A Novel Convolutional Neural Network Voiceprint Recognition Method Based on Improved Pooling Method and Dropout Idea

Wei-Zhong Sun, Jie-Sheng Wang*, Bo-Wen Zheng, and Zhong-Feng Li

Abstract—**A voiceprint recognition method based on improved pooling strategy for convolutional neural networks (CNN) is proposed. For the voiceprint recognition, the voice signal is quantized and pre-emphasized, and then the processed voice information is framed and windowed, and the voiceprint information of the self-built digital voiceprint library is converted into a score map to construct a digital voiceprint database. Then an CNN based voiceprint recognition method by introducing an improved pooling method is proposed. The new pooling method is to square the activation value after the activation function, and assign the square number probability to realize the random pooling. This method can not only preserve the feature extraction of the maximum pooling method, but also absorb the advantage of random pooling. It retains the possibility of extracting hidden features, and effectively enhancing the generalization ability of the network. The traditional convolutional neural network LeNet-5 is modified by adopting the improved pooling method and introducing the Dropout idea so as to an improved LeNet-5 network. The voiceprint recognition experiments were performed by using the gray-scale digital spectrogramon in the self-built voiceprint database through the improved LeNet-5 network. The recognition experiments were performed on the VGG-16 network using the three-channel color digital spectrum map. An automatic voiceprint recognition system was built to detect voiceprints in real-time voice information. The simulation results verify the effectiveness of the proposed method.**

Index Terms—**convolutional neural network, voiceprint recognition, pooling method**

I. INTRODUCTION

VOICEPRINT recognition is a kind of biometric recognition technology, which is applied to many programs of protected proprietary data, and can verify the identification through voiceprint information [1]. Voiceprint recognition requires only one recording device to complete the information collection process. So this simple identification method is the safest and most effective way to verify personal identity. It is different from passwords or tokens [2]. Voiceprint recognition is a branch of speech recognition [3]. As early as the 1930s, speech recognition technology has been proposed, this is the embryonic stage of speech recognition. During this period, scientists mainly focused on the research of sound recognition. Scientists observed the law of the sound, summed up the rules and proposed the concept of voiceprint [4]. Since then, the study of voiceprint has become a focus of speech recognition. From the 1940s to the early 1970s, various techniques of voiceprint recognition were gradually introduced during this period. It is also the emergence of voiceprint recognition based on pattern matching [5] and the statistical analysis of variance, so the speech recognition changes from the initial human ear to the stage of automatic recognition. Scientists and researchers all over the world have invested in voiceprint research and proposed techniques for selecting, enhancing and extracting many speech feature parameters. Linear prediction analysis and cepstrum techniques are proposed for voiceprint recognition [6]. Researchers have also gradually discovered features such as formants and fundamental frequency profiles. In the late 1970s and 1980s, linear and nonlinear factors existed in the sound personality parameters of this stage are found, and more effective model matching methods, such as the Mel cepstrum parameter [7], were also proposed in this stage. After entering the 1990s, the researchers proposed various pattern matching models, but also found corresponding shortcomings. Nowadays, people have proposed the concept of deep learning, and deep learning for speech recognition has become a research hotspot [8].

When people talk, it is a very complicated physiological process. Everyone's voice through the vocal organs is very different, which also leads to the uniqueness of people's voiceprints. Since each person's voice characteristics are relatively stable, the identification of voiceprints can effectively identify an individual's identity [9]. In this paper, a voiceprint recognition method based on convolutional neural network by adopting an improved pooling method is proposed. The voice preprocessing of voiceprint recognition is studied, which mainly includes quantization, pre-emphasis, frame-window to produce the spectrum. A digital voiceprint database is constructed. Then based on the classic convolutional neural network LeNet-5, the improved pooling method replaces the original pooling layer method and the Dropout idea is introduced to improve the network and obtain the improved LeNet-5 network. The voiceprint recognition

Manuscript received October 26, 2020; revised January 20, 2021. This work was supported by the Basic Scientific Research Project of Institution of Higher Learning of Liaoning Province (Grant No. 2017FWDF10), and the Project by Liaoning Provincial Natural Science Foundation of China (Grant No. 20180550700).

Wei-Zhong Sun is a lecturer of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. (e-mail: weizhongsun@126.com).

Jie-Sheng Wang is a professor of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. (Corresponding author, phone: 86-0412-2538246; fax: 86-0412-2538244; e-mail: wang_jiesheng@126.com).

Bo-Wen Zheng is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. (e-mail:18842038281@163.com).

Zhong-Feng Li is a senior engineer of Yingkou Institute of Technology, Yingkou, 115000, P. R. China. (e-mail: afeng0601@163.com).

experiments were performed on the self-built voiceprint database through the improved LeNet-5 network, and then the recognition experiment was performed on the VGG-16 network using the three-channel color digital spectrum map.

II. Voiceprint Recognition System

A. Feature Extraction Methods

Voiceprint recognition is a branch of speech recognition, which mainly has two key steps, one is feature extraction and the other is pattern matching. The task of feature extraction is mainly to extract the voiceprint information of the speaker. The extracted voiceprint is highly separable and stable, and the extracted features also need to have personalized features [9]. There are several methods for feature extraction.

(1) Linear prediction method

This method uses the existing samples to predict new samples, subtracts the sample values from the actual values to obtain an error, quantizes the code by error signals, and performs linear prediction to obtain linear prediction coefficients. It describes the model constructed by all poles, but there is a certain error in the description of the voiced and unvoiced sound at the zero point [10].

(2) Linear prediction cepstrum coefficient

The method eliminates the excitation in the speech signal by the cepstrum feature, and the resonance characteristics of the sound can be effectively described by using only a dozen cepstral coefficients. Therefore, it has the advantage of small computational complexity, but has poor ability in feature extraction of consonants and low noise immunity [11].

(3) Mel cepstrum coefficient

The Mel cepstrum coefficient is based on the auditory principle of the human ear and refers to the nonlinear proportional relationship between the height and frequency of speech. The speech signal is weighted by a filter to simulate a human auditory perception system. Mel cepstral coefficient feature is currently the most widely used speech recognition feature parameter method [12-13].

(4) Pitch period coefficient

The pitch period coefficient refers to the reciprocal of the vibration frequency period generated by the vibration of the voiced sound in the voice information. The feature extraction is performed by the time domain, the frequency domain and the integrated information of the pitch, mainly including the peak, valley and zero-crossing rate of the signal. It can effectively describe the sound characteristics [14].

B. Voiceprints Pattern Matching Methods

(1) Template matching method

This method uses dynamic time warping (DTW) to apply the fixed features to the training features and test features. The training process is to describe the features in the speaker's training statements. The test process is to use the same method to extract the template of the test sample on the the statements need to be tested. It's adopted methods are vector quantization and dynamic event regularization [15-16].

(2) Nearest neighbor sound method

This method is to represent all samples with k neighbors, whose main core idea is that when the k neighbors of the sample are mostly considered to belong to a category, the sample is also considered to belong to this class. It is to determine the category according to the number of samples that are closest to the sample to be classified, that is, to classify by the category around the sample to be tested. Since each vector finds the nearest k samples of the training vector at the time of recognition, it will result in a large amount of calculation and a large model storage [17].

(3) Neural network method

There are many methods to identify voiceprints through neural networks. Among them, the radial basis function (RBF) neural network and muti-layer perceptron (MLP) neural network are the best methods. This type of method can effectively distinguish between the recognizer's voice and the background sound, but the amount of training is large and the model is not perfect [18].

(4) Vector quantization method

This is a method of recognizing a speech signal after it is compressed. It is to construct a model Codebook for the speech information, and then match each frame of the speech information with the Codebook. The matching model does not depend on the chronological order of the parameters, and does not require DTW technology, so the parameter storage of this method is smaller. On the other hand, the segmentation and time problems of speech are avoided. However, this type of method only has a good recognition of long speech information, and the recognition rate will drop sharply as the speech becomes shorter [19].

(5) Hidden Markov Model Method

The hidden markov model (HMM) is a stochastic model constructed by transmission probability and transition probability. This method mainly regards speech information as a sequence of random symbols. The sound signal is identified by observing the relationship between the event and the probability distribution of the state, and the statistical correspondence between the state and the observed value. The HMM model does not require time warping for voiceprint recognition, so it can effectively reduce the amount of storage. The disadvantage is that the amount of calculation is large [20-21].

(6) Gaussian mixture model method

The Gaussian mixture model (GMM) is a process of voiceprint recognition by describing the probability density of short-term spectral feature vectors. Aiming at the different probability densities for short-term speech differences when people speak, GMM adopts Gaussian functions to describe the features by using mean and covariance matrices. GMM can effectively describe the feature distribution of speech, and the feature information does not change the speech state with time [22-23].

The voiceprint recognition system has four main steps. Firstly, information acquisition of voice signals. Secondly, pre-processing the collected voice information. Thirdly, the

voiceprint features are extracted from the voice information. Finally match the speech features extracted by the speech information with the original speech library and output the result [24]. The flowchart of the voiceprint recognition system is shown in Fig. 1.

III. Convolutional Neural Networks and Improved Pooling Method

A. Structure of Convolutional Neural Network

Convolutional neural networks are a new neural network structure evolved from Multilayer Perceptron (MLP) [25]. It is mainly composed of a convolutional layer, a pooled layer and a fully connected layer. In this paper, *C* represents the convolutional layer and *S* represents the pooled layer. The convolutional layer is to realize the feature extraction, the pooled layer is for dimensionality sampling of the extracted features, and the fully connected layer is mainly for classification. The structure of the convolutional neural network is shown in Fig. 2.

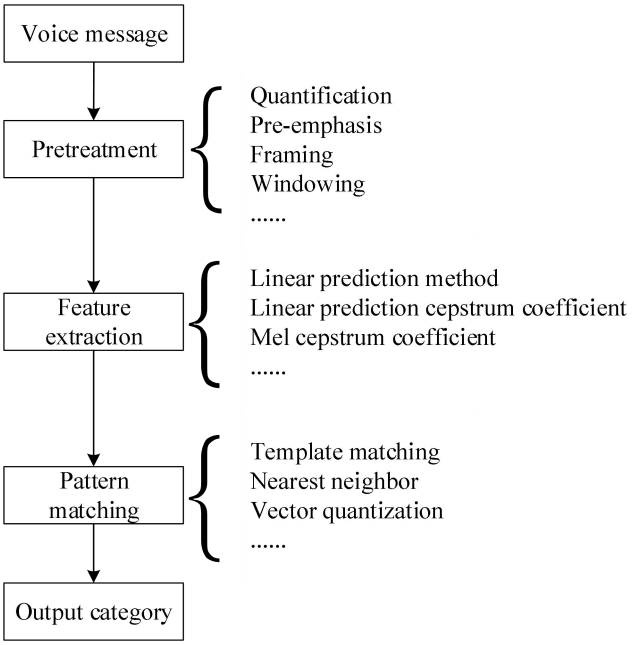


Fig. 1 The structure diagram of the voiceprint recognition system.

B. Convolutional Layer

In the convolutional neural network, a more complex features are to combine the low-level rich edge features by the high-level convolutional layers [26]. The convolutional layer operation is actually equivalent to a linear correlation operation, which can be described as follows:

$$x_j^l = \sigma\left(\sum_{i\epsilon M_j} x_i^{l-1} * w_{ij}^l + b_j^l\right) \quad (1)$$

where, x is a two-dimensional vector with a size of $i \times j$, w is the convolution kernel, b is the offset, l is the number of layers, and M_j is the characteristic map of the input.

C. Pooling Layer

The pooling layer is also called the downsampling layer, which is usually connected behind the convolutional layer and whose main purpose is to reduce the feature map. The pooling methods in convolutional neural networks mainly include Max-Pooling, Mean-Pooling, and Stochastic Pooling [27]. The mathematical expression for the pooling layer is described as:

$$x_j^l = f(\beta_j^l down\left(x_j^{l-1}\right) + b_j^l) \quad (2)$$

where, *down()* represents the function of downsampling.

D. Improved Pooling Method

By combining the max-pooling method and the stochastic pooling method, an improved pooling method shown in Fig. 3 is proposed to make each pooled sampling area not only have better ability to store information, but also better improve the generalization of neural network. Firstly, the value of each sampling window x_i is squared and then it is normalized so as to get the corresponding sampling probability P_i:

$$P_i = \frac{x_i^2}{\sum_{k\epsilon R} x_k^2} \quad (3)$$

$$\sum_{k\epsilon R} P_i = 1 \quad (4)$$

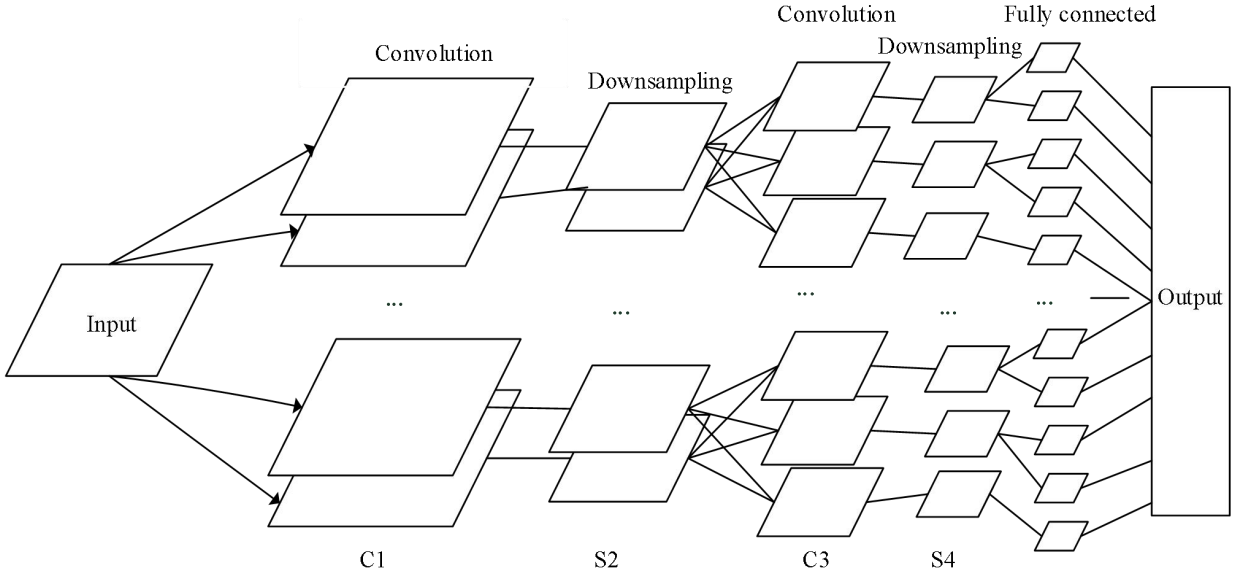


Fig. 2 The structure diagram of the convolutional neural network.

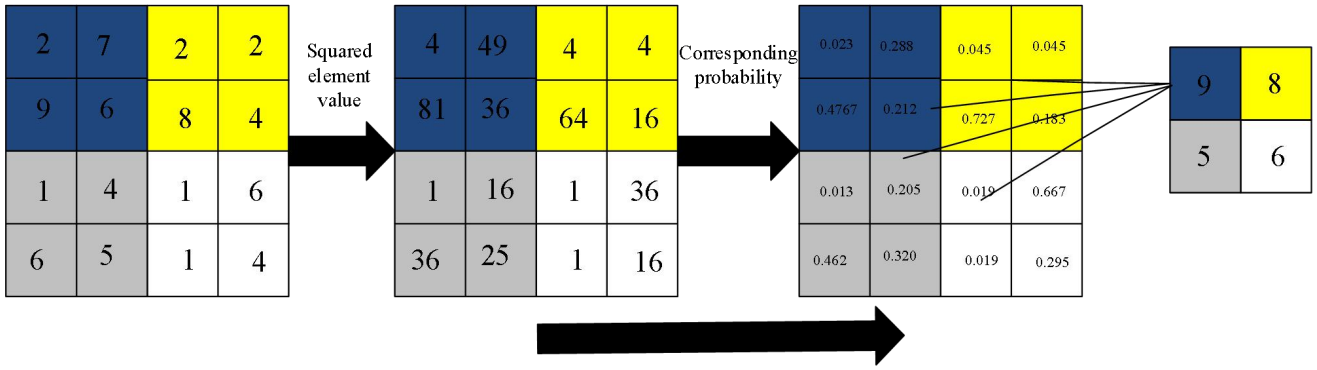


Fig. 3 2*2 Improved stochastic pooling.

After the discrete distribution probability set $P_i(i \in R)$ is obtained by calculation, γ is obtained by carrying out the randomly sampling on i according to this probability. So the pooled sample value s belonging to this the window can be defined as:

$$s = a_\gamma, \quad \gamma \sim P\{P_1, P_2, ..., P_R\} \tag{5}$$

E. Full Connection Layer

The fully connected layer is to connect all the neurons of the previous layer to the neurons of the next layer one by one. The aligned one-dimensional feature vector obtains the output for the fully connected layer l through the activation function:

$$x^l = f(u^l) \tag{6}$$

$$u^l = w^l x^l + b^l \tag{7}$$

where, u^l is the net activation of the fully connected layer, which is got by carrying out the weight and offset on the output of the previous layer, w^l is the weight coefficient of the fully connected layer, and b^l is the bias of the fully connected layer.

IV. Voiceprint Recognition Method Improved Pooling Layer Based CNN

A. Voice Information Preprocessing

Because in the voiceprint recognition of voice information, the voice information need to be converted into a picture information that can be recognized by the convolutional neural network, so a series of preprocessing operations need to performed on the voice information [28]. A spectrogram is an intuitive three-dimensional image containing the speech information. It analyzes the characteristics of speech signals through a combination of time domain and frequency domain. The X-axis of the spectrogram represents the temporal transformation of speech information, and the Y-axis represents the speech in frequency. The change in color, the difference in color is the amount of energy at that time. This form of expression is a good reflection of a person's voice information, so different people's spectral maps are adopted to identify and classify.

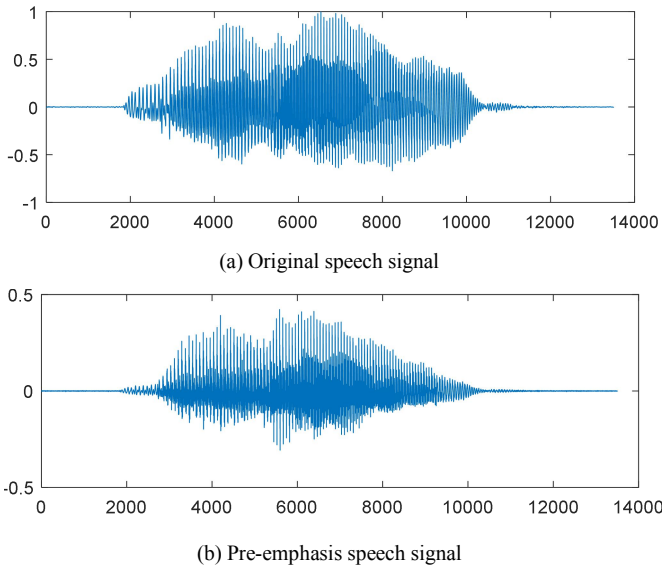
(1) Quantization and Pre-emphasis of Speech

For the collection of voice information, it is random and unstable. Therefore, the voice information need to be preprocessed before the speech recognition. The main operation is the quantization of the voice and the pre-emphasis. The speech information is an analog signal that is continuous in time and transforms in amplitude. Therefore, the speech information need to sampled and quantized. Sampling is to sample the speech information at intervals in the time domain, and the sampling frequency must be greater than the original sampling. Doubling the maximum frequency can ensure that the sampled information is not lost. By sampling the continuous signal, it is also necessary to quantify its amplitude. That is, the amplitude of the waveform of the speech signal is discretized, and the sample points having the same amplitude are dropped into the same interval, which is the quantization operation of the speech.

For the quantized speech signal, a pre-emphasis operation is required. In this way, the speech signal has a stronger high-frequency resonance value, which makes the short-time frequency smoother, and suppresses random noises, eliminates DC drift, and makes the unvoiced sound have higher energy. Through pre-emphasis processing, voice information has better vocal parameters and spectrum analysis capabilities. The pre-emphasis operation can be described as:

$$H(z) = 1 - \alpha z^{-1}, \alpha = 0.94 \sim 0.97 \tag{8}$$

where, α is the pre-emphasis coefficient. After the pre-emphasis operation of the pronunciation of the number 0, the processing effect is shown in Fig. 4.



(a) Original speech signal

(b) Pre-emphasis speech signal

Fig. 4 Pre-emphasis on the number 0.

It can be seen that after the pre-emphasis processing of the number 0, the waveform becomes smoother, which reduces the influence of noises on the speech signal. The high frequency portion of the graph has a more pronounced high frequency formant and better high frequency resolution.

(2) Speech Framing and Windowing

After the voice information is pre-emphasized, the waveform is still in an unstable state. So it need to be carried out the framin and windowing operations. The framing process is to make the transition of the speech information and the autocorrelation more smooth, so that each frame of the speech information is excessively smoother. The part that overlaps between two adjacent frames is called frame shift, and it is usually half of frame length, which is shown in Fig. 5. The framing-processed signal also requires windowing. The selection of windowing mainly follows three criteria: narrow transition zone, narrow main lobe, and low sidelobe height. According to these three standards, there are three commonly used window functions: Hamming window, Hanning window and matrix window, which are shown in Fig. 6. The calculation functions can be described as:

$$\omega(t) = \begin{Bmatrix} \frac{1}{T}, 0 \le |t| \le T \\ 0, \ |t| > T \end{Bmatrix} \tag{9}$$

$$\omega(t) = \begin{Bmatrix} \frac{1}{T}(\frac{1}{2} + \frac{1}{2}cos\frac{\pi t}{T}), |t| \le T \\ 0, \ |t| > T \end{Bmatrix} \tag{10}$$

$$\omega(n) = \begin{Bmatrix} 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right], 0 \le n \le N-1 \\ 0, \qquad other \end{Bmatrix} \tag{11}$$

These three window functions have different characteristics, so they have different advantages and disadvantages in different situations. Although the rectangular window has the narrowest main lobe, its sidelobe leakage is severe, so it is not used as a windowing function in this paper. Although the Hamming window has the smallest side lobes, it causes the side lobes to slow down and affect the decay speed. The performance of Hanning window is relatively balanced between the two, so the final choice of this paper is Hanning window to perform windowing operation. The Hanning window is framed and windowed on the voice information of digital 0, whose results are shown in Fig. 7 and Fig. 8.

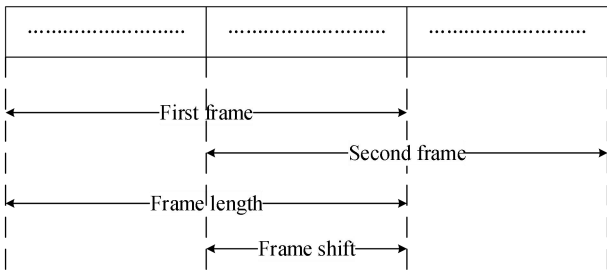
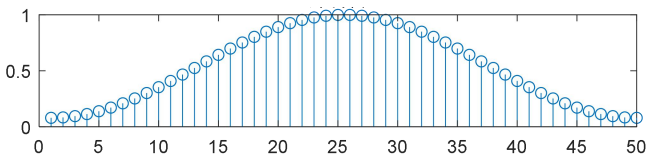
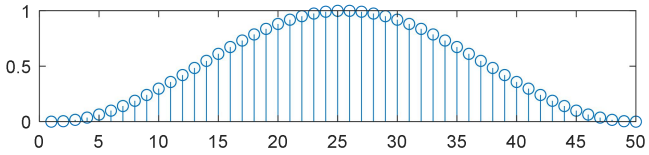


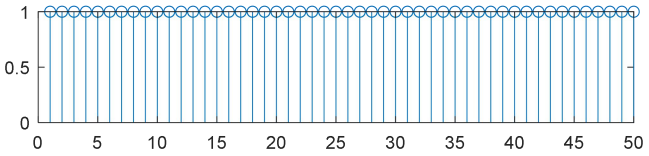
Fig. 5 Framing diagram.



(a) Hamming window



(b) Hanning window



(c) Rectangular window

Fig. 6 Three types of window function waveform.

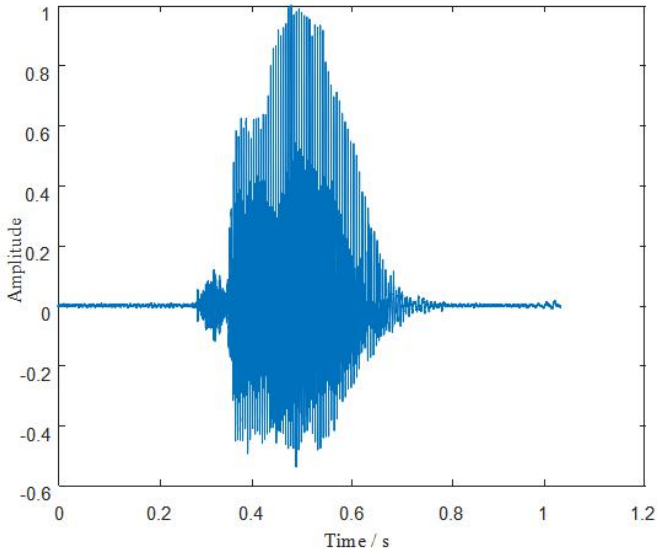


Fig. 7 Digital 9-voice waveform.

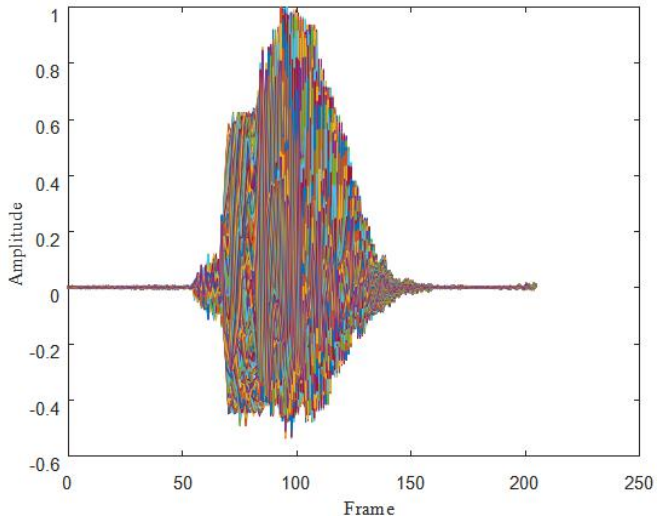


Fig. 8 The voice of the number 9 is processed by frame and window.

(3) Spectral Map

Different people's pronunciation characteristics are different, but it can't intuitively distinguish the differences in features of voice information. Therefore, in order to be able to express the characteristic parameters of the speech more intuitively, it is necessary to distinguish the features by using the spectral map. After segmentation and windowing of a piece of speech information, the short-time Fourier transform is performed on each frame, and the energy density of each frame is calculated, and then the energy density of each frame is sorted in time order. Finally, they are placed in a

two-dimensional matrix and converted into a two-dimensional images through a mapping relationship to obtain a spectral map.

Suppose the time domain signal is $x(n)$, which is a discrete sampled signal, where $n = 0,1...N-1$, and n is the sample number of the time domain signal, and N is the total length. By framing the speech, $x(n)$ is represented as $X_n(m)$, where n is the sequence number of each frame, m is the time serial number of the frame synchronization, and N is the frame length. Perform a short-time Fourier transform on the speech signal to obtain:

$$X_n(e^{jw}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-jwm} \quad (12)$$

The Fourier transform on the discrete time domain is then performed on the speech signal shown in Eq. (13) and (14).

$$X(n,e^{jw}) = \sum_{m=0}^{N-1} x_n(m)e^{-jwm} \quad (13)$$

$$X(n,k) = \sum_{m=0}^{N-1} x_n(m)e^{-j\frac{2k\pi m}{N}} \quad (14)$$

where, $0 \le k \le N-1$, $X(n,k)$ is an estimate of the short-term amplitude spectrum of $x(n)$. Then the energy density function at time m can be expressed as:

$$p(n,k) = |x(n,k)|^2 = (X(n,k) \times (conjX(n,k))) \quad (15)$$

where, $p(n,k)$ is a short-time Fourier transform of $x(n)$, and the value of k is taken as the ordinate, and the time n is taken as the abscissa. So the two-dimensional spectrogram is obtained. For example, the spectrogram of the voice of number 9 is shown in Fig. 9.

B. Voiceprint Database

There is no suitable database for voice voiceprint authentication, so in the database selection, this paper establishes a digital voiceprint database for experiments by the author B. W. Zheng. He is the a postgraduate student of University of Science and Technology Liaoning (USTL), and the related work was approved by USTL. The voiceprint speech library comes from 10 people, and each person extracts 200 sounds of 0-9 digits for a total of 20,000 samples. This is used as a speech database for the training set of voiceprint authentication. The speech spectrum of different voice information in the database is shown in Fig. 10.

C. Voiceprint Recognition Based on Convolutional Neural Network

(1) Voiceprint Recognition of Spectrogram

The spectrogram contains all the information that a person makes a sound, so the spectrogram is used as the input of the information, which has the best and most complete original information. Through the convolutional neural network, the invariance of the convolution is used to establish the sound so as to maintain the integrity and diversity of voice information. In the early studies, the focus of the processing of voice information lies on the RNN network, the focus of image processing is the CNN network. In this paper, the speech signal is converted into a two-dimensional image, and the idea of convolution is used to identify and classify the processing. Through short-term features, different people can be effectively identified and classified. In this paper, 20,000 spectral map information are extracted from 0 to 9 by 10 individuals. As shown in Fig. 11, this is the speech information spectrum of the number 1 of B. W. Zheng. In this paper, the spectral map of 20,000 voice messages with labels is used as the input of convolutional neural network, and then the network training is carried out by convolutional neural network, and 4000 of them are extracted as test sets for recognition and classification experiments.

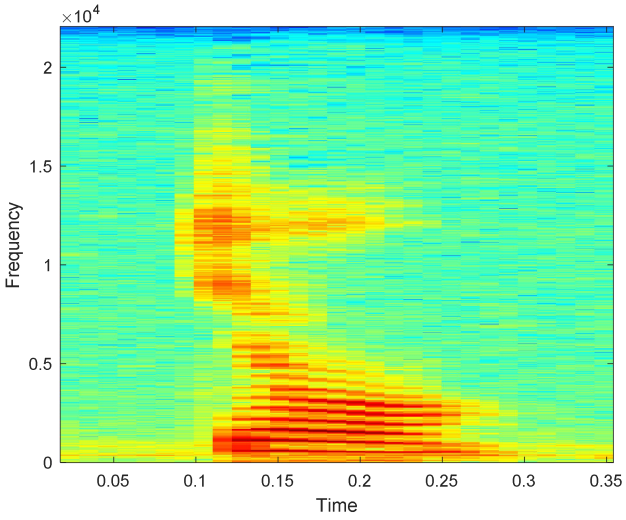
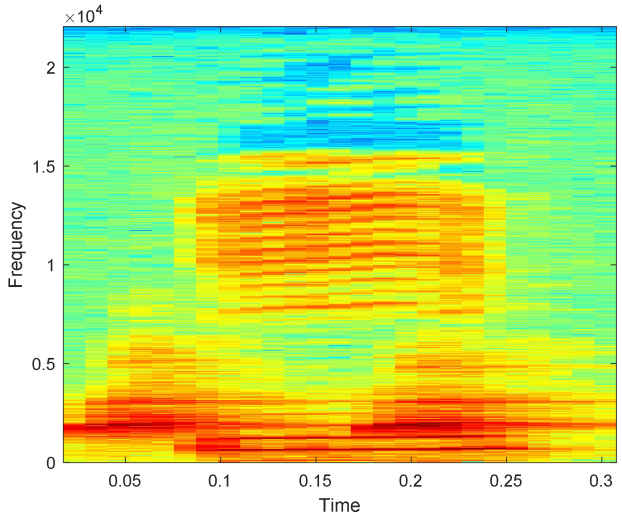
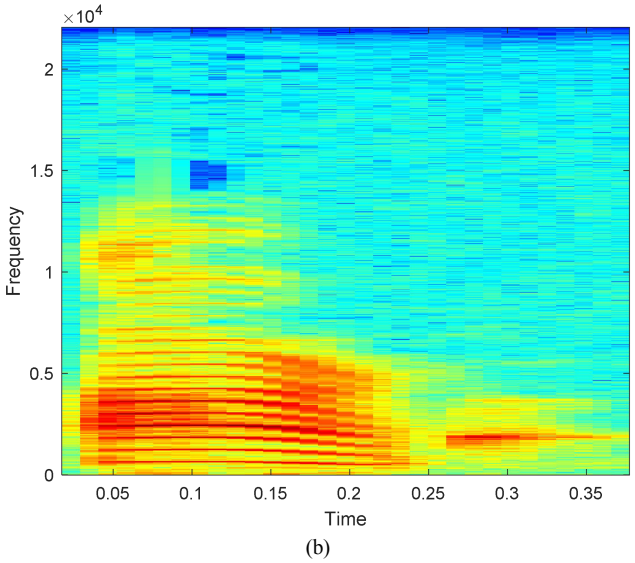


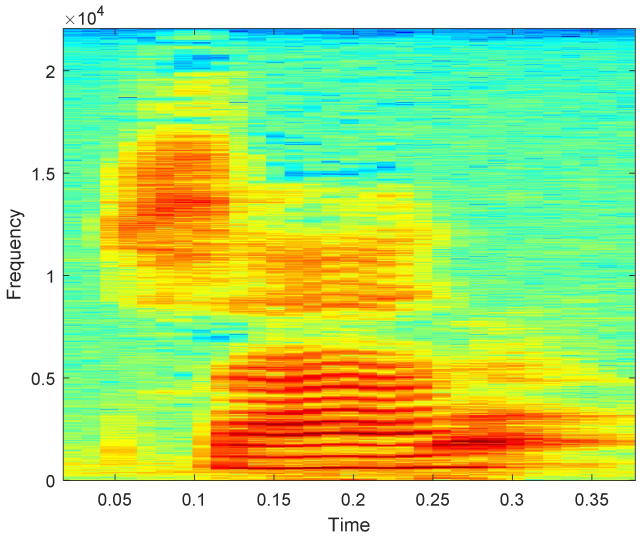
Fig. 9 Phonetic spectrogram of the number 9.



(a)



(b)



(c)

Fig.10 Spectrograms of different Numbers.

(2) Voiceprint Recognition Based on Convolutional Neural Network

In the speech recognition research, the most classic LeNet-5 network is selected to be modified. The voiceprint recognition experiments are carried out through the improved convolutional neural network. Fig. 12 is a structural diagram of the LeNet-5 network. Its network structure is relatively simple, consisting of two convolution layers, two pooling layers, and a fully connected layer and a softmax classification layer. The specifically improved LeNet-5 network structure is shown in Fig. 13.

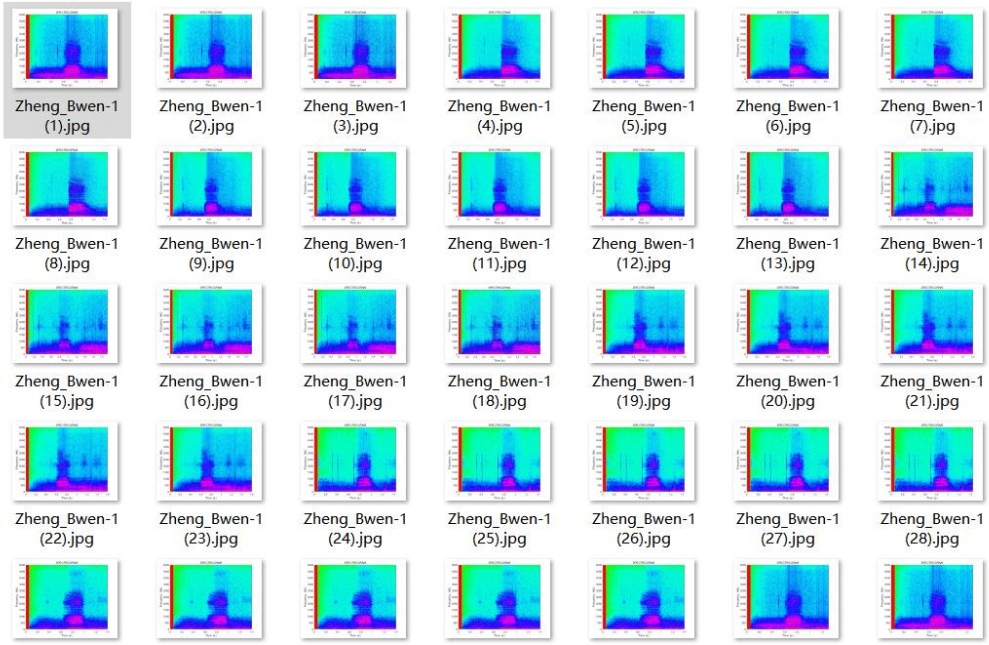


Fig. 11 The spectrogram of the number 1.

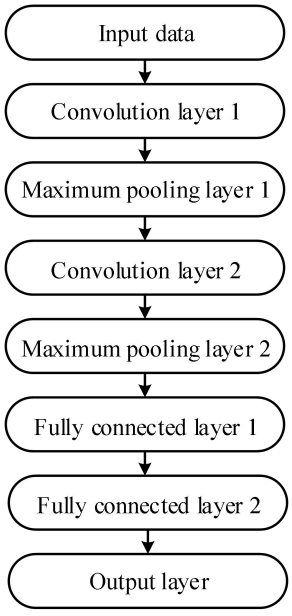


Fig. 12 Lenet-5 neural network.

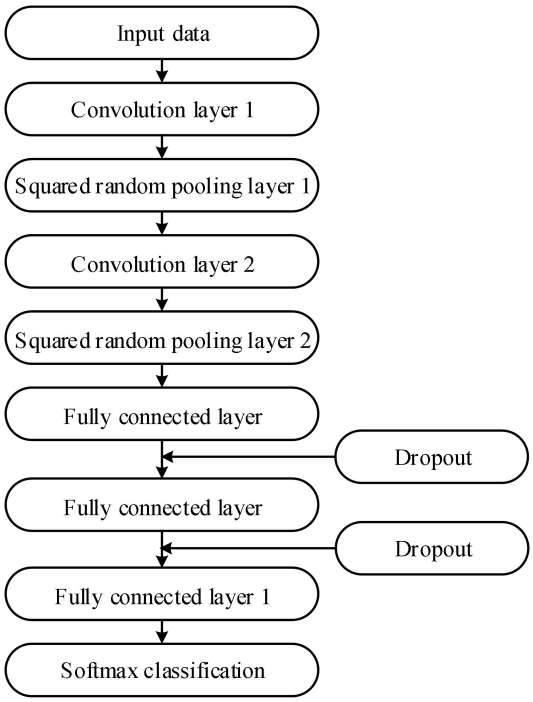


Fig. 13 Improved Lenet-5 neural network.

The proposed improved pooling method is applied to it as a pooling layer method for improving the network. The number of fully connected layers is increased, followed by the introduction of Dropout technology and Softmax as classifiers, which thereby increases the high capacity of the network model to prevent over-fitting in a small number and reduces the amount of network training. For the improved LeNet-5 network, the specific parameters are configured as shown in Table I. It can be seen from Tab. 1 that a 32*32 picture size spectrogram is entered into the improved LeNet-5 network as input of voice information. After two convolutions and pooling operations, 40 8*8 feature images are obtained. Then all the feature maps are converted into a one-dimensional image vector for full connection so as to get the one-dimensional feature vector of all classes of the output. Then the Softmax regression layer is adopted for classification in order to achieve the classification and recognition of voiceprint .

D. Voiceprint Recognition Experiment Based on Convolutional Neural Network

(1) Spectral Recognition of LeNet-5 Network

The application of LeNet-5 network in digital handwriting has been very successful, so by referring to the network method, the handwritten digits are converted into digital voiceprint images, and extracted voiceprint images are converted into 32*32 grayscale image used as input data. The LeNet-5 network digital handwriting input images are shown in Fig. 14.

TABLE. I SPECIFIC PARAMETERS OF IMPROVED LENET-5 NETWORKS

Label	Layer name	Output value	Parameter
1	Input layer	32*32*1 image	0
2	Convolution layer 1	32*32*20	（3*3*20+1）*20
3	Standardization layer 1	32*32*20	0
4	Activation function layer 1	32*32*20	0
5	Pooling layer 1	16*16*40	0
6	Convolution layer 2	16*16*40	（3*3*40+1）*40
7	Standardization layer 2	16*16*40	0
8	Activation function layer 2	16*16*20	0
9	Pooling layer 2	8*8*40	0
10	Dropout layer 1	8*8*40	0
11	Fully connected layer 1	1*1*100	（8*8*40+1）*100
12	Activation function layer 3	1*1*100	0
13	Dropout layer 2	1*1*100	0
14	Fully connected layer 2	1*1*100	100*100
15	Activation function layer 4	1*1*100	0
16	Fully connected layer 3	1*1* Number of output categories	100*100
17	Softmax layer		

The digital histogram of numbers are converted as the data input of the LeNet-5 network. The spectrom of the number 1 belonging to the same person is shown in Fig. 15. It can be seen that the digital voiceprint spectrum of the same person has obvious similarities, so the spectral map is converted into a grayscale image and then inputted into the LeNet-5 network as data to perform voiceprint recognition experiments. The data of experimental digital recognition classification results of voiceprint recognition are shown in Fig. 16 and Fig. 17.

The experiment of voiceprint recognition is carried out on a self-built database, which is a total of 20,000 data sets from 20 people. 80% of them were used as training sets, 20% of which were used as test sets. The convolutional layer in the convolutional neural network has a convolution kernel size of 3*3, and the pooled layer pooling kernel has a square random pool of 2*2. Dropout is set as 0.5, the initial learning rate is 0.1, which gradually decreases to 0.009, and the weights are initialized with the a mean of 0.

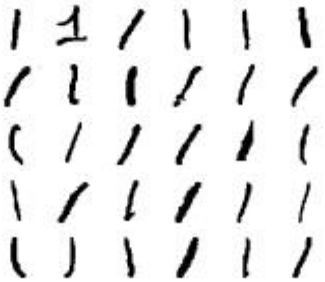


Fig. 14 Handwritten numerals 1.

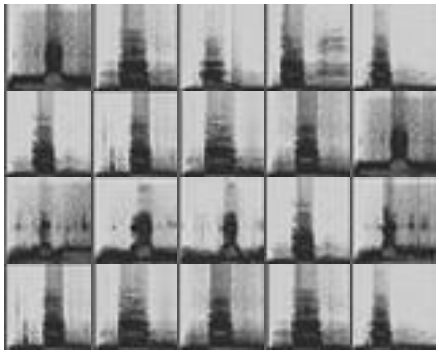


Fig.15 The spectrogram of the number 1 for the same person.

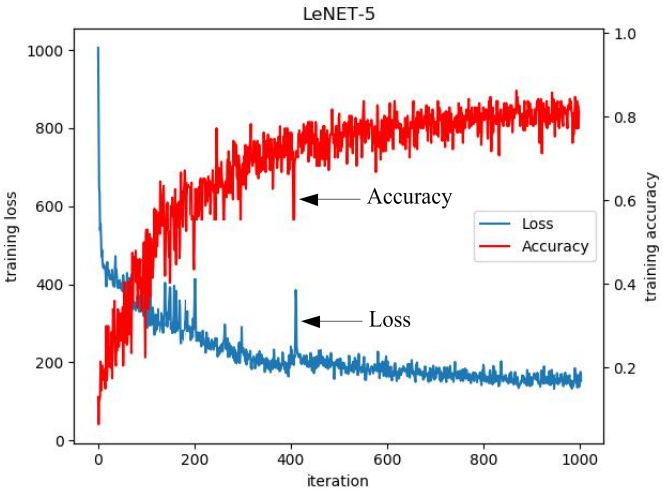


Fig. 16 The loss and recognition rate curve of improved LeNet-5 for voiceprint recognition.

```
Lenet_voiceprint_recognition
Optimization Iteration:    970, Training Accuracy:  83.5%
Optimization Iteration:    980, Training Accuracy:  82.9%
Optimization Iteration:    990, Training Accuracy:  79.9%
Optimization Iteration:   1000, Training Accuracy:  76.8%
Time usage: 00:12:51
Accuracy on Test-Set: 78.8%
```

Fig. 17 The experimental results of improved LeNet-5 for voiceprint recognition.

It can be seen from the experimental results of the voiceprint recognition that the loss value of the experiment gradually decreases and converges, eventually becoming stable, and the final loss value is about 0.2. As the number of iterations increases, the recognition rate also increases and tends to stabilize at around 80%. The recognition rate of the final test set is 78.8%. Experiment results can prove that it is feasible to convert the waveform of speech into a spectrogram for voiceprint authentication, but because of the limitations of the LeNet-5 convolutional neural network, it is necessary to convert the color spectrogram into a grayscale spectrogram, and the resolution is reduced, which makes the spectral map information not rich in originality. Although the network generalization is enhanced by changing the pooling method to reduce the possibility of over-fitting, the accuracy of voiceprint recognition is still too low. However, the digital spectrograms have a high recognition rate in the classification of voiceprints.

(2) Spectral Recognition of VGG-16 Network

Through the improved LeNet-5 network experiments, the lack of voiceprint recognition and the factors are analyzed. In this paper, the color map is preprocessed and used as input data to perform voiceprint identification experiments based on the improved VGG-16 network. The parameters of the network training are set as follows: Batch size is selected to be 256, dropout is 0.5, the initial learning rate is 0.1, which gradually decreases to 0.001, the regularization mode is Weight decay, and the weights are initialized with a mean of 0. The experimental results are shown in Fig. 18 and Fig. 19.

Through the voiceprint recognition experiment results of the VGG-16 network, it can be seen that the loss value gradually decreases and converges, and tends to be stable after 400 iterations, and the final loss value is about 0.1. The recognition rate of the test on 4000 voiceprint pictures was 91.2%. Experiments results can prove that the spectral map retains the original three-channel color information, which not only retains the original characteristics of the voiceprint, but also enriches the extracted feature information, so that the recognition accuracy of the voiceprint on the convolutional neural network is improved.

E. Automatic Voiceprint Recognition System Based on Convolutional Neural Network

(1) Algorithm Flowchart of Voiceprint Recognition System Algorithm

The voiceprint verification of the convolutional neural network method mainly includes the input of voice information, the detection of language information, the preprocessing operation of voice information, the generation of spectral maps, the training of models, and the matching and verification of features. The operational flowchart of the voiceprint verification system is shown in Fig. 20.

(1) Voice information input. Information collection is performed on the voiceprint verifier by the information collecting device.

(2) Language information detection. The collected voice information is detected to check whether it meets the input voice standard.

(3) Voice information preprocessing. quantizing, pre-emphasizing, framing, and windowing operations are performed on the collected voice information.

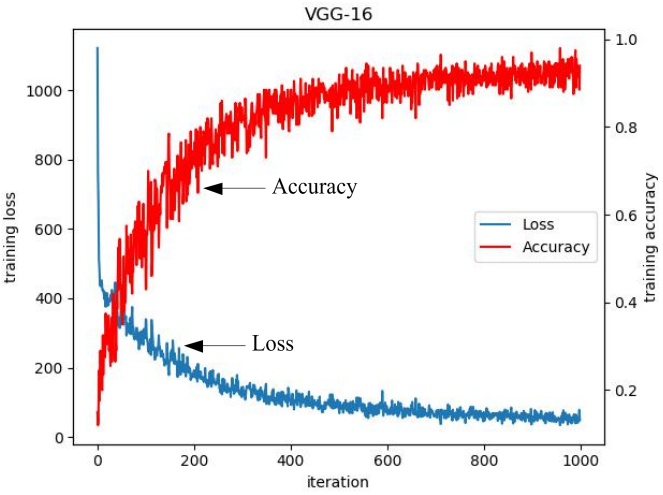


Fig. 18 The loss and the recognition rate curve of improved VGG-16 for voiceprint recognition.

```
VGG-voiceprint-recognition
Optimization Iteration:    960, Training Accuracy:  91.5%
Optimization Iteration:    970, Training Accuracy:  93.5%
Optimization Iteration:    980, Training Accuracy:  86.9%
Optimization Iteration:    990, Training Accuracy:  90.2%
Optimization Iteration:   1000, Training Accuracy:  92.5%
Accuracy on Test-Set: 91.2%  (3648/4000)
```

Fig. 19 The experimental results of improved VGG-16 for voiceprint recognition.

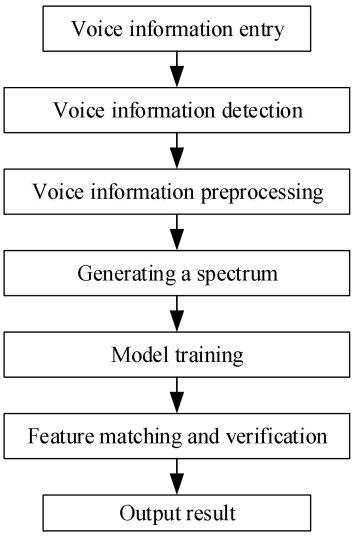


Fig. 20 The system flow chart of voiceprint recognition.

(4) Generate a spectral map. A short-time Fourier transform is used to generate a spectral map of the pre-processed speech information.

(5) Model training. The extracted spectrum map is used for network training of the model.

(6) Matching and verification of features. Match the features extracted by the user's spectral map with the information of the speech library, and verify whether it belongs to the voice information of the database and determine which user belongs to.

(2) Voiceprint Recognition System

A graphical user interface (GUI) is built for automatic speech recognition system. The system collects information on the voiceprint recognizer through recording, mainly relying on the tester to read a specific number, and performs a series of pre-processing operations on the obtained voice information, performs network training and recognition on the read voice voiceprint, matches the existing database, and finally gets the object of the voiceprint. So by changing the initial learning rate, the influence of learning rate on recognition accuracy is observed. Under the same experimental conditions, the initial learning rate of the convolutional neural network is set as 0.001, 0.01, 0.05, 0.1, 0.15, 0.2 and 0.5, respectively. Training and recognition through different learning rates can be obtained differently. The experimental results are shown in Table II.

It can be seen from Table II that the recognition accuracy of the training set and the test set is the best when the initial learning rate is 0.1, so finally 0.1 is selected as the initial learning rate of the automatic voiceprint recognition system. Then, by changing the time of information collection, the influence of voice information of different lengths of time on the recognition efficiency is observed. In the case where the other conditions of the experiment are the same, the time for collecting the voice information of the experimental object is changed to 0.5, 1, 1.5, 2, 3, 4, and 5 second, respectively. By changing the sampling time to carry out simulation experiments, the accuracy of network identification is shown in Table III.

Seen from Table III, when the time of voice acquisition is 1 second, the recognition accuracy of the training set and the test set are the highest, so 1 second is selected as the automatic voiceprint recognition system when selecting the voice acquisition time. Combined with the above experimental results, in the real-time voiceprint recognition system basedon convolutional neural network, the initial learning rate is 0.1 and the voice information acquisition time is 1 second.

The specific operation of the real-time voiceprint recognition system is described as follows. At the beginning of the experimental test, let the subject press the “Record” button, then one second is applied for the recognition object to pronounce, and the experiment object needs to give the microphone an Arabic number between 0 and 9. At this time, the system collects the voiceprint information of the tester, and through the pre-processing operation, converts the voice information of the collected object into a waveform diagram and a spectrum map of the voice and displays it. Then click the “Start Recognition” button. At this time, the system will train and identify the collected voice information according to the preset learning rate and initialization value. Soon the system will get a recognition through the database. The name information is displayed in the recognition result box. If you need to make the next recognition, you need to click to clear the reset, the information displayed by the system will be cleared and you can start the next test.

Through simulation experiments, the accuracy of the automatic speech recognition system is about 80.4%, which is much lower than the accuracy of 91.2% obtained through direct testing of the voice library. The main reason is that the automatic speech recognition system is in an unrestricted environment. When information is collected, it is affected by the noise of the external environment and is also affected by the equipment for information collection. Therefore, it is still lacking to carry out the experiment by using the preprocessing method of database voiceprint recognition, and it is necessary to obtain improvement, so as to enhance the interference of external factors such as noise to obtain better voiceprint information and improve the recognition accuracy.

V. Conclusion

In this paper, the speech information is quantized, pre-emphasized, framing, windowed, and the processed speech information is converted into a spectrogram for classification and recognition, and a digital voiceprint database is established. Then the traditional LeNet-5 network is adopted on the voiceprint digital recognition. The grayscale images are obtained by the color digital spectrum, and then the proposed improved pooling method is used to replace the original pooling method. On the other hand, the idea of Dropout is introduced and added after the two fully connected layers, and the Softmax classifier is adopted to realize the classification. The constructed improved LeNet-5 network is adopted to carry out the digital sound experiments ant the related experiment results show that the recognition rate was not ideal. Then the three-channel color digital voiceprint spectrum was tested on the VGG-16 network, and the experimental result with an accuracy of 91.2% was obtained.

TABLE. II THE INFLUENCE OF NETWORK INITIAL LEARNING RATE ON RECOGNITION EFFECT

Learning rate	0.001	0.01	0.05	0. 1	0.15	0.2	0.5
Convergence	Slow	Slow	Faster	Fast	Faster	Faster	Concussion
Recognition rate (training set)	71.1%	78.5%	90.5%	91.5%	89.5%	85.4%	73.5%
Recognition rate (test set)	71.9%	70.2%	88.5%	91.2%	89.4%	83.0%	70.4%

TABLE. III THE EFFECT OF SPEECH ACQUISITION TIME ON RECOGNITION PERFORMANCE

Acquisition time	0.5s	1s	1.5s	2s	3s	4s	5s
Recognition rate (Training set)	88.3%	91.5%	87.2%	86.2%	85.9%	83.8%	78.3%
Recognition rate (Test set)	85.1%	91.2%	88.2%	86.4%	86.0%	84.6%	78.5%

REFERENCES

[1] S. M. Siniscalchi, "Combining Speech Attribute Detection and Penalized Logistic Regression for Phoneme Recognition," *Neurocomputing*, vol. 93, pp. 10-18, 2012.

[2] F. Narducci, and S. Barra, "Biometric Recognition in Surveillance Scenarios: a Survey," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 515-541, 2016.

[3] R. Lippmann, "Review of Neural Networks for Speech Recognition," *Neural Computation*, vol. 1, no. 1, pp. 1-38, 2014.

[4] J. S. Wang, Y. L. Ruan, B. W. Zheng, and S. Z. Gao, "Face Recognition Method Based on Improved Gabor Wavelet Transform Algorithm," *IAENG International Journal of Computer Science*, vol. 46, no. 1, pp. 12-24, 2019.

[5] L. J. Wang, and Y. Y. Zhao, "Design and Implementation of Diamond Press's Pressure Hammer Protection Based on Voiceprint Recognition," *Applied Mechanics & Materials*, vol. 539, pp. 156-160, 2014.

[6] X. Kou, Y. Cao, Q. Wang, and H. Qiao, "Sub-assembly Recognition Algorithm and Performance Analysis in Assembly Sequence Planning," *The International Journal of Advanced Manufacturing Technology*, vol. 107, no. 3, pp. 971-981, 2020.

[7] C. Y. Huang, Q. Z. Zhang, and B. Lai, "Research on MEL Parameter Extraction from Voiceprint Recognition," *Chinese Journal of Power Sources*, vol. 35, no. 4, pp. 433-435, 2020.

[8] Z. Zhang, Geiger, Jürgen, J. Pohjalainen, E. D. Mousa, W. Jin, and B. Schuller, "Deep Learning for Environmentally Robust Speech Recognition: an Overview of Recent Developments," *ACM Transactions on Intelligent Systems & Technology*, vol. 9, no. 5, pp. 1-28, 2017.

[9] Y. Lavner, J. Rosenhouse, I. Gath, "The Prototype Model in Speaker Identification by Human Listeners," *International Journal of Speech Technology*, vol. 4, no. 1, pp. 63-74, 2001.

[10] G. Nancie, M. Diarmid, M. I. Fergus, and J. Mervyn, "Usability Evaluation of Voiceprint Authentication in Automated Telephone Banking: Sentences Versus Digits," *Interacting with Computers*, vol. 23, no. 1, pp. 57-69, 2011.

[11] H. S. Bae, H. J. Lee, and S. G. Lee, "Voice Recognition-based on Adaptive MFCC and Deep Learning for Embedded Systems," *Journal of Institute of Control Robotics & Systems*, vol. 22, pp. 1542-1546, 2011.

[12] Z. C. Wang, "A Voiceprint Recognition Algorithm Based on MFCC and Formants," *Computer Knowledge & Technology*, vol. 12, no. 4, pp. 188-190, 2016.

[13] H. M. Yin, J. M. Wang , and H. H. Li, "Voice Print Recognition Based on Data Training Missing Feature Detection and Reconstruction," *Computer Engineering and Applications* , vol. 52, no. 22, pp. 159-163, 2016.

[14] B. Prabhu, and M. Pradeep, "Implementation of Voice Recognition Wireless Home Automation System with ZigBee," *Social Science Electronic Publishing*, vol. 3, no. 4, pp. 1-14, 2016.

[15] L. G Kersta, "Voiceprint Classification," *The Journal of the Acoustical Society of America*, vol. 37, no. 6, pp. 1217, 1965.

[16] Q. H. He, H. H. Zhu, and Z. Y. Zhu, "Design of Cancelable Template Based on Random Transformation and Dynamic Projection,"*Journal of South China University of Technology*, vol. 41, no. 1, pp. 1-7, 2013.

[17] T. Tuncer, and S. Dogan, "Novel Dynamic Center Based Binary and Ternary Pattern Network Using M4 Pooling for Real World Voice Recognition," *Applied Acoustics*, vol. 156, pp. 176-185, 2019.

[18] L. G. Kersta, "Voiceprint Classification for an Extended Population," *Journal of the Acoustical Society of America*, vol. 39, no. 6, pp. 1239-1240, 1966.

[19] H. B. Kekre, Vaishali, and Kulkarni, "Speaker Identification by Using Vector Quantization," *International Journal of Engineering Science and Technology*, vol. 2, pp. 1325–1331, 2010.

[20] A. Mcleod, M. Steedman, "HMM-based Voice Separation of MIDI Performance," *Journal of New Music Research*, vol. 45, no. 1, pp. 1–10, 2016.

[21] X. Li, and Z. Wang. "A HMM-based Mandarin Chinese Singing Voice Synthesis System," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 2, pp. 192-202, 2016.

[22] X. Yu, S. He, Y. S. Peng, and W. Zhou, "Pattern Matching of Voiceprint Recognition Based on GMM," *Communications Technology*, vol. 48, no. 1, pp. 97-101, 2015.

[23] Y. P. Li, L. Lin, and D. Y. Tao, "Research on Identification of Electronic Disguised Voice Based on GMM Statistical Parameters," *Computer Technology & Development*, vol. 27, no. 1, pp. 103-106, 2017.

[24] X. Ting, W. Wei, and Z. Xin, "Voice Print Recognition Algorithm Based on the Natural Language Processing," *Computer Simulation*, vol. 30, no. 6, pp. 244-247, 2013.

[25] C. Zhang, X. Pan, H. Li, A. Gardiner, I. Sargent, J. Hare, and P. M. Atkinson, "A Hybrid MLP-CNN Classifier for Very Fine Resolution Remotely Sensed Image Classification," *ISPRS Journal of Photogrammetry & Remote Sensing*, vol. 140, pp. 133-144, 2018.

[26] P. P. Liu, L. D. Shi, Z. Miao, B. X. Jin, and Q. Z. Zhou, "Relative Distribution Entropy Loss Function in CNN Image Retrieval," *Entropy*, vol. 22, no. 3, pp. 321, 2020.

[27] A. Khatami, M. Babaie, H. R. Tizhoosh, A. Khosravi, T. Nguyen, and S. Nahavandi, "A Sequential Search-space Shrinking Using CNN Transfer Learning and a Radon Projection Pool for Medical Image Retrieval," *Expert Systems with Applications*, vol. 100, pp. 224-233, 2018.

[28] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition," *Procedia Computer Science*, vol. 116, pp. 523-529, 2017.

Wei-Zhong Sun received his B. Sc. degree in Electronic and Information Engineering and his M. Sc. degree in Control Science from University of Science and Technology Liaoning, China in 2004 and 2008, respectively. He is a lecturer of School of Computer Science and Software Engineering, University of Science and Technology Liaoning. His main research interest is image processing, pattern recognition and embedded system.

Jie-Sheng Wang received his B. Sc. And M. Sc. degrees in Control Science from University of Science and Technology Liaoning, China in 1999 and 2002, respectively, and his Ph. D. degree in Control Science from Dalian University of Technology, China in 2006. He is a professor and Doctor's supervisor in School of Electronic and Information Engineering, University of Science and Technology Liaoning. His main research interest is intelligent control and computer integrated manufacturing.

Bo-Wen Zheng is with School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. His main research interest is modeling methods of complex process and intelligent optimization algorithms.

Zhong-Feng Li is a senior engineer of Yingkou Institute of Technology, Yingkou, 115000, P. R. China. His main research interest is modeling methods of complex process.