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

End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model

Han Feng, Sei Ueno, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
{feng, ueno, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

In this paper, we propose speech emotion recognition (SER) combined with an acoustic-to-word automatic speech recognition (ASR) model. While acoustic prosodic features are primarily used for SER, textual features are also useful but are error-prone, especially in emotional speech. To solve this problem, we integrate ASR model and SER model in an end-to-end manner. This is done by using an acoustic-to-word model. Specifically, we utilize the states of the decoder in the ASR model with the acoustic features and input them into the SER model. On top of a recurrent network to learn features from this input, we adopt a self-attention mechanism to focus on important feature frames. Finally, we finetune the ASR model on the new dataset using a multi-task learning method to jointly optimize ASR with the SER task. Our model has achieved a 68.63% weighted accuracy (WA) and 69.67% unweighted accuracy (UA) on the IEMOCAP database, which is state-of-the-art performance.

Index Terms: speech emotion recognition, acoustic-to-word speech recognition, end-to-end, self-attention mechanism, multi-task learning

1. Introduction

Speech communication between humans and machines is becoming more common in our daily lives. With the advancement of automatic speech recognition (ASR) technology, there is a growing demand for speech emotion recognition (SER) as well. By using emotion detection, machines can communicate and interact with humans more appropriately and naturally. Therefore, SER has become an important part in human-machine interaction [1, 2].

In SER, feature extraction from speech is an important issue. Traditional methods, such as OpenSmile [3], used some statistical functions to extract relevant emotional features from prosodic features. With the rise of deep learning, which can learn feature extraction from source data, the recurrent neural network (RNN), convolutional neural network (CNN) [4], and recently self-attention mechanism [5] have been widely applied to the SER task.

In [6], a deep neural network (DNN) with an extreme learning machine (ELM) is used to learn SER tasks. In [7], a method for feature pooling with local attention is used. In [8], convolutional recurrent neutral network (CRNN) is used to capture emotional information from speech. In [9], a spectrogram in mel-scale is used as input features for CNN and LSTM [10] with a structured self-attention mechanism [11]. In [12], self-attention and global windowing systems are used to predict emotion labels. Besides, some other researchers used not only speech data but also transcripts [13, 14, 15]. Generally, these SER models using speech and transcript data perform better than speech-only models, and achieved accuracy over 70%. In [16], an RNN is used to train the speech and text data individually, and the model predicts the label based on the concatenated vector. In [15], bi-directional LSTM (Bi-LSTM) with multi-hop attention is used to extract features from speech and text data.

In practical situations, however, input speech needs to be automatically transcribed to get the textual information. However, ASR of emotional speech is very challenging and results in high error rates. To make SER robust against ASR errors, in this paper, we propose an SER model combined with an acoustic-to-word ASR model [17]. The model is enhanced with the self-attention mechanism, and then the multi-task learning method is applied to finetune the ASR model. We propose an ASR feature, which is the hidden state of the decoder in the ASR model, to replace the textual data in the SER model. As a result, our model is a speech-only ASR-SER model but has a performance that is close to that of the models using speech and text data. Combined with the self-attention mechanism to deal with the acoustic features and the ASR features separately, our ASR-SER model is much better. By using multi-task learning, the ASR model can be finetuned, and the SER models can get more accurate ASR features to improve the recognition performance. The results reach 68.63% WA and 69.67% UA.

We describe the ASR model in Section 2, some baseline models in Section 3, and our proposed method in Section 4. Section 5 presents the system configuration and evaluation results. Finally, we give conclusions in Section 6.

2. Baseline models

To realize high-performance SER, we exploit the information from audio and transcripts. To better utilize the sequence information, we adopt a recurrent neural network. LSTM with multi-head self-attention is a sophisticated structure for feature learning for SER. We prepared three baseline models, which have a very similar structure and extract information from audio, transcripts, or both.

2.1. Audio-based, text-based, and combined model

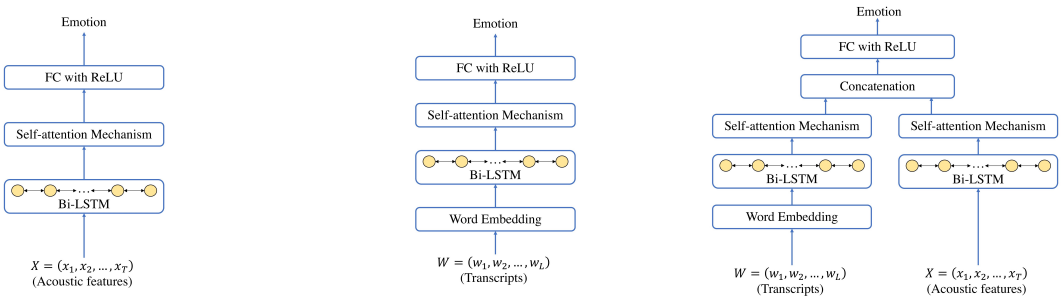
Figure 1 shows our three baseline models of SER. Figure 1(a) is the audio-based SER model with acoustic features as inputs. The Bi-LSTM network encodes the input acoustic features X and outputs a sequence of hidden states $H_X = (h_{1,X}, ..., h_{T,X})$.

$$h_{t,X} = Recurrency(h_{t-1,X}, x_t) \tag{1}$$

Then, H_X is fed into the self-attention mechanism, which we will describe in the next section. Finally, the output vector of the self-attention layer is fed into a fully connected layer with the ReLU activation function.

Similarly, Figure 1(b) shows the text-based SER model with textual data as inputs. Textual data is first fed into the word

http://dx.doi.org/10.21437/Interspeech.2020-1180



(a) audio-based model (b) text-based model (c) combined model

Figure 1: *Architecture of three baseline models for SER*

embedding layer, and then the remaining part of the model is the same as the audio-based model. $H_W = (h_{1,W}, ..., h_{L,W})$ is the output of the Bi-LSTM network in the text-based model,

$$h_{l,W} = Recurrency(h_{l-1,W}, w_l) \tag{2}$$

Figure 1(c) shows the combined SER model with both acoustic features and textual data as inputs. Based on the above two baseline models, we concatenate the two output vectors from the self-attention layer of the textual data and the acoustic features, and then feed the concatenated vector into a fully connected layer with the ReLU activation function.

2.2. Multi-head self-attention model

Li et al. [9] demonstrated that the self-attention structure is effective for SER. We adopt a structured self-attention network [11] following the Bi-LSTM layer to extract features from hidden states H and output fixed-length vector M. We set W_1 and w_2 as trainable parameters, and the attention mechanism outputs a vector of weights a_i:

$$a_i = softmax(w_{2,i} tanh(W_1 H^T)) \tag{3}$$

In this work, we use a multi-head structure. We concatenate all of the weighted sums of hidden states to get fixed-length vector M as output.

$$M = Concat(head_1, ..., head_n) \tag{4}$$

$$head_i = a_i H \tag{5}$$

3. Acoustic-to-word ASR model

End-to-end speech recognition is a sequence-to-sequence problem. Generally, there are two main methods: connectionist temporal classification (CTC) [18] and the attention-based encoder-decoder model [19, 20, 21]. In this study, we use the hidden state of the ASR decoder as input for the SER part, and the attention-based encoder-decoder model is more suitable than the CTC model. The acoustic-to-word ASR model can provide the representation of word-level recognition most related to the content of speech. Thus, we choose the acoustic-to-word model to pre-train the ASR model.

We denote a length T sequence of input acoustic features as $X = (x_1, x_2, ..., x_T)$, a length L sequence of the output word labels as $W = (w_1, w_2, ..., w_L)$. The encoder transforms the acoustic features into context vector $H = (h_1, h_2, ..., h_T)$. The decoder transforms the context vector into the target words. At the l-th decoding step, hidden state s_l of the decoder is:

$$s_l = Recurrency(s_{l-1}, g_l, w_l) \tag{6}$$

where g_l denotes the context vector, and w_l means the current predicted label, which will be utilized in the recurrent network. The formula that calculates the context vector is as follows:

$$g_l = \sum_{t=1}^{T} a_{l,t} h_t \tag{7}$$

where a is the location-based attention, which is formulated as follows:

$$f_l = F * a_{l-1} \tag{8}$$

$$e_{l,t} = z^T tanh(Z s_{l-1} + V h_t + U f_l + b) \tag{9}$$

$$a_{l,t} = softmax(e_{l,t}) \tag{10}$$

where F is the parameters of a 1-dimensional convolution, z^T,Z,V,U are the parameters of fully connected layer. Based on the formula above, we can predict the next label w_l:

$$w_l \sim Generate(g_l, s_l) \tag{11}$$

4. End-to-end ASR and SER

4.1. Integration of acoustic-to-word ASR model to SER

The performance of SER will be better when we utilize the information from both the audio and transcripts. We use the pre-trained ASR model to replace the text input in the SER combined model in Figure 1(c). Instead of using word embedding as the input of the SER model, we use the hidden state of the ASR decoder s_l, which we call ASR feature. In the acoustic-to-word ASR, they are fed into the last fully connected layer to calculate the probability of each word. In our model, we use the ASR features as the input of the Bi-LSTM layers to retrieve the textual information. The ASR features can be more robust than the text output of ASR, which is error-prone, and they have a close representation to the ground-truth text. The formula of the Bi-LSTM encoder is changed as follows:

$$h_{l,W} = Recurrency(h_{l-1,W}, s_l) \tag{12}$$

As shown in Figure 2, the framework of our proposed model includes two Bi-LSTMs, a self-attention mechanism, and several fully connected layers with the function of concatenating ASR features with the acoustic features. The ReLU activation function and softmax nodes are used for the final decision.

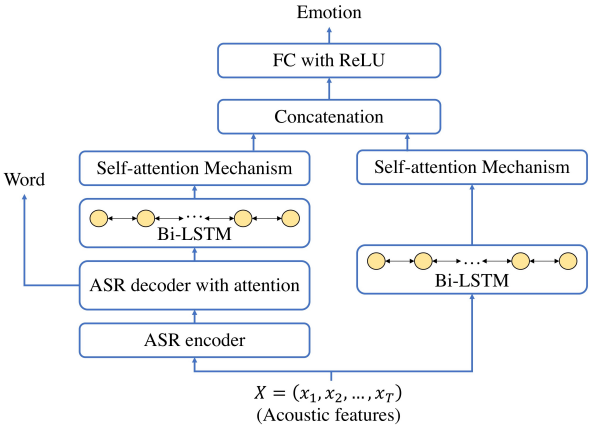


Figure 2: *Our proposed end-to-end speech emotion recognition combined with acoustic-to-word speech recognition model*

The process can be done step by step. First, the acoustic-to-word ASR model is pre-trained. Then, we combine the ASR model with the SER model. The SER model gets input from acoustic features and ASR features, and the Bi-LSTM and self-attention mechanism process these two kinds of features separately. Finally, after concatenating these two output vectors from the self-attention mechanism, we use the softmax output layer to predict the probability of emotion categories. The entire network is trained with the cross-entropy criterion.

4.2. Joint learning of ASR and SER

The acoustic-to-word ASR model can be finetuned on the matched dataset to adjust to the new environment, and we adopt multi-task learning to optimize the joint loss:

$$L = \lambda L_{ASR} + (1 - \lambda) L_{SER} \tag{13}$$

We denote L_{ASR} and L_{SER} as the losses for speech recognition and emotion classification, and λ represents the weights of the first task. We use the cross entropy as a loss function to calculate the losses for L_{ASR} and L_{SER}. When λ is close to 0, the model will pay more attention to SER.

5. Experiments

5.1. Database

We used two databases in our experiments: the interactive emotional dyadic motion capture database (IEMOCAP) [22] and Librispeech database [23]. IEMOCAP contains approximately 12 hours of speech, which consists of improvised and scripted scenarios. There are 10 actors (5 males and 5 females) to perform 5 dyadic sessions, with 10 emotions (angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other), which have been evaluated by at least three different annotators. Librispeech is a corpus of approximately 1000 hours of speech sampled at 16 kHz, whose training data contains three subsets, with an approximate size of 100, 360, and 500 hours. We utilize the total 960 hours of training data for pre-training the acoustic-to-word ASR model.

5.2. Data preprocessing

For the IEMOCAP database, to compare our work with that of others, we combine happy and excited emotions into the happy category, so we have 4 categories of happy, sad, neutral, and angry. Some utterances longer than 20 seconds are too long for multi-task learning, thus we do not use them. Therefore, we use a total of 5515 utterances, and the numbers of emotional utterances of happy, sad, neutral, and angry are 1633,1074,1707, and 1101 respectively. We also prepare another dataset that consists of improvised speech data with happy, sad, neutral, and angry categories, a total of 2274 utterances. During the experiments, we use 5-fold cross-validation to train our model in keeping with prior work.

We use a 120-dimensional feature vector of 40-channel log mel-scale filter bank (lmfb) outputs as input features. For features used in ASR, we apply non-overlapping frame stacking to them. For features used in SER, we do not apply frame stacking to the acoustic features, which constitute a 40-dimensional feature vector of 40-channel lmfb. Besides, following [9] we set the maximal length of the utterance to 7.5 seconds. Long utterances (over 7.5 seconds) are cut at 7.5 seconds from the start, and short utterances are padded with zeros.

5.3. System configuration

The encoder of the acoustic-to-word ASR model has five layers of Bi-LSTM with 320 cells. The decoder consists of a location-based attention mechanism, a hidden layer with tanh activation function, one layer of LSTM with 320 cells, and an output with softmax function. During the decoding step, we use a beam search method with a beam width of 4.

In the ASR-SER model, we use the ASR features and acoustic features as input. Our SER model consists of 2 layers of Bi-LSTM with 256 cells, a self-attention mechanism with 8 heads and 512 nodes to retrieve information from the ASR features and acoustic features separately, 1 fully connected layer with 2048 nodes that uses concatenated vector from the self-attention layers, a ReLU activation layer, and an output layer with softmax function.

We adopt the Adam method to optimize the parameters, whose learning rate is 10^{-4} and weight decay is 10^{-5}. The gradients are clipped with a threshold of 1.0. The dropout of each Bi-LSTM is 0.2. The batch size is set to 20, and the maximum number of epochs is 60. We set λ to 0.2 in the multi-task learning, which means we pay more attention to the SER task.

The parameter settings in the three baseline SER models are the same as that in the ASR-SER model, except that the word embedding layer in the text-based model and combined SER model has 300 dimensions.

5.4. Results

Table 1 shows the results of our baseline SER models introduced in Section 3. We used a 5-fold cross-validation method to train the SER model, which is speaker-independent. Besides, we have used a pipeline method for the text-based model and combined SER model, which used the output text of the pre-trained ASR model instead of the ground-truth text. We used the word error rate (WER) to evaluate the ASR performance and used the weighted accuracy (WA) and unweighted accuracy (UA) to evaluate the performance of SER.

The WER of the ASR model on the Librispeech test set is 14.9%. When we apply the ASR model directly to the IEMOCAP dataset, the WER is 40.7% because the IEMOCAP dataset is spontaneous, emotional, and mismatched to Librispeech .

The audio-based SER model achieves a WA of 55.7%, text-based SER model achieves a WA of 63.7%, and combined SER model achieves a WA of 68.9%. The combined SER model

Table 1: *SER model results comparison*

Model	WA	UA	WER
Audio-based	55.7%	57.0%	-
Text-based	63.7%	63.6%	-
Text-based (ASR transcripts)	51.0%	50.7%	40.7%
Combined	**68.9%**	69.4%	-
Combined (ASR transcripts)	60.1%	60.7%	40.7%
ASR-SER model	68.6%	**69.7%**	35.7%

Table 2: *ASR-SER model results comparison*

Model	WA	UA	WER
Mirsamadi et al.[7] (2017)	63.5%	58.8%	-
Luo et al.[8] (2018)	60.4%	64.0%	-
Dai et al.[24] (2019)	65.4%	66.9%	-
Tarantino et al.[12] (2019)	68.1%	63.8%	-
Ours			
ASR-SER model	**68.6%**	**69.7%**	35.7%

shows a significant performance gain for emotion classification of 13.2% and 5.1% with respect to the audio-based and text-based SER model, respectively. The combined model apparently extracts more information from the input data. When evaluated on ASR transcripts, however, the accuracies of text-based SER model and combined SER model are significantly degraded to 51.0% and 60.1%. The low performance of the ASR model may lead to a lower performance of the SER model when using the ASR transcripts directly. On the other hand, using the proposed ASR features instead of the ASR transcripts as shown in the last row is effective.

Our ASR-SER model achieves a WA of 68.6%, and a UA of 69.7%, while the WER decreases from 40.7% to 35.7%. This means the finetuning is successful, and with more data, the WER will be further reduced. The ASR-SER model performs much better than the combined SER model using ASR transcripts, and the results are very close to those of the combined model using the ground-truth text. The ASR feature is robust against ASR errors, which has a close representation to the feature of the ground-truth label, and replaces the role of word embedding to some extent. Table 2 shows a comparison with prior works. [7] proposed a method for feature pooling with local attention. [8] adopted CRNN structure to learn the SER task. [24] was a method that learned discriminative features from variable length spectrograms by combining the softmax cross-entropy loss with the center loss. [12] used the self-attention and global windowing methods.

Table 3 shows the confusion matrix of the results of our ASR-SER models on the IEMOCAP dataset. The classification of the neutral emotion has a low performance of around 59.1%; however, all of the other three emotions have a WA of over 70%. It is easy to misclassify the neutral emotion to the happy emotion, and it is also easy to misclassify the happy emotion to the neutral emotion.

In Table 4, we also show the results for the improvised dataset, which consists of improvised utterances only of the IEMOCAP, and the experiment also used four categories: happy, neutral, sad, and angry. Note that the performance of our ASR-SER model is better than that of major prievious works [25, 26].

Table 3: *Confusion matrix of results of our ASR-SER model (the dataset has a total of 5515 utterances)*

Ground Truth	Prediction			
	Happy	Sad	Neutral	Angry
Happy	**1159**	81	280	113
Sad	83	**752**	197	42
Neutral	373	207	**1009**	118
Angry	105	18	113	**865**

Table 4: *Improvised utterances only results comparison*

Model	WA	UA	WER
Lee et al.[25] (2015)	62.8%	**63.9%**	-
Satt et al.[26] (2017)	68.8%	59.4%	-
Ours			
ASR-SER model	**69.7%**	63.1%	54.5%

Table 5: *Speaker-dependent experiment of SER models*

Model	WA	UA	WER
Audio-based	66.8%	67.0%	-
Text-based	66.5%	66.6%	-
Text-based (ASR transcripts)	52.4%	52.1%	40.7%
Combined	73.2%	73.6%	-
Combined (ASR transcripts)	65.6%	66.0%	40.7%
ASR-SER model	**75.5%**	**76.4%**	29.8%

Furthermore, we also performed an experiment on speaker-dependent setting. The results are shown in Table 5. In this case, the training dataset is chosen randomly from the whole of the database in accordance with the mean proportion of each emotion category. These results have a higher performance in both WA, UA, and WER, and the proposed ASR-SER model achieves the best performance.

Through Table 1 and Table 5, note that the WA and UA in our baseline models and those in the ASR-SER model are very close, which means our models are more stable than others, even though the number of emotional utterances in each category is unbalanced.

6. Conclusion

In this paper, we have proposed an end-to-end speech emotion recognition (SER) model combined with an acoustic-to-word ASR model. Our model benefits from the ASR feature and robustly works with speech data only. The results show our method achieves a performance that is better than that of the conventional pipeline method and is close to the ground-truth performance. It shows better performance than other prior works on the IEMOCAP dataset. The code is available.[1]

7. References

[1] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, pp. 572–587, 2011.

[1] https://github.com/Kyoto-University-Speech-and-Audio/feng-asr-ser

[2] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, pp. 90 – 99, 2018.

[3] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," *ACM Multimedia 2001*, pp. 1459–1462, 2010.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[6] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH*, 2014.

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

[8] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *INTERSPEECH*, 2018.

[9] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *INTERSPEECH 2019*, 2019.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," 1997.

[11] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *ArXiv*, vol. abs/1703.03130, 2017.

[12] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *INTERSPEECH 2019*, 2019.

[13] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *INTERSPEECH*, 2018.

[14] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 112–118, 2018.

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

[16] S. Tripathi and H. S. M. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," *CoRR*, vol. abs/1804.05788, 2018. [Online]. Available: http://arxiv.org/abs/1804.05788

[17] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition," in *INTERSPEECH*, 2017.

[18] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML '06*, 2006.

[19] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, 2016.

[20] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *NIPS*, 2015.

[21] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4945–4949, 2016.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[24] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. M. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7405–7409, 2019.

[25] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *INTERSPEECH*, 2015.

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