3DACRNN Model Based on Residual Network for Speech Emotion Classification

Zhangfang Hu, Shanshan Tang, Yuan Luo, Fang Jian, Xingtong Si

Abstract—**Speech emotion recognition(SER) is extremely challenging due to the problem of disappearing or exploding gradients and weak spatiotemporal correlations. To address this issue, a new approach is proposed the 3D attentional convolutional recurrent neural networks based on residual networks (Res3DACRNN) model to learn emotion deep features. The Res3DCNN model extracts deep-level multiscale spectral-temporal features of emotional speech from spectrograms. The introduction of a residual network allows compensation for the missing features of traditional CNNs in the convolution process to prevent the problem of gradient disappearance or explosion. An attention-based recurrent neural network (ARNN) then extracts the long-term dependencies of these features, improving the weak spatiotemporal correlation of the problem. To reduce the computational complexity, this paper improves the forget gate of LSTM and proposes a novel post-forgetting gate structure. Finally, a softmax layer is utilized for emotion classification. The experimental results of the proposed model on the EMO-DB and IEMOCAP emotional corpus show that the performance is significantly improved compared with the current mainstream deep learning methods.**

Index Terms—**Convolutional neural network, recurrent neural network, residual network, post-forget gate**

I. Introduction

LANGUAGE is the direct medium of human information exchange [1]. The field of human-computer interaction based on speech interaction, such as speaker recognition [2] and speech recognition, is of great significance. Therefore, speech emotion recognition (SER) is the most representative in terms of practicality and wide application. One of the key aspects is the extraction of feature sets from speech signals that can characterize human emotions. And so far, there are still no systematic feature sets [3].

Many previous studies have directly extracted low-level descriptors (LLDs) from speech and then used traditional machine learning methods to classify emotions[4]-[6]. However, it is still difficult to find salient feature sets from LLDs to recognize distinct emotions due to the presence of contextual and environmental factors. With the development of science and technology, image processing has become easy to implement [7], so the new trend in SER research is to convert speech signals into spectrograms as SER recognition objects. Many researchers have carried out SER technology research based on spectrograms and achieved good results. Tarunika et al. used deep neural networks (DNNs) to extract high-level emotional feature representations from the amplitude spectrum and showed better performance compared to traditional acoustic features [8]. Han et al. used the highest-energy fragments to train a DNN model to extract effective emotional information [9].

In recent years, CNN and RNN have been widely used in the field of SER. The deep convolutional model can keep the spectral-time translation invariance of speech signals, and RNN has excellent performance in processing time-series information [10]. Neumann et al. integrated the learning representation from an unsupervised automatic encoder into the CRNN sentiment classifier to improve the recognition accuracy [11]. However, this method of a CNN learning features from spectrograms is only a fusion of CNN features from a single spectrogram; thus, the link between spectra and time is often overlooked. Therefore, some studies have proposed a three-dimensional convolutional model for SER, which can better capture the spectral-temporal relationship of feature representation. Peng et al. used the spectrogram as the input of 3DCRNN, 3D convolutional layers for extracting high-level multiscale spectral-temporal representations, and recurrent layers for extracting long-term dependencies of these features [12]. To address the problem of interference of redundant information in SER, Chen et al. proposed an attention-based 3D convolutional recurrent neural network (ACRNN) model for learning the discriminative features of SER [13]. The introduction of an attention mechanism effectively reduces the influence of redundant information such as silent frames.

However, as the number of convolutional layers using the CNN model increases, the original features are gradually lost, with gradient disappearance or explosion problems, resulting in a decrease in the recognition accuracy [14]. For this problem, this paper proposes a 3D attention convolution

Manuscript received July 3, 2020; revised February 17, 2020. This work was supported in part by National Natural Science Foundation Youth Fund Project (Grant No. 61703067), Chongqing Basic Science and Frontier Technology Research Project (Grant No. Cstc2017jcyjAX0212), and Chongqing Municipal Education Commission Science and Technology Research Project (KJ1704072)

Zhangfang Hu is a Professor of the Key Laboratory of Optical Information Sensing and Technology, School of Optoelectronic Engineering, Chongqing University Posts and Telecommunications, Chongqing 400065 China (e-mail: 495075688@qq.com).

Shanshan Tang is a Master's degree candidate of the Department of Electronic Science and Technology, School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 1439789101@qq.com)

Yuan Luo is a Professor of the Key Laboratory of Optical Information Sensing and Technology, School of Optoelectronic Engineering, Chongqing University Posts and Telecommunications, Chongqing 400065 China (e-mail: 2217793866@qq.com)

Fang Jian is a Master's degree candidate of the Department of Electronic Science and Technology, School of Optoelectronic Engineering, Chongqing University Posts and Telecommunications, Chongqing 400065 China (e-mail: 1172599506@qq.com)

Xingtong Si is a Master's degree candidate of the Department of Electronic Science and Technology, School of Optoelectronic Engineering, Chongqing University Posts and Telecommunications, Chongqing 400065 China (e-mail: 1091253738@qq.com)

recurrent neural network based on ResNet to extract affective salient features based on previous studies. The CNN, after adding ResNet, improves the recognition rate by transmitting the original feature information to different levels of convolution structures, thus effectively solving the gradient disappearance or explosion problem. The ARNN model has made a great breakthrough in processing the structural relationship between input and output sequences [15]. To reduce the computational complexity, this paper improves the forget gate in the LSTM structure and proposes a novel post-forget gate structure to reduce the computations by reducing the number of parameters.

II. 3DACRNN Model Based on Residual Network for Speech Emotion Classification

A. Spectrogram

The spectrogram can avoid the cumbersome process of manual feature extraction and reduce the workload in the modelling and training process. In addition, it can reflect not only the energy characteristics of the speech signal but also the textural features of the rhythmic changes in the speech signal [16]. The horizontal axis of the spectrogram is time, and the vertical axis is frequency. The magnitude of the energy value is represented by color, and the darker the point is, the stronger the energy [17]. However, the traditional spectral graph is two-dimensional data, and the information characteristics are not obvious, while the input of 3DCNN is three-dimensional data. Therefore, in this paper, the speech signal is processed to extract the spectral-time representation.

In this paper, we extract the spectral-temporal representation using the signal processing steps shown in Fig. 1 [18]. We first apply a pre-emphasis filter to the signal to amplify the high frequency, so that the signal can be well recognized, and then use normalization to eliminate the difference. Second, a set of 32 Gammachirp filters simulate cochlear processing to filter the emotional speech signal s(n). Then, Hilbert transform is used to calculate the instantaneous amplitude of the channel signal to extract the time envelope. In addition, to get the connection between time and spectrum, we use a low-pass filter and five bandpass filters to extract spectral-time modulation signals [19]. The formats of the input and output data are set to "D×H×W", where D, H, and W indicate the acoustic frequency channel (depth), modulating frequency channel (height) and time series (width), respectively. Specifically, the input size is 32×6×6,000.

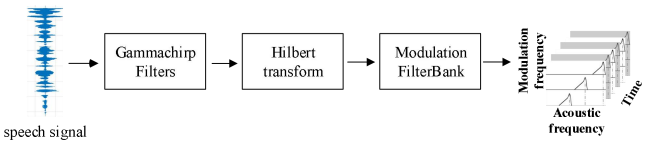

Fig. 1. Steps of extracting spectral-temporal representation

B. 3D Convolutional Neural Networks Based on ResNet

ResNet structure

The main idea of the ResNet structure [20] is to input shallow information into the deep network, which can alleviate the gradient disappearance problem caused by increasing the depth in the deep neural network. The principle is shown in Fig. 2. The residual block in the convolutional neural network directly passes the input x to the output, and the output result is $H(x)=F(x)+x$. ResNet changes the learning objective from the complete output to the difference between the target value $H(x)$ and x, which is also called the residual. Therefore, the goal of the training is to approximate the residual result to zero, and accuracy can be maintained as the network deepens.

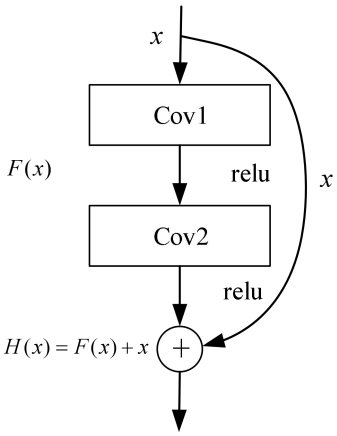

Fig. 2. Residual blocks

3D Convolutional Neural Networks Based on ResNet

CNN is widely used in image processing [21], and the 3DCNN framework was first applied in motion recognition [22]. The structure of the 3DCNN is shown in Fig. 3. The size of the 3D convolutional kernel is N×N×M, and L indicates the number of output channels. In this structure, each feature map is connected to multiple adjacent consecutive frames in the previous layer to capture temporal and spatial association information in speech signal.

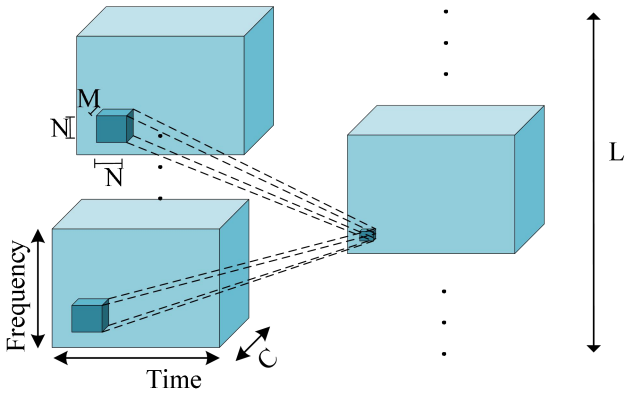

Fig. 3. The structure of the 3DCNN

Deep neural networks can extract more abstract features and are useful for speech emotion classification effects. However, it is prone to the problem of gradient disappearance or explosion. Li Wuke proposed a deep learning framework based on a multilayer extreme learning machine (ELM) to address this problem [23]. Inspired by this, we use residual networks. The shortcuts of residual networks can make the gradients propagate more efficiently. Many experiments have been conducted to show that the use of just one residual block is not very effective, generally

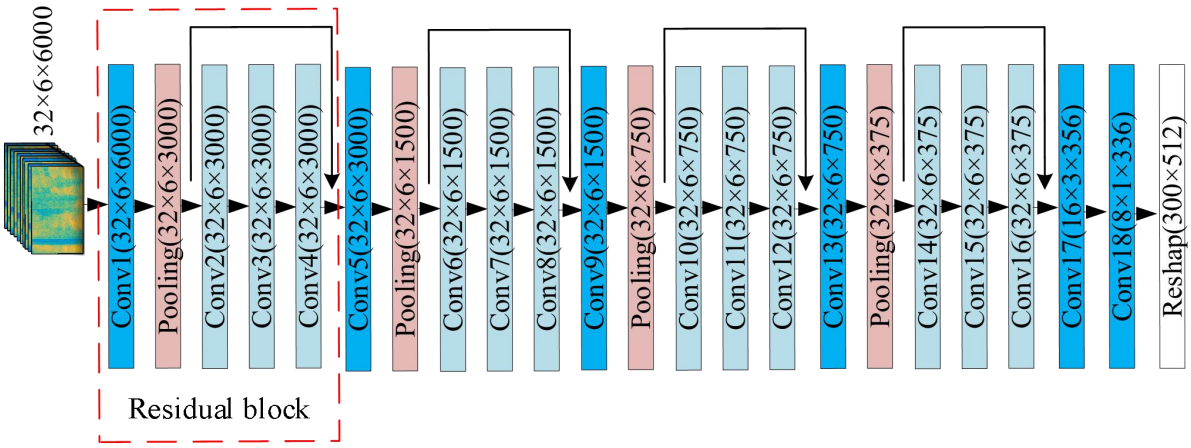

Fig. 4. The structure of 3D convolutional network based on ResNet

using more than two blocks [24].

As shown in Fig. 4, the Res3DCNN structure designed in this paper consists of four residual blocks, each containing four convolutional layers and one pooling layer. The size of the convolution kernel of the first layer is 1×1×1, and the size of the remaining three convolution layers is 3×3×3, which is used to extract 3D deep features that are composed of acoustic frequency, modulation frequency, and short-time windows. The size of the pooling layer is 1×1×2, and the step size is 1×1×1. The purpose of this setup is to pool in time, and as much as possible, to preserve the frequency and channel characteristics. After the four residual blocks, two convolutional layers are added. The size of the convolution kernel is 2×2×20, and the step size is 2×2×1. The setup further convolves in the three directions. Finally, the output data are reshaped into two-dimensional data of 300×512 size as the input of the ARNN module.

Each convolution layer is followed by a batch normalization layer (BN) and a ReLU activation function layer to normalize the data distribution and make the gradient more predictable and stable while being able to speed up training. The residual network transfers shallow feature information to different levels of the convolutional structure to compensate for the features lost in the deep CNN convolution process, which can solve the problem of exploding or disappearing gradients and has a lower error rate in deeper networks.

C. Attention-based recurrent neural networks

The post-forget gate

LSTM is a special type of RNN [25] that can learn long-term dependent information. The network structure is shown in Fig. 5a, which is mainly composed of the forget gate, input gate and output gate. The forget gate of the LSTM unit is used to determine what information should be discarded in the unit state at the previous moment. The calculation formula of the forget gate is shown in formula (1):

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

Where h_{t-1} represents the output of the hidden layer at the previous moment, x_t the input at the current moment, W_f the weight of the forgetting gate obtained by training, b_f the bias value of the neural network, σ the logical sigmoid function. The forget gate updates the current state by calculating the weighted sum of the new and old unit states. The update algorithm of the unit state is related to the output of the hidden layer at the previous time and the input at the current time. However, when the network structure is complex and the quantity of data is large, the computational complexity increases and the recognition efficiency decreases. In response to this problem, an attention gate was proposed by Xie et al. to replace the forget gate of traditional LSTM and the experimental results show that the recognition efficiency of the modified LSTM improved [26]. Inspired by this, we modified the forget gate of the traditional LSTM to reduce the number of calculations without sacrificing performance. The formula is as follows:

$$f_t = \sigma(W_f \times C_{t-1} + b_f) \tag{2}$$

Where C_{t-1} represents the unit state at the previous moment, $W_f \in R^{N \times N}$ the parameter to be trained, N the number of hidden units. Here, we use the post-forget gate to directly discard some unimportant features from C_{t-1}, and then add the information learned from x_t and h_{t-1} as the

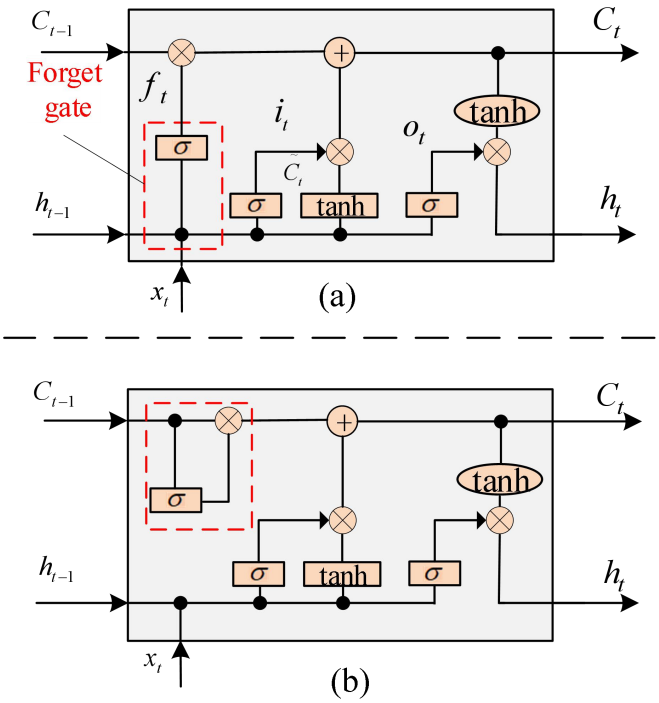

Fig. 5. Traditional LSTM and improved LSTM network structure

unit state at the next moment. Because the new information was learned from the input at the previous moment but did not go through the forget gate, we choose to add the required content at the first moment and forget it at the next moment. The improved LSTM network is shown in Fig. 5(b). Compared with formula (1), it is found that the dimension here is smaller, because the formula is not used x_t and h_{t-1} participate in the calculation. We refer to the modified gate structure as the post-forget gate. The experimental results show that using formula (2) to update the cell state will not affect the performance of the final LSTM mode.

Attention-based recurrent neural networks

The network block diagram of the attention-based recurrent neural network (ARNN) model is shown in Fig. 6. In this paper, a bidirectional LSTM network structure is used, which enables the model to maintain contextual links during the recognition process. The core idea of the attention mechanism is that the human brain pays unbalanced attention to the whole picture, with some weight distinction [27]. Therefore, an attention layer is added after the bidirectional LSTM network in this paper to focus on emotionally relevant parts and generate a discourse-level representation for SER [28].

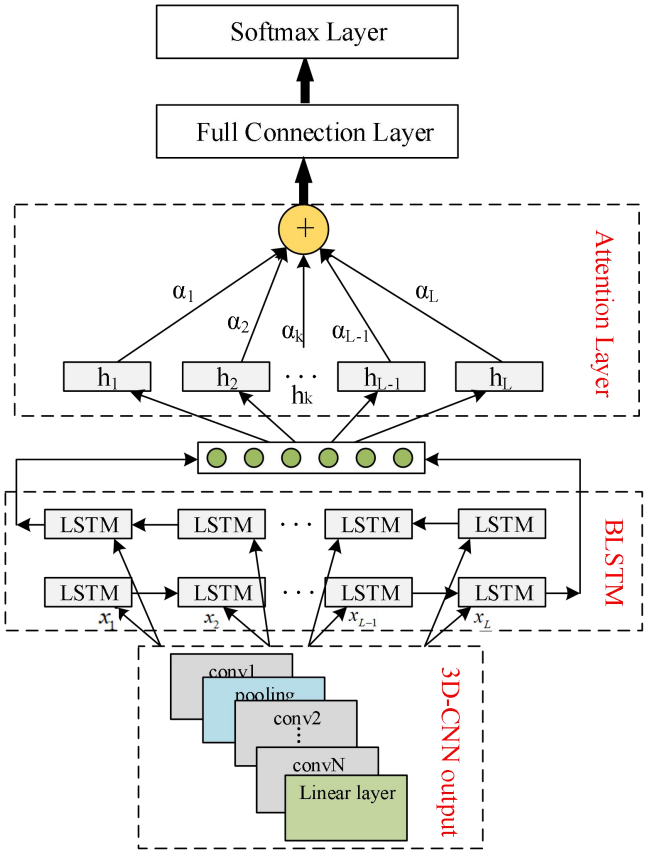

Fig. 6. The ARNN model

In this model, the BLSTM has 512 bidirectional hidden units. Then, we create a new sequence of shape L×1024, place it in the attention layer, and finally generate a new sequence h.

D. The proposed method

The proposed SER system consists of three components: the spectrogram extraction module, a Res3DCNN module and an ARNN module. The general framework is shown in Fig. 7. The spectrogram extraction module processes the speech signal and extracts the spectral-temporal representation as the input of the Res3DCNN. Res3DCNN is used to extract deep spectral-temporal features from the spectral-temporal representation as input to the ARNN module. BLSTM is the process of extracting the long-term dependencies of contextual links in speech signals and then reducing the weight of the silent frame and emotion-independent frame through the attention layer. Finally, the feature information is input into the softmax layer for classification via a fully connected layer.

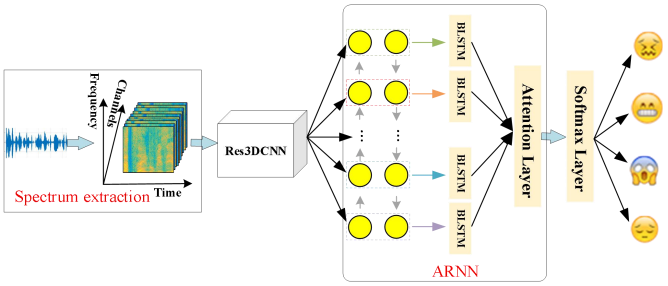

Fig. 7. The general system of speech emotion recognition

III. Experiment and Experiment Analysis

In this part, the database and experimental environment are introduced firstly. Then, the experiment and parameters analysis of the proposed model are demonstrated in detail. Finally, the parameters and experimental results are compared with other models.

A. A Experimental database of emotional speech and experimental setup

Experimental database of emotional speech

For testing the performance of our model, we used the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [29] and the Berlin Emotion database (EMO-DB) [30]. EMO-DB is a German emotional speech database recorded by the Berlin University of Technology. It was recorded by 10 actors in a professional recording studio to imitate 7 emotions on 10 statements. It contains a total of 535 sentences. The mean length of all the audio is 2.7 s. As the CNN input length must be equal to all samples, so we set the maximal length to 3.5 s (mean duration plus standard deviation). A longer audio was cut at 3.5 s, and shorter turns were filled with zero. IEMOCAP was recorded from the

TABLE I
IEMOCAP AND EMO-DB DATABASE SENTIMENT DISTRIBUTION

Emotion	Anger	Happiness	Sadness	Neutrality	Boredom	Disgust	Fear	Total
IEMOCAP	1103	1636	1084	1708	-	-	-	5531
EMO-DB	127	71	62	79	81	46	69	535

conversations of two professional actors, including five conversations by two speakers (a male and a female). These data have an uneven distribution of emotion classes, so happiness and excitement were combined into pleasure classes. After processing in a similar way, we extract the 5,531 conversations in it for experiments, including four emotions. The mean length of all the audio segments is 4.55 s. We set the maximum length to 7.5 s(mean duration plus standard deviation). A longer audio was cut at 7.5 s, shorter turns were filled with zero.The sentiment distribution of the two databases is shown in Table I.

Experimental setup

The experimental hardware conditions were an Intel Xeon E5 CPU and NVIDIA 2080ti GPU. The system environment is Ubuntu 16.04 LTS, and the experimental tool was MATLAB 2016b. In the experiment, the TensorFlow toolkit [31] was used to complete the construction of the proposed network model and the implementation of the training algorithm. For all random weight initializations, we chose L2 regularization. Using the Adam optimization method, all parameters of the model were simultaneously optimized to minimize the chances of having a cross-entropy objective. The detailed training parameters used in the training verification are shown in Table II. A 10-fold cross validation technique is used to verify the model designed in this paper.

TABLE II
TRAINING PARAMETERS

Parameter	Decay	Learning rate	Batch size	Epoch number
Value	1e-4	0.01	100	1,000

B. *Model performance testing and analysis*

For testing the recognition accuracy, convergence speed, and loss value of the residual network and the improved LSTM structure, four sets of experiments were conducted on the EMO-DB dataset. The models were 3DACRNN, improved 3DACRNN, 3DACRNN with ResNet, and improved 3DACRNN with ResNet. Fig. 8, Fig. 9, and Fig. 10 show the changes in training accuracy, test accuracy, and loss value with the number of iterations before and after adding a residual network and before and after improving LSTM.

It can be clearly seen from Fig. 8 and Fig. 9 that the training accuracy and testing accuracy of the 3DACRNN model increased after the addition of the residual structure. When the iteration is about 400 times, the testing recognition accuracy of the improved 3DACRNN with ResNet enters a steady state, which is faster than the other three models. Experimental comparison before and after improvement of LSTM structure shows that the post-forget gate structure proposed in this paper not only retained the performance of the traditional LSTM but also excelled in reducing computational complexity. From the results of the 3DACRNN and Res3DACRNN experiments, it can be seen that the introduction of the residual network improved the training accuracy. As shown in Fig. 10, the convergence speed of 3DACRNN with the residual network was significantly faster, and the loss value is small. The improved LSTM had a smaller improvement in the loss value and convergence speed compared with the traditional

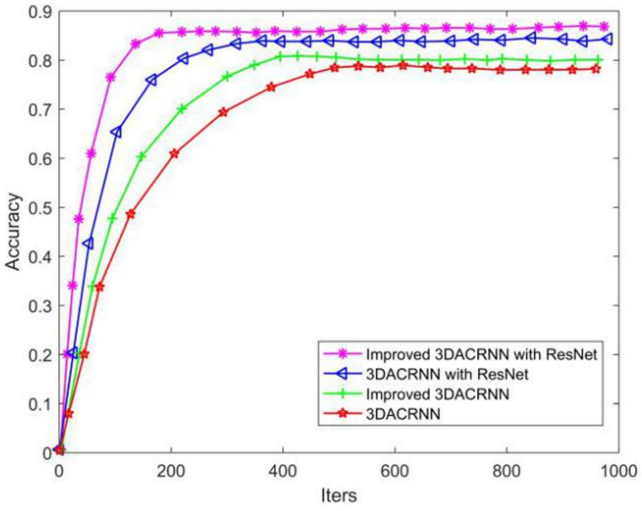

Fig. 8. Comparison of training accuracy

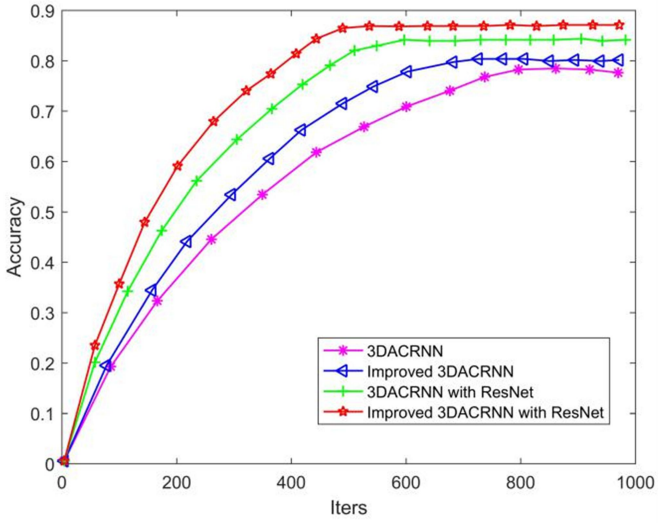

Fig. 9. Comparison of test accuracy

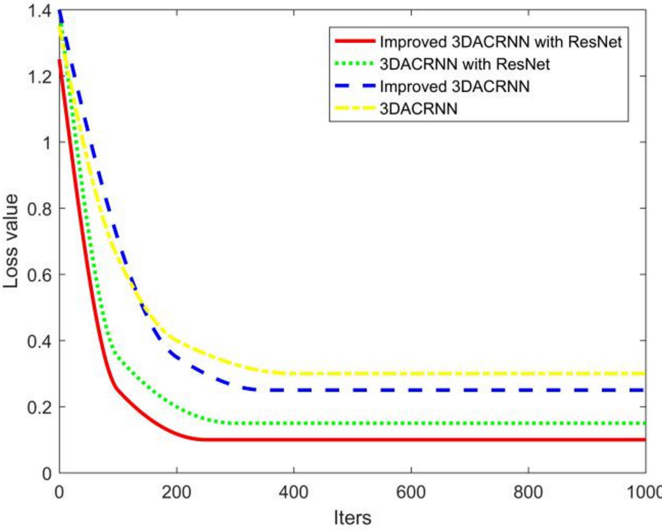

Fig. 10. Comparison of loss value

LSTM. In summary, the performance of 3DACRNN with the residual network was better than that of 3DACRNN without the residual network, and the improved LSTM network had a faster convergence speed than the unimproved LSTM network. The superiority of the model was verified.

C. Experimental results

From the above experiments, it can be seen that both the residual network and the improved LSTM structure can improve the performance of SER. In order to further analyse the performance of the Res3DACRNN model, it is tested on the EMO-DB and IEMOCAP databases, and The statistical results of each emotion are shown in Fig. 11 and Fig. 12. It shows that the Res3DACRNN model has a good recognition effect on neutrality and sadness, and the recognition accuracy on the EMO-DB and IEMOCAP databases was over 90% and 70%, respectively. It had the best recognition accuracy for the sad emotion, even reaching 97.96% on the EMO-DB database, which is due to the more obvious energy value characteristics of this emotion. The lowest recognition accuracy was observed for the happy emotion, not reaching 50% on both EMO-DB and IEMOCAP databases. In the EMO-DB, 45.32% of the happy samples were misidentified as angry. On the IEMOCAP, 47.58% of the happy samples were misidentified as angry and 12.85% were misidentified as neutral. The reason for this phenomenon is that the two emotions, happiness and anger, have similar activation levels, while neutrality is in the middle of the activation valence space.

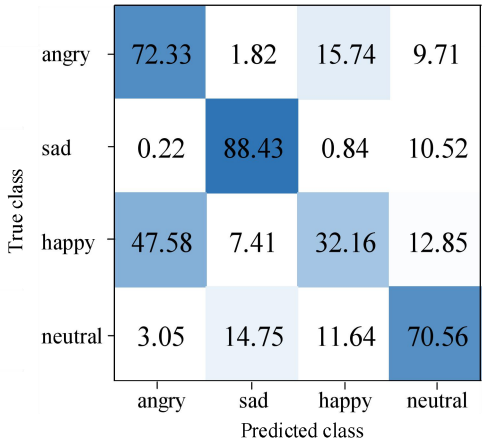

Fig. 11. Comparison of training accuracy on IEMOCAP database

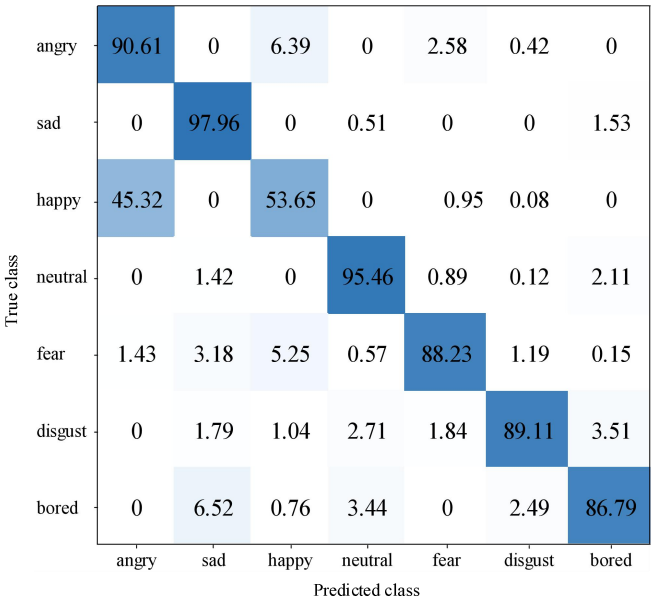

Fig. 12. Comparison of training accuracy on EMO-DB database

D. Experimental results

In order to furtherly illustrated the superiority of the proposed model, we compared the recognition of different models. The compared results were as shown in Table III. Due to the unbalanced distribution of the various emotion classes, we evaluated each model using the unweighted average recall (UAR) to obtain reliable results.

The literature [32] used a convolutional neural network consisting of three convolutional layers to filter the input spectrogram in both time and frequency dimensions, while the literature [33] used the CRNN model to extract features. A comparison between the two showed that the CRNN model performed better than the CNN. In [34], the authors used an autoencoder to learn affective features directly from the linguistic signal and used a bidirectional LSTM network to identify the final affective states. In [35], the authors conducted experiments using an LSTM attention mechanism, and the recognition performance was better than that of the LSTM alone model, demonstrating that the attention mechanism can indeed simulate the human eye to achieve attention focused on emotionally intense areas. In [11], the authors integrated unsupervised autoencoder learning representations into an attention-based CRNN emotion classifier to improve the recognition accuracy. It is also evident from the comparison between [12] and [13] that attention mechanisms are important in the field of SER. The comparison between [11] and [13] shows the better performance of the 3DCNN compared to the traditional CNN. Compared to our experimental results, our proposed model performs better, which proves that the application of residuals and improved LSTM in SER experiments will give better identification results.

TABLE III
COMPARISON OF PROPOSED METHOD AND OTHER METHODS ON IEMOCAP DATABASE AND EMO-DB DATABASES

Literature	Method	UAR(%)	
		EMO-DB	IEMOCAP
Ref[32]	CNN	75.39	
Ref[33]	CRNN	80.00	
Ref[34]	BLSTM		52.80
Ref[35]	BLSTM+A		58.80
Ref[11]	ACRNN		59.54
Ref[12]	3DCRNN		60.93
Ref[13]	3DACRNN	82.82	64.74
Proposed method	Res3DACRNN	85.93	66.87

IV. CONCLUSION

In this paper, a 3D attention convolution recurrent neural network based on a residual network (Res3DACRNN) model was proposed for speech emotion recognition, which reduces the gradient disappearance or explosion due to the increase in neural network layers and the weak spatiotemporal correlation problem in speech emotion recognition. First, the spectrum-time representation is extracted from the speech signal as the Res3DCNN input, and the depth spectrum-time feature is extracted. Then, the output of Res3DCNN is input into an improved bidirectional LSTM network to extract long-term dependencies. Finally, the attention layer is used to focus on the prominent part of emotion representation, which can reduce the impact of redundant information on SER and generate discourse-level

emotion feature representation for SER. The improved LSTM reduces the computational complexity and improves the training speed. Experiments on the IEMOCAP and Emo-DB databases are compared with other similar models. The results show that the proposed Res3DACRNN model is better than other deep learning models with better performance and can effectively improve the recognition of SER.

While the model presented in this paper is useful for the study of speech emotion recognition, there are some areas for improvement:

(1) Extracting the spectral-temporal representation is very time-consuming, and subsequent improvements are needed to improve the real-time performance.

(2) The computational load of the 3DCNN model is very large, and although this paper has been improved this problem, the computational load is still very large, and subsequent research needs to be carried out to address the problem of large computational load.

(3) The current model still has a very low recognition rate for pleasure.

REFERENCES

[1] Masdiyasa, I. Gede Susrama, I. Ketut Eddy Purnama, Mauridhi Hery Purnomo. "A New Method to Improve Movement Tracking of Human Sperms," IAENG International Journal of Computer Science, vol.45, no.4, pp. 531-539, 2018.

[2] Sung-Hyun Yoon, Min-Sung Koh, and Ha-Jin Yu, "Fuzzy Restricted Boltzmann Machine based Probabilistic Linear Discriminant Analysis for Noise-Robust Text-Dependent Speaker Verification on Short Utterances," IAENG International Journal of Computer Science, vol. 47, no.3, pp. 468-480, 2020

[3] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 5135-5139.

[4] Z. Zeng, J. Tu, B. M. Pianfetti and T. S. Huang, "Audio-Visual Affective Expression Recognition Through Multistream Fused HMM," in IEEE Transactions on Multimedia, vol. 10, no. 4, pp. 570-577, June 2008.

[5] A. D. Dileep and C. C. Sekhar, "GMM-Based Intermediate Matching Kernel for Classification of Varying Length Patterns of Long Duration Speech Using Support Vector Machines," in IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 8, pp. 1421-1432, Aug. 2014.

[6] J. Umamaheswari and A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 177-183.

[7] Santosa and Stefanus, "Wood Types Classification using Back-Propagation Neural Network based on Genetic Algorithm with Gray Level Co-occurrence Matrix for Features Extraction," IAENG International Journal of Computer Science, vol.46, no.2, pp. 149-155, 2019.

[8] K. Tarunika, R. B. Pradeeba and P. Aruna, "Applying Machine Learning Techniques for Speech Emotion Recognition," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bangalore, 2018, pp. 1-5.

[9] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition using Deep Neural Network and Extreme Learning Machine," in Proceedings of Interspeech, 2014.

[10] Zahra Berradi, Mohamed Lazaar, Hicham Omara, and Oussama Mahboub, "Effect of Architecture in Recurrent Neural Network Applied on the Prediction of Stock Price," IAENG International Journal of Computer Science, vol. 47, no.3, pp. 436-441, 2020.

[11] M. Neumann and N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 7390-7394.

[12] Z. Peng, Z. Zhu, M. Unoki, J. Dang and M. Akagi, "Auditory-Inspired End-to-End Speech Emotion Recognition Using 3D Convolutional Recurrent Neural Networks Based on Spectral-Temporal Representation," 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, 2018, pp. 1-6.

[13] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrentneural networks with attention model for speech emotion recognition," IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440-1444, Oct. 2018.

[14] P. Yang, W. Zhao, R. Ni, and Y. Zhao, "Source camera identification based on content-adaptive fusion residual network," Pattern Recognition Letters, vol. 119, pp. 195-204, 2017.

[15] Huang, Che Wei, and S. S. Narayanan. "Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition." INTERSPEECH 2016, pp. 1387-1391.

[16] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognitionusing deep convolutional neural network and discriminant temporal pyramid matching," IEEE Trans. Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018.

[17] D. Wu, Z. Tao, Y. Wu, C. Shen, Z. Xiao, and X. Zhang, "Speech endpoint detection in noisy environment using Spectrogram Boundary Factor," 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, 2016, pp. 964-968.

[18] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang and M. Akagi, "Speech Emotion Recognition Using 3D Convolutions and Attention-Based Sliding Recurrent Networks With Auditory Front-Ends," in IEEE Access, vol. 8, pp. 16560-16572, 2020.

[19] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Modulation Spectral Features for Predicting Vocal Emotion Recognition by Simulated Cochlear Implants," in Proceedings of interspeech, 2016, pp. 262-266.

[20] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo and T. Liu, "Residual Networks of Residual Networks: Multilevel Residual Networks," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 6, pp. 1303-1314, June 2018.

[21] Ahmad Alzu'bi, Abbes Amira, and Naeem Ramzan, "Learning Transfer Using Deep Convolutional Features for Remote Sensing Image Retrieval," IAENG International Journal of Computer Science, vol. 46, no.4, pp. 637-644, 2019.

[22] Rastgoo, Razieh, Kourosh Kiani, and Sergio Escalera, "Hand sign language recognition using multi-view hand skeleton," Expert Systems With Applications vol. 150, pp. 113336-113348, July 2020.

[23] Li Wuke, Yin Guangluan, and Chen Xiaoxiao, "Application of Deep Extreme Learning Machine in Network Intrusion Detection Systems," IAENG International Journal of Computer Science, vol. 47, no.2, pp. 136-143, 2020

[24] Hashida, Shuichi, Keiichi Tamura and Tatsuhiro Sakai. "Classifying Tweets using Convolutional Neural Networks with Multi-Channel Distributed Representation," IAENG International Journal of Computer Science, vol. 46, no. 1, pp. 68-75, 2019.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput, vol. 9, pp. 1735-1780, 1997.

[26] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou and B. Schuller, "Speech Emotion Classification Using Attention-Based LSTM," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1675-1685, Nov. 2019.

[27] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower and S. Kim, "IEMOCAP: interactive emotional dyadic motion capture database." Language Resources and Evaluation, 42.4(2008), pp. 335-359.

[28] K. Cheng, Y. Yue and Z. Song, "Sentiment Classification Based on Part-of-Speech and Self-Attention Mechanism," in IEEE Access, vol. 8, pp. 16387-16396, 2020.

[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in Interspeech, vol. 5, Lisbon, Portugal, 2005, pp. 1517-1520.

[30] Y. Chen, R. Chen, M. Liu, A. Xiao, D. Wu and S. Zhao, "Indoor Visual Positioning Aided by CNN-Based Image Retrieval: Training-Free, 3D Modeling-Free," Sensors, 18, 2692, 2018.

[31] M. Aqib, R. Mehmood, A. Alzahrani, I. Katib and A. Albeshri, "Altowaijri, S.M. Smarter Traffic Prediction Using Big Data, In-Memory Computing, Deep Learning and GPUs," Sensors, 19, 2206,2019.

[32] N. Vrebčević, I. Mijić and D. Petrinović, "Emotion Classification Based on Convolutional Neural Network Using Speech Data," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 1007-1012.

[33] S. Basu, J. Chakraborty and M. Aftabuddin, “Emotion recognition from speech using convolutional neural network with recurrent neural network architecture,” 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2017, pp. 333-336.
[34] S. Ghosh, E. Laksana, L.P. Morency, and S. Scherer, “Representation learning for speech emotion recognition,” in Proc. Interspeech, 2016, pp. 3603–3607.
[35] S. Mirsamadi, E. Barsoum and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2227-2231.