Chinese Journal of Electronics
Vol.32, No.3, May 2023

Improved Cross-Corpus Speech Emotion Recognition Using Deep Local Domain Adaptation

ZHAO Huijuan[1,2], YE Ning[3], and WANG Ruchuan[3,4]

(1. *College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*)

(2. *College of Computer and Software, Nanjing Vocational University of Industry Technology, Nanjing 210023, China*)

(3. *College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*)

(4. *Jiangsu High Technology Research Key Laboratory for Wireless Sensor Network, Nanjing 210003, China*)

Abstract — Due to the scarcity of high-quality labeled speech emotion data, it is natural to apply transfer learning to emotion recognition. However, transfer learning-based speech emotion recognition becomes more challenging because of the complexity and ambiguity of emotion. Domain adaptation based on maximum mean discrepancy considers marginal alignment of source domain and target domain, but not pay regard to class prior distribution in both domains, which results in the reduction of transfer efficiency. In order to address the problem, this study proposes a novel cross-corpus speech emotion recognition framework based on local domain adaption. A category-grained discrepancy is used to evaluate the distance between two relevant domains. According to research findings, the generalization ability of the model is enhanced by using the local adaptive method. Compared with global adaptive and non-adaptive methods, the effectiveness of cross-corpus speech emotion recognition is significantly improved.

Key words — Transfer learning, Domain adaptation, Maximum mean discrepancy, Cross-corpus speech emotion recognition.

I. Introduction

Speech is one of the important natural ways of human-to-human and human-computer interaction [1], [2]. Speech emotion recognition (SER) is an active direction in the field of speech research and it is also an important prerequisite for harmonious human-computer interaction. At present, SER has attracted widespread attention from academia and industry, and has made great progress. In most current SER researches, training data and testing data are sampled from the same corpus, but in practical cases, training data and testing data usually belong to different databases [3].

The success of cross-corpus SER is one of the important factors affecting the wide application of SER [4]. Besides recognition accuracy, high generalization ability is also an important objective of model optimization. Furthermore, it is obviously resource-consuming to train a new model from scratch for each task. Emotion generation and perception are affected by attributes such as speaking style and speaker's country, gender, and age [2], [5]. The subjectivity and uncertainty of emotion make cross-corpus SER more challenging.

At present, researches on cross-corpus SER are often combined with transfer learning and deep learning. We divide the cross-corpus SER methods into two categories according to the learning process: 1) Assisted learning: Source task is used to assist target task in order to improve the performance of the target task. The key is that source task has the knowledge to transfer and both homogeneous transfer learning and heterogeneous transfer learning are covered. The relevant information can be attributes of the speaker, speech script, the dimensional information of emotion. This type is mainly based on supervised transfer learning, including multitask learning [6], [7] and pretrain-finetuning [8]. 2) Dependent learning: The target task depends on the source task, usually the two domains perform domain adaptation [9], [10], including supervised domain adaptation and unsupervised domain adaptation. This work focuses on the second type of methods, which aims to

Manuscript Received May 28, 2021; Accepted Sept. 28, 2021. This work was supported by the National Natural Science Foundation of China (61572260), Postgraduate Research & Practice Innovation Program of Jiangsu Province (46035CX17789), and Research Project of Nanjing Vocational University of Industry Technology (YK20-05-08).

 DOI:10.23919/cje.2021.00.196

improve the performance of cross-corpus SER. Our main work and contributions are summarized as follows:

1) We propose a new cross-corpus speech emotion recognition model based on the local domain adaptation (L-DA).

2) We design a weighted maximum mean discrepancy based on emotion category to measure the distance between source data and target data.

3) We conduct comprehensive experiments on two public corpora and verify the effectiveness of the proposed method.

II. Related Work

In this section, we present the most relevant work related to our approach, including domain adaptation and maximum mean discrepancy (MMD).

1. Domain adaptation

Domain adaptation [11]: Given a source domain $D_{\mathrm{s}} = \{(x_i^{\mathrm{s}}, y_i^{\mathrm{s}})\}_{i=1}^{n}$ and a target domain $D_{\mathrm{t}} = \{(x_j^{\mathrm{t}})\}_{j=1}^{m}$ assuming that their feature spaces and category spaces are the same, but the marginal distributions of these two domains are different, i.e., $P_{\mathrm{s}}(X^{\mathrm{s}}) \neq P_{\mathrm{t}}(X^{\mathrm{t}})$. The purpose of domain adaptation is to learn a classifier $f(\cdot)$ from source domain D_{s} and target domain D_{t}. Usually, the source data and target data are mapped into a common feature space where the source domain and the target domain are as close as possible [12], so that the trained source domain model can predict target labels. Conceptually, cross-corpus learning and domain adaptation are different. Cross-corpus learning refers that training data and testing data are sampled from different distributions. Domain adaptation is a branch of transfer learning, which emphasizes adaptive learning in source domain and target domain.

2. Maximum mean discrepancy

According to the definition of transfer learning, the similarity between the source domain D_{s} and the target domain D_{t} is the basis of transfer learning. Currently, there are many ways to measure the difference between two distributions. Direct similarity measures include mutual information [13] and Jaccard correlation coefficient. Distance metrics include Kullback-Leibler divergence [14], Jensen-Shannon distance, maximum mean discrepancy, etc.

MMD has been widely used in transfer learning, especially in the applications of domain adaptation [15], [16]. Mapping features from a low-dimensional space to a high-dimensional reproducing kernel Hilbert space. In the high-dimensional space, the distance between two distributions is used to evaluate their similarity. When this metric is used in loss function, the goal of optimization is to minimize this distance. Suppose the source data X^{s} and target data X^{t} are $X^{\mathrm{s}} = \{x_1^{\mathrm{s}}, x_2^{\mathrm{s}}, \ldots, x_{n_{\mathrm{s}}}^{\mathrm{s}}\}$ and $X^{\mathrm{t}} = \{x_1^{\mathrm{t}}, x_2^{\mathrm{t}}, \ldots, x_{n_{\mathrm{t}}}^{\mathrm{t}}\}$ respectively, and they are sampled from two different distributions. Then the definition of MMD of the two distributions can be expressed as

$$\begin{aligned}
&\mathrm{MMD}^2\left(X^{\mathrm{s}}, X^{\mathrm{t}}\right) \\
&= \left\| \frac{1}{n_{\mathrm{s}}} \sum_{i=1}^{n_{\mathrm{s}}} \phi\left(x_i^{\mathrm{s}}\right) - \frac{1}{n_{\mathrm{t}}} \sum_{j=1}^{n_{\mathrm{t}}} \phi\left(x_j^{\mathrm{t}}\right) \right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n_{\mathrm{s}}^2} \sum_{i=1}^{n_{\mathrm{s}}} \sum_{j=1}^{n_{\mathrm{s}}} k\left(x_i^{\mathrm{s}}, x_j^{\mathrm{s}}\right) + \frac{1}{n_{\mathrm{t}}^2} \sum_{i=1}^{n_{\mathrm{t}}} \sum_{j=1}^{n_{\mathrm{t}}} k\left(x_i^{\mathrm{t}}, x_j^{\mathrm{t}}\right) \\
&\quad - \frac{2}{n_{\mathrm{s}} n_{\mathrm{t}}} \sum_{i=1}^{n_{\mathrm{s}}} \sum_{j=1}^{n_{\mathrm{t}}} k\left(x_i^{\mathrm{s}}, x_j^{\mathrm{t}}\right)
\end{aligned} \tag{1}$$

where $\phi(\cdot) : X \longmapsto \mathcal{H}$, and the kernel function k means $k(x^{\mathrm{s}}, x^{\mathrm{t}}) = \langle \phi(x^{\mathrm{s}}), \phi(x^{\mathrm{t}}) \rangle_{\mathcal{H}}$.

Set $k_{\mathrm{s,t}} = \sum_{i=1}^{n_{\mathrm{s}}} \sum_{j=1}^{n_{\mathrm{t}}} k\left(x_i^{\mathrm{s}}, x_j^{\mathrm{t}}\right)$, then formula (1) can be transformed into the following matrix form:

$$\mathrm{MMD}\left(X^{\mathrm{s}}, X^{\mathrm{t}}\right) = \mathrm{tr}\left(\boldsymbol{K} \times \boldsymbol{N}\right)$$

where

$$\boldsymbol{K} = \begin{bmatrix} k_{\mathrm{s,s}} & k_{\mathrm{s,t}} \\ k_{\mathrm{t,s}} & k_{\mathrm{t,t}} \end{bmatrix}$$

$$\boldsymbol{N} = \begin{bmatrix} \dfrac{1}{n_{\mathrm{s}}^2} & -\dfrac{1}{n_{\mathrm{s}} n_{\mathrm{t}}} \\ -\dfrac{1}{n_{\mathrm{s}} n_{\mathrm{t}}} & \dfrac{1}{n_{\mathrm{t}}^2} \end{bmatrix}$$

MMD is first proposed to neural network to minimize the gap of deep features across domains in [17], [18]. When multiple kernel functions are used in MMD, we called MK-MMD, which achieves better performance than single kernel MMD in applications [19], [20].

In the field of SER, Song *et al.* used MMD to evaluate the similarity of different corpora, and used non-negative matrix factorization for computation [21]. Liu *et al.* proposed a deep domain adaptive convolutional neural network for cross-corpus SER and compared the performance of different models [22]. Both researches focused on how to reduce the global level differences between training and testing data, but did not consider the differences between training and testing data at the emotion category level.

III. Proposed Method

In this section, we propose a framework for speech emotion recognition based on L-DA, which using local maximum mean discrepancy (LMMD). As shown in Fig.1, the framework is represented in the top rectangle box, and the contents in the rounded rectangle at the bottom is the implementation of the corresponding

module. According to the pipeline, the raw speech signals in the source and target domains are preprocessed firstly and then fed into the model for feature learning. After feature representation is acquired, local domain adaptation and emotion classifier are followed. Here the LMMD operation is added to the fully connected layer. The main objective of the framework is to predict the emotion category according to the target utterance.

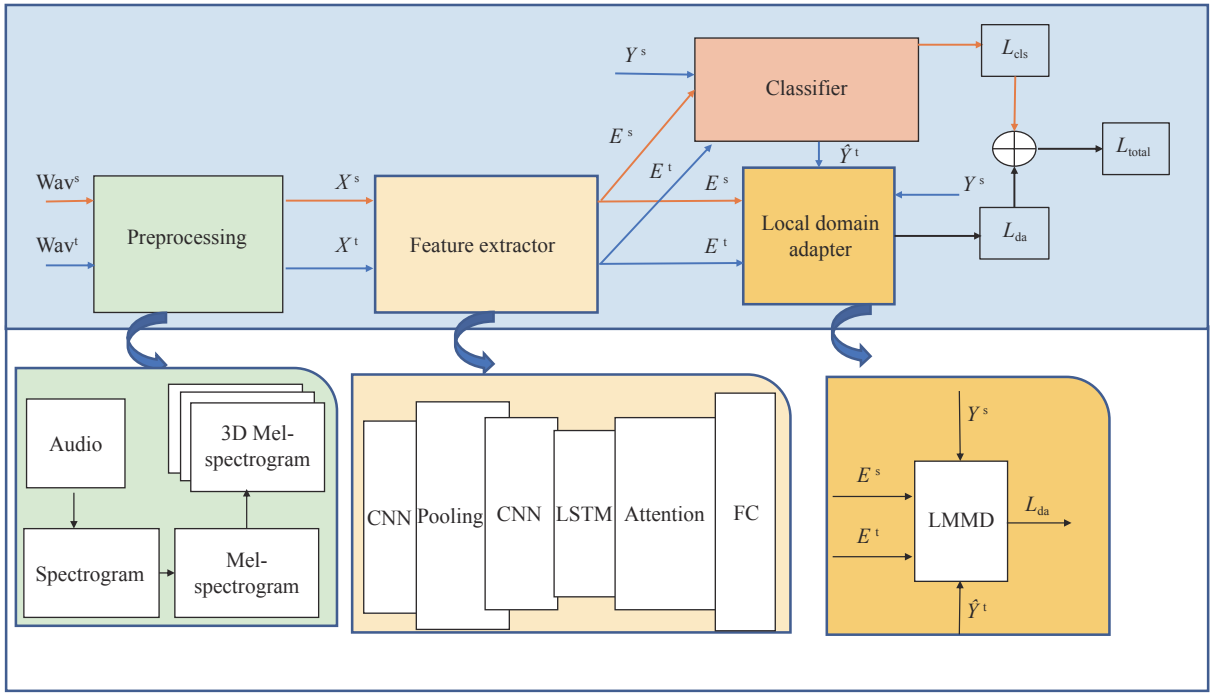

Fig. 1. Speech emotion recognition framework based on L-DA.

1. Cross-corpus SER framework based on the L-DA

The model mainly consists of three core components, namely feature extractor, domain adapter and emotion classifier. The feature extractor is used to learn the embedding of the input, which are shared by the source domain and target domains.

The domain adapter calculates the similarity between source and target domains at the emotion category level. Therefore, the domain adaptation loss function is usually defined by a distance metric, which is used to measure the similarity of two distributions. Corresponding to L-DA, the adaptation in which the source and target domains within the whole domain without any division is named global domain adaptation (G-DA). The classifier obtains the emotion category through cross-entropy and the activation function.

The model consists of emotion classification task and domain adaptation task with a parameter λ as the tradeoff of the two tasks. The total loss function is defined as (2), where $L_{\rm cls}$ represents the emotion classification loss function shown as (3) and $L_{\rm da}$ represents the domain adaptation loss function. L-DA can be used in multiple layers. In this work, L is the total number of the adaptation layers, and $L_{\rm da}$ is the sum of the adaptation loss functions in each adaptation layer, as shown in (4).

$$L_{\rm total} = L_{\rm cls} + \lambda L_{\rm da} \tag{2}$$

$$L_{\rm cls} = \frac{1}{n_{\rm s}} \sum\nolimits_{i=1}^{n_{\rm s}} J\left(f\left(x_i^{\rm s}\right), y_i^{\rm s}\right) \tag{3}$$

$$L_{\rm da} = \sum\nolimits_{l=1}^{L} L_{{\rm da}_l} \tag{4}$$

The method of local maximum mean discrepancy is described in Section III.2. After the model converges, the model can be used to predict the emotion category of the target domain.

2. Local maximum mean discrepancy

In domain adaptation for SER, the distribution of source data is different from that of target data and the domain shift should be reduced. It is difficult to achieve fine-grained knowledge transfer only using the G-DA. LMMD divides the data distribution into different subdomains, and aligns the data in the sub-domains to achieve fine-grained domain adaptation. Good results have been achieved in computer vision applications [23], [24]. However, there are few relevant studies in the field of SER.

We use LMMD to implement L-DA in SER framework. First, the source domain and the target domain

are divided into different subdomains according to emotion categories. Then, the maximum mean discrepancy with category weight between the source data and the target data are calculated for each specific emotion category until all categories are calculated. The square distance between the source domain and the target domain is expressed as follows:

$$\begin{aligned}&\mathrm{LMMD}^2\left(X^{\mathrm{s}}, X^{\mathrm{t}}\right)\\&=\frac{1}{c}\sum_{c=1}^{C}\left\|\sum_{x_i^{\mathrm{s}}\in D_{\mathrm{s}}} w_i^{\mathrm{sc}}\phi\left(x_i^{\mathrm{s}}\right)-\sum_{x_j^{\mathrm{t}}\in D_{\mathrm{t}}} w_j^{\mathrm{tc}}\phi\left(x_j^{\mathrm{t}}\right)\right\|_{\mathcal{H}}^2\end{aligned}\tag{5}$$

where C is the number of emotion categories, meanwhile is the number of sub-domains, w_i^{sc}, w_j^{tc} represent the probability that the feature belongs to the source domain and the target domain for category c respectively, $\phi(\cdot)$ is for feature learning, $\{e_i^{\mathrm{s}}\}_{i=1}^{n_{\mathrm{s}}}$ and $\{e_j^{\mathrm{t}}\}_{j=1}^{n_{\mathrm{t}}}$ are the outputs of the feature extractor, $\mathrm{LMMD}^2(X^{\mathrm{s}}, X^{\mathrm{t}})$ means that the LMMD algorithm and the distance of the feature distribution are converted to the form of kernel function representation as follows:

$$\begin{aligned}&\left\|\sum_{x_i^{\mathrm{s}}\in D_{\mathrm{s}}} w_i^{\mathrm{sc}}\phi\left(x_i^{\mathrm{s}}\right)-\sum_{x_j^{\mathrm{t}}\in D_{\mathrm{t}}} w_j^{\mathrm{tc}}\phi\left(x_j^{t}\right)\right\|_{\mathcal{H}}^2\\&=\sum_{i=1}^{n_{\mathrm{s}}}\sum_{j=1}^{n_{\mathrm{s}}} w_i^{\mathrm{sc}}w_j^{\mathrm{sc}}k(e_i^{\mathrm{s}}, e_j^{\mathrm{s}})+\sum_{i=1}^{n_{\mathrm{t}}}\sum_{j=1}^{n_{\mathrm{t}}} w_i^{\mathrm{tc}}w_j^{\mathrm{tc}}k(e_i^{\mathrm{t}}, e_j^{\mathrm{t}})\\&\quad-2\sum_{i=1}^{n_{\mathrm{s}}}\sum_{j=1}^{n_{\mathrm{t}}} w_i^{\mathrm{sc}}w_j^{\mathrm{tc}}k(e_i^{\mathrm{s}}, e_j^{\mathrm{t}})\end{aligned}\tag{6}$$

It can be seen from (6) that LMMD calculation is based on the implementation of MMD, and the corresponding weights are added before each kernel function. The weight is essentially the probability that the emotion category appearing in the field. The probability calculation depends on the labels in the source and target domains. Since the labels in the target domain are unknown during training process, pseudo-labels are obtained to calculate LMMD values. Instead of using the neural network to predict the target domain label directly from the input, probabilities are used as the target labels for different categories of the given input.

In this work, Gaussian function is chosen as the kernel function, which is defined as $k(x, y) = \exp(-|x-y|^2)$. It can map the finite-dimensional input to the infinite-dimensional space and has achieved good performance in the applications of MMD [19].

3. Algorithm description

Algorithm 1 shows our L-DA training process, which includes feature extraction, domain adaptation, network backpropagation and emotion classification.

Algorithm 1 L-DA for cross-corpus speech emotion recognition

Input λ; $D_{\mathrm{s}} = \{(x_i^{\mathrm{s}}, y_i^{\mathrm{s}})\}_{i=1}^{n_{\mathrm{s}}}$; $D_{\mathrm{t}} = \left\{\left(x_j^{\mathrm{t}}\right)\right\}_{j=1}^{n_{\mathrm{t}}}$.
Output: classifier $f(\cdot)$.
Initialize: network parameters θ, training epoches: *epoches*, learning rate: lr, batch size: bz.
for $t = 1, 2, \ldots,$ *epoches* do
  Sample $\{(x_i^{\mathrm{s}}, y_i^{\mathrm{s}})\}_{i=1}^{bz}$ and $\left\{\left(x_j^{\mathrm{t}}\right)\right\}_{j=1}^{bz}$ from D_{s} and D_{t};
  Compute $\hat{y}^{\mathrm{s}}$ and $\hat{y}^{\mathrm{t}}$ in minibatch;
  Compute L_{cls} according to (3);
  Compute L_{da} according to (4) and (5);
  Compute L_{total} according to (2);
  Backpropagate with the objective L_{total} in (2);
  Update network parameters θ;
end for

IV. Experiments

In order to evaluate the effectiveness of the proposed cross-corpus SER method, we design four sets of experiments on the public corpora IEMOCAP [25] and Emo-DB [26]. Three methods are used in each group of experiments, and then the experimental results were comprehensively analyzed.

1. Emotion corpus

IEMOCAP is a multimodal emotion corpus that has been widely used in SER research. The corpus consists of two subsets of scripted data and spontaneous data. Scripted data are collected when the subjects are asked to perform the selected scripts, and spontaneous data are collected when the subjects are asked to improvise based on some hypothetical scenarios. Although the speaker's attributes are the same, due to the different speaking scenarios, different recognition results will be produced even under the same model. Some SER studies treat the two subsets differently, or only use one subset [27]. In this work, we choose two subsets of data as two different domains to evaluate domain adaptation algorithms.

Emo-DB is a German corpus in which ten actors express their emotions according to ten pre-defined sentences. A total of 535 emotional expressions, covering seven emotional categories of happiness, anger, disgust, fear, sadness, neural and surprise. The sample distribution of the selected corpus in the experiments is shown in Table 1.

2. Data preprocessing

Data preprocessing plays an important role for deep learning. The preprocessing of this work mainly includes data filtering, data augmentation, data segmentation, data transformation and data normalization.

The first step is emotion category selection. Com-

Table 1. The corpora information for the cross-corpus SER experiments

Flag	Description	Happy	Anger	Neural	Sad	Sum	Related transfer tasks
IS	IEMOCAP-scripted	311	814	609	476	2210	IS→II, II→IS
II	IEMOCAP-improved	284	289	1099	608	2280	IS→II, II→IS
I	IEMOCAP	595	1103	1708	1084	4490	I→E, E→I
E	Emo-DB	71	127	79	62	339	I→E, E→I

mon emotion categories are selected from source and target domains (partial domain adaptation is not considered in this work). Since the amount of data in Emo-DB corpus is small, we use the vocal tract length perturbation (VTLP) technology for data augmentation with factors ranging from 0.9 to 1.1. VLTP was proposed for speaker normalization [28], and recently it has been used for automatic speech recognition [29].

Secondly, each utterance is split into 30-ms segments using Hamming window with length of 25 ms and shift of 10 ms. Then the short time Fourier transform is used to convert the information from the time domain to the frequency domain to form a spectrogram. It is necessary to transform spectrogram to Mel-spectrum through Mel-scale filter banks when considering the auditory characteristics of human hearing.

According to [5], using deltas features can reduce the difference between speakers and improve the emotion recognition performance. We take the logarithm of the Mel spectrum information, and then perform the first-order and second-order differences to form three-dimensional data. The data is fed into the model after normalization.

3. Setup

We choose three methods for the transfer learning tasks. The cross-corpus SER without domain adaptation (denoted as N-DA) is set as the base method. The source dataset and most of the target dataset are used for training, and the other target data are used for testing. The training and testing split ratio of the target data is 8:1. SER based on G-DA and L-DA are observation tasks. Two performance metrics including unweighted average recall (UAR) and accuracy rate (Acc) are selected to measure the effectiveness of the method, which have been widely used in the field of SER. In this work, UAR is used for utterance and Acc for segment.

This part is model design and parameters description. We use ACRNN [27] as the main model structure. There are six convolutional neural network (CNN) layers, one recurrent neural network (RNN) layer, one attention layer and one classifier from the input layer to the output layer. For each CNN layer, the kernel size is (5,3) and the stride is 1. Unlike the other five CNN layers, the first CNN layer is followed by a max pooling layer with a kernel size of (2,2) and a stride of (2,4). The RNN layer is implemented as a bidirectional long short-term memory layer with 256 units. The batch size is 64 for tasks between IS and II, and 40 for tasks between I and E. The optimizer is Adam, the initial learning rate is 0.01, the exponential decay rate of the first-order moment estimation is 0.9, the exponential decay rate of the second-order moment estimation is 0.999, The tradeoff parameter λ for LMMD loss function is in the range of [0.1,0.9] with step 0.2 for grid search. In addition, we also use manual selection. The model has achieved the best recognition performance when the value of lambda is 0.01 for tasks between IS and II and is 0.1 for tasks between E and I. We use leave-one-speaker-out cross-validation and repeat three times in total. All programs are implemented in PyTorch.

4. Results and analysis

Speech emotion recognition experiment results based on different configurations are shown in Table 2 and Table 3. Table 2 shows the total accuracy in four transfer tasks which are using three different methods, N-DA, G-DA and L-DA. Table 3 shows the accuracy for each emotion category in all tasks when using method L-DA.

According to the comprehensive analysis of the experimental results, we have the following findings:

1) Cross-corpus SER based on L-DA is more effective than N-DA and G-DA.

2) The cross-corpus SER performances between II and IS are better than that between E and I. This is because that the first two corpora are much similar than that of the second one.

Table 2. Performance UAR (%) and Acc (%) of cross-corpus SER tasks under different methods

Method	Task							
	II→IS		IS→II		E→I		I→E	
	UAR (%)	Acc (%)	UAR (%)	Acc (%)	UAR (%)	Acc (%)	UAR (%)	Acc (%)
N-DA	51.75	60.71	49.68	47.02	42.15	31.40	37.33	37.21
G-DA	53.67	**62.50**	57.02	44.60	43.89	32.09	40.58	**44.19**
L-DA	**54.49**	61.01	**60.88**	**49.43**	**51.17**	**47.51**	**45.78**	39.53

Note: The results in bold indicate that these values are statistic better than the others.

Table 3. UAR (%) of four emotion categories for cross-corpus SER tasks based on L-DA

Task	UAR (%)				
	Happy	Anger	Neural	Sad	Total
II → IS	5.56	85.45	78.00	48.94	54.49
IS → II	48.28	73.33	35.43	86.49	60.88
E → I	12.77	78.57	38.33	75.00	51.17
I → E	42.86	0.00	54.55	85.71	45.78

3) The accuracy of the E→I task is lower than that of the I→E task. This is because the sample size of Emo-DB is smaller than that of IEMOCAP. The available supervision information is less than the task I→E.

V. Conclusions

In this article, we investigated how to use domain adaptation to solve the cross-corpus speech emotion recognition problem. We use local maximum mean discrepancy to measure the distribution difference between source data and target data, and build a new speech emotion recognition framework based on local domain adaptation. Compared with the basic domain adaptation method, our local domain adaptation studies more fine-grained data alignment, which makes domain adaptation achieve better performance. Our future work includes multi-source domain adaptation using local maximum mean discrepancy.

References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, " Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol.44, no.3, pp.572–587, 2011.

[2] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol.61, no.5, pp.90–99, 2018.

[3] M. S. Fahad, A. Ranjan, J. Yadav, *et al.*, " A survey of speech emotion recognition in natural environment," *Digital Signal Processing*, vol.110, article no.102951, 2021.

[4] K. X. Feng and T. Chaspari, " A review of generalizable transfer learning in automatic emotion recognition," *Frontiers in Computer Science*, vol.2, article no.9, 2020.

[5] H. J. Zhao, N. Ye, and R. C. Wang, "Speech emotion recognition based on hierarchical attributes using feature nets," *International Journal of Parallel, Emergent and Distributed Systems*, vol.35, no.3, pp.354–364, 2020.

[6] B. Zhang, E. M. Provost, and G. Essl. "Cross-corpus acoustic emotion recognition from singing and speaking: A multitask learning approach," in *Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, pp.5805—5809, 2016.

[7] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proceedings of Interspeech 2017*, Stockholm, Sweden, pp.1103—1107, 2017.

[8] S. H. Liu, M. Y. Zhang, M. Fang, *et al.*, "Speech emotion recognition based on transfer learning from the FaceNet framework," *The Journal of the Acoustical Society of America*, vol.149, no.2, pp.1338–1345, 2021.

[9] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, pp.5058—5062, 2015.

[10] J. Deng, Z. X. Zhang, F. Eyben, *et al.*, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol.21, no.9, pp.1068–1072, 2014.

[11] S. J. Pan and Q. Yang, " A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol.22, no.10, pp.1345–1359, 2010.

[12] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp.1180—1189, 2015.

[13] P. Viola and W. M. Wells ó, " Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol.24, no.2, pp.137–154, 1997.

[14] T. Van Erven and P. Harremos, " Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol.60, no.7, pp.3797–3820, 2014.

[15] K. Saito, K. Watanabe, Y. Ushiku, *et al.*, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp.3723—3732, 2018.

[16] W. W. Lin, M. W. Mak, and J. T. Chien, " Multisource I-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.26, no.12, pp.2412–2422, 2018.

[17] E. Tzeng, J. Hoffman, N. Zhang, *et al.*, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint*, arXiv: 1412.3474, 2014.

[18] M. S. Long, Y. Cao, J. M. Wang, *et al.*, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp.97–105, 2015.

[19] A. Gretton, B. Sriperumbudur, D. Sejdinovic, *et al.*, "Optimal kernel choice for large-scale two-sample tests," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe, NV, USA, pp.1205–1213, 2012.

[20] M. S. Long, H. Zhu, J. M. Wang, *et al.*, " Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning*,

Sydney, NSW, Australia, pp.2208–2217, 2017.

[21] P. Song, W. M. Zheng, S. F. Ou, *et al.*, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol.83, pp.34–41, 2016.

[22] J. T. Liu, W. M. Zheng, Y. Zong, *et al.*, "Cross-corpus speech emotion recognition based on deep domain-adaptive convolutional neural network," *IEICE Transactions on Information and Systems*, vol.E103.D, no.2, pp.459–463, 2020.

[23] H. L. Yan, Y. K. Ding, P. H. Li, *et al.*, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp.945—954, 2017.

[24] Y. C. Zhu, F. Z. Zhuang, J. D. Wang, *et al.*, "Deep subdomain adaptation network for image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol.32, no.4, pp.1713–1722, 2021.

[25] C. Busso, M. Bulut, C. C. Lee, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol.42, no.4, pp.335–359, 2008.

[26] F. Burkhardt, A. Paeschke, M. Rolfes, *et al.*, "A database of German emotional speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, pp.1517–1520, 2005.

[27] M. Y. Chen, X. J. He, J. Yang, *et al.*, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol.25, no.10, pp.1440–1444, 2018.

[28] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, vol.6, no.1, pp.49–60, 1998.

[29] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, GA, USA, 2013.

ZHAO Huijuan received the M.S. degree in computer software and theory from Nanjing University of Posts and Telecommunications. She is a teacher of Nanjing Vocational University of Industry Technology. She is currently pursuing the Ph.D. degree with the Nanjing University of Posts and Telecommunications. Her research interests include affective computing and deep learning. (Email: zhaohj86@126.com)

YE Ning received the B.S. degree in computer science from Nanjing University in 1994, the M.S. degree from the School of Computer and Engineering, Southeast University, in 2004, and the Ph.D. degree from the Institute of Computer Science, Nanjing University of Posts and Telecommunications, in 2009, where she is currently a Professor. In 2010, she was a Visiting Scholar and Research Assistant with the Department of Computer Science, University of Victoria, Canada. Her research interests include wireless networks and Internet of Things. (Email: yening@njupt.edu.cn)

WANG Ruchuan (corresponding author) researched on graphic processing with the University of Bremen and on program design theory with Ludwig Maximilian Muenchen Unitversitaet from 1984 to 1992. He has been a Professor and Supervisor of Ph.D. candidates with the Nanjing University of Posts and Telecommunications since 1992. His major research interests include wireless sensor networks and information security. (Email: wangrc@njupt.edu.cn)