INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Adaptive Domain-Aware Representation Learning for Speech Emotion Recognition

Weiquan Fan[1], *Xiangmin Xu*[1,3], *Xiaofen Xing*[1,*], *Dongyan Huang*[2]

[1]School of Electronic and Information Engineering, South China University of Technology, China
[2]UBTECH Robotics Corp, China
[3]Institute of Modern Industrial Technology of SCUT in Zhongshan, China

weiquan.fan96@gmail.com, xmxu@scut.edu.cn, xfxing@scut.edu.cn, dongyan.huang@ubtrobot.com

Abstract

Speech emotion recognition is a crucial part in human-computer interaction. However, representation learning is challenging due to much variability from speech emotion signals across diverse domains, such as gender, age, languages, and social cultural context. Many approaches focus on domain-invariant representation learning which loses the domain-specific knowledge and results in unsatisfactory speech emotion recognition across domains. In this paper, we propose an adaptive domain-aware representation learning that leverages the domain knowledge to extract domain aware features. The proposed approach applies attention model on frequency to embed the domain knowledge in the emotion representation space. Experiments demonstrate that our approach on IEMOCAP achieves the state-of-the-art performance under the same experimental conditions with WA of 73.02% and UA of 65.86%.

Index Terms: speech emotion recognition, human-computer interaction, domain-aware representation learning

1. Introduction

Speech emotion recognition (SER) has attracted attention of various academic researchers and industrial engineers in human-computer interaction areas. However, understanding emotions in speech is an extremely difficult task for computers due to the different emotion expressions of people across gender, age, languages and social cultural context. Recently, various research works have been conducted in SER, and a great progress has been made.

In general, a SER system consists of two parts: feature extraction and classification. Low-level descriptors (LLD) like prosody and spectral features extracted by openSMILE toolkit [1] are widely used in SER. The INTERSPEECH 2013 COMPARE feature set [2] contains 6373 emotion-related features and promotes the development in SER. In [3], the basic acoustic features sets, Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and the extended GeMAPS are proposed. In recent years, it has been reported that the spectrogram shows an excellent performance in emotion recognition [4, 5].

In order to make full use of these features, numerous researches have been carried out on classifiers. In [6], Schuller has applied Hidden Markov Model (HMM) to SER. Lee [7] has also made some progresses using decision tree. Stuhlsatz [8] has proposed a generalized discriminant analysis method based on Restricted Boltzmann Machines (RBM), which raises the recall rate. With the development of deep learning, LSTM [9, 4] and CNN [10, 11] , which have better ability to capture high level semantics, have become the most common classifiers. Sahu [12, 13] has built a model based on adversarial autoencoders (AAE) and achieved further performance. In addition, some studies [14, 15, 16] have benefited from multi-task learning structures to improve performance.

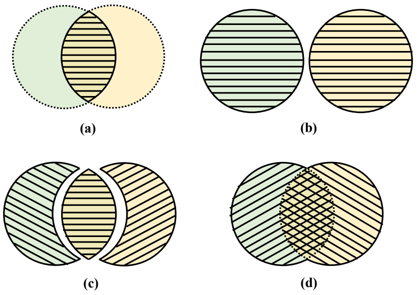


Figure 1: *Different schemes for domain A (green) and domain B (yellow). (a) Map data of the two domains to a common shadow domain to find domain-invariant features. (b) Train a model for each domain separately. (c) Find domains which are orthogonal to the common domain base on (a), and map data to domain-invariant and domain-specific features according to the given domain knowledge. (d) Automatically select domain A or B according to the learned domain knowledge.*

Speech features vary greatly according to the special factors, such as gender, age, language and social cultural context. Despite a lot of progress have been made, SER is still very challenging. Lots of researchers have attempted to find domain-invariant features as Figure 1 (a). For instance, [17] used adversarial multitask training to extract domain share features. However, these methods are limited due to lack of domain knowledge. In the field of depression detection, [18] considers male and female as two domains and has trained two models for each domain separately as Figure 1 (b). In [19], three LSTMs are used to model a common domain (domain-invariant) and two orthogonal domains (domain-specific) for text emotion classification as Figure 1 (c). Nevertheless, in the inference phase, these models need to know in advance which domain the test data belongs to so as to select the corresponding domain model and it does not work with unseen domains.

In [20], a prefrontal cortex-like (PFC-like) module is proposed to learn context-specific mappings, which modulate the representation of features using contextual information. Inspired by [20], we propose an adaptive domain-aware representation Learning (ADARL) in this paper, which consists of a multi-task representation learning by combining domain classification and emotion recognition, and an extracting domain aware representation through the domain knowledge embed-

[*] corresponding author
This work was done by UBTECH-SCUT Joint Research Lab.

http://dx.doi.org/10.21437/Interspeech.2020-2572

ding module, which adapts features to its own domain as Figure 1 (d). Results demonstrate the effectiveness of our approach on the IEMOCAP database [21].

2. Adaptive Domain-Aware Representation Learning

In this section, we present our adaptive domain-aware representation learning (ADARL) in details. The schematic diagram of our approach is shown in Figure 2. Firstly, we introduce a multi-task learning framework including domain pathway and emotion pathway. Then we describe how to implement domain-aware attention module (the yellow part in Figure 2) and the implementation in details.

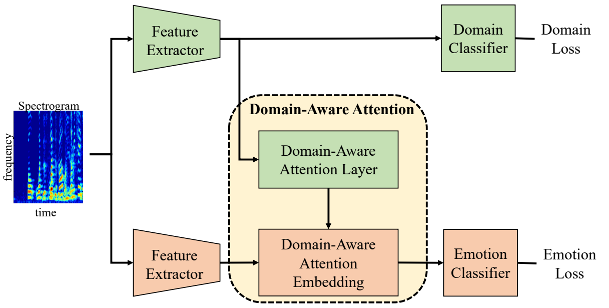


Figure 2: *Adaptive Domain-Aware Representation Learning. At the inference time, the domain classifier is not needed.*

2.1. Multi-task Learning

As shown in Figure 2, ADARL adopts the multi-task learning framework, consisting of domain classification and emotion recognition. The input of both pathways is speech spectrogram, which can be taken as a picture representing the magnitudes of energy over continuous time and frequency. For each pathway, a feature extractor with four convolution layers is used to extract high level semantic features separately. The outputs of the extractor for the domain pathway and the emotion pathway are denoted as F_4^D and F_4^E, respectively.

For the domain pathway, the feature maps F_4^D are flatten to a vector, followed by two fully connected layers to predict the domain label. In general, the first fully connected layer is followed by the ReLU layer and batch normalization layer.

$$Out^D = w_{fc2}^D \times BN^D(ReLU(w_{fc1}^D \times F_4^D(:))) \quad (1)$$

where w_{fc2}^D and w_{fc2}^D are the weights of two fully connected layers, and $(:)$ is the flatten operation.

For the emotion pathway, the domain knowledge is obtained through the domain-aware attention layer and is embedded into F_4^E to get the feature maps F_{emb}^E (see Section 2.2). Then, an average pooling along time axis (TP) is applied to aggregate and smooth the time information, which outputs $F_{TP}^E = TP^E(F_{emb}^E) \in \mathbb{R}^{C\times H}$. Next, a channel-wise fully connected(CFC) layer is set to further synthesize the frequency information. Specifically, given C weights $w_{CFC}^k \in \mathbb{R}^{1\times H}, (k \in \{1,2,3,...,C\})$, the inputs $F_{TP}^E = \left\{f_{TP}^{k,i}\right\}$ are mapped into $F_{CFC}^E = \left\{f_{CFC}^k\right\} \in \mathbb{R}^C$.

$$f_{CFC}^k = w_{CFC}^k \times f_{TP}^{k,:} \quad (2)$$

Lastly, two more fully connected layers are added to predict the emotion.

$$Out^E = w_{fc2}^E \times BN^E(ReLU(w_{fc1}^E \times F_{CFC}^E)) \quad (3)$$

where w_{fc2}^E and w_{fc2}^E are the weights of two fully connected layers.

2.2. Domain-Aware Attention

Figure 3 describes the detail of domain-aware attention module. The feature maps in the domain pathway $F_4^D \in \mathbb{R}^{C\times H\times W}$ are transformed by the domain-aware attention layer to acquire the domain attention knowledge, which will be embedded into the emotion feature space through the domain-aware attention embedding.

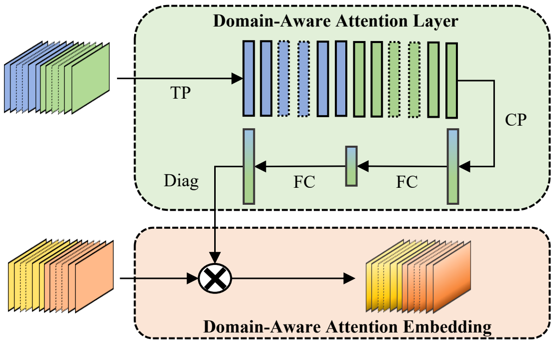

Domain-Aware Attention Layer
TP
CP
Diag
FC
FC
Domain-Aware Attention Embedding

Figure 3: *The details of the Domain-Aware Attention module. Here, TP, CP and FC are time pooling, channel pooling and fully connected layer, respectively.*

Domains, such as gender or age, are closely related to frequency in spectrogram. Thus, we keep the frequency information, and utilize the time and channel pooling layers. The time pooling layer smooths the time information to make the module insensitive to time, followed by the channel pooling layer to get more comprehensive depth features.

The resulting features retains frequency characteristics. Similar to [22], a transformation including two fully connected layers is applied to obtain the final domain knowledge. It $I^D \in \mathbb{R}^H$ can be denoted as

$$I^D = \sigma(w_{fc2}^I \times (ReLU(w_{fc1}^I \times CP^I(TP^I(F_4^D))))) \quad (4)$$

where TP^I and CP^I are the average pooling along the time axis and the channel axis, respectively. w_{fc2}^I and w_{fc2}^I are the weights of two fully connected layers, and σ is the sigmoid activation function.

Domain knowledge can be used to predict which domain the current sample belongs to, which reflects the sensitivity of the domain to the frequency. With domain-aware attention, the features in the emotion pathway can contain both emotion knowledge and auxiliary domain knowledge. Given the domain knowledge $I^D \in \mathbb{R}^H$, the emotion feature maps $F_4^E = \left\{f^k\right\} (k = 1,2,3,...,C)$ are refined through frequency attention . Therefore, the resulting feature maps $F_{emb}^E = \left\{f_{emb}^k\right\}$ can be calculated as Equation 5, where $diag(.)$ diagonalizes a vector to a diagonal matrix.

$$f_{emb}^k = diag(I^D) \times f^k \quad (5)$$

2.3. Loss

There are two pathways in ADARL, one for predicting domain and the other for emotion recognition. In the training phase, for the domain pathway, only the most basic cross entropy (CE) loss function is used. Define the ground-truth as p and the prediction as $\hat{p}$, and the CE loss is described as below.

$$\mathcal{L}^D = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} p_i^k \log \hat{p}_i^k \tag{6}$$

where N is the batch size of the training set and K is the number of categories.

For the emotion pathway, define the ground-truth as y and the prediction as $\hat{y}$. Similarly, CE loss for classification tasks $\mathcal{L}^E$ is also applied.

$$\mathcal{L}^E = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_i^k \log \hat{y}_i^k \tag{7}$$

In addition, considering the diversity of emotional expressions, center loss is used to cluster the features of the intra-class samples. which [23] solves this problem and achieves intra-class compactness by restricting feature distance, as formulated in Equation 8.

$$\mathcal{L}_{cen}^E = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{c}_{y_i}||_2^2, \tag{8}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is the deep feature with d dimension of the i-th sample, and $\mathbf{c}_{y_i} \in \mathbb{R}^d$ is the center of y_i-th category, initialized by xavier [24]

Lastly, the final function can be describe as

$$\mathcal{L} = \gamma_1\mathcal{L}^D + \gamma_2\mathcal{L}^E + \gamma_3\mathcal{L}_{cen}^E \tag{9}$$

where γ_1, γ_2 and γ_3 are coefficients to balance the importance among these losses. It should be noted that in order to improve domain prediction performance, additional training data with domain label can be introduced.

2.4. Implementation of ADARL

Due to the uncertainty of the sentence length, we divide the sentences longer than 3 seconds into multiple non-overlapping sub-segments with 3 seconds. The probability of the whole sentence is obtained from the average of each sub-segment. For each sub-segment, STFT with Hamming window length of 20 ms, 40 ms and the window shift of 10 ms is performed, followed by logarithmic operation and z-score normalization. For the frequency axis, only the 0-4 kHz band (400 points) with significant information is reserved. For the time axis, zero padding is applied to get 300 points. Therefore, a $3 \times 400 \times 300$ spectrogram is drawn by MATLAB for each sub-segment.

The parameters of our approach are shown in Table 1. Considering the temporal uncertainty of speech, we use multi-scale CNN [25] to extract high level features, whose basic unit is multi-scale convolution. Specifically, convolution kernels with different scales are applied to convolve the input. The feature maps after convolution are then concatenated along the channel dimension. Similarly, the batch normalization layer and ReLU layer are followed closely behind. Finally, an average pooling layer is used for downsampling. Therefore, the output of l-layer F_l is described as

$$F_l = p_{avg}(ReLU(BN_l(\{w_{1l} * F_{l-1}, w_{2l} * F_{l-1}\}))) \tag{10}$$

where F_0 is the original spectrogram. w_1 and w_2 are the different convolution kernel weights. $\{,\}$, $*$ and p_{avg} are the channel concatenation, convolution and average pooling operation, respectively.

Table 1: *Specifications of our ADARL with convolution (C^{in}, C^{out}, K), AvgPool (K) and FC ((C^{in}, C^{out}), where C^{in} is input channels, C^{out} is output channels, K is kernel size. The stride of AvgPool is the same as its kernel size.*

Domain pathway	Emotion pathway	Domain-aware attention
Conv (3, 16, (3, 3)) + BN + ReLU		-
Conv (3, 16, (3, 3)) + BN + ReLU		
AvgPool ((2, 2))		
Conv (32, 16, (3, 3)) + BN + ReLU		-
Conv (32, 16, (3, 3)) + BN + ReLU		
AvgPool ((2, 2))		
Conv (32, 16, (3, 3)) + BN + ReLU		-
Conv (32, 16, (3, 3)) + BN + ReLU		
AvgPool ((2, 2))		
Conv (32, 32, (3, 3)) + BN + ReLU		-
Conv (32, 32, (3, 3)) + BN + ReLU		
AvgPool ((6, 6))		
FC (3072, 64)	TP AvgPool ((1,6))	TP AvgPool ((1,6))
ReLU + BN	CFC 64×$FC(8,1)$	CP AvgPool ((1,64))
FC (64, 2)	FC (64,64)	FC (8,4) + ReLU
-	ReLU + BN	FC (4,8) + Sigmoid
-	FC (64,4)	-

3. Experiments

3.1. Database

The performance of the proposed ADARL is evaluated on the the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [21], which is one of the most commonly used databases in SER. It records approximately 12 hours of audiovisual dialog data for ten actors, which are divided into five sessions with two speakers each. Each sentence is labeled with one of the 10 emotions. Similar to [26, 27, 28, 29, 30], we use four categories of improvised data, including neutral, happy, sad and angry.

3.2. Experimental Setup

We chose gender classification as our domain pathway task, and the gender label can be easily acquired in various databases. All convolution operations are initialized by Xavier. In our experiments, we train the model for 50 epochs using SGD optimizer with weight decay of 0.00004 and batch size of 32. The learning rate is initialized to 0.2, and it decreases to 70% after every epoch. In addition, we set $\lambda_1 = \lambda_2 = 1$ to maintain the balance of the two pathways and $\lambda_3 = 0.5$. When evaluating, as pointed out by recent research [31], the random folds method has a more optimistic estimation than the by-speaker folds method because of personal information leakage. We choose the common by-speaker folds method: Leave-One-Speaker-Out (LOSO) cross-validation, which is evaluated in turn with a new speaker. As for the evaluation metrics, we choose the widely used weighted accuracy (WA, the overall classification accuracy) and unweighted accuracy (UA, the averaged accuracy of each category).

3.3. Experiment Results and Analysis

3.3.1. Comparative Experiments

Table 2 summarises the performance of different methods on IEMOCAP. There are a lot of researches in SER in recent years

[26, 27, 28, 29, 30]. We can find that our ADARL outperforms other methods with WA of 73.02% and UA of 65.86%, which validates the performance of our approach.

Table 2: *The performance of different methods.*

Model	**WA**	**UA**
FCN + Attention Model [26] (2018)	70.4%	63.9%
Fixed-Length Model [27] (2018)	68.86%	57.45%
Variable-Length Model [27] (2018)	71.45%	64.22%
SegCNN-ELM [28] (2019)	62.34%	64.53%
audio-BRE [29] (2019)	64.60%	65.20%
Fusion Model [30] (2020)	72.34%	58.31%
Ours	**73.02%**	**65.86%**

3.3.2. Ablation Studies

To verify the validity of the domain-aware attention module, we adjust the weight of the loss $\mathcal{L}^D$ when carrying out the ablation study. When $\lambda_1 = 0$, it is equivalent to the backbone network with only emotion pathway.

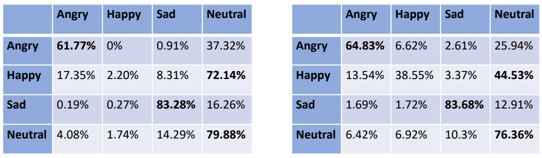
As Table 3 shows, the backbone network has achieved acceptable performance with WA of 69.96% and UA of 56.78%. The domain-invariant model with adversarial multitask training as Figure 1 (a) has achieved good results. With the domain-aware attention, the performance of the model is further improved significantly. This is because the domain-aware attention solves the interference problem caused by the domain diversity through adaptive embedding domain knowledge to the emotion feature space. In our observation, even during the testing phase, the domain's prediction accuracy is close to 100%.

Table 3: *The results of ablation studies.*

Model	**WA**	**UA**
Backbone	69.96%	56.78%
Domain-invariant Model	71.47%	61.91%
ADARL	**73.02%**	**65.86%**

The confusion matrices of backbone network and domain-aware network are shown as Figure 4. Each fold experiment has a confusion matrix, and the final confusion matrix is obtained by averaging them. We can find that the samples of ”angry”, ”sad” and ”neutral” achieve good performance. Especially for “sad” samples, only few samples are mispredicted, reaching an accuracy of 83.68% with domain-aware network. Intuitively, “sad” samples have obvious voice characteristics such as low pitch and slow speed. In addition, after the domain-aware attention, our performance has been significantly improved, especially for ”happy” samples. However, there are still some ”happy” samples were mispredicted as ”neutral”. We speculate that the reason is that many people don't react too much in voice when they are happy, but instead respond to facial features such as smiles. In the future, we will try multi-modal fusion to improve the “happy” recognition performance. Overall, these results demonstrate the comparable performance of our ADARL.

We further plot the 2-dimensional feature distribution before the last fully connected layer by t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm as Figure 5 shows. We can observe three obvious clusters of ”angry” (left), ”neutral” (middle) and ”sad” (right). Note that the samples of female and male are clustered in similar positions for each category with ADARL. We compute the L2 distance between male and female for each category center as Table 4 shows. It can be found that each distance between male and female is reduced with domain-aware attention, which means the problem of domain diversity has been improved.



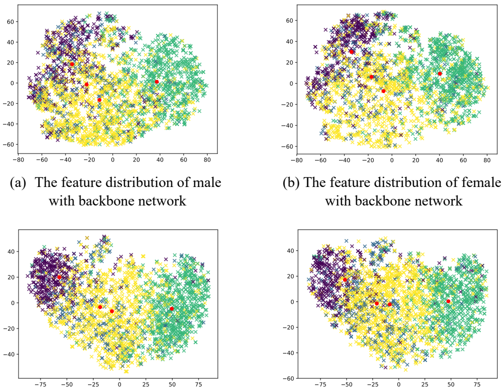
	Angry	Happy	Sad	Neutral
Angry	**61.77%**	0%	0.91%	37.32%
Happy	17.35%	2.20%	8.31%	**72.14%**
Sad	0.19%	0.27%	**83.28%**	16.26%
Neutral	4.08%	1.74%	14.29%	**79.88%**

(a) The confusion matrix of backbone model

	Angry	Happy	Sad	Neutral
Angry	**64.83%**	6.62%	2.61%	25.94%
Happy	13.54%	38.55%	3.37%	**44.53%**
Sad	1.69%	1.72%	**83.68%**	12.91%
Neutral	6.42%	6.92%	10.3%	**76.36%**

(b) The confusion matrix of ADARL

Figure 4: *Confusion matrices in ablation studies.*



(a) The feature distribution of male with backbone network

(b) The feature distribution of female with backbone network

(c) The feature distribution of male with ADARL

(d) The feature distribution of female with ADARL

Figure 5: *The feature distribution where purple, blue, green, and yellow represent angry, happy, sad, and neutral emotions, respectively. The red dots are the data centers in each category.*

Table 4: *The L2 distance between genders for each category.*

Model	Angry	Happy	Sad	Neutral
Backbone	11.58	9.16	8.61	10.14
ADARL	6.32	3.23	5.41	4.66

4. Conclusions

In our paper, we propose ADARL for speech emotion recognition. The approach is based on the multi-task learning framework and the domain knowledge is embedded to the emotion features through domain-aware attention, so that the model can adaptively learn the task aware features from the whole feature space. The results on IEMOCAP demonstrate the comparable performance of our ADARL to the state-of-the-art methods.

5. Acknowledgements

The work is supported in part by Key-Area Research and Development Program of Guangdong (2019B010154003); the National Natural Science Foundation of China (U1801262, U1636218, 61702192); Science and Technology Project of Zhongshan (2019AG024); Guangzhou Key Laboratory of Body Data Science (201605030011).

6. References

[1] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[2] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[3] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

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

[5] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 27–33.

[6] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2. IEEE, 2003, pp. II–1.

[7] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[8] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 5688–5691.

[9] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, 2008, pp. 597–600.

[10] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[11] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan, "Improvement on speech emotion recognition based on deep convolutional neural networks," in *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, 2018, pp. 13–18.

[12] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," *arXiv preprint arXiv:1806.02146*, 2018.

[13] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," *arXiv preprint arXiv:1806.06626*, 2018.

[14] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[15] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning." in *Interspeech*, 2017, pp. 1103–1107.

[16] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning," *arXiv preprint arXiv:1708.03920*, 2017.

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

[18] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 45–51.

[19] P. Liu, X. Qiu, and X. Huang, "Adversarial multi-task learning for text classification," *arXiv preprint arXiv:1704.05742*, 2017.

[20] G. Zeng, Y. Chen, B. Cui, and S. Yu, "Continual learning of context-dependent processing in neural networks," *Nature Machine Intelligence*, vol. 1, no. 8, pp. 364–372, 2019.

[21] 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.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[23] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[24] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[25] W. Fan, Z. He, X. Xing, B. Cai, and W. Lu, "Multi-modality depression detection via multi-scale temporal dilated cnns," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 73–80.

[26] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1771–1775.

[27] 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. Interspeech*, 2018, pp. 3683–3687.

[28] S. Mao, P. Ching, and T. Lee, "Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition," *Proc. Interspeech 2019*, pp. 1686–1690, 2019.

[29] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2822–2826.

[30] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7189–7193.

[31] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6484–6488.