

Received November 21, 2020, accepted December 14, 2020, date of publication December 25, 2020, date of current version January 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047395

# Attention-Based Convolution Skip Bidirectional Long Short-Term Memory Network for Speech Emotion Recognition

HUIYUN ZHANG<sup>1,2,3</sup>, HEMING HUANG<sup>1,2,3</sup>, AND HENRY HAN<sup>1,4</sup>

<sup>1</sup>School of Computer Science, Qinghai Normal University, Xining 810008, China

<sup>2</sup>Key Laboratory of Tibetan Information Processing, Ministry of Education, Xining 810008, China

<sup>3</sup>Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province, Xining 810008, China

<sup>4</sup>Department of Computer and Information Science, Fordham University, New York City, NY 10023, USA

Corresponding author: Heming Huang (huanghm@qhnu.edu.cn)

This work was supported in part by the National Science Foundation of China (NSFC) under Grant 62066039.


**ABSTRACT** Speech emotion recognition is a challenging task in natural language processing. It relies heavily on the effectiveness of speech features and acoustic models. However, existing acoustic models may not handle speech emotion recognition efficiently for their built-in limitations. In this work, a novel deep-learning acoustic model called attention-based skip convolution bi-directional long short-term memory, abbreviated as SCBAMM, is proposed to recognize speech emotion. It has eight hidden layers, namely, two dense layers, convolutional layer, skip layer, mask layer, Bi-LSTM layer, attention layer, and pooling layer. SCBAMM makes better use of spatiotemporal information and captures emotion-related features more effectively. In addition, it solves the problems of gradient exploding and gradient vanishing in deep learning to some extent. On the databases EMO-DB and CASIA, the proposed model SCBAMM achieves an accuracy rate of 94.58% and 72.50%, respectively. As far as we know, compared with peer models, this is the best accuracy rate.

**INDEX TERMS** Emotion recognition, attention mechanism, weighted pooling, skip connection.

## I. INTRODUCTION

The emotional state is an important element in the interactions of human beings. It influences many aspects of communication such as facial expressions, voice characteristics, and semantic contents [1]. As we all know, emotion is an inseparable component of speech, and it plays an important role in recognizing, interpreting, and responding to the emotions expressed in speech for a human-machine interface [2]. Therefore, speech emotion recognition (SER) is an essential component in natural language processing (NLP). SER consists of the following main steps: corpus construction, signal preprocessing, feature extraction, and acoustic modeling, etc. [3]. Among which, the acoustic model is the core component of an SER system. It deciphers the relationship between an audio input signal and linguistic elements through knowledge discovery models.

Traditionally, emotional features are input into the model, and the recognition results are obtained through various

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang .

acoustic models such as hidden Markov models (HMM) [4], Gaussian mixture models (GMM) [5], support vector machines (SVM) [6], and so on. HMM is a parametric representation of time-varying features that simulate human language processing and needs a large number of samples for time-consuming training [7]–[9]. GMM is a probability density estimation model that can fit all probability distribution functions, but it depends heavily on data and it is sensitive to data noise [10]–[12]. SVM maps the feature vectors from input space to a high-dimensional Hilbert space by using kernel tricks at first and then seeks an optimal hyperplane in the high-dimensional space to classify samples. But it cannot solve the problems of large-scale training samples that lead to a large or prohibitively huge kernel matrix [13]–[16].

With the rise of deep learning, a variety of artificial neural networks (ANNs) [17] are introduced for acoustic modeling. Compared with traditional methods, these neural networks based on deep learning have better performance for their capabilities in learning when handling large scale data. However, different deep learning acoustic models have their own pros and cons. For example, recurrent neural networks (RNN)

are good at dealing with time series information [18]–[20], convolution neural networks (CNN) do well in capturing spatial information [21]–[23], and the deep residual network (DRN) can tackle the problems of gradient exploding or gradient vanishing which become popular with the deepening of the network layers [24]–[26]. Some representative deep learning acoustic models are summarized in detail as follows.

RNN is normally used as a dynamic model for sequential input, whose output is related not only to the current input but also to the output of the previously hidden layer. RNN can successfully predict the subsequent information when the context length is small. However, it may not predict well due to the problem of gradients vanishing or exploding caused by its training algorithm BPTT (Back Propagation through Time) [27]–[30].

To solve the problems of gradient exploding or gradient vanishing, long short-term memory (LSTM) is used as the basic recurrent unit of RNN and it uses memory cells and gates to control whether the input information is to be memorized, output, or forgotten [31]–[34].

LSTM only makes good use of information of the previous time step, in contrast, Bi-LSTM (Bidirectional LSTM) presumes that the state of the current time step relies not only on the information of the previous time step, but also on that of the future time step. It enables the network to make full use of context information and make more accurate judgments. This presumption, however, makes the network focus mainly on memorizing a large amount of input information and weakens its modeling capability [35], [36]. To make up for this deficiency, skip connections [37]–[39], the core technique of DRN [40], is introduced especially for deeper Bi-LSTM networks, because each neuron node in the skip connections makes use of the information of the previous hidden layer and enhances the modeling ability of the network.

Furthermore, Bi-LSTM cannot deal with spatial information in emotion recognition and its computation is more complicated. These problems are handled well by introducing convolution and pooling, the core operations of CNN [41]–[45].

Some other techniques are proposed to handle the challenges that may affect the recognition accuracies of acoustic models based on deep learning. For example, the masking operation is introduced to reduce the amount of calculation [46]–[49]. Similarly, weighted pooling based on attention over time is proposed to tackle the problems caused by a long silence, pause, or non-speech filler of the input voice, because it focuses mainly on specific regions of a speech signal that are more emotionally salient [50]–[52].

Given all that, a novel acoustic model SCBAMM is proposed to handle the challenges in speech emotion recognition. It has eight hidden layers, namely, two dense layers, convolutional layer, skip layer, mask layer, Bi-LSTM layer, attention layer, and pooling layer. This novel model makes good use of spatiotemporal information and captures emotion-related features effectively. In addition, it, to some extent, solves the problems of gradient exploding and gradient vanishing

in deep learning. It demonstrates its superiority to the peer models on the EMO-DB [53] and CASIA [54] corpus.

The remaining of the paper is organized as follows. Section 2 describes the details of the proposed model. Section 3 describes the experimental results. Section 4 discusses future research directions and concludes this study.

## II. METHODS

The development path of the proposed models will be unveiled in this section in the sequence of CBAM, SCBAM, and SCBAMM.

### A. CBAM: ATTENTION-BASED CONV-BiLSTM

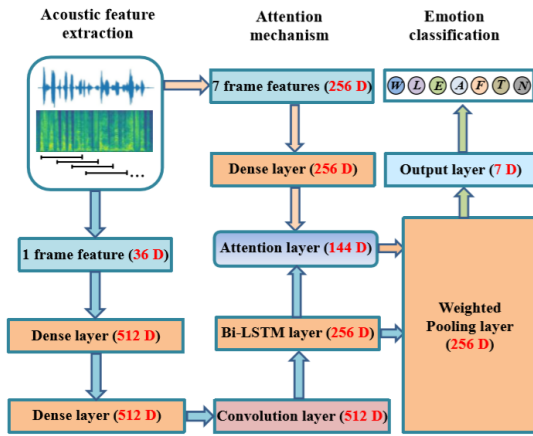
The visual attention mechanism is a special brain signal processing mechanism of human vision. Human vision scans the global image quickly and then obtains a target area. To obtain more detailed information, more attention would be invested in the target area. In the meantime, it suppresses other useless or irrelevant information [55]–[60].

The attention mechanism of deep learning, first proposed by DeepMind for image classification, is similar to that of human vision [61]. It enables the neural network to focus more on the relevant parts of the input and less on the irrelevant parts. Since then, the attention mechanism is widely used in many NLP fields, especially, in speech emotion recognition to extract features [62].

To extract the temporal information of speech more effectively, Bi-LSTM is first introduced because it can simultaneously use the information of previous time and future time. CNN is then used to extract the spatial information of speech signals. Furthermore, the attention mechanism is employed to select the features that can best represent emotions.

Based on the above analysis, a model called attention-based Conv-BiLSTM, abbreviated as CBAM, is developed. Figure 1 depicts the flow chart of the proposed CBAM. There are six hidden layers, namely, two dense layers, a convolutional layer, a Bi-LSTM layer, an attention layer, and a weighted pooling layer. The convolutional layer is used to extract spatial information, the Bi-LSTM layer is used to extract contextual information, the attention mechanism is employed to learn the weights of each time sequence, and then the weighted pooling is computed as the representation of the whole utterance. In this way, the proposed model CBAM can learn to assign weights to different time steps from data and it is especially efficient in emotion recognition. A *Softmax*( $\cdot$ ) function is finally employed to classify emotions based on the fused features of the output of the CBAM. The model parameters are optimized by minimizing the cross-entropy loss objective function. The following subsections present the details of the proposed CBAM model layer by layer.

The input layer receives features of speech frames. In this study, they are 36 dimensional acoustic features including 34-dimensional spectral features, 1-dimensional pitch, and 1-dimensional harmonic to noise ratio (HNR).



**FIGURE 1.** The CBAM network topology that consists of six hidden layers, namely, two dense layer, a convolutional layer, a Bi-LSTM layer, an attention layer, and a weighted pooling layer.

The first hidden layer of the CBAM model is a dense layer, its output is recorded as  $h^1$  and it is calculated as:

$$h^1 = f \left( \sum_{i=1}^d w^1 x + b \right), \quad (1)$$

where  $b = [b_1, b_2, \dots, b_{36}]$  is the bias,  $x$  is the 36-dimensional feature vector ( $d = 36$ ).

$$x = [x_1, x_2, \dots, x_{36}]^T, \quad (2)$$

$w_{ij}^1$  is the element of the weight matrix  $w^1$  and it represents the weight of  $i$ -th node of input layer connected to  $j$ -th node of the first dense layer (512 nodes), where  $i = 1, 2, \dots, 36$  and  $j = 1, 2, \dots, 512$ . The matrix  $w^1$  is defined as:

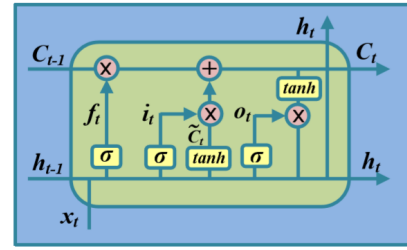
$$w^1 = [w_{ij}^1]_{i \times j}^T \quad (3)$$

$f(\cdot)$  is a LeakyReLU activation function and it is defined as:

$$f(x) = \begin{cases} x, & x \geq 0 \\ \alpha x, & x < 0, \end{cases} \quad (4)$$

where  $\alpha$  is a hyperparameter. When  $\alpha = 0$ , it is the ReLU function. For negative input, both the output of ReLU and its first derivative is always 0, which makes the neuron unable to update the parameters. When  $\alpha > 0$ , it is the LeakyReLU. For negative input, both the output of LeakyReLU and its first derivative is non-zero, which solves the gradient problems in deep learning to some extent and solves the problem that neurons do not learn when the ReLU function enters the negative interval. The value of  $\alpha$  in this paper is 0.01.

In this case, the obtained  $h^1$  is used as the input of the next dense layer. The calculation of  $h^2$  is similar to that of  $h^1$ , see formula (1). Similarly, the calculated  $h^2$  is used as the input of the convolution layer. Valid convolution only considers the case that the length of a one-dimensional tensor can completely cover the convolution kernel, that is, the convolution kernel moves inside the one-dimensional tensor. The output



**FIGURE 2.** The LSTM cell.

$conv^3$  of the valid convolution is input to the Bi-LSTM layer, and it is defined as:

$$conv^3 = f \left( \frac{h^2 * F}{S} \times N \right), \quad (5)$$

where  $F = [k_1, k_2, \dots, k_{512}]$  represents the convolution kernel,  $N$  is the number of filters and it is set 512, and  $S$  represents the stride and it is set 1.

To the Bi-LSTM layer, it has three inputs: the first one is  $conv^3$ , which comes from lower layer at the current time  $t$ ; the second one is  $h_{t-1}$ , which is the output of the same hidden state at time  $t-1$ ; and the third one is  $h_{t+1}$ , which is the output of the same hidden state at time  $t+1$ .

The gating mechanism of memory cell is used to control information flow. Figure 2 shows the LSTM cell. There is a cell state  $C_t$  to memorize information and it is updated as:

$$C_t = C_{t-1} \odot f_t + \tilde{C}_t \odot i_t, \quad (6)$$

where  $\odot$  represents Hadamard product,  $C_{t-1}$  represents the cell state of previous time series.  $f_t$  is the output of forgetting gate at time  $t$  and calculated as:

$$f_t = \sigma \left( W_f h_{t-1} + U_f o_{conv}^3 + b_f \right). \quad (7)$$

It represents the forgetting probability of the hidden state of the previous time sequence.

The output of  $f_t$  is a three-dimensional array, where the first element, which is set to 32, represents the dimension of the batch size vector; the second element, which is set to 144, represents the dimension of the time step vector; and the third element, which is set to 128, represents the number of hidden states. In the following formula, the dimensions of  $i_t, \tilde{C}_t, o_t$  are equal to that of  $f_t$ .

The  $U = (r_{ij})_{m \times n}$  represents the weight matrix between the convolution layer and Bi-LSTM cell states, where  $m = 1, 2, \dots, 512, n = 1, 2, \dots, 128$ . The dimensions of matrices  $U_f, U_i, U_{c''},$  and  $U_o$  are the same as those of  $U$ , where  $U_f$  is the forgetting weight matrix,  $U_i$  and  $U_{c''}$  are the input weight matrices, and  $U_o$  is the output weight matrix.

The symbol  $W_{n \times n}$  represents the weight matrix between the hidden states at adjacent time steps. The dimensions of matrices  $W_f, W_i, W_c,$  and  $W_o$  all are equal to that of  $W_{n \times n}$ , where  $W_f$  is the connection weight matrix between the former hidden state and current time forgetting gate,  $W_i$  is the connection weight between the former hidden state and the

current time input gate,  $W_c$  is the connection weight between the former hidden state and the current time cell state, and  $W_o$  is the connection weight between the former hidden state and the current time output gate.

The parameter  $h_{t-1}$  in equation (7) is a 128-dimensional vector of the hidden state,  $b_f = [b_f^1, b_f^2, \dots, b_f^{128}]$  is the bias, and  $\sigma(\cdot)$  is a Sigmoid function defined as:

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \in [0, 1] \quad (8)$$

The inputting gate, responsible for processing the input information of the current sequence position, consists of two parts:  $i_t$  and  $\tilde{C}_t$ , and they are multiplied to update the cell state. The parameter represents the output of the activation function Sigmoid:

$$i_t = \sigma(W_i h_{t-1} + U_i o_{conv}^3 + b_i), \quad (9)$$

Correspondingly, the parameter  $C''$  represents the output of the activation function  $\tanh$ :

$$\tilde{C}_t = \tanh(W_c h_{t-1} + U_c o_{conv}^3 + b_c), \quad (10)$$

where

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \in [-1, 1], \quad (11)$$

The update of the hidden state  $h_t$  is the Hadamard product of  $o_t$  and  $h_{t-1}$ , that is:

$$h_t = o_t \odot \tanh(C_t), \quad (12)$$

where the dimensions of  $h_t$  equal to that of  $h_{t-1}$  and  $o_t$  is computed as:

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o), \quad (13)$$

Finally, the output  $y_o^t$  of the current sequence is calculated as:

$$y_o^t = \sigma(V h_t + b_t), \quad (14)$$

where  $b_t = [b_t^1, b_t^2, \dots, b_t^{128}]$  is the bias vector, and  $V$  is the connection weight matrix between the cell hidden state and output that has the same dimensions as  $W_{n \times n}$ . Similarly, the dimensionality of  $y_o^t$  is the same as that of  $f_t$ .

Because Bi-LSTM processes a sequence of information from both forward and backward directions at the same time, the final output  $y_B$  of Bi-LSTM, which is also the input of the attention layer, is a three-dimensional array, where the first element, which is set to 32, represents the dimension of the batch size vector; the second element, which is set to 1024, represents the dimension of the time step vector; and the third element, which is set to 256, represents the number of hidden states.

In the attention layer,  $\text{Softmax}(\cdot)$  is used to learn the attention parameters of an input frame feature. It computes the final weights for the frames which sum to unity.  $u$  represents a 256-dimensional vector calculated as:

$$u = \text{Softmax}(W_A i_A + b_A), \quad (15)$$

where  $i_A$ , as the input of attention layer, is a two-dimensional array: the first element, which is set to 32, represents the dimension of the batch size, and the second element, which is set to 256, represents the dimension of the time step. In addition,  $W_A = [w_{ij}]_{2m \times 2n}$  is the weight matrix and  $b_A = [b_1, b_2, \dots, b_{2n}]$  is the bias vector.  $\text{Softmax}(\cdot)$  is an activation function which maps the original output to the interval (0,1) and the sum of the values is 1. It could be understood as the probability and the node with the highest probability should be selected as the focus of attention.  $\alpha$  is the probability of the sequence features passing through the attention layer, which is calculated as:

$$\alpha = \text{Softmax}(u \cdot y_B), \quad (16)$$

where  $\cdot$  represents the dot product operation. It is noted to take the last dimension of  $u$  and  $y_B$  for the dot product operation to calculate the probability through the  $\text{Softmax}(\cdot)$  function. The vector corresponding to the maximum probability is the target of attention mechanism that has the same dimension as that of  $i_A$ .

In the weighted pooling layer, in order to get the utterance-level representation  $z_p$ , the weighted pooling operation is performed on the sentence and take the value on the horizontal axis of  $\alpha$  and  $y_B$  for the dot product operation, that is

$$z_p = \alpha \cdot y_B. \quad (17)$$

On the top of the CBAM model, there is an output layer and it calculates the probability through  $\text{Softmax}(\cdot)$  function to perform classification:

$$y_{nk} = \text{Softmax}(z_p). \quad (18)$$

To find the optimal weight and bias, the cross-entropy loss function is employed to train the CBAM network. The cross-entropy  $L_{CE}$  is calculated as:

$$L_{CE} = -\frac{1}{N} \sum_n \sum_k t_{nk} \log y_{nk}, \quad (19)$$

where  $N$  denotes the total number of samples,  $n$  denotes the  $n$ -th sample,  $k$  denotes the  $k$ -th class,  $t_{nk}$  denotes the label of sample. It is worthwhile to point out that  $t_{nk}$  denotes the ground probability of the  $n$ -th sample belongs to the class  $k(k=0,1,2,\dots)$ . In addition,  $y_{nk}$  is the output of the neural network and represents the predicted probability of the  $n$ -th sample belonging to the class  $k$ .

### B. SCBAM: CBAM WITH SKIP CONNECTIONS

The CBAM network focuses on memorizing a large amount of input information. By adding a skip connection [37] between the first hidden layer and the convolution layer, a new model called SCBAM is developed to enhance the modeling capability of deep learning networks in this study. Figure 3 illustrates its topology. The skip connection introduced in this model makes the network focus not only on memorizing a large amount of input information but also target to improve the modeling ability.

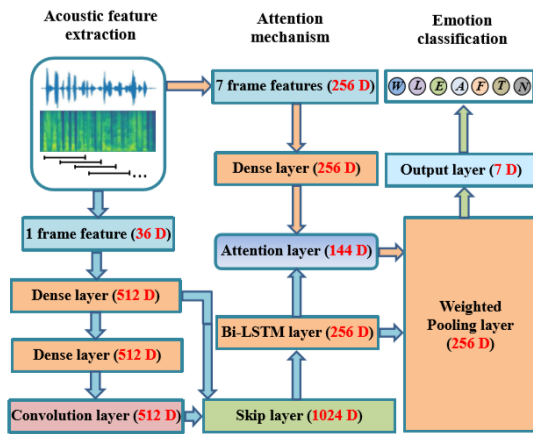


FIGURE 3. The SCBAM network topology. Compared with CBAM model, SCBAM model has a skip layer.

Furthermore, the SCBAM network can avoid the gradient exploding or gradient vanishing problems as the network is deepened. The reason behind this is that SCBAM fuses the feature vectors  $conv^3$  and  $h^1$ , where  $conv^3$  represents the feature vector extracted from the convolution layer and  $h^1$  represents the feature vector extracted from the dense layer. The fused feature  $F_c$  is calculated as:

$$F_c = concatenate(conv^3, h^1), \quad (20)$$

where the  $concatenate(\cdot)$  function concatenates the two features. The dimension of  $F_c$  is the sum of the dimensions of  $conv^3$  and  $h^1$  because of the concatenation. That is, there are 1024 neuron nodes in the skip layer. The calculation procedures of other layers in the SCBAM network are exactly as same as those in the CBAM network. Furthermore, in implementation, both CBAM and SCBAM networks employ the LeakyReLU [63] activation function and RMSprop [64] optimizer.

### C. SCBAMM: SCBAM WITH MASKING OPERATIONS

The SCBAM network focuses on the specific region of a speech signal that is emotionally salient. To extract the features of the target region more effectively, a mask layer [49] is added between the convolution layer and the Bi-LSTM layer of SCBAM to build a new model named SCBAMM.

Figure 4 illustrates the system diagram of SCBAMM. The function of masking operation is to extract the features of the interest region. These features are obtained by multiplying the feature mask of interest with the features to be processed. The inner image value of the interest region remains unchanged while the outer value is 0.

When inputting the sample features, 0s are padded to align the dimensions of all the sample features. To the Bi-LSTM network model, all 0s in  $F_c$  need to be masked, that is, all 0s in  $F_c$  do not participate in the calculation. The mask operation  $y_m$  can be represented as:

$$y_m = Mask(F_c, 0) \quad (21)$$

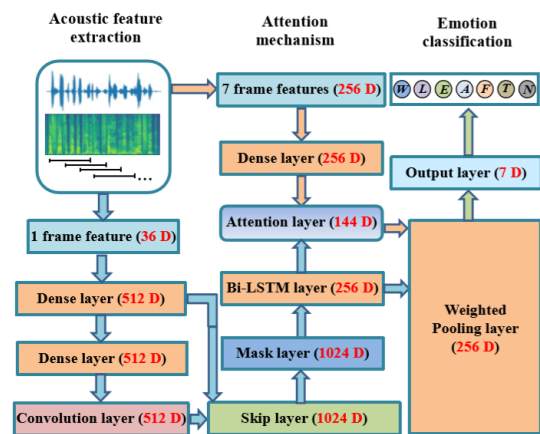


FIGURE 4. The SCBAMM network topology. Compared with SCBAM model, SCBAMM model has a mask layer.

where  $(F_c, 0)$  means all 0s in  $F_c$  and do not need to be calculated. Similarly, the calculation procedures of other layers in the SCBAMM network are exactly as same as those in the SCBAM network.

To prevent possible data over-fitting [65], during the training stage of CBAM, SCBAM, and SCBAMM, dropout [66] is implemented in all layers but the attention layer and weighted pooling layer. The dropout rate is set to be 0.1 generally unless specified. At the same time, the batch size is assigned 32, the number of cross-validation is assigned 10, and the epoch is set as 100. In addition, the optimizer and the activation functions are the RMSprop and LeakyReLU, respectively.

### III. EXPERIMENTAL RESULTS

The performances of the proposed CBAM, SCBAM, and SCBAMM are validated on the EMO-DB corpus [54] and CASIA corpus [55].

EMO-DB is a German emotion database made up of 10 actors (5 males and 5 females) to simulate 7 classes of emotions, namely, anger (W), boredom (L), disgust (E), fear (A), joy (F), sadness (T), and neutral (N). The sample numbers of these classes are 127, 81, 46, 69, 71, 62, and 79, respectively. Totally, the corpus contains 535 emotional speech sentences with a sampling rate of 48-kHz and 16-bit quantification. Randomly, one male and one female are selected as testing subjects. The data from other subjects is used as validation data to check if the system needs to be stopped as soon as possible. The 36-dimensional feature vector consists of 34D magnitude FFT vectors, harmonic to noise ratio (HNR), and pitch (F0). The feature extraction is performed within a 25ms window with a shifting step size of 10ms. The acoustic feature sequence is Z-normalized within each utterance [54].

The CASIA speech emotion database was recorded by the Institute of Automation, Chinese Academy of Sciences. It is recorded by actors (2 men and 2 women) in six different

TABLE 1. The parameters of the proposed models.

Layer	Nodes	Shape
Input layer	36	32 × 1024 × 36
Dense layer	512	32 × 1024 × 512
Dense layer	512	32 × 1024 × 512
Convolution layer	512	32 × 1024 × 512
Skip layer	1024	32 × 1024 × 1024
Mask layer	1024	32 × 1024 × 1024
BiLSTM layer	256	32 × 144 × 256
Attention layer	144	32 × 144
Pooling layer	256	32 × 256
Output layer	7	32 × 7

emotions, namely, anger (A), fear (F), happy (H), neutral (N), sad (Sa), surprise (Su). The signal-to-noise ratio (SNR) is about 35 dB, and data acquisition is complemented in a pure recording environment with 16bit quantization and 16KHz sampling rate. The publicly available CASIA dataset contains 1200 utterances; each actor speaks 300 words in the same text, and each person recites six emotions. The average length of an audio file is about 1.9s [55]. The 20-dimensional MFCC features are extracted, and the high-level statistical functions, namely mean, variance, and maximum, of the MFCC features are calculated. The feature extraction is performed within a 25ms window with a shifting step size of 10ms. The acoustic feature sequence is Z-normalized within each utterance [54].

The experiments are conducted on a powerful PC with 64G RAM running under Windows 10, the benchmark speed of the CPU is 2.10 GHz, the core is 40, the logic processor is 80, and two RTX 2080 Ti GPUs are also employed for calculation speedup.

The CBAM, SCBAM and SCBAMM architectures are implemented with TensorFlow toolkit. The parameters of the proposed models are shown in Table 1. The optimizer is Rmsprop and the initial learning rate is set to 0.001. When training the neural network, if the learning rate is very large, it is likely that more neurons in the network are ‘dead’, and LeakyReLU retains some values of the negative axis so that all information of the negative axis will not be lost. Thus, the network can be better trained. Rmsprop with bias correction accelerates convergence and decreases possible oscillations in training [64]. It is more robust when the gradient becomes sparser.

The confusion matrix and five evaluation measures are employed to evaluate the performances of each model. In a confusion matrix, each row represents the prediction categories of each emotion, each column represents the actual categories of each emotion, and each number on the diagonal indicates the correct number of identified samples. The five evaluation measures include accuracy, precision, weighted average recall, unweighted average recall (UAR), and F1-score, respectively. The accuracy refers to the probability of correct predictions among predictions, i.e., the proportion of correct predictions. The precision represents the number of positive samples predicted to be positive; the recall evaluates how many positive samples in the total samples are predicted correctly. F1-score is the weighted average of recall rate and

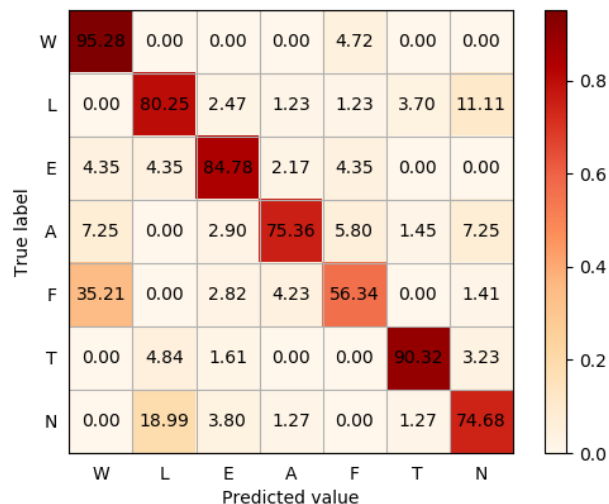


FIGURE 5. Confusion matrix of CBAM on EMO-DB dataset.

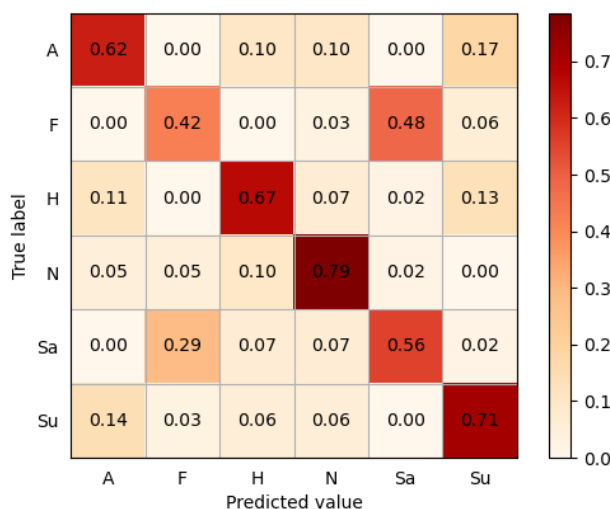
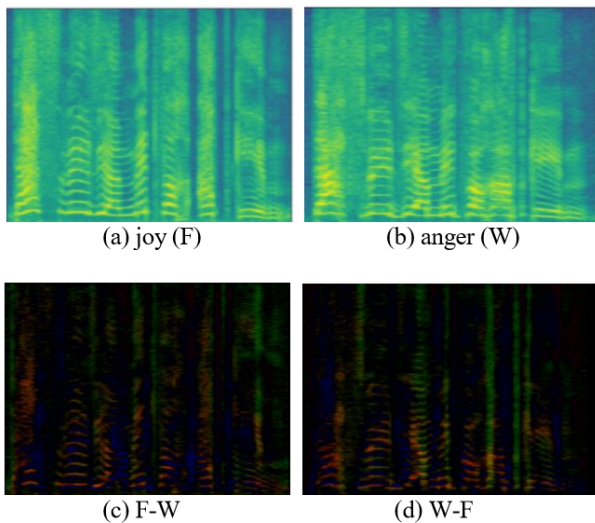


FIGURE 6. Confusion matrix of CBAM on CASIA dataset.

precision rate. In the case of imbalanced data, the recall rate can be biased, therefore, to evaluate the experimental performance comprehensively, both weighted average recall (WAR) and unweighted average recall (UAR) are used.

A. PERFORMANCE OF CBAM

Confusion matrix and 10-fold cross-validation are employed to verify the performance of CBAM. Figure 5 and Figure 6 are the best confusion matrices of CBAM on the databases EMO-DB and CASIA, respectively. It can be seen that: Firstly, the average accuracy rates are 80.75% on the EMO-DB and 63.33% on the CASIA, respectively. Secondly, 95.28% of anger (W) samples are predicted correctly on the EMO-DB dataset, which is a very considerable recognition result. Thirdly, there is a situation where one class is easily predicted to be another. Take joy (F) emotion as an example, only 56.34% of its samples are identified correctly, and 35.21% of its samples are predicted to anger (W), and 4.23% of its samples are predicted to fear (A), etc. Finally, W and



**FIGURE 7.** The emotion spectrograms. (a) The spectrogram of joy (F); (b) The spectrogram of anger (W); (c) The spectrogram of F-W; and (d) The spectrogram of W-F.

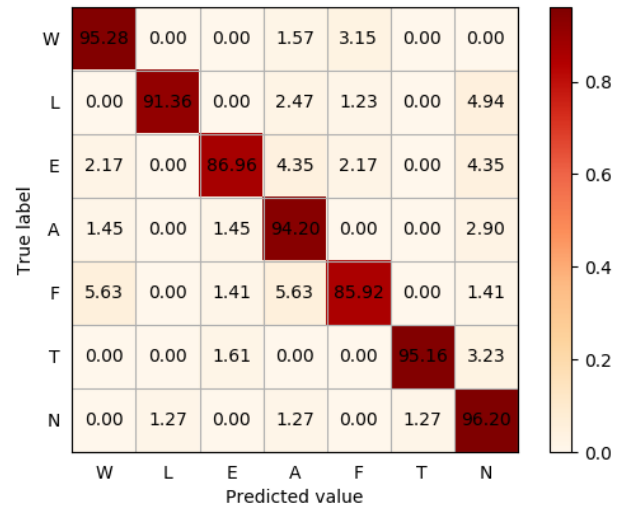
F, L and N, are easily confusing emotion class pairs. Here, the spectral subtraction of two samples is used to demonstrate their similarity more intuitively. Figure 7 shows the spectrograms of W, F, F-W, W-F. It can be found that the spectrograms of joy (F) and anger (W) are similar. F-W and W-F reflect the difference between F and W. The darker the spectrogram of F-W or W-F, the more similar are classes F and W.

**B. PERFORMANCE OF SCBAM**

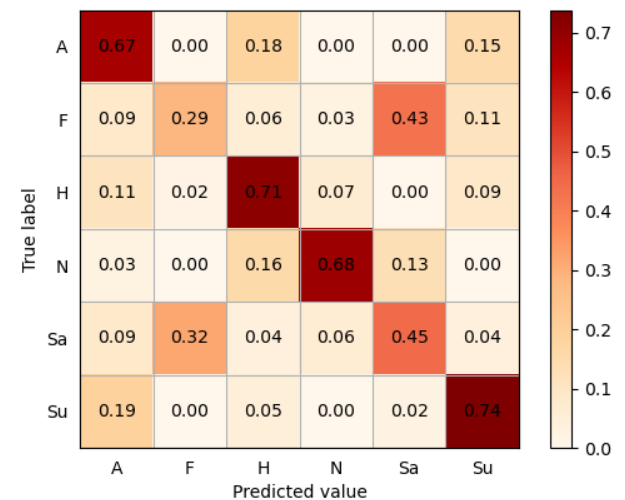
Figures 8 and 9 show us the best confusion matrices of SCBAM on the databases EMO-DB and CASIA. It can be seen that: Firstly, the average accuracy rates are 92.71% on the EMO-DB and 70.00% on the CASIA under the same computing environment of CBAM. Secondly, the accuracy of SCBAM is 11.96% higher than that of CBAM on the EMO-DB dataset. The reasons behind that are as follows. The skip connections in SCBAM make the network focus not only on memorizing a large amount of input information, but also on the promotion of the modeling ability; in addition, it can deal with the problems of the gradient exploding or gradient vanishing. Thirdly, except the classes of joy (F) and disgust (E), the samples of the other five types of emotions can be well recognized. 5.63% of joy (F) samples are predicted to be another. Once again, it is proved that the samples of class F are easily predicted as class W.

**C. PERFORMANCE OF SCBAMM**

Figure 10 and Figure 11 illustrate the best confusion matrices of SCBAMM on the databases EMO-DB and CASIA. It is obvious that: Firstly, the classification accuracy of the proposed model is 94.58% on EMO-DB and 72.90% on CASIA under the same computing environment of CBAM and SCBAM. Secondly, the accurate rate of SCBAMM for



**FIGURE 8.** Confusion matrix of SCBAM on EMO-DB dataset.



**FIGURE 9.** The confusion matrix of SCBAM on the CASIA dataset.

each kind of emotion reaches 90.00% on the EMO-DB dataset, which indicates that the SCBAMM model has good robustness. Finally, the accuracy rate of SCBAMM is 13.83% and 1.87% higher than that of CBAM and SCBAM, respectively. The reason is that the masking operation in SCBAMM is good at extracting the effective features of the target regions, which contributes to detecting different emotion states.

**D. COMPARISON OF CBAM, SCBAM, AND SCBAMM**

Figure 12 shows the improvements in terms of accuracy of the proposed models CBAM, SCBAM and SCBAMM in the 10-fold cross-validation on the EMO-DB database. It is easy to come to the following conclusions. Firstly, the average accuracy of SCBAMM is optimal (orange squares in the box) among that of CBAM, SCBAM, and SCBAMM. Secondly, the results obtained by SCBAMM

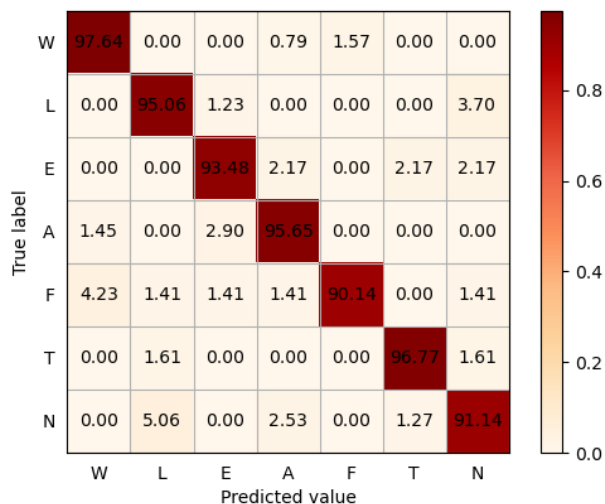


FIGURE 10. The confusion matrix of SCBAMM on the EMO-DB dataset.

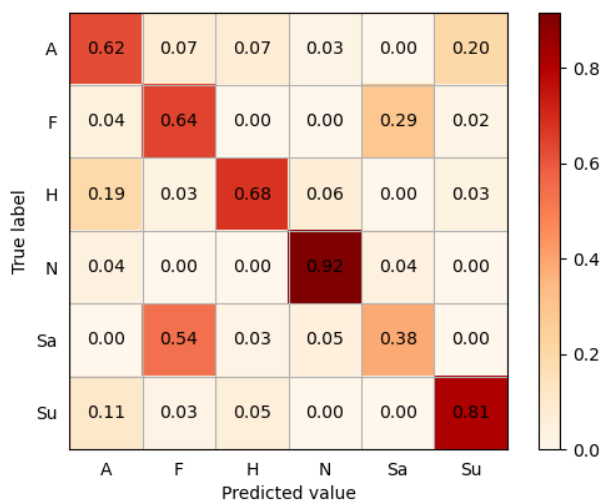


FIGURE 11. The confusion matrix of SCBAMM on the CASIA dataset.

in the 10-fold cross-validation are relatively concentrated (longitudinal height of the box), which indicates that SCBAMM has better stability and robustness. Finally, red solid circles indicate outliers.

Table 2 summarizes the improvements in terms of accuracy of the proposed models CBAM, SCBAM and SCBAMM to the peer models.

Firstly, SCBAMM is superior to SCBAM and CBAM on both datasets EMO-DB and CASIA in evaluation measures such as accuracy, UAR, precision, and F1-score.

Secondly, SCBAMM is superior to previous research results other than reference [75] on the EMO-DB dataset, no matter which evaluation index is measured. The accuracy rate of reference [75] is as high as 98.00%. The reason behind that is reference [75] just selects a subset of the EMO-DB database, which contains only four types of emotions, each containing 30 emotional sentences.

Thirdly, SCBAMM demonstrates a strong prediction capability in emotion recognition. Mathematically, the weight

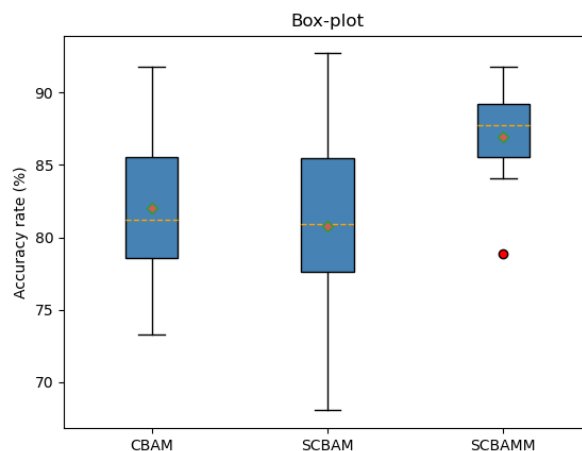


FIGURE 12. The boxplots of the classification accuracies of the CBAM, SCBAM, and SCBAMM models on the EMO-DB dataset under the 10-fold cross-validation.

TABLE 2. Performance comparisons (%) of the proposed models to those of other peer models on the EMO-DB corpus.

Datasets	Models	Acc	WAR	UAR	Prec	F1-score
EMODB	SVM[67]	/	81.74	/	/	/
	GerDA[68]	84.60	/	/	/	/
	RDBN[69]	82.32	81.64	80.52	81.24	80.88
	SVM[70]	/	88.17	/	/	/
	ACRNN[29]	82.43	82.68	82.82	83.53	83.17
	SVM[71]	89.90	/	/	/	/
	SVM[72]	88.60	88.61	87.83	89.07	88.45
	HMM[73]	78.40	/	/	/	/
	UAE[74]	/	/	62.00	/	/
	SVM[75]	98.31	/	/	/	/
	PCRN[76]	/	86.44	84.53	/	/
	CBAM	80.75	81.61	79.57	82.07	80.79
SCBAM	92.71	93.66	92.15	92.75	92.45	
SCBAMM	97.58	93.32	92.30	94.00	93.14	
CASIA	GABEL[77]	/	38.55	38.55	/	/
	SVM[67]	/	43.50	43.50	/	/
	SVM[69]	/	48.50	48.50	/	/
	SVM[76]	/	58.25	58.25	/	/
	CBAM	63.55	59.00	59.00	58.30	58.65
	SCBAM	70.00	62.83	62.83	62.47	62.65
	SCBAMM	72.50	67.50	67.50	71.36	69.37

matrix of SCBAMM can be more representative and faster than those of the peer models for its customized optimizations

Firstly, SCBAMM is superior to SCBAM and CBAM on both datasets EMO-DB and CASIA in evaluation measures such as accuracy, UAR, precision, and F1-score.

Secondly, SCBAMM is superior to previous research results other than reference [75] on the EMO-DB dataset, no matter which evaluation index is measured. The accuracy rate of reference [75] is as high as 98.00%. The reason behind that is reference [75] just selects a subset of the EMO-DB database, which contains only four types of emotions, each containing 30 emotional sentences.

Thirdly, SCBAMM demonstrates a strong prediction capability in emotion recognition. Mathematically, the weight matrix of SCBAMM can be more representative and faster than those of the peer models for its customized optimizations



#### IV. CONCLUSION AND FUTURE WORKS

SCBAMM, a novel acoustic model based on deep learning, is proposed for speech emotion recognition. To achieve better performance, several techniques, namely, attention mechanism, skip connection, mask operation, and integration of spatial and time series information all are proposed. It demonstrates obvious advantages over the peer models on the benchmark datasets EMO-DB and CASIA. Experimental results suggest that SCBAMM seems to be much fitter for emotion recognition than its peers. The reason behind that is, SCBAMM makes good use of spatiotemporal information and captures emotion-related features effectively.

It can be interesting for us to further prove the superiority of the proposed model from machine learning theory. For example, it is highly likely that the weight matrix sequence in SCBAMM learning can be sparse and meaningful than those of its peer model. Such a study would be useful to know whether the specific operations proposed in SCBAMM would be redundant for some special datasets (e.g. imbalanced data). To further verify the effectiveness of SCBAMM, it will be applied to other emotion classification databases. In addition, it will be extended to speech recognition and image classification.

#### REFERENCES

- [1] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 116–125, Jan. 2012.
- [2] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: Challenges," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 1, pp. 16–28, Jan. 2016.
- [3] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, Apr. 2019.
- [4] S. Mao, D. Tao, G. Zhang, P. C. Ching, and T. Lee, "Revisiting hidden Markov models for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6715–6719.
- [5] I. Shahin, A. B. Nassif, and S. Hamsa, "Emotion recognition using hybrid Gaussian mixture model and deep neural network," *IEEE Access*, vol. 7, pp. 26777–26787, 2019.
- [6] Z. Teng, F. Ren, and S. Kuroiwa, "Emotion recognition from text based on the rough set theory and the support vector machines," in *Proc. Int. Conf. Natural Lang. Process. Knowl. Eng.*, Aug. 2007, pp. 36–41.
- [7] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003.
- [8] M. Song, C. Chen, and M. You, "Audio-visual based emotion recognition using tripled hidden Markov model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, pp. 877–880.
- [9] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 142–156, Feb. 2012.
- [10] I. J. Tashve, Z.-Q. Wang, and K. Godin, "Speech emotion recognition based on Gaussian mixture models and deep neural networks," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2017, pp. 1–4.
- [11] J. Jiang, Z. Wu, M. Xu, J. Jia, and L. Cai, "Comparison of adaptation methods for GMM-SVM based speech emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2012, pp. 269–273.
- [12] H. K. Vydana, P. P. Kumar, K. S. R. Krishna, and A. K. Vuppala, "Improved emotion recognition using GMM-UBMs," in *Proc. Int. Conf. Signal Process. Commun. Eng. Syst.*, Jan. 2015, pp. 53–57.
- [13] H. Hu, M.-X. Xu, and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. ICASSP*, Apr. 2007, pp. 413–416.
- [14] C. Caihua, "Research on multi-modal mandarin speech emotion recognition based on SVM," in *Proc. IEEE Int. Conf. Power, Intell. Comput. Syst. (ICPICS)*, Jul. 2019, pp. 173–176.
- [15] F. Dong, G. Zhang, Y. Huang, and H. Liu, "Speech emotion recognition based on multi-output GMM and SVM," in *Proc. Chin. Conf. Pattern Recognit. (CCPR)*, Oct. 2010, pp. 1–4.
- [16] V. Fernandes, L. Mascarehnas, C. Mendonca, A. Johnson, and R. Mishra, "Speech emotion recognition using mel frequency cepstral coefficient and SVM classifier," in *Proc. Int. Conf. Syst. Model. Advancement Res. Trends (SMART)*, Nov. 2018, pp. 200–204.
- [17] X. Mao, L. Chen, and L. Fu, "Multi-level speech emotion recognition based on HMM and ANN," in *Proc. WRI World Congr. Comput. Sci. Inf. Eng.*, 2009, pp. 225–229.
- [18] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.
- [19] X. Chen, W. Han, H. Ruan, J. Liu, H. Li, and D. Jiang, "Sequence-to-sequence modelling for categorical speech emotion recognition using recurrent neural network," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact. (ACII Asia)*, May 2018, pp. 1–4.
- [20] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 190–195.
- [21] S. Zhang, A. Chen, W. Guo, Y. Cui, X. Zhao, and L. Liu, "Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition," *IEEE Access*, vol. 8, pp. 23496–23505, 2020.
- [22] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5115–5119.
- [23] S. Wang, J. Li, T. Cao, H. Wang, P. Tu, and Y. Li, "Dance emotion recognition based on laban motion analysis using convolutional neural network and long short-term memory," *IEEE Access*, vol. 8, pp. 124928–124938, 2020.
- [24] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using WaveNet encoder and residual networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 265–274, May 2019.
- [25] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6675–6679.
- [26] Z. Li, J. Li, S. Ma, and H. Ren, "Speech emotion recognition based on residual neural network with different classifiers," in *Proc. IEEE/ACIS 18th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2019, pp. 186–190.
- [27] H. Zhao, Y. Xiao, J. Han, and Z. Zhang, "Compact convolutional recurrent neural networks via binarization for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6690–6694.
- [28] 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.
- [29] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
- [30] T. Zhang and J. Wu, "Speech emotion recognition with i-vector feature and RNN model," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Jul. 2015, pp. 524–528.
- [31] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 11, pp. 1675–1685, Jul. 2019.
- [32] F. Tao and G. Liu, "Advanced LSTM: A study about better time dependency modeling in emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2906–2910.
- [33] K.-Y. Huang, C.-H. Wu, T.-H. Yang, M.-H. Su, and J.-H. Chou, "Speech emotion recognition using autoencoder bottleneck features and LSTM," in *Proc. Int. Conf. Orange Technol. (ICOT)*, Dec. 2016, pp. 1–4.
- [34] B. T. Atmaja and M. Akagi, "Speech emotion recognition based on speech segment using LSTM with attention model," in *Proc. IEEE Int. Conf. Signals Syst. (ICSigSys)*, Jul. 2019, pp. 40–44.

- [35] A. Yadav and D. K. Vishwakarma, "A multilingual framework of CNN and bi-LSTM for emotion classification," in *Proc. 11th Int. Conf. Comput., Commun. New. Technol. (ICCCNT)*, Jul. 2020, pp. 1–6.
- [36] A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia, "Facial emotion recognition using light field images with deep attention-based bidirectional LSTM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3367–3371.
- [37] J.-Y. Liu and Y.-H. Yang, "Denoising auto-encoder with recurrent skip connections and residual regression for music source separation," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 773–778.
- [38] D. Yoon, Z. Yeoh, and J. Byun, "Seismic data reconstruction using deep bidirectional long short-term memory with skip connections," *IEEE Geosci. Remote Sens. Lett.*, early access, May 25, 2020, doi: [10.1109/LGRS.2020.2993847](https://doi.org/10.1109/LGRS.2020.2993847).
- [39] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "LCSCNet: Linear compressing-based skip-connecting network for image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 1450–1464, 2020.
- [40] E. Chandra and J. Y.-J. Hsu, "Deep learning for multimodal emotion recognition-attentive residual disconnected RNN," in *Proc. Int. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Nov. 2019, pp. 1–8.
- [41] R. Taniguchi, K. Hoshiba, K. Itoyama, K. Nishida, and K. Nakadai, "Signal restoration based on bi-directional LSTM with spectral filtering for robot audition," in *Proc. 27th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2018, pp. 955–960.
- [42] 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.
- [43] Y. Cao, S. Ma, and H. Pan, "FDTA: Fully convolutional scene text detection with text attention," *IEEE Access*, vol. 8, pp. 155441–155449, 2020.
- [44] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "End-to-end continuous emotion recognition from video using 3D convlstm networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6837–6841.
- [45] E. Franti, I. Ispas, and M. Dascalu, "Testing the universal baby language hypothesis—automatic infant speech recognition with CNNs," in *Proc. 41st Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2018, pp. 424–427.
- [46] M. Kokuier and P. Jancovic, "Incorporating mask modelling for noise-robust automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 3929–3932.
- [47] L. Zao, D. Cavalcante, and R. Coelho, "Time-frequency feature and AMS-GMM mask for acoustic emotion classification," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 620–624, May 2014.
- [48] G. Zhan, Z. Huang, D. Ying, J. Pan, and Y. Yan, "Improvement of mask-based speech source separation using DNN," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Oct. 2016, pp. 1–5.
- [49] J. Barker and X. Shao, "Energetic and informational masking effects in an audiovisual speech recognition system," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 3, pp. 446–458, Mar. 2009.
- [50] A. R. Avila, J. Monteiro, D. O'Shaughnessy, and T. H. Falk, "Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2017, pp. 360–365.
- [51] Y. Zhang, Z.-R. Wang, and J. Du, "Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
- [52] A. R. Avila, Z. A. Momin, J. F. Santos, D. O'Shaughnessy, and T. H. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Trans. Affect. Comput.*, early access, 2018, doi: [10.1109/TAFFC.2018.2858255](https://doi.org/10.1109/TAFFC.2018.2858255).
- [53] M. Dan Zbancioc and S. M. Feraru, "A study about the automatic recognition of the anxiety emotional state using emo-DB," in *Proc. E-Health Bioeng. Conf. (EHB)*, Nov. 2015, pp. 1–4.
- [54] L. Chen, W. Su, M. Wu, W. Pedrycz, and K. Hirota, "A fuzzy deep neural network with sparse autoencoder for emotional intention understanding in human-robot interaction," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 7, pp. 1252–1264, Jul. 2020.
- [55] Q. Wei, G. Zhai, C. Hu, and X. Min, "Visual attention analysis and prediction on human faces with mole," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2016, pp. 1–4.
- [56] N. Li, J. Zhao, P. Jiang, and C. Li, "Medical image enhancement method based on visual attention mechanism," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 452–456.
- [57] N. Li, J. Zhao, and P. Jiang, "Fabric defects detection via visual attention mechanism," in *Proc. Chin. Autom. Congr. (CAC)*, Oct. 2017, pp. 2956–2960.
- [58] F. Guo, J. Zhao, and P. Jiang, "Target search via feature cutting strategy of visual attention mechanism," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2019, pp. 5650–5654.
- [59] S. Yoon, S. Dey, H. Lee, and K. Jung, "Attentive modality hopping mechanism for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3362–3366.
- [60] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Hierarchical attention transfer networks for depression assessment from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7159–7163.
- [61] D. Zoran, S. Chrzanowski, P.-S. Huang, S. Goyal, A. Mott, and P. Kohli, "Towards robust image classification using sequential attention models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9480–9489.
- [62] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2526–2530.
- [63] T. Jiang and J. Cheng, "Target recognition based on CNN with LeakyReLU and PReLU activation functions," in *Proc. Int. Conf. Sens., Diag., Prognostics, Control (SDPC)*, Aug. 2019, pp. 718–722.
- [64] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, "A sufficient condition for convergences of adam and RMSProp," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11119–11127.
- [65] K. Liu, J. Song, W. Zhang, and X. Yang, "Alleviating over-fitting in attribute reduction: An early stopping strategy," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit. (ICWAPR)*, Jul. 2018, pp. 190–195.
- [66] J. Wang and L. Liu, "A neural network sparseness algorithm based on relevance dropout," in *Proc. IEEE 6th Int. Conf. Ind. Eng. Appl. (ICIEA)*, Apr. 2019, pp. 480–484.
- [67] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local hu moments for speech emotion recognition," *Biomed. Signal Process. Control*, vol. 18, pp. 80–90, Apr. 2015.
- [68] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5688–5691.
- [69] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, "Random deep belief networks for recognizing emotions from speech signals," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–9, Mar. 2017.
- [70] H. Tao, R. Liang, C. Zha, X. Zhang, and L. Zhao, "Spectral features based on local hu moments of Gabor spectrograms for speech emotion recognition," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 8, pp. 2186–2189, 2016.
- [71] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs turn-level: Emotion recognition from speech considering static and dynamic processing," in *Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact.*, Sep. 2007, pp. 139–147.
- [72] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic recognition of speech emotion using long-term spectro-temporal features," in *Proc. 16th Int. Conf. Digit. Signal Process.*, Jul. 2009, pp. 1–7.
- [73] L. Fu, X. Mao, and L. Chen, "Speaker independent emotion recognition using HMMs fusion system with relative features," in *Proc. Ist Int. Conf. Intell. Netw. Intell. Syst.*, Nov. 2008, pp. 608–611.
- [74] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, "Univer-sum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.
- [75] K. Amol T. and R. M. R. Guddeti, "Multiclass SVM-based language-independent emotion recognition using selective speech features," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2014, pp. 1069–1073.
- [76] P. X. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90376, 2019.
- [77] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, Oct. 2018.

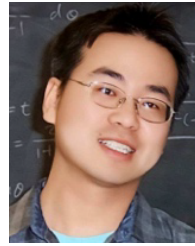


**HUIYUN ZHANG** was born in Huanxian, Gansu, China, in 1993. She is currently pursuing the doctor's degree with the Computer Science and Technology, Qinghai Normal University, China. Her research interests include pattern recognition and intelligence systems. Her research interests also include speech emotion recognition and machine learning.



**HEMING HUANG** was born in Ledu, Qinghai, China, in 1969. He received the B.S. degree in mathematics from Shaanxi Normal University, the M.S. degree in computer application technology from Lanzhou University, and the Ph.D. degree in pattern recognition and intelligence system from Southeast University, China.

He is currently a Professor of computer science and technology with Qinghai Normal University and a Doctoral Supervisor of pattern recognition and intelligent system. He is also a member of the China Computer Federation (CCF) and the Association for Computing Machinery (ACM).



**HENRY HAN** received the Ph.D. degree from the University of Iowa, in 2004.

He is currently a Professor of computer science with the Department of Computer and Information Science, Fordham University. He is also the Director of the Laboratory of Big Data and Analytics. His current research interests include AI, data science, big data, bioinformatics/health informatics, fintech, and cybersecurity. He has published nearly 80 articles in leading journals and conferences in data science fields. He was the Founding Director of Fordham University's master program in cybersecurity besides Department Associate Chair. He has been supervising about a total of 60 undergraduate, master students, and Ph.D. students, since 2005. His research has been supported by NSF, NIH, and research contracts from the industry.

...