Applied Acoustics 214 (2023) 109658

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

MPAF-CNN: Multiperspective aware and fine-grained fusion strategy for speech emotion recognition

Guoyan Li [a,*], Junjie Hou [a], Yi Liu [a], Jianguo Wei [b]

[a] School of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300384, China
[b] School of Software, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Keywords:
SER
Multi perspective
Feature fusion
Attention mechanism

ABSTRACT

Speech emotion recognition (SER) is a crucial and challenging task in affective computing due to the intricacy and variability inherent in speech. In this paper, a novel method (i.e., MPAF-CNN) combines a convolutional neural network (CNN)-based multiperspective aware module (MPAM) and a frame-level fine-grained fusion strategy (FFS) for SER by utilizing speech information. MPAM perceives the emotional information embedded in speech from three main perspectives: local, frame-level, and global. Specifically, this module introduces the multiscale idea of perceiving multi-granular emotional information under different local sensory fields from the local perspective; a novel frame-level aggregated attention is proposed in this paper, aiming to learn the intrinsic emotional associations of intermediate features from the frame-level perspective, enhance the model's attention to emotionally informative frames, and improve the emotional expression of intermediate features; in the global perspective, multiple layers of global intermediate features are aggregated from the time domain, frequency domain, or channel to enhance the model's ability to extract and express global feature information. A new frame-level fine-grained fusion strategy is proposed to employ an attention mechanism to model the interaction of emotional representations from different acoustic features at the frame level, capturing their underlying relationships and thus further improving the overall performance of the model. The experimental results show that our method has excellent performance in recognizing speech emotions, and MPAF-CNN obtains 72.19% and 72.88% recognition accuracy on the IEMOCAP database.

1. Introduction

The field of affective computing [1,2] has garnered increasing interest from researchers with the continuous breakthroughs in computing software, hardware, and artificial intelligence. An important branch of affective computing is emotion recognition, which analyzes emotions from human biological features such as facial expressions, speech, physiological signals, and body movements [3]. Among all biometric features, speech is a particularly effective source of emotional information. Speech emotion recognition has a vast range of applications in human–computer interaction, such as medical assistance [4,5] and artificial customer service [6]. However, the variability and ambiguity of speech signals and the presence of noise and background interference pose significant obstacles to accurately recognizing emotions from speech [7].

To accurately recognize the emotional state of the speaker, appropriate acoustic features from the speech must be extracted. Generally, there are two types of acoustic features in speech emotion recognition: frame-level features and segment-level features. Frame-level features refer to manually designed features on multiple short segments of a speech segment, such as energy, zero crossing rates, Mel-frequency cepstral coefficients (MFCCs) [8–10], etc. Segment-level features provide a visual representation of the features of speech segments, such as speech spectrograms and log-mel spectrograms [11].

Hand-designed features are commonly used in speech emotion recognition, but obtaining accurate emotion states through direct classification using these features is challenging. To address this issue, researchers have explored various algorithms aiming to extract high-level emotional embedding representations from these features, which can improve classification accuracy [12,13]. Before the emergence of neural networks, traditional algorithms such as the hidden Markov model (HMM) and Gaussian mixture model (GMM) were commonly used to identify emotions. HMM is a probabilistic model that can effectively model the temporal dynamics of emotions. Lin Y L. et al. applied HMM

* Corresponding author.
E-mail addresses: ligy@tcu.edu.cn (G. Li), 1831485231@qq.com (J. Hou), liuyi@tcu.edu.cn (Y. Liu), jianguo@tju.edu.cn (J. Wei).

https://doi.org/10.1016/j.apacoust.2023.109658
Received 29 June 2023; Received in revised form 26 August 2023; Accepted 16 September 2023
Available online 26 September 2023

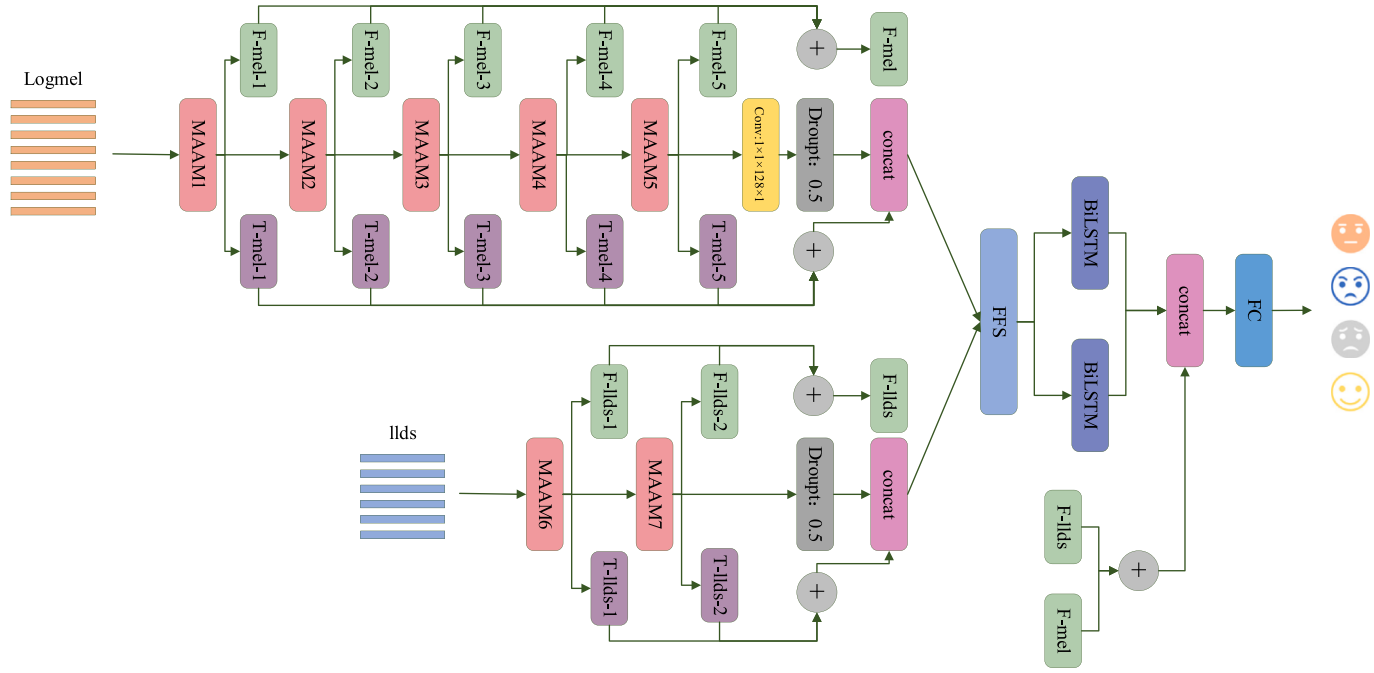

Fig. 1. The overall architecture of the model presented in this paper. "MPAM": multiperspective aware module, "FFS": fine-grained fusion strategy, "BiLSTM": bidirectional long short-term memory network layer, "FC": fully connected layer, "+": add.

to a Danish emotional speech database and achieved promising results. GMM is a specific type of continuous HMM [14]. Mishra H K. et al. proposed a variational Gaussian mixture model that was further improved based on GMM. The results demonstrated that the variational GMM outperformed GMM when using the same amount of data, achieving high accuracy in emotion recognition [15]. While traditional algorithms have made notable strides in improving emotion recognition accuracy, they still face certain limitations. The advent of neural networks has opened new avenues for advancing emotion recognition, with researchers discovering that using neural network algorithms can lead to higher precision rates. Consequently, present-day SER research has prioritized enhancing and integrating diverse deep learning models [16]. For instance, in [17], deep belief networks (DBNs) were utilized to extract high-level representations of emotions from amplitude spectrograms, demonstrating superior performance compared to traditional models.

Researchers have developed numerous deep learning models to provide effective emotional representations. CNNs have gained significant popularity in SER due to their ability to process data with grid-like topologies, such as time series and image data [16,18]. Araño, K.A. acquired emotional representations by integrating MFCCs with image features extracted from spectrograms using pretrained CNNs [19]. These combined features were then fed into a classifier for emotion classification. Basic CNNs generally acquire feature information by learning fixed-scale features [20,21]. However, emotions are manifested through various scales of prosodic variations during speaker speech. To address these challenges, researchers have incorporated multiscale concepts into the SER field. For instance, a multiscale global awareness model was proposed in [22] to extract emotion information from various scales and subsequently utilized a global awareness module to integrate the emotion information from multiple scales. Although the incorporation of multiscale concepts enhances the diversity of affective information, it also results in an increased amount of redundant information, which complicates the differentiation between affective and nonaffective information in speech. Consequently, devising methods to minimize redundant information while augmenting the diversity of affective information remains a critical issue that warrants further investigation.

It is important to note that most previously mentioned methods rarely consider the issue of losing emotional information embedded in intermediate features at various levels as the network model deepens, which causes interference in SER. A parallel network of ResNet-CNN-transformer encoders is proposed in [23] to solve the emotion information loss problem. X. Jiang et al. [24] improved resnet34 so that the network focuses on extracting local critical information and reduces the loss of affective information. Although resnet reduces the problem of the loss of intermediate features at each level, it only focuses on the previous intermediate features of the current layer, and there is still the problem of the loss of affective information in the intermediate features at each level as the network deepens. And the emotional signal in speech with context-sensitive dependences and the commonly employed base modules, such as CNN and bidirectional long short-term memory (BiLSTM), in the SER domain focus on local feature learning, which makes it challenging to capture the global context information of the emotional signal in speech.

We implemented fusion strategies on the segment-level and frame-level acoustic features to significantly enhance the SER classification performance. The fusion strategy plays a pivotal role in multifeature linguistic emotion recognition, and researchers have proposed various approaches to combine distinct types of features effectively. For instance, the study in [25] utilized a fully convolutional network (FCN) to extract spatial features and a BiLSTM to extract temporal features. These features were then concatenated and input into an attention mechanism, ultimately achieving speech emotion recognition through a fully connected layer. A confidence-based fusion strategy was developed in [26], which integrate the power of different classifiers in recognizing different emotional states. Most existing fusion schemes fuse multiple classifier results or directly splice and sum fusion of emotion representations and have yet to consider the potential relationships between different emotion representations. There is variability in the emotion representations extracted from different acoustic features, and the emotion information expressed in each frame is not necessarily the same, thus necessitating fine-grained interaction, i.e., frame-frame modeling.

The above analysis reveals that most models suffer from the problem of considering only a single perspective and not focusing on the potential relationships between different emotion representations. This paper

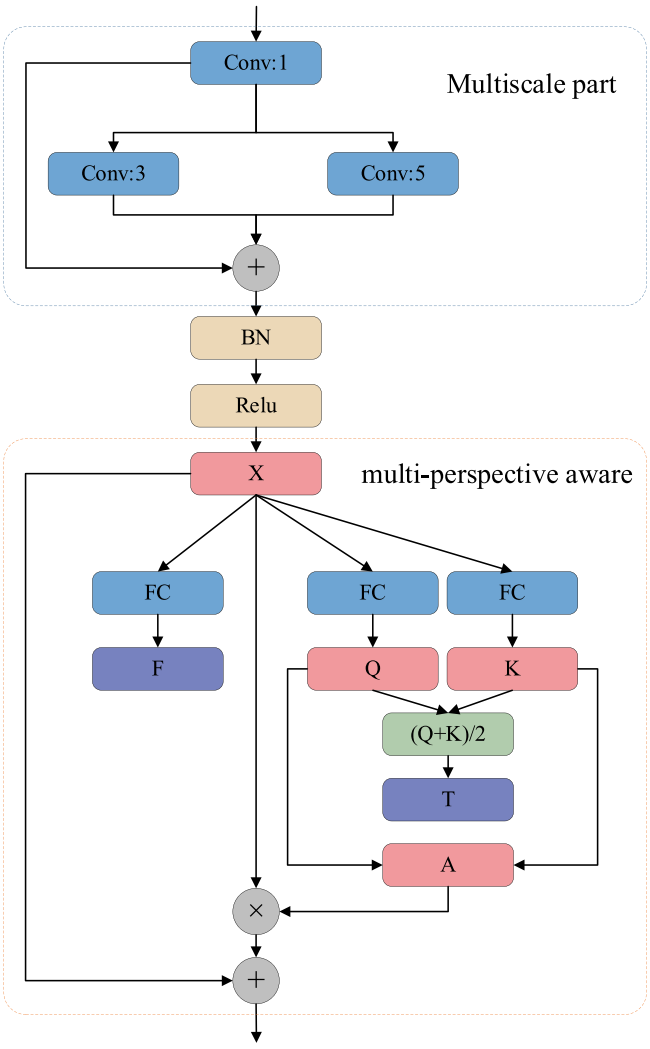

Fig. 2. The overall architecture of MPAM.

proposes a novel MPAF-CNN framework to address these issues, and the framework contains two main modules, namely the multiperspective aware module and the frame-level fine-grained fusion strategy. MPAM aims to improve the model's emotional representation from three perspectives: local, frame-level, and global. Specifically: (1) From the local perspective, by introducing the idea of multiscale, we get multi-granularity emotional information from the different local receptive fields and increase the richness of emotional information. (2) In the frame-level perspective, this paper proposes a novel frame-level aggregated attention mechanism to learn the intrinsic emotion associations in intermediate features, focusing on frames in which intermediate features are relevant to affect and minimizing redundant information unrelated to affect. (3) In the global perspective, the module obtains multi-layer global intermediate feature representations from the time domain, the frequency domain, or the channel, and learns the global representations of the intermediate features while preventing the loss of emotion information embedded in the intermediate features at all levels. Finally, a novel frame-level fine-grained fusion strategy based on an attentional mechanism is proposed in this paper to infer potential relationships between emotion representations from different acoustic features and thus improve the overall performance of the model.

The main contributions of this paper are summarized as follows:

(1) A novel multiperspective aware module is proposed to capture rich emotional information from local, frame-level, and global perspectives.
(2) A novel frame-level fine-grained fusion strategy based on attentional mechanisms is proposed to capture the potential relationships between different emotional representations and explicitly model the interactions between different emotional representations at the frame level.
(3) Based on the experimental results, our approach achieved recognition accuracies of 72.19% and 72.88% on the IEMOCAP database.

The rest of this paper is structured as follows. Section 2 describes the method presented in this article in detail. In Section 3, we set the parameters for the experiment. In Section 4, we analyze the experimental results. Finally, conclusions and future work are presented in Section 5.

2. Proposed method

The proposed scheme is shown in Fig. 1. The model takes both frame-level low-level descriptors (LLDs) and segment-level Mel-spectrograms (MS) of the same utterance as input. The proposed scheme consists of an MPAM based on a 2DCNN for MS feature extraction and an MPAM based on a 1DCNN for LLD feature extraction. After extracting the features from two subnetworks, two feature vectors are fused using a fine-grained fusion strategy and fed to BILSTM to extract temporal features. Finally, the output is fed into the classifier to recognize emotions. The following sections describe each module in detail.

2.1. Input feature maps

The input acoustic embeddings are divided into segment-level features (i.e., MS) and frame-level features (i.e., LLDs) [11,27–29]. Specifically, this paper first calculates the short-time zero crossing rate, log energy, short-time energy, and MFCC which consists of its static features and first- and second-order differentials for each LLD. Next, the MS is calculated to obtain the input segment-level features.

The MS is a spectrogram using the Mel scale, a visual representation of the sound intensity or energy over time for all frequencies present in an audio signal. MS is good at analyzing a particular audio signal's frequency components and intensities and simulates the human ear's perception of speech at different frequencies. Furthermore, MFCC is a compressed representation of mel filter banks, containing information about the rate variation in different spectral bands and concisely describes the shape of the spectral envelope. Furthermore, short-term energy is derived by applying a linear filter to the squared speech signal, while log energy is obtained by taking the logarithm of the short-term energy.

Short-term energy and log energy are temporal domain features, whereas MFCC parameters are perceptual features based on human auditory perception and belong to the frequency domain. In this study, the calculation of MFCC does not incorporate short-term energy; instead, only the MFCC coefficients are directly utilized. The limited correlation between these two features can be attributed to their representation of distinct characteristics within the audio signal. By combining these features, a more comprehensive and informative set of emotion-related information can be extracted [30,31].

2.2. MPAM

A novel multiperspective awareness module for the extraction of emotional information is proposed in this paper. The module comprises three key components: a multiscale layer, an attention layer, and a global representation layer. The architecture of the module is shown in Fig. 2.

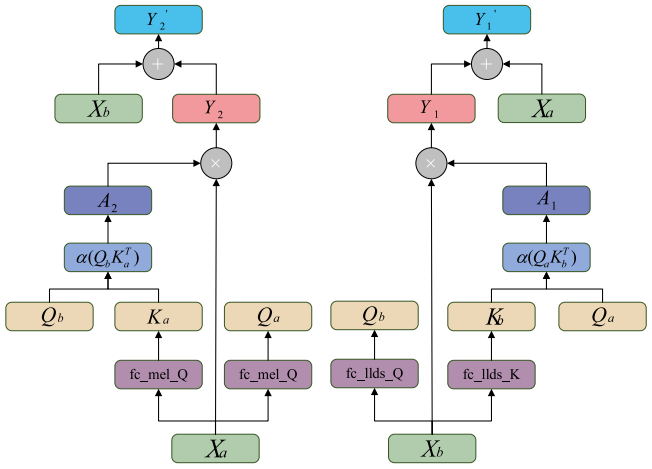

Fig. 3. The overall architecture of the fine-grained fusion strategy.

Table 1
The number of statements per emotion category of IEMOCAP.

Emotion	Angry	Happy	Sad	Neutral	Total
Sample number	1103	1636	1084	1708	5531

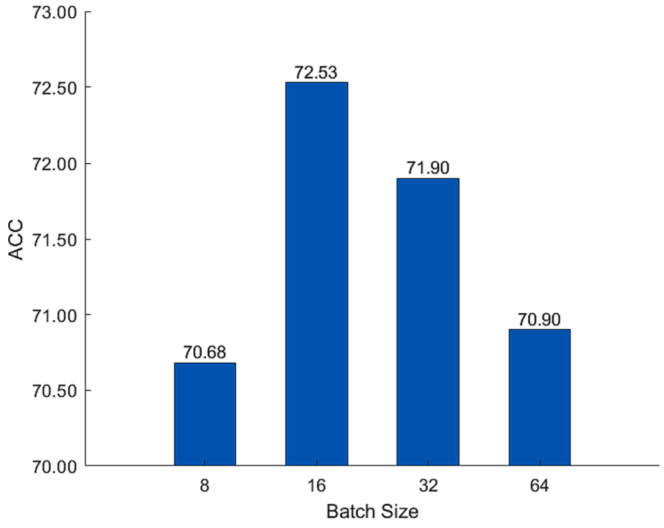

Fig. 4. The results of the model proposed in this paper are obtained at different batch sizes.

Table 2
Ablation study for the individual module on the IEMOCAP dataset. Note: Bold font is the model with the best results.

	Module	WA(%)	UA(%)	ACC(%)
M1	**Ours**	**72.19**	**72.88**	**72.53**
M2	a subnetwork with MS as input	66.03	67.55	66.79
M3	a subnetwork with LLDs as input	66.66	68.23	67.44
M4	Ours without MPAM	69.80	70.47	70.13
M5	Ours without muti-scale layer	70.56	72.05	71.30
M6	Ours without attention layer	70.74	72.54	71.64
M7	Ours without TDGR	68.29	69.34	68.81
M8	Ours without FDGR and CDGR	71.37	72.08	71.72
M9	Ours without fine-grained fusion strategy	70.92	71.41	71.16

2.2.1. Multiscale layer

This paper proposes a multi-scale layer. And residual mapping and multi-scale convolution are used. Multi-scale convolution aims to extract multi-granularity emotion information under different feeling fields by introducing the multi-scale idea. Residual mapping is introduced to avoid the gradient vanishing problem. The convolutional kernel size is $n,n \in \{3,5\}$. The rectified linear unit (RELU) activation function was selected for this paper, as it effectively addresses the issue of gradient vanishing. The batch normalization (BN) layer normalizes the activations of the convolutional layer at each batch and improves the performance and stability of deep networks.

2.2.2. Attention layer

The attention mechanism is utilized in the attention layer to focus on salient frames related to the emotional output of the multiscale layer. The goal is to learn the emotional associations of intermediate features and reduce redundant information. Specifically, attention is calculated as follows:

$$X = AM \tag{1}$$

where $M \in R^{d\times t}$ represents the feature after passing through the multi-scale layer, A denotes the intrinsic emotional relevance, and $X \in R^{d\times t}$ represents the feature after the attention layer. Each parameter will be explained in detail below.

We employ two distinct feature mapping functions to obtain Q and K from x. To learn the global features of each time unit and reduce the number of required training parameters, we compress the features of each time unit to a single vector of dimension one using a learnable mapping vector. The formula for this process is as follows:

$$Q = MW_Q \tag{2}$$

$$K = MW_K \tag{3}$$

where $W_Q \in R^{d\times 1}$ and $W_K \in R^{d\times 1}$ is the learnable mapping vector.

In this study, we utilize the product of Q and K as an intrinsic measure of emotional relevance, and since the input dimensions for Q and K correspond to 1, the result of using the scale parameter remains 1. Hence it is not explicitly written in the formula. expressed by the following formula:

$$A = softmax(QK) \tag{4}$$

where *softmax* is the aggregation function.

2.2.3. Global representation layer

The global representation layer comprises three components: time domain global representation (TDGR), frequency domain global representation (FDGR), and channel domain global representation (CDGR). TDGR is calculated using Q and K in b to obtain a global representation of each temporal unit. FDGR and CDGR represent the long-term time-varying features of the speech emotional signal. Specifically, FDGR is extracted from MS, while CDGR is extracted from LLDs. The implementation formulas for each component are as follows:

$$\mathrm{F}_{TDGR} = (Q+K)/2 \tag{5}$$

$$\mathrm{F}_{FDGR} = XW_{\mathrm{F}} \tag{6}$$

$$\mathrm{F}_{CDGR} = XW_c \tag{7}$$

$\mathrm{F}_{TDGR} \in R^{t\times 1}$ is the time domain global representation, $\mathrm{F}_{FDGR} R^{d_1\times 1}$ is the frequency domain global representation, $\mathrm{F}_{CDGR} \in R^{d_2\times 1}$ is the channel domain global representation, and $W_{\mathrm{F}} \in R^{t_1\times 1}, W_C \in R^{t_2\times 1}$ is the trainable feature mapping vector.

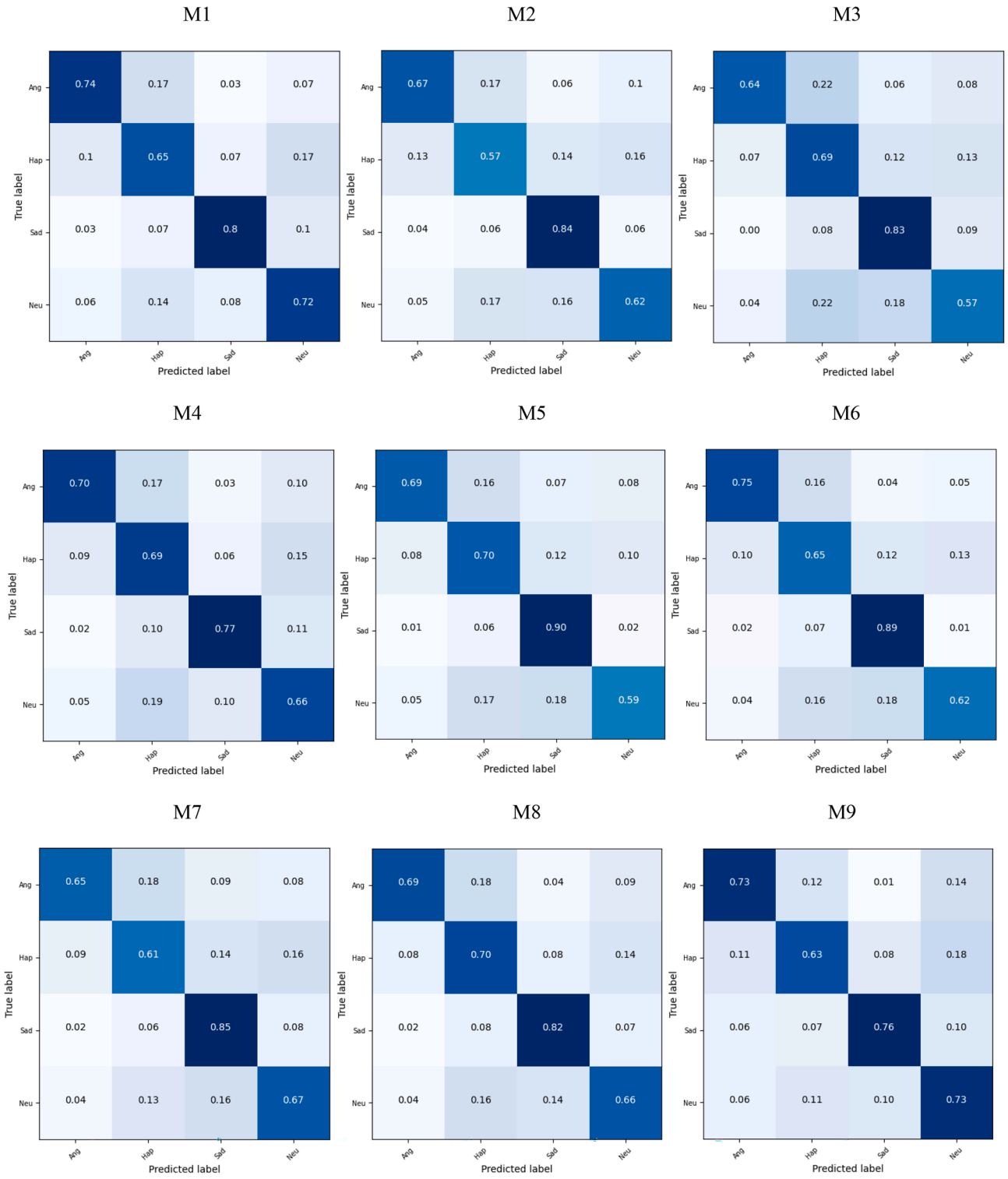

Fig. 5. Confusion matrix of classification results for six models. The diagonal line represents the number of correct emotion predictions for each category.

2.3. *Fine-grained fusion strategy*

This paper proposes a frame-level interactive attention fusion strategy to effectively fuse different types of features and capture the complementary information between different representations. The specific implementation steps are illustrated in Fig. 3.

Let $X_{MS} = \{x_{MS}^1, x_{MS}^2, \ldots, x_{MS}^{T_2}\} \in R^{T_1 \times d_{X_{MS}}}$ represent the features from MS, where T_1 refers to the temporal dimension and $d_{X_{MS}}$ represents the dimension of each temporal unit. Similarly, let $X_{llds} = \{x_{llds}^1, x_{llds}^2, \ldots, x_{llds}^{T_2}\} \in R^{T_2 \times d_{X_{llds}}}$ denote the features from LLDs.

We employ the summation fusion strategy to combine F_{TDGR}. By doing so, we obtain F_{TDGR}^{MS} and F_{TDGR}^{llds}. We then fuse F_{TDGR}^{all} with X_{MS} and X_{llds} to obtain the new X_a and X_b. The formula for this process is shown below:

$$X_a = concat\left(\left[X_{MS}, F_{TDGR}^{MS}\right], \dim = T_1\right) \tag{8}$$

Table 3
Classification performance of different advanced methods for emotion recognition using speech on IEMOCAP datasets. Note: Bold font is the model with the best results.

Model	WA(%)	UA(%)
(Li et al.) [18]	56.14	57.84
(Yao et al.) [26]	57.10	58.30
(Chen et al.) [36]	69.22	70.51
(Chen et al.) [37]	68.73	70.56
(Liu et al.) [38]	70.27	66.27
Ours	**72.19**	**72.88**

$$X_b = concat\left(\left[X_{llds}, \mathrm{F}_{TDGR}^{llds}\right], \dim = T_2\right) \tag{9}$$

where $X_a \in R^{T_1 \times d_a}(d_a = d_{X_{MS}} + 1)$,$X_b \in R^{T_2 \times d_b}\left(d_b = d_{X_{llds}} + 1\right)$.

To establish the correlation between the different types of features, a linear rectification layer is employed to transform the feature sequences into two new sequences, one for query and the other for keys. This transformation is achieved through the following equation:

$$Q_a, Q_b = X_a W_a^Q, X_b W_b^Q \tag{10}$$

$$K_a, K_b = X_a W_a^K, X_b W_b^K \tag{11}$$

where $Q_n, K_n \in R^{T \times 1}(n \in \{a, b\}, T \in \{T_1, T_2\})$ represent the query and key, respectively.$W_m^Q, W_m^K \in \mathrm{R}^{d \times 1}$ $(m \in \{a, b\}, d \in \{d_a, d_b\})$ is the learnable projection vector.

In Eq. (10) and Eq. (11), a learnable mapping vector compresses the features of each time unit into one dimension, representing all features of each time unit. This process reduces the number of training parameters while maintaining the integrity of the features. To obtain a new matrix, perform a matrix multiplication operation between Q_b of X_b and the transpose of K_a of X_a. Each row element in the matrix represents the similarity between a single frame of X_a and all frames of X_b. Afterward, a set of weights A_1 is obtained by performing a normalized aggregation using the softmax activation function. Similarly, A_2 is obtained using the same operation. It is important to note that the dimensions used to normalize and aggregate A_1 and A_2 differ. Specifically, A_1 is normalized in the T_1 dimension, which allows us to determine the degree of influence that X_a has on each time unit in X_b. Similarly, A_2 is normalized in the T_2 dimension, which enables us to determine the degree of influence that X_b has on each time unit in X_a.

$$A_1 = \alpha\left(Q_b K_a^T\right)\left(A_1 \in R^{T_1 \times T_2}\right) \tag{12}$$

$$A_2 = \alpha\left(Q_a K_b^T\right)\left(A_2 \in R^{T_2 \times T_1}\right) \tag{13}$$

where α represent softmax activation function.

The weights obtained for X_a and X_b are multiplied in matrix form to produce results reflecting fine-grained interactions between the two at the frame level. The final results are obtained through a combination of summation and fusion methods. The formula for this process is presented below.

$$Y_1 = W_{Y_1}(A_1 X_b) + X_a\left(Y_1 \in R^{T_1 \times d_a}\right) \tag{14}$$

$$Y_2 = W_{Y_2}(A_2 X_a) + X_b\left(Y_2 \in R^{T_2 \times d_b}\right) \tag{15}$$

where $W_{Y_1} \in R^{d_b \times d_a}$ and $W_{Y_2} \in R^{d_a \times d_b}$ are learnable mapping matrices to the values mapped to the same dimensions as X_a and X_b, respectively.

The proposed strategy emphasizes the fine-grained fusion of various features, thereby preventing the loss of complementary information across different feature types. This approach enhances the network's ability to fuse features and improves the model's robustness to noise.

2.4. Classifiers

The last step of classification for speech emotion recognition is to obtain the classification score of emotion by learning the obtained high-dimensional features through a fully connected layer. As shown in Fig. 1, we utilize summation fusion to fuse the F_{FDGR} from MS and F_{CDGR} from LLDs of each layer to obtain F_{FDGR}^{MS} and F_{CDGR}^{llds}, sum F_{FDGR}^{MS} and F_{CDGR}^{llds}, connect them with the emotional representation after BILSTM, and send them to the classifier to identify the current emotional state together. The summation operation aims to reduce the number of training parameters. We choose the cross-entropy loss function during training.

3. Experiments

3.1. Dataset

The Interactive Emotion Dyadic Motion Capture (IEMOCAP) dataset is used for the experiments in this paper [32]. The experiment in this paper uses the IEMOCAP dataset, the most widely used dataset in the SER field. This dataset contains 12 hours of emotional speech recorded by five men and five women in pairs from the University of Southern California Department of Drama. The dataset is divided into two parts, improvisation and script, and the recorded utterances are labeled with a total of nine emotions: anger, happiness, excitement, sadness, neutral, frustration, fear, surprise, and other. Because the amount of data for each emotion is different, to avoid the impact of unbalanced data distribution, researchers often use five emotions: neutral, happiness, excitement, sadness, and anger. Among them, happiness and excitement have certain similarities, and researchers often combine the two emotions to increase the amount of data [33–35]. In this paper, 5531 utterances (1636 happy, 1103 angry, 1084 sad, and 1708 neutral) from the IEMOCAP dataset are used, where the dataset details are shown in Table 1.

3.2. Evaluation metrics

This paper uses weighted accuracy (WA) and unweighted accuracy (UA) for evaluation, two evaluation criteria widely used in the SER field-related literature. WA and UA do not necessarily reach the maximum value simultaneously in the same model; therefore, the average of WA and UA represented using ACC is calculated as the final evaluation metric. WA, UA, and ACC detailed calculations:

$$\mathrm{WA} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}} N_{\mathrm{c}}^{(\mathrm{i})}}{N} \tag{16}$$

$$\mathrm{UA} = \frac{1}{k}\sum_{i=1}^{k}\frac{N_c^{(i)}}{N_o^{(i)}} \tag{17}$$

$$ACC = \frac{WA + UA}{2} \tag{18}$$

where $N_c^{(i)}$ denotes the number of samples correctly identified by class i, $N_o^{(i)}$ denotes the total number of samples of class i, and k represents the number of sample categories to be recognized.

3.3. Experimental setup

There is no uniform way to divide the dataset in the SER field. Therefore, this paper randomly split the dataset into a training set (80% of the data) and a test set (20% of the data). Each utterance was divided into 2-second segments with 1.6 s of overlap between segments. Since speech is utterance-level data and needs to be tested based on utterance, this paper takes the average of the prediction results of all speech segments in the same utterance as the final prediction result of this utterance [36]. This paper uses cross-entropy as the final objective function,

and the Adam algorithm with a learning rate of 0.0001 is used to optimize the model. In this paper, the model is trained with 50 epochs.

4. Results and discussion

4.1. Ablation study

4.1.1. Impact of batch size on model performance

In model construction and training, batch size setting is an essential factor that affects the model's performance. In the experiments, it is easy to find that setting different batch sizes leads to different experimental results. As the testing results show in Fig. 4, the best choice of batch size for the IEMOCAP dataset with a large data volume is 16. Therefore, the batch size is set to 16 in this paper.

4.1.2. Ablation experiments

To evaluate the effectiveness of the proposed network in this paper, ablation experiments are performed in this section. A detailed analysis of the results is shown in Table 2 and Fig. 5.

(1) Model 1 (M1): This is our proposed model.
(2) Model 2 (M2): This is a subnetwork that extracts emotional embedding from MS.
(3) Model 3 (M3): This is a subnetwork that extracts emotional embedding from LLDs.
(4) Model 4 (M4): This comes from M1 but removes the MPAM.
(5) Model 5 (M5): This comes from M1 but removes the multiscale layer.
(6) Model 6 (M6): This comes from M1 but removes the attention layer.
(7) Model 7 (M7): This comes from M1 but removes the TDGR.
(8) Model 8 (M8): This comes from M1 but removes the FDGR and CDGR.
(9) Model 9 (M9): This comes from M1 but removes the fine-grained fusion strategy.

First, to verify the effectiveness of the combination of MS and LLDs, we compared the performance of M1, M2 and M3. M1 is the proposed framework MPAF-CNN, M2 only takes MS features as model input, and M3 only takes LLDs as model input. The results in Table 3 demonstrate a significant performance advantage of M1 over M2, with improvements of 6.16% and 5.33% on WA and UA, respectively. Similarly, M1 outperforms M3 with a large margin, showing improvements of 5.53% and 4.65% on WA and UA, respectively. Compared with M2 and M3, the SER method that combines MS and LLDs as model inputs exhibits superior classification performance.

Second, to verify the effectiveness of the MPAM, this study compares the proposed framework (M1) with the model obtained by removing MPAM from M1(i.e., M4). The experimental results in Table 3 show that M1 outperforms M4 with a large margin and improves by 2.39% and 4.80% on WA and UA, respectively. MPAM significantly improves emotion recognition performance by extracting multiperspective speech emotion embedding.

Third, to verify the effectiveness of the three perspectives, we compared the performance of M1, M5, M6, M7 and M8. M5 is the model obtained by removing the local perspective from M1. M6 is the model obtained by eliminating the frame-level perspective from M1. M7 is the model obtained by removing TDGR. M8 is the model obtained by removing FDGR and CDGR. The results presented in Table 3 demonstrate that M1 outperforms M5, M6, and M7 by a significant margin, resulting in improvements of 1.63% and 0.83% in WA and UA of M1 compared to M5, the effectiveness of the local perspective has been demonstrated; the WA and UA of M1 improved by 1.45% and 0.34%, respectively, compared to M6, the effectiveness of the frame-level perspective has been demonstrated; the WA and UA of M1 improved by 3.9% and 3.54%, respectively, compared to M7, Furthermore, M1 also outperforms M8 with improvements of 0.82% and 0.80% in WA and UA, respectively. The global perspective's effectiveness was validated by comparing M1, M7, and M8.

Finally, to verify the effectiveness of the fine-grained fusion strategy, we compared the performance of M1 and M9. M9 is a model derived from M1 by removing the fine-grained fusion strategy. The experimental results in Table 3 show that M1 outperforms M9 with a large margin and improves by 1.27% and 1.47% on WA and UA, respectively, validating the proposed fine-grained fusion strategy. This strategy demonstrates the interaction by modeling different emotional representations and capturing their underlying relationships. It increases the amount of emotion-related information, thereby enhancing classification performance.

Fig. 5 shows the confusion matrix of the model classification results, which shows the advantages of the proposed model in this paper. It can be found that the proposed model in this paper has a low recognition effect in the happy emotion, which only reaches 65.00%, and the best recognition effect in the sad emotion, which reaches 80%. It can be seen from Fig. 5 that on the IEMOCAP database, it is easy to misclassify anger and neutral emotions as neutral, and happiness and sadness emotions as happiness. This phenomenon contradicts the user's emotional state and should not occur in human–machine interactions.

4.2. Comparison with other approaches

To verify the effectiveness of the proposed method, the model presented in this paper is compared with some other advanced networks. In contrast to papers that use a combination of a convolutional neural network (CNN) and long short-term memory (LSTM), such as Li et al.'s BLSTM and CNN stacking architecture [18] and Liu et al.'s deep neural network consisting of CNN and ABLSTM [38], some papers propose different approaches for improving emotion recognition. Chen et al. proposed ANSNet, which uses multiscale ideas and attention mechanisms to improve performance [36]. Yao et al. proposed a framework that effectively integrates three distinctive classifiers to fuse multiple features for emotion recognition [26]. Chen et al. also proposed the dual attention-BLSTM, which combines attention mechanisms with BLSTM to improve performance [37].

The experimental results of different methods are shown in Table 3. The model in this paper improves the WAR by 1.85% and the UAR by 2.32% compared with the best model in Table 3 [18,26,34–38], and these results prove that the model proposed in this paper has good classification performance.

There are two main reasons for the method's superiority in this paper. On the one hand, the multiperspective awareness module used in this paper can effectively improve the inadequacy of emotional representation extraction and provide rich emotional representation for the SER method. On the other hand, the fine-grained fusion of frame-level interactions enhances the effectiveness of fusion between different features, improving the whole model's performance.

5. Conclusion

In this paper, we propose a deep learning model combined with a multiperspective awareness module and a fine-grained fusion strategy for SER. Our proposed method uses a multiperspective-aware module to obtain rich emotional information from speech. In addition, an attention mechanism is utilized to focus on the salient features, and a fine-grained fusion strategy is used to fuse the different features. The effectiveness of the proposed method has been verified under a series of comparative experiments and ablation studies on IEMOCAP. Comparing the models in Table 3, the WA and UA of the proposed method achieve 72.19% and 72.88% with absolute increments of more than 1.85% and 1.68%, respectively. In the future, we will try to learn discriminative features and robust representations by using more feature information.

The authors contributed equally to the preparation of the manuscript

and the concept of the re-search. The writing of the draft was by Y.L. and J.H.; the review and editing of the draft were done by G.L. and J.W.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant number 61876131].

References

[1] Waern A. Rosalind picard: affective computing. User Model User-Adap Inter 2002; 12:85–9. https://doi.org/10.1023/A:1013324906380.
[2] Picard, R.W., Picard, R.: Affective Computing. In: EEG-detected olfactory imagery to reveal covert consciousness in minimally conscious state. Brain injury. 29,1729–1735(1997).
[3] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, et al. Emotion recognition in human-computer interaction. IEEE Signal Process Mag 2001;18(1):32–80.
[4] N. JIA and C. Zheng: Emotion Recognition of Depressive Patients Based on General Speech Information. 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP). 618-621(2021). https://doi.org/10.1109/ICSP51882.2021.9408759.
[5] S. Harati, A. Crowell, H. Mayberg and S. Nemati: Depression Severity Classification from Speech Emotion.2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).2018 5763-5766(2018). https://doi.org/10.1109/EMBC.2018.8513610.
[6] B. Waelbers, S. Bromuri and A. P. Henkel: Comparing Neural Networks for Speech Emotion Recognition in Customer Service Interactions.2022 International Joint Conference on Neural Networks (IJCNN). 1-8(2022). https://doi.org/10.1109/IJCNN55064.2022.9892165.
[7] de Lope J, Graña M. An ongoing review of speech emotion recognition. Neurocomputing 2023;528:1–11. https://doi.org/10.1016/j.neucom.2023.01.002.
[8] Liu Z-T, Wu M, Cao W-H, Mao J-W, Xu J-P, Tan G-Z. Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing 2018;273:271–80.
[9] Nancy AM, Kumar GS, Doshi P, Shaw S. Audio based emotion recognition using mel frequency cepstral coefficient and support vector machine. J Comput Theor Nanosci 2018;15(6):2255–8.
[10] Origlia A, Cutugno F, Galatà V. Continuous emotion recognition with phonetic syllables. Speech Comm 2014;57:155–69. https://doi.org/10.1016/j.specom.2013.09.012.
[11] Ververidis D, Kotropoulos C. Emotional speech recognition: Resources, features, and methods. Speech Comm 2006;48(9):1162–81. https://doi.org/10.1016/j.specom.2006.04.003.
[12] Akçay MB, Oğuz K. Kaya Oğuz: Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Comm 2020;116:56–76.
[13] Jahangir R, Teh YW, Hanif F, Mujtaba G. Correction to: deep learning approaches for speech emotion recognition: state of the art and research challenges. Multimed Tools Appl 2021;80(16).
[14] Yi-Lin Lin and Gang Wei: Speech emotion recognition based on HMM and SVM.2005 International Conference on Machine Learning and Cybernetics.8, 4898-4901(2005). https://doi.org/10.1109/ICMLC.2005.1527805.
[15] H. K. Mishra and C. C. Sekhar. Variational Gaussian Mixture Models for Speech Emotion Recognition.2009 Seventh International Conference on Advances in Pattern Recognition. 183-186(2009). https://doi.org/10.1109/ICAPR.2009.89.
[16] Al-Dujaili MJ, Ebrahimi-Moghadam A. Ebrahimi-moghadam, a: speech emotion recognition: a comprehensive survey. Wirel Pers Commun 2023;129(4):2525–61.
[17] E. M. Schmidt and Y. E. Kim: Learning emotion-based acoustic features with deep belief networks. 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 65-68(2011). https://doi.org/10.1109/ASPAA.2011.6082328.
[18] Li D, Sun L, Xu X, Wang Z, Zhang J, Du W. BLSTM and CNN stacking architecture for speech emotion recognition. Neural Process Lett 2021;53(6):4097–115.
[19] Araño KA, Gloor P, Orsenigo C, Vercellis C. When old meets new: emotion recognition from speech signals. Cogn Comput 2021;13(3):771–83.
[20] Pawar MD, Kokate. R.D: convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. Multimed Tools Appl 2021;80:15563–87. https://doi.org/10.1007/s11042-020-10329-2.
[21] Y. Zhang, J. Du, Z. Wang, J. Zhang and Y. Tu: Attention Based Fully Convolutional Network for Speech Emotion Recognition. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),1771-1775(2018). https://doi.org/10.23919/APSIPA.2018.8659587.
[22] W. Zhu and X. Li: Speech Emotion Recognition with Global-Aware Fusion on Multi-Scale Feature Representation, ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6437-6441(2022). https://doi.org/10.1109/ICASSP43922.2022.9747517.
[23] S. Han, F. Leng and Z. Jin: Speech Emotion Recognition with a ResNet-CNN-Transformer Parallel Neural Network. 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), 803-807(2021). https://doi.org/10.1109/CISCE52179.2021.9445906.
[24] X. Jiang, Y. Guo, X. Xiong and H. Tian, "A Speech Emotion Recognition Method Based on Improved Residual Network," 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 2021, pp. 539-542, doi: 10.1109/IAECST54258.2021.9695727.
[25] Zhao Z, Zheng Y, Zhang Z, et al. Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition. conference of the international speech communication association,272-276(2018).
[26] Yao Z, Wang Z, Liu W, Liu Y, Pan J. Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. Speech Commun. 2020;120:11–9. https://doi.org/10.1016/j.specom.2020.03.005.
[27] Liu Zhen-tao Xu, Min J-P, et al. A review of speech emotion feature extraction and dimension reduction methods. Chinese J. Computers 2018;41(12):2833–51.
[28] Xu X, et al. Survey on discriminative feature selection for speech emotion recognition. the 9th. Int Symp Chin Spoken Lang. Processing 2014;345–349. https://doi.org/10.1109/ISCSLP.2014.6936641.
[29] C. -W. Huang and S. S. Narayanan: Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition.2017 IEEE International Conference on Multimedia and Expo (ICME), 583-588(2017). https://doi.org/10.1109/ICME.2017.8019296.
[30] Mohmmad S, Sanampudi SK. Tree Cutting Sound Detection Using Deep Learning Techniques Based on Mel Spectrogram and MFCC Features[C]//Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022. Singapore: Springer Nature Singapore; 2023. p. 497–512.
[31] D. Ververidis C. Kotropoulos Emotional speech recognition: Resources, features, and methods[J] Speech communication 48 9 2006 1162 1181.
[32] Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Evaluation 2008;42(4):335–59.
[33] Li P, Song Y, McLoughlin IV, et al. An attention pooling based representation learning method for speech emotion recognition. Interspeech 2018;3087–3091.
[34] Zhao Z, Bao Z, Zixing, Zhang.et, al.. Attention enhanced connectionist temporal classification for discrete speech emotion recognition. Interspeech 2019;206–210.
[35] Michael Neumann and Ngoc Thang Vu. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 7390–7394(2019).
[36] Chen Z, Li J, Liu H, Wang X, Wang Hu, Zheng Q. Learning multi-scale features for speech emotion recognition with connection attention mechanism. Expert Syst Appl 2023;214. https://doi.org/10.1016/j.eswa.2022.118943.
[37] Chen Q, Huang G. A novel dual attention-based BLSTM with hybrid features in speech emotion recognition. Eng Appl Artif Intel 2021;102:104277. https://doi.org/10.1016/j.engappai.2021.104277.
[38] Liu Z-T, Han M-T, Bao-Han Wu, Rehman A. Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning. Appl Acoust 2023;202:109178.