Hindawi
Advances in Multimedia
Volume 2022, Article ID 2910859, 13 pages
https://doi.org/10.1155/2022/2910859

Research Article

The Improving Effect of Intelligent Speech Recognition System on English Learning

Qi Luo

Inner Mongolia Vocational and Technical College of Communication, Chifeng 024005, China

Correspondence should be addressed to Qi Luo; 20200062@stu.nun.edu.cn

Received 24 November 2021; Revised 22 December 2021; Accepted 10 January 2022; Published 10 March 2022

Academic Editor: Qiangyi Li

To improve the effect of English learning in the context of smart education, this study combines speech coding to improve the intelligent speech recognition algorithm, builds an intelligent English learning system, combines the characteristics of human ears, and studies a coding strategy of a psychoacoustic masking model based on the characteristics of human ears. Moreover, this study analyzes in detail the basic principles and implementation process of the psychoacoustic model coding strategy based on the characteristics of the human ear and completes the channel selection by calculating the masking threshold. In addition, this study verifies the effectiveness of the algorithm in this study through simulation experiments. Finally, this study builds a smart speech recognition system based on this model and uses simulation experiments to verify the effect of smart speech recognition on English learning. To improve the voice recognition effect of smart speech, band-pass filtering and envelope detection adopt the gammatone filter bank and Meddis inner hair cell model in the mathematical model of the cochlear system; at the same time, the masking effect model of psychoacoustics is introduced in the channel selection stage to prevent noise. Sex has been improved, and the recognition effect of smart voice has been improved. The analysis shows that the intelligent speech recognition system proposed in this study can effectively improve the effect of English learning. In particular, it has a great effect on improving the effect of oral learning.

1. Introduction

For a long period of time in the future, the dominant mode of college English teaching has gradually become clear. That is, college English teaching should complete the transformation from a teacher-centered teaching model that only teaches language knowledge and skills to a student-centered teaching model that not only imparts general language knowledge and skills, but also pays more attention to the cultivation of language use ability and autonomous learning ability [1]. The new teaching model will be supported by multimedia and network technology, supplementing and perfecting a single classroom teaching based on teacher lectures. It is necessary to implement the combination of teachers' classroom teaching and tutoring with computer-based and classroom-based English multimedia teaching and vigorously develop individualized learning and autonomous learning [2].

The signal classification algorithm based on statistical characteristics mainly obtains a series of relevant statistical characteristics or classification characteristic values by calculating the statistical characteristics of the data set and then determines the signal category according to the relevant empirical threshold [3]. The acquisition of this threshold comes from the calculation and analysis of statistical characteristics of a large number of relevant data. After the main distribution range of the corresponding statistical characteristics is obtained, the range is used as the empirical threshold for signal classification, thereby realizing signal classification. Signal classification algorithms based on statistical characteristics also need to calculate feature values of the data. However, unlike other classification algorithms, methods based on statistical characteristics do not have a mature classification model to model feature data, but only rely on data features and the size of the empirical threshold is compared to determine the signal category.

The existing international standards for speech/audio hybrid coding adopt different signal classification and coding method selection algorithms. Among them, the 3GPP AMR-WB + standard includes two frameworks: closed-loop mode and open-loop mode. Different modes use different signal classification logics. However, both modes cannot have both the time complexity of the signal classification algorithm and the classification accuracy rate. Good performance—the closed-loop mode has high coding quality and high computational complexity; the open-loop mode has low computational complexity, but the coding quality is also significantly damaged.

This study analyzes in detail the basic principles and implementation process of the psychoacoustic model coding strategy based on the characteristics of the human ear and completes the channel selection by calculating the masking threshold. In addition, this study verifies the effectiveness of the algorithm in this study through simulation experiments. Finally, this study builds a smart speech recognition system based on this model and uses simulation experiments to verify the effect of smart speech recognition on English learning.

The innovation and contribution of this paper is to improve the speech recognition effect of intelligent speech. In the mathematical model of cochlear system, gamma tone filter bank and Meddis inner hair cell model are used for band-pass filtering and envelope detection. At the same time, psychoacoustics is introduced in the channel selection stage, which improves the masking effect model in terms of anti-noise and improves the recognition effect of intelligent speech.

2. Related Work

With the increasing requirements for the accuracy and time complexity of speech and English classification, in some real-time systems, researchers pay more attention to the balance between classification ability and computational complexity when selecting audio features. Literature [4] analyzed the classification ability and computational complexity of a variety of short-term features and long-term features. The results show that although short-term features are relatively easy to calculate, they have poor classification capabilities, while long-term features are more conducive to classification accuracy, but calculations are relatively time-consuming. Literature [5] tries to use the combination of short-term and long-term features to achieve the best classification efficiency. The study sets up a 250-frame buffer. While calculating the short-term features, once the buffer is filled, the corresponding long-term features are calculated and the classifier for the long-term features is started. Although the research has good classification accuracy, it still cannot meet the demand in terms of time complexity. Literature [6] regards the perceptual linear prediction cepstrum coefficient as one of the short-term features. It uses different classifiers for short-term features and long-term features and finally gets better classification efficiency. However, the algorithm in this study is not applicable to smaller frame lengths, when audio data are divided into smaller frame lengths. Literature [7] borrowed a low-dimensional identity feature from the field of speaker recognition and language recognition. This feature has achieved great success in the above two areas. Moreover, in the study of speech and English classification, this feature is easy to calculate due to its low-dimensional characteristics, but for signals with a shorter frame length, this feature cannot meet the requirements of high classification accuracy and low time complexity at the same time. Literature [8] proposes a hybrid combination of feature extraction methods. In addition to feature extraction from the perspective of one-dimensional audio signal processing, this method also draws on the content of two-dimensional image processing to extract the gray spectrum features of the signal. However, due to the need to calculate more complex two-dimensional features, the time cost of its calculation has increased.

Literature [9] studied the speech and English signal classification algorithms based on statistical methods and neural network methods and designed experiments to implement the Bayesian classifiers and multilayer perceptrons to classify speech and English signals. The experimental results show that the latter has a higher classification accuracy rate, and correspondingly, its computational time complexity has also increased to a certain extent; literature [10] has implemented three algorithms for speech and English signal classification, namely multilayer perceptual neural network, radial basis function neural network, and hidden Markov model, and the experimental results show that the multilayer perceptual neural network has the best classification accuracy among the three network models; literature [11] uses the feedforward neural network in the framework of AMR-WB + Next and classifies the audio data at the 20 ms frame level. Literature [12] proposed a new centroid neural network based on the Bhattacharya kernel and used it for the classification of audio signals. The kernel function maps the input signal to a higher latitude space and uses the Bhattacharya distance as a metric for samples. The experimental results of this algorithm on multiple data sets show that its classification accuracy is better than traditional methods such as self-organizing mapping algorithm; literature [13] proposes a deep network, which is limited by binary random units. The Boltzmann machines are stacked, and the results show that the deep network model has strong generalization ability in the classification of binary speech and English signals; literature [14] introduces the method of artificial neural network into lung sound signals Classification problems are used to help diagnose respiratory diseases. The performance of neural networks on the same data set is comparable and in some cases is better than other classification methods such as Gaussian mixture models; literature [15] is based on deep confidence networks (DBNs) that constructed a speech English classifier, and the network can perform deeper mining of audio features. The research is based on experiments to select the number of hidden layers of DBN and the number of units of hidden layers and finally obtains a better classification accuracy than SVM. Literature [16] is proposed to use a cyclic neural network to identify complex excitations generated by nonperiodic signals and to determine its category in a short period of time; literature

[17] uses a cyclic neural network to learn time-series information, and to predict the future time node, the prediction accuracy of this method when predicting ECG data is higher than that of the traditional regression model.

Literature [18] analyzes the short-term energy of the speech signal, extracts speech features with frame length characteristics, and performs hidden Markov modeling on the entire speech. Model training uses a training set recorded by many speakers and uses statistical theory to solve the difference between the individual and the whole, so that the speaker-independent single-sentence hidden Markov modeling is robust. Literature [19] uses a probability model to describe the pronunciation of statistics. The continuous-state transition in the hidden Markov model can be regarded as the utterance of a short speech segment, that is, a connected HMM, which represents the speech segment. HMM is a method mainly used for speech recognition in recent years. Literature [20] regards HMM as the core of speech recognition. In the process of speech recognition, the system uses the Viterbi algorithm to decode to find out the correct recognition result. Using the hidden Markov modeling on a single voice can describe the relevance of words in each voice. Under the condition that the speaker is fully trained, the speaker-independent short English speech modeling can be achieved more accurately. However, HMM requires a priori statistical knowledge of the speech signal and weak classification decision-making ability, including the complexity calculation of the Viterbi algorithm and the probability distribution problem in the Gaussian mixture model. These shortcomings make it difficult to further improve the recognition performance of a single HMM [21]. Most of the literature in the field of speech recognition has improved the clustering algorithm in HMM and used it as a method of pattern classification to estimate the parameters of the optimized model, but the effect of speech recognition is not ideal. For English speech with a large amount of data and more complex pronunciation changes, the shortcomings of HMM are more obvious, and its recognition time is longer [22]. Literature [23] tried to combine the clustering algorithm with the hidden Markov model based on a single hidden Markov model and applied it to English speech recognition. According to the characteristics of English voice and the similarity between voices, the data set of English voice is divided into several groups, and each group is composed of some voices with similar voice characteristics. Therefore, when recognizing English speech, there is no need to calculate all the speech in the Viterbi decoding, only the HMM parameters in the selected group to which the input speech belongs. In the case of selecting a suitable clustering group, the system will save a lot of calculation and the recognition performance will be greatly improved. This not only provides a new voice recognition reference method for real-time small device applications, but also lays the foundation of voice recognition for a new English voice evaluation system [24].

Through the above research, we can see that the existing English intelligent speech recognition methods have major drawbacks. Compared with English words, more feature data and more complex pronunciation changes make the speech recognition of English speech more difficult. First of all, the vocabulary of English pronunciation is larger, and there are no obvious pauses between the pronounced words. In other words, there is no clear boundary between words. Secondly, the pronunciation of each word in English pronunciation is usually more natural, and the pronunciation of related words is more casual than the pronunciation of isolated words, so the effect of co-pronunciation is more serious. In addition, affected by the context, in the process of English pronunciation, English pronunciation may have differences in pronunciation, rhythm, intonation, stress, and speaking speed. The prosody characteristics are also different. Therefore, this study studies a coding strategy of a psychoacoustic masking model based on the characteristics of the human ear, analyzes in detail the basic principles and implementation process of the coding strategy of a psychoacoustic model based on the characteristics of the human ear, and completes the channel selection by calculating the masking threshold to improve English voice intelligent recognition effect.

3. Intelligent Speech Recognition Based on Human Ear Bionics

The main function of the cochlear basement membrane is to filter the speech signal and decompose the frequency. Biomedical experiments have proved that the sound pressure waves cause vibration in the tympanic membrane of the outer ear, which is then transmitted to the inner ear fluid through the middle ear and propagates along the basement membrane through the fluid. During the propagation of sound waves on the basement membrane, the sound waves of different frame rates correspond to different positions of the basement membrane. Moreover, the amplitudes of sound waves of different frequencies also correspond one to one with the amplitudes of different vibration positions of the basement membrane. Therefore, the basement membrane can decompose the different frequency components and their corresponding amplitudes in the sound and complete the analysis of the sound frequency and intensity of the cochlea.

The density distribution of the characteristic frequency of the basement membrane is nonuniform. Different positions of the human ear basement membrane correspond to different characteristic frequencies. The lower the frequency of the input signal, the closer the vibration position is to the wormhole. The low-frequency band below 800 Hz is basically linearly distributed. However, the higher the frequency of the input sound wave, the closer the peak is to the bottom of the basement membrane, and the high-frequency range is distributed along a logarithmic relationship on the basement membrane. The corresponding relationship between the characteristic frequency f and the basement membrane position x is expressed as follows:

$$f = A\left(10^{a(L-x)} - k\right). \tag{1}$$

Among them, f is the characteristic frame rate, L is the length of the basement membrane, and x is the distance from

the bottom end of the basement membrane to the position of the characteristic frequency f (the bottom end of the basement membrane corresponds to $x = 0$). When the physiological constants $A = 165.4$, $k = 0.88$, $a = 0.06$, and $L = 35$ mm (the length of the basement membrane of the human ear) are introduced into equation (1), the relationship between the characteristic frequency f and the position x of the basement membrane can be obtained, as shown in Figure 1.

It can be seen from Figure 1 that the characteristic frequency at the bottom of the basement membrane is the largest, while the characteristic frequency at the top is the smallest. The position of the basement membrane and the characteristic frequency are distributed in a nonlinear relationship, which is similar to the perception characteristics of the human ear.

The cochlear basilar membrane is similar to a group of parallel band-pass filter banks with different center frequencies. The basilar membrane can decompose the sound signal transmitted to the human ear in the frequency domain according to frequency bands, which is equivalent to multiple band-pass filters. Therefore, the basement membrane can be regarded as a spectrum analyzer functionally, and its mathematical model is shown in Figure 2.

It is known that the impulse response function of the gammatone filter is as follows:

$$g_m(t) = t^{n-1}\exp(-2\pi B_m t)\cos(2\pi f_m t + \phi_m)u(t), \quad 1 \le m \le M. \tag{2}$$

In the formula, when $t < 0$, $u(t) = 0$. On the contrary, when $t > 0$, $u(t) = 1$. Among them, n represents the filter order. A large number of research experiments show that the $n = 4$th-order gammatone filter can better simulate the filtering characteristics of the cochlear basilar membrane. M is the number of filters, and gm represents the initial phase of the filter. f_m is the center frequency of each filter, that is, the characteristic frequency of the basement membrane. The parameter B_m represents the transformation frequency of the characteristic frequency f in the equivalent rectangular bandwidth (ERB) domain, and this parameter determines the attenuation speed of the impulse response. The relationship between f_m and B_m is as follows:

$$B_m = 1.019 \times ERB(f_m). \tag{3}$$

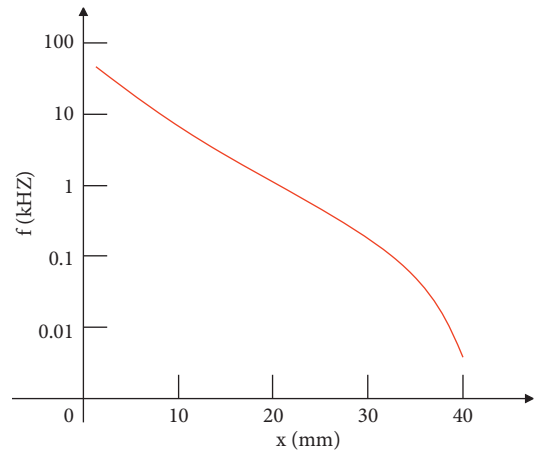

Figure 1: Relationship curve between characteristic frequency f and basement membrane position x.

In the cochlear physiological system, the equivalent rectangular bandwidth $ERB(f_m)$ of each filter can be expressed as follows:

$$ERB(f_m) = 24.7 \times \left(4.37 \times \frac{f_m}{1000} + 1\right). \tag{4}$$

Equation (4) has a nonlinear characteristic, indicating the logarithmic relationship between the equivalent rectangular bandwidth and the center frequency f_m, which is consistent with the auditory characteristics of the human ear.

In the time-domain expression of the gammatone function, if the gain and initial phase are ignored, the gammatone filter can be converted from the continuous domain to the discrete domain. That is, under the premise of not affecting the performance of the filter, we set $b = 2\pi B_m$ and $w_0 = 2\pi f_m$. The impulse response function of the gammatone filter can be simplified as follows:

$$g(t) = t^{n-1}e^{-bt}\cos(w_0 t). \tag{5}$$

Through the Laplace transform of the simplified gammatone function, the s domain (continuous domain) transfer function of the zero-pole form of the fourth-order gammatone function is obtained as follows:

$$G(s) = \frac{[s + b + (\sqrt{2} + 1)w_0][s + b - (\sqrt{2} + 1)w_0][s + b + (\sqrt{2} - 1)w_0][s + b - (\sqrt{2} - 1)w_0]}{[(s + b + jw_0)][(s + b - jw_0)]^4}. \tag{6}$$

Then, the mapping relationship from the s domain to the z domain (discrete domain) is $z = e^{sT}$, where T is the sampling period. We set $a_1 = \cos(\omega_0 T)$, $a_2 = \sin(\omega_0 T)$, $a_3 = e^{-bT}$. From the impulse response invariant method, the corresponding z domain transfer function can be obtained as follows:

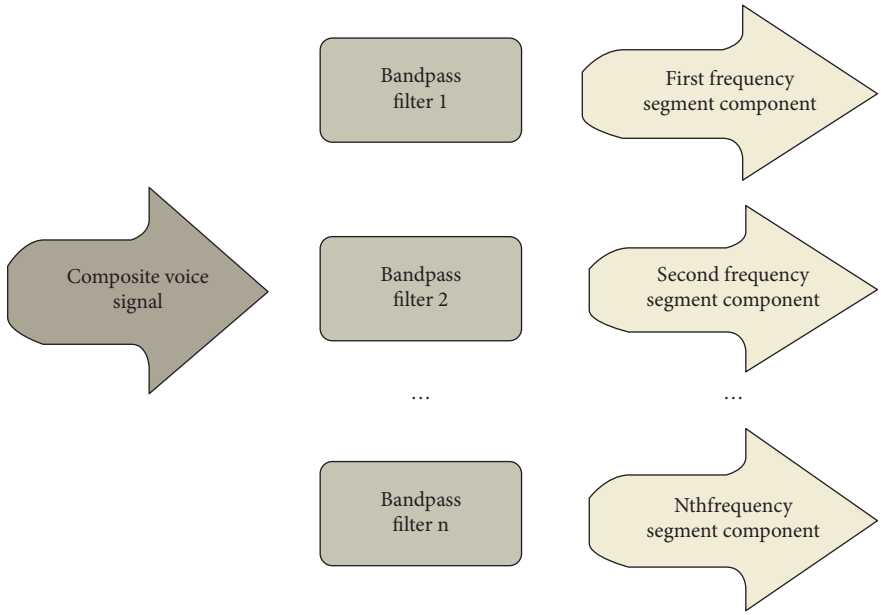

Figure 2: Mathematical model of basement membrane.

$$G(z) = \frac{T - Ta_3\left[a_1 + (\sqrt{2}+1)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}} \times \frac{T - Ta_3\left[a_1 - (\sqrt{2}+1)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}} \times \frac{T - Ta_3\left[a_1 + (\sqrt{2}-1)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}} \times \frac{T - Ta_3\left[a_1 - (\sqrt{2}-1)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}}. \quad (7)$$

Observing the order of the abovementioned denominator, it can be seen that a 4th-order gammatone filter can be realized through an 8th-order z domain transfer function. Then, each gammatone filter is further decomposed into 4 second-order transfer function cascade forms, namely $G(z) = H_1(z) \times H_2(z) \times H_3(z) \times H_4(z)$. The mathematical expressions of the 4 second-order transfer functions are as follows:

$$H_1(z) = \frac{T - Ta_3\left[a_1 + (\sqrt{2}+1)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}},$$

$$H_2(z) = \frac{T - Ta_3\left[a_1 - (\sqrt{2}+1)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}},$$

$$H_3(z) = \frac{T - Ta_3\left[a_1 + (\sqrt{2}-1)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}}, \quad (8)$$

$$H_4(z) = \frac{T - Ta_3\left[a_1 - (\sqrt{2}-1)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}}.$$

Each $H(z)$ expression can be simplified to a standard IIR second-order nodal structure, and its expression is as follows:

$$H(z) = \frac{m_0 + m_1z^{-1}}{1 - n_0z^{-1} - n_1z^{-2}}. \quad (9)$$

Converting the transfer function of equation (9) into a difference equation, we can get the following:

$$y(n) = m_0x(n) + m_1x(n-1) + n_0y(n-1) + n_1y(n-2). \quad (10)$$

Among them, $\{x(n)\}$ is the input sequence, $\{y(n)\}$ is the output sequence, and n and m are the coefficients of the filter. $x(n)$ is the input sample at the nth time, and $y(n)$ represents the output of the second-order nodal filter at the nth time.

After the input sample is processed by the gammatone filter, the solution process of its output can be expressed in Figure 3.

Without affecting the settings of other filters, the advantage of this cascade structure is that the poles and zeros of a certain filter can be individually adjusted. Therefore, this structure is not only convenient for accurately setting the poles and zeros of the filter, but also can adjust the frequency response characteristics of the filter. At the same time, fewer storage units are required for hardware implementation, which greatly reduces the resource requirements for hardware.

The impulse response function gammatone function is directly subjected to Fourier transformation, and the amplitude-frequency response characteristics of the filter can be

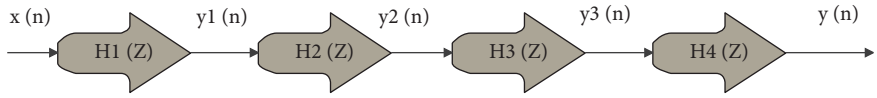

Figure 3: Mathematical solution process diagram of the output sequence of the gammatone filter.

obtained. The amplitude-frequency response curve of the fourth-order gammatone filter under different center frequencies is shown in Figure 4.

It can be found from Figure 4 that the gammatone filter has the largest amplitude at the center frequency, and there are steep edges on both sides of the center frequency, and different center frequencies have different bandwidths. This shows that the gammatone filter has sharp frequency selection characteristics, so the amplitude-frequency response characteristics of the gammatone filter are similar to the filtering characteristics of the basement membrane.

In the gammatone filter, the center frequency determines the bandwidth of the filter. In fact, the bandwidth of the filter needs to be determined according to the actual situation. In the human ear physiological system, the frequency band from 20 Hz to 16 kHz is usually divided into 24 frequency groups, and the critical bandwidth and upper and lower limit frequencies are shown in Table 1.

There are many models of inner hair cells, and the accepted model is the inner hair cell function model proposed by Meddis. The Meddis model describes the process by which cochlear hair cells transmit sound to the auditory nerve. The model includes a firing factory and a free transmission well, which transmit neurotransmitters to the synaptic cleft. This model is very close to the results of real physiological experiments. The inner hair cell model is shown in Figure 5.

The left side of the dotted line in the figure is the inside of the inner hair cell, and the right side is the outside of the inner hair cell.

The permeability $k(t)$ of the inner hair cell permeable membrane is variable. The permeability reflects the ability of neurotransmitters to transfer from inner hair cells to the gap. The permeability of the cell membrane is determined by the amplitude of the instantaneous input sound intensity, which rectifies the output signal of the basement membrane. The mathematical expression is as follows:

$$k(t) = \begin{cases} \dfrac{g[x(t) + A]}{[x(t) + A + B]}, & [x(t) + A] > 0, \\ 0, & [x(t) + A] < 0. \end{cases} \tag{11}$$

In the formula, $x(t)$ is the instantaneous amplitude of the input sound wave; $k(t)$ is the permeability of the cell membrane. g, A, and B are cell parameters. This is a nonlinear process, which describes how the release of neurotransmitters from the free transfer well to the synaptic cleft is affected by the amplitude of the sound waves. When the instantaneous amplitude of the input sound wave is O, $k(t) = \mathrm{Ag}/(A + B)$, which represents the spontaneous response of the cell membrane. The relationship between the instantaneous amplitude $x(t)$ of the acoustic wave and the permeability $k(t)$ is shown in Figure 6.

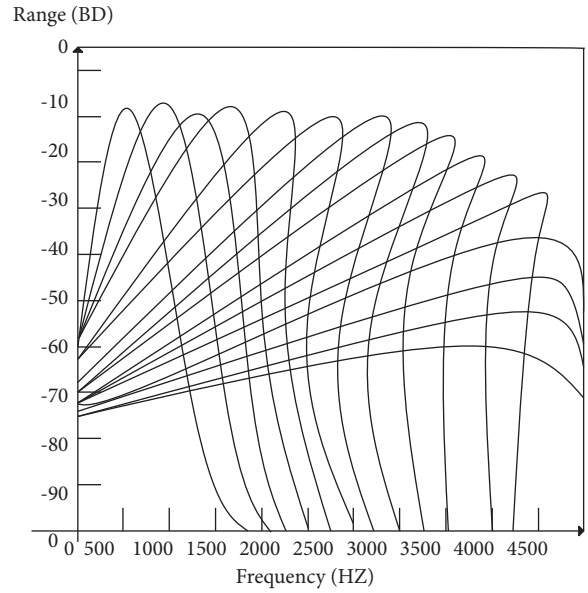

Figure 4: Gammatone function amplitude-frequency response curve at different center frequencies.

3.1. The Neurotransmitter Quality in Inner Hair Cells. The neurotransmitter $q(t)$ in the inner hair cell always exists, and its rate of change over time is shown as follows:

$$\frac{\mathrm{d}q(t)}{\mathrm{d}t} = y(1 - q(t)) - k(t)q(t) + xw(t). \tag{12}$$

Among them, $y(1 - q(t))$ represents the amount of neurotransmitters that the factory slowly replenishes to the inner hair cells, and $k(t)q(t)$ is the amount of neurotransmitters lost to the synaptic cleft by the inner hair cells. $\mathrm{xw}(t)$ is the amount of neurotransmitter that continuously flows back to the inner hair cells through the warehouse per unit time. Together, they determine the rate of change in neurotransmitter in inner hair cells over time.

3.2. The Neurotransmitter Quality in the Synaptic Cleft. The change rate of the neurotransmitter $c(t)$ in the synaptic cleft over time can be described by the following equation:

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = k(t)q(t) - lc(t) - rc(t). \tag{13}$$

Among them, $k(t)q(t)$ has the same meaning as the above formula, $\mathrm{lc}(t)$ is the amount of neurotransmitter lost to other nervous systems in the synaptic cleft, and $\mathrm{rc}(t)$ is the amount of neurotransmitter that returns to the warehouse from the synaptic cleft. These three items together determine the rate of change in neurotransmitter in the synaptic cleft over time.

Table 1: Filter bandwidth parameter table (frequency: Hz).

Serial number	Center frequency	Bandwidth	Lower limit frequency	Up and down frequency
1	50.0	80.0	20.0	100.0
2	150.0	100.0	100.0	200.0
3	250.0	100.0	200.0	300.0
4	350.0	100.0	300.0	400.0
5	450.0	110.0	400.0	510.0
6	570.0	120.0	510.0	630.0
7	700.0	140.0	630.0	770.0
8	840.0	150.0	770.0	920.0
9	1000.0	160.0	920.0	1080.0
10	1170.0	190.0	1080.0	1270.0
11	1370.0	210.0	1270.0	1480.0
12	1600.0	240.0	1480.0	1720.0
13	1850.0	280.0	1720.0	2000.0
14	2150.0	320.0	2000.0	2320.0
15	2500.0	380.0	2320.0	2700.0
16	2900.0	450.0	2700.0	3150.0
17	3400.0	550.0	3150.0	3700.0
18	4000.0	700.0	3700.0	4400.0
19	4800.0	900.0	4400.0	5300.0
20	5800.0	1100.0	5300.0	6400.0
21	7000.0	1300.0	6400.0	7700.0
22	8500.0	1800.0	7700.0	9500.0
23	10500.0	2500.0	9500.0	12000.0
24	13500.0	3500.0	12000.0	15500.0

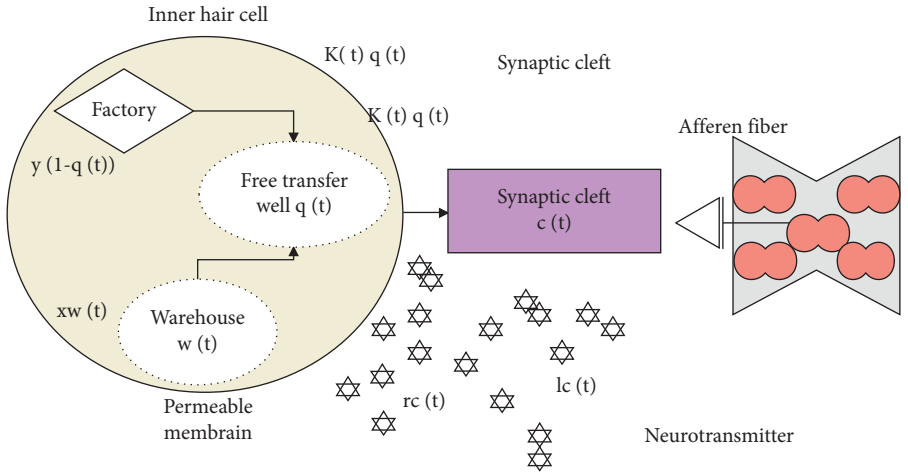

Figure 5: Inner hair cell model.

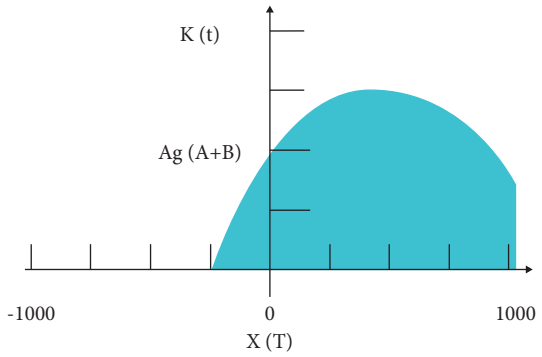

Figure 6: Cell membrane permeability change curve with instantaneous acoustic wave amplitude.

Table 2: Meddis model parameters of inner hair cells.

Parameter	Describe	Unit	Numerical value
A	Penetration constant	1/s	2.00
B	Penetration constant	1/s	300.00
L	Loss rate	1/s	2500.00
R	Recovery rate	1/s	6580.00
X	Reprocessing rate	1/s	66.31
Y	Replenishment rate	1/s	8.00
G	Release rate	1/s	2000.00

3.3. The Quality of Neurotransmission in the Warehouse. The change rate of the neurotransmitter quality $w(t)$ in the warehouse with time can be expressed by the following equation:

$$\frac{dw(t)}{dt} = rc(t) - xw(t). \tag{14}$$

The biological experiment gives the physiological constant parameters in the above formulas, as shown in Table 2. This table is the result of the actual measurement of the human ear, and the parameters are adjusted according to the specific signal waveform conditions in the application. Therefore, the Meddis model of inner hair cells can be described by formulas (11), (12), (13), and (14).

In a quiet environment, the minimum sound intensity that the human ear can just feel is called silent listening, or absolute listening threshold, which changes with frequency. The frequency range of human ears is about 20 Hz~18 kHz. The frequency resolution of the human ear is better at low frequencies than high frequencies. It is especially sensitive to sound signals near 2 kHz to 4 kHz, and it is dull to sounds that are too low or too high. Therefore, the absolute listening threshold is correspondingly smaller in this frame rate range. When a single-frequency tone with an intensity of 60 dB and a frequency of 1 kHz appears, the masking threshold curve will change, as shown in Figure 7.

It can be seen from the above figure that the parts below 0.5 kHz and above 5 kHz are relatively far away from 1 kHz, and the auditory masking threshold is not affected and remains unchanged. In other words, the total masking threshold curve coincides with the absolute listening threshold curve. Between 0.5 kHz and 5 kHz, a new masking curve will be formed. The sound events below the masking curve cannot be heard because they are masked by the 60 dB strong 1 kHz signal, so they do not need to be processed. At this time, we call the 1 kHz single-frequency tone as the masking tone, and the sound events below the masking curve are called the masked tone. If you want to hear the abovementioned 1 kHz signal and another 2 kHz signal at the same time, it can be seen from the figure that the signal strength of 2 kHz must be above 40 dB.

The coding strategy of the psychoacoustic model based on the characteristics of the human ear is to select the speech signal of N frequency bands with the relevant meaning of the speech information in each cycle. To further understand the psychoacoustic model of the coding strategy, we take the implementation process of the two channels Zi and Zj as an example to briefly introduce, as shown in Figure 8. Firstly,

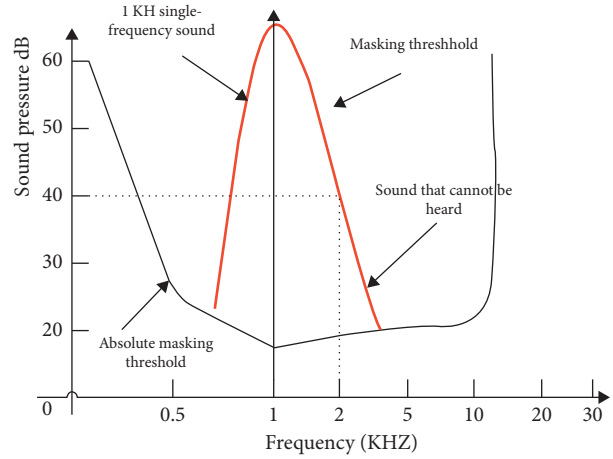

Figure 7: Spectrum masking characteristics.

the masking threshold of each channel is calculated separately, and then, the masking thresholds and absolute listening thresholds of these different channels are nonlinearly superimposed to calculate the combined masking threshold, which forms a psychoacoustic masking model.

After the introduction of the psychoacoustic masking effect model, a new N-of-M-type strategy emerged. This new strategy's selection of N channels is different from the previously introduced ACE strategy. The basic principle of the speech coding strategy based on the psychoacoustic model of the human ear is shown in Figure 9.

It can be seen from the principle block diagram that the processing process of the previous frequency band division and envelope detection is different from the method adopted by the ACE strategy. Band-pass filtering and envelope detection are implemented using gammatone filter bank and Meddis inner hair cell model, respectively.

When the sound is so weak that human ears can just hear it, we call the sound intensity at this time the "listening valve." The absolute listening valve refers to the minimum value of pure tone that can be heard by human ears in a quiet environment. The researchers found that the listening valve changes with the frequency of the sound. The function calculation formula of the absolute listening valve Tabs(f) is shown as follows:

$$\mathrm{Tabs}(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6((f/1000)-3.3)^2} + 10^{-3}\left(\frac{f}{1000}\right)^{4}. \tag{15}$$

In the formula, f represents the center frequency, and the unit of Tabs(f) is dB.

The sensitivity of the human ear to different frequencies varies greatly. Among them, it is most sensitive to signals in the range of 2 kHz to 4 kHz, and even signals with very low amplitude can be heard by the human ear. In the low-frequency and high-frequency areas, the signal amplitude that can be heard by the human ear needs to be much higher. In this experiment, we use the function Las(z) to represent the

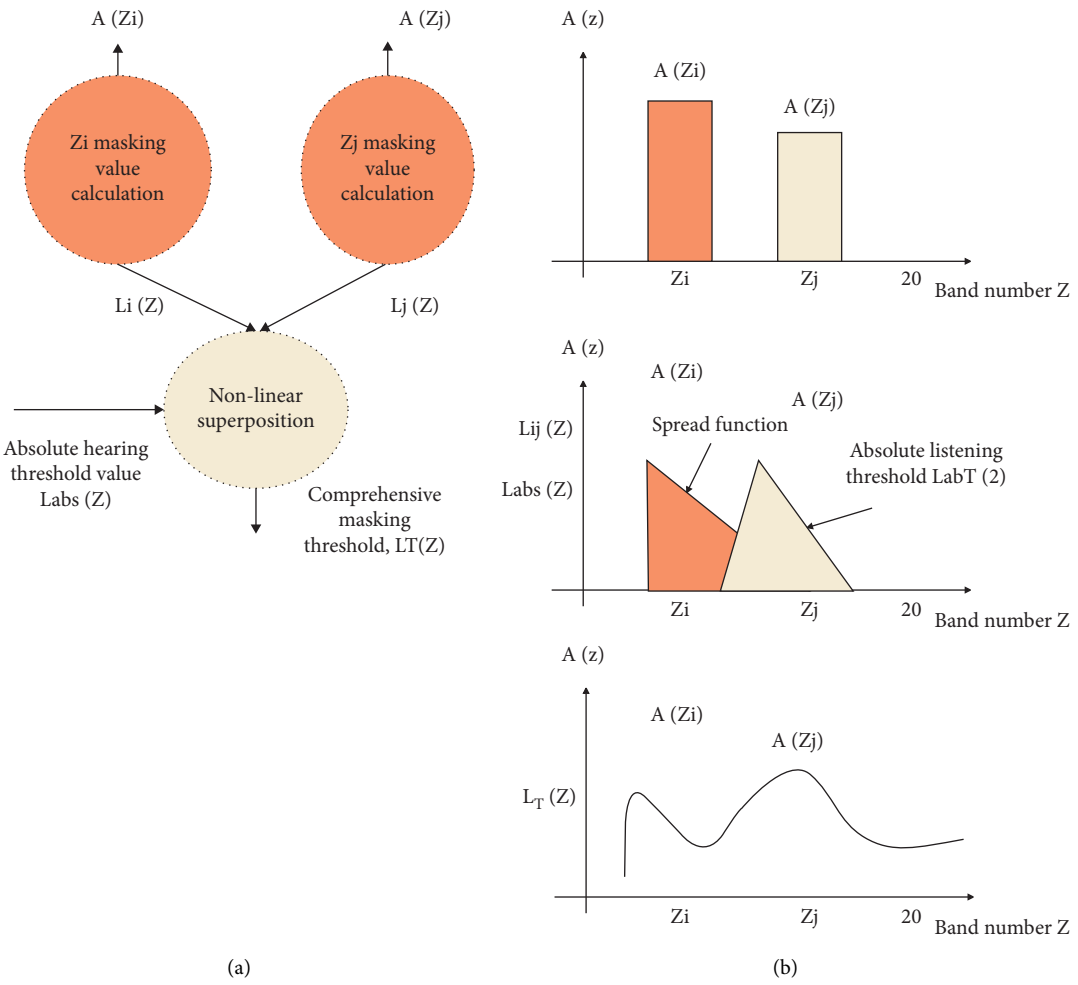

Figure 8: Psychoacoustic masking model.

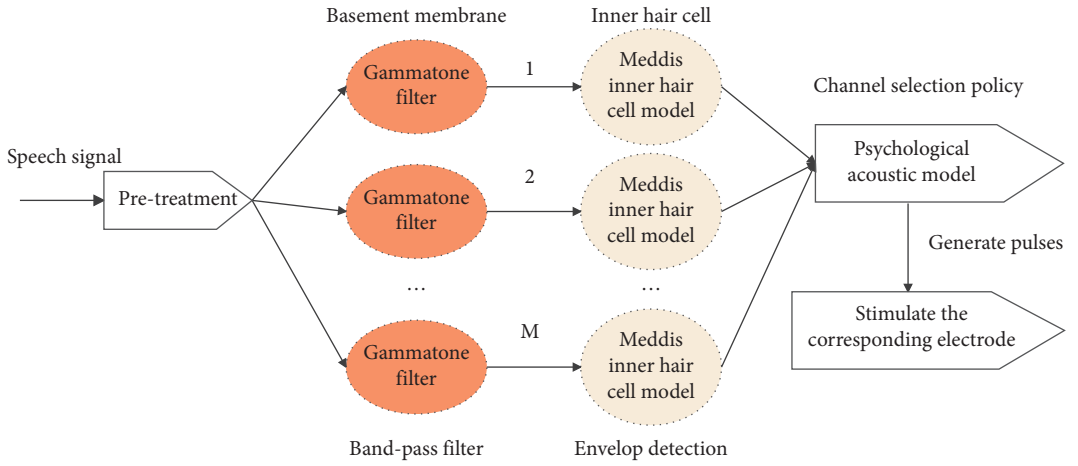

Figure 9: Principle block diagram of a psychoacoustic model coding strategy based on the characteristics of the human ear.

Table 3: English speech recognition effect.

Number	Method of this article	Method of literature [22]	Number	Method of this article	Method of literature [22]
1	93.5	83.86	24	96.0	82.96
2	95.8	83.86	25	94.7	90.79
3	93.5	79.63	26	93.6	81.53
4	94.8	89.32	27	95.5	90.83
5	96.6	86.24	28	93.7	91.45
6	96.2	87.95	29	93.6	80.09
7	96.3	86.55	30	95.9	90.93
8	94.6	91.37	31	96.3	85.16
9	95.9	93.98	32	95.5	93.20
10	93.9	83.89	33	94.6	86.83
11	95.0	85.50	34	96.4	88.37
12	95.3	90.00	35	96.0	81.89
13	95.9	82.42	36	95.2	86.71
14	96.6	93.77	37	96.7	89.92
15	94.5	90.59	38	96.9	93.67
16	95.3	93.01	39	93.0	82.91
17	96.4	88.78	40	94.4	87.92
18	94.5	92.23	41	94.5	80.77
19	93.9	86.58	42	95.6	88.83
20	95.8	87.30	43	95.4	91.67
21	94.6	87.76	44	93.2	89.45
22	94.7	83.42	45	95.3	82.93
23	96.2	90.50			

absolute threshold value in the 22 different center frequency channels z, which is determined by the specific center frequency in Table 3.

The function expression of the extension function $L(z)$ is as follows:

$$L_i(z) = \begin{cases} A(z_i) - a_v - s_i \cdot (z_i - z), & z < z_i, \\ A(z_i) - a_v - s_i \cdot (z_i - z), & z \geq z_i. \end{cases} \tag{16}$$

Among them, z represents the number of critical bands, $1 \leq z \leq M$, and i represents the number of selected channels. The spread function $L(z)$ of the frequency band z is shown in Figure 10.

A power law superposition model is proposed and applied to the superposition of different masking thresholds to express nonlinear superposition. This power law model is determined by the parameter α, $0 < a \leq 1$. If α is 1, the superposition is linear. If α is less than 1, then the superposition is nonlinear. In this experiment, $\alpha = 0.25$. Then, the nonlinear superposition function $I(z)$ of sound intensity is shown as follows:

$$I_T(z) = \left[[I_{abs}(z)]^{\alpha} + \sum_i [I_i(z)]^{\alpha} \right]^{1/\alpha}. \tag{17}$$

Among them, the functional expression of sound intensity is as follows:

$$\begin{aligned} I_{abs}(z) &= 10^{L_{ahs}(z)/10}, \\ I_i(z) &= 10^{L_i(z)/10}. \end{aligned} \tag{18}$$

In formula (18), $I_{abs}(z)$ represents the sound intensity of the absolute listening valve, and $I_i(z)$ represents the sound intensity of the zth frequency band. Finally, the superimposed masking threshold is denoted by $L_T(z)$:

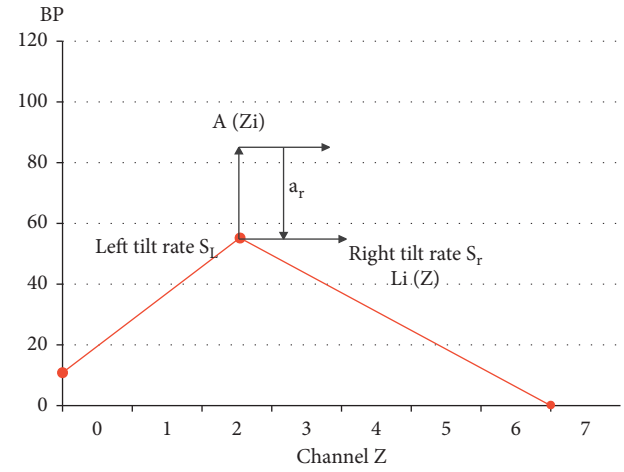

Figure 10: Spread function of frequency band z.

$$L_T(z) = 10 \log_{10}(I_T(z)). \tag{19}$$

3.4. Channel Selection. The coding strategy selects N channels in a loop iteration, and the psychoacoustic model is used in each iteration calculation. The principle is shown in Figure 11.

After the speech signal passes through the band-pass filter and envelope detection, the envelope information $A(z)$ ($z = 1,\ldots,M$) of $M = 22$ frequency band signals is obtained. In the first iteration, there is no masking threshold, and we do not consider the absolute listening valve. The selection of the first channel is based on the maximum amplitude. In this channel, the masking threshold $L(z)$ ($z = 1,\ldots,M$) is calculated according to the psychoacoustic model. In the

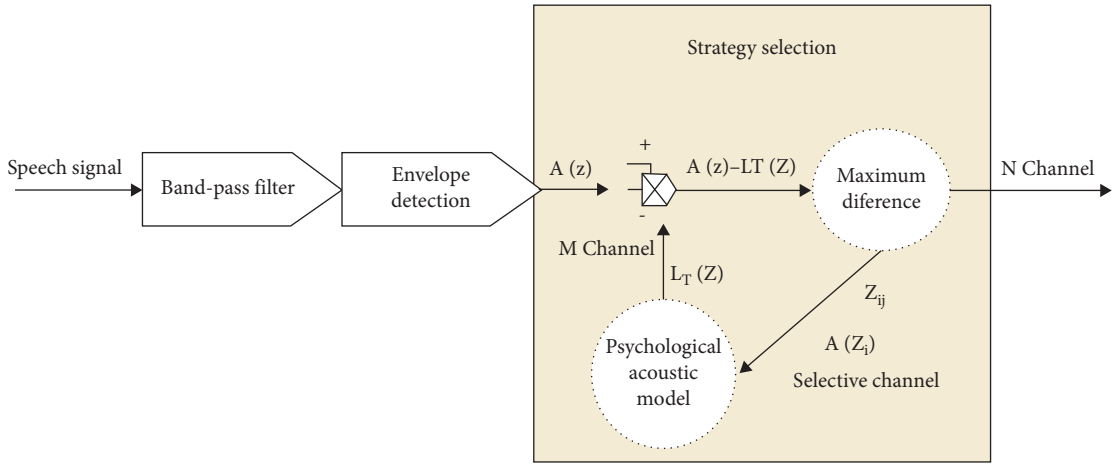

Figure 11: Channel selection.

remaining $M-1$ channels, we will iteratively select the channel z_i according to the maximum difference in the following equation:

$$z_i = \arg\max\left(A(z) - L_T(z)\right), \quad z = 1, \ldots, M. \tag{20}$$

The masking threshold $L(z)$ of each frequency band is related to the previous masking threshold. The masking threshold $L(z)$ is obtained after multiple iterations. After a series of cycles, N channels are finally selected to generate pulses.

4. The Improving Effect of Intelligent Speech Recognition System on English Learning

The system drives the intelligent mechanical equipment to perform corresponding operations through the control module. The implementation process of system voice recognition control is shown in Figure 12.

The structure of the hardware part of the embedded English speech recognition control system is shown in Figure 13.

After obtaining the above intelligent speech recognition system, the system is constructed through the simulation system, and the reliability of the system proposed in this study is explored through experimental teaching methods, and the speech recognition effect of the intelligent speech recognition system proposed in this study is verified, and the results are compared with the literature [22], and the results are as follows.

From the above research, we can see that the intelligent speech recognition system proposed in this study can effectively recognize English speech in students' English learning. After that, this study analyzes the improving effect of the system on English learning and obtains the results shown in Table 4.

From the above research, we can see that the intelligent speech recognition system proposed in this study can effectively improve the effect of English learning. In particular, it has a great effect on improving the effect of oral learning.

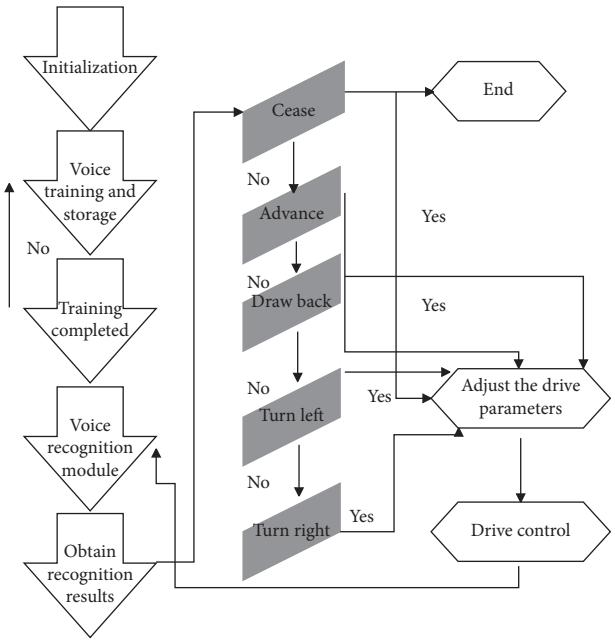

Figure 12: Realization process of system speech recognition control.

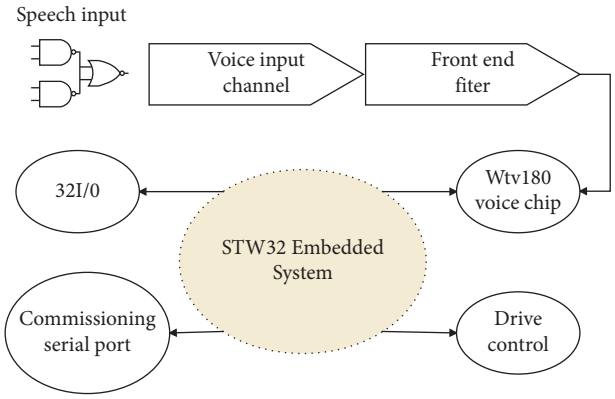

Figure 13: System hardware part structure.

Table 4: Evaluation of the improving effect of English learning.

Number	Method of this article	Method of literature [22]	Number	Method of this article	Method of literature [22]
1	65.7	62.85	24	72.2	65.52
2	73.8	63.99	25	67.9	63.79
3	73.2	65.13	26	64.1	61.85
4	70.1	66.50	27	73.3	66.68
5	62.0	60.10	28	71.0	65.94
6	71.5	69.47	29	61.1	56.02
7	76.2	66.47	30	68.2	59.31
8	69.2	61.39	31	69.8	62.97
9	71.1	63.97	32	61.9	60.59
10	76.1	66.69	33	72.5	62.31
11	72.4	64.01	34	62.3	58.93
12	65.7	61.06	35	75.8	65.03
13	68.3	60.71	36	74.5	67.25
14	73.6	62.65	37	62.8	53.76
15	71.3	67.66	38	72.9	65.87
16	61.1	53.13	39	67.4	65.27
17	67.5	59.81	40	76.9	68.21
18	73.3	69.40	41	68.9	65.60
19	72.3	64.48	42	70.6	69.09
20	66.2	62.98	43	64.1	55.47
21	64.3	59.26	44	64.8	63.17
22	70.2	66.63	45	64.6	56.28
23	66.5	64.07			

5. Conclusion

With the continuous development of informatization in colleges and universities, English teaching has also gradually realized informatization. In this environment, it is very important to realize computer-assisted teaching. Aiming at the defects of the current mainstream English speech recognition system, this study proposes an improved algorithm based on human ear simulation for the realization of the English speech recognition system. The experimental results show that the system improved according to the method in this study not only improves the recognition rate of the system, but also can accurately and automatically recognize English speech. Through the planning of the program, the purpose of automatic recognition of English speech rationality is finally achieved, and an automatic recognition model of English speech rationality has been successfully created. Moreover, through experiments, it can be confirmed that the model has a higher recognition rate. The analysis shows that the intelligent speech recognition system proposed in this study can effectively improve the effect of English learning. In particular, it has a great effect on improving the effect of oral learning.

In addition to information in the time domain and frequency domain, the audio signal also includes the phase relationship. The electronic cochlear coding strategy proposed in this study does not discuss the phase relationship of the signal. The introduction of phase information into the future electronic cochlear coding strategy will be more in line with the characteristics of human hearing. Due to the quantization error of the hardware system and other issues, the experimental data results are worse than the actual effect, which affects the speech recognition effect to a certain extent. The algorithm of this article can be further expanded in the follow-up to increase the scope of research.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This study was sponsored by the Inner Mongolia Vocational and Technical College of Communication.

References

[1] P. H. Kumar and M. N. Mohanty, "Efficient feature extraction for fear state analysis from human voice," *Indian Journal of Science & Technology*, vol. 9, no. 38, pp. 1–11, 2016.

[2] R. Rhodes, "Aging effects on voice features used in forensic speaker comparison," *International Journal of Speech Language and the Law*, vol. 24, no. 2, pp. 177–199, 2017.

[3] M. Sarria-Paja, M. Senoussaoui, and T. H. Falk, "The effects of whispered speech on state-of-the-art voice based biometrics systems," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, no. 1, pp. 1254–1259, Halifax, Canada, May 2015.

[4] A. Leeman, H. Mixdorff, M. O'Reilly, M.-J. Kolly, and V. Dellwo, "Speaker-individuality in Fujisaki model f0 features: implications for forensic voice comparison," *International Journal of Speech Language and the Law*, vol. 21, no. 2, pp. 343–370, 2015.

[5] A. K. Hill, R. A. Cárdenas, J. R. Wheatley et al., "Are there vocal cues to human developmental stability? Relationships between facial fluctuating asymmetry and voice

attractiveness," *Evolution and Human Behavior*, vol. 38, no. 2, pp. 249–258, 2017.

[6] M. Woźniak and D. Połap, "Voice recognition through the use of Gabor transform and heuristic algorithm," *Nephron Clinical Practice*, vol. 63, no. 2, pp. 159–164, 2017.

[7] T. Haderlein, M. Döllinger, V. Matoušek, and E. Nöth, "Objective voice and speech analysis of persons with chronic hoarseness by prosodic analysis of speech samples," *Logopedics Phoniatrics Vocology*, vol. 41, no. 3, pp. 106–116, 2015.

[8] S. S. Nidhyananthan, K. Muthugeetha, and V. Vallimayil, "Human recognition using voice print in LabVIEW," *International Journal of Applied Engineering Research*, vol. 13, no. 10, pp. 8126–8130, 2018.

[9] F. L. Malallah, K. N. Y. M. G. Saeed, S. D. Abdulameer, and A. W. Altuhafi, "Vision-based control by hand-directional gestures converting to voice," *International Journal of Scientific & Technology Research*, vol. 7, no. 7, pp. 185–190, 2018.

[10] S. Morgan, "Contact effects on voice-onset time in Patagonian Welsh," *Acoustical Society of america Journal*, vol. 140, no. 4, p. 3111, 2016.

[11] G. Mohan, K. Hamilton, A. Grasberger, A. C. Lammert, and J. Waterman, "Realtime voice activity and pitch modulation for laryngectomy transducers using head and facial gestures," *Journal of the Acoustical Society of America*, vol. 137, no. 4, p. 2302, 2015.

[12] T. G. Kang and N. S. Kim, "DNN-based voice activity detection with multi-task learning," *IEICE - Transactions on Info and Systems*, vol. E99.D, no. 2, pp. 550–553, 2016.

[13] H.-N. Choi, S.-W. Byun, and S.-P. Lee, "Discriminative feature vector selection for emotion classification based on speech," *The Transactions of The Korean Institute of Electrical Engineers*, vol. 64, no. 9, pp. 1363–1368, 2015.

[14] C. T. Herbst, S. Hertegard, and D. Zangger-Borch, "Freddie Mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics," *Logopedics Phoniatrics Vocology*, vol. 42, no. 1, pp. 1–10, 2016.

[15] J. Al-Tamimi, "Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: implications for formal representations," *Laboratory Phonology*, vol. 8, no. 1, pp. 1–40, 2017.

[16] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[17] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 24, no. 7, pp. 1315–1329, 2016.

[18] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[19] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[20] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[21] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: a survey," *Speech Communication*, vol. 56, no. 3, pp. 85–100, 2014.

[22] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[23] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, no. 3, pp. 535–557, 2017.

[24] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.