Speech Emotion Recognition based on Multi-Level Residual Convolutional Neural Networks

Kai Zheng, ZhiGuang Xia, Yi Zhang, Xuan Xu,Yaqin Fu

***Abstract*—Speech emotion recognition, using the convolutional neural networks (CNN) model, is challenging due to the problem of features loss and the decrease of recognition accuracy. To address this issue, a Multi-level residual CNN model is proposed is this paper. In this model, the speech signals are converted into spectrogram, then the multi-level residual identity maps are introduced to compensate the missing features in the CNN during the convolution process, so as to improve the recognition accuracy of speech emotion. The research results show that the Multi-level residual CNN can achieve 74.36% recognition accuracy on the EMO-DB dataset, which has better performance than traditional deep CNN method.**

***Index Terms*—CNN, residual network, speech emotion recognition**

I. Introduction

Speech signal is the most frequently used in human-computer interaction[1]. With the rapid development of artificial intelligence technology, new requirements for high-performance and fast Speech Emotion Recognition (SER) systems have been put forward in different application scenarios. For example, hospital nurse service robots require 24 hours to monitor the emotional state of the patient in the ward without interruption, so that the doctor can make a comprehensive judgment and provide better treatment for the patient's condition. The vehicle monitoring system needs to monitor the driver's emotional state to avoid the driver's emotional excitement which will cause wrong judgment [2]. There are three main challenges to speech emotion recognition. Firstly, the selection of emotional features is tough, as determining which features can represent emotions is hard [3]. Secondly, the recognition of speech emotions is influenced by varieties of factors such as the accent of speaker, the speech style. and the poor generalized model [2], [3]. Finally, the speech signal is affected by noise and multiple sources in the environment, which reduce the efficiency of emotion recognition.

With the advent of high-performance computers, image-processing has become easy to implement. A new hotspot of speech emotion recognition research is transforming speech signals into speech spectrogram for the recognition object of speech emotion recognition. This method can avoid the cumbersome process of manual feature extraction, reducing the workload in the modeling and training process. Moreover, it can not only reflect the energy characteristics of the speech signal, but also the texture features reflecting the rhythm changes in the speech signal. In recent years, many researchers have begun to study the speech emotion recognition based on the spectrogram, and achieved good results. Wang K-C et al. first obtained the spectrogram of the four emotions of anger, sadness, happiness and fear in the database, then enhance the contrast, and use Laws masks to represent the types of emotions. they achieved recognition accuracy of 73.06%, 77.42%, and 65.20% on the EMO-DB, eNTERFACE, and KHUSCEmoDB databases by SVM [4]. Zhang S et al. applied the Mel spectrogram and the AlexNet deep learning model in SER system and extract features by discriminating temporal pyramid mapping. The results revealed that the pre-trained AlexNet deep learning model performed well and achieved 72.35 % on the EMO-DB dataset [5]. Prasomphan S et al. applied ANN (Artificial Neural Networks) and the score map to identify five emotions (anger, sadness, happiness, neutrality and fear). They divided the spectrogram into 16 small areas and judge the scores of each area, and obtained a good recognition rate [6]. Badshah AM et al. proposed a deep CNN (Convolutional Neural Network) which contains three convolutional layers and two fully connected layers, the recognition accuracy reached 73.8% on the EMO-DB dataset [7]. Fayek, H. M et al. illustrated the influence of deep learning structure on the rate of speech emotion recognition. They explored the different structures of ANN, RNN (Recurrent Neural Network), CNN, LSTM-RNN for anger, happiness, sadness, neutral. The result indicated that two convolutional layers and the two fully connected layers have the best performance, and the recognition rate achieved 64.78% on the IEMOCAP dataset [8], [9].

However, as the number of convolution layer increases of

Manuscript received July 15, 2019; revised February14, 2020. This work was supported in part by Chongqing Research Program of Basic Research and Frontier Technology (Grant No. cstc2017jcyjAX0151), the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN201800634).

Kai Zheng is with Engineering Research Center for Information Accessibility and Service Robots, Chongqing University Posts and Telecommunications, Chongqing 400065 China (e-mail: zhengkai2001@163.com)

ZhiGuang Xia, Yi Zhang, Xuan Xu,Yaqin Fu are with Engineering Research Center for Information Accessibility and Service Robots, Chongqing University Posts and Telecommunications, Chongqing 400065 China(15140013568@163.com,1107529848@qq.com, 2295129267@qq.com, 2508174350@qq.com)

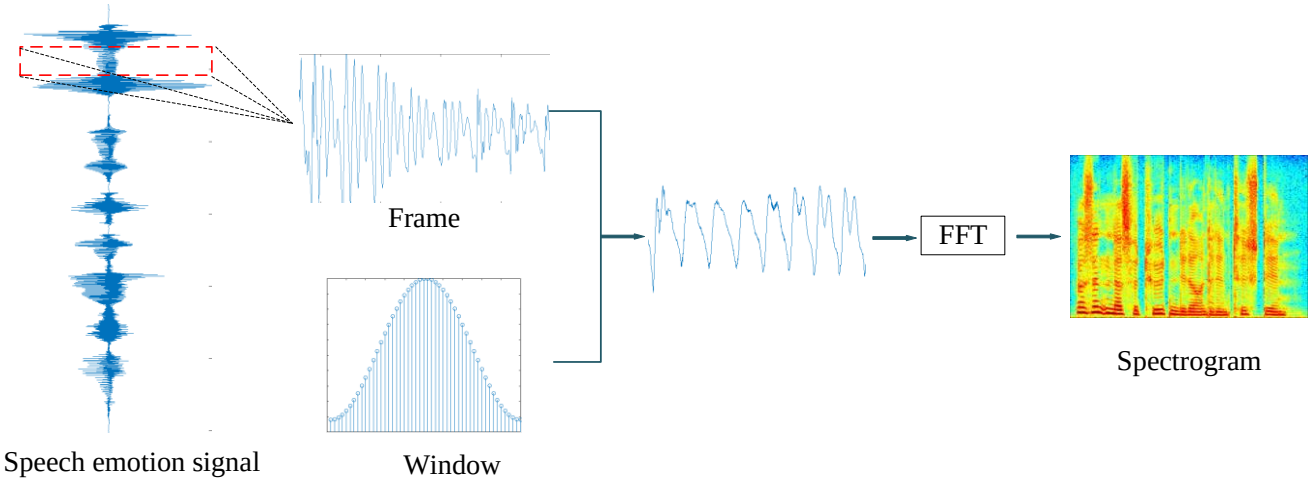

Fig. 1. Process of make spectrogram

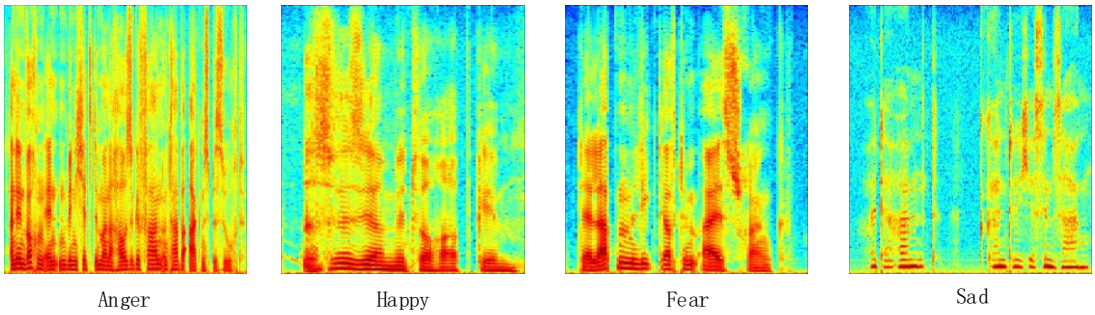

Fig. 2. Spectrogram samples

using CNN model, the original features will be gradually lost, resulting in a decrease in the recognition accuracy [10], [11]. Aiming at this problem, this paper proposed a Multi-level residual CNN structure [12]. Multi-level residual CNN improves the recognition rate by transmitting the original feature information to different levels of convolution structure [13], thereby compensating for the missing feature information in the CNN convolution process.

This paper is organized as followed: Section 2 introduces the proposed model in detail. Section 3 introduces the preparation work. Section 4 demonstrates the experiment and analysis. Finally, conclusions are drawn in Section 5.

II. Methods

A. *Spectrogram*

The spectrogram is wildly used in representing three-dimensional information in a two-dimensional image. In the spectrogram, the horizontal axis represents time while the vertical axis represents the frequency value of the speech signal [14]. The energy value of the speech signal is represented by color, here the low energy is represented by dark blue while the stronger energy is red. The frequency distribution and energy change of the speech signal with emotion often have certain rules, so the spectrogram contains the characteristic information such as the fundamental frequency, unvoiced sound and popping sound [15].

The process of obtain the spectrogram is shown in fig. 1.

In this process. The basic idea is that the speech signal can be regarded as a smooth linear signal in each frame of the signal [16]. The resolution of the spectrogram [17] will have a greater impact on the training of the network model. Higher resolution will result in slower training and difficulty in fitting while too low resolution will result in lower recognition accuracy. Therefore, the resolution of the spectrogram is set to 512x512 [18]. The sample of the emotional speech spectrogram used in this paper shown in fig. 2.

B. *Convolutional Neural Network*

The convolutional neural network is a deep learning network structure that extracts the regional features of an image layer by the convolution kernel and the pooling layer. By setting the size and step size of the convolution kernel [19], it can be traversed in the image as a small block area. The pooled layer is employed to reduce the dimension of the convolved image, which cut down the number of network weights and improve the training efficiency. Since each convolution kernel of the convolutional layer is equivalent to a feature extractor, CNN can extract low/mid/high-level features in the image [20]. The more layers of the network, the more features can be extracted to different levels. The realization of convolutional neural networks is mainly based on three important ideas: sparse interaction, parameter sharing and isomorphic representation [21]. The calculation

rule between CNN different connection layers is as shown in formula (1):

$$h^{(l)} = y^{(l-1)} * w^{(l)} + b^{(l)} \quad (1)$$

Where $w^{(l)} \in R^{(m*j*k)}$ represents the tensor of convolution kernel, j , k respectively the width and height of the convolution kernel m the number of convolution kernels, l the convolutional layer, $b^{(l)}$ the bias value of the neural network, and $y^{(l-1)}$ the previous convolution output of the convolutional layer. The convolved value also needs to be pooled [22,23]. The pooling layer can compress the input feature map, simplifying the network complexity, and extract the main features.

C. Residual structure

The main idea of the residual structure [24] is inputting the original information directly into the convolutional layer. The principle is shown in Fig. 3. Assuming that the expected output after two layers of convolution is $H(x)$, the optimization learning goal of the next layer is $F(x) = H(x) - x$ after the congruent mapping, where x is the output value of the upper layer of the convolution network. Therefore, the optimization target is $H(x) - x$, which converted into the difference between the output and the input. The residual can speed up the training and improve the training effect [25].

D. Proposed model

CNN can extract the deep feature information [26], however, with the deepening of the convolutional layer, the extracted feature information becomes less and less semantic. Consequently, the problem of losing original feature [27] information is easy to occur, reducing the recognition accuracy. Since the recognition accuracy can be effectively

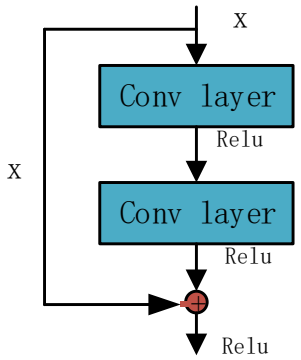

Fig. 3. Residual structure

improved as the residual structure can connect the original information. Therefore, this paper proposes a new model by adding the residual structure to the CNN structure, which is defined as Multi-level residual CNN.As it is shown in Fig.4.

This model is a deep neural network containing multiple convolution pooling layers and residual structures. The spectrogram containing the original information is first traversed by the convolution kernel of the convolutional layer, and the extracted feature map [28] information is treated as the input information of the next layer convolution. After the convoluting the multiple convolution layers, the original information is gradually lost [29]. The extracted feature information is more and more difficult to characterize the information of the spectrogram [30]. By introducing the residual structure at the appropriate position, the feature information can be compensated, while the depth of the convolutional layer can be deepened [31]. The high-dimensional features extracted by multiple residual structures and convolutional layers need to be calculated by the fully connected neural network [32], and the final results

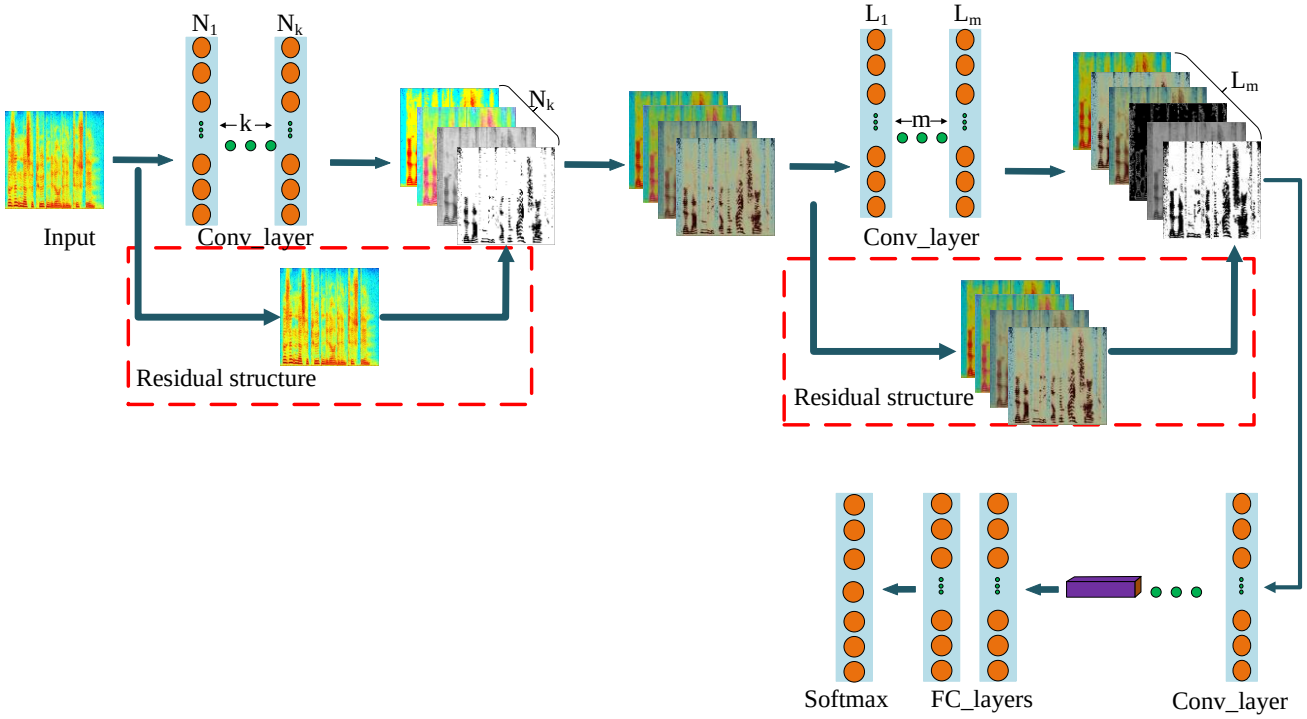

Fig. 4. Multi-level residual CNN

are classified and output by the softmax [33] classifier. The loss function used in this paper is cross entropy loss function [34], which is defined as in formula (2).

$$C = -\sum_{K=1}^{K} \hat{y}_K \log(y_K^L) \tag{2}$$

Where C is value of the loss function, K representing the different categories, $\hat{y} \in \{0,1\}^K$ the label changing from 1 to K , $y^{(L)}$ the actual output of the network model. Deep neural network [35] training is achieved by a suitable optimizer [36] to minimize the loss function. where the gradient descent [37]is the most commonly method. The RMSProp algorithm [38] employed in this paper is one of the variants of the gradient descent algorithm which has the characteristics of full parameter self-adaptation and it is important for improving the training speed. RMSProp is demonstrated by formula (3) and (4). Where r is the slip rate of the squared value of the gradient, w the decay rate, and α the learning rate.

$$r := \eta r + (1-\eta)(\partial C / \partial w)^2 \tag{3}$$

$$w := w - \alpha \frac{\partial C / \partial w}{\sqrt{r} + \varepsilon} \tag{4}$$

The number of layers of the CNN, the size of the convolution kernel, the connection order of the residual network and the number of neurons in the fully connected layer on the recognition effect will be explored in Part5.

E. Preparation work

The database, the preprocessing of speech signals, the experimental platform and the parameters settings are introduced in this section.

The performance of the model is tested based on the Berlin Database (EMO-DB). EMO-DB is composed of seven different emotions (anger, boredom, fear, disgust, happiness, sadness, neutrality) which expressed by actors. The sampling frequency of the EMO-DB database is 16Khz. The detailed description of the database is shown in Table I.

The EMO-DB database applied in the experiment contains of 535 speech emotion signals. The dataset is expanded courtesy of the difficulty of obtaining a better model that the sample size is too small. The expansion of the dataset mainly includes random cropping, random flipping, graying, adding noise points [39], etc. and the final get 50000 spectrograms, each emotion contains approximately 7000 spectrograms.

In the experiment, the TensorFlow [40] is used to complete the construction of the proposed network model and the implementation of the training algorithm. The parameters of

TABLE II
PLATFORM CONFIGURATION

Devices	GPU	CPU	CUDA	System
Name	GTX1080	E5-2665	3584	Ubuntu16.04

TABLE III
TRAINING PARAMETERS

Parameter	Learning rate	Decay	Iteration	Batch size
Value	0.001	1e-6	1e5	50

TABLE I
DATABASE DISTRIBUTION

Emotion	Anger	Disgust	Fear	Happiness	Sadness	Boredom	Neutrality	Total
Samples	127	46	69	71	62	81	79	535

TABLE IV
EXPERIMENTAL RESULTS

Structures	Accuracy
$Conv(16\times10\times10)+pool+Conv(32\times10\times10)+pool+FC(716)\times2+soft\max$ Reference[7]	73.57
$Conv(120\times11\times11)+pool+Conv(256\times5\times5)+pool+Conv(384\times3\times3)+pool+FC(1024)+FC(2048)+soft\max$ Reference[8]	71.51
$Conv(32\times4\times4)+pool+Conv(64\times8\times8)+pool+Conv(64\times10\times10)+pool+FC(716)\times2+soft\max$	70.63
$Conv(32\times4\times4)+pool+Conv(64\times8\times8)+pool+Conv(64\times10\times10)+pool+FC(716)\times2+soft\max$	72.18
$Conv(32\times4\times4)+pool+Conv(64\times8\times8)+pool+Conv(64\times10\times10)+pool+FC(716)\times2+soft\max$	73.14
$Conv(32\times4\times4)+pool+Conv(64\times8\times8)+pool+Conv(64\times10\times10)+pool+FC(716)\times2+soft\max$	70.72
$Conv(16\times2\times2)+pool+Conv(16\times4\times4)+pool+Conv(32\times8\times8)+pool+Conv(64\times10\times10)+pool+FC(716)+FC(1024)+soft\max$	74.36
$Conv(16\times2\times2)+pool+Conv(16\times4\times4)+pool+Conv(32\times8\times8)+pool+Conv(64\times10\times10)+pool+FC(716)+FC(1024)+soft\max$	69.08
$Conv(16\times2\times2)+pool+Conv(32\times4\times4)+pool+Conv(32\times8\times8)+pool+Conv(64\times8\times8)+pool+FC(716)+FC(1024)+soft\max$	73.29
$Conv(16\times2\times2)+pool+Conv(32\times4\times4)+pool+Conv(32\times8\times8)+pool+Conv(64\times10\times10)+pool+FC(716)+FC(1024)+soft\max$	74.14
$Conv(32\times4\times4)+pool+Conv(64\times8\times8)+pool+Conv(64\times10\times10)+pool+Conv(64\times10\times10)+pool+FC(716)\times2+soft\max$	60.37

server platform and detailed training parameters used in the training verification are shown in Table II and Table III.

III. Results and Result Analysis

In this part, the experiment and parameters analysis of the proposed model are demonstrated in detail. The parameters and comparative study with other method are analyzed.

A. Experiments

The experiment first reproduced the results of the CNN model proposed in [7] and [8]. Then the training and testing of the Multi-level residual CNN model is also introduced. The experimental results were shown in Table IV. The results unveil that the change of the structure has a direct relationship with the recognition accuracy. The recognition accuracy begins to increase when the residual structure is added as well as the number of the convolutional layer is three. As the original information that is before the first convolutional layer is gradually introduced into the deeper convolutional layer, the recognition rate begins to increase slowly. When the original input information is completely introduced before the fully connected layer, the recognition rate starts to decrease because the residual value is too large to fit. When the convolution layer is deepened to the fourth layer, the recognition rate is further improved. The difference is that the maximum value of the recognition rate of the four convolutional layers does not appear in the structure that further introduces the original information into the deep layer. When the original information is completely introduced before the fully connected layer, its recognition rate is greatly reduced. Therefore, when the convolutional layer reaches four layers, continuing to deepen the network does not improve the recognition rate.

By analyzing Table IV, it can be found that the Multi-level residual CNN can reach up to 74.36%, and its structure is composed of four convolutional layers, two fully connected layers and three residual structures.

B. Model performance testing and analysis

Convergence speed and accuracy

In order to test the recognition accuracy and convergence speed of Multi-level residual CNN, the experiment tested the Multi-level residual CNN and the subsequent network model before adding the residual structure. The results are illustrated in Fig. 5 and Fig. 6. It shows the variation of the train accuracy and the test accuracy with iterations before and after of adding the residual structure.

It can be clearly seen from Fig. 5and Fig. 6 that the model recognition rate rises surging after adding the residual structure. When the iteration is about 40,000 times, the test recognition accuracy enters a steady state. On the other hand, the recognition rate of the Multi-level residual CNN model without residual fluctuated greatly and was very unstable. As shown in Fig. 7, Multi-level residual CNN without residual convergence process is relatively stable, and there is no large fluctuation of the loss value. Multi-level residual CNN has the fluctuation of the loss function value, but the convergence speed is significantly faster. In summary, the performance of Multi-level residual CNN is better than the CNN without residual structure. The model performance also shows in Fig.7, which verifies the superiority of the proposed model.

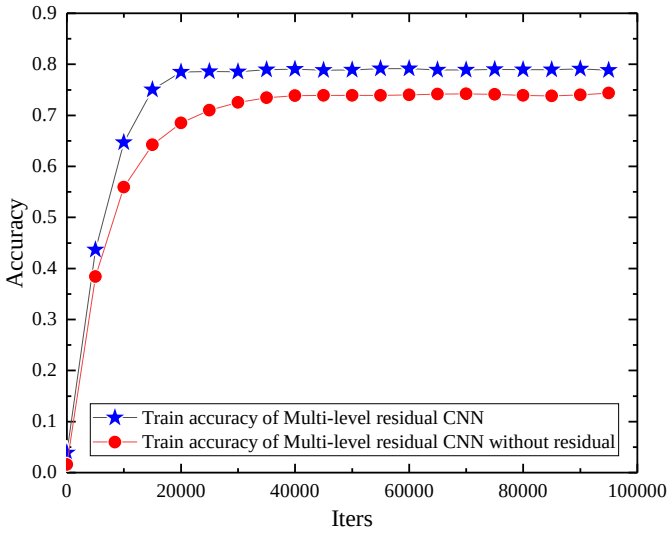

Fig. 5. Train accuracy

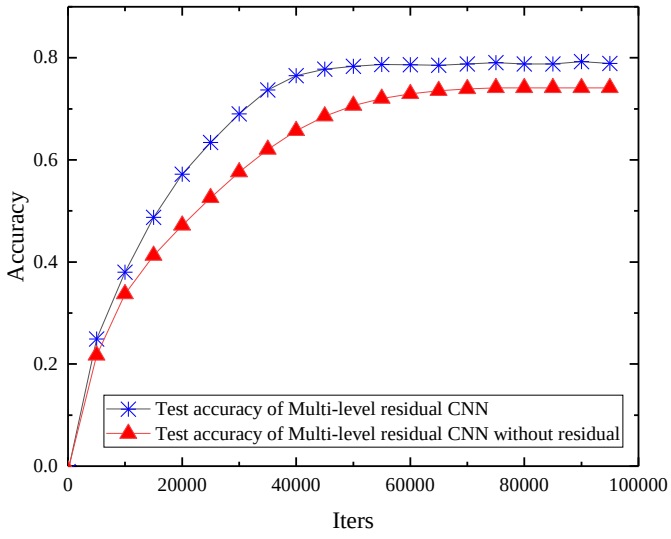

Fig. 6. Test accuracy

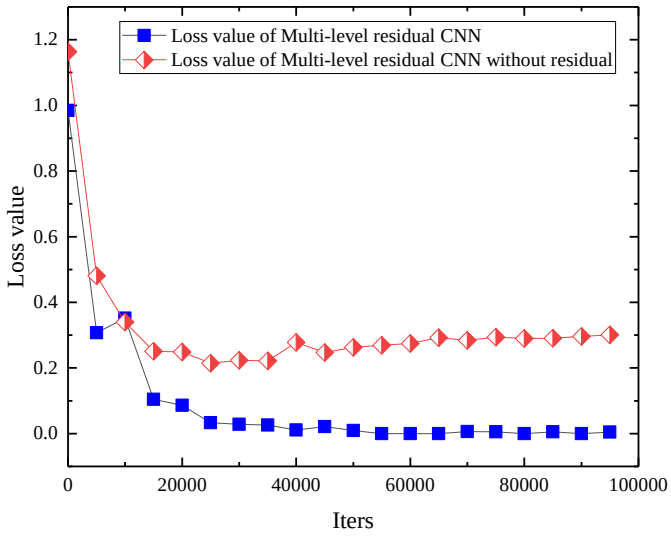

Fig. 7. Loss value

Recognition accuracy for different emotions

From the above experiments it can be concluded that Multi-level residual CNN has better performance of identifying different emotions. For furtherly analyzing the performance of Multi-level residual CNN, it is tested on the EMO-DB dataset. The statistical results are shown in Fig. 8.

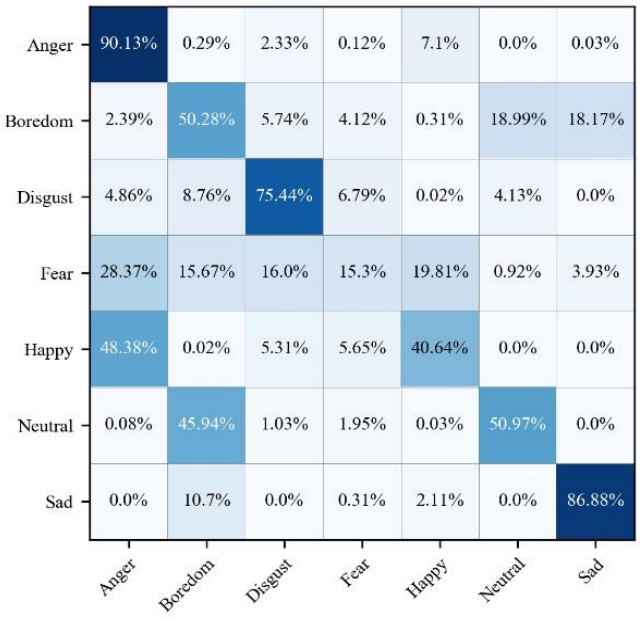

Fig. 8. Confusion matrix

It indicates that Multi-level residual CNN has a good recognition effect on Anger, Boredom, Disgust, Neutral, and Sad, and the recognition accuracy is higher than 70%. Among them, it has the best recognition accuracy for the emotion of Anger and Sad, and the recognition accuracy is higher than 80% owing to the more obvious of the energy value characteristics of these two emotions. However, the recognition rate of Fear and Happy is poor, which is lower than 50%. This because that the two emotional energy values are unstable and difficult to identify. It should be noted that the recognition accuracy is also related to the language type, speech style, and cultural background of the dataset.

Performance comparison

In order to furtherly illustrated the superiority of the proposed model, the recognition effects of different emotions are tested. Table V is a comparison of the overall recognition rate of the model. It can be seen that the overall recognition rate of Multi-level residual CNN is higher than [7] and [8], but the recognition rate without adding a residual structure is slightly lower than [7], and higher than [8]. Fig.9 is a statistical comparison of the rate of recognition of each of the three different models.

We can find that Multi-level residual CNN have higher recognition of Anger, Disgust, and Sad than the other two methods while the recognition rate of the Fear is lower than the other two methods. The recognition rate of Happy, Boredom and Neutral is higher than [7]. The proposed model is lower than [8]. In general, the recognition rate of different types of emotions in the model proposed by this study is similar to [7]. That is, the recognition rate of Anger, Disgust, and Sad is higher while the recognition effect of Fear and Happy is poor. The [8] has a better recognition effect for Boredom and Neutral. In summary, Multi-level residual CNN increases the recognition rate to 74.36%, which can better identify the emotions in the EMO-DB database, proving that the proposed method has better performance.

TABLE V
ACCURACY COMPARISON

Methods	Accuracy
Proposed model in [7]	73.57%
Proposed model in [8]	70.57%
Multi-level residual CNN	74.36%
Multi-level residual CNN without residual	72.15%

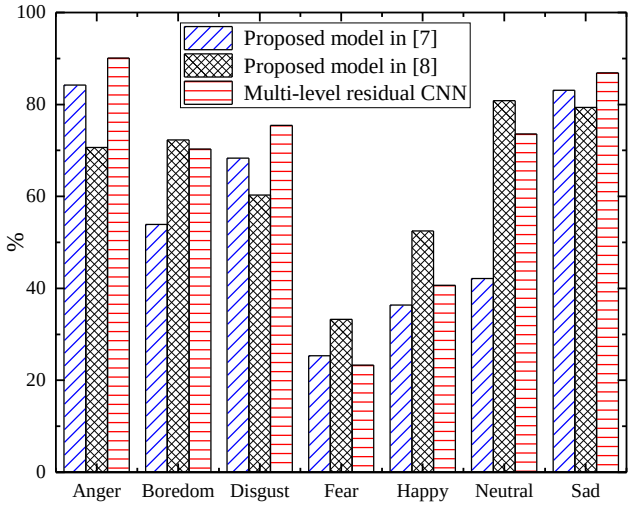

Fig. 9. Different emotion accuracy

IV. CONCLUSION

This paper proposed a Multi-level residual CNN model, which can compensate the missing features of the CNN. In this study, we first turn the speech emotion signal into a spectrogram, and then use the spectrogram to train the Multi-level residual CNN model, and explore the performances with different size and number of CNN convolution kernel, the convolution layer, the connection position of the residual structure, etc. And then compared it with the related other CNN models. The experiment results indicate that the proposed Multi-level residual CNN has better performance than the other deep CNN model and can effectively improve the recognition rate of speech emotion.

The model proposed in this paper maybe helpful for the study of speech emotion recognition using spectrogram and deep learning while there are some improvements we ought to notice:

(1) The process of transforming speech signal into speech spectrogram is relatively time-consuming. Improve the real-time performance of the whole recognition process is the next research direction.

(2) The speech signal is a sequence with strong timing. Solve the sample sequence is the next stage of research.

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] Schuller. B, Rigoll. G, Lang. M, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector," IEEE International Conference on Acoustics ,1, 1: I - 577-80, 2004.

[3] Ayadi. M. E, Kamel, M. S, Karray. F, "Survey on speech emotion recognition: features, classification schemes, and databases," Pattern Recognition, 44(3), 572-587, 2011.

[4] Wang K-C, "Speech emotional classification using texture image information features," International Journal of Signal Processing Systems, 3,1–7, 2015.

[5] Zhang. S, Zhang. S, Huang. T, Wen. G, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," IEEE Transactions on Multimedia, PP (99), 1-1,2017.

[6] Prasomphan. S, "Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. In Systems, Signals and Image Processing (IWSSIP)," International Conference on 2015, pp. 73–76.

[7] Badshah. A. M, Ahmad. J, Rahim. N, Baik. S. W, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," International Conference on Platform Technology & Service. IEEE 2017.

[8] Fayek. H. M, Lech. M, Cavedon. L, "Evaluating deep learning architectures for speech emotion recognition," Neural Networks, 92,60-68, 2017.

[9] Kutlu. H, Avcı. E, "A Novel Method for Classifying Liver and Brain Tumors Using Convolutional Neural Networks, Discrete Wavelet Transform and Long Short-Term Memory Networks," Sensors, 19, 1992，2019.

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

[11] Zhou. S, Hu. Y.-H, Jiang. H, "Multi-View Image Denoising Using Convolutional Neural Network," Sensors, 19, 2597, 2019.

[12] Hashida, Shuichi, Keiichi Tamura, 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.

[13] Bogale. T. E, Vandendorpe. L, "Linearly combined signal energy-based spectrum sensing algorithm for cognitive radio networks with noise variance uncertainty," International Conference on Cognitive Radio Oriented Wireless Networks,80-86, 2013.

[14] Kingsbury. B. E. D, Morgan. N, "Greenberg, S. Robust speech recognition using the modulation spectrogram," Speech Communication,25(1), 117-132,1998.

[15] Gramacki. Artur, "FFT-Based Algorithms for Kernel Density Estimation and Bandwidth Selection," Nonparametric Kernel Density Estimation and Its Computational Aspects, pp, 85-118, 2018.

[16] Grant. A, "Enhancing the Resolution of the Spectrogram Based on a Simple Adaptation Procedure," IEEE Transactions on Signal Processing,60(10),5566-5571, 2012.

[17] Hijmans. R. J, Cameron. S. E, Parra. J. L, Jones. P. G, Jarvis. A. "Very high-resolution interpolated climate surfaces for global land areas," International Journal of Climatology, 25(15), 1965-1978,2010.

[18] Wang. Z, Zhuang. Y, Yang. J, Zhang. H, Dong. W, Wang. M, Hua. L, Liu. B, Shi. L, "A Double Dwell High Sensitivity GPS Acquisition Scheme Using Binarized Convolution Neural Network," Sensors, 18, 1482, 2018.

[19] Rachmadi, Reza Fuad. "A Systematic Evaluation of Shallow Convolutional Neural Network on CIFAR Dataset," IAENG International Journal of Computer Science, vol. 46, no. 2, pp.365-376, 2019.

[20] Santosa, 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.

[21] Kim. H, Jeong. Y.-S, "Sentiment Classification Using Convolutional Neural Networks," Appl. Sci, 9, 2347, 2019.

[22] Bologna. G, "A Simple Convolutional Neural Network with Rule Extraction," Appl. Sci, 9, 2411, 2019.

[23] Giusti. A, Dan. C. C, Masci. J, Gambardella. L. M, Schmidhuber. J, "Fast image scanning with deep max-pooling convolutional neural networks," IEEE International Conference on Image Processing,4034-4038, 2013.

[24] Lei. H, Tao. H, Feng. Z, Zhen. Y, Jing. Q, Elazab. A, "A deeply supervised residual network for hep-2 cell classification via cross-modal transfer learning," Pattern Recognition, 79,290-302, 2018.

[25] Baraniuk. R. G, Flandrin. P, Janssen. A. J. E. M, Michel. O. J. J, "Measuring time-frequency information content using the renyi entropies," IEEE Transactions on Information Theory, 47(4), 1391-1409, 2001.

[26] Shafiee. M. J, Barshan. E, Li. F, Chwyl. B, Karg. M, Scharfenberger. C, "Learning Efficient Deep Feature Representations via Transgenerational Genetic Transmission of Environmental Information During Evolutionary Synthesis of Deep Neural Networks," IEEE International Conference on Computer Vision Workshop, pp, 979-986, 2017.

[27] Pearlmutter. B.A, "Gradient calculations for dynamic recurrent neural networks: a survey," IEEE Transactions on Neural Networks, 6(5), 1212-1228, 1995.

[28] Vedaldi. A, Zisserman. A, "Efficient additive kernels via explicit feature maps," IEEE Transactions on Pattern Analysis & Machine Intelligence, 34(3), 480-492, 2012.

[29] Kitamura, Takuya, and T. Sekine, "A Novel Method of Sparse Least Squares Support Vector Machines in Class Empirical Feature Space," International Conference on Neural Information Processing, pp, 475-482, 2012.

[30] Shi. X, Zhou. F, Liu. L, Zhao. B, "Textural feature extraction based on time–frequency spectrograms of humans and vehicles," Radar Sonar & Navigation Iet, 9(9), 1251-1259, 2015.

[31] Zhang. Y. D, Dong. Z, Chen. X, Jia. W, Du. S, "Muhammad, K. Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation," Multimedia Tools and Applications, 3,1-20, 2017.

[32] Chen, Rung-Ching. "Using Deep Learning to Predict User Rating on Imbalance Classification Data," IAENG International Journal of Computer Science, vol. 46, no. 1, pp. 109-117, 2019.

[33] Wu. Y, Li. J, Kong. Y, Fu. Y, "Deep Convolutional Neural Network with Independent SoftMax for Large Scale Face Recognition," Acm on Multimedia Conference,1063-1067, 2016.

[34] Shore. J, Johnson. R, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," IEEE Transactions on Information Theory, 26(1), 26-37, 1980.

[35] Han. S, Liu. X, Mao. H, Pu. J, Pedram. A, Horowitz. M. A. Eie, "Efficient inference engine on compressed deep neural network," Acm Sigarch Computer Architecture News, 44(3), 243-254,2016.

[36] Fachrurrozi, Muhammad, Arief Wijaya, Muhammad Naufal Rachmatullah. "New Optimization Technique to Extract Facial Features," IAENG International Journal of Computer Science, vol.45, no.4, pp.523-530,2018.

[37] Zeng. J, Yin. W, "On nonconvex decentralized gradient descent," IEEE Transactions on Signal Processing, 66,11, 2018.

[38] Mukkamala. M. C, Hein. M, "Variants of rmsprop and adagrad with logarithmic regret bounds," ICML'17 Proceedings of the 34th International Conference on Machine Learning,70,2545-2553, 2017.

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

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