INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Learning Mutual Correlation in Multimodal Transformer for Speech Emotion Recognition

Yuhua Wang, Guang Shen, Yuezhu Xu, Jiahang Li, Zhengdao Zhao*

College of Computer Science and Technology, Harbin Engineering University, Harbin, China
{wangyuhua, shenguang, xuyuezhu, lizhi01, 1099361413}@hrbeu.edu.cn

Abstract

Various studies have confirmed the necessity and benefits of leveraging multimodal features for SER, and the latest research results show that the temporal information captured by the transformer is very useful for improving multimodal speech emotion recognition. However, the dependency between different modalities and high-level temporal-feature learning using a deeper transformer is yet to be investigated. Thus, we propose a multimodal transformer with sharing weights for speech emotion recognition. The proposed network shares the weights across the modalities in each transformer layer to learn the correlation among multiple modalities. In addition, since the emotion contained in a speech generally include audio and text features, both of which have not only internal dependence but also mutual dependence, we design a deep multimodal attention mechanism to capture these two kinds of emotional dependence. We evaluated our model on the publicly available IEMOCAP dataset. The experimental results demonstrate that the proposed model yielded a promising result.

Index Terms: Speech emotion recognition, Transformer, Sharing weights, Multimodal attention

1. Introduction

Emotion is a common trait of all human beings across different nations and cultures. Psychological research shows that people of different cultures have the similar ways of emotional expression. Emotion, as a kind of human psychological activity, represents the most real idea in the human heart and also reflects the human health state. The cross-cultural and cross-racial characteristics of emotion make the research of emotion recognition attract increasing attention from academia and industry.

Emotion recognition infers the emotional state through the comprehensive analysis of human facial expression, voice, and speech content by computer, aiming to achieve a better human-computer interaction experience. Emotion recognition has a wide application scenario. It can be used to detect the emotional state of drivers and passengers, judging whether there is a dispute or not [1]. It can also automatically detect a suspect's psychological state in criminal investigation, assisting lie detection in time. Various customer service systems can also improve user experience by detecting user's emotional states. It is also used to help autistic children to solve their difficulties in recognizing and expressing emotions. Besides, emotion recognition is widely used in various voice assistants and wearable devices.

Emotional recognition research has developed from the 1970s to the present. Deep learning algorithm has improved the performance of unimodal emotion recognition [2, 3, 4, 5, 6]. Voice emotion recognition, facial expression recognition, text emotion recognition, and other modalities have been significantly developed. However, the way people express their emotions is multimodal, inspiring researchers to use multimodal information to improve emotion recognition accuracy [7, 8, 9, 10, 11, 12]. Emotion information in speech and semantics is an essential resource, a necessary part of human communication. In the same sentence, because of the speaker's different intonation, emotion will be other. The same intonation says additional content, emotional expression will also be different due to the varying textual. Because speech and semantics are more accessible to obtain than video, using multimodal information of audio and text to achieve emotion recognition is essential for emotion research in the future.

2. Related work

Recently, deep learning has been widely used to resolve emotion recognition [13, 14, 15, 16, 17]. Most studies usually extracted engineered low-level features or high-level statistical features and applied a classifier for unimodal speech emotion recognition. It shows that deep learning can extract high-level features from raw data or low-level features.

Multimodal approaches combining audio and text information have been widely used in speech emotion recognition. In [18], Yenigalla *et al.* generate phoneme embedding, combining spectrogram and phoneme embedding sequence based on CNN for emotion classification. Some studies employed recurrent neural networks (RNNs) for emotion classification due to the sequential structure of speech signals and text [19]. Yoon *et al.* [20] utilize an audio encoder, a text encoder, and a multimodal fusion network, including an attention mechanism for emotion classification. In [21], Yoon *et al.* proposed an efficient framework using two bi-directional BLSTM for obtaining hidden representations of the utterance. Besides, they apply multi-hop attention to computes the relevant segments of the textual data corresponding to the audio signal, automatically inferring the mutual correlation between the modalities. Gu *et al.* [22] proposed a dyadic fusion network that mainly relies on attention mechanisms to extract contextual features and fuse the audio and data information. Pepino *et al.* [23] studied different approaches for classifying emotions from speech using acoustic-text features and proposed to obtain contextualized word embeddings with BERT to represent the information contained in speech transcriptions. Atmaja *et al.* [24] proposed two methods to predict emotional attributes from audio and visual data using multitask learning and a fusion strategy. He employed multitask learning by adjusting three parameters for each attribute to improve the recognition rate and proposed multistage fusion to combine results from various modalities' final predictions.

The latest studies confirm that learning the temporal correlation among multiple modalities at a fine-granular level is vital for a more accurate SER. In [25], Xu *et al.* employed an atten-

* corresponding author.

http://dx.doi.org/10.21437/Interspeech.2021-2004

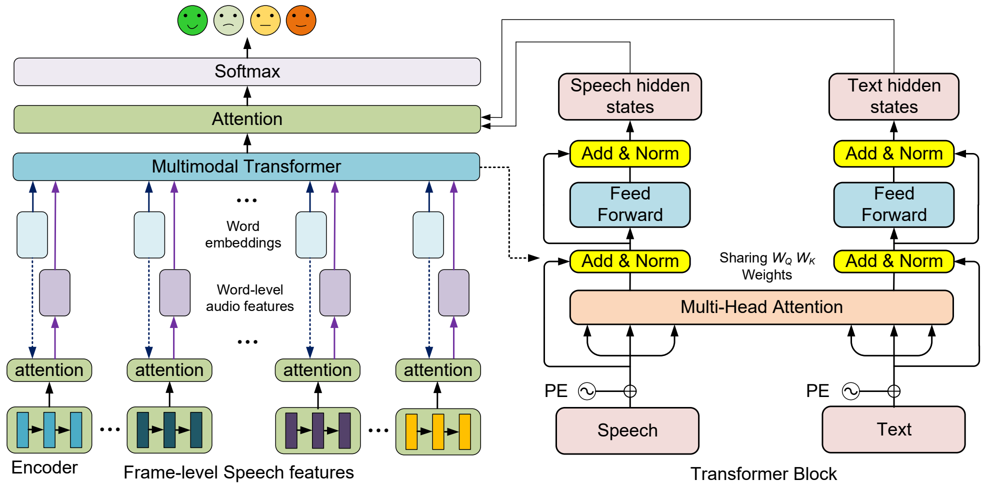

Figure 1: *The architecture of our proposed model. The right part illustrates the multimodal transformer fusion with weights sharing; the left part illustrates the word-level speech representation and the attention-based interaction mechanism.*

tion mechanism to learn the alignment weights between speech frames and text words. Shen and Lai [26] study the problem of temporal correlation among multiple modalities. The key idea is to apply an interaction mechanism to dynamically fuse audio and text features at the word level, modeling the evolution of the emotion contained in an utterance.

In this paper, we construct a novel deep multimodal transformer network with sharing weights, which can preserve the whole state's multimodal property to facilitate the complex correlation learning between multiple modalities. We put forward an original attention-based deep multimodal fusion framework to obtain the co-dependent between different modal feature sequences, leading to more accurate SER. The experimental results, on the public benchmark IEMOCAP dataset, show the superiority than other fusion strategies.

3. The Model Framework

Given an audio signal, we can use the ASR system to get the corresponding text, then use forced alignment methods to get the timestamps of each word and use attention mechanism to get the related word-level features of speech, subsequently, use transformer to learn the consistency correlation between the two multiple modalities, and finally, use the self-attention mechanism to fuse two modal features for emotion classification. The overall architecture of our framework is illustrated in Figure 1.

3.1. Multimodal Transformer with Sharing Weights

In this work, we perform transform fusion for continuous emotion recognition. The transformer [27] model learns emotional long-term temporal dependencies with the self-attention mechanism. What's more, we achieve word-level audio text modalities fusion with the multi-head attention module. Since speech and text have a constant relationship in emotional expression, in order to learn this consistency relationship, we introduce a transformer model with sharing Q and K weights. The transformer's self-attention mechanism is realized by Q, K, V vectors. V represents the value of the input sequence. Q and K can score each node of the sequence. There is a complementary relationship between the audio and text features. For example, the emotional features of a word may be weak, but the corresponding text features are relatively strong. In order to simulate the emotional evolution of two modes in a fine-grained way, we use a multimodal transformer model with sharing Q and K weights. The two modes simulate the evolution of emotion over time by the joint influence on the two weights.

The first step of the model is to obtain word-level features of speech and text. We can get the corresponding text from a speech signal through the ASR system in real time, then use the forced alignment method to get the corresponding time stamp of each word, obtaining the corresponding speech segment frames of each word. We use the Embedding layer in PyTorch to transform text words into 100 dimensional word-embedding vectors. In this way, the text sequence contained in a sentence is represented as $[t_1, t_2, \cdots, t_N]$, where N represents the number of words contained in the sentence. At the same time, the corresponding acoustic word-level feature sequence is represented as $[a_1, a_2, \cdots, a_N]$. We discuss how to generate the word-level speech features a_i in the next section. Note that the speech and text sequence features have the same length.

Before introducing the multimodal transformer fusion mechanism, we firstly use transformer to model the text and audio sequences at the word level. We mainly introduces the self attention mechanism of transformer. The formulas, excluding the bias terms, are as follows:

$$Q = W_Q^* X \tag{1}$$

$$K = W_K^* X \tag{2}$$

$$V = W_V X \tag{3}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{4}$$

where X indicates the input audio and text feature sequences, each modality in the multimodal transformer has its own weights, hidden states h_{a_i}, h_{t_i}.

In order to explicitly learn the word-level correlation between the audio and text, we introduced a multimodal transformer network with sharing weights across the modalities. As shown in Fig.1, the network builds a transformer for each modality while sharing weights W_Q, W_K. Therefore, the network is allowed to behave uniquely and learn separate temporal features for different modalities. This is convenient for handling asynchronous emotion reflection among the modalities. It also provides more interpretation and training efficiency by concentrating on one modality's input.

When building the deep multimodal transformer network, exclusive states preserve the multimodal property, thereby facilitating complex correlation learning among the modalities. At the same time, the cross-modal correlation is explicitly learned by sharing the weights W_Q^*, W_K^*, and not W_V, because, compared with Q and K, V contains the main emotion information from each modality, representing the emotional information that each transformer should retain. By sharing W_Q, W_K across the modalities, the multimodal transformer can interact with each other during the learning processes. Provided that W_Q, W_K are also shared across the time steps, they play an important role in learning the temporal correlation among the different modalities.

3.2. Word-level Speech Representation

Referring to the WISE model, we propose a hierarchical speech emotion representation model to extract meaningful high-level emotion features from the basic speech signal. Based on the hierarchical property of speech, the feature vectors of frames are aggregated to construct word-level features. In a segment of speech, the contribution of different words' pronunciation to the emotional expression of speech is different, and the contribution of frames to words is also different. The importance of frames depends on the context. In order to describe this situation, we apply the attention mechanism to capture this difference.

Similar to [26], we use openSMILE [28] to generate frame-level audio features. Each audio signal will be divided into frames of 25ms window with 40% overlap. A frame-level feature vector contains 39-dimensional Mel-Frequency Cepstral Coefficients (MFCCs), including 13 cepstral coefficients, 13 delta coefficients and 13 acceleration coefficients.

To construct the word-level feature of the i-th word in a sentence, we represent the frame sequence in the word as $[f_{i1}, f_{i2}, \cdots, f_{iN_i}]$, where f_{ij} is the feature of the j-th frame in the i-th word and N_i denotes the number of frames in the word. A gated recurrent unit (GRU) network is used to capture the contextual information between frames. It can be represented as:

$$h_{ij} = \mathrm{GRU}(f_{ij}), j \in \{1, 2, \cdots, N_i\}, \tag{5}$$

where h_{ij} is the contextual hidden state of the j-th frame. We apply the attention mechanism with textual attention information to extract the informative frames. We represent the word embedding vector as a query vector to calculate the frame feature sequence's weight distribution. The formulas are as follows:

$$\alpha_i = \frac{h_{ij}^T \omega_a h_i^{(e)}}{\sum_j h_{ij}^T \omega_a h_i^e} \tag{6}$$

$$a_i = \sum_{j=1}^{N_i} \alpha_i h_{ij}, \tag{7}$$

where $h_i^{(e)}$ represents the i_{th} word-embedding in the sentence. We form the word-level acoustic vector a_i by taking a weighted sum of contextual state h_{ij} of the frame and the corresponding frame level attention distribution α_i specifically.

3.3. Multimodal Cooperation and Interaction

With the aforementioned multimodal transformer and word-level speech representation, we are ready to present the attention-based multimodal fusion framework. In [29], Chen *et al.* achieve excellent results in capturing the mutual reliance between different feature sequences with an incidence matrix. On this basis, we apply the attention model to capture the mutual dependence between audio and text hidden states. We introduce a different relation vector from [29] in applying the relation vector and obtaining the temporal correlation between different modal feature sequences.We represent the text and audio hidden states $[h_{a_1}, h_{a_2}, \cdots, h_{a_M}]$ and $[h_{t_1}, h_{t_2}, \cdots, h_{t_M}]$ as H_a and H_t

First, for the text hidden states H_t, we calculate the relation vector $r_{t,i}$ between the part h_{t_i} and the audio hidden states H_a as follows:

$$r_{t,i} = H_a^T tanh(h_{t_i}^T W_h^t H_a), \tag{8}$$

where W_h^t are trainable parameters. We then integrate $r_{t,i}$ into the part h_{t_i}, and the co-dependent text emotion features h_t is the sum of all parts, as follows:

$$h_t = \sum_{i=1}^{M} tanh(W_t h_{t_i} + r_{t,i}), \tag{9}$$

where the W_t is the weight parameters, M denotes the number of words contained in a sentence.

Similarly, for the audio hidden states H_a, we calculate the relation vector $r_{a,i}$ between the part h_{a_i} and the text hidden states H_t, as follows:

$$r_{a,i} = H_t^T tanh(h_{a_i}^T W_h^a H_t), \tag{10}$$

where W_h^a are trainable parameters. We then integrate $r_{a,i}$ into the part h_{a_i}, and the co-dependent audio emotion features h_a of is the sum of all parts, as follows:

$$h_a = \sum_{i=1}^{M} tanh(W_a h_{a_i} + r_{a,i}), \tag{11}$$

Finally, we concatenate the audio feature sequences h_t and the text feature sequences h_a to obtain the utterance level emotion representation p_c, as follows:

$$p_c = ReLU(W_u[h_a; h_t] + b_u), \tag{12}$$

The final emotion representation p_c will be fed into the fully connected layers for emotion classification. The fully connected layers contain two linear layers, the first two of which are followed by a rectified linear unit (ReLU) layer and a dropout layer.

$$p_1 = ReLU(\omega_1^\top r_1 p_c) \tag{13}$$

$$\hat{y}_k = \mathrm{softmax}(\omega_2^\top p_1), \tag{14}$$

where ω_1, ω_2, are trainable parameters with ω_i $(i \in \{1, 2\})$ being the weight of each linear layer, r_1 are the dropout vectors, and $\hat{y}_k$ is the output of the softmax function, which represents the final prediction result. The cross-entropy loss for K-class classification is used as the loss function:

$$\mathcal{L} = \sum_{k=1}^{K} y_k \log(\hat{y}_k). \tag{15}$$

4. Experimental Evaluation

4.1. Datasets

We evaluated our model on published datasets: interactive emotional dyadic motion capture (IEMOCAP) dataset. IEMOCAP is a multimodal emotion dataset including visual, audio, and text data (Busso et al, 2008 [30]). It was referred to as a standard benchmark dataset widely used for SER. It includes five sessions of utterances for 10 unique speakers, along with the corresponding labeled speech text (at both phoneme and word levels). For each sentence, we use the label agreed on by the majority (at least two of the three annotators). In this study, we use the ground-truth transcripts to generate word-level textual features, we evaluate 4-category (happy, sad, anger, and neutral) for emotion classification [26]. The final dataset consists of 1635 happiness data, 1103 anger data, 1084 sadness, 1708 neutral data. Similar to the setting in [20, 18], we perform 5-fold cross-validation to do the evaluation. We compare our solution with four state-of-the-art multimodal methods that achieve the best performance on IEMOCAP: LSTM+Attn [25], MHA-2 [21], WISE [26] and dual RNNs [20].

4.2. Implementation Details

We implement our model in PyTorch. Each transformer model contains two transformer blocks. We set the maximum number of frames contained in a word to 64 because most words contain less than 64 frames, the maximum number of words in a sentence to 128. We use 50 hidden units in a GRU for word-level audio representation, 50 hidden units in the text embedding. We set the output dimension of the attention layer to 100 and the output dimensions of the two linear layers to 50, 4 respectively. The proposed model is optimized with the Adam optimizer. We set the learning rate to 0.001 and the decay rate of the learning rate to 0.001. We use L_1 and L_2 regularization with the regularization coefficient of 0.0001, 0.0002 and 0.5 dropout rate to address overfitting.

4.3. Results and discussion

Following [20, 18], the performance of our system is measured in widely used evaluation metrics: weighted accuracy (WA) is the overall classification accuracy, and unweighted accuracy (UA) that is the average recall over the emotion categories. Table 1 presents the performance of our approach for emotion recognition compared with the state-of-the-art methods. First, the experimental results show that most multimodal systems achieve better performance than unimodal systems. Because of integrating more emotion information, multimodal approaches significantly improve the performance of emotion recognition. Second, the WISE model achieves better performance. It confirms that leverage word-level interactions between audio and text features is indeed beneficial for speech emotion recognition. Finally, our proposed model achieves the best WA scores. It outperforms the best state-of-the-art methods by absolute 0.3%WA increase. But it has an absolute 0.5%UA decrease over the best methods.

Table 1: *Performance comparison between our solution and state-of-the-art multimodal models on the IEMOCAP dataset*

Methods	**WA**	**UA**
Audio only	0.665	0.657
Text only	0.692	0.701
LSTM+Attn [25]	0.725	0.709
WISE [26]	0.759	0.764
Dual RNNs [20]	0.737	0.753
MHA-2 [21]	0.765	0.776
ours	**<u>0.768</u>**	**<u>0.771</u>**

Table 2: *Performance of different variants of our proposed model*

Methods	**WA**	**UA**
MT+WR	0.751	0.739
MT+MI	0.735	0.729
MI+WR	0.743	0.751
MI+MT+WR*	0.762	0.756
ours	0.768	0.771

To prove the benefits of the different components consisted in our framework, we perform an ablation study. We study and compare the following variants of our model to further evaluate and understand our model's performance. The MT+WR method uses only the multimodal transformer with sharing weights and the word level speech representation. The MT+MI method use the multimodal transformer with sharing weights and Multimodal Cooperation and Interaction. The MI+WR method use only the co-attention layer to capture the mutual dependence of text and audio features, obtaining further high-level emotion representation. In addition, it uses an independent transformer to model the text and audio sequences at the word level separately instead of sharing W_Q, W_K weights. Table 2 reports the main results of the above four variants and the originally proposed model on IEMOCAP datasets. It shows that the complete model we proposed outperforms the four variants, which demonstrates its validity. The results of MT+WR are obviously worse than those of the complete model, which indicates that the lack of aggregation between audio and text is not comprehensive. The MI+MT method takes out WR. By comparing its performance with the complete model, it can be seen that WR is important to distill emotion signals from audio data. The performance difference between MI+WR and the complete model suggests that MT indeed helps more effectively capture emotion evolutions by learning the temporal correlation among the different modalities, leading to improved performance. WR* is a simplified version of WR without the textual attention layer. We can observe that using text information to generate word-level emotional features of speech allows removing the emotionally irrelevant information from original audio features.

5. Conclusions

In this paper, we studied the problem of SER by finding the mutual correlation and integrating them between the information from audio and text. In order to learn the dependency and high-level temporal correlation between different modalities, we proposed a novel deep multimodal transformer network reinforced with weight sharing that learn separate temporal features and handle asynchronous emotion reflection among the modalities, modeling the emotion evolution in a utterance alone timeline. In addition, since the emotion contained in audio and text features have internal and mutual dependence, we design a deep multimodal attention mechanism to capture the emotional dependence. The experimental results show that the proposed model achieved a promising results. In future, we plan to extend the proposed network to additional applications that simultaneously utilize video and other modalities.

6. References

[1] Y. Xing, Z. Hu, Z. Huang, C. Lv, and E. Velenis, "Multi-scale driver behaviors reasoning system for intelligent vehicles based on a joint deep learning framework," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020.

[2] P. Tzirakis1, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5089–5093.

[3] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 3673–3677.

[4] 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.

[5] 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 (TMM)*, vol. 20, no. 6, pp. 1576–1590, 2018.

[6] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 272–276.

[7] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 247–251.

[8] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 2247–2256.

[9] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 3454–3466.

[10] E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6720–6724.

[11] J. Sebastian and P. Pierucci, "Fusion techniques for utterance-level emotion recognition combining speech and transcripts," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 51–55.

[12] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1936–1940.

[13] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. Schuller, "Towards temporal modelling of categorical speech emotion recognition," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 932–936.

[14] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 152–156.

[15] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, and H. Meng, "Speech emotion recognition using capsule networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6695–6699.

[16] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7390–7394.

[17] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2578–2582.

[18] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 3688–3692.

[19] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

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

[21] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," *IEEE*, 2019.

[22] Y. Gu, X. Lyu, W. Sun, W. Li, S. Chen, X. Li, and I. Marsic, "Mutual correlation attentive factors in dyadic fusion networks for speech emotion recognition," in *Proceedings of the 27th ACM International Conference on Multimedia (MM)*, 2019, pp. 157–166.

[23] 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.

[24] B. T. Atmaja and M. Akagi, "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," *IEEE*, 2020.

[25] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 3569–3573.

[26] G. Shen, R. Lai, and R. Chen, "WISE: word-level interaction-based multimodal fusion for speech emotion recognition," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 369–373.

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

[28] 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 (MM)*, 2010, pp. 1459–1462.

[29] W. Chen, F. Cai, H. Chen, and M. Rijke, "A dynamic co-attention network for session-based recommendation," in *the 28th ACM International Conference*, 2019.

[30] 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.