Voiceprint recognition based on BP Neural Network and CNN

Dong Zhipeng¹, Wei Jingcheng¹, Xu Yumin¹, Meng Qingmin¹,², Wang Xiaoming¹,²

¹Key Laboratory of Broadband Wireless Communication and Sensor Network Technology (Nanjing University of Posts and Telecommunications), Ministry of Education, College of Telecommunications & Information Engineering, Nanjing 210003, China
²National Mobile Communications Research Lab, Southeast University, Nanjing 210096, China

e-mail: dongzp.kwan@qq.com, wjcyx2009@qq.com, {mengqm,xmwang}@njupt.edu.cn

Abstract. At present, speech recognition has become a key technology of human-computer interaction, which can be used in semantic recognition and speaker identification and other related applications. This paper focuses on a speaker identification scene and the technical design. Because artificial neural network has the ability to distinguish complex classification boundaries, plenty of work on speech recognition had studied multi-layer perceptual networks to improve the accuracy of classification where back propagation method (BP algorithm) has been used. In the studied scheme, when the test objects (speakers) speak the same isolated word, the identification system can judge who the speaker is by identifying the voice voiceprint of different people. Our design not only considers the recognition based on BP as well as its variant, but also explores the voiceprint recognition based on convolution neural network (CNN). Secondly, the network structure and performance are also compared in detail.

1. Introduction
With the improvement of the ability of computer to process data, people hope to realize the interaction between human and computer efficiently [1], so speech recognition has been developed rapidly in the past 20 years. In the early research of speech recognition, a lot of attention was paid to the recognition of speech content. These methods can not identify the speaker according to the timbre of the speech itself. In some occasions, it is particularly important to identify the speaker, thus the research on voiceprint recognition has also attracted the attention of researchers [2].

Voice voiceprint recognition is a biometric technology, which refers to the use of different voice characteristics of different people to distinguish or identify them [3]. With the development of computer technology, voice voiceprint or voiceprint recognition technology is widely used in artificial intelligence and information security and other fields. The dynamic time warping model matching (DTW) was proposed in the 1970s, and some linear prediction methods were subsequently proposed [4]. In the 1980s, a variety of voiceprint recognition methods were proposed in turn, such as feature extraction method based on Mel frequency cepstrum coefficient (MFCC) [5], Vector quantization (VQ) [6] and Hidden Markov Model (HMM) [7] based on probability Model. Up to now, MFCC is still applied to speech feature extraction. By the 1990s, Gao Si's mixed model-General background Model (GMM-UBM) had become a mainstream method [8]. In recent years, researchers have also proposed improved speech...
recognition algorithms, such as weighted clustering recognition SVM and spectral and DNN voiceprint recognition [2] [9].

In this paper, the task of speaker recognition is regarded as a pattern matching problem of images [2], and the voiceprint recognition based on convolution neural network (CNN) (method 1) is studied in detail. Secondly, the single-layer BP neural network (method 2) is used as a reference to discuss different voiceprint recognition systems, and the problem of directly using the original PCM signal as the input to the BP neural network (method 3) is also considered. Detailed computer experiments have been completed and the results are helpful to compare different voiceprint recognition methods.

2. Data Sets and Features

2.1. Preparation of data sets
Four artificial voice data sets, "dl", "wjc", "hy" and "tzl" are used, which are the initials of the names of the subjects (students). Each type of dataset contains 2000 wav audio files, each of which is an isolated word "circuit" that lasts one second, for a total of 8000 audio files. At the same time, in the training of CNN network, the data set is further split into 75% training set, 12.5% verification set and 12.5% test set. In addition, in order to obtain a more robust parameter set, Gao Si random background noise with variance of 0.01 is added in the speech dataset.

2.2. Feature extraction

2.2.1. Feature extraction of method 2
For method 2 based on BP neural network, the feature engineering based on MFCC is adopted. The human ear perceives speech signals of different frequencies in a variety of ways, and the relationship between the input and output of the perception is nonlinear. In order to model this feature, Mel frequency scale is used. The experimental results show that the auditory system of human ear has high resolution at low frequency and low resolution at high frequency [10]. In the experiment studied here, the sample time interval of each audio signal is set to 1s, the order of Mel filter is set to 24, the length of FFT transform is set to 256, and the sampling frequency is set to 8000Hz. In addition, the sliding stride on the timeline is 10 Ms and the mobile window size is 30 Ms, so each audio file will form (1000 - 30) / 10 + 1 = 98 frames. Secondly, the spectrum on the Mel spectrum is analyzed, that is, a method of doing discrete cosine transform (DCT) is employed to obtain the two-dimensional matrix of 98 x 24. The specific process of the feature project is shown in Fig.1. In order to reduce the influence of edge noise and accelerate the convergence of the supervised learning, when using MFCC extraction features as the input, only 3920 audio files are used, that is, 980 audio files for each category, and at the same time, for simplicity, the labels are set to "1", "2", "3" and "4", respectively.

![Fig.1 MFCC feature extraction progress](image)

2.2.2. Feature extraction of method 3
In addition to extracting the feature vector with the help of MFCC, the method of whether the audio signal sampled by 8000Hz can be directly input into the BP neural network is also discussed, that is, each audio is trained with 8000 features. The feasibility of method 3 or a variant of method 2 is evaluated by the following experiments. If the original PCM signal is directly used as the input of the BP network, too large feature number will result in taking too much time to learning. Our improvement is to...
downsample 8000 features to compress them into 1000 features. Finally, the experimental results show that the effect is good.

2.2.3. Feature extraction of method 1

When using CNN neural network for training, the method of obtaining the features of the speech spectrum is chosen. The spectrum is a visual representation of frequency change in a speech signal. The common format of the spectrum is an image with three dimensions, in which the horizontal axis and the vertical axis represent time and frequency, respectively, while the third dimension represents the amplitude of a particular frequency at a particular time.

In the experiment, the spectrogram function is used in the Matlab routine, which uses the short-time Fourier transform (STFT). Therefore, the spectrum of the input audio signal \( x(t) \) can be calculated by the square amplitude of its own STFT, and its mathematical expression is as follows:

\[
\text{Spectrogram}(t, \omega) = |\text{STFT}(x(t))|^2
\]

\[
\text{STFT}(x(t)) = \int_{-\infty}^{\infty} x(t) \omega(t) e^{-j\omega t} dt
\]

Here, \( \omega(t) \) is a window function, usually a hamming window or Gaussian window, and \( \tau \) is a time constant.

Fig.2 shows a one-dimensional time domain signal (above) and a two-dimensional audio spectrum (below) generated using an acoustic spectrum.

3. BP Neural Network

3.1. Input characteristics of methods 2 and 3

Method 2 and method 3 both have BP neural network, but the input characteristics of the two methods are different. The speech recognition modeling process includes BP neural network construction, BP neural network training and BP neural network classification walk. The algorithm flow is shown in Fig. 3. Let's discuss the construction of the network in some detail. The structure of BP neural network needs to be determined according to the data characteristics of the input and output of the system. For method 2, because the feature signal extracted by MFCC has 24 dimensions and there are 4 types of speech signals to be classified, the structure of BP neural network is set to 24-25-4, that is, there are 24 nodes in the input layer and 25 nodes in the hidden layer. The output layer has four nodes. For method 3, because of the compression of the features, the original 8000 is changed to 1000, so its structure becomes 1000-1001-4, that is, there are 1000 nodes in the input layer, 1001 nodes in the hidden layer and 4 nodes in the output layer.
Consider two different types of data here. For method 1 and method 2, 3920 audios are prepared in advance, and 2940 audios are randomly selected as the training data to train network and the rest are selected as the test data. For method 3, 2000 audio files are prepared for each category, a total of 8000 audio files, from which 6000 were randomly selected as the training data to train network and 2000 as the test data.

3.2. BP neural network classification

The classification process is defined as: the trained BP neural network is used to classify the categories to which the test data belong, and the correct rate of each class is calculated according to the classification results.

When training the data in this network model, the default learning rate is set to 0.1 and the unipolar Sigmoid function is used as the transfer function of the output layer. The formula is as follows:

\[
\begin{align*}
 f(net) &= \frac{1}{1+e^{-net}}, 0 < f(net) < 1 \\
(f')(net) &= \alpha f(net)[1 - f(net)]
\end{align*}
\]

(3)

Unipolar Sigmoid function, usually \( \alpha = 1 \)

Here, BP neural network uses gradient correction method as the learning updating algorithm of weights and thresholds. Specifically, it is to modify and update the weights and thresholds from the negative gradient direction of the network classification error. Because the accumulation of previous experience is not taken into account in the method, so the learning and convergence is slow. In order to solve this problem, additional momentum method is adopted, that is, the previous weight or threshold is attached to a momentum, the formula is as follows:

\[
\begin{align*}
\omega(k+1) &= \omega(k) + \eta \Delta \omega(k) + \alpha [\omega(k) - \omega(k-1)] \\
T(k+1) &= T(k) + \eta \Delta T(k) + \alpha [T(k) - T(k-1)]
\end{align*}
\]

(4)

(5)

Where \( \omega(k-1), \omega(k) \) and \( \omega(k+1) \) are the weights of \( k-1 \), \( k \) and \( k+1 \), respectively. \( T(k-1), T(k) \) and \( T(k+1) \) are the thresholds of \( k-1 \), \( k \) and \( k+1 \), respectively, \( \eta \) is the default learning rate, and \( \alpha \) is the momentum learning rate. In order to speed up the convergence of learning, the momentum learning rate is set to 0.01.

For the entire network, define training sample set: \( \{X_1, D_1\}, \{X_2, D_2\}, \ldots, \{X_M, D_M\} \), the global error of the network is represented as:

\[
E = \frac{1}{2 \sum_{t=1}^{M} \sum_{k=1}^{l}(d_k(t) - o_k(t))^2}
\]

(6)

where \( X_M \) denotes the M sample, \( d_k(t) \) denotes the output in \( t \)th iteration and \( o_k \) represents the expected output.
4. CNN

Let’s move on to method 1. The input of convolution neural network (CNN) is a two-dimensional image array. A common method is used: convert one-dimensional audio signal into two-dimensional image array. An image is regarded as a two-dimensional array or matrix in which the pixel value of each point in the original image or the normalized gray value is stored. By using the spectrum of audio, audio with a duration of 1s is converted into a spectrum signal array of 40 x 98, and the spectrum is used as the input feature. For simplicity, a convolution neural network with five convolution layers is designed, including input layer (Input Layer), convolution layer (Conv Layer), pooling layer (Pooling Layer), dropout layer and full connection layer (Full Connected Layer). The function of the input layer is to input the spectrum map; the convolution layer comprises a plurality of filters to obtain different features; the pooling layer divides the feature layer into several regions and reduces the parameters by taking the maximum or average value, the dropout layer aims to temporarily discard the neural network unit from the network according to a certain probability in the training process of the deep learning network, so as to avoid overfit and improve the performance. The full connection layer is that each node is connected to each node in the upper layer, thus synthesizing the output characteristics of the previous layer. The parameters of our convolution layer are shown in Table 1.

| Type   | Number | Filters | Size  | Stride |
|--------|--------|---------|-------|--------|
| Conv1  | 16     | 3x3     | 1     |        |
| Conv2  | 32     | 3x3     | 1     |        |
| Conv3  | 64     | 3x3     | 1     |        |
| Conv4  | 64     | 3x3     | 1     |        |
| Conv5  | 64     | 3x3     | 1     |        |

5. Adjustment of CNN Parameters in Experiments

In the experiment of the CNN, the rule of the special normal distribution for the initialization of the weight is followed, which has a zero mean value and a specified standard deviation.

5.1. Network training batch size

In network training, the size of training batch will have a significant impact on the learning efficiency of the network. A stochastic optimization method is used for adaptive moment estimation of (Adam). Adam is a method based on gradient descent, but the learning step size of each iteration parameter has a certain range, which will not lead to a large learning step because of a large gradient, and the value of the parameter is relatively stable. In the network, the minimum batch size is set to 128.

5.2. Learning rate

In the simulation, by trying a variety of schemes, it is concluded that the combination of the two learning rates is better. That is to say, 25 epochs are trained, in the scheme, and the learning rate of the first 20 epochs is set to 0.0005, while the learning rate of the next 5 epochs reduce tenfold to 0.00005. All in all, this combination can obtain good convergence.

5.3. Cross entropy

In the CNN neural network structure, the softmax classifier with cross entropy loss L is used to estimate each output tag in the last layer. The formula is as follows:

$$L_i = -\log\left(\frac{e^{f_{iL}}}{\sum_j e^{f_{jL}}}\right)$$

(7)
Here, \( f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \), \( f_j(z) \) is a softmax function that represents the value of the \( j \)th neuron.

5.4. Regularization

Dropout layer can reduce the overfit in neural network, and the dropout method is applied to regularization technology. In the experiment, the probability of dropout is set to 0.2, that is, each neuron has a probability of 0.2 being randomly discarded.

6. Test Results and Comparision

According to the different feature extraction methods, three kinds of methods are considered. Method 2 is a BP neural network using Mel frequency cepstrum coefficient (MFCC) to extract features; method 3 is a BP neural network that directly uses audio signals as the input; method 1 is a CNN neural network using audio acoustic spectrum features. Their training accuracy is shown in Table 2.

| Method | dl(1) | hy(2) | tzl(3) | wjc(4) |
|--------|-------|-------|--------|--------|
| 2      | 0.740 | 0.556 | 0.623  | 0.716  |
| 3      | 0.996 | 0.992 | 0.996  | 0.970  |
| 1      | 1.00  | 1.00  | 1.00   | 1.00   |

6.1. Accuracy comparision

The MFCC feature extraction dimension of method 2 is 24 dimensions, and the input dimension of method 3 is 1000 dimensions. In training, the accuracy of method 3 is significantly higher than that of method 2. In Method 3, the accuracy of training is close to 1. Referred to Table 2, it is observed that the more feature dimensions, the better the performance of BP neural network training. Secondly, the accuracy of method 1 is "1", which indicates that the performance of CNN neural network is better than that of BP neural network.

6.2. Multi-classification confusion matrix

The confusion matrixes of multiple classifications are drew in Fig. 4(a) and 4(b). Fig. 4(a) and 4(b) represent the confusion matrix corresponding to methods 2 and 3 respectively. Each row of the matrix represents the true attribution category of the sample, and the sum of the data of each row represents the number of sample instances of that category, and each column represents the prediction category, the sum of each column represents the number of samples predicted for that category, and the values in each
column indicate that the sample is predicted to be the number of that class, therefore, the values on the diagonal of the confusion matrix represent the correct number of samples for each class.

7. Conclusion
The speech data designed in this paper contains a single isolated word and only considers four categories of voiceprint recognition. Two different neural network models are used: BP neural network and CNN neural network. The two neural networks use different feature extraction methods. BP neural network (method 2) uses Mel cepstrum coefficient (MFCC), and CNN neural network (Method 1) uses spectrogram method. Method 3 is a BP neural network method which uses original audio signals as the input. Therefore, three methods to identify voiceprint are compared in detail. The experimental results show that the CNN neural network model is superior to the BP neural network model, but the BP neural network of method 3 can improve the performance by increasing the feature dimension, but at the expense of time efficiency. In the future, voiceprint recognition with more categories and based on BP neural network will be discussed in order to further evaluate the method of using the original PCM signal as the input.

Acknowledgments
This work was partially supported by the National Natural Science Foundation of China (No. 61801240), the Natural Science Foundation of Jiangsu Province (NoBK20180753) and the open research fund of National Mobile Communications Research Laboratory Southeast University (No. 2019D16). Corresponding author: Meng Qingmin.

References
[1] Z Huang, SM Siniscalchi, CH. Lee, "A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition," Neurocomputing, pp. 448-459, 2016.
[2] Yuan Xue, Luping Wang, Linxuan Li, Zhiqi Liu, Jialin Liu, “Matlab-Based Intelligent Voiceprint Recognition System,” 2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), pp. 303-306, 2016
[3] Penghua Li, Minglong Chen, Fangchao Hu, Yang Xu, “A spectrogram-based voiceprint recognition using deep neural network,” The 27th Chinese Control and Decision Conference (2015 CCDC), pp. 2923 – 2927, 2015
[4] J.H. Liu, Y.X. Zou, Y.C. Huang, "An effective voiceprint based identity authentication system for Mandarin smartphone users", International Conference on Pattern Recognition, pp. 1077-1082, 2016.
[5] K. I. Molla and K. Hirose, "On the effectiveness of MFCCs and their statistical distribution properties in speaker identification," 2004 IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2004. (VCIMS), Boston, MA, USA, 2004, pp. 136-141.
[6] H. Jialong, L. Liu, G. Palm, "A discriminative training algorithm for VQ-based speaker identification," Speech & Audio Processing IEEE Transactions on, vol. 7, no. 3, pp. 353-356, 1999.
[7] M. Inman, D. Danforth, S Hangai et al., "Speaker identification using hidden Markov models," IEEE International Conference on Signal Processing, vol. 1, pp. 609-612, 1998.
[8] A. Douglas, T.F. Quatieri, R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, pp. 19-41, 2000.
[9] Yang Wu ; Lihong Xu ; Yandong Chen ; Xueyang Zhang, “Research on voiceprint recognition based on weighted clustering recognition SVM algorithm,” 2017 Chinese Automation Congress (CAC), pp. 1144 – 1148, 2017
[10] Wu Yaozhen, "Studies on physiological perception feature extraction methods in underwater target-radiated noise," China Xi’an: Northwestern Polytechnical University, 2014.