FRONTEND ATTRIBUTES DISENTANGLEMENT FOR SPEECH EMOTION RECOGNITION

Yu-Xuan Xi[1], *Yan Song*[1], *Li-Rong Dai*[1], *Ian McLoughlin*[1,2], *Lin Liu*[3]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China.
[2]ICT Cluster, Singapore Institute of Technology, Singapore.
[3]iFLYTEK Research, iFLYTEK CO., LTD, Hefei, China.

ABSTRACT

Speech emotion recognition (SER) with limited size dataset is a challenging task, since a spoken utterance contains various disturbing attributes besides emotion, including speaker, content, and language. However, due to a close relationship between speaker and emotion attributes, simply fine-tuning a linear model is enough to obtain a good SER performance on the utterance-level embeddings (*i.e.*, i-vector and x-vectors) extracted from the pre-trained speaker recognition (SR) frontends. In this paper, we aim to perform frontend attributes disentanglement (AD) for SER task, using a pre-trained SR model. Specifically, the AD module consists of attribute normalization (AN) and attribute reconstruction (AR) phases. The AN filters out the variation information using instance normalization (IN), and AR reconstructs the emotion-relevant features from the residual space to ensure high emotion discrimination. For better disentanglement, a dual space loss is then designed to encourage the separability of emotion-relevant and emotion-irrelevant spaces. To introduce the long-range contextual information for emotion related reconstruction, a time-frequency (TF) attention is further proposed. Different from the style disentanglement of the extracted x-vectors, the proposed AD module can be applied on frontend feature extractor. Experiments on IEMOCAP benchmark demonstrate the effectiveness of the proposed method.

Index Terms— speech emotion recognition, convolutional neural network, style transformation, disentanglement

1. INTRODUCTION

Speech emotion recognition (SER) is the automatic identification of human emotion from analysis of spoken utterances. This field has attracted increasing research interest in recent years, largely due to the rapid growth of speech-based human-computer interaction applications, such as intelligent service robotics, automated call centres, remote education and so on.

Traditional SER techniques typically follow a conventional pattern recognition pipeline. This mainly focuses on robust and discriminative feature extraction, an effective classifier, and often a combination of both. The frontend feature extractor processes frame level features to obtain high-level semantic information, which is then transformed into segment level features for training the emotion classifier. Recently, motivated by the success of deep learning techniques, SER methods based on deep neural networks (DNN), convolutional neural networks [1, 2], and recurrent neural networks (RNN) [3] have been proposed.

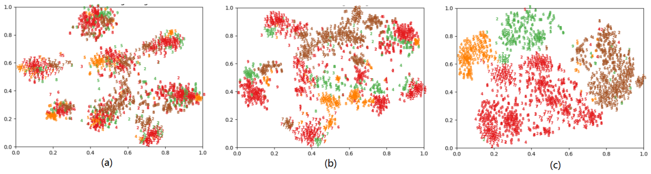

Fig. 1. The t-SNE visualization results of the extracted embeddings of IEMOCAP. Note: the colors represent different emotion labels, and the numbers indicate speaker ids. (a) Embeddings from the feature extractor, pre-trained using VoxCeleb1 corpus. (b) Embeddings from the feature extractor, fine-tuned with IEMOCAP corpus by adding a simple linear model as [4]. (c) Embeddings from the feature extractor + Attribute Disentanglement (AD) module, fine-tuned with IEMOCAP corpus by adding a simple linear model.

Despite good progress, current SER methods mainly suffers from the lack of a large-scale well-labeled data. Speech emotion is generally psychological in nature, and as such may be affected by the attributes like speaker, content, and language. In particular, the difficulty of collecting and labeling sentiment data has led to the fact that sentiment training sets often have low data volume and with ambiguous labels. It is challenging to improve the SER performance using advanced deep learning architectures like ResNet and DenseNet. Some previous research used data augmentation or generative adversarial networks (GAN) to create new data [5, 6, 7, 8]. Other researchers focused on applying the cross-corpus transfer learning methods within the scope of emotion data[9, 10]. In [11], it was shown that the annotation of emotion can be transferred from the visual (faces) to speech domain (voices) through cross-domain distillation on the VoxCeleb dataset. In [12, 13, 14, 15], the authors investigated using speaker characteristics like gender and age. Due to the close connections between speaker recognition (SR) and SER tasks, many researchers [4, 6] choose to use speaker datasets for pre-training the frontend feature extractor. In [4], it is shown that using a simple linear model is enough to obtain good SER performance on the features extracted from pre-trained models such as the x-vector model. The performance can be further improved by finetuning for emotion classification. In our previous work [16], we also studied the domain adaption method from speaker to emotion recognition tasks using residual adapters. However, the introduction of speaker data sets may bring some new problems. As shown in the Fig. 1a, features extracted by a network pre-trained with speaker corpus will be clustered according to the speaker tag of the sample. After

Yan Song is the corresponding author. This work was supported by the Leading Plan of CAS (XDC08030200)

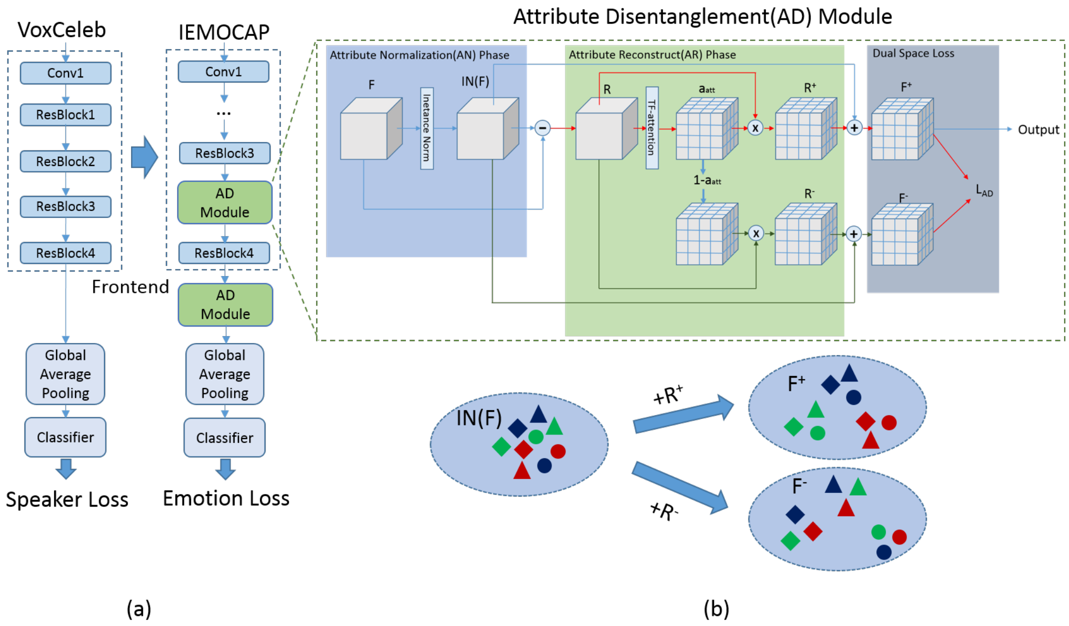

Fig. 2. illustration of the frontend attributes disentanglement framework for SER framework.

fine-tuning with emotional data (Fig. 1b), although the distribution of features has gathered according to emotion, there are still strong speaker characteristics, interfering with SER performance.

Several domain-adaptation and disentangle methods have been proposed to reduce the interference of other factors on emotion classification [17, 18, 19]. Li et al. [20], used a method similar to gradient reverse to remove the influence of speaker information. Williams et al. [21], used a structure combining two parallel auto-encoders to remove unwanted speaker information in utterance-level representations such as i-vectors and x-vectors. The entangled information about speaker and emotion is pushed apart by using the auxiliary classifiers that take one of the two latent subspaces.

In this paper, we propose to perform attributes disentanglement (AD) to improve SER performance, which mainly focuses on frontend feature extraction, as shown in Fig. 2a. Following the transfer learning framework in [4], the overall SER system utilize the frontend feature extractor which is a backbone using deep residual network (ResNet) pre-trained on a large scale SR corpus, *i.e.*, VoxCeleb1. A linear model is added after frontend feature extractor for training and finetuning the emotion classifier. The main difference lies in that we propose to insert AD module to improve the effectiveness of embedding learning .

Specifically, the AD module consists of attribute normalization (AN) and attribute reconstruction (AR) phases as shown Fig. 2b. The AN phase aims to filter out the interfering attributes other than emotion, such as speaker, speech content and language, by using instance normalization (IN). However, it is shown that simply applying IN instead of batch normalization (BN) may lead to performance degradation due to the loss of the discriminative information. In AR phase, the emotion-relevant and irrelevant spaces from the residual space are derived from the output of AN. Furthermore, motivated by [22], a time-frequency (TF) attention mechanism is introduced to exploit the long-range contextual information for both time and frequency axes. Finally, to improve the separation of two spaces for better disentanglement, the dual space loss is proposed, as detailed in 3.1.3.

2. OVERVIEW OF SPEECH EMOTION RECOGNITION ARCHITECTURES

SER research is constrained by some common problems. Firstly, increasing network depth generally benefits performance, but the limited scale of emotion corpora greatly limits complexity in practice. Most deep learning SER models contain only a few layers, and need to be specifically designed for emotion corpora. Hence, some powerful models, such as ResNet, which achieve great success in related tasks, cannot effectively be utilized for SER due to these training data limitations. Secondly, existing methods mainly focus on cross-emotion corpus learning, but due to the difficulty and cost of labelling emotion data, limitations remain in cross-corpus methods.

In this paper we therefore 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, this aims to tackle the lack of labelled corpus data, and transfer information from VoxCeleb to a specific SER target dataset. Using Resnet18 as the feature extractor, outputs are classified with a full connection layer. Firstly, VoxCeleb data is used to pre-train the network with speaker labels, then the classification layer is removed and replaced with an FC layer for emotion classification. The emotion data set is used to train the new classification layer, and the entire feature extractor is fine tuned.

After every ResBlock the AD module is used to find information that is helpful to emotion recognition, and to remove information that interferes with the emotion recognition task. The structure of the model is shown in Fig. 2 and described further in Section 4.2.

It is worth noting that the frontend AD is complementary to the style entanglement in [21] and finetuning in [4]

3. METHODS

3.1. Attribute Disentanglement (AD) module

The AD module can be used to perform disentanglement on frame level features to separate helpful and irrelevant information for SER. For an AD module, we denote the input (which is a feature map) by $F \in \mathbb{R}^{C\times T\times F}$ and the output by $F^{+} \in \mathbb{R}^{C\times T\times F}$, where C, T, F denote the size of channel, time, and frequency. The module is mainly divided into three steps:

3.1.1. *Attribute Normalization Phase*

In AD, we first apply Instance Normalization to cut down the domain differences on the input features:

$$IN(F) = \gamma(\frac{F - \mu(F)}{\sigma(F)}) + \beta \tag{1}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation computed across spatial dimensions independently for each channel and each sample. $\gamma, \beta \subseteq \mathbb{R}^{c}$ are parameters learned from data. Unlike more common normalizations such as BatchNorm, IN normalizes each sample to greatly reduce the difference between samples. It can remove style information [23], but it will also lose part of the feature information that is useful for classification.

3.1.2. *Attribute Reconstruction Phase.*

IN may inevitably removes the discriminative information. And then Attribute Reconstruction (AR) phase reconstruct the emotion-relevant features from the variation space to ensure high discrimination. We regain useful lost information by distilling it from residual feature $R = F - IN(F)$, the difference between the original input feature and the style normalized feature. Using residual information R, we aim to separate task-related information from irrelevant information. Hence an attention mask a_{att} is learned to obtain emotion-relevant, R^{+} and emotion-irrelevant features R^{-}:

$$\begin{aligned} R^{+} &= a_{att}R \\ R^{-} &= (1 - a_{att})R \end{aligned} \tag{2}$$

By adding the distilled emotion-relevant feature R^{+} to the style normalized feature IN(F), we obtain the output reconstructed feature F^{+} of the AD module as

$$\begin{aligned} F^{+} &= IN(F) + R^{+} \\ F^{-} &= IN(F) + R^{-} \end{aligned} \tag{3}$$

3.1.3. *Dual Space Loss*

In order to facilitate the disentanglement of emotion-relevant feature and emotion-irrelevant feature, we need to achieve that F^{+} should carry information which relates to classification tasks, and F^{-} should carry information irrelevant to classification task.

In SER task, due to the feature extractor being trained with speaker labels, a large amount of task-irrelevant speaker information will be incorporated. Therefore, we use a speaker classification loss function when training the AD module.

For the sample (x, y_e, y_s), where x is the input data and y_e and y_s are the emotion and speaker label sample respectively. After x is input to the feature extractor and AD module, F^{+} and F^{-} are obtained. By defining the logits obtained by transforming the two features as $f^{+} = FC(pool(F^{+})), f^{-} = FC(pool(F^{-}))$, the loss is defined as:

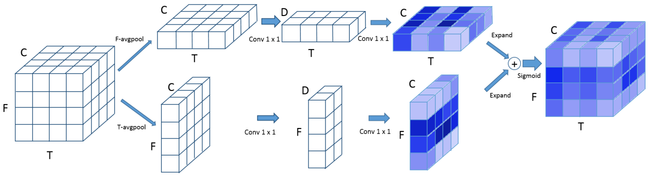

Fig. 3. Structure of the TF-attention for AR

$$\begin{aligned} L_{AD} &= L_{AD}^{+} + L_{AD}^{-} \\ L_{AD}^{+} &= CrossEntropy(f^{+}, y_e) \\ L_{AD}^{-} &= CrossEntropy(f^{-}, y_s) \end{aligned} \tag{4}$$

In this way, F^{-} can carry as much speaker information as possible, so as to remove the speaker information in F^{+} and assist in the emotion recognition task.

The overall loss is then $L = L_{Emo} + \lambda L_{AD}$. L_{Emo} denotes the emotion classification loss.

3.2. Time-Frequency attention

A self attention module needs to be used in AD. Of course, an SE-like module [22] can be used to calculate the attention weight in the channel dimension. However, in speech tasks, there are also points to pay attention to in the time and frequency dimensions, and the keynote time and frequency of emotional and speaker information may be different. Strip pooling [24] is an improvement over the generic SE module. It is intended to identify the uneven length and width of image regions.

The structure of TF-attention is shown in the Fig. 3. Given feature map $R \in \mathbb{R}^{C\times T\times F}$, we first apply average pooling in the F and T dimensions separately, and then use a 1×1 convolution layer to reduce the feature to D dimensions.

$$\begin{gathered} R_{i,j}^{F} = \frac{1}{F}\sum_{0<k\leq F} R_{i,j,k}, R_{i,k}^{T} = \frac{1}{T}\sum_{0<j\leq T} R_{i,j,k} \\ R^{FD} = Conv_{C\to D}^{1\times 1}(R^{F}), R^{TD} = Conv_{C\to D}^{1\times 1}(R^{T}) \end{gathered} \tag{5}$$

In the time dimension, each time point can derive a D-dimensional feature vector. The feature vector changes with time and aims to encapsulate the phoneme information at that time point. It is also the same in the frequency dimension. Then we use a 1×1 convolution layer to change it back to the C dimension, and use a sigmoid to map the value between 0 and 1. Hence we derive a channel dimension attention weight at each time and frequency point,

$$\begin{aligned} R^{TF} =& Expand(Conv_{D\to C}^{1\times 1}(R^{FD})) \\ &+ Expand(Conv_{D\to C}^{1\times 1}(R^{TD})) \\ a_{att} =& Sigmoid(R^{TF}) \end{aligned} \tag{6}$$

Thus, the attention mask a_{att} used in Eqn. 2 is obtained.

4. EXPERIMENTS

4.1. Dataset and acoustic features

We use the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) for all experiments. IEMOCAP contains approxi-

Table 1. ResNet parameters and dimensions.

Layer name	Parameter
Conv1	$7 \times 7, 16$
Stage1	$\begin{bmatrix} 3 \times 3, 16 \\ 3 \times 3, 16 \end{bmatrix} \times 2$
Transition1	$3 \times 3, 32$, stride 2
Stage2	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 2$
Transition2	$3 \times 3, 64$, stride 2
Stage3	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
Transition3	$3 \times 3, 128$, stride 2
Stage4	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
Transition4	$3 \times 3, 128$, stride 2

Table 2. SER results (weighted F1) for ResNet and VGG.

Backbone	Method	F1-score
ResNet18	Baseline	55.47%
	Pre-train	62.24%
	Pre-train+AD	68.19%
VGG5	Baseline	61.84%
	AD	64.34%

Table 3. F1-score comparison with existing models.

Model	F1-score
ICON [25]	63.0%
MCED [26]	60.1%
COPYPASTE [6]	63.78%
Proposed	68.19%

mately 12 hours of audiovisual data recorded by 10 skilled actors. The database has 5 sections, each containing data from one male and one female actor. Each utterance in the dataset is annotated by multiple annotators into 8 emotion labels. Following previous works, we choose 4 emotion types for experiments (neutral, happy, angry and sad) from the improvised speech for study – since scripted data may contain undesired contextual information. Adopting the methodology of previous works, we performed a 5-fold cross-validation using a leave-one-out strategy. In each training process, 8 speakers are used as training data, one of the remaining speakers is used as verification data and the final speaker as the test data.

For pre-training, we use the VoxCeleb1 dataset which contains 1152 speakers. 41-dim filter-banks are utilized as input features, extracted over 40 ms Hamming windows with a 10 ms shift between windows, and a VAD applied. Emotional data were processed in exactly the same way.

4.2. System description

As noted, we adopt a ResNet18 model for our experiments, with the detailed network parameters of the network listed in Table 1. The CNN training makes use of the PyTorch deep learning framework. When using emotional data fine-tuning, the optimization method is Adam with a mini-batch size of 128. The CNNs are trained over 50 epochs with an initial learning rate for the emotion classifier of 0.01, reducing by a factor of 10 at the 21, 31 and 41 epochs. The learning rate of the feature extractor is one hundredth of that of the classifier.

In order to increase the reliability of the experiment, we replaced the feature extractor with VGG5 network. Here, VGG5 contains five 3×3 convolution layers, and the output channel dimensions are 16, 32, 48, 64 and 80 respectively.

Although AD module can be added at any position of the feature extractor, we only use AD module at the end of the feature extractor, which results in best performance. In the AD module, the reduction dimension D is 8 and the coefficient of L_{AD}, λ is set to 0.5.

4.3. Results and analysis

Table 2 lists SER results in terms of weighted F1-scores for ResNet and VGG models. The baseline system has random initialization to learn emotional data directly, and no AD module. It is evident that using emotional data directly to train ResNet18 yields very poor results, because of the combination of high network complexity and lack of data. Performance is much improved by pre-training with VoxCeleb and fine tuning the network parameters. After adding the AD module, the performance is improved again, demonstrating that our proposed method is effective. We conducted similar experiments using VGG. Although other data sets were not used for pre-training, AD also yields an improvement.

We also give the t-SNE visualization result of the embedding processed by AD module. As shown in the Fig. 1c, we note that the AD module does help us gather the data of different speakers and the same emotion, so as to make the emotion recognition effect better.

Note that experimenting with more AD modules in the network layer did not perform well. According to our analysis, adding AD modules to more ResNet18 network layers causes the network structure to change such that many of the pre-training parameters lose their meaning. As for VGG, additional AD modules greatly increases network complexity and worsens the training effectiveness.

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

We compare the proposed model to the state-of-the-art published results in Table 3. Due to the existence of many ways of using IEMOCAP data, we only listed results which adopted a comparable evaluation methodologies. Among these models, the proposed time-frequency attention mechanism outperforms other more complex systems.

5. CONCLUSION

This paper aims to overcome the significant difficulties associated with insufficient labeled SER training data via an effective method to utilize large scale related-domain data, namely speaker recognition data, to pre-train large, high performing networks. Different from the existing methods [4, 21], which transferred the knowledge from the utterance-level representations via fine-tuning or style disentanglement, we proposed the AD module on a pre-trained frontend feature extraction. The AD module takes advantage of the Instance Normalization (IN) to filter out interference from attributes other than emotion, and to reconstruct high discriminative features for SER. To effectively disentangle the emotion-relevant and emotion-irrelevant spaces, a dual space loss is designed. Furthermore, a time-frequency (TF) attention is proposed to introduce the long-range contextual information. Evaluation on the IEMOCAP benchmark demonstrates the effectiveness of the proposed method, achieving a four class recognition performance (F1-score) of 68%, outperforming baseline models by about 6%.

6. REFERENCES

[1] Aharon Satt, Shai Rozenberg, and Ron Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. INTERSPEECH*, 2017, pp. 1089–1093.

[2] Pengcheng Li, Yan Song, Ian McLoughlin, Wu Guo, and Lirong Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proc. INTERSPEECH*, 2018.

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

[4] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP*, 2020, pp. 7169–7173.

[5] Mingke Xu, Fan Zhang, Xiaodong Cui, and Wei Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *ICASSP*, 2021, pp. 6319–6323.

[6] Raghavendra Pappagari, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak, "Copypaste: An augmentation method for speech emotion recognition," in *ICASSP*, 2021, pp. 6324–6328.

[7] Fang Bao, Michael Neumann, and Ngoc Thang Vu, "CycleGAN-Based Emotion Style Transfer as Data Augmentation for Speech Emotion Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2828–2832.

[8] Georgios Rizos, Alice Baird, Max Elliott, and Björn Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *ICASSP*, 2020, pp. 3502–3506.

[9] Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps, "Transfer learning for improving speech emotion classification accuracy," in *INTERSPEECH 2018*, 2018, pp. 257– 261.

[10] Peng Song, Wenming Zheng, Shifeng Ou, Yun Jin, Wenming Ma, and Yanwei Yu, "Joint transfer subspace learning and feature selection for cross-corpus speech emotion recognition," in *ICASSP 2018*, 2018, pp. 5504– 5508.

[11] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM International Conference on Multimedia*, New York, NY, USA, 2018, MM '18, p. 292–301, Association for Computing Machinery.

[12] Bagus Tris Atmaja and Masato Akagi, "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," in *ICASSP 2020*, 2020, pp. 4482– 4486.

[13] Biqiao Zhang, Emily Mower Provost, and Georg Essl, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *ICASSP*, 2016, pp. 5805– 5809.

[14] Rui Xia, Jun Deng, Bj´orn Schuller, and Yang Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *ICASSP*, 2014, pp. 990–994.

[15] Zhong-Qiu Wang and Ivan Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *ICASSP*, 2017, pp. 5150–5154.

[16] Yuxuan Xi, Pengcheng Li, Yan Song, Yiheng Jiang, and Lirong Dai, "Speaker to emotion: Domain adaptation for speech emotion recognition with residual adapters," in *Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2019.

[17] Jun Deng, Zixing Zhang, Florian Eyben, and Bj´orn Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE SIGNAL PROCESSING LETTERS*, vol. 21, pp. 1068–1072, 2014.

[18] Jun Deng, Xinzhou Xu, Zixing Zhang, and Sascha Frühholz, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 26, pp. 31–43, 2018.

[19] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE SIGNAL PROCESSING LETTERS*, vol. 24, pp. 500–530, 2017.

[20] Haoqi Li, Ming Tu, Jing Huang, Shrikanth Narayanan, and Panayiotis Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *ICASSP*, 2020, pp. 7144–7148.

[21] Jennifer Williams and Simon King, "Disentangling Style Factors from Speaker Representations," in *Proc. Interspeech 2019*, 2019, pp. 3945–3949.

[22] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu, "Squeeze-and-excitation networks," in *Computer Vision and Pattern Recognition(CVPR)*, 2018.

[23] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1510–1519.

[24] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4003–4012.

[25] Jilt Sebastian and Piero Pierucci, "Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts," in *Proc. Interspeech 2019*, 2019, pp. 51–55.

[26] Ruichen Li, Jinming Zhao, and Qin Jin, "Speech Emotion Recognition via Multi-Level Cross-Modal Distillation," in *Proc. Interspeech 2021*, 2021, pp. 4488–4492.