Hindawi
Mobile Information Systems
Volume 2022, Article ID 3192892, 11 pages
https://doi.org/10.1155/2022/3192892

Research Article

The Application of Intelligent Speech Analysis Technology in the Spoken English Language Learning Model

Min Zhu

School of Foreign Languages, Xi'an Aeronautical University, Xi'an 710077, China

Correspondence should be addressed to Min Zhu; 201307015@xaau.edu.cn

Received 24 February 2022; Revised 17 March 2022; Accepted 16 April 2022; Published 9 May 2022

Academic Editor: Chia-Huei Wu

In order to improve the effect of spoken English processing, it is necessary to improve the spoken English processing technology from the perspective of the characteristics of spoken English, combined with intelligent algorithms. This paper combines the intelligent speech analysis technology to improve the spoken English recognition technology and combines the actual and needs of English learning to improve the system algorithm. Moreover, this paper combines the intelligent speech analysis to construct the intelligent spoken English learning model structure and combines the statistical method and the intelligent evaluation method to analyze the model effect. After obtaining the system function structure, this paper designs experiments to verify the effect of the model proposed in this paper. From the experimental analysis results, it can be seen that the intelligent English speech analysis model proposed in this paper can play an important role in the learning of spoken English.

1. Introduction

Spoken English signal processing has been widely used in many aspects such as English oral enhancement, English oral recognition, English oral coding, and English oral communication. It provides great help for people to obtain information conveniently, accurately, and quickly [1]. Spoken English enhancement technology, as an extremely important and basic branch of the signal processing of spoken English, has a positive impact on spoken English recognition and spoken English coding. Moreover, it has developed rapidly in the past few decades, and many smart spoken English devices use English spoken language enhancement technology to reduce noise in noisy spoken English that is disturbed by noise. The enhanced performance of these smart spoken English devices directly affects the further processing of spoken English [2]. However, in the actual environment, the types of noise are different and constantly changing. For a long time, it has not been determined which spoken English enhancement technology has the best function and the best objective evaluation method for spoken English. Therefore, researching a reliable and practical spoken English evaluation technology has become a very important topic in the signal processing of spoken English [3].

Oral English is usually the main medium for the exchange of information between people or between people and machines. However, in real life, the spoken English signal will inevitably be interfered by background noise. These noises include the noise of the surrounding environment, the noise of the transmission medium, the noise in the equipment, and the interference of other speakers' spoken English. It is the existence of these noises that have largely destroyed the acoustic characteristics and model parameters of the original spoken English signal and caused the performance of the spoken English transmission system to deteriorate sharply, resulting in that the signal received at the receiving end is not the pure English that people want. The spoken signal is a noisy spoken English signal mixed with noise. For example, when communication equipment is used in public places such as noisy hypermarkets, bustling streets, and train stations, useful spoken English signals are often submerged in the noise around these places, thereby greatly

reducing the quality of spoken English. It affects the normal oral English communication; in addition, if you are in a high-intensity background noise environment for a long time, the listener will also experience auditory fatigue. Therefore, in order for people to communicate normally and the spoken English signal processing system can work effectively when the background noise is widespread, it is necessary to perform special processing on the received noisy spoken English signal, that is, noise suppression.

This paper combines intelligent speech analysis to study the spoken English learning model, constructs an intelligent model, and designs experiments to verify the effect of the intelligent speech model, which provides a theoretical reference for subsequent spoken English recognition and spoken English learning.

2. Related Work

The noise in the shortwave channel greatly reduces the quality of the voice signal passing through the shortwave channel and also limits the effectiveness of the shortwave communication. Therefore, it is necessary to perform voice enhancement on the shortwave voice signal with noise. The purpose of voice enhancement is to reduce the noise of shortwave voice signals with noise, minimize voice signal distortion, and improve voice quality and intelligibility. Currently, the most widely used voice enhancement algorithm is spectral subtraction [4]. The pure voice signal is obtained by subtracting the voice of the noisy voice signal from the voice of the noisy voice signal. According to the investigation of literature [5], the characteristics of noise mainly come from the silent period of the voice signal, and it is easier to estimate the spectral amplitude of the noise signal than its amplitude and phase. The spectral subtraction method is to estimate by subtracting the silent period of the noisy voice signal. The noise amplitude spectrum is used to obtain the short-term amplitude spectrum of the voice signal. Other methods apply the spectral subtraction filter to the noisy voice signal, so that the amplitude spectrum of the enhanced voice signal can be obtained [6]. Appropriate selection of the filter parameters can minimize the difference between the enhanced voice and the original clean voice. Spectral subtraction is mainly used to reduce additive noise, but when the signal-to-noise ratio is low, the disadvantage of spectral subtraction is that it introduces musical noise. Literature [7] all proposed enhancement algorithms based on speech generation models. The former uses the maximum a posteriori criterion to estimate the all-pole parameters, and the latter improves on the former, adding restrictions on speech in the iterative process of the algorithm, further improving the noise reduction effect. Literature [8] proposed a voice enhancement algorithm based on signal subspace. The former uses orthogonal transform to divide the vector space of noisy voice signals into noise subspace and signal plus noise subspace. Finally, in the signal, the voice signal is estimated in the noise-added subspace. According to this principle, the human ear auditory model calculates the masking threshold, and the human ear can tolerate the noise below the masking threshold. Therefore, spectral subtraction based on masking threshold came into being [9]. Aiming at the voice signal in a non-stationary environment, literature [10] proposed an improved voice enhancement algorithm based on logarithmic spectrum estimation, which can effectively improve the voice quality, but the computational complexity is relatively high. Most of the above voice enhancement methods can improve the quality of the voice signal while minimizing the loss of the intelligibility of the voice signal. However, after the voice signal passes through the shortwave channel, multipath fading will occur, and the single-channel voice enhancement effect still cannot meet the actual demand under the shortwave channel. Therefore, it is necessary to adopt new technologies to improve the voice quality. In shortwave wireless communication, diversity combining technology can effectively combat multipath fading. The prerequisite for multichannel voice signal merging is multichannel voice synchronization. Because there is currently no relevant literature to merge multiple channels of analog voice signals, there is currently no literature on the synchronization of multiple voice signals. There are many methods of diversity combining, among which the most widely used are maximum ratio combining, selective combining, and equal gain combining [11]. These merging methods are equivalent to performing linear dimensionality reduction operations [12], and the three merging methods have their own advantages and disadvantages. When the channel estimation is accurate, the performance of maximum ratio combining is the best, and the output signal-to-noise ratio is the largest, but the complexity is high. It is necessary to continuously estimate the fading and phase of each branch signal; equal gain combining is compared with maximum ratio combining and only needs to estimate the instantaneous phase of each branch the complexity is lower than the maximum ratio combination, easy to implement, and the performance is slightly inferior to the maximum ratio combination, but when the quality of the voice signals of each channel is uneven, the performance is poor [13]; choose the combination structure. The simplest is that you only need to continuously monitor the performance of each branch and select the branch with the best performance as the output, but only one channel of the signal is selected for merging, and the performance is poor, and the remaining receivers do not contribute, resulting in a waste of resources [14]. In addition to the above-mentioned simple merging methods, diversity merging methods also include mixed merging methods [15]. The hybrid combining method combines two or more simple combining methods to combine signals. Compared with selective combining, the hybrid combining method has better combining performance, but compared with the maximum ratio combining, it has poor performance and computational complexity. Higher [16].

3. Intelligent English Speech Analysis Technology

There are currently three typical voice perception methods: energy detection, matched filter detection, and cyclostationary detection. First, we build a system signal model. There

are only two possibilities for the result of the judgment: existence and nonexistence. H_0 represents the absence of the signal, and H_1 represents the presence of the signal. The specific signal model construction is expressed as [17]

$$r(t) = \begin{cases} n(t), & H_0, \\ x(t) + n(t), & H_1, \\ t = 1, 2, \cdots . N. \end{cases} \quad (1)$$

In the formula, $r(t)$ is the received signal; $x(t)$ is the signal with a center frequency spectrum of fc and an average power of Es; $n(t)$ is the Gaussian white noise with a mean value of 0 and a bilateral power spectral density of N_0; N is the number of sampling points of the received signal.

As for how to perceive whether the primary user is working, the false alarm probability, missed detection probability, detection probability, etc., are generally used as the basis for detection. We usually use D_0 to indicate that there is no authorized user in the detection frequency band. D_1 is used to indicate the presence of authorized users in the detection frequency band. Similarly, we use H_0 to indicate that there are no authorized users in the actual frequency band and use H_1 to indicate that there are authorized users in the actual frequency band.

The probability of detection formula is

$$P_d = P(D_1|H_1). \quad (2)$$

That is, the original main user works in the frequency band allocated by him, and the working status of the main user can be accurately judged through the detection algorithm. Generally, it is stipulated to use the detection probability P_d to describe. In order to improve the accuracy of the perception system, the detection probability P_d should be increased accordingly.

The formula for false alarm probability is [18]

$$P_f = P(D_1|H_0). \quad (3)$$

That is, the original main user did not work in the frequency band allocated by him, and the working status of the main user was incorrectly determined by the detection algorithm, which resulted in a "false alarm" judgment. In order to improve the effective use of voice in the perception system, the detection probability P_f should be reduced accordingly.

The formula for the probability of missed detection is

$$P_m = P(D_0|H_1). \quad (4)$$

That is, the original main user works in the frequency band allocated by himself, but the working state of the main user is not detected by the detection algorithm, which leads to a "missing detection" judgment and interferes with the normal communication of authorized users. In order to reduce the interference to authorized users, the probability of missed detection P_m should be reduced accordingly.

Figure 1 shows the overall process of energy detection. The collected signal $r(t)$ first uses a band-pass filter to filter out the frequency bands that cognitive users need to detect and then enters the energy meter to obtain the energy value. After a period of time, the required statistics are collected and accumulated. By comparing the final value with the previously specified threshold, it can be judged whether the main user is working.

The principle formula of the energy detection method is as follows [19]:

$$E\left\{(x(t) + n(t))^2\right\} = E\left\{x(t)^2\right\} + E\left\{n(t)^2\right\} \geq E\left\{n(t)^2\right\}. \quad (5)$$

Because when the main user is in the working state, the collected signal energy must theoretically exceed the energy when the main user is in the idle state; that is, the collected signal is only noise. When the final statistic is higher than the preset threshold, it indicates that the current primary user signal is in the allocated frequency band. If it is lower than the preset threshold, then it indicates that the current primary user signal is not in the allocated frequency band.

In the additive white Gaussian voice (AWGN) channel, it is assumed that the channel noise is Gaussian white noise with bilateral power spectral density N_0 and bandwidth W. The test statistic V for energy detection is [20]

$$V(k) = \frac{1}{N_0} \int_0^T r^2(t) dt. \quad (6)$$

According to Shannon's sampling theorem, the formula for noise is

$$n(t) = \sum_{i=\infty}^{\infty} a_i \sin\left[c(2Wt - i)\right]. \quad (7)$$

Among them,

$$a_i = n\left(\frac{i}{2W}\right) \quad \sin cx = \frac{\sin \pi x}{\pi x}. \quad (8)$$

On $(0, T)$, $n(t)$ can be approximated by the number of sampling points of $2TW$:

$$n(t) = \sum_{i=t}^{2TW} a_i \sin\left[c(2Wt - i)\right], 0 < t < T. \quad (9)$$

Therefore, its energy on $(0, T)$ can be expressed as [21]

$$\int_0^T n^2(t) dt = (1/2W) \sum_{i=1}^{2TW} a_i^2. \quad (10)$$

Similarly, for signal $x(t)$, there is:

$$x(t) = \sum_{i=t}^{2TW} \alpha_i \sin\left[c(2Wt - i)\right], 0 < t < T \quad (11)$$

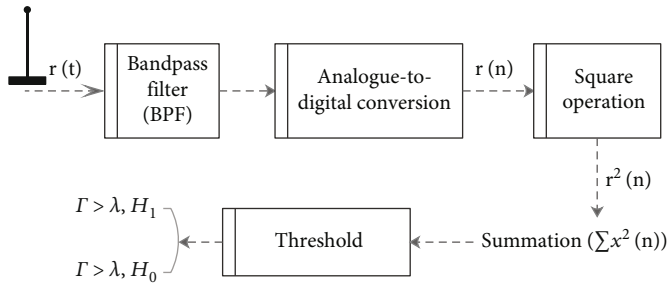

Figure 1: Block diagram of the energy detector.

Among them,

$$\alpha_i = x\left(\frac{i}{2W}\right). \tag{12}$$

Therefore, the energy of $x(t)$ on $(0, T)$ can be expressed as

$$\int_0^T x^2(t)dt = (1/2W)\sum_{i=1}^{2TW} \alpha_i^2. \tag{13}$$

$b_t = a_t/\sqrt{2WN_0}$, and under hypothesis H_0, the test statistic V can be expressed as

$$V = \frac{1}{N_0}\int_0^T n^2(t)dt = \sum_{i=1}^{2WT} b_i^2 \sim \chi_{2TW}^2. \tag{14}$$

$\beta_i = \alpha_i/\sqrt{2WN_0}$, and under hypothesis H_1, the test statistic V can be expressed as

$$V = \sum_{i=1}^{2TW} (b_i + \beta_i)^2 \sim \chi_{2TW}^2(\gamma). \tag{15}$$

In summary, in the energy detection algorithm, the detection statistic V has the following distribution:

$$V \sim \begin{cases} \chi_{2TW}^2, & H_0, \\ \chi_{2TW}^2(\gamma), & H_1. \end{cases} \tag{16}$$

In the above formula, χ_{2TW}^2 is the central chi-square distribution of the degree of freedom $2TW$, and $\chi_{2TW}^2(\gamma)$ is the noncentral chi-square distribution of the degree of freedom $2TW$ and the noncentral parameter γ, where γ represents the instantaneous signal-to-noise ratio.

Then, the probability density function of the detection statistics is

$$f_V(v) \sim \begin{cases} \frac{1}{2^u\Gamma(u)} v^{u-1} \exp\left(-\frac{v}{2}\right), & H_0, \\ \frac{1}{2}\left(\frac{v}{2\gamma}\right)^{u-1/2} \exp\left(-\frac{2\gamma + v}{2}\right) I_{u-1}\left(\sqrt{2\gamma v}\right), & H_1. \end{cases} \tag{17}$$

Among them, TW represents the product of the observation time and the bandwidth of the frequency band, also known as the product of the time bandwidth, and $u = TW$, u takes an integer. $\Gamma(\bullet)$ is a complete Gamma function, and $I_v(\bullet)$ is a Bessel function of order v of the first kind.

For a given threshold V_T', the false alarm probability p_f is

$$p_f = P\left\{V > V_T' | H_0\right\} = p\left\{\chi_{2TW}^2 > V_T\right\}. \tag{18}$$

Similarly, the detection probability p_d is

$$p_d = P\left\{V > V_T' | H_1\right\} = p\left\{\chi_{2TW}^2(\gamma) > V_T'\right\}. \tag{19}$$

Because energy detection belongs to blind inspection, the blind inspection here means that there is no need to know the signal modulation method, power, and other characteristic information, so the threshold value of the algorithm decision cannot be obtained according to the inspection probability, and the threshold value of the algorithm decision can only be determined by the false alarm probability. The decision threshold can be set as

$$\lambda = \delta_n^2 \chi_{pf}^2(N). \tag{20}$$

In the formula, δ_n^2 is the noise power and N is the sampling point. $\chi_{pf}^2(N)$ is the upper quantile function of the standard chi-square distribution. If the judgment is based on the above formula, the judgment process is complicated and not suitable for practical research. The number of signal points N collected in practical research is relatively large, and the above-mentioned random statistics can be

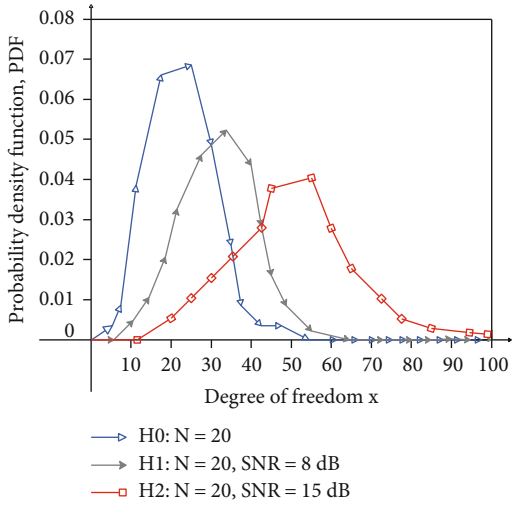

Figure 2: The variation curve of probability density with the change of signal-to-noise ratio SNR.

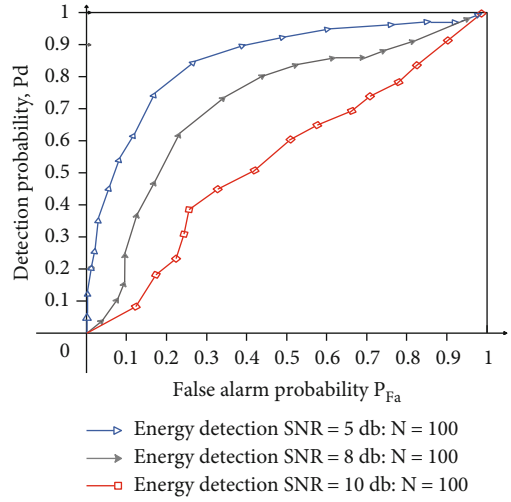

Figure 3: The influence of signal-to-noise ratio on inspection performance.

approximated to a normal distribution through the central limit theorem.

$$V \sim \begin{cases} N\left(N\delta_n^2, 2N\delta_n^4\right), & H_0, \\ N\left(N\left(\delta_n^2+\delta_s^2\right), 2N\left(\delta_n^2+\delta_s^2\right)^2\right), & H_1. \end{cases} \tag{21}$$

In the formula, δ_s^2 is the power of Gaussian signal. The actual detection does not require the actual signal to obey the Gaussian distribution; only the characteristics of Gaussian white noise are used.

According to the analysis of the energy detection method, when the collected signal does not have the main user signal, the statistics obey the χ^2 distribution. When the main user of the collected signal is in the working state, the inspection quantity obeys the noncentral χ^2 distribution. Therefore, we draw the probability density curve of the statistics according to the formula (17), as shown in Figure 2.

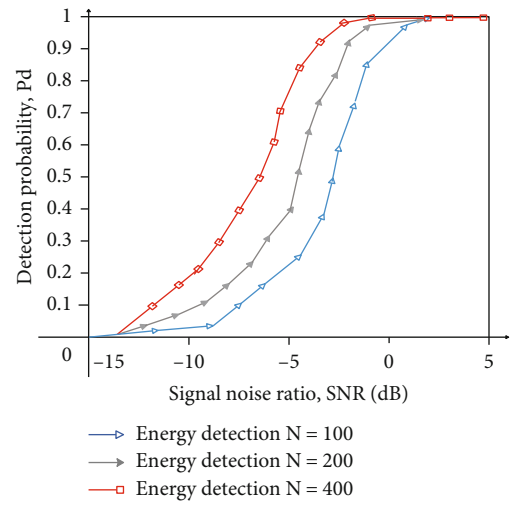

Figure 4: The influence of the number of sampling points on the detection performance.

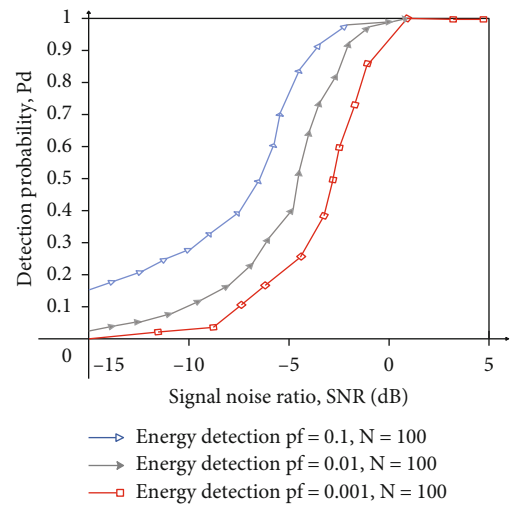

Figure 5: The impact of false alarm probability on detection performance.

The simulation result shown in Figure 2 shows that the signal-to-noise ratio and the probability density curve are inevitable. When the signal-to-noise ratio is larger, the H_1 probability density curve deviates from the H_0 probability density curve. Therefore, the greater the signal-to-noise ratio, the greater the degree of distinction between the signal and the noise, the easier it is to check the existence of the signal, and the lower the error-prone rate. Next, this article conducts simulation experiments on the impact factors, respectively. Figure 3 shows a simulation diagram of the influence of signal-to-noise ratio on the perception algorithm. In the algorithm simulation, the main user signal is in the v channel, the bandwidth is 5^*10^4 Hz, the sampling frequency is $2 * 5 * 10^4$ Hz, the sampling point is 100, and the number of repetitions is 5000. Three different signal-to-noise ratio values are selected in the detection simulation model to obtain three corresponding ROC curves.

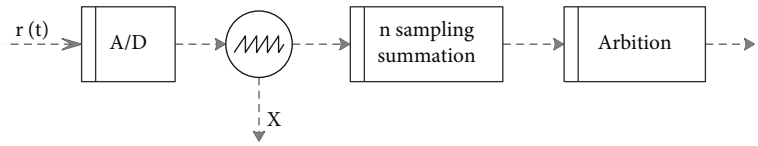

Figure 6: Block diagram of matched filter.

As shown in the simulation diagram shown in Figure 3, it can be concluded that in the sensing process, the detection probability P_d cannot be made low, which will affect the normal communication of the main user. Moreover, the probability of false alarms cannot be increased too high, resulting in low voice usage.

Figure 4 shows a simulation diagram of the impact of the number of sampling points on the accuracy of the perception algorithm. In the algorithm simulation, the main user signal is set in the AWGN channel, the bandwidth is $5 * 10^4$ Hz, the sampling frequency is $2 * 5 * 10^4$ Hz, and the number of repetitions is 5000. The experiment uses 3 different sampling points for simulation and obtains different experimental curves.

The simulation results shown in Figure 4 can be concluded that when the signal-to-noise ratio of the received signal is lower than -15 db, the credibility of the algorithm is basically lost. Moreover, the increase in the number of sampling points cannot play a significant role, which shows that the signal-to-noise ratio plays a decisive role at this stage. By comparing the curves of three different sampling points, it is found that under the same signal-to-noise ratio, the number of sampling points is proportional to the detection performance.

Figure 5 shows a simulation diagram of the effect of false alarm probability on the accuracy of the perception algorithm. In the simulation process, the authorized user signal is set under the AWGN channel, the bandwidth is $5 * 10^4$ Hz, the sampling frequency is $2 * 5 * 10^4$ Hz, the sampling point is 100, and the number of repetitions is 5000. In this paper, three different false alarm probabilities are selected for experimental comparison, and different experimental curves are obtained.

As can be seen from the simulation results shown in Figure 5, in the three simulation curves, under the same signal-to-noise ratio, the greater the false alarm probability (decision threshold), the higher the accuracy of the algorithm.

If the cognitive user can obtain the characteristic quantity of the main user, the matched filter algorithm has the best perception performance. The matched filter algorithm must obtain a priori information of the authorized signal before detection and generate a similar signal for comparison based on its characteristics. Moreover, it matches the signal obtained by the cognitive user with similar signals to determine whether the main user is working normally, that is, whether the frequency band is in working state, as shown in Figure 6.

We use the detection model shown in Figure 6 and set $n(t)$ to be additive white Gaussian noise (AWGN) with a mean value of 0 and a variance of δ^2. As shown in Figure 6, the characteristic quantity of the main user signal has been given, and the received signal $r(n)$ is multiplied by $x(n)$ for accumulation operation, and the statistic is expressed as

$$T = \sum_{n=0}^{N-1} x(n)r(n). \tag{22}$$

In the formula, T is the output of the perception algorithm. N is the length of the collected data. Because the set signal obeys the Gaussian distribution, the statistic T of the matched filter algorithm also obeys the Gaussian.

$$T = \begin{cases} N\left(0.\sigma^2\varepsilon\right), & H_0, \\ N\left(\varepsilon.\sigma^2\varepsilon\right), & H_1. \end{cases} \tag{23}$$

P_d and P_f are

$$\begin{aligned} P_d &= P(T > \lambda | H_1) = Q\left(\frac{\lambda - \varepsilon}{\sqrt{\sigma^2\varepsilon}}\right), \\ P_f &= P(T > \lambda | H_0) = Q\left(\frac{\lambda}{\sqrt{\sigma^2\varepsilon}}\right). \end{aligned} \tag{24}$$

In the formula, Q is the Marcum function Q, which is defined as follows:

$$Q(a, b) = \int_t^{\infty} tI_0(at) \exp\left[-\frac{\left(x^2 + a^2\right)}{2}\right] dx, \tag{25}$$

where I_0 is the modified 0-order Bessel function.

The determination of the threshold is similar to the energy detection algorithm. Figure 7 shows the simulation experiment, the signal is in the AWGN channel, the bandwidth is $5 * 10^4$ Hz, the sampling frequency is $2 * 5 * 10^4$ Hz, the sampling point is 100, and the number of repetitions is 5000.

According to the simulation curve shown in Figure 7, the following can be obtained: for the three curves, under the same false alarm probability condition, the signal-to-noise ratio is proportional to the algorithm performance. For the same curve, the false alarm probability is also proportional to the accuracy of the algorithm.

The realization process of cyclostationary feature detection is shown in Figure 8.

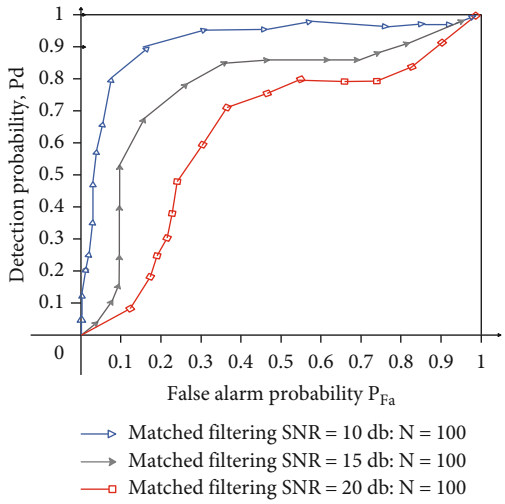

Figure 7: Matched filter detection curve under the same false alarm probability and different signal-to-noise ratio.

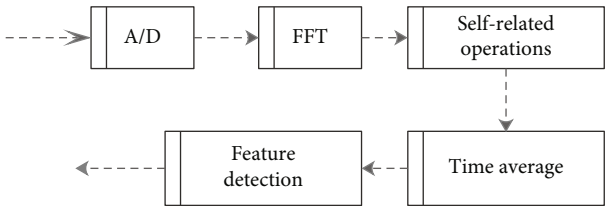

Figure 8: Block diagram of cyclostationary feature detection principle.

We assume that $x(t)$ has a cyclostationary characteristic with a period of T_0, so the following formula is satisfied

$$m_x = E[x(t)] = m_x(t + T_0),$$
$$R_x(t,\tau) = E\{x(t+\tau/2)^* x^*(t-\tau/2)\} = R_x(t+T_0,\tau). \tag{26}$$

$R_x(t,\tau)$ is transformed into a Fourier series.

$$R_x(t,\tau) = \sum_{\alpha} R_x^{\alpha}(\tau) e^{j2\pi\alpha t}. \tag{27}$$

$R_x(t,\tau)$ is the Fourier series coefficient, so (27) is transformed into

$$R_x^{\alpha}(\tau) = \frac{1}{T_0}\int_{T_0} R_x(t,\tau) e^{-j2\pi\alpha t} dt. \tag{28}$$

In the above formula, $R_x^{\alpha}(\tau)$ is called the cyclic autocorrelation function (CAF). Then, the CAF function is Fourier transformed, so

$$S_x^{\alpha}(f) = \int_{-\infty}^{\infty} R_x^{\alpha}(\tau) e^{-j2\pi f t}. \tag{29}$$

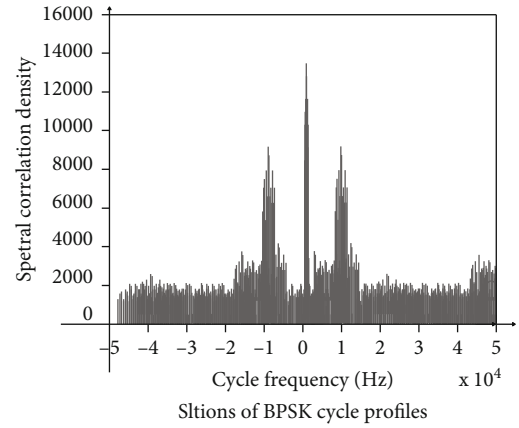

Figure 9: The cycle spectrum of BPSK signal plus noise.

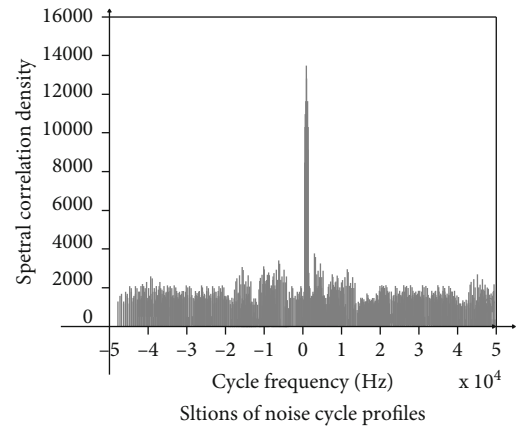

Figure 10: Noise cycle spectrum.

$S_x^{\alpha}(f)$ is the cyclostationary spectrum density function (CSD) of the signal.

$$S_x^{\alpha}(f) = \begin{cases} S_n^0(f), & \alpha = 0, \text{ the signal does not exist,} \\ |H(f)|^2 S_x^0(f) + S_n^0(f), & \alpha = 0, \text{ signal presence,} \\ 0, & \alpha \neq 0, \text{ the signal does not exist,} \\ H\left(f + \frac{\alpha}{2}\right) H \cdot \left(f - \frac{\alpha}{2}\right) S_x^{\alpha}(f), & \alpha \neq 0, \text{ signal presence.} \end{cases} \tag{30}$$

It can be obtained from the above formula that when there is only noise in the system, at $\forall\alpha \neq 0$, the cyclic spectral density is zero, that is, $R_x^{\alpha}(\tau) = 0$, $S_x^{\alpha}(f) = 0$. When there are periodic signals in the system, at $\forall\alpha \neq 0$, the cyclostationary spectrum density is not zero, that is, $R_x^{\alpha}(\tau) \neq 0$ and $S_x^{\alpha}(f) \neq 0$. According to this characteristic, we can distinguish the signal and noise by the value of the cyclic spectral density. Since in the real system, the signal sequence is not infinitely long, and there must be truncation, which causes the noise to have a certain value at the cycle $\alpha_0(\alpha_0 \neq 0)$ frequency $S_n^{\alpha}(f)$. That is, the *CAF* function $R_x^{\alpha}(\tau)$ and *CSD* function calculated by

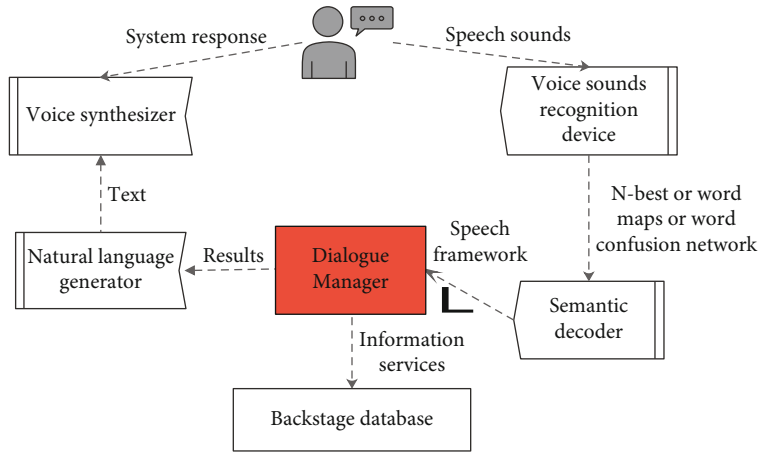

Figure 11: Oral English learning model.

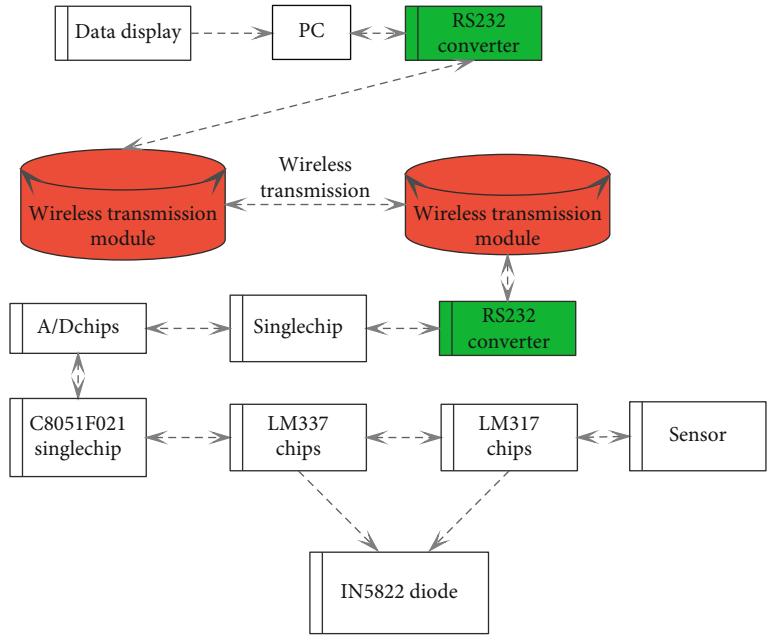

Figure 12: The hardware architecture of the spoken English learning model.

noise will not be completely 0, but most of the noise has been filtered out at this time.

Cyclostationary algorithm has its own unique advantages for speech perception, but due to the large amount of calculation, it needs to be optimized. Aiming at the characteristics of cyclostationary spectrum, this article mainly studies the optimization algorithm for FFT accumulation algorithm (FAM algorithm).

The FAM algorithm can quickly estimate the cyclostationary spectrum of the communication modulation signal, and its expression is as follows:

$$S_{xT}^{\alpha_l+q\Delta\alpha}\left(n, f_j\right)_{\Delta t} = \sum_r X_T(rL, f_k) X_T^*(rL, f_I) g_c(n-r) e^{-j2\pi qr/P}. \tag{31}$$

Among them, $q = -P/2, \cdots, P/2-1, L$ is the decimation factor, P is the number of columns of the channel matrix, $LP = N\$, \$f_k = k(f_k/N'), k = -N'/2, \cdots, N'/2-1$, the frequency coordinate is $f_j = (f_k + f)/2 = ((k+l)/2)(f_s/N')$, and the cyclic frequency coordinate is $\alpha_i = f_k - f_l = (k-l)(f_s/N')$.

Principle of FAM algorithm: the algorithm starts to extract the signal into P segments, and then performs N -point Hamming windows on the P-segment signal to reduce the influence of noise on the signal, and then performs the first N' point FFT. Then, in order to obtain the baseband signal, the operation of the twiddle factor exp $(-j 2\pi kLl/N')$ is required. Finally, the obtained baseband signal is subjected to conjugate operation, and the second P-point FFT is performed, and finally, the cyclic spectral density of the signal is obtained.

Next, we analyze the BPSK signal recognition of cyclostationary. For the BPSK signal, since it is a cyclostationary signal, the verification method is also briefly explained above, so I will not elaborate on it here. Therefore, it will have an impact at the nonzero cycle frequency α_0, and distinguish between BPSK signal and noise by checking whether there is an impact at a special cycle frequency point. According to relevant conclusions in the literature, the BPSK signal cyclostationary spectrum has an impact signal at the cycle frequency $\alpha = kf_0$ and $\alpha = \pm 2f_c + kf_0$. According to the relevant conclusions in the literature, the BPSK signal cycle spectrum has an impact signal at the cycle frequency and place. Therefore, it is possible to find the cycle frequency point α_i that maximizes $|R_x^\alpha(\tau_0)|$ or $|S_x^\alpha(f)|$ except for the cycle frequency $\alpha = 0$, that is, $\alpha_i = 2f_c$. Finally, the carrier frequency $f_c = \alpha_i/2$ of the BPSK signal is obtained.

Figure 9 shows the cycle spectrum of the BPSK signal plus noise. The BPSK carrier frequency is generated by MATLAB, the sampling rate is 5000 Hz, and the number of sampling points is 4096. This simulation diagram can clearly get the carrier frequency $F_c = a_i/2 = 10000/2 = 5000$ Hz of the BPSK signal.

Figure 10 shows that MATLAB produces the same number of noise as above. It can be concluded from the simulation diagram that although the signal sequence is not infinite, the absolute value of CSD is not equal to 0 when 0, but it does not have the characteristic of impact on the carrier frequency when the BPSK signal is 0. Therefore, according to this characteristic, BPSK signals can be identified in a noisy environment.

4. Oral English Learning Model Based on Intelligent Speech Analysis

In this paper, the intelligent speech analysis algorithm is researched, and the corresponding experiments are combined to verify the effectiveness of the algorithm. On this basis, the spoken English learning model is constructed. The oral learning model is shown in Figure 11.

The system mainly includes 5 modules: speech recognizer, semantic decoder, dialogue manager, natural language generator, and speech synthesizer. The hardware structure of the spoken English learning model is shown in Figure 12.

Human-computer interaction can realize information exchange between humans and computers and can effectively improve user experience. This system is based on C ++ language to create business logic components and implement practice strategies. The system software is developed under the Windows 10 operating system environment, and the database server and application server are deployed at the same time. In order to clearly describe the learning history of students, the data grid control is integrated in the system, which uses two-way binding to record events during the exercise. On this basis, it is necessary to create a complete configurable back-end service system, so the software design of the system should give priority to the adaptive training of practice strategies. After completing the formulation of the practice strategy, the system should provide targeted oral training projects based on the students' ability to effectively improve the students' oral expression skills.

After constructing the system model of this article, the effect of the model of this article is verified, and the effect of English speech analysis and spoken English learning effect of the model of this article is counted. The statistical effect of English speech analysis is shown in Table 1 and Figure 13.

Table 1: English speech analysis effect based on intelligent speech analysis.

Number	Speech analysis	Number	Speech analysis	Number	Speech analysis
1	87.11	22	87.08	43	87.86
2	89.92	23	92.07	44	87.07
3	90.21	24	89.96	45	88.14
4	90.02	25	87.86	46	90.21
5	90.35	26	88.81	47	92.72
6	87.30	27	89.76	48	87.47
7	87.85	28	89.69	49	90.76
8	92.51	29	91.78	50	87.69
9	89.87	30	88.76	51	87.32
10	88.89	31	90.04	52	87.36
11	91.68	32	87.72	53	88.58
12	91.59	33	89.36	54	89.41
13	89.41	34	91.00	55	91.68
14	91.72	35	88.69	56	91.24
15	87.01	36	92.38	57	93.09
16	87.29	37	88.45	58	87.60
17	88.58	38	87.50	59	89.16
18	92.11	39	87.56	60	89.13
19	88.15	40	90.08	61	89.84
20	90.42	41	87.41	62	93.15
21	93.24	42	92.28	63	93.88

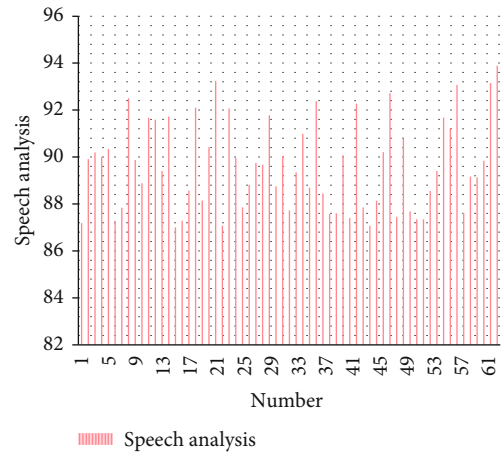

Figure 13: Statistics on the effect of English speech analysis.

Table 2: Evaluation of the effect of spoken English learning.

Number	Oral learning	Number	Oral learning	Number	Oral learning
1	87.11	22	87.08	43	87.86
2	89.92	23	92.07	44	87.07
3	90.21	24	89.96	45	88.14
4	90.02	25	87.86	46	90.21
5	90.35	26	88.81	47	92.72
6	87.30	27	89.76	48	87.47
7	87.85	28	89.69	49	90.76
8	92.51	29	91.78	50	87.69
9	89.87	30	88.76	51	87.32
10	88.89	31	90.04	52	87.36
11	91.68	32	87.72	53	88.58
12	91.59	33	89.36	54	89.41
13	89.41	34	91.00	55	91.68
14	91.72	35	88.69	56	91.24
15	87.01	36	92.38	57	93.09
16	87.29	37	88.45	58	87.60
17	88.58	38	87.50	59	89.16
18	92.11	39	87.56	60	89.13
19	88.15	40	90.08	61	89.84
20	90.42	41	87.41	62	93.15
21	93.24	42	92.28	63	93.88

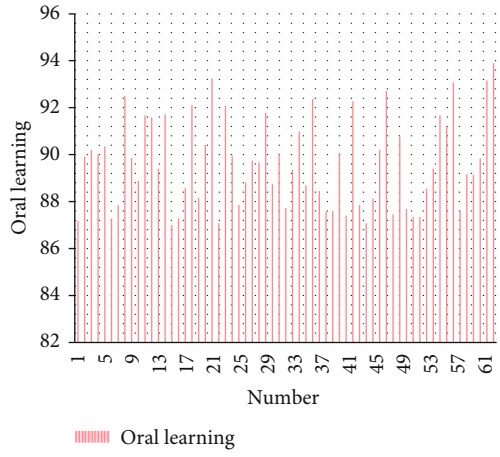

Figure 14: Statistics of learning effect.

It can be seen from the above research that the intelligent speech analysis model proposed in this paper can play a certain effect in English speech analysis. On this basis, the effect of the spoken English learning model based on intelligent speech analysis is evaluated, and the results shown in Table 2 and Figure 14 are obtained.

From the above research, it can be seen that the spoken English learning model based on intelligent speech analysis proposed in this paper can effectively improve the effect of spoken English learning.

5. Conclusion

The intelligibility of spoken English focuses on the level of comprehension and recognition of spoken English. The objective experimental evaluation method of spoken English is easy to operate and convenient, but it has obvious shortcomings. The evaluation result can only be infinitely close to people's subjective perception of spoken English, rather than a realistic reflection of subjective perception characteristics. The objective evaluation of spoken English is related to linguistics and digital signal processing and is closely related to physiology, psychology, and dialect characteristics. Generally speaking, in recent years, scientific research and technical personnel have been searching for an objective evaluation method that can quickly and accurately reflect the quality and intelligibility of spoken English through experiments. Most of these methods use the time domain amplitude, frequency domain amplitude, and transform domain of the spoken English signal as evaluation criteria and rarely involve other factors that affect the quality and intelligibility of spoken English, such as sound, intonation, and grammar. This article combines intelligent speech analysis to study the spoken English learning model, constructs an intelligent model, and designs experiments to verify the effect of the intelligent speech model. The research results show that the spoken English learning model based on intelligent speech analysis proposed in this paper can effectively improve the effect of spoken English learning.

Data Availability

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

Conflicts of Interest

The author declare no competing interests.

Acknowledgments

This study is sponsored by Xi'an Aeronautical University.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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, 2016.

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

[6] 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.

[7] M. Sleeper, "Contact effects on voice-onset time in Patagonian Welsh," *Acoustical Society of America Journal*, vol. 140, no. 4, pp. 3111–3111, 2016.

[8] 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, pp. 2302–2302, 2015.

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

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

[11] C. T. Herbst, S. Hertegard, D. Zangger-Borch, and P. Å. Lindestad, "Freddie mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics," *Logopedics Phoniatrics Vocology*, vol. 42, no. 1, pp. 29–38, 2017.

[12] J. Al-Tamimi, "Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: implications for formal representations," *Laboratory Phonology*, vol. 8, no. 1, pp. 28–40, 2017.

[13] O. Abdel-Hamid, A. 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.