Anti Noise Speech Recognition Based on Deep Learning in Wireless Communication Networks

Yanning Zhang(✉), Lei Ma, Hui Du, and Jingyu Li

Beijing Polytechnic, Beijing 100016, China
witgirl316@126.com

Abstract. As a new high-tech industry, the application of speech recognition technology is becoming more and more competitive, with a wide range of application fields and application prospects, and has far-reaching significance for the development of science and technology. The communication environment of wireless communication network will bring various types of noise to speech, so an anti noise speech recognition method based on deep learning of wireless communication network is designed to achieve anti noise speech recognition in this environment. The voice signal of wireless communication network is preprocessed by anti aliasing filtering, analog-to-digital conversion, pre emphasis, framing and windowing, endpoint detection, etc. A series of denoising processes are implemented for the voice signal of wireless communication network, and different speech preprocessing methods are adopted for different characteristics of noise. A speech signal feature extraction method based on improved EMD is designed and implemented. The speech recognition model is designed based on the regression neural network in deep learning, and the anti noise speech recognition of wireless communication network is realized. Test results show that the lowest word error rate of this method is 0.156, and the word error rate is also low.

Keywords: Deep Learning · Improve EMD · Hidden Markov Model · Anti Noise Speech Recognition

1 Introduction

With the development of society, wireless communication has been used in many scenarios, and many new communication networks have emerged. The emergence and application of wireless ad hoc networks, wireless mesh networks, wireless sensor networks and other networks have improved people's living standards. In wireless communication networks, communication speech often produces noise, which affects people's auditory judgment. Speech is the most natural, effective and convenient means for human information exchange. Therefore, it is of great significance to study speech recognition in wireless communication networks [1].

The research work of speech recognition began in the 1950s. AT&T Bell Laboratories invented the first speech recognition system - Audry system, which can recognize

L. Yun et al. (Eds.): ADHIP 2023, LNICST 547, pp. 152–168, 2024.
https://doi.org/10.1007/978-3-031-50543-0_11

10 English numbers. In the late 1960s and early 1970s, with the development of computer technology, speech recognition technology has also made substantial development and become an important subject in the research field [2]. Firstly, the development of computer technology provides hardware and software conditions for the realization of speech recognition. Secondly, "Dynamic programming method" has effectively solved the problem of difficult alignment of speech signals on the time axis. At the same time, speech recognition has proposed two technologies, dynamic time warping technology and linear prediction coding technology, which provide effective methods for unequal length matching of speech signals and feature extraction of speech signals. The use of these two technologies can basically achieve the speech recognition of individual isolated words. At the same time, the hidden Markov model I5 technology has also been preliminarily applied and the vector quantization L0 theory has been proposed. In the 1980s, the research on speech recognition gradually deepened, and the focus of research turned to non-specific, large vocabulary continuous word speech recognition, and many new speech recognition algorithms emerged. Another breakthrough development is the proposal of the model technology based on statistics. Compared with the template matching technology, this technology does not require the refinement of speech features, but constructs a speech recognition system from the perspective of the overall analysis of speech signals. In addition, the successful application of hidden Markov model and artificial neural network in speech recognition system has made a great breakthrough in continuous speech recognition. At present, hidden Markov model has become the mainstream technology of speech recognition. Among them, Sphinx system developed by CMU in 1988 is a typical speech recognition system. It is the first non-specific, large vocabulary continuous speech recognition system with high performance in the world. Its database has 99 words and 4200 consecutive sentences, and the system's recognition rate can reach 95.8% at the highest. Since the 1990s, many developed countries in Europe and the United States and some famous technology companies have invested huge funds in the field of speech recognition to realize the practicality of speech recognition, and have begun to provide relevant products for the market. The recognition rate of the speech recognition system has been greatly improved, and the speech recognition technology has further matured and started to enter the practical application from the laboratory. At the same time, the recognition of Chinese speech has been paid more and more attention.

At present, the pure speech recognition technology has developed more maturely. In a quiet environment, the speech acquisition and template training can match the speech to be recognized very well. Therefore, the existing speech recognition system can complete the recognition very well, and the recognition efficiency is also very high. However, in the wireless communication network environment, speech will contain noise. Noise interference affects the matching between the speech to be recognized and the training template, which further affects the recognition rate of the system. Therefore, anti noise becomes one of the key technologies in modern speech recognition research.

For the research of anti noise speech recognition, some scholars proposed an anti noise speech recognition technology based on RBF neural network: to solve the problem of poor performance of speech recognition system in noisy environment at present, RBF neural network has the best approximation performance, fast training speed and other characteristics. The anti noise speech recognition system based on RBF neural network is

implemented by clustering and fully supervised training algorithms respectively. The hidden layer training of the clustering algorithm adopts the K-means clustering algorithm, and the output layer learning adopts the linear least square method; The adjustment of all parameters in the fully supervised algorithm is based on the gradient descent method, which is a supervised learning algorithm and can select parameters with good performance. Experiments show that the fully supervised algorithm has higher recognition rate than the clustering algorithm under different signal-to-noise ratios. Other scholars have proposed an anti noise speech recognition technology based on 3F speech enhancement distortion compensation: in order to improve the robustness of speech recognition system based on hidden Markov model in noisy environments, this paper studies an anti noise speech recognition algorithm based on speech enhancement distortion compensation. At the front end, speech enhancement effectively suppresses background noise. Thus, the signal to noise ratio of the input signal is improved, and the spectral distortion and residual noise caused by speech enhancement are adverse factors for speech recognition. The influence will be compensated by parallel model merging in the recognition phase or cepstrum mean normalization in the feature extraction phase. The experimental results show that this algorithm can significantly improve the recognition accuracy of the speech recognition system in the noisy environment in a very wide SNR range, and the improvement of the system performance is especially obvious in the low SNR situation, such as 5 dB white noise. Compared with the baseline recognizer, this algorithm can reduce the error rate by 67.4%. There are complex nonlinear relationships in speech recognition, such as time-varying speech and noise interference. When using the above methods for noise resistant speech recognition in wireless communication networks, it is difficult to capture these nonlinear relationships, resulting in poor recognition performance. And deep learning models have strong nonlinear modeling capabilities, which can better adapt to complex speech features and improve recognition performance. Therefore, a wireless communication network anti noise speech recognition method based on deep learning is designed.

2 Design of Anti Noise Speech Recognition Method for Wireless Communication Network

2.1 Speech Signal Pre-processing

The voice signal of wireless communication network is preprocessed by anti aliasing filtering, analog-to-digital conversion, pre emphasis, framing and windowing, endpoint detection, etc.

The anti aliasing filter is selected as the band-pass filter. The power frequency interference of 50 Hz power supply is suppressed through its high pass filtering part, and all the components in the frequency domain components of the input signal that exceed the following formula are suppressed through its low pass part to prevent aliasing interference.

$$h = \frac{k_s}{2} \tag{1}$$

In formula (1) k_s it refers to the sampling frequency.

Human pronunciation is a continuous analog signal, which cannot be processed by computer. Therefore, it is necessary to convert analog voice signals into digital signals, that is, analog-to-digital conversion. Since the average power spectrum of speech signal is affected by glottal excitation and nose and mouth radiation, the high-frequency end drops significantly above 800 Hz, so when calculating the spectrum of speech signal, the higher the frequency is, the smaller the corresponding component is, and the spectrum of high-frequency part is more difficult to find than that of low-frequency part. Therefore, pre emphasis processing should be carried out in preprocessing. The purpose of pre emphasis is to improve the high frequency part so that the spectrum of the signal becomes flat. The pre emphasis part is realized by a digital filter to improve the high-frequency characteristics, which is generally a first-order digital filter:

$$g(s) = 1 - \nu s^{-1} \tag{2}$$

In formula (2) ν is the pre weighting coefficient; s it is a voice signal.

Among ν The value of is generally between 0.9 and 1, and the typical value is 0.9375.

Speech signal is a typical non-stationary signal, its characteristics change with time. However, the formation process of speech is closely related to the movement of the vocal organs. This physical movement is much slower than the speed of sound vibration. Therefore, speech signals can often be assumed to be stable in a short time, that is, in a period of 10–30 ms, its spectral characteristics can be seen as nearly unchanged. In this way, the analysis and processing method of stationary process can be used. From this assumption, various short-time processing methods based on frames are derived, and various speech processing methods discussed later are based on this assumption. In order to smooth the transition between frames and maintain their continuity, the method of overlapping segments [3] is adopted here. The overlapping part of the previous frame and the next frame is called frame shift, which is generally 0–1/2 of the frame length. The specific method is shown in Fig. 1.

In order to reduce the Gibbs effect caused by truncation after framing, it is usually necessary to windowing each frame signal. Each 10 ms–30 ms frame of the voice is analyzed in turn. This operation is called windowing. The window slides on the voice signal to frame the voice signal. When adding windows to voice signals, use Hamming windows.

The primary problem of speech signal processing is to determine the pure noise segment of a noisy speech, the noisy speech segment, and the start and end points of each speech segment, that is, endpoint detection in signal processing. Endpoint detection is the basis of speech signal processing. If the endpoint detection of a speech recognition system is done well, it can not only reduce the amount of calculation, but also improve the recognition rate of the system [4]. The calculation process of the short-term average zero crossing rate is shown in Fig. 2, that is, first process the speech signal sequence in pairs to check whether there is sign transformation, and if there is, it means zero crossing once; Then make a first-order difference and take the absolute value; Finally, low-pass filtering is performed to output short-term average zero crossing.

Since the short-term average zero crossing rate can reflect the frequency to a certain extent, while the energy of voiced voice is concentrated in the low frequency band and the energy of unvoiced voice is concentrated in the high frequency band, the zero crossing rate is generally low in the voiced voice band and high in the unvoiced voice band.

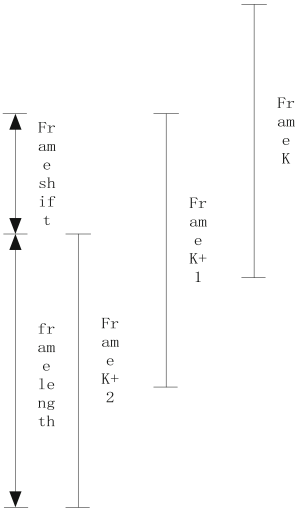

Fig. 1. Framing Diagram

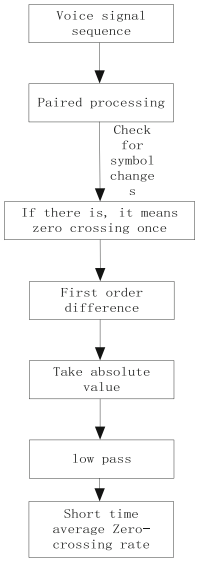

Fig. 2. Calculation of short-term average zero crossing rate

2.2 Speech Signal Denoising

A series of denoising processes are implemented for the voice signal of wireless communication network, and different speech preprocessing methods are adopted for different characteristics of noise.

First, build a noisy speech model:

$$x(t) = q(t) + w(t) \tag{3}$$

In formula (3) $x(t)$ is noisy speech; $q(t)$ is pure voice; $w(t)$ it is a noise signal.

In this model, the pure speech signal can be regarded as a speech segment, and its physical characteristics and spectrum characteristics are regarded as invariable. The short-term spectrum analysis of speech is relatively stable, which is a time-varying, non-stationary random process. The noise signal is an additive noise, which is locally stationary, statistically independent or uncorrelated with the speech signal, and has no other reference signal.

Spectral subtraction is an effective method to deal with single frequency noise, and is the simplest method of speech enhancement. The spectral subtraction method is improved to overcome the shortcomings of spectral subtraction in broadband noise processing.

The improved method of spectral subtraction algorithm: after spectral subtraction of noisy speech frames and pure noise frames, the residual noise will produce music noise. How to eliminate music noise is the key to improve spectral subtraction algorithm. The analysis shows that if the estimated frame of music noise can be obtained, music noise can be eliminated.

The method of obtaining music noise frame: subtract the square of the amplitude spectrum of any two pure noise frames, which can be used as the estimation value of the square of the amplitude spectrum of music noise. Subtract the square of the amplitude spectrum of the same frame of pure noise (standard pure noise) from the square of the amplitude spectrum of multiple frames of pure noise to obtain the estimated square of the amplitude spectrum of different music noises. The specific methods are as follows:

First spectral subtraction: subtract the square of the amplitude spectrum of the noisy speech from the square of the amplitude spectrum of the standard pure noise:

$$P = S(x(t))^2 - S(q(t))^2 \tag{4}$$

In formula (4) $S(x(t))$ it is the amplitude spectrum of noisy speech; $S(q(t))$ it is a pure speech amplitude spectrum.

Second spectrum subtraction: subtract the square of the amplitude spectrum of the first frame of music noise from the result of the first spectrum subtraction:

$$P' = P - S(f_1(t))^2 \tag{5}$$

In formula (5) $S(f_1(t))$ it is the amplitude spectrum of music noise in the first frame.

Third spectrum subtraction: subtract the square of the amplitude spectrum of the second frame of music noise from the result of the second spectrum subtraction:

$$P'' = P' - S(f_2(t))^2 \tag{6}$$

In formula (6) $S(f_2(t))$ it is the amplitude spectrum of music noise in the second frame.

By analogy, it is called cascading spectral subtraction.

The fluctuation noise is eliminated by the adaptive filter denoising method, and the wavelet threshold denoising algorithm is selected. The algorithm is mainly divided into four steps:

The first step is to select the wavelet basis function, and select a wavelet function to conduct discrete wavelet transform on the speech signal. Generally, db5 wavelet is selected as the wavelet function, which is decomposed into 3–5 layers of wavelet coefficients.

The second step is to determine the threshold value, which determines the final denoising effect of the denoising method [5]. If the threshold value is too small, the denoising effect will be unsatisfactory. If the threshold value is set too large, the useful part of the signal will be removed, resulting in the distortion of the denoised signal. The threshold selection method is as follows: the maximum variance minimizes the threshold to produce an extreme value of the minimum mean square error, which minimizes the maximum mean square error.

In the third step, threshold function is used to process the wavelet coefficients. The threshold function used is hard threshold:

$$l_i = \begin{cases} l_i & |l_i| \geq \gamma \\ 0 & |l_i| < \gamma \end{cases} \tag{7}$$

In Formula (7) l_i represents wavelet coefficients; γ is a hard threshold.

The fourth step is wavelet reconstruction, which uses discrete wavelet inverse transform to reconstruct the wavelet coefficients of threshold processing.

For impulse noise, the denoising method adopted is the optimized VMD decomposition and wavelet threshold denoising algorithm for speech signal denoising. Combining GWO optimized VMD algorithm with adaptive decomposition of signal and wavelet with the best approximation efficiency of one-dimensional signal, a speech signal denoising algorithm based on optimized VMD decomposition and wavelet threshold denoising is proposed. This method can eliminate the impulse noise interference of speech signal to a great extent.

When GWO algorithm is used to optimize VMD parameters, the fitness function uses permutation entropy. The complexity of the signal can be seen from the permutation entropy by calculating the fitness function. If the signal is more complex, the calculated permutation entropy is larger, and vice versa. After speech signal is decomposed by VMD, if there are many noise components included in IMF components, the higher the signal complexity, the greater the permutation entropy; If there are few noise components included in the IMF component, the more regular the signal is, the less complex the signal is, and the lower the permutation entropy is. Once the component K and penalty factor are determined α, VMD is used to decompose it, and the component with the smallest entropy value is the component with the best feature information of voice signal. Therefore, the objective of parameter optimization is to minimize the permutation entropy as the fitness value.

The specific implementation steps of the proposed optimized VMD decomposition and wavelet threshold denoising methods are as follows:

Input: voice signal with noise

Output: clean voice signal after denoising

Start.

Step 1: Use GWO algorithm to find the optimal combination of decomposition mode number and penalty factor for VMD to decompose noisy speech signal [k0, α0];

Step 2: Use the optimal combination found by the GWO algorithm to decompose the noisy voice signal into VMD, and obtain a finite number of modal components:

$$R = \{r_1, r_2, \ldots, r_k\} \tag{8}$$

In formula (8) r_k it refers to the No k Modal components.

Step 3: use correlation coefficient to select noisy modes from finite modal components;

3.1 Calculate the correlation coefficient between each modal component and the original signal;

3.2 Calculate the screening threshold through the screening principle of correlation coefficient f. When the modal component meets the following formula:

$$r_k > f \tag{9}$$

It can be considered that the correlation between the modal component and the original voice signal is good, and it needs to be retained; Otherwise, the corresponding modal components are taken out for wavelet denoising.

Step 4: carry out wavelet threshold processing for noisy modes;

4.1 Select appropriate wavelet bases, wavelet decomposition levels and wavelet thresholds. The wavelet base of sym8 is selected in this paper, and the number of decomposition layers is 6. Wavelet threshold uses the threshold of wavelet threshold de-noising algorithm;

4.2 Use the selected wavelet basis function to carry out wavelet transform on IMF components to obtain the corresponding.

A set of wavelet coefficients;

4.3 Compare the wavelet threshold with the wavelet coefficient obtained. If the threshold is less than the wavelet coefficient, it can be considered that the useful signal constitutes the wavelet coefficient, and then the wavelet coefficient needs to be retained; On the contrary, it is considered that the wavelet coefficient is composed of noise signals, so the wavelet coefficient needs to be discarded [6].

4.4 For the reserved signal, the denoised IMF component can be obtained only after reconstruction;

4.5 The selected IMF components must be operated from 4.2 to 4.4;

Step 5: Reconstruct the mode and effective mode after wavelet threshold processing to obtain the denoised speech signal.

End.

The specific process is shown in the following Fig. 3.

2.3 Feature Extraction

A speech signal feature extraction method based on improved EMD is designed to implement the feature extraction of voice signals in wireless communication networks.

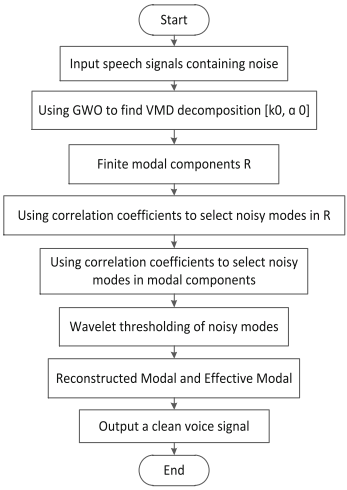

Fig. 3. VMD decomposition and wavelet threshold denoising flowchart

The traditional MFCC extraction algorithm is improved. The speech signal is decomposed by the improved empirical mode decomposition algorithm, and then the eigenmode components obtained by decomposing the speech signal are classified. The eigenmode components with a high proportion of the original signal are reconstructed and their MFCC is extracted. The proposed method saves as many components of the original signal as possible while eliminating noise signals. Since the noise of the signal can be decomposed into IMF components of each layer to enhance the signal-to-noise ratio of the weak signal layer, the improved method can be obtained as follows:

(1) EMD is decomposed into N IMF components and differentiated into reconstructed signals

The processed voice signal is decomposed into several IMF components by EMD method, as follows:

$$V = \{v_1, v_2, \ldots, v_j\} \tag{10}$$

In Eq. (10) v_k it refers to the No j IMF quantity.

Then autocorrelation processing is performed on the IMF components obtained. According to the autocorrelation function waveform, the dividing point between the useful original signal and the noise signal is derived, and the obtained eigenmode components are divided into two categories: the IMF component dominated by the noise signal and the IMF component dominated by the original signal. And reconstruct the main intrinsic mode components of the original signal to obtain the signal $E(t)'$.

(2) Fast Fourier Transform

The transformation of speech signal in the time domain can hardly see the characteristics of the signal. Usually, the speech signal is converted to the frequency domain to observe the characteristics of the speech signal. Therefore, the speech signal spectrum is obtained through fast FFT.

(3) Mel filter bank

The obtained voice signal is then passed through a set of Mel scale triangular filter banks, and the central frequency of the filter bank is $c(b)$ The transfer function satisfying each bandpass filter is:

$$c(b)=\begin{cases} 0 \; u < c(b-1) \\ \frac{u-c(b-1)}{c(b)-c(b-1)} \; c(b-1) \le u \le c(b) \\ \frac{c(b+1)-u}{c(b+1)-c(b)} \; c(b) \le u \le c(b+1) \\ 0 \; u > c(b-1) \end{cases} \tag{11}$$

The sampling rate is set to 8000 Hz. Here, the number of filters is set to M = 24.

(4) Take logarithmic energy

It is found that logarithmic energy is suitable for speech signal feature extraction. Compared with traditional entropy, logarithmic energy entropy can describe information more carefully. At present, most of the traditional entropy algorithms are used to describe the global information and column direction information. However, logarithmic energy can not only describe the overall global information of the signal. When describing the signal, the one-dimensional feature vector of the original signal can be changed into a single eigenvalue, so that the dimension of the original signal can be used to analyze and identify future information. Logarithmic energy is insensitive to a certain extent. In practice, logarithmic energy can be obtained by taking logarithm of the processed energy spectrum $o(m)$.

(5) DCT transformation

Discrete cosine transform (DCT) is a transform strongly related to Fourier transform. The discrete cosine transform is similar to the discrete Fourier transform. Because most of the collected information is noisy, if the data is not cleaned, direct prediction will produce data disturbance, greatly reducing the prediction accuracy. The discrete cotransform process uses only real numbers [7]. From the perspective of frequency domain, many natural signals are concentrated in the low-frequency part of the signal after discrete cosine transform. Convert logarithmic energy $o(m)$ MFCC coefficient is obtained through discrete cosine transformation:

$$\delta(z)=\sum_{m=0}^{M-1} o(m) COS \frac{\pi(m-0.5)}{L} \tag{12}$$

In Eq. (11) M refers to the number of signals;L it refers to the maximum value of the energy spectrum.

(6) Dynamic and static characteristics

The traditional cepstrum parameter MFCC can accurately reflect the static characteristics of speech signals, but the dynamic characteristics of parameters can not be well understood. Therefore, this paper proposes to obtain the dynamic characteristic parameters of MFCC from the difference of static MFCC, and then combine the dynamic

characteristic parameters and static characteristic parameters to form a mixed MFCC parameter as the training feature of speech signal.

2.4 Speech Recognition

A speech recognition model is designed based on the regression neural network in deep learning to realize the noise resistant speech recognition in wireless communication networks.

The designed speech recognition model includes three core parts: acoustic model, language model and decoder. In the DNN-HMM based acoustic model modeling, the DNN function is to replace the original GMM model and estimate the verification probability after HMM status.

The Kaldi open source speech recognition system is selected as the carrier to implement the DNN-HMM model. In Kaldi, the script is called quickly to make verification easier [8].

In the DNN-HMM acoustic model, the DNN network structure consists of one input layer, six hidden layers and one output layer.

For the input layer, 39 dimensional MFCC features are obtained by extracting voice information, 11 frames of voice information are used, and there are 429 input nodes. According to the number of triphone states corresponding to the target output of the network, that is, the number of clustered triphone state IDs, the output layer node is set to 1462 [9].

DNN network uses back-propagation algorithm to adjust parameters, so monitoring information is needed for training. For voice signals, it is necessary to know the phoneme state corresponding to each frame. After the forced alignment recognition result, the trisyllon state corresponding to the original voice information is obtained as the training supervision information [10].

The training criteria of the acoustic model should be able to be simply calculated and have a high correlation with the task. The improvement of the criteria should finally be reflected in the completion level of the task. Therefore, the minimum expected loss function should be selected as the training criterion of model parameters [11].

$$\theta_{EL} = U(\theta(\sigma, \varsigma O\rho)) \tag{13}$$

In Eq. (13) U is a statistical expectation operator; $\theta(\sigma, \varsigma 0\rho)$ is the loss function, where σ, ς is a model parameter, O to observe the vector, ρ is the corresponding output vector.

Given training criteria, model parameters σ, ς the famous error back propagation algorithm can be used for learning, and the chain rule can be used for deduction. In its simplest form, the model parameters are optimized using the first derivative information. The gradient of the top weight matrix relative to the training criteria depends on the training criteria [12].

The function of the decoder is to extract the decoding map from the trained model, and use this decoding map to search and match the test speech signal, and output the word sequence with the highest probability. The decoding diagram consists of four parts:

1) Grammar, which is the receiver of coding grammar or language model;
2) Phonetic dictionary, which is used to convert phonemes into words;
3) Context relation, which is used to output phonemes according to the most possible combination of phonemes in the context window;
4) HMM is defined to transform PDF ID (PDF index assigned by decision tree clustering) into phonemes representing context.

Firstly, context sensitive HMM acoustic model, ternary grammar language model and voice dictionary are integrated into a weighted finite state converter [13], which is optimized by deterministic algorithm and minimization algorithm to build a search space. Then send the test voice into the search space through the following four steps:

(1) Initialize search path:

$$\xi = \xi_0 \tag{14}$$

(2) Use acoustic model and language model to re judge the path score;

(3) Cut out the path with lower scores;

(4) The optimal path is obtained by backtracking, and the optimal result is obtained after the search.

The language model uses the pre training language model PLMs, which is composed of the Embedding layer, Pre training and Fine tuning. In the pre training process, a large number of unmarked corpus sets are used for unsupervised pre training, and then the weights in Fine tuning are initialized to the weights obtained in the pre training process [14]. All the pre training parameters will participate in the training when fine-tuning, and marked corpus is used in the training process. In different downstream tasks, their respective BERT models will be created as needed.

When using the Pre training Model for feature representation, there are generally two types of strategies: feature based strategy and fine-tuning based strategy. The traditional standard LM is the single way mode, which leads to many limitations when selecting language architecture [15]. BERT attempts to break through the one-way restrictions in the past standard language models by using the masked language model. The specific approach is to use [mask] to mask a random number of characters each time, and predict the words to be [mask] through the objective function. At the same time, BERT is also committed to the NSP task, that is, to determine whether the two sentences given are adjacent sentences in the text segment, so that the model can predict sentence level information [16].

The embedding layer is the embedding layer. BERT has three embedding layers, namely, Token Embedding, Segment Embedding and Position Embedding. The word elements in the input text complete vector conversion, sentence segmentation and word segmentation after three layers of embedding.

3 Experimental Test

3.1 Experimental Data Set

For the designed anti noise speech recognition method based on deep learning in wireless communication networks, its performance is tested through experimental data sets. The voice data, language model and dictionary used in this experiment are all from the "THCHS30 2015" dataset. This Chinese speech dataset contains 25 h of training data (30 speakers), 6 h of test data (10 speakers), and corresponding annotations. The voice data contains a total of 1000 sentences, and the coverage of binary phonemes and ternary phonemes reaches 71.5% and 14.3% respectively. All voice data are voice data under the wireless communication network, the sampling frequency is 16000 Hz, and the number of data bits is 16.

3.2 Experimental Process

First, the experimental data set is preprocessed by anti aliasing filtering, analog-to-digital transformation, pre emphasis, framing and windowing, endpoint detection, etc.

Then, a series of denoising processes are implemented on the experimental data set, and feature extraction is implemented through the speech signal feature extraction method based on improved EMD. In feature extraction, the order of Mel filter is defined as 24, the length of fast Fourier transform is 256, and the sampling frequency is 16000 Hz. Normalize Mel filter bank coefficients, divide speech signals into frames, and calculate MFCC parameters of each frame. Fast Fourier transform is performed to calculate first-order difference coefficient and second-order difference coefficient, and then MFCC parameters and first-order difference parameter MFCC are combined. After MFCC feature extraction, the relationship between the dimension and amplitude of the voice feature, and the relationship between the number of frames and amplitude can be drawn. After feature extraction, the voice feature parameter of one time pronunciation of each English number is 500×24.

Finally, through the speech recognition model based on regression neural network, the anti noise speech recognition of wireless communication network is realized. The script used for acoustic model training is shown in Table 1.

The word error rate and word error rate in the speech recognition of the design method are tested. In the test, the anti noise speech recognition technology based on RBF neural network and the anti noise speech recognition technology based on 3F speech enhancement distortion compensation are used as the comparison test methods, which are represented by technology 1 and technology 2 respectively.

3.3 Experimental Results

Table 2 shows the test results of word error rate in design method and technology 1 and technology 2 speech recognition.

According to the above table, the lowest word error rate of the design method is only 0.156, and the word error rate in speech recognition is far lower than that of technology

Table 1. Script used for acoustic model training

S/N	Script Name	Specific role
1	decode_fmllr.sh	Decoding the speaker adaptive model
2	train.sh	Training depth neural network model
3	pretrain_dbn.sh	Deep neural network pre training foot
4	align_si.sh	Align the specified data as input to the new model
5	decode.sh	Decode and generate word error rate results
6	mkgraph.sh	Establish identification network
7	train_sat.sh	Speaker adaptive training based on maximum likelihood linear regression in feature space
8	train_deltas.sh	Training a context sensitive three phoneme model
9	train_mono.sh	Training monophone hidden markov model

Table 2. Word error rate in speech recognition

Number of tests	Word error rate (%)		
	Design method	Technology 1	Technology 2
5	0.156	0.525	0.317
10	0.157	0.522	0.313
15	0.159	0.524	0.317
20	0.150	0.555	0.315
25	0.141	0.557	0.335
30	0.156	0.553	0.337
35	0.165	0.555	0.335
40	0.162	0.594	0.356
45	0.161	0.596	0.357
50	0.162	0.590	0.359

1 and technology 2, which shows that the design method has better noise resistance and more accurate speech recognition.

The test results of word error rate in the design method and technology 1 and technology 2 speech recognition are shown in Fig. 4.

According to the above figure, the word error rate in the design method speech recognition is also far lower than that in technology 1 and technology 2, which shows that its speech recognition performance is better. At the same time, it can be seen that the word error rate in speech recognition is higher than the word error rate as a whole.

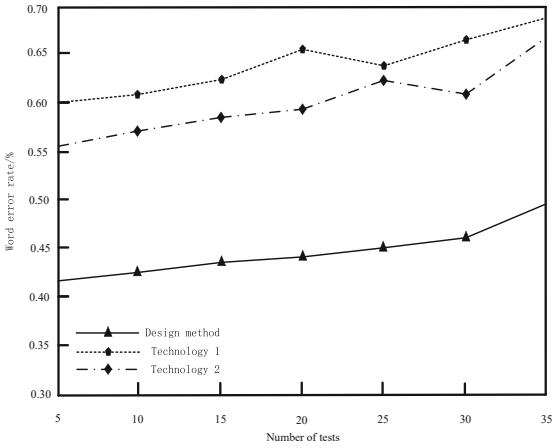

Fig. 4. Word Error Rate Test Results

To further validate the practicality of the design method, comparative tests were conducted using speech recognition time as the experimental indicator, and the test results are as follows (Fig. 5).

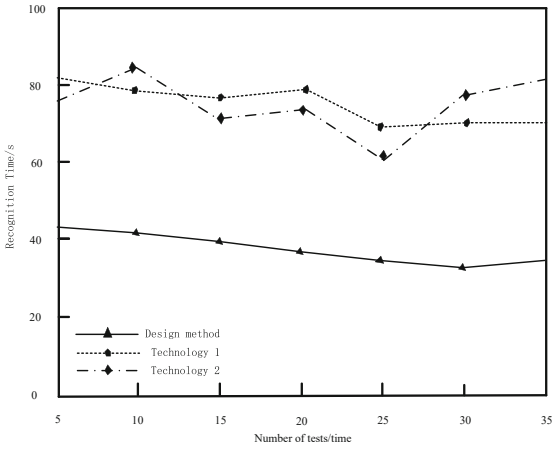

Fig. 5. Identify time test results

According to the above figure, it can be seen that the longest recognition time of the design method is 42 s, while the longest recognition times of the technology 1 and technology 2 methods are 81 s and 84 s, respectively. The recognition efficiency of the design method is significantly higher than that of the comparison method, indicating that the method proposed in this paper is practical.

4 Conclusion

The communication environment of wireless communication network will bring various types of noise to speech, so an anti noise speech recognition method based on deep learning of wireless communication network is designed to achieve anti noise speech recognition in this environment. In the wireless communication network, a series of preprocessing operations are carried out on the speech signal, and the speech signal features are extracted based on the improved EMD. According to the extracted feature signals, the speech recognition model is constructed by using the deep learning regression neural network, and the anti-noise speech recognition task in the wireless communication network is realized. However, the design method also has some limitations and some problems, and will continue to be optimized in the future.

References

1. Michelsanti, D., Tan, Z.H., Zhang, S.X., et al.: An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Trans. Audio Speech Lang. Process. **PP**(99), 1 (2021)
2. Goehring, T., Archer-Boyd, A.W., Arenberg, J.G., et al.: The effect of increased channel interaction on speech perception with cochlear implants. Sci. Rep. **11**(1), 1–9 (2021)
3. Datta, H., Hestvik, A., Vidal, N., et al.: Automaticity of speech processing in early bilingual adults and children– CORRIGENDUM. Bilingualism: Language and Cognition **24**(2), 1 (2021)
4. Kapnoula, E.C., Mcmurray, B.: Individual differences in speech perception: evidence for gradiency in the face of category-driven perceptual warping. J. Acoustical Soc. Am. **149**(4), A54–A54 (2021)
5. Kyaw, W.T., Sagisaka, Y.: Studies on association characteristics between vowels and visual colors using multiple speakers' speech. Acoust. Sci. Technol.. Sci. Technol. **42**(4), 161–169 (2021)
6. Abryutina, A., Ponomareva, A.: German-English Interference in the Field of Vocalism (Based on the Speech of Germans who Study English as a Foreign Language). Izvestia of Smolensk State University **1**(53), 128–143 (2021)
7. Qiang, H.: Consumption reduction solution of TV news broadcast system based on wireless communication network. Complexity **2021**(23), 1–13 (2021)
8. Dong, N., Lv, W., Zhu, S., et al.: Anti-noise model-free adaptive control and its application in the circulating fluidized bed boiler. Proc. Inst. Mech. Eng. Part I J. Syst. Control Eng. **235**(8), 1472–1481 (2021)
9. Wang, Q., Jiang, X., Weng, B., et al.: A 3D curvature attribute analysis method with excellent anti-noise property suitable for high steep formation. Geophys. Prospect. Petroleum **56**(4), 559–566 (2022)
10. Guan, Y., Hu, Z., Chen, C., et al.: An anti-noise transmission algorithm for 5G mobile data based on constellation selection and channel joint mapping. Alex. Eng. J. **60**(3), 3153–3160 (2021)
11. Basak, S., Agrawal, H., Jena, S., et al.: Challenges and limitations in speech recognition technology: a critical review of speech signal processing algorithms, tools and systems. Comput. Model. Eng. Sci. **2023**(5), 1053–1089 (2023)
12. Hadwan, M., Alsayadi, H.A., AL-Hagree, S.: An end-to-end transformer-based automatic speech recognition for qur'an reciters. Comput. Mater. Continua **2023**(2), 3471–3487 (2023)

13. El-Bialy, R., Chen, D., Fenghour, S., et al.: Developing phoneme-based lip-reading sentences system for silent speech recognition. CAAI Trans. Intell. Technol. **8**(1), 129–138 (2023)
14. Kamal, M.B., Khan, A.A., Khan, F.A., et al.: An innovative approach utilizing binary-view transformer for speech recognition task. Comput. Mater. Continua **2022**(9), 5547–5562 (2022)
15. Alsulami, N.H., Jamal, A.T., Elrefaei, L.A.: Deep learning-based approach for Arabic visual speech recognition. Comput. Mater. Continua **2022**(4), 85–108 (2022)
16. Nisar, S., Khan, M.A., Algarni, F., Wakeel, A., Irfan Uddin, M., Ullah, I.: Speech recognition-based automated visual acuity testing with adaptive mel filter bank. Comput. Mater. Continua **2022**(2), 2991–3004 (2022)