Voice Print Recognition Check-in System Based on Resnet

Ruxin Zheng *, Yanyan Fang, Jie Dong

Wuhan University of Technology, Wuhan, China

* Corresponding Author Email: 298693@whut.edu.cn

Abstract. With the development of modern technology and the rise of artificial intelligence, the application scenarios of identity authentication technology are becoming more and more complex, especially the current situation of the spread of the novel coronavirus, the traditional identity authentication technology can no longer meet people's actual needs, and the society urgently needs a secure and convenient identity authentication technology. Voice print recognition technology is an identity recognition method that uses specific feature extraction methods to extract the features that can identify the speaker's identity from the original speech input, and then uses these features to identify the speaker's identity. Aiming at the above problems, this paper proposes a deep learning-based voicing recognition algorithm, which is based on the theoretical knowledge of deep learning. The research work includes the following aspects: providing convolutional network to extract features; The "speech feature template library" is established by massive training. Research on matching and recognition algorithm; Research on check-in system based on voice print recognition. Based on this algorithm, this paper designs and implements a voiceprint recognition check-in system with clear interactive interface. The system has the functions of adding members, refreshing the voice library, checking in with the voice print, clearing the results and so on. The system interface will display the data of the voice print library and the historical check-in record, as well as the current recognition results, accuracy and check-in time. The average recognition rate of the system is about 95%, which can meet the requirement of practical application.

Keywords: Voiceprint recognition, Deep learning, Convolution network.

1. Introduction

1.1. Project Background

In daily life, we transmit a variety of information to our surroundings through speech. A speech not only conveys the language information about the text content, but also conveys a lot of information reflecting the characteristics of the speaker, such as gender, health status, emotional state and identity information.

Voiceprint recognition is a combination of computer, acoustics and life science. It is biometric technology. Organs due to the pronunciation of the individual differences of the physiological characteristics and the degree of protection for pronunciation organs, the day after tomorrow speech organ growth environment of its impact and pronunciation habits behavior characteristics, voice print as well as fingerprint, iris, facial features, has uniqueness, give voice print features extracted from speech, brought may identify the speaker identity. As an indispensable way of communication in daily life, it is convenient to use voice for user identification. Dynamic password is used for identification, which can effectively avoid fake voice such as audio recording and splicing in advance, and the system has high security. And the process of voice collection involves less user privacy data, which is easy to be accepted by users.

Before the 21st century, voice recognition technology has not been widely used due to low accuracy and poor computing ability of equipment. Since the beginning of this century, the computing technology and integration ability of computers have been rapidly developed, and the voice print recognition technology is constantly updated and iterated, and the cost is gradually reduced, which brings opportunities for the promotion and application of voice print recognition. Voiceprint recognition technology is an important branch of artificial intelligence.

1.2. Research Significance

The task of voice print recognition can guide the privacy protection of individuals and enterprises. The research in this field has both theoretical value and practical value.

Check-in system based on voice print recognition can greatly convenient enterprises, schools, such as the management of the subordinate personnel, in today's new crown epidemic situation is particularly serious, to a voiceprint recognition instead of traditional identification methods, such as face recognition, fingerprint recognition, realize voice print sign in through sign-in page, and other functions, managers need to look at only the page access to sign in, adjust the management policy in time according to the actual situation. The system will provide great convenience for managers and their staff at the same time, improve enterprise efficiency, campus management efficiency, and improve the enthusiasm of employees for work and the enthusiasm of students for learning.

1.3. Research status of voiceprint recognition technology

Foreign voiceprint recognition technology started earlier. In 2006, ABN AMRO, a Dutch bank, first used Voice Vault, a voice-print recognition system based on users' prerecorded privacy questions. In 2009, Andew NG et al. applied deep learning to voice-print feature extraction for the first time, and achieved good results. Some researchers use deep convolutional neural networks to extract the feature vectors of speech signals, or use the spectrogram of speech signals as the input of convolutional neural networks to classify and recognize speaker speech. In 2018, Google applied GE2E Loss to voiceprint recognition, and compared the value obtained by each update with the value of multiple individuals to further improve the recognition accuracy. Foreign scholars systematically studied and improved the voiceprint recognition technology, and obtained rich research results.

The research work of voice print recognition started late in China. From 2003 to 2004, in the internationally renowned NIST voice recognition competition, the Two-speaker voice recognition system was proposed for the first time by a team led by Dai Beibei from the University of Science and Technology of China, and won the second and third place for two consecutive years. In 2008, the voice print recognition team of IFlytek won the first prize in the SRE Voice Print Recognition Competition with its self-developed USTC-IFLY system. In 2012, Deng Li from the acoustics research team of Microsoft applied the following Deep Neural network to voice print recognition for the first time, which greatly improved the recognition accuracy, and also led to the research upsurge of DNN-HMM(Deep Neural Networks-Hidden Markov Model) Acoustic Model. In 2017, the Baidu voice recognition team proposed a Deep Speaker using Residual network Convolutional Neural network (ResCNN) and GRU architecture. The recognition accuracy is 60% higher than that of the I-vector benchmark method based on Deep Neur Al Net Works (DNN).

1.4. Research status of check-in system

In the traditional attendance management, the more commonly used methods are paper check-in, punch check-in and fingerprint check-in. The traditional check-in method has the disadvantages of low efficiency and substitutability, which makes it difficult to guarantee the attendance rate. The way of punching in can not only be replaced, but also a waste of resources to set up check-in personnel at each punching point. Fingerprint and face recognition methods can be guaranteed to be irreplaceable, but there are problems such as a large amount of data collection before installation, the normal use of fingerprint or face recognition, and the epidemic prevention effect cannot be guaranteed during the epidemic. Code scanning has the problem of remote signal vulnerability, which cannot fully guarantee the success of attendance.

2. Related technology introduction

The working process of voice print recognition system is mainly divided into two stages: training and testing. In the train phase. The system will collect the original voice signal through a series of preprocessing operations, and then according to the system set parameters. The extraction method

extracts the corresponding feature parameters for each speech. The speech features are input to the recognition network for training. The speaker model is obtained and stored in the sample library for subsequent matching. In the test phase, the speech samples in the test sample set are preprocessed and feature extracted, and the features of the speech samples to be tested are matched with the registered speaker model to determine the identity of the speaker to be tested. Voice print recognition. The system framework is shown in Figure 1.

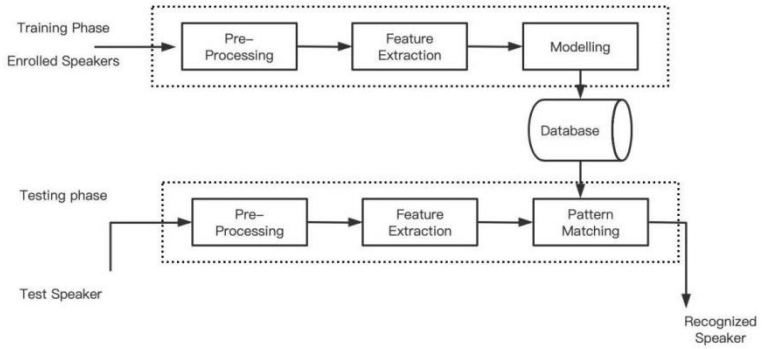

Figure 1. Frame of voiceprint recognition system

2.1. Speech signal preprocessing

On the mask voice signal. The purpose of speech signal preprocessing is to remove the environment and channel noise as much as possible. And convert the original speech signal into subsegments that facilitate subsequent feature extraction operations. Preprocessing includes preweighting of speech signals, speech framing, short frame speech windowing and other operations, specific.

2.1.1 preemphasis

The purpose of the preweighting operation is to enhance the energy amplitude of the input speech in the high frequency band. In essence, the frequency domain of the speech is multiplied by a coefficient related to the frequency of the speech, which can be simply understood as a high-pass filter:

$$s_{n'} = s_n - \mu s_{n-1}$$

2.1.2 framing

Frame segmentation refers to the segmentation of continuous speech samples into sub-frames of equal length, which is conducive to short-term signal analysis. Speech signal is a time-varying signal, the amplitude of which fluctuates over time. Therefore, in speech processing related tasks, speech is usually first divided into a series of 10-30ms short frames, and the corresponding features are extracted based on the short frames. At the same time, some adjacent frames overlap with one third to one half of the frame length to avoid the loss of voice information caused by subsequent windowing operations.

2.1.3 add window

Windowing operation is to multiply each frame of speech waveform obtained by the original speech frame with the window function point by point. The window is added to the short frame speech, so that the energy at both ends of the signal can gradually decay close to the zero value, so that the energy of the signal is mainly distributed in the main lobe, which is closer to the real spectrum. The common window function is Hamming window, and the corresponding window function formula is shown in equation:

$$s'(n) = w(n) * s(n)$$

$$W(\mathrm{n},\alpha)=(1-\alpha)-\alpha*\cos\left[\frac{2n\pi}{N-1}\right],0\le n\le N-1$$

Where N is the frame length, α is generally 0.46.

2.2. Feature Extraction

The purpose of feature extraction is to compute information from the original audio signal that can be used to characterize speech for a specific recognition task. Mel frequency cepstral coefficient is a commonly used speech feature in voice print recognition system.

Mel Frequency Cepstral Coefficient (MFCC) is designed according to the characteristics of human auditory model. Studies on human auditory system have found that the human ear has different perceptual sensitivity to sound signals of different frequencies, and there is not a linear mapping relationship between the perceptual response degree to sound signal frequency and the actual frequency of the signal. The actual audio frequency and the frequency perceived by the human ear meet the nonlinear mapping relationship from the linear frequency scale to the MEL frequency scale, as shown in equation:

$$F_{mel}(f)=2595*\lg\left(1+\frac{f}{700}\right)$$

MFCC feature extraction consists of two main steps: conversion to MEL frequency and cepstrum analysis. Conversion to Meir frequency: The cochlea plays an extremely important role in the ability of the human ear to correctly distinguish speech in noisy silence. It essentially functions as a filter bank on a frequency scale. Cepstral analysis: the signal is decomposed, and the convolution and sum of the two signals are transformed into the sum of the sum by logarithm operation. The MFCC extraction process is shown in Figure 2.

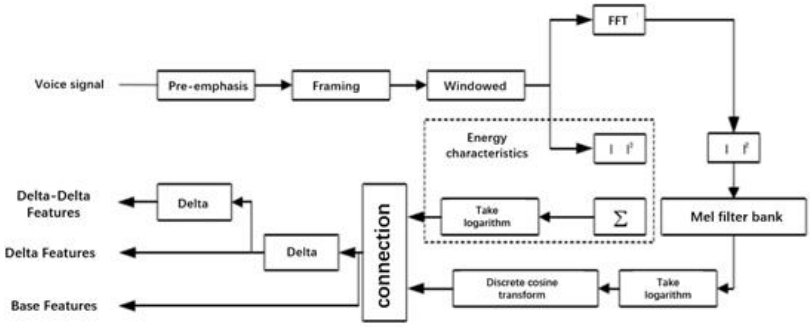

Figure 2. MFCC feature extraction process

The calculation steps of MFCC are as follows:

1) The original speech signal is divided into frames, and each frame is multiplied by a window function to make the signals of adjacent frames more continuous and prevent spectrum leakage.

2) Perform fast Fourier transform on each frame of speech after frame segmentation and windowing to calculate the power spectrum.

3) Map the power spectrum onto the MEL scale using the triangle filter bank of the MEL scale.

4) Transform the filtered power spectrum to the logarithmic domain and execute the discrete cosine transform.

5) The amplitude value of spectrum obtained by difference operation is the MFCC characteristic parameter.

The design of MFCC is based on the auditory characteristics of the human ear, which can effectively extract the signal features in the low-frequency domain. It is a classic feature used in speaker recognition algorithms.

2.3. Voice print recognition model

2.3.1 Concept of convolutional neural network

Convolutional neural network is the mainstream neural network for image processing related tasks. In order to avoid the problem of gradient disappearance or gradient explosion caused by too deep network, deep residual structure is introduced in this paper to improve the performance of voiceprint recognition check-in system.

2.3.2 Resnet Feature Extraction Network

Deep residual network is proposed to solve the performance degradation problem caused by too deep network layers. For convolution neural network, the role of each layer of learning a mapping from input to output function, in the residual structure, however, is that part of the learning an input to the output layer of residual function, F (x) = H (x) - x, the depth of the convolution of the neural network gradient disappear, and network degradation by the transformation of the residual structure has been effectively alleviated. The structure of Resnet residual learning module is shown in the figure.

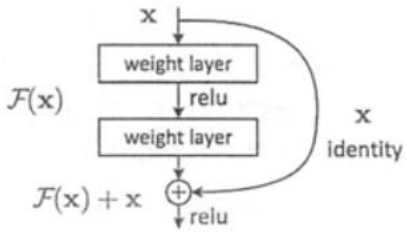

Figure 3. Residual learning module of Resnet

In this paper, we hope to use a deep network structure to extract multi-level and high-level information from the input spectrogram features, avoid the problem of performance degradation caused by two deep layers of convolutional neural network, and improve the recognition performance of voice print recognition system. Based on the above considerations, we use ResNet50 as the back-end recognition model. ResNet50 has two basic blocks named Conv Block and Identity Block. The dimensions of Conv Block input and output are different, so Conv Block cannot be consecutively connected. Its function is to change the dimension of network. Identity Block Input dimension and output dimension are the same and can be concatenated to deepen the network. The overall network structure is as follows:

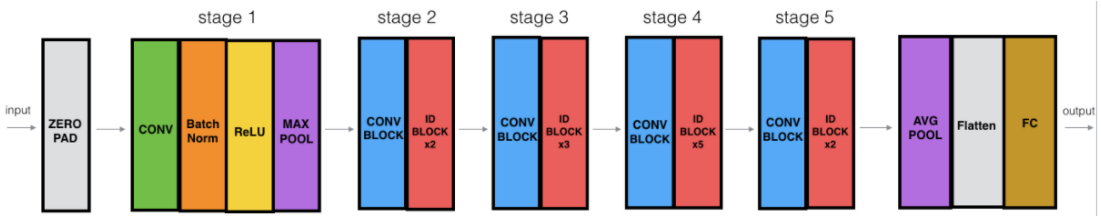

Figure 4. Overall structure of ResNet50

3. Voiceprint recognition check-in algorithm based on deep learning

3.1. Data set processing

This project will conduct experiments on the Free ST Chinese Mandarin Corpus dataset, which contains 102600 voice data of 855 people. First create a data list in the format of < Voice file path \ T Speech classification label > to facilitate later reading and reading using other speech data sets. By writing the corresponding function to generate the data list, the different speech data sets are written in the same data list, so that we can directly generate the TFRecord file in the next step. Finally, component analysis is carried out to investigate the effectiveness of multi-channel convolutional neural network design.

3.2. Obtaining training data

Using the above data list, voice data can be directly converted into training data. Firstly, the preprocessing, such as preweighting, windowing and framing are analyzed, and implemented by Python programming language. Then, the Merle frequency cepstral coefficient is used to extract the voice print features. Firstly, the acquisition of voice print information, that is, the recording, is completed by Pyaudio module in different backgrounds. In the process of conversion, the mute part of the audio is cut out, which can reduce the noise of training data and provide training accuracy. After creating the TFRecord file, in order to be able to read the TFRecord file in training, they create a program to read the training data. The known main function is used to solve the loop through the feature table, and then the similarity operation is performed on it, so that the variable parameter pro is maximized in all the similarities, and Pro is returned, which is the final similarity.

3.3. Model training

To build a ResNet50 classification model, input_shape is set to (128, none, 1) mainly to accommodate input of other audio lengths and to predict input of arbitrary size. Class_dim is the total number of classifications, and the previously trained weights can be used to initialize the model. Starting the training, the data reshaping needs to be the previous shape before entering the data into the model. Test and save models, including prediction models and network weights, are performed every 200 batches of training.

3.4. Achieve voiceprint contrast

The voice_Compare.py program is created so that when the model is loaded. See the input and output names for each layer by using Netron. Then write two functions, the classification is to load the data and perform the prediction function. The data processed by prediction are the eigenvalues of speech. Using the above two functions, speech recognition can be carried out: input two sounds, obtain their features by using the prediction function, calculate their diagonal cosine value by using the feature data, and then obtain their similarity. Based on the above comparison, voice_recognition.py is created to realize voiceprint recognition. The main recognition function is the recognition() function, which is to compare the input speech with the speech in the speech library one by one. With the above voiceprint recognition function, you can complete the voiceprint recognition by recording. The voice in the voice database folder is voice_db. After the user enters the voice_db folder, the program records the voice automatically and uses the recorded audio for voice print recognition to match the voice in the voice database and obtain the user's information.

3.5. Speaker spectrum

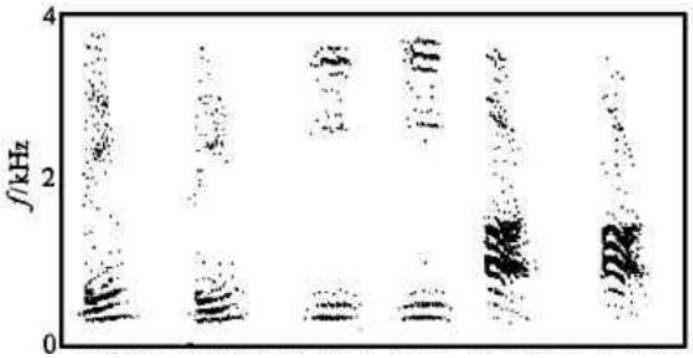

Figure 5. Spectrogram of the same person

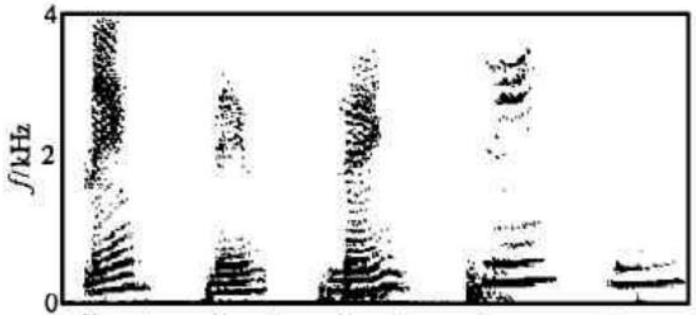

Figure 6. Spectrogram of "0" voiced by five different speakers

Through experimental comparison, the figure 5 shows the spectrogram of the same speaker voicing "0", "1" and "2" twice. The figure 6 shows the spectrogram of the sound of "0" for different speakers. Obviously, different speakers' spectrograms have different voice prints, and the personality characteristics of voice prints have little to do with the content of speech.

4. Implementation of voiceprint recognition check-in system based on deep learning

4.1. System requirement analysis

The purpose of this paper is to design and implement a voice print recognition check-in system with both speech detection function and speaker identity verification function for the application of voice print recognition system in practical application scenarios. The primary task of realizing the voice print recognition check-in system is to analyze the system requirements, which is also the basis and important basis of the system design.

Voiceprint recognition sign in system is proposed in this paper in the service of real application situations of access control of unknown users purposes, in order to ensure the system in the application scenario can be a normal execution system of all functions, and as much as possible to achieve efficient response to the user's request, the voiceprint recognition sign in system in terms of performance set the relevant requirements of real time response. The system should give timely and effective feedback when implementing the user's functional request. The performance requirements of this system as follows, the user, after sending the access, please input your own voice, pretreatment, multi-resolution system to accomplish the front-end speech spectrum feature extraction function such as the execution time of the operation should not be more than 1 second, the system to complete the input speech detection and the return of the speaker function such as authentication response process of recognition results should not be more than 2 seconds.

4.2. System workflow

The system is divided into three subsystems according to function: front-end user interaction subsystem, preprocessing and feature extraction subsystem and model subsystem. Each subsystem contains the required functional components, and the data transmission and function connection between subsystems are carried out through the hierarchical relationship. The hierarchical structure ensures the characteristics of high cohesion and low coupling of the whole system. The front-end user interaction subsystem mainly provides the functions of user registration and login, speech collection, and recognition result display. The preprocessing and feature extraction subsystem mainly provides the functions of speech preprocessing, length normalization, and spectrogram feature extraction. The recognition model subsystem includes the functions of deep speech representation extraction and speaker identity verification. The system structure diagram is shown in Figure 7.

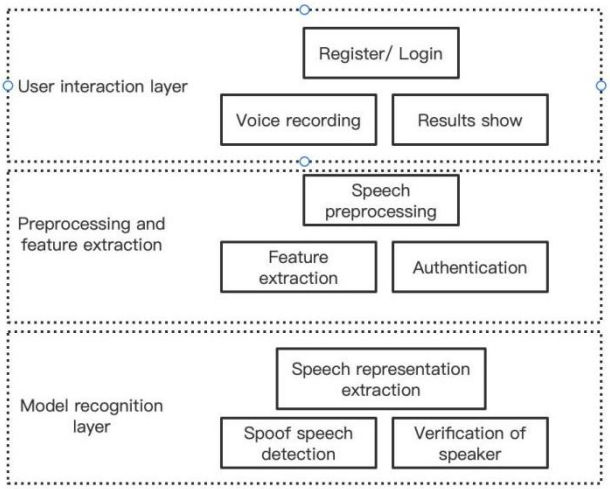

Figure 7. System structure diagram

4.2.1 User interaction layer

This layer provides a service for users to interact with each other by voiceprint recognition check-in system. When the user requests to add members, the system requires the user to input the user name as the unique identity of the user, and collects the user's voice input to the subsequent module for feature extraction and speaker modeling. When a user sends a voice print check-in request, the system collects a speech for identity determination. The specific process is shown in the figure 8.

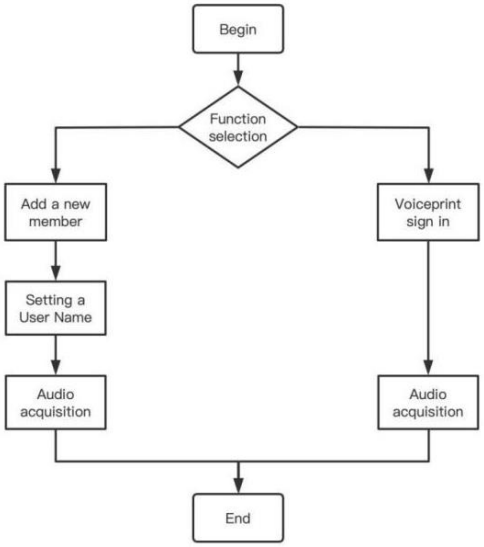

Figure 8. Flowchart of user interaction layer

4.2.2 Preprocessing and feature extraction layer

The speech data collected by the system is sent to the preprocessing and feature extraction layer, where the main tasks are: preweighting, normalization, frame and window, fast Fourier transform, and multi-resolution spectrogram feature generation. After the above operations, the obtained multi-resolution spectrogram is input to the recognition layer for judgment. The specific process is shown in Figure 9.

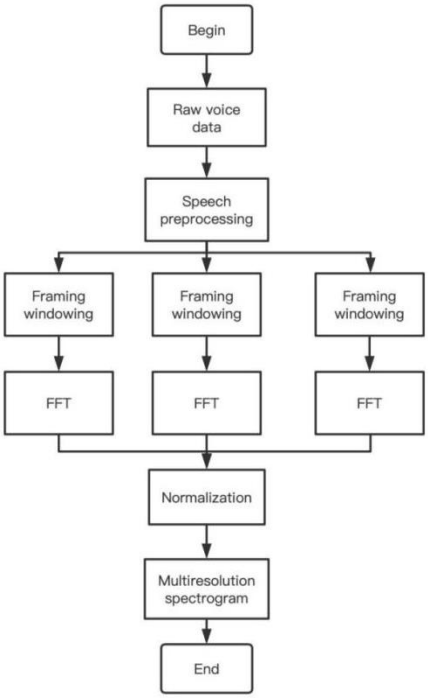

Figure 9. Flowchart of preprocessing and feature extraction layer

4.2.3 Model identification layer

The multi-resolution spectrogram features obtained after feature extraction are fed into the model recognition layer, where in deep residual network is used to extract high-level representation of speech, and then speaker identity verification is performed on speech representation. After the above processing, the identification result is returned to the front end to show to the user. The process of this layer is shown in Figure 10.

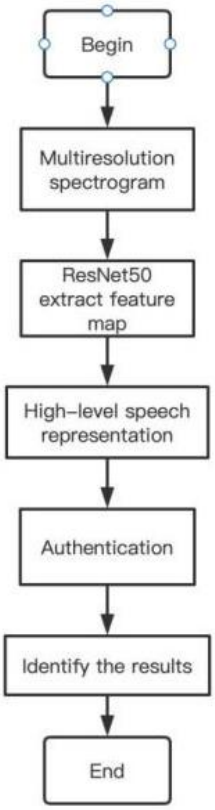

Figure 10. Flowchart of model identification layer

4.3. Test Results

Figure 11 and Figure 12 show the experimental results of the system on the real speech data training set and test set, respectively.

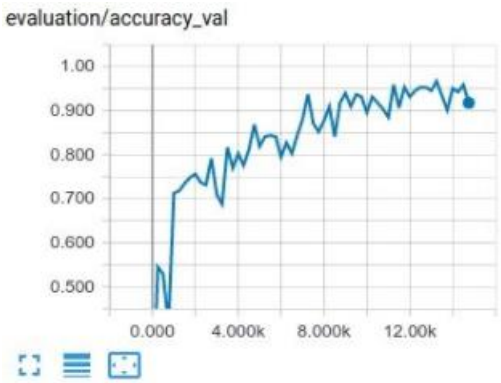

Figure 11. Accuracy diagram of real speech training set 1

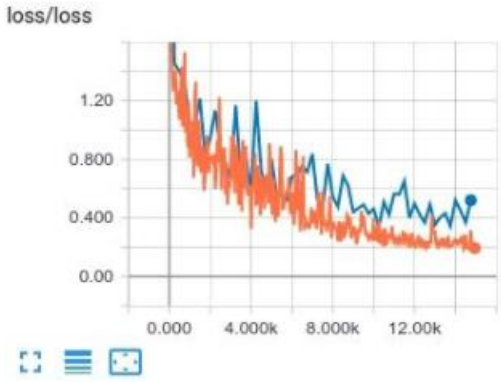

Figure 12. Loss value diagram of real speech training set 1

It can be seen that the best recognition accuracy of Resnet model on the real speech test set is 96.2%, and the fluctuation after convergence basically remains around 95%.

5. Summary of this study

In this paper, the research background and significance of voiceprint recognition technology are briefly introduced, and then the research progress, research status and key technologies in the field of voiceprint recognition and check-in system are analyzed. Based on the actual application scenario, this paper designs and extracts multiple spectrograms of different resolutions as input, so that the system can obtain richer and global information from the original speech. On the basis of multi-channel input, a deep learning-based voice print recognition check-in algorithm is proposed. Then, on the basis of this algorithm, a deep learning-based voice print recognition check-in system is designed and implemented in this paper, and the process of system module design and implementation is introduced.

In summary, the work of this paper can be divided into the following parts:

(1) A multi-resolution spectrogram extraction module is designed and implemented. Aiming at the problem that it is difficult to detect multiple types of speech by using a single feature in traditional voice print detection research, this paper proposes a multi-resolution spectrogram feature extraction method. By extracting multiple spectrograms with different resolution in time domain and frequency domain from the original speech, it is beneficial to extract more sufficient and global information from the speech.

(2) Proposed and built a voiceprint recognition check-in model based on deep residual network. The model uses multi-resolution spectrogram as input, and builds deep residual network, which is conducive to extracting higher level representation from speech, and has better recognition performance for unknown speech detection tasks.

(3) A voiceprint recognition check-in system based on deep learning model is designed and implemented. System back-end TensorFlow implementation using deep learning framework, system to collect the voice of the speaker under test. First of all, voice and data preprocessing spectra and extract the window set up multiple different language features, and then sent to the trained voice print recognition model, treatment of voice judgment results, final results will show on the front page.

References

[1] Rabiner, Lawrence, and Biinghwang Juang. "An introduction to hidden Markov models." IEEE ASSP Magazine 3.1 (1986): 4-16. .

[2] Campbell W M. A SVM/HMM system for speaker recognition [C] / 2003 IEEE, International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). IEEE, 2003, 2: II-209.

[3] Matsui T, Furui S. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's [J]. IEEE Transactions on speech and audio processing, 1994, 2(3): 456-459.

[4] Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2017). Generalized end-to-end loss for speaker verification. ArXiv preprint arXiv: 1710.10467.

[5] Kumar, R., Yeruva, V., & Ganapathy, S. (2018). On Convolutional LSTM Modeling for Joint Wake-Word Detection and Text Dependent Speaker Verification. Proc. Interspeech 2018, 1121-1125.

[6] Lei Y, Burget L, Scheffer N. A noise robust i-vector extractor using vector taylor series for speaker recognition [C] /2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 6788-6791.

[7] Tanprasert C, Wutiwiwatchai C, Sae-Tang S. Text-dependent speaker identification using neural network on distinctive Thai tone marks. InIJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339) 1999 Jul 10 (Vol.5, pp.2950-2953). IEEE.