Hindawi
Journal of Mathematics
Volume 2024, Article ID 9819732, 1 page
https://doi.org/10.1155/2024/9819732

Retraction

Retracted: A Convolutional Network-Based Intelligent Evaluation Algorithm for the Quality of Spoken English Pronunciation

Journal of Mathematics

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] X. Zhan, "A Convolutional Network-Based Intelligent Evaluation Algorithm for the Quality of Spoken English Pronunciation," *Journal of Mathematics*, vol. 2022, Article ID 7560033, 9 pages, 2022.

Hindawi
Journal of Mathematics
Volume 2022, Article ID 7560033, 9 pages
https://doi.org/10.1155/2022/7560033

Research Article

A Convolutional Network-Based Intelligent Evaluation Algorithm for the Quality of Spoken English Pronunciation

Xia Zhan

School of Foreign Languages, Changchun Institute of Technology, Changchun 130012, China

Correspondence should be addressed to Xia Zhan; wy_zx@ccit.edu.cn

Received 30 November 2021; Revised 22 December 2021; Accepted 27 December 2021; Published 18 January 2022

Academic Editor: Naeem Jan

Aiming at the problems of long time consumption and low accuracy of traditional spoken English pronunciation quality assessment algorithms, a convolutional network-based intelligent assessment algorithm for spoken English pronunciation quality is proposed. The convolutional neural network structure is given, the original data of the spoken English pronunciation voice signal are collected by multisensor detection, and the spoken English pronunciation voice signal model is constructed. Based on audio and convolutional neural network learning and training, it realizes the feature selection and classification recognition of spoken English pronunciation. The PID algorithm is used to extract the emotional elements of spoken English at different levels to achieve accurate assessment of the quality of spoken English pronunciation. The experimental results show that the average correct rate of spoken English pronunciation of the algorithm in this paper is 94.58%, the pronunciation quality score is 8.52–9.18, and the detection time of 100 phrases is 2.4 s.

1. Introduction

As a widely used language, English has attracted more and more people's attention. English is becoming more and more important in daily life. People appreciate American TV shows and Hollywood movies. They need to use English when traveling abroad, and they need to use it for import and export transactions. When it comes to English, English is required for academic research communication, and English is also required for industrial production, programming, and viewing technical documents [1]. So, being able to speak English is becoming more and more important for Chinese people. For Chinese people, dumb English has always been the number one problem in learning English. With the development of speech signal processing technology, the use of speech signal recognition methods to intelligently evaluate the quality of spoken English, combined with speech information processing technology to improve the quality of spoken English pronunciation, is of great significance in improving the effectiveness of spoken English teaching. The intelligent assessment of spoken English pronunciation quality evaluates and calculates pronunciation quality and detects pronunciation errors [2]. The related intelligent assessment algorithm of spoken English pronunciation quality has a great role in promoting the standardization of spoken English pronunciation, and it has also received great attention from people.

Wen [3] proposed the design of an automatic correction system for English pronunciation errors based on the dynamic time warping (DTW) algorithm. Relying on the optimized design of the speech recognition sensor and the improved design of the pronunciation recognition processor, the hardware design of the system is completed; the software design of the system is completed based on the design of the English pronunciation acquisition program and the extraction of English pronunciation error signal parameters. This method can accurately assess the pronunciation quality of spoken English, but the assessment takes a long time. Luo et al. [4] proposed an automatic evaluation technology algorithm for spoken English based on deep neural networks. Based on the verification experiment conducted on the real scene data of the large-scale unified oral English test in junior and senior high schools, the proposed automatic evaluation method has a greater performance advantage than the traditional method based on goodness of pronunciation (GOP). The evaluation of this method takes less time, but the detection accuracy still needs to be improved.

When the user mistakenly pronounces one phoneme into another phoneme in the phoneme set, this hypothesis can be a good approximation to the true posterior probability value, but when the user's pronunciation is different from any standard pronunciation in the phoneme set, maximum number of multiple candidates differs from sum. Therefore, in some cases, this assumption will seriously reduce the accuracy of the confidence calculation. Aiming at the problems of the above methods, this paper proposes an intelligent assessment algorithm for spoken English pronunciation quality based on convolutional networks. Deep learning attempts to learn a better representation of data from large-scale unlabeled data, so deep learning is also called representation learning or unsupervised feature learning algorithm. One of the most commonly used scenarios of deep learning is to use unsupervised or semisupervised algorithms to automatically learn features to replace manually designed features. The convolutional neural network structure in deep learning is used to train the features of spoken English pronunciation signals, and based on audio to realize the screening of spoken English pronunciation features and classification and recognition, the proportional-integral-derivative (PID) algorithm is used to extract the emotional elements of speech, and the quality of spoken English pronunciation can be accurately measured.

The arrangement of the paper is as follows: Section 1 is the introduction and literature review. In Section 2, the structure of the convolutional neural network (CNN) is explained in detail. Moreover, the voice signal model of spoken English pronunciation features is given. Finally, the extraction of spoken English pronunciation features is carried out. Section 3 presents an intelligent assessment algorithm for the quality of spoken English pronunciation. In addition, the screening and classification of spoken English pronunciation features are done. In order to validate the proposed algorithm, Section 4 carries out the experiments and analyses their outcomes for the purpose of comparison. Lastly, Section 5 concludes the paper.

2. Spoken English Pronunciation Feature Extraction Based on the Convolutional Neural Network

In this section, the structure of the convolutional neural network (CNN) is explained in detail. Moreover, the voice signal model of spoken English pronunciation features is given. Finally, the extraction of spoken English pronunciation features is carried out.

2.1. Convolutional Neural Network Structure. The deep convolutional neural network is mainly composed of the input layer, hidden layer, and output layer. The hidden layer is composed of repeated and alternating multilevel convolutional layers and pooling layers, and its structure is shown in Figure 1.

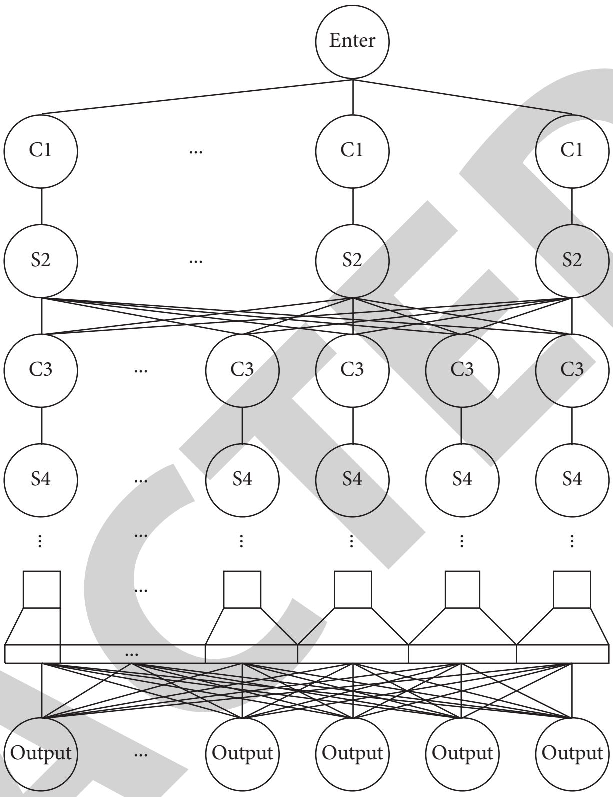

Figure 1: Deep convolutional neural network structure.

The initial data without feature extraction are input into the input layer, the input data are convolved through the convolution kernel in the convolution layer (C1), the corresponding convolution feature map is obtained, and the convolution is pooled through the pooling layer (S2). From the feature map obtained in the layer, the corresponding pooling feature map is obtained [5], and the operation is repeated in the hidden layer (C3, S4) imitating C1 and S2. By setting the convolution and pooling of the network, the extraction of data features can be effectively achieved, and the detection model can improve the degree of tolerance of the image that satisfies the distortion invariance [6]. At the same time, the resolution of the image is reduced, and the feature images are increased to obtain a large amount of feature data. The input information outputs the final detection result through the fully connected layer [7].

2.1.1. Convolutional Layer. The preprocessed acceleration sensor x, y, z data (depth is 3) are taken as the input data. In order to ensure the same size of the input and output, the data need to be filled with 0. During the convolution operation, the transformation of the same convolution kernel does not affect its weight, and the weight is shared with the x-axis data. This feature can effectively reduce the number of parameters of deep convolutional neural networks and accelerate network training [8].

All convolution kernels in the deep convolutional neural network have the function of automatic feature extraction. The acceleration sensor x, y, z data are convolved through the convolution kernel, and various details can be extracted by each convolution kernel [9].

Let the height and width of the convolution kernel be f_h and f_w, respectively, to obtain a two-dimensional convolution:

$$y_{n,m} = A\left\{\begin{matrix} x_{n,m} & x_{n+1,m} & \cdots & x_{n+f_w,m} \\ x_{n,m+1} & x_{n+1,m+1} & \cdots & x_{n+f_w,m+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,m+f_h} & x_{n+1,m+f_h} & \cdots & x_{n+f_w,m+f_h} \end{matrix}\right\}. \tag{1}$$

The activation function uses the ReLU function to get the input and output of the total convolutional layer:

$$Y = \mathrm{ReLU}\left(\sigma(WB + b)\right). \tag{2}$$

2.1.2. Maximum Pooling Layer. The significance of the existence of the pooling layer is mainly to select and reduce the dimensionality of the output. The maximum pooling strategy is applied, the pooling core is 2 × 2, let s be the step size, and the height and width of the pooling core are p_h and p_w, respectively, to get the maximum pooling:

$$y_{i,j} = \max\left\{\begin{matrix} x_{is,js} & x_{is,js+1} & \cdots & x_{is,js+p_w} \\ x_{is+1,js} & x_{is+1,js+1} & \cdots & x_{is+1,js+p_w} \\ \vdots & \vdots & \ddots & \vdots \\ x_{is+p_h,js} & x_{is+p_h,js+1} & \cdots & x_{is+p_h,js+p_w} \end{matrix}\right\}. \tag{3}$$

Through the pooling layer, the dimensionality of the data and the corresponding training parameters can be reduced to a great extent, and the speed of network training can be accelerated.

2.1.3. Fully Connected Layer and Output Layer. The deep convolutional neural network is connected to the fully connected layer below its hidden layer, and the number of connected fully connected layers is greater than or equal to one. The existence of a fully connected layer is equivalent to a multilevel perceptron, in which all neurons of the same level are connected to all neurons in the upper layer, and the difference between the convolutional layer and the pooling layer can also be significant in this layer. Part of the information is fused. Taking the ReLU function as the activation function of the fully connected layer can effectively improve the performance of the deep convolutional neural network structure. The output layer receives the output value from the bottom fully connected layer and connects to different classifiers according to the required target. In order to prevent the overfitting situation in the traditional training of small-scale datasets, regular applications are often applied to the fully connected layer. The randomness of this method leads to the fact that the corresponding network structure of the dataset transmitted every time is not consistent, but the weights of all network structures are shared. This method greatly improves the stability of the detection model and makes every nerve less complicated when the elements adapt to each other [10].

The deep convolutional neural network convolutional layer applies a weight sharing method while reducing the parameters and difficulty of its structure and preventing the model from overfitting in the early stage so that it has better generalization ability, through pooling. To ensure the stability of the model, the network has a variety of characteristics that make it maintain the translation, scaling, and distortion when the transformation occurs. Deep convolutional neural networks have strong expression effects and scalability and can be well applied to various difficult problems.

2.2. The Voice Signal Model of Spoken English Pronunciation. In order to realize the quality assessment of spoken English pronunciation based on the convolutional neural network, firstly, the spoken English pronunciation voice signal model is given, and the multisensor detection method is adopted to collect the original data of spoken English pronunciation voice signals, and then the collected spoken English pronunciation voice signals are collected. Scale decomposition and feature extraction are carried out [11], spoken English pronunciation quality assessment and feature detection are carried out, and the mathematical model expression of the spoken English pronunciation speech signal is given as

$$z(t) = s(t) + js(t) \otimes h(t) = s(t) + j\int_{-\infty}^{+\infty} \frac{s(u)}{t-u}\,\mathrm{d}u. \tag{4}$$

In the formula, $a(t)$ is called the spoken English pronunciation voice signal-received signal amplitude at the nth array element, sometimes called the envelope, $\phi(t)$ is called the phase of the multiuniform linear wideband array, $Z(f)$ can be obtained by the Fourier transform of $S(f)$, and $H(f)$ is the step transfer function of the spoken English pronunciation voice signal. Based on the convolutional neural network, the spoken English pronunciation speech signal modeling and detection and recognition are carried out, and the array element distribution of the speech information sampling is $v_m, m \in [1, n]$. The result of the separation of the phonetic features of spoken English pronunciation is calculated as

$$y(t) = \iint_{a,b} \rho(a,b) f\left(\frac{t-b}{a}\right). \tag{5}$$

In the formula, $f(t - b/a)$ is the instantaneous frequency estimation value of the received spoken English speech signal, $\rho(a,b)$ is the delay component of the broadband signal incident on the array element, a is the high-order statistical characteristic information of the signal, and b is the frequency shift distribution. The feature components of spoken English pronunciation information are calculated as

$$X_p(u) = \int_{-\infty}^{+\infty} K_p(t,u)x(t)\mathrm{d}t. \tag{6}$$

The fusion weight is updated, and the output signal component $\pi/2$ obtained can be expressed as

$$X_p(u) = \begin{cases} x(u), & \alpha = 2n\pi, \\ x(-u), & \alpha = (2n \pm 1)\pi. \end{cases} \tag{7}$$

In the formula, p is the order of the best received polarization vector, which can be any real number, and the phase of voice detection is $\alpha = p\pi/2$. When $\pi/2$ is reached, it rotates to the frequency axis, thus realizing oral English modeling the statistical information of the articulated speech signal.

2.3. Extraction of Spoken English Pronunciation Features. In order to extract the pronunciation features of spoken English, the basic network based on the deep convolutional neural network is ResNet101; in order to better extract the subtle features of spoken English pronunciation, in the middle of the convolutional layer and the pooling layer, batches are added to layer by layer through ResNet. The residual block adjusts the information transmission strategy while accelerating the network training speed and promotes the optimization of the network [12].

The batch normalization algorithm is applied to the batch normalization layer, which integrates the processing operations of the network layer input into the spoken English pronunciation detection and processes the spoken English pronunciation feature samples through microbatch normalization.

The batch normalization is expressed as

$$\widehat{X} = \mathrm{norm}(x, X). \tag{8}$$

In the formula, x describes all the vectors that are input to a certain layer in the deep convolutional neural network, and X represents a certain value for the overall training sample. The output of the batch-normalized network can be judged by using the input vector of the previous layer and the overall value. The network input of each layer of the training set is obtained from the output of the previous layer, and the parameters of the model will also limit the input vector.

When optimizing the network parameters, the backpropagation algorithm is used to obtain the Jacobian matrix corresponding to the batch normalization of the input vector and the overall training sample value. The formula is

$$\begin{gathered} \frac{\partial norm(x, X)}{\partial x}, \\ \frac{\partial norm(x, X)}{\partial X}. \end{gathered} \tag{9}$$

Batch normalization is a big project to process the input of all layers, it needs to calculate the matrix of covariance, and it takes a long time. In this regard, the following two simplified improvement methods are proposed:

(1) The joint normalization processing of each dimension data is replaced with the data of each dimension delivered by the independent batch normalization processing, and the formula is as follows:

$$\widehat{X}(k) = \frac{x_i(k) - E[x(k)]}{\sqrt{[x(k)]}}. \tag{10}$$

In the formula, the k dimension of the input sample is described by $x(k)$, the expectation is described by $E[x(k)]$, and the variance is described by $\mathrm{var}[x(k)]$. Independent batch normalization can effectively speed up the convergence speed of network training, but it does not guarantee the stability of the initial description of each layer of the network, resulting in the initial output characteristics that cannot be fully described by the input. In order to maintain the constant change of the added batch normalization process, parameters $\lambda(k)$ and $\beta(k)$ are added to the k dimension of each input sample to obtain the formula

$$y(k) = \lambda(k)\widehat{X}(k) + \beta(k). \tag{11}$$

In the formula, $\beta(k)$ and $\mathrm{var}[x(k)]$ are equal, both are descriptions of the input standard deviation, which means the k dimension of the input sample after scale transformation, $\beta(k)$ and $E[x(k)]$, is equal, and both are the expected input, which Indicates that the input sample after translation is k. Using this parameter together with each parameter in the model for network training can effectively ensure the description level of the model.

(2) Stochastic gradient training of deep convolutional neural networks is carried out through microbatch samples, the average value and variance of each layer are estimated by calculating each sample, and the aforementioned operation can be used to realize the reverse direction propagation of the gradient.

Suppose the microbatch sample is denoted as B, its sample size is described as m, a certain dimension input to a certain level is denoted as x, and the dimension-wise normalization is expressed as

$$BN_{\lambda,\beta}\colon x_1, \ldots, x_m \longrightarrow y_1, \ldots, y_m. \tag{12}$$

Through the above content, the feature extraction of spoken English pronunciation based on the convolutional neural network is realized.

3. Intelligent Assessment Algorithm of Spoken English Pronunciation Quality

For the quality of spoken English pronunciation, this section presents an intelligent assessment algorithm. In addition, the screening and classification of spoken English pronunciation features are done.

3.1. Screening and Classification of Spoken English Pronunciation Features. The current spoken English pronunciation assessment algorithms mostly rely on language signals for judgment, ignoring the role of signals in pronunciation error correction. For this reason, an audio-based method for screening and classifying spoken English pronunciation features is proposed. Using the convolutional neural network learning method, the feature screening and classification of spoken English pronunciation signals are performed. Assuming that the input spoken English pronunciation speech signal is a single-frequency signal $\cos 2\pi f_0 t$, where f_0 is the spoken English pronunciation frequency, the reference component of the spoken English pronunciation signal detected by the first array element is set to construct the error feature screening of the spoken English pronunciation. The model uses the time-frequency feature transformation method for dynamic detection and feature selection of spoken English pronunciation signals, and the mth block sparse feature quantity is

$$s_m(t) = \cos\{2\pi f_0 [t + \tau_m(\theta)]\}. \tag{13}$$

The target source signal detection method is used to monitor the characteristics of spoken English pronunciation speech signals, and the characteristic distribution of spoken English pronunciation errors is obtained as

$$l(t) = u_m \cos(2\pi f_0 t) - v_m \sin(2\pi f_0 t), \tag{14}$$

From this, the eigenvalues of spoken English pronunciation speech signals are extracted, and the beam-forming method is used to focus on the characteristics of spoken English speech signals. Therefore, the deep neural network detection method is used to detect the error characteristics of spoken English speech signals. The output is

$$y_1(t) = A_1(t) \exp\{j2\pi [F(t - t_a) - F \ln Dt + f_{e1} t]\}. \tag{15}$$

The output feature quantity of the pronunciation error of harmonic spoken English is expressed as

$$y_2(t) = A_2(t) \exp\{j2\pi [F(t - t_a) - F \ln Dt + f_{e2} t]\}. \tag{16}$$

In the formula, f_{e1} is the beam domain cutoff frequency, and f_{e2} is the harmonic cutoff frequency. The statistical feature analysis method is used to separate the features of spoken English pronunciation errors, and the output information of spoken English pronunciation errors is

$$y(t) = s(t) + n(t). \tag{17}$$

The spectrum of mispronunciation messages in spoken English is

$$\begin{aligned} Y_p(u) &= F^a[y(t)] \\ &= F^a[s(t) + n(t)] \\ &= F^a[s(t)] + F^a[n(t)]. \end{aligned} \tag{18}$$

When the prior probability of the signal satisfies the convergence condition, the time width of the spoken English speech signal is calculated:

$$T^2 = \int_{-\infty}^{+\infty} (t - t_m)^2 |x(t)|^2 \mathrm{d}t. \tag{19}$$

The frequency-domain characteristics of spoken English pronunciation speech signals are described as

$$B^2 = \int_{-\infty}^{+\infty} (v - v_m)^2 |X(v)|^2 \mathrm{d}v. \tag{20}$$

According to the Bayesian formula, the characteristics of the spoken English pronunciation signal are screened, and the detection output is

$$s(t) = \sqrt{s}\, u[s(t - \tau_0)]. \tag{21}$$

3.2. Intelligent Assessment Algorithm of Spoken English Pronunciation Quality. In order to solve the problem that the existing system only considers, intonation and rhythm when evaluating the quality of spoken English pronunciation, but does not take into account the effect of speech emotion, which leads to the poor effect and inefficient evaluation of spoken English pronunciation, the PID algorithm is used to extract the emotional elements of the spoken language at different levels. Taking full account of the imbalance of corpus evaluation data, the data of various elements that affect the pronunciation of spoken English are extracted [13]. Since the traditional system has researched and extracted conventional indicators such as intonation and rhythm, the PID algorithm is used on the basis of the existing methods to extract the emotional elements of spoken English at different levels [14] in order to extract English accurate assessment of the quality of spoken English pronunciation.

PID is the most common algorithm for remote operation. Suppose that the actual output value of the intelligent evaluation algorithm for spoken English pronunciation quality based on the convolutional network is $c(t)$, the fixed value is $r(t)$, and the operation deviation calculation formula of the evaluation algorithm is

$$e(t) = c(t) - r(t). \tag{22}$$

The differential (D), proportion (P), and integral (I) of the scoring deviation of the spoken English pronunciation quality scoring system are linearly combined to form the operation volume of the laboratory experiment remote operating system, and each pronunciation element is scored, which is called the PID algorithm. In the virtual reality-based English-speaking pronunciation quality scoring system, according to the standard rules of spoken English pronunciation and pronunciation characteristics, the P, I, and D operation rules are appropriately combined to complete the extraction of speech emotion elements [15]. The law calculation formula is

$$u(t) = \left[e(t) + T_D + \frac{1}{T_1} \int_0^t e(t) \mathrm{d}t \right] \times K_P. \tag{23}$$

In the formula, K_P represents the proportional coefficient of the emotional elements in the spoken pronunciation; T_1 represents the validity of the voice emotional index; T_D represents the differential time constant for the completion of the operation; t represents the time required for extraction. Since the characteristic data recognized by the traditional scoring system are limited and cannot be operated continuously on the characteristic data, the PID algorithm is used to discretize the information data in the scoring system. The calculation formula for the discretization is

$$u(k) = K_P\left[e(t) + \frac{T}{T_1}\sum_{i=1}^{k} e(i) + \frac{T_D}{T}\right] + u_0. \tag{24}$$

In the formula, u_0 represents the initial value when the score deviation is 0; T represents the sampling period of speech emotion elements. After discretizing the data information through PID algorithm, the continuous operation of the system is realized, and the effective extraction of voice emotion elements is guaranteed [16].

According to the extraction results of speech emotion elements, the quantitative recursive analysis method comprehensively evaluates the quality of spoken English pronunciation and finally obtains the scoring results. The panel data for the evaluation of spoken English pronunciation quality are established, and the method of combining quantitative analysis and fuzzy prediction is used to obtain the statistical regression analysis results of panel data for the evaluation of spoken English pronunciation quality as follows:

$$q^{w}\left\{\lambda_w \eta^{w} - D^{-1}\right\} \geq 0. \tag{25}$$

In the formula, w represents the mean value of the feature; η represents the standard deviation of the pronunciation; λ represents the ambiguity feature amount of the speech.

Combining the minimum cost and the best balanced method of teaching quality [17], the game balance control of the English pronunciation quality score is carried out, and the optimization level is selected as the dependent variable, and the statistical detection quantity is obtained as

$$q^{A} = \frac{\left(\beta c_n\right)^2}{16\beta(1-\beta)} + \frac{\left(1-c_n\right)^2}{16}. \tag{26}$$

In the formula, β represents the phoneme competition subset; c represents the independent threshold; n represents the voice recording rate.

Therefore, a panel data statistical analysis model for the evaluation of spoken English pronunciation quality is constructed, and a game model for the evaluation of spoken English pronunciation quality is obtained, which is defined as

$$V_i = \frac{X^{i}_{\max} - X^{i}}{q^{A}\left(X^{i}_{\max} - X^{i}_{\min}\right)}. \tag{27}$$

In the formula, V represents the factors that affect pronunciation evaluation; X represents the correct vowels and words entered. In summary, the quantitative regression analysis method and the full-sample regression test analysis method are used to achieve the scoring of the quality of spoken English pronunciation.

4. Experimental Analysis

In order to test the performance of the algorithm in this paper in realizing the intelligent evaluation of spoken English pronunciation quality, a simulation experiment was carried out. The experiment was designed with MATLAB 7 simulation software to verify the correct rate of spoken English pronunciation, the score of spoken English pronunciation quality, and the algorithm response time. The effectiveness of the results and the method of Wen [3] and Luo et al. [4] are used as experimental comparison methods.

4.1. Experimental Data Preparation. This study selects the spoken Arabic digit dataset as the experimental dataset, which contains a large amount of spoken English pronunciation data. In order to reduce the difficulty of the experiment, a 16 KHz sampling rate was used to randomly select 13,500 data in the spoken Arabic digit dataset. The specific experimental data information is shown in Table 1.

The number of nodes sampling the spoken English pronunciation signal is 120, the resolution of feature extraction is 200 KHz, the length of the output spoken English pronunciation signal is 1200, the number of sources to be measured is 20, and the interference signal-to-noise ratio is −20 dB.

4.2. Analysis of Experimental Results. Based on the experimental data prepared above and the determined experimental evaluation indicators, an intelligent evaluation experiment for the quality of spoken English pronunciation is carried out. The analysis process of the specific experimental results is shown below.

4.2.1. Analysis of the Correct Rate of Pronunciation Errors in Spoken English. The correct rate data of spoken English pronunciation error detection obtained through experiments are shown in Table 2.

As shown by the comparison of data in Table 2, in the process of 10 experiments on spoken English pronunciation, the algorithm in this paper has a high error detection rate of spoken English pronunciation, the highest is 96.2%, the lowest is 92.5%, and the average is 94.58%, which is much higher than the references' comparison method. Because the method in this paper uses the convolutional neural network to train the spoken English pronunciation data, it improves the correct rate of pronunciation error detection. The experimental results show that the designed intelligent assessment algorithm of spoken English pronunciation quality has better error detection performance.

4.2.2. Analysis of the Quality of Spoken English Pronunciation. After applying the designed intelligent assessment algorithm for spoken English pronunciation quality, the calibrated pronunciation quality score data are shown in Table 3.

Table 1: Experimental data information table.

Experimental data information		
Type of data	Open testing	Close testing
Number of data	4500	9000
Number of speakers	5	10
Proportion of male to female speakers	3 : 2	1 : 1
Data format	16 bit/PCM	

Table 2: Data table of correct rate of pronunciation error detection in spoken English.

	Correct rate of pronunciation error detection in spoken English (%)		
Number of experiments	Wen's [3] algorithm	Luo et al.'s [4] algorithm	The proposed algorithm
1	75.2	70.6	92.5
2	73.1	70.5	93.5
3	74.8	69.7	94.5
4	75.6	68.5	95.3
5	74.9	72.2	95.1
6	69.5	71.0	94.8
7	68.4	65.2	95.2
8	72.4	69.3	94.8
9	71.8	68.5	96.2
10	71.6	65.7	93.9

Table 3: Spoken English pronunciation quality score data table.

Number of experiments	Wen's [3] algorithm	Luo et al.'s [4] algorithm	The proposed algorithm
1	6.58	6.31	8.52
2	6.95	6.83	9.01
3	6.78	6.95	9.12
4	6.80	7.09	8.67
5	7.10	7.18	8.95
6	7.08	6.68	8.74
7	6.95	6.95	9.02
8	6.45	7.41	9.15
9	7.01	7.35	8.95
10	6.89	7.25	8.89

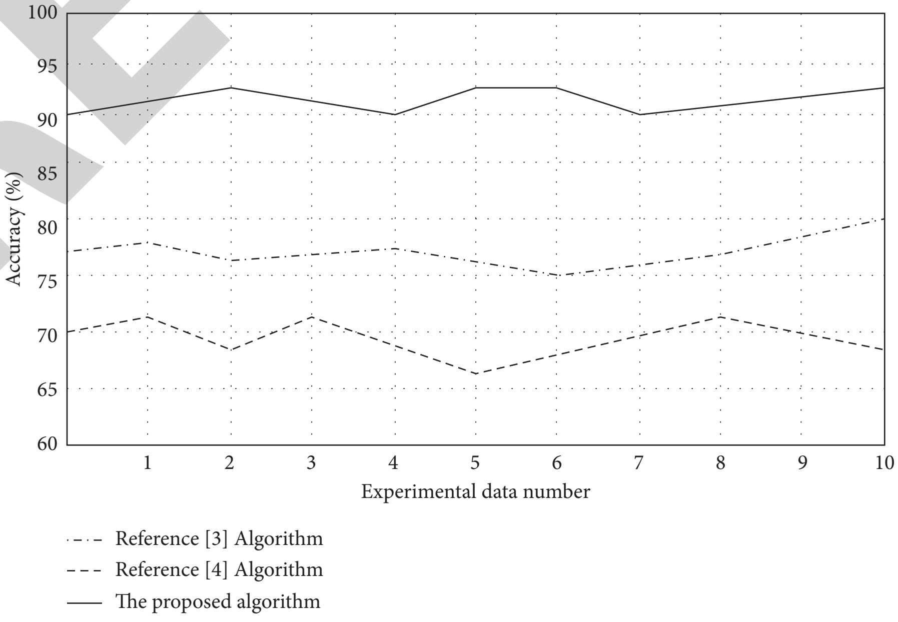

Figure 2: Accuracy rate of the spoken English pronunciation quality score.

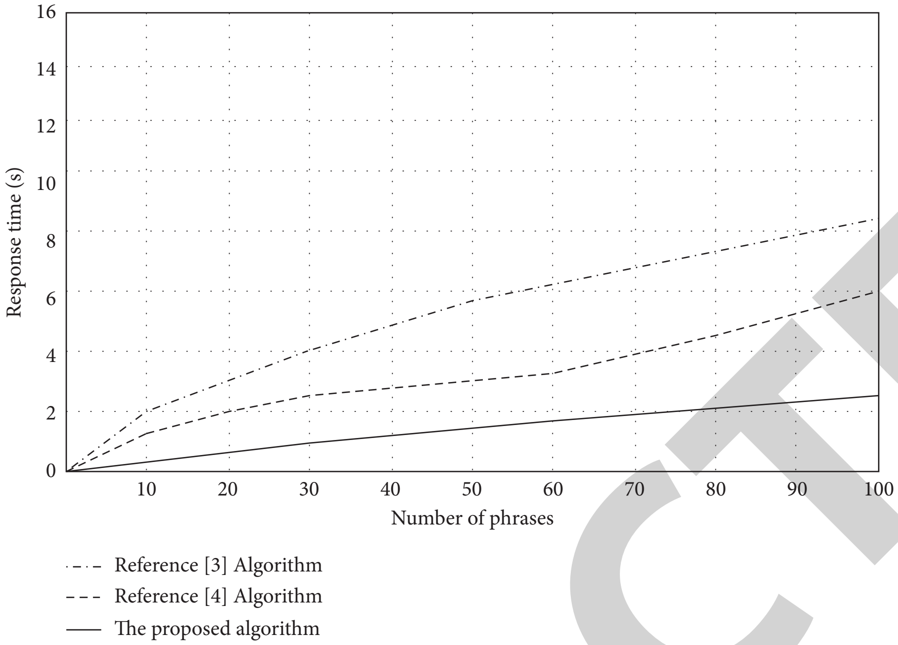

Figure 3: Comparison results of response time.

As shown in Table 3, the pronunciation quality score of the algorithm in this paper is 8.52 points to 9.18 points, the pronunciation quality score of the algorithm in [3] is 6.45 points to 7.10 points, and the pronunciation quality score of the algorithm in [4] is 6.31 points to 7.35 points which shows that the algorithm in this paper has a better effect on scoring spoken English pronunciation quality. The accuracy rate of the spoken English pronunciation quality score is shown in Figure 2.

Analyzing Figure 2 shows that, in the course of 10 spoken English pronunciation quality experiments, the average accuracy rate of the spoken English pronunciation quality score of the algorithm in this paper is 93.5%, and the average accuracy rate of the spoken English pronunciation quality score of the algorithm in [3] is 78.5%. The average accuracy rate of spoken English pronunciation quality scores based on the algorithm in [4] is 71.5%. The experimental results show that the accuracy of the spoken English pronunciation quality score of the algorithm in this paper is higher.

4.2.3. Comparison of Algorithm Response Time. The intelligent assessment algorithm for spoken English pronunciation quality requires extremely high performance for responsive time, and the trainer's pronunciation recording should quickly output the words that need to be corrected. Therefore, the response time is also one of the key indicators of the detection system performance. The experiment uses 100 individual word data as the test data and does not include the collected time. From the initial input to the end of the spoken English pronunciation quality evaluation, the entire process is used. The results of the test comparison are shown in Figure 3.

Analyzing Figure 3, it can be seen that, in the process of the oral English pronunciation test of 100 phrases, the response time of the spoken English pronunciation quality score of the algorithm in this paper is 2.4 s, and the response time of the spoken English pronunciation quality score of the algorithm in [3] is 8.2 s. In [4], the response time of the algorithm's spoken English pronunciation quality score is 6.0 s. The experimental results show that the response time of the algorithm in this paper is shorter, and the accuracy of its spoken English pronunciation quality score is higher, and it can efficiently and accurately realize the intelligent assessment of spoken English pronunciation quality.

5. Conclusion

This paper proposes a convolutional neural network intelligent assessment algorithm for spoken English pronunciation quality, selects a more complex GMM-HMM model than softmax in the original CNN for training and recognition, and builds a CNN-GMM-HMM speech recognition model system. Through audio recognition, the feature screening and classification recognition of spoken English pronunciation are realized, and the PID algorithm is used to extract the emotional elements of spoken English pronunciation, so as to realize the accurate assessment of the quality of spoken English pronunciation. Experiments have proved that the intelligent assessment algorithm of spoken English pronunciation quality based on the convolutional neural network can improve the correct rate of oral English pronunciation error detection and obtain efficient and accurate pronunciation quality assessment results.

Data Availability

The data used to support the findings of this study are available upon request to the author.

Conflicts of Interest

The author declares that he has no conflicts of interest.

References

[1] W. J. Hwang, T. M. Tai, B. T. Pan, B. T. Pan, T. Y. Lou, and Y. J. Jhang, "An intelligent QoS algorithm for home networks," *IEEE Communications Letters*, vol. 23, no. 4, 2019.

[2] X. Li, "Characteristics and rules of college English education based on cognitive process simulation," *Cognitive Systems Research*, vol. 57, no. 8, pp. 11–19, 2019.

[3] Y. Wen, "Design of DTW algorithm based automatic correction system for English pronunciation mistakes," *Modern Electronics Technique*, vol. 43, no. 10, pp. 124–126, 2020.

[4] D. Luo, L. Xia, C. Zhang, and L. Wang, "Automatic scoring of L2 English speech spoken by Chinese middle school students based on," *Deep Learning*, vol. 16, no. 2, pp. 100–104, 2021.

[5] N. Hu, S. Wu, and Y. Zhang, "Loop closure detection for visual SLAM based on convolutional neural network," *Computer Simulation*, vol. 37, no. 5, p. 5, 2020.

[6] B. D. Barkana and A. Patel, "Analysis of vowel production in Mandarin/Hindi/American- accented English for accent recognition systems," *Applied Acoustics*, vol. 162, no. 5, Article ID 107203, 2020.

[7] B. Wu, J. Zhou, H. Yang et al., "An ameliorated deep dense convolutional neural network for accurate recognition of casting defects in X-ray images," *Knowledge-Based Systems*, vol. 226, no. 6, Article ID 107096, 2021.

[8] J. Jendeberg, P. Thunberg, and M. Lidén, "Differentiation of distal ureteral stones and pelvic phleboliths using a convolutional neural network," *Urolithiasis*, vol. 27, no. 2, pp. 1–9, 2021.

[9] T. Zhang, Y. Yang, J. Wang et al., "Comparison between atlas and convolutional neural network based automatic segmentation of multiple organs at risk in non-small cell lung cancer," *Medicine*, vol. 99, no. 34, Article ID e21800, 2020.

[10] K. Seddiki, P. Saudemont, F. Precioso, N. Ogrinc, and A. Droit, "Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification," *Nature Communications*, vol. 11, no. 1, pp. 1–10, 2020.

[11] L. Deng, "Design of automatic evaluation system for spoken English pronunciation quality," *Automation & Instrumentation*, no. 6, pp. 175–179, 2019.

[12] G. Luo and F. Zhao, "Design of an automatic scoring system for oral English test based on sequence matching," *Automation & Instrumentation*, no. 6, pp. 87–90, 2020.

[13] C. V. Trappey, A. J. C. Trappey, and S. C.-C. Lin, "Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies," *Advanced Engineering Informatics*, vol. 45, no. 4, Article ID 101120, 2020.

[14] M. Alghamdi, "Optimizing Arabic speech distinctive phonetic features and phoneme recognition using genetic algorithm," *IEEE Access*, vol. 8, pp. 395–411, 2020.

[15] Q. Xu, Q. Guo, C. X. Wang et al., "Network differentiation: a computational method of pathogenesis diagnosis in traditional Chinese medicine based on systems science," *Artificial Intelligence in Medicine*, vol. 118, no. 7724, Article ID 102134, 2021.

[16] L. Minmin, H. Jiang, H. Yule et al., "A systematic review on botany, processing, application, phytochemistry and pharmacological action of Radix Rehmnniae," *Journal of Ethnopharmacology*, vol. 285, Article ID 114820, 2021.

[17] W. U. Xin-Yu, "Research on the problems and solutions of the evaluation of the quality of undergraduate English classroom teaching in colleges and universities," *Education Teaching Forum*, no. 50, pp. 251-252, 2019.