IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 April 2023, accepted 16 May 2023, date of publication 19 May 2023, date of current version 30 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3278106

RESEARCH ARTICLE

Speech Emotion Recognition Based on Attention MCNN Combined With Gender Information

ZHANGFANG HU, KEHUAN LINGHU, HONGLING YU, AND CHENZHUO LIAO
Key Laboratory of Optoelectronic Information Sensing and Technology, Chongqing University of Posts and Telecommunications (CQUPT), Chongqing 400065, China

Corresponding author: KeHuan LingHu (s210431053@stu.cqupt.edu.cn)

This work was supported in part by the National Natural Science Foundation Youth Fund Project under Grant 61801061, and in part by the National Natural Science Foundation of China under Grant 51905065 and Grant 52005070.

ABSTRACT Emotion recognition is susceptible to interference such as feature redundancy and speaker gender differences, resulting in low recognition accuracy. This paper proposes a speech emotion recognition (SER) method based on attention mixed convolutional neural network (MCNN) combined with gender information, including two stages of gender recognition and emotion recognition. (1) Using MCNN to identify gender and classify speech samples into male and female. (2) According to the output of the first stage classification, a gender-specific emotion recognition model is established by introducing coordinated attention and a series of gated recurrent network units connecting the attention mechanism (A-GRUs) to achieve emotion recognition results of different genders. The inputs of both stages are dynamic 3D MFCC features generated from the original speech database. The proposed method achieves 95.02% and 86.34% accuracy on EMO-DB and RAVDESS datasets, respectively. The experimental results show that the proposed SER system combined with gender information significantly improves the recognition performance.

INDEX TERMS SER, convolutional neural network, gender information, attention, GRU.

I. INTRODUCTION

Speech is the most direct and natural way of human communication and is most easily accessible without the influence of other factors, allowing effective transmission of information. Speech with emotion makes everyone and human-computer communication efficient and dynamic. Speech emotion recognition (SER) is a relatively active research direction in human-computer interaction and digital signal processing [1], and speech emotion recognition systems based on deep learning have made great contributions in many aspects of artificial intelligence, such as in healthcare, education, service industries, in-car driving systems, and transportation. However, there is still a large gap between the performance of currently available speech emotion recognition technologies in practical applications and the emotional information perceived by human hearing [2]. On the one hand, emotion perception is quite subjective [3] and varies by individual listeners, such as gender, age, and culture, leading to complexity and uncertainty in labeling emotion labels, and on the other hand, not all features in the original speech signal are valid for emotion recognition and contain many silent and blank frames that are not related to emotion [4]. Despite the strong subjectivity of emotions, it is undeniable that robots improve the relevance and accuracy of human-machine plus care through speech emotion recognition, which lays the foundation for realizing machine emotion bionics.

The acoustic features commonly used in speech emotion recognition systems are Mel-Frequency Cepstral Coefficient (MFCC), amplitude, over-zero rate, fundamental frequency, resonance peak, and short-time energy, etc., among which the best performance is MFCC [5], through the evaluation of voice quality, the physiological and psychological information of the speaker can be obtained and differentiation of their emotional state. Speech signals represented as speech spectrograms are extracted by deep learning networks, such as convolutional neural networks (CNN) [6], recurrent neural networks (RNN) [7], and convolutional recurrent networks (CRNN), which combine both, are also often used to extract feature sequences and capture temporal dependencies.

The associate editor coordinating the review of this manuscript and approving it for publication was Mounim A. El Yacoubi.

Moreover, the use of attentional mechanisms for SER has been shown to have higher recognition rates than traditional speech emotion recognition methods [8], because attentional mechanisms are able to focus on salient emotional features in speech periods and can handle emotional features of different granularity.

Many factors will affect the performance of speech emotion recognition system. Human gender is a factor leading to average physiological differences, and differences in physiological characteristics can cause differences in acoustic characteristics. Compared with women, the pitch difference of male speech is relatively small, but the pitch difference between male and female speech is quite significant. For example, a happy male voice may be confused with a calm female voice, leading to difficulties in speech emotion recognition. As speech emotion recognition technology continues to evolve, the use of deep learning algorithms to leverage gender information into SER systems can be summarized in two ways: one is to create a separate emotion model for each gender. The second is to create a dependency model using gender information as an augmented feature vector. The former gender information does not need to be represented and is divided into gender recognition and sentiment recognition, which are separately trained to go for sentiment classification, while the latter aims to provide a priori information about the speaker's gender. Both approaches can make use of the speaker's gender information and thus improve the accuracy of SER.

This study considers the gender differences of speakers and is inspired by the successful application of convolutional neural networks and attention mechanisms to SER to improve system performance. A hybrid convolutional model based on attention and gender information is proposed for speech emotion recognition. The main contributions of our proposed system are as follows:

Firstly, in order to analyze the time variation of speech data, we add velocity and acceleration features to the static two-dimensional MFCC and stack them to generate dynamic three-dimensional MFCC as network input for pre-training.

Secondly, we propose a mixed convolutional neural network (MCNN) model with dilated convolution for gender recognition and classify speech samples into male and female. Then, in order to better capture emotional location features and context information, we introduce coordinate attention (CA) and a series of gated recurrent network units connecting the attention mechanism (A-GRUs) into the original MCNN architecture to establish an emotion recognition model. Finally, according to different genders, a specific emotion recognition model is used to complete speech emotion recognition. Experiments show that our proposed model achieves the best recognition effect in both classification tasks.

The rest of this article is structured as follows: The second section gives a SER recommendation method that combines gender information. The third section shows the experimental results of our method on relevant databases and compares it with other published works. The fourth section summarizes the work of this paper.

II. RELATED WORK

In the past research on speech emotion recognition algorithms, including speaker gender information has been shown to improve SER accuracy. Gender-based emotion recognizers produce better results than gender-independent emotion recognizers [9]. Bisio et al. [10] proposed an emotion classification algorithm based on pitch features to build gender recognition, aiming to provide a priori information about the speaker's gender, and used SVM as a classifier with a recognition rate of 81.5% on the EMO-DB dataset. Liu and Zhang [11] proposed two models based on adversarial multi-task learning with emotion recognition as the main task and by adding noise recognition and gender recognition as auxiliary tasks respectively, the accuracy rate reached 89.13% on the AVEC database. Similarly, Liu et al. [12] proposed a multi-task learning SER model with CNN, attention-based bidirectional long and short-term memory network (ABLSTM) and gender as an auxiliary task, which achieved 70.27% and 66.27% WAR and UAR, respectively, on the IEMOCAPS database. Kanwal and Asghar [13] proposed a density-based noise genetic algorithm applied to spatial clustering (DGA) and combined with gender information for optimization using SVM as a classifier with recognition rates of 89.6%, 82.5%, and 77.7% on EMO-DB, RAVDESS, and SAVEE datasets, respectively. Sun [14] proposed a new emotion recognition algorithm that does not rely on any acoustic features and combines residual convolutional neural network (R-CNN) and gender information blocks, using a deep learning algorithm to select important information from the original speech signal for the classification layer to complete emotion recognition, and the results showed that the proposed algorithm achieved recognition rates of 84.6%, 90.3% and 71.5% on CASIA, EMODB and IEMOCAP datasets, respectively.

Recently, Falahzadeh et al. [15] proposed a 3D CNN model with gender information to compute the three-dimensional reconstructed phase spaces (3D RPS) from the original speech signal and convert it into a 3D tensor for emotion recognition on the EMO-DB and eNTERFACE'05 datasets with recognition rates of 94.42% and 88.47%, respectively. Zhang et al. proposed a gender classification-based emotion recognition algorithm [16], where the original speech is classified by gender using a multilayer perceptron (MLP) at the front end and different genders are recognized using a CNN-BLSTM model at the back end. The recognition rates of the proposed algorithm are 84.72% and 87.91% on the RAVDESS and CASIA datasets, respectively.

Although the above studies considered the effect of gender information on SER performance, only gender information was used as a secondary task or required to construct respective sentiment feature sets by gender, and only traditional CNN or RNN models were used in the sentiment recognition stage, and the sentiment recognition results were not fully

improved. Therefore, we will focus on improving the performance of traditional CNN for extracting features and capturing sentiment information using attention mechanism with RNN variant module while considering gender information to improve the recognition accuracy.

III. PROPOSED METHOD

In this section, we establish an overall SER framework that incorporates gender information, as shown in Figure 1. It consists of two stages: (1) gender recognition (2) emotion recognition. In the first stage, we use the dynamic 3D MFCC features generated from the speech database as the input of the gender recognition network, use MCNN to identify gender and divide the speech samples into male and female. In the second stage, based on the output of the first stage classification, dynamic 3D MFCC features are extracted from male and female speech samples respectively and input into the established emotion recognition model to realize the emotion recognition of different genders. In addition, in the second stage of the emotion model, we introduce the coordinate attention and A-GRUs model to better capture emotional features and context information.

A. SPEECH SIGNAL PREPROCESSING

Speech feature is an important topic in speech emotion recognition. Traditionally, a large number of studies on SER use only a single feature as input [17]. A recent study on device classification shows that performance can be improved by stacking multiple features extracted from sequential data or combining multiple input features [18]. Based on the characteristics of human ear, cochlea and basement membrane, MFCC has a nonlinear correspondence with the actual frequency, which makes its cepstrum coefficient more similar to the nonlinear human auditory system. MFCC mean and fundamental frequency F0 can classify speech features of different genders more accurately [19]. However, the speech signal is a dynamically changing signal, and the MFCC feature alone does not consider the relationship between time changes in the speech signal. In order to reflect the dynamic characteristics of the speech signal, we add the MFCC speed (delta MFCC) and acceleration (double-delta MFCC).

In this study, the dynamic 3D MFCC is used as the input of the proposed network model, and the extraction process is shown in Figure 2. The first step is to extract MFCC features. The given speech signal is divided into frames (about 20 ms), and the length of time between consecutive frames is 5-10 ms. Before performing Fourier transform on each frame signal, a Hamming window is used, and the window length is equal to the frame length. Short-time Fourier transform is performed on each frame, and the power spectrum is obtained by summing the squares. The short-term power spectrum is a comprehensive representation of speech noise characteristics, including 2D spatial information in frequency domain and time domain. Logarithmic scaling is commonly used in Mel frequency spectra to adapt to human auditory factors, showing a linear distribution below 1000 Hz and a logarithmic growth above 1000 Hz [20]. MFCC features are obtained by discrete cosine transform of logarithmic Mel spectrum [21].

Since the original MFCC features are static, in order to add dynamic information to the static MFCC features, we add delta features and double-delta features to form multi-dimensional dynamic features, which are performed by local estimation of the differential operation of the input MFCC features along the time axis. Incremental features and double incremental features provide dynamic information of the original features over time. Assuming that the MFCC at the frame t is C_t, the corresponding delta spectral feature D_t is defined as follows [22]:

$$D_t = C_{t+m} - C_{t-m} \tag{1}$$

where m represents the number of adjacent frames D_t represents the delta MFCC coefficient at frame, which is calculated by the static coefficients C_{t+m} and C_{t-m}. Similarly, double-delta MFCC is defined based on the subsequent delta operation of delta MFCC. The extracted MFCC, delta MFCC and double-delta MFCC are combined to obtain a dynamic three-dimensional MFCC. The final input feature shape of the proposed network model is $224 \times 224 \times 3$.

B. PROPOSED MIXED CONVOLUTION BLOCK

Since the standard convolution alone cannot obtain a large receptive field, the separate dilated convolution has intervals that may not allow all inputs to participate in the operation, resulting in feature discontinuity. In order to solve this problem, we propose a hybrid convolutional layer for gender and emotional feature extraction on the basis of reducing computational complexity, as shown in Figure 3.

The mixed convolution layer combines dilated convolution [23] and standard convolution in the same layer, and can use the same convolution kernel to obtain larger spatial information in the dynamic three-dimensional MFCC spectrum without adding parameters. The mixed convolution layer is formed as follows:

$$[\sigma(\omega_s); \sigma(\omega_d)] \tag{2}$$

where ω_s and ω_d are the parameters of standard convolution and dilated convolution respectively, σ is a combination of GN and ReLUs, which hides the bias for simplicity. In order to obtain fine-grained features, we add GN and ReLUs after the convolution layer, and determine the mixed convolution layer, GN and ReLUs as a mixed convolution block. GN is used to normalize the feature maps with deep features on the Mel spectrogram, so that the model can be stably normalized according to the number of training samples, regardless of the batch size. The ReLUs function is used to activate the model, and the corresponding output characteristic is to set some of the outputs to 0, thereby reducing the mutual dependence and alleviating the gradient disappearance caused by the deepening of the network layer. The features extracted by standard convolution and dilated convolution are connected

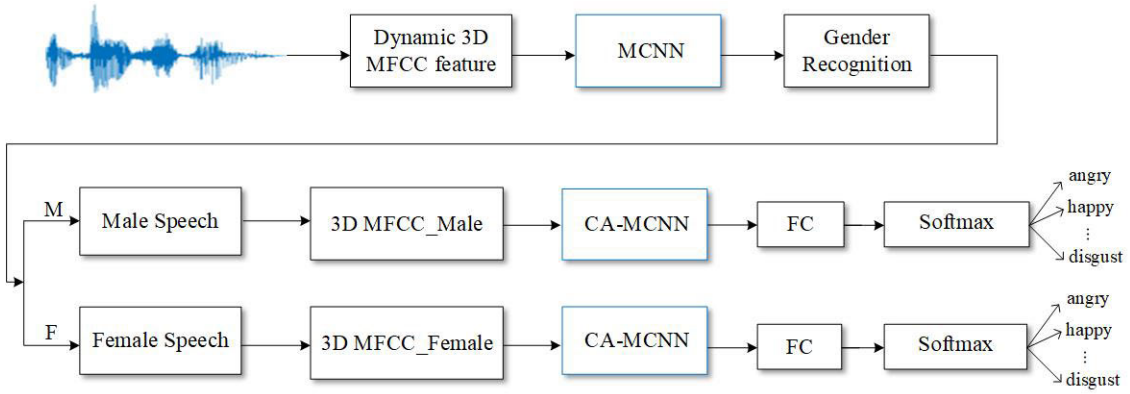


FIGURE 1. SER network architecture combined with gender information.

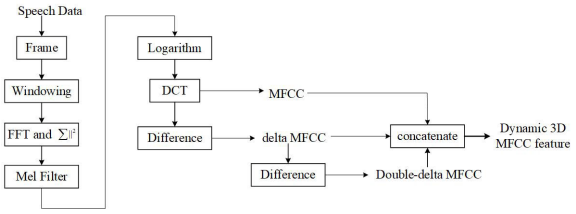


FIGURE 2. Preprocessing extraction of three-dimensional dynamic MFCC feature flow.

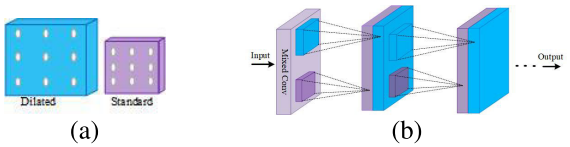


FIGURE 3. Mixed convolution formation and extraction diagram: (a) Two types of convolution modules. (b) Mixed convolution extraction.

together, and then we apply GN and ReLUs to these features. More specifically, it is assumed that the output of the mixed convolution layer has Q channels, the first P channels are obtained by dilated convolution, and the remaining $Q - P$ channels are obtained by standard convolution, The channel ratio of standard convolution to dilated convolution is 3:1.

In order to simplify the problem and achieve a fair comparison, we introduce the following constraints in the configuration of the network architecture:

(1)All mixed convolution layers have the same number of convolution kernels.

(2)Two convolution operations (standard convolution and dilated convolution) have convolution kernels of the same size.

(3)The expansion factor of each dilated convolution is 2.

The details of each mixed convolutional layer for gender recognition and emotion recognition are shown in Table 1.

C. GENDER IDENTIFICATION MODEL

The first stage of the hybrid convolutional network model architecture for gender recognition is shown in Figure 4.

TABLE 1. Details of each mixed convolution layer.

Layers	Kernel_Size	Stride	Channel_In	Channel_Out
Mixed-Conv1	(2,2)	2	3	32
Max pooling	(2,2)	2	32	32
Mixed-Conv2	(1,1)	1	32	32
Mixed-Conv3	(1,1)	1	32	96
Mixed-Conv4	(2,2)	2	96	96
Mixed-Conv5	(1,1)	1	96	96
Mixed-Conv6	(2,2)	2	96	288
Mixed-Conv7	(2,2)	2	288	288
Average pooling	(2,2)	2	288	288

We use five mixed convolution blocks, one pooling layer and two fully connected layers to complete gender classification. The first fully connected layer (FC1) consists of 1000 neurons, and the last fully connected layer (FC2) is a classification layer with two neurons corresponding to male or female.

D. EMOTION RECOGNITION MODEL

The proposed MCNN model for emotion recognition in the second stage is shown in Figure 5. A coordinate attention module is added between the third and fourth layers of the hybrid convolutional layer, the fifth and sixth layers, and the seventh layer and the average pooling layer. With the idea of residual neural network, the residual path directly connects the input to the output, which can speed up the training speed and avoid overfitting. In addition, this paper constructs a multi-layer GRUs network work. This architecture can model the forward and reverse information of the input at each timestamp, so that the past and future information is saved, and the attention mechanism is combined to pay attention to important context information.

1) COORDINATE ATTENTION

The attention mechanism [24] is an algorithm based on the importance of different parts of a certain thing, that is, assign

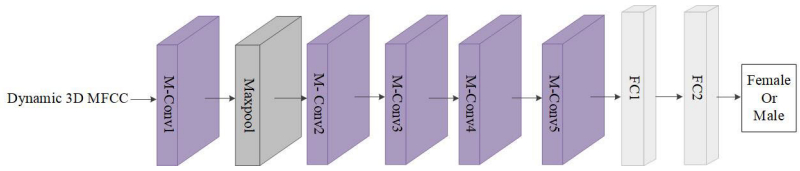


FIGURE 4. The proposed MCNN architecture for gender recognition.

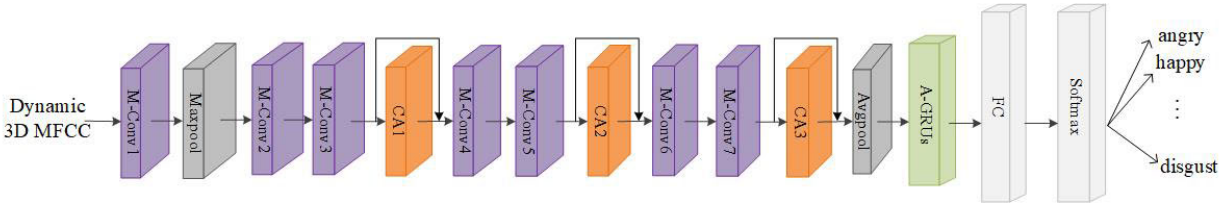


FIGURE 5. The proposed CA-MCNN architecture for emotion recognition.

more attention to the key parts of the thing, and assign more weight to a key part by calculating the probability distribution of attention.

At present, the most popular framework in the SER field is the convolutional block attention module (CBAM) [25]. Although it considers channel and spatial location information, it uses large-scale pooling to use location information to only capture local correlations, and it is difficult to model long-term dependencies. Therefore, we introduce a relatively new method: coordinate attention [26]. The specific calculation process is shown in Figure 6. The specific steps of coordinating attention can be divided into two parts: coordinate information embedding and coordinate attention generation. The former encodes channel information in horizontal and vertical coordinates, and the latter captures location information and generates weight values.

Step1 coordinate information embedding: for a given input element $X = [x_1, x_2, \ldots, x_c] \in \mathbb{R}^{C\times H\times W}$, the pooled kernels with sizes $(H, 1)$ and $(1, W)$ are used to encode information from different channels along the horizontal and vertical directions, respectively. The output process of the characteristics of the C channel at the height H is:

$$z_c^h(h) = \frac{1}{\mathrm{W}} \sum_{0\le i\le w} x_c(h, i) \tag{3}$$

Similarly, the output of channel C at width W is:

$$z_c^w(w) = \frac{1}{H} \sum_{0\le i\le \mathrm{H}} x_c(j, w) \tag{4}$$

The formula (3) and formula (4) generate a pair of directional perception feature maps to realize the embedding of coordinate information.

Step 2 coordinate attention generation: connect two coding features in the spatial dimension, and the length becomes $(H + W)$. Then we use the shared convolution transform function F_1 to obtain:

$$f = \delta(F_1(\left[z^h, z^w\right])) \tag{5}$$

Among them, $\left[z^h, z^w\right]$ represents the series operation along the spatial dimension, δ is a nonlinear activation function, $f \in \mathbb{R}^{c/r\times(H+W)}$ is an intermediate element mapping, which is used to encode the spatial information in the horizontal and vertical directions, where r is the control block size reduction rate, generally 32, with formula (6) to reduce the number of channels of f.

$$C_{out} = \max(8, C_{in}/r) \tag{6}$$

F is decomposed into two independent tensors along the spatial dimension: $f^h \in \mathbb{R}^{c/r\times H}$ and $f^w \in \mathbb{R}^{c/r\times W}$, using two convolution transforms for and so that they preserve tensors with the same number of channels as X input. Then, the sigmoid activation function is used to obtain g^h and g^w, which are realized by Formula (7) and Formula (8):

$$g^h = \delta(F_h(f^h)) \tag{7}$$

$$g^w = \delta(F_w(f^w)) \tag{8}$$

where F_h and F_w are two 1×1 convolutions, g^h and g^w are two-dimensional weights. Finally, g^h and g^w are fused with the input feature X to obtain the output of the coordinate attention module:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{9}$$

The coordination attention module embeds position information into channel attention, which increases the spatial range of attention and avoids a lot of computational overhead. The module uses two parallel one-dimensional feature coding to integrate channel and spatial coordinate information into the generated attention map, and eliminates the problem of location information loss caused by two-dimensional global pooling in CBAM. The module is a plug-and-play model, flexible and lightweight. It not only considers the channel and

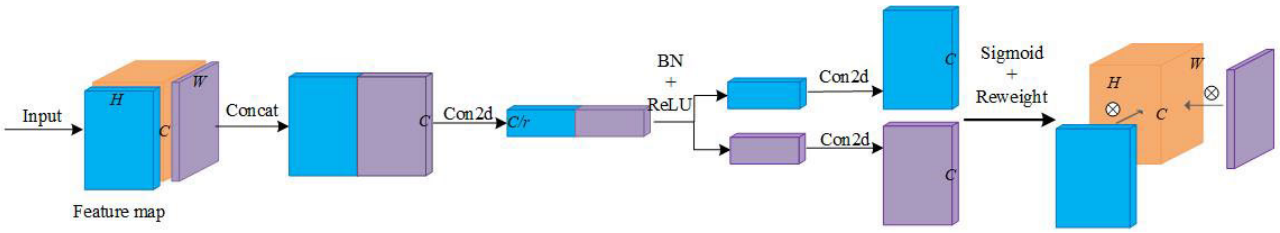


FIGURE 6. Coordinate attention calculation process.

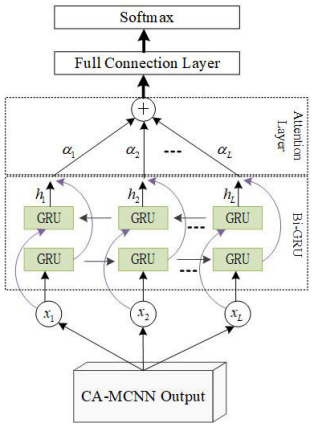


FIGURE 7. Network block diagram of A-GRUs model.

space in parallel, but also solves the long-term dependence problem.

2) A-GRUs STRUCTURE

The network block diagram of the A-GRUs model is shown in Figure 7. The gate recurrent unit (GRU) used in this paper is a special form of RNN [27]. Each GRU unit has the ability to learn time context information, which propagates information through hidden states. The datasets in the field of SER are generally small, and there is little difference in performance between GRU and LSTM in heavy training tasks. Compared with the long short-term memory network (LSTM), GRU has only one hidden state, the structure is simpler, and contains fewer parameters [28], so it can converge quickly. The attention mechanism weights the input sequence information and distinguishes the importance of the feature information according to the weight. In this paper, we use bidirectional gate recurrent unit(Bi-GRU) to extract context information, use CA-MCNN to extract deep salient features from speech samples, and pass them to Bi-GRU network to capture order information, and add attention layer to focus on the emotional related part of speech feature information.

In this model, the set Bi-GRU has 512 bidirectional hidden units, and then a new sequence with a shape of $L \times 1024$ is created and placed in the attention layer, where $H = \{h_1, h_2, \cdots, h_L\}$, L is the length of time (frame) and d is the size of the Bi-GRU hidden layer. The specific implementation of the attention layer is as follows:

$$e_i = \tanh(W_j h_i + b_j) \tag{10}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_i \exp(e_i)} \tag{11}$$

$$h_i' = \alpha_i h_i \tag{12}$$

where W_j and b_j are trainable parameters, α_i is the attention weight $\sum_i \alpha_i = 1$, h_i' is the characteristic value weighted by h_i.

TABLE 2. Emotional analogy and gender distribution of EMO-DB and RAVDESS datasets.

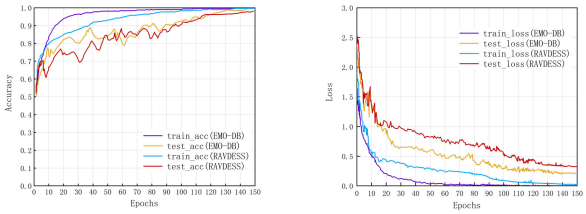
Dateset	Emotions	Sample	Male	Female
EMO_DB	Angry, Disgust, Fear, Happy, Bored, Neutral,Sad	523	223	302
RAVDESS	Angry, Disgust, Fear, Happy, Clam,Neutral, Sad, Surprise	1440	720	720

IV. EXPERIMENTAL RESULTS AND COMPARISON

A. EMOTIONAL SPEECH DATABASE

The Berlin Database of Emotional Speech(EMO-DB) [29] contains the following seven emotional categories: anger, boredom, neutral, disgust, fear, happiness, sadness. The sound was recorded by five male and five female actors aged 20-30. The speech corpus consists of 10 German phrases of different lengths, and the total number of speech files is 535. The sound file is captured at 16 kHz sampling frequency, 16 bit resolution and single channel. The average time length of each audio file is three seconds.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [30] contains 24 skilled actors, 12 males and 12 females, with neutral North American accents clearly expressing two words similar sentences. Speech data includes eight emotions, namely, calm, happiness, sadness, anger, fear, surprise, disgust and neutral. The dataset contains audio-only, audio-visual and video-only files, where we only select audio voice files for our experiments, and the pure audio voice sample sampling frequency is 48 kHz. The sentiment category and gender distribution of the two datasets are as follows Table 2.



(a) (b)

FIGURE 8. Visualization results on EMO-DB and RAVDESS datasets for gender classification: (a) Accuracy of training and testing. (b)Loss rate of training and testing.

B. EXPERIMENTAL SETTINGS

The experimental hardware conditions are Intel Xeon E5 CPU and NVIDIA 2080ti GPU. The system environment is Ubuntu 16.04 LTS, and Python is selected as the basic development language. In the experiment, the TensorFlow2.0 toolkit [31] was used to complete the construction of the proposed network model and the implementation of the training algorithm. In order to avoid overfitting, we choose the early stopping method in the experiment, and use the cross entropy error function as the training objective function to minimize the cross entropy loss to train the network. Adam algorithm is used for optimization. The learning rate of the network is 0.001, the batch size is 10, and the epoch number is 150. For each experiment, we divided the data in a ratio of 8:2, with 80% for training and 20% for testing.

The training of gender recognition, male emotion recognition and female emotion recognition is completed separately using the MCNN model architecture. Different from the gender recognition model, the male and female gender emotion recognition model adds a coordinated attention module and A-GRUs to better capture the emotional cues in the features. The mixed samples of men and women are provided to the gender classifier for training and the female emotion classifier and the male emotion classifier are trained independently using only female samples and only male samples respectively. The output of the gender recognition model determines that the gender-specific emotion recognition model is used. Therefore, for a given audio sample, only two models are used. However, the input of these three models is the dynamic three-dimensional MFCC spectrum obtained after feature extraction.

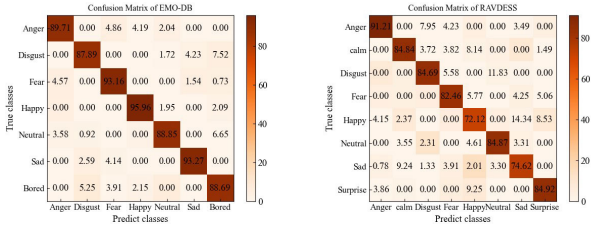
C. EXPERIMENTAL RESULTS

1) GENDER IDENTIFICATION RESULTS

In order to evaluate the accuracy of identifying men and women, the accuracy and loss rate are trained and tested on two datasets of EMO-DB and RAVDESS. The visualization results are shown in Figure 8. Through testing, the accuracy of gender classification in EMO-DB dataset can reach nearly 100%, and the accuracy of RVADESS dataset can reach nearly 99.5%.

TABLE 3. Compares the recognition rate of gender information on RAVDESS and EMO-DB datasets.

Method	Database(Average Accuracy %)					
	EMO-DB			RAVDESS		
	Male	Female	comprehensive	Male	Female	comprehensive
Without Gender	\	\	91.08	\	\	82.47
With Gender	95.79	94.24	95.02	86.53	85.24	86.34



(a) (b)

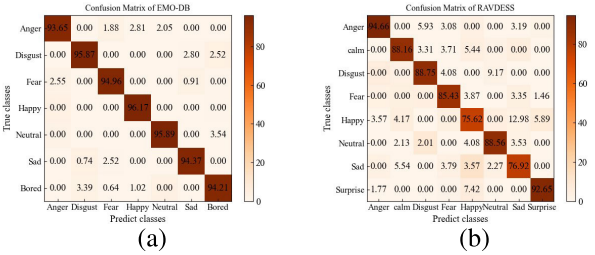
FIGURE 9. Confusion matrix results without gender information: (a) Results on EMO-DB dataset. (b)Results on RAVDESS dataset.

2) SPEECH EMOTION RECOGNITION RESULTS

In order to consider the influence of gender information on sentiment classification, this paper trains and tests the models with and without gender information on EMO-DB and RAVDESS datasets respectively, and obtains the recognition results as shown in Table 3.

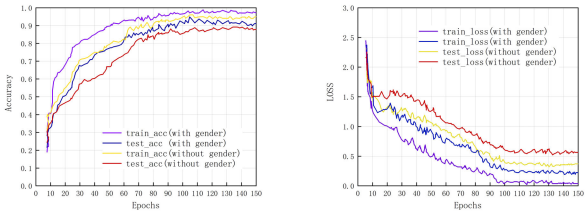
It can be seen from Table 3 that in the EMO-DB dataset, the comprehensive recognition rate of gender classification is 3.94% higher than that of no gender classification. In the RAVDESS dataset, the comprehensive recognition rate of gender classification is 3.87% higher than that without gender classification. In addition, the average accuracy of male speech recognition is significantly higher than that of female, which also shows that the changes of male speech emotion characteristics are more easily recognized by the model. Comparing the confusion matrix of the two data sets without gender recognition model and with gender recognition model is shown in Figure 9 and Figure 10.

In order to analyze the performance of the proposed model more intuitively, the accuracy and loss rate of training and testing on EMO-DB and RAVDESS datasets are given. The visualization results are shown in Figure 11 and Figure 12.



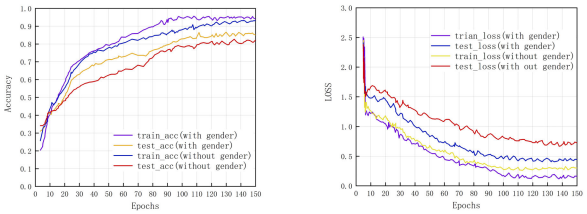
(a) (b)

FIGURE 10. Confusion matrix results with gender information: (a) Results on EMO-DB dataset. (b)Results on RAVDESS dataset.



(a) (b)

FIGURE 11. Visualization results on EMO-DB dataset with or without gender information: (a) Accuracy of training and testing. (b)Loss rate of training and testing.



(a) (b)

FIGURE 12. Visualization results on RAVDESS dataset with or without gender information: (a) Accuracy of training and testing. (b)Loss rate of training and testing.

Through the analysis of Figure 11 and Figure 12, it can be seen that the training and testing accuracy of the proposed model considering gender information on EMO-DB and RAVDESS datasets is better than that without considering gender information. At the same time, the training and test loss rate considering gender information is significantly lower than that without considering gender information. Therefore, the method in this paper can obtain higher accuracy and lower loss value in training and testing, which proves the effectiveness and feasibility of the method.

D. COMPARISON AND ANALYSIS OF NETWORK ARCHITECTURE

In this section, we will compare the proposed MCNN with the baseline CNN model, and the comparison results are shown in Table 4. In addition, in order to further illustrate the superiority of the proposed model, we compare the algorithms based on deep learning (without gender information) and gender information. The comparison results are shown in Table 5 and Table 6.

1) COMPARISON WITH BASELINE CNN MODEL

The convolutional layers and corresponding parameter settings of CNN without dilated convolution are consistent with the proposed MCNN, and the coordinate attention and A-GRUs position remain unchanged. Since the ultimate goal of this paper is emotion recognition regardless of whether the speaker is male or female, we directly consider the final recognition results rather than the recognition results of each stage to evaluate its performance.

TABLE 4. Comparison of the proposed MCNN model with the baseline CNN model.

Recognition Architecture	Database(Average Accuracy %)	
	EMO-DB	RAVDESS
Baseline CNN	93.80	85.27
Proposed MCNN	95.02	86.34

As can be seen from Table 4, the MCNN model with the addition of the dilated convolutional channel in both databases improves the average recognition accuracy by nearly 1.2% over the baseline CNN model. It can be seen that dilated convolution is beneficial to improve the recognition rate by covering a larger receptive field to obtain more spatial information. At the same time, the baseline CNN model still has high recognition accuracy, which also proves that our proposed network architecture has better recognition performance for sentiment classification.

2) ACCURACY COMPARISON BASED ON DEEP LEARNING ALGORITHMS

In [32], three convolutional layers are used to extract features from the input spectrogram in time and frequency, and support vector machine (SVM) is used for sentiment classification. In [33], RNN is added to CNN to extract features. Both show that the performance of CRNN model is better than that of CNN model. In [34], the author designed a 2D CNN model using a visual attention module with channels and spaces to learn emotional features from og-mel spectrograms. In [35], based on the traditional CNN model, the author used the method based on redial based function network (RBFN) to select the key sequence of the speech spectrogram and feed it to the bidirectional long short-term memory network (BLSTM) to identify the final time information of the emotional state. In [36], the author proposed an artificial intelligence-assisted deep step convolutional neural network (DSCNN) to extract emotional features from the spectrogram, using a special stride to down-sample the feature map, and finally learning the global discriminant features in the fully connected layer. In [37], the author used deep neural network (DCNN) to learn the high-level features of each segment divided in the 3D log-mels spectrogram, and input the learned segment-level features into discriminative time pyramid matching (DTPM) to form a global discourse-level feature representation for sentiment classification. In [38], the author divided the CBAM module into channels and spatial attention modules and added them to the traditional CNN to form a residual attention convolutional neural network (RACNN) architecture to extract more emotional details from the spectrogram.

From the comparison between [33] and [35], it is obvious that the attention mechanism is beneficial to the SER system to improve the recognition performance. Compared with References [36] and [37], the recognition accuracy of the proposed method on EMO-DB and RAVDESS databases is improved by 3.77% and 2.97%, respectively. This paper

TABLE 5. The proposed method is compared with other methods on EMO-DB and RAVDESS datasets.

Literature	Method	Database(Average Accuracy%)	
		EMO-DB	RAVDESS
Ref [32]	CNN+SVM	64.33	-
Ref [33]	CRNN	80.00	-
Ref [34]	VACNN	-	74.31
Ref [35]	CNN+BLSTM	85.57	77.02
Ref [36]	DSCNN	-	79.50
Ref [37]	DCNN+DTPM	87.31	-
Ref [38]	RACNN	-	81.76
Proposed	CA-MCNN+A-GRUs	91.08	82.47

TABLE 6. The proposed method is compared with other methods on EMO-DB and RAVDESS datasets.

Literature	Method	Database(Average Accuracy%)	
		EMO-DB	RAVDESS
Ref [13]	DGA+SVM	89.65	82.50
Ref [14]	Raw Speech+R-CNN	90.03	-
Ref [16]	MFCC+CNN+BLSTM	-	84.72
Ref [15]	3D RPS+3D-CNN	94.42	-
Proposed	3D MFCC+CA-MCNN	95.02	86.34

improves the baseline CNN network by introducing mixed convolution blocks and coordinate attention model, and pre-trains the voice network to improve the generalization ability of cross-database SER research. In addition, by adding the A-GRUs model, the improved attention-based MCNN model has better spatial-temporal feature learning ability and effectively improves the accuracy of emotion recognition.

3) COMPARISON OF THE ACCURACY OF ALGORITHMS BASED ON GENDER INFORMATION

It can be seen from Table 6 that 3D MFCC can not only use acoustic information and real individual differences to better express gender information, but also retain information about effective emotions and their changes. The experimental results show that the SER model designed in this paper has good generalization ability. In addition, classification based on gender information helps to improve the accuracy of emotion recognition and make the model more robust.

V. CONCLUSION

This study proposes a speech emotion recognition method based on attention MCNN combined with gender information. This method has two functional stages: gender classification and sentiment classification. The output of the first stage of gender classification determines the use of the gender-specific sentiment classification model in the second stage. Thus, the problem of low accuracy of emotion recognition due to the existence of pitch differences between genders that are difficult to overcome is solved. Firstly, we extract 3D MFCC spectra from the original speech signal as network input. Then, we designed and pre-trained a MCNN network to identify gender and male and female speech signal classification, and extracted the corresponding 3D MFCC spectra from male and female speech samples, respectively. Finally, by introducing the coordinate attention model and A-GRUs, emotional features are extracted from gender-specific spectrograms to provide emotional recognition results for different genders. The improved MCNN model is tested on RAVDESS and EMO-DB databases. The results show that the three-dimensional MFCC features of speech signals can effectively identify gender features and speaker 's emotional state, which is helpful to improve the robustness of the model. Compared with other similar methods, the proposed method has better performance and can effectively improve the recognition rate of SER.

The value of the method proposed in this paper in speech emotion recognition research is that the difficulty of model recognition can be reduced by distinguishing gender. However, the proposed algorithm cannot run in real time, the computational load is large, and it is difficult to integrate into mobile devices, which reduces the application scenario. In future work, we intend to combine more modes for training and testing under more conditions, such as images and ages.

REFERENCES

[1] W. Dai, D. Han, Y. Dai, and D. Xu, "Emotion recognition and affective computing on vocal social media," *Inf. Manage.*, vol. 52, no. 7, pp. 777–788, Nov. 2015.

[2] U. Garg, S. Agarwal, S. Gupta, R. Dutt, and D. Singh, "Prediction of emotions from the audio speech signals using MFCC, MEL and Chroma," in *Proc. 12th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Sep. 2020, pp. 87–91.

[3] H. Chou and C. Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5886–5890.

[4] S. Lugovic, I. Dunder, and M. Horvat, "Techniques and applications of emotion recognition in speech," in *Proc. 39th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2016, pp. 1278–1283.

[5] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in *Proc. Interspeech*, Sep. 2005, pp. 493–496.

[6] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[7] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2227–2231.

[8] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6319–6323.

[9] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016, pp. 1–4.

[10] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarrone, "Gender-driven emotion recognition through speech signals for ambient intelligence applications," *IEEE Trans. Emerg. Topics Comput.*, vol. 1, no. 2, pp. 244–257, Dec. 2013.

[11] L. Yunxiang and Z. Kexin, "Design of efficient speech emotion recognition based on multi task learning," *IEEE Access*, vol. 11, pp. 5528–5537, 2023.

[12] Z.-T. Liu, M.-T. Han, B.-H. Wu, and A. Rehman, "Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning," *Appl. Acoust.*, vol. 202, Jan. 2023, Art. no. 109178.

[13] S. Kanwal and S. Asghar, "Speech emotion recognition using clustering based GA-optimized feature set," *IEEE Access*, vol. 9, pp. 125830–125842, 2021.

[14] T. Sun, "End-to-end speech emotion recognition with gender information," *IEEE Access*, vol. 8, pp. 152423–152438, 2020.
[15] M. R. Falahzadeh, E. Z. Farsa, A. Harimi, A. Ahmadi, and A. Abraham, "3D convolutional neural network for speech emotion recognition with its realization on Intel CPU and NVIDIA GPU," *IEEE Access*, vol. 10, pp. 112460–112471, 2022.
[16] L.-M. Zhang, Y. Li, Y.-T. Zhang, G. W. Ng, Y.-B. Leau, and H. Yan, "A deep learning method using gender-specific features for emotion recognition," *Sensors*, vol. 23, no. 3, p. 1355, Jan. 2023.
[17] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, vol. 492, pp. 245–263, Jul. 2022.
[18] J.-G. Kim and B. Lee, "Appliance classification by power signal analysis based on multi-feature combination multi-layer LSTM," *Energies*, vol. 12, no. 14, p. 2804, Jul. 2019.
[19] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," in *Proc. Speech Prosody*, Boston, MA, USA, May/Jun. 2016, pp. 84–88.
[20] S. D. H. Permana and K. B. Y. Bintoro, "Implementation of constant-Q transform (CQT) and MEL spectrogram to converting bird's sound," in *Proc. IEEE Int. Conf. Commun., Netw. Satell. (COMNETSAT)*, Jul. 2021, pp. 52–56.
[21] G. Liu, C. Sun, and Y. Yang, "Target feature extraction for passive sonar based on two cepstrums," in *Proc. 2nd Int. Conf. Bioinf. Biomed. Eng.*, May 2008, pp. 539–542.
[22] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 4784–4787.
[23] C. Gan, L. Wang, Z. Zhang, and Z. Wang, "Sparse attention based separable dilated convolutional neural network for targeted sentiment analysis," *Knowl.-Based Syst.*, vol. 188, Jan. 2020, Art. no. 104827.
[24] Y. Wu, Y. Yang, F. Tian, and L. Yang, "Robust target feature extraction based on modified cochlear filter analysis model," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2013, pp. 1–5.
[25] S. Woo, J. Park, and J. Lee, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2018, pp. 3–9.
[26] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.
[27] Y. Deng, L. Wang, H. Jia, X. Tong, and F. Li, "A sequence-to-sequence deep learning architecture based on bidirectional GRU for type recognition and time location of combined power quality disturbance," *IEEE Trans. Ind. Informat.*, vol. 15, no. 8, pp. 4481–4493, Aug. 2019.
[28] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS*, Dec. 2014, pp. 1–9.
[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1–4.
[30] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.
[31] M. Aqib, R. Mehmood, A. Alzahrani, I. Katib, A. Albeshri, and S. M. Altowaijri, "Smarter traffic prediction using big data, in-memory computing, deep learning and GPUs," *Sensors*, vol. 19, no. 9, p. 2206, May 2019.
[32] N. Vreb, "Emotion classification based on convolutional neural network using speech data," in *Proc. 42th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. (MIPRO)*, May 2019, pp. 1007–1012.
[33] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in *Proc. 2nd Int. Conf. Commun. Electron. Syst. (ICCES)*, Oct. 2017, pp. 333–336.
[34] M. Seo and M. Kim, "Fusing visual attention CNN and bag of visual words for cross-corpus speech emotion recognition," *Sensors*, vol. 20, no. 19, p. 5559, Sep. 2020.
[35] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.
[36] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019.
[37] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.
[38] Z. F. Hu, L. Wang, Y. Luo, Y. L. Xia, and H. Xiao, "Speech emotion recognition model based on attention CNN Bi-GRU fusing visual information," *Eng. Lett.*, vol. 30, no. 2, pp. 427–432, Jun. 2022.

ZHANGFANG HU received the master's degree from the University of Electronic Science and Technology, Sichuan, China, in 1994. She was a Visiting Scholar with Zhejiang University, China. She is currently a Professor with the College of Photoelectrics, Chongqing University of Posts and Telecommunications (CQUPT). Her research interests include photoelectric sensing, photoelectric information processing, and speech processing.

KEHUAN LINGHU received the B.S. degree from the Chongqing University of Science and Technology, Chongqing, China, in 2021. She is currently pursuing the master's degree with the College of Photoelectrics, Chongqing University of Posts and Telecommunications (CQUPT). Her current research interests include computer vision, speech emotion recognition, and pattern recognition.

HONGLING YU received the B.S. degree from the Chongqing University of Arts and Sciences, Chongqing, China, in 2021. She is currently pursuing the master's degree with the College of Photoelectrics, Chongqing University of Posts and Telecommunications (CQUPT). Her current research interests include computer vision, image processing, and visual tracking.

CHENZHUO LIAO received the B.S. degree from the College of Photoelectrics, Chongqing University of Posts and Telecommunications (CQUPT). His current research interests include pattern recognition and data processing.