Collaborative Enhanced Attention for Speech Emotion Recognition Based on Multimodal Acoustic Information Fusion

Ping Lin1, Can Bu1 and Yongming Huang1()

1 School of Automation, Southeast University, Nanjing, China  
huang\_ym@seu.edu.cn

**Abstract.** Speech emotion recognition (SER) aims to infer human emotions based on speech information. Effectively and comprehensively combining the various pieces of information in the audio for accurate judgment remains a significant challenge. In this paper, we propose a novel SER network based on a Collaborative Enhanced Attention (CEA) mechanism. Initially, we apply wavelet transform to translate the raw speech into the frequency domain, obtaining log spectrograms. We then use EfficientNet and Mamba to extract acoustic features. Finally, CEA and Co-attention are used to fuse these features with high-level acoustic information embedded through the WavLM model. The CEA is specifically designed to refine the acoustic information within WavLM features by accentuating different aspects of the signal, thereby enhancing the feature representation from multiple perspectives. We evaluate our method on the IEMOCAP dataset, where the weighted accuracy and unweighted accuracy improved by 4.96% and 4.71%, respectively. Our code is available on https://github.com/Chara2001/CEA-SER.git.

**Keywords:** Speech emotion recognition, Collaborative enhanced attention, Wavelet transform, Multimodal fusion.

1. Introduction

Speech emotion recognition (SER) has become an important component of human-computer interaction. The emotional modality in speech is applied in some spoken dialogue systems, such as call center conversations, in-vehicle driving systems, and others [1]. Despite the emergence of some multimodal methods that have achieved high recognition accuracy in emotion recognition [2-4], these models are not widely applied, and there may be redundancy and interference of information between different modalities. In contrast, the speech modality is easier to obtain, so we focus on the research of SER tasks.

The SER method mainly consists of two stages: feature extraction and feature classification. In the feature extraction stage, Chen et al. [5] proposed a deformable speech transformer called DST to adaptively discover and focus on valuable embedded information in speech; Zhang et al. [6] introduced an autoencoder with emotional embedding to extract deep emotional features. In the feature classification stage, Ong et al. [7] proposed an improved multi-scale visual transformer (MViTv2) to effectively model interactions between labels in spatiotemporal structures; Ye et al. [8] introduced a time-aware bi-directional multi-scale network (TIM-Net) to learn multi-scale contextual emotional representations from various temporal scales. Although these methods have improved the accuracy of SER, they only use one type of speech information, which inevitably leads to the loss of other speech feature information.

To address this issue, Zou et al. [9] proposed a SER method based on multi-level acoustic information, which simultaneously uses three acoustic features: Mel-Frequency Cepstral Coefficients (MFCC), log spectrogram, and features extracted by the Wav2Vec2 model. He et al. [10] employed a cross-attention transformer to process dual-source inputs based on these three acoustic features. Pan et al. [11] replaced the Wav2Vec2 model with the Hubert model and proposed a Global Perception Feature Representation Network (GFRN-SEA) for speech emotion analysis. These studies all introduced different attention mechanisms to process the extracted features. However, since general-purpose models like Wav2Vec2 may extract features that contain not only emotional information but also a lot of other information, this could affect the model's emotional recognition performance. Therefore, a mechanism is needed that can collaborate with different feature information and the advanced acoustic information embedded in the WavLM model, in order to enhance its emotional information.

In this study, we propose a SER method based on the Collaborative Enhancement Attention (CEA) module for multimodal acoustic information fusion. In this method, we select three types of features as network inputs: logarithmic spectrogram, MFCC, and large-model speech features. However, unlike traditional methods, we apply the wavelet transform to convert the original audio signal into the frequency domain. This transformation effectively captures the time-varying characteristics of speech signals, thereby enhancing the ability to extract emotional features. Next, we employ EfficientNet for further feature extraction from the spectrogram. We introduce the Mamba [12] model to extract further feature to MFCC, which effectively captures both short-term and long-term dependencies, significantly improving the model's emotion recognition performance. Additionally, we utilize the advanced large-scale speech model WavLM [13] to directly extract features from the raw audio signal. In the feature fusion stage, we designed the CEA mechanism, which dynamically learns attention weights to effectively enhance emotion-related information in the WavLM features (WLM). The CEA mechanism also applies weighted adjustments to each frame of the signal using spectrogram and MFCC features, compensating for time-frequency information that may be lost during training. Finally, the processed features are concatenated and fed into a classifier composed of three fully connected layers to achieve emotion classification.

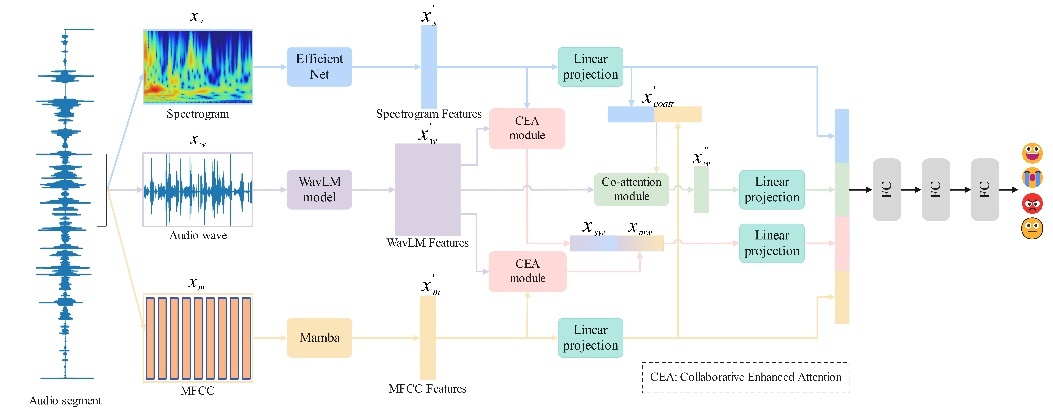
Our main contributions are summarized as follows:

* We propose an SER network that integrates logarithmic spectrogram, MFCC, and self-supervised large model features, with EfficientNet, Mamba, and WavLM employed respectively to further extract emotional information from these three features.
* We innovatively apply the Continuous Wavelet Transform (CWT) to process the original audio, effectively capturing the emotional characteristics of short-term speech in the frequency domain.
* We design a novel Collaborative Enhancement Attention mechanism, which enhances emotional information in features through dual pathways and compensates for missing information, enabling the network to perform more comprehensive emotion recognition.

1. Proposed Method

This section introduces the proposed SER method. In the following subsections, we will provide a detailed description of the feature extraction module and the proposed collaborative enhancement attention mechanism.

* 1. Overall Architecture



**Fig. 1.** The main framework of our proposed method

The overall SER method is shown in **Fig. 1**. We represent the MFCC, log-spectrogram, and WavLM obtained from the same audio segment as ,  and , respectively. This method first uses EfficientNet [14] to extract frequency features from the log-spectrogram obtained via wavelet transform. The Mamba module is then used to extract timbral and temporal features from the MFCC. Additionally, WavLM, a large pre-trained model, is employed to directly extract feature sequences from the raw audio signal, without the need for fine-tuning. Next, the extracted MFCC features and spectrogram features are passed through linear layers for transformation, in preparation for fusion with WLM. The three features are further processed using co-attention [9] mechanisms and collaborative enhancement attention mechanisms to capture dependencies between different parts of the features, thus enhancing the model's ability to learn emotional features. Finally, the MFCC features, spectrogram features, and the features processed by the two attention mechanisms are concatenated and fed into a classifier composed of three fully connected layers for emotion recognition.

* 1. MFCC Processing

MFCC is a core feature in speech signal processing, designed to simulate the nonlinear frequency perception characteristics of the human auditory system. We process the MFCC sequence through the Mamba module, with a state sequence factor of 64, a local convolution kernel width of 4, and a block expansion factor of 2. As shown in **Fig. 1**, the Mamba module first obtains the MFCC input embedding through linear projection. Then, a convolution operation is applied before the Selective State Space Model (Selective SSM). Afterward, the linearly projected state after SSM processing is multiplied by the activated linear projection, and finally passed through another layer of linear projection to generate the output. The calculation process of SSM is as follows:

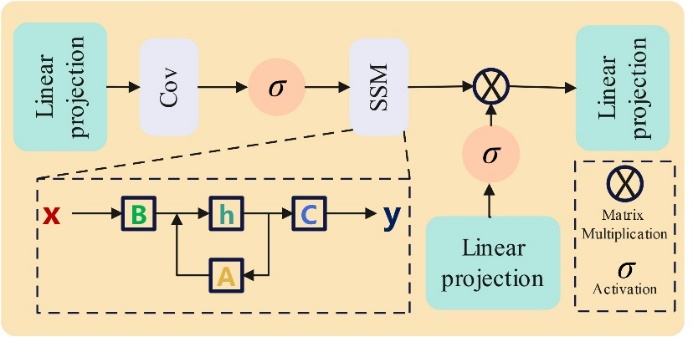
 (1)

here,  represents the hidden state at the current time step,  is the input at the current time step,  is the updated hidden state,  is the output at the current time step, and ,  and  are learnable parameters.

Selective SSM expands the model's capability by introducing a selective mechanism, allowing it to better capture the complex dynamic behaviors in MFCC and extract features that influence these dynamics. At the same time, selective SSM focuses only on the states that are important for emotion recognition, significantly reducing the computational load and improving computational efficiency. The structure of SSM is shown in **Fig. 2**. The processed vectors are then fed into a linear layer, with Rectified Linear Unit (ReLU) as the activation function and a dropout rate of 0.1, in order to obtain:

 (2)

where ,  represents the linear transformation and activation operation applied to the output features of Mamba.



**Fig. 2.** The structure of selective SSM

* 1. Log-spectrogram Processing

SER tasks require extracting emotion-related features such as pitch, speech rate, and volume variations from speech signals. Thanks to the time-frequency localization property of wavelet transform, this method can more accurately capture short-term fluctuation features of speech signals, thereby improving the accuracy of emotion recognition. For emotional speech with rapid and instantaneous changes, features extracted by Fourier transform may not effectively reflect the dynamic changes in emotion. Therefore, this study uses CWT to obtain the log-spectrogram of the original speech, as shown in the following formula:

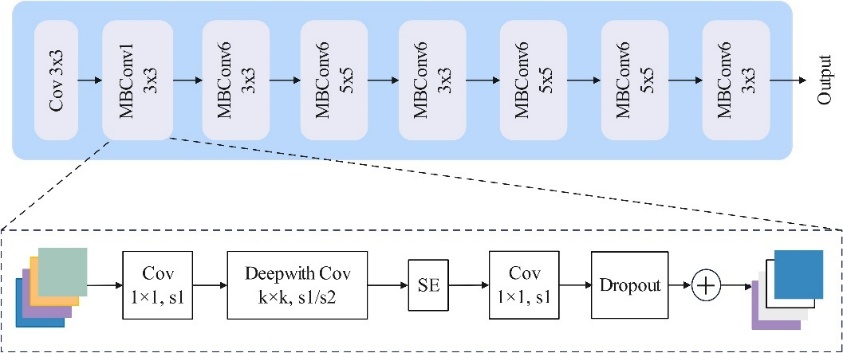
 (3)

where  is the wavelet transform coefficient,  is the original signal,  is the scaled (scale) and translated (translation) mother wavelet function,  is the scale parameter, which controls the frequency resolution,  is the translation parameter, which controls the time resolution,  represents the complex conjugate of the mother wavelet. In this paper, we use the Morlet wavelet as the mother wavelet.

Next, we encode the spectrogram features using the pre-trained EfficientNet (network structure detailed in **Fig. 3**). EfficientNet is an efficient CNN architecture proposed by the Google Brain team, which achieves better performance with fewer parameters. We then follow the same processing steps as in MFCC, ultimately obtaining:

 (4)

where ,  represents the linear transformation and activation operation applied to the output features of EfficientNet.



**Fig. 3.** The structure of EfficientNet

* 1. WavLM Feature Processing

Through pretraining on large-scale unlabeled speech data, WavLM can extract speech features that contain rich fine-grained semantic and acoustic information. Due to the advanced structure of WavLM, we use WavLM to directly extract features from the raw audio like following:

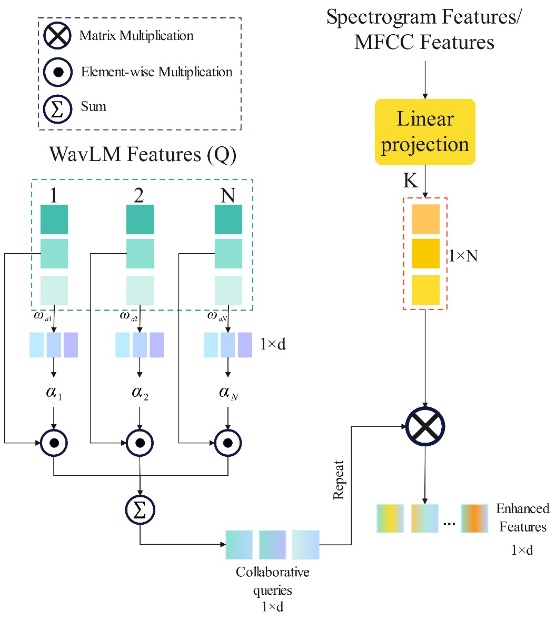
 (5)

where ,  represents the linear transformation and activation operation applied to the WLM.

* 1. Collaborative-Enhanced-Attention Based Fusion

Related researches indicate that weighting the different frames of features extracted by large models can provide guidance for SER tasks [9, 11]. We use WavLM as a feature extractor for the raw audio input, where the output of WavLM represents the input audio signal (typically raw audio waveform) at multiple time steps. In addition to emotion-related information in the audio, these representations also contain information useful for other downstream tasks. We refer [9] to previous work and incorporate a co-attention mechanism into the network. However, we believe that simply using co-attention alone cannot effectively assist in the fusion of the three types of features and cannot fully extract the emotional information from the WLM. In this context, we propose CEA for the SER task, as shown in the computational flow in **Fig. 4**. In CEA, we use  to represent ,  to represent  and  to represent .  and  will be used to collaborate with , considering both frequency and time perspectives. First, , as the vector to be collaborated and enhanced, is multiplied by a learnable weight matrix to learn the enhanced attention weights, thus generating the enhanced attention vector :

 (6)



**Fig. 4.** Collaborative-Enhanced-Attention module

Then, based on the learned attention weights, the features in  are enhanced, allowing the model to more effectively focus on emotion-related information, resulting in a single collaborative queries :

 (7)

Since the collaborative queries  obtained after this process is derived by summing all the frames of , it clearly loses some information along the sequence dimension. Therefore, we use  and  to perform collaboration along the sequence dimension on the stacked  matrix formed by repeating the  vector, thereby supplementing the lost information. In this process,  collaborates with  in the frequency domain to obtain , while  collaborates with  in the temporal domain to obtain:

 (8)

We concatenate  and , and input them into a linear layer to obtain:

 (9)

where .

Finally, we concatenate the MFCC features, logarithmic spectrogram features, and the WavLM features processed by both the co-attention and collaborative enhancement attention mechanisms. The final emotion prediction output is:

 (10)

where  represents the WLM features processed by the co-attention mechanism.

1. Experiments
   1. Dataset

We validate the proposed method on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [15] dataset. This dataset is widely used in various emotion recognition tasks and consists of 10 actors, with 5 different sessions, each performed by two native English-speaking actors (one male and one female). Following previous work [9, 16], we merge the “excited” and “happy” categories into a single “happy” category. We consider a total of 5531 voice samples from four emotional categories: anger, sadness, happiness, and neutral. To compare with baseline methods, we test our model using 10-fold leave-one-speaker cross-validation and evaluate performance using the commonly used weighted accuracy (WA) and unweighted accuracy (UA) metrics.

* 1. Experiment Setup

The raw audio in the dataset has a sampling rate of 16 kHz. In our experiments, each audio sample is segmented into several 3-second clips. If the segment length is shorter than 3 seconds, we pad it with zeros. We use the librosa library to extract 40-dimensional MFCCs from the raw audio segments. For frequency-domain transformation, we apply CWT using the morlet wavelet as the mother wavelet, with a frequency resolution set to 200. The power spectrum of the signal is obtained by calculating the square of the wavelet coefficients' magnitude. We then apply a logarithmic transformation to the power spectrum, converting it into a logarithmic scale spectrogram, with 300 time steps set. WavLM features are extracted from the pre-trained WavLM model, corresponding to the output of the last hidden layer of the WavLM model.

The optimizer used for model training is AdamW, with a learning rate set to . The batch size for training is 32, and early stopping is configured with a patience of 10 epochs.

* 1. Results and Comparison

Table 1 compares our method with baseline methods in terms of weighted accuracy (WA) and unweighted accuracy (UA) using both 10-fold leave-one-speaker-out and 5-fold leave-one-session-out methods. The methods in references [9-11, 16, 17] adopt the 10-fold leave-one-speaker-out approach, while those in references [9, 18, 19] adopt the 5-fold leave-one-session-out method. SMW\_CAT employs a co-attention mechanism with a Cross-Attention Transformer (CAT) to process dual-source inputs. GFRN-SEA designs a new convolutional module to extract information from log spectrograms and MFCCs and utilizes multi-layer cross-attention to fuse three types of acoustic information. Reference [17] reduces one type of acoustic information, designs a Discriminative Channel Weighting (DCW) module to weight MFCC, and introduces Two-stream Pooling Attention (TsPA) to fuse two features. MFHCA also reduces one type of acoustic information, designs a Multi-Spatial Fusion module (MF) to extract features from the spectrogram, and uses hierarchical attention for feature fusion. IS09-classification explores whether SER can benefit from Transformer-based self-attention and global windowing. GLNN establishes a speech graph based on feature similarity and introduces a novel graph neural network architecture that employs an LSTM aggregator and weighted pooling. Experimental results show that our method achieves the best performance, with WA reaching 76.60% and UA reaching 77.41% in the 10-fold leave-one-speaker-out method. In the 5-fold leave-one-session-out method, WA reaches 74.46% and UA reaches 74.27%. These results demonstrate that leveraging wavelet transform for log-spectrogram extraction, using the Mamba module for MFCC feature extraction, and employing the CEA module for feature fusion significantly enhance SER performance.

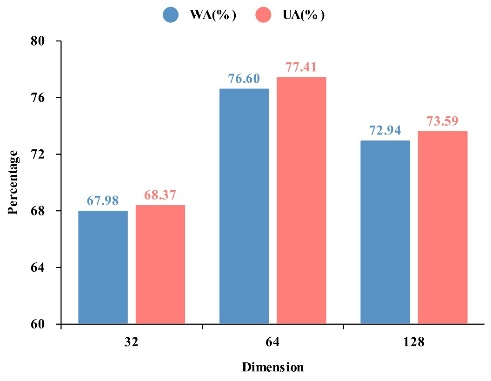
**Table 1.** Compare with the state-of-the-art models evaluated on the IEMOCAP dataset, where 'L' denotes the logarithmic spectrogram, 'M' represents the MFCC, and 'A' signifies the audio waveform.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reference** | **Model** | **Features** | **WA(%)** | **UA(%)** |
| Zou et al. [9] | Co-attention | L + M + A | 71.64 | 72.70 |
| He et al. [10] | SMW\_CAT | L + M + A | 73.80 | 74.25 |
| Liu et al. [17] | DCM + TsPA | M + A | 73.18 | 74.26 |
| Jiao et al. [16] | MFHCA | M + A | 74.24 | 74.57 |
| Pan et al. [11] | GFRN-SEA | L + M + A | 75.29 | 75.57 |
|  | **CEA(Ours)** | L + M + A | 76.60 | 77.41 |
|  | Improvement |  | 1.31 | 1.84 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Zou et al. [9] | Co-attention | L + M + A | 69.80 | 71.05 |
| Tarantino et al. [18] | IS09 - classification | A | 68.10 | 63.80 |
| Li et al. [19] | GLNN | A | 71.83 | 65.39 |
|  | **CEA(Ours)** | L + M + A | 74.46 | 74.27 |
|  | Improvement |  | 2.63 | 8.88 |

* 1. Ablation Study

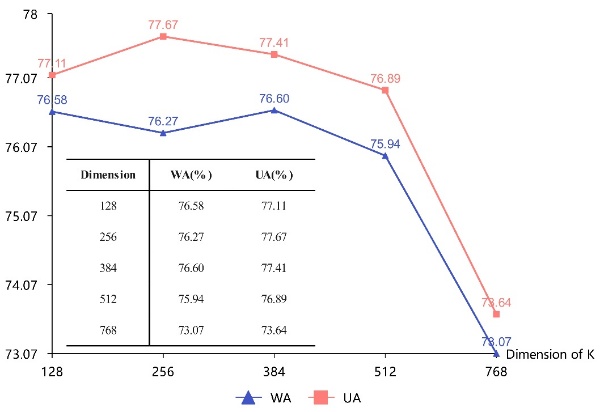
The results of our ablation experiments are shown in Table 2, which fully demonstrate the effectiveness of the proposed method. In the table, EfficientNet represents the log-spectrogram features, Mamba represents the MFCC features, and WavLM represents the WLM features. In the experiments without CEA, we only use the co-attention mechanism to ensure comparability. The first two rows and the last four rows in the table present the results of different feature extraction module combinations. By comparing these three combinations, it can be observed that the Mamba module and WavLM model help improve emotion recognition performance. By comparing the emotion recognition accuracy with and without wavelet transform on the log-spectrogram features, the results indicate that the wavelet transform can slightly improve the emotion recognition accuracy. In row 12 of Table 2, the accuracy slightly decreases when the wavelet transform is applied, possibly because the co-attention mechanism is not particularly suitable for processing features extracted from wavelet-transformed spectrograms. The following two rows present the combinations of WLM with log-spectrogram and MFCC features and compare the results with or without CEA. From the comparison, it can be observed that CEA can effectively utilize the cooperation between other acoustic information and WLM, resulting in approximately a 1% improvement in recognition accuracy for both types of acoustic information. From the last four rows, it can be seen that the CEA module effectively focuses on the emotion-related information in WLM and better integrates the three features, improving emotion recognition accuracy. **Fig. 5** shows the impact of changes in the hidden layer dimension of Mamba on model performance. It can be seen that when the hidden layer dimension of Mamba is set to 64, the model achieves optimal performance. **Fig. 6** illustrates the effect of changes in the dimension  on the performance of CEA. It can be observed that the best experimental results are obtained when  is set to 384.



**Fig. 5.** Ablation experiment on Mamba hidden layer dimension

**Table 2.** Ablation Study on Network Components.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Wavelet** | **CEA** | **WA(%)** | **UA(%)** |
| EfficientNet + LSTM + wav2vec2 |  |  | 70.22 | 71.02 |
|  |  | 69.50 | 71.28 |
| EfficientNet + LSTM + WavLM |  |  | 72.99 | 72.88 |
|  |  | 74.17 | 74.96 |
| EfficientNet + WavLM |  |  | 73.72 | 74.49 |
|  |  | 74.76 | 75.09 |
| Mamba + WavLM |  |  | 73.49 | 74.40 |
|  |  | 75.27 | 76.47 |
| EfficientNet + Mamba + WavLM |  |  | 74.10 | 75.45 |
|  |  | 73.69 | 74.28 |
|  |  | 74.84 | 75.74 |
|  |  | **76.60** | **77.41** |



**Fig. 6.** The changing trend of WA and UA with the dimension change of K

**Fig. 7** shows the feature distribution with and without CEA, visualized using t-SNE. It can be seen that after applying CEA, points of the same class are more tightly clustered, and the decision boundaries of the features are clearer.

|  |  |
| --- | --- |
| 1. Final features w/o CEA | 1. Final features w/ CEA |

**Fig. 7.** The feature distribution visualization using t-SNE. Figure (a) shows the final features without CEA, and Figure (b) shows the final features with CEA.

1. Conclusion

This paper presents a novel SER method that utilizes CEA for the collaboration and enhancement of multiple acoustic features. In this method, we apply wavelet transform to convert the raw audio into the frequency domain and use multiple feature extraction modules, such as Mamba, to extract features from different acoustic information. Furthermore, CEA facilitates the collaboration between the spectrogram, MFCC, and WLM, enhancing the emotional information in the WavLM features. Experimental results on the IEMOCAP dataset demonstrate the effectiveness of our approach.

Acknowledgments.This research was financially supported by the Jiangsu Provincial Key Research and Development Program (BE2022135, BE2022154, BE2023030).

References

1. Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H., Alhussain, T.: Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access. 7, 117327–117345 (2019).
2. Zhang, C., Zhang, Y., Cheng, B.: RL-EMO: A Reinforcement Learning Framework for Multimodal Emotion Recognition. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10246–10250 (2024).
3. Georgescu, A.-L., Chivu, G.-I., Cucu, H.: Exploring Fusion Techniques for Multimodal Emotion Recognition. In: 2024 15th International Conference on Communications (COMM). pp. 1–6 (2024).
4. Zhao, Z., Wang, Y., Wang, Y.: Knowledge-Aware Bayesian Co-Attention for Multimodal Emotion Recognition. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023).
5. Chen, W., Xing, X., Xu, X., Pang, J., Du, L.: DST: Deformable Speech Transformer for Emotion Recognition. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023).
6. Zhang, C., Xue, L.: Autoencoder With Emotion Embedding for Speech Emotion Recognition. IEEE Access. 9, 51231–51241 (2021).
7. Ong, K.L., Lee, C.P., Lim, H.S., Lim, K.M., Alqahtani, A.: Mel-MViTv2: Enhanced Speech Emotion Recognition With Mel Spectrogram and Improved Multiscale Vision Transformers. IEEE Access. 11, 108571–108579 (2023).
8. Ye, J., Wen, X.-C., Wei, Y., Xu, Y., Liu, K., Shan, H.: Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023).
9. Zou, H., Si, Y., Chen, C., Rajan, D., Chng, E.S.: 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). pp. 7367–7371 (2022).
10. He, Y., Minematsu, N., Saito, D.: Multiple Acoustic Features Speech Emotion Recognition Using Cross-Attention Transformer. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023).
11. Pan, L., Wang, Q.: GFRN-SEA: Global-Aware Feature Representation Network for Speech Emotion Analysis. IEEE Access. 12, 161500–161518 (2024).
12. Gu, A., Dao, T.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces,
13. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F.: WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE J. Sel. Top. Signal Process. 16, 1505–1518 (2022).
14. Tan, M., Le, Q.V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. CoRR. abs/1905.11946, (2019).
15. Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. 42, 335–359 (2008).
16. Jiao, X., Wang, L., Yu, Y.: MFHCA: Enhancing Speech Emotion Recognition Via Multi-Spatial Fusion and Hierarchical Cooperative Attention. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–5 (2024).
17. Liu, K., Wang, D., Wu, D., Feng, J.: Speech Emotion Recognition Via Two-Stream Pooling Attention With Discriminative Channel Weighting. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023).
18. Tarantino, L., Garner, P.N., Lazaridis, A.: Self-Attention for Speech Emotion Recognition. In: Interspeech 2019. pp. 2578–2582 (2019).
19. Li, Y., Wang, Y., Yang, X., Im, S.-K.: Speech emotion recognition based on Graph-LSTM neural network. EURASIP J. Audio Speech Music Process. 2023, 40 (2023).