МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

МАТЕРИАЛЫ

IX-й Международной научной конференции

«МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ИНФОРМАЦИОННЫХ, ТЕХНИЧЕСКИХ И ЭКОНОМИЧЕСКИХ СИСТЕМ»

Томск, 26–28 мая 2022 г.

Под общей редакцией
кандидата технических наук И.С. Шмырина

Томск
Издательство Томского государственного университета
2022

11. *Yang J.* Enhancing Action Recognition of Construction Workers Using Data-Driven Scene Parsing // Journal of Civil Engineering and Management. – 2018. – №7 (24). – P. 568–580.

12. *Luo X. [и др.].* Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks // Automation in Construction. – 2018. – (94). – P. 360–370.

13. *Yin Y. [и др.].* Method for detection of unsafe actions in power field based on edge computing architecture // Journal of Cloud Computing. – 2021. – №1 (10). – P. 17.

14. *Oh S. [и др.].* A large-scale benchmark dataset for event recognition in surveillance video. – 2011. – P. 3153–3160.

15. *Demir U., Rawat Y.S., Shah M.* TinyVIRAT: Low-resolution Video Action Recognition. – 2021. – P. 7387–7394.

16. *Tirupattur P. [и др.].* TinyAction Challenge: Recognizing Real-world Low-resolution Activities in Videos // arXiv:2107.11494 [cs]. – 2021.

17. *Xu Y. [и др.].* ARID: A New Dataset for Recognizing Action in the Dark Communications in Computer and Information Science / под ред. X. Li [и др.]., Singapore: Springer Singapore, 2021. – P. 70–84.

18. *Hara K., Kataoka H., Satoh Y.* Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? – 2018. – P. 6546–6555.

19. *Chen R. [и др.].* DarkLight Networks for Action Recognition in the Dark. – 2021. – P. 846–852.

20. *Monfort M. [и др.].* Moments in Time Dataset: One Million Videos for Event Understanding // Ieee Transactions on Pattern Analysis and Machine Intelligence. – 2020. – №2 (42). – P. 502–508.

21. *Chen C.-F. [и др.].* Deep Analysis of CNN-based Spatio-temporal Representations for Action Recognition // arXiv:2010.11757 [cs]. – 2021.

22. *Kay W. [и др.].* The Kinetics Human Action Video Dataset // arXiv:1705.06950 [cs]. – 2017.

23. *Goyal R. [и др.].* The «something something» video database for learning and evaluating visual common sense. – New York: Ieee, 2017. – P. 5843–5851.

24. IP камеры видеонаблюдения [Электронный ресурс]. URL: https://securityrussia.com/cctv/ip-kamery/ (дата обращения: 24.05.2022).

25. *Kong Y., Fu Y.* Human Action Recognition and Prediction: A Survey // arXiv:1806.11230 [cs]. – 2018.

26. *Bobick A.F., Davis J.W.* The Recognition of Human Movement Using Temporal Templates // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2001. – №3 (23). – P. 257–267.

27. *Dalal N., Triggs B.* Histograms of oriented gradients for human detection. – 2005. – P. 886–893.

28. *Laptev I. [и др.].* Learning realistic human actions from movies Anchorage. – AK, USA: IEEE, 2008. – P. 1–8.

29. *Herath S., Harandi M., Porikli F.* Going deeper into action recognition: A survey // Image and Vision Computing. – 2017. – (60). – P. 4–21.

30. *Tran D. [и др.].* Learning Spatiotemporal Features with 3D Convolutional Networks. – New York: Ieee, 2015. – P. 4489–4497.

31. *Carreira J., Zisserman A.* Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset/ – New York: Ieee, 2017. – P. 4724–4733.

32. *Zhu Y. [и др.].* A Comprehensive Study of Deep Video Action Recognition // arXiv:2012.06567 [cs]. – 2020.

33. *Simonyan K., Zisserman A.* Two-Stream Convolutional Networks for Action Recognition in Videos (под ред. Z. Ghahramani [и др.].). – La Jolla: Neural Information Processing Systems (nips), 2014.

34. *Wang L. [и др.].* Temporal Segment Networks: Towards Good Practices for Deep Action Recognition (под ред. B. Leibe [и др.].). – Cham: Springer International Publishing Ag, 2016. – P. 20–36.

35. *Zhou B. [и др.].* Temporal Relational Reasoning in Videos (под ред. V. Ferrari [и др.].) – Cham: Springer International Publishing Ag, 2018. – C. 831–846.

36. *Feichtenhofer C. [и др.].* SlowFast Networks for Video Recognition. – New York: Ieee, 2019. – P. 6201–6210.

DOI: 10.17223/978-5-907572-27-0-2022-37

SPEECH EMOTION RECOGNITION BASED ON DEEP CONVOLUTIONAL NEURAL NETWORKS

Chen Jin
Tomsk Polytechnic University
czin2@tpu.ru

Introduction

In recent years, information processing and decision making in human-computer interaction have received a great deal of attention. Intelligent information technologies, especially human-computer interaction systems, have developed considerably. The effectiveness of these systems depends to a large extent on the quality of recognition of the information coming from the users of the automated systems and the purposefulness of the human influence on the object of study. Since 1985, Professor Minsky of MIT [1] discussed the "emotional problem of artificial intelligence". The problem of artificial intelligence emotion has become one of the main research areas in artificial intelligence technology. The main problem of research on how to make computers have higher and more comprehensive intelligence is how to make computers have the ability to understand, recognize, and generate various emotional characteristics similar to human beings. With the discovery of research, it is only possible to achieve the purpose of conversational interaction between the computer and the user if most of the features of the speech stream generated during human-computer interaction are taken into account. In speech contains not only phonetic information, but also emotional information. For example, Russian contains about 40% of words with emotional overtones. In addition, emotions are encoded by certain acoustic parameters in the speech signal. Understanding these features of the acoustic coding of emotions makes it possible to understand the underlying mechanisms of emotion perception and expression.

Experts studying the problem of identifying emotions through acoustic features of speech have found that a speaker's emotional state is naturally reflected in the acoustic features of his speech and voice, which in turn are the objective basis for the listener's fully subjective perception of the speaker.

Despite all this work, some issues remain unresolved, especially those related to the level of accuracy of human recognition of basic emotions based on acoustic parameters of speech. It has been demonstrated that the emotional information conveyed is conveyed by intonation features, but the question of how these features enable the recognition of emotional expressions in human speech is still not well understood. Addressing this issue is of great importance because it allows us to separate the semantic and intonational components of speech information. Therefore, this paper does a study on how to recognize emotional information in speech.

1. Feature Extraction

In the research of speech emotion recognition, the selection of speech emotion features is always a crucial part. According to the current research status and the comparison of the advantages and disadvantages of the effect, the MFCC feature is more direct to the representation of speech emotion, and it performs well in the field of speech emotion recognition. However, there are two types of 13-dimensional Mel-frequency cepstral coefficients and 39-dimensional Mel-frequency cepstral coefficients that have good performance in the recognition effect. Therefore, the main research content of this chapter is to build a VGG-like one-dimensional convolutional neural network to conduct speech emotion analysis on the 13-dimensional Mel frequency cepstral coefficients and the 39-dimensional Mel frequency cepstral coefficients extracted from the speech signal respectively. Identify the experiments, compare the performance differences between the two, and analyze the experimental results. Then, the speech features with better performance among the two are selected as the features of subsequent experiments.

1.1. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a feature that accurately describes the variation of vocal tract deformation in the short-term power spectrum of speech. It is a feature that is more suitable for the human hearing mode, because the human ear has a certain difference in the auditory inspiration in sensing sound waves of different frequencies, and the MFCC has a linear change in the logarithmic energy spectrum of the nonlinear Mel scale. The characteristics are closer to the human auditory system than the spectrum, so MFCC can better characterize the sound signal from multiple angles.

1.2. Feature extration

The MFCC feature is the cepstral data obtained by performing a series of calculations under the Mel scale after a series of processing of the speech signal. The extraction process is:

pre-emphasis, framing, windowing, fast Fourier transform, Mel filtering, logarithmic operation, discrete cosine transform, and finally the MFCC feature is obtained.

Among them, pre-emphasis, framing and windowing have been described in detail above, and the subsequent steps are mainly described in this chapter.

After the speech signal is preprocessed, because the characteristics of the time domain are not obvious, the speech signal is converted from the time domain to the frequency domain signal through the fast Fourier transform. Compared with the time domain features, the frequency domain signal is more star-like, and its energy distribution can be observed through the energy spectrum, and the power spectrum after the frequency domain is modulo squared can be effectively analyzed. Therefore, many researchers also use the spectrogram as one of the research characteristics. The function expression for conversion is (1):

$$S_i(k) = \sum\nolimits_{n=1}^{N} s_i(n)\omega(n)e^{\frac{-j2\pi kn}{N}}, \ 0 \le k \le K. \tag{1}$$

Mel filtering and the acoustic frequency (*f*) of the human ear exhibit a nonlinear relationship, which can be approximated by (2):

$$\mathrm{Mel}(f) = 2595\lg\left(1+\frac{f}{700}\right). \tag{2}$$

According to the characteristics of the human ear to the speech signal, when extracting features, the low-frequency part should be denser, and the high-frequency part should be relatively sparse. Mel filtering is implemented by triangular filtering. The energy spectrum after passing the FFT is sent to a Mel-scale triangular filter bank (generally consists of 26 triangular filters), as shown in Fig. 1.

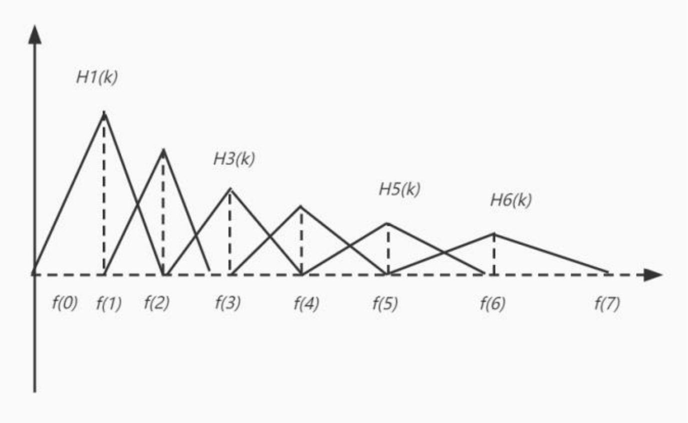

Fig. 1. Schematic of a triangular filter bank

The triangular filter expression is as follows:

$$H_n = \begin{cases} 0, \ k < f(m-1), \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, \ f(m-1) \le k \le f(m), \\ 1, \ k = f(m), \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, \ f(m) \le k \le f(m+1), \\ 0, \ k > f(m+1), \end{cases} \tag{3}$$

$\sum\limits_{m=0}^{M-1} H_m(k) = 1$, *f*(*m*) is the center frequency of the mth triangular filter.

A triangular filter will smooth out the frequencies, making the original formants more pronounced. Therefore, the triangular filter will reduce the interference of the pitch level on the emotional characteristics of the MFCC, and avoid the objective influence of the pitch on the emotion.

After the smoothing of the triangular filter, the logarithmic calculation of the result of each filter is required, and the logarithmic energy can be obtained from the formula (4):

$$s(m) = \ln\left(\sum_{k=0}^{N-1} | X_a(k) |^2 H_m(k)\right),\ 0 \le m \le M\ . \tag{4}$$

After the logarithmic energy is obtained, a discrete cosine transform can be performed to obtain the MFCC, the formula is as (5):

$$C(n) = \sum_{m=0}^{N-1} s(m) \cos\left(\frac{\pi n(m-0.5)}{M}\right),\ n = \overline{1, L}\ . \tag{5}$$

M is the number of triangular filters, and *L* is the order of Mel cepstral coefficients.

In general, *L* is usually 12, but in order to ensure data integrity, a frame of logarithmic energy spectrum and a 12-dimensional MFCC will be spliced to obtain a 13-dimensional MFCC.

The 12-dimensional and 13-dimensional MFCC features obtained above are only the static features of the original voice signal, and the differential spectrum obtained by the low-dimensional MFCC through differential can describe the corresponding dynamic features of the voice signal. Therefore, high-dimensional MFCC features (including first-order and second-order differences) are generated, and the difference parameter extraction formula is shown in (6):

$$d(t) = \begin{cases} C_{t+1} - C_t,\ t < K, \\ \dfrac{\sum_{k=1}^{K} k\left(C_{t+k} - C_{t-k}\right)}{\sqrt{2\sum_{k=1}^{K} k^2}},\ \text{others}, \\ C_t - C_{t-1},\ t \ge Q - K. \end{cases} \tag{6}$$

1.3. Pre-experimental model

Convolutional Neural Network (CNN) is a kind of feedforward neural network with deep structure that includes convolutional computation and is one of the representative algorithms of deep learning [2]. Convolutional neural network can perform shift-invariant classification of input information according to its hierarchical structure [3], so it is also called "Shift-Invariant Artificial Neural Networks". According to the characteristics of CNN, it is often used for further feature extraction and representation, in order to better complete the training of deep learning.

MFCC is most commonly used as a feature for speech emotion recognition is 13 and 39 dimensional two, in order to choose the most suitable feature in speech emotion recognition. In this study, I used a more basic VGG-like CNN network model for the experiment. And the experimental results are compared.

The schematic diagram of the CNN network structure with VGG-like structure used in this experiment is shown in Fig. 2.

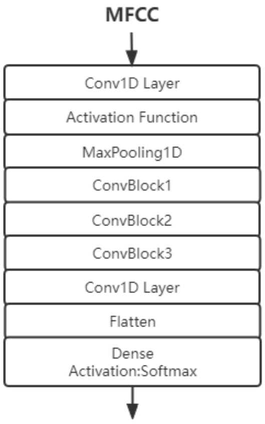

Fig. 2. CNN model

The ConvBlock in Fig. 2 is a convolution block, and its composition is shown in Fig. 3.

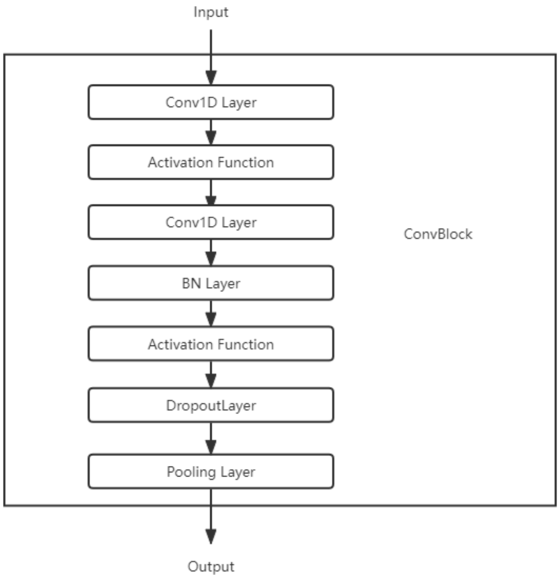

Fig. 3. ConvBlock

Since the processing of speech is carried out according to the frame, the corresponding feature vector can be obtained after each frame is processed. Assuming that a piece of speech has *N* frames, after processing this piece of speech, an MFCC feature matrix with *M* rows and *N* columns can be obtained (*M* represents the feature dimension, and *N* represents the number of frames). Therefore, in this CNN model, a 1-dimensional convolution Conv1D layer is used, the activation function is ReLu, and the pooling layer is MaxPooling1D. The construction of a simple speech emotion recognition network is realized through the CNN network similar to VGG structure. A comparison experiment of speech emotion recognition is carried out on the 13-dimensional MFCC and the 39-dimensional MFCC.

In this experiment, the dataset is selected as the SAVEE sentiment dataset. The selected data set is randomly divided into training set and test set according to the ratio of 8:2, and cross-validation is performed.

The data for the network model are shown in Tab. 1.

Table 1

Network parameters

Item	Parameters
Conv1D Layer	Filters =256, kernel_size=5
Activation Function	ReLu
Maxpooling1D	4
Dropout	0.25
ConvBlock1	Filters=256,kernel_size=5
ConvBlock2	Filters=128,kernel_size=5
ConvBlock3	Filters =64,kernel_size=5

The training optimizer uses Adam, and its parameters are: lr = 0.001, beta_1 = 0.9, beta_2 = 0.99. The loss function used the categorical cross-entropy function.

1.4. Results and Analysis

For the results of this experiment, multiple evaluation criteria were selected. In addition to the accuracy, precision, recall and F1-score mentioned above, there is also a confusion matrix.

Confusion matrix is an evaluation benchmark that directly expresses experimental data results in matrix form. Since the two dimensions of the confusion matrix are represented as the true label category and the predicted result category, the confusion matrix is usually represented as a two-dimensional *N*-order matrix with *N* rows and *N* columns. The advantage of the confusion matrix is that the recognition situation of each classification can be observed intuitively, and it is easy to draw more accurate judgments and make corresponding adjustments quickly.

The average accuracy, precision, recall and F1-score of its 13-dimensional MFCC and 39-dimensional MFCC are shown in Tab. 2.

Table 2

Comparison of experimental standards

Category Standard	13-dimensional MFCC	39-dimensional MFCC
Accuracy	85.94%	84.90%
Precision	87.63%	84.81%
Recall	84.90%	84.38%
F1-score	86.23%	84.59%

Among them, the mixture matrix of the verification set of speech emotion recognition performed by 13-dimensional MFCC features is shown in Tab. 3.

Table 3

13-dimensional MFCC mixing matrix

Predicted True	neutral	sad	angry	surprise	disgust	fear	happy	Precision
neutral	22	0	0	0	0	0	2	91.67%
sad	0	22	0	0	0	2	0	91.67%
angry	0	0	22	0	0	0	2	91.67%
surprise	2	0	0	22	0	0		91.67%
disgust	2	3	2	0	39	2	0	81.25%
fear	3	0	1	0	2	18	0	75.00%
happy	0	0	2	2	0	0	20	83.33%
Recall	75.86%	88.00%	81.48%	91.67%	95.12%	81.81%	83.33%	84.90%/87.63%

The mixture matrix for speech emotion recognition on 39-dimensional MFCC features is shown in Tab. 4.

Table 4

39-dimensional MFCC mixing matrix

Predicted True	neutral	sad	angry	surprise	disgust	fear	happy	Precision
neutral	21	0	0	0	0	1	2	87.50%
sad	0	20	0	0	3	1	0	83.33%
angry	0	2	20	0	0	2	0	83.33%
surprise	0	0	0	22	0	2	0	91.67%
disgust	4	2	0	2	40	0	0	83.33%
fear	0	0	0	0	4	20	0	83.33%
happy	0	0	2	0	2	0	20	83.33%
Recall	84.00%	83.33%	90.91%	91.67%	81.63%	76.92%	90.91	84.38%/84.81%

First, for the results shown in Tab. 2. From the four evaluation benchmarks of accuracy, precision, recall, and F1-score, it is obvious that the 13-dimensional MFCC may be more suitable for emotion recognition than the 39-dimensional MFCC. Except for the recall rate of these four indicators, the other three evaluation benchmarks, the 13-dimensional MFCC is 1–3% higher than the 39-dimensional MFCC.

It can be seen from Tab. 3 that the 13-dimensional MFCC has higher recognition accuracy in the four emotions of neutral, sad, angry, and surprise, but has lower recognition accuracy for disgust, fear, and happy. At the same time, the recall rate of sad, surprise and disgust is higher, while the other emotions are lower.

From the experimental data of the 39-dimensional MFCC in Tab. 4, it can be clearly seen that the accuracy of the seven emotion recognition is not very different. Compared with the 13-dimensional MFCC, the accuracy of various emotions is more average.

In terms of the recognition accuracy of various emotions, the 39-dimensional MFCC performs better on the SAVEE dataset than the 13-dimensional MFCC, and can identify various emotions more accurately. The reason should be the advantages brought by the dynamic characteristics represented by the first-order difference spectrum and the second-order difference spectrum in the 39-dimensional MFCC. However, from the overall accuracy, precision, recall and F1-score of the four evaluation benchmarks, the 13-dimensional MFCC is better. It may be due to the better representation of the static features characterized by the 13-dimensional MFCC. Therefore, the next experiments will select 13-dimensional MFCC as the feature of speech emotion recognition.

2. Speech Emotion Recognition Based on MS-ResNet

For the model improvement and innovation of neural networks, there are two main ideas for building network models at this stage, one is to deepen the number of network layers, and the other is to widen the network structure. The representatives of these two methods are ResNet [4] and GoogLeNet [5]. These two models have outstanding performances in image processing, text recognition and speech recognition, so many researchers use these two models The model serves as the model basis for related research.

2.1. ResNet

Although ResNet and traditional neural network retain the main structure of traditional convolutional neural network in terms of structure, there is a big structural difference between ResNet and traditional neural network architecture due to the introduction of its residual block. ResNet network introduces residual network structure into neural network. Through this structure, the problem of gradient disappearance or gradient explosion can be dealt with to a certain extent. The core idea is to directly skip one or more layers by introducing a residual structure. This residual structure of the current layer and the previous layer is generally described as "shortcuts". Among them, the residual module is shown in Fig. 4.

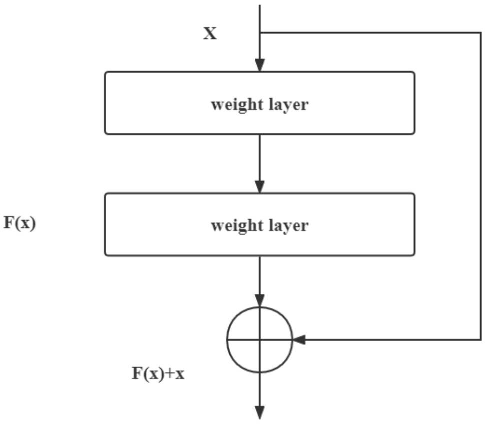

Fig. 4 Residual block diagram

In the figure, the input signal *X*is first sent to the first weight layer and passed through the activation function to obtain the intermediate original equation of the neural network, which is $F(x) = H(x) - x$, and the value of *F*(*x*) shown in (7):

$$F(x) = W_2 \cdot \mathrm{Re}lu\left(W_1 \cdot X\right). \tag{7}$$

Through the residual block, the output function is obtained as:

$$H(X) = F(X) + X\ , \tag{8}$$

which is:

$$H(X) = W_2 \cdot \mathrm{Re}lu\left(W_1 \cdot X\right) + X\ . \tag{9}$$

In the residual block, it should be noted that after the input *x* passes through the first weight and activation function, after passing through the second weight, it is first added to *x* and then input to the second activation function. Through the residual block structure *H*(*x*) in Fig. 4, the adaptive force of the network can be improved, thereby solving the problem of network degradation.

The following table shows the configuration of the ResNet network structure proposed in [4]. The residual structure in the table gives the size of the convolution kernel and the number of convolution kernels on the main branch. **N* in the table indicates that the residual network structure is repeated *N* times.

Table

Residual Structure Parameters

Layer name	Output size	18-layer	34-layer	50-layer	101-layer	152-layer
Conv1	112*112	7*7,64,stride 2				
Conv2_x	56*56	3*3 max pool, stride 2				
		$\begin{bmatrix}3*3 & 64\\3*3 & 64\end{bmatrix}*2$	$\begin{bmatrix}3*3 & 64\\3*3 & 64\end{bmatrix}*3$	$\begin{bmatrix}1*1 & 64\\3*3 & 64\\1*1 & 64\end{bmatrix}*3$	$\begin{bmatrix}1*1 & 64\\3*3 & 64\\1*1 & 256\end{bmatrix}*3$	$\begin{bmatrix}1*1 & 64\\3*3 & 64\\1*1 & 256\end{bmatrix}*3$
Conv3_x	28*28	$\begin{bmatrix}3*3 & 128\\3*3 & 128\end{bmatrix}*2$	$\begin{bmatrix}3*3 & 128\\3*3 & 128\end{bmatrix}*4$	$\begin{bmatrix}1*1 & 128\\3*3 & 128\\1*1 & 128\end{bmatrix}*4$	$\begin{bmatrix}1*1 & 128\\3*3 & 128\\1*1 & 512\end{bmatrix}*4$	$\begin{bmatrix}1*1 & 128\\3*3 & 128\\1*1 & 256\end{bmatrix}*8$
Conv4_x	14*14	$\begin{bmatrix}3*3 & 256\\3*3 & 256\end{bmatrix}*2$	$\begin{bmatrix}3*3 & 256\\3*3 & 256\end{bmatrix}*6$	$\begin{bmatrix}1*1 & 256\\3*3 & 256\\1*1 & 256\end{bmatrix}*6$	$\begin{bmatrix}1*1 & 256\\3*3 & 256\\1*1 & 1024\end{bmatrix}*23$	$\begin{bmatrix}1*1 & 256\\3*3 & 256\\1*1 & 1024\end{bmatrix}*36$
Conv5_x	7*7	$\begin{bmatrix}3*3 & 512\\3*3 & 512\end{bmatrix}*2$	$\begin{bmatrix}3*3 & 512\\3*3 & 512\end{bmatrix}*3$	$\begin{bmatrix}1*1 & 512\\3*3 & 512\\1*1 & 512\end{bmatrix}*3$	$\begin{bmatrix}1*1 & 512\\3*3 & 512\\1*1 & 2048\end{bmatrix}*3$	$\begin{bmatrix}1*1 & 512\\3*3 & 512\\1*1 & 2048\end{bmatrix}*3$
	1*1	Average pool, 1000-d fc, sofrmax				

It can be seen from the above table that ResNet has various network structures ranging from 18 layers to 152 layers. From the theoretical analysis, with the increase of the number of layers, the corresponding network structure is gradually deepened, and the more thorough the feature extraction is, the better the effect is obtained. However, as the number of layers increases and the number of training parameters increases, the resources and time required for training will gradually increase. Therefore, when selecting the required number of network layers, it is necessary to select the best depth within a reasonable range of resources. In this paper, ResNet-18 is selected as one of the comparison models for speech emotion recognition comparison experiments.

2.2. GoogLeNet

GoogLeNet is a high-performance neural network model proposed by Google in 2014. GoogLeNet is the same as ResNet, and also introduces the idea of basic block, that is, the Inception module. The core structure of the Inception module GoogLeNet can not only improve the performance of the neural network while increasing the depth and width of the network, but also ensure the efficiency of computing resources. Since then, the Inception structure has been continuously updated and optimized for the main problems that limit the performance of deep neural networks, and different versions have been iterated, which are introduced below.

Google first proposed the original version of Inception, as shown in Fig. 5. Its main idea is to design a parallel network structure, using multiple convolution kernels and pooling kernels of different sizes in each layer to process the input. This structure can increase the adaptability of the network to a certain extent. However, the parameter quantity of each layer of Inception module is the sum of all branch parameters. If the multi-layer Inception structure is superimposed, the final parameter quantity of the model will be too large, which will result in greater dependence on computing resources.

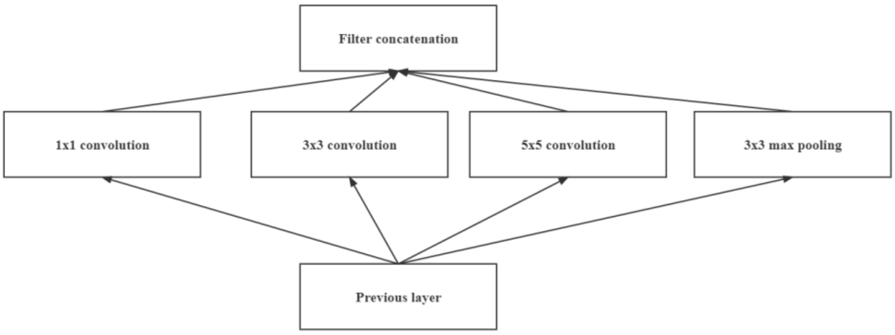

Fig. 5. Original Inception version

In order to reduce the number of parameters, a reduced-dimensional version of the Inception structure was proposed later. The main difference between the two is that the latter performs this 1x1 convolution operation before the 3x3, 5x5 convolution kernel and after max pooling, respectively. The operation can ensure that the parameter quantity and complexity of the model are reduced without losing the representation ability of the model. The specific structure is shown in Fig. 6.

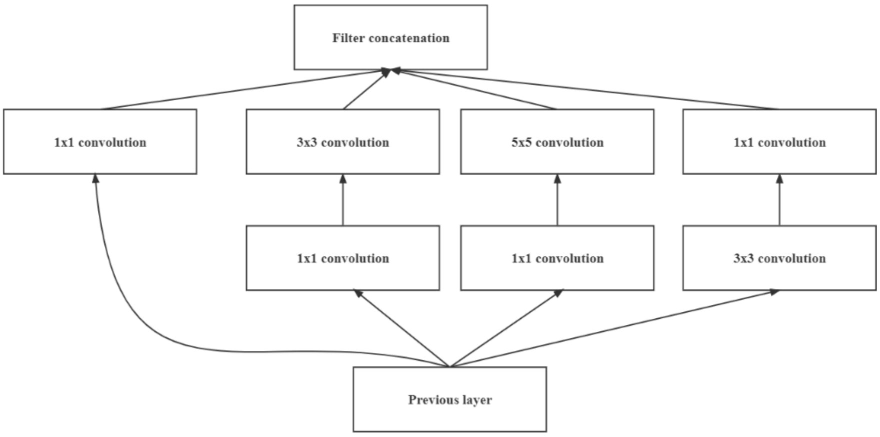

Fig. 6. Inception V1

A 22-layer GoogLeNet is constructed based on this Inception structure. The network adopts the Inception modular structure to facilitate changes to the model.

Once the Inception structure was proposed, it has attracted widespread attention due to its excellent performance. Google has further improved it and proposed Inception V2, V3 [6], [7]. Its core ideas have two main points. First, for The Batch Normalization (BN) method is proposed to solve the Internal Covariate Shift problem in the process of neural network training. During neural network training, the input distribution of each layer is always changing, making it difficult to train the model. BN is an effective regularization method. It performs a normalization operation on the data in a mini-batch to ensure the output. It is N(0,1), which can increase the robustness of the model. Second, the convolution kernel is decomposed, and a large convolution kernel is decomposed into multiple small convolution kernels. For example, two 3x3 convolution kernels are used to replace the 5x5 convolution kernels in the Inception module. A large number of experiments have proved that this approach will not lead to a decline in the model's expressive ability. In addition, the convolution kernel is asymmetrically decomposed, and a larger 2D convolution is split into two smaller 1D convolutions, as shown in Fig. 7. Asymmetric decomposition of convolution enables the network to handle richer spatial features.

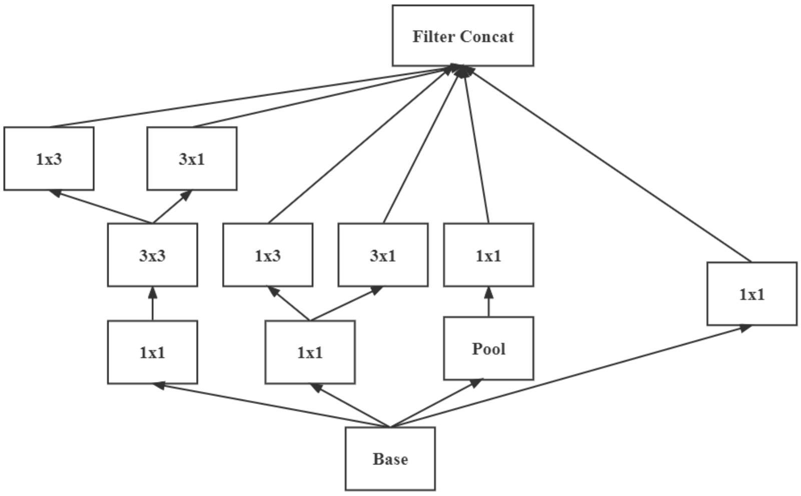

Fig. 7. Inception V2

2.3. MS-ResNet

MFCC is a frame-level feature based on speech signal, so the features contained in each frame may be different. To address this problem, we propose MS-ResNet, a multiscale recognition model based on residual networks, which fuses features from multiple scales to obtain the potential features of MFCC.

In this paper, the scaling of neural network depth and width is considered. Parallel multi-branch network mechanism based on GoogLeNet. Based on the combined ResNet-18 network model, a parallel multi-branch network is constructed by introducing a multi-scale mechanism. This network model is shown in Fig. 8.

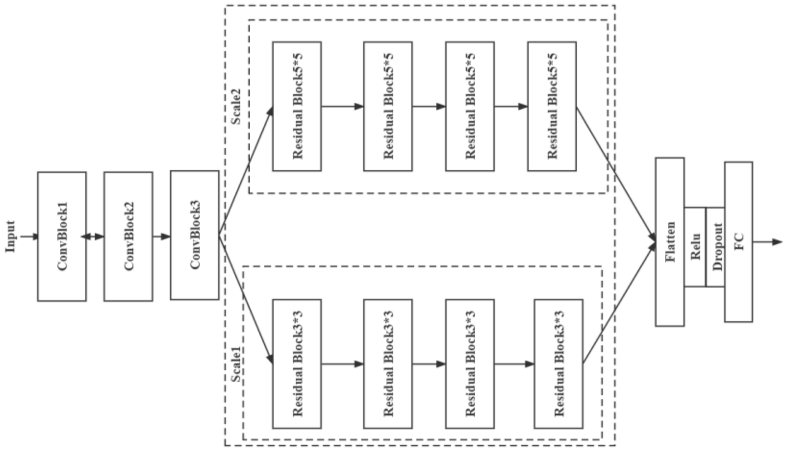

Fig. 8 MS-ResNet model

In Fig. 6, ConvBlock is a 1-dimensional convolution block, and its model is shown in Fig. 8, where ConvBlock1, ConvBlock2, and ConvBlock3 are convolution blocks with filters of 256, 128, and 64, respectively, and the convolution kernel of 7*7. Scale1 and Scale2 are based on ResNet-18, respectively, and the changed convolutional network, the convolution kernel is 3*3 and 5*5 respectively.

Among them, the residual network adopts the method of "shortcut connection", and the same feature map will have different expressions in two different scale spaces to achieve the purpose of information complementation. Therefore, the two scale spaces are fused in this way to obtain feature parameters with better emotional information, and global information can be obtained. The correlation between adjacent frames can also be obtained. The correlation between non-adjacent frames can be obtained. If the output of the scale1 network is $f^{s_1}(x)$ and the output of the scale2 network is $f^{s_2}(x)$, the output of the network after fusion is $f^{s_1}(x)+f^{s_2}(x)$.

2.4. Result

The experimental environment used in this chapter is exactly the same as the experimental environment in the experimental part of the third chapter of this paper. The datasets are compared with SAVEE and EMO-DB [8]. In addition, the selected dataset is divided into training set and test set in a ratio of 8:2. At the same time, use the StratifiedShuffleSplit function in the sklearn library for cross-validation. The object is the merger of StratifiedKFold and ShuffleSplit, and returns a hierarchical random stack. where stacking is done by the percentage of samples in each class, the percentage is 20%.

This chapter mainly builds MS-ResNet network for speech emotion recognition. In addition, the ResNet structure and the GoogLeNet structure are introduced in this chapter, so this chapter will use the 13-dimensional MFCC described above to conduct comparative experiments around the above models.

Combined with hardware factors such as the experimental environment, ResNet-18 has relatively few parameters and can perform better experiments. Therefore, this paper selects ResNet-18 as the benchmark model for experiments.

Table

Experimental results on the SAVEE dataset

	Accuracy	Precision	Recall	F1-score
GoogLeNet	81.25%	81.12%	80.73%	80.92%
ResNet-18	84.90%	85.38%	84.90%	85.14%
MS-ResNet	86.46%	87.40%	86.46%	86.91%

The line chart of the training accuracy of MS-ResNet on the two datasets is as follows.

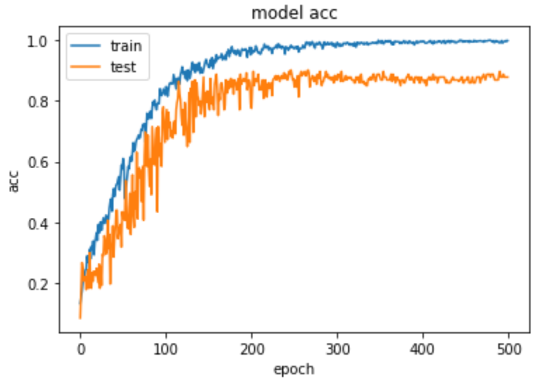

Fig. On SAVEE acc curve graph

Table

Experimental results on Emo-DB dataset

	Accuracy	Precision	Recall	F1-score
GoogLeNet	87.85%	87.85%	87.85%	87.85%
ResNet-18	85.98%	86.20%	86.20%	86.20%
MS-ResNet	87.85%	88.57%	88.19%	88.97%

The line chart of the training accuracy of MS-ResNet on the two datasets is as follows.

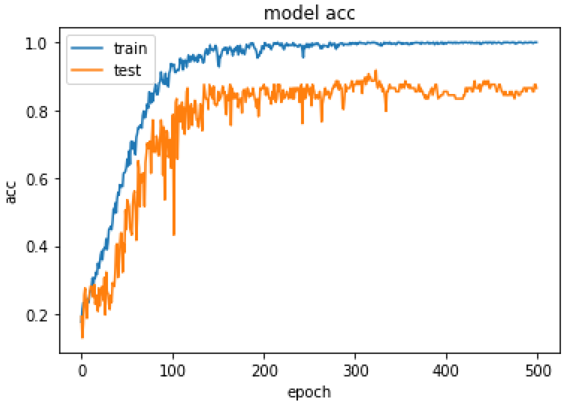

Fig. On EMO-DB acc curve graph

From the comparison of experiments on the two datasets, it can be clearly seen that MS-ResNet has improved compared with GoogLeNet and ResNet-18 in the four evaluation indicators of accuracy, precision, recall and F1-score.

Combined with the experiments in this paper, it can be seen that for MS-ResNet, compared with the GoogLeNet and ResNet benchmark models, the performance on the MFCC feature set has a certain improvement. From the results, on the SAVEE dataset, the four evaluation indicators, MS-ResNet has an improvement of 1%–2% compared with the ResNet network model, and an improvement of about 5%–6% compared with GoogLeNet. Compared with the benchmark ResNet-18, the improvement of MS-ResNet on Emo-DB is more obvious.

It can be seen from this that the multi-scale learning mechanism can more effectively extract the hidden information in the speech information, and fuse the speech features between adjacent frames and non-adjacent frames, compared with the single-scale deep convolutional neural network. The network model has improved. Therefore, the multi-scale residual network provides a research idea for the extension of the deep learning network model, depth and width.

Conclusion

In the research of this paper, a speech recognition model based on multi-scale residual convolutional neural network is mainly developed. The algorithm mainly adopts a deep learning approach to improve the accuracy of speech emotion recognition.

In this study, a VGG-like neural network is first constructed with the aim of targeting the structural features of the Mel Spectral Coefficients (MFCC) speech emotion features that have performed well in current speech emotion recognition research. It was used to test MFCC with two different dimensions of 13 and 39 dimensions in the SAVEE emotion database.By visualizing and comparing the mixture matrix, it was found that the 39-dimensional MFCC had stronger performance in speech emotion recognition. Therefore, the 39-dimensional MFCC was selected as the follow-up feature.

Then, a multiscale residual neural network model based on ResNet network was constructed based on two different research ideas in the field of neural network research. Experiments were conducted on EMO-DB and SAVEE datasets using this model and compared with ResNet and GoogleNet networks. From the experimental results, it can be concluded that the MS-ResNet model constructed in this paper has high accuracy in speech emotion recognition. Finally, it can be concluded that multi-scale can obtain the hidden emotion information in speech signals from MFCC.

REFERENCE

1. *Minsky M.* Society of mind. – Simon and Schuster, 1988.

2. *Gu J, Wang Z, Kuen J, et al.* Recent advances in convolutional neural networks // Pattern Recognition. – 2018. – 77. – P. 354–377.

3. *Zhang W., Tanida J., Itoh K., et al.* Shift-invariant pattern recognition neural network and its optical architecture // Proceedings of annual conference of the Japan Society of Applied Physics. – 1988: – P. 2147–2151.

4. *He K., Zhang X., Ren S., et al.* Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – P. 770–778.

5. *Szegedy C., Liu W., Jia Y., et al.* Going deeper with convolutions // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2015. P. 1–9.

6. *Ioffe S., Szegedy C.* Batch normalization: Accelerating deep network training by reducing internal covariate shift // International conference on machine learning. – PMLR. – 2015. – P. 448–456.

7. *Szegedy C., Vanhoucke V., Ioffe S., et al.* Rethinking the inception architecture for computer vision // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – P. 2818–2826.

8. *Burkhardt F., Paeschke A., Rolfes M., et al.* A database of German emotional speech // Interspeech. – 2005. – 5. – P. 1517–1520.