Multimodal Speech Emotion Recognition using Cross Attention with Aligned Audio and Text

Yoonhyung Lee, Seunghyun Yoon, Kyomin Jung

Department of Electrical and Computer Engineering,
Seoul National University, Seoul, South Korea

cpi1234@snu.ac.kr, mysmilesh@snu.ac.kr, kjung@snu.ac.kr

Abstract

In this paper, we propose a novel speech emotion recognition model called Cross Attention Network (CAN) that uses aligned audio and text signals as inputs. It is inspired by the fact that humans recognize speech as a combination of simultaneously produced acoustic and textual signals. First, our method segments the audio and the underlying text signals into equal number of steps in an aligned way so that the same time steps of the sequential signals cover the same time span in the signals. Together with this technique, we apply the cross attention to aggregate the sequential information from the aligned signals. In the cross attention, each modality is aggregated independently by applying the global attention mechanism onto each modality. Then, the attention weights of each modality are applied directly to the other modality in a crossed way, so that the CAN gathers the audio and text information from the same time steps based on each modality. In the experiments conducted on the standard IEMOCAP dataset, our model outperforms the state-of-the-art systems by 2.66% and 3.18% relatively in terms of the weighted and unweighted accuracy.

Index Terms: speech emotion recognition, multimodal learning, deep learning, attention mechanism

1. Introduction

In developing human-computer interaction systems, Speech Emotion Recognition (SER) technology is considered as an essential element to provide proper response depending on a user's emotional state [1]. Many machine learning models have been built for SER, in which the models are trained to predict an emotion among the candidates such as happy, sad, angry, or neutral for a given speech [2, 3, 4, 5]. Recently, researchers have adopted multimodal approaches in SER considering that emotions can be expressed in various ways such as facial expressions, gestures, texts, or speech [6, 7]. In particular, the text modality has been frequently used in addition to the speech in many SER studies, because human speech inherently consists of the acoustic features and the linguistic contents that can be expressed using text [8, 9].

The major issue in SER using both the audio and text modalities is how to extract and combine the information that each audio and text carries. For example, if someone says, *"Thank you for being with me."* in a very calm voice, the emotional information is contained mostly in the linguistic contents while it sounds neutral based on the acoustic features. Previous studies have approached this issue by designing their models to encode the audio and text independently and fuse the results using attention mechanisms, which help their models effectively capture the locally salient regions from given signals. In these attention mechanisms, the separately encoded audio and text information operated as each other's query and key-value pair. Yoon et al. [8] used the last hidden state of a recurrent modality encoder as a query and used the other encoded modality as a key-value pair in the attention mechanism. In another research, Xu et al. [9] designed their model to learn the alignment between the audio and text by itself from the attention mechanism.

Table 1: *An example of the alignment information provided in the IEMOCAP dataset. The numbers in the table represent the timing when the each uttering of the words begins and ends in the speech. The values are expressed in 10 milliseconds unit. $\langle s \rangle$, $\langle /s \rangle$, and $\langle sil \rangle$ are special tokens meaning the start and end of a sentence, and the silence.*

start	end	word
0	51	$\langle s \rangle$
52	75	i
76	88	like
89	140	$\langle sil \rangle$
141	143	apple
144	177	$\langle /s \rangle$

However, letting the model learn the complex interaction between the different modalities without any constraints can make the training more difficult. Using the last hidden state of a recurrent encoder as a query as in [8] can lead to temporal information loss in the attention as pointed out in [5]. Besides, learning the alignment between the audio and text signals relying on the attention mechanism as in [9] is a challenging task unless additional prior knowledge is provided as in [10, 11].

To overcome these limitations, we propose a novel SER model called Cross Attention Network (CAN) that effectively combines the information obtained from aligned audio and text signals. Inspired by how humans recognize speech, we design our model to regard the audio and text as temporarily aligned signals. In the CAN, each audio and text input is separately encoded through its own recurrent encoder. Then, the hidden states obtained from each encoder are independently aggregated by applying the global attention mechanism onto each modality. Furthermore, the attention weights extracted from each modality are directly applied to each other's hidden states in a crossed way, so that the information at the same time steps is aggregated with the same weights.

In order to make the cross attention work properly, we propose an aligned segmentation technique that divides each audio and text signal into the same number of parts in an aligned way. In the aligned segmentation technique, the text signal is segmented into words. Following the text, the audio signal is segmented using alignment information as shown in Table 1, where the start- and end-time for each word are used to determine the partitioning points in the audio signal. The aligned segmentation technique enables our model to successfully combine the information from the aligned audio and text signals without having to learn the complex attention between different modalities as in the previous works.

To evaluate the performance of the proposed method, we conduct experiments on the IEMOCAP dataset. Firstly, we compare the CAN with other state-of-the-art SER models that use additional text modality. The results show that our model outperforms the other models in both weighted and unweighted accuracy by 2.66% and 3.18% relatively. Furthermore, ablation studies are conducted to see the actual effectiveness of the components such as aligned segmentation, stop-gradient operator, and additional loss. In the ablation studies, we observe the independent contribution of each component for improving the model performance.

2. Related work

After the classical machine learning models such as the hidden markov model or the support vector machine [2, 3], models using neural networks have been actively studied in Speech Emotion Recognition (SER). To improve the model performance, researchers proposed various methods to effectively capture the locally salient regions over the time axis from a given speech. Bertero et al. [12] proposed a model consisting of the convolutional neural network (CNN) that captures local information from given acoustic feature frames. Mirsamadi et al. [5] used the global attention mechanism to make their model learn where to attend to capture the locally salient features. Sahoo et al. [13] proposed to train a CNN-based model with audio segments that are segmented from an utterance with equal length, which improved the model by forcing it to learn to capture the locally salient emotional features in a more elaborated manner.

Recently, multimodal models that use the audio and text together for SER have attracted much attention [8, 9, 14, 15]. Since the audio and text signals contain different information, it has been a major issue of how to design the models to effectively extract information from each modality and combine them. In the previous studies, attention mechanisms were frequently used to combine the information [8, 9], where the hidden states obtained separately from the audio and text signals were used as each other's query or key-value pair. The attention mechanisms were expected to help their models learn to combine the information of each modality by themselves. However, none of these studies used proper constraints of prior knowledge to ease the difficulty of learning the complex interaction between the audio and text signals.

3. Methodology

In this section, we propose a novel Speech Emotion Recognition (SER) model called Cross Attention Network (CAN). First, we explain a preprocessing of the text and audio data, which is necessary for the CAN to work properly. The purpose of the preprocessing is to make the text and audio have the same number of time steps while the same time steps of the sequential signals cover the same time span. Then the CAN is explained, which is a model utilizing the cross attention mechanism that enables the CAN to focus on the salient features of the aligned text and audio signals with a different perspective of each modality.

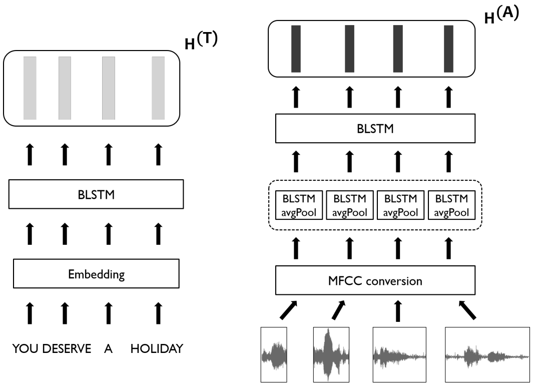

Figure 1: *Text encoder and Audio encoder. The BLSTM modules inside the dotted box share their weights. For better understanding, we represent the audio input using raw waveforms but we convert the wavs into MFCC segments in advance and use them as audio inputs.*

3.1. Data preprocessing

3.1.1. Text data

In this study, we consider a text input as a word sequence, so the text input is represented as $X = \{x_1, x_2, ..., x_L\},\ X \in \mathbb{R}^{L \times V}$, where L is the number of words, V is the size of the vocabulary, and each x_i is a one-hot vector representing the corresponding word. Then, $\mathbf{E}^{(T)} \in \mathbb{R}^{L \times D_e}$, an embedded text input, is obtained after the X passes through a trainable Glove embedding layer [16], where D_e is the dimension of the embedding layer.

3.1.2. Audio data

Let $Y = \{y_1, y_2, ..., y_T\},\ Y \in \mathbb{R}^T$ be the 1-dimensional audio data and $D = \{d_1, d_2, ..., d_L\}$ be its alignment information, where T is the audio length and each $d_i = (s_i, e_i)$ represents the start and the end of each word. To prevent information loss about the correlation, we make the neighboring d_is have 10% overlap.

Using the Y and D, we obtain a segmented audio data $\mathbf{E}^{(A)} \in \mathbb{R}^{L \times (T' \times D_f)}$; the audio Y is first segmented into audio segments $Y' = \{y_{s_1:e_1}, y_{s_2:e_2}, ...,\ y_{s_L:e_L}\}$, and then each segment is converted into a MFCC feature and stacked into the $\mathbf{E}^{(A)}$ with zero-padding. Here, D_f is the number of the MFCC coefficients and T' is the length of the longest MFCC.

3.2. Model architecture

3.2.1. Text encoder

The embedded text data $\mathbf{E}^{(T)}$ is fed into the text encoder consisting of the bidirectional long short-term memory (BLSTM) [17] as represented at the left side of Figure 1, which leads to the hidden states $\mathbf{H}^{(T)} \in \mathbb{R}^{L \times D_h}$ obtained from the equations below:

$$\overrightarrow{h_i} = f_\theta(\overrightarrow{h_{i-1}},\ \mathbf{E}_i^{(T)}), \tag{1}$$

$$\overleftarrow{h_i} = f'_\theta(\overleftarrow{h_{i+1}},\ \mathbf{E}_i^{(T)}), \tag{2}$$

$$\mathbf{H}^{(T)} = \{[\overrightarrow{h_1}; \overleftarrow{h_1}],\ [\overrightarrow{h_2}; \overleftarrow{h_2}],\ ...,\ [\overrightarrow{h_L}; \overleftarrow{h_L}]\}, \tag{3}$$

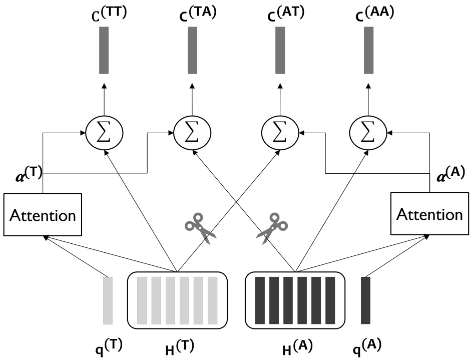

Figure 2: *Cross Attention Network. The scissors represent the stop-gradient operator that cuts the gradient flow during back propagation.*

where f_θ, f'_θ are the forward and backward LSTMs having D_h hidden units with parameter θ. Additionally, h_i represents the hidden state at i-th time step and $\mathbf{E}_i^{(T)}$ represents the i-th embedded word vector of the text data.

3.2.2. Audio encoder

The audio encoder consists of two bidirectional LSTM layers as represented at the right side of Figure 1. The bottom LSTM layer encodes each MFCC segment $\mathbf{E}_i^{(A)} \in \mathbb{R}^{T' \times D_f}$ independently and outputs a vector from each segment using average pooling. The BLSTM modules inside the dotted box in Figure 1 share their weights. The upper LSTM layer encodes the audio features obtained from the bottom layer and outputs the hidden states $\mathbf{H}^{(A)} \in \mathbb{R}^{L \times D_h}$, which has the same time steps L with the $\mathbf{H}^{(T)}$.

3.2.3. Cross attention

In the cross attention, attention weights obtained from one modality are used to aggregate the other modality as shown in Figure 2, while conforming to the constraint that the audio and text are temporarily aligned. Since the salient regions can be different depending on what modality the prediction is based on, the aggregation of the modalities happens twice based on each modality in the cross attention as follows:

$$\alpha_i^{(T)} = \frac{\exp(\ (\mathbf{q}^{(T)})^\intercal\ \mathbf{H}_i^{(T)}\ )}{\sum_j \exp(\ (\mathbf{q}^{(T)})^\intercal\ \mathbf{H}_j^{(T)}\ )}, \quad (i = 1, ..., L), \quad (4)$$

$$\alpha_i^{(A)} = \frac{\exp(\ (\mathbf{q}^{(A)})^\intercal\ \mathbf{H}_i^{(A)}\ )}{\sum_j \exp(\ (\mathbf{q}^{(A)})^\intercal\ \mathbf{H}_j^{(A)}\ )}, \quad (i = 1, ..., L), \quad (5)$$

$$\mathbf{c}^{(TT)} = \sum_i \alpha_i^{(T)}\ \mathbf{H}_i^{(T)}, \quad (6)$$

$$\mathbf{c}^{(TA)} = \sum_i \mathbf{sg}(\alpha_i^{(T)})\ \mathbf{H}_i^{(A)}, \quad (7)$$

$$\mathbf{c}^{(AA)} = \sum_i \alpha_i^{(A)}\ \mathbf{H}_i^{(A)}, \quad (8)$$

$$\mathbf{c}^{(AT)} = \sum_i \mathbf{sg}(\alpha_i^{(A)})\ \mathbf{H}_i^{(T)}, \quad (9)$$

where $\mathbf{q}^{(T)}$ and $\mathbf{q}^{(A)}$ are the global queries used to decide which parts of the aligned signals to focus on based on each modality perspective. Additionally, $\mathbf{c}^{(xy)}$s are context vectors, where the x represents the modality used as a query and the y represents the modality used as a key-value pair. To prevent the CAN from learning attention based on the other modality, we introduce a function $\mathbf{sg}$; stop-gradient operator as shown in the equations (7) and (9). It cuts the gradient flow through its argument during the backpropagation.

3.3. Training objective

In the training, the CAN makes three different predictions using the context vectors.

$$\hat{y} = \text{softmax}([c^{(TT)}; c^{(TA)}; c^{(AA)}; c^{(AT)}])^\intercal\ \mathbf{W} + \mathbf{b}), \quad (10)$$

$$\hat{y}^{(T)} = \text{softmax}((c^{(TT)})^\intercal\ \mathbf{W}^{(T)} + \mathbf{b}^{(T)}\ ), \quad (11)$$

$$\hat{y}^{(A)} = \text{softmax}((c^{(AA)})^\intercal\ \mathbf{W}^{(A)} + \mathbf{b}^{(A)}\ ), \quad (12)$$

where the $\mathbf{W}$s and $\mathbf{b}$s are trainable weights. $\hat{y}$ is made based on all the context vectors and each $\hat{y}^{(T)}$ and $\hat{y}^{(A)}$ is made based on a context vector that uses either the text or the audio modality. Using the predictions, we calculate loss terms as follows:

$$\mathcal{L}_{align} = CE(\hat{y}^{(T)},\ y) + CE(\hat{y}^{(A)},\ y), \quad (13)$$

$$\mathcal{L}_{total} = CE(\hat{y},\ y) + \alpha \cdot \mathcal{L}_{align}, \quad (14)$$

where CE represents the cross-entropy loss, y is the true emotion labels, and α is a weight for the additional loss term $\mathcal{L}_{align}$, of which optimal value is found using the validation dataset. The additional loss terms in $\mathcal{L}_{align}$ are added to help the global attention attend to the salient features based on each modality better.

$$\hat{y}^{\text{final}} = (\hat{y}) \cdot (\hat{y}^{(T)})^\alpha \cdot (\hat{y}^{(A)})^\alpha \quad (15)$$

After the training, the final prediction $\hat{y}^{\text{final}}$ is calculated following the equation (15).

4. Experiments

In this section, we describe the experimental setup and the results conducted on the IEMOCAP dataset. First, we compare the CAN to other SER models for the weighted accuracy (**WA**) and the unweighted accuracy (**UA**), where the CAN shows the best performance. In addition, we conduct several analyses on our model to see how each component described in Section 3 affects the performance of the CAN.

4.1. Dataset

In the experiments, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [18] dataset which provides the speech and text dataset including the alignment information as represented in Table 1. Each utterance in the dataset is labeled as one of the 10-class emotions, where we do not use the classes with too few data instances (fear, disgust, other) so the final dataset contains 7,486 utterances in total (1,103 angry, 1,040 excite, 595 happy, 1,084 sad, 1,849 frustrated, 107 surprise and 1,708 neutral). In the experiments, we perform 10-fold cross-validation, and in each validation, the total dataset is split into 8:1:1 training set, validation set, and test set, respectively.

4.2. Experimental setup

For text input, we use a sequence of words in Table 1 as our text input and the 300-dimensional GloVe word vectors [16] are used as the embedding vectors. In this step, we remove the special tokens such as '$\langle s \rangle$', '$\langle sil \rangle$', '$\langle /s \rangle$' and their durations are equally divided into the neighboring words. For audio input, we use the zero-padded MFCC segments as our audio input, which are obtained as described in Section 3.1.2. In the MFCC conversion, we use 40 MFCC coefficients and the frames are extracted while sliding the hamming window with 25ms frame size and 10ms hopping. We use the bidirectional LSTMs with 128 hidden units followed by the dropout layer with 0.3 dropout probability. For the cross attention, multi-head global attention with four heads is used to view the inputs from various perspectives so enrich the aggregated information [19]. During the training, we use the validation dataset as a criterion of early stopping with the patience 10. We use the batch size of 64 and use the Adam optimizer [20] with a learning rate of 1e-3, and the gradients are clipped with a norm value of 1.0. The weight of the additional loss term α is set to 0.1, which is obtained from the cross-validation.

4.3. Results

4.3.1. Performance comparison

Table 2: *Comparison of the models. The accuracy values are represented as an average of the 10-fold validations and the standard deviations are written next to them.*

Model	WA	UA
TextModel [5]	0.513 ± 0.015	0.443 ± 0.015
AudioModel [5]	0.431 ± 0.017	0.323 ± 0.015
Yoon et al. [8]	0.564 ± 0.020	0.472 ± 0.017
Xu et al. [9]	0.560 ± 0.028	0.450 ± 0.028
CAN (ours)	**0.579** ± **0.019**	**0.487** ± **0.017**

Table 2 shows the performance of the CAN and the other SER models. Each 'TextModel' and 'AudioModel' uses a single modality by encoding it with a simple bidirectional LSTM with the global attention following [5]. The other two multimodal models are [8] and [9] proposed in the previous studies, where the attention weights are obtained based on the interaction between the audio and the text modalities. In the experiments, we re-implement all the models and obtain the accuracy values as described in Section 4.2. As the Table 2 shows, our CAN outperforms the other models for both **WA** and **UA** including the previous state-of-the-art model [8]. To analyze the causes of the performance gain, we conduct further experiments to see the effectiveness of the components in our methodology, which are described in the next sections.

Table 3: *Comparison of the segmentation policies.*

Segmentation	WA	UA
Aligned	**0.579** ± 0.019	**0.487** ± **0.017**
Equal	0.568 ± 0.022	0.467 ± 0.021

4.3.2. Segmentation policy

In order to demonstrate the superiority of the aligned segmentation, we compare it to the segmentation where a 1-dimensional audio signal is segmented into the segments of equal length, which has been widely used in the previous studies [13, 21]. In the experiment, the aligned segmentation outperforms the equal segmentation for both **WA** and **UA**. The results in Table 3 imply that our aligned segmentation actually has effectiveness in combining the information in the cross attention.

4.3.3. Ablation study

Table 4: *Accuracy comparison in the ablation studies*

Model	WA	UA
CAN	**0.579** ± **0.019**	**0.487** ± **0.017**
- stop-gradient	0.570 ± 0.018	0.469 ± 0.023
- $\mathcal{L}_{align}$	0.573 ± 0.015	0.484 ± 0.017
- stop-gradient, $\mathcal{L}_{align}$	0.563 ± 0.015	0.479 ± 0.024
- cross attention	0.556 ± 0.013	0.458 ± 0.015

As supportive evidence of the components in our model, we conduct ablation studies for four variant models while removing each of the components in Section 3. When we remove the stop-gradient operator and additional loss term $\mathcal{L}_{align}$, the accuracy of the CAN decreases and the worst performance for the **WA** is observed when both components are removed. Furthermore, we even remove the whole cross attention in the CAN, where the prediction is based only on a concatenation of the $c^{(\mathrm{TT})}$ and $c^{(\mathrm{AA})}$ and the stop-gradient operator and the $\mathcal{L}_{align}$ are not used. In that case, the performance decreases even more compared to the other variants.

5. Conclusion

In this paper, we propose a Cross Attention Network (CAN) for Speech Emotion Recognition (SER) task. It uses the cross attention to combine information from the aligned audio and text signals. Inspired by the way humans recognize speech, we align the text and audio signals so that the CAN regards each modality to have the same time resolution. In the experiments conducted on the IEMOCAP dataset, the proposed system outperforms the state-of-the-art systems by 2.66% and 3.18% relatively for the weighted and unweighted accuracy. To the best of our knowledge, this is the first study that shows the improvement using the aligned audio and text signals in SER. In order to apply our system to the real-world scenario where only the speech signal is available, the text and alignment information are required for the CAN to work properly. In future work, we plan to extend our research by integrating the CAN with the automatic speech recognition system which outputs the text and alignment information given a speech signal.

6. Acknowledgments

K. Jung is with ASRI, Seoul National University, Korea. This work was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under the Industrial Technology Innovation Program (No. 10073144). This research was results of a study on the "HPC Support" Project, supported by the 'Ministry of Science and ICT' and NIPA.

7. References

[1] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wrobel, "Emotion recognition and its applications," in *Human-Computer Systems Interaction: Backgrounds and Applications 3*. Springer, 2014, pp. 51–62.

[2] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[3] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.

[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] 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.

[6] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 92–103.

[7] S. Yoon, S. Dey, H. Lee, and K. Jung, "Attentive modality hopping mechanism for speech emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3362–3366.

[8] 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.

[9] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *arXiv preprint arXiv:1909.05645*, 2019.

[10] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2837–2846.

[11] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," *arXiv preprint arXiv:1910.10288*, 2019.

[12] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5115–5119.

[13] S. Sahoo, P. Kumar, B. Raman, and P. P. Roy, "A segment level approach to speech emotion recognition using transfer learning," in *Asian Conference on Pattern Recognition*. Springer, 2019, pp. 435–448.

[14] J. Sebastian and P. Pierucci, "Fusion techniques for utterance-level emotion recognition combining speech and transcripts," in *Proc. Interspeech*, 2019, pp. 51–55.

[15] J. Liang, S. Chen, J. Zhao, Q. Jin, H. Liu, and L. Lu, "Cross-culture multimodal emotion recognition with adversarial learning," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4000–4004.

[16] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

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

[18] 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.

[19] 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.

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

[21] 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.