Design of a Virtual Reality-Based Learning System for Spoken English

https://doi.org/10.3991/ijet.v17i24.35655

Yang Wang[1(✉)], Ala Mughaid[2]
[1]Public Teaching Department, Changchun Automobile Industry Institute, Changchun, China
[2]Department of Information Technology, Faculty of Prince Al-Hussien bin Abdullah for IT, The Hashemite University, Zarqa, Jordan
yang_wang130@yeah.net

Abstract—Virtual reality is an important technology that is fast gaining global attention in different spheres of life particularly in the education sector. In view of this, this study designs a distance learning system for spoken English based on virtual reality, firstly, the overall design of the teaching system and the hardware and software of the system are designed, then a double-supervised signal convolutional neural network algorithm is proposed for the speech data recognition function of the system, and finally the testing of the system performance and the simulation analysis of the algorithm are carried out. The results show that the step response curve of the system designed in this study is gradually stabilized after 11s of operation, although there are certain fluctuations in the initial stage; the speaking scoring function of the system is more influenced by the sampling period T. When T is at 3 and 4, the speaking scoring speed of the teaching system is 33s~42s, which is significantly better than other intervals. The number of information submission and feedback was approximately the same and the interaction activity was very high after students used the system designed in this study, reflecting that student were more motivated to learn spoken English after using the system. The final loss rate using Goog Le Net is smaller and more convergent compared to the loss rate of the other three CNN models trained. The convolutional neural network algorithm constructed in this study has a very high accuracy rate in the recognition of English speech data, which is significantly better than other recognition models. To a certain extent, this study can provide guidance for the construction of English-speaking distance learning system, and more needs of users can be considered in future research.

Keywords—virtual reality, spoken English, remote learning system design

1 Introduction

In recent years, with the continuous development of Internet technology, virtual reality technology has also been developed significantly, providing a better teaching mode for English speaking distance learning [1]. Virtual reality technology has become an important research direction in the field of education, and the technology can not only realize resource sharing but also remote control, which has a better development

prospect in the field of education [2]. The research of speech recognition technology has high commercial value. How to apply this technology to business and daily life more naturally, reliably and at low cost is its development direction. In the future, people can realize some things more conveniently and quickly through voice interaction and enjoy more modern services [3]. With the development of globalization, the relationship between countries in the world is becoming closer and closer, and economic and trade integration is gradually realized. China's reform, opening up and development have achieved great success. It has demonstrated its strength in the world. Its cultural and economic communication and exchanges with other countries are increasing day by day. As the most widely used language in the world, English is of great importance to learn and master. Learning English well is not only a means to understand and learn foreign culture and knowledge, but also an indispensable tool for future work. More and more foreign enterprises are established in China. Fluent English can reduce the estrangement between you and your foreign partners. However, the study of spoken English has always troubled Chinese people. In daily English vocabulary, reading, listening and oral learning, oral learning is the biggest problem. In order to effectively improve the efficiency of teaching spoken English, an important research on distance learning system of spoken English is needed. Virtual reality technology is able to simulate the real environment in 3D dynamics, allowing users to break through the time and space limitations and be fully immersed in it [4]. When virtual reality technology is applied to English speaking teaching, it can provide users with a virtual learning environment of "independent learning and human-computer interaction", which can not only effectively increase students' interest in learning English, but also improve their learning effect [5]. Therefore, this study aims to build a virtual reality-based distance learning system for spoken English, firstly, the overall design of the teaching system, followed by the design of the hardware and software of the system, and then the double-supervised signal convolutional neural network algorithm is proposed to optimize the speech data recognition problem of spoken English in the system. The virtual reality-based English-speaking distance learning system breaks the traditional English-speaking teaching concept and allows students to learn modular English speaking in a virtual environment, and students enhance their English-speaking application skills by conversing with virtual tasks to achieve a diversified teaching approach [6].

2 Related work

At present, the development of teaching systems based on virtual reality has become a trendsetter in the development of the education field, and the teaching of spoken English has also received attention from many scholars, especially the integration of virtual reality technology into the distance learning of spoken English has become the focus of English education work nowadays. Vieiramonteiro et al. explored the application of virtual reality in the teaching of foreign language vocabulary, and the participants used English placement tests, vocabulary tests, and exposure to virtual environments, and the results showed that the technology stimulated students' interest in learning and immersed them in authentic situations [7]. Pack et al. investigated English language learners' perceptions of using a prototype virtual reality learning environment (VRLE),

and the results showed that learners' perceptions of Goehle G presented a course on common calculus topics based on "virtual reality", including The description of how the lecture was implemented in a virtual reality and augmented reality hardware system, and a quiz and survey to describe students' reactions to the lesson, summarize some general impressions of this emerging technology [9]. retention with similar effects and VR does not provide any additional retention performance [10]. Alemi et al. analyzed the effect of virtual reality-assisted pronunciation training on English learners' pronunciation and the results showed significant differences in learners' performance before and after the training [11]. Lai et al. investigated the effects of virtual reality and personal computer games on language learners' vocabulary learning and emotion perception, and the results showed that both the virtual reality and computer groups were able to gain vocabulary knowledge in translation and recognition tests, and the virtual reality group scored significantly higher on average than the computer group in the post-vocabulary translation delay test [12]. Chen used transfer learning as technical support to study English speech emotion recognition, by comparing the performance of English speech emotion recognition models based on CNN neural networks, the results showed that migration learning has some advantages over other algorithms in English speech emotion recognition [13]. Mustaqeem et al. proposed a new technique for speech emotion recognition, namely, one-dimensional extended convolutional Neural Network (1D-DCNN), and the results showed that the proposed model was evaluated on three benchmark datasets including IEMOCAP, EMO-DB and RAVDESS, and its accuracy reached 72.75%, 91.14% and 78.01%, respectively [14]. qian et al. explored knowledge extraction from a normal full-precision floating-point model to a compressed binary model and simulated it on standard switch speech recognition, and the results showed that the proposed binary neural network can provide a 3–4 times speedup over the normal full-precision depth model [15]. Zheng et al. proposed a binarized convolutional neural network (BCNN) based based speech recognition processor that integrates an on-chip self-learning mechanism to compensate for the accuracy loss due to low precision, and the results show that the processor consumes 2.5 times less energy per neuron and 8.0 times less energy per speech frame compared to state-of-the-art speech recognition implementations [16].

In the field of education, there is still less research on the teaching system based on virtual reality, and there is almost no research on the speech recognition of the teaching system. Therefore, this study introduces virtual reality technology into the English-speaking distance learning system and optimizes the English-speaking recognition by improving the CNN algorithm, aiming to improve the performance of the teaching system.

3 Research on distance learning system of spoken English based on virtual reality

3.1 Overall design of teaching system and software and hardware design

Virtual reality technology can simulate the real environment into a three-dimensional dynamic model, so that users can immerse themselves in the environment without

being limited by space or time. When applied to the education industry, by creating a teaching environment of "independent learning and human-computer interaction", students' interest in learning can be enhanced and their ability to master knowledge can be enhanced. Therefore, an English teaching system based on virtual reality technology is designed. The system breaks the traditional teaching concept, converts students' passive learning into active learning, and converts modular English knowledge into dynamic virtual scenes; Students can improve their English application ability through dialogue with virtual characters; Human computer interaction allows students to ask questions in the teaching system in real time, and the system feeds back questions to students or teachers, realizing diversified teaching methods. The virtual reality-based English-speaking distance learning system consists of five main modules, which are the main interface, login, power circuit, command receiving and forwarding, and central operation module [17]. The detailed parameters of the system are mainly: memory over 200 GB, system temperature above 0° and below 50°, hard disk capacity over 100 GB, and CPU of 7.15 GHz. The overall design of the virtual reality-based English-speaking distance learning system is shown in Figure 1. In this study, the hardware of the five modules of the virtual reality-based English-speaking distance learning system is designed. The main interface module is designed to provide a guidance function for users, including content, help, navigation and various title bars, and the content of this module is selected by users themselves, and when users enter the main interface, the system function tips and consultation will appear.

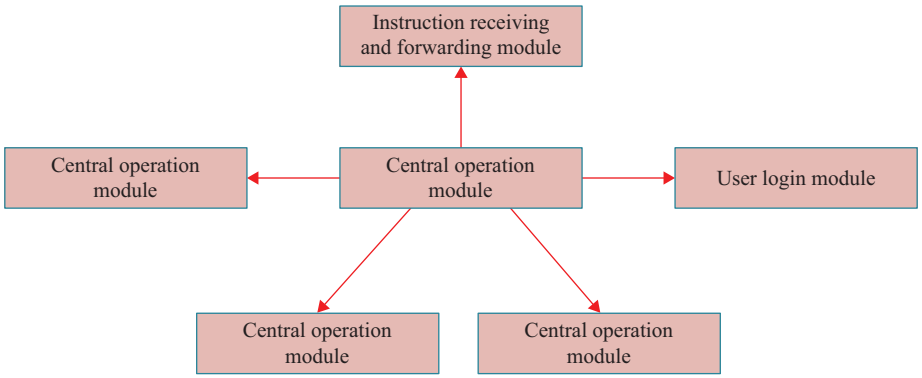
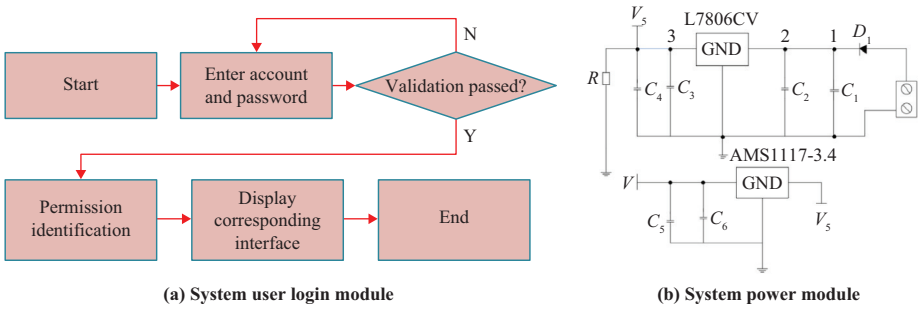


Fig. 1. The whole framework of virtual reality oral English distance teaching system

For the login module design, it is mainly to verify the identity of the logged-in user, the user needs to enter his identity information in the login interface, and the system will judge the user's authority in the system based on the user's identity information, so as to provide its corresponding function page, and the user can modify the content displayed on the page according to his own needs, see Figure 2(a). The design of the power supply circuit module belongs to the most important content of the hardware design of the virtual reality-based English-speaking distance learning system, mainly because the design of the power supply circuit is related to the incidence of system failure and the incidence of other instability factors. The power supply circuit of this research system

is filtered at C_1, C_2 by L7806CV, where the system core power supply voltage is 2V, and the system chip is powered by AMS1117-3.4 when the system voltage reaches 5V. A large number of capacitors are connected to the system I/O ports to filter the external interference, and finally the capacitors at the power supply output are operated to flatten the wave, see Figure 2(b) for the design.



(a) System user login module **(b) System power module**

Fig. 2. Design of system user login module and system power supply module

The instruction receiving and forwarding module is mainly to send and receive instructions for the whole remote system monitoring end. After the user performs the operation at the monitoring end, the system will generate the corresponding instruction, and the central control module will generate the corresponding file for the generated instruction, and then perform the processing operation, and finally extract the content in the file for identification, and execute the corresponding function of the system through the corresponding instruction. Finally is the design of the central operation module, the main function of this module is to analyze and process all kinds of instructions received by the system, belongs to the main line of the program of the controlled end of the system, mainly by listening to the port to achieve the receipt of instructions, and timely processing after the successful receipt of instructions, and continue to receive other instructions after the end of the operation, the central operation module will continue to carry out this operation until the system is shut down.

The software design of the English-speaking distance learning system based on virtual reality mainly includes two aspects, which are virtual reality education environment construction and human-computer interaction implementation. The virtual reality education environment is built mainly by applying 3DMAX software, firstly uploading the scanned data information in the computer, and then using the error function to adjust the error of each model by simulating the physical model of the surface technology components. The error function for the adjustment of the model in this study is shown in equation (1).

$$\delta = \Delta p / s = (\lambda \omega_i - s) / s \tag{1}$$

In Eq. (1), δ is the relative error value of the error function, λ is the coefficient, Δp is the absolute error, and ω_i is the independent virtual model consisting of i sub-models obtained after plotting. According to the teaching requirements of the English-speaking distance learning system, the parameters of the constructed models are fine-tuned, and

virtual models of various resolutions are obtained to meet the virtual display requirements of the system. If the design model has x_i and y_i pixels in the horizontal and vertical directions respectively, the total pixels of the design model can be expressed as $z = x_i y_i$. Therefore, the pixel resolution of the module is adjusted as shown in equation (2).

$$\begin{cases} f_1 = l_1 \cos\theta tn \cdot z \\ f_2 = l_2 \cos\theta tn \cdot z \end{cases} \tag{2}$$

In Eq. (2), f_1 and f_2 denote the model pixels in the forward and reverse directions; l_1 and l_2 denote the model lengths in the forward and reverse directions; n denotes the number of surfaces constituted by the model in each direction; t denotes the illumination duration of each stage; and θ is the illumination angle. According to equation (2), a virtual reality-based English-speaking distance learning scenario is constructed in this study, as shown in Figure 3. In the virtual reality English speaking teaching environment, students can simulate many English-speaking conversation scenarios and have conversations with virtual characters for the purpose of practicing English speaking.

Fig. 3. Virtual teaching environment

The design of the human-computer interaction function of the virtual reality-based English-speaking distance learning system is the core work of the system. Firstly, the video player is fixed, the rotation angle of the lens is set as needed, the simulation device is set through VRML browser, and the 3D coordinates of the device are set for the purpose of virtual space conversion. According to the above settings, set the demonstration of English-speaking teaching scene in the system to get the virtual teaching scene set in advance, and import the scene in VRML to realize the interaction function of the system. The function of behavioral interaction module is added to the virtual character and the flowchart approach is used to decide the sequence of interaction operation. Then the framework of the English-speaking distance learning system constructed in this study is implemented to run the teaching system, and the response function is used to send and receive instructions, interact with data and change the page. The system response function $g(w)$ is calculated as shown in equation (3).

$$g(w) = mq - k_i \tag{3}$$

In equation (3), k_i is the system teaching page response coefficient, q is the speaking teaching level, and m is the coefficient of system response. The system response function can trigger the operation level of students when they are learning spoken English, so that they can quickly enter into the virtual reality teaching scenario.

3.2 Double-supervised signal convolutional neural network algorithm

In a virtual reality-based distance learning system for spoken English, the recognition of spoken English speech data is particularly important. Therefore, this study introduces a double-supervised signal convolutional neural network algorithm to improve the spoken English speech recognition function of the system. Currently, the common structures of convolutional neural networks include AlexNet structure, GoogLeNet structure and Vgg-16Net structure, which has five convolutional operations, each of which contains 2–3 convolutional layers, and uses a maximum pooling layer at the end of each convolutional operation to reduce the size of the feature map. The first segment of the structure contains 64 convolutions, the second segment contains 128 convolutions, the third segment contains 256 convolutions, the fourth segment 512 convolutions, and the fifth segment 512 convolutions. Since the 2nd–5th segments appear multiple 3×3 are convolutional layers, which are equivalent to a 5×5 convolutional operation, the size of the sensory field is 5×5. Three 3×3 convolutional layers are equivalent to a 7×7 convolution, but the number of parameters is only half of the 7×7 convolution. The AlexNet structure contains 5 convolutional and pooling layers, 3 fully-connected layers, and softmax loss output in the last layer, and a nonlinear correction function activates the data after the output of each convolutional and fully-connected layer. The GoogLeNet structure is more complex, and the core idea is to approximate the optimal local sparse structure in the convolutional network, based on Hebbian's law and multi-scale processing to optimize performance, and to increase the width and depth of the network by process design under the condition that the network computational resources remain unchanged.

In this study, two supervised signals, softmax loss and center loss, are trained on the built GoogLeNet structure, where the role of center loss is to cooperate with softmax loss so that samples of the same species can be as close to the sample center as possible by penalizing the sample center of each species and the sample offset. center loss supervises the function of the signal as shown in equation (4).

$$L_S = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{4}$$

In equation (4), denotes the $W_i j$ th column of the parameter matrix of the fully connected layer, and x_i denotes the i th sample feature belonging to the y_i th category. softmax loss serves to map the output of multiple neurons into the (0,1) interval, minimizing the cross-entropy and probability of the true distribution of the classification. softmax loss supervises the signal as a function of the signal, as shown in equation (5).

$$L_C = \frac{1}{2}\sum_{i=1}^{m} \| x_i - c_{y_i} \|_2^2 \tag{5}$$

In equation (5), c_{yi} denotes the feature center of the sample of the y_i species. The overall framework of the algorithm for the softmax loss and center loss dual-supervised signal convolutional neural network thus constructed is shown in Figure 4.

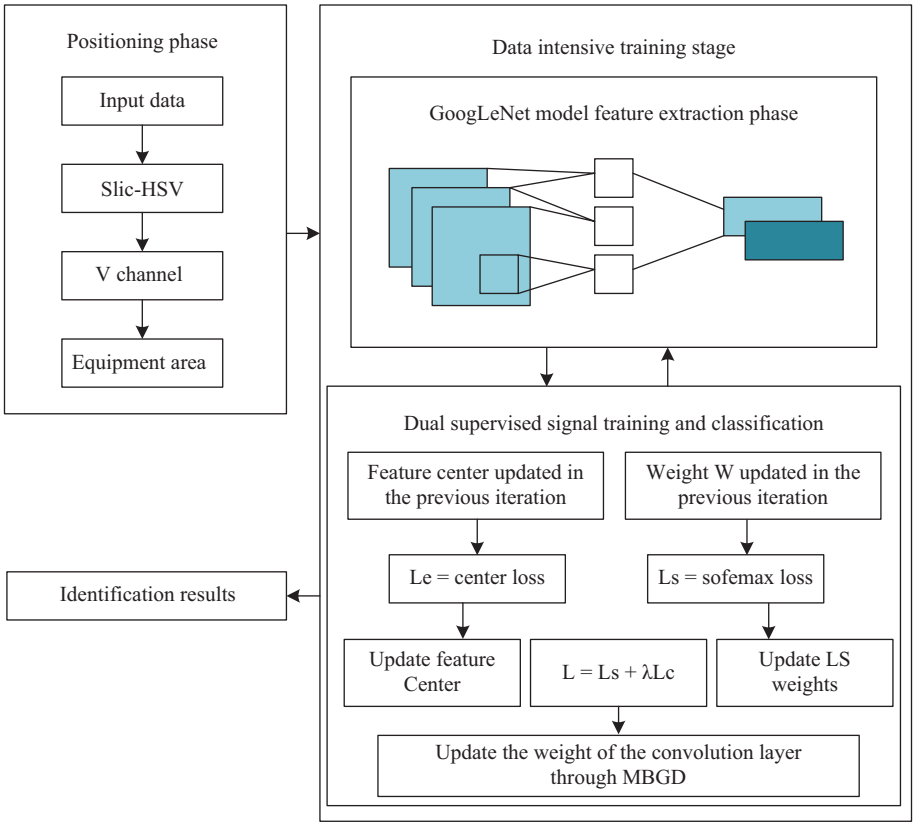


Fig. 4. Overall framework of algorithm

By MBGD (small batch gradient descent), the weights of the convolutional layer are updated by combining the feature centers L_C and their weights W, and the weights of the convolutional layer θ_c are obtained by the formula $L = L_S + \lambda L_C$. After each iteration, the algorithm updates the feature centers again. The distance between the current feature and the feature center is calculated at each iteration, and the distance is superimposed on the feature center in the form of a gradient, as shown in Eqs. (6) to (8).

$$c_j^{i+1} = c_j^t - \alpha \Delta c_j^t \tag{6}$$

$$\frac{\partial L_C}{\partial x_i} = x_i - c_{yi} \tag{7}$$

$$\Delta c_j = \frac{\sum_{i=1}^{m} \delta(y_i = j)(c_j - x_i)}{1 + \sum_{i=1}^{m} \delta(y_i = j)} \tag{8}$$

The α in Eq. (6) can control the learning rate of feature center. In Eq. (8), $\delta(y_i = j)$ is the indicator function. When the current sample category is true, $y_i = j$, i.e., the indicator function takes the value of 1. Conversely, when the current sample category is not true, $y_i \neq j$, and the indicator function takes the value of 0.

In order to balance the ratio between the two losses, we introduced the parameter λ to obtain the final loss function.

$$L = L_S + \lambda L_C = -\sum_{i=1}^{m} lo\,\mathrm{g}\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} + \lambda \frac{1}{2}\sum_{i=1}^{m} \| x_i - c_{y_i} \|_2^2 \tag{9}$$

When softmax loss and center loss are jointly supervised training, the weights of the initialized convolutional layer θ_c, parameters λ, loss layer weights W and learning rate μ are first obtained, and the weights are calculated according to the formula θ_c. If the network does not converge, the parameters and center vector positions are updated. The updated weights are $\theta_c^{t+1} = \theta_c^t - \mu^t \frac{\partial L^t}{\partial x_i} \cdot \frac{\partial x_i^t}{\partial \theta_C^t}$.

4 System performance testing and algorithm simulation analysis

4.1 System performance testing

In order to verify the effectiveness of the virtual reality-based English-speaking distance learning system designed in this study, the performance of the system was tested. Under the premise of satisfying the users' needs, the most common system performance of the users was first evaluated, which were the scoring accuracy of spoken English, the scoring efficiency of spoken English, and the step response capability of the system, and the system designed in this study was compared with the systems designed in the literature [18] and literature [19], respectively. Firstly, the step response capability of different systems was compared, as shown in Figure 5. If the step response coefficient of the system is higher, then the working elapsed time is less.

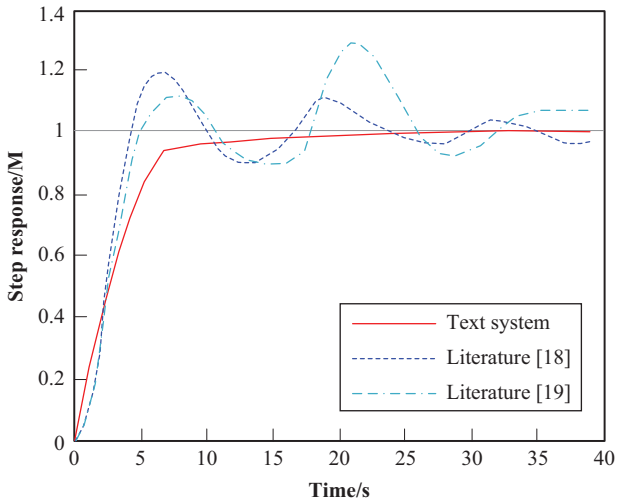


Fig. 5. Comparative analysis of step response of three different systems

As can be seen from Figure 5, the step response coefficients of all three systems quickly reached from 0M to near 1M in the early stage of operation. Among them, the step response curve of the designed system in this study fluctuates to some extent in the initial stage but gradually stabilizes after 11 s of operation, but the step response curves of the designed systems in the literature [18] and [19] fluctuate very much. This result indicates that the step response of the present study system is significantly better than the designed systems in the literature [18] and literature [19]. Analyzing the reason, it may be that the processing of instructions in this study mainly uses the central operation module, which belongs to the main line of the program at the controlled end, and can change the output of the system from 0M to 1M within a very small elapsed time and maintain it smoothly, further reflecting the superiority of the designed system in this study.

The efficiency of English-speaking scoring in a virtual reality-based English-speaking distance learning system was tested with the main purpose of testing the scoring speed corresponding to the sampling period (T) of English-speaking pronunciation data, and the results are shown in Table 1.

Table 1. Scoring speed of collection cycle T in different collection intervals

Experiment No	Scoring Speed/s	Acquisition Cycle (T)
1	26	2
2	33	3
3	42	4
4	48	5
5	53	6

As can be seen from Table 1, the speaking scoring function of the virtual reality-based English-speaking distance learning system is more influenced by T. When t is between 3 and 4, the oral scoring speed of the teaching system is between 33S and 42s,

which is significantly better than other intervals. This result reflects that the system designed in this study has excellent timeliness and confirms the feasibility of the oral English distance teaching system based on virtual reality.

The accuracy of English-speaking evaluation of the different systems was then compared, and the results are shown in Figure 6. The results show that the scoring accuracy of the virtual reality-based English-speaking distance learning system is significantly higher than the other two systems. The main reason is that the system in the literature [18] works in a more complicated way and has more algorithms, so that the amount of operations is too large, which greatly affects the evaluation accuracy of spoken English; while the way of evaluating spoken English in the literature [19] only uses the central control station and other interfaces to realize the transmission of instructions, and then does not carry out the analysis and processing operation of the instructions, which is easy to make errors during the transmission of instructions. Therefore, the accuracy rate of spoken English evaluation in this system is low.

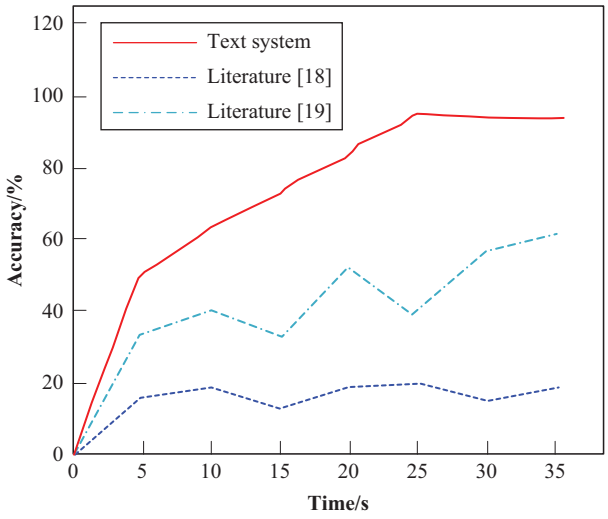


Fig. 6. Comparison of the accuracy of oral English evaluation in different systems

The traditional English-speaking teaching system in the literature [20] was selected to compare with the virtual reality-based English-speaking distance learning system for interactivity evaluation. One hundred college students in a university were randomly selected to conduct interactivity evaluation tests on students using the two systems, and the results are shown in Figure 7.

As can be seen from Figure 7(a), students have approximately the same number of information submissions and feedback after using the virtual reality-based English-speaking distance learning system, and the interaction is very active, reflecting that students are more motivated to learn English speaking after using the system. However, the results in Figure 7(b) show that the active and passive interaction curves fluctuated less after students used the traditional English-speaking teaching system, reflecting the low interaction performance of the system and the low motivation of students.

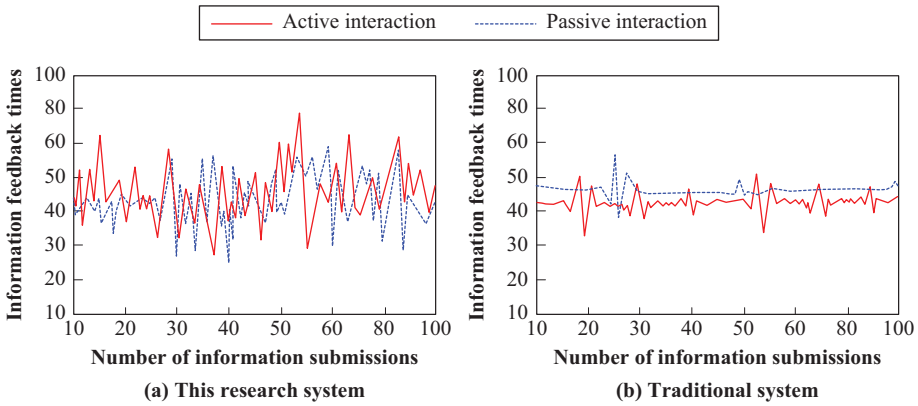


Fig. 7. Interactive comparison results of the two systems

The above comparative analysis shows that the virtual reality-based English-speaking distance learning system has better step response coefficient, speaking evaluation accuracy, efficiency and interactive performance, which can effectively enhance students' motivation to learn spoken English.

4.2 Algorithm performance simulation analysis

Compared to the way of recognizing speech based on manual features, the principle of CNN is to extract features automatically by forward transmission of feature extraction techniques, while reverse transmission updates the convolutional kernel weights according to the nature of the loss function. In order to verify the role of CNN algorithm in English spoken speech recognition, a set of speech data is randomly selected from the test set in the experimental dataset, and the CNN algorithm recognition and judgment process, as shown in Figure 8.

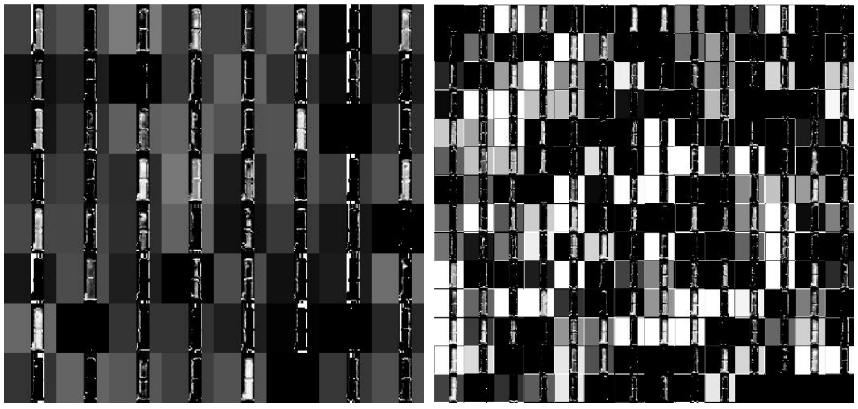


Fig. 8. Convolution processing of layer 1 and layer 2

When the English speech data to be recognized is input in the system, the CNN algorithm first extracts the features in the speech data, and after further convolutional processing in the first layer of the algorithmic network, Figure 8(a) is obtained, and the feature map of 64 English speech data is shown in Figure 8(a); after the second layer of convolutional processing, Figure 8(b) can be obtained, and the feature map of 196 English speech data. These two convolutions mainly extract the feature information such as edge information of English speech data, and the features obtained become more and more comprehensive as the number of convolutions increases.

In this study, four deep learning models, VggNet, improve_cnn, GoogLeNet, and Alex Net, were used to train and test the speech dataset. The number of iterations in the training set was set to 20, the learning rate u was set to 0.001, and the number of iterations was 1500. The loss curves derived from the training of the IMPROVE_CANN, SOFTMAX and the four different CNN models were tested, as shown in Figure 9.

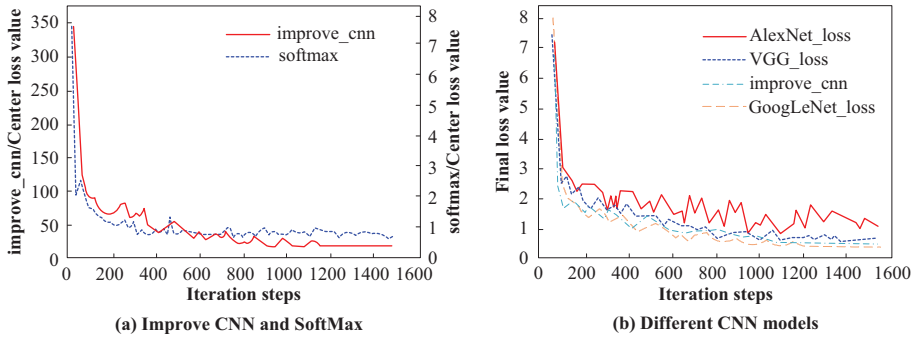


Fig. 9. Center loss curve under different modes

The vertical coordinates in Figure 9 are the loss values of the training samples; the smaller loss values represent the higher degree of network convergence. From Figure 9, it can be seen that as the number of iterations increases, the smaller the loss value of the model training, the higher the degree of network convergence; the total loss rate of IMPROVE_CANN is slightly higher than that of the Google Net model using soft max loss alone; the final loss rate using Google Net is smaller compared to the other three CNN models training, and the degree of convergence is higher.

The accuracy profiles of different CNN models trained are shown in Figure 10. An accuracy test is performed after every 500 iterations. As can be seen from Figure 10, the accuracy of the network at the double-supervised signal training rises about 2% compared to the original model. This may be because after adding the center loss, the distance between the current sample feature and the feature center is additionally calculated at each iteration of the network training, and the distance is used as the basis for updating the feature center for the next step of the calculation process. The Slic-HSV algorithm applies the idea of clustering to the iterative process, thus reducing the possibility of misclassification of features and greatly improving the accuracy of judgment.

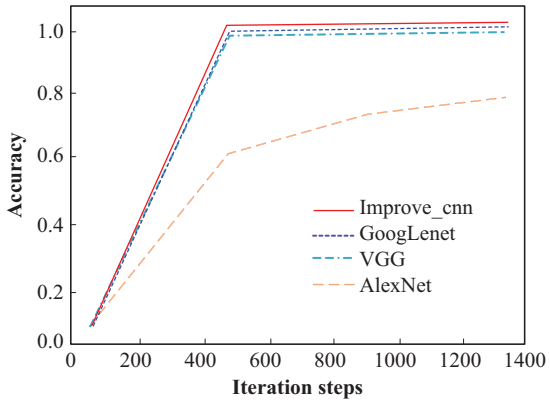


Fig. 10. Accuracy of different CNN models

Four different CNN models were used for recognition of English speech data, firstly to identify the target to be recognized, after that to extract the features of the target data, and finally to recognize the data type based on the data features. Comparing Vgg Net, Google Net, Alex Net and the algorithm used in this study to recognize eight types of speech data, the results are shown in Figure 11.

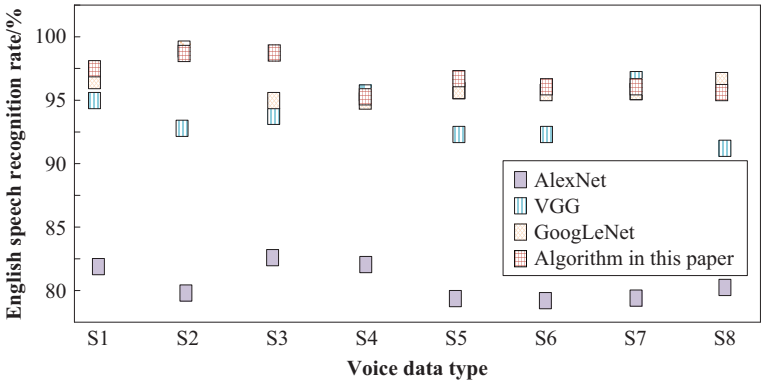


Fig. 11. Recognition of English speech data using different CNN models

As can be seen from Figure 11, the convolutional neural network algorithm constructed in this study has a very high accuracy rate in the recognition of English speech data, which is significantly better than other recognition models. In the recognition of individual data, the accuracy rate of the CNN algorithm is slightly lower than that of the Google Net model, but the accuracy rate of the Google Net model is slightly lower than that of the algorithm proposed in this study in the accuracy of most types of recognition Therefore, the performance of the proposed algorithm is better.

5 Conclusion

Aiming at the design problem of oral English distance teaching system based on virtual reality, this paper applies double supervised signal convolution neural network algorithm to oral speech recognition technology, and verifies the performance of the system and algorithm. The results show that after students use the system designed in this study, the number of information submission and feedback is roughly the same, and the interaction activity is very high, which reflects that students' enthusiasm for oral English learning is greater after using the system; The convolution neural network algorithm constructed in this study has a very high accuracy in the recognition of English speech data, which is obviously superior to other recognition models. With the increase of iteration times, the smaller the loss of model training, the higher the degree of network convergence; improve_ The total loss rate of CNN is slightly higher than that of Google Net model using soft max alone; Compared with the other three CNN models, the final loss rate using Google Net is smaller and the convergence degree is higher; The accuracy of the network at the double supervised signal training station is about 2% higher than that of the original model. The step response coefficient, accuracy, efficiency and interactivity of oral English remote teaching system based on virtual reality can be better, which can effectively improve students' enthusiasm in learning oral English. Due to the limited time and ability, this study will add more student samples in the future work to analyze the impact of the teaching system on students' psychological indicators.

6 References

[1] W. Zuo, "Design and implementation of digital art teaching system based on interactive virtual reality," *IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association*, vol. 30, pp. 587–594, 2018.

[2] N. Zhang, L. Tan, F. Li, et al., "Development and application of digital assistive teaching system for anatomy", *Virtual Reality & Intelligent Hardware*, vol. 3, pp. 315–335, 2021. https://doi.org/10.1016/j.vrih.2021.08.005

[3] Y. Zhang, Y. Min, X. Chen, "Teaching chinese sign language with a smartphone", *Virtual Reality & Intelligent Hardware*, vol. 3, pp. 248–260, 2021. https://doi.org/10.1016/j.vrih.2021.05.004

[4] A. Hao, J. Cui, S. Li, et al., "Personalized cardiovascular intervention simulation system", *Virtual Reality & Intelligent Hardware*, vol. 2, no. 2, pp. 104–118, 2020. https://doi.org/10.1016/j.vrih.2020.04.001

[5] P. Y. Chiang, H. Y. Chang, Y. J. Chang, "Pottery go: A virtual pottery making training system", *IEEE Computer Graphics and Applications*, vol. 38, pp. 74–88, 2018. https://doi.org/10.1109/MCG.2018.021951634

[6] X. D. Rao, H. B. Yu, J. H. Wu, et al., "Practical experience of virtual acupuncture and moxibustion teaching system in the operation teaching course of Acupuncture Sciences", *Chinese Acupuncture & Moxibustion*, vol. 40, pp. 877–879, 2020.

[7] Ana Maria Vieira Monteiro, Patrícia Nora de Souza Ribeiro, "Ribeiro virtual reality in English vocabulary teaching: An exploratory study on affect in the use of technology", *Trabalhos em Linguística Aplicada*, vol. 59, pp. 1310–1338, 2020. https://doi.org/10.1590/010318137569316202007 16

[8] A. Pack, A. Barrett, H. N. Liang, et al., "University EAP students' perceptions of using a prototype virtual reality learning environment to learn writing structure", *International Journal of Computer-Assisted Language Learning and Teaching*, vol. 10, pp. 27–46, 2020. https://doi.org/10.4018/IJCALLT.2020010103

[9] G. Goehle, "Teaching with virtual reality: Crafting a lesson and student response", *The International Journal for Technology in Mathematics Education*, vol. 25, pp. 35–45, 2018.

[10] Y. Gürkan, S. Yldrm, E. Dolgunsz, "The effect of VR and traditional videos on learner retention and decision making", *World Journal on Educational Technology Current Issues*, vol. 11, pp. 21–29, 2019. https://doi.org/10.18844/wjet.v11i1.3983

[11] M. Alemi, S. Khatoony, "Virtual reality assisted pronunciation training (Vrapt) for young Efl learners", *JET (Journal of English Teaching)*, vol. 20, pp. 59–81, 2020.

[12] K. Lai, H. Chen, "A comparative study on the effects of a VR and PC visual novel game on vocabulary learning", *Computer Assisted Language Learning,* vol. 33, pp. 1–34, 2021. https://doi.org/10.1080/09588221.2021.1928226

[13] X. Chen, "Simulation of English speech emotion recognition based on transfer learning and CNN neural network", *Journal of Intelligent and Fuzzy Systems*, vol. 40, pp. 2349–2360, 2021. https://doi.org/10.3233/JIFS-189231

[14] K. S. Mustaqeem, "1D-CNN: Speech emotion recognition system using a stacked network with dilated CNN features", *Computers, Materials and Continua*, vol. 67, pp. 4039–4059, 2021. https://doi.org/10.32604/cmc.2021.015070

[15] Y. Qian, X. Xiang, "Binary neural networks for speech recognition". *Frontiers of Information Technology & Electronic Engineering*, vol. 20, pp. 701–715, 2019. https://doi.org/10.1631/FITEE.1800469

[16] S. Zheng, P. Ouyang, D. Song, et al., "An ultra-low power binarized convolutional neural network-based speech recognition processor with on-chip self-learning", *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, pp. 4648–4661, 2019. https://doi.org/10.1109/TCSI.2019.2942092

[17] W. L. Liao, T. T. Lee, W. W. Jiang, et al., "Augmented reality teaching system based on cognitive theory of multimedia learning – an example system on four-agent soup", *Applied Science and Management Research*, vol. 6, pp. 54–69, 2019.

[18] Y. Li, D. Zhang, H. Guo, et al., "A novel virtual simulation teaching system for numerically controlled machining", *International Journal of Mechanical Engineering Education*, vol. 46, pp. 64–82, 2018. https://doi.org/10.1177/0306419017715426

[19] X. Zhang, "Design and analysis of music teaching system based on virtual reality technology", *IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association*, vol. 30, pp. 196–202, 2018.

[20] B. Yin, "Practical teaching reform of art design based on virtual reality technology". *IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association*, vol. 30, pp. 703–709, 2018.

7 Authors

Yang Wang obtained her BA in English from JNU, Siping, China in 2007. She received her M.Eng. in NNU, Changchun, China in 2012. Presently, she is working as a teacher in the Public Teaching Department of Changchun Automobile Industry Institute, Changchun, China. She has published articles in more than 10 reviewed journals and books. Her areas of interest include ESP English teaching, college teaching in CALL, second language acquisition.

Ala Mughaid was born in Irbid, Jordan, in 1984. He received the BSC degree in Computer Science from Jordan University of Science and Technology (JUST), Jordan, in 2006, and the MSC in Engineering degree in engineering Computer Network from the Western Sydney University, Sydney, Australia, in 2010. Dr. Mughaid received the Ph.D. degree in Computer Science from Newcastle University – Sydney, Australia, in 2018. In 2018, Dr. Mughaid joined the Department of Computer Science, The Hashemite University, as an assistant professor, Zarqa, Jordan. Dr. Mughaid current research interests include but not limited to Cyber Security, Cloud Computing, Network Security, Artificial Intelligence, Virtual reality, Data ining. He is working voluntarily in many social services. E-mail: ala.mughaid@hu.edu.jo

Article submitted 2022-09-27. Resubmitted 2022-11-03. Final acceptance 2022-11-05. Final version published as submitted by the authors.