IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 1, 2019, accepted July 4, 2019, date of publication July 8, 2019, date of current version July 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927384

Parallelized Convolutional Recurrent Neural Network With Spectral Features for Speech Emotion Recognition

PENGXU JIANG[1], HONGLIANG FU[1], HUAWEI TAO[1], PEIZHI LEI[1], AND LI ZHAO[2]
[1]College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China
[2]Laboratory of Underwater Acoustic Signal Processing, Southeast University, Nanjing 210096, China

Corresponding author: Pengxu Jiang (px20115c@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61673108 and Grant 61601170, in part by the Henan Provincial Science and Technology Research Project under Grant 192102210101, in part by the Natural Science Project of Henan Education Department under Grant 19A510009, and in part by the Start-up Fund for High-level Talents of Henan University of Technology under Grant 31401148.

ABSTRACT Speech is the most effective way for people to exchange complex information. Recognition of emotional information contained in speech is one of the important challenges in the field of artificial intelligence. To better acquire emotional features in speech signals, a parallelized convolutional recurrent neural network (PCRN) with spectral features is proposed for speech emotion recognition. First, frame-level features are extracted from each utterance and, a long short-term memory is employed to learn these features frame by frame. At the same time, the deltas and delta–deltas of the log Mel-spectrogram are calculated and reconstructed into three channels (static, delta, and delta–delta); these 3-D features are learned by a convolutional neural network (CNN). Then, the two learned high-level features are fused and batch normalized. Finally, a SoftMax classifier is used to classify emotions. Our PCRN model simultaneously processes two different types of features in parallel to better learn the subtle changes in emotion. The experimental results on four public datasets show the superiority of our proposed method, which is better than the previous works.

INDEX TERMS Speech emotion recognition, parallelized convolutional recurrent neural network, convolutional neural network, long short-term memory.

I. INTRODUCTION

Language contains a wealth of emotional information. People can capture the change of their emotional state from the language because we can perceive the information that reflects the emotional state from the speech signal. Speech emotion recognition is to simulate the process of people's perception of emotion by using a machine to mine the information contained in the speech. In the past decades, speech emotion recognition has attracted worldwide attention of relevant researchers and has made great achievements in many related fields [1]. With the development of Artificial Intelligence, the interaction between human and computer becomes more comfortable and convenient, how to make better use of artificial intelligence to recognize speech emotion has become the focus of the next generation of artificial intelligence development [2]. Therefore, the research on speech emotion recognition has strong theoretical value and practical significance.

Feature extraction is the first and most important step in speech signal processing. So far, a variety of hand-designed features have been used for speech emotion recognition [3], [4]. The spectral feature is a popular feature in recent years. Compared with traditional hand-designed features, spectral features can extract more emotional information by considering both frequency and time axes. Due to the above advantages of spectral features, many scholars have done a lot of related research [5]–[7].

However, these manual features are low-level; these features still unable to express the emotions contained in an utterance well. So how to extract more abundant emotional details from each utterance is the first problem we need

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa Rahimi Azghadi.

to solve. In recent years, because the neural network [8] exhibits outstanding performance in feature learning, it provides us with a possible solution to this problem. Deep learning proposes a method for the machine to automatically learn features, which can automatically extract specific feature representations from a large number of learning tasks and incorporate feature learning into the process of building models. Compared with hand-designed features, deep learning features reduce the incompleteness caused by artificial designed features.

Two typical deep learning models are convolutional neural network (CNN) [9] and Long Short-Term Memory (LSTM) [10]. LSTM is suitable for time series data because it can maintain the dependence between the front and back of the data, and CNN is suitable for image data processing by perceiving the local field of view of data. In recent years, several related networks have been successfully applied to feature learning in speech emotion recognition. Mao *et al.* [11], using CNN to train the emotional significance feature from the spectrogram. CNN training network consists of two steps. In the first stage, unlabeled data is used to train Local Invariant Features (LIF). In the second stage, LIF as input to feature extractors, Salient Discriminative Feature Analysis (SDFA) to get recognition features. The proposed feature has achieved better recognition results on three datasets. Zhang *et al.* [12] propose a Discriminant Temporal Pyramid Matching (DTPM) algorithm to pool deep features learned by CNN for speech emotion recognition. Chen *et al.* [13] proposed a 3-D attention-based convolutional Recurrent Neural Networks (ACRNN) for speech emotion recognition, they combine CNN with LSTM, and 3-D spectral feature of segments are used as input. Finally, an attention layer was used to produce utterance-level features. Trigeorgis *et al.* [14] proposed a convolutional recurrent model for speech emotion recognition; they used the original audio samples directly and divided each speech into equal-length segments as input to the model. The experiment has achieved good results.

The above works have promoted the progress in the research of speech emotion recognition, and the successful application of related neural networks has prompted us to use related networks to extract deep features of the speech signal. To achieve this, three issues need to be addressed. First, speech signals may have a various time of duration, and most models require a fixed input size. Some of the above works [12]–[14] split the speech signal into equal-length segments, for speech signal with fixed length is easier to design models, but fewer emotional details may not be able to train a robust network model, Second, the ability of individual neural network model to recognize emotion is limited. Many previous [13], [14] work combined different networks serially for emotional recognition. One of the advantages of this structure is that it is easier to design models suitable for segment-level features. However, the serial model structure may lose some emotional information due to the inheritance relationship between models. Designing the appropriate input to meet the needs of different models is the last issue.

To address these challenges, a Parallelized Convolutional Recurrent Neural Network (PCRN) with spectral features is proposed for speech emotion recognition. Firstly, log Mel-spectrograms are extracted from acoustic features set, and then we calculated deltas and delta-deltas for the log Mel-spectrogram to compose 3-D data as input of CNN. We use 2-D convolution to learn these data, 2-D convolution has more parameters, which can better capture the time-frequency correlation information in spectral features. Besides, we put the whole utterance into CNN to preserve all the emotional details of each data, so that the model has a stronger ability to feature learning. At the same time, we put each frame-level features into LSTM to learn frame by frame, the dimension of frame-level feature varies with speech length, speeches with different length has the same degree of detail, we can use LSTM model to learn all the emotional details of speech signals with different durations. Then we fuse the high-level features learned in LSTM and CNN models. The proposed PCRN model makes full use of the time-frequency correlation of spectral features; 2-D convolution is employed to capture more detailed time-frequency correlation but also learns the subtle change of emotion in different frequencies frame by frame without losing data details. Finally, these fused features are Batch Normalized [15] and classified by SoftMax classifier.

The main contributions of this paper can be summarized as:

1) The proposed PCRN model combines CNN and LSTM. Different from the traditional cascade connection mode, the parallel connection mode is adopted in our model and multi-feature as input to learn the complete emotional details in different functional features at the same time.
2) The proposed model utilizes the time-frequency correlation of spectral features in parallel processing to capture emotional changes in the time-frequency domain. Experimental results show that our PCRN model has better recognition effect than other previous works.

II. RELATED WORK

Emotion classification and feature extraction are two key steps in speech emotion recognition. In this section, we first briefly introduce the speech emotion classifier, and then focus on feature extraction, because it is more relevant to our work.

A. EMOTIONAL CLASSIFIER

Classifier plays a role in classifying and recognizing features. Common emotional classifiers include Hidden Markov Models (HMM) [16], Gaussian Mixture Model (GMM) [17], Support Vector Machine (SVM) [18], K-Nearest Neighbor (KNN) [19] and Softmax function [20]. Among these classifiers, SVM and Softmax are most widely used in speech-related recognition methods. Each classifier has its advantages and disadvantages.

B. FEATURE EXTRACTION

Feature extraction is the key to emotional recognition. It extracts information related to emotional types from

speech signals. The quality of features directly determines the result of emotion recognition. Speech features can usually be classified into four categories: acoustic features, linguistic features, context information, and hybrid features. Among them, acoustic features are most frequently used in affective recognition.

Acoustic features, as one of the most popular emotional features, can be subdivided into four categories: prosodic features, speech quality features, spectral correlation features, and other features. Prosodic features are also called super tone quality features or Supersegmental features. Prosodic features are one of the most important forms of speech and emotional expression. The Prosodic feature is a phonological structure of language. It is closely related to linguistic structures such as information structure and syntax. It is also a typical feature of human languages, such as pitch down and pauses. Common prosodic features include zero-crossing rate, fundamental frequency, logarithmic energy, etc. In emotional recognition, prosodic features play an important role. Scherer [21] studied the importance of envelope curves in different emotional contexts. Zhou *et al.* [22] extracted 79 prosodic features for emotion recognition. Dai *et al.* [23] extracted the characteristics of peak and energy for emotion recognition. Speech quality is a common feature in affective recognition. When people are emotionally agitated and difficult to control, their voices are often accompanied by choking, tremolo, and other acoustic manifestations. In emotional recognition, commonly used speech quality features include harmonic noise ratio, glottal parameters, etc. Kane *et al.* [24] studied the application of speech quality features in emotional speech synthesis. Kächele *et al.* [25] and others integrate prosodic features with speech quality information to identify speech emotion. Traditional linear spectral correlation features include: Linear Predictor Coefficient (LPC), Log-Frequency Power Coefficient (LFPC), Cepstrum features include: Linear Predictor Cepstral Coefficient (LPCC), Mel-Frequency Cepstral Coefficient (MFCC), etc. In recent years, scholars have conducted in-depth research on spectral correlation features. Wu *et al.* [26] proposed a modulation spectrum feature (MSFs). MSFs feature uses acoustic filter banks and modulation filter banks to process speech signals and extract time-frequency information of signals. Deng *et al.* [27] proposed using improved Modified group delay features and All-pole group delay features to identify the emotion

III. GENERATION OF PCRN INPUT

Since spectral features can better describe the time-frequency correlation of emotional details, more and more researchers have applied spectral features to speech emotion recognition [28], [12], [13]. The advantage of spectral features is to model the speech spectrum as an image and extract emotional information from the spectral using image feature descriptors. The frequency axis and time axis of the spectrum are considered synthetically to extract more information related to emotion. However, speech signals have different durations, but most neural networks require a fixed length of the input. Many previous works [13], [14] has divided the speech signal to equal-length segments as input and Connected different models in cascade. It is easier to design models in this way, but incomplete feature representations in each utterance may lose emotional details. To address this issue, we extract 3-D log Mel-spectrograms and frame-level features from the original speech signal as inputs to our PCRN model to prevent the loss of emotional information in time-frequency domain learning.

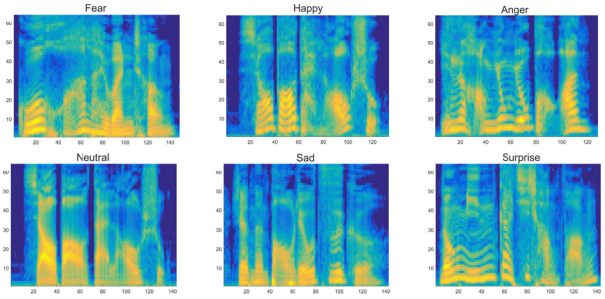

FIGURE 1. Sample speech spectrograms.

In this paper, for each speech signal, we adopt 64 Mel-filter banks to obtain frame-level features with Hamming windows size of 25 ms and 15 ms overlapping; each frame has the same degree of detail. The size of log Mel-spectrograms varies with the size of frame-level features. Sample spectrograms are shown in Fig. 1. The X-axis represents the number of frames, and the Y-axis represents the number of Mel-filter banks. After obtaining 2-D static Mel-spectrogram, we calculated the delta and delta-delta coefficients along the time axis of the static Mel-spectrogram, 3-D feature representation as one of the inputs of the PCRN model:

$$Mels \in R^{T \times F \times C}, \tag{1}$$

where T is the frame number, because the length of each utterance may be different, the size of T may also be different. F is the number of Mel-filter banks, in this experiment, we set F to 64. C represents the number of channels. We set C to 3 to represents static, delta and delta-delta. At the same time, we use frame-level features as another input to our PCRN model. Frame-level features can be represented as:

$$Frames \in R^{T \times F}, \tag{2}$$

the values of T and F are the same as those in Mels. We use two different features as the input of the model to obtain more emotional information in each speech.

IV. PROPOSED FRAMEWORK

A. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural networks are usually composed of convolution layer, pooling layer, and fully connected layer. Convolution kernel is used to extract the deep information

of the features automatically; each convolution kernel is connected to the local region of the upper feature map. Convolutional layers can be composed of multiple feature maps; each feature maps can be represented as:

$$x_j^l = f\left(\sum_{i\in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right), \tag{3}$$

where x_j^l represents the j-th feature maps in Convolutional Layer l, it is activated by convoluting all the feature maps of the upper layer and adding bias. f(·) is an activation function, usually sigmoid and tanh. k_{ij}^l and b_j^l are weights and biases, respectively.

Pooling layer is used to sample the feature maps and reduce the parameters. Each output feature maps can be represented as:

$$x_j^l = f(u_j^l), \tag{4}$$
$$u_j^l = down\,(x_j^{l-1}), \tag{5}$$

down (·) represents the down-sampling function, computing the feature map x_j^{l-1} of the upper layer with a certain sampling size. For example, the size of 2×2 or 4×4.

The fully connected layer is usually located at the end of the network. Each neuron is fully connected to all the neurons in the upper layer; it can integrate local information with category discrimination in convolution or pooling layers. The output x_j^l of the fully connected layer can be obtained by weighted summation of inputs:

$$x_j^l = f\left(u_j^l\right), \tag{6}$$
$$u^l = w^l x^{l-1} + b^l. \tag{7}$$

Local connection, weight sharing, and down-sampling in CNN can effectively reduce the complexity of the network model and the number of training parameters. The input data are abstracted into high-level features representation by algorithmic operations between layers. Finally, the feature-to-task target mapping is used as the end.

B. LONG SHORT-TERM MEMORY

Long Short-Term Memory is a variant to solve the problem of long-term dependence in the recurrent neural network; it implements a more refined internal processing unit to store and update context information effectively. The input of LSTM at each time is the input value x_t at the current time, the output value h_{t-1} at the previous time and the unit state c_{t-1} at the last time, the output is the current time h_t and the current state c_t, respectively. The forget gate f_t to determine the information that cells discard:

$$f_t = \sigma\left(W_f\,[h_{t-1}, x_t] + b_f\right), \tag{8}$$

σ represent the activation functions Sigmoid, W and b represent weight and bias, respectively. The forget gate f_t outputs a value between '0' and '1' by reading the input x_t and the cell state h_{t-1} of the previous moment. '1' means complete reservation, '0' means complete abandonment. Then the cell decides the value to be updated:

$$i_t = \sigma(W_i\,[h_{t-1}, x_t] + b_i), \tag{9}$$
$$C_t^{\sim} = tanh\,(W_c\,[h_{t-1}, x_t] + b_c). \tag{10}$$

The sigmoid layer determines what values we will update; a tanh layer creates a new candidate value vector. The cell then updates the status from C_{t-1} to C_t and eventually outputs h_t:

$$C_t = f_t * C_{t-1} + i_t * C_t^{\sim}, \tag{11}$$
$$h_t = o_t * tanh\,(C_t). \tag{12}$$

Since the expanded LSTM model can obtain repetitive network structure and share parameters among each network, the training parameters are reduced, and the model can be extended to different length sequences, this allows the LSTM model to be used for sequences of varying lengths.

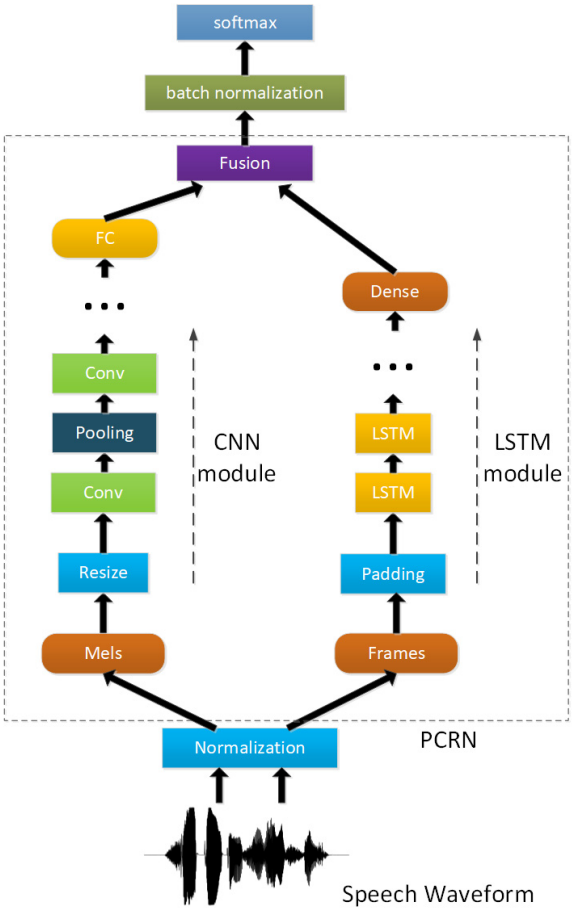

FIGURE 2. Model structure.

C. STRUCTURE OF PCRN

The network structure designed in this study is presented in Figure 2. To improve the convergence speed of the model, we first normalize the original speech waveform. Then we

extract two different feature representations and send them to different modules in the PCRN to learn the details of emotional features in the time-frequency domain.

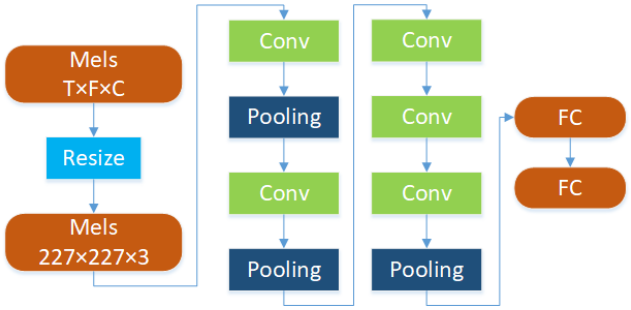

FIGURE 3. CNN module.

First, the CNN module is used to learn the details of emotion from time and frequency of the Mels features we extracted. To get a better training effect, we use AlexNet trained on ImageNet dataset as our initial CNN model. The model structure of CNN is shown in Figure 3. There are five convolution layers, three pooling layers, and two fully connected layers in our model, max pooling is adopted in pooling layer, and we deleted the last fully connected layer. Because most convolutional neural networks need fixed input size, we need to resize Mels $\in R^{T\times F\times C}$ into $227 \times 227 \times 3$. We perform the resize operation with bilinear interpolation.

At the same time, we use the LSTM model to learn the temporal changes of emotional details. The LSTM network can handle variable length features by feeding it one frame at a time, but the number of frames of frame-level features varies with time. To meet the requirements of the model, these features should be zero-padded into the same dimension. When the time steps are outside the range of the actual data length, the LSTM internal parameters will stop updating. The model structure of LSTM is shown in Figure 4. To improve the stability of the model, we average the output of each frame.

Then we integrated two different types of high-level features. To improve the convergence speed and avoid gradient diffusion of training, the output of PCRN was normalized by batch normalization. Finally, a SoftMax classifier is used to classify emotions. All modules are trained at the same time to ensure that the model can learn the emotional integrity of each utterance.

V. EXPERIMENTS

A. SPEECH EMOTION DATABASE

To verify the effectiveness of our proposed model, we tested it on four common datasets, including the Institute of Automation, Chinese Academy of Sciences emotion dataset (CASIA) [29], Berlin EMO-DB [30] German Emotional Voice Library, the Airplane Behavior Corpus (ABC) [31] and Surrey Audio-Visual Expressed Emotion (SAVEE) [32].

The CASIA speech emotion database was recorded by Institute of Automation, Chinese Academy of Sciences. It is

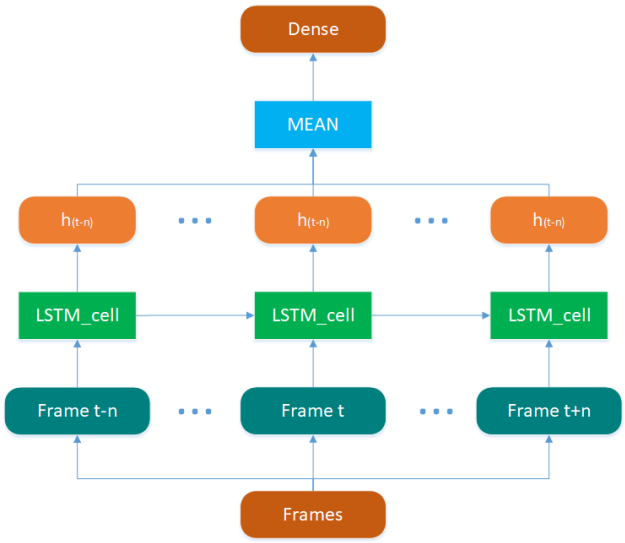

FIGURE 4. LSTM module.

a Chinese database for emotional phonetics. It is recorded by 4 actors (2 men and 2 women) in 6 different emotions (anger, fear, happy, neutral, sad, surprise). The signal-to-noise ratio (SNR) is about 35 dB for data acquisition in a pure recording environment, 16bit quantization, and 16 KHz 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. Each audio file averages about 1.9 seconds long.

The EMO-DB database is a German-language speech emotion database recorded by the University of Berlin. The database consists of 535 utterances with seven different emotions, anger (127), boredom (81), disgust (46), fear (69), happiness (71), neutral (79), and sadness (62). The database displayed by ten professional actors (5 males and 5 females), everyone speaks 49, 58, 43, 38, 55, 35, 61, 69, 56, 71 utterances respectively. Each audio file averages about 2.7 seconds long.

The ABC speech emotion dataset is a German database; this database is an evoked database. To induce every clear emotion, everyone has to experience a scene through a hidden script. Eight gender balance subjects (4 men and 4 women) participated in the recording. Each speaks 45, 59, 60, 47, 71, 61, 39, 48 texts. A total of 11.5 hours of video recording were recorded, the average length of the data set is 10s. Three experts cut the data and demarcate the emotion. It contains six emotions, aggressive (95), cheerful (105), intoxicated (33), nervous (93), neutral (79), tired (25).

The SAVEE database consists of recordings from 4 male actors in 7 different emotions, anger (60), disgust (60), fear (60), happiness (60), sadness (60), surprise (60), and neutral (120). 480 British English utterances in total. Each speaker says 120 utterances; the average length of the data set is 4s. The sampling rate was 44.1 kHz. The data were recorded in a visual media lab with high-quality audio-visual equipment, processed and labeled.

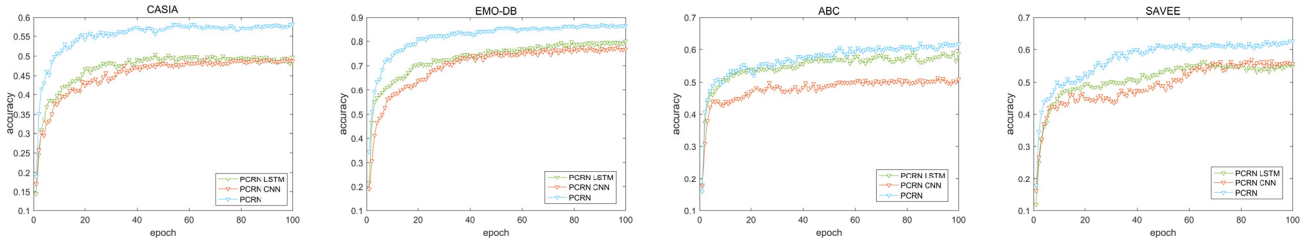

FIGURE 5. Experimental results of different models on four database.

B. EVALUATION METHODS

Leave-One-Speaker-Out (LOSO) [33] cross-validation strategy is used in this experiment, which is more realistic and challenging. In this strategy, each time an emotional speech sample of one person is selected from the data set as a test set of experiments, the remaining emotional speech samples are used as a training set. Each person's voice takes turns as a test set. Finally, the average of several trials is calculated as a result.

As classes are unbalanced, weighted average recall (WA) [34] and unweighted average recall (UA) are used to evaluate experimental results. WA is the ratio of identifying the correct number of test samples to the number of all test samples. UA is the sum of the accuracy of all classes divided by the number of classes, without considering the number of samples per class.

C. EXPERIMENTAL PARAMETERS

To get the emotional details contained in the speech better, we extract two different spectral features *M*els and *Frames* of each speech as our PCRN inputs. Parallel operation of CNN and LSTM in the model of PCRN, the parameters of the PCRN model are shown in Table 1.

TABLE 1. Parameters of PCRN.

Network	Layer	Shape
PCRN_CNN	Input	227×227×3
	Conv_1	11×11×96
	Conv_2	5×5×256
	Conv_3	3×3×384
	Conv_4	3×3×384
	Conv_5	3×3×256
	Fc_6	2048
	Fc_7	2048
PCRN_LSTM	Input	64
	Hidden units	2048
	Dense	2048
Output		4096

When we get the fused features of the output of the PCRN model, we pass these features through a batch normalization layer and then use a SoftMax classifier to classify emotions. The PCRN architecture is implemented with TensorFlow toolkit. The model parameters are optimized by minimizing the cross-entropy objective function. We used Adam optimizer in our experiments, and the initial learning rate is set to 0.00001. To prevent data over-fitting during training, we added Dropout into the PCRN model and set it to 0.8.

D. MODEL COMPARISON

To illustrate the effect of our parallel model on the experimental results, experiments were carried out on single-feature model PCRN_CNN and PCRN_LSTM, that is, removing the LSTM part of the model and the CNN part of the model separately, and identifying emotions using single-feature with the remaining parameters unchanged. Through 100 epoch iterative training, the experimental results are shown in Fig 5, from left, on the CASIA, EMO-DB, ABC, and SAVEE database. Table 2 provides detailed data comparison. Because the number of samples in each category was equal, the experimental results have the same values of WA and UA on the CASIA database.

TABLE 2. Accuracy (%) comparison of different architectures of PCRN model.

Database	Model	WA	UA
CASIA	PCRN_CNN	48.66	48.66
	PCRN_LSTM	49.67	49.67
	PCRN	58.25	58.25
EMO-DB	PCRN_CNN	76.64	72.85
	PCRN_LSTM	80.01	78.37
	PCRN	86.44	84.53
ABC	PCRN_CNN	50.97	44.90
	PCRN_LSTM	58.77	54.88
	PCRN	61.63	57.59
SAVEE	PCRN_CNN	55.62	49.64
	PCRN_LSTM	54.79	49.76
	PCRN	62.49	59.40

As shown in Table 2, the performance of PCRN was demonstrated on four databases; the weighted average recall was 58.25%, 86.44%, 61.63%, 62.49%, respectively. The performance of PCRN model has a significant improvement.

We observed that in the ABC dataset, the improvement of the PCRN model to the LSTM model was relatively small; this may be because the average length of each utterance in the ABC dataset is about 10s, we use 25ms window size with 15ms overlapping to obtain frame-level features. About $(10000\text{ms} - 25\text{ms}) \div 10ms + 1 \approx 998$ frames

in 10 seconds, and CNN requires a fixed input size. Large-length texts may not be able to learn the emotional information of each speech in convolution fully. On the other hand, because of the increase in the number of speech frames, our LSTM module can learn more abundant time-related information. Using the advantage of parallel training with multi-features as input, the PCRN model can balance the differences of emotional information between modules, so that it can finally learn the whole emotional information of each utterance.

It can also be observed in Fig. 5. Through 100 epoch training, all data sets converge. Compared with the CASIA and EMO-DB database, the ABC and SAVEE database has greater fluctuation in iteration convergence maybe for the following reasons. First, LOSO strategy is used in this experiment, we need to average each person's test results rather than individual test results, and some volatile test results will affect the average recognition rate. Second, the ABC database has 430 samples, and SAVEE database has 480 speech samples, while the CASIA database has 1200 speech samples. Lack of sufficient training data makes it more difficult for databases with fewer samples to converge in iteration than a database with larger samples, which will cause large fluctuations. Third, the number of samples in each emotion was not equal. In the ABC database, the number of 'cheerful' is 105, and the number of 'tired' is only 25. The imbalance in the number of categories may cause huge fluctuations in convergence. Fourth, the average sample length of CASIA and EMO-DB datasets is less than 3s, and both databases are recorded in professional studios. The average sample length of ABC database is 10s, and it is an evoked database. The long duration of speech may affect the model's ability to discriminate emotions, and the noise will also have a certain degree of interference. Although the average sample length of SAVEE database is 4s, the sampling rate was 44.1 kHz, which means that more frames need to be calculated.

TABLE 3. T-test on test results.

Database	Models	P-Value
CASIA	(PCRN, CNN)	< 0.0001
	(PCRN, LSTM)	< 0.0001
EMO-DB	(PCRN, CNN)	< 0.0001
	(PCRN, LSTM)	< 0.0001
ABC	(PCRN, CNN)	< 0.0001
	(PCRN, LSTM)	0.015
SAVEE	(PCRN, CNN)	< 0.0001
	(PCRN, LSTM)	< 0.0001

A significance test was carried out on the experimental results to investigate the improvement of the recognition effect of different modules on four databases. The results of T-test are shown in Table 3; it uses t-distribution theory to infer the probability (P-Value) of difference occurrence to judge whether there is a significant difference between two groups of data. When P-Value is less than 0.05, the difference between data is significant. As we can see from Table 3, all P-Values are less than 0.05 on four databases. Therefore, compared with CNN and LSTM module, the performance of PCRN model has a significant improvement.

Furthermore, the performance of the proposal is compared with some state-of-art works as well. We only compare with works also using the same setting. Table 4 shows a comparison between our proposed method and other methods.

TABLE 4. Accuracy (%) Comparison of different features.

Database	Refs	Features	WA	UA
CASIA	[35]	GA-BEL	38.55	38.55
	[36]	HuWSF	43.50	43.50
	[39]	RDBN	48.50	48.50
	Ours	PCRN	58.25	58.25
EMO-DB	[36]	HuWSF	81.74	/
	[37]	Acoustic	81.90	79.10
	[39]	RDBN	82.32	/
	[40]	LNCMSF	/	74.46
	[13]	ACRNN	/	82.82
	Ours	PCRN	86.44	84.53
ABC	[37]	Acoustic	61.50	56.10
	[33]	Acoustic	61.40	55.50
	[38]	ComParE	/	56.11
	[40]	LNCMSF	/	52.26
	Ours	PCRN	61.63	57.59
SAVEE	[35]	GA-BEL	44.18	/
	[36]	HuWSF	50.00	/
	[39]	RDBN	53.60	/
	Ours	PCRN	62.49	59.40

From Table 4, we can see that our method is very competitive to the state-of-the-art results. The best performance of our method in the CASIA and SAVEE database and our method obviously outperforms [35], [36], [39]. The recognition rate is at least 9.75% and 8.89% higher than that of all comparative experiments. Effects of the experiment are conspicuous. On the ABC dataset, our method also outperforms [33], [37], [38], [40] in term of WA. And we report an UA of 57.59% on the ABC dataset, on which outperforms all the four compared works, i.e., 56.1% by [37], 55.5% by [33], 56.11% by [38], 52.26% by [40]. On the EMO-DB dataset, our method also clearly outperforms all the four compared works. The experimental results in four databases demonstrate the feasibility of using features of different functions as model input simultaneously. And the recognition effect of high-level features extracted by the proposed PCRN model outperforms other comparative experiments.

To further investigate the recognition accuracy, we present the confusion matrix to analyze the performances of our PCRN model.

Fig.6 shows that on the CASIA dataset, our model has excellent recognition results for 'anger' and 'sad', the classification accuracy of 'anger' and 'sad' are 75% and 72%, respectively. The classification accuracy of 'neutral' is

FIGURE 6. Confusion matrix of PCRN on the CASIA dataset.

FIGURE 7. Confusion matrix of PCRN on the EMO-DB dataset.

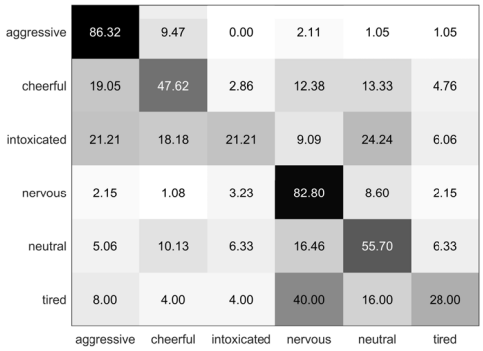

FIGURE 8. Confusion matrix of PCRN on the ABC dataset.

61.5%, and the other three emotions are identified with accuracies lower than 50%. Fig.7 shows that on the EMO-DB dataset, 'anger' and 'sadness' are classified with accuracies higher than 90%, the classification accuracy of 'boredom' and 'neutral' are 87.65% and 88.61%, respectively. The recognition rate of the other three emotions is less than 80%. Fig.8 shows that on the ABC dataset, 'aggressive' also obtains the highest recognition rate, and the accuracy of 'nervous' is also over 80%, the classification accuracy of 'intoxicated' and 'tired' is less than 30%, and the other two emotions can be recognized with accuracies of 47.62%, 55.70%, respectively. Fig.9 shows that on the SAVEE dataset, 'neutral' is identified with the highest accuracy of 84.17%, and the accuracy of the other six emotions was less than 70%.

FIGURE 9. Confusion matrix of PCRN on the SAVEE dataset.

From the confusion matrix, we can see that the difference between the highest recognition rate 'aggressive' and the lowest recognition rate 'intoxicated' in the ABC database is more than 60%, the difference between the highest recognition rate 'neutral' and the lowest recognition rate 'sadness' in SAVEE database is 44.17%. In CASIA and EMO-DB databases, there would not be such a big difference. On the one hand, speech samples in the CASIA database are almost three times as many as those in ABC and SAVEE database, and the EMO-DB database also has more samples than ABC and SAVEE databases, adequate training data will improve the performance of the model. And on the other hand, the number of samples in each category is not balanced. The two emotions 'intoxicated' and 'tired' with the lowest recognition rate in the ABC database also have the lowest proportion in the database. 21.21%, 18.18%, 24.24% 'intoxicated' sample are misclassified as 'aggressive', 'cheerful', 'neutral', 40% 'tired' sample are misclassified as 'nervous', where the four emotions that account for the largest proportion in the database. The highest recognition rate 'neutral' in SAVEE database also accounts for the highest proportion of samples and 45% 'sadness' samples are misclassified as 'neutral'. And LOSO strategy is adopted in this experiment, speed, voice line, and style of speech of different speakers also have different effects on the experiment.

VI. CONCLUSION

This paper presents a network model PCRN for speech emotion recognition. Different from previous works on speech emotion recognition, we use 3-D log Mel-spectrograms and frame-level features, two features of two different types as input simultaneously. Variable length frame-level features preserve the time information of speech completely, and 3-D log Mel-spectrograms involves more parameters to capture more detailed temporal-frequency correlations. Two kinds of features are processed simultaneously by using the PCRN model to extract high-level features with different emotional details, and the parallel internal structure balances the differences of emotional information between modules. The experiment shows that the PCRN model can effectively mine the emotional information from these spectral features, and

experiments on four databases show the superiority of our proposed approach compared with the state-of-the-art.

REFERENCES

[1] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. Di Natale, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 314–327, Jul./Sep. 2017.

[2] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Commun.*, vol. 83, pp. 34–41, Oct. 2016.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[4] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, 2018.

[5] 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.

[6] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5150–5154.

[7] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. ICASSP*, Mar. 2017, pp. 2227–2231.

[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] 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.

[12] S. Zhang, S. Zhang, T. Huang, and W. Geo, "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.

[13] 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.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1–9.

[16] 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, 2003.

[17] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Amsterdam, The Netherlands, Jul. 2005, pp. 1500–1503.

[18] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, Montreal, QC, Canada, May 2004, p. I-577.

[19] T. L. Pao, *et al.*, "A comparative study of different weighting schemes on KNN-based emotion recognition in mandarin speech," in *Proc. Intell. Comput. Int. Conf. Adv. Intell. Comput. Theories Appl.* Springer-Verlag, 2007.

[20] N. Andrew, N. Jiquan, and Y. F. Chuan. *Softmax Regression[EB/OL]*. Accessed: Apr. 10, 2013. [Online]. Available: http://ufldl.stanford.edu/wiki/index.php/Softmax_Regression

[21] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, nos. 1–2, pp. 227–256, 2003.

[22] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in *Proc. Int. Conf. Inf. Eng. Comput. Sci. (ICIECS)*, Wuhan, China, Dec. 2009, pp. 1–4.

[23] K. Dai, H. J. Fell, and J. MacAuslan, "Recognizing emotion in speech using neural networks," *Telehealth Assistive Technol.*, vol. 31, pp. 38–43, Apr. 2008.

[24] J. Kane, S. Scherer, M. Aylett, L.-P. Morency, and C. Gobl, "Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 7982–7986.

[25] M. Kächele, D. Zharkov, S. Meudt, and F. Schwenker, "Prosodic, spectral and voice quality feature selection using a long-term stopping criterion for audio-based emotion recognition," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, Stockholm, Sweden, Aug. 2014, pp. 803–808.

[26] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011.

[27] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Exploitation of phase-based features for whispered speech emotion recognition," *IEEE Access*, vol. 4, pp. 4299–4309, 2016.

[28] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[29] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "CHEAVD: A Chinese natural emotional audio–visual database," *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 6, pp. 913–924, 2017.

[30] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Eur. Conf. INTERSPEECH-EUROSPEECH DBLP*, 2005, pp. 1517–1520.

[31] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Honolulu, HI, USA, vol. 2, Apr. 2007, pp. 733–736.

[32] H. Sanaul, P. J. B. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition," in *Proc. AVSP*, 2009, pp. 53–58.

[33] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Merano, Italy, Nov./Dec. 2009, pp. 552–557.

[34] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brighton, U.K., 2009, pp. 312–315.

[35] 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.

[36] 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.

[37] 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)*, Prague, Czech Republic, May 2011, pp. 5688–5691.

[38] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4990–4994.

[39] 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, Feb. 2017, Art. no. 1945630.

[40] 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.

PENGXU JIANG is currently pursuing the master's degree with the College of Information Science and Technology, Henan University of Technology, China. His research interests include speech emotion recognition and machine learning.

HONGLIANG FU received the B.E., M.S., and Ph.D., degrees from the Nanjing University of Posts and Telecommunications, China, in 1986, 1989, and 2006, respectively. He is currently a Professor with the Henan University of Technology, China. His research interest includes signal processing.

HUAWEI TAO received the Ph.D. degree in information and communication engineering from Southeast University, China, in 2017. He is currently a Lecturer with the Henan University of Technology. His current research interests include speech emotion recognition and deception detection.

PEIZHI LEI is currently pursuing the master's degree with the College of Information Science and Technology, Henan University of Technology, China. His research interests include deception detection and machine learning.

LI ZHAO received the B.E. degree from the Nanjing University of Aeronautics and Astronautics, China, in 1982, the M.S degree from Suzhou University, China, in 1988, and the Ph.D. degree from the Kyoto Institute of Technology, Japan, in 1998. He is currently a Professor with Southeast University, China. His research interests include speech signal processing and pattern recognition.