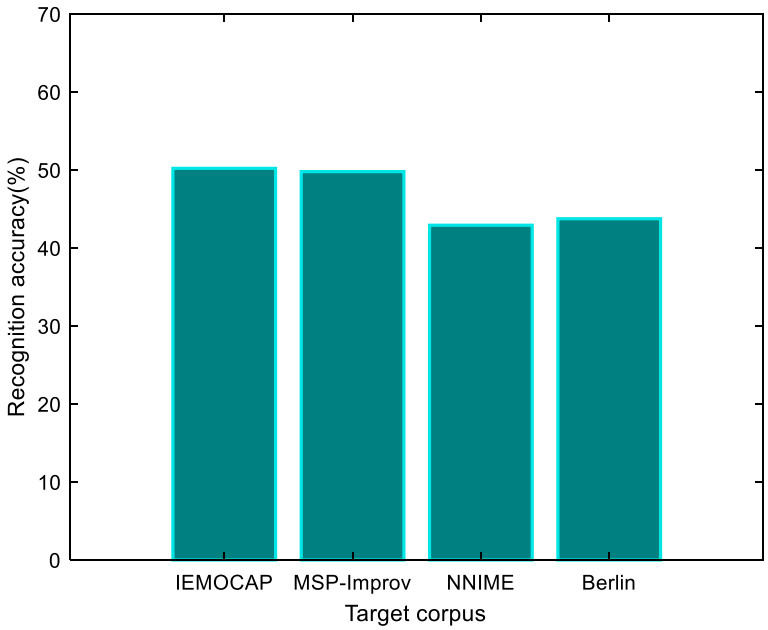

Fig. 8 Recognition accuracy of MSP-PODCAST as source corpus

performance of the proposed method using IEMOCAP and MSP-Improv is better than that of other corpora.

3.3.5 Parameters analysis

The influence of different parameters on the recognition performance of the proposed method is analyzed in this section. The analyzed parameters include the number of the nearest neighbors K, the embedding regularization parameter μ, and the information entropy regularization parameter λ. Different recognition accuracy can be obtained by selecting different values of parameters.

First of all, the nearest neighbor number K is analyzed, which is used to identify the number of neighboring frames of the current frame. The complexity of the algorithm is affected by K. The smaller K is, the fewer neighboring frames are identified, and the less feature is provided. While the larger K is, the more neighboring frames are identified, the more feature is provided. However, if K is set large, some frames which are not useful for recognition may be identified as neighboring frames, which may lead to high algorithmic complexity. So, the range of K is set from 3 to 9 in this paper. In different cases, the recognition accuracy of different K is shown in Fig. 9. From Fig. 9, we can find that the proposed method achieves a good recognition accuracy when $K = 6$. However, it is not enough to only use the recognition accuracy to measure the recognition performance under different corpus settings. Therefore, variances of recognition accuracy are introduced in parameter analysis to measure the recognition performance under different corpus settings at the same time in this paper. For K, variances under different corpus settings are shown in Fig. 10. It can be seen from Fig. 10 that, although the variances of recognition accuracy achieve the maximum when $K = 6$, there is a small difference when K takes different values. Therefore, considering the algorithmic complexity and recognition performance, K is selected as 6 in this paper.

Then, the embedding regularization parameter μ is analyzed, which is used to control the value of the embedded coordinates. The range of μ is set as $\{-1/2, -1/3, -1/4, 1/4, 1/3, 1/2, 1\}$ in this paper. In different cases, the recognition accuracy of the proposed method with different μ is shown in Fig. 11. From Fig. 11, it is clear that the proposed method can achieve a good recognition accuracy when $\mu = 1/4$. The variance of recognition accuracy with different μ under different corpus settings is shown in Fig. 12. Although the variance of recognition accuracy is very small when $\mu = 1$, the recognition accuracy is significantly lower than that under other conditions. Therefore, in consideration of recognition accuracy and variance of recognition accuracy, $\mu = 1/4$ is chosen in this paper.

Finally, the information entropy regularization parameter λ is analyzed, which controls the weight of the information entropy. The range of λ is set as $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ in this paper. In different cases, the recognition accuracy of the proposed

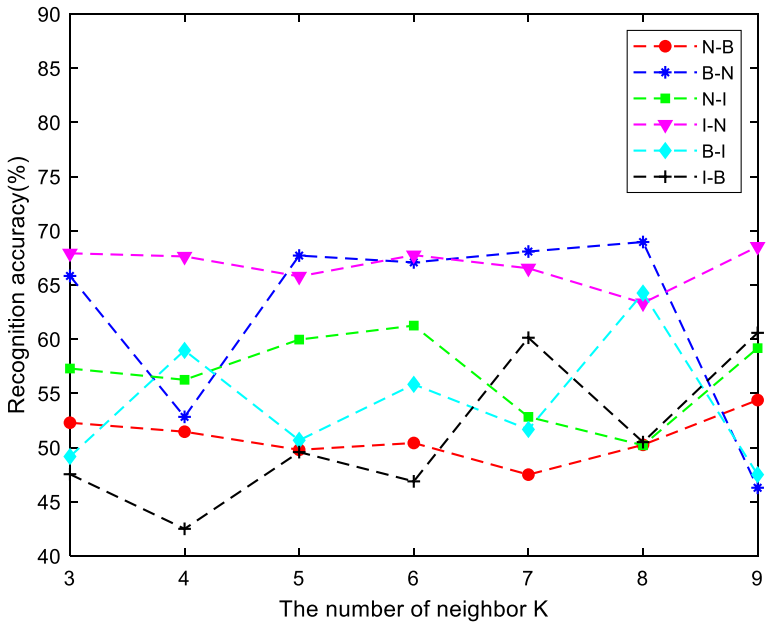

Fig. 9 Recognition accuracy with different *K*

method with different λ is shown in Fig. 13. As shown in Fig. 13, when $\lambda = 100$ and $\lambda = 1000$, the changes in the recognition accuracy are great. Although when $\lambda = 100$, the recognition accuracy in both N-I and B-I cases exceeds 70%. However, it is not stable in these two cases as shown in Fig. 14. Therefore, considering recognition accuracy and variance of recognition accuracy in a compromise, $\lambda = 10$ is chosen in this paper.

3.4 Complexity analysis

For the performance evaluation of a method, both recognition accuracy and model complexity should be considered.

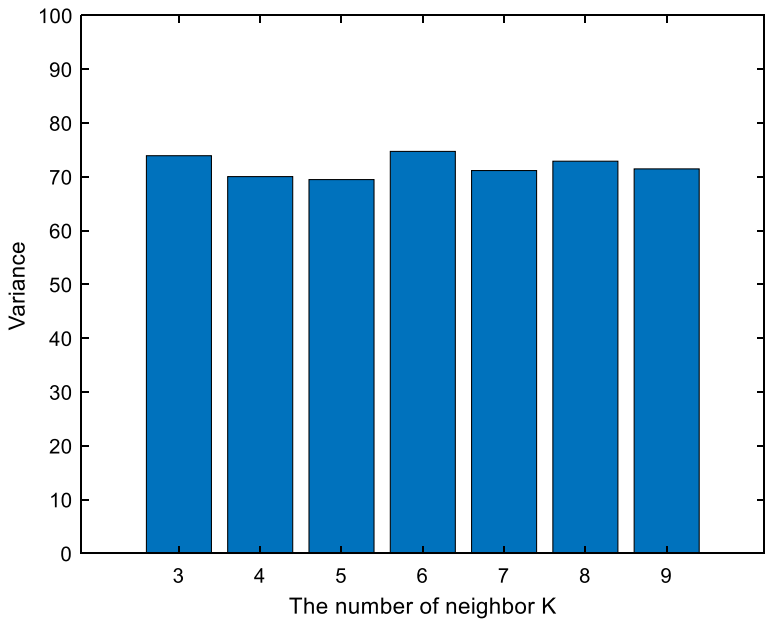

Fig. 10 Variance of recognition accuracy with different *K*

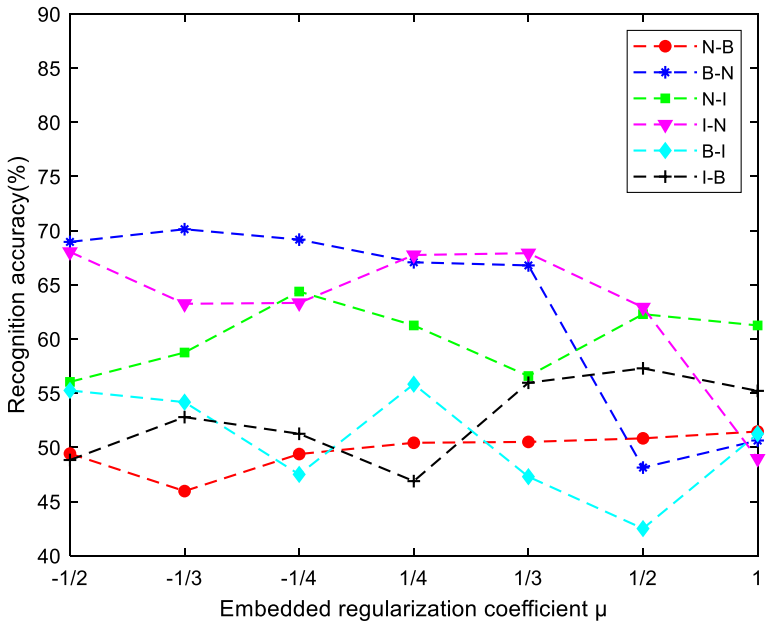

Fig. 11 Recognition accuracy with different μ

For the deep learning-based method, the complexity of the model is determined by the network structure and the number of parameters. Therefore, some complexity analysis of the proposed method and reference methods are given in this subsection. For the CNN-based method, the feature encoder consists of two convolution layers and a max pooling layer, and the emotion classifier consists of fully connected layers and softmax. On this basis, the ADDoG model adds a critic composed of full connection layers. With the increase of the input MFBs, the calculation amount and trainable parameter amount of each layer will increase more. In addition, during training, when the number of samples in the source domain and target domain increases, the computational complexity of the

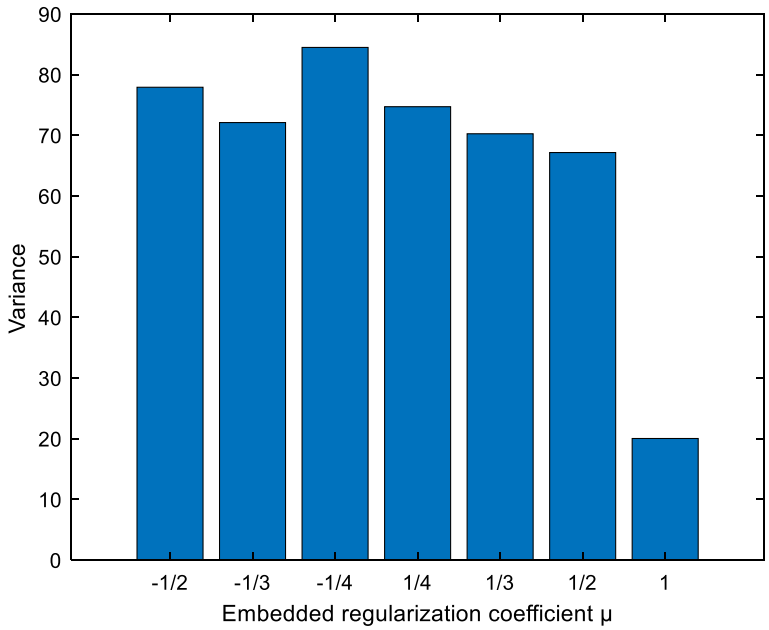

Fig. 12 Variance of recognition accuracy with different μ

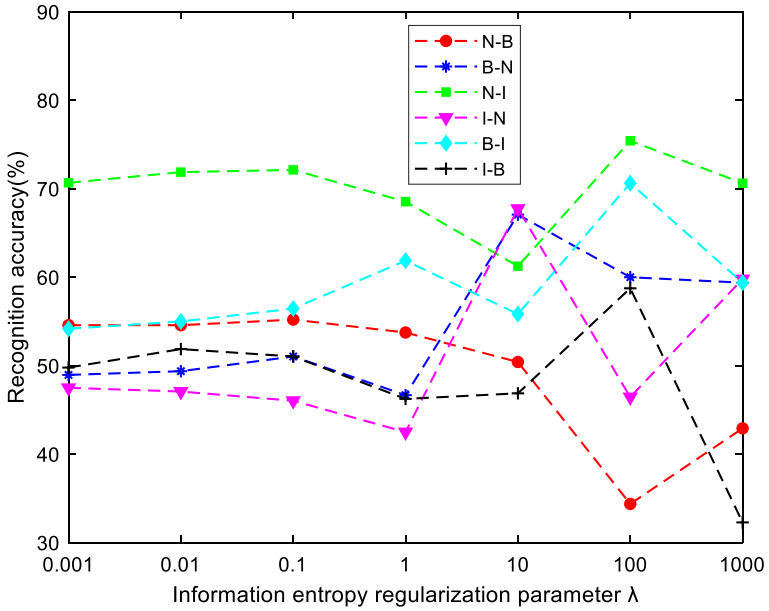

Fig. 13 Recognition accuracy with different λ

loss function and iteration times increase. Although there is a user-defined maximum number of iterations for the proposed method, convergence can be achieved by an average of 50 or fewer iterations under each experimental setting. In summary, the proposed method requires relatively few adaptation steps compared to the needing of fine-tuning whole deep neural network.

4 Conclusion

In this paper, a cross-corpus speech emotion recognition method is proposed using subspace learning and domain adaptation. In the subspace learning part, the Hessian matrix is introduced to locally embed the features in both source and target domains to form the feature subspace. In the domain adaption part, the

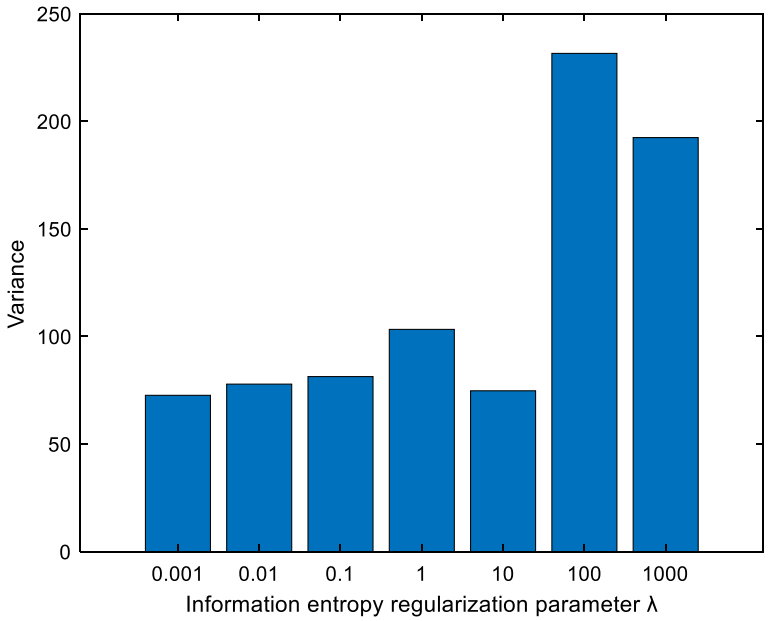

Fig. 14 Variance of recognition accuracy with different λ

mapping relationship is constructed based on information entropy. Then, the common space of both the source and target domains is obtained, which reduces the discrepancy in feature distribution between the source and target domains. Extensive experiments on datasets in three different languages are conducted to verify the performance of the proposed method.

Acknowledgements
This work was supported by the National Natural Science Foundation of China under Grants (61971015), Beijing Natural Science Foundation (No. L223033), and the Cooperative Research Project of BJUT-NTUT (No. NTUT-BJUT-110-05).

Authors' contributions
CX performed the whole research and wrote the paper. JM provided support to the writing and experiments. The authors read and approved the final version of the paper.

Funding
This work was supported by the National Natural Science Foundation of China under Grants (61971015) and the Cooperative Research Project of BJUT-NTUT (No. NTUT-BJUT-110-05).

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China. [2]Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei, Taiwan.

Received: 20 August 2022 Accepted: 14 December 2022
Published online: 27 December 2022

References

1. S. Zhao, G. Jia, J. Yang, G. Ding, K. Keutzer, Emotion recognition from multiple modalities: fundamentals and methodologies. IEEE Sign. Process. Magazine **38**(6), 59–73 (2021)
2. X. Wu, S. Hu, Z. Wu, X. Liu, H. Meng, in *2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP)*. Neural architecture search for speech emotion recognition (2022), pp. 1–4
3. C.-C. Lee, K. Sridhar, J.-L. Li, W.-C. Lin, S. Bo-Hao, C. Busso, Deep representation learning for affective speech signal analysis and processing: preventing unwanted signal disparities. IEEE Sign. Process. Magazine **38**(6), 22–38 (2021)
4. J.S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, H. Xiao, Y.-H. Yang, E. Gómez, Music emotion recognition: toward new, robust standards in personalized and context-sensitive applications. IEEE Sign. Process. Magazine **38**(6), 106–114 (2021)
5. W. Chung-Hsien, W.-B. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans. Affect. Comput. **2**(1), 10–21 (2011)
6. J.-H. Hsu, M.-H. Su, C.-H. Wu, Y.-H. Chen, Speech emotion recognition considering nonverbal vocalization in affective conversations. IEEE/ACM Trans. Audio, Speech Lang. Process. **29**, 1675–1686 (2021)
7. B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, D. Zhang, Multimodal emotion recognition with temporal and semantic consistency. IEEE/ACM Trans. Audio, Speech Lang. Process. **29**, 3592–3603 (2021)
8. B.T. Atmaja, A. Sasou, M. Akagi, Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. Speech Commun. **140**, 11–28 (2022)
9. Y. Jin, P. Song, W. Zheng, L. Zhao, Novel feature fusion method for speech emotion recognition based on multiple kernel learning. J. South. Univ. (English Edition) **29**(2), 129–133 (2013)
10. U. Garg, S. Agarwal, S. Gupta, R. Dutt, D. Singh, in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*. Prediction of emotions from the audio speech signals using MFCC, MEL and Chroma (2020), pp. 87–91
11. N.P. Jagini, R.R. Rao, in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. Exploring emotion specific features for emotion recognition system using PCA approach (2017), pp. 58–62
12. S.R. Krishna, R.R. Rao, in *2017 International Conference on Communication and Signal Processing (ICCSP)*. Exploring robust spectral features for emotion recognition using statistical approaches (2017), pp. 1838–1843
13. J.A. Russell, A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)
14. J. Posner, J.A. Russell, B.S. Peterson, The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. **17**(3), 715–734 (2005)
15. A. Mehrabian, *Basic dimensions for a general psychological theory* (Oelgeschlager, Gunn& Hain, Incorporated, Cambridge, 1980), pp. 39–53
16. R.F. Bales, *Social interaction systems: theory and measurement* (Transaction Publishers, Piscataway, 2001), pp. 139–140
17. Y. Zhou, X. Liang, Y. Gu, Y. Yin, L. Yao, Multi-classifier interactive learning for ambiguous speech emotion recognition. IEEE/ACM Trans. Audio, Speech Lang. Process. **30**, 695–705 (2022)
18. Y. Pan, P. Shen, L. Shen, Speech emotion recognition using support vector machine. Int. J. Smart Home **6**(2), 101–108 (2012)
19. S. Mao, D. Tao, G. Zhang, P.C. Ching, T. Lee, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*. Revisiting hidden Markov models for speech emotion recognition (2019), pp. 6715–6719
20. H. Hu, M. Xu, W. Wu, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. GMM supervector based SVM with spectral features for speech emotion recognition (2007), pp. IV-413–IV-416
21. Y.-C. Kao, C.-T. Li, T.-C. Tai, J.-C. Wang, in *2021 9th International Conference on Orange Technology (ICOT)*. Emotional speech analysis based on convolutional neural networks (2021), pp. 1–4
22. C.-H. Park, D.-W. Lee, K.-B. Sim, in *2002 International Conference on Machine Learning and Cybernetics*. Emotion recognition of speech based on RNN, vol 4 (2002), pp. 2210–2213
23. S. Wang, X. Ling, F. Zhang, J. Tong, in *2010 International Conference on Measuring Technology and Mechatronics Automation*. Speech emotion recognition based on principal component analysis and back propagation neural network (2010), pp. 437–440
24. K.H. Lee, H. Kyun Choi, B.T. Jang, D.H. Kim, in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*. A study on speech emotion recognition using a deep neural network (2019), pp. 1162–1165
25. X. Wu et al., in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Speech emotion recognition using sequential capsule networks, vol 29 (2021), pp. 3280–3291
26. L. Yi, M.-W. Mak, Improving speech emotion recognition with adversarial data augmentation network. IEEE Trans. Neural Netw. Learn. Syst. **33**(1), 172–184 (2022)
27. S. Mao, P.C. Ching, T. Lee, Enhancing segment-based speech emotion recognition by iterative self-learning. IEEE/ACM Trans. Audio, Speech Lang. Process. **30**, 123–134 (2022)
28. N. Liu et al., Transfer subspace learning for unsupervised cross-corpus speech emotion recognition. IEEE Access **9**, 95925–95937 (2021)
29. P. Song, W. Zheng, Feature selection based transfer subspace learning for speech emotion recognition. IEEE Trans. Affect. Comput. **11**(3), 373–382 (2020)
30. H. Luo, J. Han, Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition. IEEE/ACM Trans. Audio, Speech Lang. Process. **28**, 2047–2060 (2020)
31. P. Song, Transfer linear subspace learning for cross-corpus speech emotion recognition. IEEE Trans. Affect. Comput. **10**(2), 265–275 (2019)
32. J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Universum autoencoder-based domain adaptation for speech emotion recognition. IEEE Sign. Process. Lett. **24**(4), 500–504 (2017)
33. Y. Zong, W. Zheng, T. Zhang, X. Huang, Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. IEEE Sign. Process. Lett. **23**(5), 585–589 (2016)

34. J. Gideon, M.G. McInnis, E.M. Provost, Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). IEEE Trans. Affect. Comput. **12**(4), 1055–1068 (2021)
35. M. Abdelwahab, C. Busso, Domain adversarial for acoustic emotion recognition. IEEE/ACM Trans. Audio, Speech Lang. Process. **26**(12), 2423–2435 (2018)
36. W. Zhang, P. Song, Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition. IEEE/ACM Trans. Audio, Speech Lang. Process. **28**, 307–318 (2020)
37. D.L. Donoho et al., Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. U. S. A. **100**(10), 5591–5596 (2003)
38. Lianbo Zhang, D. Tao and Weifeng Liu, in Proceedings of the 16th International Conference on Communication Technology. Supervised Hessian Eigenmap for dimensionality reduction (IEEE, Hangzhou, China, 2015), pp.903–907.
39. F. Asano, Y. Suzuki, D.C. Swanson, Optimization of control source configuration in active control systems using Gram-Schmidt orthogonalization. IEEE Trans. Speech Audio Process. **7**(2), 213–220 (1999)
40. F. Nie, H. Huang, X. Cai, et al, in Proceedings of the 24th Annual Conference on Neural Information Processing Systems. Efficient and Robust Feature Selection via Joint ℓ2, 1-Norms Minimization (NIPS, Vancouver, BC, Canada, 2010), pp.1–9
41. R. He, T. Tan, L. Wang, W. Zheng, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. ℓ2, 1 regularized correntropy for robust feature selection (2012), pp. 2504–2511
42. Y. Shi, F. Sha, in Proceedings of the 29th International Conference on Machine Learning. Information-Theoretical Learning of Discriminative Clusters for Unsupervised Domain Adaptation (IMLS, Edinburgh, United kingdom, 2012), pp.1079–1086
43. B. Gholami, P. Sahu, O. Rudovic, K. Bousmalis, V. Pavlovic, Unsupervised multi-target domain adaptation: an information theoretic approach. IEEE Trans. Image Process. **29**, 3993–4002 (2020)
44. Y. Tu, M. Mak, J. Chien, Variational domain adversarial learning with mutual information maximization for speaker verification. IEEE/ACM Trans. Audio, Speech Lang. Process. **28**, 2013–2024 (2020)
45. D. Xin, T. Komatsu, S. Takamichi, H. Saruwatari, in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual TTS (2021), pp. 6608–6612
46. X. Wang, L. Yan and Q. Zhang, in Proceedings of the International Conference on Computer Network, Electronic and Automation. Research on the Application of Gradient Descent Algorithm in Machine Learning (IEEE, Xi'an, China, 2021), pp. 11–15
47. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, in Proceedings of the Interspeech. A database of German emotional speech (ISCA, Lisbon, Portugal, 2005), pp. 1517–1520
48. H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, C.-C. Lee, in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. NNIME: the NTHU-NTUA Chinese interactive multimodal emotion corpus (2017), pp. 292–298
49. C. Busso, M. Bulut, C.C. Lee, et al., IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resourc. Eval. **42**(4), 335–359 (2008)
50. C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, E.M. Provost, MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. IEEE Trans. Affect. Comput. **8**(1), 119–130 (2017)
51. R. Lotfian, C. Busso, Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. IEEE Trans. Affect. Comput. **10**(4), 471–483 (2019)
52. Fan, Weiquan, et al, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing. LSSED: a large-scale dataset and benchmark for speech emotion recognition (IEEE, Toronto, Canada,2021), pp. 641–645
53. J. Haitsma, T. Kalker, in Proceedings of the 3rd International Conference on Music Information Retrieval. A highly robust audio fingerprinting system (ISMIR, Paris, France, 2002), pp. 107–115
54. Y.C. Du, W.C. Hu, L.Y. Shyu, in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. The effect of data reduction by independent component analysis and principal component analysis in hand motion identification (2004), pp. 84–86
55. S. Ji, J. Ye, Generalized linear discriminant analysis: a unified framework and efficient model selection. IEEE Trans. Neural Netw. **19**(10), 1768–1782 (2008)
56. D. Cai, Spectral Regression: A Regression Framework for Efficient Regularized Subspace Learning. (Doctoral dissertation, University of Illinois at Urbana-Champaign), 2009
57. D. Cai, X. He, J. Han, Speed up kernel discriminant analysis. Int. J. Very Large Data Bases **20**(1), 187–191 (2011)
58. D. Cai, X. He, J. Han, in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. Spectral regression: a unified approach for sparse subspace learning (2007), pp. 73–82
59. B. Gong, Y. Shi, F. Sha, K. Grauman, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Geodesic flow kernel for unsupervised domain adaptation (2012), pp. 2066–2073
60. B. Fernando, A. Habrard, M. Sebban, T. Tuytelaars, in *2013 IEEE International Conference on Computer Vision*. Unsupervised visual domain adaptation using subspace alignment (2013), pp. 2960–2967
61. J. Wang, W. Feng, Y. Chen, et al, in Proceedings of the ACM Multimedia Conference. Visual Domain Adaptation with Manifold Embedded Distribution Alignment (ACM, Seoul, Korea, 2018), pp. 402–410
62. M. Long, J. Wang, G. Ding, J. Sun, P.S. Yu, in *2013 IEEE International Conference on Computer Vision*. Transfer feature learning with joint distribution adaptation (2013), pp. 2200–2207
63. S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis. IEEE Trans. Neural Netw. **22**(2), 199–210 (2011)
64. J. Wang, Y. Chen, S. Hao, W. Feng, Z. Shen, in *2017 IEEE International Conference on Data Mining (ICDM)*. Balanced distribution adaptation for transfer learning (2017), pp. 1129–1134
65. M. Long, J. Wang, G. Ding, J. Sun, P.S. Yu, in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Transfer joint matching for unsupervised domain adaptation (2014), pp. 1410–1417

Publisher's Note

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com