INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Continuing Engineering Education and Life-Long Learning

ISSN online: 1741-5055 - ISSN print: 1560-4624
https://www.inderscience.com/ijceell

Automatic speech recognition based on adaptive parameters technology in English MOOC teaching system

Juan Qian

DOI: 10.1504/IJCEELL.2025.10073313

Article History:

Received:	20 January 2025
Last revised:	07 April 2025
Accepted:	26 May 2025
Published online:	10 October 2025

Automatic speech recognition based on adaptive parameters technology in English MOOC teaching system

Juan Qian

College of Foreign Language and International Education,
Anhui Xinhua University,
Hefei, 230088, Anhui, China
Email: qianjuan0307@126.com

Abstract: With the rise of MOOCs and the popularity of the internet, more and more English learners are beginning to study independently online. In order to accomplish system platform adaptation, enhance system compatibility, and increase the scoring mechanism's accuracy and dependability, this study studies a scoring approach based on adaptive parameters (AP). The system acquires the formant information of the learner's pronunciation and the standard reference pronunciation following pre-processing, FFT transformation, formant extraction, and other procedures. It contains a separate scoring parameter creation module to create adaptive parameters before speech scoring. The learner pronounces multiple voices in the scoring parameter generating module, and the expert assigns a score based on the learner's pronouncing experience. The experimental class's pre-test and post-test scores are in agreement with the design sample t-test result. This study can increase students' motivation to learn and enhance their practical English skills.

Keywords: internet of things; IoT; automatic speech recognition; ASR; English MOOC teaching; humidity sensor; performance evaluation.

Reference to this paper should be made as follows: Qian, J. (2025) 'Automatic speech recognition based on adaptive parameters technology in English MOOC teaching system', *Int. J. Continuing Engineering Education and Life-Long Learning*, Vol. 35, No. 8, pp.199–216.

Biographical notes: Juan Qian received her Doctor's degree from the Lyceum of The Philippines University, Philippines. Currently, she works in the College of English and International Education, Anhui Xinhua University. Her research interest includes education, literature and cross-culture communication.

1 Introduction

Although many vibrant internet of things (IoT)-based English learning programs are available, most English reading software is internet-dependent. It has heavy demands on the user's network environment. In addition to its stunning graphical user interface and robust data processing capabilities, the platform can rapidly and easily construct applications due to its extensive development components and complete openness. The English MOOC teaching system training system can train learners to correctly pronounce

words using sound and animation, allowing learners to receive feedback on their pronunciation and make necessary corrections to raise their proficiency. An English-assisted pronunciation training system built on the IoT's automatic speech recognition (ASR) platform is soon to be developed to accommodate users' varying reading needs in various contexts. The IoT connects multiple terminals to the internet, which can effectively collect and analyse data, thus providing the possibility for personalised learning for students. ASR technology is based on deep learning. Through analysing speech signals, it can achieve accurate speech conversion to text and understanding of complex semantic structures. The organic integration of the two into the English MOOC platform can significantly enhance students' learning experience through real-time feedback and interactive exercises, which is of great significance to promoting educational reform and the development of online teaching towards a more intelligent development.

The system can read local and online documents, which is convenient for users with different reading needs in different environments. Users can read English articles in the local SD card's content and access resources on the network. Simultaneously, to aid users, the system also suggests reading material, which significantly enhances the reading material and is highly practical. The benefit of offering both online and local reading materials is that, without affecting the user experience, readers can continue to utilise the English-assisted speech training system, which is based on the IoT and an ASR platform, in an offline mode. With the vigorous development of MOOC and IoT technology in education, more and more innovative learning methods are being applied to learning scenarios outside the classroom. MOOC enables scholars and students worldwide to transcend the boundaries of time and space and enjoy high-quality educational resources. At the same time, IoT technology provides a broader space for personalised and interactive learning by integrating smart hardware and networked devices. The introduction of ASR technology in English teaching can identify and correct students' pronunciation errors in real-time, simulate an authentic language environment, and improve the interactivity and accuracy of learning. Integrating various technologies provides English learners with a more convenient and efficient learning platform. It dramatically enhances the fun and practicality of learning, bringing new opportunities for the innovation and development of English education.

In order to achieve system platform adaptation, enhance system compatibility, and increase the scoring mechanism's accuracy and dependability, this study suggests a scoring approach based on adaptive parameters (AP). This work builds an anti-noise voice recognition module based on humidity sensors to address the challenges of speech recognition in the current noisy environment and the noise insensitivity of humidity sensors. The primary function of the pronunciation formant's image display module is to qualitatively represent the variations in the pronunciation of the two by graphically illustrating how the learner's pronunciation and the standard reference pronunciation change over time.

2 Related work

ASR has been widely used, but its recognition effect is still problematic, and it is prone to recognition errors in noise, dialects, and long conversations. Dhanjal and Singh (2024) emphasised the neural network-based speech recognition technology, datasets, toolkits,

and evaluation metrics used in the past, providing solutions for improving accuracy. ASR systems can be trained with a large amount of manually transcribed speech, and Aldarmaki et al. (2022) identified a model that can achieve fully unsupervised ASR. Sun (2023) employed an explanatory sequential design to examine the impact of utilising ASR technology with peer correction on improving second language pronunciation and speaking skills among English as a foreign language learners. Oruh et al. (2022) proposed an enhanced deep learning LSTM recurrent neural network to solve the problem that existing algorithms cannot accommodate the multiple computing units required to process continuous input streams. For low-resource languages, Mukhamadiyev et al. (2022) proposed an end-to-end deep neural network-hidden Markov model speech recognition model and a hybrid connectionist temporal classification (CTC)-attention network, which improved the accuracy of speech recognition. Kheddar et al. (2024) believed that deep learning poses a significant challenge for ASR, and enabling adaptive systems can improve ASR performance in dynamic environments. ASR technology provides a foundation for MOOC teaching.

Existing MOOC education is mainly combined with intelligent algorithms. Among them, Zhang (2023) designed an English MOOC question-answering system based on intelligent algorithms, which uses a combination of web pages and local question-answering libraries to retrieve, organise, and manage questions and answers, playing a specific auxiliary role in teachers' classroom teaching. Chen and Li (2023) used edge computing to build an English MOOC online teaching platform, solving the network load problem caused by large data transmission in online teaching systems. Fu et al. (2025) studied the speech recognition algorithm of the interactive artificial intelligence system, applied the speech recognition algorithm based on the interactive artificial intelligence system to the English video teaching system, and built a system with strong recognition ability and interactivity. In order to shorten the integration time of English teaching resources, Wang and Bi (2024) proposed a MOOC English online and offline hybrid teaching resource integration model based on a convolutional neural network to achieve intelligent integration of teaching resources. These studies provide a research basis for this paper. In order to optimise the problems in MOOC teaching, this paper uses adaptive parameter technology to construct a MOOC teaching system.

3 Research method

3.1 Design of MOOC English teaching system

In order to effectively demonstrate the performance of the English MOOC teaching system, a display platform for the English MOOC teaching system is designed (Huang, 2024). The system includes user information management, a dataset display, and a user feedback mechanism. MySQL is used as a database to store user data, student course selection records, and other related information. Regarding front-end display, HTML, Echarts, and ElementUI are used to build the user interface, while the back-end functional logic is implemented through the Flask framework.

3.1.1 I/O module design

The system uses the audio record class method to record the speech signal and, simultaneously, chooses the Audio Track class method to play the corresponding speech signal.

3.1.2 Scoring module design

The Shared Preferences component generates an XML file in the same Android package directory. The file stores several key-value pairs. The stored data can exist forever if the file does not disappear. This storage method can ensure the permanent storage of function parameters. Only one adaptive parameter generation process is needed for the system, and this scoring function can be used permanently for pronunciation scoring.

3.1.3 Realisation of word pronunciation practice

Clicking the 'view formant diagram' button in the pronunciation score box may cause the system to display a comparison chart between the learner's pronunciation and the standard pronunciation to give him a more detailed understanding of pronunciation.

3.1.4 User feedback mechanism

The user feedback mechanism is designed to collect users' opinions and suggestions on the ASR system in various ways to improve and optimise the system continuously. Users submit suggestions on speech recognition accuracy, operating interface, and user experience through the system's built-in feedback window. At the same time, the system also regularly collects users' usage experience and demand changes through questionnaires. Feedback data is used to adjust the system's algorithms, interface design, and function settings to ensure that the system better meets user needs and improves the system's intelligence and personalisation level. The system can be continuously improved through continuous user feedback and data analysis to provide more accurate and efficient speech recognition and pronunciation scoring services.

3.2 Pronunciation quality evaluation

An analysis of the features and capabilities of the Android platform's English pronunciation training system leads to the conclusion that the system's pronunciation scoring algorithm must be highly accurate and dependable to assess learners' pronunciation performance more precisely (Dai, 2024). The intelligent English pronunciation training system developed on the Android platform can use sound and animation to enhance learners' pronunciation of the English language while meeting system and real-time computing needs. This enables students to get feedback on their pronunciation and make any corrections (Kennedy et al., 2023). It is required to achieve real-time pronunciation scoring of vowel phonetic symbols and words. It simply calls speech to text (STT) to convert the user's voice into text and then compares it. For example, if the original one is Text, the user gives a high score if the text identified by STT is text, a normal score if it is test, a low score if it is Texas/Taxi, and a zero score if it is identified as a word that has no similarity at all or cannot be identified at all. The scoring mechanism in the pronunciation quality evaluation system primarily examines the

mapping relationship between matching distance and pronunciation score. It provides a way to calculate the relationship between the two.

$$\mathrm{d} = \frac{D(N, M)}{N} \tag{1}$$

Among them, N is the frame length of the test template. Take m1 and m2 as the two focal points of the ellipse, and use d_{12} as the distance between the two focal points of the ellipse. When m3 is located outside the ellipse, the average matching distance of the frame of pronunciation:

$$\mathrm{d} = \left(d_{13} + d_{23}\right)/3 \tag{2}$$

When m3 is located inside the ellipse, these three pronunciations are already very similar, but d cannot be infinitely smaller due to the limitation of d_{12}. At this time, take the minimum of d_{13} and d_{23} as the average frame matching distance:

$$\mathrm{d} = \min\left(d_{13} + d_{23}\right) \tag{3}$$

The distance must be converted into a pronunciation score after using the dual-standard pronunciation reference template to determine the average matching distance of the learner's pronunciation. The solution determined through experiments is as follows: when $d = d_{12}$, set the score to 90 points; when $d > d_{12}$, the score decays exponentially, and the parameters of the decay rate are adjustable; when $d < d_{12}$, d has a linear relationship with the score.

The scoring method based on the single reference template is simpler and requires less computation than the scoring method based on the dual reference template. Pattern matching often uses mathematical operations to translate the average matching distance of frames into pronunciation scores. Given that the matching similarity and the pronunciation score are inversely proportional, the following conversion procedure can be used if the pronunciation score range is set to 0–100 points:

$$score = \frac{100}{1 + a(d)^b} \tag{4}$$

The system can translate d, the average frame matching distance among them, into a score between 0 and 100. The link between the expert experience score and the matching distance can be used to derive the scoring criteria a and b. Once the scoring parameters have been established, this computation can fairly match the pronunciation scores to the score range of 0–100. The weighted sum of the scores for each feature parameter serves as the actual scoring formula when the scoring system incorporates numerous feature parameters:

$$score = w \cdot \frac{100}{1 + a_1\left(d_1\right)^{b_1}} + w \cdot \frac{100}{1 + a_2\left(d_2\right)^{b_2}} + \cdots + w \cdot \frac{100}{1 + a_n\left(d_n\right)^{b_n}} \tag{5}$$

w is the weight of each characteristic parameter, and a, b is the corresponding conversion parameter. Research has indicated that the most effective and significant impact on pronunciation scoring is attributed to MFCC feature parameters. As a feature extraction algorithm that simulates the hearing characteristics of the human ear, MFCC has been

widely used in speech signal processing. It converts the speech signal into a spectrum, and then performs logarithmic operations and discrete cosine transforms on the Mel scale to obtain coefficients that can reflect speech characteristics. These coefficients can well reflect speech's short-term power spectrum characteristics (Rai et al., 2024). This system chooses and employs the most efficient MFCC parameter as the sole characteristic parameter for scoring in order to reduce computation and enhance real-time performance.

The single-reference template scoring approach is enhanced by the AP-based scoring method. The AP-based scoring algorithm can automatically optimise the scoring model according to data characteristics and user performance. It introduces a continuous learning feedback mechanism to make the scoring more accurate and fair. This method can dynamically adjust according to individual differences to ensure that the evaluation results reflect the actual situation and adapt to environmental factors to achieve a more personalised and accurate scoring strategy. The definition of the AP-based scoring algorithm in this context is:

$$score = \frac{100}{1 + x(d)^{Y}} \tag{6}$$

In this way, for each pronunciation, the MFCC frame matching distance and the corresponding expert score are one-to-one correspondence. Suppose the set of average matching distances of all MFCC frames for training pronunciation is:

$$A = \{d1, d2, \ldots di, \ldots dn\} \tag{7}$$

Thus, n pairs of data of frame matching distance and expert score are obtained, and they satisfy the following relationship:

$$s_1 = \frac{100}{1 + x(d_1)^{y}} \tag{8}$$

$$s_2 = \frac{100}{1 + x(d_2)^{y}} \tag{9}$$

$$s_n = \frac{100}{1 + x(d_n)^{y}} \tag{10}$$

The least squares curve fitting approach can be used to determine the parameters x and y and the optimal values for x and y. The system can optimise the scoring model for different environmental conditions by adjusting parameters in real-time. It can dynamically adjust the scoring criteria according to the user's pronunciation characteristics to provide a fairer and more accurate evaluation. Combined with environmental data, it can automatically adapt to noise interference to ensure the stability and reliability of the scoring results (Bachiri et al., 2024). Theoretically, the fitting function obtained is more accurate with greater sample space. The pronunciation scoring module and the scoring parameter generating module are located on the same mobile phone device. The parameters for the pronunciation scoring computation are created using expert scoring training before pronunciation learning to reflect the features of the current system hardware platform (Yang, 2022). This leads to high closeness between the scoring results and expert experience scores. Firstly, a single reference template scoring

mechanism in the AP-based method in the pronunciation scoring module is introduced to estimate the corpus. This paper presents an AP-based scoring method, which improves the single-reference template scoring method. The system mainly trains and evaluates algorithms using high-quality datasets (Tieying, 2023). The dataset contains rich and diverse pronunciation samples, covering different speech characteristics, accents, and pronunciation errors to improve the algorithm's generalisation ability. Then, appropriate features are selected to model the pronunciation, using pitch, pitch, rhythm, and phoneme accuracy to describe the user's pronunciation. According to the specific situation, features can be processed through dimensionality reduction, standardisation, etc. to extract more informative features, thereby ensuring high reliability and accuracy of the pronunciation scoring algorithm. In addition, the proposed method can adaptively control the trade-off between noise reduction and speech distortion and accurately estimate the noise by utilising the information provided based on the AP. The intelligent English pronunciation training system based on Android can achieve English pronunciation training through animation and sound, give feedback to learners' pronunciation, correct pronunciation, and improve their English pronunciation level (Wu et al., 2022). The system designed in this study also adds the function of speech input and output and acquisition. Users mainly use voice input and output to read sentences and record. Users can also play this recording to hear their spoken reading level. Reading content is also extended from local files to network resources. There are also various reading methods to choose from, including SD card content reading and Web content reading. In addition, the system provides different training modes, such as follow-up mode, dialogue mode, listening comprehension mode, etc. to meet users' needs in different scenarios. Users can choose the appropriate mode based on their reading goals and environment. Implementing the system would be more convenient for users who can read English anytime and anywhere. Therefore, the AP-based method has strong adaptability, high accuracy, and reliability while significantly improving the system's compatibility. The scoring method, which is based on AP, mainly provides personalised scoring standards for each user based on pronunciation characteristics and voice data. The system can adjust scoring parameters based on users' pronunciation habits and accents, making the scoring more accurate and objective. In addition, the adaptive parameter scoring method comprehensively considers multiple pronunciation features and speech data, allowing the system to analyse multiple aspects of speech, such as pitch, pitch, rhythm, and phoneme accuracy, and more comprehensively evaluate users' pronunciation levels.

4 Experimental design

4.1 Experimental steps

The primary purpose of this experiment is to verify the effectiveness of the adaptive parameter scoring method in the ASR system and the performance of the system in different usage scenarios.

After the functional module test is completed, the experiment begins to test the system performance, simulating multi-user concurrent access and using software simulation tools to gradually increase the number of concurrent users, ranging from 10 to 200, to test the response time and stability of the system under high concurrency conditions. The English pronunciation data in the Mozilla open source speech dataset is

used to test for the speech recognition part. The audio data is first pre-processed in the experiment, including endpoint detection, signal box selection, window function windowing, feature extraction, and other steps. Then, the extracted feature data is input into the system based on the adaptive parameter scoring method for scoring and compared with the scoring results of the traditional English pronunciation training system. During the test, this paper focuses on evaluating the similarity between the scores of the two systems and the manual scores of experts, records the score differences of each test sample, and further analyses the advantages of the adaptive parameter scoring method in terms of accuracy, reliability, and compatibility.

4.2 Data collection

During the data collection process, this paper relies on automatic data acquisition and recording tools combined with the system's log function and external test equipment. For concurrent access testing, simulation tools are used to automatically simulate user behaviour, record the response time, request processing time, and system denial of service rate at each concurrency level, and ensure the system's performance can be thoroughly evaluated. For voice scoring experiments, all voice inputs and scoring results are recorded in real-time and stored in the database for subsequent statistical analysis. The experiment compares the similarity between the system's automatic scoring and the expert's manual scoring and uses statistical analysis tools such as SPSS to analyse the test results and calculate the mean, standard deviation, and similarity percentage of the score. In the experiment, the processing results of the pronunciation formant and speech signal are collected in real-time to ensure the system can accurately evaluate every detail of the user's pronunciation. The collected data ultimately provides a basis for subsequent system optimisation and parameter adjustment.

5 Results

5.1 System performance and functions

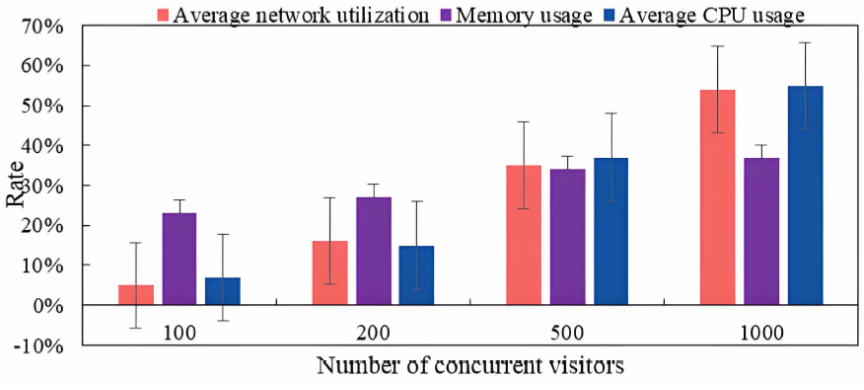
It is necessary to test the system's specific performance after the functional module test is finished. To mimic users' access to the system, software is applied. In order to determine whether the system can handle user demands, the software is used to simulate increased concurrent access and conduct virtual visits during testing. One crucial measure of system performance is the server's capacity to handle user demands at high concurrency. The purpose of adding concurrent access through simulation in system performance testing is that by simulating hundreds or thousands of users, repeatedly executing and running tests, one can identify performance bottlenecks and optimise and tune the application to find them. The concurrent test results are shown in Table 1.

Table 1 Concurrent test results

Number of customers	*10*	*50*	*100*	*200*
Average response time	0.035	0.792	1.053	2.531
Denial of service rate	0%	0%	1.2%	2.7%

Figure 1 displays the results of the performance test. It has been demonstrated through testing that this system can still meet the fundamental demands of users in high concurrency situations and accomplish the aim, even when the number of users gradually increases along with the response time. The system can meet the fundamental performance requirements according to the test.

Figure 1 Performance test results (see online version for colours)



In order to evaluate the stability of the core functions (speech recognition and scoring) of the MOOC system when users access it at high concurrency, this study simulated the calls to the system API with different numbers of concurrent users (50 to 500) through a stress testing tool. The test indicators include the API response success rate and average latency. The results are shown in Table 2.

Table 2 Comparison of speech recognition API performance in high concurrency scenarios

Number of concurrent users	*API response success rate (%)*	*Average latency (ms)*	*Server resource utilisation (CPU/RAM)*
50	99.8	238	32%/45%
200	98.1	352	68%/72%
500	85.4	612	93%/89%

In the scenario of 50 concurrent users, the system performed well, with an API response success rate of nearly 100% and a stable latency of 238 ms, which met the expected performance. When the number of concurrent users increased to 200, the success rate dropped to 98.1%, the latency increased to 352 ms, and the server resource usage increased significantly (CPU 68%). However, it is still within the controllable range. When there are 500 concurrent users, the system faced a resource bottleneck (CPU 93%), causing the success rate to drop sharply to 85.4%. This result shows that the current hardware configuration needs to be further optimised or a load-balancing strategy needs to be introduced to support larger-scale services.

To verify the system's resource loading capabilities under different network environments, this study simulated three scenarios: 5G (20 Mbps), Wi-Fi (100 Mbps), and low bandwidth (2 Mbps) to test the loading time and success rate of course resources (audio, video, and text), the results are shown in Table 3.

Resource loading efficiency is optimal in Wi-Fi environments, with a video loading time of only 2.1 seconds, which meets the MOOC platform's requirements for fluency. The system adapts to high-bandwidth scenarios through dynamic compression technology. In low-bandwidth environments, video loading time increases significantly (12.6 seconds), and the overall success rate is only 76.3%. This problem requires improving user experience by optimising cache strategies or providing offline modes. Text loading is stable in all scenarios (< 1.1 seconds), indicating that the system's shared preferences component effectively reduces the basic resource load.

Table 3 Resource loading performance in multiple network environments

Network type	*Audio loading time (s)*	*Video loading time (s)*	*Text loading time (s)*	*Overall success rate (%)*
5G (20 Mbps)	1.2	3.8	0.4	99.5
Wi-Fi (100 Mbps)	0.7	2.1	0.3	99.9
Low bandwidth (2 Mbps)	4.5	12.6	1.1	76.3

5.2 Adaptive parameter effects

To verify the effectiveness of adaptive parameter-based methods in improving system compatibility, accuracy, and reliability, this paper compares them with existing English pronunciation training systems on the market. This paper uses the Mozilla Voice Dataset as the test set. The Mozilla Open Source Voice Dataset is an open-source, multilingual voice dataset that contains 1488 hours of English voice data. This paper first pre-emphasises, frames, and windows the pronunciation and standard reference pronunciation in the test set and pre-processes endpoint detection. Then, feature extraction and pattern matching calculations are performed. Finally, the pronunciation training system based on the adaptive parameter scoring method and the traditional English pronunciation training system is used to score the pronunciation in the test set automatically. The similarity between the scoring results of the two types of systems and expert manual scoring is compared to verify the role of the adaptive parameter scoring method in improving the system's compatibility, accuracy, and reliability. The composition of the test set case data is shown in Table 4.

The test case data in this paper mainly includes vowel phonemes, consonant phonemes, pitch, tone, rhythm, syllables, linking, and stress. The similarity comparison between the scoring results of the two systems and the expert manual scoring is shown in Figure 2.

Figure 2 presents that the adaptive parameter scoring method has an ideal effect in improving the system's compatibility, accuracy, and reliability. In this system, the highest similarity between the adaptive parameter scoring results and expert manual scoring is 96.1%, with an average similarity of 92.5%. The similarity between the traditional system scoring results and expert manual scoring is only the highest, with an average similarity of 81.1%. The comparison results show that the scoring method based on AP has higher accuracy and reliability. By comprehensively considering multiple pronunciation features and speech data, an adaptive parameter scoring method can effectively improve the reliability of English pronunciation systems. Based on analysing the pitch, pitch, rhythm, and phoneme accuracy and accurately extracting feature information of speech, the system's compatibility, accuracy, and reliability have been

effectively improved by adjusting scoring parameters based on the user's pronunciation characteristics and speech data through AP.

The comparison results with four similar systems are shown in Table 5.

Table 4 Test set case data

Sequence	*Task*	*Number*
1	Vowel phoneme	20
2	Consonant phoneme	28
3	Pitch	7
4	Tone	2
6	Rhythm	5
7	Syllable	200
8	Connected reading	50
9	Stress	15

Figure 2 Similarity comparison results (see online version for colours)

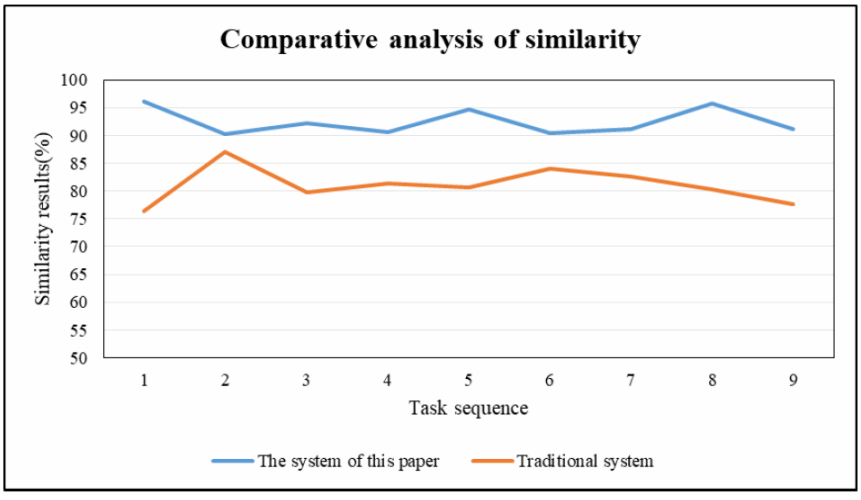


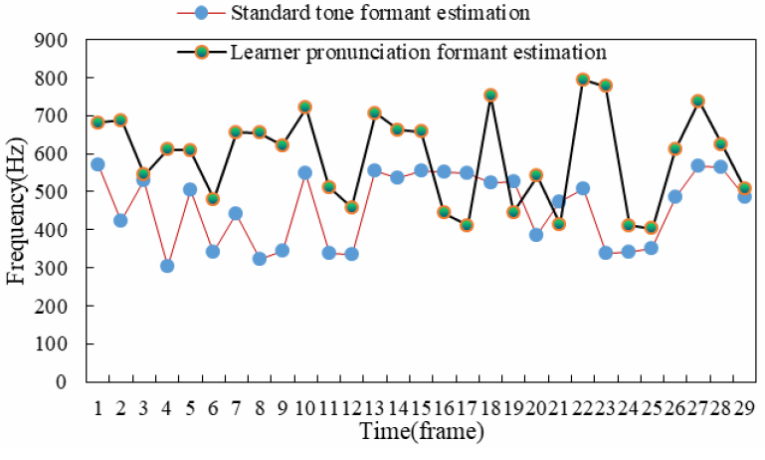
Table 5 shows that the pronunciation training system of the adaptive parameter scoring method has a significant advantage in the similarity with the expert manual scoring, reaching the best, with the highest similarity reaching 96.10% and the average similarity being 92.50%. The highest similarity of the Google speech recognition system is 90.20%; the highest similarity of the iSpeech pronunciation training system is 85.30%; and the similarities of the SpeechAce pronunciation scoring system and the Rosetta Stone pronunciation training system are 89.70% and 88.40% respectively. This shows that the adaptive parameter scoring method has advantages in the accuracy, reliability and compatibility of pronunciation scoring. When dealing with the pronunciation characteristics of different users, it can dynamically adjust the scoring parameters, provide more accurate feedback, and significantly improve the scoring accuracy and reliability of the system.

Table 5 Comparison results with four similar systems

System	*Highest similarity (%)*	*Average similarity (%)*
Adaptive parameter scoring method pronunciation training system	96.10%	92.50%
Google speech recognition system	90.20%	86.70%
iSpeech pronunciation training system	85.30%	82.10%
SpeechAce pronunciation scoring system	89.70%	85.40%
Rosetta Stone pronunciation training system	88.40%	83.20%

During processing, the system divides the voice input into several frames, each of which is only 20–30 ms long and can be viewed as a brief constant signal. The shape of the mouth and the position of the tongue are thought to be relatively constant during each speech frame. Thus, each frame's pronunciation formant reflects the current pronunciation's mouth shape and tongue location. The changing mouth shape and tongue location characteristics throughout time are likewise reflected in the altered corresponding formants between successive frames. Therefore, the system uses the pronunciation formant contrast on the Android platform to reflect the mouth shape changes of the entire pronunciation, as shown in Figure 3. The horizontal axis is time (frame), and the vertical axis is frequency (Hz). Each point on the broken line represents the frequency of the pronunciation formant at this time. The entire broken line changes with time, reflecting the change of the formant of the speech and, thus, the movement of the mouth shape and tongue position during pronunciation. The system uses blue and red broken lines in the figure to represent the formant fluctuations of the standard reference pronunciation and the learner's test pronunciation respectively. The coincidence of the two broken lines also reflects, to a certain extent, the mouth shape similarity between the standard pronunciation and the learner's pronunciation. Meanwhile, it is possible to have an intuitive understanding of the pronunciation time of the two.

Figure 3 Comparison of pronunciation formants (see online version for colours)



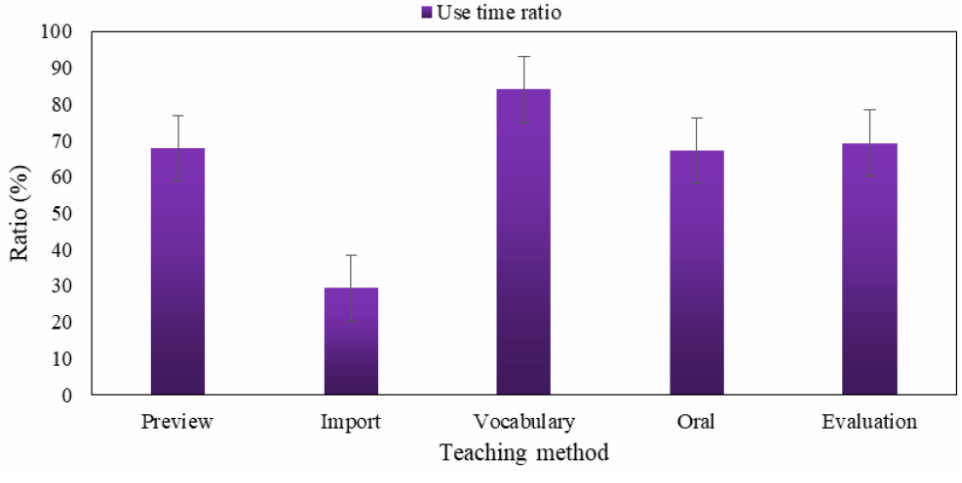
5.3 Role of intelligent speech technology in English teaching

In order to more intuitively reflect the role of intelligent speech technology in English teaching, the paper took four English class hours as an example (40 minutes per class hour, 160 minutes in total for four class hours). In recording classroom teaching activities, the intelligence is also recorded in detail. The statistical results of the frequency and duration of speech technology used in the course are shown in Table 6.

Table 6 Statistical results

Teaching link	*Cumulative time*	*Cumulative time of using intelligent voice technology*	*Cumulative use of intelligent voice technology*	*Intelligent voice technology use time ratio*
Preview	12 m 48 s	8 m 41 s	4	67.839%
Import	9 m 42 s	2 m 51 s	2	29.381%
Vocabulary teaching	38 m 34 s	32 m 25 s	42	84.054%
Sentence pattern teaching	19 m 2 s	9 m 33 s	8	50.175%
Oral communication teaching	28 m 40 s	19 m 16 s	12	67.209%
Review and consolidate	36 m 41 s	20 m 18 s	23	55.338%

Figure 4 Different teaching methods (see online version for colours)



In the classroom teaching process, teachers use intelligent voice technology in almost every link, and use intelligent voice technology to evaluate students' pronunciation in the classroom evaluation. Among them, vocabulary teaching is used the most frequently, 42 times in total for the four class hours, mainly using intelligent speech technology tools and the 'personalised reading pen card' made by the teacher to lead the teaching of vocabulary. Figure 4 displays that the use time ratio of intelligent speech technology in word teaching and evaluation is relatively high, reaching 84.05% and 69.30%

respectively. In the evaluation process, in addition to allowing students to evaluate themselves and others, teachers also ask students to read words into the microphone and use intelligent voice technology tools to evaluate the students' pronunciation.

5.4 System comprehensive performance evaluation

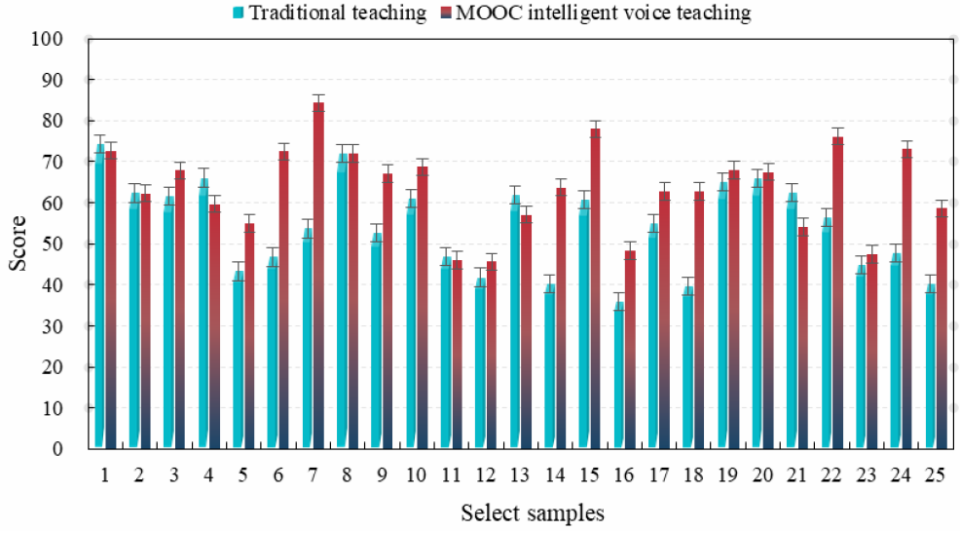
The paper took a class paper test for students in two classes. After participating in the batching process, the experimental and control classes' test results are collected. After SPSS data analysis, the average post-experiment scores of the two classes are shown in Table 7.

Table 7 The average post-experiment scores of the two classes

Class	*Mean*	*N*	*Standard deviation*
A	86.79	62	10.3816
B	82.86	65	11.365

The comparative analysis of the results is shown in Figure 5. The results of the paired design sample t-test for the pre-test and post-test scores of the experimental class are $t = -8.764$, the degrees of freedom df = 51, and the two-sided test P value (Sig.) = 0.000 < 0.005, indicating that there are significant differences in the test scores before and after the experimental class; control The pre-test and post-test scores of the paired design sample t-test result is $t = 1.646$, the degree of freedom df = 50, and the two-sided test P value (Sig.) = 0.106 > 0.005, indicating that although the pre-and post-test scores of the control class have improved, there is no significant differences, the above experimental results show that compared with the traditional English teaching mode, the MOOC smart phonetic English teaching mode can improve students' English performance as a whole.

Figure 5 Comparative analysis of results (see online version for colours)



In the performance evaluation section, in order to comprehensively evaluate this system's performance, this paper adds evaluation indicators such as speech recognition speed, noise resistance and user satisfaction to reflect the performance of the system in different practical application scenarios more comprehensively. The speech recognition speed reflects the system's efficiency in processing user input, the noise resistance test verifies the stability and reliability of the system in a noisy environment, and user satisfaction measures the system's acceptance in actual use through questionnaires and feedback data. The comprehensive performance evaluation results are shown in Table 8.

Table 8 Comprehensive performance evaluation results

System	*Speech recognition speed (ms)*	*Noise immunity (accuracy %)*	*User satisfaction (points)*
Adaptive parameter scoring method pronunciation training system	252	95.40	8
Google speech recognition system	303	88.70	7
iSpeech pronunciation training system	355	83.20	4
SpeechAce pronunciation scoring system	409	85.60	5
Rosetta Stone pronunciation training system	388	82.10	3

Table 8 presents that the pronunciation training system of the adaptive parameter scoring method performs well in all evaluation dimensions. Its speech recognition speed is 252 milliseconds, significantly better than other similar systems and shows high processing efficiency. In the noise resistance test, the recognition accuracy of this system is as high as 95.40%, far exceeding other systems, proving that it can still work stably in complex noise environments and ensure high-quality pronunciation scoring. In terms of user satisfaction, this system received a high evaluation of eight points, reflecting users' good experience and recognition during use. Overall, the advantages of this system in terms of speech recognition speed, noise resistance, and user satisfaction make it more adaptable and reliable in actual application scenarios.

In order to expand the application scope of the pronunciation training system of the adaptive parameter scoring method, this paper verifies the performance of multiple languages such as Chinese, French, German, and Spanish. The verification performance of different languages is shown in Table 9.

Table 9 Verification performance of different languages

Languages	*Speech recognition speed (ms)*	*Noise immunity (accuracy %)*	*User satisfaction (points)*
English	252	95.40	8
Chinese	265	94.20	7
French	280	93.10	7
German	290	92.80	6
Spanish	275	94.50	7

In Table 9, the system performs best in English, with a speech recognition speed of 252 milliseconds, a noise immunity accuracy of 95.40%, and a user satisfaction score of eight points, showing excellent performance. The performance of Chinese, French, German, and Spanish is also very close, with speech recognition speeds and noise immunity accuracy rates within acceptable ranges of 265 milliseconds, 280 milliseconds, 290 milliseconds, and 275 milliseconds, respectively, and noise immunity accuracy rates of 94.20%, 93.10%, 92.80%, and 94.50%, respectively. User satisfaction is generally seven points, indicating that the system in this paper has strong cross-language adaptability, can run stably in a multilingual environment, and provides users with a high-quality pronunciation training experience.

6 Discussion

The intelligent speech recognition system based on AP proposed in this study has shown significant advantages in the English MOOC teaching scenario. Experimental results show that the adaptive scoring method of the system is superior to the traditional system in the accuracy and reliability of pronunciation evaluation, with the highest similarity, with expert scoring reaching 96.1% and the average similarity being 92.5%. This advantage stems from the dynamic adjustment ability of the AP algorithm to multi-dimensional speech features (such as fundamental frequency, rhythm, and phoneme accuracy). In addition, the system's multilingual adaptability verification further proves its technical framework's versatility and provides a potential solution for cross-language education scenarios.

However, there are still challenges in practical applications. The current system has limited integration of deep learning algorithms and mainly relies on traditional signal processing and expert experience models. In the future, neural networks must be introduced to improve the robustness of complex semantic understanding. From the perspective of educational practice, the system significantly improves students' participation and learning efficiency through real-time feedback and personalised scoring mechanisms. The average score of the post-test of the experimental class verifies the positive impact of intelligent voice technology on English oral teaching. It is worth noting that the frequency of use of intelligent voice technology in vocabulary teaching accounts for as high as 84.05%, indicating its core role in basic language skills training. This technology-driven teaching model reduces traditional classrooms' dependence on teacher resources and provides students in remote areas with equal access to high-quality educational resources.

7 Conclusions

The system designed in this research also adds voice input, output, and collection functions. Voice input and output are mainly used by users to read and record sentences. Users can also play this recording to listen to their oral reading level. After joining the evaluation system in the future, it is convenient for learners to find mistakes and continuously improve. The system provides users an interactive interface for reading English text before the oral assessment. The feature of this system is that it is suitable for all kinds of people who love English. The reading content is also extended from local

files to network resources. There are also various reading methods to choose from, including SD card content reading and web content reading. Implementing the system is more convenient for users to use, and it allows them to read English anytime and anywhere. There is much room for development and improvement of the system in the future, and of course, there are also many challenges. After joining the evaluation system, it is bound to greatly help learners in English reading and speaking. At the same time, the interactivity of the interface should be more humane, and the functions should be more prominent. Moreover, applying more advanced artificial intelligence algorithms, such as deep learning and neural networks, in the speech recognition module further improves the accuracy and robustness of speech recognition. The application scenarios of ASR technology in the English MOOC teaching system can be further expanded to online tutoring and self-study platforms, providing personalised pronunciation feedback and real-time correction to help learners improve their oral English skills anytime and anywhere. This expansion direction not only helps to improve learning efficiency but also expands its commercial value, promotes the innovative development of the educational technology industry, and provides equal and convenient learning opportunities for most English learners, which has a far-reaching social impact.

Acknowledgements

This work was supported by quality project of Anhui Province in 2022 – Curriculum Ideological and Political Demonstration Course (2021kcszsfkc187) and Anhui Xinhua College High Talent Training Project (bs2025kyqd045).

Declarations

There is no potential conflict of interest in our paper and all authors have seen the manuscript and approved to submit to your journal. We confirm that the content of the manuscript has not been published or submitted for publication elsewhere.

References

Aldarmaki, H., Ullah, A., Ram, S. and Zaki, N. (2022) 'Unsupervised automatic speech recognition: a review', *Speech Communication*, April, Vol. 139, pp.76–91, https://doi.org/10.1016/j.specom.2022.02.005.

Bachiri, Y.A., Mouncif, H., Bouikhalene, B. and Hamzaoui, R. (2024) 'Integrating AI-based speech recognition technology to enhance reading assessments within Morocco's Tarl Program', *Turkish Online Journal of Distance Education*, Vol. 25, No. 4, pp.1–15.

Chen, X. and Li, H. (2023) 'Speech detection in English MOOC online teaching platform based on edge calculation', *International Journal of System Assurance Engineering and Management*, pp.1–8, https://doi.org/10.1007/s13198-023-02029-5.

Dai, R.Q. (2024) 'Overcoming cross-language communication barriers with speech recognition technology', *Applied and Computational Engineering*, Vol. 74, pp.53–58.

Dhanjal, A.S. and Singh, W. (2024) 'A comprehensive survey on automatic speech recognition using neural networks', *Multimedia Tools and Applications*, Vol. 83, No. 8, pp.23367–23412.

Fu, M., Guan, X., Wang, Y. and Chen, Q. (2025) 'Application of speech recognition algorithm based on interactive artificial intelligence system in English video teaching system', *Entertainment Computing*, January, Vol. 52, p.100859, https://doi.org/10.1016/j.entcom.2024.100859Commented.

Huang, A. (2024) 'Speech recognition based on mobile biosensor networks and quality evaluation of university political education', *International Journal of High Speed Electronics and Systems*, p.2540128, https://doi.org/10.1142/S0129156425401287.

Kennedy, D., Halim, E., Condrobimo, A.R., Syamsuar, D. and Ferdianto, F. (2023) 'Overcoming language barriers in MOOCs with artificial intelligence: an AI-based approach for multilingual education', in *2023 Eighth International Conference on Informatics and Computing (ICIC)*, IEEE, pp.1–6.

Kheddar, H., Hemis, M. and Himeur, Y. (2024) 'Automatic speech recognition using advanced deep learning approaches: a survey', *Information Fusion*, September, Vol. 109, p.102422, https://doi.org/10.1016/j.inffus.2024.102422.

Mukhamadiyev, A., Khujayarov, I., Djuraev, O. and Cho, J. (2022) 'Automatic speech recognition method based on deep learning approaches for Uzbek language', *Sensors*, Vol. 22, No. 10, p.3683.

Oruh, J., Viriri, S. and Adegun, A. (2022) 'Long short-term memory recurrent neural network for automatic speech recognition', *IEEE Access*, Vol. 10, pp.30069–30079, https://doi.org/10.1109/ACCESS.2022.3159339.

Rai, A.K., Jaiswal, S.D. and Mukherjee, A. (2024) 'A deep dive into the disparity of word error rates across thousands of NPTEL MOOC videos', in *Proceedings of the International AAAI Conference on Web and Social Media*, May, Vol. 18, pp.1302–1314.

Sun, W. (2023) 'The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation', *Frontiers in Psychology*, Vol. 14, p.1210187, https://doi.org/10.3389/fpsyg.2023.1210187.

Tieying, Z. (2023) 'Application of android voice assistant based on parallel storage in multimedia English teaching platform', *Soft Computing*, pp.1–12, https://doi.org/10.1007/s00500-023-08539-5.

Wang, K. and Bi, D. (2024) 'Integrating MOOC online and offline English teaching resources based on convolutional neural network', *International Journal of Business Intelligence and Data Mining*, Vol. 25, Nos. 3–4, pp.271–291.

Wu, F., Chen, Y. and Han, D. (2022) 'Development countermeasures of college English education based on deep learning and artificial intelligence', *Mobile Information Systems*, Vol. 2022, No. 1, p.8389800.

Yang, Z. (2022) 'Natural language enhancement for English teaching using character-level recurrent neural network with back propagation neural network based classification by deep learning architectures', *Journal of Universal Computer Science (JUCS)*, Vol. 28, No. 9, p.984.

Zhang, Q. (2023) 'An English MOOC answering system based on intelligent algorithms', in *International Conference on Cognitive based Information Processing and Applications*, pp.223–232, Springer Nature Singapore, Singapore.