ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

An Interactive English Listening and Speaking Self-study System Based on Computer Vision

Shouren Wu[1] *and Qin Zhang*[2,*]

[1] School of Foreign Languages, Shaoyang University, Shaoyang 422000, Hunan, China
[2] Department of Basic Courses, Hunan Polytechnic of Water Resources and Electric Power, Changsha 410131, Hunan, China

SUMMARY: *With the development of modern society, many people have become accustomed to life under IoT and machine learning, and now interaction has become an important factor in affecting English listening and speaking skills. The more interaction the learners have among themselves, the more active they will be and the better the results will be. Vision is a way for people to look at the world and know about it. People use their eyes and brains to acquire, process and understand visual information. At present, the series of problems for mobile English listening and speaking learning are still being addressed, such as non-optimal functions and poor operability. Based on Smart Sensing and Communication, we fully utilize computer vision technology to design a set of computer vision modules for an interactive English self-study system, and have completed user demand analysis, overall architecture design, functional applications, etc. The System will be able to play sound and video generally. Add images, mind maps, art words and other multimedia to increase the learners' enthusiasm and interest in learning, and provide an intuitive and vivid experience for users. Based on the above experiments, it is clear that the system interaction is more convenient and pleasant for English learners. The system can address some deficiencies of the existing system by adding new functions to the current system, improving the user experience, and achieving an effective voice recognition accuracy of over 95 per cent. The studies in this paper provide necessary support for the application of both IoT networks and machine learning.*

KEYWORDS: *Computer Vision; Interactive English Listening and Speaking; Self-study System; Recognition Rate*

1 Introduction

In recent years, research on Internet of Things (IoT) networks and machine learning has been carried out by researchers in industry and academia. A typical type of IoT network and machine learning. With the spread of globalisation, people around the world need English to talk. The development of computer network technology has been progressing at a high speed, and many learning materials for learners now incorporate multi-media to create excellent language environments and better external learning conditions. With the development of the IT industry and mobile network technology in recent years, the traditional way of learning English is likely to be replaced by mobile learning mode. Many English learning systems with different functions have appeared in recent years. English is primarily used to learn how to speak and listen to learn English. Therefore, the listening-speaking ability of English learners

*zhangqinhnslsd@163.com
https://doi.org/10.65102/is2026936
Published: 30 April, 2026

in daily life is relatively high, and it also serves as the main objective of English teaching. One of the purposes of learning English is to help students improve their listening and speaking skills in English; at the same time, they should be able to communicate freely in English in their daily study and work, and enhance their independent study abilities. Computer vision began in the 1950s and, through computers, now extracts three-dimensional data about environments from captured pictures and videos.

Given the Internet of Things and machine learning algorithms, this study is based on the above circumstances and integrates computer vision technology with the characteristics of English listening and speaking. And then there are pictures, audio and video to develop an English listening and speaking self-study system; users can practice English listening and speaking anytime, anywhere, to boost their interest in learning, and after some practice, their English listening and speaking skills will improve effectively.

Based on the foundation of IoT and machine learning, the innovations in this paper are as follows: It can be seen that the clear page of the system in the computer repository is adjusted by the picture quality of the video; the computer vision-based system for capacity installation and unloading operations is simple and fast; by analyzing data from a self-learning system using computer vision, intelligent detection and effective monitoring can be achieved.

2 Related Work

At present, many scholars have conducted research and analysis on various different application areas of computer vision, as follows:

Barbu A thinks that many computer vision and medical image problems are now being addressed by learning from large-scale datasets with millions of observations and features in IoT networks and machine learning. He put forward a new high-efficiency learning scheme that reduces the sparsity constraints by gradually removing variables according to a standard and a schedule. It offered a theoretical basis for the convergence and selection consistency. Experiments on both real and synthetic data have shown that the proposed method is relatively simple in terms of computation and can be scaled well compared with other high-performing methods for regression, classification and sorting [1].

IoT Networks and Machine Learning. Rathore M. pointed out that computers are increasingly being used as essential work tools. Computer vision syndromes refer to a set of eye and other eye problems caused by extended use of computers. He thinks that an all-encompassing patient history and eye examination can help to identify problems through vision and refractive tests, correction, tear-film function tests, etc. Maintain good eye health and use eye exercises, alternation, rest, appropriate workplace lighting, ergonomics and corrections to the computer use method to prevent and treat this condition [2].

IoT networks and machine learning.Decost B L will conduct an all-encompassing analysis and comparison of the initial research on computer vision methods for plant species identification. He identified 120 peer-reviewed studies with multi-stage screening published in the last 10 years. Based on the above research, he presented the actual application methods of classification for the parts of plants and explored characteristics such as shape, texture, colour, margin, and vein structure. Based on the classification accuracy of public datasets, he also compared the methods. His work provides material for ecology and computer vision research [3].

Kadir presented a simplified computer vision application of a multi-layer perceptron (MLP)-based artificial neural network (ANN) for the accurate classification of wheat particles as bread or ented carbide. IPT was employed to obtain the principal visual features of the four dimensions, three colours and five textures. He copied a total of 21 visual features from the 12

main features to increase the diversity of the input population for training and testing the ANN model. The visual feature set of this paper was used as the input to the neural network model [4].

Wldchen J has developed a system for powder raw material characterisation in metal additive manufacturing (AM) based on computer vision and machine learning. For eight sets of commercial raw material powders, the system correctly identifies the material system based on powder images with an accuracy of more than 95 per cent. Based on the above results, it has been proposed that powder variation can be quantified according to processing history, linked to microstructural features of the powder, and thereby correlated with performance variations of Additive Manufacturing (AM), enabling the definition of objective material indices from visual observations. The above are the general advantages of a computer-vision approach: autonomy, objectivity and reproducibility [5].

Lopez-Fuentes L tends to focus on state-of-the-art systems that cover the same emergency he is studying and ignores important research in other areas. To show this overlap, the four directions of the survey are: types of emergency situations in computer vision research, objectives of the algorithms they can achieve, necessary hardware types, and algorithms used. Therefore, this paper will review the development of computer vision for emergency situations [6].

Khan S proposed and trained a multimodal biometric identification system based on a convolutional neural network (CNN) and k-nearest neighbors (KNN) to identify and recognise people using multimodal biometric scores. He has tested a model on noisy data and seen how badly it performs in adverse conditions. Computer simulations show that the CNN and KNN multimodal biometrics systems perform better than most of the top-performing state-of-the-art biometrics validation techniques [7].

3 Computer Vision and Operating System

3.1 Application and Development of Computer Vision

Based on Smart Sensing and Communication. With the development of computers and robots, people now hope that machines can gather information about their surroundings independently; thus, computer vision research has started to develop. A typical case of Artificial Intelligence is Computer Vision. It has a broad field of application and is used widely in various places, such as aerospace and smart home appliances for daily life [8, 9]. Since the 1970s of the last century, it has been studied continuously and is gradually being improved today. Since the 1950s, with the development of digital image technology and other technologies in recent years, computer vision has gradually entered the industrial market and now applies to various kinds of automated and intelligent production systems. After the 1990s, computer vision technology has been widely applied in recognition, and the two main forms are speech recognition and recognition based on text information. Many areas of research are directly related to it, such as artificial intelligence, neural networks, signal processing, computer graphics and images, pattern recognition, geometric calculations and statistics, etc. Its basic process is to use computers and cameras to replace human resources for perception and measurement of various objects; it is mainly three-dimensional geometry, and finally, the characteristics of the object are analyzed [10]. A number of computer vision algorithms need to ensure the integrity of the picture collection system. However, most current computer vision operating systems have a deficiency in a higher threshold. Specifically, the algorithms are not only very difficult but also exceptionally large; therefore, the mathematical theory and programming and practical skills of the developers must be excellent.

Based on Smart Sensing and Communication, some scholars have developed a general-purpose application library for this problem that can be employed to realise computer vision algorithms, thereby significantly reducing the high threshold for use of computer vision. At present, the main applications of computer vision technology are in five areas: input devices for sympathetic interaction between people and machines; automated operation and control programs, such as driverless cars and robots in the development industry; monitoring and testing systems for images and videos; physical modeling, such as biomedical engineering and medical experiments in a topological environment; and organizing information to build a database of images and their contents [11]. Computer vision technology is now capable of performing some tasks that previously required human observation, and thus the cost and time for production have been reduced. It is not as high as those for the general population.

3.2 System Requirements Modeling

Based on Smart Sensing and Communication, it aims to enhance learners' self-study ability. The learning system that has been studied in this paper is based on the mode of human-machine dialogue, which includes listening and oral training, as well as daily communication and dialogue. The above ways have a high degree of interaction and promote independent learning by students. Based on the above system, users can exercise their English listening and speaking skills according to their own circumstances. The system needs to serve the various demands of users learning independently and offer flexible options for free-time and free-place study. It allows users to select specific learning content (such as partial life or business English) freely, enables learners to exercise their own initiative in learning, utilises the advantages of mobile system learning, and offers a variety of English listening and speaking learning materials, as well as convenience and rich multimedia functions, network resource sharing, etc. [12] The general architecture diagram of the self-study based on the interactive English listening and speaking system platform is shown in Figure 1.

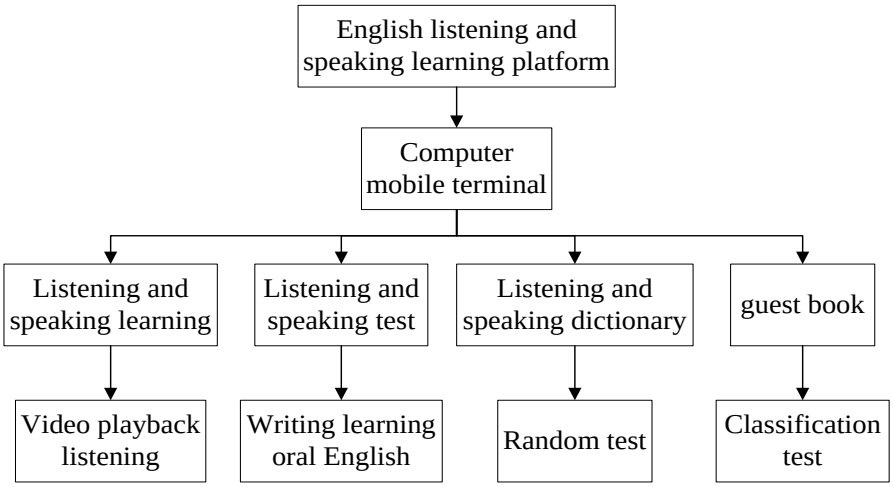


Figure 1: Structure Diagram of English Listening and Speaking Learning System

For example, when users carry out oral dialogue training, the mutual aid function of the system plays a positive role by improving the ability of English audio file sharing, thus avoiding the problem of non-standard or "dumb English" pronunciation and manifesting instead as a serious imbalance in the development of listening, speaking, reading, and writing skills, strong reading ability, and relatively weak listening and speaking ability [13]. At the

same time, in order to achieve the function of edutainment, one should also add rich multimedia elements, design some fun games for listening and speaking English, or include movie clips and dubbing; through these games and activities, subtly enhance users' English listening and speaking skills and make users feel that English listening and speaking is also a kind of fun. The interface of the system has taken into account users' habits in the design stage and is relatively convenient for them to operate, as well as providing all necessary functions [14].

3.3 Design and Implementation of the System

Based on Smart Sensing and Communication, this system in the study employs the S/C architecture and can be divided into a server side and a client side. The client is developed using network application technologies, such as Eclipse + DJK + ADT plug-in; the server side employs JavaE based on the model of the Spring framework and has a network service layer and an entity service layer, adds a Tomcat server to handle HTTP requests from multiple clients, sets up a database for the program, and runs it on the main service machine [15]. The server and client send network data using the JNOS+HTTP protocol. The whole diagram of the system is shown in Figure 2.

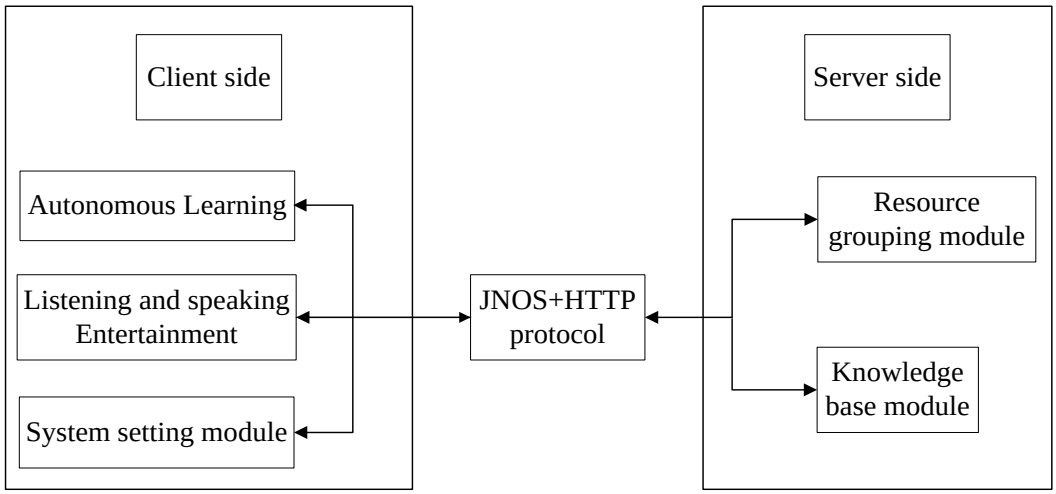


Figure 2: Platform system structure

Based on Smart Sensing and Communication, direct interaction with the user can be achieved at the module layer of the client; English learning resources can be downloaded from the network and stored in the desktop database; the functions of English listening and speaking learning in the system can be displayed; these functions are responded to in the module layer of the server, feedbacked to the client after processing, and work in cooperation with the client to help users learn [16].

Computer vision has been used to optimise and improve the self-learning system for English listening and speaking. The three modules of the system are: listening and speaking entertainment; autonomous learning and software settings; and the application functions of this system. The structure chart of the function is as follows: Figure 3.

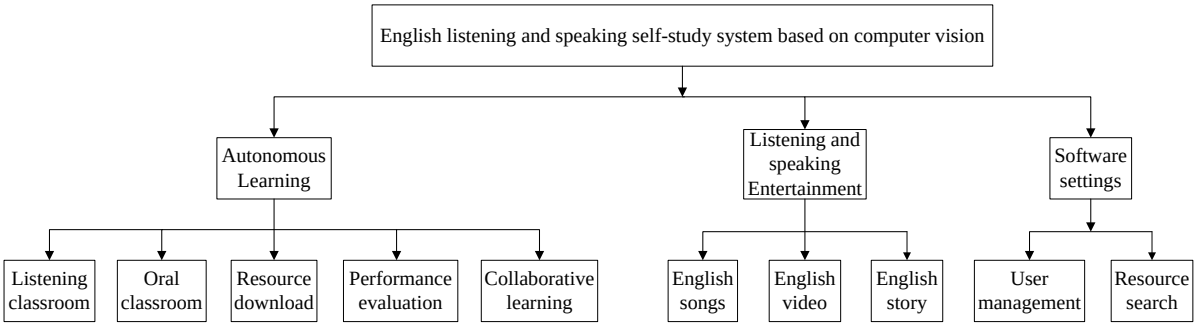


Figure 3: Client Function Structure Diagram

Self-directed learning is the learner's own setting of learning objectives and adjustments to meet the new demands of life. They can choose to learn and apply various strategies and regularly assess how well these strategies have improved learning results. Autonomous learning functions include listening and speaking classrooms, free downloading of resources, testing and grading of tests, assistance in cooperative learning, etc. The listening classroom includes exercises such as playing audio, dialogues in daily life and listening level tests, which users can carry out according to their own actual needs to improve their listening skills. The functions of the oral English classroom also include reading English classics, oral English ability tests, daily communication in oral English, and dubbing video clips from movies or TV series. Among them, English reading should be based on specific situations, display the English text in the user interface, and be able to click and play the corresponding machine voice when there is a network interruption; at that time, it can be read by a user to achieve the goal of training oral skills [17].

For Smart Sensing and Communication. Of course, before this, users can also download the corresponding voice files without a network connection, and the downloaded files are saved as a binary array in the database of devices with the system installed. Users can post their experiences in learning English listening and speaking in the collaborative learning area and learn from others. Identify and evaluate the practice results of users with the system's speech recognition function components. The column of listening and speaking entertainment includes English songs, English videos, bedtime stories and dubbing mentioned above. It has many new forms, such as pictures, videos, audio, etc., and combines them with English listening and speaking learning to arouse students' interest in learning to the greatest extent [18].

All the different auxiliary systems will be introduced in the following sections, such as the speech recognition system, automated response system, text-to-audio system, and so on.

3.4 Speech Recognition System

Smart Sensing and Communication. Speech recognition technology is a highly interdisciplinary field that includes a number of related sciences, such as the "three studies" (phonetics, acoustics, and linguistics) and the "four theories" (artificial pattern theory, digital information processing theory, signal theory, and intelligent recognition theory), among others; therefore, its application prospects are also very wide. The general way is that the machine converts speech signals sent by the speaker into the text or instructions of the corresponding words during recognition and understanding [19].

A full set of a speech recognition system is shown in Figure 4. The four parts are feature extraction, decoder feature extraction, sound and speech model exercises, etc. The first stage of the process is model practice, in which a sound model refers to specific parameters of voice

model parameters and a speech text database, and a series of syntax and semantics analysis of the speech model is performed; the second stage is feature extraction, that is to say, complicated information that does not function in speech signal recognition is deleted, and some information that shows the essential characteristics of voice is saved; then the saved information is decoded with a decoder, and the voice data is identified as text with the help of a voice and a voice model [20].

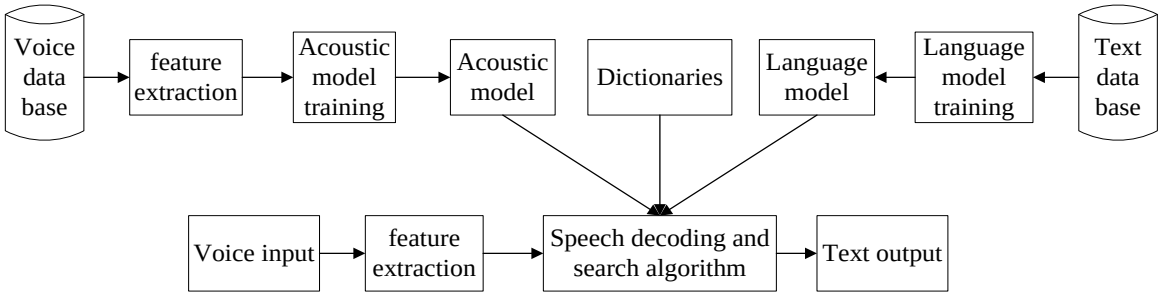


Figure 4: Block Diagram of Continuous Speech Recognition

The basic architecture expression of the speech recognition system is as follows:

$$Q^* = agr\max_q O(Q \mid T) \tag{1}$$

$$= agr\max_q \frac{O(Q \mid T)O(Q)}{O(T)} \tag{2}$$

$$\approx agr\max_q O(Q \mid T)O(Q) \tag{3}$$

T Where, Q refers to the serial number of the text, T refers to the process of speech input. Equation 1 explains the ultimate goal of speech recognition, that is, to find out the most likely text sequence number based on the established speech input. According to the Bayes equation, Equation 2 is obtained. The denominator in the equation is the probability size of the text with the most possible possibility. Because the parameter is smaller than the text sequence to be solved, Equation 3 is obtained. The left part is the probability of the most likely speech in a given text sequence, the so-called sound model; the right part is the probability of the most likely text sequence number, the so-called speech model.

The Euclidean distance expression of several vectors for similar feature extraction is:

$$s_{2.1} = \sqrt{\sum_{i=1}^{128}(y_{i1} - y_{i2})^2} \tag{4}$$

For Smart Sensing and Communication, to simplify the modelling process, it is necessary to incorporate information processing of the operation process in the system. The parameter model of the identification architecture of the system is assumed to be:

$$Z(m) = [z_1(m), z_2(m), \ldots, z_n(m)]^R \tag{5}$$

$$Q_h = [q_1(h), q_2(h), \ldots, q_n(m)]^R, h < n \tag{6}$$

$$s_h(m) = [s_1(m), s_2(m), \ldots, s_n(m)]^R, h < n \tag{7}$$

$$T_{h-1}(m) = [t_1(m), t_2(m), \ldots, t_n(m)]^R, h < n \tag{8}$$

A parameter identification method is employed to adjust the parameters of the system, and then, in combination with an adaptive fuzzy control method, the control law for the incentive source under limited conditions is obtained as follows:

$$q_1 \equiv \lambda_1 = a\tan 3(p_y, p_t) \tag{9}$$

$$q_2 \equiv \lambda_2 = a\tan 3(p_y, \mathrm{c}2p_t + sp_y - l) \tag{10}$$

Correspondingly, p_m and p_n can be found, namely:

$$A_{\mathrm{n}} = \frac{Q(m)}{Q(m)} \times \frac{Q_1 \exp(p_m - p_n)}{p_m} \tag{11}$$

Under a particular parameter-identification method for a specific model, the range of the interactive system is relatively large. Thus, the error of the system is automatically monitored and adjusted, and the final adaptive law is:

$$\Gamma(q) \bullet T_2 = \prod_3 (q_1) \tag{12}$$

Without considering the span size of the system, the equation for tracking the construction background of the system is:

$$x(k+1) = \Theta(k)x(k) + w(k) \tag{13}$$

$$z(k) = \mathrm{K}(k+1)x(k+1) + v(k+1) \tag{14}$$

The two kinds of the system environment are dynamic and static. In the two states, based on the method of global positioning and acceleration, the built control model is employed to monitor the human-computer interaction operation of the system. Vibration under a specific structural condition is as follows:

$$b_i = \sum_m Q_{ij} v_i (k+1) \tag{15}$$

Based on the perturbation characteristics of the pose distribution, the system uses its characteristic inertia at a certain point to compensate for the nonlinear features, and then the linear eigenvalue of the parameter matrix is:

$$Q_{ij} = \frac{1}{b_i} P_j v_j (u+v) \tag{16}$$

For Smart Sensing and Communication, to build a dynamic human-computer interaction control model and track it, we should first use an adaptive method to obtain the target object, then calculate the mass linear coefficient of the model, and finally obtain the deviation

tracking value of its monitoring:

$$\partial = \theta - \theta_r \tag{17}$$

The Design Distribution Value of vibration mode coordinates is:

$$\dot{\partial} = w_i - \theta_r \tag{18}$$

Select the corresponding module matrix, and the final adjustment function is as follows:

$$w_i = -c\partial + \theta_r - a_1\gamma_1 \tag{19}$$

The automatic adaptation law obtained by the base method of decreasing subsystem is:

$$\ell = \alpha_1 \lambda V^2 \mathrm{e} \tag{20}$$

Based on the above series of computational analyses, it can be concluded that the self-study system designed in this study has good stability and convergence.

Next, the system will be used to recognise the text information, such as English letters, Arabic numerals, etc. The results are as follows (see Table 1), and the actual identification rate is about 99 per cent.

Table 1: Text Recognition Results

Text content	Recognition rate	Text content	Recognition rate
Numbers 0-9	98%	Female	99%
EXIT	97%	TOILET	95%
Male	96%	NO SMOKING	94%

In order to test the actual recognition rate of the system designed in this paper, we recorded a computer headset with a specific command, played it at the same volume, and carried out many field experiments to determine the recognition accuracy at different distances. The results are as follows: Figure 5.

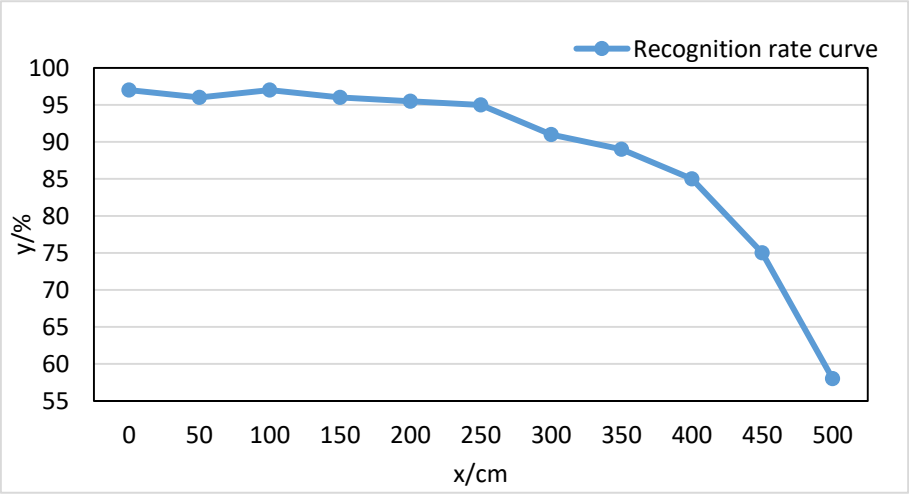


Figure 5: Recognition Rate Curve at Different Distances

Smart Sensing and Communication: The two reasons for identification inaccuracy are computer parameters and the performance of sound-acquisition facilities. A microphone has been added to the interactive system in this study for sound collection. As shown in Figure 5, when the distance is less than 350 cm, the recognition accuracy rate is over 90%; and if the distance is less than 250 cm, it is more than 97%.

3.5 Automatic Response System

For Smart Sensing and Communication. The first goal of the automatic response system is to automatically answer the voice input from a user after it has been received, and it is necessary for the language processing level. Generally speaking, search engines access a database to return many relevant data sets for users, and then these data sets are filtered. Based on the two above statements, it can be seen that an automated response system has the advantages of being less time-consuming; it can help users acquire only the minimal amount of information they need from a large number of sources [21].

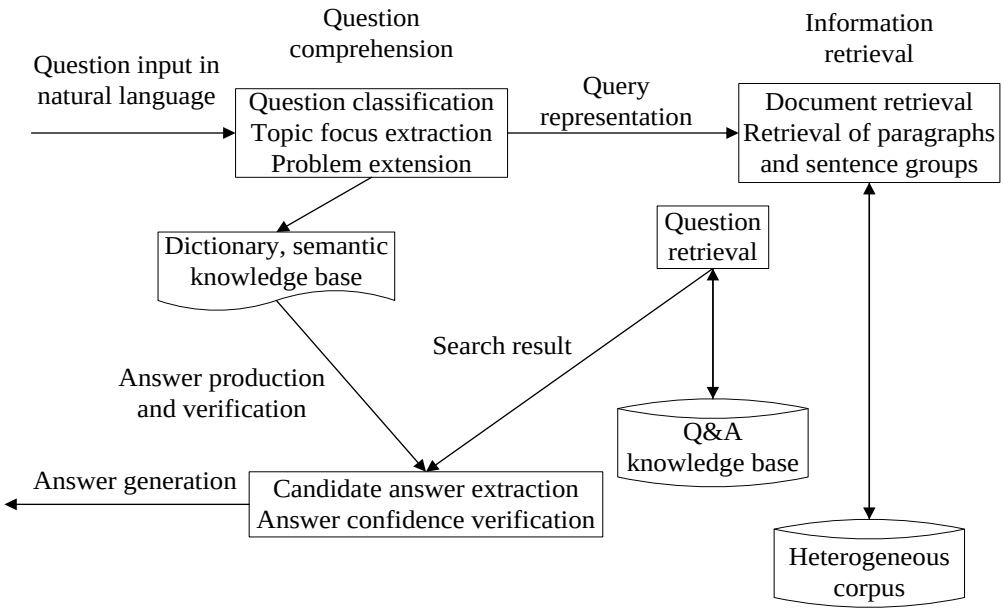


Figure 6: Processing Framework of the Response System

The first three modules of the architecture diagram for the automated response system are, in order, question analysis, text retrieval and final answer generation and verification. To ensure the high accuracy of the system, the first part (understanding questions) effectively analyzes the syntax and vocabulary in the database to perform its basic operation; the second part (text retrieval) uses traditional text retrieval techniques to answer and rank them by relevance; the third part (final answer generation and verification) refers to the text document obtained in the second step via semantic theory and then selects the highest-ranked answer for the user [22].

3.6 Text to Audio System

Text-to-audio systems are widely used for speech synthesis in the field of speech synthesis, as the name suggests, and convert text documents into audio.

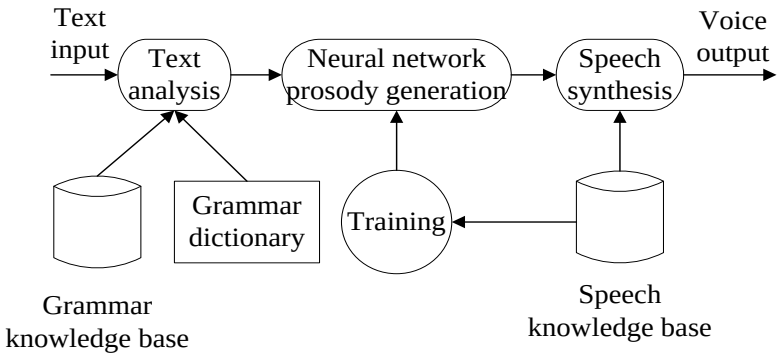
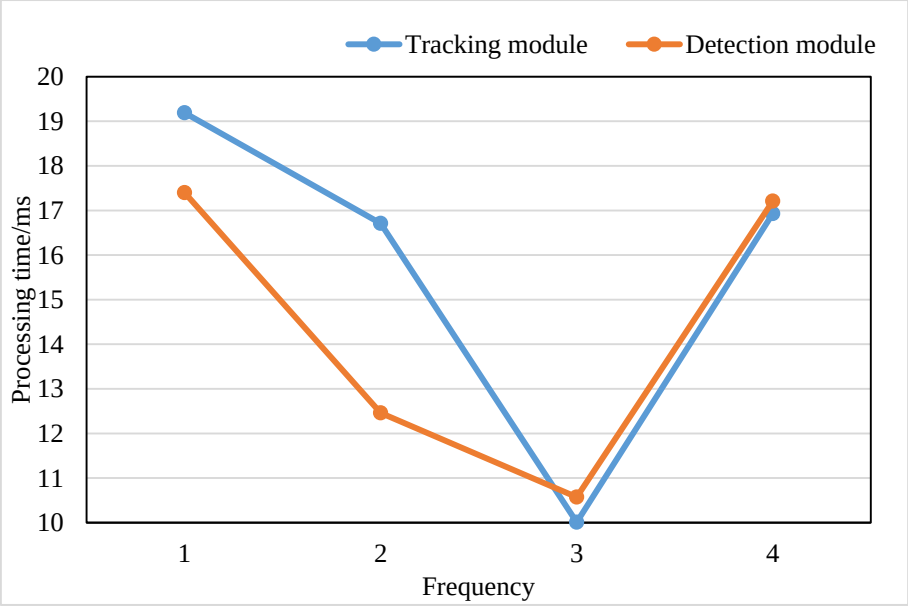


Figure 7: Frame Diagram of Text-to-Speech System

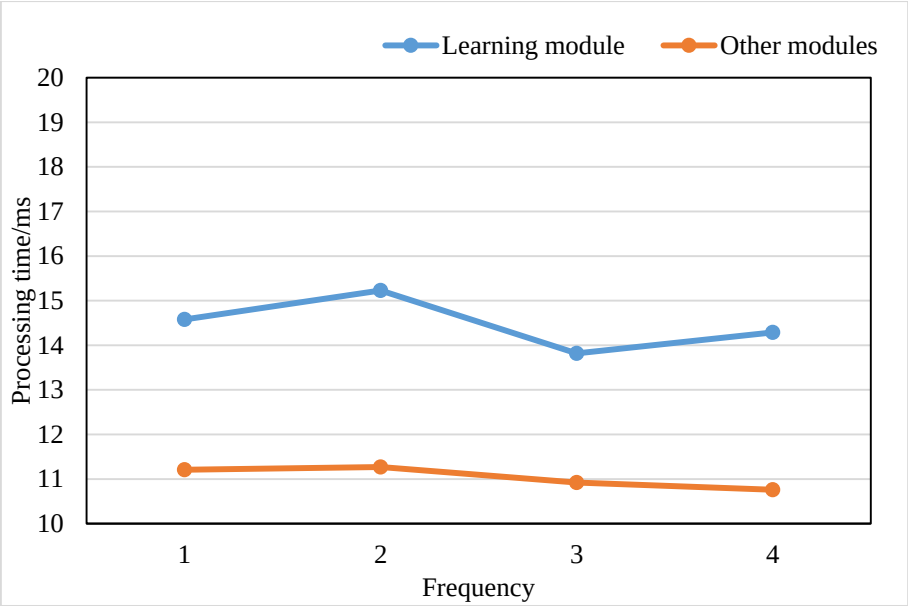
Figure 7 is the process in the system. As shown in the processing process chart, the three blocks of the system are: the processing module for text, rhythm and sound synthesis. As this study will be implemented as a self-study system based on computer vision, the practice of generating blocks of sound rhythm will be carried out on the computer side to test the system more effectively. The first block (word processing) is at the front end of the system, and the word conversion function belongs to this block; the main purpose of the second block (tone) is to analyse the rhythm architecture, light and voice tone from the text database; the third block (sound synthesis) is located at the back end of the system, and then output to the client [23].

4 Interactive System

To perform an all-encompassing study of the system, we should first do so by comparison. Therefore, we used the two computers with the highest matching rates to the system, and one of them was employed to add interactive performance for testing and analyzing the processing time of each frame during the operation of the system. The last test results are as follows: Figure 8.



A

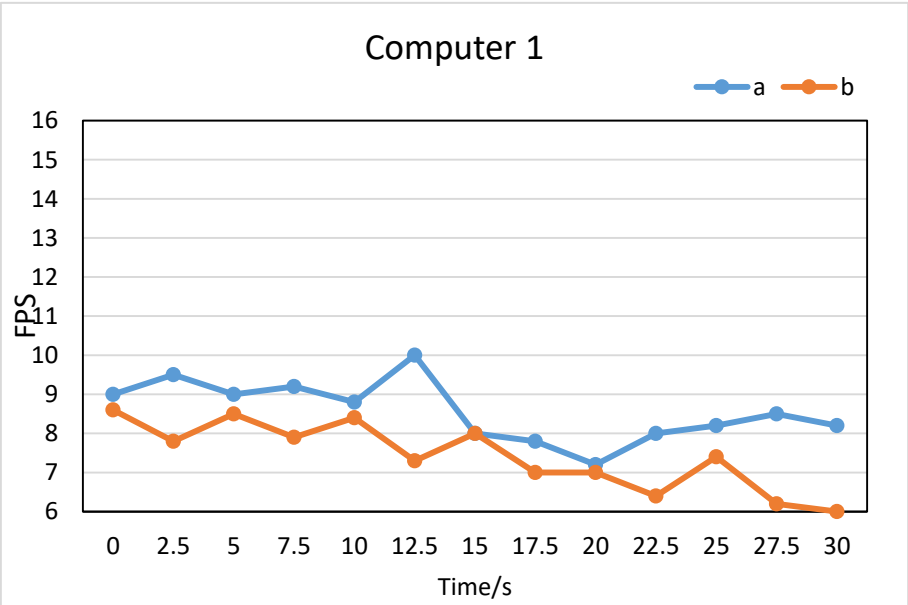


B

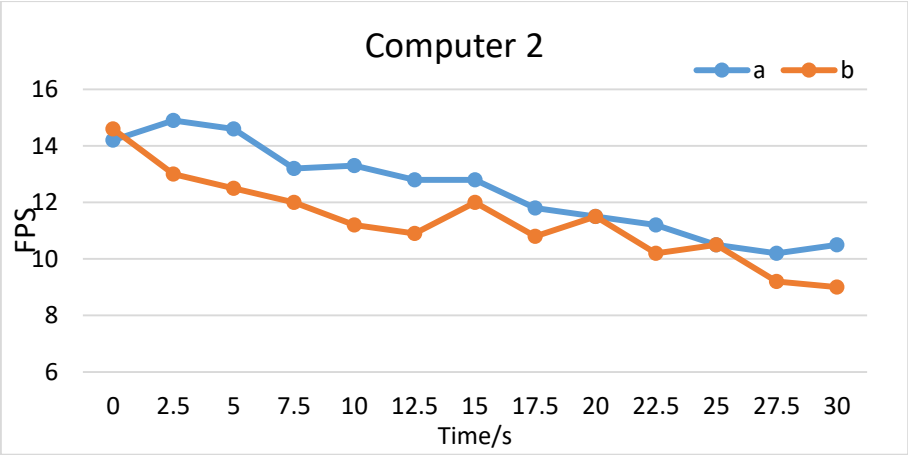
Figure 8: Processing Time of Computer Test

According to the data obtained from the two computers, it can be observed that the computers with better interactive performance have relatively short processing times, with a minimum of 10.76ms. The reason for the above result is that it has a higher-performance CPU, so it can be concluded that a stronger computing power will result in a shorter time.

Next, select the 30s process for the system start-up, and the test results of the two computers are shown in Figure 9.



A



B

Figure 9: Test Computer Frame Rate Diagram

Computers with interactive performance also have a high running frame rate, up to 15fps, and are more volatile; therefore, a CPU system with good performance is required for real-time operation.

Smart Sensing and Communication: A questionnaire survey was conducted among a group of students to better understand their application of and impact from the self-study system based on computer vision in English listening and speaking classes. A total of 388 questionnaires were distributed, 367 were collected validly, and the collection rate was 94.6%. To test the students' ability to study independently more accurately, the main content of this survey will cover their study methods and listening-speaking skills. The results of the questionnaires are shown in Table 2, and A and B are the experimental group and the control group, respectively.

Table 2: Comparison of Autonomous Learning Strategies and Listening-Speaking Levels.

project		class	percentage
Autonomous Learning Strategies and learning ability	Make a reasonable study plan	A	55. 8%
		B	38. 4%
	Complete the learning task according to the plan	A	80. 5%
		B	46. 7%
	Listen purposefully	A	76. 9%
		B	50. 4%
	Using listening skills	A	83. 2%
		B	55. 6%
English listening and speaking level	Be able to understand English teaching, and be able to discuss and speak as required	A	95. 7%
		B	89. 6%
	Be able to understand special English programs in English speaking countries	A	80. 3%
		B	65. 1%
	Be able to understand the conversation or lecture of people from English speaking countries, and understand the main points and details	A	76. 4%
		B	59. 8%
	Be able to have a simple conversation with people from English speaking countries on daily topics	A	93. 2%
		B	85. 7%
	Be able to basically express personal feelings and describe personal experience	A	86. 4%
		B	70. 5%
	Be able to make thematic reports on topics in professional fields	A	69. 9%
		B	50. 2%

As shown in the table, the self-study strategy and ability of the experimental group of students are significantly better than those in the control group; a difference of 33.8% was achieved, and their English listening and speaking levels have been improved, indicating that the self-study system based on computer vision for these students has had an extremely positive effect.

In order to learn about the results of this system, another group of students was chosen from the subjects to be tested in oral English. As shown in Figure 10, the girls' total scores were 0.46 points higher than those of the boys. Among them, there were 149 experimental class students (73 male, 76 female) and 141 students (100 female, 41 male); together, they voluntarily participated in the survey.

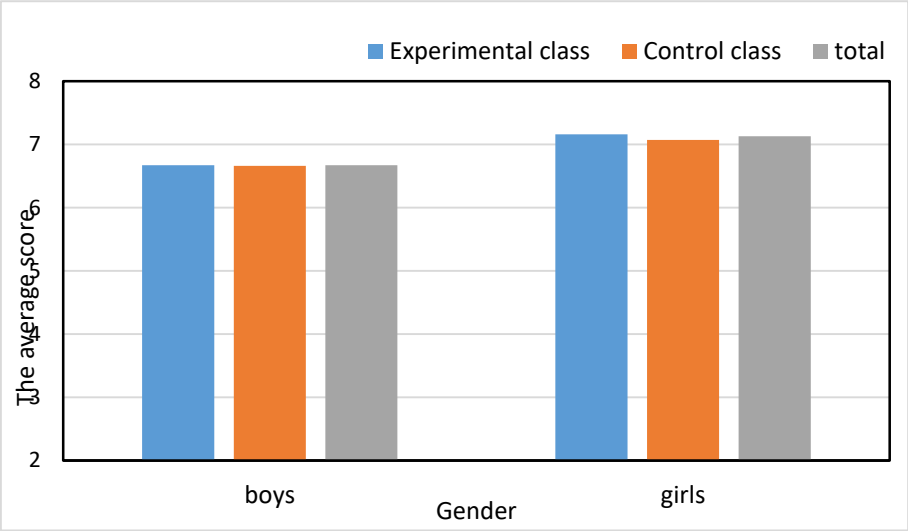
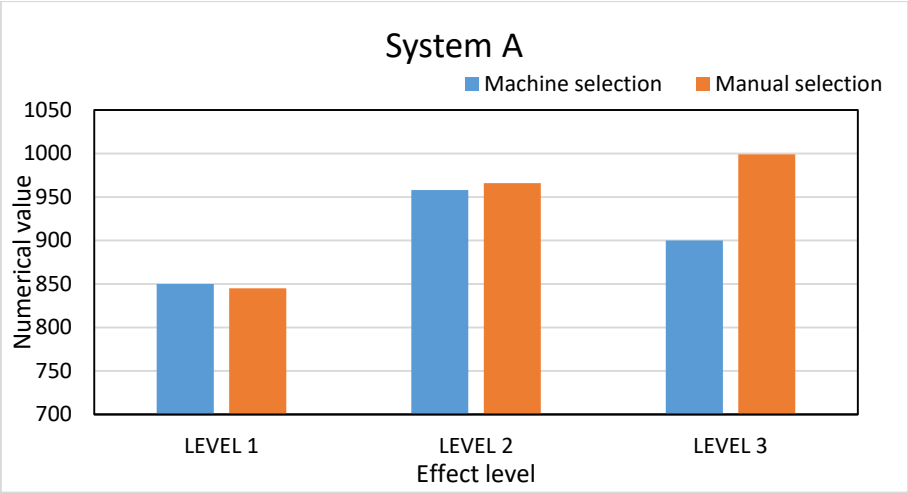
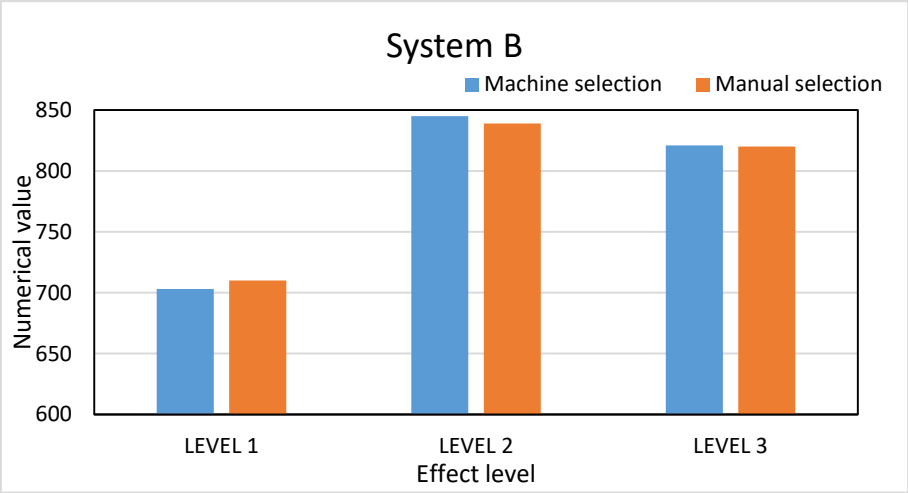


Figure 10: Oral Test Results

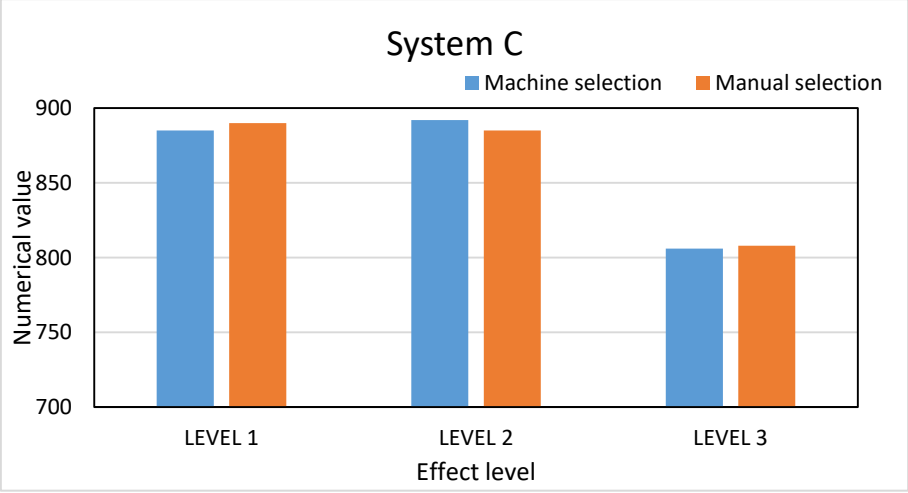
Based on the results of the above oral English tests, it can be concluded that after a semester of listening and speaking training, the average scores of the students in the experimental class are generally higher than those in the control class; thus, it can be verified that the system has had a good effect on the spoken and listening ability of English learners. At the same time, computer vision technology is employed for both the machine-based and manual detection of the system (A), and several different English listening and speaking systems (B and C systems) have been selected in this paper. The first, second and third levels of the effect are shown here, and the final sample results are in Figure 11. As shown in the figure, there is a small difference between manual and machine detection, and the minimum deviation is 1 value.



A



B



C

Figure 11: Comparison Results Based on Computer Vision

Smart Sensing and Communication: Based on the above data, the deviation is calculated and the results are shown in Figure 12. The actual results of the above test showed that the system deviation rate of the design in this study was the lowest, and its value was 1.7; it also had good stability.

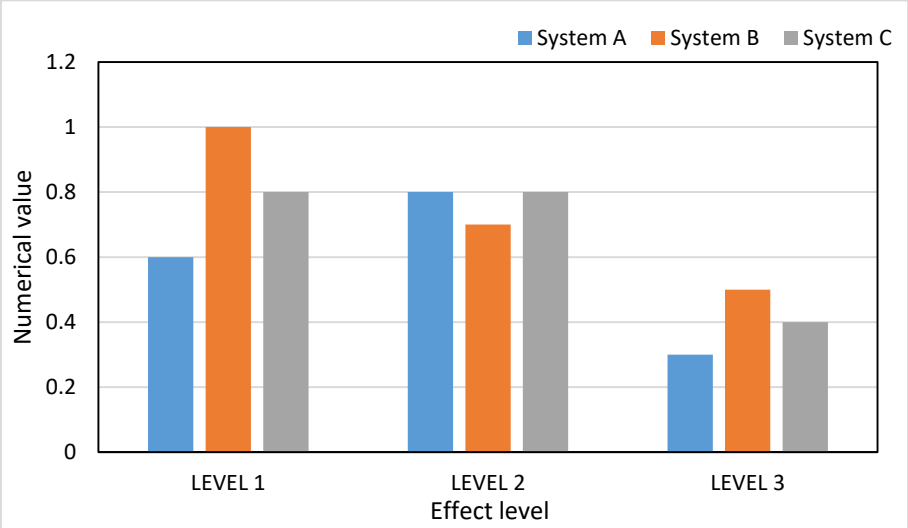
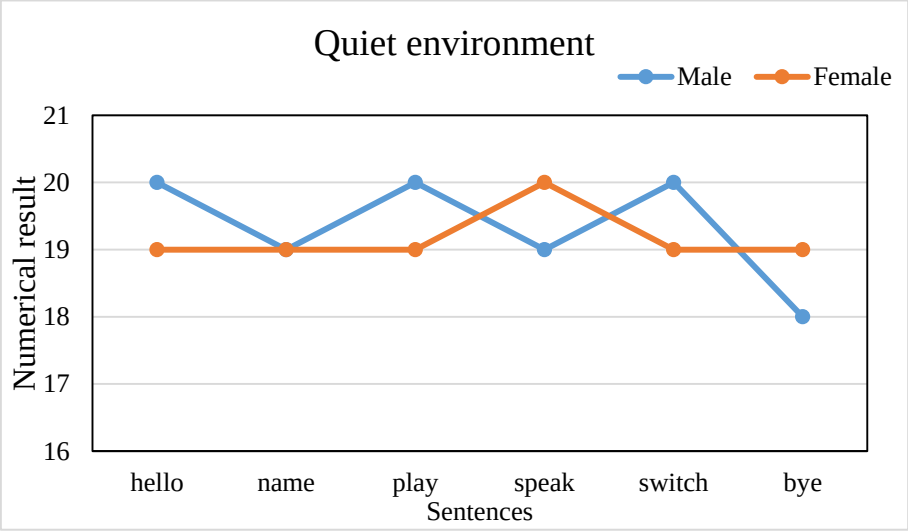
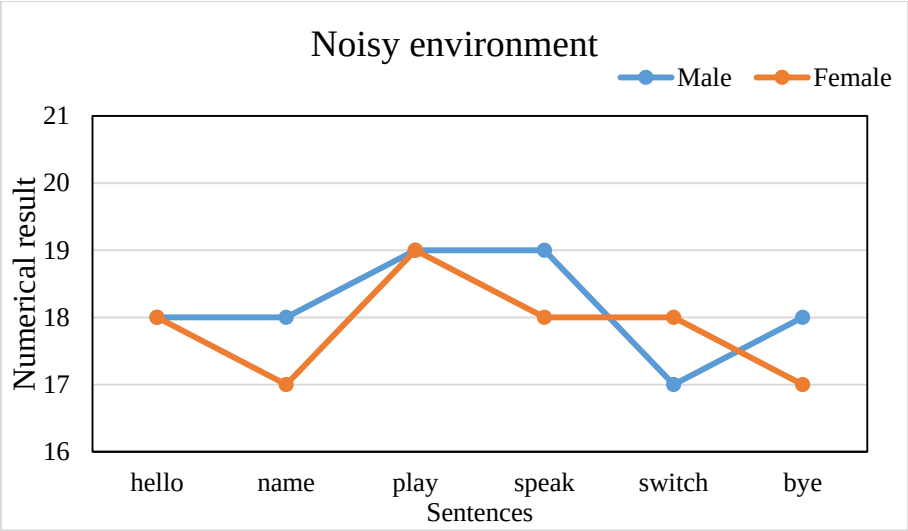


Figure 12: Deviation Comparison Results

In order to ensure a high recognition rate and good sound stability of the whole system designed in this study, we have tested the corresponding blocks by emitting the speech of the user a total of 20 times. In addition, to ensure the practicality of the experiment, two different background conditions were chosen for recording, namely a quiet one and a noisy one, and the number of successful trials was recorded. The final results are as follows: Figure 13.



A



B

Figure 13: Test Effect Statistics for Speech Recognition Module

As shown in Figure 13, in the quiet state, the system can recognise the voice of men with a rate of 96.7% and that of women at 95.8%, both exceeding 90%. In the quiet state, 90.8% of men and 89.2% of women, generally above 85%.

To realise the effect of a computer vision-based self-learning system in human-computer interaction, all-round design of the system has been carried out through experiments. Finally, it has been found that the system can effectively ensure the realisation of human-computer interaction, and is simple in calculation and strong in real-time performance.

5 Discussion

Based on Smart Sensing and Communication, according to the above investigation and experimental analysis, it can be concluded that the self-study system based on computer vision is feasible and necessary for improving students' English listening and speaking abilities. This relatively advanced learning mode is suitable for reference and dissemination by the relevant department. The English listening-speaking self-learning system built in this paper can perform the basic functions of practicing listening and speaking, and is also somewhat entertaining. Users can obtain diverse listening and speaking materials independently, and are not restricted by time or place. Based on modern technical means and the characteristics of English listening and speaking learning, this paper offers some convenience for users to promote their learning of English listening and speaking, reduces labour costs, and has strong practical value. Test experiments of the system have shown that it is user-friendly, feasible and accurate.

6 Conclusion

Research on the Internet of Things network and machine learning is a hot topic in recent years, attracting the attention of many scholars and practitioners. As a typical form of an Internet of Things network and machine learning, this paper proposes and builds an interactive self-learning system for improving the listening and speaking ability of English learners based on computer vision technology. Computer vision technology is one of the highlights of AI at

present, and many scholars are paying attention to it. The self-study system based on computer vision in the optimisation process also has some defects.

IoT and machine learning are sometimes too simple for the system; thus, an effective guarantee has not been provided through supervision mechanisms. Self-study mainly relies on the learners' own awareness and initiative, but some learners with poor self-discipline need to be added to a certain supervision system to ensure good learning results. Therefore, in the actual operation process, we need to provide reasonable direction and appropriate support to help the English learners enhance their listening and speaking abilities further. The Hardware Facility Arrangement for the computer will also affect how well the system performs in practice. Next, we will design a communication platform that can be updated dynamically at any time to expand the scope of application for its resources and gather various resources to broaden the road of English listening and speaking study. The studies in this paper provide some support for the application of both IoT networks and machine learning.

About the Author

Shouren Wu was born in Lianyuan, Hunan Province, China, in 1991. He is a Teacher at Shaoyang University. He is studying to be a doctor at Hunan Normal University. His research areas are English-language teaching and discourse analysis.

Qin Zhang was born in Changsha, Hunan, China, in 1991. She is a teacher at the Hunan Polytechnic of Water Resources and Electric Power. She received both her undergraduate and graduate degrees from Hunan University. Her Research interests are applied linguistics, second language acquisition and English teaching. E-mail: zhangqinhnslsd@163.com

References

[1] Barbu, A., She, Y., & Ding, L. (2017). Feature selection with annealing for computer vision and big data learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(2), 272-286.

[2] Rathore, M. I. (2017). Computer vision syndrome—An emerging occupational hazard. Research Journal of Science and Technology, 9(2), 293-297.

[3] Wäldchen, J., & Mäder, P. (2018). Plant Species Identification with Computer Vision Technology: A Systematised Literature Review. Archives of Computational Methods in Engineering, 25(2), 507-543.

[4] Kadir, S., Sabanci, A. (2017). Computer Vision-based Method for Classification of Wheat Grains using Artificial Neural Networks. Journal of the Science of Food and Agriculture, 97(8), 2588-2593.

[5] DeCost, B. L., Jain, H., & Rollett, A. D. (2017). Computer vision and machine learning for autonomous characterization of AM powder feedstocks. JOM, 69(3), 456-465.

[6] Lopez-Fuentes, L., Joost, V., & González-Hidalgo, M. (2017). Review on computer vision techniques in emergency situation. Multimedia Tools and Applications, 77(13), 1-39.

[7] Khan, S., Rahmani, H., & Shah, S. A. (2018). A guide to convolutional neural networks

for computer vision. Synthesis Lectures on Computer Vision, 8(1), 1-207.

[8] Elngar, A. A., Arafa, M., Fathy, A., Moustafa, B., Mahmoud, O., Shaban, M., & Fawzy, N. (2021). Image classification based on CNN: A survey. Journal of Cybersecurity and Information Management, 6(1), 18-50.

[9] Ding, S., Qu, S., Xi, Y., & Wan, S. (2019). Stimulus-driven and concept-driven analysis for image caption generation. Neurocomputing.

[10] Martynenko, A. (2017). Computer vision for real-time control in drying. Food Engineering Reviews, 9(2), 91-111.

[11] Nie, S., Meng, Z., & Qiang, J. (2018). The deep regression Bayesian network and its applications: Probabilistic deep learning for computer vision. IEEE Signal Processing Magazine, 35(1), 101-111.

[12] Tretola, M., Di Rosa, A. R., & Tirloni, E. (2017). Former food products safety: Stereomicroscopy and computer vision for evaluation of packaging remnants contamination. Food Additives & Contaminants, 34(8), 1427-1435.

[13] Zendel, O., Murschitz, M., & Humenberger, M. (2017). How good is my test data? Introducing safety analysis for computer vision. International Journal of Computer Vision, 125(1), 95-109.

[14] Lenoir, J., Cotin, S., & Duriez, C. (2017). Interactive physically-based simulation of catheter and guidewire. Journal of Preventive Medicine Information, 61(13), 2132-2141.

[15] Wang, G., Zuluaga, M. A., Li, W., et al. (2019). DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(7), 1559-1572.

[16] Boneva, L., & Linton, O. (2017). A discrete choice model for large heterogeneous panels with interactive fixed effects with an application to the determinants of corporate bond issuance. Journal of Applied Econometrics, 32(7), 1226-1243.

[17] Cobrzan, C., Schoeffmann, K., & Bailer, W. (2017). Interactive video search tools: A detailed analysis of the video browser showdown 2015. Multimedia Tools and Applications, 76(4), 5539-5571.

[18] Brigdan, M., Hill, M. D., & Jagdev, A. (2017). Novel interactive data visualization: Exploration of the ESCAPE trial (Endovascular Treatment for Small Core and Anterior Circulation Proximal Occlusion With Emphasis on Minimizing CT to Recanalization Times) data. Stroke, 49(1), 193-196.

[19] Sun, A., & Jing. (2017). The E3 ubiquitin ligase NEDD4 is an LC3-interactive protein and regulates autophagy. Autophagy, 13(3), 522-537.

[20] Liu, A. A., Xu, N., Nie, W. Z., et al. (2017). Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Transactions on Cybernetics, 47(7), 1781-1794.

[21] Pamarthi, V., Grimm, L., & Johnson, K. (2019). Hybrid interactive and didactic teaching format improves resident retention and attention compared to traditional lectures. Academic Radiology, 26(9), 1269-1273.

[22] Lam, A. T., Ma, J., & Barr, C. (2019). First-hand, immersive full-body experiences with living cells through interactive museum exhibits. Nature Biotechnology, 37(10), 1238-1241.

[23] Javad, R., Behzad, G., & Afsaneh, G. (2017). L2 motivational self-system and self-efficacy: A quantitative survey-based study. International Journal of Instruction, 11(1), 329-344.