Interspeech 2024
1-5 September 2024, Kos, Greece

Multimodal Fusion of Music Theory-Inspired and Self-Supervised Representations for Improved Emotion Recognition

Xiaohan Shi[1], *Xingfeng Li*[2], *Tomoki Toda*[1]

[1]Nagoya University, Japan
[2]Hainan University, China

xiaohan.shi@g.sp.m.is.nagoya-u.ac.jp, lixingfeng@hainanu.edu.cn,
tomoki@icts.nagoya-u.ac.jp

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 advancements in MER performance. To address these challenges, we propose a MER method using self-supervised representations and handcrafted music theory-inspired representations across different modalities to comprehensively capture emotional information. Additionally, we introduce a novel multimodal fusion method to explore modality-specific and modality-invariant relationships, thereby reducing distribution gaps between different modalities in MER. Extensive experimental validation underscores the effectiveness of our approach, with state-of-the-art results showing a 3.55% improvement compared with the baseline. These results validate the effectiveness of our proposed method, signifying a notable enhancement in MER performance.

Index Terms: multimodal emotion recognition, multimodal fusion, self-supervised learning

1. Introduction

Emotion recognition has been a growing area of focus within affective computing for interpreting a human subject's feelings, thoughts and behavioral responses. With the widespread use of social networks, more and more people tend to express their feelings by sharing speech on the internet. Therefore, a considerable amount of work has recently focused on multimodal approaches that utilize both speech and text information to explore acoustic and lexical cues in emotional states [1]. Its applications have extended across diverse domains of human-computer interaction, including call-center interfaces [2], in-vehicle dashboard systems [3], speech-to-speech translation platforms [4], and others. Despite the considerable progress achieved by these approaches, there still exist two fundamental issues for this task: 1) how to extract acoustic and lexical features that are most effective in distinguishing between emotions; and 2) how to design suitable fusion methods for integrating multiple modalities for emotion recognition. This study addresses both issues to model the recognition process of emotion in a multimodal scenario.

Most features employed in the speech and text modalities for emotion recognition are typically categorized into two main groups: self-supervised representations (SSR) and handcrafted features [5]. Within its areas of interest, SSR have recently gained considerable attention due to their advances in capturing multiple emotional characteristics comprehensively and reliably [6]. Incidentally, handcrafted features also play a significant role in emotion recognition, and are increasingly evident in providing complementary information to SSR [7]. For example, Zou et al. incorporated multiple acoustic features, including Mel-frequency cepstral coefficients, spectrograms, and Wav2vec 2.0 embeddings, for categorical emotion recognition. Their approach yielded an absolute improvement of 7.03% compared with using Wav2vec 2.0 embeddings alone [8]. Moreover, Padi et al. presented a multimodal emotion recognition (MER) framework using Mel spectrogram and fine-tuning of pre-trained BERT models, offering complementary emotional insights derived from both speech and text modalities [9].

Motivated by the findings of these studies, this paper focuses on the feature extraction of multimodal emotions by integrating SSR and handcrafted features, particularly exploring handcrafted features of music theory-inspired acoustic representation (MTAR) to enhance the effectiveness of SSR in emotion recognition. Historically and today, speech and music expression and perception have dominated the nonverbal communication of emotion [10, 11]. There is widespread evidence supporting the recognition of emotion in both stimuli, which has been proven to develop in parallel [12, 13]. Fujisawa et al. studied the relationship between emotion perception and F0 on the basis of music intervals, suggesting that the interval structure is a fruitful means for determining the emotional valence of speech [14]. In addition, Yang et al. applied harmony perception known from music to improve emotion recognition performance and confirmed that musical interval-inspired representations associated with F0 counters are important cues for affective speech content [15]. Li et al. further solidified the link between music and speech by designing an acoustic representation of emotional speech regarding vocal emotion expression (VEE) and auditory emotion perception (AEP) processes via investigating music theory contents and showing promising performance compared with acoustic traditions of speech in emotion analysis [16]. In line with these advancements, our study first introduces a robust feature extraction method applied to MER by integrating the handcrafted MTAR and SSR.

The fusion strategy is another key aspect of MER. Two fundamental fusion strategies for MER are feature and decision-level fusion [17]. For instance, Yoon et al. proposed a deep dual recurrent neural network (RNN) for encoding audio-text sequences, subsequently concatenating their outputs to predict the final emotion. It achieves an accuracy of 71.8% on IEMOCAP [18]. Moreover, Pepino et al. designed multiple dual RNNs to represent audio-text sequences, and conducted a comparative analysis between feature and decision-level fusion approaches, highlighting their comparable performance [19]. Most researchers believe that decision-level fusion is performed more easily but ignore the relevance among representations of different modalities [20]. In contrast, our study introduces a

10.21437/Interspeech.2024-2350

feature fusion strategy at the feature-level, originally integrating a multimodal fusion module to enhance the acquisition of modality-specific and modality-invariant representations.

Our contributions can be summarized as follows:

- We propose a novel feature extraction approach for learning multimodal emotion information by leveraging both the SSR and handcrafted features extracted from MTAR. To the best of our knowledge, this is the first systematic attempt to integrate music theory-inspired acoustics with SSR to complement emotion-related information.
- We propose a multimodal fusion module that comprehensively integrates modality-specific and modality-invariant emotional information from both speech and text modalities. This module is designed to fully exploit emotional cues present in diverse modalities, ensuring a holistic understanding of emotional content across different sources.
- The experimental results demonstrate that the proposed approach adeptly addresses the MER tasks, surpassing the performance of existing feature extraction and fusion methods.

2. Proposed Method

This section details our proposed MER system, which is based on the multimodal fusion method leveraging the MTAR and SSR. As illustrated in Fig. 1, the network consists of three main components: an embedding module for encoding MTAR and SSR, a multimodal fusion module for integrating modality-specific and modality-invariant emotional information, and an emotion prediction module for predicting the emotion label.

2.1. Model Description

As illustrated, raw audio utterances are fed into dedicated encoder networks designed to extract MTAR, including VEE-derived MTAR, AEP-derived MTAR, and SSR. Simultaneously, transcripts undergo a self-supervised encoder to extract text SSR. These distinct modalities of information are then harmonized utilizing the proposed multimodal fusion method, thereby facilitating the final emotion recognition process.

2.2. Embedding Module

2.2.1. Music Theory-inspired Acoustic Representations

We represent the speech sample using MTAR features, as demonstrated by Li et al. [16]. The MTAR features can be categorized into two groups: a 41-dimensional MTAR derived from VEE and a 39-dimensional MTAR derived from AEP processes. More specifically, the VEE-derived MTAR draws inspiration from five music theory subgroups and includes descriptors related to ten MIDI notes, three music dynamics, five music main intervals, nine microtonal music attributes, and 114 descriptors associated with the syntactic structure of music. Additionally, the AEP-derived MTAR predominantly captures musical interval information, including melodic and harmonic intervals. We adopt $H_A = (h_A^1, h_A^2, \cdots, h_A^m) \in \mathbb{R}^{m*d}$ and $H_V = (h_V^1, h_V^2, \cdots, h_V^m) \in \mathbb{R}^{m*d}$ to symbolize the AEP and VEE-derived MTAR, respectively, where *m* denotes the number of frames extracted from an utterance, *d* is the dimension of hidden representations.

In the VEE encoder and AEP encoder, VEE-derived MTAR and AEP-derived MTAR are processed by a Bidirectional Gated Recurrent Unit (BiGRU) [21] with tanh as the activation function and a dropout rate of 0.5.

2.2.2. Speech Representations

To obtain a comprehensive understanding of acoustic features, we employ a pretrained SSL model, WavLM [22], as our speech self-supervised encoder. WavLM utilizes a hybrid architecture that includes convolutional neural network (CNN) layers and a transformer encoder to effectively capture speech features and contextual information. It has been fine-tuned for various downstream speech tasks [23]. We denote $H_S = (h_S^1, h_S^2, \cdots, h_S^m) \in \mathbb{R}^{m*d}$ to represent the speech SSR, where *m* denotes the number of frames extracted from an utterance, *d* is the dimension of hidden representations.

2.2.3. Text Representations

To acquire comprehensive information regarding lexical features, we leverage a pretrained SSL model, RoBERTa [24], as our text encoder. RoBERTa is an extension of the bidirectional encoder representations from the transformer (BERT) model, a widely-used model in natural language processing (NLP), which is specifically designed to address challenges related to long-range dependencies and is finely tuned for various NLP tasks. Pretrained on extensive corpora, including a diverse range of texts from various sources, such as a dataset comprising 58 million tweets, RoBERTa exhibits exceptional contextual understanding, thereby enhancing text-related tasks. We denote $H_T = (h_T^1, h_T^2, \cdots, h_T^n) \in \mathbb{R}^{n*d}$ to represent the text SSR, where *n* denotes the number of tokens extracted from an utterance, *d* is the dimension of hidden representations.

2.3. Multimodal Fusion (MF) Module

On the basis of previous study [25], our MF is composed of eight multi-level fusion (MLF) blocks, and two collaborative fusion (CF) blocks. The objective is to facilitate the learning of modality-specific representations and modality-invariant representations.

In this section, we offer an in-depth explanation of the operations of the MLF and CF blocks.

MLF Block adheres to the structure of a standard transformer layer, incorporating a cross-attention module, residual connections, and a BiGRU module. Initially, we employ four MLF blocks to derive AEP-aware and VEE-aware text SSR (H_T^{A}, H_T^{V}) $\in \mathbb{R}^{n*d}$, as well as AEP-aware and VEE-aware speech SSR (H_S^{A}, H_S^{V}) $\in \mathbb{R}^{m*d}$. This is achieved by utilizing H_V (or H_A) as queries and H_S (or H_T) as keys and values within each MLF block.

$$Q = H_V\,(\text{or } H_A), K = H_S\,(\text{or } H_T), V = H_S\,(\text{or } H_T). \quad (1)$$

$$H_S^{\mathrm{A}}(\text{or} H_S^{\mathrm{V}}, H_T^{\mathrm{A}}, H_T^{\mathrm{V}}) = \mathrm{BiGRU}(\text{Cross-Attention}(Q, K, V)). \quad (2)$$

CF Block is employed to extract complementary information from the AEP and VEE in MTAR:

First, we generate $H_S^{\mathrm{MTAR}} \in \mathbb{R}^{m*m}$, $H_T^{\mathrm{MTAR}} \in \mathbb{R}^{n*n}$ from AEP features H_A and VEE features H_V using a combination of BiGRU and fully-connected (FC) layers, as follows:

$$H_S^{\mathrm{MTAR}}(\text{or } H_T^{\mathrm{MTAR}}) = \mathrm{FC}(\mathrm{BiGRU}(H_V \oplus H_A)). \quad (3)$$

Then, the speech SSR H_S or text SSR H_T are multiplied by the collaborative representation H_S^{MTAR} or H_T^{MTAR} to obtain the weighted collaborative SSR ($H'_S \in \mathbb{R}^{m*d}$, $H'_T \in \mathbb{R}^{n*d}$).

$$H'_S\,(\text{or } H'_T) = H_S(\text{or } H_T) \cdot H_S^{\mathrm{MTAR}}(\text{or } H_T^{\mathrm{MTAR}}). \quad (4)$$

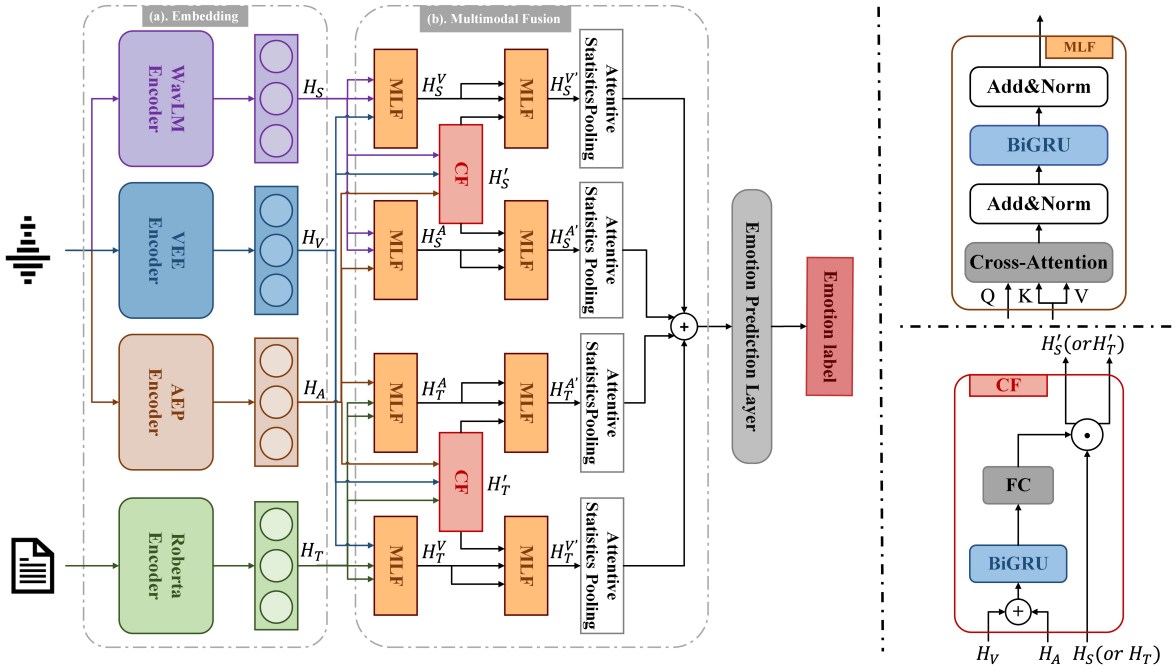

Figure 1: *The overall architecture of our proposed method.*

Next, we employ four additional MLF blocks to obtain both MTAR-aware speech SSR ($H_S^{\mathrm{A}'}$, $H_S^{\mathrm{V}'}$) $\in \mathbb{R}^{m*d}$ and MTAR-aware text SSR ($H_T^{\mathrm{A}'}$, $H_T^{\mathrm{V}'}$) $\in \mathbb{R}^{n*d}$. This is achieved by employing H_S' (or H_T') as queries and (H_S^{A}, H_S^{V}, H_T^{A}, H_T^{V}) as keys and values within each MLF block.

$$H_S^{\mathrm{A}'}(\mathrm{or}H_S^{\mathrm{V}'}, H_T^{\mathrm{A}'}, H_T^{\mathrm{V}'}) = \mathrm{BiGRU}(\text{Cross-Attention}(Q, K, V)). \tag{5}$$

Finally, we adopt attentive statistics pooling [26] to obtain a 1-dimension vector for each output. The final MTAR-aware speech and text representations ($H_S^{\mathrm{A}'}$, $H_S^{\mathrm{V}'}$, $H_T^{\mathrm{A}'}$, $H_T^{\mathrm{V}'}$) are concatenated and written as follows:

$$H_{ST}^{\mathrm{MTAR}} = H_S^{\mathrm{A}'} \oplus H_S^{\mathrm{V}'} \oplus H_T^{\mathrm{A}'} \oplus H_T^{\mathrm{V}'}. \tag{6}$$

2.4. Emotion Classification Module

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

$$P(y_{\mathrm{emo}}|H_{ST}^{\mathrm{MTAR}}) = \mathrm{SoftMax}(\mathrm{FC}(H_{ST}^{\mathrm{MTAR}})). \tag{7}$$

where y_{emo} is the predicted emotion classification.

3. Experiments

3.1. Dataset

Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is a widely used corpus in affective computing [27]. It contains approximately 12 hours of audio-visual recordings and is designed for two-person dialogs. Each dialog in IEMOCAP has been segmented into utterances with continuous labels in the Valence-Arousal dimension and category labels for specific emotional states. We consider four categorical emotions consistent with experimental protocols used in many previous studies [8, 16]: neutral, happiness, sadness, and anger. Additionally, we merge happiness and excitement into one category.

3.2. Experimental Procedure

We conduct three experiments in this study. In Experiment 1, we examine the effect of MTAR on SSR for MER, comparing it with a Mel Spectrogram. In Experiment 2, we explore the effect of fusion methods on MER, using concatenation and co-attention [8] as benchmarks from previous works. In Experiment 3, we further investigate the effect of MTAR on common speech and text-based SSR (Wav2vec 2.0 [28], Hubert [29], WavLM [22], BERT [30], DeBERTa [31], RoBERTa [24]) within the domain of emotion recognition tasks.

3.3. Implementation

Our deep learning models were developed using Python 3.7 and PyTorch 1.11.0. The model was trained and evaluated on a computer with Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz, 32GB RAM, and one NVIDIA Tesla V100 GPU.

For the speech modality, the self-supervised encoder was initialized using the WavLM model[1], resulting in speech SSR with a dimensionality of 1024. For the text modality, we utilized ground truth text from IEMOCAP as the source for the self-supervised encoder. Specifically, we employed the RoBERTa model[2], with a hidden size of 768, 12 attention layers, and 12 attention heads. Both the WavLM and RoBERTa models underwent fine-tuning during the training process. Most specifically, all speech fine-tuned SSL models utilize the large-sized model, while all text fine-tuned SSL models utilize the base-sized model.

[1]https://huggingface.co/microsoft/wavlm-large
[2]https://huggingface.co/FacebookAI/roberta-base

3.4. Evaluation

In evaluating our results on IEMOCAP, for which a standard train, dev, and test split is lacking, we adopt a common strategy: leave-one-speaker-out cross-validation, as also employed in previous works [32, 33]. The categorical MER performance is assessed using the Unweighted Average Recall (UAR) and F1 scores, which have been widely utilized in experiments with unbalanced data to evaluate performance [34, 35], across the four distinct emotional labels.

4. Results and Discussion

To analyze the effect of MTAR on SSR for emotion recognition, we compare the prediction performance characteristics of single and multimodal emotion recognition, as shown in Table 1.

Table 1: *Comparison of MER performance obtained by different SSR that integrated the proposed MTAR and Mel Spectrogram.*

Modality	Model	UAR (%)	F1 (%)
Single modal	MTAR [16]	61.92	61.47
	WavLM	70.46	70.24
	WavLM + Mel Spectrogram	71.70	71.15
	WavLM + MTAR	**73.47**	**73.01**
Multimodal	RoBERTa	71.33	71.08
	RoBERTa + Mel Spectrogram	72.12	72.08
	RoBERTa + MTAR	**75.11**	**74.79**
	RoBERTa + WavLM	76.34	76.15
	RoBERTa + WavLM + Mel Spectrogram	77.08	76.42
	RoBERTa + WavLM + MTAR	**79.89**	**79.35**

The results show that the utilization of MTAR yields superior performance compared with relying solely on SSR. The observed improvements in UAR are 3.01% for single modal approaches, 3.78% for utilizing only text SSR, and 3.55% for employing both speech and text SSR in a multimodal approach. These findings suggest that incorporating MTAR for emotional recognition outperforms relying solely on SSR. Moreover, when compared with considering MTAR and speech SSR or MTAR and text SSR, integrating of MTAR, speech and text SSR results in notable enhancements, with UAR improvements of 6.42% and 4.78%, respectively. These results underscore the significance of integrating multiple modalities of information to further enhance the performance of emotion recognition systems. Noting that, compared with the Mel spectrogram, MTAR achieved improvements of 1.77% in single modal, 2.99%, and 2.81% in multimodal approaches, respectively.

To evaluate the effect of fusion methods on emotion recognition, we contrast the predictive performance attributes of single and multimodal emotion recognition, delineated in Table 2.

Table 2: *Comparison of MER performance obtained by the proposed MF and baseline fusion methods.*

Modality	Model	UAR (%)	F1 (%)
Single modal	WavLM + MTAR (Concat)	71.85	71.12
	WavLM + MTAR [8]	72.61	71.92
	WavLM + MTAR (MF)	**73.47**	**73.01**
Multimodal	RoBERTa + MTAR (Concat)	72.83	72.24
	RoBERTa + MTAR [8]	73.50	73.35
	RoBERTa + MTAR (MF)	**75.11**	**74.79**
	RoBERTa + WavLM + MTAR (Concat)	77.78	77.29
	RoBERTa + WavLM + MTAR [8]	78.38	77.83
	RoBERTa + WavLM + MTAR (MF)	**79.89**	**79.35**

Evidently, the proposed fusion method demonstrates an enhancement in both single and multimodal emotion recognition model performance. When compared with the commonly utilized concatenate method, the observed improvements in UAR are 1.62% for the speech modality 2.28% and 2.11% for multimodal approaches. These results signify that the adoption of the MF method enhances the performance of MER. Furthermore, in comparison with previous work [8], the observed improvements in UAR are 0.86% for the speech modality 1.61% and 1.51% for multimodal approaches. These findings suggest a new feasibility for feature fusion techniques, offering a promising path for advancing emotion recognition systems.

Table 3: *The effectiveness of MTAR on speech and text-based SSR for emotion recognition.*

Modality	Model	UAR (%)	F1 (%)
Single modal	Wav2vec 2.0	68.43	65.91
	Hubert	68.47	67.72
	WavLM	70.46	70.24
	Wav2vec 2.0 + MTAR	70.06	68.08
	Hubert + MTAR	72.05	71.52
	WavLM + MTAR	**73.47**	**73.01**
Multimodal	BERT	70.74	70.69
	DeBERTa	71.45	70.96
	RoBERTa	71.33	71.08
	BERT + MTAR	73.91	73.52
	DeBERTa + MTAR	74.87	74.51
	RoBERTa + MTAR	**75.11**	**74.79**

In Table 3, we delve deeper into the effect of MTAR on common speech and text-based SSR. Our results reveal that MTAR yields notable enhancements in the efficacy of common SSR for emotion recognition. Specifically, for speech-based SSR, the observed improvements in UAR are as follows: Wav2vec 2.0 demonstrates a noteworthy enhancement of 1.63%, Hubert exhibits a substantial improvement of 3.58%, and WavLM shows a marked increase of 3.01%. Additionally, for text-based SSR, BERT showcases a commendable enhancement of 3.17%, while DoBERTa achieves a notable improvement of 3.42%. RoBERTa demonstrates a remarkable increase of 3.78%. Collectively, these results underscore the substantial performance gains observed in SSR through the incorporation of MTAR.

5. Conclusions and Future Work

In this paper, we integrated handcrafted MTAR and SSR for MER tasks. Our findings revealed superior performance when incorporating handcrafted MTAR compared to merely integrating individual modality SSR, demonstrating the effectiveness of MTAR in enhancing SSR in emotion recognition. Additionally, we introduced a novel MF method for integrating modality-specific and modality-invariant emotional information, consistently achieving higher accuracy in MER. For future research, we advocate for further exploration of the intricate relationship between music and speech to devise innovative fusion methods.

6. Acknowledgements

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.

7. References

[1] P. Singh, R. Srivastava, K. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowledge-Based Systems*, vol. 229, p. 107316, 2021.

[2] T. Deschamps-Berger, L. Lamel, and L. Devillers, "End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[3] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–30, 2020.

[4] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–10.

[5] J. Tian, D. Hu, X. Shi, J. He, X. Li, Y. Gao, T. Toda, X. Xu, and X. Hu, "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.

[6] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] N. Naderi and B. Nasersharif, "Cross corpus speech emotion recognition using transfer learning and attention-based fusion of wav2vec2 and prosody features," *Knowledge-Based Systems*, vol. 277, p. 110814, 2023.

[8] 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.

[9] 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.

[10] K. R. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[11] A. Gabrielsson and P. N. Juslin, *Emotional expression in music*. Oxford University Press, 2003.

[12] A. D. Patel, "Language, music, syntax and the brain," *Nature neuroscience*, vol. 6, no. 7, pp. 674–681, 2003.

[13] L. Jäncke, "The relationship between music and language," p. 123, 2012.

[14] T. Fujisawa, K. Takami, and N. D. Cook, "On the role of pitch intervals in the perception of emotional speech," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[15] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," *Signal processing*, vol. 90, no. 5, pp. 1415–1423, 2010.

[16] X. Li, X. Shi, D. Hu, Y. Li, Q. Zhang, Z. Wang, M. Unoki, and M. Akagi, "Music theory-inspired acoustic representation for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[17] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, pp. 345–379, 2010.

[18] 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.

[19] 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)*. IEEE, 2020, pp. 6484–6488.

[20] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 02, 2020, pp. 1359–1367.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[22] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *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.

[23] 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.

[24] 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.

[25] 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 *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 066–11 070.

[26] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[27] 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.

[28] 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.

[29] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[30] X. Qin, Z. Wu, T. Zhang, Y. Li, J. Luan, B. Wang, L. Wang, and J. Cui, "Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 492–13 500.

[31] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[32] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Enhancing two-stage finetuning for speech emotion recognition using adapters," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 316–11 320.

[33] Y. Gao, C. Chu, and T. Kawahara, "Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining," in *Proc. Interspeech*, 2023.

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

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