Hindawi
Advances in Multimedia
Volume 2022, Article ID 2424380, 9 pages
https://doi.org/10.1155/2022/2424380

Research Article

Exploring the Teaching Mode of English Audiovisual Speaking in Multimedia Network Environment

Shunlan Wang

School of Arts and Sciences, Nanning College of Technology, Nanning 530105, Guangxi, China

Correspondence should be addressed to Shunlan Wang; wsl2022@nnct.edu.cn

Received 27 May 2022; Revised 16 June 2022; Accepted 28 June 2022; Published 15 July 2022

Academic Editor: Qiangyi Li

Introducing multimedia network tools in English audiovisual teaching and building a new model of network-based multimedia teaching can make English audiovisual teaching more in line with students' cognitive thinking characteristics and processes. This can improve the overall efficiency of English teaching in schools. Computers have been widely used in language evaluation and speech recognition for language learning, and speech recognition technology is an important reflection of the level of language learning. The large amount of language signal data, complex pronunciation changes, and high dimensionality of pronunciation feature parameters in the language learning process make it difficult to identify pronunciation features. The computational volume of pronunciation evaluation and recognition is too large, which requires high hardware resources and software resources to realize high-speed processing of massive pronunciation signals. To address the problem of low recognition rate of English pronunciation, this study proposes a sound recognition algorithm based on adaptive particle swarm optimization (PSO) matching pursuit (MP) sparse decomposition. The algorithm firstly improves the parameter adaptive setting of PSO based on the particle and population evolution rate, establishes parameter adaptive PSO, and realizes the optimization of adaptive PSO optimized MP sparse decomposition. The continuous Gabor super-complete atomic set is constructed based on the continuous space search property of PSO to improve the optimal atomic matching of the evolutionary process. Finally, the recognition of English pronunciation is realized by the support vector machine (SVM) algorithm. The test results show that the misjudgement rate for different mispronunciations is less than 1% when the system is used to evaluate the English pronunciation level. It proves that the method can effectively detect the mispronunciation and has high evaluation accuracy.

1. Introduction

At the time when the postindustrial age of human society is gradually going away, the knowledge economy is advancing rapidly and getting unprecedented development and attention. In particular, the development and application of network technology and multimedia technology have gradually popularized online education in Chinese education [1]. This kind of communication and popularization is changing the traditional learning time, learning method, and learning environment of contemporary Chinese students. At present, many schools carry out English audiovisual teaching and training directly on LAN or campus network, so that network technology is spreading rapidly and gradually popularized in English teaching.

Multimedia technology relies on the Internet and applies it to English audiovisual teaching [2]. It can also simulate the audiovisual content that is difficult to be expressed by traditional teaching methods through images, sounds, videos, and other forms. This makes the expression of English teaching more visual and intuitive.

Multimedia audiovisual teaching is characterized by informationization, diversification, and autonomy of learning environment, which fully embodies the learning theory of "input hypothesis" put forward by American applied linguist Krashen. According to Krashen, the ideal learning input should have four characteristics: comprehension, interesting and relevant, nongrammatical program ordering (not grammatically sequenced), and sufficient input (I + 1). By increasing the amount of language input and reducing the emotional filter factor, students' listening and speaking ability can be improved based on high-quality language output (speaking). In the multimedia network environment, dynamic and real corpus and scene input cultivate students' interest, create a relaxed language atmosphere, effectively promote the diversification and individuation of students' learning behavior, fully mobilize students' enthusiasm and initiative to acquire knowledge, and effectively improve students' English listening and speaking ability.

As an international language, English has been valued by many countries. The enthusiasm for English learning in China is constantly rising, with various kinds of English learning software and platforms emerging in an endless stream. However, in the whole learning process, due to the lack of evaluation and feedback correction of spoken English pronunciation, most of the learning ability of listening and speaking is weak, and it is difficult to achieve standard colloquial communication. Speech recognition technology is used to assist English pronunciation learning, which effectively corrects learners' wrong pronunciation [3] such as FLUENCY foreign language pronunciation system, EduSpeak voice system, and PLASER voice pronunciation training system, which are more maturely applied at present. Different pronunciation systems provide recognition and capture of speech signals, classification based on English pronunciation, feedback scoring based on language popularity and duration, etc., but all kinds of platforms have certain defects. For example, dynamic time warping (DTW) is used to train and identify English words and sentences, which effectively reduces the matching computation. However, the similarity between words is not taken into account in pronunciation, which makes it difficult to achieve the comparison of pronunciation evaluation. The PLASER speech system is based on the confidence of the factor score of English words, which makes the speech signal very fuzzy and difficult to achieve accurate matching. These systems mainly focus on computer platform system, and it is difficult to achieve the current portable, timely training requirements.

Referring to the research results in the field of speech signal detection, most of the existing algorithms extract signal characteristics such as MCFF, short-term energy spectrum, correlation coefficient, acoustic spectrum, ESMD permutation entropy, multiband energy, and sparse synthesis NMF. α-distribution analysis and T distribution can be used to reduce the interference of the external environment. Then, traditional classifiers such as K-nearest neighbor, mixed Gaussian model, SVM, DNN, and their combination patterns can be used to detect and recognize English pronunciation. Although good recognition effect is achieved, the above algorithms often need large sample size to support training, and there are problems such as difficult to set the order reasonably, multi-hidden layer error, and strong background noise of public environment has strong interference to the feature extraction and detection and recognition of the existing algorithms.

With the rise of deep learning, deep neural network has also been applied to deal with English pronunciation recognition and classification [4]. Based on the improved deep learning convolutional network, Lin et al. [5] extracted multi-scale normalized local features with stacked decreasing convolution kernel, improved the convergence speed and stability of the algorithm with dynamic learning rate, and achieved better recognition rate. Jia et al. [6] used three-hidden layer deep neural network to identify MCFF features of acoustic signals and achieved better recognition results than SVM and GMM. Compared with traditional classifiers, deep learning network improves the detection accuracy of pronunciation, but its huge parameter requirements, complex parameter settings, and calculation requirements require further optimization and improvement in practical applications [7].

Matching pursuit (MP) achieves sparse signal decomposition and noise reduction without prior information and has a good adaptability to the acoustic signal under the interference of external environment. Zhou et al. [8] extracted acoustic signal features with the help of over-MP sparse decomposition and detected abnormal acoustic signals with DBN. Wang et al. [9] used secondary sparse decomposition and reconstruction of sound signals to eliminate background noise interference and then extracted features of reconstructed signals for recognition. Wang and Ding [10] used PCA and LDA to extract the features of acoustic signals for acoustic signal detection and recognition after MP sparse decomposition of acoustic signals.

Building on these studies, this study proposes a sound recognition algorithm based on adaptive particle swarm optimization MP sparse decomposition. Firstly, the algorithm improved the parameter adaptive setting of PSO based on the particle and population evolution rate, established the parameter adaptive PSO, and realized the optimization of MP sparse decomposition of the adaptive PSO optimization. Based on the continuous space search feature of PSO, the continuous Gabor super-complete atom set was constructed to improve the optimal atom matching degree in the evolutionary process. Finally, the recognition of English pronunciation was realized by the SVM algorithm. The results demonstrate the effectiveness and robustness of the proposed algorithm.

The innovations of this study are as follows:

(1) The adaptive setting of particle swarm parameters is improved, and the objective function of sparse decomposition based on the evolution rate of particle and population is optimized.

(2) The continuous super-complete Gabor atom set is established based on the continuous set search characteristic of the adaptive particle swarm optimization algorithm to improve the matching degree of the best matched atom and the acoustic signal and speed up the atom matching search.

(3) SVM classifier is used to realize compound feature recognition of English pronunciation.

This study consists of five main parts: the first part is the introduction, the second part is state of the art, the third part is methodology, the fourth part is result analysis and discussion, and the fifth part is the conclusion.

2. State of the Art

As most English teachers grow up under the background of examination-oriented education, their teaching philosophy is largely influenced by the examination-oriented education philosophy, which leads them to follow the traditional teacher-centered teaching mode to carry out English audiovisual teaching [11]. In the phonic class, the teacher constantly explains, plays the recording to the students, and demonstrates and leads the students to pronounce. In this teaching mode of repeated playback, mechanical imitation, and parroting, there is no effective interaction between teachers and students. Students can only passively accept phonetic knowledge and lack flexible practice opportunities. The way is boring, and it is difficult for students to develop interest in learning.

With the continuous advancement of education informationization construction in China, most schools are equipped with multimedia phonetic teaching equipment, which provides convenience for English audiovisual speaking teaching [12]. Although most teachers can use multimedia technology, it is limited to using multimedia technology for the presentation of phonetic knowledge. Unable to make good use of multimedia technology leads to the advantages of multimedia technology that cannot be fully reflected, resulting in the waste of multimedia teaching resources.

Students' listening and speaking ability is generally low. The factors causing this phenomenon are diversified, such as inappropriate learning methods, being influenced by mother tongue, and lack of language environment. [13]. This also makes a considerable number of students in the pronunciation course, when listening to the recording, and cannot accurately carry out oral pronunciation; as time goes by, students will gradually lose interest in English pronunciation learning. In the pronunciation class, those students showed low enthusiasm, were even afraid of the teacher to call on the roll to speak English, and tried to avoid all kinds of English communication activities. The existence of this problem will seriously affect the learning confidence of students and is not conducive to the improvement of students' English listening and speaking ability.

In the present situation of audiovisual English teaching in schools, students' subjectivity has not been fully reflected. Phonetic courses mainly focus on teachers' explanation of basic phonetics knowledge and test students' pronunciation learning effect by means of written knowledge assessment. However, students lack the opportunity to practice phonetic skills. This teacher-centered explanation replaces the student-centered learning mode, which has a negative impact on students' English listening and speaking ability.

Students at school have high self-esteem. Although they crave approval and praise from others, they are also afraid of making a fool of themselves in front of others, which makes many students dare not speak English in public for fear of being laughed at by other students. Therefore, few students take the initiative to speak in school English pronunciation class, and those who are asked to speak are more likely to make mistakes because of greater psychological pressure. In addition, there are some students with weak psychological quality; after making mistakes in class, they will become more inferior and dare not speak and speak English in class. This makes speech errors more difficult to correct. In this vicious cycle, their English listening and speaking ability is more difficult to improve.

3. Methodology

3.1. English Pronunciation Recognition Algorithm. Let the element a_x in set D $= \{a_x, x = 1, 2, \ldots, V\}$ be the unit vector of space $B = R^T . f = g \cdot a$, where g is the expansion coefficient and $a = \{a_1, a_2, \ldots, a_w\}$ is the sparse decomposed atom set. Among all the expressions, the minimum value of m is the sparse decomposition of $f \in$ B.

In practical application, the discreteness of atomic set and the redundancy of super-complete set are contradictory to some extent. PSO has the property of continuous space search. If it is introduced into MP sparse decomposition process, it can improve the influence of atom set discretization.

To avoid this problem, the particle needs a higher speed inheritance in the initial iteration to maintain the global search capability and a higher local search capability in the later iteration to maintain the stable solution. Based on this, a parameter adaptive adjustment strategy is proposed.

The evolution capability of particle and population during algorithm iteration is as follows:

$$\begin{cases} E_x^n = \dfrac{\left(f_m^n - f_x^n\right)}{\left(f_m^n - f_h^n\right)}, \\ \\ E_a^n = f_a^n - f_a^{n-1}, \end{cases} \tag{1}$$

where f_x^n and f_{a}^n are, respectively, the fitness values of particle x and the optimal particle in population history at n iterations. f_{m}^n and f_{h}^n were the maximum and minimum values of fitness in the population. The evolution rate of the particle can be expressed as follows:

$$W_x^{n+1} = \frac{1}{\sqrt{\left(E_a^n\right)^2 + \left(E_x^n\right)^2 + 1}}. \tag{2}$$

Evolution describes the information inheritance of a particle to the last iteration, and the adaptive setting of the inertial weight factor of the algorithm is as follows:

$$m_x^{n+1} = m_{\text{init}} - \left(m_{\text{init}} - m_{\text{end}}\right) \times W_x^{n+1}. \tag{3}$$

According to formulas (2) and (3), a larger population evolution rate indicates that after iteration, some of its particles can obtain better solutions than the current iteration; that is, they have better exploration ability. In subsequent iterations, these particles should conduct global optimization with a larger inertia factor. Otherwise, a local search is performed with a smaller value. When the E_x^n of particle x is large, its evolution rate will be affected and decreased, and the corresponding m_x^{n+1} inertia factor will be affected by the evolution rate and inherited more information of the particle in the last iteration. On the contrary, the particle m_x^{n+1} is small and is less affected by the information of the previous iteration. In formula (3), the parameter values are $m_{init} = 0.9$ and $m_{en\ d} = 0.4$.

It can be seen that the particle inertia factor of the next iteration is affected by the particle and population evolution ability of the previous iteration. The global optimization ability of a particle affects its search range, and the particle itself determines the setting of its independent inertia factor.

It can be seen that the particle inertia factor of the next iteration is affected by the particle and population evolution ability of the previous iteration. The global optimization ability of a particle affects its search range, and the particle itself determines the setting of its independent inertia factor.

Self-learning factor c_1 and social learning factor c_2 control the influence of individual and group historical optimal iteration on the subsequent information inheritance, respectively. In this study, we adjust the learning factor according to the evolutionary change rate of particles to increase their own learning ability in the early stage and social learning ability in the later stage and realize the optimal global search and precise local search. At the same time, the different evolution rates between particles due to different evolutionary abilities can further adjust the learning factor, so that the particles can adjust the learning mode according to their own iteration, thus increasing the diversity of particles. The learning factor control method in this study is as follows:

$$\begin{cases} c_{1x}^{n+1} = \dfrac{c_{1\max} - \left(c_{1\max} \cdot \sin\left(W_x^{n+1} \cdot \pi/2\right)n\right)}{N}, \\[2ex] c_{2x}^{n+1} = \dfrac{c_{2\max} - \left(c_{2\max} \cdot \sin\left(W_x^{n+1} \cdot \pi/2\right)n\right)}{N}, \end{cases} \tag{4}$$

where $c_{\max}$ and $c_{\min}$ represent the maximum and minimum value of learning, respectively. Formula (4) shows that each particle of the improved algorithm adaptively adjusts its learning factor according to its evolutionary change rate. In the initial iteration, the particle has strong self-learning ability, so its c_1 value is dominant. If the evolution rate of a particle is small at this time, the value of c_1 is slightly larger than that of other particles of the same generation, which is more conducive to global optimal search. With the deepening of iterative evolution, the social learning ability of particles becomes stronger, and its c_2 value dominates. If the evolution rate of a particle is large at this time, its c_2 value is slightly larger than that of other particles of the same generation, which is more conducive to local optimization search for the particle.

To avoid local optimization, particle recombination strategy is adopted in this study. In the iterative process of particle evolution, some particles with weak evolutionary ability are reselected to learn from the particles with strong evolutionary ability and recombine these particles. A random number $u_{x\ d}$ is generated for each dimension of each recombination particle. If $u_{\text{xd}} > U_{\text{c}}$ (U_{c} is the learning probability), the particle i_{xd} learns from the dimension i_{zd} of the particle with stronger evolutionary ability in the d-dimension. If $u_{x\ d} \le U_c$, particle i_{xd} remains unchanged, so the particle recombination process is as follows:

$$i_{\text{xd}}^{\text{New}} = \begin{cases} i_{z\text{d}}, & u_{x\text{d}} > U_c, \\ i_{x\text{d}}, & u_{x\text{d}} \leq U_c. \end{cases} \tag{5}$$

After recombination, the particles enter the iterative process together with the particles with strong evolutionary ability. The recombination strategy not only widens the search range of particles effectively but also ensures the speed and search accuracy of the algorithm.

The key to complete signal sparse decomposition is to match the signal characteristics with the sound-signal characteristics of super-complete set atoms. Based on the continuous space search property of particle swarm optimization (PSO), this study uses the improved Gabor function formula to generate super-complete connected Gabor atom set D required for sparse decomposition.

$$a_\gamma(n) = \lambda \cdot a\left(\frac{n-p}{s}\right)\cos(qn + \omega). \tag{6}$$

Here, λ is the normalized parameter. The parameter set $\gamma_x = \{s, p, q, \omega\}$ is used to describe the characteristics of the atom, and its parameter set constitutes the spectral characteristics of the signal to be sparsely decomposed. The continuous Gabor set makes its atom number far exceed that of the discrete set, which ensures the redundancy of the atom set and the matching program of the optimally matching atom to the original signal structure.

According to the calculation process of MP sparse decomposition, the fitness value $f(i_x^w(z))$ was calculated by the inner product of the residual of signal decomposition in the iterative process when the objective function was constructed by the adaptive MP sparse decomposition target algorithm in this study. As shown in formula (7), to analyze whether the particle position is optimal, the optimal fitness value is searched through the iterative update of particle velocity and position.

$$f(i_x^w(z)) = \left|R^{w-1}, a_r^w\right|, \tag{7}$$

where a_{r}^w is the time-frequency parameter group of the atom with the optimal signal decomposition. R^w is residual generated in the process of sparse decomposition. $R^0 = s$. $<\cdot,\cdot>$ is the inner product operation. $f(i_x^w(z)) \in [0,1]$ describes the degree to which the acoustic signal matches the atom. Based on formula (7), the objective function and corresponding reconstruction function of the adaptive MP sparse decomposition algorithm in this study are designed as follows:

$$\begin{aligned} Y &= \max f(i_x^w(z)) a_x \in D, \\ f_s &= \|f\| \cdot Y \cdot a_{\text{best}}. \end{aligned} \tag{8}$$

Here, a_{best} is the atom set of optimal matching. The time-frequency parameter $\gamma_{\text{best}} = \{s, p, q, \omega\}$ of a_{best} reflects the characteristics of pure acoustic signal containing noise.

According to the above description, the computational flow of the adaptive particle swarm optimization MP sparse decomposition algorithm in this study is shown as follows:

(1) Initialize the relevant parameters of the improved particle algorithm. The boundary conditions were set

as $[i_{\min}, i_{\max}]$ and $[q_{\min}, q_{\max}]$, the initial position and velocity of particles were randomly generated, and the fitness value $f[i_x(z)]$ was calculated.

(2) Update the velocity and position of particles, and limit the transgression according to the boundary values $u_{\mathrm{h}x}$ and $a_{\mathrm{h}x}$.

$$\begin{cases} q_x^{z=1} = m \cdot q_x^z + c_1 \cdot r(\cdot)\left(u_{hx} - i_x^z\right) + c_2 \cdot r(\cdot)\left(a_{hx} - i_x^z\right), \\ i_x^{z+1} = i_x^z + q_x^{z+1}, \end{cases} \tag{9}$$

where r(·) is a random function with a random value between (0,1]. m is the inertia factor of the adaptive value. If m value is too large, the particle will over speed and jump out of iteration, while if m value is too small, it is not conducive to algorithm convergence. Therefore, based on the adaptive value, the inertia factor is further adjusted as follows:

$$m = m_{\max} - z \cdot \frac{m_{\max} - m_{\min}}{Z}. \tag{10}$$

(3) Judge whether the particle velocity and position are out of bounds. If so, the boundary value is used instead. $f[i_x^w(z+1)]$ is updated to update population and individual optimality. Let iteration $z = z+1$ and go to Step (2) iterate until $z \geq$ zmax, record a_{best} and corresponding γ_{best}, and update the residual of sparse decomposition.

$$R^w = R^{w-1} - R^{w-1}, a_r^w \cdot a_r^w. \tag{11}$$

(4) Acoustic signals are reconstructed from formula (9) for subsequent detection and recognition.

After sparse decomposition, the time-frequency characteristics of the optimally matched atoms can better match the structural characteristics of the original signal. Based on literature [14], using the scaling factor and frequency factor of $p(s_\lambda, q_\lambda)$ and standard deviation of $\sigma(s_\lambda, q_\lambda)$ and ESMD permutation entropy and MFCC composite characteristics $F(\lambda) = \{p(s_\lambda, q_\lambda), \sigma(s_\lambda, q_\lambda), F_{\text{ESMD}}(\lambda), F_{\text{MFCCS}}(\lambda)\}$ At the same time, the problem of multi-node arrangement structure selection, which is closely related to classification results, is solved. An improved SVM multi-classification extension algorithm based on dichotomous SVM with maximum recognition rate as the root node of decision-oriented acyclic graph is proposed. Different decision methods are used to process different data without increasing the amount of computation so as to optimize the training and final decision accuracy. The recognition rate is defined as follows:

$$u = \left(\frac{P_1}{P}\right) \times 100\%. \tag{12}$$

Here, P and P_1 represent the total number of evaluation samples and the number of accurately classified samples, respectively.

3.2. System Design. The architecture diagram of the system is shown in Figure 1. The voice recognition engine is invoked to provide English learning services for users through the Apache server. The database mainly includes user management database, basic words, and grammar, which were used to manage the system user information (basic information, learning, curriculum information, etc.), basic word information (spelling, polysemy, and other information), and grammatical information (common syntax information and correlation information). By running the speech recognition engine and the intelligent processing middleware on the server, the accuracy of the user's English sentences can be judged according to the grammar rules.

Apache works using URL to request corresponding resources. The server will operate according to the corresponding identification algorithm of the program according to the user request, return the resources found to the client, that is, to complete a request, and then wait for the next request.

Software is mainly divided into background part and foreground part, according to the actual needs of the software design backend and frontend function modules. The background module mainly completes user management, data management, and system operation and maintenance. The foreground module is mainly customer operation module, including user login, English listening and speaking, and other functions. The functional composition structure is shown in Figure 2.

3.2.1. Background Functions. The user management module mainly completes the management personnel's operation response, including the system administrator's account, password, email, and other information. This section describes how to add and delete administrators. The system background is logged as a superadministrator and the preceding functions are performed, while a common administrator can only manage some common basic data.

The data management module mainly contains data recording, and data download two main functions. The data recording function is mainly to input the basic data required by the system, such as commonly used words and grammar rule information, mainly including textbook management, article management, and sentence management units. The data download module is to respond to the user URL request, complete the allocation of resources download on the Apache server, and return the customer request information.

The system operation and maintenance module is for the administrator to performance optimization and other work, including the maintenance of the system foreground and background interface (see Table 1).

3.2.2. Foreground Functions. After entering the user name and password, request message can be defined as < iq type = "get" id = regl><query xmlns = "jabber: iq: Register "/> </iq >, the server side after parsing back to the client side login success or not information.

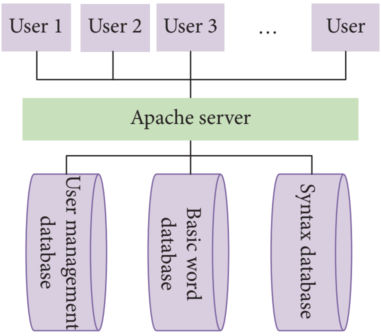

Figure 1: System framework.

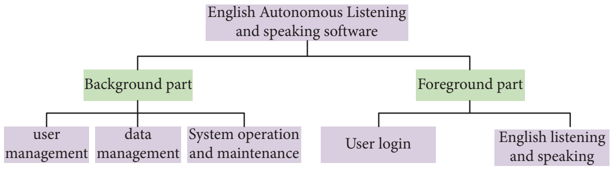
English Autonomous Listening and speaking software
Background part
Foreground part
user management
data management
System operation and maintenance
User login
English listening and speaking

Figure 2: System function module composition diagram.

Table 1: Interface detailed parameters of statement.

Parameter	Description
Id	Statement id
English	Original English
Chinese	Chinese translation
Voice URL	Audio files

Users can obtain the required data through the above interfaces.

The English listening and speaking module includes text statement selection, native statement playback, recording, voice playback, and other functions. Users select different function buttons based on their own needs, and the server responds to the user's request in combination with the voice recognition engine. As the main function of the software, this module accounts for 80% of the system functional requirements.

3.2.3. User Usage Process. The user usage flow is shown in Figure 3.

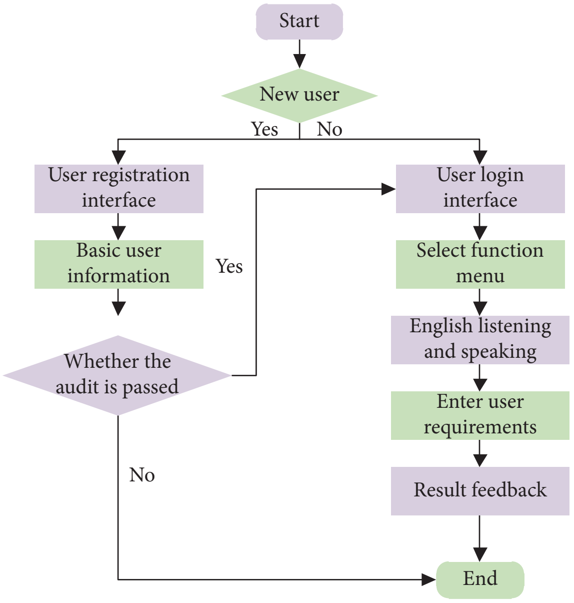

Figure 3: User usage flow chart.

4. Result Analysis and Discussion

4.1. English Pronunciation Error Detection. In the process of spectrogram extraction, FFT window size, frameshift, maximum frequency, and jump size were set to 20 ms, 10 ms, 16000 Hz, and 160.2 d, respectively. The dimensions of convolution kernel and pooling layer are 3 × 3 and 2 × 2. The dropout value during convolution is 0.5. Because the model contains CTC loss function, it is implemented by TensorFlow and Keras. Softmax of the input tag (annotated tag sequence), tag length, input length, and model output is passed to the CTC loss function to calculate the loss.

To evaluate the performance of the pronunciation error detection system, this study designs evaluation indexes by referring to the hierarchical evaluation structure developed in the literature [15]. There are four types of test results: true acceptance (AT), true rejection (RT), false rejection (RF), and false acceptance (AF), as shown in Table 2. Based on these four test results, the system performance was measured using the false acceptance rate (RFA), false rejection rate (RFR), and accuracy diagnosis (AD). RFA represents the percentage of learners' mispronunciation detected by the system as correct pronunciation. RFR represents the percentage of learners' correct pronunciation detected by the system as incorrect pronunciation. AD is the diagnostic accuracy rate of the system. The calculation formula is as follows:

$$\begin{aligned} R_{FR} &= \frac{R_F}{R_F + A_T}, \\ R_{FA} &= \frac{A_F}{A_F + R_T}, \\ A_T &= \frac{A_T + R_T}{A_F + R_F + A_F + R_T}. \end{aligned} \tag{13}$$

In the above three evaluation indexes, it is hoped to reduce the error rate of the other two types as much as possible while ensuring high diagnostic accuracy. The key is to avoid undermining learners' learning confidence by judging their correct pronunciation as incorrect pronunciation, so the experiment aims at a high diagnosis rate and a low false rejection rate. Experimental results of different models for English pronunciation error detection are listed as follows (see Table 3 and Figure 4).

By comparing the proposed model with other 5 models, the results show that the proposed model achieves better results in false rejection rate (FRR) and diagnostic accuracy (DA). Compared with the model in the literature [16–20], the model in this study has achieved better results in false acceptance rate (FAR).

Table 2: Classification of experimental results.

Experimental results	Explanation
True acceptance (AT)	AT is the number of correct pronunciation
True rejection (RT)	RT is the number of incorrect pronunciation
False rejection (RF)	RF is the number of correct pronunciation
False acceptance (AF)	AF is measured by the number of incorrect pronunciations

Table 3: Detection results of different models (%).

Acoustic model	FAR	FRR	DA
Literature [16]	30.68	19.83	80.13
Literature [17]	27.64	38.19	82.94
Literature [18]	22.25	31.93	78.26
Literature [19]	42.01	10.03	82.63
Literature [20]	15.22	7.25	87.95
Proposed	14.85	4.55	89.99

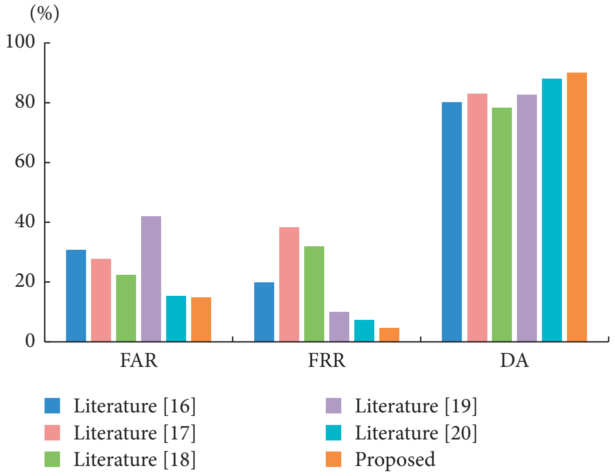

Figure 4: Test results of different models (%).

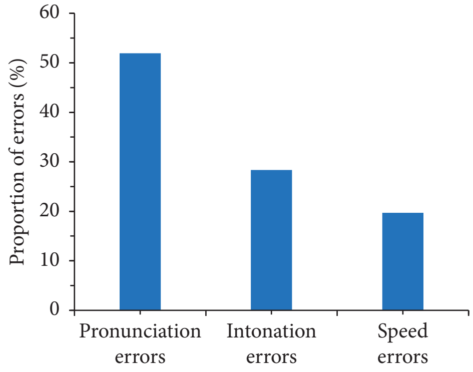

Figure 5: Detection results of three types of false pronunciation in this model.

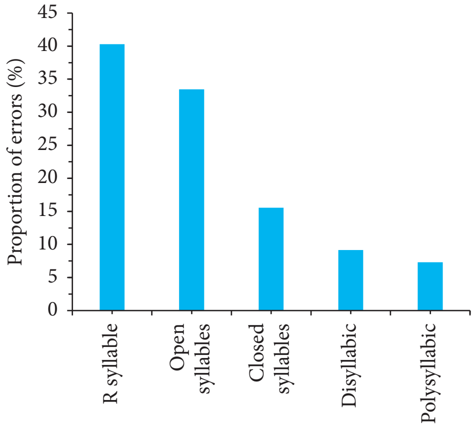

Figure 6: Detection results of pronunciation errors of 5 different syllables.

In English pronunciation errors, the 64 errors are divided into three types: pronunciation errors, intonation errors, and speed errors. Statistical results of these three types of errors are shown in Figure 5.

Although British English and American English are both English, there are obvious differences in their pronunciation systems:

(1) American English has a distinct "r" sound, while British English does not. For example, the word worker is pronounced as |'w∂:rk∂| in American English and |'w∂:k∂| in British English.

(2) The word |a:| is pronounced in British English and |æ| in American English. For example, the word pass is pronounced |pa:s| in American English and |pæs| in American English and similar words such as ask.

(3) British English reads the sound of |O|, American English reads |a:| such as the word box, British reads |bOks |, and British reads |ba:ks|, and similarly watch.

(4) British English is used to skim words, while American English is used to read each syllable in its entirety. For example, the word interesting is pronounced as |'intristiŋ| in British style and |'int∂ristiŋ| in American style.

(5) The British English pronounces |*i* | sound, and the American English pronounces | ∂ |; e.g., the word system is pronounced as |'sistim| in British style and |'sist∂m| in American style.

(6) There are some words that are pronounced completely differently in British English and American English. For example, leisure in British is |'leʒ∂| and in American is |'li:z∂r|.

Figure 6 shows the pronunciation bias of English students for words with 5 middle syllables, and the results show that they perceive the *r* syllable poorly.

Table 4: English pronunciation level evaluation results.

Voice sequence number	Tone/points	Speed/points	Intonation/points	Rhythm/points	Emotion/points	Final score/points
1	8.6	8.5	8.4	8.5	6.9	8.8
2	8.6	8.6	8.5	8.6	7.6	8.6
3	8.7	8.2	6.7	6.7	7.5	7.8
4	9.3	4.9	7.4	5.9	6.2	6.9
5	8.7	8	5.2	4.3	8	6.8
6	7.2	8.5	8.7	8	6.8	8
7	7.6	5.9	5	10.6	5.2	6.9
8	8.9	8.7	9.6	8.5	8.9	9.3

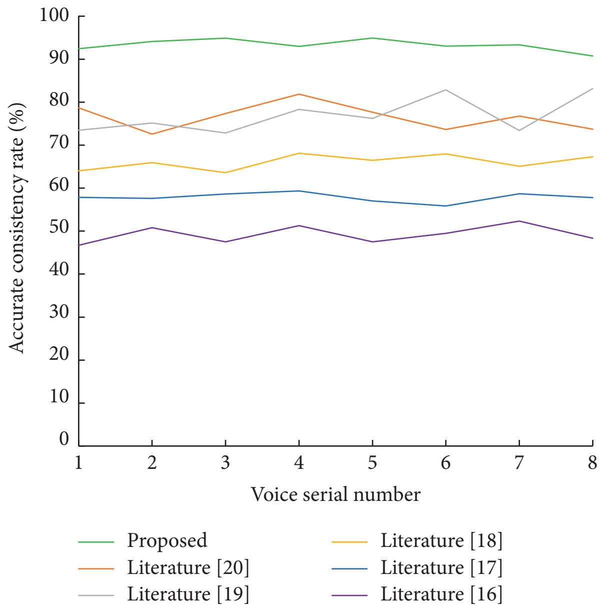

Figure 7: Accuracy comparison results.

4.2. Evaluation of English Pronunciation Quality. To test the practical validity of the English pronunciation evaluation system based on the model of this study, the following experiment was designed.

The experimental environment is as follows: a student majoring in business English in a foreign language school is selected as the experimental object, and MATLAB software is used to program this system. Eight linear FM signals of the student's spoken English pronunciation were collected, and the time width and relative bandwidth of the collected speech samples were 1.5 s and 0.5 dB, respectively, and the collected frequencies of the spoken English pronunciation signals with different vocal cords and baseband signals were 1024 kHz and 3~9 kHz, respectively.

The good classification performance of the support vector machine helps the system to evaluate the English pronunciation level accurately. Based on this, the system was used to evaluate the English pronunciation level of the collected 8 speech segments (see Table 4).

Table 4 shows that the system can effectively evaluate five indicators of English pronunciation, tone, speed, intonation, rhythm, and emotion, and use the evaluation results of each indicator to make the final evaluation of English pronunciation level. It shows that this system can evaluate learners' English pronunciation level from different directions and has high evaluation validity.

Six systems were used to evaluate the pronunciation level of the student's 8 segments of spoken English, and the comparison of the accurate agreement rate of the six systems is shown in Figure 7.

5. Conclusion

Multimedia network technology integrates text, pictures, animation, video, sound, and other media forms into one. It has the advantages of combining visual and auditory senses into one, and its application to English audiovisual teaching can provide a broader platform and more choices for students' English audiovisual learning. The famous American linguist Krashen believes that language learning is mainly done by language input. The same is true for English audiovisual teaching, which requires students to continuously make phonetic input. Aiming at the present situation of inaccurate pronunciation, the traditional English pronunciation learning lacks pronunciation evaluation and error correction guidance. In this study, an adaptive MP sparse decomposition algorithm for abnormal acoustic signal recognition is proposed. Firstly, the adaptive setting of PSO parameters was improved based on the evolution rate of particle and population, and a new objective function was constructed to realize the adaptive MP sparse decomposition. Then, the feature matching degree between the optimal atom and the acoustic signal is improved by continuous super-complete set. Finally, SVM is used to realize the accurate recognition of English pronunciation. The results show that compared with the existing algorithms, this algorithm has the best recognition rate of English pronunciation and has better recognition robustness for different pronunciation systems. In the subsequent research, more complex characteristic parameters will be used to further improve the detection accuracy of English pronunciation recognition.[21].

Data Availability

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

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This work was supported by the Nanning College of Technology.

References

[1] X. Zhu, "The integration of media technology and the change of education (IMTCE) - a study of model of the impact of media technology on education," *International Journal of Information and Education Technology*, vol. 8, no. 6, pp. 422–427, 2018.

[2] Y. Zhao and C. Zhao, "Research on College English teaching model under the computer environment," *Social Networking*, vol. 08, no. 02, pp. 104–111, 2019.

[3] S. Manoharan and N. Ponraj, "Analysis of complex non-linear environment exploration in speech recognition by hybrid learning technique[J]," *Journal of Innovative Image Processing (JIIP)*, vol. 2, no. 04, pp. 202–209, 2020.

[4] A. Shewalkar, D. Nyavanandi, and S. A Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, no. 4, pp. 235–245, 2019.

[5] Y. Lin, D. Guo, J. Zhang, Z. Chen, and B Yang, "A unified framework for multilingual speech recognition in air traffic control systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3608–3620, 2021.

[6] Y. Jia, X. Chen, and J. Yu, "Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network[J]," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1749–1757, 2021.

[7] Y. Lu and H. Li, "Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory," *Applied Sciences*, vol. 9, no. 8, p. 1599, 2019.

[8] Q. Zhou, Y. Zhang, C. Yi, J. Lin, L. He, and Q Hu, "Convolutional sparse coding using pathfinder algorithm-optimized orthogonal matching pursuit with asymmetric Gaussian chirplet model in bearing fault detection," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 18132–18145, 2021.

[9] Y. Wang, W. Wang, M. Zhou, A. Ren, and Z Tian, "Remote monitoring of human vital signs based on 77-GHz mm-wave FMCW radar," *Sensors*, vol. 20, no. 10, p. 2999, 2020.

[10] B. Wang and C. Ding, "Hierarchical frequency-domain sparsity-based algorithm for fault feature extraction of rolling bearings," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6228–6240, 2020.

[11] S. E. Kaymakamoglu, "Teachers' beliefs, perceived practice and actual classroom practice in relation to traditional (Teacher-Centered) and constructivist (Learner-Centered) teaching (note 1)[J]," *Journal of Education and Learning*, vol. 7, no. 1, pp. 29–37, 2018.

[12] R. Riskandi, H. Syarif, and N. Gistituati, "Analysis of the needs of English learning models in basic schools[J]," *International Journal of Progressive Sciences and Technologies*, vol. 23, no. 1, pp. 202–206, 2020.

[13] N. I. A. Bakar, N. Noordin, and A. B. Razali, "Improving oral communicative competence in English using project-based learning activities[J]," *English Language Teaching*, vol. 12, no. 4, pp. 73–84, 2019.

[14] B. Hu, "The evaluation method of English teaching efficiency based on language recognition technology," *International Journal of Continuing Engineering Education and Life Long Learning*, vol. 30, no. 4, p. 445, 2020.

[15] Y. Mao, F. Xu, X. Zhao, and X Yan, "A gearbox fault feature extraction method based on wingsuit flying search algorithm-optimized orthogonal matching pursuit with a compound time-frequency atom dictionary," *Journal of Mechanical Science and Technology*, vol. 35, no. 11, pp. 4825–4833, 2021.

[16] J. Cai and Y. Liu, "Research on English pronunciation training based on intelligent speech recognition," *International Journal of Speech Technology*, vol. 21, no. 3, pp. 633–640, 2018.

[17] Z. Gang, "Quality evaluation of English pronunciation based on artificial emotion recognition and Gaussian mixture model," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 4, pp. 7085–7095, 2021.

[18] D. Ran, W. Yingli, and Q. Haoxin, "Artificial intelligence speech recognition model for correcting spoken English teaching," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 2, pp. 3513–3524, 2021.

[19] W. Yipu, "The function development of network teaching system to English pronunciation and tone in the background of internet of things," *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 5, pp. 5965–5972, 2019.

[20] L. Zhang, Z. Zhao, C. Ma et al., "End-to-End automatic pronunciation error detection based on improved hybrid CTC/attention architecture," *Sensors*, vol. 20, no. 7, p. 1809, 2020.

[21] T Lindeberg, "Spatio-temporal scale selection in video data [J]," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 4, pp. 525–562, 2018.