SPEAKER-INVARIANT AFFECTIVE REPRESENTATION LEARNING VIA ADVERSARIAL TRAINING

Haoqi Li[⋆]   *Ming Tu*[†]   *Jing Huang*[†]   *Shrikanth Narayanan*[⋆]   *Panayiotis Georgiou*[⋆]

[⋆]Dept. of Electrical and Computer Engineering, University of Southern California, USA
[†]JD AI Research, Mountain View, USA

ABSTRACT

Representation learning for speech emotion recognition is challenging due to labeled data sparsity issue and lack of gold-standard references. In addition, there is much variability from input speech signals, human subjective perception of the signals and emotion label ambiguity. In this paper, we propose a machine learning framework to obtain speech emotion representations by limiting the effect of speaker variability in the speech signals. Specifically we propose to disentangle the speaker characteristics from emotion through an adversarial training network in order to better represent emotion. Our method combines the gradient reversal technique with an entropy loss function to remove such speaker information. Our approach is evaluated on both IEMOCAP and CMU-MOSEI datasets. We show that our method improves speech emotion classification and increases generalization to unseen speakers.

Index Terms— Speech emotion recognition, adversarial training, speaker invariant, affective representation

1. INTRODUCTION

Human speech signals contain rich linguistic and paralinguistic information. Linguistic information is encoded at different temporal scales ranging from phoneme to sentence and discourse levels. More importantly, speech signal encodes speaker characteristics and affective information. All information above is jointly modulated and intertwined in the human-produced speech acoustics and it is difficult to dissociate these various components simply from features, such as those from the time waveform or its transformed representations e.g., Mel filterbank energies.

Representation learning of speech [1, 2, 3], i.e., the transformation from low-level acoustic descriptors to higher-level representations, has received significant attention recently. Traditional methods focus on using supervised learning, specifically multi-task learning [4] to extract specialized representations of particular targets. However, target representations are easily contaminated by undesired factors, such as noise, channel or source (speaker) variability. These are difficult to eliminate due to the complexity and entanglement of information sources in the speech signal.

Emotion recognition systems are further greatly affected by source variability, be that speaker, ambient acoustic conditions, language, or socio-cultural context [5]. Limited domain data and labeling costs have resulted in many systems that are only evaluated within domain and are not robust to such variability. For example mismatch between training and evaluation sets, such as speaker variations [6] and domain condition incongruity [7], make it challenging to obtain robust emotion representations across different speakers and domains.

In this work, we propose an adversarial training framework to learn robust speech emotion representations. We specifically aim to remove speaker information from the representation, as that is one of the most challenging and confounding issues in speech emotion recognition (SER). Note that many SER systems have addressed this issue through normalization of features, but these ad-hoc solutions lack generalization within complex learning frameworks [8, 9].

In our work, inspired by the domain adversarial training (DAT) [7], we propose a neural network model and an adversarial training framework with an entropy-based speaker loss function to relieve speaker variability influences. Considering the adversarial training strategy and entropy-based objective function, we name our model *Max-Entropy Adversarial Network* (*MEnAN*). We demonstrate the effectiveness of the proposed framework in SER within- and across-corpora. We show that MEnAN can successfully remove the speaker-information from extracted emotion representations, and this disentanglement can further effectively improve speech emotion classification on both the IEMOCAP and CMU-MOSEI datasets.

2. RELATED WORK

Robust representations of emotions in speech signals have been investigated via pre-trained denoising autoencoders [10], end-to-end learning from raw audio [11], unsupervised contextual learning [3], and multi-task learning [12] etc. In a different way, we apply GANs based adversarial training to generate robust representations across domains (speaker to be specific) for speech emotion recognition. Among previous work on SER, GANs are mainly utilized to learn discriminative representations [13] and conduct data augmentation [14]. Our method is different in that we aim to disentangle speaker information and learn speaker-invariant representations for SER.

Recently, within speech applications, domain adversarial training (DAT) techniques have been applied on cross-corpus speaker recognition [15], automatic speech recognition [16, 17, 18] and SER [19, 20] to deal with the domain mismatch problems. Compared to the two most related studies [19, 20], our proposed MEnAN is different from DAT: 1) we argue that simple gradient reversal layer in DAT may not guarantee domain-invariant representation: simply flipping the domain labels can also fool the domain classifier however the learned representation is not necessary to be domain invariant. 2) we propose a new entropy-based loss function for domain classifier to induce representations that maximize the entropy of the domain classifier output, and we show the learned representation is better than DAT for speech emotion recognition.

3. METHODOLOGY

Our goal is to obtain an embedding from a given speech utterance, in which emotion-related information is maximized while minimizing the information relevant to speaker identities. This is achieved by our proposed adversarial training procedure with designed loss functions which will be introduced in this section.

3.1. Model structure

Our proposed model is built based on a multi-task setting with three modules: the representation encoder (ENC), the emotion category classification module (EC) and the speaker identity classification module (SC). The structure of our model is illustrated in Figure 1.
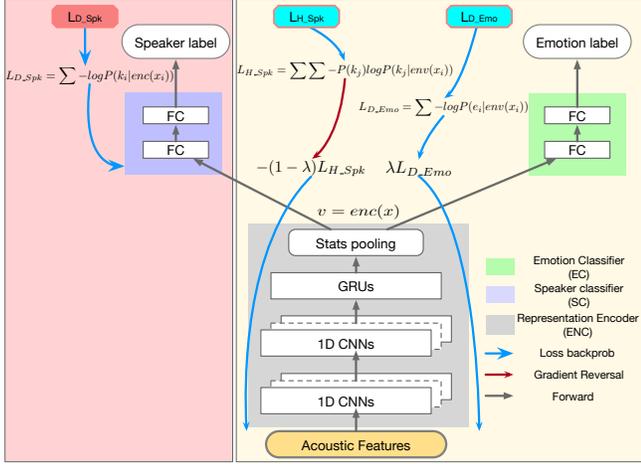

Fig. 1: Model structure with loss propagating flow

The ENC module has three components: (1) stacked 1D convolutional layers; (2) recurrent neural layers and (3) statistical pooling layers. The sequence of acoustic spectral features is first input to multiple 1D CNN layers. The CNN kernel filters shift along the temporal axis and include the entire spectrum information per scan, which is proven to have better performance than other kernel structure settings by [21]. CNN filters with different weights are utilized to extract different information from same input features and followed by recurrent layers to capture context and dynamic information within each speech segment. Then, we add the statistical pooling functions, including maximum, mean and standard deviation in our model, to map a variable speech segment into an embedding with a fixed dimension.

This fixed dimension representation embedding, as the output of ENC, is further connected with two sub-task modules: the emotion classifier (EC) and speaker classifier (SC), which are both built with stacked fully connected layers.

With normal training settings, our model can be regard as a multi-task learning framework. Moreover, our model can be regarded as a speech emotion recognition system if we only keep the EC and ENC components.

3.2. Difference with prior work

In domain adversarial training [7, 19, 20], one gradient reversal layer is usually added to the domain classifier (SC in our case) to reverse the gradient flow in order to generate the domain-invariant (speaker-invariant) features. The usage of the gradient reversal layer ensures the (desired) lower performance of domain classifier (SC in our case), however, it often fails to guarantee the domain (speaker) information has been fully removed [22]. For instance, in this approach, it is highly likely that even a lower performance of SC will only map a particular speaker to other target speakers with similar sounds instead of properly removing the speaker identity information, likely picking up the second-best speaker match.

Our proposed training method is different from the existing strategy as it attempts to completely remove all speaker information.

3.3. Emotion representation adversarial training

We now describe the adversarial training strategy and the designed loss function in detail. Our training dataset $\mathcal{D} = \{(x_1, e_1, k_1), ..., (x_N, e_N, k_N)\}$ contains N pairs of $(x_i, e_i, k_i) \in (\mathcal{X}, \mathcal{E}, \mathcal{K})$, in which speech segment x_i is produced by the speaker k_i with emotion e_i. $\mathcal{X}$, $\mathcal{E}$ and $\mathcal{K}$ are the sets of whole speech utterances, emotion labels and all speakers respectively.

Our training strategy is similar to generative adversarial networks (GANs) [23]. The system has two output paths. On one path (left Fig.1), we attempt to accurately estimate the speaker information (loss L_{D_Spk}). On the other path (right Fig.1), we attempt to estimate the emotion label (loss L_{D_Emo}) and remove speaker information (loss L_{H_Spk}). Both estimators (SC and EC) employ the same representation encoder (ENC) but that is only updated from the right-side loss back propagation.

The output of ENC, denoted as v, is the speaker-invariant emotion representation we try to obtain. We have $v = enc(x)$.

3.3.1. Training of SC

The speaker classifier (SC) can be regarded as a discriminator which is trained to distinguish speaker identities based on a given encoder output v and has no influence in the training of v. The SC is trained by minimizing L_{D_Spk}, the cross entropy function as in (1):

$$L_{D\_Spk} = \sum_{(x_i, e_i, k_i) \in \mathcal{D}} -logP(k_i|enc(x_i)) \tag{1}$$

In this training step, weights of ENC and EC are frozen. Only parameters of SC are optimized to achieve higher speaker classification accuracy from a given representation v.

3.3.2. Training of ENC and EC

Under adversarial training we need to ensure the ENC output contains emotion-related information, while it is also optimized to confuse and make it difficult for SC to distinguish speaker identities.

Thus, we need to optimize ENC to increase the uncertainty or randomness of SC's outputs. Mathematically, we want to maximize the entropy value of SC's output. The entropy of SC's output, denoted as L_{H_Spk}, is defined as

$$L_{H\_Spk} = \sum_{(x_i, e_i, k_i) \in \mathcal{D}} \sum_{k_j \in \mathcal{K}} -P(k_j)logP(k_j|env(x_i)) \tag{2}$$

Maximizing entropy would promote equal likelihood for all speakers:

$$P(k_j|env(x_i)) = P(k_q|env(x_i)) \quad \forall k_j, k_q \in \mathcal{K} \tag{3}$$

This differs, as mentioned above, from simply picking up a different speaker, since that may lead into selecting a "second-best" similar sounding speaker. Our proposed loss function removes all (correct or wrong) speaker information.

In addition, the performance of emotion classifier is optimized by minimizing the cross entropy loss L_{D_Emo} from EC's output:

$$L_{D\_Emo} = \sum_{(x_i, e_i, k_i) \in \mathcal{D}} -logP(e_i|env(x_i)) \tag{4}$$

To combine these two objective functions above together, we flip the sign of L_{H_Spk} to do a gradient reversal and minimize the weighted overall loss sums. The final objective loss function is written as:

$$L(\theta_{ENC}, \theta_{EC}) = \lambda L_{D\_Emo} - (1 - \lambda)L_{H\_Spk} \qquad (5)$$

where $\lambda \in (0, 1)$ is a parameter adjusting the weighting between two types of loss functions.

In this training step, weights of SC are frozen. Only parameters of ENC and EC are optimized. Modules with corresponding loss back propagation flows are shown in Fig.1. With this iterative training scheme, we expect the proposed model can ultimately relieve the impact of speaker variability thus improve the SER performance.

4. DATESET

Two datasets are employed to evaluate the proposed MEnAN based emotion representation learning in our work:

The **IEMOCAP** dataset [24] consists of five sessions of speech segments with categorical emotion annotation, and there are two different speakers (one female and one male) in each session. In our work, we use both improvised and scripted speech recordings and merge *excitement* with *happy* to achieve a more balanced label distribution, a common experiment setting in many studies such as [10, 25, 26]. Finally, we obtain 5,531 utterances selected from four emotion classes (1,103 angry, 1,636 happy, 1,708 neutral and 1,084 sad).

The **CMU-MOSEI** dataset [27] contains 23,453 single-speaker video segments carefully chosen from YouTube. This database includes 1000 distinct speakers, and are gender balanced with an average length of 7.28 seconds. Each sample has been manually annotated with a [0,3] Likert scale on the six basic emotion categories: happiness, sadness, anger, fear, disgust, and surprise. The original ratings are also binarized for emotion classification: for each emotion, if a rating is greater than zero, it is considered that there is presence of that emotion, while a zero results in a false presence of that emotion. Thus, each segment can have multiple emotion presence labels.

IEMOCAP provides a relatively large number of samples within each combination across different speakers and emotions, making it feasible to train our speaker-invariant emotion representation. We mainly use CMU-MOSEI for evaluation purposes, since it includes variable speaker identities, and to establish cross-domain transferability of MEnAN.

5. EXPERIMENT SETUP

Feature extraction: In this work we utilize 40 dimensional Log-Mel Filterbank energies (Log-MFBs), pitch and energy. All these features are extracted with a 40 ms analysis window with a shift of 10 ms. For pitch, we employ the extraction method in [28], in which the normalized cross correlation function (NCCF) and pitch (f0) are included for each frame. We do not perform any per-speaker/sample normalization.

Data augmentation: To enrich the dataset, we perform data augmentation on IEMOCAP. Similar to [29], we create multiple data samples for training by slightly modifying the speaking rate with four different speed ratios, namely 0.8, 0.9, 1.1 and 1.2.

General settings: To obtain a reliable evaluation of our model, we need to ensure unseen speakers for both validation and testing. Thus, we conduct 10 fold leave-one-speaker out cross-validation scheme. More specifically, we use 8 speakers as the training set, and for the remaining session (two speakers), we select one speaker for validation and one for testing. We then repeat the experiment with the two speakers switched.

In addition, considering the variable length of utterances, we only extract the middle 14s to calculate acoustic features for utterance whose duration is longer than 14s (2.07% of the total dataset) [30], since important dynamic emotional information is usually located in the middle part and lengthy inputs would have negative effect on emotion prediction [31]. For utterances shorter than 14s, we use the cycle repeat mode [31] to repeat the whole utterance to the target duration. The idea of this cycle repeat mode is to make emotional dynamic cyclic and longer, which facilitates the training process of utterances of variable duration.

5.1. Model configurations

The detailed model parameters and training configurations are shown in Table 1.

Training details:	Adam optimizer (lr=0.001) + polynomial learning rate decay ; batch size=16; epochs=300; λ=0.5
ENC	Conv1D(in_ch=43,out_ch=32, kernel size=10, stride=2, padding=0), PReLU Conv1D(in_ch=32,out_ch=32, kernel size=5,stride=2, padding=0), PReLU GRU(in_size=32, hidden_size=32, num_layers=1) Linear(in=32, out=32), PReLU Statistical Pooling[Mean, Std, Max]
EC	Linear(in=32, out=32) PReLU Linear(in=32, out=10) PReLU Linear(in=10, out=4)
SC	Linear(in=32, out=32) PReLU Linear(in=32, out=10) PReLU Linear(in=10, out=8)

Table 1: Model structure and training configuration details

6. RESULTS AND DISCUSSION

Evaluation on IEMOCAP

For comparison purposes, we also train the EC only model, multi-task learning model and DAT model [20] with regular cross entropy loss under the same configuration. Both the weighted accuracy (WA, the number of the correctly classified samples divided by the total number of samples) and the unweighted accuracy (UA, the mean value of the recall for each class) are reported. The Table 2 shows the emotion classification accuracy (%) on both validation and testing, and we also include their differences (Δ).

	WA			UA		
	Val	Test	Δ	Val	Test	Δ
EC only model	58.50	55.92	-2.56	59.94	57.45	-2.49
Multi-task model	59.24	55.90	-3.34	60.52	57.28	-3.24
DAT model	58.28	56.68	-1.60	60.16	58.48	-1.68
Proposed MEnAN	58.85	**58.62**	-0.23	60.24	**59.91**	-0.33

Table 2: Classification accuracy (%) comparison on IEMOCAP

First, we observe that our model achieves the best classification accuracy in the test case among all models. To the best of our knowledge, the best results from the literature on IEMOCAP with similar settings are generally around 60% [32, 25]. We achieve a UA of 59.91% which is comparable with the state of the art results. However, strict comparisons remain difficult because there are no standardized data selection rules or train/test splits. For example, some did not use speaker independent split [25] or only used improvised utterances. Some did not clearly specify which speaker in each session was used for validation and testing respectively [33] or performed per-speaker normalization in advance [9, 34].

Second, we notice that there is a large difference of Δ value among all four models. Compared with others, we find that the multi-task learning model can achieve a better performance on the validation set. However, the extra gain from speaker information can also lead a significant mismatch during the evaluation of unseen speakers, as indicated by the large value of Δ. Compared with DAT model, our MEnAN model gains better classification accuracy with smaller Δ. This supports our claim of the MEnAN's advantage over DAT. The small Δ in our model suggests our embedding has better generalization ability and is more robust to unseen speakers. To illustrate this, we plot t-SNE of emotion representation, i.e., $enc(x)$, on two unseen speakers.
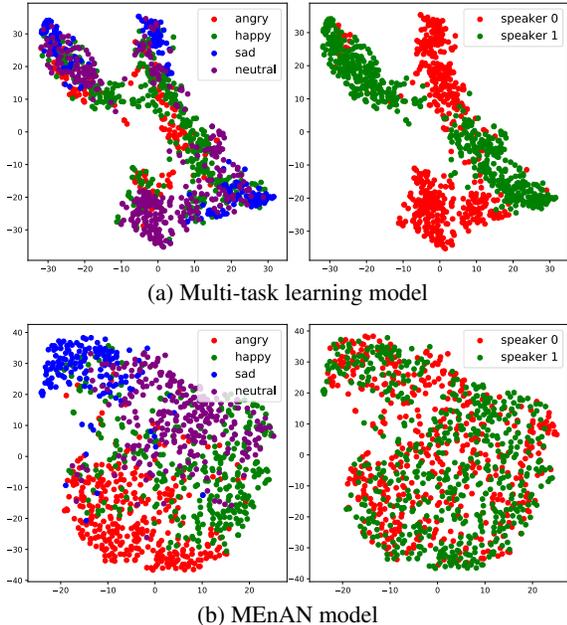

(a) Multi-task learning model

(b) MEnAN model

Fig. 2: t-SNE plot of emotion embedding with both 4 emotion labels (left) and 2 speaker labels (right) for multi-task model and our proposed MEnAN model.

As shown in Fig.2, in the multi-task learning setting, it is obvious that the speaker's characteristics and emotion information are entangled with each other, which makes this representation less generic on unseen speakers. For our proposed MEnAN, the speaker representations of different speakers on the 2D space are well mixed and independent of speaker labels; while different emotion segments are more separable in the embedding manifold. These results further demonstrates the effectiveness and robustness of proposed model.

Evaluation on CMU-MOSEI

In addition, we test our system on the CMU-MOSEI dataset (cross-corpus setting). As mentioned before, the CMU-MOSEI has a large variability in speaker identities, which is a suitable corpus to evaluate our model's generalization ability on unseen speakers. It also introduces a challenge stemming from the different annotation methodology and the inherent effect on the interpretation of labels.

To match emotion labels of IEMOCAP, we only consider samples with positive ratings in the categories of "happiness", "sadness" and "anger". Samples with zero ratings of all six emotion categories are also included with the label "Neutral". Finally, 22,323 samples are selected and four-class emotion classification evaluation are performed. The prediction is considered to be correct if the rating of that predicted emotion originally has a positive value. In Table 3, we report the mean, minimum and maximum of the classification ac-

curacy (%) evaluated on the pretrained model of each fold from the 10-fold cross validation of IEMOCAP.

	mean	min	max
EC only model	31.35	27.14	34.96
DAT model	32.34	25.91	37.85
Proposed MEnAN	**33.24**	**28.84**	**39.85**

Table 3: Emotion classification accuracy (%) on CMU-MOSEI

We observe that MEnAN model has the best performance among all three models, and it achieves better classification accuracy with 1.89% improvement on the mean value and with 4.89% on the best model compared with the EC only model. Considering that all speakers of these evaluation samples are not seen during the training, these results suggest our adversarial training framework can provide more robust emotion representation with better speaker-invariant property and achieve improved performance in the emotion recognition task.

7. CONCLUSION

Compared with other representation learning tasks, the extraction of speech emotion representation is challenging considering the complex hierarchical information structures within the speech, as well as the practical low-resource (labeled) data issue. In our work, we use an adversarial training strategy to generate speech emotion representations while being robust to unseen speakers. Our proposed framework MEnAN, however, is not limited to the emotion recognition task, and it can be easily applied to other domains with similar settings e.g., cross-lingual speaker recognition. For further work, we plan to combine the domain adaption techniques with our proposed model to employ training samples from different corpora. For example, we can utilize speech utterances from speaker verification tasks to obtain more robust speaker information.

8. REFERENCES

[1] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

[2] Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *arXiv preprint arXiv:1904.03416*, 2019.

[3] Haoqi Li, Brian Baucom, and Panayiotis Georgiou, "Unsupervised latent behavior manifold learning from acoustic features: Audio2behavior," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[4] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[5] Björn 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.

[6] Puming Zhan and Martin Westphal, "Speaker normalization based on frequency warping," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1997, vol. 2, pp. 1039–1042.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[8] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps, "Speaker normalisation for speech-based emotion detection," in *15th international conference on digital signal processing*. IEEE, 2007, pp. 611–614.

[9] Carlos Busso, Angeliki Metallinou, and Shrikanth S Narayanan, "Iterative feature normalization for emotional speech detection," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011.

[10] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer, "Representation learning for speech emotion recognition.," in *Interspeech*, 2016, pp. 3603–3607.

[11] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016.

[12] Biqiao Zhang, Emily Mower Provost, and Georg Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 85–99, 2017.

[13] Jonathan Chang and Stefan Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[14] Saurabh Sahu, Rahul Gupta, and Carol Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," *arXiv preprint arXiv:1806.06626*, 2018.

[15] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[16] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition.," in *INTERSPEECH*. San Francisco, CA, USA, 2016.

[17] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[18] Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gang, and Biing-Hwang Juang, "Speaker-invariant training via adversarial learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[19] Mohammed Abdelwahab and Carlos Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[20] Ming Tu, Yun Tang, Jing Huang, Xiaodong He, and Bowen Zhou, "Towards adversarial learning of speaker-invariant representation for speech emotion recognition," *arXiv preprint arXiv:1903.09606*, 2019.

[21] Che-Wei Huang, Shrikanth Narayanan, et al., "Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition," *arXiv preprint arXiv:1706.02901*, 2017.

[22] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan, "Transferable adversarial training: A general approach to adapting deep classifiers," in *International Conference on Machine Learning*, 2019, pp. 4013–4022.

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[24] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[25] Michael Neumann and Ngoc Thang Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[26] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost, "Progressive neural networks for transfer learning in emotion recognition," *Proc. Interspeech 2017*, pp. 1098–1102, 2017.

[27] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[28] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[29] Dengke Tang, Junlin Zeng, and Ming Li, "An end-to-end deep learning framework for speech emotion recognition of atypical individuals.," in *Interspeech*, 2018, pp. 162–166.

[30] Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[31] Jian Huang, Ya Li, Jianhua Tao, Zhen Lian, et al., "Speech emotion recognition from variable-length inputs with triplet loss function.," in *Interspeech*, 2018, pp. 3673–3677.

[32] Rui Xia and Yang Liu, "Leveraging valence and activation information via multi-task learning for categorical emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[33] Jinkyu Lee and Ivan Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[34] Saurabh Sahu, Vikramjit Mitra, Nadee Seneviratne, and Carol Espy-Wilson, "Multi-modal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription," *Proc. Interspeech 2019*, pp. 3302–3306, 2019.