MFHCA: Enhancing Speech Emotion Recognition Via Multi-Spatial Fusion and Hierarchical Cooperative Attention

Xinxin Jiao
School of Computer Science and Technology
Xinjiang University
Urumqi, China
jxx2022@xju.edu.cn

Liejun Wang[1†]
School of Computer Science and Technology
Xinjiang University
Urumqi, China
wljxju@xju.edu.cn

Yinfeng Yu
School of Computer Science and Technology
Xinjiang University
Urumqi, China
yuyinfeng@xju.edu.cn

arXiv:2404.13509v1 [cs.SD] 21 Apr 2024

***Abstract*—Speech emotion recognition is crucial in human-computer interaction, but extracting and using emotional cues from audio poses challenges. This paper introduces MFHCA, a novel method for Speech Emotion Recognition using Multi-Spatial Fusion and Hierarchical Cooperative Attention on spectrograms and raw audio. We employ the Multi-Spatial Fusion module (MF) to efficiently identify emotion-related spectrogram regions and integrate Hubert features for higher-level acoustic information. Our approach also includes a Hierarchical Cooperative Attention module (HCA) to merge features from various auditory levels. We evaluate our method on the IEMOCAP dataset and achieve 2.6% and 1.87% improvements on the weighted accuracy and unweighted accuracy, respectively. Extensive experiments demonstrate the effectiveness of the proposed method.**

***Index Terms*—Speech emotion recognition, multi-spatial fusion, hierarchical cooperative attention, Hubert features**

I. Introduction

Speech emotion recognition (SER) technology plays a crucial role in intelligent human-machine interaction systems [1], as it can identify the speaker's emotional state, thereby enhancing the naturalness of human-machine interaction. Recently, some multimodal SER methods [2], [3], [22] have achieved significantly higher accuracy than speech emotion recognition. However, the model can only extract information from the speech signal in specific scenarios, such as voice assistants and phone customer service. Therefore, we focus on extracting emotional information from speech signals.

Many SER models can be seen as a combination of deep feature extractors and classifiers in their structure. In order to attain powerful representational capabilities, numerous researchers have enhanced network models. Xu et al. [4] proposed an Attention-based CNN (ACNN) to acquire more effective features, aiming to enhance the performance of SER. Liu et al. [5] integrated multi-scale CNN with time-frequency CNN, concurrently modeling local and global information. Chen et al. [6] employed a dynamic window transformer to locate significant regions at different time scales, capturing essential information. Chen et al. [7] devised a Deformable Speech Transformer (DST) that captures multi-granular emotional information through deformable attention windows. Although these methods have achieved higher performance, the intricate structures pose challenges for computational resources.

The outstanding performance of speech self-supervised learning in downstream tasks such as automatic speech recognition (ASR) has opened up new avenues for developing SER. Xia et al. [10] fine-tuned the Wav2Vec model [11], leveraging learned features for SER tasks. He et al. [8] designed a SER method based on the cross-attention transformer that integrates three acoustic features, one of which is extracted from the Wav2Vec2 model [9]. Gat I et al. [12] proposed a speaker feature normalization framework based on self-supervised feature representation, achieving outstanding performance in SER.

In this paper, we propose a simple SER method (MFHCA) based on Multi-Spatial Fusion module (MF) and Hierarchical Cooperative Attention module (HCA), the MF is primarily composed of several Global Receptive Field block (GRF). Unlike [8], [13], which uses many acoustic features for SER, we only use features from the Hubert model [14] and log Mel spectrogram. The Hubert model is trained by performing K-means clustering on MFCC or Hubert features, serving as the training objective. This approach enables the learning of rich feature representations. The log Mel spectrogram better aligns with the auditory characteristics of the human ear. In MFHCA, we employ the MF to extract features in different scale spaces along the temporal and frequency directions, enhancing the model's focus on emotion-related features. Subsequently, the two sets of features are cascaded using the HCA. Finally, a classifier consisting of three fully connected layers is employed to classify the features, completing the emotion recognition task.

Our primary contributions can be summarized as follows:

- We propose a new SER network that combines joint self-supervised features and the log Mel spectrogram. We

† Corresponding author.

alse design a Hierarchical Cooperative Attention module (HCA) to integrate the two sets of features interactively.
- We propose a novel spectrum-based lightweight feature extraction module, denoted as Multi-Spatial Fusion module (MF), which captures dependencies and positional information in different scale spaces, aiding the network in locating emotional information.

II. Proposed method

This section provides an overview of the proposed SER method. In the following subsections, we provide a detailed description of MF, GRF, and HCA.

A. Overall Architecture

The overall structure of the end-to-end emotion recognition method we proposed is illustrated in Fig.1. The method employs two parallel encoders to extract features from the log Mel spectrogram and raw audio. Specifically, MF uses two parallel convolutional layers to capture low-level features from the log Mel spectrogram in both temporal and frequency directions. GRF extracts dependencies and positional information from the features, enhancing the model's ability to learn emotion-related features. In order to save computational time, we utilize Hubert as a feature extractor to obtain feature sequences from the audio directly without fine-tuning Hubert [12]. HCA hierarchically integrates the two sets of features, and a classifier composed of three fully connected layers utilizes the concatenated features for emotion recognition.

B. Multi-Spatial Fusion Module

Fig. 1 illustrates the structure of MF, which is primarily composed of a parallel convolutional layer, a pooling layer, and three Global Receptive Field blocks (GRF). We use processed log Mel spectrogram as input. For the first layer of the module, we employ two parallel convolutions with kernel sizes of (10, 2) and (2, 8), respectively, to extract features in the time and frequency directions. Not all information in the input features is directly relevant to the SER task. Accurately localizing emotional information enables a more comprehensive representation of emotions. To address this problem, we propose a simple block called GRF, and its structure is illustrated in Fig. 2.

GRF consists of three parts. Specifically, the first part represents the input features, denoted by X. Global pooling can capture spatial information, but compressing and preserving global spatial information through squeezing may lead to losing positional information. In the second part, we adopt encoding in both the temporal and frequency directions as a replacement for average pooling. We employ global average pooling(GAP) with kernels $(H, 1)$ and $(1, W)$ to obtain a set of encodings with dependency and positional information. The output for the c-th channel at height h is represented as $z_c(h)$, and the output for the c-th channel at width w is denoted as $z_c(w)$.

$$z_c(h) = \frac{1}{W} \sum_{0<i\leq W} x_c(h, i), \tag{1}$$

$$z_c(w) = \frac{1}{H} \sum_{0<j\leq H} x_c(j, w). \tag{2}$$

We utilize 1×1 convolutions to interactively pass positional information between $z_c(h)$ and $z_c(w)$ to guide the model in locating regions relevant to emotional information. Additionally, this allows for learning relationships between channels.

$$f = \delta(BN(F([z(h), z(w)^T]))), \tag{3}$$

where, $[z(h), z(w)^T]$ represents a cascading operation in the spatial dimension, F denotes a 1×1 convolution operation, BN stands for Batch Normalization, δ means the non-linear activation function $swish$. Subsequently, we partition feature f in the spatial dimension into two independent tensors, f_h and f_w.

$$g_h = \sigma(F_h(f_h)), \tag{4}$$

$$g_w = \sigma(F_w(f_w)), \tag{5}$$

here, F_h and F_w represent 1×1 convolutions, and their role is to adjust the number of channels in f_h and f_w to be consistent with the input X. σ denotes the sigmoid function. Finally, the outputs of these two parts are used as attention weights applied to the input X. Y_a represents the final output.

$$Y_a(i, j) = x(i, j) \odot g_h(i) \odot g_w(j), \tag{6}$$

where, $\odot$ represents hadamard product.

In the second part, We employ average pooling to perform downsampling on the input with a rate of r. Assuming the input is a 3D tensor with a shape of (C, H, W), the shape becomes $(C, H/r, W/r)$ after pooling. Equation (7) demonstrates the computation process.

$$Y_b = X \oplus Up(F(AvgPool_r(X))), \tag{7}$$

where $AvgPool_r$ represents average pooling, F is a convolutional transformation, and Up is a bilinear interpolation operator used to restore features from a smaller-scale space to the original feature space. $\oplus$ is addition. Y_b is the output. The features are transformed into low-dimensional embeddings through downsampling. After transformation by convolutional layers, the low-dimensional embeddings encompass a larger receptive field, enabling the calibration of the convolutional layers in Y_a. This communication between the convolutional layers in Y_a and Y_b expands the receptive field in spatial positions. Each spatial position can also gather context from its surroundings, and dependencies between channels can be learned. Benefiting from this internal information exchange, GRF can generate more discriminative representations. The output of GRF is denoted as Y.

$$Y = X \oplus Y_a \oplus Y_b, \tag{8}$$

where $\oplus$ represents addition.

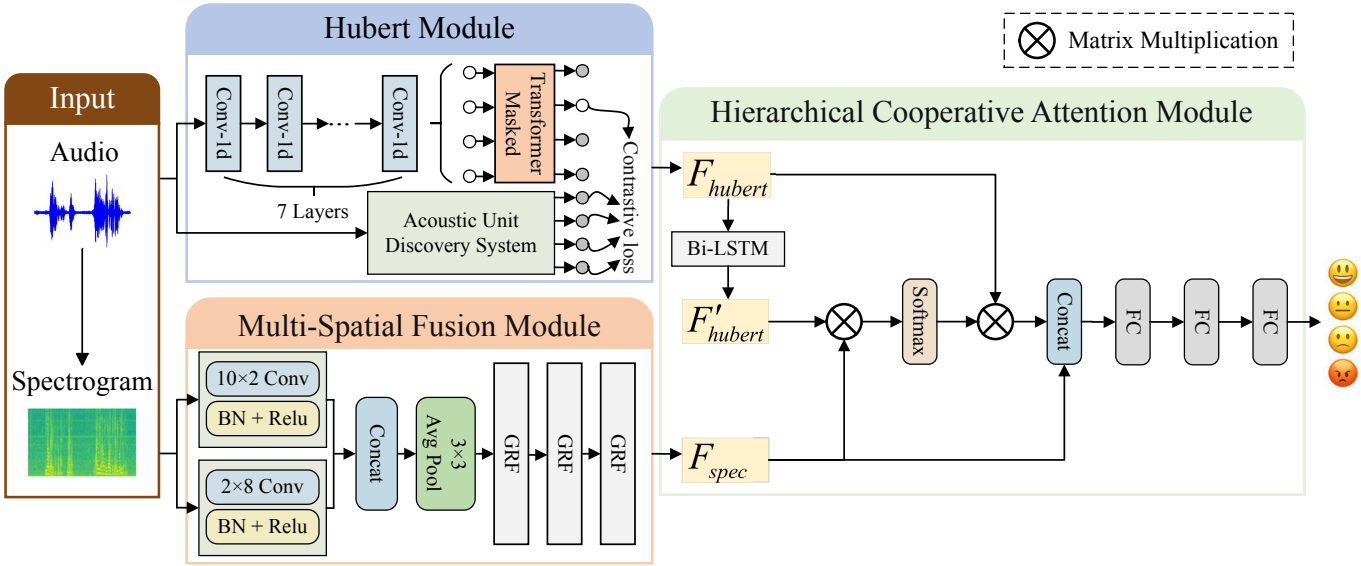


Fig. 1. The overall architecture of our proposed method.

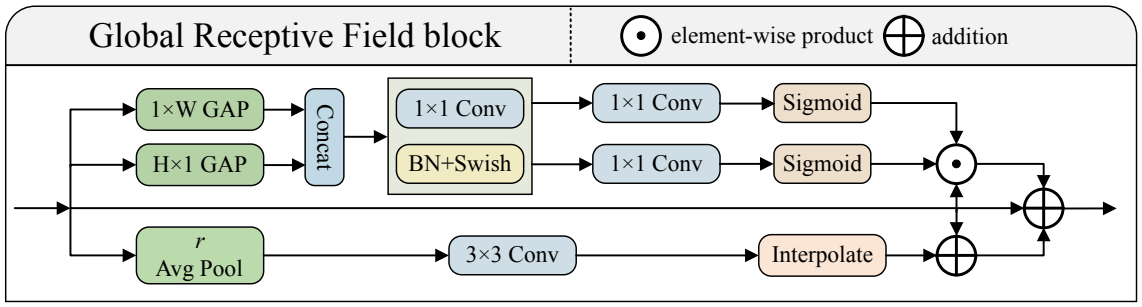


Fig. 2. Global Receptive Field block

C. Hubert Module

Hubert is a transformer-based model composed of three components: a waveform encoder, an Acoustic Unit Discovery System, and a Bert encoder. The Acoustic Unit Discovery System is an unsupervised clustering model trained to label input speech and generate discrete target sequences. The waveform encoder consists of seven layers of convolutional networks, extracting low-level features. The Bert encoder comprises multiple identical Transformer blocks, enabling it to learn contextualized representations.

Hubert conducts k-means and GMM clustering on MFCC features to discretize speech signals, creating training targets. This discretized target sequence derived from MFCC allows the model to learn more representations, such as vocabulary information related to speech recognition and features associated with speakers and emotions. This diversity of information gives it a distinct advantage in emotion recognition. The pre-trained HuBERT model excels at capturing richer, multi-layered speech information, leading to enhanced generalization performance in speech emotion recognition.

D. Hierarchical Cooperative Attention Module

Research has shown that the hierarchical fusion of different features is helpful in SER tasks [19]. We employ the Hubert model as a feature extractor without fine-tuning, and the learned features contain rich information. In addition to emotion recognition-related information, information is also relevant to other downstream tasks. Considering these aspects, we designed the HCA module. Spectrogram features are frame-level, and Hubert features contain contextual information. MF can pinpoint the location of emotional information. We use features from MF as guidance for Hubert features, assisting the network in focusing on emotional information within Hubert features and jointly hierarchically obtaining more powerful feature representations. We use f_{hubert} to denote the output of the hidden layer from Hubert, f'_{hubert} to represent the output of Hubert features after passing through a BiLSTM, and f_{spec} to denote the spectrogram. We perform the following computations:

$$f_{att} = softmax(f_{spec} \otimes f'_{hubert}), \tag{9}$$

$$f'_{att} = f_{hubert} \otimes f_{att}, \tag{10}$$

$$f''_{att} = concat(f_{spec}, f'_{att}), \tag{11}$$

where $\otimes$ represents Matrix multiplication, we append a BiLSTM to the end of the Hubert model to obtain a temporal sequence f'_{hubert} containing contextual information. A co-attention operation is performed between f'_{hubert} and f_{spec} to obtain weights f_{att}. Then, a second-level operation is conducted between f_{att} and f_{hubert}, resulting in f'_{att}. At this stage, the emotional features in f'_{att} are enhanced. Finally, f'_{att} is concatenated with f_{spec} to obtain the final feature representation. The emotion classification is achieved through a classifier consisting of three fully connected layers.

III. EXPERIMENTS AND DISCUSSION

A. Dataset and Implementation Details

We evaluated our approach on the Interactive Emotional dyadic Capture database (IEMOCAP) [15], which consists of 10 actors engaged in 5 dyadic sessions, each featuring a unique pair of male and female actors. Following the work of others [16]–[18], we merged $happy$ and $excited$ into a single category labeled as $happy$, while also considering $neutral$, sad, $happy$, and $angry$ emotions.

In the preprocessing stage, to facilitate comparison with baseline methods [13], we also segment the original audio signal into 3-second-long segments. We apply zero padding when the audio segment is less than 3 seconds. Spectrograms are extracted using a Hamming window with a window length of 40ms and a window shift of 10ms. Each windowed block is treated as a frame, and a Discrete Fourier Transform (DFT) of length 800 is applied to transform each frame into the frequency domain. The first 200 DFT points are used as input spectrogram features. Hubert features correspond to the output of the final hidden layer of the Hubert model.

The model optimizer is Adam, the learning rate is set with 1×10^{-5}, the training batch size is 32, and early stopping is configured for 10 epochs, the optimization function is Cross Entropy Loss. We employed a 10-fold leave-one-speaker-out cross-validation strategy to assess the model's performance, with evaluation metrics being Unweighted Accuracy (**UA**) and Weighted Accuracy (**WA**).

B. Results and Comparison

Table 1 compares our proposed method and the baselines on WA and UA. We compare our method with these approaches under two speaker-independent validation strategies. The experimental results demonstrate that our approach achieves the best performance. HNSD [18] extracts static and dynamic features for emotion recognition from the log-Mel filter bank feature. The model can achieve better performance when utilizing features from the pre-trained model [20]. In order to obtain suitable emotion representations, SER systems based on pre-trained models incorporate multiple speech features using methods such as co-attention-based approaches [13], cross-representation learning [21], and cross-attention transformers [8]. Moreover, our model has 54.26% fewer parameters than the baseline [13] while achieving higher performance. The experimental results demonstrate the effectiveness and lightweight of our method.

TABLE I
COMPARISON WITH KNOWN STATE-OF-THE-ART SYSTEMS ON IEMOCAP, WHERE A AND T DENOTE AUDIO AND TEXT MODALITIES.

Model	Modality	WA(↑)	UA(↑)
HNSD [18]	A	70.50	72.50
Co-attention [13]	A	71.64	72.70
SMW_CAT [8]	A	73.80	74.25
E2e ASR [20]	A+T	71.70	72.60
Cross-representation [21]	A+T	73.00	73.50
MFHCA(ours)	A	**74.24**	**74.57**

TABLE II
ABLATION STUDY ON THE PROPOSED MODEL.

Models	MF	HCA	WA(↑)	UA(↑)
Spec	✗	✗	62.13	62.25
	✓	✗	62.21	62.37
Hubert	✗	✗	69.99	70.57
Spec+W2V2	✗	✗	70.05	71.30
	✓	✗	71.73	72.32
	✗	✓	70.72	72.09
	✓	✓	**72.00**	**73.44**
Spec+Hubert	✗	✗	72.13	72.51
	✓	✗	73.72	74.53
	✗	✓	73.19	73.72
	✓	✓	**74.24**	**74.57**

TABLE III
THE IMPACT OF QUANTITY AND CHANNEL VARIATION OF GRF ON PERFORMANCE.

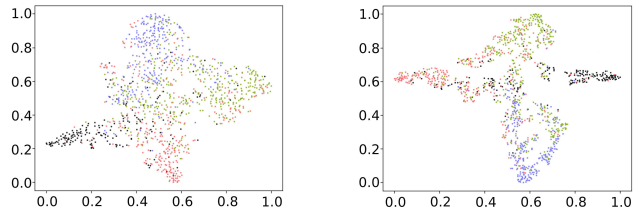
Num	Channel	WA(↑)	UA(↑)
2	16-32	73.76	74.02
3	**16-32-48**	**74.24**	**74.57**
3	16-32-64	73.60	74.49
4	16-32-48-64	73.67	74.03
4	16-32-64-128	72.75	73.78

C. Ablation Study

In Table II, we conducted ablation experiments to demonstrate the effectiveness of MF and HCA. "Hubert" denotes features extracted using Hubert, and "W2V2" represents features extracted using wav2vec2. Specifically, in the experiment involving two features, the absence of HCA implementation does not imply the exclusion of fusion methods. To ensure comparability, we employed baseline fusion methods. We initially conducted experiments using single features and found that MF contributes to the extraction of emotional information in the spectrogram. The pre-trained model used in the baseline is wav2vec2. For a more intuitive comparison, we conducted experiments based on Spec and W2V2. The results indicate that our proposed method performs better in learning emotional information from features compared to the baseline. Subsequently, we further validated our method on Spec and

TABLE IV
THE IMPACT OF SAMPLING RATIO r ON PERFORMANCE.

r	**WA(↑)**	**UA(↑)**
1/2	73.82	74.33
1/4	**74.24**	**74.57**
1/8	72.22	73.34
1/16	71.12	72.69



(a) Final features w/o HCA (b) Final features w/ HCA

Fig. 3. The t-SNE visualization of feature distribution. (a) and (b) are the final combined features without and with the proposed HCA.

Hubert, achieving improved recognition accuracy. In Table III, we employed grid search to investigate the impact of changes in the number of GRFs and channels on the experimental results. In Table IV, we similarly studied the influence of variations in the scaling factor r on the experimental results.

To further validate our approach, we employed t-SNE visualization to depict the feature distributions with and without HCA in Fig. 3. When HCA was used, the dense areas of feature distribution were reduced, and classification boundaries became more distinct.

IV. CONCLUSION

This paper presents MFHCA, a novel SER approach utilizing MF for spectrogram feature extraction. Our method capitalizes on GRF's advantages to overcome CNN's limitations in capturing global information. HCA fosters interaction between spectral diagrams and Hubert features, jointly harvesting emotional representations from both components. Experimental results on IEMOCAP substantiate the effectiveness of MF and HCA in our approach.

ACKNOWLEDGMENT

The following projects jointly supported this work: the Tianshan Excellence Program Project of Xinjiang Uygur Autonomous Region, China (2022TSYCLJ0036); the Central Government Guides Local Science and Technology Development Fund Projects (ZYYD2022C19); the National Natural Science Foundation of China under Grant 62303259.

REFERENCES

[1] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction[J]. IEEE Signal processing magazine, 2001, 18(1): 32-80.

[2] Chen W, Xing X, Xu X, et al. Key-sparse transformer for multimodal speech emotion recognition[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 6897-6901.

[3] Wang S, Ma Y, Ding Y. Exploring complementary features in multi-modal speech emotion recognition[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[4] Xu M, Zhang F, Khan S U. Improve accuracy of speech emotion recognition with attention head fusion[C]//2020 10th annual computing and communication workshop and conference (CCWC). IEEE, 2020: 1058-1064.

[5] Liu J, Liu Z, Wang L, et al. Speech emotion recognition with local-global aware deep representation learning[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7174-7178.

[6] Chen S, Xing X, Zhang W, et al. DWFormer: Dynamic Window Transformer for Speech Emotion Recognition[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[7] Chen W, Xing X, Xu X, et al. DST: Deformable speech transformer for emotion recognition[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[8] He Y, Minematsu N, Saito D. Multiple acoustic features speech emotion recognition using cross-attention transformer[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[9] Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in neural information processing systems, 2020, 33: 12449-12460.

[10] Xia Y, Chen L W, Rudnicky A, et al. Temporal Context in Speech Emotion Recognition[C]//Interspeech. 2021, 2021: 3370-3374.

[11] Schneider S, Baevski A, Collobert R, et al. wav2vec: Unsupervised pre-training for speech recognition[J]. arXiv preprint arXiv:1904.05862, 2019.

[12] Gat I, Aronowitz H, Zhu W, et al. Speaker normalization for self-supervised speech emotion recognition[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 7342-7346.

[13] Zou H, Si Y, Chen C, et al. Speech emotion recognition with co-attention based multi-level acoustic information[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 7367-7371.

[14] Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[15] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language resources and evaluation, 2008, 42: 335-359.

[16] Guo L, Wang L, Xu C, et al. Representation learning with spectro-temporal-channel attention for speech emotion recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6304-6308.

[17] Wu W, Zhang C, Woodland P C. Emotion recognition by fusing time synchronous and time asynchronous representations[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6269-6273.

[18] Cao Q, Hou M, Chen B, et al. Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6334-6338.

[19] Tian L, Moore J, Lai C. Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features[C]//2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016: 565-572.

[20] Lu Z, Cao L, Zhang Y, et al. Speech sentiment analysis via pre-trained features from end-to-end asr models[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7149-7153.

[21] Makiuchi M R, Uto K, Shinoda K. Multimodal emotion recognition with high-level speech and text features[C]//2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021: 350-357.

[22] Yu Y, Jia Z, Shi F, et al. WeaveNet: End-to-End Audiovisual Sentiment Analysis[C]//International Conference on Cognitive Systems and Signal Processing. Singapore: Springer Nature Singapore, 2021: 3-16.