INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning

Yuanchao Li[1], *Tianyu Zhao*[2], *Tatsuya Kawahara*[2]

[1]Honda Innovation Lab, Honda R&D Co., Ltd.
[2]Graduate School of Informatics, Kyoto University

liyuanchao.24x@kyoto-u.jp, {zhao, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

Accurately recognizing emotion from speech is a necessary yet challenging task due to the variability in speech and emotion. In this paper, we propose a speech emotion recognition (SER) method using end-to-end (~~E2E) multitask learning~~ with self attention to deal with several issues. First, we extract features directly from speech spectrogram instead of using traditional hand-crafted features to better represent emotion. Second, we adopt self attention mechanism to focus on the salient periods of emotion in speech utterances. Finally, giving consideration to mutual features between emotion and gender classification tasks, we ~~incorporate gender classification as an auxiliary task by~~ using multitask learning to share useful information with emotion classification task. Evaluation on IEMOCAP (a commonly used database for SER research) demonstrates that the proposed method outperforms the state-of-the-art methods and improves the overall accuracy by an absolute of 7.7% compared to the best existing result.

Index Terms: Speech emotion recognition (SER), spectrogram, end-to-end (E2E), attention mechanism, multitask learning

1. Introduction

With the rapid development of automatic speech recognition and natural language processing, a number of spoken dialogue systems (SDS) are being investigated to conduct specific tasks. Smartphone assistants and smart speakers are widely used for information retrieval in daily lives, and humanoid robots with a dialogue function are starting to serve in reception and elderly care. However, these systems usually lack the ability to deal with human emotion, which makes them perceived cold and robotic. To this end, the study of speech emotion recognition (SER) has attracted great attention in recent years.

SER basically consists of two necessary steps: feature extraction and emotion classification. The baseline feature sets for the INTERSPEECH 2009 Emotion Challenge [1], the INTERSPEECH 2013 computational paralinguistics challenge [2], and AVEC challenge [3] are commonly used. These feature sets include acoustic and prosodic features such as pitch, energy, Mel-frequency cepstral coefficients (MFCC) and so on, as well as their statistical functionals. However, directly learning the mapping from speech spectrogram has emerged as a trend in current works [4, 5, 6] , and this approach proved better in representing emotion. Traditional machine learning approaches, for example hidden Markov model, Gaussian mixture model, and support vector machines are widely adopted for classification using extracted features in previous works [7, 8, 9, 10]. However, a series of neural networks are replacing traditional machine learning approaches, such as convolutional neural networks (CNN) and recurrent neural networks (RNN) [11, 12], thanks to remarkable success of deep neural technologies in the last decade.

Despite these progresses, SER is still a challenging task because of the variability in speech and emotion. In this paper, we propose a method using end-to-end (E2E) multitask learning with self attention aiming to resolve several issues which affect the recognition accuracy. First, we extract features directly from spectrogram and then train the model using a combination of CNN and bidirectional long short-term memory (BLSTM) model in an E2E manner. Traditional hand-crafted features lack the ability to represent emotion comprehensively resulting in that there is no single feature set that well suits every speech corpus. Second, because emotion is usually expressed by salient emotion-related periods instead of the whole utterance, we adopt self attention mechanism to focus on these periods for utilizing relevant features. Finally, inspired by that emotion and gender share mutual features in classification task, we incorporate gender classification as an auxiliary task. By using multitask learning, the model could better learn the differences between signal patterns of male and female speech to improve the recognition accuracy. The major contributions of this work are summarized as follows.

1) Classifying emotion using ~~spectrogram based~~ self-attentional CNN-BLSTM model in an E2E manner.

2) Combining emotion classification and gender classification using multitask learning.

3) Achieving ~~an absolute~~ increase of overall accuracy by 7.7% compared to the best existing result.

The rest of this paper is organized as follows. We present related works on SER in Section 2. Next, we describe the IEMOCAP database in Section 3 and the proposed method in Section 4. The experiments along with the evaluation results are described in Section 5. In Section 6, we conclude with a brief summary and mention of the future work.

2. Related Works

To automate SER, researchers have applied various acoustic and prosodic features in last few decades. Traditionally, a number of low-level descriptors (LLD) built upon prior knowledge have been hand-crafted to represent emotion. Nevertheless, recent studies have shown that using spectrogram information directly for SER outperforms using hand-crafted LLDs. [5] used spectrogram with deep CNNs, and [13] used spectrogram in Mel-scale with a combination of CNN and LSTM. In this work, we follow this path.

Machine translation and speech recognition have achieved great success by using attention mechanism [14, 15]. [16] presents the effectiveness of multi-headed self attention. Emotions are usually labeled at the utterance-level while expressed by salient emotional parts of the utterance. Moreover, speech

http://dx.doi.org/10.21437/Interspeech.2019-2594

utterances in most available databases contain silence periods. Hence, attention mechanism is important to select relevant features for SER. [17] used local attention and achieved an increase in SER task. In this work, we adopt self attention in our architecture.

Multitask learning recently rose as an approach to improving SER by learning from auxiliary tasks. [18] jointly predicted valence, arousal and dominance by combining shared layers and attribute-dependent layers. However, the required information for these predictions highly overlap with each other, which makes it difficult for the model to learn extra features. [19, 20] have proved that gender classification share mutual features such as pitch and MFCC with emotion classification. Thus, we use gender classification as an auxiliary task. Considering the differences between signal patterns of male and female speech, gender classification may help identify the differences to increase SER accuracy.

3. Database Description

We use the IEMOCAP [21], which is a widely used database in SER research comprising of scripted and improvised multimodal interactions of five dyadic sessions with 10 subjects (5 male and 5 female actors). The database consists of approximately 12 hours of speech and is labeled by three annotators with the categorical emotion labels chosen from the set: ANGRY, DISGUSTED, EXCITED, FEARFUL, FRUSTRATED, HAPPY, NEUTRAL, SAD, SURPRISED, and OTHER. We first combine EXCITED and HAPPY into HAPPY in accordance with past works, and use four categories of ANGRY, HAPPY, NEUTRAL, and SAD. The total number of the utterances used in this work is 5531.

4. Proposed Method

4.1. Spectrogram Extraction

First, the maximal length of the utterances were set to 7.5s (mean duration plus standard deviation of all utterances). Longer utterances were cut at 7.5s, and shorter ones were padded with zeros. Next, Hanning windows with length of 800 were applied to the audio signal. The sampling rate was set at 16000Hz. For each frame, a short term Fourier transform (STFT) of length 800 with hop length 400 was calculated. The calculated spectrogram was mapped to Mel-scale in order to mimic the non-linear human ear perception of sound. Finally, the spectrogram with deltas and delta-deltas for better capturing personalized features [22], were extracted as the input to the neural network. We have also tried several different lengths of sampling window but the results had no clear differences.

4.2. Self-Attentional CNN-BLSTM

We denote the spectrogram of a speech segment as $X = \{x_1, x_2, \cdots, x_L\}$, where $x_i \in \mathbb{R}^{d_{\text{spec}}}$, L is the temporal length of the spectrogram, and d_{spec} is the dimension of a spectrogram feature vector.

We encode the spectrogram X as a fixed-length vector z, and conduct classifications on z. To efficiently encode salient features of X, we propose to use a self-attentional CNN-BLSTM network. As shown in Figure 1, the network mainly consists of 3 components, namely a 2-layer convolutional neural network (CNN), a 2-layer bidirectional long short-term memory (BLSTM) network, and a self attention network. The three components serve for different purposes in the encoding process.

4.2.1. CNN and Pooling Layer

The CNN integrates local contexts by applying d_{cnn} convolutions over X with a stride of 1 and a window size of n_{cw}, and produces a sequence of vectors $H^{\text{cnn}} = \{h_1^{\text{cnn}}, h_2^{\text{cnn}}, \cdots, h_L^{\text{cnn}}\}$, where $h_i^{\text{cnn}} \in \mathbb{R}^{d_{\text{cnn}}}$. The j-th element in h_i^{cnn} is the inner product of the j-th filter and the i-th window in X.

$$h_{i,j}^{\text{cnn}} = f(b_j^{\text{cnn}} + \langle W_j^{\text{cnn}}, win(X, i, n_{\text{cw}})\rangle_F) \quad (1)$$

$$\langle A, B\rangle_F = \sum_i \sum_j A_{i,j} B_{i,j} \quad (2)$$

$$win(X, i, n_{\text{cw}}) = [x_{i-\frac{n_{\text{cw}}-1}{2}}; \cdots; x_i; \cdots; x_{i+\frac{n_{\text{cw}}-1}{2}}], \quad (3)$$

where $\langle \cdot, \cdot \rangle_F$ denotes the matrix inner product operation, and $win(X, i, n_{\text{cw}})$ means the i-th window in X with a window size of n_{cw}. We choose $ReLU$ [23] as the activation function $f(\cdot)$. $W_j^{\text{cnn}} \in \mathbb{R}^{n_{\text{cw}} \times d_{\text{spec}}}$ and $b_j^{\text{cnn}} \in \mathbb{R}$ are trainable parameters of the j-th CNN filter. Notice that X is left- and right- padded before the convolution operation by repeating x_1 and x_L, respectively.

Then a temporal *MaxPooling* operation is applied to every n_{pw} CNN outputs. It keeps salient information in the inputs while reducing the length of the sequence, which makes computation in subsequent BLSTM networks faster. The pooled outputs are $H^{\text{pool}} = \{h_1^{\text{pool}}, \cdots, h_{L/n_{\text{pw}}}^{\text{pool}}\}$, where $h_i^{\text{pool}} \in \mathbb{R}^{d_{\text{cnn}}}$.

$$h_i^{\text{pool}} = MaxPooling(win(H^{\text{cnn}}, i \times n_{\text{pw}} - 1, n_{\text{pw}})). \quad (4)$$

4.2.2. BLSTM Layer

The bidirectional LSTM [24] network encodes global contexts by updating its hidden states recurrently. It takes as input H^{pool} and outputs a sequence of hidden states $H^{\text{blstm}} = \{h_1^{\text{blstm}}, \cdots, h_{L/n_{\text{pw}}}^{\text{blstm}}\}$, where $h_i^{\text{blstm}} \in \mathbb{R}^{2d_{\text{lstm}}}$ is the concatenation of the i-th forward hidden state and the i-th backward hidden state.

4.2.3. Self Attention Layer

Following the BLSTM layer, a structured self attention network [9] aggregates information from the BLSTM hidden states H^{blstm} and produces a fixed-length vector z as the encoding of the speech segment.

Given hidden states H^{blstm} as input, the network computes a vector of attentional weights a. And h^{attn} is the weighted sum of the hidden states.

$$a = softmax(w_{s2} \tanh(W_{s1} {H^{\text{blstm}}}^T)) \quad (5)$$

$$h^{\text{attn}} = a H^{\text{blstm}}, \quad (6)$$

where $W_{s1} \in \mathbb{R}^{d_{\text{attn}} \times 2d_{\text{lstm}}}$ and $w_{s2} \in \mathbb{R}^{d_{\text{attn}}}$ are trainable parameters.

The combination yielded by a may only capture one specific aspect of the input information. In order to obtain an overall representation, we can compute multiple combinations of H^{blstm} by using n_{attn} different w_{s2}s. Then we concatenate all the weighted sums to get the final encoding vector $z \in \mathbb{R}^{2n_{\text{attn}}d_{\text{lstm}}}$.

$$a_j = softmax(w_{s2,j} \tanh(W_{s1} {H^{\text{blstm}}}^T)) \quad (7)$$

$$h_j^{\text{attn}} = a_j H^{\text{blstm}} \quad (8)$$

$$z = h_1^{\text{attn}} \oplus h_2^{\text{attn}} \oplus \cdots \oplus h_{n_{\text{attn}}}^{\text{attn}}. \quad (9)$$

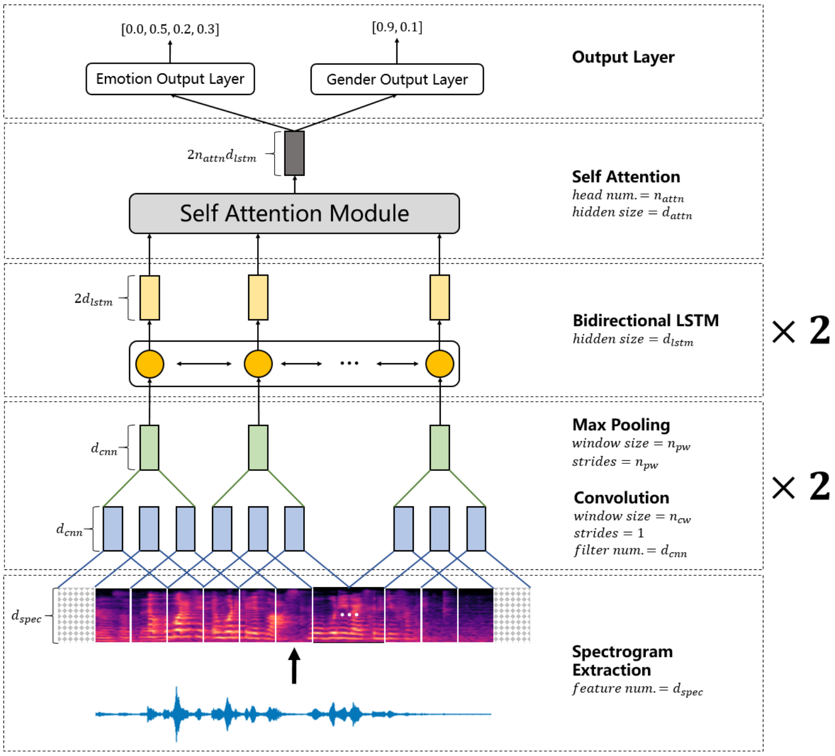

Figure 1: *Overview of the model architecture.*

4.3. Output Layers with Multitask Learning

Finally, following the self-attentional CNN-BLSTM model, two output layers generate the probability distributions over emotions and genders, respectively. Gender classification is incorporated to consider the relationship between the two tasks and better classify emotion in a multitask learning manner. The model is optimized by the following objective function.

$$\mathcal{L} = \alpha \times \mathcal{L}_{\text{emotion}} + \beta \times \mathcal{L}_{\text{gender}}, \tag{10}$$

where $\mathcal{L}_{\text{emotion}}$ and $\mathcal{L}_{\text{gender}}$ are the losses for emotion classification and gender classification, respectively. α and β represent the weights for the two tasks. We choose cross-entropy as the loss function for both tasks. We tried several weights in order for the network to focus on the main task, and found that setting both weights to 1.0 yielded the best accuracy of emotion classification.

The two output layers are fully-connected layers with hidden units of the number of emotion categories (four) and gender categories (two), respectively. We choose *softmax* as the activation function for both layers.

5. Experiments and Evaluation Results

5.1. Implementation

The self-attentional CNN-BLSTM model was implemented in Keras, and optimized using the Adam method [25] with a learning rate of 10^{-4}, a decay rate of 10^{-6} and a batch size of 40. Gradients were clipped within $[-1.0, 1.0]$. We used 100 epochs for training and saved the best one as the final model. We applied a dropout after every BLSTM layer with 0.5 dropout probability. The dimensions mentioned in Section 4 are summarized in Table 1.

Table 1: *Dimension details.*

Notation	Meaning	Value
d_{spec}	Number of spectrogram features	384
d_{cnn}	Number of convolution filters	64
d_{lstm}	Number of LSTM hidden units	60
d_{attn}	Hidden size of a attention head	512
n_{cw}	Size of a convolution window	3
n_{pw}	Size of a pooling window	3
n_{attn}	Number of attention heads	8

5.2. Experimental Settings

We used leave-one-session-out cross-validation (four sessions for training and the remaining one for testing) to train the model in accordance with prior works. We evaluated the performance using both weighted accuracy (WA) and unweighted accuracy (UA). WA is the overall accuracy of the entire test data, and UA is the average accuracy of each emotion category. We compared the result of the proposed self-attentional CNN-BLSTM model with the state-of-the-art result as well as other results reported by prior works on IEMOCAP database. The comparison results are shown in Table 2.

5.3. Evaluation Results Compared with Prior Works

The upper part of Table 2 shows results reported in prior works. The DNN-ELM model in [26] used deep neural networks (DNNs) to calculate segment-level probability distribu-

Table 2: *Comparison results.*

Method	**WA**	**UA**	**Year**
DNN-ELM [26]	54.3%	48.2%	2014
BLSTM-ELM [12]	62.8%	63.9%	2015
AE-BLSTM [27]	50.5%	51.9%	2016
CNN-BLSTM [13]	68.8%	59.4%	2017
Multichannel CNN [6]	73.9%	68.5%	2018
Ours			
Full model	**81.6%**	**82.8%**	
− Self attention	55.3%	51.1%	
− Multitask learning	70.5%	72.6%	

Table 3: *Confusion matrix of the full model predictions.*

Emotion labels	Prediction results			
	ANGRY	HAPPY	NEUTRAL	SAD
ANGRY	**939**	101	63	0
HAPPY	271	**1283**	71	11
NEUTRAL	2	107	**1300**	299
SAD	8	25	57	**994**

tion over emotions. The segment-level distributions were integrated as sentence-level features and processed by an extreme learning machine (ELM). Replacing the DNNs with BLSTMs improved the performance as reported in [12]. In [27], the authors used spectrogram features extracted from both speech and glottal flow signals. A stacked denoising autoencoder (AE) encodes the spectrogram, and a BLSTM predicts an emotion label for each input step in a sequence labeling manner. The final emotion label is decided in a voting scheme. A combination of CNN and BLSTM was proposed in [13]. The previous state-of-the-art result was achieved by a multichannel CNN model proposed in [6], which used only CNNs to encode phoneme and spectrogram features. Compared to the previous works, the proposed self-attentional CNN-BLSTM model outperforms the best reported results by a large margin. We improve WA from 73.9% to 81.6% (7.7% absolute improvement and 10.4% relative improvement), and UA from 68.5% to 82.8% (14.3% absolute improvement and 20.9% relative improvement). The confusion matrix showing detailed classification results is presented in Table 3.

5.4. Ablation Study

We also conducted ablation study to confirm the usefulness of the self attention component and multitask learning with gender classification. The lower part of Table 2 reports the results of (1) the full proposed model, (2) the model without self attention component, and (3) the model trained without gender classification loss. As the WA and UA measurements suggest, both techniques substantially contribute to the accuracies of the full model. Removing the self attention component resulted in around 30% absolute accuracy reduction. And training the model solely using SER loss resulted in around 10% absolute accuracy reduction.

5.5. Analysis of Gender Classification

While we reached new state-of-the-art accuracies on the emotion classification task, the accuracy of gender classification is only around 75%, which is not as good as common gender classification results. This condition may result from two reasons. First, because IEMOCAP is especially collected for emotion recognition research, it may not be a suitable database for gender recognition. Second, gender recognition may require more hand-crafted features such as pitch, energy, and MFCC, since prior works achieved significant results using only these features [28, 29, 30]. On the other hand, although the proposed method did not improve gender classification accuracy to the state-of-the-art level, incorporating this auxiliary task enabled emotion classification to avoid confusion differentiating male and female speech. By analyzing audio files according to emotion classification results, we found that emotion classification sometimes failed because of the lack of gender information without using multitask learning. For example, several female speech utterances with sad emotion, as well as male speech utterances with extreme anger were falsely classified as neutral emotion. This phenomenon is plausible because female speech with sadness are usually low-frequency, and male speech with extreme anger are highly aroused. The emotions of these speech are hard to determine without knowing the gender [20]. However, this issue was resolved by sharing gender information with emotion classification using multitask learning in this work.

6. Conclusions

In this paper, we propose a self-attentional CNN-BLSTM model to train the SER model with multitask learning in an E2E manner, operating on the speech spectrogram. The evaluation on IEMOCAP database demonstrates that the proposed method achieves 7.7% and 14.3% absolute improvements of WA and UA compared to the existing best result, respectively. Our method could be used for recognizing different kinds of paralinguistic information from speech at the same time in various scenarios including call center, health care, and human-robot interaction. We aim to use raw audio and incorporate more tasks, such as age recognition and speaker identification in our future work.

7. Acknowledgements

The authors would like to thank Dr. Zhichao Peng from Tianjin University for valuable discussions.

8. References

[1] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[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. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 3–8.

[4] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[5] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional

neural network," in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[6] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," *Proc. Interspeech 2018*, pp. 3688–3692, 2018.

[7] H. Hu, M.-X. Xu, and W. Wu, "Gmm supervector based svm with spectral features for speech emotion recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–413.

[8] 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.

[9] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *ICLR 2017, The 5th International Conference on Learning Representations*, 2017.

[10] Y. Li, C. T. Ishi, N. Ward, K. Inoue, S. Nakamura, K. Takanashi, and T. Kawahara, "Emotion recognition by combining prosody and sentiment analysis for expressing reactive emotion by humanoid robot," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1356–1359.

[11] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 801–804.

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

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

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[15] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[17] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

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

[19] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *2004 12th European Signal Processing Conference*. IEEE, 2004, pp. 341–344.

[20] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiaion." in *LREC*, 2006, pp. 1123–1126.

[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] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML 2010, The 27th International Conference on Machine Learning*, 2010, pp. 807–814.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[27] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition." in *Interspeech*, 2016, pp. 3603–3607.

[28] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *The journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[29] M. Sedaaghi, "A comparative study of gender and age classification in speech signals," *Iranian Journal of Electrical and Electronic Engineering*, vol. 5, no. 1, pp. 1–12, 2009.

[30] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarrone, "Gender-driven emotion recognition through speech signals for ambient intelligence applications," *IEEE transactions on Emerging topics in computing*, vol. 1, no. 2, pp. 244–257, 2013.