



Advancing Emotion Recognition via Ensemble Learning: Integrating Speech, Context, and Text Representations

Xiaohan Shi¹, Jinyi Mi¹, Xingfeng Li², Tomoki Toda³

¹Graduate School of Informatics, Nagoya University, Japan

²Faculty of Data Science, City University of Macau, China

³Information Technology Center, Nagoya University, Japan

xiaohan.shi@g.sp.m.is.nagoya-u.ac.jp, mi.jinyi@g.sp.m.is.nagoya-u.ac.jp,
xfli@cityu.edu.mo, tomoki@icts.nagoya-u.ac.jp

Abstract

Speech Emotion Recognition (SER) in real-world scenarios aims to identify a speaker's emotional states from spontaneous speech. While prior research has focused on noise reduction techniques within individual domains, integrating multi-domain noise-robust representations for SER remains underexplored. To address this challenge, we propose a novel Speech-Context-Text (SCT) model, which integrates speech, context, and text representations via ensemble learning. Specifically, we introduce the Mamba method for speech representation, employ a layer adapter to capture context representation, and adopt ASR correction to refine text representation. Extensive experiments demonstrate the effectiveness of SCT, achieving a 7.4% Macro-F1 improvement over the official baseline of the Speech Emotion Recognition in Naturalistic Conditions Challenge at INTERSPEECH 2025, securing 6th place in the competition. Additionally, SCT yields 7.37% and 7.95% gains on MSP-Podcast and IEMOCAP, respectively.

Index Terms: speech emotion recognition, text emotion recognition, ensemble learning

1. Introduction

Speech Emotion Recognition (SER) is a critical area in human-computer interaction, playing a key role in intelligent systems such as virtual assistants [1] and automated customer service [2]. Despite significant progress, SER systems remain challenged by real-world variability, which affects their robustness [3]. To address these challenges, the Speech Emotion Recognition in Naturalistic Conditions Challenge at INTERSPEECH 2025 aims to advance emotion recognition from spontaneous speech, prioritizing real-world applicability over controlled, acted scenarios [4].

In recent years, SER in real-world scenarios has garnered increasing attention. For instance, Zhao et al. [5] proposed a noise-robust SER approach by employing a weighted sparse representation model based on maximum likelihood estimation, leading to improved classification outcomes. Shi et al. [6] leveraged automatic speech recognition (ASR) models as noise-robust feature extractors to effectively remove non-vocal components from noisy speech for SER. Beyond noise reduction techniques in speech, recent research has increasingly focused on leveraging ASR models to extract text-based emotional information from noisy speech, enabling a more comprehensive representation of emotional cues. However, noisy speech in real-world scenarios can significantly degrade ASR transcriptions, necessitating the development of error detection and correction techniques to enhance model robustness. For instance, He et al. [7] introduced ASR error detection and ASR error correction to improve semantic coherence in transcriptions, along

with a novel multimodal fusion (MF) framework for learning shared representations across modalities. Similarly, Lin et al. [8] proposed an SER system that mitigates ASR inaccuracies by integrating complementary semantic information from audio, thereby improving transcription reliability. Although previous studies have made significant progress in mitigating the impact of real-world conditions on SER. However, the potential of leveraging multi-domain noise-robust representations to enhance SER robustness remains unexplored.

To address this challenge, we propose a novel Speech-Context-Text (SCT) model, which integrates speech, context, and text representations via ensemble learning for SER. Specifically, the SCT model consists of a Mamba-based WavLM model for speech emotion representation, a layer adapter-based Whisper model for context emotion representation, and a RoBERTa-based model enhanced with an ASR correction module for text emotion representation. To assess the effectiveness of the proposed SCT model, we evaluate its performance on the MSP-Podcast database and the IEMOCAP database. Furthermore, to analyze the contributions of each representation within the SCT model, we investigate the performance of different representation combinations for SER.

The contributions of this study are summarized as follows:

- We propose a novel SCT model, which integrates speech, context, and text representations via ensemble learning to enhance SER performance.
- To effectively leverage noise-robust emotion information, we incorporate the Mamba method, layer adapter method, and ASR correction within the SCT model.
- Our proposed SCT model outperforms the challenge baseline by 7.4%, ranks 6th in the competition [9], and achieves 7.37% and 7.95% gains on MSP-Podcast and IEMOCAP.

2. Proposed Method

2.1. Model Description

As illustrated in Fig. 1, the SCT model comprises three independent emotion recognition models and an ensemble learning-based model: (1) a WavLM model for capturing speech-based emotion representations, (2) a Whisper model for integrating context-based emotion representations, (3) a RoBERTa model for extracting text-based emotion representations, and (4) an ensemble learning-based model that fuses these representations to predict the final emotion labels.

2.2. Speech-based Emotion Recognition Model

To obtain a comprehensive understanding of acoustic features, we employ a pretrained self-supervised learning (SSL) model, WavLM [10], as our speech SSL model.

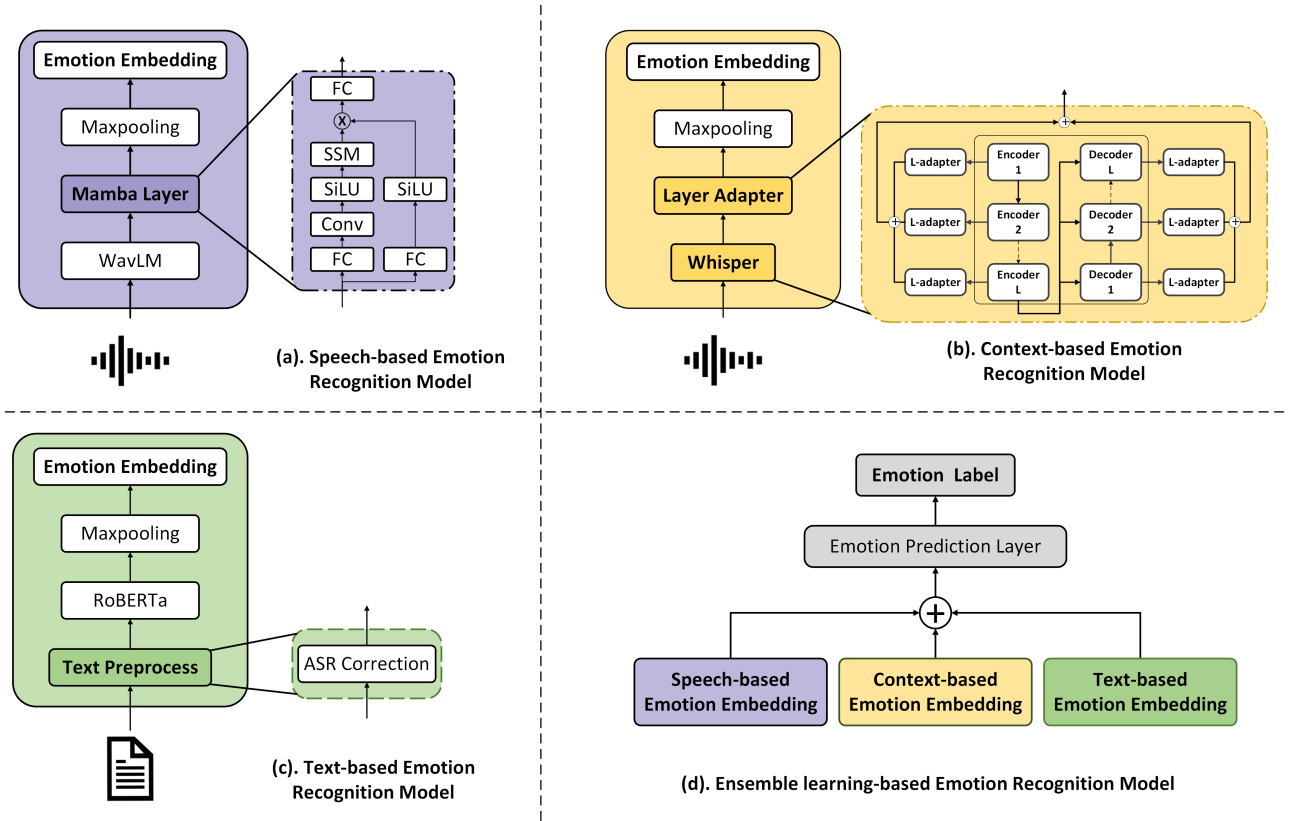


Figure 1: The overall architecture of our proposed SCT model.

WavLM employs a transformer-based architecture with a convolutional front-end, leveraging SSL on large-scale speech corpora covering diverse acoustic conditions. By training on datasets that include various noise types and reverberant environments, WavLM achieves enhanced robustness against background interference. We denote H_E as the speech SSL representations.

Mamba Layer: To capture expressive emotional representations from speech SSL, we incorporate the Mamba layer, which facilitates long-range dependency modeling and enhances information aggregation capabilities [11, 12]. The Mamba layer comprises fully connected (FC) layers, convolutional layers, SiLU activation, and a state space model (SSM). The speech SSL representations H_E involve two pathways, each starting with the FC layers:

$$h_1 = \text{FC}(H_E), \quad h'_1 = \text{FC}(H_E). \quad (1)$$

In the first pathway, h_1 is first processed by a one-dimensional convolution with a kernel size of k , followed by a SiLU activation function. In the second pathway, h'_1 is directly transformed using the SiLU activation function. Specifically, the transformations in the two pathways are defined as follows:

$$\begin{aligned} h_2 &= \text{SiLU}(\text{Conv1D}(h_1; k)), \\ h'_2 &= \text{SiLU}(h'_1). \end{aligned} \quad (2)$$

Then, long-range dependencies are further captured by processing h_2 through an SSM:

$$y_1 = \text{SSM}(h_2). \quad (3)$$

The final speech-based emotion representations are obtained by combining the outputs from the SSM and the second pathway via max-pooling:

$$H_{\text{Speech}} = \text{MaxPooling}(y_1 \odot h'_2), \quad (4)$$

where \odot denotes element-wise multiplication.

2.3. Context-based Emotion Recognition Model

To achieve a more comprehensive representation of acoustic features, we employ Whisper [13] as our context SSL model. Whisper is a supervised speech recognition framework built on an encoder-decoder transformer architecture, trained on approximately 680,000 hours of multilingual speech covering 60 languages [13]. Inspired by a previous study [6], which demonstrated the superiority of Whisper in noisy emotion recognition tasks, we model both the Whisper encoder and decoder as context-based representations, denoted as H_{Context} .

An audio utterance x is first transformed via a mapping function F_θ and then processed by a CNN layer, yielding an initial representation E_0 :

$$E_0 = \text{CNN}(F_\theta(x)). \quad (5)$$

where F_θ denotes the mapping function implemented by Whisper. Subsequently, E_0 is refined through L encoder layers:

$$E_{l+1} = \text{Encoder}(E_l) \quad \text{for } l = 0, \dots, L-1, \quad (6)$$

where $E_l = (e_l^1, \dots, e_l^m) \in \mathbb{R}^{m \times d}$ represents the hidden states at layer l , and m is the sequence length. The final encoder output E_L encapsulates the contextualized representation of the entire input sequence.

The decoder initializes with a special start-of-sequence token and generates subsequent tokens autoregressively. At each

step, it conditions on previously generated tokens and the encoder output E_L :

$$D_{l+1} = \text{Decoder}(D_l, E_L) \quad \text{for } l = 0, \dots, L-1, \quad (7)$$

where $D_l = (d_l^1, \dots, d_l^n) \in \mathbb{R}^{n \times d}$ denotes the decoder’s hidden states at step l , and n is the length of the generated output.

Layer adapter (L-adapter): To refine the intermediate representations extracted by Whisper, we introduce a layer adapter approach at each encoder and decoder layer. Each layer adapter comprises an FC layer, followed by a nonlinear activation function and layer normalization, and a mean operation forming dedicated pathways from each encoder and decoder layer. The adapted encoder representations $H_E^* \in \mathbb{R}^{1 \times d}$ and the adapted decoder representations $H_D^* \in \mathbb{R}^{1 \times d}$ as follows:

$$\alpha_E^l = \text{FC}(E_l), \quad \alpha_D^l = \text{FC}(D_l) \quad \text{for } l = 1, 2, \dots, L, \quad (8)$$

where $E_l \in \mathbb{R}^{m \times d}$ and $D_l \in \mathbb{R}^{n \times d}$ denote the encoder and decoder hidden states, respectively.

Subsequently, a weighted sum is applied to aggregate the adapted representations across all L layers:

$$H_E^* = \sum_{l=1}^L w_l \alpha_E^l, \quad H_D^* = \sum_{l=1}^L w_l \alpha_D^l. \quad (9)$$

Finally, these aggregated representations are concatenated with an FC layer to obtain the final context-based emotion representation:

$$H_{\text{Context}} = \text{FC}(\text{Concate}(H_E^*, H_D^*)). \quad (10)$$

2.4. Text-based Emotion Recognition Model

To extract comprehensive lexical features, we adopt the pre-trained SSL model RoBERTa [14] as our text-based emotion recognition model. Since noise exacerbates ASR recognition errors, potentially distorting emotion-related information in transcriptions, we first apply an ASR correction method to enhance transcription quality. Specifically, we employ *Language-Tool*¹ for ASR correction. The resulting text-based emotion representations are denoted as H_{Text} .

2.5. Ensemble learning-based Emotion Recognition Model

To integrate emotional information from different modalities, we employ ensemble learning to fusion three independent emotion recognition models. Specifically, we utilize an FC layer and a SoftMax activation function to fuse the extracted representations, formulated as follows:

$$y_{\text{Emo}} = \text{SoftMax}(\text{FC}(\text{Concate}(H_{\text{Speech}}, H_{\text{Context}}, H_{\text{Text}}))), \quad (11)$$

where y_{Emo} is the predicted emotion classification.

The loss function \mathcal{L}_{Emo} is formulated using cross-entropy:

$$\mathcal{L}_{\text{Emo}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{ic} \log(\hat{y}_{ic}), \quad (12)$$

where N denotes the 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 .

¹<https://languagetool.org/>

3. Experiments

3.1. Dataset

3.1.1. MSP-Podcast Corpus

The MSP-Podcast corpus consists of speech segments from podcast recordings that have been perceptually annotated through crowdsourcing [15]. It includes eight categorical emotion labels: Anger, Sadness, Happiness, Surprise, Fear, Disgust, Contempt, and Neutral. As part of the Speech Emotion Recognition in Naturalistic Conditions Challenge at INTERSPEECH 2025, the training partition comprises 68,119 conversational turns, while the validation set includes 19,815 utterances from 454 distinct speakers [9]. Since the ground-truth labels of the official test set are not publicly available, we use the challenge’s validation set as the test set.

3.1.2. Interactive Emotional Dyadic Motion Capture dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset is a widely used benchmark in affective computing [16]. It comprises audio-visual recordings of two-person dialogues segmented into utterances. Each utterance is annotated with both continuous labels in the valence-arousal space and categorical labels corresponding to specific emotional states. In this study, we focus on four categorical emotions—neutral, happiness, sadness, and anger—following common experimental protocols [17, 18]. Additionally, we merge the happiness and excitement categories into a single class, as done in prior research [19, 20].

3.2. Implementation

Our deep learning models were implemented using Python 3.7 and PyTorch 1.11.0. Training and evaluation were conducted on a system equipped with an Intel(R) Xeon(R) Gold 6248 CPU (2.50GHz), 32GB RAM, and a single NVIDIA Tesla V100.

The speech SSL model was initialized with *WavLM-Large*², producing 1024-dimensional speech representations. The context-aware component was based on *Whisper-large-V3*³, also yielding 1024-dimensional representations. For the text SSL model, we employed *RoBERTa-large*⁴, which comprises 24 attention layers, 12 attention heads, and a hidden dimension of 1024. During training, WavLM and RoBERTa were fine-tuned, whereas Whisper remained frozen. For ensemble learning, we incorporated an FC layer with a dropout rate of 0.5.

3.3. Evaluation

For evaluating model performance on the MSP-Podcast dataset, we adhere to the official training and validation set partitions provided by the challenge organizers [21]. For IEMOCAP, which does not have a standardized train, validation, and test split, we adopt the commonly used leave-one-session-out cross-validation strategy, following prior studies [22, 23, 24].

To assess categorical SER performance, we employ Unweighted Accuracy (UA) and Macro-F1 score, both of which are widely used metrics for evaluating models on imbalanced datasets [25, 26]. These metrics are computed over the eight emotional categories in MSP-Podcast and the four emotional categories in IEMOCAP considered in this study.

²<https://huggingface.co/microsoft/wavlm-large>

³<https://huggingface.co/openai/whisper-large-v3>

⁴<https://huggingface.co/FacebookAI/roberta-large>

Table 1: The effectiveness of the proposed SCT model in MSP-Podcast and IEMOCAP.

Database	Experiment	Modality	Representation	Method	UA (%)	F1 (%)
MSP-Podcast	Baseline	Speech	WavLM Whisper	- -	32.75 35.74	32.13 34.17
		Text	RoBERTa	-	31.04	31.79
	Proposed	Speech	WavLM Whisper	Mamba Layer Adapter	39.23 37.34	38.50 36.34
		Text	RoBERTa	ASR Correction	31.73	32.29
		Speech + Text	ALL	SCT	48.64	41.54
IEMOCAP	Baseline	Speech	WavLM Whisper	- -	70.31 71.87	70.40 71.93
		Text	RoBERTa	-	71.71	71.79
	Proposed	Speech	WavLM Whisper	Mamba Layer Adapter	71.47 72.47	71.61 72.83
		Text	RoBERTa	ASR Correction	72.01	71.70
		Speech + Text	ALL	SCT	79.89	79.88

4. Results and Discussion

To assess the effectiveness of the proposed SCT model, we evaluate its performance on the MSP-Podcast and IEMOCAP databases, as shown in Table 1.

The results demonstrate that the SCT model significantly improves SER performance on both datasets. In MSP-Podcast, for the speech modality, WavLM with Mamba achieves a 6.48% increase in UA and 6.37% in F1, compared to WavLM alone, while applying a layer adapter to Whisper yields further gains of 1.60% in UA and 2.17% in F1, compared to the Whisper encoder. For the text modality, incorporating ASR correction in RoBERTa results in 0.69% and 0.50% improvements in UA and F1, respectively, compared to using RoBERTa alone. In the multimodal setting, the proposed SCT model achieves the best performance, surpassing the best baseline (Whisper) with gains of 12.9% in UA and 7.37% in F1, highlighting the benefits of multimodal integration. A similar pattern is observed in IEMOCAP, where WavLM with Mamba improves UA by 1.16% and F1 by 1.21%, while the layer adapter for Whisper contributes additional gains of 0.6% in UA and 0.9% in F1. ASR correction further enhances the text modality, yielding increases of 0.3% in UA. Finally, in the multimodal setting, the SCT model provides an additional 8.02% increase in UA and 7.95% in F1 over the best unimodal result, underscoring the significance of the proposed SCT model in SER.

To further validate the contributions of each representation within the proposed SCT model, we compare three pairwise representation combinations (WavLM + Whisper, WavLM + RoBERTa, and Whisper + RoBERTa) against the full representation model, which integrates all three, as shown in Table 2.

The results demonstrate that each representation combination within the SCT model significantly enhances SER performance on both the MSP-Podcast and IEMOCAP datasets. On the MSP-Podcast dataset, the WavLM + RoBERTa combination achieves the best two-representation configuration, with UA and F1 scores only 0.32% and 0.31% lower, respectively, than the full representation model. On the IEMOCAP dataset, the Whisper + RoBERTa combination yields the best performance among two-representation configurations, with UA and F1 scores 0.7% and 0.56% lower, respectively, than the full representation model. Notably, for speech modality repre-

Table 2: The impact of different representations on SCT performance for SER

Database	Representation	UA (%)	F1 (%)
MSP-Podcast	Whisper + RoBERTa	45.63	37.39
	WavLM + RoBERTa	48.32	41.23
	WavLM + Whisper	44.64	39.52
	ALL	48.64	41.54
IEMOCAP	Whisper + RoBERTa	79.19	79.32
	WavLM + RoBERTa	77.96	77.95
	WavLM + Whisper	74.63	74.81
	ALL	79.89	79.88

sentations, either WavLM or Whisper, when combined with RoBERTa, achieves superior performance. The optimal choice between the two depends on the dataset, potentially due to differences in their robustness to various types and levels of noise. Furthermore, an important finding is that incorporating both WavLM and Whisper in speech modality does not degrade performance. Consequently, leveraging both representations remains beneficial for achieving optimal performance.

5. Conclusions and Future Work

In this paper, we propose a novel SCT model that integrates speech, context, and text representations via ensemble learning for SER. Our approach secured 6th place in the Speech Emotion Recognition in Naturalistic Conditions Challenge at INTERSPEECH 2025. Experimental results show that integrating speech, context, and text representations significantly enhances SER performance, particularly when combining multimodal representations. Moreover, incorporating advanced techniques, including Mamba, layer adapters, and ASR correction, improves the robustness of individual domain representations within the SCT model. For future work, we suggest exploring additional paralinguistic information (e.g., speaker identity and gender) to further improve SER performance.

6. Acknowledgements

This work was partly supported by JST SPRING, Grant Number JPMJSP2125, JST AIP Acceleration Research JPMJCR25U5, and JSPS KAKENHI Grant Number 21H05054, Japan.

7. References

- [1] G. Lugano, "Virtual assistants and self-driving cars," in *2017 15th international conference on ITS telecommunications (ITST)*. IEEE, 2017, pp. 1–5.
- [2] E. K. Harris, *Customer service: A practical approach*. Prentice-Hall, Inc., 2002.
- [3] M. S. Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digital signal processing*, vol. 110, p. 102951, 2021.
- [4] L. Goncalves, A. N. Salman, A. R. Naini, L. Moro-Velázquez, T. Thebaud, P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024 - speech emotion recognition challenge: Dataset, baseline framework, and results," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 247–254.
- [5] X. Zhao, S. Zhang, and B. Lei, "Robust emotion recognition in noisy speech via sparse representation," *Neural Computing and Applications*, vol. 24, pp. 1539–1553, 2014.
- [6] X. Shi, J. He, X. Li, and T. Toda, "On the effectiveness of asr representations in real-world noisy speech emotion recognition," *arXiv preprint arXiv:2311.07093*, 2023.
- [7] 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.
- [8] B. Lin and L. Wang, "Robust multi-modal speech emotion recognition with asr error adaptation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [9] A. R. Naini, L. Goncalves, A. N. Salman, P. Mote, I. R. Ülgen, T. Thebaud, L. Velazquez, L. P. Garcia, N. Dehak, B. Sisman, and C. Busso, "The interspeech 2025 challenge on speech emotion recognition in naturalistic conditions," in *Interspeech 2025*, vol. To appear, Rotterdam, The Netherlands, August 2025.
- [10] 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.
- [11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
- [12] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, "U-mambanet: A highly efficient mamba-based u-net style network for noisy and reverberant speech separation," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2024, pp. 1–5.
- [13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [14] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.
- [15] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.
- [16] 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.
- [17] 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*, vol. 31, pp. 2534–2547, 2023.
- [18] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Speech emotion recognition with multi-level acoustic and semantic information extraction and interaction," in *Proc. Interspeech 2024*, 2024, pp. 1060–1064.
- [19] X. Shi, Y. Gao, J. He, J. Mi, X. Li, and T. Toda, "A study on multimodal fusion and layer adapter in emotion recognition," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2024, pp. 1–6.
- [20] J. Mi, X. Shi, D. Ma, J. He, T. Fujimura, and T. Toda, "Two-stage framework for robust speech emotion recognition using target speaker extraction in human speech noise conditions," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2024, pp. 1–6.
- [21] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The msp-conversation corpus," *Interspeech 2020*, 2020.
- [22] 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.
- [23] H. Sun, S. Zhao, X. Kong, X. Wang, H. Wang, J. Zhou, and Y. Qin, "Iterative prototype refinement for ambiguous speech emotion recognition," *arXiv preprint arXiv:2408.00325*, 2024.
- [24] H. Sun, S. Zhao, S. Li, X. Kong, X. Wang, J. Zhou, A. Kong, Y. Chen, W. Zeng, and Y. Qin, "Enhancing emotion recognition in incomplete data: A novel cross-modal alignment, reconstruction, and refinement framework," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [25] H. Sun, S. Zhao, X. Wang, W. Zeng, Y. Chen, and Y. Qin, "Fine-grained disentangled representation learning for multimodal emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 051–11 055.
- [26] 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.