A Study on Multimodal Fusion and Layer Adapter in Emotion Recognition

Xiaohan Shi*,Yuan Gao†, Jiajun He*, Jinyi Mi*, Xingfeng Li‡, Tomoki Toda*
* Nagoya University, Japan.
E-mail: {xiaohan.shi, jiajun.he, mi.jinyi}@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp
† Kyoto University, Japan.
E-mail: gao.yuan.75x@st.kyoto-u.ac.jp
‡ City University of Macau, Macau.
E-mail: xfli@cityu.edu.mo

***Abstract*—Multimodal emotion recognition (MER) is a rapidly evolving field aimed at integrating information from various modalities, such as speech and text, to deepen our understanding of emotions. However, challenges in feature extraction and fusion hinder further advances in MER performance. To address these challenges, we propose different modality representations to capture emotional information comprehensively. Additionally, we introduce a novel layer adapter and multimodal fusion method to explore modality-dependent and modality-invariant interactions in MER. Extensive experimental results show the effectiveness of our approach, achieving state-of-the-art results with an absolute improvement of 1.89% over the baseline. The significant improvements validate the effectiveness of our proposed method in enhancing the MER system.**

I. Introduction

Emotion recognition is a growing field driven by the advancements in human-computer interactions (HCI) [1]. Earlier research focused mainly on using speech signals for emotion recognition, but recent studies show that unimodal systems are inadequate for complex HCI contexts. To improve emotion recognition accuracy, recent studies advocate integrating diverse modalities, including textual and visual data, to foster a comprehensive understanding of human emotional states [2]–[4]. Multimodal emotion recognition (MER) has applications in various domains, including social media analytics [5], market research [6], and enhancing customer service experiences [7]. Through its interdisciplinary implications, MER holds promise in uncovering novel insights into human emotional dynamics, thereby paving the path for innovative solutions across a myriad of practical domains. Despite substantial progress, two key challenges remain: extracting effective acoustic and lexical features for distinguishing emotions and developing suitable fusion methods to integrate multiple modalities in emotion recognition. This study endeavors to address each of these challenges, aiming to model the emotion recognition process in a multimodal context.

In MER, most features are employed in the speech and text modalities as these modalities predominantly promote emotion recognition within existing datasets. For example, Majumder et al. [8] propose a multimodal sentiment analysis approach based on hierarchical fusion and context modeling, enhancing sentiment recognition accuracy and robustness. Chuang et al. [9] introduce a multi-modal deep learning model for integrating speech and text data, achieving significant performance improvements in emotion recognition tasks.

In recent years, with the emergence of self-supervised learning (SSL) models, more researchers have shifted towards using pre-trained SSL features instead of traditional deep learning features to enhance the performance [10]. For example, Zou et al. [11] incorporated multiple acoustic features, including Mel-frequency cepstral coefficients, spectrograms, and Wav2vec 2.0 embeddings, for categorical emotion recognition, achieving a 7.03% improvement compared to using Wav2vec 2.0 embeddings alone. Additionally, Padi et al. [12] present a MER framework using mel-spectrogram and fine-tuning pre-trained BERT models, providing complementary emotional insights from the speech and text modalities.

Motivated by these findings, this paper focuses on extracting multimodal emotion features by integrating general speech and text self-supervised representations and context representations from the automatic speech recognition (ASR) module to enhance the MER system. Fusion methods are another key aspect of this task. In the literature, feature-level and decision-level fusion are widely used for multimodal systems. For instance, Yoon et al. [13] proposed a deep dual recurrent neural network to encode audio-text sequences, then concatenated their outputs to predict the emotion. Similarly, Pepino et al. [14] designed multiple dual RNNs to represent audio-text sequences and compared feature-level fusion and decision-level fusion approaches, highlighting their comparable performance. In contrast, our study introduces a fusion method at both the feature level and decision level, integrating a multi-level fusion module to extract discriminative features for MER.

Our contributions can be summarized as follows:

- We propose a novel feature extraction approach for learning multimodal emotion information by leveraging multimodal self-supervised representations.
- We introduce a multimodal fusion module that comprehensively integrates modality-dependent and modality-invariant emotional information from both speech and text modalities.

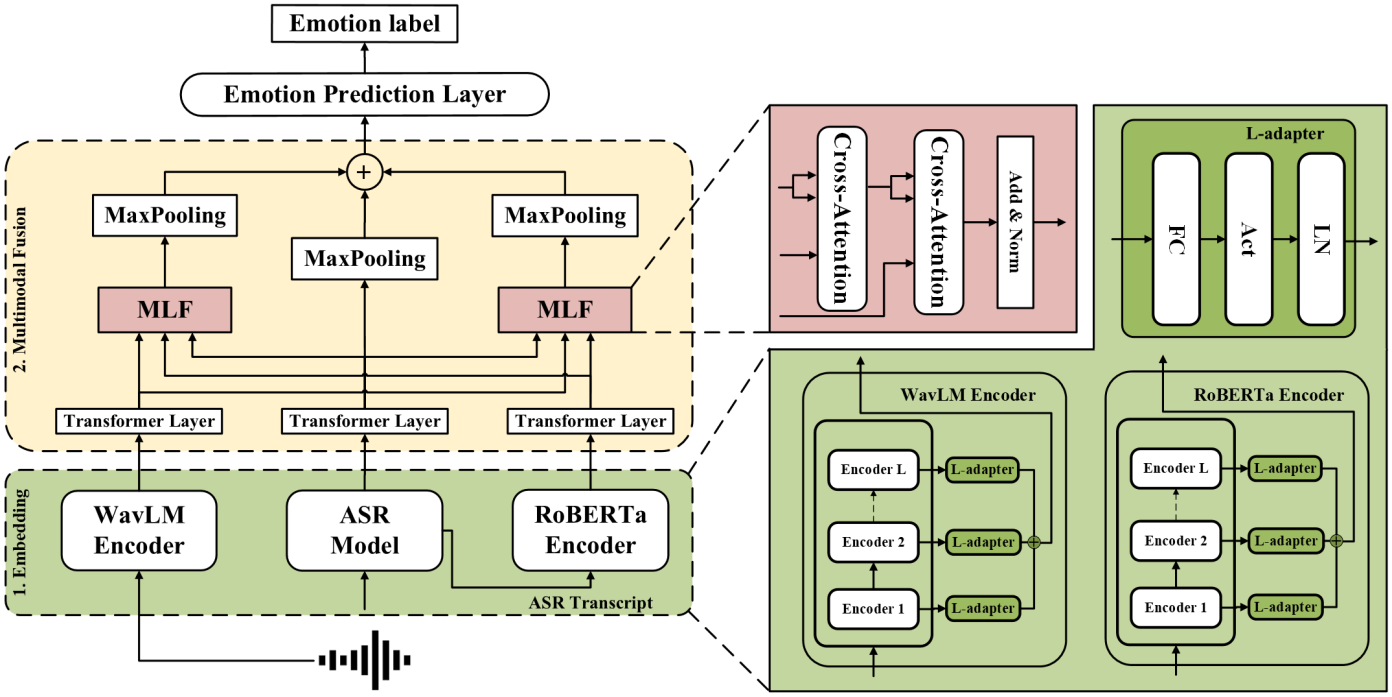

Fig. 1. The overall architecture of our proposed method.

- Our experimental results demonstrate that the proposed approach effectively addresses MER tasks.

II. Proposed Method

In this section, we outline our MER system and detail the multimodal fusion method. As depicted in Fig. 1, the network comprises four primary components: an ASR module for extracting textual and contextual information from raw speech, two embedding modules for encoding self-supervised representations, a multimodal fusion module for integrating modality-dependent and modality-invariant emotional information, and an emotion prediction module for predicting the emotion label.

A. Model Description

As depicted, raw audio utterances are directed into an acoustic feature encoder to extract self-supervised representations. Simultaneously, an ASR module is employed to obtain textual and contextual information. The transcripts then undergo a textual feature encoder to extract text self-supervised representations. These different modality representations are subsequently combined using the proposed multimodal fusion method, streamlining the final emotion recognition process.

B. Problem Formulation

The overall system can be expressed as the function $f(S, T, C) = L$, where each of the speech and context features $S = (s_1, s_2, \cdots, s_m)$ and $C = (c_1, c_2, \cdots, c_m)$ consists of m frames extracted from an utterance. The text modality $T = (t_1, t_2, \cdots, t_n)$ represents the original ASR hypotheses of an utterance, comprising n tokens. The model output $L \in \{l_1, l_2, \cdots, l_e\}$, where e represents the emotional categories.

C. Embedding Module

1) Speech Representations: To obtain a comprehensive understanding of acoustic features, we employ a pretrained SSL model, WavLM [15], as our speech self-supervised encoder.

WavLM employs a hybrid architecture consisting of convolutional neural network layers and a transformer encoder to capture speech features and contextual information effectively. We denote $H_S = (h_S^{(1)}, h_S^{(2)}, \cdots, h_S^{(m)})$ to represent the speech self-supervised representations, where m denotes the number of frames extracted from an utterance.

2) Context Representations: To acquire comprehensive information regarding contextual information from speech, we employ ASR representation as context features. In this study, we leverage Whisper [16] as our ASR module. Whisper is a supervised model specifically designed to transcribe spoken language into text. It has been trained on a diverse dataset comprising approximately 680,000 hours of speech collected from the web.

Whisper adopts an encoder-decoder transformer architecture; the encoder encodes the mel spectrogram information, and the decoder generates text as a sequence of words. In whisper, first, the raw audio inputs are converted to a log-mel spectrogram by the action of the feature extractor. Then, the Transformer encoder encodes the spectrogram to form a sequence of encoder-hidden states. Finally, the decoder

autoregressively predicts text tokens, conditional on both the previous tokens and the encoder's hidden states.

We denote $H_C = (h_C^{(1)}, h_C^{(2)}, \cdots, h_C^{(m)})$ to represent the context representations from whisper encoder layer, where *m* denotes the number of frames extracted from an utterance.

3) Text Representations: To acquire comprehensive information regarding lexical features, we leverage a pretrained SSL model, RoBERTa [17], as our text encoder. RoBERTa is an extension of the bidirectional encoder representations from the transformers model, which is specifically designed to address challenges related to long-range dependencies, and finely tuned for various natural language processing tasks. Pretrained on extensive corpora, including a dataset comprising 58 million tweets, RoBERTa exhibits exceptional contextual understanding, thereby enhancing text-related tasks. We denote $H_T = (h_T^{(1)}, h_T^{(2)}, \cdots, h_T^{(n)})$ to represent the text self-supervised representations, where *n* denotes the number of tokens extracted from an utterance.

4) L-adapter module (LA): To harness intermediate representations from the initial fine-tuning stages, layer adapters establish pathways from each encoder from the speech and text self-supervised model. Each layer adapter comprises a Fully connected (FC) layer, succeeded by a non-linear activation function and layer normalization, as illustrated in Fig. 1b. The application of the layer adapter results in adapted representations, computed as follows:

$$\alpha_S^l(\text{ or } \alpha_T^l) = \text{FC}((H_S^l) \text{ or } H_T^l) \tag{1}$$

for $l = 1, 2, \ldots, L$ in the encoder, and the weighted sum of the adapted representations is computed as:

$$H_S^*(\text{ or } H_T^*) = \sum_{l=1}^{L} w_l \alpha_S^l(\text{ or } \alpha_T^l) \tag{2}$$

This is fed into the emotion recognition classification, where w_l are learnable weights.

D. Multimodal Fusion (MF) Module

Inspired by [18], our MF is composed of two Multi-Level Fusion (MLF) blocks. The objective is to facilitate the learning of modality-dependent representations and modality-invariant representations.

In this section, we provide an in-depth explanation of the operation of each MLF block.

MLF Block adheres to the structure of a standard transformer layer, incorporating two cross-attention modules, and residual connections.

Initially, we employ three multi-head transformer layers to derive speech, text, and context representations.

$$H_S'(\text{ or } H_T', H_C') = \text{Transformer}(H_S^*(\text{ or } H_T^*, H_C)) \tag{3}$$

Then, two MLF blocks are used to derive speech, text self-supervised representations, and context representation. This is achieved by utilizing H_S' (or H_T') as queries and H_T' (or H_S') as keys and values within the first cross-attention block.

$$Q_1 = H_S'(\text{ or } H_T'), K_1 = H_T'(\text{ or } H_S'), V_1 = H_T'(\text{ or } H_S'). \tag{4}$$

$$H_S^{\text{T}}(\text{ or } H_T^{\text{S}}) = \text{Cross-Attention}(Q_1, K_1, V_1). \tag{5}$$

Next, the first cross-attention block output is used as queries and H_C' as keys and values within the second cross-attention block.

$$Q_2 = H_S^{\text{T}}(\text{ or } H_T^{\text{S}}), K_2 = H_C', V_2 = H_C'. \tag{6}$$

$$H_C^{\text{ST}}(\text{ or } H_C^{\text{TS}}) = \text{Cross-Attention}(Q_2, K_2, V_2). \tag{7}$$

Finally, we adopt Max pooling to obtain a 1-dimensional vector for each output. The final speech, context and text representations (H_C^{ST}, H_C^{TS}, H_C') are concatenated and written as follows:

$$H_{STC} = H_C^{\text{ST}} \oplus H_C^{\text{TS}} \oplus H_C'. \tag{8}$$

E. Emotion Classification Module

Emotion classification is conducted using the output representations H_{STC} of the MF module, which is subsequently passed through a FC layer and a SoftMax activation function.

$$P(y_{\text{emo}}|H_{STC}) = \text{SoftMax}(\text{FC}(H_{STC})). \tag{9}$$

where y_{emo} is the predicted emotion classification.

The loss function $\mathcal{L}_{\text{Emotion Classification}}$ is formulated using cross-entropy:

$$\mathcal{L}_{\text{Emotion Classification}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}) \tag{10}$$

where N denotes the total number of samples, C is the count of emotion classes, y_{ic} is the ground truth label for sample i and class c, and $\hat{y}_{ic}$ is the predicted probability of class c for sample i. This loss function penalizes the model based on the discrepancy between predicted probabilities and ground truth labels across all samples and classes, facilitating accurate classification.

III. Experimental Setup

A. Dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is a widely utilized corpus in MER [19]. It contains 12 hours of audiovisual data, including audio, video, and textual transcriptions from 10 speakers. In each session, one male and one female performed a series of scripts or improvisational scenarios. For each utterance, three annotators assigned the categorical labels. Following common practice, we merged 'happy' and 'excited' into a single emotion class labeled 'happy'. We implemented the common practice of merging "happy" and "excited" into one emotion class "happy"; thus, the emotion labels in this dataset are happy (29%), sad (20%), angry (20%), and neutral (31%).

B. Experimental Procedure

We conduct two experiments in this study. Experiment 1 examines the impact of different modality representations on MER by comparing our approach with previous studies:

- **SAWC** [20]: This method modifies the importance weights based on confidence measures, thereby reducing the impact of ASR errors by focusing on relevant speech segments.
- **RMSER-AEA** [21]: This approach incorporates complementary semantic information, adjusts for ASR errors via an auxiliary task, and combines text and acoustic representations for SER.
- **SAMS** [22]: This approach utilizes high-level emotion representations as supervisory signals to establish a multi-spatial learning framework for each modality, facilitating cross-modal semantic learning and the exploration of fusion representations.

Meanwhile, we provide ablation experiments for each modality representation by visualizing the feature distributions using t-distributed stochastic neighbor embedding (t-SNE) [23], as shown in Fig. 2.

In Experiment 2, we explore the performance of the proposed fusion methods, namely Layer Adapter (LA) and Multimodal Fusion (MF), across different combinations of modality representations.

C. Implementation

Our deep learning models were developed using Python 3.7 and PyTorch 1.11.0. The speech self-supervised encoder was initialized using the WavLM-base model[1], producing speech self-supervised representations with a dimensionality of 768. For the text self-supervised encoder, we utilized the RoBERTa-base model[2], which has a hidden size of 768, 12 attention layers, and 12 attention heads. The context encoder was initialized using the Whisper-small model[3], resulting in context representations with a dimensionality of 768. The word error rate (WER) is 33.87% on the IEMOCAP dataset. the WavLM and RoBERTa models were fine-tuned, while the Whisper model was frozen during the training process. We used the Adam optimizer [24], with a dropout rate of 0.5 and a weight decay of 1e-5 to avoid overfitting and ensure model generalization. During training, we used a batch size of 32 and trained for 30 epochs.

D. Evaluation

To evaluate our results on the IEMOCAP dataset, which lacks a standard train/dev/test split, we adopt a leave-one-section-out cross-validation approach, consistent with prior studies [25]–[27]. We assess categorical MER performance using Unweighted Average Recall (UAR) and F1 scores, metrics commonly employed in experiments with imbalanced data to gauge performance [28]–[30] across the four discrete emotional labels.

[1]https://huggingface.co/microsoft/wavlm-base

[2]https://huggingface.co/FacebookAI/roberta-base

[3]https://huggingface.co/openai/whisper-small.en

IV. Results and Discussion

To analyze the different modality representations for MER, we compare the recognition performance of single modality and multimodal system, as shown in Table 1.

TABLE I
COMPARISON RESULTS OF OUR PROPOSED METHOD AND PREVIOUS MER STUDIES.

Modality	Model	UAR (%)	F1 (%)
Single modal	Speech (WavLM)	63.55	64.26
	Text (RoBERTa)	69.29	68.77
	Context (Whisper)	71.66	70.85
Multimodal	SAWC [20]	76.6	-
	RMSER-AEA [21]	76.4	-
	SAMS [22]	76.6	-
	Proposed	**77.73**	**77.24**
	w/o Speech	74.81	74.71
	w/o Text	67.21	67.01
	w/o Context	76.31	76.29

The effectiveness of incorporating features from multiple modalities is evaluated. The results demonstrate the effectiveness of multi-modal feature extraction. The improvements in UAR and F1 scores were 14.18% and 12.98% respectively, when compared to using only speech representation. When compared to using only text representation, the improvements are 8.44% and 8.47%, respectively, and compared to using only context representation, the improvements are 6.07% and 6.39%, respectively. Furthermore, compared with the MER previous studies, our proposed model achieves a 1.13% absolute improvement in UAR. Ablation results indicate that the most critical representation is the text representation, with its removal causing a 10.52% and 10.23% decrease in UAR and F1 scores, respectively. On the other hand, the context representation has the least impact on performance, leading to a 1.42% and a 0.95% decrease in UAR and F1 scores, respectively. These results highlight the significant role of multimodal representations in enhancing MER performance. The text modality is particularly influential. The minor impact of context representation suggests that its contribution is not as pivotal as that of speech and text. Furthermore, the feature distributions under the different modality representations and the proposed method are visualized using t-SNE, as depicted in Fig. 2. This visualization provides additional evidence supporting the effectiveness of our proposed method.

To analyze the effects of MF and LA on MER, we compared their performance across different modalities in detail, as shown in Table 2.

The results indicate that the combination of MF and LA yields superior performance compared to using only either MF or LA. Specifically, when utilizing speech and text representations, the observed improvements in UAR and F1 scores are 2.45% and 2.64%, respectively, compared to using only MF. With LA, these improvements are 1.75% and 3.08%,

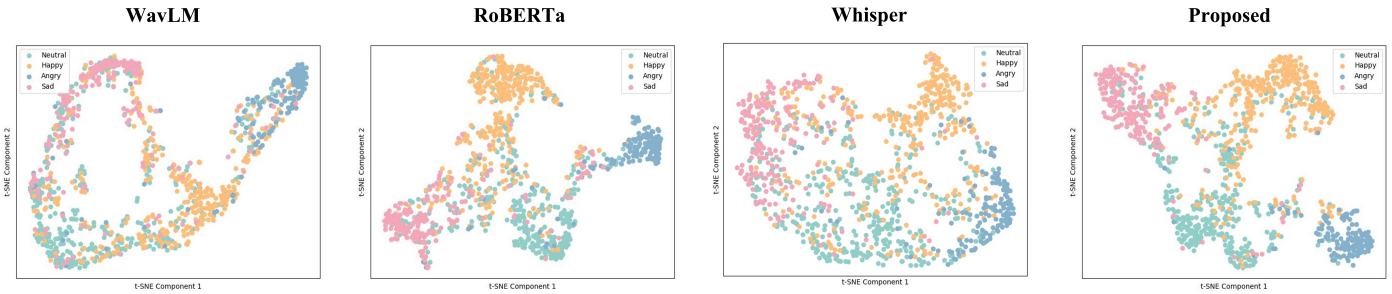

Fig. 2. The t-SNE visualization of different modality representations on IEMPOCAP session 2.

TABLE II
COMPARISON OF MER PERFORMANCE OBTAINED BY THE PROPOSED MF AND LA FUSION METHODS.

Model	Method	UAR (%)	F1 (%)
Speech + Text	-	72.27	72.06
	MF	73.86	73.65
	LA	74.56	73.21
	MF + LA	**76.31**	**76.29**
Speech + Context	-	65.04	64.39
	MF	66.89	66.59
	LA	66.74	66.94
	MF + LA	**67.21**	**67.01**
Text + Context	-	72.65	72.20
	MF	73.86	73.65
	LA	73.90	73.65
	MF + LA	**74.81**	**74.71**
Speech + Text + Context	-	75.84	75.59
	MF	76.83	76.24
	LA	77.61	77.05
	MF + LA	**77.73**	**77.24**

respectively. In the speech and context representations, the enhancements in UAR and F1 scores are 0.32% and 0.42%, respectively, when using only MF, and 0.47% and 0.07%, respectively, with LA. For text and context representations, the improvements in UAR and F1 scores are 0.95% and 1.06%, respectively, with MF, and 0.91% and 1.06%, respectively, with LA. Finally, when employing speech, text, and context representations, the observed enhancements in UAR and F1 scores are 0.9% and 1%, respectively, with MF, and 0.12% and 0.19%, respectively, with LA. In summary, our analysis highlights that the combination of MF and LA consistently improves MER performance across various modality combinations.. This finding highlights the significant potential of integrating multimodality representation to achieve robust and reliable emotion recognition.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a multimodal self-supervised representation extraction and fusion approach to capture emotional information comprehensively. In the proposed method, we introduce a novel layer adapter to explore modality-dependent and modality-invariant interactions in MER. Our findings reveal that the proposed representations across different modalities perform better than single-modality representations. Moreover, our novel layer adapter and modality fusion method for integrating modality-invariant emotional information consistently achieves higher accuracy in MER.

For future research, we advocate for further exploration of innovative modality fusion methods to further enhance the accuracy of MER.

VI. ACKNOWLEDGMENT

This work was financially supported by JST SPRING, Grant Number JPMJSP2125, and in part by JST CREST Grant Number JPMJCR19A3, Japan, and JSPS KAKENHI Grant Number 21H05054.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Speech emotion recognition with multi-level acoustic and semantic information extraction and interaction," in *Interspeech 2024*, 2024, pp. 1060–1064.

[3] H. Sun, S. Zhao, X. Wang, W. Zeng, Y. Chen, and Y. Qin, "Fine-grained disentangled representation learning for multimodal emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 051–11 055.

[4] J. Tian, D. Hu, X. Shi, *et al.*, "Semi-supervised multimodal emotion recognition with consensus decision-making and label correction," in *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing*, 2023, pp. 67–73.

[5] N. Andalibi and J. Buss, "The human in emotion recognition on social media: Attitudes, outcomes, risks," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–16.

[6] S. Lugović, I. Duner, and M. Horvat, "Techniques and applications of emotion recognition in speech," in *2016 39th international convention on information and communication technology, electronics and microelectronics (mipro)*, IEEE, 2016, pp. 1278–1283.

[7] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan, "Ordinal learning for emotion recognition in customer service calls," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2020, pp. 6494–6498.

[8] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-based systems*, vol. 161, pp. 124–133, 2018.

[9] Z.-J. Chuang and C.-H. Wu, "Multi-modal emotion recognition from speech and text," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, 2004, pp. 45–62.

[10] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, "Using Semi-supervised Learning for Monaural Time-domain Speech Separation with a Self-supervised Learning-based SI-SNR Estimator," in *Proc. INTERSPEECH 2023*, 2023, pp. 3759–3763. DOI: 10.21437/Interspeech.2023-85.

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

[12] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models," *arXiv preprint arXiv:2202.08974*, 2022.

[13] 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.

[14] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6484–6488.

[15] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," *arXiv preprint arXiv:2307.03183*, 2023.

[17] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[18] J. He, X. Shi, X. Li, and T. Toda, "Mf-aed-aec: Speech emotion recognition by leveraging multimodal fusion, asr error detection, and asr error correction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 066–11 070.

[19] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[20] S. Dutta and S. Ganapathy, "Multimodal transformer with learnable frontend and self attention for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6917–6921.

[21] B. Lin and L. Wang, "Robust multi-modal speech emotion recognition with asr error adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[22] M. Hou, Z. Zhang, C. Liu, and G. Lu, "Semantic alignment network for multi-modal emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 5318–5329, 2023.

[23] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

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

[25] X. Shi, X. Li, and T. Toda, "Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation," in *Proc. Interspeech*, 2023, pp. 765–769.

[26] H. Sun, S. Zhao, X. Kong, *et al.*, "Iterative prototype refinement for ambiguous speech emotion recognition," in *Interspeech 2024*, 2024, pp. 3200–3204.

[27] X. Shi, X. Li, and T. Toda, "Multimodal fusion of music theory-inspired and self-supervised representations for improved emotion recognition," pp. 2024–2350, 2024.

[28] Y. Gao, L. Wang, J. Liu, J. Dang, and S. Okada, "Adversarial domain generalized transformer for cross-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2023.

[29] X. Shi, S. Li, and J. Dang, "Dimensional emotion prediction based on interactive context in conversation.," in *INTERSPEECH*, 2020, pp. 4193–4197.

[30] X. Li, X. Shi, D. Hu, *et al.*, "Music theory-inspired acoustic representation for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2534–2547, 2023.