INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Focus-attention-enhanced Crossmodal Transformer with Metric Learning for Multimodal Speech Emotion Recognition

Keulbit Kim[1,2], *Namhyun Cho*[1]

[1]Speech AI Lab., NCSOFT Corporation, South Korea
[2]Department of Electrical and Electronic Engineering, Yonsei University, South Korea
{keulbit, cnh2769}@ncsoft.com

Abstract

Recognizing emotions in speech is essential for improving human-computer interactions, which require understanding and responding to the users' emotional states. Integrating multiple modalities, such as speech and text, enhances the performance of speech emotion recognition systems by providing a varied source of emotional information. In this context, we propose a model that enhances cross-modal transformer fusion by applying focus attention mechanisms to align and combine the salient features of two different modalities, namely, speech and text. The analysis of the disentanglement of the emotional representation various multiple embedding spaces using deep metric learning confirmed that our method shows enhanced emotion recognition performance. Furthermore, the proposed approach was evaluated on the IEMOCAP dataset. Experimental results demonstrated that our model achieves the best performance among other relevant multimodal speech emotion recognition systems.

Index Terms: speech emotion recognition, multimodal emotion recognition, multimodal sentiment analysis, focus-attention mechanism, metric learning

1. Introduction

Emotions are an essential aspect of human communication. Accurate emotion recognition is critical for effective communication. Voice-based human-machine interactions are becoming increasingly common and, with the advent of popularized chatbots and large-scale language models, the importance of speech emotion recognition (SER) based on various modalities is increasing. Emotions play a crucial role in how humans interact with computers and technology. The ability to detect emotions accurately can improve the user experience, enhancing the usability of computer systems [1]. For example, SER can be used by mental health professionals to track and monitor the emotional state of patients, particularly those suffering from depression or anxiety for which emotional regulation is crucial for successful treatment [2]. Additionally, using SER a virtual assistant can detect emotions from the user's voice and offer more appropriate help or guidance to resolve the issue. However, despite significant progress in this field, emotion recognition remains a challenging task because emotions are complex, subjective, and their expression differs significantly among individuals.

The utilization of multiple modalities, such as textual, acoustic, and visual, facilitates the inference of human emotions or sentiments, as each modality captures different facets of human emotional expression. Consequently, effectively fusing multimodal information is crucial for improving the accuracy and comprehensiveness of predictions by leveraging the complementary nature of diverse modalities [3, 4, 5]. Incorporating text information is also essential, as speech containing semantic information is inherently multimodal and the text modality can provide the necessary contextual information for multimodal emotion recognition. Recent studies [6, 7, 8, 9, 10, 11, 12, 13] have demonstrated that leveraging multimodal features is more effective than relying on a single modality. Yoon et al. [6] use multimodal dual recurrent encoder networks to integrate information from both audio and text. In another study [7] Yoon et al. propose a multi-hop attention model to identify the relevant segments of textual data that correspond to the audio signal. Gu et al. [8] proposed a dyadic fusion network that primarily relied on attention mechanisms to extract contextual features and fuse the audio and textual information.

In other words, in situations where two modalities interact sequentially and convey emotion information, it is crucial to construct models that can accurately and efficiently perform both fusion and alignment. This requires explicitly modeling the interactions between the modalities and taking into account the dynamics within each modality. For instance, the occurrence of a negative word at the middle of an utterance may cause the preceding speech to become louder. By capturing important area in modality these inter-modal interactions, the model can better analyze multimodal sequential data. In [14], Chen et al. propose the key-sparse transformer that assigns attention weight only to emotion-related features when aligning two modalities. However, humans instinctively identify the important segment in each modality when perceiving emotions. To this end, we proposed a novel crossmodal transformer architecture that incorporates an focus-attention (FA) mechanism. The FA mechanism was employed in the document summary task detect salient information using focal bias [15]. In this work, FA detects the dominant segment for SER in each modality to achieve accurate and efficient alignment between speech and text modalities.

Meanwhile, self-supervised learning (SSL) using pre-trained models has demonstrated remarkable performance in various research fields and tasks [16], including natural language processing (NLP) [17, 18] and automatic speech recognition (ASR) [19]. Existing studies typically apply various methods for multimodal fusion and use transfer-learning approaches based on a SSL model [9, 20, 14]. Yang et al. [9] fine-tune the textual/acoustic pre-trained SSL models on sentiment analysis and emotion recognition task before bi-modal model. Zhao et al. [20] proposes a multi-level fusion framework that combines the SSL model embeddings to address the issue of data sparsity in multimodal emotion recognition. Although SSL models are widely used knowledge transfer models trained on large-scale datasets, they are designed to preserve the maximum information content of the input, resulting in feature representations that may include non-emotional information. In this context, we ar-

10.21437/Interspeech.2023-555

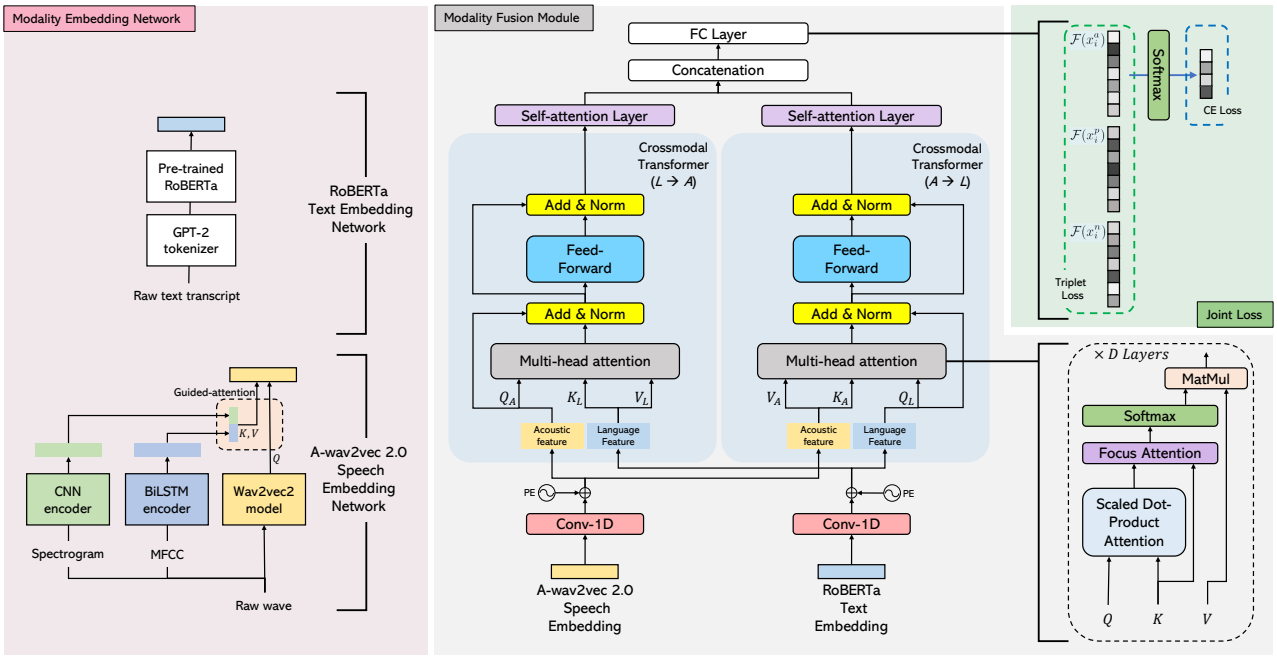

Figure 1: *Architecture of the proposed method, including (a) the modality-embedding network, (b) modality fusion module, and (c) joint loss with the metric loss.*

gue that the disentanglement of the emotion embedding space can play a crucial role in classifying emotions using deep metric learning (DML).

2. Proposed Framework

The proposed framework depicted in Figure 1 consists of three main parts. The modality-embedding network is used to learn acoustic and language features of the input. The modality fusion module includes a crossmodal transformer [3] enhanced with a FA mechanism for learning align of two-modality. The joint loss uses a combination of metric and cross-entropy loss for back-propagation during training. Further elaboration of process will be provided in the subsequent subsections.

2.1. Modality-embedding Network

2.1.1. Text Embedding Network

RoBERTa [18] is a pre-trained language model optimized for a variety of natural language processing tasks. It is based on the same transformer-based architecture as BERT [21], but with several modifications to improve its performance. It is trained on a larger and more diverse corpus of text data, use a larger batch size during training, and do not use BERT's next sentence prediction object during pre-training. In our work, the text transcript of speech is tokenized using the GPT-2 tokenizer [22] and send to the pre-trained RoBERTa models. The text embedding outputs have the dimension of 1024 and maximum sequence length of 512.

2.1.2. Acoustic Wav2vec 2.0 Feature Representation

In another input pipeline, simultaneously, the model receives the speech signal, while the text script is fed into the system. However, relying solely on the wav2vec 2.0 feature is not enough to accurately capture all the prosodic information necessary for recognizing emotions in speech [23]. Therefore, inspired by [24] that utilized different encoders to incorporate multi-level acoustic information in the SER system, we introduce a new structure that combines acoustic information with wav2vec 2.0 features. As illustrated Figure 1(a), after the raw speech signal is pre-processed using windowing, two types of acoustic information features - the spectrogram and MFCC - are separately fed into their respective feature encoder networks, which consist of CNN and BiLSTM encoders. The extracted MFCC and spectrogram features from each encoder are concatenated. Then, they are combined with the wav2vec 2.0 embedding using guided-attention manner [25] to obtain the final speech embedding.

2.2. Crossmodal Transformer with FA Mechanism

Inspired by crossmodal transformer networks [3] developed for tri-modal emotion recognition, we apply the same concept to our bi-modal tasks. We employ a crossmodal transformer block for the bi-modal fusion network. This transforms latent information from one modality to another by iteratively enhancing the features of one modality using the features of the other, and vice versa, through a multi-head attention mechanism. The FA is an attention mechanism that enables neural network models to selectively attend to the most important parts of an input sequence by employing a Gaussian distribution-based focal bias [15]. In our study, the FA mechanism at the crossmodal transformer block detects the salient information in another modality during encoding. For example, when the crossmodal transformer block provides a latent adaptation from text (L) modality to speech (A) modality (denote $L \rightarrow A$), it helps to focus on which words (speech frame; the opposite case $A \rightarrow L$) are important for emotion classification. This mechanism models a focal bias by adding a regularization term determined by the center position and coverage scope to the attention score.

As shown in Figure 1(b), the input to the modality fusion

module consists of the embedding of each modality that were obtained in the previous subsection. We denote the speech and text embedding as $X_A = \{x_i^A\}_{i=1,...,T_A} \in \mathbb{R}^{T_A \times d_A}$ and $X_L = \{x_i^L\}_{i=1,...,T_L} \in \mathbb{R}^{T_L \times d_L}$, respectively. In the process of computation, three matrices query $Q_A \in \mathbb{R}^{T_A \times d_k}$, key $K_L \in \mathbb{R}^{T_L \times d_k}$, and value $V_L \in \mathbb{R}^{T_A \times d_v}$ are obtained firstly by the linear projection. To obtain the focal bias, at the i-th sequence step, the center position scalar $\mu_i \in \mathbb{R}$ and coverage scope scalar $\sigma_i \in \mathbb{R}$ are calculated using two linear projection processes:

$$\mu_i = U_c^T \tanh(W_p K_i + W_g G), \tag{1}$$

$$\sigma_i = U_d^T \tanh(W_p K_i + W_g G), \tag{2}$$

where $W_p \in \mathbb{R}^{d_k \times d_k}$ and $W_p \in \mathbb{R}^{d_k \times d_k}$ are two learnable shared weights, and $U_c \in \mathbb{R}^{d_k}$ and $U_d \in \mathbb{R}^{d_k}$ are two different linear projection weight vectors. T_L is the length of one modality. $G = \frac{1}{T_L}\Sigma_{i=1}^{T_L} K_i$ is the mean vector to provide complementary information. Moreover, we regulate the value of μ_i and σ_i to closed interval $[0, T_L]$,

$$\widetilde{\mu}_i = T_L * sigmoid(\mu_i), \tag{3}$$

$$\widetilde{\sigma}_i = T_L * sigmoid(\sigma_i). \tag{4}$$

From the definition of Gaussian distribution, the focal bias for the i-th step $f_{i,j} \in \mathbb{R}^{T_L \times T_A}$ is obtained with $\tilde{\mu}_i$ and $\tilde{\sigma}_i$ as follows:

$$f_{i,j} = -\frac{(P_j - \widetilde{\mu}_i)^2}{(\widetilde{\sigma}_i)^2/2}, \tag{5}$$

where $i \in \{1, 2, ..., T_L\}$, $j \in \{1, 2, ..., T_A\}$, P_j is the absolute position of speech embedding vector x_j^A in speech modality. Eventually, this focal bias is added to the crossmodal attention ($L \rightarrow A$) weight before softmax function.

$$\begin{aligned} Attention(Q_A, K_T, V_T) &= softmax\left(\frac{Q_A K_L^\top}{\sqrt{d_k}} \oplus f^\top\right) V_L \\ &= softmax\left(\frac{Z_A W_{Q_A} W_{K_L}^\top Z_L^\top}{\sqrt{d_k}} \oplus f^\top\right) Z_L W_{V_L} \\ Z_{\{A,L\}} &= Conv1D(X_{\{A,L\}}) + PE(T_{\{A,L\}}), \end{aligned} \tag{6}$$

where $Z_{\{A,L\}}$ is the position-aware feature, PE is positional embedding, and $\oplus$ denotes element-wise summation.

2.3. Joint Loss

In our experiment, The triplet loss is used as the metric loss. The triplet loss function is built to optimize the embedding space by minimizing the distance between an anchor and a positive sample (i.e., with the same emotion label as that of the anchor), while maximizing the distance between the same anchor and a negative sample (i.e., different emotion label). Thus, our method uses the dissimilarities between the samples to enhance feature discrimination. The embedding vector of input x, which is the output of the last FC layer is represented as $\mathcal{F}(x) \in \mathbb{R}^d$. Subsequently, the goal of the training process is to minimize the dissimilarity between an anchor utterance x_i^a and all positive utterances x_i^p, while maximizing the dissimilarity between x_i^a and any negative utterance x_i^n. The final embedding space is defined by the following equations:

$$\begin{aligned} \|\mathcal{F}(x_i^a) - \mathcal{F}(x_i^p)\|_2^2 + \mathcal{M} < \|\mathcal{F}(x_i^a) - \mathcal{F}(x_i^n)\|_2^2, \\ \forall(\mathcal{F}(x_i^a),\ \mathcal{F}(x_i^p),\ \mathcal{F}(x_i^n)) \in \Gamma, \end{aligned} \tag{7}$$

where $\mathcal{M}$ denotes the margin which is enforced between positive and negative pairs and Γ is the set of all possible triplets $(\mathcal{F}(x_i^a), \mathcal{F}(x_i^p), \mathcal{F}(x_i^n))$ in the training set with cardinality N. The metric loss to be minimized is defined as follows:

$$\begin{aligned} D_{pos} &= \|\mathcal{F}(x_i^a) - \mathcal{F}(x_i^p)\|_2^2, \\ D_{neg} &= \|\mathcal{F}(x_i^a) - \mathcal{F}(x_i^n)\|_2^2, \\ L_{metric} &= \tfrac{1}{N}\textstyle\sum_{i \in \Gamma}^{N} [D_{pos} - D_{neg} + \mathcal{M}]. \end{aligned} \tag{8}$$

During model training, both triplet and cross-entropy losses are weighted using the coefficients α and β, respectively. The joint loss L is calculated as follows:

$$L = \alpha L_{metric} + \beta L_{ce}, \tag{9}$$

where L_{ce} is the cross-entropy loss, which is defined as follows:

$$L_{ce} = -\sum_{k=1}^{K} \hat{y}_k \log(y_k), \tag{10}$$

where y denotes the softmax output and $\hat{y} \in \{0, 1\}^K$ is a one-of-K label vector.

3. Experiments and Evaluation Results

3.1. Dataset

The performance of our model is evaluated using the IEMOCAP dataset [26], which is designed to replicate natural dyadic interactions between actors based on theatrical theory. The IEMOCAP dataset comprises five sessions, with each session featuring utterances from two speakers (one male and one female), resulting in ten unique speakers. We apply categorical evaluations with majority agreement and considered only four emotional categories: happy, sad, angry, and neutral - to compare the performance of our model with that of previous researches using the same categories [6, 14, 13]. To ensure a consistent comparison with the results of previous studies [6, 14, 13], we combine the excitement dataset with the happiness dataset. Our final dataset include 5,531 utterances (1,636 happy, 1,084 sad, 1,103 angry, and 1,708 neutral).

3.2. Experimental Setup

In our study, we use two SSL models to embed each modality, specifically, the wav2vec 2.0 and RoBERTa models. The pre-trained "wav2vec2-base" model [1] and RoBERTa[2] are available online. The MFCC is a 40-dimensional feature structure including human auditory characteristics based on HTK-style Mel frequencies extracted from raw speech segments using the librosa library [27]. To obtain the spectrogram, we applied a series of 40 ms Hamming windows with a hop length of 10 ms and treated each windowed block as a frame. The SER system is implemented using PyTorch. We use Adam as the optimizer, with a learning rate of $1e-5$ and a training batch size of 32.

[1] https://huggingface.co/facebook/wav2vec2-base

[2] https://github.com/facebookresearch/fairseq

Table 1: *Results of the ablation study for the individual components of our framework on the IEMOCAP dataset.*

Model	Method	WA	UA
A	Full model	**0.774**	**0.777**
B	Full model without FA mechanism	0.762	0.755
C	Full model without Metric Loss	0.752	0.730

Table 2: *Impact on performance of various weights for the metric and CE loss evaluated using WA and UA. α and β are the weights of the metric and cross-entropy loss, respectively.*

	Weight		Metric	
Loss	α	β	WA	UA
CE Loss	0	1	0.752	0.730
Joint Loss	0.5	1	**0.771**	**0.763**
	1	0.5	0.764	0.756
	1	1	0.761	0.762

Table 3: *Experimental results on the effect of the position of embedding applying the metric loss.*

Embedding Space	WA	UA
Speech + Text Embedding	0.771	0.763
Speech Embedding	0.774	0.777
Text Embedding	0.756	0.749

Table 4: *Classification performance of the proposed and other relevant multimodal SER models on the dataset IEMOCAP.*

Model	WA	UA
LSTM+Attn [11]	0.725	0.709
KS-transformer [14]	0.743	0.753
WISE [13]	0.759	0.764
LM-MulT [12]	0.768	0.771
MHA-2 [7]	0.765	0.776
Ours	**0.774**	**0.777**

A dropout with $p = 0.25$ was used to alleviate over-fitting. The PML (pytorch-metric-learning) library [28] is used to apply metric loss.

To evaluate the performance of our system, we calculate the weighted accuracy (WA), which measures the overall classification accuracy, and unweighted accuracy (UA), which measures the average recall across various emotion categories.

3.3. Ablation Study

To assess the individual impact of each component of our method, we create two additional comparison systems via ablation. By doing so, we aim to better understand the significance of each component and its contribution to the overall model performance. Table 1 presents the results of this analysis. Model A is the proposed model. Models B and C are derived from model A by removing the FA mechanism from the crossmodal transformer and metric loss, respectively. To verify that the impact of focus-attention mechanism that aligns modalities in crossmodal transformer, we compare the performances of models A and B. The experimental results presented in Table 1 show that model A outperforms model B, reporting improvements of 1.2% and 2.2% for WA and UA, respectively. We can conclude that the proposed method enables improved detection of salient parts in other modality when cross-modal information provides latent adaptation across modalities. To verify the effect of metric loss, we comparatively assess the performance of models A and C. Compared to model C, model A reported a significant improvement of 2.2% and 4.74% for WA and UA, respectively. This outcome indicates that an appropriate choice of the triplet loss enhances the performance of a classifier.

Table 2 shows the outcomes of a study that examined the influence of weights on the effectiveness of metric and cross-entropy loss in a combined loss task. The model shows the best performance when α is 0.5 and β is 1. The performance is better when the CE and triplet losses are learned jointly than when the CE loss is used alone.

Table 3 reports the experimental results for the various embedding spaces on which the proposed metric loss method is applied. We identify three distinct embedding spaces suitable for our model. The first is the embedding in which the text-enhanced speech and speech-enhanced text representations are concatenated, which is passed through each of the crossmodal transformers. The second and third embedding spaces are the speech and text embedding, respectively, considered before they are sent as input to the crossmodal transformers. We observe that the application of the metric loss to speech embedding resulted in a slight improvement in performance compared with when the metric loss was applied to fused speech and text embedding. Conversely, when the metric loss was applied solely to text embedding, we observed a slight decrease in the performance compared with that of the metric loss based on fused speech and text. This may be caused by the lack of emotion data to disentangle the text embedding into the emotion embedding space by using metric loss.

3.4. Comparison with the Other Relevant Multimodal SER Models

Additional experiments were conducted to validate the efficacy of the proposed model by comparing it with other advanced SER methods using the IEMOCAP dataset. Table 4 presents the experimental results of the comparison, showing that the proposed approach outperforms the other relevant multimodal SER models in both WA and UA. This is due to the effective joint loss and the exceptional ability of the multimodal fusion method enhanced by FA to attend to the salient parts of each modality. These results confirm that the proposed approach can be used to improve the performance of SER.

4. Conclusions

In this paper, to enhance the efficiency of a multimodal SER system, we introduced a novel cross-modal transformer architecture that incorporates an FA mechanism and jointly metric loss method, demonstrating a superior emotional embedding efficacy. Extensive experiments using the IEMOCAP dataset showed that the proposed model outperforms other relevant multimodal SER methods, achieving an average performance of 77.4% and 77.7% for WA and UA, respectively. Additionally, we analyzed the effects on the model performance of various emotion embedding spaces by using the metric loss.

5. References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] F. Haider and S. Luz, "An automated mood diary for older user's using ambient assisted living recorded speech," *Proc. Interspeech 2022*, pp. 1961–1962, 2022.

[3] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[4] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," *Proc. Interspeech 2020*, pp. 364–368, 2020.

[5] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, and L.-P. Morency, "Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences," in *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 2021*, 2021.

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

[7] 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.

[8] 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*, 2019, pp. 157–166.

[9] W. Yang, S. Fukayama, P. Heracleous, and J. Ogata, "Exploiting fine-tuning of self-supervised learning models for improving bi-modal sentiment analysis and emotion recognition," *Proc. Interspeech 2022*, pp. 1998–2002, 2022.

[10] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," *Proc. Interspeech 2020*, pp. 3755–3759, 2020.

[11] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *Proc. Interspeech 2019*, pp. 3569–3573, 2019.

[12] Y. Wang, G. Shen, Y. Xu, J. Li, and Z. Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition." in *Interspeech*, 2021, pp. 4518–4522.

[13] G. Shen, R. Lai, R. Chen, Y. Zhang, K. Zhang, Q. Han, and H. Song, "Wise: Word-level interaction-based multimodal fusion for speech emotion recognition." in *Interspeech*, 2020, pp. 369–373.

[14] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6897–6901.

[15] Y. You, W. Jia, T. Liu, and W. Yang, "Improving abstractive document summarization with salient information modeling," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2132–2141.

[16] W. Yang and J. Ogata, "Stronger baseline for robust results in multimodal sentiment analysis," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, 2021, pp. 41–50.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[20] Z. Zhao, Y. Wang, and Y. Wang, "Multi-level fusion of wav2vec 2.0 and bert for multimodal emotion recognition," *arXiv preprint arXiv:2207.04697*, 2022.

[21] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[23] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech 2021*, pp. 3400–3404, 2021.

[24] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7367–7371.

[25] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6281–6290.

[26] 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, pp. 335–359, 2008.

[27] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[28] W. De Vazelhes, C. Carey, Y. Tang, N. Vauquier, and A. Bellet, "metric-learn: Metric learning algorithms in python," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5447–5452, 2020.