Research on Speech Recognition Method in Multi Layer Perceptual Network Environment

Kai Zhao[1], Dan Wang[2*]
[1]Image and Network Investigation Department, Railway Police College, Zhengzhou 450000, China;
zhaokai966@163.com
[2]School of Intelligent Engineering, Zhengzhou University of Aeronautics, Zhengzhou 450000, China;
wangdancn2021@163.com

Received: February 6, 2021. Revised: July 25, 2021. Accepted: August 16, 2021. Published: August 24, 2021.

Abstract—Aiming at the problem of low recognition rate in speech recognition methods, a speech recognition method in multi-layer perceptual network environment is proposed. In the multi-layer perceptual network environment, the speech signal is processed in the filter by using the transfer function of the filter. According to the framing process, the speech signal is windowed and framing processed to remove the silence segment of the speech signal. At the same time, the average energy of the speech signal is calculated and the zero crossing rate is calculated to extract the characteristics of the speech signal. By analyzing the principle of speech signal recognition, the process of speech recognition is designed, and the speech recognition in multi-layer perceptual network environment is realized. The experimental results show that the speech recognition method designed in this paper has good speech recognition performance.

Keywords—Multi-layer perception; Network environment; Speech recognition; Sample collection; Speech signal features.

I. Introduction

As the most convenient way of human-computer interaction at present, speech has the characteristics of high efficiency, direct and natural, and it is also one of the most convenient and quick ways of communication between human beings [1]. At present, the biggest significance of speech interaction in multi-layer sensing network environment is that it can completely liberate people's hands and reduce the learning cost [2, 3]. As an important branch of pattern recognition, the main purpose of speech recognition method is to make the machine understand the meaning of the content said by human, so as to realize the natural communication between human and machine, so as to get rid of the limitations of the existing interaction in the form of text input [4, 5]. There are four main performance indicators of a speech recognition system: Vocabulary range: This refers to the range of words or phrases that the machine can recognize. If no restrictions are made, the vocabulary range can be considered unlimited. Speaker Restriction: Is it possible to recognize only the speech of a given speaker or all the speech of any speaker? Training requirements: whether to train before use, that is, whether to let the machine first "listen" to a given speech, and how many times to train. Correct identification rate: The average percentage of correct identification, which is related to the previous three indicators. However, in practical application, the voice is always affected by the environment noise interference or the transmission medium, cause the sound quality is damaged, affect the normal language information implied in transmission, so how to deal with the multilayer perception of voice network environment, reduce the influence of noise and interference, and to improve the robustness of speech recognition method is crucial.

A lot of research has been done on this, Jiang *et al.* [6] proposed a speech emotion recognition method based on convolutional neural network feature representation, the optimal solution of LBG algorithm in Discrete Hidden Markov speech recognition system is to break the dependence on the initial codebook, and obtain the optimal codebook by vectorizing speech feature parameters through artificial bee colony algorithm. On this basis, an isolated word speech recognition method is proposed. ABC is improved to become DHMM. To be specific, on the basis of extracting the characteristic parameters of speech signal, each food source is used to represent a codebook by ABC algorithm, and artificial bee colony evolution is used to iterate to make the initial codebook become the optimal codebook, and then the code vector symbols of the optimal codebook are recognized and trained in DHMM model. Experiments show that this method is effective DHMM speech recognition method improved by ABC, which can recognize speech effectively with strong robustness; Gu [7] thinks that under the condition of traditional speech recognition, the effect is better in a quiet environment, but in the real environment, it is often affected by noise, which makes the measured data not very reliable and the accuracy of recognition is low. Therefore, a new speech recognition method, namely multi-core learning combination algorithm, is proposed, which effectively combines multi-core learning and projection algorithm. On the basis of different bandwidth, the multi band probability model is obtained, which can suppress the noise and improve the speech recognition ability, so that the speech can be effectively recognized in the noise; Song *et al.* [8] propose a new English speech recognition method based on multi feature deep learning. In the proposed method, the gaussian mixture model and weighted k-means clustering are the technologies have been applied in the spectral domain in order to reduce the dimension of feature space and clustering of centroid vector and covariance matrix elements is considered to be one of the auxiliary characteristic vector of each frame attributes, in order to evaluate the effectiveness of this method, the new feature vectors of main phoneme recognition in TIMIT database are tested. The results show that the recognition rate of different phoneme sets is significantly improved by using the proposed two-level feature vectors compared with MFCC features. Compared with MFCC features, the recognition rate of voiced stops is increased by 5.9% by using wkm clustering, and by using GMM clustering, the average recognition rate is improved Compared with MFCC features, the biggest improvement of wkm clustering is about 7.4%. Compared with the methods in reference, the designed speech recognition is based on the speech machine, the processing of natural sounds and the sound conduction. In the past many years, the problems of speech recognition mainly focus on the transformation of speech sequences into more speech signals after recognition, among which the key and difficulty is whether the speech signals can be changed [9]. And the speech waveform changes in many ways. For example, in acoustic variables, the same phoneme may be pronounced differently in different texts, and the speaking style of the same person may be divided into many kinds, including normal voice, Shouting or whispering. Phoneme variables in the same text are the differences caused by the different accents of people from different regions [10]. However, the quality of speech signal has been changed due to noise and channel distortion. So, this problem exists, especially in the absence or distortion of noise, which not only affects the acoustic model of the signal, but also changes the source information of the speech.

II. Design of Speech Recognition Method in Multilayer Perceptual Network Environment

A. Preprocessing Speech Signal in Multi-layer Perceptual Network Environment

The feature extraction process of speech signal recognition is shown in Fig. 1.

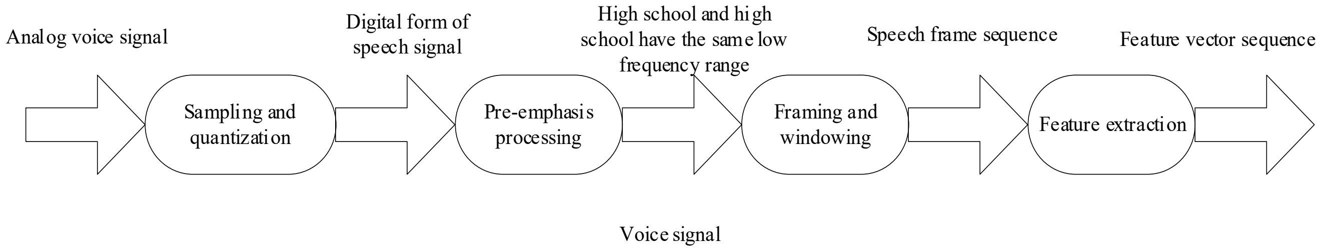

Fig. 1 Feature extraction process of speech signal recognition

In speech signal pre emphasis processing, hardware and software methods can be used [11]. Generally, speech signal is processed in a filter with (1-Az^{-1}) feature, and the transfer function of the filter is expressed as:

$$H(z)=1-aZ^{-1} \tag{1}$$

$$\hat{S}(n)=S(n)-aS(n-1) \tag{2}$$

where, *a* represents the pre emphasis coefficient, *S(n)* represents the speech signal before the pre emphasis processing, and $\hat{S}(n)$ represents the speech signal after the pre emphasis filtering processing.

The voice signal is windowed and divided into frames, and the frame length is recorded as *N*. In the processing process, the method of overlapping segmentation is adopted [12], as shown in Fig. 2.

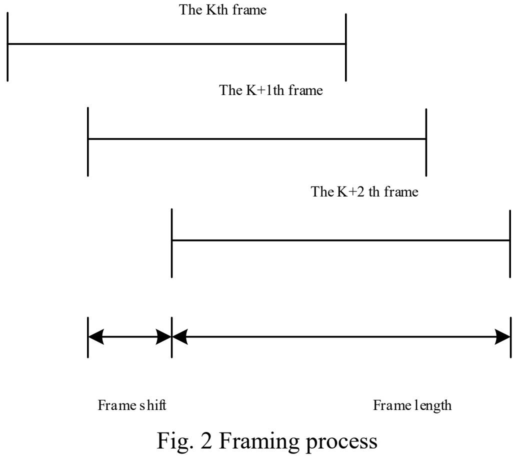

Fig. 2 Framing process

In the process of speech recognition, silence will inevitably appear. In order to remove the silence and preserve the signal segment of speech features, the speech endpoint detection method is introduced to remove the silence part [13]. After the voice semaphore is windowed and divided into frames, the average energy of the voice semaphore is calculated and less than zero percent and energy threshold are obtained. The average energy is calculated as follows:

$$E(i)=\sum_{n=1}^{N}\left(\left|X_i(n)\right|\right) \tag{3}$$

$$E(i)=\sum_{n=1}^{N}\left(X_i^2(n)\right) \tag{4}$$

$$E(i)=\sum_{n=1}^{N}\left(\log X_i^2(n)\right) \tag{5}$$

where, N is the frame length of speech signal and $X_i(n)$ is the amplitude energy of speech signal. The calculation formula of less than zero percent is as follows:

$$Z_n=\frac{1}{2}\sum_{m=-\infty}^{\infty}(|\operatorname{sgn}[X(m)]+\operatorname{sgn}[X(m-1)]|\,w(n-m)) \tag{6}$$

The endpoint frame part of the voice semaphore is detected by using the duplicato threshold value endpoint detection algorithm [14]. The specific flow is shown in Figure 3.

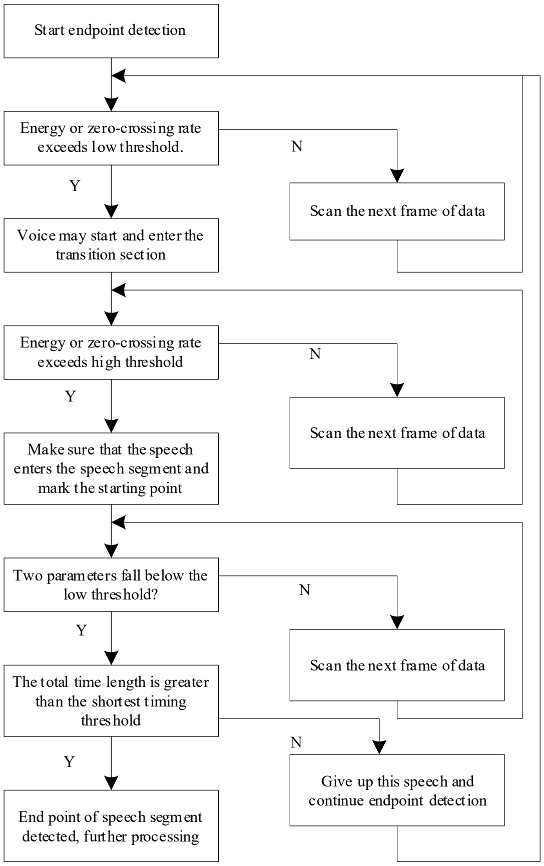
Start endpoint detection
Energy or zero-crossing rate exceeds low threshold.
N
Y
Scan the next frame of data
Voice may start and enter the transition section
Energy or zero-crossing rate exceeds high threshold
N
Y
Scan the next frame of data
Make sure that the speech enters the speech segment and mark the starting point
Two parameters fall below the low threshold?
N
Y
Scan the next frame of data
The total time length is greater than the shortest timing threshold
N
Y
Give up this speech and continue endpoint detection
End point of speech segment detected, further processing

Fig. 3 Flow chart of double threshold endpoint detection algorithm

B. Speech Feature Extraction

Speech features have many kinds of feature parameters. In this paper, we choose MFCC and the first order difference of MFCC as feature parameters [15]. MFCC is designed based on linear prediction cepstrum coefficient. The flow chart of linear prediction of cepstrum coefficient is shown in Fig. 4.

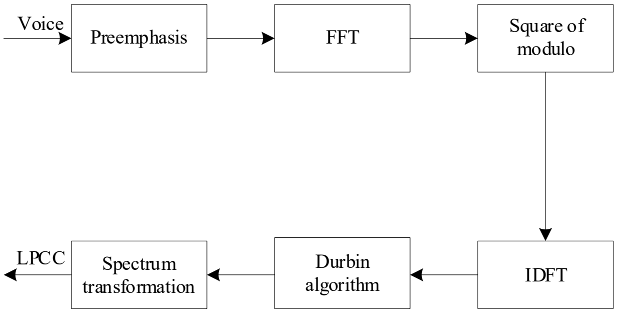

Fig. 4 Flow of linear prediction of cepstrum coefficient

The cepstrum coefficient of linear prediction is a very important parameter. The linear prediction coefficient is obtained by recursion [16], the iterative relationship is as follows:

$$\begin{aligned} c_0 &= \log G^2 \\ c_m &= a_m+\sum_{k=1}^{m-1}\frac{k}{m}c_k a_{m-k}, 1\le m\le p \\ c_m &= \sum_{k=1}^{m-1}\frac{k}{m}c_k a_{m-k}, m>p \end{aligned} \tag{7}$$

Based on formula (7), the conversion relationship between MFCC and actual frequency of speech signal can be obtained as follows:

$$f_{\mathrm{nw}}=2595\log_{10}(1+\frac{f}{700}) \tag{8}$$

The calculation process of Mel frequency cepstrum coefficient is shown in Fig. 5.

According to the calculation process of MFCC, the calculation steps of MFCC are as follows: firstly, the number of points of each frame speech sampling sequence is determined, and it is pre emphasized. After discrete fast fourier transform transformation [17], taking the square of the modulus to obtain the discrete power spectrum $S(n)$ of speech samples and the discrete power spectrum $S(n)$ is calculated by using M $H_m(n)$. The discrete cosine transform (DCT) of speech samples is obtained, and the result of DCT is taken as the MFCC.

Human ears are more sensitive to the dynamic features of voice [18]. Therefore, this paper uses the first-order difference of MFCC and MFCC as parameters to describe. The solution principle of the first-order difference can be realized by formula (9):

$$d(n)=\frac{1}{\sqrt{\sum_{i=-k}^{k}i^2}}\sum_{i=-k}^{k}ic(n+i) \tag{9}$$

where, k is a constant and c is a speech parameter. However, in the actual speech recognition system, the speech feature parameters can better reflect the dynamic characteristics of speech.

In addition to static features, speech signals also have transient features between consecutive frames. The first-order difference of MFCC and MFCC is described as a parameter, the static feature and dynamic feature can be organically combined to improve the recognition rate of speech recognition.

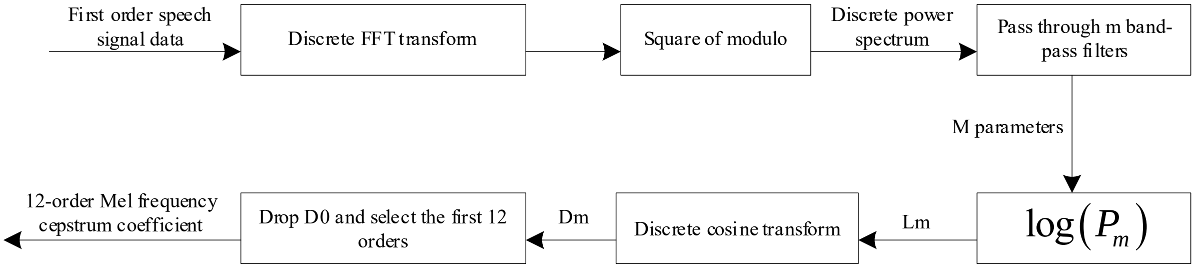
First order speech signal data
Discrete FFT transform
Square of modulo
Discrete power spectrum
Pass through m band-pass filters
M parameters
$\log(P_m)$
Lm
Discrete cosine transform
Dm
Drop D0 and select the first 12 orders
12-order Mel frequency cepstrum coefficient

Fig. 5 Calculation process of Mel frequency cepstrum coefficient

C. *Design Speech Recognition Process*

Speech recognition process can be roughly divided into are two stages, the first stage is to extract speech features, which includes digital sampling of speech signal and acoustic signal analysis. This paper uses spectrum analysis technology to extract and analyze the content of speech signal [19]; The second stage is to identify the phoneme, phoneme group and words of speech signal, and there are many methods to use in this stage. There are many kinds of recognition methods, such as expert system, artificial neural network, artificial bee colony algorithm, dynamic time warping, deep neural network, hidden Markov model, etc. [20], some recognition methods may try to understand the real content of the speech signal, that is, to convert the speech signal into the actual meaning that the speaker wants to express, or to realize the meaning of what the speaker says. The flow of speech recognition is shown in Fig. 6.

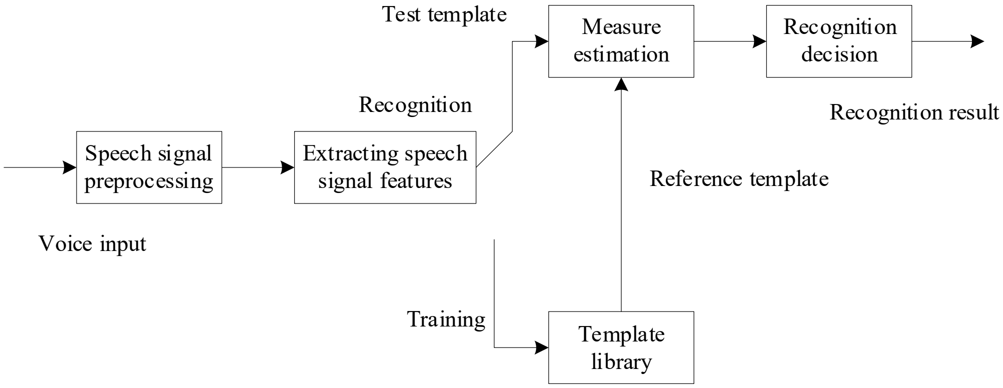

Fig. 6 Speech recognition process

In the multi-layer awareness network environment, the hierarchical situation element collection model is used to analyze the voice situation elements. The voice hierarchical situation recognition framework under the neural network propagation operation is shown in Fig. 7.

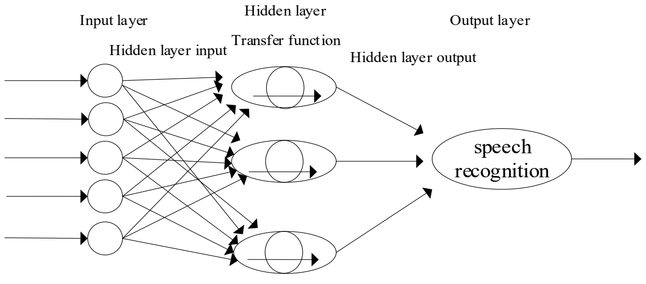

Fig. 7 Framework diagram of speech hierarchical situation recognition based on neural network propagation operation

Neural network is an optimized neural network which uses wavelet function to replace the hidden layer transfer function of neural network. Through supervised learning method, the weights are gradually modified by the lowest speed descent method of deviation forward transmission, so as to achieve the goal of speech recognition. Neural network can deal with speech recognition and other problems well, and has good generalization performance, so it has a good space in many analog speech recognition.

To sum up, using two methods of hardware and software, using the transfer function of the filter, the speech signal processing in the filter, according to the framing process, to add Windows and framing the voice signal, in order to remove mute of speech signal, calculates the average energy and zero crossing rate of speech signals, with double threshold endpoint detection algorithm, extraction of speech signal characteristics, by analyzing the principle of speech signal recognition, the process of speech recognition is designed, and the recognition of speech signal is realized.

III. Experiment Analysis

A. Establishment of Speech Sample Database

In the multi-layer perceptual network environment, five people are randomly selected, and the proportion of men and women is balanced. The collected speech is taken as the main source of speech samples. The randomly selected five people must pronounce clearly, so as to complete the speech collection work in the multi-layer perceptual network environment. In the process of simulation analysis, the acquisition frequency of speech samples is set to 8kHz, and the coding bits are 16 bits. In the multi-layer perceptual network environment, the speech samples are 1-10 Chinese digit sounds. Each person collects 10 groups of collected speech signals, each group of speech signals contains 1-10 Chinese digit sounds, and the ten numbers are recorded separately.

The simulation analysis experiment of single-layer perceptual network model is realized on ordinary computer, and multi-layer perceptual network model is realized on MATLAB. C language is used to realize the deep belief of network model. The operating system selected is Windows 7, the main frequency of CPU is 2.5GHz, and the load of software program is about 25%.

B. Single-layer Perceptual Network

1) Improved single-layer perceptual network model

Based on the application points of single-layer perceptual network, this paper attempts to improve the single-layer perceptual network algorithm to make the convergence speed of single-layer perceptual network model faster in the training process, the specific steps are as follows:

Step 1: Normalize input vector

The relationship between the modified weight coefficient and the $P(i)$ value of the input data is positive correlation, and the input vector is normalized to the interval of $[-1.0,1.0]$, which can avoid entering the saturation region of the excitation function;

Step 2: Randomly select the floating-point number whose initial value is in the interval of $[-1.0,1.0]$ in the network connection matrix, so that the single-layer perceptual network has a certain stability in convergence.

Step 3: Use batch processing to learn speech samples

The total error of each group of speech sample data is used as the basis of weight adjustment, so that the gradient descent has an average effect.

Step 4: Introduce adaptive learning rate

The perceptual network is a new perceptual mode of the Internet of Things proposed in the context of ubiquitous mobile perceptual devices. For the specific problem of speech recognition, we need to choose an appropriate learning rate first. Usually we use experiments to obtain the learning rate or rely on experience to obtain the learning rate, but it is not necessarily suitable for the learning rate with good effect in the early stage of learning. Therefore, it needs to automatically adjust the learning rate in the process of network learning. The adjustment criterion of the learning rate is to detect whether the correction result of the weight leads to the decrease of the error function. If the error function decreases, it means that the selection of the learning rate is too small. In this way, an appropriate amount can be added. If the error function does not decrease, the value of the learning rate should be reduced appropriately.

$$\eta(t+1)=\begin{cases}1.05\cdot\eta(t), & E(t+1)<E(t)\\ 0.75\cdot\eta(t) & E(t+1)>1.04\cdot E(t)\\ \eta(t) & \text{others}\end{cases} \tag{10}$$

Step 5: Limit the number of cycles

Due to the small number of speech samples collected in the simulation analysis experiment, it is difficult to achieve a good error between the expected output and the actual output. In the learning process of the single-layer perceptual network model, the number of cycles is the only cycle judgment condition.

2) Identification performance analysis

In the process of simulation analysis, the characteristic parameters used include first-order difference of MFCC and MFCC. The two coefficients are compressed and regularized into four frames of data, and the dimension of one frame parameter is set to 12.

According to the different input parameters, it can be divided into two speech recognition modes, that is

Mode 1: MFCC is as the only characteristic.

$$X_n=M_{FCC1}+M_{FCC2}+M_{FCC3}+M_{FCC4} \tag{11}$$

Mode 2: MFCC and the first-order MFCC are combined as the mixed characteristics.

$$\begin{aligned}X_n^{'}&=M_{FCC1}+M_{FCC2}+M_{FCC3}+M_{FCC4}\\ &+\Delta M_{FCC1}+\Delta M_{FCC2}+\Delta M_{FCC3}+\Delta M_{FCC4}\end{aligned} \tag{12}$$

(1) The influence of hidden layer unit on the performance of single-layer perceptual network

In the first mode, MFCC is regarded as the only feature of speech classification and recognition. The number of nodes in the output layer of the single-layer perceptual network model is 4, and the number of input nodes is 48. The network used is a single hidden layer. Then the influence of the number of hidden layer nodes on the single-layer perceptual network model is shown in Table 1.

Table 1 The influence of the number of hidden layer nodes in mode 1 on the single-layer perceptual network model

Number of hidden layer nodes	Study time (s)	Recognition accuracy /%
10	17	80.4
15	29	80.8
20	39	82.0
25	52	81.2

It can be seen from the results in Table 1 that at the beginning of the simulation test, with the number of hidden layer nodes increasing, the recognition accuracy of the single-layer perceptual network model is getting higher and higher. However, when the number of hidden layer nodes is 25, the recognition accuracy of the single-layer perceptual network model begins to decline. There is a positive correlation between the learning time and the number of hidden layer nodes. In mode 1, when the number of hidden layer nodes is 20, the performance of single-layer perceptual network model is the best.

Next, taking the mixed parameters in mode 2 as the only feature, the number of output layer nodes of the single-layer perceptual network model is still set to 4, and the number of input nodes is 96. Then the influence of the number of hidden layer nodes on the single-layer perceptual network model is shown in Table 2.

Table 2 The influence of the number of hidden layer nodes in mode 2 on the single-layer perceptual network model

Number of hidden layer nodes	Study time (s)	Recognition accuracy /%
10	26	82.0
20	53	83.6
30	81	87.2
40	106	86.4
50	129	86.0

It can be seen from the results in Table 2 that the performance of the single-layer perceptual network model of the two modes is basically consistent in the change trend. In the early stage of simulation test, with the number of hidden layer nodes increasing, the recognition accuracy of the single-layer perceptual network model is getting higher and higher. However, when the number of hidden layer nodes increases from 40 to 50, the recognition accuracy of the single-layer perceptual network model begins to decline, and the learning efficiency is low. There is a positive correlation between time and the number of hidden layer nodes. When the number of hidden layer nodes in mode 2 is 30, the performance of single-layer perceptual network model is the best.

(2) The influence of characteristic parameters on the performance of single-layer perceptual network

From the results in Table 1 and Table 2, the number of hidden layer nodes with the best recognition effect in the single-layer perceptual network model is used. Combined with the two patterns, the single-layer perceptual network model is used for training and recognition. The results are shown in Table 3.

Table 3 The best network performance under different characteristic parameters

Characteristic parameter scheme	MFCC	MFCC+△MFCC
Number of hidden layer units	20	30
Learning practice / s	39	81
Number of perceived nodes	10	10
Recognition accuracy /%	82.0	87.2

It can be seen from the results in Table 3 that the accuracy of speech classification and recognition reaches 87.2% when the single-layer perceptual network model uses the mixed parameters of mode 2 as the unique feature, while the accuracy of speech classification and recognition is only 82% when the mixed parameters of mode 1 are used as the unique feature, which indicates that the former has better learning effect. But the learning time of the former is longer, about twice that of the latter. It shows that with more and more first-order difference parameters of MFCC, the representativeness of speech signal feature parameters will be stronger, which significantly improves the effect of speech recognition.

C. Multi-layer Perceptual Network

1) Experimental settings

In the simulation experiment, multi-layer perceptual network is used to recognize the input speech signal, and the first-order difference of the MFCC and the MFCC are used as the extraction features of the speech signal. The original input matrix is expanded horizontally, converted into a 96 dimensional vector, and normalized, so that the size of the input data is within [0, 1]. According to the different types of speech data set, the number of units at the top level is 1 ~ 10. The model used in the experiment contains a hidden layer, that is, a restricted Boltzmann machine model needs to be pre trained.

2) Building a restricted Boltzmann machine model

(1) Setting RBM parameters

In order to learn better and improve the effect of RBM model, if the RBM parameter setting is unreasonable, it is not easy for some speech datasets and RBM to be modeled correctly by RBM model. Therefore, it is necessary to strengthen the understanding of RBM parameter setting rules. The parameter setting rules used in the simulation experiment are as follows:

- Update parameters. In order to improve the computational efficiency, the speech sample set is divided into several small batches of data, and then batch learning is used. The training set in the experimental testing process contains 10 categories. RBM uses small batch processing method in learning, and each batch of speech sample data contains 1~10 sample data, so as to reduce the sampling error of gradient estimation.
- Initial value of parameter. In the simulation experiment, Gaussian distribution is used to assign the connection weights, and the unit bias in the hidden layer and output layer of the network is set to 0. For the visible unit, it is easy to use the hidden layer unit in the early stage of the experiment, so that the speech sample data activates the *i*-th eigenvalue with probability *pi*, and the initial bias value here is not zero.

(2) Softmax classifier

After learning RBM model, a layer of classifier is added to the top layer of multi-layer perceptual network. Softmax is used to expand logic analysis. The speech sample data is divided into multiple categories, and each category is mutually exclusive. The specific calculation method is as follows:

$$S_i = g_\theta(x) = \frac{e^{g_i}}{\sum_{i=1}^{d} e^{g_i}} \tag{13}$$

In the formula, $g_\theta(x) = WX + g_i$, $\theta = \{W, g_i\}$, X is the specific state of each cell in the hidden layer, θ is the parameter set, and g_i is the offset value of the output layer.

$r \in [0,1]^d$ is used to represent the real classification of speech sample data. In order to minimize the error between the actual output and the expected output, the error function is as follows:

$$H(r,S) = -\sum_{i=1}^{d}\left(r_i \log S_i + \left(1-r_i\right)\log\left(1-S_i\right)\right) \quad (14)$$

The minimum error function is used to train the parameters of multi-layer perceptual network.

$$\theta^{*} = \operatorname{argmin}_{\theta} H(r,S) \quad (15)$$

After solving the partial derivative of the above formula, we can get the following results:

$$\frac{\partial H(r,S)}{\partial \theta} = -\sum_{i=1}^{d}\left(r_i - S_i\right)\frac{\partial g_i}{\partial \theta} \quad (16)$$

The partial derivatives of $H(r, S)$ for W and g_i can be expressed as:

$$\frac{\partial H(r,s)}{\partial W} = (S-r)^{T} X \quad (17)$$

$$\frac{\partial H\left(r_r S\right)}{\partial b} = S - r \quad (18)$$

Using gradient descent method, the process of updating weights is as follows:

$$W^{'} = W - \eta\left((S-r)^{T} X + \lambda W\right) \quad (19)$$

$$b^{'} = b \quad \eta(S \quad r + \lambda b) \quad (20)$$

In the formula, η is the learning rate and λ is the attenuation factor of the weight.

(3) Result analysis

When the number of iterations is as follows, the number of iterations is 1000 and the number of hidden layers is 2000.

Table 4 Comparison of multi-layer perceptual network models with different number of hidden layer units

	Single hidden layer 2000	Hidden layer 500
Average reconstruction error	3.717	3.564
Recognition rate /%	87.6	50
Study time / s	910	324

From the above results, it can be seen that the input data of pattern 2 after time warping is used to expand the dimension of the data through the hidden layer, and the speech recognition is constructed at the output end. The more the number of network hidden layers, the better performance can be obtained when the sample size of voice data is relatively small, but the time is more and more. To find the ideal number of hidden layer nodes, the approximation ability and generalization ability of the network are guaranteed, and the requirements of high-precision approximation are met. Then, the search interval is extended according to the golden section principle, and the number of hidden layer nodes with stronger approximation ability is obtained by searching for the optimization in the interval.

RBM reconstruction performance is shown in Fig. 8:

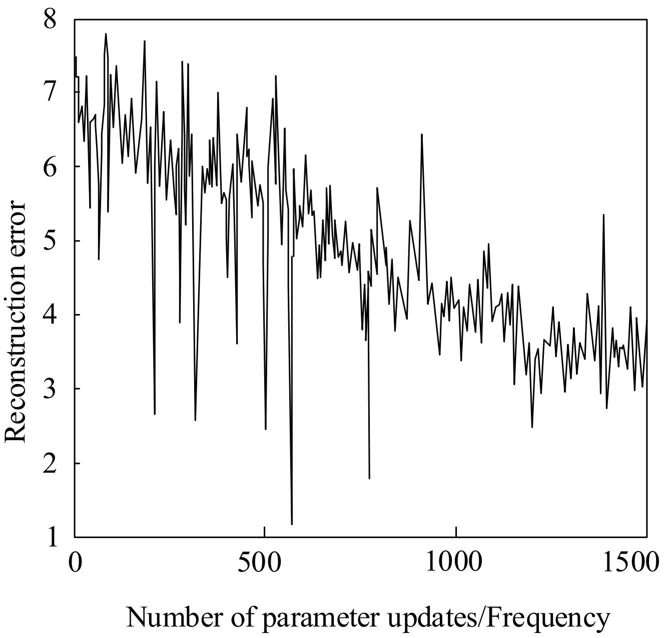

Fig. 8 RBM reconstruction performance

Reconstruction error is one of the simplest methods to evaluate RBM model. Because RBM model often has two-layer structure, the value of visible and hidden elements in each reconstruction will produce butterfly effect. The larger the number of reconstruction elements is, the more likely it is to obtain accurate activation value, but it will also become more and more complex and lack of stability. It can be seen clearly from Fig. 8 that the reconstruction error changes violently in a short time, but on the whole, it decreases, but tends to be stable in the later stage. In fact, it is always balanced at a certain level, indicating that the RBM model will be in a stable state with the constant updating of parameters.

D. Comparative Analysis of Efficiency of Different Methods

In order to further verify the feasibility of speech classification and recognition method in multi-layer sensing network environment, an experiment is carried out by comparing efficiency indexes. Efficiency refers to the amount of work a speech recognition method does per unit of time, or the ratio of the results achieved by a speech recognition method to the time and labor it takes to complete the recognition. In this paper, the speech recognition method in the single-level perception network environment is compared with the speech recognition method in the multi-level perception network environment, in order to ensure the fairness of the experiment, 200 hidden layer units are selected, and the results are shown in Table 5.

Table 5 Efficiency comparison results of different methods

-	Method of this paper	Reference [6] method	Reference [7] method	Reference [8] method
Number of hidden layer units/Quantity	200	200	200	200
Number of identification	196	164	178	182

units/Quantity				
Number of unrecognized units/Quantity	4	36	22	18
Recognition rate /%	98%	82%	89%	91%

It can be seen from the comparison results in Table 5 that contrast method in this paper, the reference [6] method, the reference [7] method, and the reference [8] method, speech recognition rate is higher, and the literature method, respectively, 82%, 89% and 91%, the reason is that the designed multi-level perception speech recognition method in the network environment, the multi-level perception network hidden layer extends the voice sample data, the characteristics of the input The recognition rate is increased by increasing the dimension of the voice sample data characteristics.

IV. DISCUSSION

Speech recognition technology is a rich connotation, widely used human-computer interaction key technology, improve the recognition accuracy, combined with linear prediction cepstrum coefficient and Meyer frequency cepstrum coefficient, calculate the average energy of speech signal and statistics of zero crossing rate, extract the characteristics of speech signal. Considering the deficiency of the traditional method in recognition rate and recognition speed, this paper makes some improvements to it, and through the simulation calculation, it is verified that this improved method can improve the recognition rate and recognition speed of the network.

V. CONCLUSIONS

In this paper, a method of speech recognition in multi-layer perceptual network environment is proposed. In the multi-layer perceptual network environment, the speech signal is preprocessed first. By extracting the features of speech signal, the speech recognition process is designed, and the speech recognition is realized. The results show that the recognition method can effectively improve the speech recognition rate. The speech recognition method is designed to understand what people are saying and communicate with them verbally, eliminating the need for typing. It will also be easier for people with different languages to communicate. This is also the basis of speech recognition technology. Speech recognition can change syllables, sounds and phrases into words, symbols or controls, responses, etc. Speech recognition has been widely used in industry, commerce, culture and finance, especially in computer, communication electronic system, information processing and other aspects.

As the number of units in the multilayer perception networks environment more and more, the computational complexity of multilayer perception networks model is becoming more and more big, so that the recognition rate also increased, but how to balance two indicators, also need to further enhance understanding, in future studies, whether need to focus on the speech recognition rate increased with the increase of model complexity increases, In this paper, speech recognition is carried out in the multi-layer sensing network environment, only small batch of isolated words are recognized, and continuous speech recognition in the context of large vocabulary is further studied. Through further research, a more suitable learning method for speech recognition is found, and the speed of recognition is improved more greatly.

ACKNOWLEDGMENT

This work was supported in part by the Basic scientific research project of Central University of Railway Police College 2020TJJBKY015, in part by the Scientific and Technological Key Project of Henan Province under Grant 202102310291, in part by the Research project of Ministry of public security 2020JSYJC29, in part by the Scientific Key Research Project of Henan Province for Colleges and Universities under Grand 20A520033.

References

[1] A. Valiyavalappil Haridas et al., "Taylor-DBN: A New Framework for Speech Recognition Systems", International Journal of Wavelets, Multiresolution and Information Processing, vol. 12, no. 9, pp. 26-35, 2020. https://doi.org/10.1142/S021969132050071X

[2] E. Owusu et al., "Face Detection Based on Multilayer Feed - forward Neural Network and Haar Features", Software: Practice and Experience, vol. 49, no. 1, pp. 120-129, 2019. https://doi.org/10.1002/spe.2646

[3] F. E. Ayo et al., "Machine Learning Techniques for Hate Speech Classification of Twitter Data: State-of-the-art, Future Challenges and Research Directions", Computer Science Review, vol. 38, p. 100311. https://doi.org/10.1016/j.cosrev.2020.100311

[4] R. Ghosh et al., "A Modified Grey Wolf Optimization Based Feature Selection Method from EEG for Silent Speech Classification", Journal of Information and Optimization Sciences, vol. 40, no. 8, pp. 1639-1652, 2019. https://doi.org/10.1080/02522667.2019.1703262

[5] M. Malcangi and P. Grew, "Evolving Connectionist Method for Adaptive Audiovisual Speech Recognition", Evolving Systems, vol. 8, no. 1, pp. 85-94, 2017.

[6] P. X. Jiang et al., "Feature Characterization Based on Convolution Neural Networks for Speech Emotion Recognition", Chinese Journal of Electron Devices, vol. 42, no. 4, pp. 998-1001, 2019. https://doi.org/10.1007/s12530-016-9156-6

[7] H. H. Gu, "Multi-band Anti-Noise Speech Recognition Method Simulation Based on Multi-Core Learning", Computer Simulation, vol. 36, no. 10, pp. 364-367, 395, 2019.

[8] Z. Song, "English Speech Recognition Based on Deep Learning with Multiple Features", Computing, vol. 102, no. 3, pp. 663-682, 2020. https://doi.org/10.1007/s00607-019-00753-0

[9] J. R. C. de Lara et al., "A Method to Compensate the Influence of Speech Codec in Speaker Recognition", International Journal of Speech Technology, vol. 21, no. 4,

pp. 975-985, 2018. https://doi.org/10.1007/s10772-018-9547-0

[10] G. M. Sapijaszko and W. B. Mikhael, "Facial Recognition System Using Mixed Transform and Multilayer Sigmoid Neural Network Classifier", Circuits, Systems, and Signal Processing, vol. 39, pp. 6142-6161, 2020. https://doi.org/10.1007/s00034-020-01453-3

[11] E. Gourdin et al., "Design of Reliable Communication Networks", Annals of Telecommunications, vol. 73, no. 1-2, pp. 1-3, 2018. https://doi.org/10.1007/s12243-017-0624-1

[12] A. M. Elsayad et al., "Diagnosis of Hepatitis Disease with Logistic Regression and Artificial Neural Networks", Journal of Computer Science, vol. 16, no. 3, pp. 364-377, 2020. https://doi.org/10.3844/jcssp.2020.364.377

[13] H. Hadizadeh et al., "A Perceptual Distinguishability Predictor For JND-Noise-Contaminated Images", IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2242-2256, 2019. https://doi.org/10.1109/TIP.2018.2883893

[14] C. Sui et al., "A Cascade Gray-stereo Visual Feature Extraction Method for Visual and Audio-visual Speech Recognition", Speech Communication, vol. 90, pp. 26-38, 2017. https://doi.org/10.1016/j.specom.2017.01.005

[15] L. M. Lee et al., "Improved Hidden Markov Model Adaptation Method for Reduced Frame Rate Speech Recognition", Electronics Letters, vol. 53, no. 14, pp. 962-964, 2017. https://doi.org/10.1049/el.2017.0458

[16] S. Khajehasani and L. Dehyadegari, "Speech Recognition Using Elman Artificial Neural Network and Linear Predictive Coding", Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), vol. 12, no. 16, pp. 65-72, 2020. https://doi.org/10.2174/2213275912666190411113728

[17] I. Speck et al., "Comparison of Speech Recognition and Localization Ability in Single-sided Deaf Patients Implanted With Different Cochlear Implant Electrode Array Designs", Otology & Neurotology, vol. 42, no. 6, pp. 98-105, 2021. https://doi.org/10.1097/MAO.0000000000002864

[18] A. Kumar and R. K. Aggarwal, "Discriminatively Trained Continuous Hindi Speech Recognition Using Integrated Acoustic Features and Recurrent Neural Network Language Modeling", Journal of Intelligent Systems, vol. 30, no. 1, pp. 165-179, 2020.

[19] S. P. S. Bibin et al., "A Low Latency Modular-level Deeply Integrated MFCC Feature Extraction Architecture for Speech Recognition – ScienceDirect", Integration, vol. 76, pp. 69-75, 2021. https://doi.org/10.1016/j.vlsi.2020.09.002

[20] Y. Wren et al., "A Systematic Review and Classification of Interventions for Speech- sound Disorder in Preschool Children", International Journal of Language & Communication Disorders, vol. 53, no. 5, pp. 446-467, 2018. https://doi.org/10.1111/1460-6984.12371

Kai Zhao was born in Zhengzhou, Henan, P.R. China, in 1982.
He received the Master degree from Northwestern Polytechnical University, P.R. China. Now, he works in Image and Network Investigation Department, Railway Police College, His research interests include digital image processing, artificial intelligence and information security.
E-mail: zhaokai966@163.com

Dan Wang was born in Shenyang, Liaoning, P.R. China, in 1981.
She received the Master degree from Northwestern Polytechnical University, P.R. China. Now, she works in school of Intelligent Engineering, Zhengzhou University of Aeronautics. Her research interests include digital image processing and artificial intelligence.
E-mail: wangdan612@163.com

Creative Commons Attribution License 4.0 (Attribution 4.0 International , CC BY 4.0)