Interspeech 2024
1-5 September 2024, Kos, Greece

Enhancing Multimodal Emotion Recognition through ASR Error Compensation and LLM Fine-Tuning

Jehyun Kyung[1], *Serin Heo*[2], *Joon-Hyuk Chang*[1,2,*]

[1]Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea
[2]Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea
{jehyunkyung, serinheo, jchang}@hanyang.ac.kr

Abstract

Multimodal emotion recognition (MER), particularly using speech and text, is promising for enhancing human-computer interaction. However, the efficacy of such systems is often compromised by inaccuracies introduced during the automatic speech recognition (ASR) process. Addressing this, we present a comprehensive MER system that incorporates ways to make up for errors in ASR-generated text. Our system capitalizes on the strengths of speech signals and ASR-generated text, employing a cross-modal transformer (CMT) to blend these modalities effectively. We introduce a novel error compensation technique to counteract the detrimental effects of ASR inaccuracies and employ preference learning to fine-tune a large language model (LLM), thus improving its ability to distinguish slight emotional nuances in text. Performance of our proposed MER system is evaluated on the IEMOCAP dataset, demonstrating significant advancements in emotion recognition accuracy over conventional methods.

Index Terms: multimodal emotion recognition, automatic speech recognition (ASR), large language model, cross-modal transformer, ASR error compensation

1. Introduction

Multimodal emotion recognition (MER) is crucial in enhancing human-computer interaction, with applications spanning from personalized recommendations to mental health assessment [1–3]. In recent years, integrating multimodal data sources, such as speech, text, and video, has significantly advanced the accuracy and reliability of emotion recognition systems [4, 5]. In particular, the fusion of speech and automatic speech recognition (ASR)-generated text has shown promise in the realm of MER [6, 7]. Speech and text offer rich and complementary cues that are essential in understanding emotional states. Nevertheless, converting speech to text through ASR often introduces errors, adversely affecting the overall performance of the system [8]. This paper presents a novel MER system that effectively leverages both speech and ASR-generated text. Our system incorporates error compensation mechanisms to mitigate the impact of ASR inaccuracies and employs preference learning [9] to fine-tune large language models (LLM), thereby capturing emotional nuances more effectively.

The study of emotion recognition has been extensive, exploring both speech and text modalities individually. Systems based on speech analyze acoustic features, such as pitch, energy, and rate, which convey emotional states [10]. Conversely, text-based approaches examine linguistic cues, including word choice and sentence structure, to infer emotions [11]. Recent advances have underscored the value of integrating these modalities, demonstrating that a synergistic approach to speech and text analysis can enhance emotion detection robustness. Despite the promising potential of ASR to bridge the gap between these two modalities, its application in emotion recognition is not without challenges [8]. The inaccuracies inherent in ASR-generated text, often resulting from homophones and background noise, can obscure the emotional cues derived from speech. The emergence of pre-trained LLM, such as BERT [12] and GPT [13], has marked a transformative phase in natural language processing, extending to emotion detection from text. These models, capable of capturing deep semantic relationships and contextual nuances, facilitate a more sophisticated understanding of emotional expressions in text [14]. Nonetheless, the endeavor to fine-tune these models for improved sensitivity to emotional nuances remains an ongoing field of exploration.

The primary contribution begins with a comprehensive system integrating speech and ASR-generated text for emotion recognition. A novel error compensation mechanism is introduced for the system to minimize the negative impact of ASR inaccuracies on the emotion recognition process. To achieve this, we utilize Kullback-Leibler divergence (KLD) loss to align ASR-derived text embeddings more closely with ground truth text embeddings, thereby reducing the distributional differences between the ASR-generated and ground truth text embeddings. This approach mitigates the impact of ASR inaccuracies and lays a foundational layer for further refinements in emotion recognition. Furthermore, we integrate preference learning within the RankNet framework [15] to fine-tune an LLM, significantly enhancing its ability to distinguish emotional content more clearly in text. We thoroughly evaluate our system on the IEMOCAP dataset [16], achieving notable enhancements in emotion recognition accuracy over existing methodologies.

2. Related works

Integration of preference learning into emotion recognition tasks brings a novel dimension to the field. One approach that is especially noteworthy is talked about in the work by Naini *et al*. [9]. They presented a multi-task learning (MTL) framework that combined preference learning with regression formulations, using a RankNet framework [15] to predict emotional attributes from speech. This study demonstrated that these formulations were not mutually exclusive but that their combination could lead to more flexible and accurate models. The MTL approach allowed the models to simultaneously predict an emotional attribute score and establish preferences between speech samples, offering a robust tool for emotion analysis in speech.

Utilization of pre-trained LLM for emotion detection in text has improved the field of natural language processing [17]. The

*Corresponding author.

10.21437/Interspeech.2024-2364

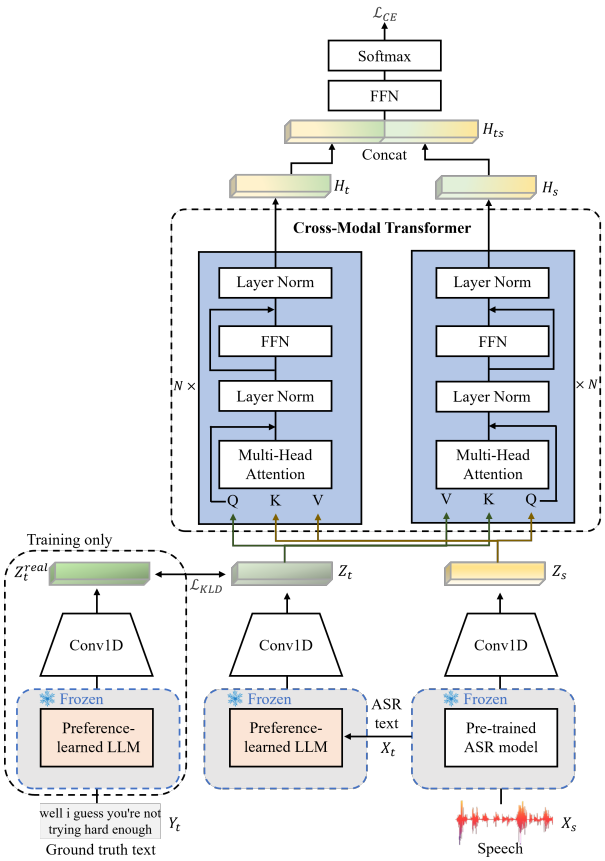


Figure 1: *Architecture of the proposed multimodal emotion recognition system incorporating a pre-trained ASR model, a preference-learned LLM, and a cross-modal transformer. The system utilizes both the ASR-generated text and speech signals to compute latent embeddings, which are then processed through the cross-modal transformer for emotion prediction. KLD loss is applied during training to minimize the discrepancy between ASR-derived and ground truth text embeddings.*

appearance of models like BERT and GPT has provided deep insights into semantic relationships and contextual understanding, which are crucial for detecting nuanced emotional expressions in the text. Fine-tuning these models to promote their emotional sensitivity remains a significant area of exploration. The preference learning methods discussed above offer promising instruments to enhance the ability of LLM to discriminate emotional content, leading to more sophisticated and accurate emotion recognition systems.

Our work builds upon these foundations, proposing an MER system that not only corrects ASR text errors but also employs a preference learning strategy to fine-tune LLM, drawn from the methodology inspired by Naini *et al*. [9]. This combination addresses the dual challenge of error-ridden ASR text and the need for nuanced emotional understanding in LLM, making it easier to advance in affective computing and human-computer interaction.

3. Proposed method

3.1. Speech and text data processing

Figure 1 shows the overall framework of our proposed MER system with the pre-trained ASR model and the preference-learned LLM model. Our approach begins with the processing of speech signals, denoted as X_s, through the pre-trained ASR model. This process yields two outputs: the text corresponding to the speech signal, X_t, and the latent embedding of the speech signal, Z_s. The extracted text X_t is further processed through the pre-trained LLM to obtain its latent embedding, Z_t. Similarly, the ground truth text, Y_t, is also subjected to the same Language model to produce the latent embedding of the accurate text, Z_t^{real}.

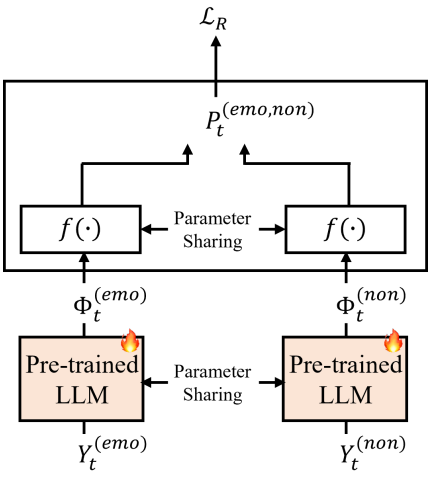


Figure 2: *Schematic of the RankNet-based preference learning framework utilizing a pre-trained LLM. Emotional* (emo) *and non-emotional* (non) *text pairs are processed through shared-parameter LLMs to generate embeddings, which are compared by the shared feed-forward network* $f(\cdot)$ *to calculate the preference probability* $P_t^{(emo,non)}$*. The system is optimized using the RankNet loss* $\mathcal{L}_R$ *to enhance the ability of the LLM to discriminate emotional content in text.*

3.2. Error compensation for text modality

To align the ASR-derived text embeddings Z_t closer to the ground truth text embedding Z_t^{real}, we use Kullback-Leibler divergence (KLD) loss to minimize the distributional differences between Z_t^{real} and Z_t. The KLD loss, $\mathcal{L}_{KLD}$, is defined as follows:

$$\mathcal{L}_{KLD}(Z_t \parallel Z_t^{real}) = \sum Z_t^{real} \log \frac{Z_t^{real}}{Z_t}, \qquad (1)$$

where this loss function acts as a corrective measure, aiming to align the ASR-extracted text embedding Z_t closer to its ground truth text Z_t^{real} and thus mitigate the effects of ASR-induced errors. This process is only demanded during training and is omitted during the inference process. The 1D convolution (Conv1D) block, which follows LLM, is trained to alleviate errors in ASR text.

3.3. Preference learning for enhanced LLM

To further refine the text embedding Z_t, we employ a preference learning [9] strategy on the pre-trained LLM. Using a RankNet framework [15] for preference learning, we fine-tune the LLM to distinguish between emotional and non-emotional content. This approach adapts the LLM to be more sensitive to emotional nuances in speech-related text. The pre-trained LLM is trained separately, as shown in Figure 2, and then merged into the overall system as preference-learned LLM, as shown in Figure 1.

Table 1: *Comparative analysis of emotion recognition accuracies for different methods and modality combinations. The table demonstrates the improvement trajectory from the baseline method to the proposed system, highlighting enhancements achieved through ASR text error compensation and preference learning in LLM. The results underscore the superior performance of the proposed method, which integrates speech and ASR-generated text for emotion detection.*

Method	**Modality**	**Accuracy (%)**			
		Dev WA	Dev UA	Test WA	Test UA
Baseline	Speech	69.24	69.48	66.62	68.21
	ASR text	66.84	67.18	66.10	66.88
	Speech + ASR text	73.13	74.26	74.21	74.52
+ ASR text error compensation	Speech + ASR text	73.33	74.39	75.10	76.44
+ Preference-learned LLM	Speech + ASR text	74.20	75.72	75.44	75.92
Proposed	Speech + ASR text	**75.28**	**75.48**	**76.11**	**77.16**

To fine-tune the LLM with the preference learning method using the RankNet framework, emotional text, $Y_t^{(emo)}$, and non-emotional text, $Y_t^{(non)}$, are given as pairs. The pair of data is given as the input to LLMs that share weights, respectively, and the process of extracting the latent embeddings is as follows:

$$\Phi_t^{(emo)} = \text{LLM}(Y_t^{(emo)}), \tag{2}$$

$$\Phi_t^{(non)} = \text{LLM}(Y_t^{(non)}), \tag{3}$$

where $\Phi_t^{(emo)}$ and $\Phi_t^{(non)}$ denote the latent embeddings of emotional and non-emotional texts, respectively. To model the probability that a given text is more emotional than other text, RankNet uses a sigmoid function as follows:

$$P_t^{(emo,non)} = \frac{1}{1 + e^{-\sigma(f(\Phi_t^{(emo)}) - f(\Phi_t^{(non)}))}}, \tag{4}$$

where $P_t^{(emo,non)}$ is the probability of preferring one sample ($Y_t^{(emo)}$) over the other ($Y_t^{(non)}$) and $f(\cdot)$ is a feed-forward network (FFN) with shared weights. The loss function used to optimize the parameters of LLM and $f(\cdot)$ is as follows:

$$\mathcal{L}_R = -\bar{P}_t^{(emo,non)} \log P_t^{(emo,non)} - (1 - \bar{P}_t^{(emo,non)}) \log(1 - P_t^{(emo,non)}), \tag{5}$$

where $\bar{P}_t^{(emo,non)}$ and $P_t^{(emo,non)}$ are the expected and actual probability, respectively. LLM and $f(\cdot)$ functions are trained to distinguish between emotional and non-emotional text.

3.4. Cross-modal transformer model

Further processing involves the integration of Z_s and Z_t through a cross-modal transformer (CMT) model, which generates H_s and H_t, the transformed embeddings of speech and text, respectively.

$$(H_s, H_t) = \text{CMT}(Z_s, Z_t), \tag{6}$$

where $\text{CMT}(\cdot,\cdot)$ denotes corss-modal transformer model. These embeddings are concatenated to form H_{st}, which is then passed through a feed-forward layer (FFN) and a softmax function to predict the emotion probabilities, denoted as $\hat{Y}_e$.

$$\hat{Y}_e = \text{Softmax}(\text{FFN}(H_{ts})), \tag{7}$$

where H_{ts} is obtained by the concatenation operation of H_t and H_s and $FFN(\cdot)$ denotes feed-forward layer.

To optimize the network of the entire model structure except for pre-trained ASR and LLM in Figure 1, we design a cross-entropy loss that minimizes the discrepancy between the predicted emotion $\hat{Y}_e$ and the ground truth emotion label Y_e. The cross-entropy loss is defined as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} Y_e^{i,c} \log(\hat{Y}_e^{i,c}), \tag{8}$$

where N and C denote the total number of data points and classes, respectively, and i and c denote the indices over all data points and possible classes, respectively. The overall training objective is defined by the total loss function, $\mathcal{L}_{total}$, as below:

$$\mathcal{L}_{total} = L_{CE} + \lambda L_{KLD}, \tag{9}$$

where λ serves as a balancing factor between the cross-entropy loss and the KLD loss.

4. Experiments setup

4.1. Dataset and preparation

All experiments were conducted on the IEMOCAP [16] dataset, which is widely used in emotion recognition recipes [18–20]. The IEMOCAP dataset consists of 12 hours of audio, text, and video recordings with nine emotion labels. The dataset consists of 10 sessions, each containing two actors (five male and five female) having a natural conversation in English. We chose utterances with happy, sad, angry, and neutral emotion labels for the four-class emotion classification. To make a fair comparison with previous studies, we merged the excited emotion label with the happy emotion label. There were a total of 5,531 utterances in the sessions used in all experiments.

4.2. Training and inference procedure

Similar to the experimental setup in the previous study [18, 21], we applied five-fold cross-validation on the IEMOCAP dataset, using three sessions for training, one for validation, and one for testing. We used the Adam optimizer [22] with a learning rate of $1.5e$-4 to train the proposed MER model and fine-tune the LLM. We trained the MER and fine-tuned the LLM model for up to 100 epochs using an early-stop strategy based on the evaluation we performed on the validation set. We used a batch size 32 to train the MER model and 16 to fine-tune the LLM model. We set the hyperparameter to λ = 0.4 in Eq. (9). The proposed method was implemented using PyTorch [23], and four NVIDIA GeForce RTX 3090 GPUs with 24 GB memories were used for the experiments.

Table 2: *Evaluation of emotion recognition performance using different ASR models, comparing the baseline and proposed methods. The table illustrates the relationship between the Word Error Rate (WER) of the ASR model and the respective accuracies, emphasizing the enhanced capability of the proposed system in handling inaccuracies in ASR to improve emotion detection.*

Model	# Parameters	WER (%)	Method	Accuracy (%)	
				Dev WA	Test WA
HuBERT [29]	317M	44.87	Baseline	61.45	60.28
			Proposed	67.98	67.27
W2V2 [30]	317M	39.65	Baseline	67.82	67.94
			Proposed	70.47	71.49
WavLM [31]	316M	27.65	Baseline	69.27	70.11
			Proposed	72.08	73.44
Whisper [24]	1.5B	14.57	Baseline	73.13	74.21
			Proposed	**75.28**	**76.11**

4.3. Model specifications

The speech signals sampled at 16 kHz were converted to the 80-dimensional log-mel spectrogram using the 25 ms Hann window with a hop size of 10 ms. Accordingly, the log-mel spectrogram was used as the input feature for the ASR model.

For the ASR model, we adopted the pre-trained Whisper large-v3 model [24] from the OpenAI downloaded from Hugging Face[1]. The Whisper large-v3 consists of a transformer-based encoder-decoder model, also referred to as a sequence-to-sequence model. The Whisper large-v3 model was pre-trained on one million hours of weakly labeled audio and four million hours of pseudo-labeled speech [24]. The hidden dimensionality of the latent embedding of the speech extracted by the encoder was 1280. To align the hidden dimensionality with the latent embedding of the text to 768, the following consisted of a Conv1D block with kernel size three and stride one. For the LLM model, we also applied the BERT large model [12] from the Google downloaded from Hugging Face[1]. BERT large model is a transformer model pre-trained on a large corpus of English data in a self-supervised fashion [25]. It consisted of 24 layers, 16 attention heads, and 768 hidden dimensions. Similarly, the Conv1D block with kernel size three and stride one follows to align the hidden dimension of the latent embedding to 768. For the CMT architecture, we applied a cross-modal attention-based transformer model [26] consisting of four layers ($N = 4$), eight attention heads, and 768 hidden dimensions. Each layer has a multi-head attention and feed-forward network followed by layer normalization [27] and dropout [28] layers.

4.4. Evaluation metrics

We evaluated the accuracy of our proposed MER system. Since the IEMOCAP dataset suffers from emotion class imbalance [18], we used weighted accuracy (WA) and unweighted accuracy (UA) as evaluation metrics: weighted accuracy represents the accuracy for all classes, and unweighted accuracy is the average value of the accuracy for each class.

5. Results and analysis

5.1. Overall performance analysis

In our experimental evaluation, the proposed MER system, integrating speech and text data through a cross-modal transformer model, significantly improved emotion recognition accuracy. As summarized in Table 1, the proposed system achieved a weighted accuracy for the test dataset (Test WA) of 76.11% and an unweighted accuracy for the test dataset (Test UA) of 77.16%. These results outperform the performance of baseline systems, highlighting the effectiveness of our method in leveraging the complementary strengths of speech and text modalities.

[1]https://github.com/huggingface/transformers.

5.2. Comparison with baseline systems

The comparative analysis, shown in Table 1, illustrates the incremental benefits of each component of our proposed method. Adding ASR text error compensation method and preference-learned LLM progressively enhanced emotion recognition accuracy, starting with a baseline that uses either speech or ASR text modality. Notably, the integration of speech and ASR text modalities, coupled with error compensation method and preference-learned LLM, topped in the highest accuracy, underscoring the value of our comprehensive approach to multimodal emotion recognition.

5.3. Impact of ASR model accuracy

In Table 2, we present a comparative analysis of the performance of the proposed method in situations where different ASR models are used. The results clearly show that the choice of ASR model has a significant impact on the overall accuracy of the system. The Whisper model outperforms the other models by a large margin, which emphasizes the importance of accurate speech-to-text conversion in multimodal emotion recognition. In particular, a notable phenomenon is that when using ASR models with high word error rate (WER), i.e., in situations where the ASR text is error-prone, the accuracy ratio of the proposed method over the baseline system is significantly higher. This phenomenon suggests that the proposed ASR text error correction method is particularly effective in error-prone text. The effectiveness of the error correction on error-rich ASR text is further evident by the larger percentage of performance improvement for models with higher WERs. This provides an important insight that the proposed system is robust to errors in ASR, and that its effectiveness is particularly increased for error-rich text. These results highlight the important role that ASR text error correction plays in improving emotion recognition accuracy, and suggest that it is an important consideration in the design and optimization of multimodal emotion recognition systems.

6. Conclusions

This study introduced a novel multimodal emotion recognition system leveraging integrated speech and text modalities, enhanced by ASR text error compensation and preference learning for LLMs. Our comprehensive evaluation on the IEMOCAP dataset demonstrated the superior performance of our proposed method, achieving significant improvements in both weighted and unweighted accuracy metrics compared to baseline systems. Notably, our approach excelled in environments with high ASR error rates, showcasing the effectiveness of our error compensation mechanism in mitigating the detrimental effects of inaccurate speech-to-text conversion on emotion recognition accuracy. Our research contributes a significant step forward in the field of emotion recognition, providing a robust framework for effective multimodal analysis.

7. Acknowledgment

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00302424) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)).

8. References

[1] H. Xu, X. Wu, and X. Liu, "A measurement method for mental health based on dynamic multimodal feature recognition," *Frontiers in Public Health*, vol. 10, p. 990235, 2022.

[2] N. Shanthi *et al.*, "An integrated approach for mental health assessment using emotion analysis and scales," *Healthcare Technology Letters*, 2022.

[3] L. Tavabi, "Multimodal machine learning for interactive mental health therapy," in *Proc. International Conference on Multimodal Interaction* (ICMI), 2019, pp. 453–456.

[4] N. Ahmed, Z. Al Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, vol. 17, p. 200171, 2023.

[5] K. Ezzameli and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," *Information Fusion*, p. 101847, 2023.

[6] S. Kakuba, A. Poulose, and D. S. Han, "Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model," *IEEE Access*, 2023.

[7] D. Priyasad *et al.*, "Dual memory fusion for multimodal speech emotion recognition," in *Proc. INTERSPEECH*, 2023, pp. 4543–4547.

[8] 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," *arXiv:2401.13260*, 2024.

[9] A. R. Naini, S. Subramanium, S.-G. Leem, and C. Busso, "Combining relative and absolute learning formulations to predict emotional attributes from speech," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), 2023, pp. 1–8.

[10] C. Hema and F. P. G. Marquez, "Emotional speech recognition using cnn and deep learning techniques," *Applied Acoustics*, vol. 211, p. 109492, 2023.

[11] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021.

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

[13] J. Achiam *et al.*, "GPT-4 technical report," *arXiv:2303.08774*, 2023.

[14] S. Hosseini, M. Yamaghani, and S. Poorzaker Arabani, "A review of the methods of recognition multimodal emotions in sound, image and text," *International Journal of Applied Operational Research-An Open Access Journal*, vol. 12, no. 1, pp. 29–41, 2024.

[15] C. Burges *et al.*, "Learning to rank using gradient descent," in *Proc. International Conference on Machine Learning* (ICML), 2005, pp. 89–96.

[16] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[17] S. Lei *et al.*, "InstructERC: Reforming emotion recognition in conversation with a retrieval multi-task llms framework," *arXiv:2309.11911*, 2023.

[18] J. Kyung, J.-S. Seong, J.-H. Choi, Y.-R. Jeoung, and J.-H. Chang, "Improving joint speech and emotion recognition using global style tokens," in *Proc. INTERSPEECH*, 2023, pp. 4528–4532.

[19] Y. Xia *et al.*, "Temporal context in speech emotion recognition." in *Proc. INTERSPEECH*, 2021, pp. 3370–3374.

[20] H. Zhao, B. Li, and Z. Zhang, "Speaker-aware cross-modal fusion architecture for conversational emotion recognition," in *Proc. INTERSPEECH*, 2023, pp. 2718–2722.

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

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

[23] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[24] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. International Conference on Machine Learning* (ICML), 2023, pp. 28 492–28 518.

[25] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Proc. Advances in Neural Information Processing Systems* (NeurIPS), 2019.

[26] J. Yan *et al.*, "Cross modal transformer: Towards fast and robust 3d object detection," in *Proc. IEEE/CVF International Conference on Computer Vision* (ICCV), 2023, pp. 18 268–18 278.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

[28] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] W.-N. Hsu *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (TASLP), vol. 29, pp. 3451–3460, 2021.

[30] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Advances in Neural Information Processing Systems* (NeurIPS), 2020, pp. 12 449–12 460.

[31] S. 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.