




## Optimization of Speech Recognition System of English Education Industry based on Machine Learning

Shuping Du<sup>1</sup> 

<sup>1</sup> Xingtai University, Xingtai, Hebei 054001, China, dushuping118@sina.com

Corresponding author: Shuping Du, dushuping118@sina.com

**Abstract.** At present, English speech recognition is still in the research stage, and most of the researches have certain deficiencies in terms of fault tolerance. Based on this, this study constructed an English speech recognition system based on machine learning from the perspective of machine learning. Simultaneously, in this paper, the English language model was optimized by the method of sub-word modeling, which alleviates the problem of sparseness and robustness of the traditional whole-word language model brought by the super-large vocabulary of adhesive words. In addition, a machine learning-based English speech recognition system was built, and the performance of the system was analyzed. The research shows that the proposed algorithm has certain performance and can provide theoretical reference for related research.

**Keywords:** machine learning; English phonetics; recognition; feature recognition

**DOI:** <https://doi.org/10.14733/cadaps.2020.S1.124-136>

### 1 INTRODUCTION

Chinese students mostly study Chinese for several years and they have a certain Chinese foundation before they begin to learn English. Therefore, English pronunciation is influenced by dialect and Mandarin. Students who have the correct fluent pronunciation when they first learn English are rare. In addition, there are some similarities between English and Chinese in pronunciation. Therefore, whether the standard of the student's Mandarin will directly affect whether it can have accurate English pronunciation [1].

All languages in the world consist of three elements: vocabulary, grammar, and speech. As a tool for conversation, the language is first and foremost, that is, the voice is the first. The purpose of learning English is to use English as a tool for conversation, to accurately express one's thoughts and to understand others' thoughts. Standard speech is fundamental to spoken expression and listening comprehension. Only English with a good voice system can learn English well [2].

The smallest unit of speech that is based on the natural nature of the speech is the phoneme. The phoneme is acoustically the smallest unit of speech divided from the perspective of sound quality. In terms of physiological properties, a phoneme is formed by a pronunciation action. For example: /eg/ contains two pronunciations of /e/ and /g/, which are two phonemes. The same phoneme is the sound formed by the same pronunciation action, and the different phonemes are

the sounds formed by different pronunciation actions. For example, the two /d/ pronunciations in /hænd/ and /bænd/ are the same, so they are the same phonemes; the /b/ and /z/ are different in pronunciation, so they are different phonemes. The analysis of phonemes is generally based on pronunciation actions. For example, the pronunciation action of /m/ is: the upper lip and the lower lip are closed, the vocal cords vibrate, and the airflow rushes out of the nasal cavity, which is classified as a lip nasal [3].

The symbol used to record phonemes in phonetics is called the phonetic symbol, which is the phonetic symbol. For example, international phonetic symbols are generally marked with "/". Its formulation principle is: "A phoneme is represented by only one phonetic symbol, and a phonetic symbol represents only one phoneme. The International Phonetic Alphabet (IPA) is developed by the International Phonetic Association and strictly stipulates the principle of 'one note and one tone', that is, 'one phoneme corresponds to one symbol and one symbol corresponds to one phoneme'. It is based on the Latin alphabet and is supplemented by a method of changing glyphs and borrowing letters from other languages "[4]. In order to take care of the habit, most symbols on the pronunciation still retain the original sound of Latin or other languages.

It is generally believed by phonetics that speech can be divided into two categories: vowels and consonants, depending on whether the airflow is blocked by the vocal organs when exhaled from the lungs. When speaking, the unobstructed phoneme of the airflow is a vowel, and the phoneme whose airflow is obstructed is a consonant. The difference between a vowel and a consonant is that the airflow is unobstructed in the channel when the vowel is emitted, and the airflow is hindered to varying degrees on the channel when the consonant is emitted, that is, the airflow is closed or the airflow is narrowed. In addition, we can distinguish vowels and consonants in other ways: 1) Airflow. When the vowel is pronounced, the airflow is weaker. In contrast, the airflow is stronger when the consonant is emitted, especially when the Voiceless consonant is emitted. 2) The tension of the articulated organ. When vowels are pronounced, the vocal organ tension is average. However, when a consonant is made, the tension of the vocal organ occurs at a moment of the pronunciation and occurs at a certain part of the utterance, while the other vocal organs do not exhibit a state of tension. 3) Sound. From the perspective of acoustic phonetics, vowels have greater loudness and intensity than consonants. 4) Music and noise [5].

Domestic research on the theory of mother tongue migration has been more than 10 years later than abroad. In the 1950s, the domestic English education community began to study how Chinese dialects affect the pronunciation of English learners. The overall research in China has shown a clear upward trend, and the results are relatively fruitful. This study attempted to search for papers on the influence of Chinese dialects on English pronunciation since the 1950s and summarized the current research status of dialects on English pronunciation. This study uses the "dialects", "English pronunciation" and "migration" as the theme to search the research results of China Knowledge Network and Wanfang Data Knowledge Service Platform from 1957 to 2016 and obtained more than 300 related papers [6].

In general, the research and development in this field can be roughly divided into two stages: the first stage is from 1957 to 1999. At this stage, the research results are few and the growth is slow. There are only 12 related academic papers published in the past 42 years, including 8 core journal papers and 4 general journal papers, but there are no master's thesis and general journal papers. Fang Shuzhen's "Comparative Analysis of English and Guangzhou Dialects" published in "Western Languages" in 1957 is the first paper in this field. The paper makes a detailed comparative analysis of the pronunciation methods and phonemes of the two phonetic systems of Guangzhou dialect and English [7].

Singh S summed up the prominent dialect of Henan Province and first discussed the influence of Hefang dialect on English phonetic learning, and investigated the English pronunciation of students in Xiang dialect area and analyzed the reasons for the mistakes [8]. The research at this stage is more general and there is no relevant literature on primary and secondary schools.

Gundogdu K investigated the English pronunciation of 30 Chinese students studying in the UK and found that people who confuse the sidetones l and nasals n in Chinese dialects cannot grasp

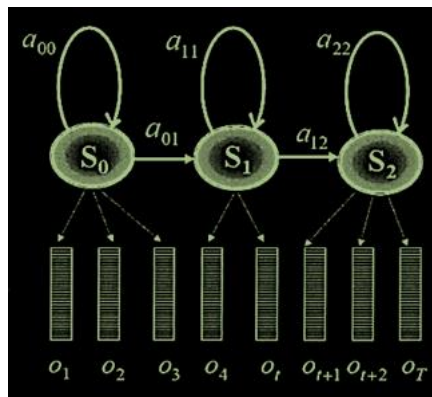
the correct pronunciation of /l/ and /n/ in English [9]. Calder L A investigated and analyzed the English pronunciation errors of middle school students in Changsha dialect area and compiled a correction manual [10]. Zhang L compared the differences between the dialects of Jinan, Chengdu and Pingxiang in the dialects and English, and revealed the influence of dialects on English pronunciation, and gave suggestions for error correction [11]. Pal M analyzed the similarities and differences between Benxi dialect, Mandarin Chinese and English, and obtained the influence of Benxi dialect on English phonetic acquisition [12]. Cesari U explored the influence of dialects in the Ordos region on English phonetics learning for junior high school students [13]. BOCCHIERI ENRICO studied the negative migration of Qingyu dialect to the English pronunciation of 6 students in the Senior grade one and Sophomore year of Qingyi Middle School [14].

Through the above analysis, it can be seen that the current English speech recognition is still in the research stage, and most of the research has certain deficiencies in the fault tolerance rate. Based on this, this study constructed an English speech recognition system based on machine learning from the perspective of machine learning and studied and analyzed the performance of the system.

## 2 HIDDEN MARKOV (HMM) ACOUSTIC MODELING

The conversion formula for the Mel domain frequency and the ordinary linear frequency is:

$$f_{mel} = 2959 \log_{10} \left( 1 + \frac{f}{700} \right) \tag{1}$$



**Figure 1:** Schematic diagram of the first-order HMM acoustic model.

As shown in Figure 1, a typical HMM can be mathematically described using five sets of parameters, namely [15]:

$$M = \{O, \Omega, \pi, A, B\} \tag{2}$$

In equation (2),  $O$  represents the sequence of observation vectors  $\{o_1, o_2, \dots, o_T\}$  from time 1 to time  $T$ , and  $\Omega$  represents the set  $\{s_1, s_2, \dots, s_K\}$  of finite implicit state sequences contained in the  $K$  HMMs. The model parameters that need to be determined in the HMM are represented by three sets of parameters:  $\lambda = \{\pi, A, B\}$ .  $\pi = \{\pi_1, \pi_2, \dots, \pi_K\}$  denotes the distribution of  $K$  states in which the HMM is at the starting time,  $A = \{a_{ij}\}_{K \times K}$  denotes the state transition probability

matrix, and B denotes the probability distribution function  $\{b_i(o)\}$  under different states. In speech recognition systems, the state output density function is usually characterized by Gaussian Mixture Model (GMM), that is [16]:

$$b_i(o_t) = \sum_{k=1}^k w_{ik} \cdot \frac{1}{\sqrt{(2\pi)^D |\Sigma_{ik}|}} \exp\left[-\frac{1}{2}(o_t - \mu_{ik})\right] \quad (3)$$

In the figure,  $a_{ij}$  represents the transition probability between states jumping from state i to state j.  $b_i(o_t)$  represents the probability of outputting the observation vector  $O_t$  when jumping from state i.

The parameters of the HMM must satisfy the following conditions [17]:

$$\begin{aligned} \pi_i \geq 0, a_{ij} \geq 0, b_i(o_t) \geq 0 \\ \sum_{i=1}^N \pi_i = 1 \quad \sum_{j=1}^N a_{ij} = 1 \quad \int b_i(o_t) do_t = 1 \end{aligned} \quad (4)$$

Considering the practical application of HMM, HMM should have two important assumptions: One is the first-order Markov hypothesis, that is, the state  $s_t$  of the current time t is only related to the state  $s_{t-1}$  at the previous moment  $t-1$  and has nothing to do with any state at any other time. The formula is expressed as:

$$p(s_t | s_1^{t-1}) = p(s_t | s_{t-1}) \quad (5)$$

The other is the output-independent hypothesis, which means that the output value at the current time is only governed by the probability density of the current state and is independent of other output values and states that have already been generated. The formula is expressed as [18]:

$$p(x_t | x_1^{t-1}, x_1^t) = p(x_t | s_t) \quad (6)$$

In order for HMM to better serve the speech recognition system, there are three classic problems that we need to solve:

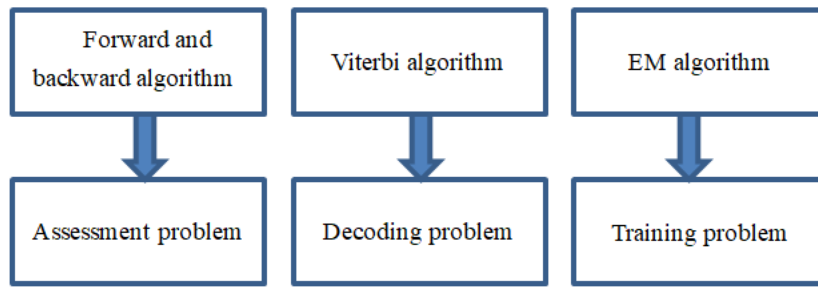
(1) Probability calculation problem: In the case of given model parameters, how the probability  $p(O | \lambda_{mm})$  of the observed vector O appears is calculated.

(2) Code problem: Under the premise of given observation vector and model parameters, how optimal the state sequence Q is calculated, so that the joint probability  $p(O, Q | \lambda_{mm})$  is the largest.

(3) Estimated problem of model parameter  $\lambda_{mm}$ : After giving enough observation vector sequences, how can we estimate the model parameters  $\lambda_{mm}$  using the existing data so that the estimated hidden Markov model is the most probable model for generating a given observation vector.

According to the HMM, any state sequence S may generate the observation vector O with a certain probability, so  $p(O | \lambda_{mm})$  should be the cumulative sum of the probabilities corresponding to each possible state sequence, that is [19]:

$$p(O | \lambda) = \sum_s p(O, s | \lambda) = \sum_s p(s | \lambda) p(O, s | \lambda) \quad (7)$$



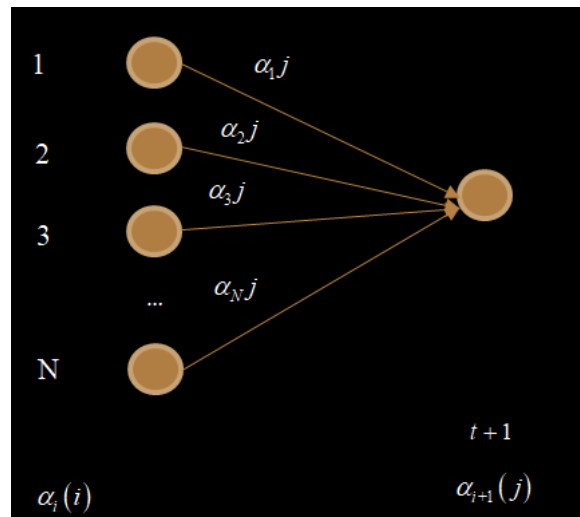
**Figure 2:** Corresponding image of the problem and the solution algorithm.

Then, for an HMM with a total of  $M$  states, if the length of time is  $N$ , the number of all possible state sequences is  $M^N$ . Moreover, the computational complexity increases exponentially as the number of states and the length of time increase, which is unacceptable. The algorithm that effectively solves this complexity problem is the Forward-Backward Algorithm, which will be discussed below.

Definition of forward probability:

$$\alpha_i(i) = p(o_1, o_2, \dots, o_i, s_i | \lambda) \tag{8}$$

Its meaning is that the output vector at time 1 to  $t$  is  $o_1, o_2, \dots, o_t$  and the state at time  $t$  is the probability of  $i$ .



**Figure 3:** Diagram of forward probability.

First, it is initialized:

$$\alpha_1(1) = 1 \quad \alpha_1(j) = 0 (j \neq 1) \tag{9}$$

Second, the recursive is taken:

$$\alpha_{t+1}(j) = \left[ \sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}) \quad 1 \leq i \leq N \quad 1 \leq t \leq T-1 \tag{10}$$

Finally, the output probability is obtained:

$$p(O|\lambda) = \sum_{i=1}^N p(o_1, o_2, \dots, o_T, s_T = s_i | \lambda) = \sum_{i=1}^N \alpha_T(i) \quad (11)$$

Definition of backward probability:

$$\beta_t(i) = p(o_{t+1}, o_{t+2}, \dots, o_T | s_t = s_i, \lambda) \quad (12)$$

Its meaning is the probability that the output vector is  $o_{t+1}, o_{t+2}, \dots, o_T$  from  $t+1$  to  $T$  under the premise that the state at time  $t$  is  $i$ .

First, it is initialized:

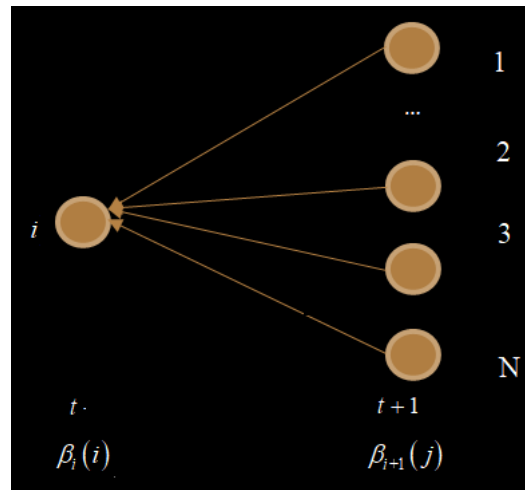
$$\beta_T(N) = 1 \quad \beta_T(j) = 0 (j \neq N) \quad (13)$$

Second, the recursive is taken:

$$\beta_t(j) = \sum_{i=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(i) \quad 1 \leq i \leq N \quad 1 \leq t \leq T-1 \quad (14)$$

Finally, the output probability is obtained:

$$p(O|\lambda) = \sum_{i=1}^N p(o_1, o_2, \dots, o_T | s_1 = s_i, \lambda) = \sum_{i=1}^N \beta_1(i) \quad (15)$$



**Figure 4:** Diagram of backward probability.

The complexity of the calculation can be greatly reduced by the forward and backward algorithm.

In the case where the observation vector  $O$  and the hidden Markov model parameter  $\lambda$  are given, the decoding problem refers to how to determine the most likely state sequence of the output observation vector  $O$  from the model parameters  $\lambda$ . Specifically, if an HMM has  $M$  states and the time length is  $N$ , the number of all possible state sequences is  $M^N$ . Then, the decoding problem is to determine the most likely sequence of states from which the observation vector  $O$  is generated from the sequence of states of the  $M^N$  number.

The formula is described as:

$$s^* = \arg \max_s p(O, s | \lambda) \quad (16)$$

Then, this problem can be solved by the Viterbi algorithm. First, we need to define the function:

$$\delta_t(i) = \max_{s_1, s_2, \dots, s_{t-1}} p(s_1, s_2, \dots, s_{t-1}, s_t = i, o_1, o_2, \dots, o_t | \lambda) \quad (17)$$

The above equation shows the probability of the maximum probability path with state  $i$  at time  $t$ .

First, it is initialized:

$$\delta_1(i) = \pi_i b_i(o_1) \quad (18)$$

Second, the recursive is taken:

$$\delta_t(i) = \max_{1 \leq j \leq N} [\delta_{t-1}(j) a_{ji}] b_j(o_t) \quad 2 \leq t \leq T \quad 1 \leq j \leq N \quad (19)$$

Finally, the optimal sequence output probability is determined:

$$p^{\max} = \max_{1 \leq i \leq N} [\delta_T(i)] \quad (20)$$

The biggest difference between the Viterbi algorithm and the forward algorithm is the difference between the summation and the maximum value. The forward algorithm is summation, the Viterbi algorithm is to find the maximum value, and the output probability of the general optimal path will account for more than 99.5% of the output probability of all paths. The training problem is how to obtain the specific value of the HMM model parameters through a better algorithm when the

observation sequence  $O$  is obtained. Moreover, that the likelihood probability  $p(O|M)$  of the observation vector  $O$  is largest under this model parameter is satisfied.

Derivation of the EM algorithm:

(1) Define the function:

$$\zeta_t(i, j) = p(s_t = s_i, s_{t+1} = s_j | O, M) = \frac{a_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N a_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \quad (21)$$

It refers to the probability that the time  $t$  is the state  $i$  and the time  $t+1$  is the state  $j$  Under the premise that the model input parameters and observation vectors are given.

(2) Define the function:

$$\gamma_t(i) = \sum_{j=1}^N \zeta_t(i, j) \quad (22)$$

It represents the probability that state  $t$  is the state  $i$  when the model parameters and the observation vector are given.

(3) Define the function:

$$\sum_{t=1}^{T-1} \gamma_t(i) \quad (23)$$

It indicates the probability that the vector is in state  $i$  in the case where the model parameters and the observation vector are given.

(4) Define the function:

$$\sum_{t=1}^{T-1} \zeta_t(i, j) \quad (24)$$

It represents the probability of jumping from state  $i$  to state  $j$  in the case where model parameters and observation vectors are given.

(5) Under the above definition, the model parameters are updated as follows:

$$\pi_i = \gamma_t(i) \quad (25)$$

The above equation represents the probability of being in state  $i$  at time  $t=1$ .

$$a_{ij} = \frac{\text{Probability of state } i \text{ transitioning to state } j}{\text{Probability of state } i} = \frac{\sum_{t=1}^{T-1} \zeta_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (26)$$

The above formula represents the weight update between nodes.

$$\hat{b}_{ij}(o_k) = \frac{\text{Probability of sample } o_k \text{ in } j \text{ state}}{\text{Probability of state } j} = \frac{\sum_{\substack{t=1 \\ o_t=o_k}}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(i)} \quad (27)$$

The above equation represents the likelihood probability update of the  $j$  state.

Assuming  $M = \{A, B, \pi\}$  is the original model parameter and  $\hat{M} = \{A, B, \hat{\pi}\}$  is the model parameter after revaluation, then the following conclusions can be proved:

1. When the model is already optimal, there must be:

$$M = \hat{M} \quad (28)$$

2. The parameters of the new model make:

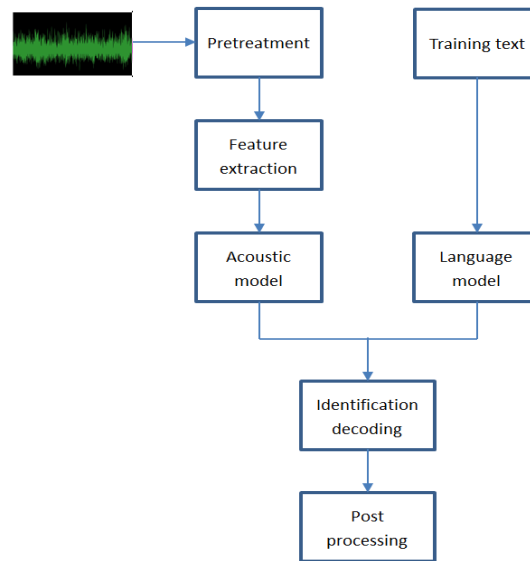
$$P(\hat{M}/M) > P(O/M) \quad (29)$$

It can be seen that as the iteration continues,  $\hat{M}$  will converge to the optimal parameters.

### 3 MODEL BUILDING

The data input to the speech feature extraction module according to FIG. 5 is waveform data obtained by pre-processing the simulated speech signal, and then it is sent to the feature extraction module. There are various differences between different speakers, such as age, gender, pronunciation habits and feelings, which will change with time, so people will express different voice signals when expressing the same content. Moreover, to a large extent, acoustic features represent speech signals. Good acoustic characteristics should try to meet the following three conditions: First, the acoustic characteristics should be well differentiated, the differences between the similarities should be as small as possible, and the differences between different classes should be as large as possible, which makes it easier to identify different information, and also facilitates more accurate modeling of different acoustic modeling units. Secondly, the extraction of speech features can be regarded as the compression coding process of speech information, and it is necessary to retain the information related to the speech content and eliminate other factors that are not closely related to the content. Moreover, it is necessary to reduce the dimension of the parameter when the information is retained enough, that is, the feature dimension should be moderate, so as to accurately and efficiently enter the training of the acoustic model. Finally, reliability and independence need to be considered, and there must be anti-interference ability against environmental noise.





**Figure 5:** System framework of speech recognition.

### (1) Spectrogram

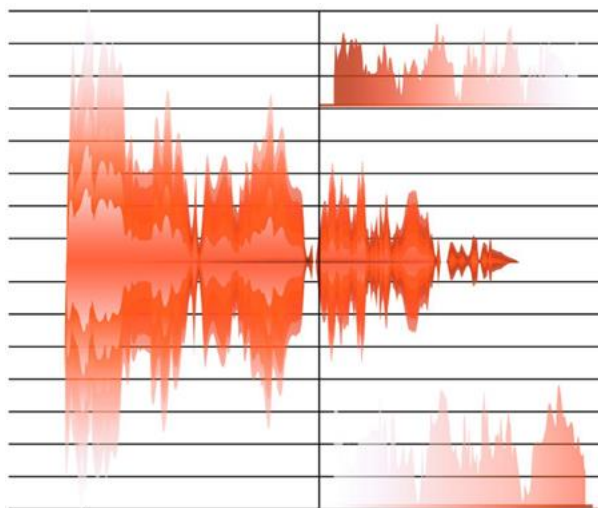
For a speech signal, it is usually analyzed from the time domain or the frequency domain. However, both methods have limitations: Time domain analysis does not have an intuitive understanding of the frequency characteristics of speech signals, and frequency domain analysis does not have a relationship of speech over time.

The spectrogram is a map that expresses three-dimensional information in a two-dimensional plane. Its abscissa indicates time, the vertical scale indicates frequency, and the gray value of each pixel reflects the energy of the corresponding time and corresponding frequency. Through the spectrogram, the properties of the phoneme can be well observed. The formant (the thicker bar in the spectrogram) carries the distinguishability of the speech, and the speech can be better recognized by observing the variation of the formant.

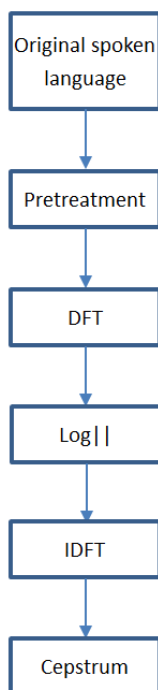
### (2) Feature extraction

Commonly used for speech recognition features include: Linear Prediction Coefficients (LPC), Mel-Frequency Cepstral Coefficients (MFCC), and Perceptual Linear Prediction (PLP). Among them, linear predictive analysis means that the speech signal at a certain moment can be linearly represented by the combination of signals at several previous moments. The basic problem is to directly derive a set of linear prediction coefficients based on the speech signal. When the mean square error (MSE) between the linear prediction estimate and the sampled value of the speech signal reaches a minimum value, the linear prediction coefficient can be extracted. The most important speech feature parameters are feature parameter extraction based on cepstrum analysis, and the cepstral coefficients are implemented based on the homomorphic processing method, and the homomorphic processing method is a processing method that transforms the nonlinear problem into a linear problem. First, the original speech signal (actually a convolutional signal) is subjected to Discrete Fourier Transform (DFT) to obtain the spectrum (at this time, it becomes a multiplicative signal, and the convolution of the time domain is equivalent to the product of the frequency domain.). The discrete spectrum then takes the logarithm to turn the multiplicative signal into an additive signal. Finally, it is restored to a convolution signal by Inverse

Discrete Fourier Transform (IDFT) to obtain cepstral coefficients. This method of finding cepstral coefficients can obtain relatively stable speech feature parameters.



**Figure 6:** Spectrogram corresponding to the time domain signal waveform.



**Figure 7:** Flow chart for calculating the cepstrum coefficient.

## 4 ANALYSIS AND DISCUSSION

The Language Model (LM) is a customary way of describing human language. It mainly reflects the intrinsic relationship between words and words in the organizational structure, and the language model determines which word sequence is more likely, and several words are known to predict the next word. The language model that can accurately describe the law of language change directly affects the efficiency and performance of decoding, which is directly related to the overall performance of speech recognition. According to different production methods, language models can be summarized into two categories, namely, statistical-based language models and rule-based language models. The statistical-based language model is obtained by training a large amount of training data. It describes the language model from the mathematical point of view and can calculate the probability of occurrence of each sentence in the original language. Rule-based language models require linguistic scholars and experts based on linguistically relevant knowledge and are artificially compiled in conjunction with real-life language situations. However, this language model has limitations in dealing with large-scale real text, so it is generally not used.

Choosing the right acoustic model modeling unit is the key and the first problem encountered in acoustic modeling. A suitable granularity modeling unit can be of great help to the performance improvement of the speech recognition system. A good acoustic modeling unit should have the following three attributes: consistency, trainability, and sharing. Consistency means that the same modeling unit in different speech instances requires the basic consistency of acoustic pronunciation. Trainability means that each modeling unit can correspond to enough modeling instances. Sharing means that the practice of different modeling units can share common training examples.

Acoustic modeling units commonly used in speech recognition have syllables, phones, and tri-phones. As a modeling unit, the consistency of syllables is weak, but its training is strong, and it is suitable for application scenarios of digital string recognition. As a modeling unit, the consistency of phonemes is general, but its trainability is relatively strong and it is suitable for isolated word recognition application scenarios. The ternary phoneme is very consistent as a modeling unit, but it is weakly trainable and suitable for large vocabulary and large-scale speech recognition applications. In the speech recognition, the phenomenon of co-pronunciation is considered, that is, each pronunciation may be distorted by the influence of adjacent sounds, and the context-dependent acoustic unit is usually selected as the modeling unit of the acoustic model. If only the influence of the previous note on the current note is considered, it is called Bi-phone. If both the previous note and the influence of the next note on the current note are considered, it is called Tri-phone. In this paper, context-sensitive triphones are used as the modeling unit in English speech recognition.

Unlike cepstral properties and linear prediction, the Mel cepstral coefficients and perceptual linear predictions are based, to some extent, on the mechanism of human auditory perception. 提取 The MFCC process is to first convert the signal from the time domain to the frequency domain by FFT, and then convolve the logarithmic energy spectrum with a set of triangular filters uniform in the Mel frequency domain. Finally, the output of the filter bank is converted by the discrete cosine transform method, and then some coefficients after DCT are taken as MFCC. At this time, a series of cepstrum vectors can be used to describe the speech, and each vector is the MFCC feature vector of each frame. Numerous studies have shown that MFCC parameters are superior to other parameters for the performance improvement of speech recognition systems.

## 5 CONCLUSION

Speech recognition is a very representative cutting-edge technology in the field of artificial intelligence, which is directly related to the future life experience of our human beings. The combination of deep learning and speech recognition will definitely push artificial intelligence forward. This study constructed an English speech recognition system based on machine learning from the perspective of machine learning, and studied and analyzed the performance of the

system, which will promote the further development of artificial intelligence. At the same time, the subject was based on deep learning, and firstly studied the acoustic modeling based on DNN, and introduced the network structure and algorithm of DNN. Then, the DNN-HMM-based acoustic model was trained with 300 hours and 500 hours of English speech data, respectively. In addition, this paper optimized the English language model by sub-word modeling method, which alleviates the problem of sparseness and robustness of the traditional whole-word language model brought by the super-large vocabulary of adhesive words.

## 6 ORCID

Shuping Du, <https://orcid.org/0000-0003-1781-4003>

## REFERENCES

- [1] Simmons, S. M.; Caird, J. K.; Steel, P.: A meta-analysis of in-vehicle and nomadic voice-recognition system interaction and driving performance, *Accident Analysis & Prevention*, 106, 2017, 31-43, <https://doi.org/10.1016/j.aap.2017.05.013>
- [2] Xiaojun, Z.; et al: Pathological voice source analysis system using a flow waveform-matched biomechanical model, *Applied Bionics and Biomechanics*, 2018, 2018:1-13, <https://doi.org/10.1155/2018/3158439>
- [3] Enajeh, S. M. A.; Cavus, N.; Ibrahim, D.: Development of a voice recognition-based system to help physically disabled people use the Facebook, *Quality & Quantity*, 52(2), 2018, 1343-1352, <https://doi.org/10.1007/s11135-018-0709-6>
- [4] Swanepoel, L.; Van den Heever, D.; Dellimore, K.: Development of a gesture and voice controlled system for burn injury prevention in individuals with disabilities, *Conf Proc IEEE Eng Med Biol Soc*, 2017, 3429-3432, <https://doi.org/10.1109/EMBC.2017.8037593>
- [5] Xiang, J.; Patrick, P. L.: Development and application of a classification system for voice intelligent agents, *International Journal of Human – Computer Interaction*, 2018, 1-9, <https://doi.org/10.1080/10447318.2018.1496969>
- [6] Lu, Z.; et al: Design of voice interaction-based training system for cerebral palsy rehabilitation, *Chinese High Technology Letters*, 2017, <https://doi.org/10.3772/j.issn.1002-0470.2017.03.011>
- [7] Köster, F.; et al: Towards degradation decomposition for voice communication system assessment, *Quality & User Experience*, 2(1), 2017, 4, <https://doi.org/10.1007/s41233-017-0006-5>
- [8] Singh, S.; et al: Information communication technology for extension: a mobile phone based voice call system for dissemination of cotton production technologies, *Journal of Agricultural & Food Information*, 2018, 1-9, <https://doi.org/10.1080/10496505.2018.1436442>
- [9] Gundogdu, K.; Bayrakdar, S.; Yucedag, I.: Developing and modeling of voice control system for prosthetic robot arm in medical systems, *Journal of King Saud University - Computer and Information Sciences*, 2017, S1319157817300216, <https://doi.org/10.1016/j.jksuci.2017.04.005>
- [10] Calder, L. A.; et al: The feasibility of an interactive voice response system (IVRS) for monitoring patient safety after discharge from the ED, *Emergency Medicine Journal*, 2017, emermed-2016-206192, <https://doi.org/10.1136/emered-2016-206192>
- [11] Zhang, L.; Gong, L.: Design of real-time voice over internet protocol system under bandwidth network, *5th International Conference on Computer-Aided Design, Manufacturing, Modeling and Simulation (CDMMS 2017) - Busan, South Korea (22–23 April 2017)*, 2017, 1834:040026, <https://doi.org/10.1063/1.4981622>
- [12] Pal, M.; Saha, G.: Spectral mapping using prior re-estimation of i-vectors and system fusion for voice conversion, *IEEE/ACM Transactions on Audio Speech & Language Processing*, PP(99), 2017, 1-1, <https://doi.org/10.1109/TASLP.2017.2743620>

- [13] Cesari, U.; et al: Voice disorder detection via an m-health system: design and results of a clinical study to evaluate Vox4Health, BioMed Research International, 2018, 1-19, <https://doi.org/10.1155/2018/8193694>
- [14] Bocchieri, E.; Caseiro, D. A.; Dimitriadis, D.; System and method for speech recognition modeling for mobile voice search, Jersey Citynj Uspphiladelphiapa Uschathamnj Us, 47(10), 2017, 4888 - 4891, <https://doi.org/US20120232902>
- [15] Culbertson, G.; et al: Facilitating development of pragmatic competence through a voice-driven video learning interface, CHI Conference on Human Factors in Computing Systems, ACM, 2017, 1431-1440, <https://doi.org/10.1145/3025453.3025805>
- [16] Hasan, M. H.; Ahlstedt, M.; Yihun, Y.: Design of inverse dynamics tracking control system and voice-based interfacing mechanism to the WMRA, International Journal of Dynamics and Control, 2017, <https://doi.org/10.1007/s40435-017-0363-1>
- [17] Merks, I.: Hearing assistance system with own voice detection, Journal of the Acoustical Society of America, 2017, <https://doi.org/10.1121/1.4861509>
- [18] Tabibian, S.: A voice command detection system for aerospace applications, International Journal of Speech Technology, 20(4), 2017, 1049-1061, <https://doi.org/10.1007/s10772-017-9467-4>
- [19] Winda, A.; et al: Motorcycle start-stop system based on intelligent biometric voice recognition, IOP Conference Series: Materials Science and Engineering, 187(1), 2017, 012039, <https://doi.org/10.1088/1757-899X/187/1/012039>