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

Exploring the Teaching Mode of Applying Artificial Intelligence to Enhance College English Listening and Speaking Ability

Chunneng Zhao[1,*]

[1] Hebei University of Technology, Langfang, Hebei, 065000, China

SUMMARY: *Aiming at the problem of improving university English listening and speaking ability, this paper designs an English learning system based on deep learning and artificial intelligence technology, which is used as the kernel of the English teaching model to explore the teaching effectiveness. The system assists students in listening and speaking ability training in the form of intelligent conversation, embedded with the improved EMD-FD speech signal feature extraction algorithm and the multi-parametric English pronunciation quality evaluation model (MPEPQE). The results show that the MFCC extraction algorithm combined with the EMD-FD extraction algorithm is able to extract the features of the high-frequency region of the speech signal more completely than the single MFCC extraction algorithm, which in turn improves the speech recognition rate. Meanwhile, compared with the CNN+LSTM scoring model, the MPEPQE scoring model shows better fitting effect and adaptive ability, and its three index values of Pearson correlation coefficient, accuracy and average score difference are 0.706, 84.52% and 0.613 respectively, which all achieve better results. In addition, a teaching model based on the AI English learning system was constructed, the use of which significantly improved students' English listening and speaking skills compared to the traditional teaching model (P<0.05). This study makes some necessary attempts and explorations for the full-scale promotion of the use of AI-enabled university English learning system in order to effectively use it to improve students' listening and speaking abilities.*

KEYWORDS: *artificial intelligence; English listening and speaking ability; pronunciation quality evaluation; EMD-FD; English learning system*

1 Introduction

With the vigorous development of “Internet Plus” and the gradual promotion of the new curriculum reform, the traditional English teaching mode is undergoing a profound change [1]. At present, some useful attempts have been made to apply artificial intelligence (AI) technology in the English subject. For example, more than 20 provinces in China have actively piloted the “human-computer dialogue” listening and speaking test mode in the English test, and the results of listening and speaking are included in the total score of the test instead of the results of listening, which highlights the importance and orientation of the combination of AI and listening and speaking ability test [2, 3]. At present, in the teaching field, affected by the diversity of teaching activities, the complexity of the teaching process and the mechanical nature of test-oriented education, the application of AI technology in the English language subject is still in its infancy, mainly in the higher education stage and some areas of individual primary and secondary schools, but has not yet formed a standard theoretical system and large-

*Y9874561238@163.com
https://doi.org/10.65102/is2026846
Published: 30 April, 2026

scale application [4-7]. 2024 implemented the reform of the English language teaching materials, the new version of English language textbook The unit content of the new version of English teaching materials is specially added to the special content of oral expression, which puts forward clear requirements for the cultivation of students' oral output ability and listening ability, and at the same time, it also puts forward higher standards for the teachers' listening and speaking teaching methods and teaching ability [8, 9]. And the traditional English listening and speaking teaching mode can't meet the problems of students' individualized needs, teachers' limited resources, the subjectivity of oral scoring, and the limitations of the language environment [10, 11]. Therefore, exploring the AI-driven teaching mode of college English listening and speaking ability can help improve the teaching effect and students' listening and speaking ability.

AI technology in English listening and speaking teaching can perform tasks such as error correction, assessment, and material generation to achieve personalized and precise teaching. Literature [12] developed an AI English speaking error correction system based on speech recognition and synthesis technology, which is a mechanism to improve learners' pronunciation accuracy through automatic assessment and targeted error correction, which can effectively improve learners' oral pronunciation level. Literature [13] findings emphasize that AI-assisted oral assessment eliminates the subjectivity of traditional assessment, effectively improves grammar, vocabulary, intonation and fluency, and enhances learners' willingness to communicate in English in different contexts. Literature [14] reported that AI-generated adaptive listening materials can meet learners' individual needs and emphasized the importance of instructional integration and teacher guidance in realizing their potential. Literature [15] designed indicators for assessing college students' English reading and listening abilities based on scenario-based teaching and introduced a back-propagation neural network improved by genetic algorithm to assess students' listening and reading abilities.

AI-driven innovative design of English speaking and listening teaching mode. Literature [16] constructed an AI-based interactive English speaking teaching platform, integrating an all-round speaking corpus, a teaching monitoring assistance system, and a scoring system, which improved students' speaking performance. Literature [17] personalized oral English teaching based on students' oral characteristics through natural language processing technology, which improved students' comprehensive oral ability and learning motivation, as well as their comprehension and participation. Literature [18] created an AI English adaptive learning system using recurrent neural networks and natural language processing to personalize teaching through sequence modeling and real-time feedback, which significantly improved students' engagement and oral clarity and grammatical correctness. Literature [19] created an AI-driven English role-playing instructional model that significantly optimizes student performance and reduces linguistic errors, while emphasizing its value in facilitating positive learning interactions through corrective feedback to improve students' oral proficiency. Literature [20] showed that the AI-driven English listening and speaking flipped classroom model enhanced students' academic performance significantly, improved their speaking skills and self-managed learning. Literature [21] examined the effectiveness of English listening flipped classroom based on AI and particle swarm algorithm, which combined with animated videos significantly improved students' performance and learning motivation, and close to 70% of the students showed a strong willingness to this teaching mode. Literature [22] provides a personalized learning environment, instant feedback, and classroom interaction into English listening teaching with the help of AI-based voice recognition, virtual tutor, and other tools, which significantly improves students' listening comprehension. Literature [23] proposed a strategy to improve English listening ability based on AI wireless network, which effectively improved

students' listening performance and learning interest, and students' listening ability was significantly improved under the learning of this strategy. In addition, literature [24] studied the role of AI in English listening and speaking teaching through meta-analysis, analyzed 19 empirical studies, pointed out that AI teaching has a significant effect on the improvement of speaking ability, while listening has a positive trend but not statistically significant, and emphasized the need to deepen the research on the impact of listening in the future. Literature [25] found that the innovative teaching modes of AI-based English IV listening material generation, human-computer interactive speaking training, PPT presentation and video production, etc., can enhance the learning interest and authenticity of expression, and improve students' listening and speaking ability, but they also reveal the problem of over-dependence that may lead to the problem, and they provide practical references for the digital transformation of teaching. The above research proves that AI technology can realize personalized and precise English listening and speaking teaching, which helps to improve the level of students' listening and speaking ability, but over-reliance should be avoided.

In this paper, a university English learning system with AI intelligent dialog as the core is designed by combining the improved EMD-FD feature extraction algorithm, MFCC extraction algorithm, deep learning speech recognition system and multi-parameter pronunciation quality evaluation model MPEPQE. In order to test the effectiveness of the algorithmic modules of the system, simulation experiments were conducted. Following this, the system was used in the design of university English teaching mode, and teaching control experiments were organized to analyze the superiority of the experimental group that experimented with the teaching mode supported by the system in improving students' English listening and speaking abilities compared with the control group that used the traditional teaching mode.

2 Design of English Learning System Combined with Artificial Intelligence Technology

In order to improve college students' English listening and speaking ability, this paper combines artificial intelligence technology to design an English learning system.

2.1 Principles of Speech Recognition and Evaluation

2.1.1 Speech signal preprocessing

The speech data acquired during the design process cannot be used directly as experimental data, this is due to the fact that the acquired speech data contains many impurities and redundant information. Therefore, the speech signal used for the experiment must be processed.

The purpose of speech signal preprocessing is to optimize each frequency band of the speech information, and to improve the speech resolution of the high and low frequency bands while ensuring that the spectrum of the speech signal is smooth and smooth. The pre-processing mainly includes three parts: pre-emphasis, windowing and end-point detection.

(1) Pre-emphasis

The high-frequency band signals of voice information are prone to significant attenuation, while the attenuation speed is faster, which makes the voice signal incomplete before it is put into use. In order to solve the problem, pre-emphasis processing is carried out on the speech signal, and the essence of the processing realized is the filtering processing, which can be expressed as follows:

$$H(z) = 1 - \alpha z^{-1}, 0.9 \le \alpha \le 1.0 \tag{1}$$

where α denotes the pre-emphasis coefficient, taking a value close to 1.

(2) Split-frame windowing

The processing of time-varying signals is of high difficulty, and speech signal is exactly an unstable time-varying signal. However, due to the special characteristics of human voice information generation, it can be processed by using the method of short-time analysis. The premise of speech signal processing is to segment the complete speech information into several smooth segments for subsequent processing, which is referred to as the frame-splitting operation. The premise of the frame-splitting operation is that the integrity of the speech information must be ensured. The split-frame processing of the speech signal is essentially the windowing process, which can be expressed as:

$$x_a(m) = x_a(m) * \omega(m) \tag{2}$$

There are various kinds of windowing functions corresponding to the windowing processing, and the functions used for the sub-frame processing include: Hamming window, Hanning window, and rectangular window. Since the spectrum needs to be analyzed in this study, the Hamming window with the smallest spectral leakage is chosen. The corresponding formula is:

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left[2\pi n / (N-1)\right] & 0 \le n \le N-1 \\ 0 & \text{Other} \end{cases} \tag{3}$$

where N denotes the sub-frame length.

(3) Spectral entropy method endpoint detection

There are many impurities in the acquired speech signal, including silent segments, noise segments and so on. This part of the useless information will increase the amount of data processing calculations, but also affect the effect of voice signal processing. Endpoint detection is one of the effective methods to solve this problem, and the principle of this method is to eliminate the useless information segments by locating the position of useful information.

In this study, the spectral entropy method is chosen for endpoint detection. Firstly, any frame of the speech signal data for denoising is transformed by FFT to obtain the spectral energy spectrum $Y_i(k)$ of the corresponding frequency component, and then the probability density function of the corresponding frequency component can be obtained:

$$p_i(k) = \frac{Y_i(k)}{\sum_{l=0}^{N/2} Y_i(l)} \tag{4}$$

where N is the FFT length.

The short-time spectral entropy of each speech frame is:

$$H_i = -\sum_{k=0}^{N/2} p_i(k) \log p_i(k) \tag{5}$$

The following relationships exist:

$$H(P) = H(p_1, p_2, \cdots, p_q) = H(1/q, 1/q, \cdots, 1/q) = \log q \tag{6}$$

where $P=(p_1, p_2, \cdots, p_q)$ is the q-dimensional vector.

In order to further strengthen the performance of the spectral entropy method, reasonable settings are made:

1) In order to make it possible to distinguish the effective speech segment and invalid speech segment more clearly in the process of endpoint detection, reasonable settings are made. The frequency range is set to 250Hz~3400Hz, if the frequency of any one spectral line is f_k, then there is a corresponding:

$$Y_i(k)=0\left(f_k<300\text{ Hz or }f_k>250\text{ Hz}\right) \tag{7}$$

2) Due to the low value of the spectral entropy corresponding to some except special noise, it is impossible to remove this part of the noise signal when using the spectral entropy method for denoising. The above problem can be solved by setting the normalized spectral probability density, i.e., the upper limit of the density is set to:

$$p_i(k)=0 \quad p_i(k)>0.9 \tag{8}$$

2.1.2 Speech signal feature extraction

(1) Basic feature extraction method

MFCC speech feature parameters are parameters that are changed by filtering and time domain, while the perceived frequency is also introduced in the transformation process, the function transformation relationship between the transformed parameters and the actual frequency is:

$$F_{Mel}=2595\times\log(1+f/700) \tag{9}$$

where F_{Mel} is the perceived frequency and f is the actual frequency.

The human ear has an innate ability to perceive sound signals, and if the ordinary frequency domain analysis can be replaced by perceptual frequency domain analysis in the process of speech feature extraction, the extraction accuracy will be greatly improved. The actual MFCC speech feature extraction flowchart is shown in Figure 1.

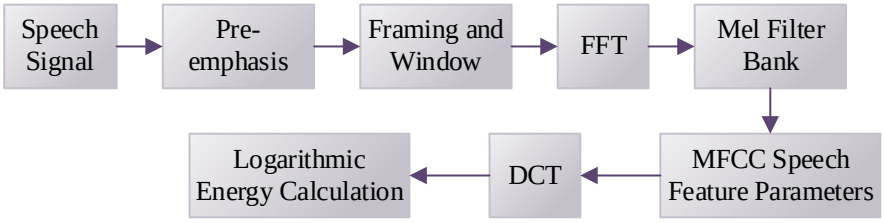

Figure 1: The extraction process of MFCC speech feature parameters

The L-order MFCC speech feature parameters are:

$$C_n=\sum_{k=1}^{M}\log x(k)\cos\left(\pi(k-0.5)n/M\right), n=1,2,\cdots,L \tag{10}$$

(2) Improved EMD-FD speech feature parameter extraction algorithm

When carrying out feature extraction of speech information, the basic MFCC speech feature

parameter extraction has shortcomings in the processing of high-frequency segment speech signals.

The traditional speech extraction algorithm cannot process the high-frequency segments of speech signals in a reasonable way, in order to solve this problem, the EMD-FD extraction algorithm is selected for speech feature extraction and further optimized.

The algorithm first carries out the collection of continuous dynamic change trajectories of speech feature vectors, and the collection process is realized by the feature differentiation method, which can be expressed as:

$$D\_Feature(j)_i = Feature(j)_i - Feature(j)_{i-1} \tag{11}$$

where $Feature$ denotes the EMD-FD speech feature parameter, $D\_Feature$ denotes the sequence of differential feature vectors, P is the feature order, and N is the number of feature vector sequences.

The feature vectors with different orders are reorganized with the feature order as the classification basis, and the optimized EMD-FD speech feature parameters can be expressed as:

$$EMD-FD = \begin{cases} Feature_i & i = 0,1,2,\cdots,P-1 \\ D\_Feature_{(i-P)} & i = P, P+1,\cdots,2P-1 \end{cases} \tag{12}$$

2.1.3 Language models

The essence of the language model as a probabilistic model is to find the joint probability of the characters $w_1, w_2, \ldots, w_t$ in this sentence, and Eq. (13) can be derived using Bayes' formula:

$$\begin{aligned} P(W) &= P(w_1, w_2, \cdots, w_t) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_t|w_1\cdots w_{t-1}) \end{aligned} \tag{13}$$

where $P(W)$ is the probability of the whole sentence. It can be shown that the joint probability can be expressed as the conditional probability of the occurrence of the tth word given the first $t-1$ words, and that the probability of the whole sentence can be obtained by multiplying them together.

And these conditional probabilities can also be obtained by simple computation. Assuming that a corpus of definite size is already available, computing the conditional probabilities $P(w_t|w_1\cdots w_{t-1})$ is simply a matter of calculating the $w_1\cdots w_{t-1}$ occurrences, and the number of times $w_1\cdots w_{t-1}w_t$ occurs, and dividing the two counts gives the conditional probability.

Assuming that the corpus contains V words, a conditional probability of a word appearing after about n word-length characters needs to be computed, then there are V^n free parameter variables to compute this conditional probability.

Assuming that a word is related to its first $n-1$ occurrences, n can take different integer values, usually $n = 1, 2, 3$.

Language models are usually trained using maximum likelihood estimation and finally normalized. The above three values are commonly used, and the most common case is when $n = 3$, which is the ternary language model.

Another problem that arises when training language models is data sparsity. Therefore, data smoothing is needed, and common smoothing methods include: additive smoothing, Good-

Turing smoothing, linear interpolation smoothing, Katz smoothing and so on.

In traditional speech recognition system, the results obtained from acoustic model and language model are decoded together, and an optimal path algorithm is used to select a path with the highest integrated probability as the final recognition result, which is output to the user.

2.1.4 Deep learning based speech recognition system

This section briefly introduces a deep learning network structure, LSTM, which is commonly used in the field of speech recognition.The principle of Long Short-Term Memory Network (LSTM) is very much like the habit of human beings to learn in normal times, where some things need to be memorized for a long time while some short-term memories will be forgotten. In LSTM, a neuron is replaced by a cellular CELL. in each CELL, the most critical is the cellular state C_t, which interacts with each cell in only one linear way, and therefore preserves the information very well.

LSTM has three gates: forgetting gate, input gate, and output gate, which determine and update the state of the cell. The so-called gates are a way to allow information to be selected to pass through, and they contain a sigmoid function as well as an arithmetic operator. By tuning the parameters, each gate can control how much of the input information can be output.

(1) Oblivion gates:

$$f_t = \sigma\left(W_f * \left[h_{t-1}, x_t\right] + b_f\right) \tag{14}$$

where f_t represents the output of the forgetting gate, and the inputs to the forgetting gate at moment t are the previous cell output h_{t-1}, and the input x_t at moment t, and the amount of this information passing through is controlled by the weights W_f and the bias b_f, i.e., the information in the cell state is selectively forgotten.

(2) Input gate:

$$\begin{cases} i_t = \sigma\left(W_i * \left[h_{t-1}, x_t\right] + b_i\right) \\ \tilde{C}_t = \tanh\left(W_c * \left[h_{t-1}, x_t\right] + b_c\right) \end{cases} \tag{15}$$

where i_t is the output of the input gate, h_{t-1} is the output of the previous cell, x_t is the input at moment t, W_i and b_i are the weights and biases of the sigmoid neuron, and W_c and b_c are the weights and biases of the tanh neuron, respectively.

The input gates need to decide what information to keep for the new input. First, the sigmoid neuron will compute the output of the input gate, which determines what values will be updated. Then, the tanh neuron computes and obtains the new candidate value $\tilde{C}_t$, and subsequently multiplies the two vectors $i_t * \tilde{C}_t$ to update into the cell state.

(3) Cell state update:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{16}$$

where C_t is the cell state at moment t after the update, f_t is the output of the forgetting gate, C_{t-1} is the cell state at moment $t-1$, and $i_t * \tilde{C}_t$ is the product of the output of the input gate and the candidate value.

When the information passes through the forgetting gate and the input gate, the cell state will update the cell state according to the results of the two gates, and the previous cell state C_{t-1} will be multiplied by the result of the forgetting gate, f_t, and added to the result of the input gate, $i_t * \tilde{C}_t$, to get the current t-moment cell state, C_t.

(4) Output gate:

$$\begin{aligned} o_t &= \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \\ h_t &= o_t * \tanh\left(C_t\right) \end{aligned} \tag{17}$$

where o_t is the output of the sigmoid neuron, and h_t is the output of that cell at moment t.

After updating the cell state, a decision needs to be made about what to output to the next cell in the chronology, and to the next layer of connected cells.

The properties of LSTM make it suitable for speech recognition tasks as it preserves the correlation between temporal sequences well.

2.1.5 Multi-parameter pronunciation quality evaluation

Taking college students' spoken English as the research object, this paper establishes the Multi-Parametric English Pronunciation Quality Evaluation Model (MPEPQE) for college students as shown in Fig. 2, based on the realization of the evaluation of multiple pronunciation quality indicators such as intonation, speech rate, rhythm and intonation, and comprehensively analyzing the relationship between the indicators, with an emphasis on considering the weights of the individual indicators in the evaluation of the overall pronunciation quality.

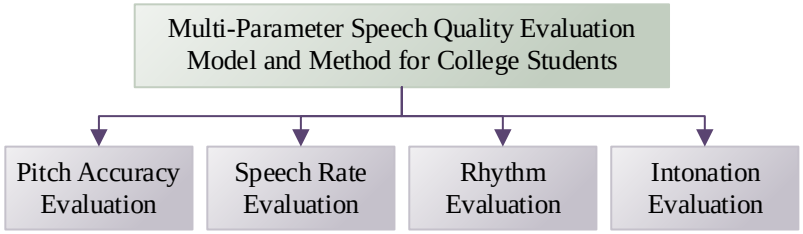

Figure 2: The multi-parameter pronunciation quality evaluation model for college students

(1) Pronunciation Evaluation

Pitch evaluation mainly examines whether the content information of the pronounced sentence is complete and accurate, whether the pronunciation is clear and fluent, and whether there are pronunciation errors. In this paper, MFCC coefficients based on the human ear hearing model are used as the evaluation parameters of pitch, and a speech recognition model is built through deep belief network for speech recognition to determine whether the content is complete and correct. At the same time, we calculate the correlation coefficient between the standard utterance and the MFCC features of the input utterance to determine whether the pronunciation is clear and fluent, and combine the two to provide pitch evaluation and feedback on the quality of English pronunciation. The process of pitch evaluation is shown in Figure 3.

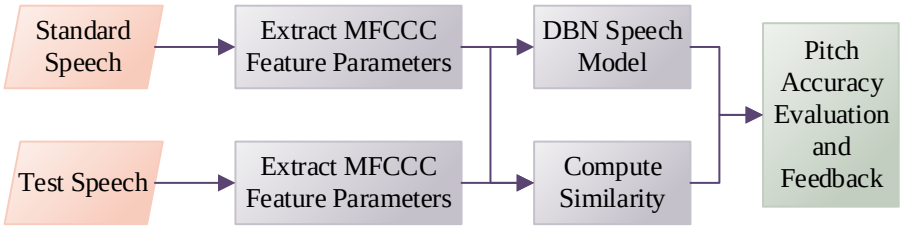

Figure 3: Pitch accuracy evaluation

(2) Speaking Speed Evaluation

The speed of speech usually refers to the speed of articulation, which is a measure of how fast or slow a speaker pronounces words, and can be reflected by calculating the number of syllables N he/she speaks in a unit of time T, which can be roughly measured in terms of the total duration of speech including pauses.

In this paper, we use speech speed evaluation based on speech duration by calculating the duration ratio φ between the test utterance and the standard utterance:

$$\varphi = \frac{Len_{Std}}{Len_{Test}} \tag{18}$$

where Len_{Std} is the duration of the standard utterance and Len_{Test} is the duration of the test utterance.

Further, φ is compared with the set utterance rate threshold, and the utterance rate evaluation process is shown in Figure 4. It is worth noting that the durations are preprocessed using the dual-threshold endpoint detection method of short-time energy and short-time average over-zero rate, which can effectively exclude the noise interference of the voiceless segment.

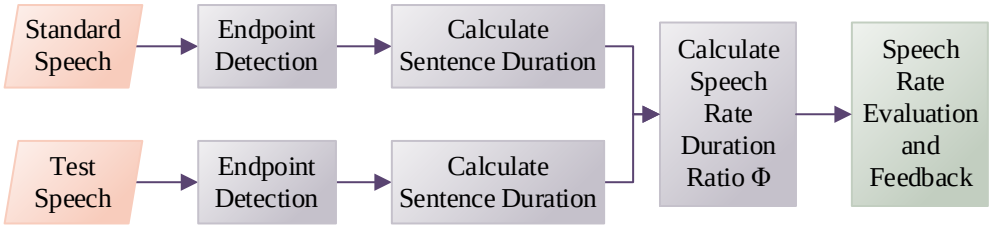

Figure 4: Evaluation of speaking speed

(3) Rhythm Evaluation

The rhythm of a language is the differences and similarities in the height, weight, length and slowness of the speech, and appears in regular alternating cycles with each other in some kind of class of speech unit fragments.

The rhythm evaluation mechanism is shown in Figure 5, which includes the following steps:

1) Extract the short-time energy value of speech to form a speech intensity graph. The loudness of stressed syllables in a sentence is directly reflected in the energy intensity in the time domain, i.e., the speech energy intensity of stressed syllables is large. According to the definition of short-time energy of speech signal $s(n)$:

$$E_n = \sum_{m=-\infty}^{\infty} \left[s(n)\omega(n-m) \right]^2 \tag{19}$$

Short-time energy values are extracted for speech sentences to form an intensity profile.

2) Regularize the sentences. In order to facilitate data processing and obtain more objective evaluation results, before evaluating the test utterances, the duration of the test utterances needs to be proportionally regularized to a degree similar to the standard utterances.

3) The improved dynamic time normalization (DTW) algorithm is used to calculate the intensity profile match between the standard utterances and the input utterances.

The basic principle of the DTW algorithm is dynamic time regularization, which matches the otherwise mismatched lengths of time between the test template and the reference template. The disadvantage of the traditional DTW algorithm is that when template matching is performed, all frames have the same weight and all templates must be matched, and the computation is relatively large, especially when the number of templates increases faster, the amount of arithmetic grows especially fast.

In this paper, the intersection points that need to be operated are limited to the parallelogram by setting the matching boundary. The R and T are divided into N and M frames in equal time, which can be divided into three paths $(1, X_a)$, (X_a+1, X_b), (X_b+1, N) to calculate the distances, and according to the coordinate calculation, we can get $X_a = \frac{1}{3}(2M-N)$ and $X_b = \frac{2}{3}(2N-M)$, X_a, X_b is taken as the closest integer. When the constraints $2M-N \geq 3$, $2N-M \geq 2$ are not satisfied, dynamic matching is not performed, which reduces the system expenditure.

The frame matching between each frame on X axis and $[y_{\min}, y_{\max}]$ on Y axis, $y_{\min}$, $y_{\max}$ are calculated as follows:

$$y_{\min} = \begin{cases} \frac{1}{2}x & x \in [0, X_b] \\ 2x + (M-2N) & x \in (X_b, N] \end{cases} \tag{20}$$

$$y_{\max} = \begin{cases} 2x & x \in [0, X_a] \\ \frac{1}{2}x + \left(M - \frac{1}{2}N\right) & x \in (X_a, N] \end{cases} \tag{21}$$

If $X_a > X_b$, the matched paths can be categorized as $(1, X_b)$, (X_b+1, X_a), (X_a+1, N). For each frame forward of X-axis, although the number of frames corresponding to Y-axis is different, the regularization property is the same, and the cumulative distance is:

$$D(x,y) = d(x,y) + \min \begin{Bmatrix} D(x-1, y) \\ D(x-1, y-1) \\ D(x-1, y-2) \end{Bmatrix} \tag{22}$$

where D and d denote the cumulative distance and frame matching distance, respectively.

(4) Setting the accent threshold and non-accent threshold as the double threshold of the feature as well as the rereading vowel duration for the division of the accent unit and determining the number of accented pronunciations.

In this paper, the double threshold comparison method is used for accent endpoint detection, and the thresholds are set after a large number of experimental verifications, as in Eqs. (23)~(24):

$$\text{Stressed syllable threshold } T_u = \left(\max\left(sig\_in\right) + \min\left(sig\_in\right)\right)/2.5 \tag{23}$$

$$\text{Unstressed syllable threshold } T_l = \left(\max\left(sig\_in\right) + \min\left(sig\_in\right)\right)/10 \tag{24}$$

In the double threshold comparison method, the maximum speech energy value $S_{\max}$ in the sentence which is greater than the stress threshold T_u is searched one by one according to the energy value of the sentence, and then the speech energy values S_l and S_r equal to the non-stress threshold T_l are searched to the left and right of $S_{\max}$, then the stress signal of the sentence can be set to S_l and S_r and at the same time the energy value between S_l and S_r is set to 0 to avoid repeated search between S_l and S_r. In this paper, the smallest unit of the stressed syllable unit is set as an approximate stressed vowel duration, which is 100ms.

Through the above steps, the division of sentence stress units is completed.

5) The improved dPVI parameter is used to calculate the respective rhythmic correlations between the standard utterance and the input utterance.

Using successive syllable unit segment durations to calculate the pairwise variation index (PVI) yields syllable duration variability and can be used as a measure of speech rhythmic correlation.

Based on the variability characteristics of English speech unit durations, this paper adopts the improved dPVI parameter calculation formula to compare the syllable unit fragment durations of standard and test utterances separately, and uses the converted parameters as the basis for systematic evaluation:

$$dPVI = 100 \times \left(\sum_{k=1}^{m-1} \left| d1_k - d2_k \right| + \left| d1_t - d2_t \right| \right) \Big/ Len \tag{25}$$

where d is the length of the speech unit segment divided into sentences, $m = \min\left(\text{Number of } Std \text{ units, Number of } Test \text{ units}\right)$, and Len is the length of the standard utterance. Since the length of the test statement has been regularized to be comparable to the length of the standard statement before the PVI operation, the calculation can only use Len as the calculation unit.

6) Comprehensively compare the number of accents, intensity curve matching and dPVI parameters of test utterances and standard utterances for rhythmic evaluation and feedback of English pronunciation quality.

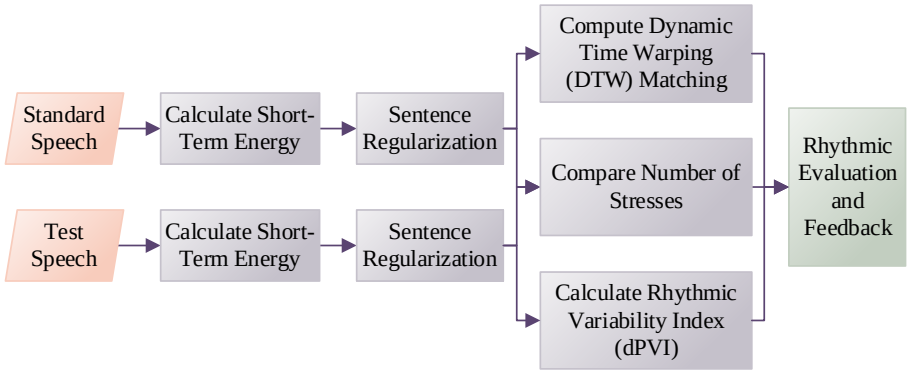

Figure 5: Rhythm evaluation

(4) Tone of voice evaluation

Intonation, i.e., the accent of speaking, is the configuration and change of the speaker's voice tone in terms of height, depression, and weight. The process of tone evaluation in this paper is shown in Figure 6. Firstly, the whole speech data is divided into frames, and then the posterior analysis is carried out in frames, using the autocorrelation function method (ACF) in the time domain to extract the pitch of each frame of data corresponding to an English sentence, further setting the range of pitch values to exclude unstable speech frames with abnormal pitch values, followed by smoothing the whole pitch by setting the median filter, and finally, using the DTW algorithm to calculate the pitch of the standard utterance with the Finally, the DTW algorithm is used to calculate the degree of intonation fit between the standard utterance and the input utterance, so as to provide intonation evaluation and feedback on the quality of English pronunciation.

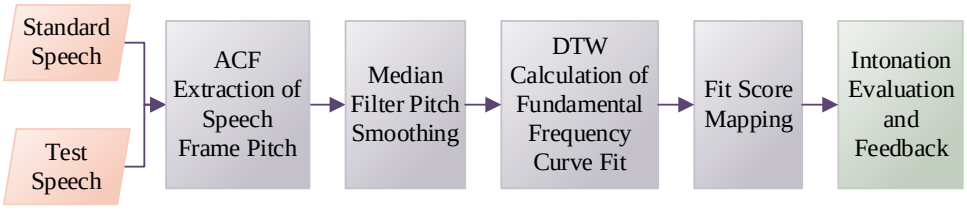

Figure 6: Intonation evaluation

The autocorrelation function method uses an autocorrelation function to compute the similarity between a sound frame $s(i)$, $i = 0,1,2,\cdots,n-1$ and itself:

$$acf(\tau) = \sum_{i=0}^{n-1-\tau} s(i)s(i+\tau) \tag{26}$$

where n is the length of a frame of speech data and τ is the amount of time delay, the pitch of this frame can be calculated by first finding the value of τ that will keep $acf(\tau)$ in a reasonably specific interval.

2.2 Intelligent Conversation System for College English Listening and Speaking

Combining the various techniques mentioned above, an intelligent session system is designed in this paper. When a learner interacts with an AI character in a specific environment or approaches an AI character in a free activity scene, a session is opened, and the flow of the session is shown in Fig. 7.

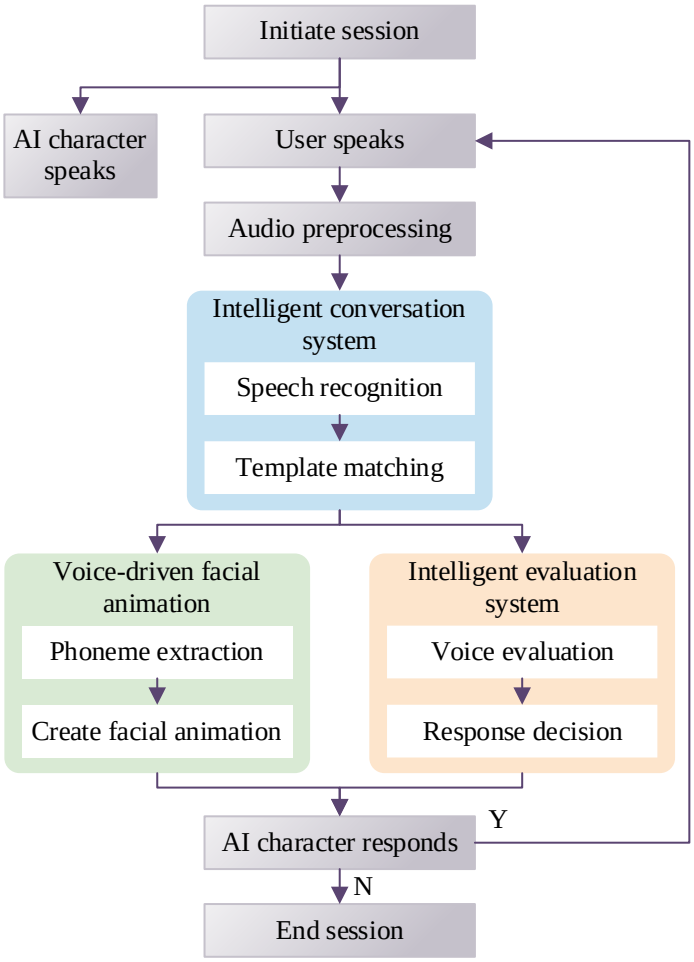

Figure 7: Conversation flow

2.2.1 Speech recognition

Before speech recognition, the system preprocesses the audio stream with the aim of noise reduction and removing noises. The system aims to exercise learners' conversational and English application skills, and does not exclude learners with a certain degree of accent. To this end, this paper designs and constructs a spoken English corpus, SELL, for Chinese English learners, and based on the SELL corpus, the trained speech recognition system will be more targeted to help the system to recognize the learners' speech.

2.2.2 Pronunciation assessment

Pitch accuracy, Speed of sound, Rhythm, Intonation, 0.3, 0.25, 0.25, 0.2, Pitch accuracy, Speed of sound, Rhythm, Intonation

In order to cooperate with the training of spoken English, a comprehensive evaluation of learners' pronunciation is made. The pronunciation evaluation system in this system contains four scoring criteria: pitch accuracy, speed of sound, rhythm, and intonation, and there are four sub-scores for each of the four criteria, and then the total score is subsequently calculated according to Equation (27):

$$S_{total} = S_{acc} * 0.3 + S_{spe} * 0.25 + S_{rhy} * 0.25 + S_{int} * 0.2 \tag{27}$$

where S_{acc}, S_{spe}, S_{rhy}, and S_{int} denote the pitch score, pitch score, rhythm score, and

intonation score, respectively. When the whole training is completed, the system automatically calculates the average score of all the assessment results in the session, as well as the total score, and when it is completed, the system displays all the scoring details and the total score to the learner, and records the scoring details in the learner's personalized learning information in order to track the learning progress.

2.2.3 Intelligent dialog

The application scenario in this system is mainly for the AI character to guide and help learners in English listening and speaking training, which belongs to the task-oriented intelligent conversation system.

The detailed flow of the whole intelligent conversation system is shown in Fig. 8, the user's voice is captured, noise reduced and packaged by the microphone, and then speech recognition is carried out, and the result of the recognition is transmitted on one side to the pronunciation evaluation module, which compares the recognition result with the reference text and gives the evaluation result, and meanwhile, the result of the recognition enters the template matching to get the valid matching result, and then the response is given by considering the evaluation result comprehensively. If the evaluation result is invalid or the matching fails, the result of this processing is discarded, and if it succeeds, the AI character responds to the learner according to the matching result.

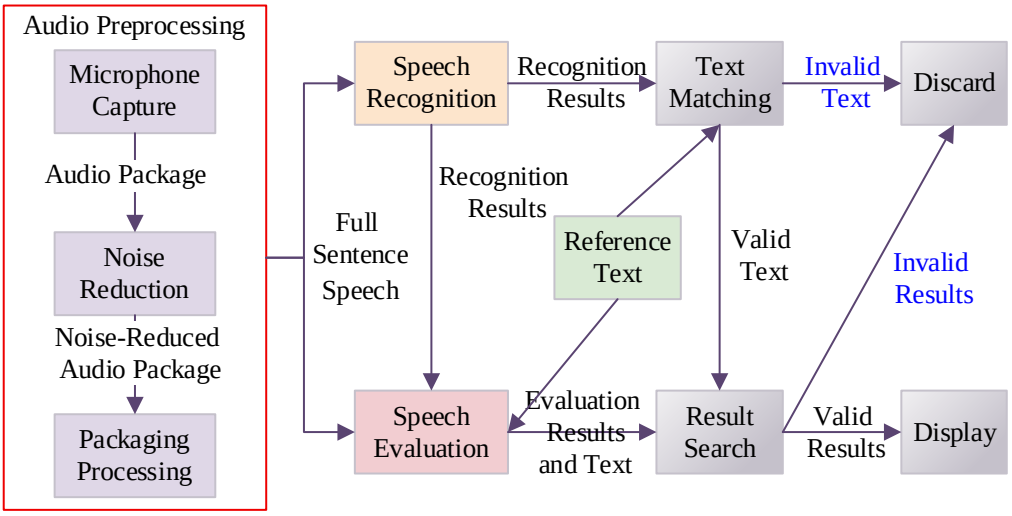

Figure 8: Detailed process of the intelligent dialogue system

2.3 Experimental results and analysis

In this section, the relevant modules in the designed AI English learning system are experimentally evaluated to verify the effectiveness of the system.

2.3.1 Speech feature parameter simulation analysis

(1) Comparison of feature extraction algorithms

Taking a piece of speech data as an example, the EMD-FD speech feature parameters of this piece of speech are extracted by the improved EMD-FD extraction algorithm, and the EMD-FD speech feature parameter curves of this piece of speech data are obtained as shown in Fig. 9. The speech spectra before and after the improved EMD-FD extraction algorithm are shown in Fig. 10, where (a) and (b) are the original speech spectrum and the improved EMD-FD algorithm speech spectrum, respectively.

It can be seen that in the entire 2400-5000 Hz frequency range, the speech spectrum of the improved EMD-FD extraction algorithm is basically consistent with the original speech map, so the improved EMD-FD extraction algorithm can be used to extract the features of the speech high-frequency region, while for the low-frequency region and the middle-frequency region, due to the large amount of computation of the improved EMD-FD extraction algorithm, and the MFCC extraction algorithm is relatively simple, so this paper adopts the hybrid speech feature parameters of MFCC and EMD-FD extraction algorithm as the speech recognition feature parameters.

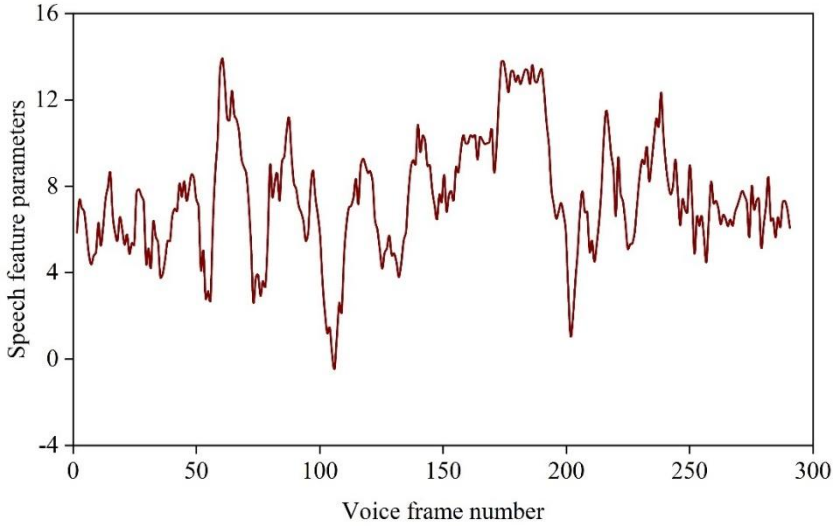

Figure 9: EMD-FD speech feature parameters

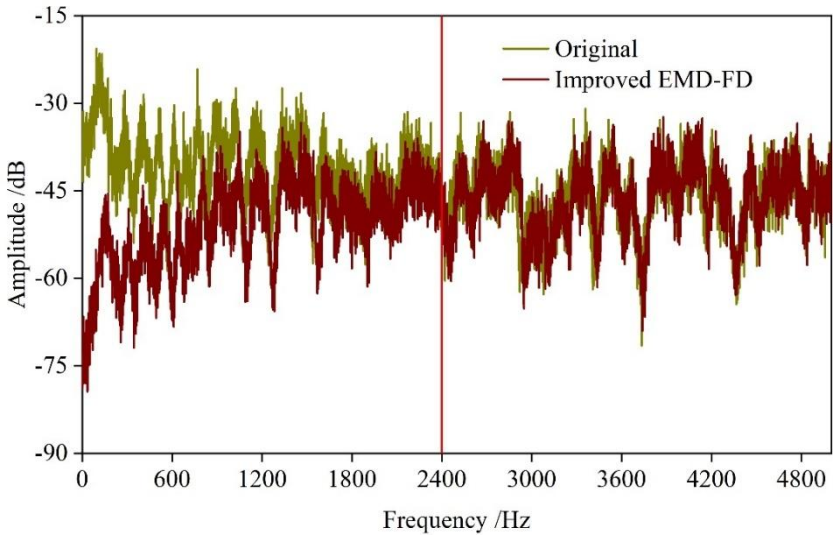

Figure 10: Calculation of EMD-FD extraction algorithm for voice data

(2) Simulation tools

The speech feature parameter extraction algorithms are all done using MATLAB and LIBSVM toolbox. Among them, LIBSVM toolbox is an open-source library based on support vector machine, which mainly solves pattern recognition, classification and regression problems, supports multi-language development, and provides a simulation and analysis interface with MATLAB software. Since it is an open source library, it contains various source codes, which can be supplied to developers for their own modification and improvement.

(3) Simulation conditions

Simulation presets: the sampling frequency of the speech signal is 8kHz, the pre-emphasis coefficient is 0.975, the sub-frame length is 512, the window function selects the Hamming window, the endpoint detection utilizes the spectral entropy method in this paper, and the number of filters contained in the final Mel filter bank is 24. In this paper, the SELL corpus is selected as a source of speech data, which contains a large amount of speech data. Four groups of speech data are selected from the SELL corpus for simulation, which are female speech w_1, w_2, w_3, w_4, male speech m_1, m_2, m_3, m_4, and noise-containing female speech w_1, w_2, w_3, w_4, w_5 and noise-containing male speech m_1, m_2, m_3, m_4, m_5, and the total number of sentences in each group is 600. Then the 24-dimensional LPCC and MFCC speech feature parameters of each speech sentence were extracted by LPCC and MFCC extraction algorithms, respectively, and then the 1-dimensional EMD-FD speech feature parameters were extracted by EMD-FD extraction algorithm.

(4) Simulation results and analysis

The extracted 1-dimensional EMD-FD and 24-dimensional MFCC speech feature parameters form a 25-dimensional hybrid speech feature parameter to compare the recognition effect with the single LPCC and MFCC speech feature parameters. The recognition simulation results of female speech, male speech, female speech with noise, and male speech with noise are shown in Tables 1~4, respectively.

According to the analysis of Tables 1~4, the average equal error rate of MFCC speech feature parameters decreased by 0.57%~1.74% compared to LPCC speech feature parameters. According to the analysis of Table 1 and Table 2, the average equal-error rate of the mixed EMD-FD and MFCC speech feature parameters compared to the single MFCC speech feature parameter decreased in 2.12% and 1.69% in female speech and male speech, respectively. According to the analysis in Table 3 and Table 4, the average equal-error rate of the speech feature parameter of the mixture of EMD-FD and MFCC did not change at all in the case of speech containing noise. In summary, LPCC speech feature parameter speech recognition effect is the worst, while MFCC speech feature parameter recognition effect is better than LPCC speech feature parameter.MFCC and EMD-FD mixed speech feature parameter not only ensures the integrity of the speech, but also shows the nonlinear characteristics of the speech, which improves the speech recognition rate.

Table 1: Equal error rate of female speech recognition

Characteristic parameters	Equal error rate /%				
	w_1	w_2	w_3	w_4	Average value
LPCC	18.49	17.58	18.37	18.64	18.27
MFCC	17.98	17.12	18.16	17.53	17.70
MFCC+EMD-FD	15.71	14.65	16.19	15.77	15.58

Table 2: Equal error rate of male speech recognition

Characteristic parameters	Equal error rate /%				
	m_1	m_2	m_3	m_4	Average value
LPCC	18.15	20.26	19.89	18.93	19.31
MFCC	17.74	18.78	18.26	16.32	17.78
MFCC+EMD-FD	15.93	16.41	16.07	15.94	16.09

Table 3: Equal error rates of female speech recognition with noise

Characteristic parameters	Equal error rate /%					
	w_1	w_2	w_3	w_4	w_5	Average value
LPCC	20.77	20.09	20.86	19.48	20.75	20.39
MFCC	18.63	19.31	18.44	18.76	18.09	18.65
MFCC+EMD-FD	15.71	16.12	16.19	15.16	14.73	15.58

Table 4: Equal error rates of male speech recognition with noise

Characteristic parameters	Equal error rate /%					
	m_1	m_2	m_3	m_4	m_5	Average value
LPCC	20.16	19.97	20.83	19.68	20.54	20.24
MFCC	18.81	19.32	18.41	18.59	18.15	18.66
MFCC+EMD-FD	15.92	16.51	16.87	15.79	15.36	16.09

2.3.2 Performance Analysis of Pronunciation Evaluation Models

(1) Performance metrics of pronunciation evaluation system

In order to make a comprehensive assessment of the model's scoring performance, we will quantitatively analyze the experimental results with the help of some rubrics.

1) Pearson's correlation coefficient

It is mainly used to reflect the linear correlation between two sequences, and its mathematical expression is as follows:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{28}$$

where the two sequences X and Y represent system ratings and teacher ratings respectively, $\operatorname{cov}(X,Y)$ is the skewed variance of X and Y, and σ_X and σ_Y are both standard deviations. The value range of $\rho_{X,Y}$ is $[-1,1]$, if the result is positive it means that there is a positive correlation between the two series, and if it is negative it means that there is a negative correlation between the two series. The closer the absolute value of the coefficient is to 1, the higher the correlation between the two sequences.

(2) The average score difference between human and machine scoring

This index is mainly used to describe the degree of deviation between machine scoring and human scoring. We use the letter d to denote this indicator, and its calculation formula is as follows:

$$d = E\left|S_{Machine} - S_{Human}\right| \tag{29}$$

3) Accuracy

This paper defines whether the scoring results are accurate by establishing a human-machine scoring error maximum. If the score difference between human scoring and machine scoring is less than this error maximum, the scoring result is considered accurate. Finally, the ratio of the number of accurate scoring results to the total number of samples in the test dataset is calculated as the accuracy index of the scoring model. In the practical application environment, we consider any human-machine scoring error within 1 point to be an accurate machine prediction result.

(2) Analysis of model scoring results

In this paper, we test the Multi-Parametric English Pronunciation Quality Evaluation (MPEPQE) model and the CNN+LSTM scoring model using 200 pieces of test data, and calculate 3 kinds of evaluation indexes to comprehensively evaluate the performance of the 2 scoring models.

The prediction results of the MPEPQE scoring model and CNN+LSTM scoring model on the total speaking score are shown in Figures 11 to 12, respectively. It can be found that the MPEPQE model exhibits a better fit than the CNN+LSTM model. Also, from the lowest student scores, we are able to see that the MPEPQE neural network exhibits better adaptability in the face of some extreme values.

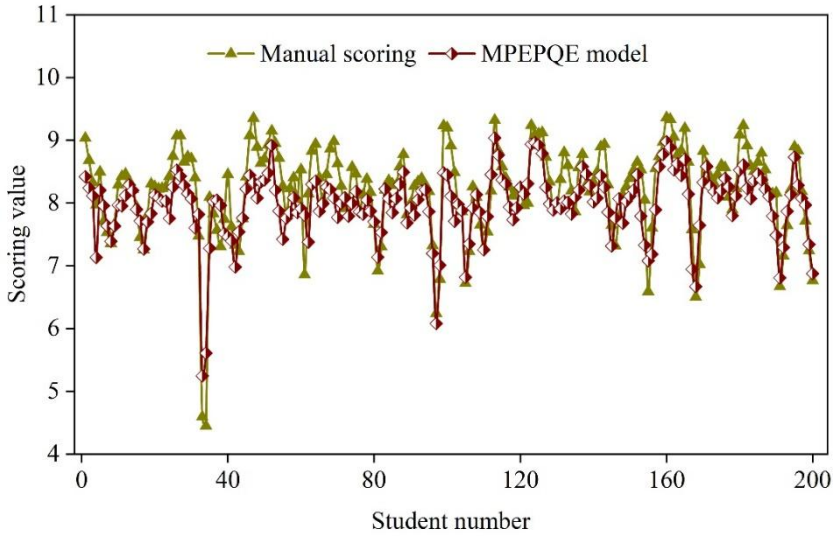

Figure 11: Scoring results of the MPEPQE model

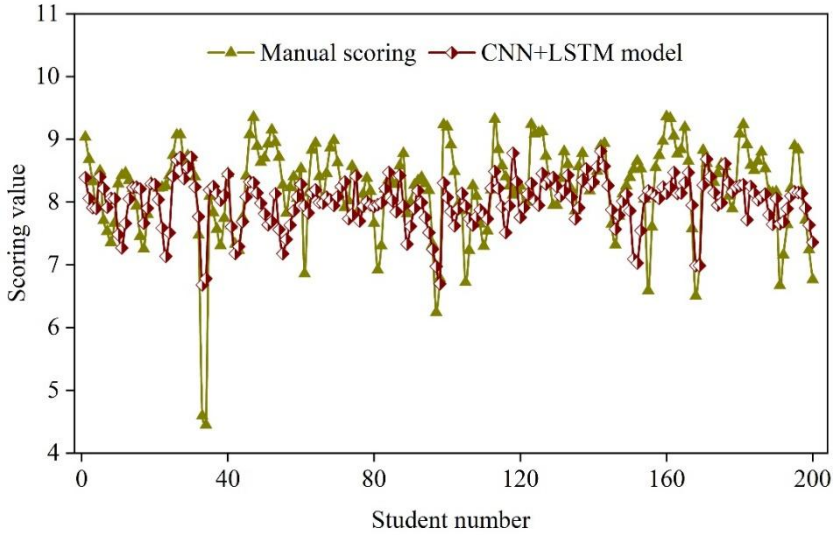

Figure 12: Scoring results of the CNN+LSTM model

The performance evaluation results of the two scoring models are shown in Table 5, the MPEPQE model outperforms the CNN+LSTM model in the three indexes of Pearson's correlation coefficient, accuracy and average score difference, and the gap between the two is more obvious in terms of human-computer scoring correlation (0.706>0.547). Overall, the Multi-Parametric English Pronunciation Quality Evaluation (MPEPQE) model constructed in this paper performs better and can be applied to the task of speech quality assessment in AI

English learning systems.

Table 5: Performance evaluation of scoring models

	Pearson correlation	Average score difference	Accuracy rate
Between manual scoring	0.775	0.498	-
MPEPQE model	0.706	0.613	84.52%
CNN+LSTM model	0.547	0.649	81.34%

3 Experimental research on improving English listening and speaking skills based on AI learning system

This study utilizes an English learning system based on AI technology to conduct a teaching experiment, aiming to investigate the effect of this teaching mode in improving English listening and speaking skills.

3.1 Experimental Program Design

3.1.1 Experimental hypotheses

The implementation of listening and speaking skills development strategies using the AI English Learning System and teaching activities in the sophomore experimental class can develop students' listening and speaking skills in English.

3.1.2 Experimental variables

(1) Experimental Variables

This study utilizes the listening and speaking skill development strategy supported by the AI English learning system in the teaching activities of the experimental class. Regular teaching methods without any AI English learning system support are utilized in the control class.

(2) Dependent Variables

Changes in students' English scores and changes in students' English listening and speaking skills.

(3) Interference Variables

Students' pre-existing cognitive experience, the psychological orientation of the classroom teachers, and the psychological state of the experimental and control groups.

(4) Control of experimental interfering variables

The experimental class and the control class are selected as far as possible as two classes similar in all aspects. The same teacher teaches the experimental class and the control class. The learning content and progress of the two classes are kept the same. Students in the two classes maintain their original learning status and study normally, and students are unaware of the experimental research being conducted by the teacher and the fact that their class is an experimental class, so as to avoid increasing the psychological burden and learning burden on the students. The experimental and control classes used the same test instruments.

3.1.3 Experimental modalities

In this study, an unequal experimental group control pre and post measurement calibration experimental design was used, and the experimental design is shown in Table 6.

Table 6: Experimental design

Simultaneous pre-test	Pre-test results	Group of subjects	Experimental processing	Simultaneous post-test
R_1	$R_1 \approx R_2$	Experimental group	Accept	R_3
R_2		Control group	Not accepted	R_4

3.1.4 Experimental subjects

In this study, two sophomore classes in college A were selected as research subjects, in which class A, as an experimental class, utilized the listening and speaking ability development strategy supported by the AI English learning system to carry out teaching activities, and class B, as a control class, used the original traditional teaching method to carry out teaching activities. In order to reduce the influence of interfering variables, the experimental class and the control class were taught by the same teacher, with the same teaching content but in different forms. The basic situation of the experimental subjects is shown in Table 7.

Table 7: Basic information of the experimental subjects

Class	Natural class	Number	Class
Experimental class	1	47	Class A
Control class	1	46	Class B

3.1.5 Experimental tools

(1) Student Questionnaire

The student questionnaire was compiled according to the listening and speaking ability, through which the changes in students' language perception ability, language comprehension ability, language evaluation ability, language application ability and information acquisition ability were analyzed. First of all, the students' listening and speaking ability was pre-tested. 47 questionnaires were distributed in the experimental class before the experiment was carried out, and 46 valid questionnaires were returned, with an effective rate of 97.87%. A total of 46 questionnaires were distributed in the control class and 44 valid questionnaires were returned with an effective rate of 95.65%. At the end of the experiment, 47 questionnaires were distributed to the experimental class and 45 valid questionnaires were retrieved with an effective rate of 95.74%. Forty-six were distributed to the control class and 43 valid questionnaires were retrieved, with an effective rate of 93.48%.

(2) Test scores

The English test questions are prepared to compare the students' achievements in improving their listening and speaking ability by using the listening and speaking ability cultivation strategy supported by the AI-based English learning system through the comparative analysis of the scores, and to verify the degree of influence of the university English listening and speaking ability cultivation strategy based on the AI-based English learning system on the cultivation of the students' listening and speaking ability.

3.2 Experimental results and analysis

3.2.1 Impact of AI English Learning System on Student Achievement

The statistical results of the two English test scores of the control and experimental groups at the beginning and end of the experiment are shown in Table 8. It can be seen that the average score of the first test of the experimental class was lower than that of the control class ($64.27<66.31$) while the second was higher than that of the control class ($78.54>70.62$).

Table 8: Comparison of the average English scores of the two classes

	Class	N	Mean	SD	Standard error
Pretest	Experimental class	47	64.27	8.451	0.628
	Control class	46	66.31	7.982	0.642
Pretest	Experimental class	47	78.54	10.457	0.735
	Control class	46	70.62	9.653	0.759

Then, based on the comparison of the differences between the mean English scores of the two classes, the corresponding differences in the overall mean scores were tested for significance. The results of the independent samples t-test for the two tests between the experimental and control classes are shown in Table 9.

It can be observed that in the first test before the experiment, the lower and upper 95% confidence intervals of the difference between the listening scores of the control class and the experimental class were -1.274 and 3.239, respectively, containing zero, with a t-value of 1.724 and a degree of freedom of 91. The probability of significance of the two-tailed t-test was 0.173>0.05, which indicated that there was no significant difference between them. In contrast, in the second test after the experiment, the listening scores of the experimental class were significantly higher than those of the control class, and the difference between the English scores of the control class and the experimental class was significant in both cases ($p=0.004<0.05$). This indicates that after receiving different teaching methods and approaches for a period of 16 weeks, the progress of English scores of students in the experimental class is greater than that of the control class, which proves the effectiveness of the AI English learning system designed in this paper in improving the performance of college students' listening and speaking skills.

Table 9: T-tests of English scores before and after the experiment

	Class	Mean	SD	95% Confidence Interval of the difference		t	df	Sig (2-tailed)
				Lower	Upper			
Pretest	Experimental class - Control class	-2.04	3.412	-1.274	3.239	1.724	91	0.173
Posttest	Experimental class - Control class	7.92	3.935	-4.817	-0.317	-4.569	91	0.004

3.2.2 Changes in Students' English Listening and Speaking Skills Before and After the Experiment

The changes in the mean of the indicators of English listening and speaking ability of the students in the experimental class and the control class after receiving different teaching styles and methods are shown in Table 10. The t-test of the difference between the mean scores of students' English listening and speaking ability in the two classes before and after the experiment is shown in Table 11.

It can be seen that the difference between the mean scores of students' English listening and speaking ability in the two classes before the experiment ranges from 0.04 to 0.33, while the difference between the mean scores of the two classes after the experiment ranges from 0.81 to 1.32. The t-test shows that, in the first survey before the experiment, the control class and the experimental class have the same mean scores in the indicators of English listening and speaking ability in the categories of language perception, language comprehension, language evaluation, language use, information acquisition and so on. Acquisition ability and other indicators of English listening and speaking ability are not significant ($P>0.05$). And in the

second survey after the experiment, there was a significant difference between the indicators of the two classes ($P<0.05$). This indicates that the students who use AI English learning system are significantly better than those who do not use AI English learning system in terms of English listening and speaking ability, i.e., the teaching mode of applying AI-enabled learning system can significantly improve the English listening and speaking ability of college students.

Table 10: Comparison of the English listening and speaking ability of the two classes

Indicator		Class	N	Mean	SD	Standard error
Language perception ability	Pretest	Experimental class	46	14.27	1.735	0.592
		Control class	44	14.38	1.623	0.588
	Posttest	Experimental class	45	15.89	1.247	0.534
		Control class	43	14.75	1.159	0.475
Language comprehension ability	Pretest	Experimental class	46	15.61	1.372	0.575
		Control class	44	15.28	1.295	0.564
	Posttest	Experimental class	45	16.74	1.138	0.531
		Control class	43	15.42	1.751	0.493
Language evaluation ability	Pretest	Experimental class	46	15.22	1.603	0.694
		Control class	44	15.18	1.624	0.685
	Posttest	Experimental class	45	16.48	1.862	0.557
		Control class	43	15.36	1.415	0.509
Language application ability	Pretest	Experimental class	46	14.17	1.648	0.541
		Control class	44	14.29	1.572	0.538
	Posttest	Experimental class	45	15.54	1.926	0.553
		Control class	43	14.41	1.617	0.494
Information acquisition ability	Pretest	Experimental class	46	9.28	1.944	0.426
		Control class	44	9.47	1.815	0.417
	Posttest	Experimental class	45	10.42	1.515	0.408
		Control class	43	9.61	1.234	0.335

Table 11: T-tests of English listening and speaking ability before and after the experiment

	Class	Mean	SD	95% Confidence Interval of the difference		t	df	Sig (2-tailed)
				Lower	Upper			
Pretest	Experimental class - Control class	-0.11	0.214	-1.452	3.657	0.924	88	0.275
	Experimental class - Control class	0.33	0.206	-1.384	3.541	1.352	88	0.064
	Experimental class - Control class	0.04	0.198	-2.241	3.057	0.346	88	0.782
	Experimental class - Control class	-0.12	0.175	-2.341	4.671	0.459	88	0.105
	Experimental class - Control class	-0.19	0.221	-1.687	3.254	0.578	88	0.224
Posttest	Experimental class - Control class	0.81	1.435	0.115	3.427	2.214	86	0.037
	Experimental class - Control class	1.14	1.527	0.142	3.196	2.267	86	0.008
	Experimental class - Control class	1.32	1.638	0.038	3.162	2.048	86	0.003
	Experimental class - Control class	1.12	1.452	0.342	2.859	1.657	86	0.005
	Experimental class - Control class	1.13	1.559	0.089	1.875	2.259	86	0.005

4 Conclusion

In this paper, a university English learning system is designed by combining artificial intelligence technology. The system realizes the auxiliary training of students' listening and speaking ability by loading modules such as improved EMD-FD special extraction algorithm, deep learning speech recognition system and multi-parameter pronunciation quality evaluation model MPEPQE, which can be used as an innovative tool for English teaching mode.

By using the improved EMD-FD algorithm to extract features in the high frequency region of speech from 2400-5000 Hz, combined with the MFCC algorithm to extract features in the low and middle frequency regions, the effective extraction of features of speech signals is realized.The speech feature parameters of the mixture of EMD-FD and MFCC are lower than those of the single MFCC speech feature and LPCC speech feature parameters in terms of the average equal-error rate, which can ensure the integrity of speech. It can ensure the integrity of speech, show the nonlinear characteristics of speech, and improve the speech recognition rate.

The fitting effect of MPEPQE model is significantly better than that of CNN+LSTM model, and it shows better adaptive ability. At the same time, the MPEPQE model outperforms the CNN+LSTM model in the three metrics of human-computer score correlation, accuracy and average score difference. Overall, the MPEPQE model performs better and can be applied to the task of speech quality assessment in AI English learning systems.

In this paper, we address the question of “whether the university English listening ability development strategy based on AI learning system is effective in developing students' listening ability”, and use quasi-experimental research to conduct teaching experiments. The results of the study show that the students who use the university English listening and speaking ability cultivation strategy based on the AI learning system improve their English listening and speaking ability, and the improvement effect is significantly better than that of the control group who uses the traditional mode ($P<0.05$). It indicates that the English listening and speaking ability cultivation mode based on intelligent speech system is effective.

This study takes the cultivation of students' listening and speaking ability as the starting point, designs teaching activities, applies cultivation strategies and carries out teaching practice. It is hoped that the research experience of this study can provide some reference for subsequent studies. In later studies, researchers can try to apply new technology tools in the classroom, break the traditional teaching mode, realize the combination of in-class and out-of-class, make full use of AI technology, and carry out new teaching methods.

About the Author

Chuneng Zhao (1973.10-), female, Han ethnicity, Tangshan, Hebei Province, scholar, holds a postgraduate degree and serves as a lecturer. Her research focuses on English linguistics, college English teaching, college English writing, and translation.

References

[1] Wu, B. (2018). Construction of Ecological Teaching Model for College English Course under the Background of Internet plus. Educational Sciences: Theory & Practice, 18(6).

[2] Wu, B., Zhang, H., & Wu, X. (2024, December). The Application of Artificial Intelligence in College English Course and CET-4 Mock Examination——A Case Study of “iTEST” and “iWRITE”. In Proceedings of the 2024 2nd International Conference on

Information Education and Artificial Intelligence (pp. 674-682).

[3] Abdellatif, M. S., Alshehri, M. A., Alshehri, H. A., Hafez, W. E., Gafar, M. G., & Lamouchi, A. (2024). I am all ears: listening exams with AI and its traces on foreign language learners' mindsets, self-competence, resilience, and listening improvement. Language Testing in Asia, 14(1), 54.

[4] Almutairi, A. F., Gegov, A., Adda, M., & Arabikhan, F. (2020). Conceptual artificial intelligence framework to improving English as second language. WSEAS Transactions on Advances in Engineering Education, 17, 87-91.

[5] Sharadgah, T. A., & Sa'di, R. A. (2022). A systematic review of research on the use of artificial intelligence in English language teaching and learning (2015-2021): What are the current effects?. Journal of Information Technology Education: Research, 21.

[6] Alshumaimeri, Y. A., & Alshememry, A. K. (2023). The extent of AI applications in EFL learning and teaching. IEEE Transactions on Learning Technologies, 17, 653-663.

[7] Zhu, M., & Wang, C. (2024). A systematic review of artificial intelligence in language education from 2013 to 2023: Current status and future implications. Available at SSRN 4684304.

[8] Zeng, T. (2025). Innovative Practices and Case Studies of Senior High school English Listening Teaching in the Context of the New Curriculum Standards. Literature, Language and Cultural Studies, 1(2), 73-82.

[9] Sukying, A., & Barrot, J. S. (2025). Friend or Foe? Investigating the Alignment of English Language Teaching (ELT) Textbooks with the National English Curriculum Standards. The Asia-Pacific Education Researcher, 34(2), 793-801.

[10] Yuldashev, S., Taryanikova, M., & Shayakubov, S. (2025). PROBLEMS OF TEACHING ENGLISH LISTENING SKILLS. International Journal of Artificial Intelligence, 1(3), 1756-1765.

[11] Somé, K. J. (2025). Teachers' Agency and Teaching Challenges in Multilingual Spaces: An Exploratory Study of How EFL Teachers Address Listening and Speaking Skills Needs in Centralized Systems. TESL-EJ, 28(4), n4.

[12] Jing, W. (2024). Speech recognition sensors and artificial intelligence automatic evaluation application in English oral correction system. Measurement: Sensors, 32, 101070.

[13] Abdulhussein Dakhil, T., Karimi, F., Abbas Ubeid Al-Jashami, R., & Ghabanchi, Z. G. (2025). The Effect of Artificial Intelligence (AI)-Mediated Speaking Assessment on Speaking Performance and Willingness to Communicate of Iraqi EFL Learners. International Journal of Language Testing, 1-18.

[14] Hamid, S. F., & Marpaung, D. V. (2025). Creating Adaptive Listening Materials with AI-Generated Voices: A Need Analysis in Intensive Listening Classrooms. PROCEEDING AISELT, 10(1).

[15] Gao, M., Song, W., Zhang, C., Zhou, Q., & Su, P. (2022). Situational teaching based evaluation of college students' English reading, listening, and speaking ability. International Journal of Emerging Technologies in Learning (iJET), 17(8), 140-154.

[16] Zhou, J. (2019, July). Construction of artificial intelligence-based interactive oral English teaching platform based on application problems of present intelligent products. In IOP Conference Series: Materials Science and Engineering (Vol. 569, No. 5, p. 052055). IOP Publishing.

[17] Li, Y. (2022). Teaching mode of oral English in the age of artificial intelligence. Frontiers in Psychology, 13, 953482.

[18] Ma, Y., Tang, X. J., & Huang, X. (2025). AI-powered adaptive English language learning systems: leveraging deep learning algorithms and natural language processing for personalized teaching approaches. IEEE Access.

[19] Darmawansah, D., Hwang, G. J., Lin, C. J., & Febiyani, F. (2025). An artificial intelligence-supported GFCA learning model to enhance L2 students' role-play performance, English speaking and interaction mindset. Educational technology research and development, 1-29.

[20] Wu, S., & Wang, F. (2021). Artificial intelligence-based simulation research on the flipped classroom mode of listening and speaking teaching for English majors. Mobile Information Systems, 2021(1), 4344244.

[21] Hu, B. (2021). English listening teaching model in flipped classroom based on artificial intelligence fusion control algorithm. Mathematical Problems in Engineering, 2021(1), 6005359.

[22] Sahito, J. K. M., Panwar, A. H., & Ramzan, I. (2025). EXPLORING THE IMPACT OF ARTIFICIAL INTELLIGENCE (AI) ON THE LISTENING SKILLS OF ENGLISH AS A SECOND LANGUAGE (ESL) LEARNERS. Journal of Applied Linguistics and TESOL (JALT), 8(1), 1059-1067.

[23] Hu, N. (2025). English listening and speaking ability improvement strategy from Artificial Intelligence wireless network. Wireless Networks, 31(2), 1071-1080.

[24] Jantakoon, T., Jantakun, T., Jantakun, K., Pongpanich, W., Pasmala, R., Wannapiroon, P., & Nilsook, P. (2025). The effectiveness of artificial intelligence in English instruction for speaking and listening skills: A meta-analysis. Contemporary Educational Technology, 17(4), ep596.

[25] Pang, Y. (2025). Research on the Theory and Practice of AI-enabled Digitalization in College English Listening and Speaking Courses. International Journal of Educational Development, 2(1), 39-49.