Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Intelligent English Classroom Interaction Design and Teaching Method Optimization Based on Deep Reinforcement Learning

Yajuan Zuo[1,✉]

[1] *Basic Teaching Department, Shanxi College of Applied Science and Technology, Taiyuan, Shanxi, 030062, China*

ABSTRACT

At present, the evaluation of spoken English in domestic universities is affected by the evaluation teachers' personal cognition, preference, time, energy and other factors, and it is difficult to unify the standard of oral evaluation in the implementation, and the evaluation frequency and timeliness are insufficient to meet the students' willingness to improve their oral language. In this paper, multimodal speech recognition technology is utilized to firstly collect students' speech signals through microphone arrays, secondly extract acoustic and linguistic features of speech, and construct multimodal feature vectors by combining visual information such as students' lip movements and facial expressions. Subsequently, the feature vectors are input into a deep neural network model for training and recognition, fusing LSTM network with attention mechanism to analyze the speech emotion and capture the emotional changes in speech. Meanwhile, the interaction behavior in speech is analyzed by combining temporal convolutional network. Construct a deep reinforcement learning model, introduce a user item interaction layer, design a user interaction simulator, and obtain user feedback on the smart English classroom. Using multimodal speech recognition technology, the temporal waveform of classroom speech is analyzed for sound pressure value, and the normalized sound pressure value range fluctuates around [-1.5,1.5].The average recognition rate of the six emotions rises to 67.86% with the joint effect of LSTM and attention mechanism. By comparing the experiment, analyzing the difference between the experimental class and the control class before and after the reading aloud ability, the average score of the experimental class is 23.945, and the average score of the control class is 21.464, at the same time, the post-test of reading aloud ability corresponding to the experimental class and the control class $P=0.005<0.05$. It can be seen that the intelligent interactive classroom of English language constructed in this paper has a facilitating effect in the process of teaching reading aloud in the aspect of reading aloud ability of students The classroom can be seen that the intelligent English interactive classroom constructed in this paper has a promoting effect in the process of teaching reading aloud in terms of students' reading ability.

✉ Corresponding author.
E-mail address: yajuan_zuo@163.com (Y. Zuo).
Received 15 May 2024; Revised 01 August 2024; Accepted 20 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-046

Keywords: multimodal speech recognition, deep neural network model, LSTM network, user interaction simulator, English interactive classroom

1. Introduction

With the continuous development of modern information technology, the informatization construction of China's college and university education undertakings has entered a brand new stage of development. Big data, the Internet of things and artificial intelligence and other digital technologies and the deep integration of professional disciplines teaching work in colleges and universities, the classroom teaching environment has undergone significant changes, from the original multimedia classroom gradually transitioned to a modern smart classroom, which will also become a normalized environment for classroom teaching in colleges and universities in the future [1-3]. In the process of comprehensively promoting “Internet + education” and striving to create a new form of digital education, the learning purpose and teaching concept of English courses in colleges and universities will also change. In the past, the learning purpose of the English classroom mainly stayed in the shallow level of literacy, mimeography and understanding of English subject knowledge, while in the college English wisdom classroom, more attention is paid to the transformation of higher-order abilities to the depth of intensive learning, such as the transfer of knowledge, application, solving complex problems, and reflection and evaluation of the teaching process, etc. [4-7]. By constructing a college English wisdom classroom depth-enhanced learning model, teachers can create a harmonious teaching environment that supports wisdom classroom depth-enhanced learning, optimize multi-level teaching objectives by enriching the English teaching content and other ways, and carry out classroom teaching reform and diversified teaching evaluation [8-11]. It not only improves the practicality and efficiency of English smart classroom teaching, but also stimulates students' enthusiasm for independent learning, thus promoting students' deep-enhanced learning and thinking development, and improving students' learning effect and thinking ability [12-14]. Therefore, whether or not to accurately construct the depth-enhanced learning mode and conduct scientific evaluation in the English smart classroom in colleges and universities has become an important measure to improve the quality of talent cultivation in Chinese universities. Targeted strategies should be adopted to address the problems such as low motivation of students' participation, poor cognitive ability and cultivation effect, etc., which exist in the teaching of college English smart classroom at this stage, in order to construct a brand-new depth-enhanced learning mode of smart classroom [15-17].

In this paper, key technologies such as multimodal speech recognition, speech emotion analysis, and speech interaction behavior are comprehensively utilized to comprehensively assess the learning status of students and provide teachers with personalized teaching decision support. A neural network is introduced on the basis of the DQN algorithm, and the neural network is used to infinitely approximate the value function therein, and then the value function therein is updated to complete the construction of the deep reinforcement learning model. It is proposed to construct an interactive network to explicitly construct the interactions between students and English classroom features, use the embedding matrix of the embedding layer to map the user attribute information into vectors, and input the mapped vectors into the fully connected layer with the activation function of ReLU to extract the user attribute feature interrogation quantities, and then subsequently, the mapped interrogation quantities are fed into the fully connected layer with the activation function of ReLU, and the user interaction simulator is designed. The designed user interaction learning model is applied to an intelligent English classroom and the impact of its interaction is analyzed in controlled experiments.

2. Intelligent Speech Interaction Based Learning Recognition Method for English Teaching and Learning

2.1. Overview of the methodology

The English teaching learning recognition method based on intelligent voice interaction makes comprehensive use of key technologies such as multimodal speech recognition, voice emotion analysis, and voice interaction behavior analysis, aiming to objectively and comprehensively assess students' learning status and provide teachers with personalized teaching decision support. The implementation process of the method is divided into three steps: first, the multimodal speech recognition technology is used to accurately recognize the content of the voice interaction between students and the intelligent teaching system, and to obtain the dialogue data in text form. Second, the speech sentiment analysis technology is used to analyze the students' speech data in terms of sentiment and attitude, and to judge the students' learning emotional state. Third, the voice interaction behavior analysis technology is used to analyze the learning behavior characteristics of students from the dimensions of students' voice interaction patterns, speech speed, and volume. Through these three steps, the cognitive state, emotional state and behavioral performance of students in the learning process can be dynamically tracked, which helps teachers grasp the actual learning needs of students to optimize teaching strategies.

2.2. Key technologies

2.2.1. Multimodal speech recognition technology. Multimodal speech recognition technology integrates acoustic, linguistic, visual and other modal information, which can effectively improve the accuracy and robustness of speech recognition [18]. First, students' speech signals are captured by microphone arrays and subjected to a series of pre-processing. Second, the acoustic and linguistic features of the speech are extracted and combined with the visual information such as the students' lip movements and facial expressions to construct a multimodal feature vector. Finally, the feature vectors are input into a deep neural network model for training and recognition. In the recognition process, a continuous density Hidden Markov model is used, with the observation sequence as $O=\{o_1,o_2,\cdots,o_T\}$ and the hidden state sequence as $Q=\{q_1,q_2,\cdots,q_T\}$. Where T denotes the length of the sequence. The transfer probability matrix $A=[a_{ij}]_{C\times C}$ describes the transfer relationship between states, where a_{ij} denotes the probability of transferring from state i to state j and C denotes the number of matrix columns. The optimal state sequence is solved by Viterbi algorithm to realize speech recognition with the expression:

$$\hat{Q}=\arg\max_Q P(Q\mid O)=\arg\max_Q \frac{P(O\mid Q)P(Q)}{P(O)} \tag{1}$$

where: $\bar{Q}$ is the optimal state sequence, $P(O)$ is the edge probability of the observation sequence, $P(Q)$ is the prior probability of the state sequence, and $P(Q\mid O)$ is the probability of generating the observation sequence under the state sequence Q. The use of multimodal speech recognition technology can accurately transform the content of students' voice interactions into text form, providing a reliable data base for the analysis of the learning situation in English language teaching English language teaching.

2.2.2. Speech Sentiment Analysis Technology. Speech is rich in emotional information, and sentiment analysis of students' speech data is an important part of emotion recognition in English teaching. The article adopts a speech sentiment analysis technique based on Long Short-Term Memory (LSTM) network and attention mechanism, which can effectively capture the emotional changes and elements in speech [19].

First, the student's speech signal is segmented into a sequence of frames, each of which has a length of L and a frame shift of S. Second, the acoustic features of each frame are extracted, including fundamental frequency, energy, resonance peaks, etc., and normalized. Again, the feature sequence is input into the LSTM network, and the modeling of long and short-term dependencies is achieved through the gating mechanism. In the output layer of the LSTM, the eliciting attention force mechanism adaptively focuses on the frames that have a greater impact on the emotion judgment to generate the emotion representation vector. Finally, the representation vector is passed into the Softmax layer for sentiment classification to obtain the probability distribution of each sentiment category. The loss function uses cross entropy with the expression:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{2}$$

where: L is the cross-entropy loss, N is the number of emotion categories, y_i is the one-hot coding of real emotion labels, and $\hat{y}_i$ is the predicted emotion probability. During the training process, the model parameters are updated by back propagation algorithm to continuously improve the accuracy of emotion recognition. Using this technology, we can analyze students' emotional state in the teaching process in real time, such as positive, neutral, negative, etc., to provide teachers with emotional feedback and optimize teaching strategies. At the same time, combined with the trend of students' emotional changes, it can predict students' learning interest and concentration level, providing reference for personalized learning.

2.2.3. Voice Interaction Behavior Analysis. Behavioral analysis of speech interaction is another important dimension of academic recognition in English language teaching. Unlike the previous analysis focusing on semantic content, this technique focuses on the behavioral features during students' speech interaction, such as speech rate, volume, pause, and repetition. These behaviors contain key learning information such as students' cognitive input and emotional state. In order to capture the temporal patterns of speech behaviors, the article proposes an analysis method based on Temporal Convolutional Networks (TCNs) [20].TCNs employ causal convolution and null convolution, which can expand the feelings while maintaining the temporal information.

The students' speech segments are first preprocessed to extract low-order descriptive features such as speech rate and volume. Then, the feature sequences are infused into TCN for higher-order temporal modeling.The core of TCN is causal convolution operation with the expression:

$$F(x_t) = (x_d^* f)(t) = \sum_{i=0}^{k-1} f(i) \cdot x_{t-d \cdot i} \tag{3}$$

where: $F(x_t)$ is the output feature of the input sequence x at the current time step t, x is the input sequence, f is the convolution kernel function, k is the convolution kernel size, d is the null factor, $*$ is the convolution operation, t is the current time step, and i is the index of the convolution kernel function f. By stacking multiple causal convolutional modules, the TCN can capture speech behavior patterns at different scales while ensuring causality. Finally, the output features of the TCN are fed to the classifier to identify the types of student interaction behaviors, e.g., focused, distracted, and puzzled. Through voice interaction behavior analysis, teachers can dynamically gain insights into students' participation and cognitive status in the classroom and adjust the teaching pace and mode in a timely manner. The study of the association between behavioral characteristics and learning performance also provides new ideas for teaching optimization.

2.3. Deep reinforcement learning model construction

The DQN algorithm approximates the value function by introducing a neural network viewed as a parameterized approximation and expressing the weights of the three convolutional layers and the two fully connected layers in terms of their magnitudes [21]. The value function is represented by Equation $Q(s,a;\theta)$, where θ denotes the set of network parameters. One of the main advantages of using neural networks for value function approximation is that it is able to handle large state spaces, which would otherwise be infeasible to store in a Q-value table. In addition, DQN uses an empirical replay mechanism to train the network, which samples transitions from a buffer to break the correlation between successive samples and reduce the variance of updates.The network structure used for DQN is shown in Figure 1.

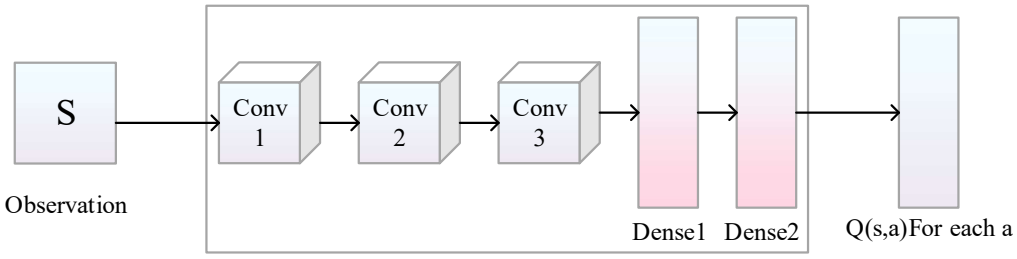

Fig. 1. DQN network structure diagram

However, instability and non-convergence are prevalent problems when deep neural networks are utilized to approximate value functions. These limitations lead to significant challenges in achieving stable convergence, which in turn limits the feasibility of neural network-based approximation of the value function in reinforcement learning applications. The reason for the above phenomenon is mainly the strong correlation between the data sample tuples. To address this issue, an empirical playback technique is utilized in DQN.

Experience replay techniques have become an essential component of deep reinforcement learning algorithms such as DQN.

At each time step, the algorithm stores experience tuples $e_t=(s_t,a_t,r_t,s_{t+1})$ in a data pool $D_t=\{e_1,e_2,e_3\cdots e_t\}$ and randomly samples from this pool to train the network parameters θ. By utilizing previous experience, the method makes more efficient use of the data and allows for multiple weight updates. This is particularly useful in complex environments with high-dimensional state and action spaces. The use of random sampling techniques effectively removes correlations in the middle of consecutive samples, which in turn reduces the variance of the updates and avoids overfitting. Notably, the learning strategy is designed in such a way that the current network parameters determine the next set of data samples to be used for training, thus ensuring a more efficient and effective training process.

A separate target network is set up in the DQN to perform separate processing operations on the TD deviations contained in the TD algorithm. The neural network is then used to perform infinite approximation of the value function therein, and then the value function therein is updated, which is essentially an update of the network parameter θ, which is updated by the gradient descent method with the formula shown in the following equation:

$$\theta_{t+1}=\theta_t+\left[r+\gamma_{a'}^{max}Q(s',a';\theta^-)-Q(s,a;\theta)\right]\nabla Q(s,a;\theta) \tag{4}$$

where $r+\gamma_{a'}^{\max}Q(s',a';\theta)$ is the TD optimization objective and the network parameter used in computing the value of $\max_{a'}Q(s',a';\theta)$ is θ.

The so-called TD network is the neural network being used to compute the TD objective, and the TD network is a key component of the DQN algorithm. Prior to the advent of the DQN algorithm, the

approximation of the value function using neural networks was achieved by using the same set of parameters θ in both the state-action value function and the gradient computation for the TD objective. However, this approach led to the problem of strong correlation of the parameter data, resulting in unstable training. To overcome this limitation, it has been proposed to use a separate set of objective network parameters, denoted as θ^{-}, for the TD optimization objective, while the estimated network parameters θ are used to compute the value function. In this way, the network parameters θ used to compute the value function are updated at each step, whereas the network parameters θ^{-} used to compute the TD optimization objective are updated only after a fixed number of steps C. So the update function θ_{t+1}, the loss function $L(\theta)$ and the objective function Q_{target} are given by:

$$\theta_{t+1} = \theta_t + \alpha[r + \gamma_{a'}^{\max} Q(s', a'; \theta^{-}) - Q(s, a; \theta)] \nabla Q(s, a; \theta) \tag{5}$$

$$L(\theta) = E_{s,a,r,s'}[Q_{target} - Q(s, a; \theta)]^2 \tag{6}$$

$$Q_{target} = E_{s'}[r + \gamma_{a'}^{\max} Q(s', a'; \theta^{-}) \mid s, a] \tag{7}$$

And the loss function $L(\theta)$ and objective function Q_{target} for DQN with negative feedback are given by:

$$L(\theta) = E_{s,a,r,s'}[Q_{target} - Q(s_+, s_-, a; \theta)]^2 \tag{8}$$

$$Q_{target} = E_{s'}[r + \gamma_{a'}^{\max} Q(s_{+'}, s_{-'}, a'; \theta^{-}) \mid s_+, s_-, a] \tag{9}$$

The specific training process diagram of DQN is shown in Fig. 2.

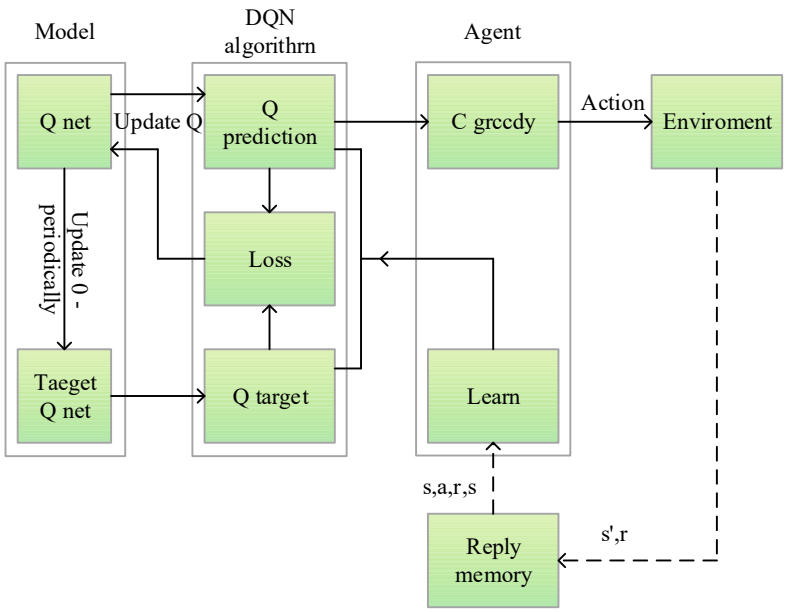

Fig. 2. Specific training process of DQN

2.4. User Project Interaction Layer

Deep reinforcement learning recommender systems are designed to provide action suggestions based on the current state. In previous research, such as the work presented in, it has been demonstrated that using a model of the interaction between the user and the English classroom can lead to an improvement in the performance of the recommender system. Therefore, it is proposed to construct a network of interactions that explicitly models the interactions between user and English

classroom features. The user attribute information is first mapped into vectors using the embedding matrix of the embedding layer, and the mapped vectors are fed into a fully connected layer whose activation function is ReLU to extract the user attribute feature askameters. Then, these mapped askameters are fed into the fully connected layer with ReLU activation function to extract the user attribute feature vectors in the case of Gender:

$$u_{ge} = ReLU(W_i^T u_i + b), u_i = E_i u_{ge} \tag{10}$$

where, W_i^T denotes the weight matrix, T denotes the transpose, b denotes the bias term, u_i denotes the embedding vector, E_i denotes the embedding matrix, and u_{ge} denotes the high-dimensional sparse matrix of the user's gender. Then the vector representations of other attributes of the user are obtained in the same way, and finally the embedding vectors of the user are spliced to obtain the embedding vectors of the user:

$$u_i = concat(u_{ge} + u_b + u_p) \tag{11}$$

The recommended and skipped courses are treated as negative feedback, and the multimodal feature vectors of the courses are extracted from n user interaction history and then transformed through a maximum pooling process to aggregate the most informative features. The resulting vectors are subsequently modeled with user embedding vectors to capture user interactions with the course. Finally, the user comment vectors on the course are extracted using bert to obtain c_i. The user comment vectors, user embedding vectors, interaction vectors, and the maximum pooling results of the course are stitched together to obtain state s_{i+}, which likewise yields the negative feedback state s_{i-} as specified below:

$$m_{\max} = \{\max(w_i m_i) \mid i = 1, \cdots, n\} \tag{12}$$

$$N_{uv} = u_i \otimes m_{max} \tag{13}$$

$$s_{t+} = concat(c_i + m_{max} + N_{uv} + u_i) \tag{14}$$

$$s_{t-} = concat(m_{max} + N_{uv} + u_i) \tag{15}$$

Here, n denotes the number of courses with the highest user rating in the interaction history, w_i is the maximum pooled weight, $\otimes$ is the element-by-element product operation, and N_{uv} denotes the user-course interaction vector.

2.5. User interaction simulator design

The construction of the task environment is a key factor that significantly affects the efficiency of deep reinforcement learning model training. The development of a well-designed task environment is a key prerequisite in the quest to achieve optimal performance. The aim of this work is to build a task environment containing a user interaction simulator that is responsible for simulating user feedback during model training. Such a simulator ensures a comprehensive learning experience that closely mimics real-world user interactions, thus improving the effectiveness of model training.

Reinforcement learning algorithms are fundamentally concerned with capturing and learning from user preferences and interests by adapting to their feedback behavior. In this regard, the inherent characteristics of the reinforcement learning framework make it necessary to simulate the user's feedback behavior in response to recommended actions. To achieve this goal, a user interaction simulator must be designed based on the best features provided by the current behavioral state S_t. This simulator will then produce a simulated feedback score r that approximates the user's real-time response to the recommended action.

Traditional recommendation algorithms based on collaborative filtering are based on the assumption that users with similar interests tend to make similar choices when they receive recommendations for similar products. Inspired by this notion, a model using historical state-action pairs to identify user behaviors and similarities is proposed. The user feedback is randomly simulated based on the calculated similarity between the current state-action pairs and the history, thus enabling the simulation of the interaction between the user and the recommender system. According to this approach, a user simulation memory M is initialized to store the history records (s,a,r), and the similarity between the behavior vector U_t, the recommended actions a_t, and the history records in the memory pool is computed $Cos(p_t,m_i)$. Negative feedback is inferred when the similarity is close to zero, as shown in the following equation:

$$Cos(p_t,m_i)=\alpha\frac{u^t u_i^T}{\|u^t\|}+(1-\alpha)\frac{a^t a_i^T}{\|a^t\|} \tag{16}$$

where α the loss rate of the current behavioral state, p_t is the similarity with the m_i record in the memory pool, and $u_i a_i$ is the historical behavior vector and the recommended behavior vector in the m_i record in the memory pool. The similarity between the current state and the value state r_i value of the m_i record in the memory pool is calculated based on $Cosine(p_t,m_i)$ and normalized as shown in (17):

$$P(p_t\rightarrow r_i)=\frac{Cos(p_t,m_i)}{\sum m_j\in Cos(p_t,m_i)} \tag{17}$$

M represents all the records in the memory pool for the records in m_j and r_i is the ith value. Get the total value r_t value in the current state as shown in equation (18):

$$r_t=\sum m_j\in M\ P(p_t\rightarrow r_i)*r_i \tag{18}$$

3. Analysis of recognition results based on multimodal speech interaction

3.1. Recognition Performance Analysis

3.1.1. Error Recognition Comparison. Using Sennheiser PC 3 CHAT headset with mono microphone with noise-cleaning technology and a separate external sound card, training recordings of spoken English pronunciation were made in an intelligent English interactive classroom with different algorithms, where the trainers were professional teachers of spoken English. The training recordings must be verified by visual waveform playback to ensure the quality. 100 trainers (50 men and 50 women) were collected. Individual data sampling points were digitized into 32 bits and the acquisition frequency was 16 KHz.

Experiments were conducted to test the ability of the training recordings recognition of the model of this paper to correctly recognize pronunciations as mispronunciations. The experiment uses 28 American consonants, 20 vowels, and 2500 words. The spoken pronunciation recordings of 100 trainers were input into the model of this paper, the DTW model and the B/S structure for comparison and analysis, and the comparison is shown in Figure 3.

According to Fig. 3, the number of errors identified by the other two models is higher than that of this paper's model, no matter it is the vowel error recognition or consonant error recognition. As this paper's Deep Reinforcement Learning model adopts the multimodal speech recognition technology to extract the features of the English spoken pronunciation, standardize and fill in the results, and then do the error extinction recognition calculation according to the attribute planning, which

effectively reduces the number of mispronunciations as the correct pronunciation. The number of errors in recognizing vowels and consonants is reduced to about 125.

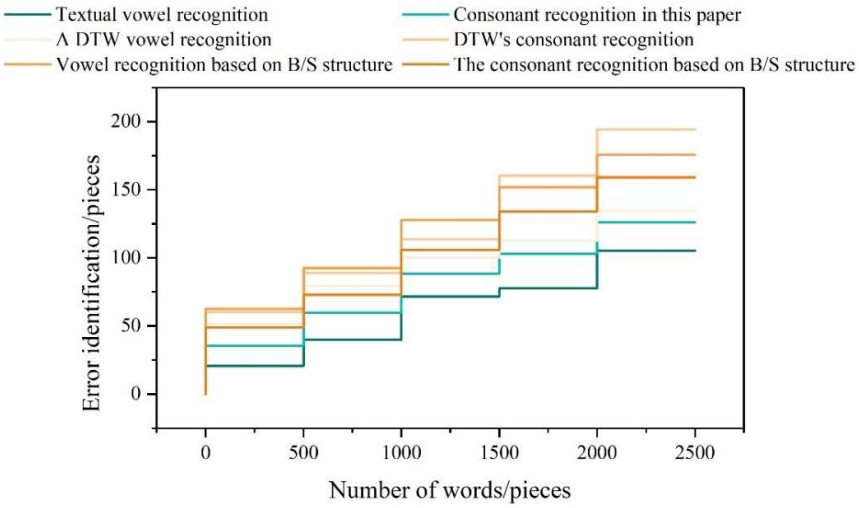

Fig. 3. Error identification contrast

3.1.2. Number of effective corrections. In order to further verify the validity of the correction results of this paper's model, the energy consumption of spoken pronunciation correction is used to confirm that this paper's model is better than other models. The smaller the energy consumption value, the better the correction performance of the system. 500 mispronounced words are selected for the experimental data and input into other models and this paper's model respectively, and Fig. 4 shows the comparison of the number of effective corrections.

It can be seen that the energy consumption curve of this model is lower than that of other models, and the correction process in this paper uses a simpler program to reduce the energy consumption to a greater extent, with the highest consumption of only 67.8733J, which is much lower than that of other models. For this reason, the model in this paper uses the least amount of energy to correct spoken pronunciation, which confirms that the model in this paper has better performance in correcting spoken pronunciation.

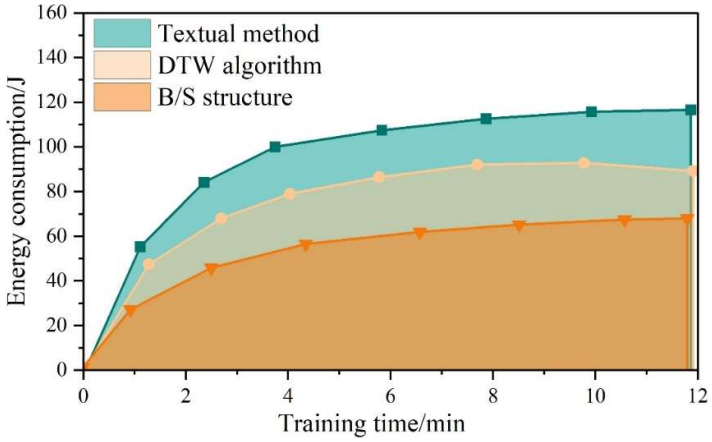

Fig. 4. The effective correction is compared

3.2. Classroom Speech Recognition

Using multimodal speech recognition technology, a speech is recorded for the interaction in the English classroom. Fig. 5 shows the time waveform with respect to the number of sampling points

and normalized sound pressure values. As can be seen from the figure, the maximum number of sampling points is 3×10^5. The normalized sound pressure values range around [-1.5,1.5].

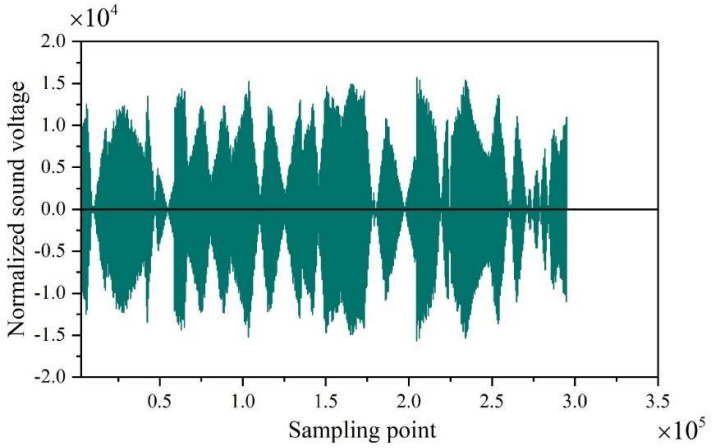

Fig. 5. Time waveform

In order to clearly see the fundamental tone period, you can narrow down the number of frames collected, replace the frame-length with 200, then you can get the fundamental tone period as shown in Figure 6. As shown in Fig. 6, the fundamental tone period can be clearly seen, and the fundamental tone period is about 40 frames. From the simulation of the value of the fundamental tone cycle can be seen on the simulation of the fundamental tone cycle graph, can clearly observe the fundamental tone cycle of the speech signal, for students, the waveform is very intuitive, so that students clearly recognize that the speech signal is divided into clear and turbid, turbid has the characteristics of the cycle, in the speech signal processing can be used to learn.

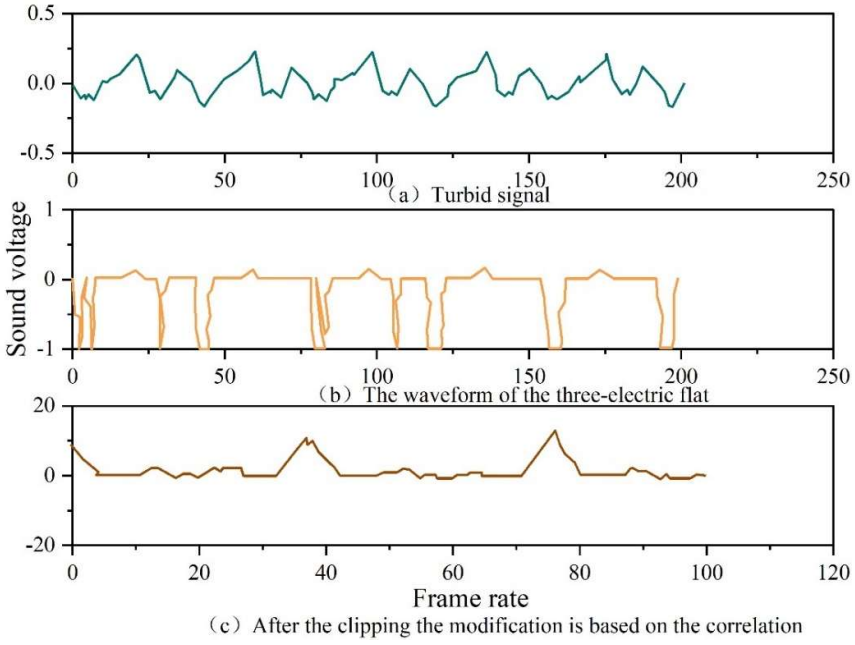

Fig. 6. 200 frame pitch period

3.3. Speech Recognition Sentiment Information Analysis

Fig. 7 shows the LSTM-based speech sentiment analysis, according to which the recognition rate of the happiness emotion decreases after the introduction of the attention mechanism of LSTM because

the error rate of the estimated probability of the LSTM and the attention mechanism accumulates, which reduces the recognition rate of the happiness emotion. The recognition rate of fear and disgust emotions increased significantly after the merger, and the average recognition rate of the six emotions increased to 67.86%.

A comprehensive analysis of the results of the verification of speech recognition text information and emotion information recognition of 100 students participating in the experiment revealed that a small number of students' spoken linguistic information was well combined with emotions, and the use of emotions was better to give more content to the recapitulation of the short text; for the majority of the students, the pursuit of accurate and fluent expression was the goal. The results of the above analysis will be informed in detail to the students who participated in the experiment, and collectively listen to the voice of the students who have a good combination of linguistic information and emotion, and on this basis, point out the shortcomings and the direction of improvement, the experimental analysis of the teaching has also been praised by the students.

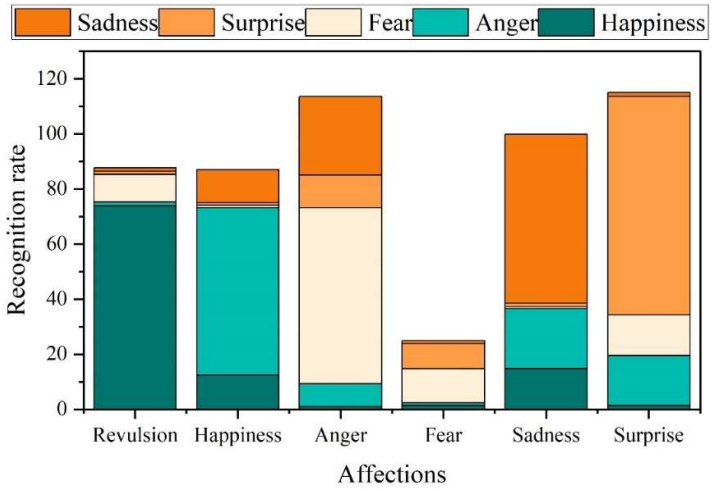

Fig. 7. LSTM based voice analysis

4. The Impact of Intelligent English Classroom Interaction on English Language Learning

4.1. Research process

The experimental period of this study was from September 2022 to December 2022, totaling 16 weeks. The whole research process was divided into four phases, namely, the experimental preparation phase, the pre-test phase, the experimental phase and the post-test phase.

After ensuring the feasibility of the formation of the voice visualization interactive classroom, the formation of the voice visualization classroom was started with the multimedia classroom management teacher and the information technology teacher. Firstly, the students' and teachers' computers were installed with the model constructed in this paper. Second, connect the headset to the computer and debugging, and finally open the Praat software to test whether the recording effect is normal and whether the electronic classroom system can be used normally. At this point, the voice visualization classroom is basically completed.

4.1.1. Pre-testing phase. Firstly, in the first week of the school year, paper questionnaires were distributed to all the students in the experimental and control classes as well as to 18 English teachers in the English group of the school to conduct a questionnaire survey. Its purpose was to find out the current status of teaching English phonological visualization in junior high school.

For this pre-test questionnaire survey, a total of 100 questionnaires were distributed to the students in the experimental class (a total of 50 students, 25 of each gender) and the control class (a total of 50 students, 25 of each gender), and a total of 100 questionnaires were finally collected. A total of 18 questionnaires were distributed to the teachers of the English group in the school and a total of 18 were eventually recovered. The recovery rate of both teacher and student questionnaires was 100%.

During the pre-testing stage, an English phonological test was administered to both the experimental and control classes. At the end of each student's test, the students' scores were recorded to prepare for the data analysis at the end of the study.

In order to understand the gap between the students' pronunciation and the standard pronunciation, as well as to verify that there was no significant difference between the experimental class and the control class in terms of the overall pronunciation level. Recordings of the pronunciation of the unit tones were organized for all students in the experimental and control classes during the pre-test phase, followed by extracting the first and second resonance peak frequencies of the unit tones through the use of Praat software and calculating their average values. In the pre-test stage of the experiment, the values of the experimental and control classes were mainly compared with each other to prove that there was no significant difference between the English speech pronunciation of the students in the two classes.

4.1.2. Post-testing phase. After a semester of teaching experiment, an English phonics test was conducted again for the experimental and control classes. This time, the text in Unit 4, Section 3 of a textbook was chosen for the speech test. The testing tool is the multimodal speech recognition technology constructed in this paper, and the testing method is exactly the same as the pre-test. At the end of the test, the post-test scores were compared and analyzed with the scores of the pre-test stage, and its purpose was to verify whether the use of Praat software for English speech teaching was effective.

4.2. Pronunciation skills

The results of the comparative analysis of the post-test data of the male experimental class are shown in Table 1. As can be seen from Table 1, the mean values of the first and second resonance peak frequencies of the 11 single vowels of the 25 male students in the control class at the end of the experiment compared with the resonance peak frequencies of the accepted British male pronunciation, the p-values of which were greater than 0.05 only for the second resonance peak F2 of the vowel [i:], the first resonance peak F1 of the vowel [3:] and the first and second resonance peaks F1, F2 of the vowel [u], were respectively 0.826, 0.1259, 0.856, and 0.759. The differences were not statistically significant, and their articulations were slightly improved from the preexperiment. Other than that, the mean values of the first and second resonance peak frequencies of the other vowels, compared with the resonance peak frequencies of the recognized pronunciations of British males, have p-values less than 0.05, which are significantly different, and their pronunciations are significantly improved compared to the preexperimental ones.

Table 1. Analysis of data comparison of unit sound resonance peak(male)

Vowel		Formant	Male RP value	Post-test value	Standard deviation	T	P
Pre-vowel	[i:]	F1	273	293.16	12.59	5.499	0.000
		F2	2220	2222.36	12.23	0.026	**0.826**
	[i]	F1	375	364.12	12.15	-6.48	0.000
		F2	1954	2015.96	10.95	18.915	0.000
	[e]	F1	558	594.15	10.75	14.269	0.000
		F2	1786	1769.36	15.26	-8.161	0.000
	[æ]	F1	730	743.96	11.68	3.915	0.0015
		F2	1525	1579.45	11.03	20.439	0.000
Mid vowel	[3:]	F1	510	516.36	11.43	1.625	**0.1259**
		F2	1370	1367.44	13.63	-2.455	0.023
	[ʌ]	F1	685	734.95	11.19	20.521	0.000
		F2	1223	1324.15	10.26	34.068	0.000
Back vowel	[ɔ:]	F1	450	519.26	17.69	29.162	0.000
		F2	641	705.62	15.48	20.826	0.000
	[ɔ]	F1	591	618.36	10.15	8.66	0.000
		F2	861	923.41	11.36	23.618	0.000
	[u:]	F1	300	314.15	11.57	4.726	0.000
		F2	1130	1095.15	10.59	10.315	0.000
	[u]	F1	412	415.62	10.26	0.149	**0.856**
		F2	1050	1049.15	12.85	-0.428	**0.759**
	[a:]	F1	685	729.64	14.69	13.348	0.000
		F2	1073	1145.59	9.82	23.948	0.000

4.3. Ability to read aloud

4.3.1. Pre-test differences in reading aloud ability between experimental and control classes. Table 2 shows the comparison of the pre-test differences between the dimensions of reading aloud ability between the experimental and control classes, with significant values p=0.6485>0.05 corresponding to the accuracy dimension, p=0.5615>0.05 corresponding to the fluency dimension, p=0.4658>0.05 corresponding to the emotional expression dimension before the experiment. Therefore, there was no significant difference in the dimensions of reading aloud ability between the students of the experimental and control classes before the beginning of the experiment. Taken together, the overall reading aloud ability and the variability of the dimensions of reading aloud between the experimental and control classes were not significant, indicating that there was no significant difference in reading aloud ability between the two samples at the time of the pre-test.

Table 2. Presurvey difference

Analysis item	Term	Sample size	Mean value	Standard deviation	T	Df	P
Accuracy	Experimental class	50	9.9155	2.1345	-0.4254	89	0.6485
	Comparison class	50	10.0544	1.7295			
Fluency	Experimental class	50	5.9154	1.4585	-0.519	89	0.5615
	Comparison class	50	6.1245	1.1695			
Emotional expression	Experimental class	50	5.2984	1.2154	-0.648	89	0.4658
	Comparison class	50	5.4265	1.0658			

4.3.2. Comparison of differences in posttests of reading aloud ability between experimental and control classes:

1) Comparison of the differences between the experimental class and the control class in overall reading aloud ability posttests. Table 3 shows the comparison of the differences in the post-test of reading aloud ability between the experimental and control classes. The average score of the experimental class was 23.945, while the average score of the control class was 21.464, and the average score of reading aloud ability of the students in the experimental class was higher than the average score of reading aloud ability of the students in the control class. P=0.005<0.05 for the post-test of reading aloud ability corresponding to the experimental and control classes.Therefore, there is a significant difference between the overall reading aloud ability of the experimental and control classes at the end of the experiment.

2) Comparison of the differences between the experimental and control classes in each dimension of reading aloud ability. The p-value corresponding to the dimension of reading accuracy = 0.005<0.05, the p-value corresponding to the dimension of reading fluency = 0.0024<0.05, and the p-value corresponding to the dimension of expression of emotion in reading aloud = 0.000<0.05 between the experimental and control classes; therefore, there is a significant difference in the dimensions of reading aloud ability between the experimental and control classes at the end of the experiment.

Taken together, the differences between the overall reading aloud ability and the dimensions of reading aloud of the students in the experimental and control classes are significant, therefore, it is preliminarily determined that the intelligent English interactive classroom designed by the study based on the constructed in this paper has a facilitating effect in the process of teaching reading aloud in terms of students' reading aloud ability.

Table 3. Post-test difference comparison

Analysis item	Term	Mean value	Standard deviation	T	Df	P
Reading ability	Experimental class	23.945	2.569	2.746	89	0.005**
	Comparison class	21.464	3.458			
Accuracy	Experimental class	11.168	1.365	2.915	89	0.005**
	Comparison class	10.126	1.618			
Fluency	Experimental class	6.458	1.136	0.915	89	0.0024**
	Comparison class	6.158	1.069			
Emotional expression	Experimental class	6.364	0.685	3.769	89	0.000***
	Comparison class	5.536	0.9948			

* p<0.05 ** p<0.01.

5. Conclusion

In this paper, we use multimodal speech recognition, speech emotion and speech interaction behavior analysis techniques respectively, construct deep reinforcement learning models, design user interaction simulators, and introduce user item interaction layers. Based on the multimodal speech interaction recognition technology, the students' English oral training recordings are recognized.

The experiment uses 28 American consonants, 20 vowels and 2500 words. The spoken pronunciation recordings of 100 trainers are input into the model of this paper, the DTW model and the B/S structure for comparative analysis. Due to the deep reinforcement learning model constructed in this paper using the multimodal speech recognition technology, the number of mispronunciations treated as correct pronunciation is effectively reduced, and the recognition errors

of vowels and consonants are reduced to about 125. On the basis of the LSTM algorithm, the attention mechanism is introduced and the correct rate of speech emotion recognition is analyzed, and the experimental results show that the average recognition rate of the six emotions rises to 67.86%.

The pre- and post-tests and controlled experiments are integrated to assess the pronunciation and reading ability of the intelligent English classroom for students. Comparing the resonance peak frequencies of the 11 vowels, the p-values of the subject students were greater than 0.05 only for the second resonance peak F2 of the vowel [i:], the first resonance peak F1 of the vowel [3:], and the first and second resonance peaks F1 and F2 of the vowel [u], with p-values of 0.826, 0.1259, 0.856, and 0.759, respectively, and p-values of the first and second resonance peaks of the rest of the vowels were all less than 0.05, there is a significant difference, and their pronunciation is significantly improved compared with the preexperiment. In the post-test experiment, the average scores of the overall reading aloud ability of the experimental class and the control class were 23.945 and 21.464, respectively, and $p=0.005<0.05$. The intelligent English interactive classroom constructed in this paper has a facilitating effect in the process of reading aloud teaching in terms of students' reading aloud ability.

Funding

This research was supported by the teaching reform and innovation project of Shanxi Province in 2024: "Research on English Teaching Innovation in Universities driven by Information Technology" (J20241671).

References

[1] Yu, H., Shi, G., Li, J., & Yang, J. (2022). Analyzing the differences of interaction and engagement in a smart classroom and a traditional classroom. *Sustainability*, 14(13), 8184.

[2] Cheng, X., & Fan, Y. (2022). Research and design of intelligent speech equipment in smart english language lab based on Internet of Things technology. *Procedia Computer Science*, 198, 505-511.

[3] Alam, A. (2022). Employing adaptive learning and intelligent tutoring robots for virtual classrooms and smart campuses: reforming education in the age of artificial intelligence. *In Advanced computing and intelligent technologies: Proceedings of ICACIT 2022* (pp. 395-406). Singapore: Springer Nature Singapore.

[4] Zhang, D. (2024, April). Using deep Reinforcement Learning to Optimize the Motivational Incentive Mechanism of Online English Learners. *In Proceedings of the International Conference on Decision Science & Management* (pp. 179-183).

[5] Xu, J. (2023). Optimizing English Education in the Information Era: A Multimodal Approach Based on BOPPPS Teaching Model. *Journal of Combinatorial Mathematics and Combinatorial Computing*, 118, 33-48.

[6] Han, C. (2024). English information teaching resource sharing based on deep reinforcement learning. *International Journal of Continuing Engineering Education and Life Long Learning*, 34(1), 53-65.

[7] Xu, J., Liu, Y., Liu, J., & Qu, Z. (2022). Effectiveness of English online learning based on deep learning. *Computational Intelligence and Neuroscience*, 2022(1), 1310194.

[8] Guo, H., & Jiang, X. (2024). English teaching evaluation based on reinforcement learning in content centric data center network. *Wireless Networks*, 30(5), 4145-4155.

[9] Susanto, B., Wajdi, M., Sariono, A., & Sudarmaningtyas, A. E. R. (2022). Observing English classroom in the digital era. *Journal of Language and Pragmatics Studies*, 1(1), 6-15.

[10] Liu, H. (2021, December). Analysis of E-learning English Teaching Path Based on Reinforcement Learning. *In International conference on Smart Technologies and Systems for Internet of Things* (pp. 688-696). Singapore: Springer Nature Singapore.

[11] Hu, J., & Jin, G. (2024). An Intelligent Framework for English Teaching through Deep Learning and Reinforcement Learning with Interactive Mobile Technology. *International Journal of Interactive Mobile Technologies*, 18(9).

[12] Ma, X. (2023). The Application Research of Distributed Interface Coordination Based on Deep Reinforcement Learning in English MOOC Platform. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[13] Nai, R. (2022). The design of smart classroom for modern college English teaching under Internet of Things. *Plos one*, 17(2), e0264176.

[14] Niu, J., & Liu, Y. (2022). The Construction of English Smart Classroom Teaching Mode Based on Deep Learning. *Computational Intelligence and Neuroscience*, 2022(1), 9037010.

[15] Dong, Y., Yu, X., Alharbi, A., & Ahmad, S. (2022). AI-based production and application of English multimode online reading using multi-criteria decision support system. *Soft Computing*, 26(20), 10927-10937.

[16] Li, B. (2023). Design and research of computer-aided english teaching methods. *International journal of humanoid robotics*, 20(02n03), 2240004.

[17] Zhang, D. (2022). Affective cognition of students' autonomous learning in college English teaching based on deep learning. *Frontiers in psychology*, 12, 808434.

[18] Kamil Skowroński,Adam Gałuszka & Eryka Probierz. (2024). Polish Speech and Text Emotion Recognition in a Multimodal Emotion Analysis System. Applied Sciences(22),10284-10284.

[19] Wang Yufei. (2023). Research on emotion classification technology of movie reviews based on topic attention mechanism and dual channel long short term memory. PeerJ. Computer sciencee1295-e1295.

[20] Xia Lei,Tang Jianfeng,Li Guangli,Fu Jun,Duan Shukai & Wang Lidan. (2024). Time Series Classification Based on Forward Echo State Convolution Network. Neural Processing Letters(4),

[21] Huanhuan Hou & Azlan Ismail. (2024). EETS: An energy-efficient task scheduler in cloud computing based on improved DQN algorithm. Journal of King Saud University - Computer and Information Sciences(8),102177-102177.