Speaker to Emotion: Domain Adaptation for Speech Emotion Recognition with Residual Adapters

Yuxuan Xi, Pengcheng Li, Yan Song, Yiheng Jiang, Lirong Dai
National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China
E-mail: {xyxah96, pclee, jiangyh}@mail.ustc.edu.cn, {songy, lrdai}@ustc.edu.cn

***Abstract*—Despite considerable recent progress in deep learning methods for speech emotion recognition (SER), performance is severely restricted by the lack of large-scale labeled speech emotion corpora. For instance, it is difficult to employ complex neural network architectures such as ResNet, which accompanied by large-sale corpora like VoxCeleb and NIST SRE, have proven to perform well for the related speaker verification (SV) task. In this paper, a novel domain adaptation method is proposed for the speech emotion recognition (SER) task, which aims to transfer related information from a speaker corpus to an emotion corpus. Specifically, a residual adapter architecture is designed for the SER task where ResNet acts as a universal model for general information extraction. An adapter module then trains limited additional parameters to focus on modeling deviation for the specific SER task. To evaluate the effectiveness of the proposed method, we conduct extensive evaluations on benchmark IEMOCAP and CHEAVD 2.0 corpora. Results show significant improvement, with overall results in each task outperforming or matching state-of-the-art methods.**

I. Introduction

Speech Emotion Recognition (SER) aims to automatically analyze emotional categories from speech utterances. Over recent years, SER has drawn increasing attention in line with the rapid growth in demand of applications such as telephone call centres, educational systems and intelligent robotics. Recently, deep learning based systems have achieved significant progress for SER, but to be successful, sufficient labeled data is needed, particularly due to the complexity of emotional information. However, existing corpora, such as IEMOCAP [15], CHEAVD [14], FAU-AIBO [29], and EMO-DB [30], are generally size-limited, in part due to annotation cost, and also suffer label ambiguity.

One possible solution is to utilize emotion information from multiple corpora. Based on this approach, several transfer learning and multi-task learning (MTL) based methods have been proposed [1], [2], [3], [4]. Transfer learning focuses on adapting knowledge from available auxiliary resources to the target domain. For example, *Latif et al.* exploited transfer learning across several corpora via a Deep Belief Network (DBN) model [1]. *Song et al.* [2] proposed a joint transfer subspace learning and feature selection (JTSLFS) algorithm. MTL treats different training and evaluation corpora as multiple target tasks. *Zhang et al.* [3], [4] experimented with several multi-task learning methods including single-task (ST), multi-task feature selection/learning (MTFS/MTFL), group multi-task feature selection/learning (GMTFS/GMTFL) on sung emotion recognition task and SER.

Existing research mainly focuses on cross-corpus methods in the scope of emotion data. However, due to the limited size of emotion corpora, SER performance is far from satisfactory and it is still difficult to apply successful deep learning architectures like ResNet and DenseNet to further improve performance. In [28], it was shown that the annotation of emotion can be transferred from the visual domain (faces) to speech domain (voices) through cross-domain distillation on the VoxCeleb dataset [13]. In [3], [4], [16], [17], the authors investigated using speaker characteristics like gender and age. These two areas of research strongly suggest a relationship between emotion and speaker characteristics.

Based on this view, we propose a domain adaptive model which can utilize a common representation between emotion and speaker identity to further improve SER accuracy, using ResNet as a backbone architecture. Specifically, the proposed method aims to tackle the lack of labeled corpus by employing a residual adapter model [12] to transfer the information from VoxCeleb to a specific SER target dataset. The residual adapter resembles ResNet [10], with the major difference that all convolutional layers are replaced by adapter modules. Each module contains two types of parameter; one type are domain-agnostic, and act as a universal model for general information extraction, the other are domain-specific parameters used for adaptation. The domain-agnostic parameters are trained on the initial task, then fixed to reduce the model complexity during domain adaptation on the target task. In this paper, the residual adapter model is trained using VoxCeleb2 data with speaker labels, then emotion corpora are used to train the domain-specific parameters, and different fully-connected layers are used to predict the classification score, as shown in Fig.1.

There are also some semi-supervised learning based methods, prompted by the wide availability of unlabeled speech data. These commonly use auto-encoders to reap benefits from a combination of labeled and unlabeled data [5], [6], [7]. In [5], a self-contained semi-supervised auto-encoder (SSAE) framework was proposed, which integrates a supervised path and an unsupervised auto-encoder. In [11], *Parthasarathy et al.* presented ladder network based semi-supervised learning, which can outperform the auto-encoder based methods. Other methods include domain adaptive least squares regres-

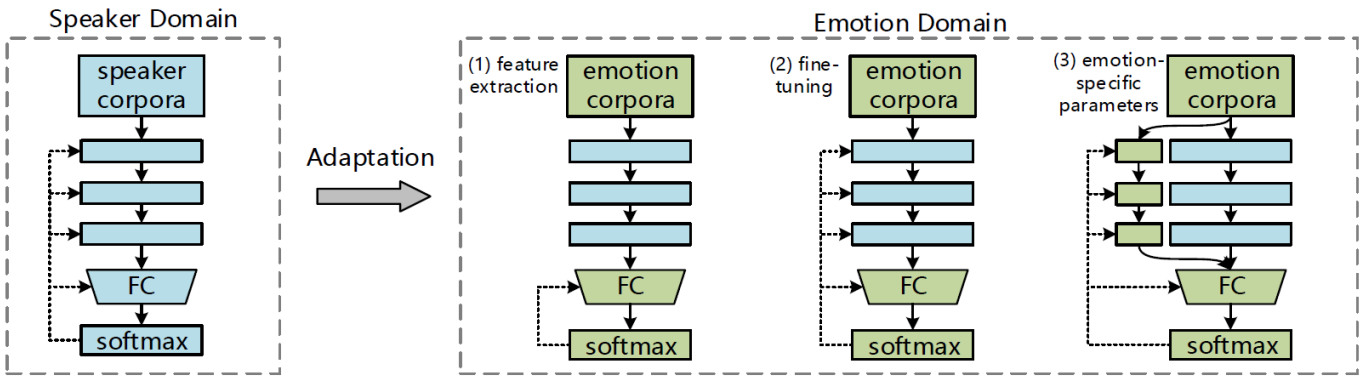

Fig. 1. illustration of speaker to emotion domain adaption framework. There are three different adaptation methods to utilize the pre-trained speaker network, including feature extractor, fine-tuning and domain adaptation, the dashed line indicates network training. "FC" means fully connected layers.

sion (DaLSR) [8] and domain-adaptive subspace learning (DoSL) [9]. The main difference lies in that the proposed residual adapter utilizes supervised learning to exploit the relationship between speaker and emotion data.

To prove the effectiveness of our method, we first use ResNet that is trained with emotion data only as a baseline system, and then conduct a series of experiments as shown in Fig.1, including: (1) A ResNet trained by VoxCeleb2 data as the feature extractor, then the classifier trained for SER. (2) The same ResNet fine-tuned with emotion data – the common practice in transfer learning. (3) The proposed residual adapter method, furthermore testing the adapter module alone, aiming to demonstrate that features learned from the speaker classification task can be beneficial to SER. The results of (3) on IEMOCAP improvised and CHEAVD both significantly exceed baseline systems, establishing the effectiveness of the proposed method.

II. OVERVIEW OF SPEAKER-TO-EMOTION DOMAIN ADAPTATION FRAMEWORK

SER encompasses some existing problems. (1) Deep learning based methods have become prevalent in recent years [20], [21], [22], owing to the powerful representation learning ability of neural networks. In general, increasing network depth benefits performance, but the limited scale of emotion corpora greatly restricts the network complexity in practice. In previous studies, most deep learning models contain only a few layers and need to be specifically designed for emotion corpora. Some powerful deep learning models, such as ResNet, which achieve great success in related tasks, cannot effectively be utilized due to the limited training data. (2) Existing methods mainly focus on cross-corpus learning among emotion corpora, but due to the difficulty and cost of labeling emotion data, cross-corpus methods still have limitations.

By contrast, labeled data is abundant for traditional speech tasks. Although the corpora from other speech domains lack emotion labels, they may still contain some related information which can assist in SER model training. Speaker-labeled corpora are potential choices, as described in Section I, where speaker characteristics such as age and gender can influence SER results. This fact indicates that there is some shared representation between speaker characteristics and emotion. On the other hand, the scale of speaker corpora are much larger than those for emotion, a fact that aids in training a deep neural network. Based on those motivations, the VoxCeleb2 [13] corpus is selected in this paper for initial model training.

In order to utilize speaker labels, a complex network is first trained by speaker corpora, and then adapted to the target emotion corpora. We therefore explore three possible adaptation methods, as shown in Fig.1. The first method is a feature extractor, which constructs an emotion classifier by retraining the topmost FC layer. The second method is fine-tuning, which takes the same structure as the feature extractor, then all network parameters as well as the FC classifier, are fine-tuned using the emotion corpora. Although these two methods can exploit the information from both speaker and emotion data, they have some obvious problems. Firstly, the network parameters are pre-trained for speaker verification, which may be quite different from SER, therefore directly utilizing the model may not be a appropriate choice. Secondly, due to the limited scale of the target emotion corpora, training the whole network may be difficult and may cause an over-fitting problem. On the other hand, the fine-tuning stage may result in forgetting the source corpora, reducing the benefit of the auxiliary information.

To address these issues, this paper attempts to establish the third method, a new domain adaptation. In this method, the deep learning model is first trained using VoxCeleb2 data with speaker labels as usual. But during the adaptation stage, some extra emotion-specific parameters are added to the original model, then the emotion corpora are utilized to only fine-tune the additional parameters which coexist alongside the previously trained parameters. Through the proposed framework, the information forgetting problem is avoided, and because the emotion corpora is only utilized to fine-tune a part of the network, the over-fitting problem may be mitigated. Based on our motivation, we exploit the residual adapter model, as demonstrated in the following section.

III. RESIDUAL ADAPTER MODEL

A. Model design

The motivation for a residual adapter model is to dynamically fit multiple domains. The idea can be accomplished by incorporating domain-specific parameters into the deep learning model. However, incorporating too many additional parameters for each domain may cause over-fitting, therefore the residual adapter model aims to reduce the number of domain-specific parameters while sharing more domain-agnostic parameters. The basic idea of constructing an adapter module is to linearly parameterize the convolutional filter group, which is the same as introducing an intermediate convolutional layer. Assuming there is a convolutional filter group with size $H \times W$, applied on T input feature maps, written as $F \in \mathbb{R}^{H \times W \times T}$. The K filter bank added with introduced parameters can be written as a linear combination $G = \sum_{k=1}^{K} (\alpha_{:k} + F_k)$, where $\alpha \in \mathbb{R}^{T \times K}$ represents additional parameters. By introducing these additional parameters, the convolutional layer can be seen as the combination of domain-specific parameters α and domain-agnostic parameters F. Applying the filter group to input x, we can obtain

$$G * x = \sum_{k=1}^{K} (\alpha_{:k} + F_k) * x = \alpha * x + F * x \tag{1}$$

in implementation α is a $1 \times 1 \times T \times K$ filter bank. For the training progress, firstly the model is trained on the initial task, using a large corpus. Next, the parameters are fixed and other domain adapters are trained using the target domain corpus.

B. Adapter module and network structure

The residual adapter module modifies the basic residual block in ResNet. The original residual block contains two 3×3 convolutional layers, a shortcut connection is linked between the input and output, which can be denoted as in the following equation:

$$ReLU(x + \omega_2 * ReLU(\omega_1 * x)) \tag{2}$$

where ω_1 and ω_2 are parameters of the convolutional layers, the activation function is ReLU. In order to make the network able to dynamically fit multiple domains, for each convolutional layer, a bank of 1×1 filters is applied on the input tensor, then the output is added to the original convolutional output, the output dimension of 1×1 filters is maintained unchanged. The modified convolutional layers can be written as in the following equation:

$$f(x) = \omega * x + \alpha * x \tag{3}$$

Since a bank of 1×1 filters has far fewer parameters, the complexity of the domain-specific part is reduced, which further prevents the over-fitting problem. The adapter module in this paper is shown in Fig. 2.

Batch normalization (BN) [19] plays an important role in deep neural network training, and BN layers can re-scale the feature distribution. In the adapter module, BN is applied after the feature sum for each domain, and because BN has learnable parameters, they also provide some domain-specific parameter adaptation. In this paper, the residual adapter model is constructed on the ResNet20 model with a network structure outlined in Table I. Three basic blocks are stacked separately for each stage, thus there are 20 convolutional layers counting the first 7×7 layer and last fully-connected (FC) layer. The stride of the last convolutional layer of each stage is set to 2 and the output dimension is doubled. Finally the network outputs a 256-dimensional feature map with size 25×13, followed by global average pooling (GAP) to downsample the feature size to 1×1. FC and softmax layers are then utilized to predict the final emotion label, and a cross-entropy loss function is used.

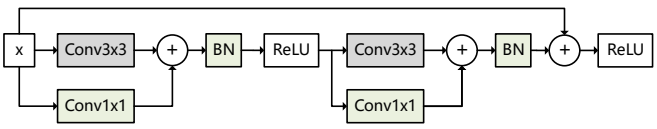

Fig. 2. Adapter module where the 3×3 filters are domain-agnostic. The 1×1 filters and BN layers are domain-specific.

TABLE I
THE RESNET20 NETWORK STRUCTURE USED IN OUR METHOD

Layer name	Output	Parameter
Conv1	400×200	7×7, 32, stride = 1
Max Pooling	200×100	3×3, stride = 2
Stage1	100×50	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$
Stage2	50×25	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$
Stage3	25×13	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 3$
Average Pooling	1×1	25×13
Fully Connected	256:classes	

IV. EXPERIMENTS AND ANALYSIS

A. Data description and pre-processing

In this study, both speaker data and emotion data are utilized. For speaker-labeled data, we choose the VoxCeleb2 corpus [13]. VoxCeleb2 is a large-scale speaker-labeled database, prevalent for SV tasks, that was collected from more than 6000 celebrities on YouTube. VoxCeleb2 consists of 2442 hours, with more than a million speech utterances, covering different ages, genders, accents and scenes. Due to the diversity of the data source, VoxCeleb2 data is likely to contain rich emotion information, but in this paper we only utilize the speaker labels.

For the emotion part, we select Interactive Emotional Dyadic Motion Capture (IEMOCAP) [15] and Chinese Natural Audio-Visual Emotion Database (CHEAVD) [14] 2.0 databases. IEMOCAP is performed by 10 skilled actors and divided into 5 sections where each section contains two actors. IEMOCAP has scripted and improvised parts, depending on the recording scenarios. We choose the improvised data part

TABLE II
THE WEIGHTED ACCURACY (WA) AND UNWEIGHTED ACCURACY (UA) OF ALL SYSTEMS (IN %)

Methods	CHEAVD		IEMOCAP	
	UA	WA	UA	WA
Plain	32.27(±3.60)	42.66(±1.33)	59.00(±2.53)	65.65(±1.85)
Feature Extractor	32.91(±2.90)	40.39(±0.25)	57.85(±1.25)	66.52(±0.25)
Fine-tuning	31.23(±3.59)	41.45(±0.91)	62.83(±1.59)	68.02(±0.53)
Res. Adapter	**34.08(±1.84)**	**43.96(±0.54)**	**67.58(±2.41)**	**72.73(±0.58)**
Adapter	32.10(±1.25)	41.54(±0.76)	57.62(±3.27)	68.89(±0.36)

in order to exclude undesired contextual information. Labels of neutral, angry, happy and sad are used.

CHEAVD 2.0 is a Chinese emotion corpus, the official data of the Multimodal Emotion Recognition Challenge (MEC) 2017. CHEAVD contains data selected from Chinese movies, soap operas and TV shows. It contains 8 emotion labels (angry, happy, sad, worried, anxious, surprise, disgust, neutral). The corpus is divided into training, validation and testing sets. We use the training/validation split for performance evaluation, the hyper-parameter tuning is based on validation set, keeping the evaluation that same as in [25].

Magnitude spectrograms are utilized as input features, with the spectrograms extracted over 40 ms Hamming windows with a 10 ms window shift and 1600 FFT points. Then 0-4000 Hz spectrogram are utilized since human vocal expression is mainly located in this frequency range. The speech utterances are cut into 2 s portions with 1 s overlap, and zero-padding applied for utterances shorter than 2 s. Thus the input spectrograms have a size of 400×200. For each spectrogram, we then apply a μ-law expansion, as used and described in our previous paper [18].

B. Experiment setup

For VoxCeleb2 data, we randomly choose 50 speakers to train the ResNet20 with adapters. For IEMOCAP improvised data, we conduct a 5-fold cross-validation, where 4 sections are used to train the network and the remaining 2 speakers are used as validation and test data. For network training, we make use of the PyTorch deep learning framework with SGD and Nesterov momentum update utilized, starting at 0.9. We train the network over 30 epochs for each dataset. For VoxCeleb2, the initial learning rate is set to 0.05, then divided by 10 at the 21^{st} epoch, with a weight decay of 0.0001. For emotion corpora, the initial learning rate is 0.01, divided by 10 at the 11^{th} and 21^{st} epochs. As suggested in [12], we use a large weight decay for emotion corpora, set to 0.001 and 0.005 for CHEAVD and IEMOCAP respectively. The loss function is cross-entropy loss and all experiments are run 5 times with the results averaged.

C. Results and analysis

All experimental results are listed in Table II in terms of weighted and unweighted accuracy (WA and UA respectively).

Baseline: We use IEMOCAP and CHEAVD to train a plain ResNet20 with results in the top row of Table II. Obviously, emotion data is insufficient to train a ResNet, so UA, WA are unsatisfactory for IEMOCAP and CHEAVD, in line with our expectations.

Fine-tuning: Using a large corpus to train a deep network, then using a small corpus to fine-tune is the common practice of transfer learning. We use VoxCeleb2 data to pre-train a plain ResNet20 then, after training, the FC layer of the network is replaced and the whole network fine-tuned by the target emotion corpus. The result is not significantly better than baseline, likely to be because the number of parameters is too large for the smaller extent of emotion data to train. On the other hand, forgetting the learned speaker information may be another problem which would reduce the accuracy.

Feature extractor: When fixing the parameters learned by the primary domain, the network becomes a feature extractor. In this experiment we fix all ResNet20 parameters and train the FC layer with emotion corpora. The performance is worse than the fine-tuning method, this indicates utilizing only speaker information is not appropriate for SER.

Residual Adapter: We next evaluate the residual adapter model. We use VoxCeleb2 data to train the same ResNet20 with adapter modules. During the adapting process, all of the parameters of the 3×3 filters are fixed, then the adapters are trained using emotion data. The result significantly outperforms the baseline system, especially for IEMOCAP, where the UA and WA achieve 67.58% and 72.73%. On CHEAVD they achieve 34.08% and 43.96%. We attempted to increase the number of speakers during residual adapter training, but the performance did not benefit from this, perhaps because a more complex model is needed.

Evaluation of adapters: Finally, we want to clarify if the improvement in SER performance has benefited from domain-agnostic parameters learned by VoxCeleb2, or simply because adapters have fewer parameters so the model can be trained by emotion corpora. To answer the question, we keep the same experiment configuration with the residual adapters, retain the model but set all 3×3 convolutional filter weights to 0, so the domain-agnostic parameters will offer no information. As a result, the accuracy significantly drops, which proves the necessity of domain-agnostic parameters.

TABLE III
COMPARISON TO EXISTING MODELS ON IEMOCAP AND CHEAVD (IN %)

Corpus	Model	UA	WA
IEMOCAP	LSTM-ELM [23]	63.89	62.85
	CNN-LSTM [21]	62.00	67.30
	CNN-GRU [24]	64.22	71.45
	CNN-Att.pooling [18]	66.38	70.18
	Our model	**67.58**	**72.73**
CHEAVD	MEC 2017 baseline [26]	27.20	39.90
	LSTM-FCN [25]	31.80	**46.30**
	Our model	**34.08**	43.96

D. Comparison to state-of-the-art systems

We compare our model to some existing published results. The reported models include LSTM-ELM [23], CNNLSTM [21], [25], CNN-GRU [24], we also evaluate the CNN-Att.pooling model in our previous paper [18], their results are listed in Table III. For IEMOCAP, one author [27] reported higher performance, but their model utilized phoneme information, which is different from the above methods. Compared to [25], our method does not exceed their WA but has better UA, indicating that the performance of data-limited small classes is improved. In fact these results show that the proposed residual adapter model can effectively utilize speaker characteristic information from the VoxCeleb2 training data, yet also provide discrimination ability for the SER task. In future we believe there is potential to exploit a deeper network for SER to further improve performance.

V. CONCLUSIONS

This paper has proposed a novel domain adaptation method to transfer the related information from a speaker corpus to an emotion corpus. This method can effectively address the lack large-scale labeled emotion training data by exploiting universal information from VoxCeleb2. Specifically, a residual adapter architecture is designed, in which ResNet20 acts as a universal model for general information extraction, and the adapter module uses limited parameters which focus on modeling the deviation for specific SER tasks. Evaluations on benchmark IEMOCAP and CHEAVD2.0 tasks demonstrate the effectiveness of the proposed domain adaptation method, with overall results outperforming or matching state-of-the-art methods. Furthermore, this method demonstrates the general potential for utilizing speech related large-scale data to improve SER performance.

ACKNOWLEDGMENT

The work was supported by the National Natural Science Foundation of China grant no U1613211.

REFERENCES

[1] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer Learning for Improving Speech Emotion Classification Accuracy" in *Proc. of INTERSPEECH 2018,* pp. 257-261.

[2] P. Song, W. Zheng, S. Ou, Y. Jin, W. Ma, and Y. Yu, "Joint transfer subspace learning and feature selection for cross-corpus speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2018, pp. 5504-5508.

[3] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING,* vol. 10, pp. 85-99, 2019.

[4] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2016, pp.5805-5809. pp. 257-261.

[5] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE SIGNAL PROCESSING LETTERS,* vol. 21, pp. 1068-1072, 2014.

[6] J. Deng, X. Xu, Z. Zhang, and S. Fruhholz, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING,* vol. 26, pp. 31-43, 2018.

[7] J. Deng, X. Xu, Z. Zhang, and S. Fruhholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE SIGNAL PROCESSING LETTERS,* vol. 24, pp. 500-504, 2017.

[8] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive leastsquares regression," *IEEE SIGNAL PROCESSING LETTERS,* vol. 23, pp. 585-589, 2016.

[9] N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, and J. Zhu, "Unsupervised cross-corpus speech emotion recognition using domainadaptive subspace learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2018, pp. 5144-5148.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016, pp. 770-778.

[11] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Proc. of INTERSPEECH,* 2018, pp. 3698-3702.

[12] S.-A. Rebuff, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *Conference on Neural Information Processing Systems (NIPS)* 2017, pp. 506-516.

[13] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. of INTERSPEECH,* 2018, pp. 1086-1090.

[14] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: a chinese natural emotional audio-visual database," *Journal of Ambient Intelligence and Humanized Computing,* vol. 8, no. 6, pp. 913-924, 2017.

[15] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation,* vol. 42, no. 4, p. 335, 2008.

[16] R. Xia, J. Deng, B. Schuller, and Y. Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2014, pp. 990-994.

[17] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2017, pp. 5150-5154.

[18] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proc. of INTERSPEECH,* 2018.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167,* 2015.

[20] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition using Deep Neural Network and Extreme Learning Machine," in *Proc. of INTERSPEECH,* 2014, pp. 223-227.

[21] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. of INTERSPEECH,* 2017, pp. 1089-1093.

[22] Z. Aldeneh and E. M. Provost, "Using Regional Saliency for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2017, pp. 2741-2745.

[23] J. Lee and I. Tashev, "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition," in *Proc. of INTERSPEECH,* 2015, pp. 1537-1540.

[24] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms," in *Proc. of INTERSPEECH,* 2018, pp. 3683-3687.
[25] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition," in *Proc. of INTERSPEECH,* 2018, pp. 272-276.
[26] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia, "Mec 2017: Multimodal emotion recognition challenges," in *Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia),* 2018.
[27] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition," in *Proc. of INTERSPEECH,* 2018, pp. 3688-3692.
[28] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *ACM Multimedia Conference,* 2018, pp. 292-301.
[29] Schuller, Steidl, and Batliner, "The interspeech 2009 emotion challenge," in *Proc. of INTERSPEECH, 2009,* 2009, pp. 312-315.
[30] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. of INTERSPEECH,* 2005, pp. 1517-1520.