Interspeech 2024
1-5 September 2024, Kos, Greece

Speech Emotion Recognition with Multi-level Acoustic and Semantic Information Extraction and Interaction

Yuan Gao, Hao Shi, Chenhui Chu, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Kyoto, Japan
gao@sap.ist.i.kyoto-u.ac.jp

Abstract

Speech emotion recognition (SER) systems can learn linguistic information by integrating automatic speech recognition (ASR). However, existing SER systems fall short in explicitly learning semantic emotional information from ASR predictions. Our proposed system addresses this problem by incorporating a semantic feature extractor for explicit emotional information extraction. Furthermore, a cross attention-based information interaction module is proposed to learn the complementary emotional information in the embeddings from both feature extractors. Within the interaction module, a temporal-aware gate fusion network is incorporated to dynamically integrate the embeddings from acoustic and semantic feature extractors and mitigate the impact of ASR errors in SER. Experimental results on IEMOCAP show that our system outperforms the existing SER systems by improving the unweighted accuracy by 3.32%.
Index Terms: speech emotion recognition, speech recognition, human-computer interaction, attention

1. Introduction

Speech is a vital component of human communication and an essential medium for expressing emotions. Consequently, speech emotion recognition (SER), which aims to discern human emotion from spoken audio, has become crucial for building empathic human-computer interactions (HCI). In recent years, learning emotional states for virtual voice assistants and chatbots has become increasingly popular, thus making SER an active research. It is also used to augment call center services by analyzing customer sentiment and aiding in detecting and treating mental health issues through vocal cues.

The inherent difficulties in collecting and annotating emotional speech data have resulted in limited training samples in existing SER corpora. Traditional deep learning-based methods have primarily focused on convolutional neural networks (CNNs) [1, 2] to identify the acoustic patterns (such as pitch, energy, and formants) and recurrent neural networks (RNNs) [3, 4] to learn latent features from speech. Moreover, the association between linguistic information and emotional cues is well-established, highlighting the critical role of ASR in enhancing the understanding of emotions. However, ASR performance of emotional speech is unsatisfactory. Therefore, instead of using the output transcription, a previous study [5] proposed to improve SER using the hidden output of the ASR model.

In recent years, self-supervised learning (SSL) models such as Wav2Vec-2.0 [6] have shown promising performance in learning both low-level acoustic features and linguistic information [7]. It is pre-trained on a large amount of unlabeled data to embed the prior linguistic knowledge, enabling it to be fine-tuned for various tasks such as SER [8], ASR, [9] and speech separation [10]. For instance, recent studies [11] confirmed the effectiveness of leveraging pre-learned Wav2Vec-2.0 in improving SER over traditional deep learning models [12]. Besides, the SSL models are also beneficial for ASR since they have already learned coarser linguistic information. Thus, finetuning Wav2Vec-2.0 for low-resource ASR achieved promising performance with only one hour of training data from the target language [13]. These advantages of SSL models fit the problem of data spareness of emotional speech. With the SSL models, the previous study [14] improved SER by training SER and ASR within the shared acoustic model (Wav2Vec-2.0). The experimental results confirmed that a better ASR-trained model yields better SER results.

Although implicitly learning linguistic information helps the model better capture the emotional state from speech [14, 15], there are two main limitations in speech-only models. Firstly, most existing SER systems lack the integration of a semantic extraction module, which explicitly learns emotional cues from the linguistic output of ASR. Secondly, the ASR-predicted text inevitably includes errors of differing severity in each sentence, typically resulting in degraded SER performance compared to ground truth text transcription. The multimodal systems, which combine the speech and ground-truth transcription as input features, usually perform better compared with speech-only unimodal systems [16, 17]. However, access to the ground-truth transcription is not practical in most application contexts. A previous study [18] explored extracting semantic information from speech only by encoding the ASR output using a semantic feature extractor (BERT). However, their system still suffers from ASR errors, which affect the training and evaluation of the semantic feature extractor. Furthermore, the complementary information in acoustic and semantic features is not well explored. These highlight the need for discriminative acoustic and semantic information learning from speech.

This paper proposes a multi-level information extraction and interaction model, which uses only speech as input, leverages the acoustic and semantic information to enhance SER. Firstly, we train the pre-trained acoustic feature extractor for SER and ASR to explicitly embed the emotional and linguistic information. Then, joint training on the acoustic and semantic feature extractor for SER is conducted. During joint training, the transcription is leveraged as the input feature of the semantic feature extractor to learn more discriminative semantic representation. After extracting the acoustic and semantic features, a cross-modal gated interaction (CmGI) module is introduced for multi-modal feature fusion. It learns the complementary information from multi-model features. Moreover, a temporal-aware gated mechanism is adopted to dynamically modulate the contribution of each feature representation, effectively mitigating the impact of ASR errors during the inference phase.

10.21437/Interspeech.2024-2385

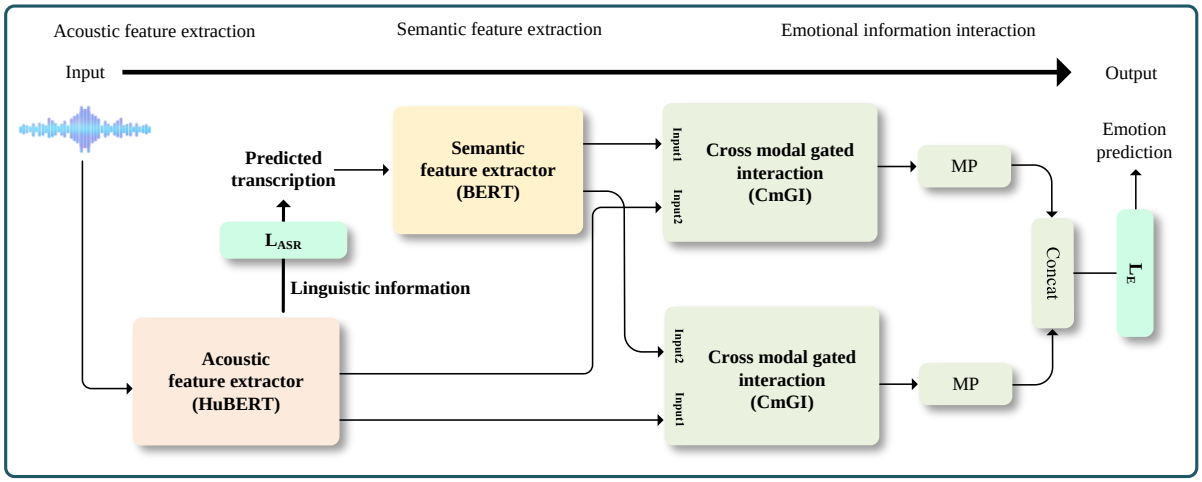

Figure 1: *Overall flowchart of the proposed information interaction system for SER.*

2. Proposed method

In the proposed system (Figure 1), acoustic and linguistic information is initially extracted from the input speech. Subsequently, emotional semantic information is derived from the linguistic data, thereby establishing a multi-level analysis framework. Finally, an information interaction module is employed to learn the complementary emotional information obtained from both acoustic and linguistic features.

2.1. Acoustic and semantic information extraction for SER

This section introduces the acoustic (HuBERT) and semantic (BERT) emotion feature extractors. Given an input speech u, we incorporate HuBERT, which consists of 7 CNN layers and 24 Transformer layers, to extract the acoustic emotion representation $x^a \in \mathbb{R}^{t\times d}$. Here, t denotes the length of the utterance, and d is the hidden dimension of the Transformer layer. For the ASR module, we fed x^a into a fully connected layer (FC layer) and applied the connectionist temporal classification (CTC) loss function to learn the linguistic information, which can be used in the semantic emotion feature extractor.

$$L_{\text{ASR}} = \text{CTC}(x^a, T) \tag{1}$$

where L_{ASR} is the loss function of ASR. During the training phase, the ground-truth transcription T is used as the input of BERT, which is composed of an embedding layer followed by Transformer layers, to extract the semantic emotion representation $x^l \in \mathbb{R}^{t\times d}$. For the inference phase, only speech is used as the input, and we decode the output of the ASR module to get predicted transcription T_{pred}. Then x^l is extracted from T^{pred} using BERT. The overall objective function L is defined as:

$$L = \alpha L_{ASR} + (1-\alpha) L_E \tag{2}$$

where α is a weight parameter, and L_E is the SER objective function, which will be introduced in the following section.

2.2. Cross-modal gated interaction

2.2.1. Intra- and inter-modal interaction learning

The feature extractors learn the emotional information from audio and text inputs independently. Acknowledging that human emotion analysis includes multiple sources, we incorporate dual branches of the cross-modal gated interaction (CmGI) module,

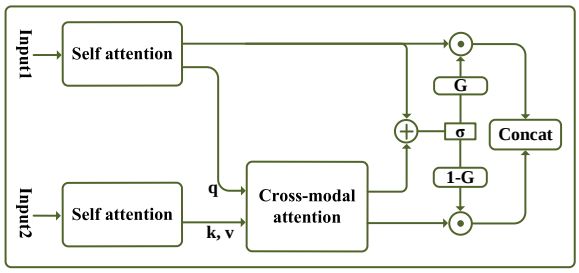

Figure 2: *Structure of cross-modal gated interaction (CmGI).*

which learns the mutual relationship by integrating emotional information of x^a into x^l and vice versa. As shown in Figure 2, our model first employs self-attention (SA) mechanisms for learning intra-modal interactions within acoustic and semantic embeddings.

$$\begin{aligned} s^a &= \text{softmax}\left((x^a W^Q)(x^l W^K)^T/\sqrt{d_k}\right)(x^l W^V) \\ s^l &= \text{softmax}\left((x^l W^Q)(x^a W^K)^T/\sqrt{d_k}\right)(x^a W^V) \end{aligned} \tag{3}$$

where $s^a \in \mathbb{R}^{t\times d}$ and $s^l \in \mathbb{R}^{t\times d}$ represents the output features of SA that can be either learned from x^a or x^l. The matrices W^Q, W^K, and W^V are the learned weights for queries, keys, and values. These layers ensure the preservation of emotional information during feature compression. Subsequently, s^a and s^l are fed into the cross-attention (CA) mechanism.

$$\begin{aligned} c^a &= \text{softmax}\left((s^a W^Q)(s^l W^K)^T/\sqrt{d_k}\right)(s^l W^V) \\ c^l &= \text{softmax}\left((s^l W^Q)(s^a W^K)^T/\sqrt{d_k}\right)(s^a W^V) \end{aligned} \tag{4}$$

where $c^a \in \mathbb{R}^{t\times d}$ ingegrates emotional information from s_l to s^a, and $c^l \in \mathbb{R}^{t\times d}$ ingegrates emotional information from s^a to s_l. This mechanism ensures the model captures the inter-modal interactions between acoustic and semantic embeddings.

2.2.2. Temporal-aware gated fusion model

In emotion recognition, the significance of information from different modalities is different. For instance, when negative emotions are expressed sarcastically (e.g., "That's great"), analysis of emotion based on linguistic information is difficult. However, the model can detect the anger of the speaker from the

Table 1: *Comparison of different approaches integrating ASR and semantic extraction module for SER. The modules for SER, ASR, and semantic information extraction used in each system are listed. "GT" represents the ground truth; "Pred" represents the ASR predicted transcription.*

Exp	Systems			Input of feature extractor			SER		ASR
				Acoustic	Linguistic				
	SER	ASR	Semantic	Speech	GT	Pred	UA	WA	WER
Speech-only Models									
1	HuBERT	✗	✗	✓			70.32	70.84	-
2	HuBERT (shared)		✗	✓			75.28	75.13	13.57
3 [19]	LSTM-LAS (shared)		✗	✓			64.40	63.10	56.40
4 [5]	LSTM	LSTM-Attention	✗	✓			69.70	68.60	35.70
5 [20]	LSTM-Attention	DNN-HMM	LSTM-Attention	✓		✓	75.90	76.10	43.50
6 [18]	Wav2Vec-2.0 (shared)		BERT	✓		✓	-	74.20	15.00
7 **(Ours)**	HuBERT (shared)		BERT	✓		✓	**77.25**	**76.86**	13.61
Text-only Models									
8	✗	✗	BERT		✓		67.13	66.85	-
Speech-Text Multi-model Models									
9	HuBERT	✗	BERT	✓	✓		72.15	73.07	-
10	HuBERT (shared)		BERT	✓	✓		77.64	77.36	13.64

acoustic features. In such scenarios, it is crucial for the model to increase the importance of x^a for emotion recognition. Prior work [21] utilized a gated-fusion model for integrating inputs.

$$\begin{aligned} G &= f(\sigma(\text{concat}(\text{mp}(x^a), \text{mp}(x^l)))) \\ output &= \text{concat}(x^a \odot G, x^l \odot (1 - G)) \end{aligned} \tag{5}$$

where $G \in \mathbb{R}^{1\times d}$ is the gated vector, and mp denotes mean-pooling (MP). The conventional method applies MP across the time dimension before fusion overlooked temporal dynamics. To overcome this limitation, we introduce a temporal-aware gated fusion model to combine SA and CA outputs:

$$\begin{aligned} G^a &= f(\sigma(\text{concat}(s^a, c^a))) \\ G^l &= f(\sigma(\text{concat}(s^l, c^l))) \\ output^a &= \text{concat}(s^a \odot G^a, c^l \odot (1 - G^a)) \\ output^l &= \text{concat}(s^l \odot G^l, c^a \odot (1 - G^l)) \end{aligned} \tag{6}$$

Let f denote the transformation layer that maps the gate vector to the same dimension with c^a and c^l, and σ is the sigmoid activation function. $G \in \mathbb{R}^{t\times d}$ is the learned gated vector, and $\odot$ denotes element-wise product. The proposed temporal-aware gated fusion enable our model to retains temporal dynamics of each frame of u and each subword of t. Finally, we concat $output^a$ and $output^l$ for SER.

$$L_E = CE(\text{concat}(\text{mp}(output^a), \text{mp}(output^l)), Y) \tag{7}$$

where Y is the emotion label.

3. Experimental Setup

3.1. Database

In this study, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [22], which contains 12 hours of audio, video, facial motion data, and textual transcriptions, to evaluate the proposed methods. Ten American English speakers record five speaker-independent dyadic sessions, and each session is performed by two speakers (one male and one female) with either a series of scripted or improvisational scenarios. We use the data with high agreement of categorical emotion label and implement the common practice of merging "happy" and "excited" into one emotion class titled "happy" [1, 23, 24, 25], resulting in 5,531 utterances with four emotions: happy (1,636), sad (1,084), angry (1,103), and neutral (1,708). The evaluation criteria are unweighted accuracy (UA) and weighted accuracy (WA). We follow the common practice [1, 26, 27] and report the performance of 5-fold speaker-independent cross-validation.

3.2. Experimental settings

We implemented our proposed models with the PyTorch framework and the Huggingface Transformers repository [28]. We use HuBERT-large [29] as the acoustic feature extractor, which is pretrained with 60,000 hours of Libri-Light. This pretrained model consists of 7 CNN layers to transform the raw waveform into latent representations and 24 Transformer layers to learn the linguistic information from speech. The semantic feature extractor used in this work is BERT-base [30], which is pretrained on BooksCorpus and the text passages in English Wikipedia. This model consists of 12 Transformer layers to learn the semantic embedding from the textual input. The dimensions of the hidden layers d for HuBERT-large and BERT-base are 1024 and 768, respectively. For arbitrary length inputs of both HuBERT and BERT, we applied sentence padding within each mini-batch. Throughout the training process, HuBERT is first finetuned with SER and ASR. Then, we froze the CNN layers in HuBERT and finetuned the Transformers in HuBERT and BERT simultaneously for SER while keeping the ASR solely on HuBERT. The learning rate is set as 10e-5, and

Table 2: *Comparison of information interaction modules.*

Model	UA	WA
Self-attention	78.10	77.58
/w gated fusion	78.22	78.37
Cross-modal attention	78.27	78.03
/w gated fusion	78.39	78.21
CmGI	**79.62**	**79.50**
w/o temporal gated fusion	78.85	78.54

Table 3: *Comparsion results with previous SER systems.*

Approach	Year	UA	WA
Pepino et al. [11]	2021	67.20	-
Santoso [20]	2021	75.90	76.10
Zou et al. [31]	2022	71.05	69.80
Ioannide et al. [32]	2023	-	74.32
Kyung et al. [33]	2023	76.30	75.10
Gao et al. [34]	2023	76.10	74.94
Ours	-	**79.62**	**79.50**

the mini-batch size was 2 with a gradient accumulation of 8. The weight parameter α is 0.1 for ASR and SER.

4. Results and Analysis

4.1. ASR and semantic information for SER

To investigate the effect of semantic information learned from ASR prediction, we first evaluated the proposed information extraction module (without the information interaction part) with previous approaches.

As shown in Table 1, when only an acoustic feature extractor is used, the performance of SER exhibits a 4.52% improvement in UA when ASR is incorporated (compared Exp.–1 with Exp.–2). This observation is in line with previous works [14]. In Exp.–3 and –4, the hidden representation of the ASR model was incorporated to improve SER. However, the improvement is not significant due to the limited ASR performance of LSTM-based models on emotional datasets. Exp.–5 [20] was the first work that encoded emotional information from ASR prediction and demonstrated the benefit of semantic feature encoding even with high WER. In Exp.–6 [18], BERT was used to learn the emotional semantic information from the ASR prediction. They utilized a pretrained SSL model and achieved better ASR performance. However, the overall performance of their approach is lower than in Exp.–5 even with better WER. While Exp.–5 explicitly trained the acoustic feature extractor before introducing the semantic feature extractor, Exp.–6 used single-stage training for both feature extractors. This indicates that finetuning HuBERT for ASR before finetuning BERT for semantic extraction is crucial. We first train the HuBERT module for SER and ASR. Then, both feature extractors are trained simultaneously for SER. Moreover, we use the ground-truth transcription to train the semantic feature extractor. These training schemes ensure that our system outperformed Exp.–6 for more than 2% on WA. Moreover, the proposed information interaction module improved more significantly over previous systems, which will be discussed in the next section.

Upon comparing Exp.–8 with Exp.–1-7, it is observed that current systems can extract more discriminative emotional information from speech than text. As depicted in Table 1, emotion recognition results using text or multimodal inputs are also provided. Comparing Exp.–7 and Exp.–10, the information extraction system achieved comparable performance to the multimodal approach using only speech input. Lastly, employing ASR in the acoustic feature extraction process enhances emotion recognition in multimodal systems (Exp. –9 vs. Exp. –10), as embedding linguistic information into the extractor improves feature extraction and benefits the final decision-making.

4.2. Evaluation of the proposed CmGI

In our experimental results in Table 2, we evaluate the performance of the proposed CmGI module.

The results demonstrate that although cross-modal attention outperforms self-attention by leveraging inter-modal interactions, its effectiveness is limited. Moreover, the integration of the traditional gated fusion mechanism showed minimal improvement, primarily due to its inability to learn the temporal dynamics within the input features. On the other hand, our CmGI module incorporates a temporal-aware gated fusion network, which preserves the time dimension of input features during fusion. This allows it to maintain the temporal dynamics of input features and better handle modality incongruity. Compared to the conventional gated fusion model combined with cross-modal attention, our method achieved a 1.35% absolute improvement in UA.

4.3. Comparison with existing SER approaches.

In this section, we benchmark the performance of the proposed approach against existing representative studies that reported using 5-fold cross-validation on the IEMOCAP. Table 3 provides a comparative summary. Our approach outperforms recent studies by more than 3.32% and 3.40% on UA and WA, respectively. The results demonstrate the superiority of our proposed multi-level information extraction and interaction system.

5. Conclusion and Future Work

We have proposed a novel SER system using only speech input. Firstly, the system encodes semantic features from the text prediction from ASR, ensuring the model explicitly extracts the acoustic and semantic information using only speech input. Secondly, the proposed cross-modal gated interaction module can effectively learn the complementary information from acoustic and semantic features and mitigate the impact of ASR error in SER by a temporal-aware gated fusion model. Finally, the comparison with existing speech-based approaches showed that our proposed system significantly outperformed state-of-the-art results. Future work will explore multi-attribute learning from speech, such as gender, to enhance feature encoding by incorporating additional emotion-related information into our system.

6. Acknowledgment

This work was supported by JST SPRING (JPMJSP2110), JST Moonshot R&D (JPMJPS2011), Grant-in-Aid for Scientific Research KAKENHI (JP19H05691, 20H00602 and 23K28144).

7. References

[1] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Proc. INTERSPEECH*, 2017, pp. 1089–1093.

[2] H. Shi, M. Mimura, L. Wang, J. Dang, and T. Kawahara, "Time-domain speech enhancement assisted by multi-resolution frequency encoder and decoder," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[3] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. Interspeech 2015*, 2015, pp. 1537–1540.

[4] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5115–5119.

[5] H. Feng, S. Ueno, and T. Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr model." in *Proc. INTERSPEECH*, 2020, pp. 501–505.

[6] 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.

[7] S. Dang, T. Matsumoto, Y. Takeuchi, H. Kudo, T. Tsuboi, Y. Tanaka, and M. Katsuno, "Using self-learning representations for objective assessment of patient voice in dysphonia," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 359–363.

[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 preprint arXiv:2401.13260*, 2024.

[9] H. Shi, M. Mimura, and T. Kawahara, "Waveform-domain speech enhancement using spectrogram encoding for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–12, 2024.

[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.

[11] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[12] Y. Gao, J. Liu, L. Wang, and J. Dang, "Domain-adversarial autoencoder with attention based feature level fusion for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6314–6318.

[13] K. Soky, S. Li, C. Chu, and T. Kawahara, "Domain and language adaptation using heterogeneous datasets for wav2vec2.0-based speech recognition of low-resource language," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning." in *Proc. INTERSPEECH*, vol. 2021, 2021, pp. 4508–4512.

[15] Y. Li, P. Bell, and C. Lai, "Fusing asr outputs in joint training for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7362–7366.

[16] P. Kumar, V. Kaushik, and B. Raman, "Towards the explainability of multimodal speech emotion recognition." in *Proc. INTERSPEECH*, 2021, pp. 1748–1752.

[17] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross-and self-attention network for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4275–4279.

[18] L. Bansal, S. P. Dubagunta, M. Chetlur, P. Jagtap, and A. Ganapathiraju, "On the Efficacy and Noise-Robustness of Jointly Learned Speech Emotion and Automatic Speech Recognition," in *Proc. INTERSPEECH*, 2023, pp. 1863–1867.

[19] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "Speech representation learning for emotion recognition using end-to-end asr with factorized adaptation." in *Proc. INTERSPEECH*, 2020, pp. 536–540.

[20] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, and T. Hiramura, "Speech emotion recognition based on attention weight correction using word-level confidence measure." in *Proc. INTERSPEECH*, 2021, pp. 1947–1951.

[21] P. Liu, K. Li, and H. Meng, "Group Gated Fusion on Attention-Based Bidirectional Alignment for Multimodal Emotion Recognition," in *Proc. Interspeech 2020*, 2020, pp. 379–383.

[22] 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, no. 4, pp. 335–359, 2008.

[23] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech." in *Proc. INTERSPEECH*, 2021, pp. 3415–3419.

[24] H. Zhang, M. Mimura, T. Kawahara, and K. Ishizuka, "Selective multi-task learning for speech emotion recognition using corpora of different styles," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7707–7711.

[25] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.

[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*, vol. 2023, 2023, pp. 765–769.

[27] 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.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

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

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

[32] G. Ioannides, M. Owen, A. Fletcher, V. Rozgic, and C. Wang, "Towards Paralinguistic-Only Speech Representations for End-to-End Speech Emotion Recognition," in *Proc. INTERSPEECH*, 2023, pp. 1853–1857.

[33] 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.

[34] 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*, pp. 3637–3641.