LETTER

# Cross-Corpus Speech Emotion Recognition Based on Deep Domain-Adaptive Convolutional Neural Network*

**Jiateng LIU**[†], **Wenming ZHENG**[†a)], **Yuan ZONG**[†b)], **Cheng LU**[††], ***Nonmembers***, ***and*** **Chuangao TANG**[†], ***Student Member***

**SUMMARY** In this letter, we propose a novel deep domain-adaptive convolutional neural network (DDACNN) model to handle the challenging cross-corpus speech emotion recognition (SER) problem. The framework of the DDACNN model consists of two components: a feature extraction model based on a deep convolutional neural network (DCNN) and a domain-adaptive (DA) layer added in the DCNN utilizing the maximum mean discrepancy (MMD) criterion. We use labeled spectrograms from source speech corpus combined with unlabeled spectrograms from target speech corpus as the input of two classic DCNNs to extract the emotional features of speech, and train the model with a special mixed loss combined with a cross-entrophy loss and an MMD loss. Compared to other classic cross-corpus SER methods, the major advantage of the DDACNN model is that it can extract robust speech features which are time-frequency related by spectrograms and narrow the discrepancies between feature distribution of source corpus and target corpus to get better cross-corpus performance. Through several cross-corpus SER experiments, our DDACNN achieved the state-of-the-art performance on three public emotion speech corpora and is proved to handle the cross-corpus SER problem efficiently.

***key words:*** *cross-corpus speech emotion recognition, deep convolutional neural network, domain adaptation*

## 1. Introduction

Speech emotion recognition (SER) is a research hotspot in the field of affective computing in recent years, which can enhance the quality of the human-computer interaction. Recognising the emotional state of speech has a broad application prospect, such as multimedia technology, criminal investigation and computer tutorial applications [1]. Researchers have put forward a number of effective methods of SER, but most are conducted on a single database. While in practice, the speech signals between training and testing speech corpora usually have a lot of differences, e.g., the speech signals are often collected in different conditions and coming from different languages, which lead to a big gap in the feature distributions between the training and testing speech sets. Therefore, it is vital to develop a more robust system which can be more resilient to discrepancies in training and testing conditions. Furthermore, the traditional speech features used for cross-corpus SER tend to be inclined to time domain or frequency domain only, such as the zero-crossing-rate (ZCR), Mel-Frequency Cepstral Coefficents (MFCC) or the fused features of above features and so on, which often lose some emotional information due to the changes in time and frequency characteristics of speech signals [2]. It is difficult to learn emotion-related features effectively through these traditional speech emotion recognition algorithm. Thus, exploring a new method to extract robust time-frequency related features is also an important part for the cross-corpus SER problem. On the whole, finding a way that can eliminate discrepancies between domains while extracting robust features from the emotional speech signal is the key to cope with the cross-corpus SER problem.

To eliminate the discrepancies between feature distributions of different speech corpora, Song et al. [3] proposed a method called transfer non-negative matrix factorization (TNMF) which utilizes the non-negative matrix factorization and the maximum mean discrepancy (MMD) criterion for similarity measurement of the cross-corpus SER problem. Meanwhile, Zong et al. [4] proposed a method using domain-adaptive least-squares regression (DaLSR) model based on MMD criterion to handle the mismatch problem of different speech corpora. In a word, a practical way to cope with the cross-corpus SER issue is to use labeled data from training corpus (source domain) along with unlabeled data from testing corpus (target domain) based on the MMD criterion to training a model.

The research on the cross-corpus SER system not only focuses on eliminating the discrepancies between feature distributions of different speech corpora, but also explores new methods to extract robust features. For example, Sun et al. [2] proposed a deep convolutional neural network model combined with deep and shallow feature fusion of speech in different levels to improve the effectiveness of speech emotion recognition. They used the speech spectrograms as the input of the neural network, which is proved to be effective to improve the speech emotion recognition performance. In [5], Badshah et al. also extracted spectrograms

Manuscript received July 11, 2019.
Manuscript revised September 4, 2019.
Manuscript publicized November 7, 2019.

[†]The authors are with the Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing 210096, China.

[††]The author is with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China.

*This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, in part by the National Natural Science Foundation of China under Grant 61921004, Grant 61572009, Grant 61902064, and Grant 81971282, in part by the Fundamental Research Funds for the Central Universities under Grant 2242018K3DN01 and Grant 2242019K40047, and in part by the Jiangsu Provincial Key Research and Development Program under Grant BE2016616.

a) E-mail: wenming_zheng@seu.edu.cn (Corresponding author)
b) E-mail: xhzongyuan@seu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2019EDL8136

from speech signal which bypassed the traditional feature exacting method, and input them to a deep convolutional neural network for emotions recognition, which is able to predict emotions efficiently. In this letter, we also utilize the spectrograms to represent time-frequency related characteristics of speech signal as the input of DCNNs. By using appropriate feature sets, the deep model can extract more valuable information on the massive speech database.

In order to solve the discrepancies between different domains and extract robust features, we will propose a novel deep domain-adaptive convolutional neural network (DDACNN) to investigate the challenging cross-corpus SER issue in this letter. The main component of the DDACNN is a common DCNN to learn salient features from the speech spectrograms. In addition, a domain-adaptive (DA) model based on the maximum mean discrepancy (MMD) criterion is added to the DCNN model to bridge the affective gap, namely calculate the MMD loss in some fully-connected (FC) layer between the data sets from source speech corpus and an additional set of unlabeled samples from target speech corpus. In this case, both the cross-entrophy loss calculated from the labeled source data sets and the MMD loss would be optimized together to train the DDACNN model.

## 2. Proposed Method

As shown in Fig. 1, the framework of the proposed DDACNN consists of two parts, including a feature extraction model based on a DCNN and a domain-adaptive (DA) layer based on the maximum mean discrepancy (MMD) criterion.

### 2.1 Feature Extraction

The first part of the DDACNN is a common deep convolutional neural network (DCNN). A typical DCNN consists of a variety of layers that are the convolution, pooling, and fully connected layers in sequence [6]. In the last, there is a Softmax layer performing the final classification task.

Motivated by the researches based on spectrograms [2], [5], the proposed framework tries to utilize feature learning schemes for spectrograms generated from speech. Spectrogram is a visual expression of time-frequency distribution of speech energy, which can connect the time-domain with frequency-domain of speech signals. The horizontal stripes of the speech spectrogram contain rich emotional information, such as pitch frequency and spectrum envelope [7]. The abscissa and ordinate of the spectrogram stand for time and frequency, respectively, and the coordinate points reflect the energy of the speech data. As shown in Fig. 2, the energy value of speech spectrogram is represented by color. Speech signals with low amplitudes are represented by dark blue colors, while those with stronger amplitudes are represented by brighter colors up through red.

The emotional features are extracted by using the speech spectrograms based on the Short-Time Fourier Transform (STFT) as the input of DCNN to study the emotional information contained in speech signal. In order to compare the effects of different sizes of DCNN on the model, we utilize two different architectures of DCNN to extract the emotional features of spectrograms of speech signals, including a LeNet [8] model and a AlexNet [9] model. The LeNet model has 5 layers totally, including 2 convolutional layers, 2 FC layers and a Softmax layer, while the AlexNet model has 8 layers totally, including 5 convolutional layers, 2 FC layers and a Softmax layer,

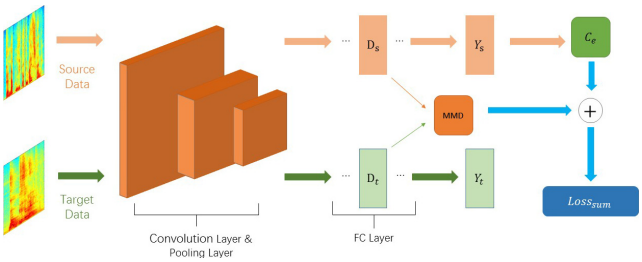
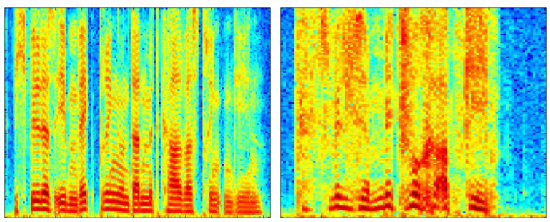



**Fig. 1** The DDACNN model



(a) (b)

**Fig. 2** Sample speech spectrograms

### 2.2 Discrepancy Measurement between Domains

To measure the discrepancies between source and target datasets, we utilize the domain-adaptive (DA) layer based on the maximum mean discrepancy (MMD) criterion [10]. While the MMD criterion is a widely used loss function of domain adaptation [3], [4], [11], [12], which has been utilized to reduce the distribution mismatch between the source and target domains. Its main purpose is to measure the discrepancies between different distributions of the source and target domains, then we can incorporate the MMD into the learning algorithm.

Assuming that $\mathbf{D}^s = [d_1^s, \ldots, d_M^s]$ and $\mathbf{D}^t = [d_1^t, \ldots, d_N^t]$ are two sample sets from distributions $P(D^s)$ and $P(D^t)$, the difference measurement can be formulated as:

$$MMD[P, Q] \triangleq \sup_{f \in \mathcal{H}} (E_{D^s}[f(\mathbf{D}^s)] - E_{D^t}[f(\mathbf{D}^t)]) \quad (1)$$

where $\mathcal{H}$ is a class of functions. If $\mathcal{H}$ is rich enough to distinguish any two distributions in an Reproducing Kernel Hilbert Space (RKHS) [13], we can express the MMD as the following forms of the distance between mean embeddings:

$$MMD[P, Q] = \|\mu_{\mathbf{D}^s}(P) - \mu_{\mathbf{D}^t}(Q)\|_{\mathcal{H}}^2 \quad (2)$$

Through comparing the square distance between the empirical kernel mean embeddings, the estimate of the MMD can

be denoted as:

$$MMD[P,Q] = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}k(D_i^s, D_j^s) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}k(D_i^t, D_j^t) - \frac{2}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}k(D_i^s, D_j^t) \quad (3)$$

where k is the kernel function. We use the gaussian function as the kernel function so as to map data onto an infinite dimensional space.

The basic method of MMD is that the statistics will be the same when the generating distributions of source and target domains are identical [11]. We expect that the DA layer based on MMD regularization embedded in training process can induce better discriminative features (the experimental results are detailed in Sect. 3.2). To be specific, as shown in Fig. 1, $\mathbf{D}_s$ and $\mathbf{D}_t$ are outputs of a batch of speech signals from source and target corpora in a FC layer of the DCNN, then the MMD is calculated as a loss function by applying Eq. (3) to the outputs, which is a part of the total loss function for optimizing during the training process. By incorporating the MMD calculated with Eq. (3) into training process for optimizing, the total loss function will go down and the discrepancies between source and target corpora can be narrowed. This is the detailed description of the DA layer.

### 2.3 DDACNN for the Cross-Corpus SER

As mentioned above, the LeNet and the AlexNet are used as the DCNN of the DDACNN, then the MMD loss is calculated in some FC layer of the DCNN.

The spectrogram images are generated from three different emotion speech corpora, including EmoDB [14], eNTERFACE [15] and CASIA [16], which were setup in different languages. We use the spectrogram of one corpus as the source set and the spectrogram of another one as the target set, then the input images were resized to $224 \times 224$ before being fed to the neural network.

The MMD loss function is calculated in some FC layer of the DCNN between the data set from source speech corpus and an additional data set of unlabeled testing samples from target speech corpus to train the DCNN model. For each spectrogram generated for the incoming speech signals, we utilize the trained model to get the final recognition accuracies.

By integrating the MMD over the domain-specific FC layers into the DCNN cross-entropy loss function, the final end-to-end objective function of DDACNN with neural network training can be obtained, which can be defined as:

$$Loss_{sum} = C_e(Y_s, y_s) + \lambda MMD_{FC}(D_s, D_t) \quad (4)$$

Where $C_e$ stands for the cross-entropy loss function of source set, $\lambda > 0$ control the tradeoff among two terms, $y_s$ is the label of source speech data, while $MMD_{FC}$ is the MMD loss function calculated in some FC layers of the DCNN.

## 3. Experiments

### 3.1 Databases and Protocols

In order to evaluate the performance of the DDACNN model, we conducted 6 experiments using three corpora from three different languages to extract their spectrograms, namely EmoDB, eNTERFACE, and CASIA. Since each emotion corpus contains different emotions, the samples containing common emotions were selected as the data onto experiment. For the experiments between EmoDB and eNTERFACE, five emotions (angry, disgust, fear, happy and sad) were chosen for validation, and 375 and 1072 audio samples were selected, respectively. While in the experiments between EmoDB and CASIA, five emotions (angry, fear, happy, disgust, sad) were chosen, which were contained in 408 and 1000 speech samples for validation. In addition, 1072 and 1000 audio samples containing five emotions (angry, fear, happy, disgust, sad) were selected for validation in the experiments between eNTERFACE and CASIA, respectively.

In addition, two accuracy measurements were utilized to evaluate the recognition performance, including the unweighted average recall (UAR) and the weighted average recall (WAR). The UAR means the ratio of the predicted accuracy per class to the whole number of classes, while the WAR stands for the ratio of correctly predicted samples to the total samples. Several state-of-the-art methods using traditional speech features were utilized for comparison purpose, including transfer component analysis (TCA) [17], transfer kernel learning (TKL) [18] and DaLSR [4]. The feature set used by these methods is provided by the INTERSPEECH 2009 Emotion Challenge (IS09) [19]. The IS09 feature set has 384 dimensions of features which contain 32 acoustic low-level descriptors (LLDs) such as Mel-frequency cepstral coefficient (MFCC) and zero-crossing-rate (ZCR), and they are extracted by the openSMILE software [20]. Meanwhile, we also utilized the original LeNet and AlexNet without a DA layer to perform the experiments. In addition, the linear SVM using the IS09 feature sets is selected as the baseline method of all experiments.

### 3.2 Results and Analysis

The final experimental results are shown on the Table 1. The L-FC1 and L-FC2 denote that the DDACNN based on LeNet with a DA layer added in its first and second FC layer, while the A-FC1 and A-FC2 mean the same situation to LeNet in AlexNet network.

According to the results, the proposed DDACNN based on LeNet with a DA layer added in its first FC layer, namely L-FC1 achieves the best results from 5 cases of all 6 experiments, and the DaLSR method achieves the best results

**Table 1** Enperimental results (UAR/WAR) for cross-corpus SER

| Exp | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Source/Target Corpus | E/B | B/E | B/C | C/B | E/C | C/E |
| SVM | 27.83/24.27 | 30.06/30.08 | 25.10/25.10 | 33.59/36.76 | 24.20/24.20 | 26.26/26.31 |
| TCA [17] | 29.70/39.20 | 26.25/26.28 | 28.10/28.10 | 37.24/37.99 | 26.10/26.10 | 24.20/24.25 |
| TKL [18] | 36.21/42.40 | 24.55/24.61 | 25.10/25.10 | 35.08/38.24 | 24.10/24.10 | 27.48/27.52 |
| DaLSR [4] | 44.41/52.27 | **36.36/36.40** | 25.40/25.40 | 23.85/26.96 | 20.00/20.00 | 22.65/24.21 |
| LeNet | 43.82/49.60 | 30.46/30.48 | 34.50/34.50 | 31.36/38.54 | 28.10/28.10 | 29.86/29.90 |
| AlexNet | 32.28/34.93 | 28.86/28.87 | 31.30/31.30 | 30.75/30.12 | 27.20/27.20 | 26.86/27.01 |
| L-FC1 | **49.93/58.13** | 34.51/34.52 | **38.10/38.10** | **46.62/48.39** | **31.90/31.90** | **31.59/31.68** |
| L-FC2 | 44.21/54.13 | 31.80/31.77 | 34.90/34.90 | 35.32/38.02 | 27.50/27.50 | 30.94/31.02 |
| A-FC1 | 35.34/45.56 | 30.29/30.28 | 33.10/33.10 | 31.01/32.09 | 28.70/28.70 | 29.53/29.53 |
| A-FC2 | 30.30/41.33 | 29.77/29.62 | 31.70/31.70 | 29.56/30.56 | 25.80/25.80 | 29.49/29.62 |

E, B and C denote the eNTERFACE, EmoDB, and CASIA corpus, respectively.

from the Exp2. Meanwhile, the results of the experimental group using LeNet are better than those using AlexNet, which show that the LeNet model with smaller structure is more effective in cross-corpus SER issue. We can find that the recognition rate of DCNN using DA layer is improved compared with these without one, which indicates that the domain adaptive method is very effective to improve the recognition accuracies of cross-corpus SER as it can eliminate the discrepancy between feature distributions of different corpora successfully. Especially in Exp1, the L-FC1 obtains 49.93%/58.13% recogniton accuracies in UAR/WAR, which is 5.52%/5.86% higher than the state-of-the-art DaLSR model. While in Exp3 and Exp4, the L-FC1 obtains 38.10%/38.10% and 46.62%/48.39% recogniton accuracies in UAR/WAR, which are 10.00%/10.00% and 9.38%/10.40% higher than the classic TCA method. In addition, the DaLSR method gets the best recognition accuracy in Exp2, which may attribute to the less contribution of DDACNN model to narrow the feature discrepancies in this case. Furthermore, the LeNet using spectrograms as input without a DA layer also achieves better performance than the SVM method using the IS09 feature sets, which shows that the way of feature extraction in DDACNN is useful for the cross-corpus SER issue.

On the basis of these results, it is apparent that the DDACNN combining LeNet with a DA layer added in its first FC layer can deal with the cross-corpus SER problems effectively in most cases than other traditional domain adaptive ways using common speech features or the common neural network methods without a DA layer.

## 4. Conclusion

In this letter, we have proposed a novel DDACNN model to handle the challenging cross-corpus SER problem. The main contribution of the proposed DDACNN is that it can eliminate the feature distribution discrepancies between the training and testing speech corpora while extracting robust features from speech signals at the same time. Through 6 extensive experiments on 3 different speech emotion corpora, the proposed DDACNN model based on LeNet with a DA layer added in its first FC layer has achieved better recognition performance than other traditional methods in most cases. Additionally, there are still many problems to cope with with the cross-corpus SER problem, such as the serious imbalance of sample sizes in different corpus. In the future, we can also pay attention to more DCNN models to enrich the DDACNN model.

### References

[1] M.E. 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] L. Sun, J. Chen, K. Xie, and T. Gu, "Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition," International Journal of Speech Technology, vol.21, no.4, pp.931–940, 2018.
[3] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," Speech Communication, vol.83, pp.34–41, 2016.
[4] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," IEEE Signal Process. Lett., vol.23, no.5, pp.585–589, 2016.
[5] A.M. Badshah, J. Ahmad, N. Rahim, and S.W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," 2017 international conference on platform technology and service (PlatCon), pp.1–5, IEEE, 2017.
[6] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "Mped: A multi-modal physiological emotion database for discrete emotion recognition," IEEE Access, vol.7, pp.12177–12191, 2019.
[7] R.V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," Science, vol.270, no.5234, pp.303–304, 1995.
[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol.86, no.11, pp.2278–2324, 1998.
[9] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," Commun. ACM, vol.60, no.6, pp.84–90, 2017.
[10] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A.J. Smola, "A kernel method for the two-sample-problem," Advances in Neural Information Processing Systems, pp.513–520, 2007.
[11] M. Long, H. Zhu, J. Wang, and M.I. Jordan, "Deep transfer learning with joint adaptation networks," Proceedings of the 34th International Conference on Machine Learning, pp.2208–2217, 2017.
[12] K. Yan, W. Zheng, T. Zhang, Y. Zong, C. Tang, C. Lu, and Z. Cui, "Cross-domain facial expression recognition based on transductive deep transfer learning," IEEE Access, vol.7, pp.108906–108915, 2019.
[13] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A hilbert space

embedding for distributions," International Conference on Algorithmic Learning Theory, vol.4754, pp.13–31, Springer, 2007.
[14] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss, "A database of german emotional speech," Ninth European Conference on Speech Communication and Technology, 2005.
[15] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," 22nd International Conference on Data Engineering Workshops (ICDEW'06), p.8, IEEE, 2006.
[16] J. Tao, F. Liu, M. Zhang, and H. Jia, "Design of speech corpus for mandarin text to speech," The Blizzard Challenge 2008 Workshop, 2008.
[17] S.J. Pan, I.W. Tsang, J.T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," IEEE Trans. Neural Netw., vol.22, no.2, pp.199–210, 2010.
[18] M. Long, J. Wang, J. Sun, and P.S. Yu, "Domain invariant transfer kernel learning," IEEE Trans. Knowl. Data Eng., vol.27, no.6, pp.1519–1532, 2015.
[19] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," Tenth Annual Conference of the International Speech Communication Association, 2009.
[20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," Proceedings of the 18th ACM international conference on Multimedia, pp.1459–1462, ACM, 2010.