Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
May-June 2022 Page No. 471-477
Article History: Received: 24 February 2022, Revised: 05 April 2022, Accepted: 08 May 2022, Publication: 30 June 2022

Speaker Recognition System based on CNN for Intelligent Attendance

Shuxi Chen, Yiyang Sun*, Jianlin Qiu, Haifei Zhang, Qinqin Liu
School of Computer and Information Engineering, Nantong Institute of Technology, Nantong, 216001, China
*Corresponding Author.

Abstract:

With the rapid development of information technology, student attendance has changed from using paper to using machine, such as taking photos, scanning QR codes, positioning, etc. These attendance work needs to turn on camera to take photos, which is slightly inefficient, or turn on the positioning service. However, many people think that turning on the positioning service will infringe on personal privacy. Therefore, we need to consider a more efficient attendance method that does not infringe on personal privacy. Voice, a signal which can be quickly obtained and contain a variety of information, can be used for class students' attendance. This paper studies, designs and implements a voiceprint recognition system based on convolutional neural networks (CNN), which can effectively recognize specific speakers. The speaker recognition system based on CNN constructed in this paper. The system mainly includes three steps: voiceprint registration stage, data training stage and speaker online recognition stage.

Keywords: *Attendance, Voiceprint recognition, CNN.*

I. INTRODUCTION

Speaker recognition integrates many research fields, such as signal and information processing, pattern recognition and psychology. It has a wide application prospect. It is changing or about to change our daily life. It also plays an important role in class students' attendance. Attendance system can not only make teachers know the attendance of students in time, but also supervise and urge students. It plays a very important role in class attendance assessment. At present, the commonly used attendance systems are mainly based on biometric technology, such as user fingerprint recognition[1], palm shape recognition[2], face recognition[3], etc. However, the above attendance methods have high requirements for external factors such as attendance object and scene light. Speaker recognition is an important biometric recognition technology to recognize the identity of the speaker according to the characteristics of the speaker's voice[4]. With its advantages of convenient collection and easy adaptation by users, it has been applied to the intelligent attendance system in recent years[5].

Speaker recognition has attracted worldwide attention for its unique performance, which is mainly reflected in the following aspects:

Convenience: when collecting data, users don't have to put their fingers on the sensor or put their eyes close to the camera. They just need to say one or two words, which is easy to accept.

Article History: Received: 24 February 2022, Revised: 05 April 2022, Accepted: 08 May 2022, Publication: 30 June 2022

Economy: fingerprint or iris recognition needs to rely on specific equipment, such as fingerprint scanner or iris scanner, which is generally expensive and difficult to maintain. Speaker recognition does not need expensive hardware equipment, but only simple sound input devices, such as microphone, telephone, mobile phone and so on.

Support remote applications: through phone and mobile device authentication, sound is probably the only available biometric.

Protect personal privacy: Face recognition or location will violate personal privacy and make people feel uncomfortable to some extent. Recognition through voice can avoid this problem.

II. RESEARCH STATUS AT HOME AND ABROAD

At present, the academic research on speaker recognition is mature, and there are many application scenarios. There are mainly voiceprint recognition for home scene and anti- deception. This paper mainly uses voiceprint recognition to check the attendance of class students.

2.1 Abroad

In 1945, L.G. KESTA, who worked in Bell Labs, made a very detailed study on the details of speech spectrum. He found that the characteristics of speech signal and the characteristics of speaker's speech signal showed a high degree of matching on the spectrum. This research result made the speech spectrum become the reference basis of voiceprint recognition and laid a foundation for future research. The voiceprint recognition technology based on GMM (Gaussian mixture model) proposed by Reynolds' team at the end of the 20th century has made great progress in the research work in this field[6]. They used TIMIT continuous speech corpus to experiment, and the model achieved good results with an accuracy of 95%. In terms of market application, the smart card developed by American communication operator at & T applies voiceprint recognition technology and has been put into the use of ATM in life. In recent years, due to the breakthrough of the bottleneck in the field of machine learning, many IT companies around the world have increased their investment in this field. In 2014, Google first used neural network to build and train the model of voiceprint recognition, and released the results two years later[7]. By further improving the network, Google researchers have improved the recognition rate and changed the extracted speech features. This result greatly encourages the confidence of relevant researchers, and makes voiceprint recognition based on deep learning gradually return to people's vision. At present, in foreign countries, voiceprint recognition technology has been widely used in military, national defense, government, finance and other fields.

2.2 Domestic

With the popularization and development of domestic Internet technology, voiceprint recognition technology is developing better and better. For example, many financial institutions in China have invested in the field of voiceprint recognition and developed a biometric authentication system based on voiceprint recognition to ensure the security of financial transactions[8]. Some domestic platforms also use voiceprint recognition as a landing method, which has achieved good results. Some companies also cooperate with

China's public security departments to arrest fugitives using voiceprint recognition technology, recognize a large number of call samples by extracting call voice, and finally lock the criminal suspect[9].

III. DESIGN OF SPEAKER RECOGNITION SYSTEM

The main components of voiceprint recognition system are shown in figure 1. The system includes: speech signal input, preprocessing, feature extraction and pattern matching.

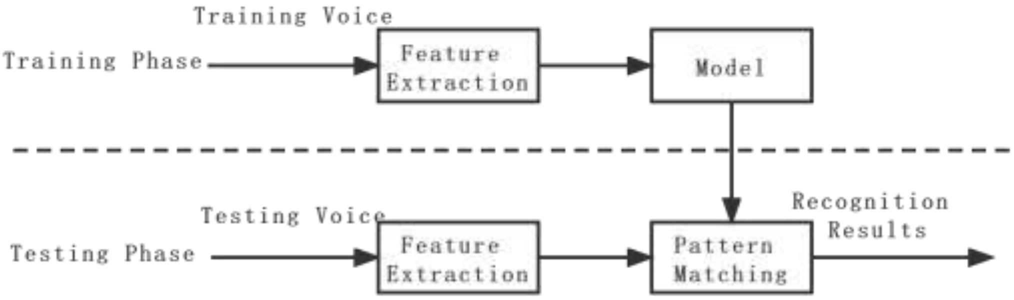

Fig1: Voiceprint recognition process

In the voiceprint recognition system, the tester first uses the microphone as a recording tool for audio acquisition, and then system decodes and converts the audio data to prepare for audio preprocessing. After preprocessing, these data are imported into the training model for training. Adjust the parameters such as batch size and iterations until the network model training is completed and saved. At this time, the tester can input the test speech for matching and complete the process of voiceprint recognition.

Ⅳ. ALGORITHM INTRODUCTION

In recent years, deep learning technology has developed rapidly. With its excellent ability of nonlinear fitting and function approximation, it has been widely used in various fields of intelligence. In this paper, the back-end uses convolutional neural network (CNN) to deeply learn the speech segments[10], and then uses softmax function to classify them. The front-end uses C/S architecture.

4.1 CNN

CNN is a special deep neural network, but its obvious difference from DNN is that its neurons are not fully connected, but connected through weight sharing, that is, the weight of local areas is the same. This structure reduces the network parameters, reduces the complexity of the model, and can effectively prevent over fitting. The weight sharing structure of CNN is more similar to the neural network of human brain. It has excellent performance in various fields of pattern recognition[11].

Convolution neural network is usually composed of convolution layer, pooling layer and full connection layer. Convolution layer and pooling layer are generally used alternately as hidden layers.

Convolution operation is to extract the features of the graph and generate the feature graph. Pooling is to reduce the sampling of the feature graph obtained by convolution. While reducing the dimension, it can strengthen the generalization ability of the network and avoid over fitting of the network. CNN network structure is shown in Figure 2.

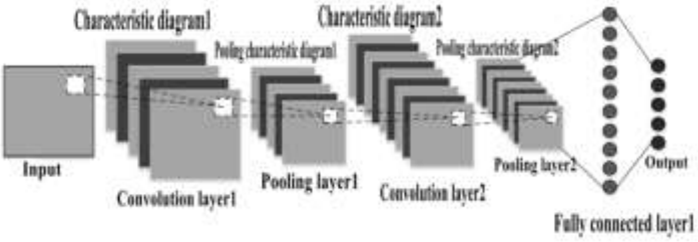

Fig 2: structure of CNN

4.2 ResNet

For deep neural networks, the deeper the network, the larger the hypothesis space, and the better the experimental effect in theory. However, in fact, it is not the case. The reason is that deeper networks are often more difficult to train and are prone to gradient disappearance and explosion, resulting in the degradation problem of the network, while ResNet can well overcome this problem and speed up the convergence of the network[12].

ResNet is formed by stacking multiple residual blocks (as shown in figure 3). The residual blocks can generally be expressed as:

$$y_l = h(x_l) + F(x_l, W_l) \tag{1}$$

$$x_{l+1} = f(y_l) \tag{2}$$

Where x_l and x_{l+1} respectively represent the input and output of the l layer. F is the residual function. ResNet is generally composed of 2 or 3 convolution operations. $h(\cdot)$ is an identity mapping function. If the dimension of the characteristic graph of x_l and x_{l+1} is the same, $h(x_l) = x_l$. If it is different, it usually needs to use 1×1 to adjust the number of characteristic graphs. $f(\cdot)$ is an activation function, generally a rectified linear unit (ReLU).

In Figure 3, BN represents batch normalization, which is a parameter optimization algorithm in deep neural network. The purpose is to forcibly normalize the activation input of hidden layer nodes to a standard normal distribution with mean value of 0 and variance of 1, so as to avoid gradient disappearance and speed up network training.

Article History: Received: 24 February 2022, Revised: 05 April 2022, Accepted: 08 May 2022, Publication: 30 June 2022

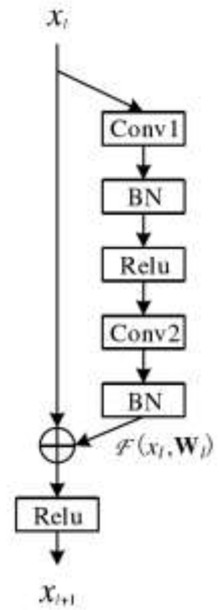

Fig3: Residual Blocks

Ⅴ. SPEAKER RECOGNITION SYSTEM BASED ON DEEP LEARNING

The main function of the speaker recognition system based on deep learning designed in this paper is to preprocess the recorded corpus, input the processed audio into the model for training. After model training, students can click "check in" on the home page of the system to collect audio again for testing. After the system completes voiceprint recognition, the results will be returned to the user through the page: “Helen checked!”.

5.1 Overall System Framework

Data layer: this layer is the bottom part of the overall system framework, which is used for data storage. It uses MySQL database to operate and store voice files, create data tables and interconnect with other layers.

Logic layer: this layer is the core part of the whole system framework. This layer calls the microphone to collect the speaker's voice, and then performs transcoding, pre emphasis and other preprocessing operations on the voice, and then input the model for training after spectrogram calculation.

Display layer: this layer realizes the interaction between the user and the system, mainly UI display, and presents the whole process of voiceprint recognition to the user. Reduce the learning cost of users and obtain better operation experience through the simple and understandable interface on the mobile phone side.

5.2 Data Sets

Considering the influence of the original speech signal quality on the recognition performance of the system, the preliminary work of this paper includes the establishment of Nantong Institute of Technology Speech Processing Researches-Speaker Detection (NIT-SD) Corpus. Dual channel recording is adopted. The sampling frequency is 48 kHz. The corpus contains 10000 phrases and 200 Q&As' answers.

5.3 Model Training

Model training module is the core part of the system. It mainly includes audio preprocessing and model training. In this link, we should mainly consider several aspects: (1) the clarity of voice files; (2) Speech preprocessing quality; (3) Speed of model training; (4) The size of the model.

The first step of speech processing is to extract the effective information contained in speech, so as to carry out subsequent processing and analysis of speech signal and achieve the purpose of recognition. Firstly, the speech signal is preprocessed, and then the characteristic parameters of the speech signal are extracted. This paper mainly extracts the spectrogram and MFCC of the speech signal.

The system uses the residual network for model training, and saves the model after the training, so as to facilitate the next voiceprint recognition work.

5.4 Voice Acquisition

Students click the "check in" button on the home page of the system to record the audio, extract the speech feature parameters after system preprocessing, and input them into the training set for speaker recognition.

5.5 Speaker Recognition

The voiceprint recognition module also calls the microphone to record the user. The voice sample is also preprocessed as a test set, matched after decoding, conversion, pre emphasis and other processing, and the final result is fed back to the tester. Voiceprint recognition module needs to meet: (1) running speed; (2) Accuracy.

Ⅵ. CONCLUSION

The speaker recognition system based on deep learning constructed in this paper can carry out class student attendance well. The following work will continue to expand the data set, and the voiceprint registration function will be introduced into the interface.

ACKNOWLEDGMENTS

This work is sponsored by: the science and technology planning project of Nantong City under Grant No. JCZ21058. The Ministry of education cooperates with universities to educate people in 2021 Project under Grant No.202102594035 and No.202102594017.

REFERENCE

[1]Chen B. Research on speaker recognition feature extraction algorithm and implementation of voiceprint attendance system. Kunming University of Technology, 2014.

[2] Non-contact Palmprint Attendance System on PC Platform. Journal of Multimedia Information System, 2018, 5(3).

[3] Mayur Surve, Priya Joshi, Sujata Jamadar, Minakshi Vharkate. Automatic Attendance System Using Face Recognition Technique. International Journal of Recent Technology and Engineering (IJRTE), 2020, 9(1).

[4] Reynolds D A. An overview of automatic speaker recognition technology. IEEE International Conference on Acoustics, 2011.

[5] Jeetvan Shah, Vikas Salunkhe, Jitendra Saturwar, Omkar Parab. Voice Input based Attendance System. International Journal of Recent Technology and Engineering (IJRTE), 2020, 9(1).

[6] Reynolds D A. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, 1995, 17(1-2): 91-108.

[7] Variani E, Lei X, McDermott E, et al. Deep neural networks for small footprint text-dependent speaker verification. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, 2014: 293-298.

[8] L Liu, X Wu, F Zheng. Application of voiceprint recognition in financial field. Security technology and application in China, 2020(05):21-26.

[9] F Xiang. Feasibility analysis of voiceprint recognition and speech recognition technology in the field of public security. Legality Vision, 2021(36):107-109.

[10] Samia Abd El-Moneim, Ahmed Sedik, et al. Text-dependent and text-independent speaker recognition of reverberant speech based on CNN. International Journal of Speech Technology, 2021, 24(4).

[11] M Xiong. Research on speaker recognition method based on deep learning. Nanchang University, 2021.

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016: 770-778.