Design of Spoken English Distance Teaching Training System Based on Virtual Reality Technology

Zhi-li Sha[1(✉)] and Yu-xiang Liu[2]

[1] School of Foreign Languages, Xichang University, Xichang 615000, China
[2] College of Tourism and Urban-Rural Planning, Xichang University, Xichang 615000, China

Abstract. At present, the spoken English long-distance teaching training system has not designed a dynamic dictionary for spoken English teaching training, which leads to a high occupancy rate of the system server hardware during the operation of the system. A virtual reality technology-based remote teaching training system for spoken English is designed. In terms of hardware design, adopt the cloud-based MVC architecture model, design the system architecture, consider the system design purpose, and system functions, individually design the system I/O module, and complete the system hardware design. In terms of software design, it sets up voice recognition function, simulates the remote teaching environment, sets up a dynamic oral English teaching training vocabulary, increases the system vocabulary, sets up remote oral English teaching training rules, and the transformation system has the function of remote teaching and training oral English. The experimental results show that: comparing the three groups of systems, operating performance and server hardware performance, the English spoken remote teaching training system designed this time has a faster operating speed and a usable server utilization rate.

Keywords: Virtual reality technology · Spoken English · Distance teaching · Teaching and training system

1 Introduction

The increasingly frequent exchanges between countries have made the role of English as the language of international communication increasingly prominent in foreign exchanges. More and more people pay attention to the learning of English, their investment in English is increasing day by day, and the effect of English teaching has also been continuously improved [1]. However, the students taught in English teaching in our country generally have poor oral expression skills and high scores. For this reason, our country has put forward new requirements for proficiency in spoken English for English teaching. After discovering the shortcomings of English teaching, some reforms have been made to language teaching at home and abroad. The combination of information

W. Fu et al. (Eds.): eLEOT 2021, LNICST 389, pp. 502–514, 2021.
https://doi.org/10.1007/978-3-030-84383-0_43

technology and foreign language discipline can make up for the shortcomings of traditional language teaching, make foreign language learning more vivid and vivid, and stimulate learners' learning motivation. At the same time, foreign language learning is endowed with situational, communicative and practical features. However, in terms of oral English learning software, multimedia technology and speech recognition technology are mainly used to design language learning systems, such as FLUENCY, EduSpeak, SPHINX, English phonetic pronunciation software, Big Mouth English, and English talk-speaking software [2, 3]. However, in the above design, the design focuses on simple pronunciation and reading exercises, accumulating all kinds of English materials together, allowing learners to learn in the environment they build, this pure use of technology to accumulate materials the method does not change the essence of oral teaching, let alone the integration of information technology in English courses, and most of the functions are similar to repeaters. Therefore, the design of oral English distance teaching and training system based on virtual reality technology, through the cloud design of oral English distance teaching and training system architecture, using audiorecord class method to record voice signals, complete the I/O module design; On this basis, the relevant speech signal processing methods are used to extract the acoustic features of the required speech. On this basis, the speech template required for speech recognition is established. According to a certain matching strategy, the similarity between the test speech template and the standard speech template is calculated. Finally, the recognition result is expressed in a certain form, Complete the software design of oral English distance teaching training system. Finally, the effectiveness of the system is verified by simulation experiments. It has certain significance and design value both in theory and practical application.

2 Hardware Design of Distance Teaching and Training System for Spoken English

2.1 System Architecture Design

In order to meet the needs of users for remote learning of spoken English, the design of a distance teaching training system for spoken English adopts the cloud-based MVC architecture model. The cloud is also called the server, which mainly provides various Chinese learning resources for the client; the client is the user interface of the system, which is mainly responsible for various interactions with the user, but the main business logic is executed on the server [4]. Therefore, the main functions of the system in the cloud are as follows: 1. To achieve various operations on the database; 2. To provide a variety of rich knowledge base for the client; 3. To provide voice recognition services; 4. To provide word recognition services; 5. To provide upgrade scripts for the client; 6. To process various business logic of the system. The main functions of the client are as follows: 1. Responsible for direct interaction with the user and provide a graphical interface for the user; 2. Implement the cloud service proxy module and provide a cloud service interface for the client. Based on the above-identified main functions of the system cloud and client, the designed system architecture is shown in Fig. 1.

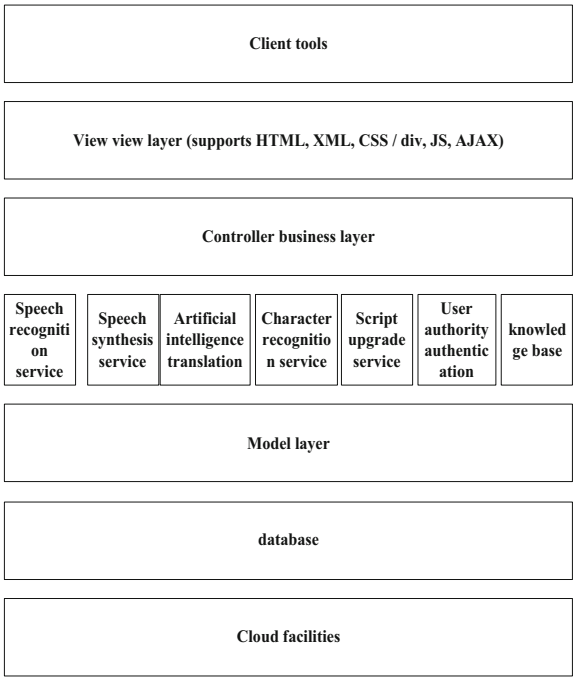

Fig. 1. System architecture

In Fig. 1, except for the client tool, everything else belongs to the server. It can be seen from Fig. 1 that the speech recognition service and text recognition service functions of the design system are placed on the system server side, simplifying the design of the client side, and achieving the cross-platform goal of the client side, thereby promoting the stability of the system client side. It runs locally on PC, android, ios and other platforms, so as to achieve the purpose of remote training of spoken English.

2.2 I/O Module Design

Based on the system architecture shown in Fig. 1, it can be found that the voice function of the English speaking distance teaching training system designed this time is an important part of the system. Therefore, the I/O module of the system is designed separately, that is, the system voice recognition and synthesis function.

The recording and playing of voice is the basis of realizing the system function [5]. As a long-distance oral English teaching and training system, voice recording and playing function is equivalent to the "ear" and "mouth" of the system, which is an indispensable part of human-computer interaction.

Since the system designed this time, its main purpose is to achieve remote training of spoken English, so the client chooses the mobile terminal that supports PC, android, ios and other systems, and the mobile terminal that runs on PC, android, ios, etc., comes with With a headset, it has achieved good results for general voice recording and playback. Therefore, the system uses the phone's built-in headset as the voice recording and playback device.

At present, the mobile client provides two implementation methods for recording, one is to use the relevant methods of mediarecorder class, the other is to use the method of audiorecord class [6]. In order to process speech signal in the future, it is necessary to determine the sampling rate, sampling bit and other parameters, and flexibly set the basic parameters of speech signal, such as the size of audio buffer, sampling rate, sampling bit, etc. Therefore, the system chooses the method of recording audio to the audiorecord class in the buffer to record the voice signal.

Consider that the AudioTrack class corresponds to the AudioRecord class and has the function of playing voice signals. Therefore, the method of the AudioTrack class is also used to play the corresponding voice signal. Based on the above analysis content, the final system audio format is determined as follows: 1. The sampling frequency is 8000 Hz; 2. The sampling channel is mono; 3. The sampling number is 16 bits.

In addition, the system has set the voice mouth demonstration function for all phonetic pronunciation, which is mainly realized by playing the built-in pronunciation animation video. The video playback uses the related methods of videoview class in Android SDK.

3 The Software Design of the Training System for Spoken English Remote Teaching Based on Virtual Reality Technology

Based on the hardware design of the spoken English distance teaching and training system, this paper designs the English speech recognition process, uses virtual reality technology to simulate the oral English teaching and training environment, sets up the dynamic vocabulary of oral English teaching and training, and formulates the training rules of oral English distance teaching, so as to make the system have the function of distance teaching and training spoken English, and complete the distance teaching of spoken English Training system software design.

3.1 Design an English Speech Recognition Process

Speech recognition includes input speech signal preprocessing and speech signal recognition. In the first stage, relevant speech signal processing methods are needed to extract the acoustic features of the required speech, and then the speech template for speech recognition is established. In the second stage, we need to extract the acoustic features of the speech signal from the processed input speech signal, and compare it with the existing speech template according to certain criteria. Then, according to certain search and matching strategies, we can find out a series of optimal templates that match the input speech, or calculate the test speech templates according to certain matching strategies Finally, the recognition results are expressed in a certain form. Its English speech recognition process is shown in Fig. 2.

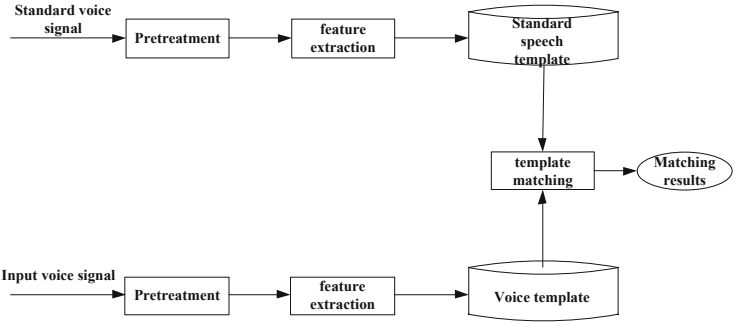

Fig. 2. English speech recognition process

Based on the English speech recognition process shown in Fig. 2, the speech signal preprocessing, feature extraction and recognition methods are designed. Considering that there will be problems such as unclear speech and dialectization in the process of English speech training, for this reason, the method of identifying speech sequence features in frames is used to recognize the speech sequence. Therefore, if the frequency domain of speech is z, the user's speech input sequence is H, n is used to represent the number of frames, and n value is taken as 20 ms. The overlapping part of the two frames is called frame shift, which is represented by m. in the process of processing the speech sequence, the generated transformation is represented by the symbol T, then the speech sequence Q_{n} at frame n is:

$$Q_{\mathrm{n}} = T[H(m)n(z-m)] \tag{1}$$

At this time (2) in the formula $n(z-m)$ can represent the minute hand hour of the speech sequence, multiplied by the amplitude. Incorporating the formula (1) into the speech sequence, after the second pre-emphasis processing is performed on the speech sequence, the Hamming window is used as the window function for speech signal processing, then:

$$w(\mathrm{z}) = \begin{cases} 0.54 - 0.46\cos\left[\frac{2z\pi}{(L-1)}\right] & 0 \le z \le L-1 \\ 0 & \text{Other} \end{cases} \tag{2}$$

In formula (2), L is the length of Hamming window function, and $w(z)$ is the w speech sequence in z frequency domain, which is the processed $H(\mathrm{z})$ speech sequence [7]. At this time, the $w(z)$ speech sequence has become smooth, and closer to the spectrum of short-term speech, improve the spectrum leakage caused by truncation effect, shield the interference information in the process of speech sequence input, and the mutual interference between speech frequency domains.

The extraction of speech features is to obtain an estimated value of a speech input sequence in the limited frequency domain of the speech recognition model. The estimated value and the actual value are analyzed for error, and the minimum square sum of

the analysis result is taken. The calculation process is (2) The speech input sequence processed by the formula is input to the speech recognition model, and the characteristics of the speech sequence are output.

$$\sum_{z} w(z)w(z-P) = \sum_{i=1}^{P} a_i \sum_{z} [w(z-P)w(\mathrm{z}-i)]^2 \tag{3}$$

In formula (3), P indicates that there are P sample values in speech sequence w, i.e. $P = 1, 2, 3, \cdots, P$. the predicted value of speech recognition model is a, the coefficient is i, and $i = 1, 2, 3 \cdots, i$. By solving equation (3), the speech feature parameters can be obtained, which reflect the characteristics of speech sequence.

Based on the formula (3), the language sequence recognition parameters are obtained, and the system speech recognition result B is obtained:

$$B_{Hw}(O) = P(O|H, w) = \frac{1}{(2H)^{P/2}\left|E_{Hw}\right|^{1/2}} \exp\left\{-\frac{1}{2}(O - e_{Hw})E_{Hw}^{-1}(O - e_{Hw})\right\} \tag{4}$$

Equation (4) is the result of system speech recognition, O is the language sequence output by navigation, e_{Hw} represents the error value between the input speech sequence and the output speech sequence, and E_{Hw} represents the mean square error between the input speech sequence and the output speech sequence Value [8].

At this time, the speech sequence input into the system is smooth and clear with obvious features, which can quickly identify the voice sequence input by users.

3.2 Simulation of Teaching Environment Based on Virtual Reality Technology

Using virtual reality technology to simulate oral English teaching and training environment, we need to rely on the actual teaching and training environment, use 3D scanning technology to scan the real image in a full range, measure the real image data, parameterize the measured real scene data, take the 3D scanning point as the original coordinate point, and look for the mapping point with the real environment in the virtual environment.

Assuming that the coordinate system of the real environment is aligned with the coordinate system in the virtual environment, the real environment is regarded as a rotation matrix R and a translation vector t. The virtual number q is used to represent the virtual environment corresponding to the real environment, w is a scalar, v = (x, y, z) is a vector, q Sitting marked as (w, v), then q satisfies $|q|=\sqrt{w^2 + x^2 + y^2 + z^2} = 1$. Then there are:

$$p = p_m(\Theta) = K\left[q|qt\right] \tag{5}$$

In formula (5), p is the matrix from the real environment to the virtual environment, Θ is the parameter vector, p_m is the matrix function of the parameter vector Θ converted

into the virtual environment, and K is the matrix calibration [9]. If (7) is brought into the equation, then:

$$\begin{cases} x = wt + x^2 \\ y = vt + y^2 \\ z = kt + z^2 \end{cases} \tag{6}$$

Through the calculation of formula (6), the coordinate points for transforming the real environment into the virtual environment can be obtained, and a point model of the virtual environment is generated, so as to construct a virtual scene by virtualizing the real world.

At this time, the operator can walk freely in the virtual environment through unique equipment, or observe the objects in the virtual environment by changing the visual position by operation, or further operation, so as to have the feeling of being in the real world. The virtual environment operation process is shown in Fig. 3.

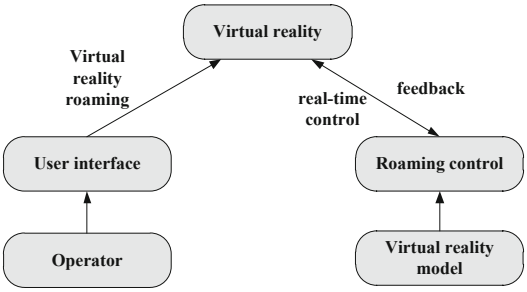

Fig. 3. Virtual environment operation process

According to the operation process of virtual environment shown in Fig. 3, the virtual environment can be adjusted in real time on the server according to the training content of oral English teaching, so that the virtual environment can be more in line with the training content of oral English teaching and improve the efficiency of distance teaching.

3.3 Setting up a Dynamic Vocabulary for Oral English Teaching and Training

At present, sphinx4 thesaurus is used in the spoken language training system. Its decoding operation module has strict requirements for the format of thesaurus. It provides six original thesauri for users to choose, namely discrete digital thesaurus TI46, continuous digital thesaurus tidigits, mini thesaurus an4, medium thesaurus RMI, large thesaurus WSJ, and super large Thesaurus hub4 [10].

However, the recognition process of sphinx4 depends on the hidden Markov model and speech model of each phoneme in the standard pronunciation in the dictionary, so the recognizer must load the selected dictionary before each recognition. The data contained in it is read into the memory, but this process takes a lot of time. The larger the capacity of the lexicon, the longer it takes to load the lexicon to start recognition. It takes about

2 s to load the RM1 thesaurus on a computer with an Intel Pentium 4 processor and a memory capacity of 2G. It takes about 5 s to load WSJ, and about 12 s to load HUB4, considering the user's computer configuration It is relatively low, and the loading of the thesaurus takes too long, which will require users to wait for a long time, which seriously affects user experience.

Therefore, the design of spoken English distance teaching training system thesaurus, choose to set the dynamic thesaurus, is to take out the vocabulary in the thesaurus, according to the words of the teaching system to dynamically rewrite thesaurus. The basic teaching module of College English teaching system is based on units. There are about 30 words to learn in a unit, so at the beginning of each unit, we will find the words to learn one by one, find the corresponding standard pronunciation and phonetic model of these words in hub4, and read these data into the dynamic new thesaurus. In this way, only the dynamic vocabulary needs to be loaded in each recognition process. The dynamic lexicon workflow is implemented by the system client, and the specific implementation of this design is shown in Fig. 4.

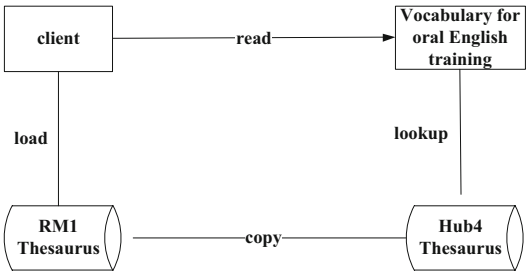

Fig. 4. Workflow of dynamic Thesaurus

3.4 Formulate Training Rules for Spoken English Distance Teaching

The user and the user should log in to the system and judge whether the students need the training together. When the user is an illegal user, the user needs to log in again; when the user is a legal user, the teacher sets up the virtual teaching environment of distance teaching, and the corresponding oral English teaching content, such as text, image, video, etc., for oral English teaching. According to the teacher's requirements, students make corresponding operation, record the spoken language training pronunciation, and compare it with the standard pronunciation. The teacher judges the result of oral English training in distance teaching and sets the corresponding teaching process.

The spoken English remote teaching training system designed this time is equipped with voice recognition function, simulating the remote teaching environment, setting up a dynamic oral English teaching training vocabulary, increasing the system vocabulary, improving the efficiency of oral English training, and formulating the training rules for oral English remote teaching. Promote the system to have the function of distance teaching and training of spoken English. At this point, the above content is converted into

system coding, which completes the design of the training system for spoken English distance education.

4 System Testing

To test the spoken English distance teaching training system designed this time, a comparative experiment will be used to verify the spoken English distance teaching training system designed this time. In this system test experiment, the designed spoken English distance teaching training system was recorded as system A; the two traditional spoken English distance teaching and training systems were recorded as system B and system C respectively.

4.1 Experimental Preparation

Test the design of spoken English distance teaching training system, set the system test environment, as shown in Table 1.

Table 1. System test environment

Test environment	Configuration	Parameter
Hardware environment	CPU	Pentium Quad Processor
	Memory Capacity	2G
	Hard drive capacity	250G
	Graphics card	Intel Integrated Graphics
Software environment	Operating system	Window XP SP2
	Programming language	Java

Based on the system test environment shown in Table 1, the test tools used to test whether the system can run normally are as follows: 1. Use LoadRunner performance test tool to test three groups of system performance and judge the system server performance; 2. Use winrunner auxiliary function test tool to check whether the system business function is correct; 3. Use TestDirector to monitor the system test process System testing process, tracking detection.

Use the three system detection tools shown above to detect the normal operation of the system. The detection process is as follows: 1. Login detection. Log in to the system and check whether the system's normal login status is consistent with the system's normal status; 2. Page detection. Test the system function interface, check whether the linked page is in a normal associated state; 3. Function detection. Check the function buttons on the page to confirm the execution result of each function button; 4. Interface detection. Check whether the system interface design is consistent with the user's usage habits and operating specifications; 5. Input processing monitoring. Check the input

search function of the system, the system makes logical processing and judgment accuracy according to the input keywords; 6. Abnormal state detection. The system detects and prompts abnormal users; 7. Business process detection. System business process execution sequence and requirements description.

Based on the above design of the system detection process, the system test experiment is detected and three groups of test systems are selected. The test results show that: in this test experiment, the selected three groups of test systems, such as login status, page display, button link, interface display, input query, abnormal status prompt, business process and other basic performance of the system, are in normal state, and can be used for system test comparative experiment.

4.2 Experimental Result

4.2.1 System Performance Test

Based on the experimental design of the system test running environment, as well as the system test experimental object, the first group of system performance test experiments are carried out. Because the oral English distance teaching training system needs to record and store a large number of English words, there is a large number of visits and so on, which needs to bear more pressure and load. Therefore, in the running performance test experiment of the system, the LoadRunner performance test tool is used to generate virtual users to carry out oral English distance teaching training and test the system performance. In this group of experiments, a total of 600 virtual users were used to perform real-time operation on the three groups of systems. The initial virtual users were set to 200, and 100 virtual users were added every 20 min. The average response time, application server utilization rate and the number of system errors of the three groups of systems were compared. And let the LoadRunner performance test tool generate the system running performance test report. The comparison results are shown in Table 2.

Table 2. System operation performance test report

Method	Number of concurrent users	Average response time	Application server usage	Number of system errors
A system	200	0.003 s	1%	0
	300	0.02 s	3%	0
	400	0.12 s	7%	0
	500	0.30 s	12%	0
	600	0.42 s	20%	0
B system	200	0.011 s	2%	0
	300	0.05 s	7%	0
	400	0.26 s	10%	1

(*continued*)

Table 2. (*continued*)

Method	Number of concurrent users	Average response time	Application server usage	Number of system errors
	500	0.34 s	15%	1
	600	0.58 s	26%	2
C system	200	0.013 s	2%	0
	300	0.07 s	5%	0
	400	0.20 s	9%	0
	500	0.41 s	18%	1
	600	0.55 s	24%	1

As can be seen from Table 2, when the virtual users are between 200 and 600, the operation response time of the three groups of systems is also maintained between 0.003 and 1 s. However, the utilization rate of application server in system B reaches 26%, and when the virtual number of users reaches 400, there will be a system operation error, so the operation performance of system B is the worst; although the operation performance of system C is better than that of system B, the utilization rate of application server also reaches 24%. When the number of virtual users reaches 500, system operation error occurs; The average response time performance of system a is the worst. Although the response time continues to extend with the increase of the number of users, it is obviously the lowest among the three groups of systems. The utilization rate of application server has been maintained below 20%, and with the increase of the number of virtual users, there is no system error. It can be seen that the design of spoken English distance teaching training system, with a faster running speed, lower server utilization, can support more users and use the system at the same time.

4.2.2 Hardware Performance Comparison of System Server

On the basis of the first set of experiments, the third set of experiments was carried out. In the first set of experiments, the number of virtual people shown in Table 2, the average response time, application server utilization rate, and the number of system errors caused by the virtual number of people in the first group of experiments were counted. The Loadrunner performance test tool is used to test the system hardware performance, and TestDirector is used to track the test process of the test system. The comparison result of the server hardware performance is shown in Fig. 5.

As can be seen from Fig. 5, system B has the lowest utilization rate of server hardware resources. When the number of virtual users reaches 330, the server resource utilization rate reaches the highest, which affects the running speed of server hardware. When the number of virtual users reaches 250, the memory presents a horizontal line. Therefore, the server hardware of system B can only bear 250 people. Once the number of virtual users exceeds this number, a curve will appear The server hardware resource utilization ratio of system C is stronger than that of system B. However, when the number of virtual

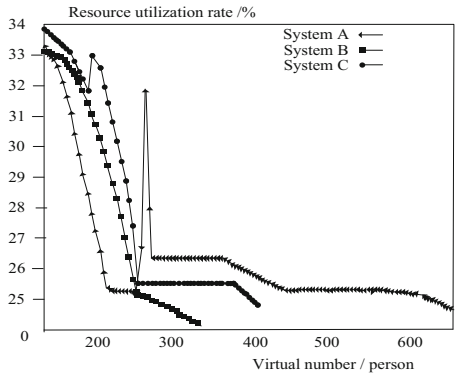

Fig. 5. Server hardware performance comparison chart

users exceeds 390, the memory line decreases, and the server resource utilization rate increases, which affects the operation speed of server hardware. Only when the number of virtual users is less than 390, the memory presents a horizontal line, which is normal In the running state, the maximum number of responders supported by system B is 390; while for system a, when the number of people exceeds 600, the memory line tends to decline, and the occupancy rate of server hardware shows an increasing state. Therefore, the maximum number of responders supported by system a is 600. From this, we can see that the number of people who visit the English system at the same time is the lowest.

5 Concluding Remarks

"Virtual reality" is a kind of computer system that can create and experience the virtual world. The various virtual environments generated by this system act on the user's vision, hearing and touch, and make the user feel immersive. The so-called virtual world is a collection of virtual environments or given simulation objects. Virtual reality is a new technology developed in recent years, At present, it has been widely used in many fields. Therefore, the design of oral English distance teaching training system, make full use of virtual reality technology, improve the system training students' oral English effect. However, the design of oral English distance teaching training system, has not been in-depth design of virtual reality scene changes, as well as its interactive function. Therefore, in the future design, we need to further design the virtual reality scene changes, as well as its interactive function, to further improve the popularity of the system, to help users overcome various difficulties in oral English learning.

References

1. Li, L., Liu, B.: Design of oral English training system based on human-computer interaction. Mod. Electron. Tech. **43**(14), 135–137 (2020)
2. Lin, H.: Study on AI virtual English speaking training system for SELL corpus. Microcomput. Appl. **36**(7), 126–129 (2020)

3. Cai, X., Zhang, Q.: The integration mechanisms of feedforward and feedback control in speech motor system. Adv. Psychol. Sci. **28**(4), 588–603 (2020)
4. Wang, X., Ma, F.: Improvement of scoring system for children speaking English based on DNN. Inf. Technol. **44**(9), 46–50 (2020)
5. Fu, C., Xu, D.: A bus riding training system for children with autism based on virtual reality. Comput. Simul. **36**(6), 209–213, 231 (2019)
6. Wang, X.: Design of English aided instruction system based on personalized recommendation. Microcomput. Appl. **35**(5), 35–38 (2019)
7. Sang, H., Sun, X.: Design of English teaching resources information integrated management system of based on artificial intelligence. Mod. Electron. Tech. **43**(10), 173–175 (2020)
8. Liu, S., Bai, W., Zeng, N., et al.: A fast fractal based compression for MRI images. IEEE Access **7**, 62412–62420 (2019)
9. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mob. Netw. Appl. **24**(1), 1–4 (2019)
10. Liu, S., Glowatz, M., Zappatore, M., Gao, H., Jia, B., Bucciero, A. (eds.): eLEOT 2018. LNICSSITE, vol. 243. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93719-9