Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 7180769, 14 pages
https://doi.org/10.1155/2021/7180769

Research Article

Study about Chinese Speech Synthesis Algorithm and Acoustic Model Based on Wireless Communication Network

Liu Shi,[1] Moyan Li,[2] Yawen Su,[3] and Yi Chen[4]

[1]*The Education University of Hong Kong, Hong Kong 999077, China*
[2]*School of Humanistic and Social Sciences, Shantou Polytechnic, 515000 Guangdong, China*
[3]*Teachers College, Jimei University, 361000 Xiamen, Fujian, China*
[4]*Skill Training Center, Shantou Polytechnic, 515000 Guangdong, China*

Correspondence should be addressed to Yawen Su; s1110624@s.eduhk.hk

Received 3 August 2021; Revised 18 September 2021; Accepted 21 September 2021; Published 4 October 2021

Academic Editor: Zhihan Lv

Chinese speech synthesis refers to the technology that machines transform human speech signals into corresponding texts or commands through recognition and understanding. This paper combines the classic VAD and GSM VAD1 algorithm simulations, improves on the above two algorithms to recognize and collect speech, and analyzes their Chinese proficiency by amplifying the signal through a filter, in order to study the adulthood of Zhengzhou University in Southeast Asian students (mother tongues are Indonesian and Thai) as the research objects, to explore the relationship between the Chinese phonetic proficiency and the acquisition motivation of Southeast Asian students. This article combines algorithm and language disciplines. According to the results of Praat and SPSS: 55-80 points account for 70%, 55 points below 20% and 80 points above 10%, we find that intrinsic motivation plays a role in CSL acquisition, a vital role. Intrinsic motivation can help mature learners from Southeast Asia to acquire Chinese better and better. The earlier you learn Chinese, the higher your motivation, and the easier it is to set your Chinese learning goals. The greater the enthusiasm for learning Chinese, the better the Chinese scores (such as HSK test scores and Chinese phonetic test scores). Therefore, the Chinese proficiency of international students has a great relationship with their interest in Chinese language, that is, the greater the interest in Chinese, the stronger their motivation to learn, and the Chinese proficiency will be very good.

1. Introduction

Language is an important tool for humans to acquire knowledge, learn, and express thoughts and emotions. Language is one of the most important and effective forms of information transmission for human beings. Language and voice are also related to people's cognitive activities, symbolizing the culture of a social country. Therefore, the study of language and phonetics is of great significance to the progress of science and the development of society. As people enter the information age, modern technologies and methods have emerged at a rapid pace, enabling them to receive, store, and process audio data more efficiently and quickly and speed up the research of audio processing technology. Text-to-speech technology is a process of processing text-like input signal sequences through a dedicated synthesizer to create natural, high-quality, and high-quality audio output. Speech synthesis acts as a machine to output human-machine voice interaction. Disciplines are related to linguistics, computer science, neuroscience, computer science, psychology, and many other subjects.

Voice and text search engines are now widely used, and image extraction applications are being tested in different search engines, but audio extraction is still in its infancy. With the widespread adoption of voice assistants, such as Apple's voice assistant Siri and Microsoft's voice assistant Xiaona, and the success of iFLYTEK's voice input method, voice recognition is gradually developing. It is widely used in business, but there are many problems with using traditional sound. Keyword recognition: first, convert all audio data into text and then perform keyword recognition. With the increase of audio data, the workload is large and the

memory efficiency is reduced. At present, the problem of multilanguage integration is not solved. Vocabulary (OOV) and resource language are excluded. Poor recognition ability: therefore, more powerful multimedia recognition technology is the foundation to support the development of the mobile Internet. In speech recognition, the flexible and variable structure of the Chinese language acoustic model and the framework of statistical training can reflect the random characteristics of speech, so it has become the mainstream of speech recognition, especially large vocabulary, continuous sound, and nonspecific speech recognition technology. The establishment of the Chinese language acoustic model is mainly to solve the state transition matrix and the characteristic observation matrix.

In fact, since the early 1950s, research on text-to-speech at home and abroad began this field. The Audry system developed by Zhou and Yu in the laboratory can automatically recognize 10-digit English numbers. This marked the beginning of formal speech recognition research. In the 1960s, improvements in computer hardware promoted the rapid development of speech recognition research. But the algorithm processing level at that time was relatively low [1]. Li et al.'s invention provides a method for compressing a neural network acoustic model. The method includes the following: dividing the row vector of the weight matrix W of the output layer of the neural network acoustic model into the number of subvectors according to the specified dimensions. First, quantize the layer vector, obtain the first-level codebook, and replace the subvectors of the matrix W with the vectors from the first-level codebook to obtain the matrix $R*$. Finally, the matrices $W*$ and $R*$ are used to represent the weight W matrix and its algorithm. This neural network is too complex and difficult to use [2]. Since the above algorithms and calculations are too complicated, we will see subsequent improvements and optimization algorithms. Zappi et al. have proposed a text-to-speech algorithm based on the statistical audio model selection unit. In the process of forming the model, it first separates the audio parameters such as the frequency spectrum and fundamental frequency of the voice data in the archive. It is then decomposed and annotated into theories in the corpus to estimate the corresponding contextual phonemes. Statistical acoustic model is prone to loss of sound accuracy [3]. This invention discloses a learning method, a text-to-speech method and device for a separate multispeech model for text-to-speech. The model learning method of this embodiment reduces the target user's voice in the process of learning the target user's voice model. The data scale requires a small amount of user voice data to form multiple personal voice models, including the voice characteristics of the target user. The speech recognition standard of this research has been significantly reduced [4]. Koguchi et al. focus on the variability of speech signals caused by the phenomenon of cooperative pronunciation in Chinese speech recognition. Therefore, a method for constructing a syllable acoustic model has been proposed. In order to alleviate the problem of scattered training data, an intrasyllable overtone syllable model is proposed to initialize syllable parameters. Subsequently, an intersyllable conversion model was introduced to solve the problem of cooperative pronunciation between syllables, but the applicability of this research is not wide enough [5]. Zhuokun et al. use this algorithm to take advantage of the powerful computing power of GPU to improve the calculation speed of matrices and vectors when learning networks. The optimized network can process multiple data streams at the same time and practice a few example sentences to speed up the training process, but the speed is quite fast and the algorithm quality is not up to the standard [6].

This article focuses on the application of neural networks in detecting voice activation. In modern voice communication, although voice activation detection methods already exist in many communication systems, with the advancement of technology, voice communication methods are becoming more and more diversified. At the same time, there are more and more application scenarios, and the application environment is becoming more and more complex. Moreover, the existing voice activation detection methods are not suitable for these complex signal-to-noise ratio conditions. In this paper, combining the user experience of neural network in image processing, using existing sound detection methods, through scaffolding design, speech processing, algorithms, and simulation experiments, a neural network-based voice activation detection is proposed. Comparison of real-time testing in real-time and nonreal-time environments: in this paper, Southeast Asian adult students (native speakers are Indonesian and Thai) in Zhengzhou University were selected as subjects to explore the relationship between Chinese phonetic level and acquisition motivation of Southeast Asian students. This proves that the voice activation detection algorithm proposed in this paper has good accuracy and speed.

2. Chinese Speech Synthesis Algorithm Theory

2.1. Classic VAD Algorithm. The voice activation algorithm based on algorithm model and the word as the modeling unit trains related keyword models and several filler models for each word. All keyword models and algorithm models together form a recognition network. The advantage is that the search network is small, and the recognition is fast, does not require language model support, and has a higher recall rate than the subword model. G. 729 is a low-bit-rate speech coding technology in the ITU-T specification. It uses a conjugate structure linear prediction coding method and a driving code to transmit at 8 kbit/s. G. 729 is also the technical specification of this figure [7]. The compressed transmission format of silent frames in this technical specification: the voice activation detection algorithm is called the standard algorithm of the voice activation detection algorithm. G. 729 discrete transmission mode supports 8 kbit/s G. 729 encoder. First, it shows the voice activation detection algorithm and then the voice activation detection algorithm, which is used to detect the audio and nonaudio segments of the input audio signal. The audio part is transmitted at the normal encoding rate, and the nonaudio part is sent with 15-bit encoding per frame, and then, the nonvoice part is replaced with comfortable sound reproduction when receiving [8, 9]. According to the above method, all the received

Table 1: Statistics of self-built voice library.

Training set			Test set
Source	Keywords only	With keywords	With keywords
Recording	5489	4896	452
The Internet	2158	2110	387
Statistics	6025	6589	715

sounds are in WAV format, monophonic, 16 kHz sampling frequency, 16-bit linear quantization; the key word (word) sound length is 1-2 s, and the sound length is two keywords (phrases) as the sound statistics of 5-10 seconds are shown in Table 1.

The function of the voice activation detection algorithm in G. 729 is to detect the audio and nonaudio parts of the audio segment and the nonaudio segment using different data rates during the call and then synthesize the voice VAD call reception and determine whether it is correct and directly affect the call quality. G. 729 parameters calculated for the speech compression algorithm in Annex B include linear spectrum frequency parameters.

Full-band power and low-band power and zero-crossing rate: the Levinson-Durbin algorithm is used to calculate the autocorrelation coefficient$\{R(i)\}$ % of the input speech signal, where $q = 12$, then convert the calculated autocorrelation coefficient into coefficient reflection, and then convert the reflection coefficient after obtaining the autocorrelation coefficient converted to linear frequency parameter {LSF}&o where $p = 10$. The calculation methods of the two parameters for the full-band energy E and low-frequency energy E are as follows:

$$E_f = 10 * \log_{10}\left[\left(\frac{1}{N}\right)R(0)\right]. \tag{1}$$

In formula (1), $N = 240$, which is the width of the LPC window.

$$E_f = 10 * \log_{10}\left[\left(\frac{1}{N}\right)h^T Rh\right]. \tag{2}$$

In formula (2), h is the impulse response of a finite impulse response filter with a cut-off frequency of 1000 Hz, R is the Topolitz autocorrelation matrix, and the autocorrelation coefficients are diagonally distributed. The calculation method of ZC zero crossing rate is shown in formula (3).

$$\mathrm{ZC} = \frac{1}{2M}\sum_{i=0}^{M-1}\{\mathrm{sgn}\,[x(i)] - \mathrm{sgn}\,[x(i-1)]\}. \tag{3}$$

In formula (3), $x(i)$ is the input signal before processing, sgn (x) is the signal function when $x < 0$, when $x = 0$, sgn $(x) = -1$, and when sgn $(x) = 0$, $x > .0$, sgn $(x) = 1$, and $M = 80$ after obtaining the required parameters. The decision does not directly depend on the given parameter value. Instead, the long-term average of these extracted parameters is used to track changes in background noise characteristics. The parameter change of each frame is calculated according to next formula. Finally, check whether the input frame is an audio signal according to the set threshold. The changes to these settings are as follows:

Spectrum harmonic distortion measurement

$$\Delta S = \sum_{i=1}^{P}(\mathrm{LSF}_i - \mathrm{LSF})^2. \tag{4}$$

Full-band energy change measurement

$$\Delta E_f = \sqrt{E_f} - E_f. \tag{5}$$

Low-band energy change measurement

$$\Delta E_I = \sqrt{E_I} - E_I. \tag{6}$$

Excess rate change measurement

$$\Delta \mathrm{ZC} = \sqrt{\mathrm{ZC}} - \mathrm{ZC}. \tag{7}$$

In Equation (4), LSF represents the average value of the linear spectrum frequency parameters, in Equation (5), E_f represents the average value of the total energy of the frequency band, and E_I in Equation (6) is the average value of the low-frequency spectrum energy, and in Equation (7), ZC represents the average value of the zero-crossing rate. The analysis process of the algorithm research program in this paper is shown in Figure 1.

2.2. GSM VAD1 Algorithm. Even the classic VAD algorithm introduced above still has certain deficiencies in algorithm and response time, so a new improved algorithm and optimization algorithm will be introduced below. The voice activation detection algorithm recommended by the European Telecommunications Standards Institute (ESTI) GSMAMR is another classic voice activation detection algorithm. It is a voice action detection algorithm based on mutual judgment of multiple parameters. This algorithm is also implemented in the 3GPP standard. Voice-activated GSM AMR is mainly composed of two branch models: one is the ENS VAD algorithm provided by Europe called VAD1, and the other is the Motorola VAD algorithm provided by Motorola, called the VAD2 part. Here, we mainly introduce VAD1. The VAD1 algorithm is an adaptive multirate voice activation detection algorithm. The algorithm structure diagram is shown in Figure 2. From Figure 2, it can be seen that the calculation of the parameter VAD1 consists of the following parts: filter channel.

Then, the signal level in each subband is calculated separately, and the subband is calculated through the subband. The frame power of the input signal is obtained from the signal level [10, 11]. Motorola VAD algorithm estimates the parameters of the domain model. Here, the decision tree parameter merging algorithm based on random segment model proposed by its algorithm is used to bundle the parameters of the transition model to improve the training

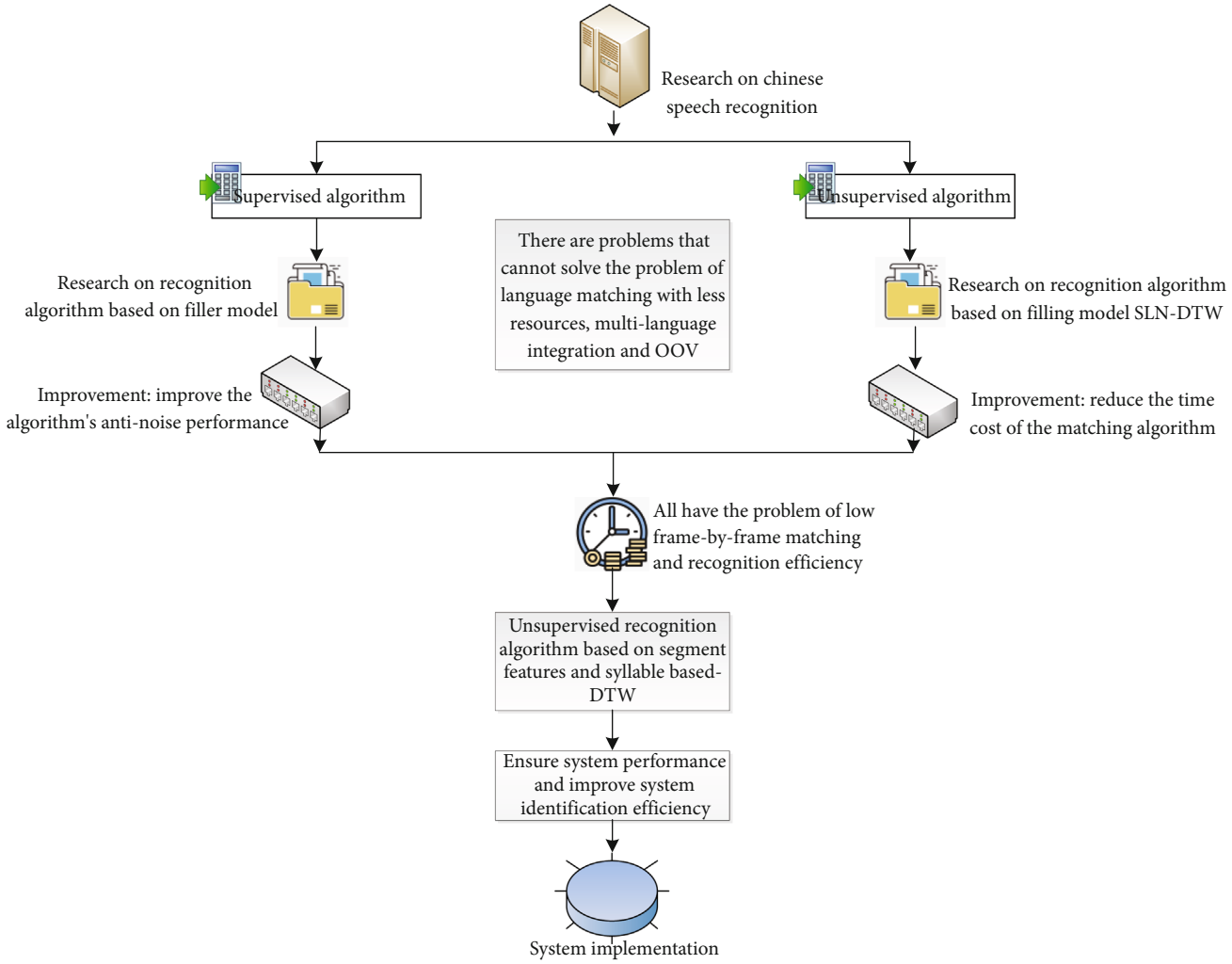

Figure 1: Algorithm research program analysis process.

effect. Pitch detection is to calculate the open-loop delay and size encoder through open-loop analysis and then use the height indicator to indicate the presence of height. Height detection makes the height analyzer pass itself, and the open-loop pitch gain obtained from the relevant calculation is used to pass the pitch. The indicator indicates the presence of the tone. The signal complexity in complex signal analysis is obtained by analyzing the pitch correlation vector, and then, the composite signal indicator is used to indicate the presence of the composite signal (such as a music signal).

Subband division and level calculation pass is a filter bank can divide the input audio signal into 9 different subbands. The structure diagram of the filter bank is shown in Figure 3. In Figure 3, the filter bank is composed of three filters 5 and 3, and each filter divides the input signal into high frequency components. The sampling rate is calculated as 2 : 1, that is, each bandwidth of the high-bandwidth section of each filter is twice the bandwidth of the low-band section. It divides the input audio signal into 9 bands according to frequency. The lower the frequency, the smaller the bandwidth of the band, and the narrower the band [12].

For sound level detection, the purpose of the pitch detection function is to detect vowels and intermittent signals. It is implemented based on the comparison of the open loop delays of the subloops calculated by the voice encoder. If the difference between the open loop delays of consecutive subbands is less than the threshold, the delay counter will be accumulated; if the sum of the count delay counters of two consecutive frames of input speech is large enough, the volume indicator will be set to 1, indicating that the field exists [13].

Sound detection: this is because the volume detection cannot detect the sound level of the audio signal. The purpose of the volume detection adjustment is to detect the volume of the input audio signal. At the same time, the height detection can also detect other signals, and the pitch detection can be realized by comparing the input pitch increment in the open loop with a defined constant threshold. If the open-loop tone expansion is greater than the set threshold, the tone flag is set to 1, indicating that a tone is detected.

Complex signal analysis and detection complex signal analysis are used to detect relevant signals after high-pass

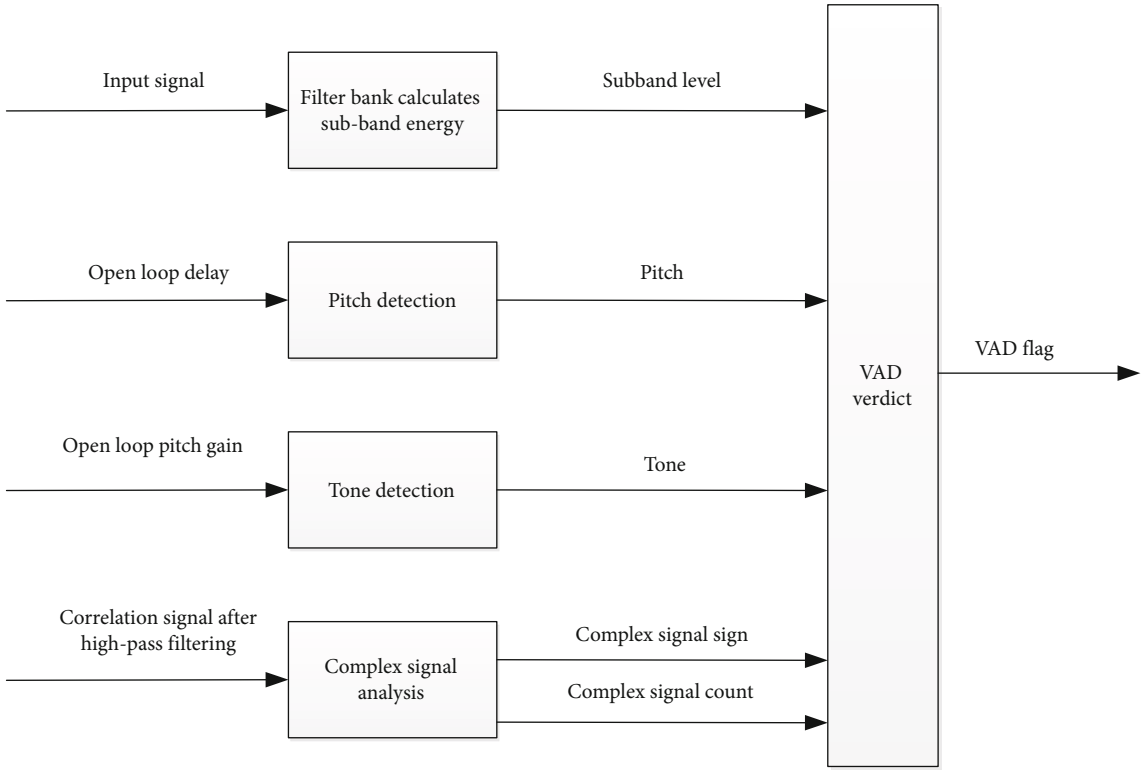

Figure 2: Block diagram of the GSM VAD1 algorithm.

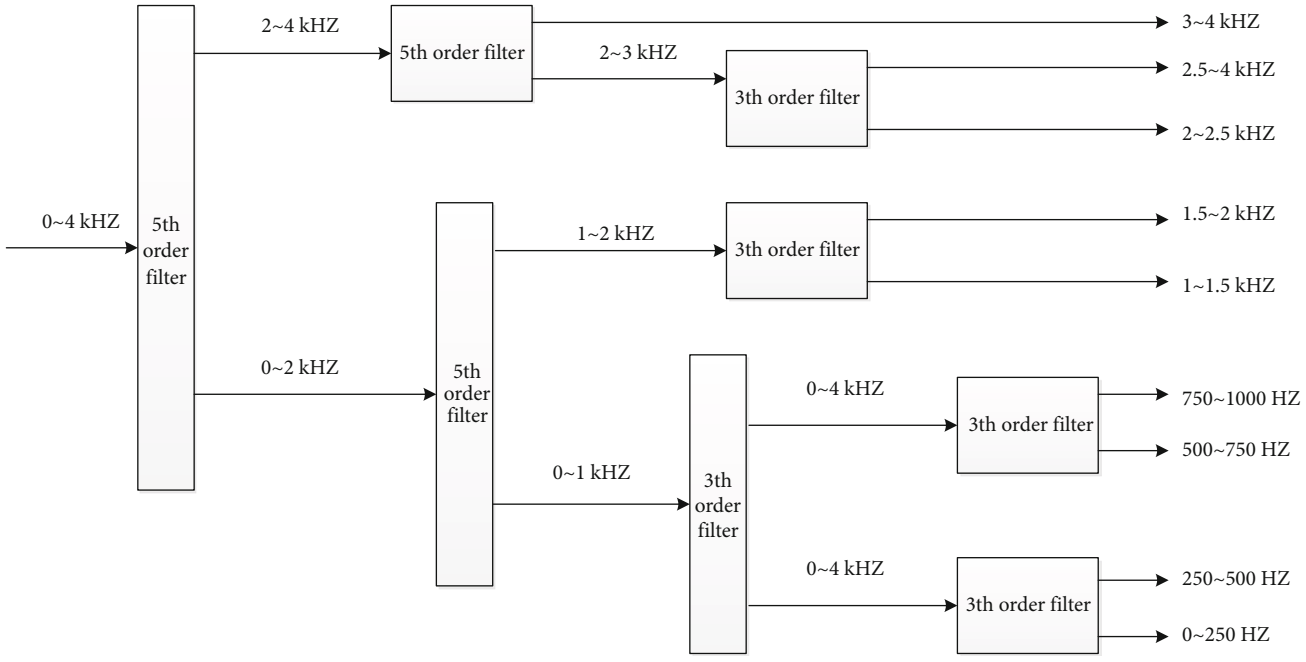

Figure 3: GSM VAD1 filter group structure block diagram.

filtering. If Comfort Noise is used to replace these signals, it cannot be accurately detected due to noise detection and distance detection. The sound will not be natural enough. If the highest normal correlation is obtained from the high-pass filtered speech signal, the position of the complex signal is marked as 1, indicating that there is a complex signal [14].

Background noise estimation: the background noise estimation is updated by the input amplitude level of the

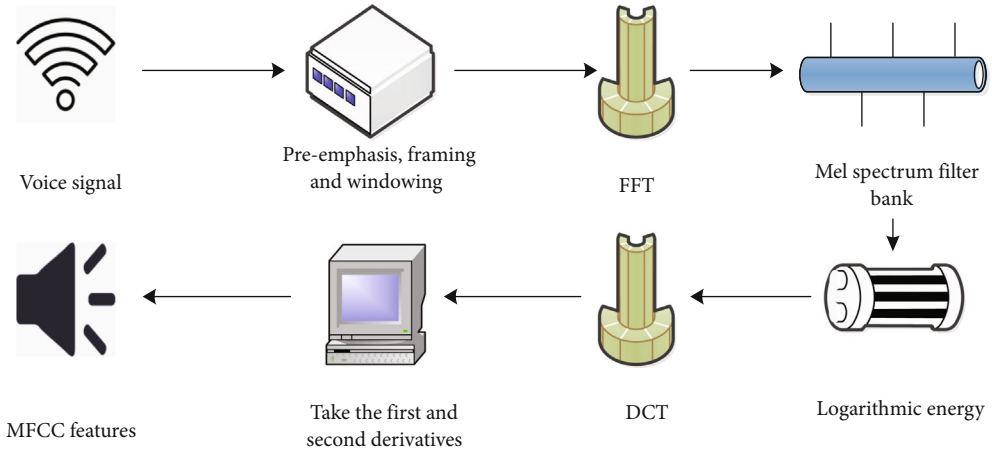

Figure 4: MFCC extraction block diagram.

previous speech signal frame. And in the noise evaluation update delay, the purpose of the frame is to avoid sudden failure of the initial speech position detection and destroy the noise detection or loudness signal. The noise estimate will not be updated.

2.3. Description of Sound Attributes. The most commonly used method to solve the problem of cooperative pronunciation in speech recognition is to establish a background-related acoustic model, which is typically a three-tone submodel, which takes into account the influence of two pronunciation units before and after on the current pronunciation unit. The three-tone submodel can achieve better results in solving the problem of cooperative pronunciation. The Mel Cepstrum Coefficient (MFCC) takes into account the characteristics of human hearing, that is, the masking effect of human hearing. Weak frequency components can be masked by neighboring stronger frequency components. It has good perception and good antinoise ability. The function diagram is shown in Figure 4.

Preemptive processing is shown in formula (8). The essence is to use a high-pass filter to process the audio signal. It has two functions: one is to improve the high-frequency part of the signal and maintain the frequency spectrum of the signal. Signal: the frequency band between the low frequency and the high frequency uses the same signal-to-noise ratio to calculate the frequency spectrum, and the second is to remove the high frequency part of the offset signal. And highlight the high frequency mode [15, 16].

$$H(Z) = 1 - \mu z^{-1}. \tag{8}$$

The audio signal can be considered a signal. The quasi-stationary can be converted into a frame and the sound in the frame after subframe processing can be regarded as a steady-state signal. In the processing of subframes, the frame interval is usually about 10 ms to 25 ms. The overlapping area is defined between adjacent images and is 1/2 or 1/3 of the image length. After framing the signal, hamming window analysis is usually selected to improve the accuracy of the analysis. If the signal after the frame is $S(n)$, $n = 0, 1, \cdots, N-1$, where N is the frame size, and $S(n)$ after windowing, the fast formulas (9) and (10) can be used.

$$S(n) = S(n) * W(n), \tag{9}$$

$$W(n, a) = (1 - a) - a * \cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1. \tag{10}$$

Among them, the Hamming window is transformed into the value of a, which is generally 0. 46. After preemphasis, framing and windowing, and fast Fourier transform (FFT) are performed on each frame of the signal, the characteristics of the speech signal must be observed in the frequency domain. This is done to get the frequency spectrum of each frame, and the power spectrum is obtained by calculating the modulus and square of the frequency spectrum. The discrete Fourier transform (DFT) of the speech signal is expressed as Equation (11).

$$X_a(k) = \sum_{0}^{N-1} x(n) e^{-j2\pi k/N}, 0 \le k \le N. \tag{11}$$

Among them, $x(n)$ is the input audio signal, and N is the number of Fourier transform points. The Mel filter bank can reflect the perceptual characteristics of the human ear. The conversion between linear frequency and Mel frequency is shown in Equation (12).

$$B(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{12}$$

The Mel filter bank consists of several bandpass filters in the frequency spectrum. As shown in Equation (13),

$$Hm(k), 1 \le m \le M. \tag{13}$$

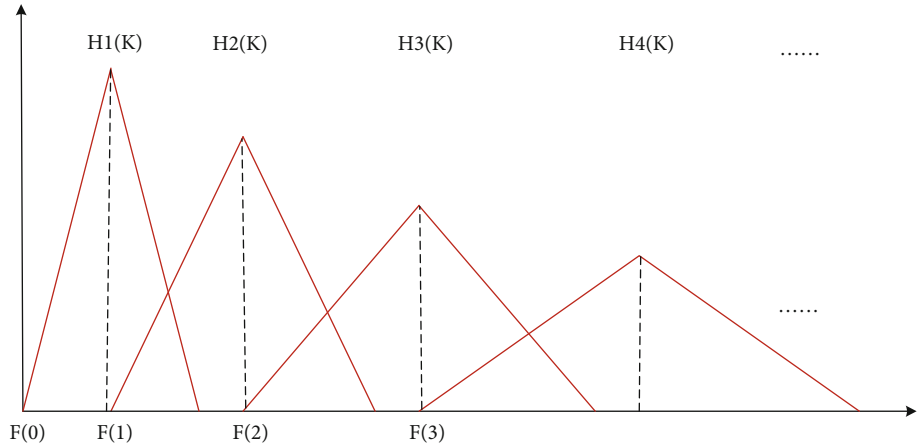

Figure 5: Mel filter bank.

Among them, M is the number of filters, usually 22 to 26. Each filter has a triangular filter characteristic, and $f(m)$ is the center frequency. The transfer function is represented by Equation (14).

$$Hm(k) = \begin{cases} 0, k < f(m-1), \\ \dfrac{2(k - f(m-1))}{(f(m+1) - f(m-1))(f(m) - f(m-1))}, f(m-1) \le k \le f(m), \\ \dfrac{2(f(m+1) - k)}{(f(m+1) - f(m-1))(f(m) - f(m-1))}, f(m) \le k \le f(m+1), \\ 0, k \ge f(m+1), \end{cases} \tag{14}$$

$$\sum_{0}^{M-1} Hm(k) = 1. \tag{15}$$

Then, the Mel filter bank is composed of a group of triangular filters, as shown in Figure 5.

The center frequency $f(m)$ can be defined in the form of Equation (16).

$$f(m) = \left(\frac{N}{F_s}\right) B^{-1}\left(B\left(f_l + m\frac{B(f_h) - B(f_l)}{M+1}\right)\right). \tag{16}$$

Among them, f_n is the highest frequency in the frequency range, f_l is the lowest frequency, N is the window width of DFT or FFT, f is the sampling frequency, and B^{-1} is the inverse function of B, which can be expressed by Equation (17).

$$B^{-1}(b) = 700\left(e^{b/2595} - 1\right). \tag{17}$$

According to Equation (18), the logarithmic energy of the filter is calculated, and the result approximate to the homomorphic transformation can be obtained.

$$s(m) = \ln\left(\sum_{K=0}^{N-1} |X_a(k)^2| Hm(k)\right), 0 \le m \le M. \tag{18}$$

The output power or amplitude of each filter channel is related. And you can use the discrete cosine transform (DCT) to obtain the coefficients of the decorative cavity, as shown in Equation (19).

$$c(n) = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{n\pi(m-0.5)}{M}\right), 0 \le n \le L. \tag{19}$$

Among them, $c(0)$ is the 0-order MFCC, which reflects the energy spectrum, and L is the MFCC order, and the value is 12-16 during the feature separation process. Assume that different speech bubbles are not related to each other. Therefore, the information between bubbles must be relevant and evolving [17]. In actual use, the relationship between bubbles is close to the second-order and second-order differential coefficients. The ceps characteristic is called the static characteristic, and the spectral difference of the static characteristic is called the dynamic characteristic [18, 19]. The complementary advantages and disadvantages of static and dynamic attributes can improve memory performance. The calculation of the difference is shown in formula (20).

$$d_t = \begin{cases} C_{t+1} - C_t, t < T, \\ \dfrac{\sum_{m=1}^{M} k(C_{t+m} - C_{t-m})}{\sqrt{2\sum_{m=1}^{M} X^2}}, \text{others}, \\ C_t - C_{t-1}, t \ge N - M. \end{cases} \tag{20}$$

Among them, d_t is the t-th difference parameter, C_t is the coefficient to be obtained, where t and N are the sequence, and M is the time difference of the derivative. Good quality means it is easy to understand and looks easy. Good quality means you have to pay a little attention to listening. But it sounds simple: quality means moderate concentration, but you can figure out that it looks light, fatigue level, low quality means you have to work hard to figure it out, and the MOS scoring method is based on a 5-point system (34 points), as shown in Table 2.

Table 2: MOS value scoring system.

Score	Quality level
5	Excellent
4	Good
3	Middle
2	Difference
1	Inferior

The test sound quality is divided into 5 levels: excellent (5 points), good (4 points), medium (3 points), poor (2 points), and poor (1 point). As a very important parameter, it is usually set as a constant in the standard gradient algorithm. However, in practical applications, it is difficult to determine an optimal learning rate that is suitable from beginning to end. If the minibatch optimization algorithm is used, the baseline model (CPU) learning rate $(\alpha = 0.1)$ adjustment strategy is no longer suitable, test the recognition rate in the actual recognition system to observe the impact on the model performance, and verify the effectiveness of the parallel optimization algorithm.

DRT reflects the intelligibility or intelligibility of speech. The test method includes testing the pronunciation of characters or words with the same vowel. For example, the vowels of the Chinese characters "you" and "li" have the same pronunciation [20]. It will be better if the correct sound quality is distinguished. DRT score is the percentage of all testers who can obtain accurate test results from audio measurement. It is generally believed that the MOS value corresponding to DRT score exceeding 95% is 5 points, the MOS value corresponding to 85%~94% is 4 points, 75%~84% correspond to 3 points, and the MOS value is 75%~84%. 65%~74% correspond to 3 points, MOS score is 2 points, and less than 65% corresponds to MOS score 1 point. DAM and DRT type tests use percentage scores as a comprehensive assessment of speech quality, as a measure of acceptance of words from many aspects. Therefore, the processing of sound quality is extremely important. The detection of sound quality determines the accuracy of Chinese speech synthesis. In the system set up below, it is necessary to have a higher requirement for sound quality to detect and score it.

3. Algorithm Simulation and Experiment

3.1. Voice Sample Collection. A phoneme is the smallest unit in speech, and the pronunciation methods and parts of several phonemes of a speech keyword are different. Each phoneme unit is composed of multiple consecutive frames of speech, and the feature parameters of the phoneme have greater differentiation, that is, several feature vectors in the phoneme segment are similar, and vice versa, there are greater differences. Considering that traditional template matching methods also use frame as a unit for keyword recognition, when the length of the voice is long, the recognition efficiency is reduced. Therefore, this article proposes a new segment feature. Ten foreign students who came from South Asia participated in our research. There are five men and five women. In order to prove that there is no relationship between Chinese critical period acquisition and Chinese proficiency, we should consider the important role of age in CPH. The subjects were all adult international students, aged 20-30 years old, and their average age is 25.3 years old. In the second place, the language background and country should be strictly controlled to ensure its accuracy and effectiveness. It is worth noting that the Southeast Asian adult learners are most be never learned Chinese before they came to China, and if they mastered Chinese before teenagers and before they came to China, the data will be meaningless. Additionally, volunteers should come from the same nations and speak a comparable or same mother tongue to assure the experiment's rigor. The function diagram is shown in Figure 6.

Firstly, the international students who came from Southeast Asia are invited to participate in the Chinese phonetic experiment. The purpose of Chinese phonetic experiment is testing the Chinese production ability of foreigners. By the way, Chinese as the interlanguage is used in the whole experiment. We recorded the test process with tascamdr44wl and collected the recorded data. In order to accurate the Chinese phonetic production ability of Southeast Asian adult learners, the Chinese phonetic paper includes two parts: vocabulary and phrase. After they finished the Chinese phonetic production, we can acquire 10 recordings. Finally, we just need to analyze the recordings of the participants and export data. As the second step of the experiment, the subjects had to fill in a questionnaire after completing the Chinese phonetic test. More importantly, the questionnaire is used as a tool to test the Chinese learning motivation, especially in part B of the questionnaire. This research is aimed at gaining the intrinsic motivation and extrinsic motivation of foreign students, who play a more important role in the relationship between Chinese learning motivation and Chinese proficiency. Third, after completing the questionnaire, the experiment ended.

From the results in Table 3, the number of filled-in forms is gradually increasing, which improves the average keyword recall rate. However, by increasing the number of filling patterns, a single keyword is very different. The reason for this phenomenon is that the speech that forms the filling pattern is randomly shuffled into 5 parts, and each group of speech is divided into 5 parts, and the nonkeyword coverage related to the keyword is destroyed. Different and each related filling mode has different absorptive capacity for nonkeywords [11]. However, according to the evolution trend of the average recall rate, the filling mode is better when the filling mode is 5, which proves that the more speech exercises without keywords, the higher the absorption rate. The importance of training model algorithm efficiency draws the keyword mention rate error in Table 3 as a transition curve, as shown in Figure 7.

The operation process proved that the supervised recognition algorithm based on the filling model must be supported by a large number of self-annotated speech. Especially when it is necessary to train an acoustic model with sound insulation performance, relying on the data is more solid although the modeling unit can guarantee the

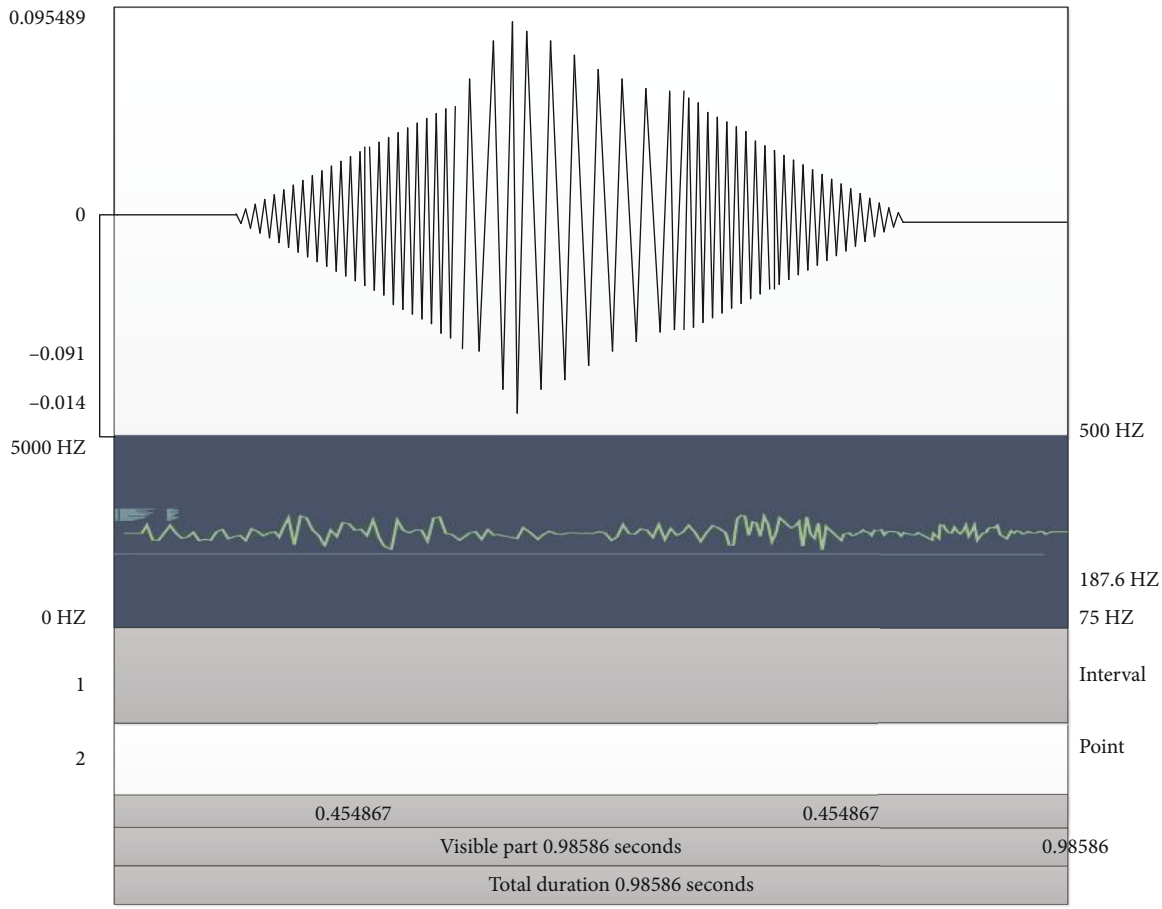

Figure 6: "Fu2" came from Southeast Asian adult learner (representative).

Table 3: The impact of the number of models on the keyword mention rate.

Number of models	3	4	5
DLDT	89.78	91.25	89.65
HLW	94.56	91.23	95.98
NBE	89.55	88.69	97.65
RMB	98.54	89.25	87.69
XZ	95.11	96.45	93.64
YDYL	89.22	83.69	95.36
ZGSH	89.67	91.94	92.62
Average mention rate	86.99	92.65	94.68

recognition rate of the algorithm. But this does not solve the OOV problem, and the identification process depends on the framework. When most words are found and the number of keywords is low, the perception and performance are low. It can be seen from the figure that when the number of states of the speech model increases to about 7, the decline in segmentation dispersion is already very weak. Therefore, from this perspective, in the speech model based on the average probability, the number of states is 6~8 is a suitable choice.

3.2. Classical VAD Algorithm Simulation. The learning rate is a very important parameter in the learning process of neural networks. In the first stage of model training, the algorithm performs simulation learning with high efficiency. As the model approaches the convergence point, the training process will not change. Before the model converges in this experiment, the learning rate will be very low [21]. The number of nodes in the hidden layer of the network is set to 600, and the training set is divided into 500 categories, so that you can fully train your model. Each model has undergone 12 rounds of training, with a batch value of 64.

This means that the network trains 64 examples of words at a time. By changing the training rate and the parameter values in the training data warehouse, we compare model convergence with changes in model performance after parallel optimization at different training rates. Select the matching learning mode. The abscissa is the simulated training cycle [22]. This article has 12 learning simulation cycles. Starting from the seventh cycle, the learning efficiency is adjusted to half of the previous one. The ordinate is each different theoretical model. Framework can eliminate the uncertainty of each collected language to a certain extent, and the PPL value in the experiment is inversely proportional to the performance of the model.

It can be seen from Figure 8 that when the value of the training phase is smaller, the model convergence effect is

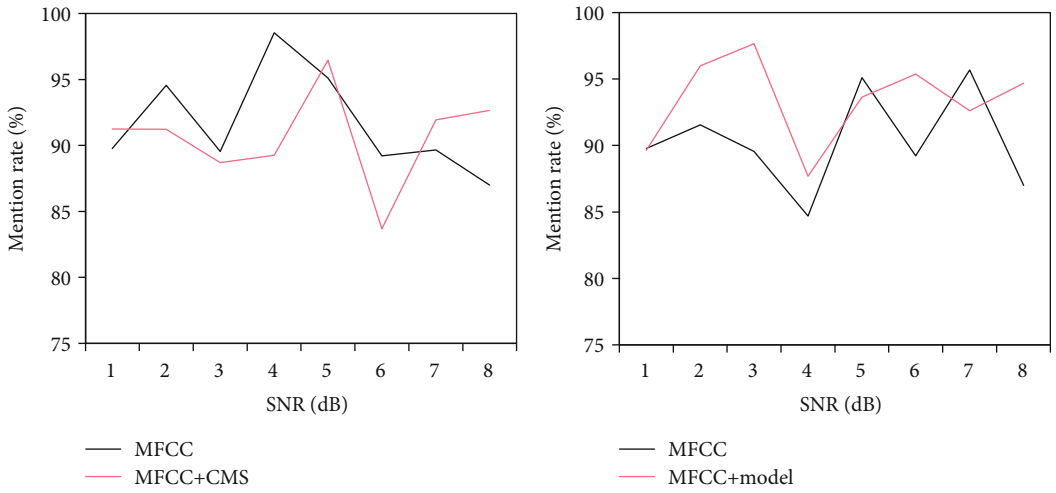

Figure 7: Error curve of the mention rate of keywords.

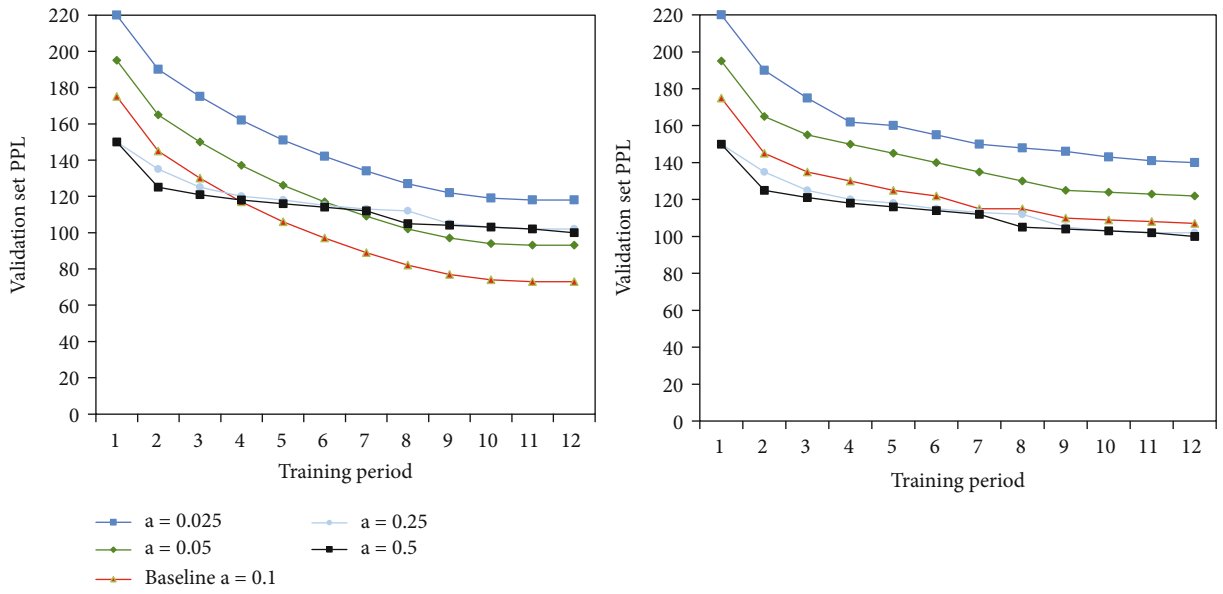

Figure 8: Convergence of the model at different learning rates after parallel optimization.

better. The difference lies in the performance of the model. When the learning rate increases, the network can learn better and better. It can be seen that the value gradually increased to a certain level during this period. The performance of the model tends to deteriorate. This indicates that a higher learning rate will over-learn the network and affect the overall capacity of the network. This article also tries to use a higher learning rate.

This can prevent the model from converging. In the experiment, when the experimental value is set to 1.3, a good experimental effect is obtained, but it is very different from the setting value of 0.1; especially when the batch value is 32, the total learning the efficiency is 1.3. When the learning efficiency is evenly divided among other keyword phrases, it can still be used as a training cycle to achieve learning progress, so there will be no problem of low learning rate.

The results are shown in Table 4, and the RNN-like model based on CPU is trained. It can be seen from Table 1 that when the setting value is 1, the training efficiency of the GPU-based model is increased by more than 2 times compared with the training. The CPU and system recognition rate will not drop too much. Demonstrated a successful implementation, the RNN classroom training on the GPU was successful, and some accelerated training results were achieved. When the batch value is 64, the number of words processed by the network per second increases by nearly 19 words compared to batch = 1 [23, 24]. Multiply the performance on the processor driver by 38. The system recognition rate after actual recognition optimization is relatively close to the system recognition rate in the basic model. Although there has been a decrease, the magnitude is not large. When inserting a 3G model, the recognition rate

Table 4: Changes in system recognition rate and model training rate after parallel optimization.

Experimental configuration	WER/%			Speed/(words·s^{-1})	Speed increase multiple
	n-gram	RNN	RNN+n-gram		
Baseline (CPU), $a = 0.1$	23.45	24.5	22.08	468	_
Batch = 1 (GPU), $\alpha = 0.1$	23.45	25.6	21.66	859	2.12
Batch = 64 (GPU), $\alpha = 1.5$	23.45	25.6	22.04	18466	37.56
Batch = 64 (GPU), $\alpha = 1$	23.45	26.5	23.14	23451	39.54

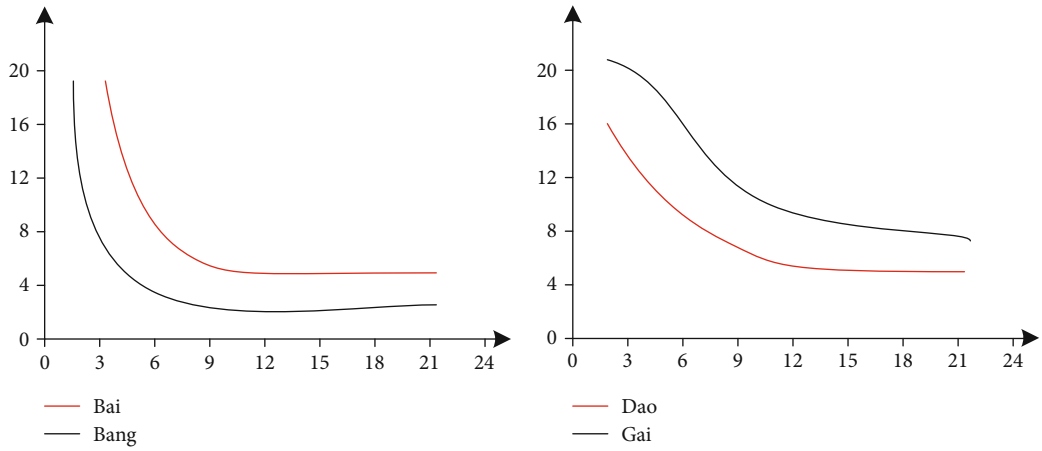

Figure 9: "bai," "bang," "dao," and "gai" in different state optimal segmentation deviation curves.

decreases to a lesser extent and significantly improves training efficiency, which makes RNN can be used to train large data warehouses. Based on the recognition algorithm of LVCSR with words as the modeling unit, the speech signal is recognized and converted into text form, and then, keyword recognition is realized by text search.

3.3. GSM VAD1 Algorithm Simulation. This article has established a voice database by itself, with a total of 10 Chinese keyword set. The voice sources include recording and webcasting. The voice types include keywords only and keywords with keywords. Record in a quiet environment and draw up the content of sentences containing keywords. Each keyword involves 3~5 sentences, which are read several times. However, how to choose the number of states of the speech model for the speech feature vector? That is, how many eigenvectors are more appropriate? As the number of cluster centers increases, the change trend of the classification gap can be seen. When the number of states is L, if the number of states continues to increase and the classification gap d is not much different, there is a classification gap d. The statistics that continue to increase are not important. Based on Fisher's algorithm, this paper calculates different Chinese diffusers in different states. The four curves in Figure 9 represent the state numbers of the sounds "bai," "bang," "dao," and "gai." The distribution of the graph can be seen as the number of states of the speech model increases to an approximate value [25, 26]. Seven, the fragmentation of descending word segmentation is already very small, so from this perspective, the number of states in the speech model is 6-8 according to the average probability, which is the correct choice.

Table 5: Recognition results after adding the transition model%.

Model type	WER	Model type	WER
Triphone	14.63	Syllable+transition model	12.98
Syllable	13.89		

When the threshold is low, the next step is to train a model with fewer training samples. But these models do not have enough training samples to fully train them. This not only does not optimize the parameters of the model, but also produce the opposite effect, so the accuracy of the system is low. When the threshold is increased to about 400, the recognition accuracy is the highest. Due to the ever-increasing threshold, some models have training examples that allow the model to be fully trained. But because the sample size is less than the defined threshold, the model is not optimized. Therefore, the recognition rate is reduced. Table 5 shows the recognition results after adding the intersyllable transformation model to the context-free syllable model system [10]. Compared with the context-independent syllable form, after adding the change model, the system word error rate is further reduced. However, since the transition model considers the problem of covoting between syllables, the effect of enhancement is not clear. There are two reasons: On the one hand, syllable pronunciation is not as serious as the pronunciation. On the other hand, it may be necessary to improve the accuracy of the

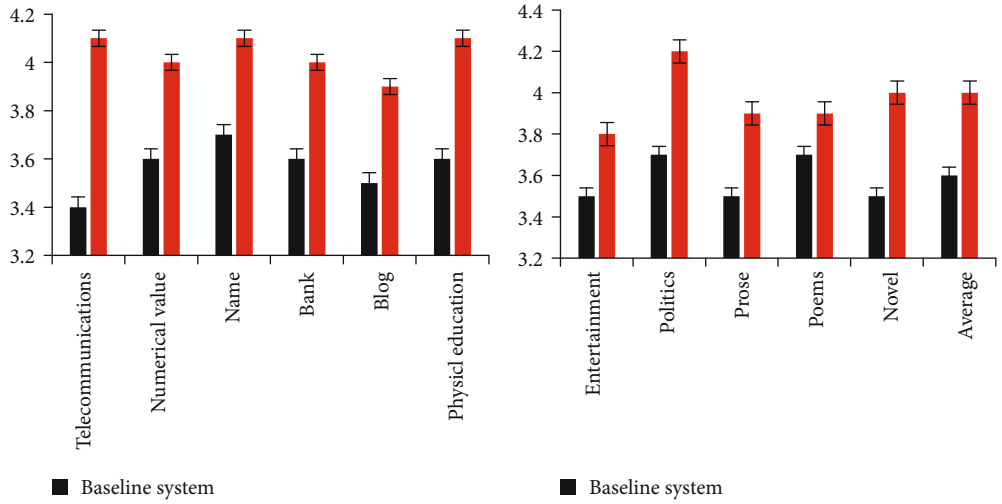

Figure 10: Comparison of evaluation results between the baseline system and the algorithm in this paper.

SSM adapting to the possible trajectory of the transition zone [27, 28].

4. Experimental Results and Analysis

4.1. Analysis of Algorithm Results. Because in the speech signal, the speech segment within the same phoneme can cover multiple consecutive frames of speech and the pronunciation has a certain degree of continuity. Therefore, in this paper, the average value of the frame features of the phoneme segment is represented as the segment feature. The detailed theory of the segment feature is described above. As already introduced in the content, this section mainly describes the process of extracting segment features. The basic system we use for comparison is a splicing synthesis system based on traditional cost functions that select units and use the same sound library. The cost table related to the system is manually edited by audio experts. Choose from 11 common text-to-speech application areas. Select 20 sentences in each field. They are combined with two synthetic systems, respectively, and the combined results are evaluated by 5 auditors with a MOS score of 1 to 5, all of whom are professional assessors outside the system. Researchers do not know the content of the pretest. During the test, the synthesized sounds of the two systems were played randomly. You can view the statistical data and final average scores of the two scoring systems on different aspects of the text.

(1) After using the statistical model to select the unit, the comprehensive results of all aspects have been continuously improved, and the average MOS score has increased by about 0.5 points. We also combined the results of the two systems to perform a T-test. And the results also proved that the improvement of this effect is significant ($p < 0.05$)

(2) *Comparison of Speech MOS Scores.* After text synthesis in different fields uses the unit selection method based on noise model statistics, the maximum difference between them is reduced from 0.312 points of the basic system to 0.228 points, better stability of synthesis effect.

(3) Due to the uneven distribution of training data and too few training samples corresponding to some syllables, sufficient training cannot be performed. Therefore, the recognition accuracy of the syllable model directly trained with the HTK tool is much lower than other models

The experimental results show that the training of the class-based RNN model in the CPU is used as the baseline. It can be seen that when the batch value is 1, the training efficiency of the model on the GPU is increased by more than 2 times compared with the training on the CPU. However, the recognition rate of the system has not greatly decreased, indicating that the class-based RNN training on the GPU has been successfully implemented, and a certain acceleration training effect has been achieved: when the batch value is 64, the number of words processed by the network per second is comparable compared to batch = 1.

4.2. Rational Analysis of Foreign Student's Foreign Language Ability. We analyzed the Hanyu Pinyin data through Praat and then used SPSS to analyze the connection between the results of the Hanyu Pinyin experience and the motivation to learn Chinese. On the other hand, in Praat Basics, the pronunciation lines of Southeast Asian Chinese learners are similar to those of standard Chinese learners. With the same digital analysis result as you can see, the sound curve has the same direction. (Because we have too much pinyin information, the researchers chose the most representative number, as shown in Figure 10.)

This paper believes that Praat has produced valuable data. Besides, we also calculate the accuracy of Chinese phonetic production, and the results are shown in the figure below. As shown in the histogram of Chinese phonetic production results, 70% of the subjects scored between 55 and 80, 20% below 55 and 10% above 80. HSK means Chinese Proficiency Test, and the full name is Hanyu Shuiping

Kaoshi. The Chinese Proficiency Test is an international Chinese proficiency standardized test established to test the Chinese proficiency of nonnative Chinese speakers (including foreigners, overseas Chinese, ethnic Chinese, and Chinese minority candidates). The results of speech test and HSK can be used as parameters to measure the Chinese proficiency of foreign students. From the interview, I learned that my least favorite teaching method is to let the teacher teach simple and repetitive. It can be seen that at the significance level of 0.05, the Sig values of the independent sample T-test of the twelve motivation types are all greater than 0.05, so it is believed that whether to learn other languages is not significantly different in the twelve motivation types. Although there is no significant difference, the learning motivation of learners who have not learned other languages except Chinese is generally stronger than that of learners who have learned other languages. This shows that learners who have only learned Chinese as a foreign language have a strong purpose in Chinese learning and a strong desire to learn Chinese well.

This makes students tired and easily distracted. The boring content of classroom teaching leads to a dull classroom atmosphere, and it is difficult for students to grasp the key points of learning and make students lose interest in classroom learning. Students who lose interest in classroom learning believe that the problem is not caused by the traditional teacher-centered teaching method, but it also includes a list-based method to describe the teacher's contempt for the course or the student. Because some teachers also pay attention to lectures, there are few interactive links, but the content of the tutorial is concise and clear, and there are also key points and highlights, which can help students master the textbook and stay focused. Most of these types of problems occur in the mid-to-long term. When you are learning Chinese in the primary stage of China, the teacher will focus on practicing in the elementary school class. It is too often used to describe educational materials and has more interactive links.

5. Conclusion

In this paper, by combining the classic VAD and GSM VAD1 algorithm simulations, it is concluded that the HSK score is positively correlated with the voice experimental test score. The initial age of Chinese acquisition is negatively correlated with intrinsic motivation (interest). The intrinsic motivation (interest) of Chinese learning is positively correlated with HSK scores. In Chinese acquisition, there is an inverse relation between initial age and extrinsic motivation (status, etc.). It seems that the relations are complex; however, actually there is a kind of connection: the initial age of Chinese acquisition affects the motivation of Southeast Asian students to learn Chinese. As far as the intrinsic motivation of Chinese acquisition is concerned, the earlier foreign students come into contact with Chinese, the more obvious the impact of intrinsic motivation on HSK and Chinese phonetic test scores. The more enthusiasm and interest they have in learning Chinese, the better their Chinese performance will be. On the contrary, they have low motivation and interest in learning Chinese, and their Chinese performance is poor. In terms of external motivation, the starting age of Chinese acquisition has an impact on the money, status, purpose, and other external motivation of Southeast Asian students, but the impact is not significant. According to the result of Praat and SPSS, we found that the intrinsic motivation plays a crucial role in CSL acquisition, and the intrinsic motivation can help mature learners who came from Southeast Asia acquire Chinese better and better. The earlier they learn Chinese, the higher their motivation will be, and the better it will be for them to set up their Chinese learning goals. The more motivated they are to acquire Chinese, the better their Chinese scores (such as HSK test scores and Chinese phonetic test scores); the higher their interest in acquiring Chinese, the better their Chinese performance/performance; the external motivation has little influence on the Chinese acquisition of Southeast Asian foreign students, while the internal thing has a profound influence on the Chinese acquisition of Southeast Asian foreign students. Occasionally, external incentive alone is insufficient to impact overseas students' Chinese phonic level. For instance, encouragement and recognition may not be major influences. If learners of CSL desire it, their intrinsic drive will be critical. External motivation and intrinsic motivation, on the other hand, can be changed into one another, and external motivation can also transform to pressure, affecting learners' study. In conclusion, the intrinsic motivation plays a vital role in Southeast Asian mature students' CSL acquisition. How Chinese teachers should rethink on how to use internal motivation to increase the skill and level of CSL learners and especially how to help students develop self-confidence are problems that need long-term research.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] W. Zhou and S. Yu, "Research on the communication method of mobile network shadow fading based on interference alignment algorithm," *The Journal of Supercomputing*, vol. 72, no. 7, pp. 2891–2909, 2016.

[2] Y. Li, S. Ding, and Z. Li, "Dictionary learning with the cosparse analysis model based on summation of blocked determinants as the sparseness measure," *Digital Signal Processing*, vol. 48, no. C, pp. 298–309, 2016.

[3] V. Zappi, A. Vasudevan, and S. Fels, "Towards real-time two-dimensional wave propagation for articulatory speech synthesis," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 2010–2010, 2016.

[4] K. Hongo, T. Nose, and A. Ito, "Spectral and pitch modeling with hybrid approach to singing voice synthesis using hidden semi-Markov model and deep neural network," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2962–2962, 2016.

[5] J. Koguchi, S. Takamichi, M. Morise, H. Saruwatari, and S. Sagayama, "DNN-based full-band speech synthesis using GMM approximation of spectral envelope," *IEICE Transactions on Information and Systems*, vol. E103.D, no. 12, pp. 2673–2681, 2020.

[6] D. U. Zhuokun, W. Shao, and W. Qin, "Research progress and application of retention time prediction method based on deep learning," *Se pu=Chinese journal of chromatography/Zhongguo hua xue hui*, vol. 39, no. 3, pp. 211–218, 2021.

[7] J. I. Changming, L. I. Chuangang, L. I. Xiaoyong, W. A. Boquan, and Z. H. Pei, "Research and application of dynamic programming algorithm in reservoir operation based on functional analysis," *Journal of Hydraulic Engineering*, vol. 47, no. 1, pp. 1–9, 2016.

[8] Z. Lv, D. Chen, and Q. Wang, "Diversified technologies in internet of vehicles under intelligent edge computing," *IEEE transactions on intelligent transportation systems*, vol. 22, no. 4, pp. 2048–2059, 2021.

[9] L. Yang, J. Qiu, X. Sun, and J. Xing, "Research and application on strength model of cemented backfill pillar for stage subsequent filling mining method," *Zhongnan Daxue Xuebao(Ziran Kexue Ban)/Journal of Central South University(Science and Technology)*, vol. 49, no. 9, pp. 2316–2322, 2018.

[10] N. Jiang and T. Liu, "An improved speech segmentation and clustering algorithm based on SOM and *K*-means," *Mathematical Problems in Engineering*, vol. 2020, no. 1, Article ID 3608286, 19 pages, 2020.

[11] T. Takara and A. Higa, "Generative model of spectra for a word using Fujisaki's model and genetic algorithm," *Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2965–2965, 2016.

[12] C. Li, H. J. Yang, F. Sun, J. M. Cioffi, and L. Yang, "Adaptive overhearing in two-way multi-antenna relay channels," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 117–120, 2016.

[13] N. Adiga and S. Prasanna, "Acoustic features modelling for statistical parametric speech synthesis:a review," *IETE Technical Review*, vol. 36, no. 2, pp. 130–149, 2019.

[14] C. H. He, "An introduction to an ancient Chinese algorithm and its modification," *International Journal of Numerical Methods for Heat&Fluid Flow*, vol. 26, no. 8, pp. 2486–2491, 2016.

[15] A. K. Kaliyev and S. V. Rybin, "Acoustic modeling for Kazakh speech synthesis," *Scientific and Technical Journal of Information Technologies Mechanics and Optics*, vol. 19, no. 5, pp. 951–954, 2019.

[16] Y. Aizawa, M. Kato, and T. Kosaka, "Many-to-many voice conversion using hidden Markov model-based speech recognition and synthesis," *Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2964-2965, 2016.

[17] K. S. Lee, "Speech synthesis using acoustic Doppler signal," *The Journal of the Acoustical Society of Korea*, vol. 35, no. 2, pp. 134–142, 2016.

[18] G. Dartmann, H. Song, and A. Schmeink, *Big Data Analytics for Cyber-Physical Systems: Machine Learning for the Internet of Things*, Elsevier, 2019.

[19] R. W. Mill and G. J. Brown, "Utilising temporal signal features in adverse noise conditions:detection, estimation, and the reassigned spectrogram," *Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 904–917, 2016.

[20] J. Qi, X. Feng, E. Nilot, and X. Li, "Joint inversion of magnetotelluric and fullwaveform seismic data based on alternating cross-gradient structural constraints," *Global Geology*, vol. 23, no. 2, pp. 55–66, 2020.

[21] J. H. Lee, "Biomimetic idealization of a mechanically coupled acoustic sound sensing mechanism," *Simulation*, vol. 94, no. 2, pp. 131–143, 2018.

[22] B. Han, X. Yang, Z. Sun, J. Huang, and J. Su, "OverWatch: a cross-plane DDoS attack defense framework with collaborative intelligence in SDN," *Security and Communication Networks*, vol. 2018, Article ID 9649643, 15 pages, 2018.

[23] M. R. Bai, Y. Li, and Y. H. Chiang, "Modeling of reverberant room responses for two-dimensional spatial sound field analysis and synthesis," *Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1953–1964, 2017.

[24] S. O. Gheibi, A. Fallah Shojaei, A. Khorshidi, and S. M. Hosseini-Golgoo, "Synthesis, characterization, and gas sensing properties of Ni-Cr-Al LDH," *Applied Physics A*, vol. 127, no. 8, pp. 1–7, 2021.

[25] R. M. Borik and M. A. Hussein, "Synthesis, molecular docking, biological potentials and structure activity relationship of new quinazoline and quinazoline-4-one derivatives," *Asian Journal of Chemistry*, vol. 33, no. 2, pp. 423–438, 2021.

[26] M. Tokareva, H. Ohar, S. Tokarev, and Y. Stetsyshyn, "Synthesis, structure and properties of the grafted peptidomimetic polymer brushes based on poly(N-methacryloyl-L-proline)," *Chemistry and Chemical Technology*, vol. 15, no. 1, pp. 26–32, 2021.

[27] J. Jibson, "Vowel identification and goodness based on level of formant detail," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2505–2505, 2020.

[28] M. Rau and J. O. Smith, "A comparison of nonlinear modal synthesis using a time varying linear approximation and direct computation," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2909–2909, 2019.